# Data Understanding, Data Analysis, and Data Science Course Notes

## Volume 1: Prelude to Data Understanding

Patrick Boily

October 2023

Quadrangle | Idlewyld Analytics and Consulting Services

This one goes out to the "Welsh" contingent: Elowyn, Llewellyn, and Gwynneth. Your world is going to be a whole lot different than mine was; maybe data can even help make some of it better. But one thing's for sure: data is not going away any time soon – better be prepared.

# Series Preface

The *first* thing to know about *Data Understanding, Data Analysis, and Data Science* (DUDADS) is that it isn't really a "book". It makes more sense to think of it as **course notes**, or as a **reference manual** and a source of examples and application.

I borrow some of its contents from authors who do a better job of explaining things than I could hope to do; I also sometimes modify their examples and code to better suit my pedagogical needs.* Major influences include [1, 2, 3, 4, 5, 6, 8] – be sure to give these masterful works the attention they deserve!

The *second* thing to know about DUDADS is that it isn't really "a" book. It makes more sense to think of it as **a bunch of books in a trench coat, masquerading as single one**.† No one is expected to traverse DUDADS in one sitting, or even to tackle more than a few of its assigned chapters, sections, subsections, exercises at any given time; rather, it is intended to be read in parallel with guided lectures.

The *third* thing to know about DUDADS is that the practical examples use `R` and/or `Python`, for no particular reason other than that *some* programming language had to be used to illustrate the concepts. In the text, `R` code appears in blue boxes:

```
... some R code ...
```

Whereas `Python` code appears in green boxes:

```
... some Python code ...
```

You may look at some piece of code and think to yourself: "This isn't how I would do it" or "such-and-such a task would be easier to accomplish if we used module/package ABC or programming language XYZ". That's quite possible.

But finding the optimal tool is not the point of this book. In the first place, new data science tools appear regularly, and it would be a fool's errand to try to continuously modify the book to keep up with them.‡ In the second place, I am serious about the "Understanding" part of *Data Understanding, Data Analysis, and Data Science*, and that is why I favour a **tool-agnostic** approach.

---

* In all cases, I have attempted to properly cite and give credit where it is due. Get in touch if you find omissions!
† I paid heed to this realization by splitting it into a number of volumes.
‡ I am not saying that I won't be adding examples in different languages in the future, but let's not get ahead of ourselves.

The *fourth* thing to know about DUDADS is that it is not a place to go to in order to obtain a detailed step-by-step guide on "how to solve it". In person, my answer to a vast array of data science related questions is, rather anti-climatically: "it depends". Of course, it depends; on the data, on the objectives, on the cost associated with making a mistake, on the stakeholder's appetite for uncertainty, and, perhaps more surprisingly, on the analytical and data preparation choices that are made along the way.

To some, this might smack of post-modernism: "you are saying that there is no truth, and that data analysis is pointless!" To which I respond: "analysts have agency (lots of it, it turns out), and their choices *DO* influence the results, so make sure to run multiple analyses to determine the variability of the outcomes". That is the nature of the discipline.

The *last* thing you should probably know about DUDADS is that I have made a concerted effort to focus mainly on the **story** of (learning) data analysis and data science; sometimes, that comes at the expense of rigorous exposition.

> "The early stages of education have to include a lot of lies-to-children, because early explanations have to be simple. However, we live in a complex world, and lies-to-children must **eventually be replaced** by more complex stories if they are not to become delayed-action genuine lies." [7]

Some of the concepts and notions that I present are **incomplete** by design, but remain (I hope) true-to-their-spirit, or at least true "enough" for a first pass.[§] My position is that learning is an iterative process and that important take-aways from an early stage might need to be modified to account for new developments at a later date. But all things in good time: flexibility is a friend in your learning adventure; perfectionism, not always so.

Patrick Boily
Wakefield, October 2023

The DUDADS reference manuals are available at idlewyldanalytics.com

- Volume 1: ***Prelude to Data Understanding***
- Volume 2: ***Fundamentals of Data Insight***
- Volume 3: ***Spotlight on Machine Learning***
- Volume 4: ***Techniques of Data Analysis***
- Volume 5: ***Special Topics in Data Science and Artificial Intelligence***
- ***The Practice of Data Visualization*** (with S. Davies and J. Schellinck)

[§] In the parlance of the field, let me simply say that some of the details are left as an exercise for the reader (and can also be found in the numerous references).

# Preface References


[1] C.C. Aggarwal. *Data Mining: the Textbook* . Cham: Springer, 2015.

[2] C.C. Aggarwal, ed. *Data Classification: Algorithms and Applications* . CRC Press, 2015.

[3] C.C. Aggarwal and C.K. Reddy, eds. *Data Clustering: Algorithms and Applications* . CRC Press, 2014.

[4] D. Dalpiaz. *R for Statistical Learning* . 2020.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* , *2nd ed.* Springer, 2008.

[6] G. James et al. *An Introduction to Statistical Learning: With Applications in R* . Springer, 2014.

[7] I. Stewart, J. Cohen, and T. Pratchett. *The Science Of Discworld*. Ebury Publishing, 2002.

[8] H. Wickham and G. Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* . O'Reilly, Jan. 2017.

# Contributors and Influences

A reference manual of this size could not have been compiled without the help of a multitude of individuals over the years, who provided contributions, influences, and/or inspiration:

**Kevin Cheung** *Optimization*, *Programming Primer*

**Aidan Crowther** *Programming Primer*

**Benoit Dionne** *Basics of Numerical Methods*

**Fabrizio Donzelli** *Overview of Linear Algebra*, *Multivariate Calculus for Data Analysis*

**Patrick Farrell** *Survey Sampling Methods*

**Ehssan Ghashim** *Programming Primer*

**Diane Guignard** *Basics of Numerical Methods*

**Shintaro Hagiwara** *Introductory Statistical Analysis*

**David Haziza** *Design of Experiments*

**Rafal Kulik** *Probability and Applications*, *Introductory Statistical Analysis*, *Classical Regression Analysis*, *Time Series and Forecasting*

**Gilles Lamothe** *Classical Regression Analysis*

**Dong Elle Liu** *Time Series and Forecasting*

**Landon Liu** *Time Series and Forecasting*

**Chunyun Ma** *Programming Primer*

**Jen Schellinck** *Programming Primer*, *Simulations and Modeling*

A hearty "thank you" to everyone, and to all others with whom I have crossed paths on this data adventure!

# Learning Paths

I mostly use the material found in this volume at different levels in my teaching at the University of Ottawa in the Department of Mathematics and Statistics.

In particular, here is what I cover in various courses:

- **MAT 2377** (*Probability and Statistics for Engineers*) – Chapters 6–7, Section 8.2;
- **MAT 3375** (*Regression Analysis*) – Chapter 8 (and some material from Chapter 20, in DUDADS Volume 3);
- **MAT 3377** (*Sampling and Surveys*)) – Chapter 10;
- **MAT 3378** (*Analysis of Experimental Design*)) – Chapter 11;
- **MAT 3379** (*Introduction to Time Series Analysis*) – Chapter 9.

In 3rd-year courses (and above), I further assume that students are familiar with the contents of Chapters 1–7.

# Contents


**1 Programming Primer** **1**

1.1 Programming Fundamentals . . . 1
1.1.1 Compiled vs. Interpreted . . . 1
1.1.2 Some Fundamental Concepts . . . 2
1.1.3 Code Components . . . 4
1.1.4 Designing With Pseudo-Code . . . 7
1.1.5 From Pseudo-Code to Code . . . 9
1.1.6 Debugging . . . 10
1.1.7 `R/Python` . . . 10
1.2 Introduction to `R` . . . 12
1.2.1 Why Use `R` . . . 12
1.2.2 Installing `R` / RStudio . . . 12
1.2.3 Test, Test, Test! . . . 13
1.2.4 Customizing RStudio . . . 14
1.2.5 Upgrading `R` / RStudio . . . 15
1.2.6 Basics of `R` . . . 15
1.3 More Programming in `R` . . . 27
1.3.1 Help and Documentation . . . 27
1.3.2 Simple Data Manipulation . . . 29
1.3.3 Exploring Data . . . 34
1.3.4 A Word About NAs . . . 39
1.3.5 Loops and Conditions . . . 40
1.4 The `tidyverse` . . . 40
1.4.1 Pipeline Operator . . . 41
1.4.2 Tidy Data . . . 42
1.4.3 The `dplyr` Package . . . 44
1.5 Basics of `Python` . . . 47
1.5.1 IDE for `Python` . . . 48
1.5.2 Introduction to `Python` . . . 48
1.5.3 `NumPy` and Arrays . . . 67
1.6 `Python` for Data Science . . . 72
1.6.1 `Pandas` and Data Frames . . . 72
1.6.2 Data Wrangling . . . 78
1.6.3 Data Aggregation . . . 83
1.6.4 Combining `Python` with `R` . . . 85
1.7 Getting Started with SQL . . . 86
1.7.1 Basics . . . 86
1.7.2 SQL Syntax . . . 87
1.7.3 Key Query Operators . . . 88
1.7.4 Examples . . . 96
1.8 Exercises . . . 98
Chapter References . . . 106

**2 Multivariate Calculus for Data Analysis** **107**
2.1 Points, Vectors, Coordinates . . . 107
2.1.1 One Dimension . . . 108
2.1.2 Two and Three Dimensions . . . 108
2.1.3 More Dimensions . . . 108
2.2 Functions . . . 109
2.3 Graphical Representation . . . 111
2.3.1 One Variable . . . 111
2.3.2 Two Variables . . . 111
2.3.3 Three or More Variables . . . 114
2.3.4 Scalars and Vector Fields . . . 115
2.4 Derivatives . . . 116
2.4.1 Difference Quotients . . . 116
2.4.2 Rules of Differentiation . . . 117
2.4.3 Partial Derivatives . . . 118
2.4.4 Gradients . . . 121
2.4.5 Directional Derivatives . . . 122
2.5 Optimization . . . 125
2.5.1 Critical Points . . . 125
2.5.2 Local vs. Global . . . 127
2.5.3 Local Extrema . . . 127
2.5.4 Global Extrema . . . 130
2.5.5 Lagrange Multipliers . . . 132
2.6 Riemann Integrals . . . 135
2.6.1 Local Densities, Total Sums . . . 136
2.6.2 One Variable . . . 137
2.6.3 Fundamental Theorem . . . 137
2.6.4 Finding Antiderivatives . . . 138
2.6.5 Several Variables . . . 139
2.6.6 Applications to Statistics . . . 140
2.7 Exercises . . . 143
Chapter References . . . 146

**3 Overview of Linear Algebra** **147**
3.1 Vector Spaces . . . 147
3.1.1 Practical Definition . . . 147
3.1.2 Linear Combinations . . . 149
3.1.3 Bases and Dimension . . . 150
3.1.4 Vector Subspaces . . . 151
3.1.5 Spanning Sets . . . 153
3.1.6 Dot Product . . . 153
3.1.7 Cross Product in $\mathbb{R}^3$ . . . 154
3.2 Linear Transformations . . . 155
3.3 Matrix Algebra . . . 157
3.3.1 Matrix Operations . . . 157
3.3.2 Square Matrices . . . 159
3.3.3 Determinants . . . 160
3.4 Linear Systems . . . 162
3.4.1 Gauss-Jordan Elimination . . . 165
3.4.2 Linear Systems & Matrices . . . 167
3.5 Matrix Diagonalization . . . 168

3.5.1 Eigenvalues & Eigenvectors . . . 168
3.5.2 Similar Matrices . . . 174
3.5.3 Diagonalization . . . 175
3.6 Exercises . . . 177
Chapter References . . . 180

**4 Basics of Numerical Methods** **181**
4.1 Basic Concepts . . . 181
4.1.1 Round-Off Error . . . 182
4.2 Equations With 1 Variable . . . 185
4.2.1 Bisection Method . . . 185
4.2.2 Golden Ratio Method . . . 191
4.2.3 Fixed Point Method . . . 193
4.2.4 Newton's Method . . . 203
4.2.5 Secant Method . . . 207
4.3 Systems of Equations . . . 208
4.3.1 Linear Systems . . . 208
4.3.2 Non-Linear Systems . . . 221
4.4 Exercises . . . 223
Chapter References . . . 226

**5 A Survey of Optimization** **227**
5.1 Beginnings . . . 227
5.2 Single-Objective Problems . . . 228
5.2.1 Feasible/Optimal Solutions . . . 229
5.2.2 Unsolvable Problems . . . 230
5.2.3 Possible Tasks . . . 230
5.3 Problems Types . . . 231
5.3.1 Classification . . . 231
5.3.2 Algorithms . . . 232
5.4 Linear Programming . . . 233
5.4.1 LP Duality . . . 235
5.4.2 Solving LP Problems . . . 237
5.5 Mixed-Integer LP . . . 238
5.5.1 Cutting Planes . . . 241
5.6 Useful Techniques . . . 241
5.6.1 Activation . . . 242
5.6.2 Disjunction . . . 242
5.6.3 Soft Constraints . . . 242
5.7 Software Solvers . . . 243
5.8 Data Envelopment Analysis . . . 244
5.8.1 Challenges and Pitfalls . . . 246
5.8.2 Pros and Cons . . . 247
5.8.3 DEA Solvers . . . 247
5.8.4 Case Study: Schools . . . 248
5.9 Exercises . . . 252
Chapter References . . . 252

**6 Probability and Applications** **253**
6.1 Basic Notions . . . 253
6.1.1 Sample Spaces and Events . . . 253

6.1.2 Counting Techniques . . . 254
6.1.3 Ordered Samples . . . 255
6.1.4 Unordered Samples . . . 257
6.1.5 Probability of an Event . . . 257
6.1.6 Conditionality Probability . . . 260
6.1.7 Bayes' Theorem . . . 266
6.2 Discrete Distributions . . . 272
6.2.1 Random Variables . . . 272
6.2.2 Expectation . . . 275
6.2.3 Binomial R.V. . . . 277
6.2.4 Geometric R.V. . . . 282
6.2.5 Negative Binomial R.V. . . . 282
6.2.6 Poisson R.V. . . . 283
6.2.7 Other Discrete R.V. . . . 288
6.3 Continuous Distributions . . . 288
6.3.1 Continuous R.V. . . . 288
6.3.2 Expectation . . . 294
6.3.3 Normal R.V. . . . 296
6.3.4 Exponential R.V. . . . 301
6.3.5 Gamma R.V. . . . 304
6.3.6 Binomial Approximations . . . 305
6.3.7 Other Continuous R.V. . . . 307
6.4 Joint Distributions . . . 307
6.5 CLT/Sampling Distributions . . . 313
6.5.1 Sampling Distributions . . . 313
6.5.2 Central Limit Theorem . . . 316
6.5.3 Sampling Distributions II . . . 323
6.6 Exercises . . . 327
Chapter References . . . 336

**7 Introductory Statistical Analysis** **337**
7.1 Introduction . . . 337
7.2 Descriptive Statistics . . . 337
7.2.1 Data Descriptions . . . 338
7.2.2 Outliers . . . 343
7.2.3 Visual Summaries . . . 343
7.2.4 Coefficient of Correlation . . . 346
7.3 Estimation . . . 349
7.3.1 Standard Error . . . 349
7.3.2 C.I. for $\mu$ With $\sigma$ Known . . . 351
7.3.3 Confidence Level . . . 356
7.3.4 Sample Size . . . 358
7.3.5 C.I. for $\mu$ With $\sigma$ Unknown . . . 359
7.3.6 C.I. for a Proportion . . . 362
7.4 Hypothesis Testing . . . 363
7.4.1 Generalities . . . 367
7.4.2 Critical Regions . . . 369
7.4.3 Test for a Mean . . . 372
7.4.4 Test for a Proportion . . . 378
7.4.5 Two-Sample Tests . . . 379
7.4.6 Difference of 2 Proportions . . . 383

7.4.7 Hypothesis Testing with R . . . 384
7.5 Additional Topics . . . 389
7.5.1 Analysis of Variance . . . 389
7.5.2 Analysis of Covariance . . . 394
7.5.3 Multivariate Statistics . . . 397
7.5.4 Goodness-of-Fit Test . . . 401
7.6 Exercises . . . 402
Chapter References . . . 408

**8 Classical Regression Analysis** **409**
8.1 Preliminaries . . . 409
8.1.1 Random Variables . . . 409
8.1.2 Multivariate Calculus . . . 416
8.1.3 Matrix Algebra . . . 417
8.1.4 Quadratic Forms . . . 417
8.1.5 Optimization . . . 419
8.2 Simple Linear Regression . . . 419
8.2.1 Least Squares Estimation . . . 421
8.2.2 Inference . . . 429
8.2.3 Estimation and Prediction . . . 437
8.2.4 Significance of Regression . . . 444
8.2.5 SLR in R . . . 446
8.3 Multiple Linear Regression . . . 447
8.3.1 Least Squares Estimation . . . 448
8.3.2 Inference . . . 451
8.3.3 Power of a Test . . . 460
8.3.4 Determination Coefficients . . . 461
8.3.5 Diagnostics . . . 461
8.4 Extensions of OLS . . . 468
8.4.1 Multicollinearity . . . 468
8.4.2 Polynomial Regression . . . 471
8.4.3 Interaction Effects . . . 474
8.4.4 Categorical Variables . . . 477
8.4.5 Weighted Least Squares . . . 477
8.4.6 Other Extensions . . . 480
8.5 OLS and Outliers . . . 481
8.5.1 Leverage and Extrapolation . . . 481
8.5.2 Deleted Residuals . . . 483
8.5.3 Influential Observations . . . 484
8.5.4 Cook's Distance . . . 485
8.6 Exercises . . . 486
Chapter References . . . 490

**9 Time Series and Forecasting** **491**
9.1 Introduction . . . 491
9.1.1 Simple Examples . . . 492
9.1.2 Pre-Processing . . . 493
9.1.3 Stationary Models . . . 504
9.1.4 Partial Autocorrelation . . . 508
9.2 Estimating Parameters . . . 510
9.2.1 Sample Statistics . . . 510

9.2.2 Examples . . . 511
9.3 ARMA Models . . . 516
9.3.1 Linear Processes . . . 516
9.3.2 ARMA in General . . . 518
9.3.3 Stationarity and Causality . . . 519
9.3.4 Linear Representation . . . 521
9.3.5 ACVF . . . 523
9.3.6 PACF . . . 526
9.4 Forecasting . . . 530
9.4.1 Yule-Walker Procedure . . . 530
9.4.2 Durbin-Levinson Algorithm . . . 533
9.4.3 Forecast Limits . . . 535
9.4.4 Example . . . 535
9.5 ARMA Estimation . . . 543
9.5.1 Mean: I.I.D. Case . . . 543
9.5.2 Mean: Time Series . . . 544
9.5.3 Yule-Walker Estimators . . . 545
9.5.4 Example . . . 548
9.6 Diagnostic Tests . . . 551
9.6.1 Ljung-Box Test . . . 552
9.6.2 Example . . . 553
9.7 MLE Estimation . . . 556
9.7.1 I.I.D. Random Variables . . . 556
9.7.2 Time Series Model . . . 558
9.7.3 Order Selection . . . 560
9.7.4 Examples . . . 560
9.8 Nonlinear Time Series . . . 575
9.8.1 ARCH Model . . . 575
9.8.2 GARCH Model . . . 576
9.8.3 Example . . . 577
9.9 Miscellenea . . . 580
9.9.1 Seasonality . . . 581
9.9.2 Asymptotic Normality . . . 584
9.10 Exercises . . . 590
Chapter References . . . 598

**10 Survey Sampling Methods** **599**
10.1 Background . . . 599
10.1.1 Sampling Generalities . . . 602
10.1.2 Survey Frames . . . 604
10.1.3 Fundamental Concepts . . . 604
10.1.4 Data Collection Basics . . . 607
10.1.5 Sampling Types . . . 607
10.2 Questionnaire Design . . . 610
10.2.1 Basic Concepts . . . 610
10.2.2 Question Types . . . 611
10.2.3 Design Considerations . . . 612
10.2.4 Question Order . . . 613
10.3 Simple Random Sampling . . . 615
10.3.1 Basic Notions . . . 619
10.3.2 Estimators and C.I. . . . 622

10.3.3 Sample Size . . . 635
10.4 Stratified Sampling . . . 637
10.4.1 Estimators and C.I. . . . 644
10.4.2 Sample Size and Allocation . . . 654
10.4.3 Comparison: SRS and STS . . . 661
10.5 Auxiliary Information . . . 663
10.5.1 Ratio Estimation . . . 663
10.5.2 Regression Estimation . . . 674
10.5.3 Difference Estimation . . . 681
10.5.4 Comparisons . . . 684
10.6 Cluster Sampling . . . 688
10.6.1 Estimators and C.I. . . . 688
10.6.2 Sample Size . . . 704
10.6.3 Comparison: SRS and CLS . . . 706
10.7 Special Topics . . . 707
10.7.1 Systematic Sampling . . . 707
10.7.2 Sampling with PPS . . . 713
10.7.3 Multi-Stage Sampling . . . 716
10.7.4 Multi-Phase Sampling . . . 720
10.7.5 Miscellaneous . . . 722
10.8 Exercises . . . 730
Chapter References . . . 732

**11 The Design of Experiments** **733**
11.1 Basic Notions . . . 733
11.1.1 Experiments . . . 734
11.1.2 Useful Distributions . . . 737
11.2 Hypothesis Testing . . . 740
11.2.1 Inference on $\mu$ . . . 740
11.2.2 Inference on $\mu_1 - \mu_2$ . . . 745
11.2.3 Inference on $\sigma^2$ . . . 751
11.2.4 Inference on $\sigma_1^2/\sigma_2^2$ . . . 753
11.3 One-Way Classification . . . 754
11.3.1 Randomized Designs . . . 754
11.3.2 1−Way Model . . . 756
11.3.3 Analysis of Variance . . . 757
11.3.4 Estimation of Parameters . . . 761
11.3.5 Unbalanced Designs . . . 762
11.3.6 Contrasts . . . 763
11.3.7 Multiple Comparisons . . . 765
11.3.8 Model Validation . . . 773
11.3.9 Power and Sample Size . . . 776
11.4 Random Effects . . . 778
11.4.1 Estimation of Parameters . . . 779
11.4.2 Analysis of Variance . . . 780
11.4.3 Inference on $\sigma^2$, $\sigma_T^2$, $\mu$ . . . 782
11.4.4 Power . . . 783
11.5 Randomized Block Designs . . . 784
11.5.1 Analysis of Variance . . . 785
11.5.2 Estimation of Parameters . . . 789
11.5.3 Multiple Comparisons . . . 789

11.5.4 Power and Sample Size . . . 790
11.5.5 Model Validation . . . 790
11.6 Factorial Designs . . . 791
11.6.1 2-Way Factorial Experiments . . . 791
11.6.2 Model Validation . . . 798
11.6.3 Model Without Interaction . . . 799
11.6.4 Multiple Comparisons . . . 800
11.6.5 $n$-Way Factorial Designs . . . 801
11.7 Exercises . . . 801
Chapter References . . . 802

**12 Simulations and Modeling** **803**
12.1 Introduction . . . 803
12.1.1 Static Models . . . 805
12.1.2 Dynamic Models . . . 808
12.1.3 Uses, Data, Contrast . . . 809
12.1.4 Simulation Types . . . 813
12.2 Modeling Strategies . . . 815
12.2.1 Information Gathering . . . 815
12.2.2 Conceptual Model . . . 816
12.2.3 Building the Model . . . 818
12.2.4 Verification and Validation . . . 818
12.2.5 Analysis of Results . . . 818
12.3 Practical Considerations . . . 820
12.3.1 Computational Complexity . . . 820
12.3.2 Applications . . . 820
12.3.3 Software . . . 822
12.4 Case Study: NWMO . . . 822
12.5 Exercise . . . 825
Chapter References . . . 825

# List of Figures

1.1 An example of a computer program . . . 2
1.2 Lexical rules of the programming language C . . . 4
1.3 Computer code elements in action, for the scripting language R . . . 5
1.4 The first stage of pseudo-coding . . . 8
1.5 RStudio interface . . . 14
1.6 Database diagram for the toy example . . . 87

2.1 Intervals on the real line . . . 107
2.2 The real line $\mathbb{R}$ . . . 108
2.3 The real plane $\mathbb{R}^2$ and space $\mathbb{R}^3$ . . . 108
2.4 An illustration of a simple 2D vector field . . . 115
2.5 Difference quotient and slope of the tangent at $x = a$ . . . 116
2.6 Tangent plane . . . 119

2.7 Tangent planes at a cone's vertex . . . 120
2.8 Illustration of Clairaut's theorem . . . 121
2.9 Gradient and level sets . . . 122
2.10 Illustration of gradient descent . . . 124
2.11 Critical points for continuous functions of a single real variable . . . 132
2.12 Monthly profit function for the gadgets and gizmos example . . . 133
2.13 Charts related to the gadgets and gizmos example . . . 133
2.14 Graphical illustration of the Riemann integral . . . 137

4.1 Schematics of scientific computing . . . 182

5.1 Graphical solution for the lemonade and lemon juice optimization problem . . . 234
5.2 Excel's numerical solver for unit $D$ . . . 248
5.3 Results of the re-allocation process in the Barcelona public school dataset . . . 251

6.1 Failure probability for the 2-engine and 3-engine planes . . . 262
6.2 Conceptual model of air traffic control security system . . . 262
6.3 Decomposition of $B$ *via* $A$ . . . 265
6.4 The Monty Hall set-up . . . 268
6.5 P.m.f. and c.m.f. for a discrete r.v. . . . 274
6.6 Uniform distributions . . . 277
6.7 Tabulated c.d.f. values for the binomial distribution with $n = 12$ . . . 279
6.8 P.d.f. and c.d.f. for a continuous r.v. . . . 290
6.9 P.d.f. and c.d.f. for a continuous r.v., with event $A$ . . . 291
6.10 P.d.f. and c.d.f. for a continuous r.v., with $\lambda = 0.2$ . . . 292
6.11 Probability of $X > 10.2$, for a continuous r.v. $X$, with $\lambda = 0.2$. . . . 293
6.12 Probability of $X > 10.2$, for a continuous r.v. $X$, with $\lambda = 2$. . . . 293
6.13 P.d.f. and c.d.f. for the Cauchy distribution, with area under the curve . . . 295
6.14 P.d.f. and c.d.f. for the standard normal distribution . . . 296
6.15 P.d.f. and c.d.f. for the exponential distribution . . . 302
6.16 Conditional and marginal probabilities in the dice example . . . 309
6.17 Support for the joint distribution of $X$ and $Y$ – example . . . 310
6.18 Illustration of the central limit theorem . . . 317
6.19 Chi-squared distribution with 8 degrees of freedom . . . 324
6.20 Student $t-$distribution with $r$ degrees of freedom . . . 325

7.1 Mean, median, and mode in various skewness scenarios . . . 340
7.2 Normal distributions with various means and standard deviations . . . 340
7.3 Boxplot with one (suspected) outlier . . . 343
7.4 Boxplot of positively skewed datasets . . . 344
7.5 Histogram and boxplot of the Sydney accident dataset . . . 345
7.6 Strong links that are not detected by the coefficient of correlation . . . 348
7.7 The 68-96-99.7 rule . . . 351
7.8 Frequentist interpretation of confidence intervals . . . 353
7.9 Quantiles of the standard normal distribution . . . 356
7.10 Two-sided quantiles of the standard normal distribution . . . 357
7.11 Two-sided quantiles of the standard normal distribution, for confidence level 0.05 . . . 357
7.12 Estimation error . . . 358
7.13 Critical value $t(4; 0.05)$ . . . 360
7.14 Binomial distribution for 10 trials, with probability of success $1/2$ . . . 363
7.15 Binomial distribution for 100 trials, with probability of success $1/2$ . . . 364
7.16 Critical test region, left-sided test . . . 373

7.17 Critical test region, right-sided test . . . 373
7.18 Critical test region, two-sided test . . . 373
7.19 Critical test regions for the right-sided test; $n = 10$ observations . . . 380
7.20 Confidence region for a bivariate normal random sample . . . 399

8.1 C.d.f. of Student's distribution . . . 413
8.2 C.d.f. of Fisher's distribution . . . 414
8.3 Response and predictor in the Gapminder dataset . . . 420
8.4 Illustrations of failed SLRM assumptions. . . . 421
8.5 Line of best fit and deviations . . . 421
8.6 Total deviation decomposition . . . 426
8.7 Illustration of the Spearman correlation . . . 427
8.8 Various $R^2$ for nonlinear datasets . . . 429
8.9 Confidence interval for the mean response . . . 437
8.10 Prediction interval for a new response . . . 439
8.11 Joint confidence interval for the mean response . . . 443
8.12 Examples of non-significant regressions . . . 444
8.13 Geometrical interpretation of multiple linear regression . . . 453
8.14 Power function . . . 460
8.15 Illustration of non-linearity using residuals and fitted values . . . 462
8.16 Illustration of the Brown-Forsythe test . . . 464
8.17 White House COVID-19 projections . . . 473
8.18 Model scope in two-dimensional predictor space . . . 482
8.19 $X-$outlier and $Y-$outlier in a dataset . . . 483
8.20 Influential observation in a dataset . . . 484

9.1 Simple time series examples . . . 493
9.2 Time series; CV by year . . . 495
9.4 Diagnostic plots and adjusted plots . . . 496
9.3 Continuous CV; estimation summary . . . 496

10.1 Various populations and samples in the sampling model . . . 603
10.2 Dewey vs Truman . . . 609
10.3 Schematics of various sampling designs . . . 610
10.4 2021 Census – How do I complete the questionnaire? . . . 614
10.5 Schematics of SRS . . . 615
10.6 Health and wealth of nations for the 2011 Gapminder data . . . 616
10.7 Schematics of STS . . . 644
10.8 Gapminder data with line of best fit . . . 674
10.9 Scatterplot of $X$ and $Y$ for difference estimation example . . . 682
10.10 Schematics of CLS . . . 689
10.11 Schematics of SYS . . . 707
10.12 Various populations and systematic samplings . . . 710
10.13 Schematics of SRS2S . . . 717
10.14 Schematics of SRS2P . . . 721

12.1 Example of analogical reasoning . . . 804
12.2 A to-scale architectural model . . . 807
12.3 An example of a data model . . . 807
12.4 Possible solutions of the $n-$body problem . . . 812
12.5 Harvard orrery; flight simulator . . . 814
12.6 A school of fish – the target system for a fish school simulation . . . 816

12.7 Some information about fish perceptual mechanics . . . 816
12.8 Representing the fish school in the simulation . . . 817
12.9 Representing the physical characteristics of individual fish . . . 818
12.10 Fish school simulation pseudo-code . . . 819
12.11 A screen shot of the 3D fish school simulation . . . 819
12.12 A sketch of computational complexity . . . 821

# List of Tables

5.5 Sample from the Barcelona public school dataset . . . 249

10.2 Sampling weights for Canadian provinces . . . 656

11.5 The four possible outcomes for hypothesis testing . . . 741
11.11 ANOVA table for equality of treatment means in one-way classification . . . 760
11.19 ANOVA table for equality of treatment means in a RCBD . . . 787
11.21 Two-way factorial design treatment structure . . . 791
11.25 ANOVA table for two-way factorial design . . . 794
11.29 ANOVA table for two-way factorial design with no interactions . . . 799

# Programming Primer 1

by **Patrick Boily** and **Jen Schellinck**, with contributions from **Kevin Cheung**, **Aidan Crowther**, **Chunyun Ma**, and **Ehssan Ghashim**

---

Programming languages go in and out of style. To be a strong programmer, it is important to understand not just the ins and outs of a particular programming language, but how computer languages and computing infrastructure work more generally.

In this chapter, learners are first introduced to some of the core concepts of computer programming in a language-agnostic way, before being shown the basics of `R` and `Python`, two of the most common programming languages used in modern data analysis.

1.1 Programming Fundamentals . 1
Compiled vs. Interpreted . . . 1
Some Fundamental Concepts . 2
Code Components . . . . . 4
Designing With Pseudo-Code 7
From Pseudo-Code to Code . . 9
Debugging . . . . . 10
`R/Python` . . . . . 10
1.2 Introduction to `R` . . . . . 12
Why Use `R` . . . . . 12
Installing `R` / RStudio . . . . . 12
Test, Test, Test! . . . . . 13
Customizing RStudio . . . . . 14
Upgrading `R` / RStudio . . . . 15
Basics of `R` . . . . . 15
1.3 More Programming in `R` . . . 27
Help and Documentation . . 27
Simple Data Manipulation . 29
Exploring Data . . . . . 34
A Word About NAs . . . . . 39
Loops and Conditions . . . . 40
1.4 The `tidyverse` . . . . . 40
Pipeline Operator . . . . . 41
Tidy Data . . . . . 42
The `dplyr` Package . . . . . 44
1.5 Basics of `Python` . . . . . 47
IDE for `Python` . . . . . 48
Introduction to `Python` . . . . 48
`NumPy` and Arrays . . . . . 67
1.6 `Python` for Data Science . . . 72
`Pandas` and Data Frames . . . 72
Data Wrangling . . . . . 78
Data Aggregation . . . . . 83
Combining `Python` with `R` . . 85
1.7 Getting Started with SQL . . 86
Basics . . . . . 86
SQL Syntax . . . . . 87
Key Query Operators . . . . . 88
Examples . . . . . 96
1.8 Exercises . . . . . 98
Chapter References . . . . . 106

## 1.1 Programming Fundamentals

What are **computer code** and **computer programs**? Is there a difference between these two concepts? (see [4] for a discussion on the topic).

In a nutshell, a **computer program** is an **algorithm**, written in a **computer language**, providing **instructions** to a computer for carrying out a **series of operations**. An example of a computer program is provided in Figure 1.1.

Computer programs can be **compiled** or **interpreted** as a series of hardware operations, carried out by a computer's **electrical components**.

### 1.1.1 Compiled vs. Interpreted Languages

Compilers **translate** full source code programs, written in **high-level language** (i.e., using natural languages, only "understandable" by people, as in Figure 1.1) into **machine language** (i.e., binary code, only "understandable" by computers): they are basically **grammatical** (syntactic) checkers – if the source code is error-free, it is converted into machine code, which is eventually run by an executable file. Compiled code runs quickly, and is thus favoured for **deployment phase**. Commonly-used compiled languages include C/C++/C#, COBOL, Fortran, Pascal, and Julia.

**Interpreters** execute the source code directly: as long as an individual statement is error-free (in the context of the available workspace), it can be executed every time it is called, without regard for the overall syntax of the file. Interpreters are slower, generally, and are favoured during the **development phase**. Commonly-used interpreted languages include R,[1] Python, JavaScript, and Ruby.

1: Most programmers do not consider `R` to be a programming language. If they are feeling generous, they might dub it a scripting language, at best. But it gets the job done for data analysis purposes.

```c
#include <stdio.h>
int main()
{
        double firstNumber, secondNumber, temporaryVariable;

        printf("Enter first number: ");
        scanf("%lf", &firstNumber);

        printf("Enter second number: ");
        scanf("%lf",&secondNumber);

        // Value of firstNumber is assigned to temporaryVariable
        temporaryVariable = firstNumber;

        // Value of secondNumber is assigned to firstNumber
        firstNumber = secondNumber;

        // Value of temporaryVariable (which contains the initial value of firstNumber)
        secondNumber = temporaryVariable;

        printf("\nAfter swapping, firstNumber = %.2lf\n", firstNumber);
        printf("After swapping, secondNumber = %.2lf", secondNumber);

        return 0;
}
```

**Figure 1.1:** An example of a computer program written in the computer language C. What do you suppose this program does? (Programiz .)

### 1.1.2 Some Fundamental Concepts

We have been using the terms "computer language" and "algorithm" as though they were everyday words. Let us take the time to ensure that their meanings are clear.

#### Formal Language

In a **formal language**, *words* are created by combining *letters* from a predefined **alphabet**, according to the rules provided by a *formal grammar*. Everything that is formed according to the rules is an acceptable word; anything else is not.

---

**Example:** Consider the formal language defined with

- **alphabet**: {a, b, C, D, !}
- **grammatical rules**:
  1. letters may only be placed immediately to the left or to the right of another letter
  2. a letter instance must always be accompanied by another instance of the same letter at some location either to its left and/or to its right (or both)
  3. an upper case letter must always be accompanied by a lower case letter immediately to its left or to its right

Thus, `aa` is a word in this formal language (rules 2 and 3 are clearly satisfied; rule 3 is satisfied vacuously), as is `bCaCab`, but `!aC!`, `DDaa`, and `Patrick` are not (why?).

Formal languages can sometimes seem ridiculous – of course letters may only be placed to the left or to the right of other letters... where else would they go? Well, *rule 1* officially (and formally) eliminates letters piling up on top of one another, for starter, but also *spaces* between words (for that language, the space _ is not in the alphabet of letters).

Human languages, of the sort deemed **natural** (in contrast with artificial or constructed languages) are formal, in theory. In practice, their grammars tend to be **flexible** (more so with English than French, say) – syntax evolves with cultures (in time and in space), and semantics (meaning) can be retained even when the grammar is mangled.[2]

2: But only up to a point, of course.

### Computer Language

**Computer languages** are languages constructed to provide instructions **to a computer**, in such a way that they can be compiled into low-level instructions that the computer processor can execute.

Computer languages are also called **programming languages**, for reasons that will soon become obvious. They are **formal** languages because if the grammatical rules are not followed *to the letter*, the program cannot be executed – computers cannot guess or infer what the programmer really meant when the syntax is out of sorts.

The **structure** of the formal definition of a computer language contains the following sections:

1. **Lexical Rules**
2. **Syntax Rules**
   - *Grammar Productions*
   - *Operator Associativities and Precedences*
3. **Typing Rules**
   - *Declarations*
   - *Type Consistency Requirements* (Function Definitions, Expressions, Statements)
4. **Operational Characteristics**
   - *Data* (Scalars, String Constants, Arrays)
   - *Expressions* (Order of Evaluation, Type Conversion, Array Indexing)
   - *Assignment Statements* (Order of Evaluation, Type Conversion)
   - *Functions* (Evaluation of Actuals, Parameter Passing, Return From a Function)
5. **Program Execution**

As an illustration, the **lexical rules** of C are shown in Figure 1.2.

In the lexical and syntax rules given below, BNF notation characters are written in green.

- Alternatives are separated by vertical bars: i.e., '*a* | *b*' stands for "*a* ***or*** *b*".
- Square brackets indicate optionality: '[ *a* ]' stands for an optional *a*, i.e., "*a* | *epsilon*" (here, *epsilon* refers to the empty sequence).
- Curly braces indicate repetition: '{ *a* }' stands for "*epsilon* | *a* | *aa* | *aaa* | ..."

---

## 1. Lexical Rules

| | | |
|---|---|---|
| *letter* | ::= | `a` \| `b` \| `...` \| `z` \| `A` \| `B` \| `...` \| `Z` |
| *digit* | ::= | `0` \| `1` \| `...` \| `9` |
| **id** | ::= | *letter* { *letter* \| *digit* \| _ } |
| **intcon** | ::= | *digit* { *digit* } |
| **charcon** | ::= | '*ch*' \| '**\n**' \| '**\0**', where *ch* denotes any printable ASCII character, as specified by **isprint()**, other than \ (backslash) and ' (single quote). |
| **stringcon** | ::= | "{*ch*}", where *ch* denotes any printable ASCII character (as specified by **isprint()**) other than " (double quotes) and the newline character. |
| *Comments* | | Comments are as in C, i.e. a sequence of characters preceded by **/*** and followed by ***/**, and not containing any occurrence of ***/**. |

**Figure 1.2:** Lexical rules of the programming language C Debray .

### Algorithm

Computer programs are **algorithms**, which is to say, sequences of instructions with (at least) one well-defined **stopping point** (an instruction that tells the program when to stop running).

Algorithms are not always mathematical or computer-based. In some sense, we could think of recipes as algorithms as well: the baking/cooking steps are presented in sequence, and some last step that must be completed before the end product can be eaten.

For instance, here is an algorithm to make **muffins**:[3]

3: Delicious!

1. Pour 1/2 cup of flour into a bowl.
2. Break one egg into the bowl.
3. Pour 3 tablespoons of oil into the bowl.
4. Pour 1 teaspoon of baking powder into the bowl.
5. Pour 1/4 cup of sugar into the bowl.
6. Mix with spoon until smooth.
7. Pour the mixture into muffin tins.
8. Bake for 15 minutes at 350 degrees Fahrenheit.
9. Let cool before eating.
10. Enjoy!

What is the **stopping point**? What is the **outcome**?

### 1.1.3 Code Components

Various sets of instructions, conventions, and structures are so fundamental to computer programming aims that they can be found in nearly all computer languages.

These **fundamental code elements** include:

- Variables
- Data Structures
- Operators
- Statements and Expressions
- Blocks (and Scope)

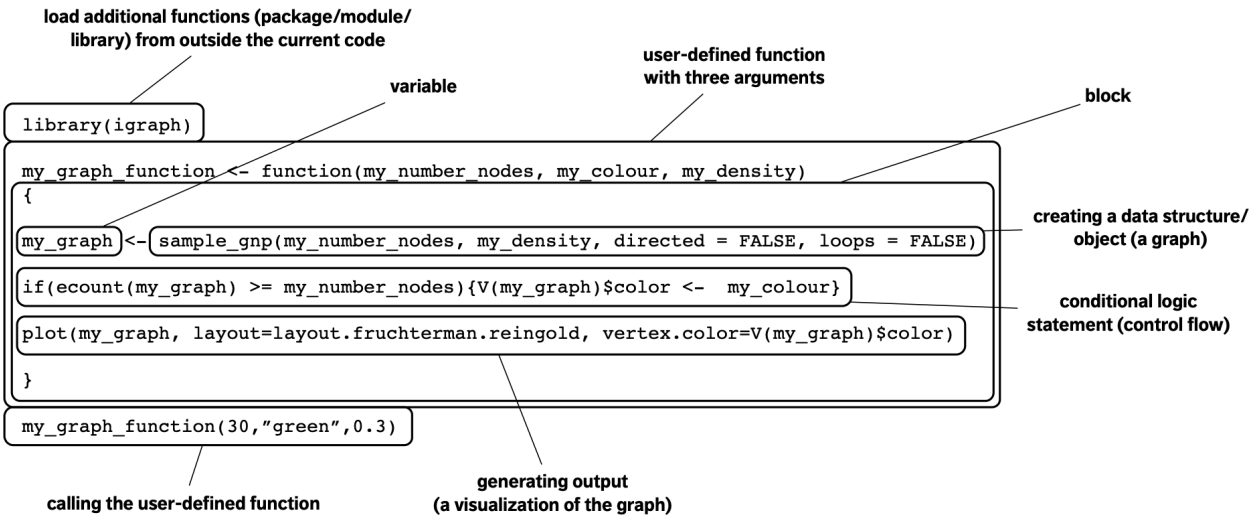



**Figure 1.3:** Computer code elements in action, for the scripting language R.

- Functions
- Logical (Control) Flow
- Libraries/Packages/Modules
- Inputs/Outputs
- Interpreters/Compilers

How these components mesh with one another depends on the syntax of the programming language under consideration (or its dialect).

In Figure 1.3, we see how this could be done in base R, for instance. This particular chunk of code uses the

> `igraph` **library** (specifically, its pre-compiled **functions** `plot()`, `sample_gnp()`, `ecount()`, and `V()`),

and builds the

> user-defined **function** `my_graph_function()` *via* a **code block**,

which takes in as

> **inputs** the **variables** `my_number_nodes`, `my_colour`, and `my_density`.

This function creates a

> graph **data structure** `my_graph`,

and colours the graph's vertices using `my_colour` as long as

> some conditional **logic statement** relating to the number of edges in the graphs and `my_number_nodes` is satisfied.

The function generates a visualization of the graph as an

> **output**,

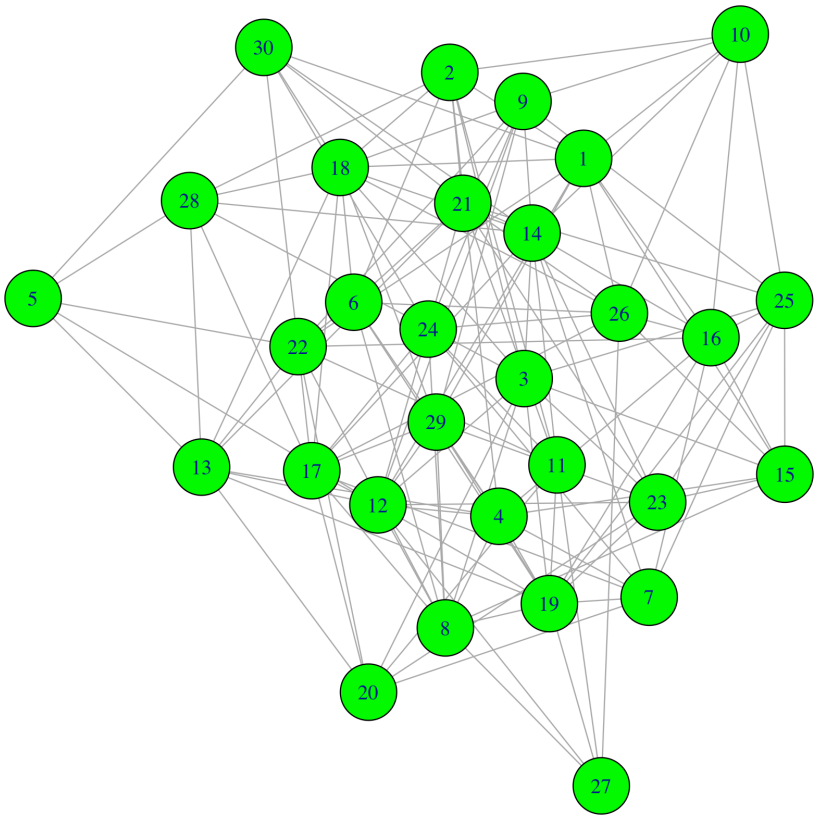
which is displayed when the **function call** is issued.

The code is seen in action below: it creates and displays a 30-node, green-coloured, non-directed, loop-free graph with probability 0.3 of there being an edge between two arbitrary nodes (we will discuss what these concepts represent in Chapter 29, *(Social) Network Data Analysis*).[4]

4: The `seed` enforces replicability.

**Creating a random graph**

```
library(igraph)

my_graph_function <- function(my_number_nodes,
                              my_colour,
                              my_density) {
           my_graph = sample_gnp(my_number_nodes,
                                 my_density,
                                 directed=FALSE,
                                 loops = FALSE)
           if(ecount(my_graph) >= my_number_nodes) {
              V(my_graph)$color <- my_colour
           }
         plot(my_graph,
              layout = layout.fruchterman.reingold,
              vertex.color = V(my_graph)$color)
}

set.seed(0)
my_graph_function(30,"green",0.3)
```



If all of this seems mysterious and opaque at first, it is important to remember that mastering a computer language requires time and practice.

Some languages are **dialects** or variants of other languages;[5] proficiency in one can make it is easier to become proficient in another. But not all programming languages follow the same paradigm: **imperative** languages (object-oriented programming) live in a different "linguistic family" than **declarative** languages (functional programming) languages.

5: Or at least, are mutually intelligible, like Swedish and Danish, say.

### 1.1.4 Designing With Pseudo-Code

Before we can start thinking about writing code (in whatever programming language whose syntax we have mastered), we need to think about what it means to **design an algorithm** (or a computer program). From a mathematical perspective, an **algorithm** is a (stochastic) function. We thus need to specify:

- the algorithm's **inputs**;
- its **outputs**, and
- the **procedure** to transform the inputs into the outputs.[6]

6: In the muffin recipe above, the ingredients are the inputs, the muffins themselves are the outputs, and the recipe instructions describe the transformation.

It is good programming practice to avoid typing up programs on the fly – code needs to be **planned**: we need to know what the program will do and how it will go about doing it before we commit it to a file, **independently of the language in which it will be implemented**.

"**Pseudo-code**" is a term used to describe a **rough sketch** of the algorithm, which indicates its expected inputs, outputs, and steps, while leaving the specifics of its functionality in "black boxes". Pseudo-code is usually designed with the main elements of code (e.g., variables, functions, logical flow, etc.), in a **language-agnostic** (i.e., human readable) manner.

**Example:** we might be interested in building an algorithm that would cluster the observations in a dataset, using a maximum number of "local" observations (see Chapters 19, *Machine Learning 101*, and 22, *Spotlight on Clustering*, for an in-depth discussion of this topic).

What might the following chunk of pseudo-code (which is part of the bigger clustering picture) do?

**Chunk of pseudo-code**

```
find_neighbours(array_of_points, max_n_neighbour_distance)
{
 for each point[i] in array_of_points
 {
  for each remaining point[j] in array_of_points
  {
   distance_between_ij = distance(point[i], point[j])
   if distance_between_ij <= max_n_neighbour_distance
   then neighbours[i] = add_to_neighbrs(point[i],point[j])
  }
 }
}
```

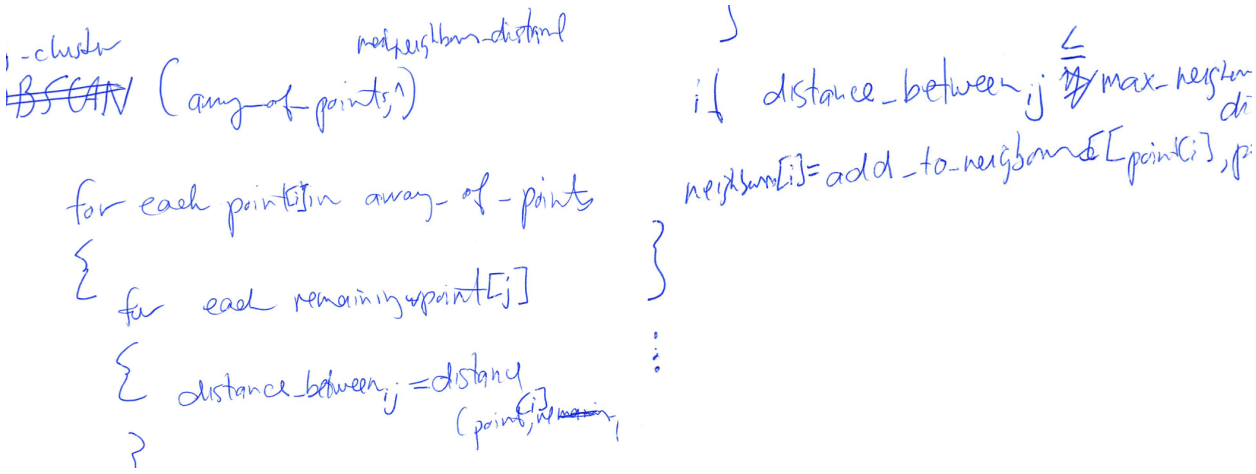


**Figure 1.4:** The first stage of pseudo-coding, in all its chicken scratch glory.

This is what is happening:

- the algorithm `find_neighbours()` takes as **inputs** a dataset `array_of_points` and a quantity `max_n_neighbour_distance`;
- for each observation `point[i]` in the dataset (`i` indexes the observations), it considers all other observations `point[j]` and computes their distances to the initial observation `point[i]` (one by one);
- when these distances are smaller than the input threshold `max_n_neighbour_distance`, it considers that the corresponding observation `point[j]` is a neighbour of observation `point[i]`, and adds the former to the neighbours of observation `point[i]`.

Evidently, this chunk of pseudo-code defines the **neighbourhood** of each observation in the dataset. Note the **black box** functions `distance()` and `add_to_neighbrs()`: their specifics are not provided,[7] but what they represent is clear. That is the **power** of pseudo-code.[8]

**Getting a feel** for the right level of pseudo code detail takes practice: should we drill down into what `add_to_neighbrs()` does? Do we need to describe what <= does? How much utility should be sacrificed in favour of understanding?

The answers to these questions depends on the **level of abstraction** of the programming language used to implement the algorithm:

- **high-level languages** (such as `R` and `Python`) contain tons of built-in functions, which allow for programming at higher levels of abstraction, whereas
- many details and functions must be programmed "by hand" in **low-level languages** (such as assembly and machine languages), which require lower levels.

The **strategy** to write useful pseudo-code is deceptively simple:

1. define the available inputs;
2. define the desired outputs, and
3. identify (and write down) a set of programmatic instructions (procedure) to transform the inputs into the outputs.[9]

7: Their eventual implementation may change depending on the computer language selected to write the program.

8: Of course, in practice, we also do not sit down and write pseudo-code on the fly... that too must be planned (see Figure 1.4).

9: This is easier said than done, obviously, and it looks an awful lot like the definition of an algorithm we provided previously, but remember that parts of the pseudo-code can be "**black boxed**", which is to say, that functionality can be described at a **high level**.

### 1.1.5 From Pseudo-Code to Code That Runs

Once we are satisfied that the pseudo-code provides a decent path to solving the problem at hand,[10] , we can start thinking about how to implement it into **real code** ("code that runs"):

1. we start by determining the **appropriate syntax** for the computer language that will be used and we re-write the pseudo-code as syntactically correct code in this language;
2. we replace all "black box" functions with real code, and
3. we determine how to connect the real code (the **software**) to the computer, so that it can be compiled/interpreted, and run by the computer (receiving inputs and generating outputs).

It might take multiple tries before this is done successfully. That is to be expected. It takes time, even for the most gifted programmer, to become an expert in a new language. The urge to feel defeated if (when?) the first few attempts fail is completely natural; as always, practice is the answer.

The process of taking the high-level code (which is really a text file) and getting it to run on a computer without a hitch requires a certain amount of infrastructure to be in place:

- libraries
- input/output + file system
- compilers/interpreters

In these notes, we are taking care of much of these issues by setting up the `R`/`Python` examples internally and running them locally (using our infrastructure); this works well for illustrating the concepts and working with pedagogical datasets, but the infrastructure conundrum must be tackled and solve before it becomes possible to produce useful and actionable data analysis results (see Chapter 17, *Data Engineering and Data Management*, for more details).

In general, there is no **single authoritative reference manual** describing how to use a particular computer language and/or how to make code run on particular hardware configurations, in no small part because coding and computer references become obsolete in the blink of an eye.[11]

Successfully coders must be embedded in a **community of coders**. Luckily, this is getting to be easier to do every day – most questions anybody could ever have about specific aspects of coding have already been answered somewhere online. Stack Exchange and similar sites can be quite useful in that regard.[12]

As a last remark on the topic, keep in mind that in the real coding world, **there is no such thing as cheating**: the objectives are to make happen the things you want to see happen. Getting help along the way is emphatically not prohibited (mind you, it is good practice to cite or acknowledge such help).

Crucially, though, we should not use code when we do not understand what it does – borrowed code may make complete sense in the context for which it was written, but may have **unintended ramifications** in a different context: **be careful!**

10: The proposed solution does not need to be final.

11: Consider the change from `Python` 2 to `Python` 3 as a cautionary tale.

12: **Fair warning:** some coder communities can be ... let us say, not overly welcoming of neophytes. It is not unusual for the answer to a question to be some variation on "look it up in the documentation". While this can be true in a general sense, such an answer is useless. We all know that things can be looked up in the documentation. And we all know that some users ask questions without taking the time to think about things, or in the hope that somebody else will do their work for them. It is in the best interest of learners to seek communities that make a concerted effort to be healthy and inclusive, to recognize that not every user has reached the same proficiency level. Such communities are plentiful online; do not waste any time and energy on gatekeepers.

### 1.1.6 Debugging

**PROGRAMMERS DRINKING SONG:**
99 little bugs in the code,
99 bugs in the code,
fix one bug, compile it again,
141 little bugs in the code.
141 little bugs in the code. . . .
(Repeat until bugs = 0)

Mistakes WILL happen. What do we do about that?

In the development phase, coding is about getting all the moving pieces to fit together, yes, but it is also about fixing the **bugs**, an "error in the source code that causes a program to produce unexpected results or crash altogether". Fixing these bugs (**debugging**) is mainly about revealing what is in memory at different points in the control flow of the code, to determine if it is actually doing what we think it ought to be doing.

As the quote at the start of the section implies, debugging is a bit of an art form, requiring the programmer to become a detective and a zen master (see The Tao of Programming). It teaches perseverance and humility, and it really helps us perfect our understanding of the language, of the code itself, and of the task at hand.

Debugging tools can help with all of this; at our level, debugging often requires running the code line-by-line until we can identify the chunk of code that is the culprit. Debugging is a **necessary** part of coding, no matter how experienced you are.

### 1.1.7 R/Python

There is only so much that can be said about programming **in general**; at some point, we need to select a computer language and get going in earnest.

At a foundational level, most programming languages are roughly **equivalent** (Turing-complete or Turing-equivalent), in the sense that anything that can be done with one can also (more or less) be done with another. But that does not mean that they are all **equally useful**.

Some are better suited to certain tasks, whether because they are less memory-intensive, or more elegant, or more intuitive, and so on. Even in the data analysis world, there are competing paradigms. In these notes, we will use two of the most popular languages (although by no mean the only ones): `R` and `Python`.

In the examples we provide, `R` code appears in blue boxes:

```
... some R code ...
```

Whereas `Python` code appears in green boxes:

```
... some Python code ...
```

**Object-Oriented Languages vs. Procedural Languages** `R` and `Python` are **objected-oriented languages**, as opposed to **procedural languages**.

> The focus of procedural programming is to break down a programming task into a collection of variables, data structures, and subroutines, whereas in object-oriented programming it is to break down a programming task into objects that expose behavior (methods) and data (members or attributes) using interfaces. The most important distinction is that while procedural programming uses procedures to operate on data structures, object-oriented programming bundles the two together, so an "object", which is an instance of a class, operates on its "own" data structure. [3]

This will make more sense if we first understand the concepts of:

- data types
- data structures
- functions

Languages have a set of built-in basic **variable types**, such as:

- integer: 5
- character: 'm'
- list: (5, 3, 9)

Other variables types can be built up out of these basic types, such as

- strings, which are list of characters: ('t', 'a', 'b', 'l', 'e')

We can also define related variables – a **data structure**:

- `struct myNames = {string firstName, string middleName, string lastName}`
- `jenNames` might be a variable of type `myNames`, with `firstName = Jen`, `middleName = Adele`, `lastName = Schellinck`.

In addition a programmer might want to be able to carry out a set of predefined instructions, or **functions**, on that data structure:

- `jenNames.print_middleName` or
- `jenNames.string_length_lastName`, say (what these functions do should be clear from their name).

Loosely speaking, an **object** is a user-defined data structure, together with a set of functions that are specific to that structure.

The **data frame object** in `R` is structured similarly to a spreadsheet:

- it has rows and columns, with associated row and column names, and
- we can carry out predefined operations (mean, count, etc.) on specific values, on selected rows, or selected columns, or the data frame as a whole.

Learners that are familiar with databases and/or languages that are more vector-focused (e.g. Java) might find the data frame implementation in `R` frustrating; those who are familiar with matrices and other mathematical concepts used in data analysis, less so.

## 1.2 Introduction to R

R is a powerful language that is widely-used for data analysis and statistical computing. It was developed in the early 90s by Ross Ihaka and Robert Gentleman, as a successor to S, a **statistical programming language**.

The inclusion of sophisticated packages (such as `dplyr`, `tidyr`, `readr`, `data.table`, `SparkR`, `ggplot2`, etc.) has made R both more powerful and more useful, allowing for smart data manipulation, visualization, and computation, using its built-in data structures and functionality.

Notably, it has gained prominence as a free and open source alternative to expensive statistical software.

### 1.2.1 Why Use R

Here are some benefits that potential users might note:

- the style of coding is intuitive;
- R is open source and free;
- more than 18,500 packages, customized for various computation tasks, are available (as of February 2022);
- the R community is overwhelmingly welcoming and useful to new users and experienced users alike;[13]
- high performance computing experience is possible (with the appropriate packages), and
- is is one of the highly sought skills by analytics and data science companies.

13: You can browse and ask questions at StackOverflow, and consult worked-out examples on R-bloggers, for instance.

### 1.2.2 Installing R / RStudio

**Note**: If you have a pre-existing installation of R and/or RStudio, you may skip this part. However, we highly recommend that both of these applications be upgraded to the most recent version, if they have not been upgraded for a while.[14]

Data analysis can be conducted using the **vanilla** (base) version of R, but also using RStudio provides a better coding experience, in our opinion.

The following steps will allow you to install R and RStudio.

1. Download and install R at https://cloud.r-project.org.
   - *Windows* users should click on **Download R for Windows**, then click on **base**, then click on the **Download R X.X.X for Windows** link, where `R X.X.X` is the version number. For example, the latest version of R as of 2022-02-07, was `R 4.1.2`;
   - *macOS* users should click on **Download R for macOS**, then on **R-X.X.X.pkg** (under "**Latest release:**"), where `R-X.X.X` is the version number. If the Mac has an Arm-based M1 chip, choose **R-X.X.X-arm64.pkg** instead;
   - *Linux* users should click on **Download R for Linux** and choose the specific distribution for more information on installing R for their actual setup.

14: Note that these instructions can quickly become obsolete; we will do what we can to stay on top of them, but you may need to consult other sources or search for "Installing R and RStudio" online. Consult *Upgrading R and/or RStudio* on 15 for details.

2. Download and install RStudio at posit.co/download/rstudio-desktop/#download .
    - look for the big blue button that says **DOWNLOAD RSTUDIO DESKTOP FOR ...**, where `...` represents the desired OS;
    - click on the button to start downloading;
    - Once downloading has completed, double-click the file to open it, and follow the installation instruction.
3. **(for *macOS* users only):** Download and install XQuartz.[15]
    - go to https://www.xquartz.org . Under "Quick Download", click on "XQuartz-2.8.1.dmg";
    - save the `.dmg` file, double-click it to open, and follow the installation instructions (you may need to restart your computer).
    - **Reminder**: you will need to re-install XQuartz when upgrading your macOS to a new major version.
4. Even with both `R` and RStudio installed, we will refrain from working directly with the `R` interface, given that RStudio provides such a "nice" **shell** over the engine that is `R`.

15: What is XQuartz and why does macOS users need it?

Once RStudio is opened, the **graphic user interface** (GUI) displays 4 panes, as in Figure 1.5.

- **Console:** bottom left; this area shows the output of code that has been run (either from the command line in the console or from the script window);
- **Script:** top left; as the name suggests, this is the area one would typically use to write code. Lines can be run by first selecting them (right-clicking) and pressing `ctrl + enter` (win) or `cmd + enter` (mac) simultaneously. Alternatively, you can click on the little 'Run' button located at the top right corner of the script window;
- **Environment:** top right; this space displays the set of external elements that have been added. This includes data set, variables, vectors, functions etc. This area allows the user to verify that data has been loaded properly;
- **Graphical Output:** bottom right; this space display the graphs created during exploratory data analysis, or embedded help on package functions from `R`'s official documentation.

### 1.2.3 Test, Test, Test!

To make sure you have installed both `R` and RStudio properly, type a simple command in the console. For example, place your cursor in the pane labelled `Console`, type `x <- 2 + 2` at the prompt, followed by `enter` or `return`, then type `x`, again followed by `enter` or `return`.

**Testing R**

```
x <- 2 + 2
x
```

You should see the value 4 printed to the screen.

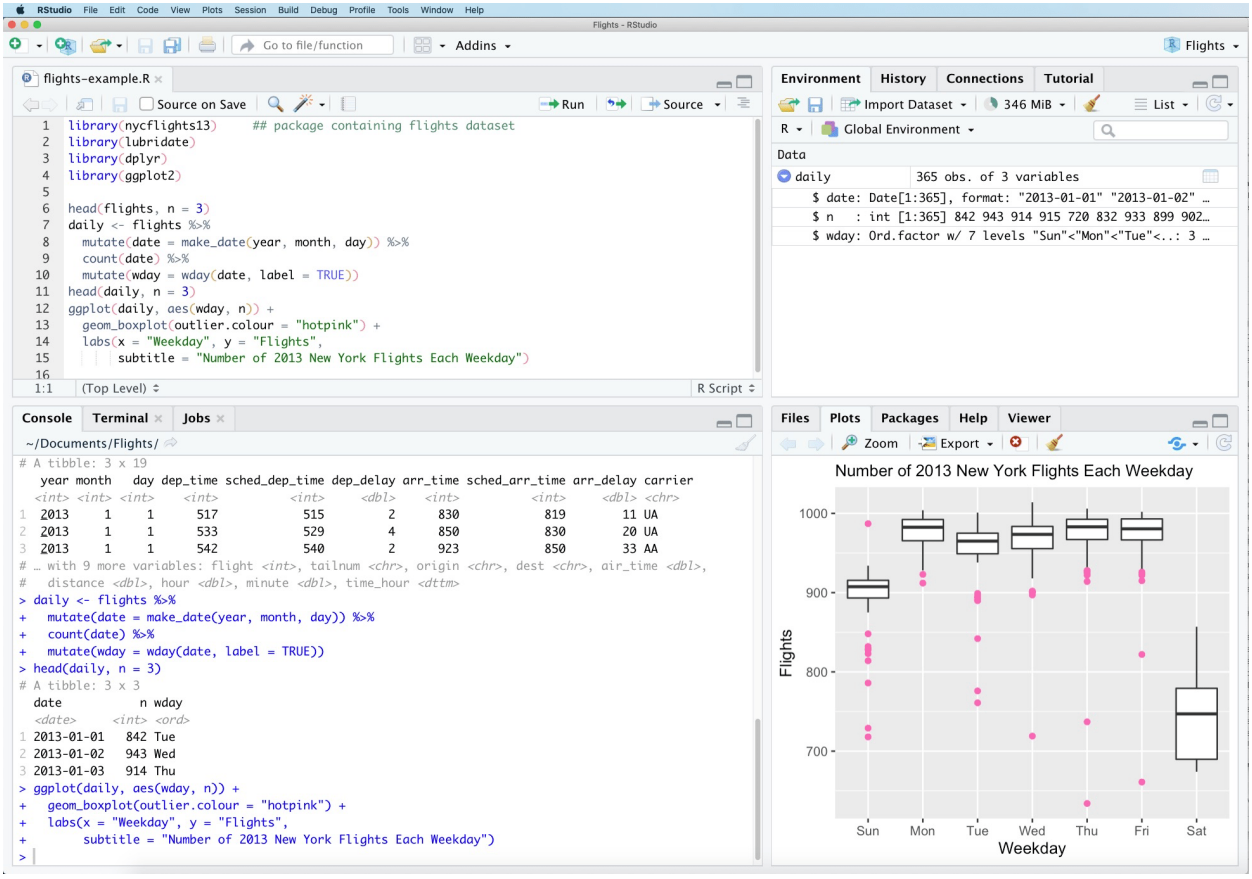
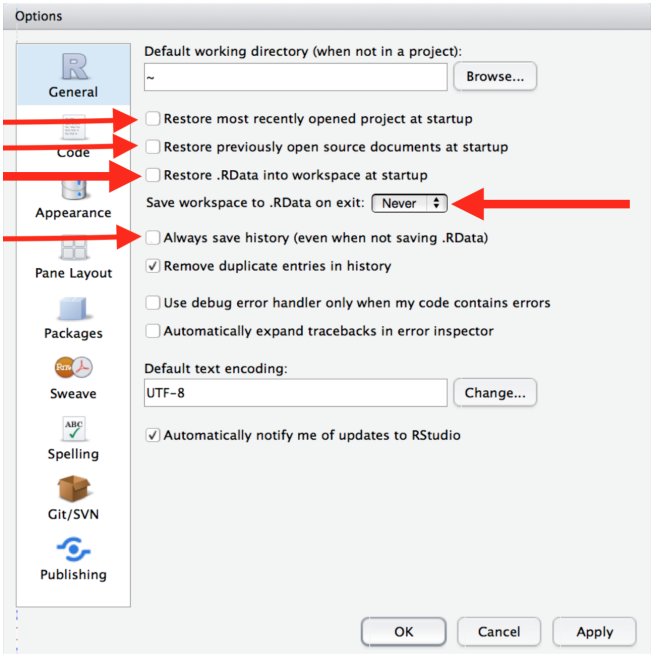



**Figure 1.5:** RStudio interface, with 4 default windows: Console, Script, Environment, and Graphical Output.

### 1.2.4 Customizing RStudio

We would like to suggest the following settings for your `R`/RStudio installation, following [16, ch.8].[16] In RStudio, go to `Tools >> Global Options`, and make the changes described below:

16: Feel free to ignore the suggestion as you wish.



> [These settings] will cause you some short-term pain, because now when you restart RStudio it will not remember the results of the code that you ran last time. But this short-term pain will save you long-term agony because it forces you to capture all important interactions in your *source code*. There's nothing worse than discovering three months after the fact that you've only stored the *results* of an important calculation in your workspace, not the calculation itself in your source code. [16]

Optionally, you could also adjust the font size via `Tools >> Global Options >> Appearance >> Editor font size`.[17]

17: By default, it is set at 12, but a larger font size may be easier on the eyes.

### 1.2.5 Upgrading R / RStudio

We suggest always working with the latest version of R and RStudio.

- To upgrade R, find out the current version of R running on your computer. You can do so from within the RStudio Console:

  **R version**

  ```
  R.version.string
  ```

  ```
  [1] "R version 4.1.3 (2022-03-10)"
  ```

  As of January 2023, the most recent version of R is `4.2.2`. If you have an older version installed on your computer, go to cloud.r-project.org and follow the steps described in on p. 12 (*Installing R / RStudio*) to install the latest version of R. You can confirm that the upgrade was successful by restarting RStudio and typing `R.version.string` in the console again.

- To upgrade RStudio from within RStudio, go to `Help > Check for Updates` to install newer version of RStudio (if available). Once both R and RStudio have been upgraded, test by typing some simple command in the console (as on p. 13, *Test, Test, Test!*).

### 1.2.6 Basics of R

How are the elements of code (introduced in *Code Components* on p. 4) implemented in R? How do they mesh with one another to form interpretable code? First, we should mention that while R is technically object-oriented, this tends to be hidden in practice; the language is thus especially well-suited for **quick**, **interactive**, and **intuitive** scripting and data exploration.

Note as well that it uses special built-in notation for statistical models, which would not usually be found in other languages (hence the "statistical programming" moniker). Some of the examples and explanations provided in the text are modified from [16, 6, 5, 11, 2, 8, 12].

The rest of this section contain information on the basic use of R; more examples are available in Section 1.3 (*More About Programming in R*) and throughout the course notes.

**Simple Computations in R** We will get familiar with the R coding environment, we start by showing how the console can be used as an interactive calculator.

Type the first line of each group in your console, followed by a carriage return to confirm that R works as we would expect of a calculator:

```
2 + 3
```

```
[1] 5
```

```
(3*8)/(2*3)
```

```
[1] 4
```

```
log(12)
```

```
[1] 2.484907
```

```
sqrt(121)
```

```
[1] 11
```

You can experiment with various combinations of calculations.

Should you want to modify or repeat a prior calculation, press the `Up Arrow` when the cursor is in the console to cycle through previously executed commands; pressing `Enter` re-runs the selected computation.

On the other hand, you can avoid scrolling through a wall of computations by creating a **variable**. In R, this is done *via* the variable assignment symbols `<-` or `=`.[18] Once a variable exists in memory, the output does not get printed directly unless it is called directly at the prompt, or if the variable assignment is surrounded with a pair of parentheses.

18: There are 3 others such symbols, but no language needs 5 assigners, let alone 2, so we will not introduce them here.

```
x <- 8 + 17
x
```

```
[1] 25
```

```
(y <- 8 + 17)
```

```
[1] 25
```

Variables can be named using any combination of alphanumeric symbols, but the name has to start with a letter (`a-z`, `A-Z`) and cannot contain spaces and punctuation marks other than periods and dashes.

**R Packages** **Packages** (or libraries) contain pre-compiled functions and objects that could be useful in specific settings.

To install a package, simply type:

```
install.packages("package_name")
```

Take note of the quotation marks. You can type this code directly in the console, followed by a carriage return, or enter it in the script window and click `Run` in the menu at the top.

The base distribution already comes with some high-priority add-on packages, namely:

| | | | |
|---|---|---|---|
| KernSmooth | MASS | boot | class |
| foreign | lattice | mgcv | nlme |
| rpart | spatial | survival | base |
| grDevices | graphics | grid | methods |
| stats | stats4 | tcltk | tools |
| cluster | nnet | datasets | splines |

These packages implement standard statistical functionality, for example linear models, classical tests, a huge collection of high-level plotting functions, and tools for survival analysis. Once a package is installed, it needs to be **loaded** before its objects (datasets, functions) can be used. This can be done by typing:[19]

19: Since entering instructions is always done in one of the ways described above, we will stop specifying where and how it must be done.

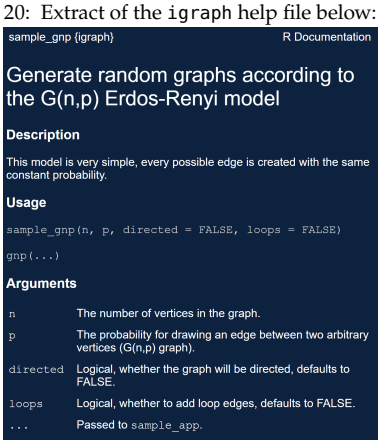
```
libary(package_name)
```

Note the absence of the quotation marks.

For instance, in *Code Components* (see p. 4), we loaded the `igraph` package to take advantage of the pre-compiled functions `sample_gnp()`, `ecount()`, `V()`, and `plot()`. The first 3 functions are not in the base distribution; the last function `plot()` does exist, but it would not know how to handle graph objects without the special instructions provided by `igraph`.

The **help file** for compiled functions can be displayed in the graphical output window by using the reserved character "`?`", as below (assuming that the `igraph` library has been loaded).[20]

20: Extract of the `igraph` help file below:


sample_gnp {igraph} R Documentation

Generate random graphs according to the G(n,p) Erdos-Renyi model

Description

This model is very simple, every possible edge is created with the same constant probability.

Usage

```
sample_gnp(n, p, directed = FALSE, loops = FALSE)
gnp(...)
```

Arguments

| | |
|---|---|
| n | The number of vertices in the graph. |
| p | The probability for drawing an edge between two arbitrary vertices (G(n,p) graph). |
| directed | Logical, whether the graph will be directed, defaults to FALSE. |
| loops | Logical, whether to add loop edges, defaults to FALSE. |
| ... | Passed to sample_app. |

```
?sample_gnp
```

In more sophisticated code, it is conceivable that we would want to load multiple libraries; because we might forget which function is associated with which library, or even that different libraries use the same name for different functions, it is **good practice** to forego explicitly loading a library in favour of directly fetching the required functionality (the package must be installed first, however). In `R`, this is done as follows:

```
package_name::function_name(function_parameters)
```

For instance, the graph code from above can be replaced by the following chunk:

```
my_graph_function <- function(my_number_nodes,
                              my_colour,
                              my_density) {
  my_graph = igraph::sample_gnp(my_number_nodes,
                                my_density,
                                directed=FALSE,
                                loops = FALSE)
  if(igraph::ecount(my_graph) >= my_number_nodes) {
            igraph::V(my_graph)$color <- my_colour
  }
  plot(my_graph, vertex.color = igraph::V(my_graph)$color)
}

my_graph_function(30,"green",0.3)
```

Note, however, that this strategy is not always optimal (in particular, when using the **pipeline operator**, see p. 41).

**R Essentials** Everything you see or create in R is an **object**: vectors, matrices, data frames, even variables (and functions) are objects.

R allows 5 basic classes of objects:

- Character
- Numeric (real numbers)
- Integer (whole numbers)
- Complex
- Logical (True / False)

Each of these classes has **attributes**. An object can have the following attributes:

- names, dimension names
- dimensions
- class
- length
- etc.

An object's various attributes can be accessed using the `attributes()` function. We will have more to say on this topic.

The most basic R object is the **vector**. An empty vector can be created using `vector()`. A vector contains various objects, but all must be of the same class.[21]

Vectors can also often created using the **combine** (or concatenate) operator `c()` (which makes it a singularly bad idea to use `c` as a variable name).

21: That can cause unforeseen difficulties as it is not always easy to visually distinguish between a real number (*numeric*) and an *integer*. Furthermore, the digits of a number can be represented as character strings in some cases.

```
(a <- c(1.8, 4.5))                    # numeric
(b <- c(1 + 2i, 3 - 6i))              # complex
(d <- c(23, 44))                      # integer
```

```
(e <- vector("logical", length = 5))  # logical
(f <- c("abc","def"))                 # character
```

```
[1] 1.8 4.5
[1] 1+2i 3-6i
[1] 23 44
[1] FALSE FALSE FALSE FALSE FALSE
[1] "abc" "def"
```

Comments can be introduced in R code *via* the # symbol: all characters following a pound symbol are ignored by R until the next line of code (so the classes in the example above would not be part of the code proper).

**R Data Types and Objects** There are various types of R objects.

**Vectors** As mentioned above, a **vector** contains objects of the same class. We may have a need to mix objects of different classes in a list – this can be done to a vector by **coercion**. This has the effect of 'converting' objects of different types to the same class. For instance:

```
# coercion to character
(vec <- c("Time", 25,TRUE,"retro", 2.22))
# coercion to numeric
(bbb <- c(FALSE, 11))
# coercion to character
(i.a <- c(215,"October"))
```

```
[1] "Time"  "25"    "TRUE"  "retro" "2.22"
[1]  0 11
[1] "215"     "October"
```

We can verify the class of these objects using the `class()` function.

```
class(vec)
class(bbb)
class(i.a)
```

```
[1] "character"
[1] "numeric"
[1] "character"
```

To convert the class of a vector, we can use the `as.` command.

```
g <- 10:16        # create a vector of 7 integers
class(g)          # find bar's class
as.numeric(g)     # convert to numeric
class(g)
as.character(g)   # convert to character
```

```
class(g)
```

```
[1] "integer"
[1] 10 11 12 13 14 15 16
[1] "integer"
[1] "10" "11" "12" "13" "14" "15" "16"
[1] "integer"
```

We can change the class of any vector using a similar approach. But be careful – while we can convert a numeric vector into a character one, going the other way will introduce NAs (conversion is subject to R's internal class rules).

**Lists** A **list** is a special type of object which can contain elements of different data types.

```
my.list <- list(254,"abab", TRUE, 0 - 3i)
my.list
```

```
[[1]]
[1] 254

[[2]]
[1] "abab"

[[3]]
[1] TRUE

[[4]]
[1] 0-3i
```

The output of a list differs from that of a vector, since all the objects are of different types. The double bracket `[[1]]` shows the index of the first element and so on. The elements of a list can be extracted by using the appropriate index:

```
my.list[[3]]
```

```
[1] TRUE
```

The single single bracket `[ ]` also has a role: it returns the list element with its index number, instead of the result above.

```
my.list[3]
```

```
[[1]]
[1] TRUE
```

**Matrices** A vector for which rows and columns are explicitly identified is a **matrix**, a 2-dimensional data structure. All the entries of a matrix have to be of the same class. The following code produces a 6 by 3 matrix consisting of the first 18 integers.

```
my.matrix <- matrix(1:18, nrow=6, ncol=3)
my.matrix
```

```
     [,1] [,2] [,3]
[1,]    1    7   13
[2,]    2    8   14
[3,]    3    9   15
[4,]    4   10   16
[5,]    5   11   17
[6,]    6   12   18
```

The dimensions of a matrix can be obtained using either the `dim()` or `attributes()` commands (the matrix dimensions are a matrix's only attributes in R).

```
dim(my.matrix)
attributes(my.matrix)
```

```
[1] 6 3
$dim
[1] 6 3
```

To extract a particular element from a matrix, simply use the appropriate indices. What might you expect to see from the following commands?

```
my.matrix[5,2]          # row 5, col 2
my.matrix[c(1,2,4),2]   # col 2, rows 1, 2, 4
my.matrix[4,2:3]        # row 4, cols 2, 3
my.matrix[,2]           # col 2
my.matrix[4,]           # row 4
my.matrix[c(1,1,4),2]   # col 2, rows 1, 1, 4
```

```
[1] 11
[1]  7  8 10
[1] 10 16
[1]  7  8  9 10 11 12
[1]  4 10 16
[1]  7  7 10
```

As an aside, it is straightforward to create a matrix from any vector, by assigning the dimensions using `dim()`.

For instance, we start by reading in a vector of ages:

```
age <- c(23, 8, 5, 44, 15, 12, 31, 19, 16)
age
```

```
[1] 23  8  5 44 15 12 31 19 16
```

Then reshape the vector as a 3 x 3 matrix:

```
dim(age) <- c(3,3)
age
class(age)
```

```
     [,1] [,2] [,3]
[1,]   23   44   31
[2,]    8   15   19
[3,]    5   12   16
[1] "matrix" "array"
```

Matrices can also be created by joining two vectors (with matching dimensions) using `cbind()` or `rbind()`:

```
x <- c(1, -2, 3, -4, 5, -6)
y <- c(200, 300, 400, 500, 600, 700)
cbind(x, y)
rbind(x,y)
```

```
      x   y
[1,]  1 200
[2,] -2 300
[3,]  3 400
[4,] -4 500
[5,]  5 600
[6,] -6 700
  [,1] [,2] [,3] [,4] [,5] [,6]
x    1   -2    3   -4    5   -6
y  200  300  400  500  600  700
```

```
class(x)
class(y)
class(cbind(x, y))
class(rbind(x, y))
```

```
[1] "numeric"
[1] "numeric"
[1] "matrix" "array"
[1] "matrix" "array"
```

We will discuss how R implements regular matrix operations (transpose, multiplication, addition, etc.) in Chapter 3 (*Overview of Linear Algebra*).

**Data Frames** The **data frame** is R's most commonly-used (and most convenient) data type, especially for data analysis tasks.

Like matrices, we can use data frames to store tabular (rectangular) data, but unlike matrices, a data frame can accommodate lists of vectors of different classes: each column of a data frame acts like a list.

When data is read into R, it is first stored as a data frame.

The following bit of code, for instance, creates a data frame with two columns, `name` and `age`:

```
df <- data.frame(
  name = c("Patrick","Brownyn","Elowyn",
           "Llewellyn","Gwynneth"),
  age = c(45,41,19,8,5)
)
df
```

```
       name age
1   Patrick  45
2   Brownyn  41
3    Elowyn  19
4 Llewellyn   8
5  Gwynneth   5
```

Here are some of `df` attributes:

```
dim(df)
str(df)
nrow(df)
ncol(df)
```

```
[1] 5 2
'data.frame':   5 obs. of  2 variables:
 $ name: chr  "Patrick" "Brownyn" "Elowyn" "Llewellyn" ...
 $ age : num  45 41 19 8 5
[1] 5
[1] 2
```

In the code above, `df` is the name of data frame, `dim()` returns its dimensions, `str()` its structure (i.e., the list of variables stored in the data frame), and `nrow()` and `ncol()`, the number of rows and number of columns in the data frame, respectively.

**Reading Data and Writing** Reading data into a statistical system for analysis, and exporting the results to some other system for report writing, can be frustrating tasks that take far more time than the statistical/data analysis itself, but the former task is required if the latter is to be undertaken in earnest.

We describe the import/export facilities available in R itself or via packages available from Comprehensive R Archive Network (CRAN).

R comes with a few **data reading** functions:

- `read.table()`, `read.csv()` for tabular data;
- `readLines()` for lines of a text file;
- `source()`, `dget()` to read R code files (inverse of `dump()` and `dput()`, respectively);
- `load()` to read-in saved workspaces;
- `unserialize()` to read single R objects in binary form.

There are, of course, numerous R packages that have been developed to read in all kinds of other datasets, and you may need to resort to one of these packages if you are working in a specific area.

**`read.table()`** The `read.table()` function is one of the most commonly-used functions for reading data. The help file[22] is worth reading if only because the function gets so much use. Its main arguments are:

22: Run `?read.table` in the console.

- `file`, the name of a file, or a connection;
- `header`, logical indicating if the file has a header line;
- `sep`, string indicating how the columns are separated;
- `colClasses`, character vector indicating the class of each column in the dataset;
- `nrows`, number of rows in the dataset;[23]
- `comment.char`, character string indicating comments;[24]
- `skip`, the number of lines to skip from the beginning of the file;
- `stringsAsFactors`, whether character variables are coded as factors or as strings.[25]

23: By default `read.table()` will read the entire file.

24: Defaults to "#".

25: Defaults to `TRUE` because back in the old days, strings represented levels of a categorical variable; now that text mining is an every day occurrence, that is not always the case.

**For small to moderately sized datasets**, you can usually call `read.table()` without specifying any other arguments

```
data <- read.table("foo.txt")
```

In this case, R will read in the file `foo.txt` an automatically:

- skip lines that begin with a #;
- figure out how many rows there are (and how much memory needs to be allocated), and
- figure what type of variable is in each column of the table.

Telling R all these things directly makes R run faster and more efficiently. The `read.csv()` function is identical to `read.table()` except that some of the defaults are set differently (such as the `sep` argument).

**With much larger datasets**, some things can be done to prevent R from choking on the data (a risk as R stores everything in RAM):

- read the help page for `read.table()`, which contains many hints;
- make a rough calculation of the memory required to store the dataset (see on the next page for an example); if the dataset is larger than the amount of RAM on your computer, it is best to stop here;
- set `comment.char = ""` if all lines in the file are uncommented;
- use the `colClasses` argument – specifying this option can make `read.table()` run MUCH faster, often twice as fast.[26] We can figure out the column classes *via* the following code:

26: In order to use this option, we must know the class of each column in the data frame; if all of the columns are "numeric", for example, then we would simply set `colClasses = "numeric"`.

```
initial <- read.table("datatable.txt", nrows = 100)
classes <- sapply(initial, class)
tabAll <- read.table("datatable.txt",
                     colClasses = classes)]
```

- set `nrows` – this doesn't make `R` run faster but it helps with memory usage (a mild overestimate is okay; the Unix tool `wc` can be used to calculate the number of lines in the file).

In general, when using `R` with larger datasets, it is also useful to know a few things about the operating system:

- how much memory is available on the system?
- what other applications are in use?[27]
- are other users logged into the same system?
- what is the operating system? (some operating systems can limit the amount of memory a single process can access).

27: Close everything that is not required.

For example, suppose we have a data frame with 2,000,000 rows and 100 columns, all of which are numeric data. Roughly speaking, how much memory is required to store this data frame?

On most computers, numeric data is stored using 64 bits of memory (8 bytes). Given that information, we have:

$$\begin{aligned} 2,000,000 \times 100 \times 8 \text{ bytes} &= 1,600,000,000 \text{ bytes} \\ &\approx 1,600 \text{ MB} = 1.6 \text{ GB}. \end{aligned}$$

Reading in a large dataset for which one does not have enough RAM is an easy way to get the computer (or the `R` session) to freeze. This is usually an unpleasant experience that requires killing the `R` process, in the best case scenario, or rebooting the computer, in the worst case.

It is always a good idea to do a rough memory requirements calculation before reading in a large dataset.

#### `txt`, `csv`, and Other Formats

- **Fixed format text files**

```
# Windows only
df = read.table("folder\\file.txt", header=TRUE)
# all OS (including Windows)
df = read.table("folder/file.txt", header=TRUE)
```

The forward slash `/` is supported as a directory delimiter on all operating systems; the double backslash `\\` is only supported under Windows. If the first row of the file includes the name of the variables, these entries will be used to create appropriate names[28] for each of the columns in the dataset. If the first row does not include the names, the `header` option can be left off (or set to `FALSE`), and the variables will be named `V1`, `V2`, ..., `Vn`.

28: Reserved characters such as '$' are changed to '.'

A limit on the number of lines to be read can be specified through the `nrows` option. The `read.table()` function also supports using a URL as a filename or browsing files interactively using `read.table(file.choose())`.

Sometimes data arrives in irregularly-shaped data files (there may be a variable number of fields per line, or some data in the line may describe the remainder of the line). In such cases, a useful generic approach is to read each line into a single character variable, then use character variable functions to extract the contents.

```
df = readLines("file.txt")
df = scan("file.txt")
```

The `readLines()` function returns a character vector with length equal to the number of lines read. A limit on the number of lines to be read can be specified through the `nrows` option. The `scan()` function returns a vector, with entries separated by white space by default. These functions read from standard input, but can also read a file or a URL.

- **Comma-separated value (CSV) files:** the `read.csv()` function takes on much the same parameters as `read.table()`.

```
df = read.csv("folder/file.csv")
```

- **Read sheets from an Excel file:** if the data is available in an Excel file, various possibilities exist, depending on the spreadsheet format.

```
df.xls = gdata::read.xls("file.xls", sheet=1)
df.xlsx = xlsx::read.xlsx("file.xlsx", sheet=1)
```

The sheet can be provided as either a number or a name.[29]

29: The appropriate packages should have been installed beforehand, however.

- **Reading datasets in other formats:** the datasets of interest sometimes comes from another software. The `foreign` library is able to do a native import for some of the most common formats: Stata, Epi Info, Minitab, Octave, SPSS, Systat, and SAS files.[30]

30: The `read.ssd()` function will only work if SAS is installed locally, however.

```
df = foreign::read.dbf("filename.dbf")
df = foreign::read.epiinfo("filename.epiinfo")
df = foreign::read.mtp("filename.mtp")
df = foreign::read.octave("filename.octave")
df = read.ssd("filename.ssd")
df = read.xport("filename.xport")
df = read.spss("filename.sav")
df = read.dta("filename.dta")
df = read.systat("filename.sys")
```

There are analogous functions for **writing data** to files:

- `write.table()` writes tabular data to text files (i.e. CSV);
- `writeLines()`, to write character data line-by-line to a file;

- `dump()`, for dumping a textual representation of multiple R objects;
- `dput()`, for outputting a textual representation of an R object;
- `save()`, for saving an arbitrary number of R objects in binary format (possibly compressed) to a file, and
- `serialize()`, for converting an R object into a binary format for outputting to a file.

There are numerous ways to store data, including structured text file formats like CSV or tab-delimited, or complex binary formats. It is important to take the time to explore the full range of functionality in order to achieve your specific aims.

## 1.3 More About Programming in R

Many software packages and libraries are available to the data analyst. R not only has the advantage that we can easily use its available packages, but it provides enough flexibility for the analyst who wants to get dirty with the data.

In this section, you will find examples and tips that highlight R's data manipulation features. It is not meant to be a complete introduction, or even necessarily a showcase of good programming practices.

### 1.3.1 Help and Documentation

R's various help files and demos can be accessed using the following commands (where function_name and search_term correspond to the desired function and/or term):

- `?function_name`
- `example(function_name)`
- `args(function_name)`
- `??search_term`

For instance, the following code would display the help file for the function `glm()` in the bottom graphical output window of RStudio:

```
?glm
```

Most help files contain examples showcasing the use of the function. These can be accessed *via* `example()`.

```
example(glm)
```

We can thus copy the code from the example file, and run it directly at the console.

```
counts <- c(18,17,15,20,10,20,25,13,12)
outcome <- gl(3,1,9)
treatment <- gl(3,3)
print(d.AD <- data.frame(treatment, outcome, counts))
```

```
glm.D93 <- glm(counts ~ outcome + treatment,
               family = poisson())
anova(glm.D93)
summary(glm.D93)
```

```
  treatment outcome counts
1         1       1     18
2         1       2     17
3         1       3     15
4         2       1     20
5         2       2     10
6         2       3     20
7         3       1     25
8         3       2     13
9         3       3     12
Analysis of Deviance Table

Model: poisson, link: log

Response: counts

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                          8    10.5814
outcome    2   5.4523         6     5.1291
treatment  2   0.0000         4     5.1291

Call:
glm(formula = counts ~ outcome + treatment, family = poisson())

Deviance Residuals:
       1         2         3         4         5
-0.67125   0.96272  -0.16965  -0.21999  -0.95552
       6         7         8         9
 1.04939   0.84715  -0.09167  -0.96656

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.045e+00  1.709e-01  17.815   <2e-16 ***
outcome2    -4.543e-01  2.022e-01  -2.247   0.0246 *
outcome3    -2.930e-01  1.927e-01  -1.520   0.1285
treatment2   1.338e-15  2.000e-01   0.000   1.0000
treatment3   1.421e-15  2.000e-01   0.000   1.0000
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10.5814  on 8  degrees of freedom
Residual deviance:  5.1291  on 4  degrees of freedom
AIC: 56.761

Number of Fisher Scoring iterations: 4
```

Similarly, the function's arguments can be accessed via `args()`.

```
args(glm)
```

```
function (formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset, control = list(...),
    model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, singular.ok = TRUE,
    contrasts = NULL, ...)
NULL
```

### 1.3.2 Simple Data Manipulation

So what can we actually do with R?

**Loading a Built-In Dataset** We can obtain a list of such datasets in the `datsets` package by calling the following function:

```
data()
```

Or those available in all installed packages *via*:

```
data(package = .packages(all.available = TRUE))
```

Let us take a look at the `swiss` built in dataset.[31] We can display the dataset by simply calling it at the prompt, like so:

31: Type `?swiss` to see the help file.

```
swiss
```

Or we can take a look at its first or last `n` entries using the functions `head()` or `tail()`.

```
head(swiss,6)
```

```
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
```

**Assigning Data** We can create, assign, and display a vector consisting of a sequence of numbers like this:

```
(x<- c(1:3))
```

```
[1] 1 2 3
```

We can also assign non-sequential numbers:

```
(w <- c(12,-9))
```

```
[1] 12 -9
```

or mixed objects:

```
(v = c(w,"pomplamoose"))
```

```
[1] "12"           "-9"           "pomplamoose"
```

or matrices:

```
(u = t(matrix(1:10,ncol=5)))
```

```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[4,]    7    8
[5,]    9   10
```

**Data Types and Conversion** We can test whether objects are of a certain type or class:

```
is.numeric(x)
```

```
[1] TRUE
```

```
is.character(x)
```

```
[1] FALSE
```

```
is.vector(x)
```

```
[1] TRUE
```

```
is.matrix(x)
```

```
[1] FALSE
```

```
is.data.frame(x)
```

```
[1] FALSE
```

```
is.character(w)
```

```
[1] FALSE
```

```
is.character(v)
```

```
[1] TRUE
```

```
is.data.frame(swiss)
```

```
[1] TRUE
```

We can also set an object to be of a specific type:

```
as.numeric(x)
```

```
[1] 1 2 3
```

```
as.character(x)
```

```
[1] "1" "2" "3"
```

```
as.vector(x)
```

```
[1] 1 2 3
```

```
as.matrix(x)
```

```
     [,1]
[1,]    1
[2,]    2
[3,]    3
```

```
as.data.frame(x)
```

```
  x
1 1
2 2
3 3
```

Or combine two vectors into a single vector:

```
c(y,w)
```

```
[1] 200 300 400 500 600 700  12  -9
```

Or convert vectors to matrices or data frames:

```
cbind(x,y)
```

```
     x   y
[1,] 1 200
[2,] 2 300
[3,] 3 400
[4,] 1 500
[5,] 2 600
[6,] 3 700
```

```
rbind(x,y)
```

```
  [,1] [,2] [,3] [,4] [,5] [,6]
x    1    2    3    1    2    3
y  200  300  400  500  600  700
```

```
data.frame(x,y)
```

```
  x   y
1 1 200
2 2 300
3 3 400
4 1 500
5 2 600
6 3 700
```

Conversely, we can convert a matrix to a vector:

```
as.vector(u)
```

```
[1]  1  3  5  7  9  2  4  6  8 10
```

or a matrix to a data frame:

```
as.data.frame(u)
```

```
  V1 V2
1  1  2
2  3  4
3  5  6
4  7  8
5  9 10
```

or a data frame to a matrix:

```
swiss_matrix=as.matrix(swiss)
head(swiss_matrix)
```

```
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
```

**Writing Functions** One of R's most advantageous feature is its flexibility: what if we want to write our own functions? The template for all functions is a block of code that looks like:

```
my.function <- function(arg1,arg2, ..., argn) {
     # what my.function does
     # typically involving the arguments
}
```

Here are some (truly) simple examples: here is a function, `my.product()`, that computes the product of two arguments $x$ and $y$.[32]

32: This is not a very interesting function as the standard multiplication $*$ is already defined in R, but this is just an illustration of the functionality.

```
my.product <- function (x,y) {
    x*y
}
```

Note that the function definition must be compiled (the code must be run) before it can be called in the code.

There are multiple ways to call `my.product()` for arguments `x=12` and `y=-2`.

```
my.product(x=12,y=-2)
my.product(y=-2,x=12)
my.product(12,-2)
my.product(-2,12)
```

```
[1] -24
[1] -24
[1] -24
[1] -24
```

The first two calls reflect better programming practices. The last of those is acceptable because multiplication is commutative, but it is risky to play with the arguments this way.

For instance, consider another simple function `my.quotient()`:

```
my.quotient <- function (x,y) {
    x/y
}
```

We call `my.quotient()` on `x=12` and `y=-2`.

```
my.quotient(x=12,y=-2)
my.quotient(y=-2,x=12)
my.quotient(12,-2)
```

```
[1] -6
[1] -6
[1] -6
```

but

```
my.quotient(-2,12)
```

```
[1] -0.1666667
```

When the parameters are not specified in the function call, their implied order reverts to the declared order in the definition (1st = $x$, 2nd = $y$).

And what might we expect to happen with this call?

```
my.quotient(12,0)
```

```
[1] Inf
```

### 1.3.3 Exploring Data

`R` is good tool for data exploration. Let us examine the `swiss` dataset in detail.

We start by displaying the first few rows of the dataset (3, in this case):

```
head(swiss,3)
```

```
            Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
```

We could also display the last few entries (6, here):

```
tail(swiss,6)
```

```
            Fertility Agriculture Examination Education Catholic Infant.Mortality
Neuchatel         64.4        17.6          35        32    16.92             23.0
Val de Ruz        77.6        37.6          15         7     4.97             20.0
ValdeTravers      67.6        18.7          25         7     8.65             19.5
V. De Geneve      35.0         1.2          37        53    42.34             18.0
Rive Droite       44.7        46.6          16        29    50.43             18.2
Rive Gauche       42.8        27.7          22        29    58.33             19.3
```

We can also get an idea as to the dataset's structure with `str()`:

```
str(swiss)
```

```
'data.frame':   47 obs. of  6 variables:
 $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
 $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
 $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
 $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
 $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
 $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

We can extract the column names with the function `colnames()`:

```
colnames(swiss)
```

```
[1] "Fertility"        "Agriculture"      "Examination"      "Education"
[5] "Catholic"         "Infant.Mortality"
```

or display a specific column of the data frame, say `Education`, with the `$` operator:

```
swiss$Education
```

```
 [1] 12  9  5  7 15  7  7  8  7 13  6 12  7 12  5  2  8 28 20  9 10  3 12  6  1
[26]  8  3 10 19  8  2  6  2  6  3  9  3 13 12 11 13 32  7  7 53 29 29
```

This cannot be done with a matrix, however – the following code will provide an error message:

```
swiss_matrix$Education
```

```
Error in swiss_matrix$Education :
  $ operator is invalid for atomic vectors
```

To extract the `Education` column from a matrix, identify its column index and use this, instead:

```
swiss_matrix[,4]
```

```
  Courtelary     Delemont Franches-Mnt      Moutier   Neuveville
          12            9            5            7           15
   Porrentruy       Broye        Glane      Gruyere       Sarine
           7            7            8            7           13
   ...
     Le Locle    Neuchatel   Val de Ruz ValdeTravers V. De Geneve
           13           32            7            7           53
  Rive Droite  Rive Gauche
           29           29
```

Just as one would expect from the behaviour of `colnames()`, `rownames()` extracts the data frame's row names:

```
rownames(swiss)
```

```
 [1] "Courtelary"   "Delemont"     "Franches-Mnt" "Moutier"
      ...
[46] "Rive Droite"  "Rive Gauche"
```

The summary statistics (5-pt summary + mean + number of missing variables for numerical variables; frequency table for others) can be obtained for all data frame's variables simultaneously:

```
summary(swiss)
```

```
   Fertility      Agriculture     Examination      Education
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
    Catholic       Infant.Mortality
 Min.   :  2.150   Min.   :10.80
 1st Qu.:  5.195   1st Qu.:18.15
 Median : 15.140   Median :20.00
 Mean   : 41.144   Mean   :19.94
 3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :100.000   Max.   :26.60
```

More in-depth statistics are available with `psych`'s `describe()`:

```
psych::describe(swiss)
```

```
                vars  n mean   sd   med trim  mad  min   max range skew kurt   se
Fertility          1 47 70.1 12.4  70.4 70.6 10.2 35.0  92.5 57.5 -0.46  0.2 1.82
Agriculture        2 47 50.6 22.7  54.1 51.1 23.8  1.2  89.7 88.5 -0.32 -0.8 3.31
Examination        3 47 16.4  7.9  16.0 16.0  7.4  3.0  37.0 34.0  0.45 -0.1 1.16
Education          4 47 10.9  9.6   8.0  9.3  5.9  1.0  53.0 52.0  2.27  6.1 1.40
Catholic           5 47 41.1 41.7  15.1 39.1 18.6  2.1 100.0 97.8  0.48 -1.6 6.08
Infant.Mortality   6 47 19.9  2.9  20.0 19.9  2.8 10.8  26.6 15.8 -0.33  0.7 0.42
```

The correlation matrix is obtained pretty much as one would expect:

```
cor(swiss)
```

```
                   F    A   Ex   Ed     C  IM
Fertility        1.0  0.3 -0.6 -0.6  0.4  0.4
Agriculture      0.3  1.0 -0.6 -0.6  0.4 -0.0
Examination     -0.6 -0.6  1.0  0.6 -0.5 -0.1
Education       -0.6 -0.6  0.6  1.0 -0.1 -0.0
Catholic         0.4  0.4 -0.5 -0.1  1.0  0.1
Infant.Mortality 0.4 -0.0 -0.1 -0.0  0.1  1.0
```

We can obtain the data frame's number of rows:

```
nrow(swiss)
```

```
[1] 47
```

or the summary of a single variable:

```
summary(swiss$Fertility)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.00   64.70   70.40   70.14   78.45   92.50
```

We can also find all observations for which a feature takes on a value greater than a certain threshold, say:

```
swiss$Fertility>50
```

```
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[37]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
```

or provide summary information for the logical vector:

```
summary(swiss$Fertility>50)
```

```
   Mode   FALSE    TRUE
logical       3      44
```

```
table(swiss$Fertility>50)
```

```
FALSE  TRUE
    3    44
```

The logical vector can be used as an index: for instance, here is the dataset only for those observations where `Fertility` was greater than 50.

```
swiss[swiss$Fertility>50,]
```

with

```
nrow(swiss[swiss$Fertility>50,])
```

```
[1] 44
```

We could also replace the threshold; for instance, here is the dataset for observations data where `Fertility` is in the top 50%:

```
swiss[swiss$Fertility>median(swiss$Fertility),]
```

```
             Fertility Agriculture Examination Education Catholic
Courtelary        80.2        17.0          15        12     9.96
Delemont          83.1        45.1           6         9    84.84
...
Le Locle          72.7        16.7          22        13    11.22
Val de Ruz        77.6        37.6          15         7     4.97
             Infant.Mortality
Courtelary               22.2
Delemont                 22.2
...
Le Locle                 18.9
Val de Ruz               20.0
```

or, solely the `Fertility` and `Education` variables for observations where `Fertility` is in the top 50%:

```
swiss[swiss$Fertility > median(swiss$Fertility),
      c("Fertility","Education")]
```

```
             Fertility Education
Courtelary        80.2        12
Delemont          83.1         9
...                ...       ...
Le Locle          72.7        13
Val de Ruz        77.6         7
```

or those observations for which `Fertility` was maximal:

```
swiss[swiss$Fertility == max(swiss$Fertility),]
```

```
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Franches-Mnt      92.5        39.7           5         5     93.4  20.2
```

### 1.3.4 A Word About NAs

`NA` values in `R` can create some havoc. Be careful!

To illustrate some of the issues, create a dataset by sampling 100 values (with replacement) among the values $\{1, 2, 3, 4, \text{NA}\}$.[33]

33: Your sample will be different.

```
test = sample(c(1:4,NA),100, replace=TRUE)
```

We can summarize `test` as follows:

```
summary(test)   # 5pt summary + mean + number of NAs
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.500   3.000   2.549   3.500   4.000      29
```

We can read the mean from the output, or we could try to compute it directly, using `mean()`:

```
mean(test)
```

```
[1] NA
```

What is happening? The function `mean()` does not know how it should handle the `NA` values; without further guidance, it elects to throw everything akimbo.

Compare with:

```
mean(test, na.rm=TRUE)
```

```
[1] 2.549296
```

### 1.3.5 Loops and Conditional Statements

R allows for **flow control** through loops and conditional statements:

- `if()` and `ifelse()` – when a condition holds, do thing 1, when it does not, do thing 2;
- `for()` – iterate a procedure for a fixed number of steps;
- `while()` – repeat steps as long as some condition holds.

High-level interpreted languages (like R) are **slower** than low-level and/or compiled languages. To get around this issue, interpreted languages will sometimes hand off[34] some operations to functions written in lower-level languages (like C).

34: "Behind the scenes", so to speak.

In order to take advantage of this, certain programming strategies are recommended when working with list, vectors, arrays, data frames, and so on, namely **vectorized** functions (see the family of `apply()` functions in R). In particular, we try to avoid cycling through each item of a list, and instead use special functions that map a chosen function or operation to every item in the list (in R, this can be done with the `apply` family of functions, among others).

This can run counter to habits gained when learning other languages, in which **for** and **while** loops, for instance, might have been emphasized. Consequently, we elect NOT to introduce such loops at this stage. The syntax is rather intuitive and will be easy to understand when we encounter it in examples.

The `ifelse()` statement is quite powerful and can speed-up and simplify data frame operations, however, and we take the time to illustrate how it can be used.

We can easily create a new `swiss` column determining whether the `Fertility` variable, say, is above a certain threshold (in which case it should take the value 1) or not (0):

```
swiss$threshold <-ifelse(swiss$Fertility>50,1,0)
```

```
 [1] 1 1 1 ... 1 1 1
[45] 0 0 0
```

There will be other opportunities to use these functions; the best way to get the hang of R is to practice and debug.

## 1.4 The `tidyverse`

R is a **functional language**, which means that it uses nested parentheses, which can make code difficult to read.[35]

35: Exhibit A: everything up to now.

### 1.4.1 Pipeline Operator

The **pipeline operator** `|>` (formerly `%>%`) and the `dplyr` package can be used to remedy the situation. Hadley Wickham[36] provided an example to illustrate how it works:

36: See [16] for everything there is to know about pipelines and tidy data.

```
hourly_delay <- filter(
   summarise(
     group_by(
       filter(
         flights,
         !is.na(dep_delay)
       ),
       date, hour
     ),
     delay = mean(dep_delay),
     n = n()
   ),
   n > 10
)
```

Without necessarily knowing how each of the internal functions works, we can still get a sense for what the overall nested structure does, and realize (albeit, with a fair amount of work) that the basic object on which we operate is the `flights` data frame.

The pipeline operator `|>` removes the need for nesting function calls, in favor of passing data from one function to the next:

```
library(dplyr)
hourly_delay <- flights |>
    filter(!is.na(dep_delay)) |>
    group_by(date, hour) |>
    summarise(delay = mean(dep_delay),n = n()) |>
    filter(n > 10)
```

It is now obvious that the flights data frame is the base object, for instance – the **gap** between pseudo-code and "code that runs" is significantly reduced. The beauty of this approach is that the block of code can now be 'read' directly: the `flights` data frame is

1. filtered (to remove missing values of the `dep_delay` variable);
2. grouped by `hours` within `days`;
3. the mean `delay` is calculated within `groups`, and
4. the mean `delay` is returned for those hours with more than `n > 10` flights.

The **pipeline rules** are simple – the object immediately to the left of the pipeline is passed as the first argument to the function immediately to its right:

- `data |> function` is equivalent to `function(data)`
- `data |> function(arg=value)` is equivalent to `function(data, arg=value)`

For instance:

```
library(dplyr)
swiss |> summary()
```

```
  Fertility      Agriculture     Examination      Education
Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
Median :70.40   Median :54.10   Median :16.00   Median : 8.00
Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
   Catholic       Infant.Mortality   threshold
Min.   :  2.150   Min.   :10.80    Min.   :0.0000
1st Qu.:  5.195   1st Qu.:18.15    1st Qu.:1.0000
Median : 15.140   Median :20.00    Median :1.0000
Mean   : 41.144   Mean   :19.94    Mean   :0.9362
3rd Qu.: 93.125   3rd Qu.:21.70    3rd Qu.:1.0000
Max.   :100.000   Max.   :26.60    Max.   :1.0000
```

The magrittr vignette provides additional information on the magrittr package, on which dplyr is based.

### 1.4.2 Tidy Data

The pipeline operator is also compatible with the tidyverse suite of packages, championed by Wickham;[37] cheat sheets are available here .

37: Including the ever popular ggplot2 (see Chapter 12, *ggplot2 Visualizations in R* in [1]).

**Tidy data** has a specific structure:

- each column represents a unique variable;
- each row represents a unique observation;
- each table represents a type of observational unit.

Two tidyr functions are used to reshape tables to a tidy format: gather() and spread() – gather() requires:

- a data frame to reshape;
- a key column (against which to reshape);
- a value column (which will contain the new variable of interest), and
- the indices of the columns that need to be collapsed.

Consider the following dataset:

```
cities <- data.frame(
  city=c("Toronto","Montreal","Vancouver",
         "Ottawa","Calgary","Edmonton",
         "Quebec City","Winnipeg","Hamilton"),
  prov=c("Ontario","Quebec","BC",
    "Ontario","Alberta","Alberta",
    "Quebec","Manitoba","Ontario"),
```

```
  pop.2016=c(6202225,4291732,2642825,
             1488307,1481806,1418118,
             839311,834678,785184),
  pop.2011=c(5928040,4104074,2463431,
    1371576,1392609,1321441,
    806406,783099,747545)
)
cities
```

```
         city     prov pop.2016 pop.2011
1     Toronto  Ontario  6202225  5928040
2    Montreal   Quebec  4291732  4104074
3   Vancouver       BC  2642825  2463431
4      Ottawa  Ontario  1488307  1371576
5     Calgary  Alberta  1481806  1392609
6    Edmonton  Alberta  1418118  1321441
7 Quebec City   Quebec   839311   806406
8    Winnipeg Manitoba   834678   783099
9    Hamilton  Ontario   785184   747545
```

It is not presented in a tidy format, because populations show up in **two** columns. In tidy format, it would instead look like:

```
cities.tidy <- tidyr::gather(cities,"year","population",
                             3:4)
cities.tidy$year <- ifelse(cities.tidy$year=="pop.2016",
                             2016,2011)
cities.tidy
```

```
          city     prov year population
1      Toronto  Ontario 2016    6202225
2     Montreal   Quebec 2016    4291732
3    Vancouver       BC 2016    2642825
4       Ottawa  Ontario 2016    1488307
5      Calgary  Alberta 2016    1481806
6     Edmonton  Alberta 2016    1418118
7  Quebec City   Quebec 2016     839311
8     Winnipeg Manitoba 2016     834678
9     Hamilton  Ontario 2016     785184
10     Toronto  Ontario 2011    5928040
11    Montreal   Quebec 2011    4104074
12   Vancouver       BC 2011    2463431
13      Ottawa  Ontario 2011    1371576
14     Calgary  Alberta 2011    1392609
15    Edmonton  Alberta 2011    1321441
16 Quebec City   Quebec 2011     806406
17    Winnipeg Manitoba 2011     783099
18    Hamilton  Ontario 2011     747545
```

`spread()`, on the other hand, generates multiple columns from two columns; it requires a data frame to reshape; a **key** column, and values in the value column to become new values.

For instance, we could reverse the "tidying" of `cities.tidy` with:

```
cities.back.to.wide <- tidyr::spread(cities.tidy,year,
                                     population)
colnames(cities.back.to.wide) <- c("city","prov",
                                   "pop.2011","pop.2016")
cities.back.to.wide
```

```
         city     prov pop.2011 pop.2016
1     Calgary  Alberta  1392609  1481806
2    Edmonton  Alberta  1321441  1418118
3    Hamilton  Ontario   747545   785184
4    Montreal   Quebec  4104074  4291732
5      Ottawa  Ontario  1371576  1488307
6 Quebec City   Quebec   806406   839311
7     Toronto  Ontario  5928040  6202225
8   Vancouver       BC  2463431  2642825
9    Winnipeg Manitoba   783099   834678
```

Other useful wrangling functions include `separate()` and `unite()`. What do you think these do?[38]

38: How could you find out?

### 1.4.3 The `dplyr` Package

The `dplyr` package provides functions to transform tabular data. Its most useful functions are compatible with the pipeline operator `|>`:

- `select()`: to extract a subset of variables from the data frame;
- `filter()`: to extract a subset of observations from the data frame;
- `arrange()`: to sort the data frame;
- `mutate()`: to create new variables from existing variables;
- `summarise()`: to create so-called pivot tables;
- `group_by()`: ... self-evident?

We will showcase these functions with the help of various examples. Try to guess what the outputs would be before looking at them.[39]

39: We do not explicitly state the `dplyr::xyz` dependency since we already had to load the `dplyr` package to gain access to the pipeline operator `|>`.

```
cities |> select(prov,pop.2016)
```

```
      prov pop.2016
1  Ontario  6202225
2   Quebec  4291732
3       BC  2642825
4  Ontario  1488307
5  Alberta  1481806
6  Alberta  1418118
7   Quebec   839311
8 Manitoba   834678
9  Ontario   785184
```

```
cities |> select(-pop.2016)
```

```
         city     prov pop.2011
1     Toronto  Ontario  5928040
2    Montreal   Quebec  4104074
3   Vancouver       BC  2463431
4      Ottawa  Ontario  1371576
5     Calgary  Alberta  1392609
6    Edmonton  Alberta  1321441
7 Quebec City   Quebec   806406
8    Winnipeg Manitoba   783099
9    Hamilton  Ontario   747545
```

```
cities |> filter(pop.2016>1000000)
```

```
       city    prov pop.2016 pop.2011
1   Toronto Ontario  6202225  5928040
2  Montreal  Quebec  4291732  4104074
3 Vancouver      BC  2642825  2463431
4    Ottawa Ontario  1488307  1371576
5   Calgary Alberta  1481806  1392609
6  Edmonton Alberta  1418118  1321441
```

```
cities |> filter(pop.2016>1000000,
                 prov %in% c("Ontario","Quebec"))
```

```
      city    prov pop.2016 pop.2011
1  Toronto Ontario  6202225  5928040
2 Montreal  Quebec  4291732  4104074
3   Ottawa Ontario  1488307  1371576
```

```
cities |> mutate(pop.increase = pop.2016/pop.2011-1)
```

```
         city     prov pop.2016 pop.2011 pop.increase
1     Toronto  Ontario  6202225  5928040   0.04625222
2    Montreal   Quebec  4291732  4104074   0.04572481
3   Vancouver       BC  2642825  2463431   0.07282282
4      Ottawa  Ontario  1488307  1371576   0.08510721
5     Calgary  Alberta  1481806  1392609   0.06405028
6    Edmonton  Alberta  1418118  1321441   0.07316028
7 Quebec City   Quebec   839311   806406   0.04080451
8    Winnipeg Manitoba   834678   783099   0.06586524
9    Hamilton  Ontario   785184   747545   0.05035015
```

```
cities |> summarise(median.2011=median(pop.2011),
variance.2011=var(pop.2011))
```

```
  median.2011 variance.2011
1     1371576  3.209519e+12
```

```
cities |> summarise(mean.2016=mean(pop.2016),
                   sum.2016=sum(pop.2016), n=n())
```

```
  mean.2016 sum.2016 n
1   2220465 19984186 9
```

```
cities |> arrange(pop.2016)
```

```
         city     prov pop.2016 pop.2011
1    Hamilton  Ontario   785184   747545
2    Winnipeg Manitoba   834678   783099
3 Quebec City   Quebec   839311   806406
4    Edmonton  Alberta  1418118  1321441
5     Calgary  Alberta  1481806  1392609
6      Ottawa  Ontario  1488307  1371576
7   Vancouver       BC  2642825  2463431
8    Montreal   Quebec  4291732  4104074
9     Toronto  Ontario  6202225  5928040
```

```
cities |> arrange(desc(pop.2011))
```

```
         city     prov pop.2016 pop.2011
1     Toronto  Ontario  6202225  5928040
2    Montreal   Quebec  4291732  4104074
3   Vancouver       BC  2642825  2463431
4     Calgary  Alberta  1481806  1392609
5      Ottawa  Ontario  1488307  1371576
6    Edmonton  Alberta  1418118  1321441
7 Quebec City   Quebec   839311   806406
8    Winnipeg Manitoba   834678   783099
9    Hamilton  Ontario   785184   747545
```

```
cities |> arrange(prov,desc(pop.2016))
```

```
         city     prov pop.2016 pop.2011
1     Calgary  Alberta  1481806  1392609
2    Edmonton  Alberta  1418118  1321441
3   Vancouver       BC  2642825  2463431
4    Winnipeg Manitoba   834678   783099
5     Toronto  Ontario  6202225  5928040
6      Ottawa  Ontario  1488307  1371576
7    Hamilton  Ontario   785184   747545
8    Montreal   Quebec  4291732  4104074
9 Quebec City   Quebec   839311   806406
```

```
cities |> group_by(prov) |>
  summarise(mean.2016 = mean(pop.2016))
```

```
# A tibble: 5 × 2
  prov     mean.2016
  <chr>        <dbl>
1 Alberta   1449962
2 BC        2642825
3 Manitoba   834678
4 Ontario   2825239.
5 Quebec    2565522.
```

```
cities |> mutate(pop.increase = pop.2016/pop.2011-1) |>
  select(city, pop.increase) |>
  arrange(desc(pop.increase))
```

```
         city pop.increase
1      Ottawa   0.08510721
2    Edmonton   0.07316028
3   Vancouver   0.07282282
4    Winnipeg   0.06586524
5     Calgary   0.06405028
6    Hamilton   0.05035015
7     Toronto   0.04625222
8    Montreal   0.04572481
9 Quebec City   0.04080451
```

`dplyr` also comes with "database" functionality (`bind_cols()`, `bind_rows()`, `union()`, `intersect()`, `setdiff()`, `left_join()`, `inner_join()`, `semi_join()`, `anti_join()`, etc.).

Do not hesitate to bookmark, consult, and borrow from the excellent [16] (and from the subsequent chapters) for more examples, and to practice, practice, practice: we learn programming by programming.

## 1.5 Basics of `Python`

`Python` is another object-oriented language (OOL). It was created in the early 90's but was not popularized until the 00's. It lends itself to writing structured, easy-to-read computer code.[40]

It is intended to be easier to understand and learn than other OOLs. One of its strength is that it has a **massive** base of open-source modules, which allow programmers to implement very sophisticated functionality simply by making a few function calls (not unlike `R`'s packages).

More information is available from the `Python` Software Foundation, on Stack Exchange (and similar sites), and in reference manuals, such as Jake VanderPlas' A Whirlwind Tour of Python or the `Python` 3 documentation.

40: **Indentation** matters in `Python`: in some of the code boxes of the next two sections, we have been forced to sometimes introduce a carriage return in order for the code to fit the width of the available box – in instances where a new line starts with indentation, it is important to verify if that line is completing code from the previous line, in which case it should be entered as a single line at the prompt.

### 1.5.1 IDE for Python

Anaconda and Jupyter are popular data science Python **integrated development environments** (IDE); Rodeo , Spyder , PyCharm , Ninja (an others) also provide RStudio-like functionality for Python. Installation instructions are available on the respective websites.

### 1.5.2 Introduction to Python

The content of the next two sections is intended to help data analysts get a better sense of how Python could be used for data analysis. They are not designed to teach the ins and outs of Python programming. Instead, they illustrate typical tasks through examples.[41]

41: Note that these examples require Python 3.5 or higher.

**Fundamentals** Let us start with the basics.

**Using Python as a Scientific Calculator** Mathematical expressions can easily be evaluated numerically in Python. For scientific calculations, one should import the math module (package/library) which contains many mathematical functions .

It is important to note that Python also provides facilities for integer arithmetic which will be covered later. In this section, only floating-point calculations are used.

Modules can be imported using the import function.

```
import math
```

We can call pre-compiled functions in a module by prepending the module name (with a period) to the function name: module.function_name() is the Python equivalent of package::function_name() in R.

For instance, there is a cos function in the math module: it is called using math.cos().

We can evaluate $\cos(\sqrt{\pi})$ with:

```
math.cos(math.sqrt(math.pi))
```

```
-0.20029354112337366
```

$\arctan(2^5/3)$ with

```
math.atan(2**5 / 3)
```

```
1.477319545636307
```

and $\ln(1 + e^4)$ with

```
math.log(1 + math.exp(4))
```

```
4.0181499279178094
```

**Using Variables to Hold Intermediate Results** It could be helpful to break complex calculations into smaller steps. Variables can be used to store intermediate results. We will see later how variables are used in algorithmic settings.

For instance, we could break down the evaluation of $\exp(\sin(\sqrt{2}+2))$ into three parts:

- $x = \sqrt{2}$
- $y = \sin(x+2)$
- $z = \exp(y)$

```
x = math.sqrt(2)
y = math.sin(x+2)
z = math.exp(y)
```

In order to display the values taken by the variables, we must call on them separately, as follows:

```
x,y,z
```

```
(1.4142135623731, -0.26925647329403, 0.7639472984402)
```

The variables are saved even when they are not displayed, however.

**Numbers as Formatted Strings** Quite often, we may want to control the way numbers are displayed (this can come in handy when reporting results). For example, we may wish to display no more than 4 decimal places for all real numbers, or we may want to pad numbers with zeros so that they all have a given width.

The following block illustrates a number of ways to obtain **formatted strings** of the number 12.3456789. For more details on the format specification mini-language, please consult the documentation .

Note that a string must be enclosed within double quotes or single quotes. We will discuss general string operations shortly.

```
x = 12.3456789
```

We can format the number as a string of width 10, with 2 decimal places:

```
"{:10.2f}".format(x)
```

```
'     12.35'
```

Or as a string with 4 decimal places:

```
"{:.4f}".format(x)
```

```
'12.3457'
```

or as a zero-padded string of width 5, with no decimal:

```
"{:05.0f}".format(x)
```

```
'00012'
```

**Fixed Decimals** **Floating-point numbers** are usually shunned as they are inherently inexact. For example, we might be bewildered to find out what the following sum amounts to:

```
2.2 + 1.1
```

```
3.3000000000000003
```

The result 3.3000000000000003 is definitely not what we would expect as a sum, namely, 3.3.

The `decimal` module allows us to express decimal numbers *exactly* (see the documentation for more information). Let's look at a few examples of working with `decimal` and `Decimal()`.

We start by defining `x` and `y` as the **fixed decimal** values 1.1 and 1.2, respectively. Note that the numbers must entered as strings.

```
import decimal
x = decimal.Decimal("1.1")
y = decimal.Decimal("2.2")
```

These computations behave as we would expect:

```
print(x+y)
print(y/x)
print(x**decimal.Decimal("3"))
```

```
3.3
2
1.331
```

If we do not enter the numbers as strings, they will be treated as floating-point numbers, and then be converted to a string, leading to unexpected results.

```
x = decimal.Decimal(1.1)
y = decimal.Decimal(2.2)

print(x+y )
```

```
3.3000000000000000266453525910
```

Rounding works as one would expect when variables are correctly declared as fixed decimals:

```
z = decimal.Decimal("3.1416")
round(z, 3)
```

```
Decimal('3.142')
```

Once fixed decimals are used, we must use mathematical functions provided by the `decimal` module in order to stay within that module (unfortunately, trigonmetric functions are not available).

For instance, if:

```
a= decimal.Decimal("0.16")
```

then

```
print(a.sqrt())
print(a.ln())
print(a.log10())
```

```
0.4
-1.832581463748310130367054424
-0.795880017344075219145044421 1
```

The same results could be obtained using the `math` module functions:

```
import math
print(math.sqrt(a))
print(math.log(a))
print(math.log10(a))
```

```
0.4
-1.8325814637483102
-0.7958800173440752
```

**List and Tuples** **Lists** and **tuples** are important objects in `Python` programming. Even though we will be mostly using `numpy` arrays and certain `pandas` objects instead of lists later on, it is useful to learn the basics of lists as some of the concepts are transferrable.

**List Creation** A **list** holds a sequence of objects, who do not all have to be the same type. One way to create a list is to enclose the elements, separated by commas, with square brackets.

Let us illustrate this concept with a simple list containing three objects.

```
x = [3,'a',5.1]
```

We can extract the elements using indices (note that the first element corresponds to index 0, the second to index 1, etc.):

```
x[0]
x[1]
x[2]
```

```
3
'a'
5.1
```

The type of each of the elements can be found using:

```
print(type(x[0]))
print(type(x[1]))
print(type(x[2]))
```

```
<class 'int'>
<class 'str'>
<class 'float'>
```

We can also "multiply" an element and transform it into a longer list:

```
['Ho']*10
```

```
['Ho', 'Ho', 'Ho', 'Ho', 'Ho', 'Ho', 'Ho', 'Ho', 'Ho', 'Ho']
```

or create a list of integers ranging from 0 to $n-1$, or from $a$ to $b-1$:

```
n = 5
list(range(n))

a=3
b=7
list(range(a,b))
```

```
[0, 1, 2, 3, 4]
[3, 4, 5, 6]
```

**Tuples** Tuples are list-like objects, but with the following differences:

- they are defined with parentheses instead of square brackets (sometimes, the parentheses can be omitted);
- they are **immutable** (once created, they cannot be modified).

For instance, if

```
t = (1,'a',4.5)
```

then we can obtain the length of `t` and print its 2nd element using

```
print(len(t))
print(t[1])
```

```
3
a
```

but we cannot change the value of the third element of `t` or append a new value to `t`: both commands in the next block of code are illegal:

```
t[2]=1
t.append(5)
```

although the same command applied to the list `x` would be legal:

```
x[2]=1
x.append(5)
print(x)
```

```
[3, 'a', 1, 5]
```

If we know the dimension of a tuple `t`, we can also use an **extract pattern** to extract the individual components, as the following examples illustrate.

```
t = (1, 'two', 3.0)
fst, snd, trd = t
print(fst, snd, trd )
```

```
two 3.0
```

We could use '_' (place holder) to extract the second component, say.

```
_, s, _ = t
print(s)
```

```
two
```

What do you think is happening on the next page?

```
days = [(0,"Sun"), (1, "Mon"), (2, "Tue"), (3, "Wed"),
        (4, "Thu"), (5, "Fri"), (6, "Sat")]
for n, d in days:
    print(d+" is represented by " + str(n))
```

```
Sun is represented by 0
Mon is represented by 1
Tue is represented by 2
Wed is represented by 3
Thu is represented by 4
Fri is represented by 5
Sat is represented by 6
```

**List Comprehension** **List comprehension** is a powerful way to create lists, based on set notation. Before we get into the technical details, let us look at some examples.

We start by importing solely the function `sqrt()` from the `math` module;[42] we also declare an index list `x`:

42: Doing so means that we will not require the prefix `math.` in order to invoke `sqrt()`.

```
from math import sqrt
x = [1, 4, 9, 16]
print(x)
```

```
[1, 4, 9, 16]
```

We can now build new lists from `x`, such as the list of the squares of the elements of `x`:

```
y = [a**2 for a in x]
print(y)
```

```
[1, 16, 81, 256]
```

the list of the square roots of the elements of `x` greater than 4:

```
z = [sqrt(b) for b in x if (b > 4)]
print(z)
```

```
[3.0, 4.0]
```

or the list of integers from 0 to 9 (equivalent to `range(10)`):

```
u = [ c for c in range(10) ]
print(u)
```

```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
```

The most basic form of list comprehension is `[f(x) for x in l]`, where `l` is a list (or an **iterable**) and `f(x)` is an expression in `x`. It creates a list obtained by applying `f` to each element or iterate in `l`.[43]

An optional **conditional** can also be present, giving the general form `[f(x) for x in l if g(x)]`, for some boolean expression `g` (taking on the values `True` or `False`) where generation of the list elements only applies to elements that satisfy the boolean expression.

Multiple lists or **iterables** can be specified in list comprehension. equal to either 'math' or 'stat'.

```
[(x,y,z) for x in [True, False] for y in range(4,7)
   for z in ['math','stat']]
```

```
[(True, 4, 'math'), (True, 4, 'stat'), (True, 5, 'math'),
 (True, 5, 'stat'), (True, 6, 'math'), (True, 6, 'stat'),
 (False, 4, 'math'), (False, 4, 'stat'), (False, 5, 'math'),
 (False, 5, 'stat'), (False, 6, 'math'), (False, 6, 'stat')]
```

We can mimic list comprehension with the help of **loops** (much less efficient); it is preferable to use the former to generate lists.

**List Operations** We illustrate various other operations that can be performed on **zero-indexed** lists in the following blocks: [44]

- sublisting
- changing values
- sorting values
- appending values
- concatenating lists
- deleting elements

Consider a given list `x`:

```
x = [3,1,7,2,5]
print(x)
```

```
[3, 1, 7, 2, 5]
```

We can find the length of the list or print the sublist from the second element to the fourth element, say:[45]

```
print(len(x))
print(x[1:4])
```

```
5
[1, 7, 2]
```

43: `range` provides an example of an iterable. One way to think of an iterable is that it provides a mechanism for generating a sequence of elements one at a time. The benefit is that `range(100000)`, for example, does not take up much computation time since no actual element is generated until it is iterated over.

44: The first element in the list has index 0.

45: Remember, ordinals start with 0, cardinals with 1.

We could also modify the second element of the list (index 1), say:

```
x[1] = 4
print(x)
```

```
[3, 4, 7, 2, 5]
```

Note that x is now permanently changed;[46] if we want to modify the last entry but we are not sure about the length of the list, for instance, we could use:

46: ... or at least, until it is modified again.

```
x[-1] = 6
print(x)
```

```
[3, 4, 7, 2, 6]
```

If we are looking to change the third last element as well, we could use

```
x[-3] = 1
print(x)
```

```
[3, 4, 1, 2, 6]
```

Finally, we could sort the resulting list:

```
x.sort()
print(x)
```

```
[1, 2, 3, 4, 6]
```

A lot of Python methods are applied using the syntax `object.method()`, in contrast to the typical R syntax that would use `method(object)`; so it is `x.sort()` instead of `sort(x)`.

Let us create another list, this time with booleans:

```
y = [3, True, False]
print(y)
```

```
[3, True, False]
```

We can append a value, say 5, at the end of this list, as follows:

```
y.append(5)
print(y)
```

```
[3, True, False, 5]
```

It is also possible to concatenate lists, using the (somewhat confusing) addition notation:

```
z = x + y
print(z)
```

```
[1, 2, 3, 4, 6, 3, True, False, 5]
```

and delete the last element of this new list:

```
del z[-1] # Delete the last element from z
print(z)
```

```
[1, 2, 3, 4, 6, 3, True, False]
```

or delete a range of elements, say from the 3rd to the 6th, from the resulting list:

```
del z[2:6] # watch out for the indices
print(z)
```

```
[1, 2, True, False]
```

**Flow Control** We will take a brief look at two ways to alter the flow of control in Python: **conditional statements** and **loops**.

**Conditional Statements** Python supports `if-elif-else` statements in various forms.

In the following example, we let x be some random integer between 1 and 12 (using function `randint()` from module `random`) and see how the results are affected.

```
import random
x = random.randint(1,12)
print(x)
```

```
9
```

(which may change from one run to another). Perhaps we want to print the string 'Helloifx' is less than 5, like so:

```
if x < 5:
    print('Hello')
```

We would see nothing here as x is 9 in this run. Perhaps we want to print 'Out of range' if x is less than 5 or greater than 9, and `Within range` otherwise?

```
if x < 5 or x > 9:
    print('Out of range')
else:
    print('Within range')
```

```
Within range
```

Finally, we might want to print 'Small' if x is positive and less than 5; otherwise, print 'Five' if x is 5; otherwise, print 'Six' if x is 6; otherwise, print +:

```
if 0 < x and x < 5:
    print('Small')
elif x == 5:
    print('Five')
elif x == 6:
    print('Six')
else:
    print('+')
```

```
+
```

Run this sequence of blocks a number of times to see the various outcomes.

**Important:** Note that the code block that follows an `if`, `else`, or `elif` statement must be **properly indented**. The custom is to use four spaces for indentation. The following example illustrates the effects of different indentations.

```
x = 4

if x < 5:
    print('Small')
else:
    print('This string will not be printed, because the
          else statement never triggers')
    print('Neither will this, for the same reason')
print('This will be printed no matter what x is, as it
      falls outside the if-else statement block')
```

```
Small
This will be printed no matter what x is, as it falls
   outside the if-else statement block
```

**Loops** Loops are useful for repeatedly executing a statement or a block. We first consider the **for loop**.

Let us start with a simple example: for each value in the list `[1,3,8]`, we print its square.

```
for i in [1,3,8]:
    print(i**2)
```

```
1
9
64
```

We could also compute sums with loops, such as $1 + 2 + \cdots + 8 + 9$:

```
sum = 0
for x in range(1,10):
    sum += x  # add the value of x to sum
print(sum)
```

```
45
```

Or print the first `n` even nonnegative integers

```
n = 5
for n in range(0,n):
    t = 2*n
    print(t)
```

```
0
2
4
6
8
```

If a for loop is used to create a list, it is probably best to rewrite it using list comprehension. The following time comparison (using `%%timeit`) illustrates the contrast when building a list of $100 \times 1000$ items.

Using a loop:

```
l = []
for i in range(100):
    for j in range(1000):
        l.append((i,j))
```

Using list comprehension:

```
l = [ (i,j) for i in range(100) for j in range(1000)]
```

**While loops** are useful for iterating until a certain condition is met. For instance, if we want to print the first 10 even positive integers, separated by a space, we could use the following block:

```
i = 0
while i < 10:  # Repeat the following block until i
               # reaches 10 or greater
    i += 1     # iterated index
    print(2*i, end=' ')
```

```
2 4 6 8 10 12 14 16 18 20
```

Or we could print the 26 lower case English alphabets letters on one line, with no separation:

```
i = 0;
while i < 26:
    print(chr(ord('a')+i), end='')
    i += 1
```

```
abcdefghijklmnopqrstuvwxyz
```

Note that `ord` returns the ordinal for a character; `chr` does the reverse.

**Functions** A **function** is a grouped sequence of code that can be called, such as `cos()` and `print()`. A function can have 0 or more **arguments**: `cos()` takes one argument, whereas `print()` can have up to five (see documentation for details).

**Named Functions** Functions facilitate code re-use. `Python` functions are defined *via* the `def` statement. In the next example, we define a function that returns a pair consisting of the sum and the product of its arguments.

```
def sumprod(x, y):
    return x+y, x*y
```

The parentheses around the tuple are optional in this context. The ouput for $x = 3$ and $y = 4$ can be obtained as follows (once the function is compiled):

```
print(sumprod(3,4))
```

```
(7, 12)
```

Functions can also have default argument values. In the following example, if the second argument is not supplied, it takes on the value 5.

```
def myIntegerList(start, end=5):
    return list(range(start, end+1))
```

Compare the results of the two calls below:

```
print(myIntegerList(2))
print(myIntegerList(7,9))
```

```
[2, 3, 4, 5]
[7, 8, 9]
```

**Anonymous (Lambda) Functions** Another way to define a function is with a **lambda statement**, which is used to define one-line functions.[47]

47: The function is anonymous because it has no name.

Anonymous functions are defined using the one-line notation:

```
lambda variables: output
```

For instance,

```
add = lambda u, v: u + v
multiply = lambda u, v: u*v
```

We can apply a bivariate function `func` to arguments `x` and `y`, in a general context, using:

```
def applyFunc(func, x, y):
    return func(x,y)
```

and apply in specific contexts (rule, inputs) as follows:

```
print(applyFunc(multiply, 3,4))
print(applyFunc(add, 7,20))
```

```
12
27
```

But we do not need to define the function prior to the call. This would also work:

```
print(applyFunc(lambda u, v: u*v, 3,4))
print(applyFunc(lambda u, v: u + v, 7,20))
```

```
12
27
```

**Strings** Text manipulation is an important part of data cleaning. Often, the raw data contains string fields that do not quite follow an expected format. For example, proper nouns could be incorrectly capitalized. Dates could have been entered under different conventions. Fortunately, `Python` offers many tools that make string manipulation rather painless. In this section, we look at some of the commonly-performed operations on strings.

Strings can be defined using single or double quotes; note that `Python` supports unicode strings.

```
a = 'First string'
b = "Second string"
c = '+*'
print(type(a), type(b), type(c))
```

```
<class 'str'> <class 'str'> <class 'str'>
```

We can use the multiplication syntax to define a string made up of identical copies of another string as illustrated below:

```
r1 = a*4
r2 = c*3

print(r1)
print(r2)
```

```
First stringFirst stringFirst stringFirst string
+*+*+*
```

Strings can be concatenated using the addition syntax:

```
d = a + c
e = r2 + a + b

print(d)
print(e)
```

```
First string+*
+*+*+*First stringSecond string
```

The character in position `i` (the **index**) of the string a can be accessed via `a[i]`. Remember that the first character's index is 0.

Negative indices can also be used:`a[-4]` returns the fourth character from the end, say. For instance, we can print the first, seventh, last, and fourth-last characters of `a` using:

```
print(a[0], a[6], a[-1], a[-4])
```

```
F s g r
```

We can obtain a **substring** of a string a using the syntax `a[i:j]` where `i` specifies the starting index and `j-1` the ending index. Note that `a[:j]` is equivalent to `a[0:j]`, and `a[i:]` is the substring starting at index `i` and reaching until the end of `a`.

```
print(a[2:4])
print(a[:3])
print(a[6:])
```

```
rs
Fir
string
```

For a string `x`, `x.split()` **splits** the string into a list of words separated by a space (by default). Note that a contiguous sequence of space characters including newline (`\n`), carriage return (`\r`), and tab (`\t`) is considered as one space.

We can also specify what separating characters to use for the splitting, instead of spaces. For example, `x.split(',')` splits `x` on commas and `x.split('--')` splits it on `--`.

Consider the examples below:

```
print('This is  a  \n\n   long   sentence with
   \r \t weird spaces separating the words.'.split())
```

```
['This', 'is', 'a', 'long', 'sentence', 'with', 'weird', 'spaces', 'separating', 'the', 'words.']
```

```
print('One,two, three ,four'.split(',')) # Note that
   # ' three ' is one of the words after separation.
```

```
['One', 'two', ' three ', 'four']
```

```
print('Five--six--ninety-four'.split('--'))
```

```
['Five', 'six', 'ninety-four']
```

In some case, it is helpful to remove leading and trailing space characters (**whitespace stripping**).

```
s = '  time   '
print(s)
print(s.strip())
```

```
  time
time
```

It is common to combine `strip()` with `split(',')`:

```
cs = 'One   , two,  three  '
print([s.strip() for s in cs.split(',')])
```

```
['One', 'two', 'three']
```

In fact, the `strip()` method can accept a string consisting of all characters to be stripped from anothe string, in any combination. For instance, we can strip any leading and trailing characters contained in `['&','#','-','.','!']` from any string as follows:

```
tostrip = '&#-.!'
t = '###.Hel#lo!?!&-'
print(t.strip(tostrip))
```

```
Hel#lo!?
```

The methods `upper()`, `lower()`, and `title()` are useful for **altering the case** of characters in a string. The following examples showcase their functionality.

```
x = "gArbagE collECtion"
print(x.upper())
print(x.lower())
print(x.title())
```

```
GARBAGE COLLECTION
garbage collection
Garbage Collection
```

The following example illustrates a function that takes a phrase and turns it into an acronym by concatenating the first letters of the words and capitalizing all the letters. Does the code make sense?

```
def acronymize(phrase):
    a = ''                  # start with empty string
    for w in phrase.split(): # iterate through words
        a += w[0]           # pick the first letter of
                            # the words and concatenate
    return a.upper()        # capitalize and return
```

```
acronymize("Be right back"), acronymize("Mr Pat Why?")
```

```
('BRB', 'MPW')
```

It can also be useful to **convert a string** representing a number to a number type, and vice versa. The following examples illustrate how these tasks can be achieved.

```
number = 12.345

s = str(number)
print( s, type(s))

f = float(s)
print(f, type(f))

i = int('345')
print(i, type(i))
```

```
12.345 <class 'str'>
12.345 <class 'float'>
345 <class 'int'>
```

We can also check if a string `t` is a substring of another string `s` via `t in s` (**pattern matching**).

```
t1 = "is"
t2 = "has"

s = "This is my car."

print(t1 in s)
print(t2 in s)
```

```
True
False
```

If we want to obtain the index at which a substring begins, we can use the `find()` method. If the substring is not found, -1 is returned.

```
print(s.find(t1))
print(s.find(t2))
```

```
2
-1
```

We shall revisit `Python` strings when we discuss *Natural Language Processing* (see Chapter 32).

**Dictionaries** A **dictionary** is a data structure for **key-value pairs** (`k:v`). To define a dictionary, simply list the key-value pairs enclosed within braces ({,}), as shown in the following examples.

The simplest dictionary is the one that is empty:

```
d = {}  # This creates an empty dictionary

print(type(d))
```

```
<class 'dict'>
```

A more interesting dictionary could be the one below:

```
days = { 'Sun': 1, 'Mon': 2, 'Tue':3, 'Wed':4, 'Thu':5,
   'Fri':6, 'Sat':7 }
print(type(days))
```

```
<class 'dict'>
```

We can **access** the value for key `k` in dictionary `d` via `d[k]`. Note that an exception will be raised if `d` does not contain the key `k`.

We can check if a key `k` is in a dictionary `d` via `k in d`.

```
print(days['Wed'])
print('Aug' in days)
```

```
4
False
```

We can **add** a new key-value pair `k:v` to a dictionary `d` via `d[k] = v`.

```
d[1]=(1,2)
d[2]= 3.45
d['three']= 'string'
print(d)
```

```
{1: (1, 2), 2: 3.45, 'three': 'string'}
```

Conversely, we can delete key `k` and its associated value from dictionary `d` via del `d[k]`.

```
del d[2]
print(d)
```

```
{1: (1, 2), 'three': 'string'}
```

We can also iterate over the keys in a dictionary using a **for loop**.

```
for key in d:
    print(type(key), type(d[key]))
```

```
<class 'int'> <class 'tuple'>
<class 'str'> <class 'str'>
```

The following code gives the same output

```
for key, value in d.items():
    print(type(key), type(value))
```

```
<class 'int'> <class 'tuple'>
<class 'str'> <class 'str'>
```

### 1.5.3 NumPy and Arrays

NumPy is a Python module that supports numerical computation on multi-dimensional arrays. It comes with many useful mathematical functions.

It is the backbone to the scientific computing library SciPy and data analysis and manipulation library pandas. Even though it is possible to do basic statisical analysis using a comprehensive statistics package without direct manipulation of NumPy arrays, knowledge of NumPy is essential for performing custom operations.

In this section, we get a taste of NumPy arrays of dimension at most two. What is covered only scratches the surface of this powerful library. A handy cheat sheet can be found here .

It is customary to use the alias np when importing the module.

```
import numpy as np
```

**Arrays** Unlike lists, NumPy arrays cannot contain elements of different types. There are various ways to create such arrays.

We can create a 1D array from a list:

```
x = np.array([1,2,3,4])

print(x.shape)
```

```
(4,)
```

shape is the method that returns the array's dimensions. We can create a 2D array from a list of lists:

```
y = np.array([[1,2,3],[4,5,6]])

print(y.shape)
```

```
(2, 3)
```

If some of the elements are not of the "right" type, they are converted automatically:

```
c = np.array(['n','u','m',15])

print(c)
```

```
['n' 'u' 'm' '15']
```

We can also define a `NumPy` array out of a range using the `arange()` function:

```
np.arange(1,5)

print(c)
```

```
array([1, 2, 3, 4])
['n' 'u' 'm' '15']
```

yields the same result as `np.array([1,2,3,4])`, but it is more efficient, from a computational perspective.

We can also obtain special arrays, composed of zeros, or composed of ones, with the functions `zeros()` and `ones()`. Here is a 3x4 2D array of 0s:

```
z = np.zeros([3,4]) # A 3-by-4 array of 0's
print(z.shape)
```

```
(3, 4)
```

and 2x1x3 3D array of 1s:

```
f = np.ones([2,3,4]) # A 2x1x3 3D array of 1's
print(f.ndim)
```

```
3
```

Note the difference between the `shape` and `ndim` methods: the former gives the actual dimensions (number of rows, columns, etc.), the latter, the number of dimensions (axes).

We can also define `NumPy` arrays containing random values; for instance, here is a 1D array of 10 random values sampled from the standard normal distribution, using the function `random.normal()`:

```
r = np.random.normal(size=10)
print(r)
```

```
[-1.10501533 -0.69929125 -0.00882625  1.12738611  0.60354054
  1.50509863  1.07440466 -0.86260135  1.12680367 -0.01988042]
```

**Arithmetic** **Adding** and **subtracting** `NumPy` arrays of the same dimensions works as we would expect. Using `x` and `y` as above, and `x2` as below, we get:

```
w = np.array([-1,-2,-3,-4])
```

```
print(x+w)
```

```
[0 0 0 0]
```

```
print(x-w)
```

```
[2 4 6 8]
```

```
print(y+y)
```

```
[[ 2  4  6]
 [ 8 10 12]]
```

**Multiplication by a scalar** also works as expected:

```
print(2*x)
```

```
[2 4 6 8]
```

However, note that **multiplication** and **division** via $*$ and `/` (resp.) are applied component-wise:

```
print(x*w)
```

```
[ -1  -4  -9 -16]
```

as is **exponentiation**:

```
print(y**3)
```

```
[[  1   8  27]
 [ 64 125 216]]
```

**Broadcasting** allows addition and substraction to be performed between arrays that do not have the same shape. There are rules governing when such operations are valid and what the effects are. Here, we provide two simple examples:

```
x + 3.5
```

```
array([4.5, 5.5, 6.5, 7.5])
```

```
y - 1
```

```
array([[0, 1, 2],
       [3, 4, 5]])
```

Can you determine what broadcasting does from these examples?

**Math Functions** `NumPy` contain some useful methods mapping arrays to a scalar.

For instance, `sum` adds up the elements in the array.

```
x.sum()
```

```
10
```

(the same result could have been obtained with `np.sum(x)`).

The usual statistical descriptions are also available as methos:

```
print(x.std(),x.var(),x.mean())
```

```
1.118033988749895 1.25 2.5
```

`NumPy` also has a collection of mathematical functions that can be applied **component-wise**, such as `abs()` and `exp()`:

```
print(np.abs(r))
```

```
[1.10501533 0.69929125 0.00882625 1.12738611 0.60354054
 1.50509863 1.07440466 0.86260135 1.12680367 0.01988042]
```

```
print(np.exp(y))
```

```
[[  2.71828183   7.3890561   20.08553692]
 [ 54.59815003 148.4131591  403.42879349]]
```

NumPy functions are more efficient when it comes to array computations; they should be used whenever possible.

**Logical Operations** Operations over arrays of boolean values can also be performed efficiently in NumPy.

Let us create a boolean array bx of the same shape as x, with bx[i] = True if and only if x[i] >= 2.5, and a boolean array by of the same shape as y, with by[i] = True if and only if y[i] >= 3.5.

```
bx = x >= 2.5
by = y >= 3.5

print(bx)
print(by)
```

```
[False False  True  True]
[[False False False]
 [ True  True  True]]
```

Comparison of two NumPy arrays of the same shape results in a boolean array, yet again of the same shape. Note that comparison is performed component-wise:

```
x2 = np.array([2,1,3,0])

print(x == x2)
```

```
[False False  True False]
```

Comparisons use the symbols ==, <, and >:

```
print(x > x2)
```

```
[False  True False  True]
```

We can perform **boolean operations** (AND, OR, NEG) on boolean arrays:

```
b = np.array([True, False, True, True])
```

AND is computed using &:

```
b & bx
```

```
array([False, False,  True,  True])
```

OR with |:

```
b | bx
```

```
array([ True, False,  True,  True])
```

NEG with ~:

```
~b
```

```
array([False,  True, False, False])
```

We can also sum over the values of a boolean array (in this case, `True` is interpreted as 1 and `False` as 0):

```
np.sum(b)
```

```
3
```

## 1.6 Python for Data Science

While `Python` remains a bona fide programming language, it is as a data science tool that its popularity has soared. Let us take a look at some of its data functionality.

### 1.6.1 Pandas and Data Frames

The `Pandas` module provides `Python` with an equivalent of `R` data frames. Essentially, it is a two-dimensional tabular data structure in which each column can be of different value types.

In this section, we cover the basics of `Pandas` data frames (and introduce a dataset found in the `Seaborn` module.[48] Comprehensive references for doing data analysis with Python include [14, 9, 7]. The pandas cheat sheet could also prove handy.

48: Which is used for data visualization (see Chapter 18 and [1]).

We start by importing the required modules, with the customary **aliases** `pd` and `sns`:

```
import pandas as pd
import seaborn as sns
```

**Loading Data** There are various ways to obtain data. One way is to use a pre-built sample dataset, such as `titanic` from `seaborn`.

```
titanic = sns.load_dataset("titanic")
type(titanic)
```

```
<class 'pandas.core.frame.DataFrame'>
```

Another way is to read a `csv` file using `pandas.read_csv()`. For instance, if the file `calculus.csv` is in the `data` folder, we would call:

```
calculus = pd.read_csv('data/calculus.csv')
```

The first rows are given using the `head()` method of a `DataFrame` object:

```
titanic.head()
```

```
   survived  pclass     sex   age  ...  deck  embark_town  alive  alone
0         0       3    male  22.0  ...   NaN  Southampton     no  False
...
4         0       3    male  35.0  ...   NaN  Southampton     no   True

[5 rows x 15 columns]
```

We can also look at the last rows using the `tail()` method,[49] such as:

49: The number of observations can also be specified in the `head()` method.

```
calculus.tail(6)
```

```
       ID Sex  Grade    GPA  Year
94  10095   F     69   6.49     1
95  10096   M     99  12.61     1
96  10097   M     40   4.17     2
97  10098   F     66   6.94     1
98  10099   M     83  10.09     1
99  10100   F     52   6.76     2
```

We get a quick **summary** of a `DataFrame` using the `describe()` method:

```
titanic.describe()
```

```
         survived      pclass         age       sibsp       parch        fare
count  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean     0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std      0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min      0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%      0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%      0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%      1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max      1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

We can also obtain a summary of a **subset** of the columns:

```
df1 = titanic[['survived', 'age', 'fare']]
df1.describe()
```

```
         survived         age        fare
count  891.000000  714.000000  891.000000
mean     0.383838   29.699118   32.204208
std      0.486592   14.526497   49.693429
min      0.000000    0.420000    0.000000
25%      0.000000   20.125000    7.910400
50%      0.000000   28.000000   14.454200
75%      1.000000   38.000000   31.000000
max      1.000000   80.000000  512.329200
```

Or specific summary statistics on the full objects or on a specific column:

```
df1.mean()
print()
df1['age'].median()
```

```
survived     0.383838
age         29.699118
fare        32.204208
dtype: float64

28.0
```

**Data Frame Operations** We continue with some basic operations on data frames. We will use another built-in dataset

```
crashes.head()
```

```
   total  speeding  alcohol  ...  ins_premium  ins_losses  abbrev
0   18.8     7.332    5.640  ...       784.55      145.08      AL
1   18.1     7.421    4.525  ...      1053.48      133.93      AK
2   18.6     6.510    5.208  ...       899.47      110.35      AZ
3   22.4     4.032    5.824  ...       827.34      142.39      AR
4   12.0     4.200    3.360  ...       878.41      165.63      CA

[5 rows x 8 columns]
```

New columns can be added to any data frame. In this example, we will generate a new column consisting of strings of the form `Cnnn` where `nnn` is a zero-padded three-digit number so that row 1, 2,... of `crashes` correspond to `C001`, `C002`, ...

```
labels = ['C'+"{:03}".format(i+1) for
                  i in range(crashes.shape[0])]
crashes['label'] = labels
```

```
crashes.head()
```

```
   total  speeding  alcohol  ...  ins_losses  abbrev  label
0   18.8     7.332    5.640  ...      145.08      AL   C001
1   18.1     7.421    4.525  ...      133.93      AK   C002
2   18.6     6.510    5.208  ...      110.35      AZ   C003
3   22.4     4.032    5.824  ...      142.39      AR   C004
4   12.0     4.200    3.360  ...      165.63      CA   C005

[5 rows x 9 columns]
```

Quite often, a particular column in a `csv` file serves as the index column. We can set this column to be an index column via the `set_index()` method:

```
df = crashes.set_index('label')
df.head()
```

```
       total  speeding  alcohol  ...  ins_premium  ins_losses  abbrev
label                            ...
C001    18.8     7.332    5.640  ...       784.55      145.08      AL
C002    18.1     7.421    4.525  ...      1053.48      133.93      AK
C003    18.6     6.510    5.208  ...       899.47      110.35      AZ
C004    22.4     4.032    5.824  ...       827.34      142.39      AR
C005    12.0     4.200    3.360  ...       878.41      165.63      CA

[5 rows x 8 columns]
```

Note that `crashes` is not affected by `set_index()`. To make the change directly to `crashes`, we would need to replace

```
df = crashes.set_index('label')
```

with

```
crashes.set_index('label', inplace=True)
```

We can **subset** a data frame by rows and columns labels via `loc[]`, as in the examples below:

```
df.loc['C010':'C013',['speeding','total']]
```

```
       speeding  total
label
C010      3.759   17.9
C011      2.964   15.6
C012      9.450   17.5
C013      5.508   15.3
```

```
df.loc['C005':'C008',:]
```

```
       total  speeding  alcohol  ...  ins_premium  ins_losses  abbrev
label                            ...
C005    12.0     4.200    3.360  ...       878.41      165.63      CA
C006    13.6     5.032    3.808  ...       835.50      139.91      CO
C007    10.8     4.968    3.888  ...      1068.73      167.02      CT
C008    16.2     6.156    4.860  ...      1137.87      151.48      DE

[4 rows x 8 columns]
```

We can also extract using position values via `iloc[]`.

```
df.iloc[1:5,0:4]
```

```
       total  speeding  alcohol  not_distracted
label
C002    18.1     7.421    4.525          16.290
C003    18.6     6.510    5.208          15.624
C004    22.4     4.032    5.824          21.056
C005    12.0     4.200    3.360          10.920
```

We can **reset** the index in a data frame via the `reset_index()` method. This has the effect of turning `label` into a data column like all other columns in the data frame `df`, for instance:

```
df.reset_index(inplace=True)
df.head()
```

```
  label  total  speeding  ...  ins_premium  ins_losses  abbrev
0  C001   18.8     7.332  ...       784.55      145.08      AL
1  C002   18.1     7.421  ...      1053.48      133.93      AK
2  C003   18.6     6.510  ...       899.47      110.35      AZ
3  C004   22.4     4.032  ...       827.34      142.39      AR
4  C005   12.0     4.200  ...       878.41      165.63      CA

[5 rows x 9 columns]
```

It is possible to use the generator `iterrows` to yield both index and row of a data frame. For instance, the next block of code will print the labels corresponding to the first five rows.

```
for index, row in df[0:5].iterrows():
    print(row['label'])
```

```
C001
C002
C003
C004
C005
```

Columns and rows can be **dropped** from a data frame via the `drop()` method. In the example below, we drop the `label` column from `df` and assign the outcome to `df2` (but note `df` itself is not changed):

```
df2 = df.drop('label', axis=1)
df2.head()
```

```
   total  speeding  alcohol  ...  ins_premium  ins_losses  abbrev
0   18.8     7.332    5.640  ...       784.55      145.08      AL
1   18.1     7.421    4.525  ...      1053.48      133.93      AK
2   18.6     6.510    5.208  ...       899.47      110.35      AZ
3   22.4     4.032    5.824  ...       827.34      142.39      AR
4   12.0     4.200    3.360  ...       878.41      165.63      CA

[5 rows x 8 columns]
```

In contrast, the `total` column is dropped from `df` (and `df` is modified as a result):

```
df.drop('total', axis=1, inplace=True)
df.head()
```

```
  label  speeding  alcohol  ...  ins_premium  ins_losses  abbrev
0  C001     7.332    5.640  ...       784.55      145.08      AL
1  C002     7.421    4.525  ...      1053.48      133.93      AK
2  C003     6.510    5.208  ...       899.47      110.35      AZ
3  C004     4.032    5.824  ...       827.34      142.39      AR
4  C005     4.200    3.360  ...       878.41      165.63      CA

[5 rows x 8 columns]
```

We can **rename** the columns of a data frame via the `rename()` method:

```
df.rename(columns={'label':'case', 'abbrev':'abbr'},
                                    inplace=True)
df.head()
```

```
   case  speeding  alcohol  ...  ins_premium  ins_losses  abbr
0  C001     7.332    5.640  ...       784.55      145.08    AL
1  C002     7.421    4.525  ...      1053.48      133.93    AK
2  C003     6.510    5.208  ...       899.47      110.35    AZ
3  C004     4.032    5.824  ...       827.34      142.39    AR
4  C005     4.200    3.360  ...       878.41      165.63    CA

[5 rows x 8 columns]
```

What would we expect the following chunk of code to do?

```
newColumnNames = {}
for name in list(df):
    newColumnNames[name] = name.capitalize()

df2=df.rename(columns=newColumnNames)
```

Rows can be **filtered** according to a given condition. In the example below, `b` and `d` are Pandas series of booleans related to the `df` data frame:

```
b = df['ins_losses'] > 160
d = df['not_distracted'] < 12
```

If we want to return the rows of `df` for which `ins_losses` is greater than 160 **AND** `not_distracted` ia less than 12, we would simply call:

```
df[b & d]
```

```
    case  speeding  alcohol  ...  ins_premium  ins_losses  abbr
4   C005     4.200    3.360  ...       878.41      165.63    CA
6   C007     4.968    3.888  ...      1068.73      167.02    CT
20  C021     4.250    4.000  ...      1048.78      192.70    MD

[3 rows x 8 columns]
```

To return the rows of `db` for which `ins_losses` is greater than 160 **OR** `abbr` is equal to `AL`, we would call:

```
df[b | (df['abbr'] == 'AL')]
```

```
    case  speeding  alcohol  ...  ins_premium  ins_losses  abbr
0   C001     7.332    5.640  ...       784.55      145.08    AL
4   C005     4.200    3.360  ...       878.41      165.63    CA
6   C007     4.968    3.888  ...      1068.73      167.02    CT
18  C019     7.175    6.765  ...      1281.55      194.78    LA
20  C021     4.250    4.000  ...      1048.78      192.70    MD
36  C037     6.368    5.771  ...       881.51      178.86    OK

[6 rows x 8 columns]
```

### 1.6.2 Data Wrangling

We now take a look at some ways to **combine** and **clean** data frames.

**Merging and Joins** Consider a fictitious test score dataset. There are two sections in the class, contained in `testA.csv` and `testB.csv`. Each row consists of a student ID, a section, and a test mark. The file `gpa.csv` contains information on the students' GPAs and their current year of study.

We start by reading in the two test score files (recall that `pd` is the alias for the `pandas` module).

```
dfA = pd.read_csv('data/testA.csv')
dfB = pd.read_csv('data/testB.csv')
```

The first entries of each sets are shown below:

```
dfA.head()
```

```
      ID Section  Mark
0  10021       A    47
1  10073       A    83
2  10084       A    51
3  10102       A    57
4  10175       A    71
```

```
dfB.head()
```

```
      ID Section  Mark
0  10011       B    97
1  10063       B    63
2  10094       B    71
3  10110       B    77
4  10133       B    81
```

We now read in the GPA information.

```
gpa = pd.read_csv('data/gpa.csv')
gpa.head()
```

```
   Student ID   GPA  Year
0       10011  12.0   3.0
1       10021   NaN   3.0
2       10063   5.6   3.0
3       10073   9.8   3.0
4       10084   6.2   3.0
```

Note that the column title for student ID is different in the test score files and in `gpa.csv`.

We now **concatenate** the two data frames of test scores into a single object using the `pandas` function `concat()`.

```
df = pd.concat([dfA,dfB])
```

We now merge the GPA data frame with this combined test score data frame.

```
df3 = pd.merge(gpa, df, left_on='Student ID', right_on='ID')
df3
```

| | Student ID | GPA | Year | ID | Section | Mark |
|---|---|---|---|---|---|---|
| 0 | 10011 | 12.0 | 3.0 | 10011 | B | 97 |
| 1 | 10021 | NaN | 3.0 | 10021 | A | 47 |
| 2 | 10063 | 5.6 | 3.0 | 10063 | B | 63 |
| 3 | 10073 | 9.8 | 3.0 | 10073 | A | 83 |
| 4 | 10084 | 6.2 | 3.0 | 10084 | A | 51 |
| 5 | 10094 | 8.1 | NaN | 10094 | B | 71 |
| 6 | 10102 | 6.9 | 2.0 | 10102 | A | 57 |
| 7 | 10110 | 8.4 | 2.0 | 10110 | B | 77 |
| 8 | 10133 | 10.4 | 2.0 | 10133 | B | 81 |
| 9 | 10145 | 5.1 | 2.0 | 10145 | B | 41 |
| 10 | 10162 | 7.2 | 2.0 | 10162 | B | 68 |
| 11 | 10175 | 6.9 | 1.0 | 10175 | A | 71 |
| 12 | 10189 | 6.1 | 1.0 | 10189 | B | 68 |
| 13 | 10190 | 11.2 | 1.0 | 10190 | A | 91 |
| 14 | 10199 | NaN | 1.0 | 10199 | A | 56 |

`merge()` performs an **inner join**, but it can also perform **outer joins**.

Let us see what happens when we merge `gpa` with `dfA`.

```
pd.merge(gpa, dfA, left_on='Student ID', right_on='ID',
             how='outer').drop('Student ID', axis=1)
```

| | GPA | Year | ID | Section | Mark |
|---|---|---|---|---|---|
| 0 | 12.0 | 3.0 | NaN | NaN | NaN |
| 1 | NaN | 3.0 | 10021.0 | A | 47.0 |
| 2 | 5.6 | 3.0 | NaN | NaN | NaN |
| 3 | 9.8 | 3.0 | 10073.0 | A | 83.0 |
| 4 | 6.2 | 3.0 | 10084.0 | A | 51.0 |
| 5 | 8.1 | NaN | NaN | NaN | NaN |
| 6 | 6.9 | 2.0 | 10102.0 | A | 57.0 |
| 7 | 8.4 | 2.0 | NaN | NaN | NaN |
| 8 | 10.4 | 2.0 | NaN | NaN | NaN |
| 9 | 5.1 | 2.0 | NaN | NaN | NaN |
| 10 | 7.2 | 2.0 | NaN | NaN | NaN |
| 11 | 6.9 | 1.0 | 10175.0 | A | 71.0 |
| 12 | 6.1 | 1.0 | NaN | NaN | NaN |
| 13 | 11.2 | 1.0 | 10190.0 | A | 91.0 |
| 14 | NaN | 1.0 | 10199.0 | A | 56.0 |

We can see that there is a row for every row in `gpa` and that only those rows for which `Student ID` is present in `dfA` have merged data (what happens if the `.drop('Student ID', axis=1)` is omitted?).

**Data Cleansing** Note that in the merged data frame `df3` (and in `gpa`), there are rows containing `NaN`. If we do not want any rows with such values, we can use the `dropna()` method.

```
df3.dropna()
```

```
    Student ID   GPA  Year     ID Section  Mark
0        10011  12.0   3.0  10011       B    97
2        10063   5.6   3.0  10063       B    63
3        10073   9.8   3.0  10073       A    83
4        10084   6.2   3.0  10084       A    51
6        10102   6.9   2.0  10102       A    57
7        10110   8.4   2.0  10110       B    77
8        10133  10.4   2.0  10133       B    81
9        10145   5.1   2.0  10145       B    41
10       10162   7.2   2.0  10162       B    68
11       10175   6.9   1.0  10175       A    71
12       10189   6.1   1.0  10189       B    68
13       10190  11.2   1.0  10190       A    91
```

We can also drop only the rows with NaN in specific columns. If we do not want to retain observations with `Year==NaN`, we would call:

```
gpa.dropna(subset=['Year'])
```

```
    Student ID   GPA  Year
0        10011  12.0   3.0
1        10021   NaN   3.0
2        10063   5.6   3.0
3        10073   9.8   3.0
4        10084   6.2   3.0
6        10102   6.9   2.0
7        10110   8.4   2.0
8        10133  10.4   2.0
9        10145   5.1   2.0
10       10162   7.2   2.0
11       10175   6.9   1.0
12       10189   6.1   1.0
13       10190  11.2   1.0
14       10199   NaN   1.0
```

Instead of dropping rows containing `NaN`, we could replace the unwanted values with some other chosen value instead (like 0, say).

```
gpa.fillna(0)
```

```
    Student ID   GPA  Year
0        10011  12.0   3.0
1        10021   0.0   3.0
2        10063   5.6   3.0
3        10073   9.8   3.0
4        10084   6.2   3.0
5        10094   8.1   0.0
6        10102   6.9   2.0
```

```
7        10110   8.4   2.0
8        10133  10.4   2.0
9        10145   5.1   2.0
10       10162   7.2   2.0
11       10175   6.9   1.0
12       10189   6.1   1.0
13       10190  11.2   1.0
14       10199   0.0   1.0
```

Note that all the `NaNs` are changed to 0.0. To change only the `GPA` volume, we can do the following (note that this will modify the original `gpa` data frame):

```
gpa.fillna({'GPA':0.0})
```

```
    Student ID   GPA  Year
0        10011  12.0   3.0
1        10021   0.0   3.0
2        10063   5.6   3.0
3        10073   9.8   3.0
4        10084   6.2   3.0
5        10094   8.1   NaN
6        10102   6.9   2.0
7        10110   8.4   2.0
8        10133  10.4   2.0
9        10145   5.1   2.0
10       10162   7.2   2.0
11       10175   6.9   1.0
12       10189   6.1   1.0
13       10190  11.2   1.0
14       10199   0.0   1.0
```

We can **apply** a function to a data frame column using the method `map()`. The following will add a `Grade` column to `dfA`, containing `Pass` or `Fail` based on the `Mark` column.

```
def markToGrade(x):
    res = 'Fail'
    if x >= 50:
        res = 'Pass'
    return res
dfA['Grade'] = dfA['Mark'].map(markToGrade)
dfA
```

```
      ID Section  Mark Grade
0  10021       A    47  Fail
1  10073       A    83  Pass
2  10084       A    51  Pass
3  10102       A    57  Pass
4  10175       A    71  Pass
5  10190       A    91  Pass
6  10199       A    56  Pass
```

### 1.6.3 Data Aggregation

Sometimes, the data in a dataset can be divided into **groups**. We might want to obtain **summary statistics** for each group. Analyses by groups and **aggregation** can help us obtain insight on groups.

**Summaries by Groups** We first illustrate obtaining simple statistics on groups using a dataset containing calculus marks (recall that `pd` is the `pandas` alias).

```
calc = pd.read_csv('data/calculus.csv')
calc.head()
```

```
      ID Sex  Grade   GPA  Year
0  10001   F     47  5.02     2
1  10002   M     57  3.82     1
2  10003   M     91  7.70     1
3  10004   M     71  4.82     1
4  10005   F     83  7.91     1
```

Suppose that we want to see separate mean grades and mean GPA based on the `Sex` variables. We can use the `groupby()` method to perform the task:

```
calc[['Sex','Grade','GPA']].groupby('Sex').mean()
```

```
         Grade       GPA
Sex
F    67.901961  6.539804
M    64.408163  5.609388
```

If we want descriptive statistics for `Grade` and `GPA` grouped by `Sex`, we can use the more general method `agg()`. Note that we first need to import `numpy` (alias `np`) to access these simple statistics functions.

```
calc[['Sex','Grade','GPA']].groupby('Sex').agg([np.mean,
                                    np.std, np.median])
```

```
         Grade                          GPA
          mean        std median      mean       std median
Sex
F    67.901961  20.162594   66.0  6.539804  3.008527   6.24
M    64.408163  16.237711   62.0  5.609388  2.756965   4.77
```

If we are interested in the `Grade` mean and the `GPA` median, grouped by `Sex`, we can use a dictionary to specify which function is applied to which column as follows:

```
calc[['Sex','Grade','GPA']].groupby('Sex').agg({'Grade':
                              np.mean, 'GPA': np.median})
```

```
         Grade   GPA
Sex
F    67.901961  6.24
M    64.408163  4.77
```

We can also build **custom aggregate functions**. The following chunk of code computes the sum of squares for the `Grade` and `GPA` columns.

```
def sumOfSq(xs):
    return np.dot(xs,xs)

calc[['Sex','Grade','GPA']].groupby('Sex').agg(sumOfSq)
```

```
         Grade        GPA
Sex
F       255471  2633.7825
M       215928  1906.6374
```

**Pivot Tables** We could also have obtained the mean `Grade` and mean `GPA` for the `Sex` groups via `pivot_table()`, as below:

```
calc[['Sex','Grade','GPA']].pivot_table(index='Sex',
                                    aggfunc=np.mean)
```

```
       GPA     Grade
Sex
F  6.539804 67.901961
M  5.609388 64.408163
```

To obtain a **pivot table** displaying the number of students in each `Year` grouped by `Sex`, we can run the following code:

```
calc[['Sex','Year']].pivot_table(index='Sex',
   columns=['Year'],aggfunc=len, margins=False)
```

```
Year  1  2  3  4
Sex
F    33  11   6   1
M    32  11   2   4
```

We can also print the margins (totals) by changing to `margins=True`.

### 1.6.4 Combining Python with R

Ask most data scientist and they will tell you that they are a Python person or a R person (or perhaps less frequently a Julia person). Python might be best for **data processing** (in terms of efficiency, especially with large datasets), while R has a package (or three!) for pretty much any **statistical and data visualization** task under the sun, but that leaves a lot of data analysis real estate that is not spoken for; frankly, it makes much more sense to be conversant with both.[50]

It is now possible to use Python within R through the reticulate package.[51] . The reticulate vignette contains detailed information on the process; for the time being, we will only give a small example detailing how this could be achieved, based on [17].

50: And anything else that comes up from this point onward.

51: There are other means, see R Interface to Python and Five ways to work seamlessly between R and Python in the same project for more information), for instance

```
library(reticulate)
```

We start by creating a variable x in the Python session:

```
x = list(range(8))
```

Once that is done, we can access the Python variable x from R; it is a column in the (reserved) py data frame:

```
str(py)
py$x
```

```
Module(__main__)
[1] 0 1 2 3 4 5 6 7
```

We can also create new variables y in the Python session from R, and pass a data frame to y:

```
py$y <- head(AirPassengers) # a built-in R dataset
```

This variable can now be displayed in the the Python session, and operated on, as needed:

```
print(y)
```

```
[112.0, 118.0, 132.0, 129.0, 121.0, 135.0]
```

It is not difficult to imagine how to expand this back and forth to more complex data analysis situations, leaving us the option of picking whatever language is best suited to a specific task.

## 1.7 Getting Started with SQL

**Structured Query Language** (SQL) is the standard language used to retrieve, modify, and add data to a **relational database**. It is implemented by all *Relational Database Management Systems* (RDMS), such as:

- MySQL [10]
- MS Access
- Oracle
- Postgres
- etc.

SQL allows users to **query** a database and manipulate the stored data using a variety of parameters. SQL code can be **embedded** into other languages in order to enable storage and processing of large datasets in an efficient manner.

The toy database with which we will work is "implemented" in Aidan Crowther's github repository . Video instructions can be found at DUDADS – How to access the toy database (04:27) | A. Crowther .[52]

52: You will need to install git, docker, and MySQL Client, and know how to open a port on Windows, MacOS, or Linux (search online if necessary).

### 1.7.1 Basics

**Table Structure** The most common form of data organization in a **relational database** is known as a **table** – it is similar to a spreadsheet. Data is stored in a **record** (row), with individual observations aligned by **fields** (columns).

**Records and Fields** Rows consist of data that fall into the categories specified by each column and that either match the **field data type** or contain a `NULL` value,[53] the **absence of data** – it is not the same as a value of zero or an empty string; `NULL` can be matched to any data type.[54]

53: Similar to R's `NA`.

54: SQL syntax often uses `ALL CAPS` in its queries to make it easier to distinguish between commands and data.

**Constraints** Data can be further restricted by Table or Field **constraints**. These constraints define rules by which the data must abide. Most commonly, these constraints are used to identify **special fields** by which data can be uniquely identified, or to ensure data matches a pattern, such as being unique, or not allowing `NULL` entries.

Here are some of common constraints (and their meanings).

- `DEFAULT`: provides a predefined default value if none is specified
- `NOT NULL`: enforces that columns can not have a `NULL` value
- `UNIQUE`: ensures that all values in a column are different
- `PRIMARY KEY`: uniquely identifies a record within a table
- `FOREIGN KEY`: uniquely identifies a record in another table
- `CHECK`: ensures all data in a field matches a restriction
- `INDEX`: used to quickly retrieve and add data to a table

Notably, **primary** and **foreign** keys allow users to create relations between tables. In addition, every table must contain **no more than one** primary key; although they do not need to be defined with a primary key, doing so is considered **bad practice**.

**Data Integrity** Data entered into a table must follow some ensuring the latter's **integrity**. The following rules exist in every **Database Management System** (DBMS).

- **Entity Integrity:** there must not be any duplicate records within a table;
- **Domain Integrity:** enforces valid entries for all fields, following restrictions on data type, format, or range;
- **Referential Integrity:** rows used by other records can not be deleted.

Essentially, we cannot enter records that can cause a table to stop being able to uniquely identify and collect data. In addition, relations between tables must never be broken through the deletion of data.

### 1.7.2 SQL Syntax

The fundamental SQL unit is the **query**, a way to manipulate and output observations from a database by following a specific set of rules.

Generally, queries are used to request data from **tables** kept within a database, but they can also be used to **modify**, **remove**, and **add** data.

The "**sentence structure**" of a SQL query is a repeated pattern of a **command** followed by a **descriptor**; the end of a query being denoted by a **semicolon** (;).[55] More information on SQL (including its syntax) is available in [10, 13, 15].

55: SQL queries read rather naturally as regular English sentences, too.

We will illustrate the various SQL query parameters with the help of a toy database with 4 tables, whose structure is shown in Figure 1.6.

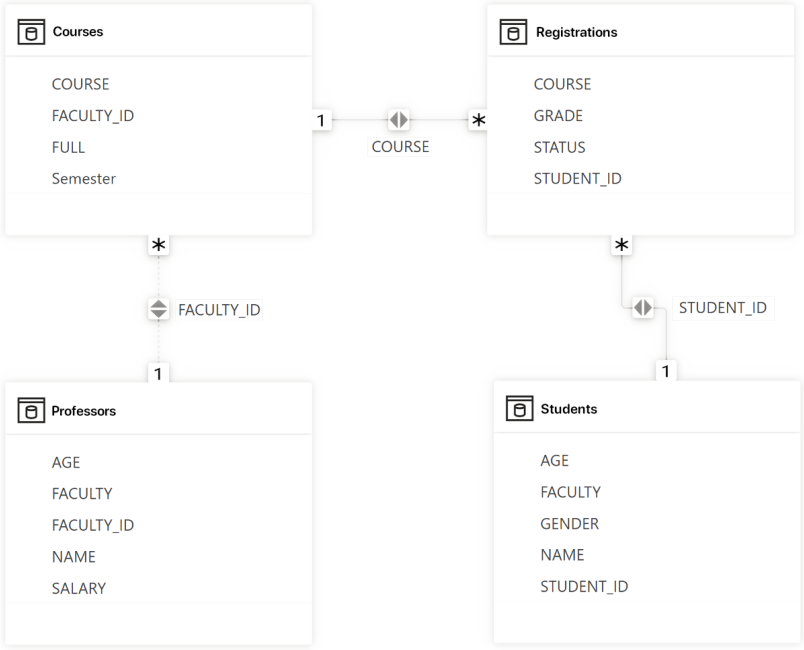



**Figure 1.6:** Database diagram for the toy example, with 4 tables. Some of the entries for 2 of the tables are shown in the Exercises. The data is also available as an Excel spreadsheet.

**Example** What would the following toy dataset query return?

**A Simple SQL Query**

```
SELECT COURSE FROM Courses WHERE FACULTY_ID=1;
```

We break down the query into its command/descriptor structure.

- `SELECT COURSE`: display only the `COURSE` identifier;
- `FROM Courses`: of the observations from the `Courses` table;
- `WHERE FACULTY_ID=4`: for which `FACULTY_ID` is 4.

This query would fetch all courses taught by the instructor #4:

```
  COURSES
1 CGSC101
2 CGSC202
```

We see that this is indeed the case in the `Courses` table:

```
   COURSE FULL Semester FACULTY_ID
1 BUSI202    1   SUMMER          6
2 CGSC101    1   SUMMER          4 <-- *
3 CGSC202    1   WINTER          4 <-- *
4 CHEM404    0   WINTER          8
5 COMP490    1     FALL          9
6 ECON101    1     FALL          1
7 ECON401    0   WINTER          1
8 MUSI101    0   SUMMER         NA
9 PHYS201    0   WINTER          2
```

### 1.7.3 Key Query Operators

#### SELECT/FROM

The `SELECT` command is nearly always used to interact with data; it is used to **request** data from a table. It is applied to **columns**, which need to be specified, using a **comma-separated list** of columns immediately after the `SELECT` **keyword**.[56] The **wildcard** character (`*`) can be used to match **all** columns.

`SELECT` also needs to be told from **which table** to retrieve data; this is accomplished with the `FROM` keyword, after columns have been specified in the query. `FROM` cannot be used without an **argument**, but only one table can be used as input.

The simplest form of a `SELECT` query takes the following form, returning all data within a table.[57]

```
SELECT * FROM Courses;
```

56: Spelling, including the case, matters.

57: In this case, the Courses table. The table is typically clear from the context.

(The output was provided at the end of the previous Section).

The `SELECT` command also allows **aggregate functions** (statistics) to be applied to the selected table columns, including `COUNT`, `SUM`, `AVG`, `MIN`, `MAX`; and more. All rows matching the field being modified will be combined into one unless combined with the `GROUP BY` clause.[58]

58: Not unlike in an Excel pivot table.

Multiple fields can be matched with aggregate functions, and multiple aggregate functions can be used in a query. This can be a useful workaround if a SQL server has **quota restrictions** on the number of queries that can be submitted, allowing multiple fields to be returned with one query.

```
SELECT AVG(SALARY), MAX(AGE) FROM Professors;
```

```
  AVG(SALARY) MAX(AGE)
1    210555.6       67
```

Evidently, the oldest professor is 67 years old, and the average salary is $210,555.60.[59]

59: Whoa! They're making a killing out there...

#### WHERE

In SQL, some queries contain **modifiers** that narrow the query **scope**. The most prevalent one of these clauses is `WHERE`. This clause is often seen used with the `SELECT` query, but can also be used to **specify targets** for other queries such as `UPDATE` and `DELETE`.

`WHERE` allows users to specify **constraints** to apply to the database **prior** to returning the results of a query. These constraints typically use **comparison operators**, such as: `>`, `<`, `=`, `NOT`, `LIKE`, `IS`, etc...

Constraints based on numerical values behave as expected, but their behaviour might be unexpected however when operating on a **strings**. Consequently, we recommend consulting the appropriate documentation in the specific database software manual.

We can determine whether a value is `NULL` by using the `IS` conditional clause to match for `NULL` type.

```
SELECT NAME FROM Professors WHERE SALARY >= 60000;
```

```
             NAME
1      Adam Smith
2     Paige Ryans
3        Alex Doe
4      Landon Liu
5 Kyra Carmichael
6    Heather Wong
7   Quine Ngyogne
8     Vikram  Das
9    Samuel Koffi
```

#### AND/OR/NOT

Clauses, such as `WHERE`, can be **chained** with other constraints in order to conduct complex queries on a database.

We can dive in further within a result when using a `WHERE` clause by combining conditions using the `AND`, `OR`, and `NOT` clauses, **Boolean operators** linking query conditions:

- `AND` returns results where **all** conditions are true;
- `OR` returns results where **at least one** condition is true, and
- `NOT` returns results where the next condition is false.

These clauses can further be organized into brackets.

```
SELECT * FROM Professors WHERE
    (SALARY>=60000 AND NOT AGE>60) OR FACULTY IS NULL;
```

```
             NAME SALARY FACULTY_ID AGE           FACULTY
1     Paige Ryans 180000          2  48           Physics
2        Alex Doe 190000          3  37              <NA>
3      Landon Liu 120000          4  34 Cognitive Science
4    Marcel Orosz     NA          5  48              <NA>
5 Kyra Carmichael 200000          6  30          Business
6    Heather Wong 200000          7  34         Economics
7   Quine Ngyogne 115000          8  55         Chemistry
8     Vikram  Das 500000          9  60  Computer Science
9    Samuel Koffi 300000         10  40 Political Science
```

#### EXISTS

The `EXISTS` keyword is used determine whether a sub-query returns any rows; it returns true if the sub-query returns at least one row; and false otherwise. It is often used in correlated sub-queries.

A **correlated sub-query** is a query that depends on values from the **outer query**; it is executed for each row of the outer query, and the results are used to **filter** the outer query (often based on some condition in the sub-query).

The syntax for a correlated sub-query is similar to a regular sub-query, but it includes a reference to the outer table in the sub-query.

```
SELECT * FROM Professors WHERE EXISTS
    (SELECT * FROM Courses WHERE
        Professors.FACULTY_ID = Courses.FACULTY_ID);
```

```
             NAME SALARY FACULTY_ID AGE           FACULTY
1      Adam Smith  90000          1  67         Economics
2     Paige Ryans 180000          2  48           Physics
3      Landon Liu 120000          4  34 Cognitive Science
4 Kyra Carmichael 200000          6  30          Business
5   Quine Ngyogne 115000          8  55         Chemistry
6     Vikram  Das 500000          9  60  Computer Science
```

The correlated sub-query identifies professors currently assigned to courses; the outer query returns the list of details for those professors.

#### HAVING/GROUP BY

The `GROUP BY` clause is used to aggregate data across multiple rows based on one or more fields.[60] It is used to group data and perform calculations on these groups. The aggregate functions include `COUNT`, `SUM`, `AVG`, `MIN`, `MAX`, etc...

60: Again, quite reminiscent of Excel pivot tables.

We can also use the `HAVING` clause to narrow grouped data further, allowing for the selection only of those results matching a supplementary set of criteria.

```
SELECT AGE, AVG(SALARY) AS AVG_SALARY FROM Professors
    GROUP BY AGE HAVING AVG(SALARY)>90000;
```

```
  AGE AVG_SALARY
1  48     180000
2  37     190000
3  34     160000
4  30     200000
5  55     115000
6  60     500000
7  40     300000
```

#### IN/BETWEEN

In addition to the use of **Boolean conditionals**, SQL has the ability to match **multiple distinct cases**, either by constraining results to a narrow value of cases specified by a **list**, or by matching within a **continuous range**.

`IN` allows a set of possible matching values to be specified – any condition contained **within this set** evaluates to true. We can also use the result of another query to specify the contents of this set *via* a `SELECT` query when specifying the set against which to match.

`BETWEEN` evaluates to true when a compared value falls strictly **within the bounds** specified by the query. This comparison is performed **inclusively**; it can also be used to match to an **alphabetically** sorted list of strings.

```
SELECT * FROM Professors WHERE FACULTY_ID IN (1, 2)
    AND NAME BETWEEN "Adam Smith" AND "Alex Doe";
```

```
        NAME SALARY FACULTY_ID AGE   FACULTY
1 Adam Smith  90000          1  67 Economics
```

#### LIMIT/ORDER BY

Some tables store a **large** number of rows, and can overwhelm a receiver; in these cases restricting the number of returned results can be **crucial**. This can be accomplished by using the `LIMIT` command, which when followed by a numerical value $n$, returns only the first $n$ results from the query.[61]

`ORDER BY` is another powerful clause, especially when used in conjunction with the `LIMIT/TOP` clause – it sorts the result set returned by the query, allowing users to specify **sorting columns** (and **directions**: `ASC` and `DESC`).[62]

61: This command can vary according to the SQL server in use – in some systems, the command is instead `TOP`.

62: This works on numerical values and strings.

```
SELECT * FROM Professors ORDER BY NAME ASC LIMIT 4;
```

```
             NAME SALARY FACULTY_ID AGE   FACULTY
1      Adam Smith  90000          1  67 Economics
2        Alex Doe 190000          3  37      <NA>
3    Heather Wong 200000          7  34 Economics
4 Kyra Carmichael 200000          6  30  Business
```

#### DISTINCT

When one of the fields being used to return results contains a large number of **duplicate** values, the `DISTINCT` clause can help narrow the returned data; multiple fields can be marked as **distinct**, which can be useful when searching for **unique** matches after performing a `JOIN` operation.

```
SELECT DISTINCT NAME From Professors;
```

```
             NAME
1       Adam Smith
2      Paige Ryans
3         Alex Doe
4       Landon Liu
5     Marcel Orosz
6  Kyra Carmichael
7     Heather Wong
8    Quine Ngyogne
9      Vikram  Das
10    Samuel Koffi
```

#### LIKE

The `LIKE` keyword is used in a `WHERE` clause to **search** for a specified pattern in a string column. It is used with the `%` and `_` wildcard characters, to match any **string** or any **single character**, respectively. The pattern provided for matching must be enclosed within **quotes**.

```
SELECT * FROM Courses WHERE COURSE LIKE 'ECON%';
```

```
   COURSE FULL Semester FACULTY_ID
1 ECON101    1     FALL          1
2 ECON401    0   WINTER          1
```

#### UNION

A **union** in SQL is a set operation which combines the result sets of two or more `SELECT` statements into a single result set.

The `UNION` command will combine the output of multiple `SELECT` queries, with a few restrictions:

- the same number of columns must be selected in all queries;
- the same data type must be used for all selections;
- the result must have the same order.

To include all rows, including duplicates, the `UNION ALL` operator can be used instead of `UNION`.

A union can be used for a wide range of purposes, such as combining data from multiple tables, aggregating data from different sources, and generating reports that require data from multiple queries.

```
SELECT NAME AS RESULTS FROM Professors WHERE FACULTY_ID=1
    UNION SELECT COURSE FROM Courses WHERE FACULTY_ID=1;
```

```
     RESULTS
1 Adam Smith
2    ECON101
3    ECON401
```

Note that a `UNION` will combine all matching results into the same column. This may require careful formatting of the selection ordering when matching multiple columns.

#### JOIN

A crucial concept of SQL is that of **combining tables** virtually in order to match related data between tables. One approach to doing so is using the `JOIN` command, which allows users to combine multiple tables into a **single virtual table** by matching like data between the two.

Multiple types of `JOIN` can be performed:

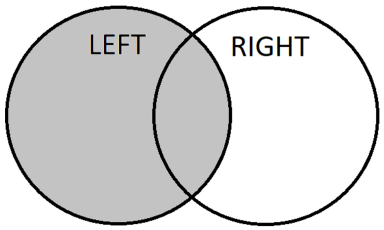
- `LEFT JOIN`
- `RIGHT JOIN`
- `INNER JOIN`
- `FULL JOIN`
- `EXCLUSIVE JOIN`

These different forms of JOIN allow data selection to be narrowed to various ranges, based on the **order** in which the tables are joined and the **type** of join used.

### LEFT JOIN

`LEFT JOIN` is a type of join operation that combines rows from two tables based on the chosen **matching condition(s)**, as well as any **unmatched** rows from the **left table**; i.e., the **first** specified table after the `FROM` clause.[63]

63: The `LEFT JOIN` is illustrated below:




The resulting table will contain all of the rows from the left table, along with any matching rows from the right table. If a row does not have a match in the **right** table, it contains only NULL values.

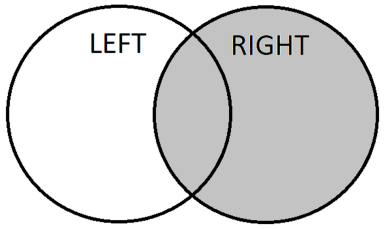
```
SELECT * FROM Professors LEFT JOIN Courses
    ON Courses.FACULTY_ID=Professors.FACULTY_ID;
```

This query will create a list of all professor-assigned-to-course matches, while also listing professors that do not teach any courses.

| | NAME | SALARY | FACULTY_ID | AGE | FACULTY | COURSE | FULL | Semester | FACULTY_ID |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Adam Smith | 90000 | 1 | 67 | Economics | ECON101 | 1 | FALL | 1 |
| 2 | Adam Smith | 90000 | 1 | 67 | Economics | ECON401 | 0 | WINTER | 1 |
| 3 | Paige Ryans | 180000 | 2 | 48 | Physics | PHYS201 | 0 | WINTER | 2 |
| 4 | Alex Doe | 190000 | 3 | 37 | <NA> | <NA> | NA | <NA> | NA |
| 5 | Landon Liu | 120000 | 4 | 34 | Cognitive Science | CGSC101 | 1 | SUMMER | 4 |
| 6 | Landon Liu | 120000 | 4 | 34 | Cognitive Science | CGSC202 | 1 | WINTER | 4 |
| 7 | Marcel Orosz | NA | 5 | 48 | <NA> | <NA> | NA | <NA> | NA |
| 8 | Kyra Carmichael | 200000 | 6 | 30 | Business | BUSI202 | 1 | SUMMER | 6 |
| 9 | Heather Wong | 200000 | 7 | 34 | Economics | <NA> | NA | <NA> | NA |
| 10 | Quine Ngyogne | 115000 | 8 | 55 | Chemistry | CHEM404 | 0 | WINTER | 8 |
| 11 | Vikram Das | 500000 | 9 | 60 | Computer Science | COMP490 | 1 | FALL | 9 |
| 12 | Samuel Koffi | 300000 | 10 | 40 | Political Science | <NA> | NA | <NA> | NA |

### RIGHT JOIN

64: The `RIGHT JOIN` is illustrated below:




`RIGHT JOIN` is identical to `LEFT JOIN`, except that the primary table in this case is the **second** ("right") table appearing after the `FROM` clause.[64] Generally, a `RIGHT JOIN` and a `LEFT JOIN` can be used **interchangeably** by altering the order in which tables are selected.

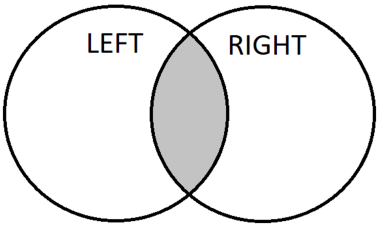
```
SELECT * FROM Professors RIGHT JOIN Courses
    ON Courses.FACULTY_ID=Professors.FACULTY_ID;
```

| | NAME | SALARY | FACULTY_ID | AGE | FACULTY | COURSE | FULL | Semester | FACULTY_ID |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Kyra Carmichael | 200000 | 6 | 30 | Business | BUSI202 | 1 | SUMMER | 6 |
| 2 | Landon Liu | 120000 | 4 | 34 | Cognitive Science | CGSC101 | 1 | SUMMER | 4 |
| 3 | Landon Liu | 120000 | 4 | 34 | Cognitive Science | CGSC202 | 1 | WINTER | 4 |
| 4 | Quine Ngyogne | 115000 | 8 | 55 | Chemistry | CHEM404 | 0 | WINTER | 8 |
| 5 | Vikram Das | 500000 | 9 | 60 | Computer Science | COMP490 | 1 | FALL | 9 |
| 6 | Adam Smith | 90000 | 1 | 67 | Economics | ECON101 | 1 | FALL | 1 |
| 7 | Adam Smith | 90000 | 1 | 67 | Economics | ECON401 | 0 | WINTER | 1 |
| 8 | <NA> | NA | NA | NA | <NA> | MUSI101 | 0 | SUMMER | NA |
| 9 | Paige Ryans | 180000 | 2 | 48 | Physics | PHYS201 | 0 | WINTER | 2 |

#### INNER JOIN

`INNER JOIN` is a type of join operation that combines rows from two tables based on the chosen matching condition(s), **omitting any unmatched rows**; the resulting table will contain only rows where both left and right tables meet the match criteria, all unmatched rows will be dropped.[65]

65: The `INNER JOIN` is illustrated below:




```
SELECT * FROM Professors INNER JOIN Courses
    ON Courses.FACULTY_ID=Professors.FACULTY_ID;
```

This query will provide a list of only those records for which there ia professor and a course match.

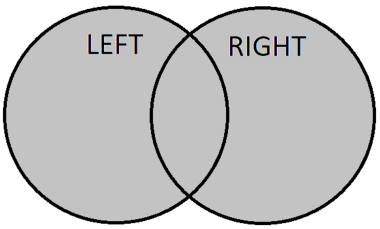
```
             NAME SALARY FACULTY_ID AGE           FACULTY  COURSE FULL Semester FACULTY_ID
1 Kyra Carmichael 200000          6  30          Business BUSI202    1   SUMMER          6
2      Landon Liu 120000          4  34 Cognitive Science CGSC101    1   SUMMER          4
3      Landon Liu 120000          4  34 Cognitive Science CGSC202    1   WINTER          4
4   Quine Ngyogne 115000          8  55         Chemistry CHEM404    0   WINTER          8
5     Vikram  Das 500000          9  60  Computer Science COMP490    1     FALL          9
6      Adam Smith  90000          1  67         Economics ECON101    1     FALL          1
7      Adam Smith  90000          1  67         Economics ECON401    0   WINTER          1
8     Paige Ryans 180000          2  48           Physics PHYS201    0   WINTER          2
```

#### FULL JOIN

`FULL JOIN` returns **all** rows based on the matching condition(s), including the **rows from both right and left tables**, replacing missing values with `NULL`; the input rows of both tables will be present in the output.

MySQL does not inherently support the `FULL JOIN` as this function is largely "syntactic sugar"; we can emulate it using `UNION` in conjunction with a `LEFT JOIN` and `RIGHT JOIN`.[66]

66: The `FULL JOIN` is illustrated below:



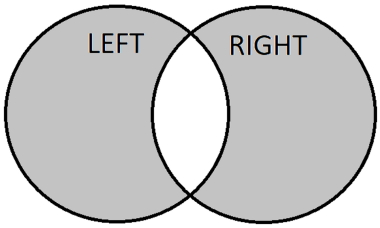

```
SELECT * FROM Courses LEFT JOIN Professors
    ON Courses.FACULTY_ID=Professors.FACULTY_ID
    UNION SELECT * FROM Courses RIGHT JOIN Professors
    ON Courses.FACULTY_ID=Professors.FACULTY_ID;
```

```
    COURSE FULL Semester FACULTY_ID            NAME SALARY FACULTY_ID AGE           FACULTY
1  BUSI202    1   SUMMER          6 Kyra Carmichael 200000          6  30          Business
2  CGSC101    1   SUMMER          4      Landon Liu 120000          4  34 Cognitive Science
3  CGSC202    1   WINTER          4      Landon Liu 120000          4  34 Cognitive Science
4  CHEM404    0   WINTER          8   Quine Ngyogne 115000          8  55         Chemistry
5  COMP490    1     FALL          9     Vikram  Das 500000          9  60  Computer Science
6  ECON101    1     FALL          1      Adam Smith  90000          1  67         Economics
7  ECON401    0   WINTER          1      Adam Smith  90000          1  67         Economics
8  MUSI101    0   SUMMER         NA            <NA>     NA         NA  NA             <NA>
9  PHYS201    0   WINTER          2     Paige Ryans 180000          2  48           Physics
10    <NA>   NA     <NA>         NA        Alex Doe 190000          3  37             <NA>
11    <NA>   NA     <NA>         NA    Marcel Orosz     NA          5  48             <NA>
12    <NA>   NA     <NA>         NA    Heather Wong 200000          7  34         Economics
13    <NA>   NA     <NA>         NA    Samuel Koffi 300000         10  40 Political Science
```

#### EXCLUSIVE JOIN

An EXCLUSIVE JOIN is a **syntactic concept**; the WHERE clause is appended to a JOIN command specifying to only return rows from the left table if **no matching data exists** in the right table. This modification effectively only return results that are **unique to each table**, but otherwise operate exactly as before.[67]

67: The EXCLUSIVE JOIN is illustrated below:




```
SELECT * FROM Courses RIGHT JOIN Professors
    ON Courses.FACULTY_ID=Professors.FACULTY_ID
    WHERE Courses.FACULTY_ID IS NULL;
```

This query will return a list of all professors not teaching a course.

```
  COURSE FULL Semester FACULTY_ID         NAME SALARY FACULTY_ID AGE           FACULTY
1   <NA>   NA     <NA>         NA     Alex Doe 190000          3  37              <NA>
2   <NA>   NA     <NA>         NA Marcel Orosz     NA          5  48              <NA>
3   <NA>   NA     <NA>         NA Heather Wong 200000          7  34         Economics
4   <NA>   NA     <NA>         NA Samuel Koffi 300000         10  40 Political Science
```

### 1.7.4 Examples

**A Representative SQL Query** Typical SQL queries tend to be more complicated than the few examples we have seen so far. The following example can be seen as representative of the level of sophistication/complexity we might encounter.[68]

68: We display the SQL keywords in lower case to make it easier to parse the query; the table and variable names have to be spelled exactly as they appear in the database, however. In practice, it might be a better idea to store the database variables and table names in lower case or camel case, and retain ALL CAPS for the SQL keywords. But you do you.

```
select NAME from
    (Professors left join Courses
    on Professors.FACULTY_ID=Courses.FACULTY_ID)
inner join
    (select COURSE, sum(STATUS in ('DNF', 'FAILED'))
        as Failing_Students
    from Registrations
    where STATUS in ('DNF', 'FAILED')
    group by COURSE order by Failing_Students desc limit 2)
as T on Courses.COURSE=T.COURSE;
```

```
            NAME
1      Adam Smith
2 Kyra Carmichael
```

It can be easier to understand a query if it is broken down from the innermost sub table.

1. We start by noting that we work on the Registrations table, and select only rows that contain a STATUS value of DNF or FAILED.

```
from Registrations
    where STATUS in ('DNF', 'FAILED')
```

2. In the sub-query, we select two fields: the `COURSE` field is returned as it appears in the data, and the count of instances where `STATUS` is `DNF` or `FAILED` (using the aggregation function `SUM`), which was saved as `Failing_Students`, now available to the outer query.

```
select COURSE, sum(STATUS in ('DNF', 'FAILED'))
    as Failing_Students
```

3. The sub-query groups the output by the `COURSE` field, ordered by the count of `Failing_Students` in each course, but limited to the two largest instances.

```
group by COURSE order by Failing_Students
    desc limit 2
```

4. We can now go to the primary query, in which the Professors table is joined to the Courses table to create a mapping of professors to the courses they teach.

```
(Professors left join Courses
    on Professors.FACULTY_ID=Courses.FACULTY_ID)
```

5. The sub-query is assigned the table identifier `T`, which is inner joined with the primary query table, returning a table with the information of the two professors with the most "failing" students.

```
inner join
...
as T on Courses.COURSE=T.COURSE;
```

6. Finally, the resultant rows are isolated and only the `NAME` field is outputted, ultimately returning the names of the two professor with the most failing students.

```
select NAME from
...
```

**SQL in `R`** It will not come as a surprise, especially after the reticulate detour of Section 1.6.4, that we can write SQL queries in `R`, with the appropriate library.[69]

**SQL in R**

```
# install required library
library(RMySQL)

# connect to the database
mysqlcon = dbConnect(RMySQL::MySQL(),
    dbname='school', host='ayyws.com', port=3000,
    user='Ruser', password='Ruser')
```

```
[1] "Courses"       "Professors"    "Registrations" "Students"
```

69: The `dbname`, `host`, `port`, `user`, and `password` arguments are those of a test server where the toy example database can be accessed. For obvious reasons, this is a read-only situation. Just as obviously, the arguments would be different when working with a real database; contact your DBA (database admin) and consult the video linked to at the start of this section for more information and troubleshooting.

```
# test the connection by listing all tables
dbListTables(mysqlcon)

# submit a query to the database
x = dbSendQuery(mysqlcon, "select * from Registrations;")

# convert the result to an R data frame, and display
data.frame = fetch(x)
print(data.frame)
```

```
  STUDENT_ID  COURSE GRADE     STATUS
1        100 ECON401    NA Registered
2        100 ECON101 10.00     Passed
3        101 ECON101  2.45     Failed
4        102 BUSI202    NA Registered
5        102 ECON101    NA        DNF
6        104 CHEM404    NA Registered
7        104 COMP490  9.80     Passed
8        101 BUSI202  3.52     Failed
```

## 1.8 Exercises

1. Write pseudo-code that will sort a list of numbers. Identify the inputs and the outputs, and solve the problem "procedurally" on a definite example before generalizing to a general list. You may need to "black box" the manipulation of individual numbers and group of numbers within the list.
2. Write pseudo-code that will enumerate all strings of up to `n` characters taken from the set `A-Z`, with no repeated character. Identify the inputs and the outputs, and solve the problem "procedurally" on a definite example before generalizing. Use "black boxes" as needed.
3. Use `R` to calculate the following quantities:
   a) The sum of 1.001, 22.9, and -73.78
   b) The square root of 64
   c) Calculate the base 10 logarithm of 90, and multiply the result with the cosine of $\pi$.[70]
4. Type the following `R` code, which assigns numbers to objects `x`, `y`.

   ```
   x<-252
   y<-5.5
   ```

   a) Calculate the product of `x` and `y`
   b) Store the result in a new object called `z`
   c) Inspect your workspace by typing `ls()`, and by clicking the Environment tab in RStudio, and find the three objects you created
   d) Make a vector of the objects 'x', 'y', and 'z'.
5. You have measured seven cylinders. Their lengths are: 2.1, 10.8, 5.5, 6.6, 9.7, 8.2, 8.1, and the diameters are: 0.4, 0.3, 1.2, 0.9, 0.3, 0.2, 0.1. Read these data points into two vectors (give the vectors

70: Hint: see `?log` and `?pi` for information about how to use.

appropriate names). Use R to calculate the volume of each cylinder ($V$ = length $\times \pi \times$ (diameter/2)$^2$).

6. Input the following data, related to space shuttle launch damage prior to the Challenger explosion. The set covers 6 launches out of 24 that were included in the pre-launch charts used to decide whether to proceed with the launch or not

| Temp | Erosion | Blowby | Total |
|---|---|---|---|
| 53 | 3 | 2 | 5 |
| 57 | 1 | 0 | 1 |
| 63 | 1 | 0 | 1 |
| 70 | 1 | 0 | 1 |
| 70 | 1 | 0 | 1 |
| 75 | 0 | 2 | 1 |

Enter these data into a R data frame, with column names `temperature`, `erosion`, `blowby`, and `total`.

7. Read the following data into R (number of honeyeaters seen at a site in a week). Give the resulting data frame a reasonable name. Type it into Excel or text file and save it as a CSV file or txt.

| Day | nbirds | Day | nbirds |
|---|---|---|---|
| Sunday | 3 | Thursday | 8 |
| Monday | 2 | Friday | 1 |
| Tuesday | 5 | Saturday | 2 |
| Wednesday | 0 | | |

Enter the following data as new observations of a different week starting on Sunday: 4, 3, 6, 1, 9, 2, 0.

8. Read the data from the space shuttle launch (from the previous question) data into R.
9. Read the following data set (various Australian populations since 1917) into an R object. Write the object into a text file, from R.

| Year | NSW | Vic. | Qld | SA | WA | Tas. | NT | ACT | Aust. |
|---|---|---|---|---|---|---|---|---|---|
| 1917 | 1904 | 1409 | 683 | 440 | 306 | 193 | 5 | 3 | 4941 |
| 1927 | 2402 | 1727 | 873 | 565 | 392 | 211 | 4 | 8 | 6182 |
| 1937 | 2693 | 1853 | 993 | 589 | 457 | 233 | 6 | 11 | 6836 |
| 1947 | 2985 | 2055 | 1106 | 646 | 502 | 257 | 11 | 17 | 7579 |
| 1957 | 3625 | 2656 | 1413 | 873 | 688 | 326 | 21 | 38 | 9640 |
| 1967 | 4295 | 3274 | 1700 | 1110 | 879 | 375 | 62 | 103 | 11799 |
| 1977 | 5002 | 3837 | 2130 | 1286 | 1204 | 415 | 104 | 214 | 14192 |
| 1987 | 5617 | 4210 | 2675 | 1393 | 1496 | 449 | 158 | 265 | 16264 |
| 1997 | 6274 | 4605 | 3401 | 1480 | 1798 | 474 | 187 | 310 | 18532 |

10. What do you think the following R calls do?

```
swiss$var1 <- swiss[,1]>median(swiss[,1])
swiss$var4 <- swiss[,4]>median(swiss[,4])
```

```
table(swiss$var1); table(swiss$var4)
table(swiss$var1,swiss$var4)
```

11. What do you think the following R calls do?

```
median(test, na.rm=TRUE)
min(test, na.rm=TRUE)
max(test, na.rm=TRUE)
quantile(test, na.rm=TRUE)
```

12. In Python:
    a) evaluate $\lfloor 10001/4 \rfloor$ and $\arcsin(\pi/4)$;
    b) obtain the value of $s$ in the following: $a = \pi(1+\ln 5)$, $b = \frac{1}{3+\sqrt{4}}$ and $s = a + b$;
    c) obtain a formatted string of $\sin(\pi^2)$ of width 8, with 5 decimal places;
    d) turn the value of $\sqrt{3}$ into a fixed decimal with 8 decimal places.

13. In Python:
    a) create a list of integers from -10 to 5;
    b) use list comprehension to create a list `(x,y)` so that `x+y > 8` where `x` can be any nonnegative integer at most 10 and `y` can be any positive integer at most 7;
    c) use list comprehension to create a list `(x,y)` so that `y` is the square of `x` and `x` is from 1 to 10;
    d) write one line of code that returns a list obtained from

```
x = ['one', 2, 3, 'four', 5, 6, 'seven', 8, 9, 10,
          'eleven', 12, 13, 'fourteen']
```

    by moving all the elements of type `str` to the end of the list.[71]

14. Write an if statement in Python that prints "odd" if x is odd and prints "even" if x is even where x is a random integer between -100 and 100, inclusive.[72]

```
import random
x = random.randint(-100,100)
```

15. Use a single while loop in Python to print all pairs `(x,y)` such that `x+y=100` and `x` ranges from 0 to 50.
16. Write a Python function `myFunc()` that returns the square of `x` if `x` is of type `int` and returns `None` otherwise.[73]

```
def myFunc(x):
    res = None
    ## Your code here
    return res
```

Verify that the function behaves as expected:

```
assert(myFunc(5) == 25)
assert(myFunc('five') is None)
```

71: Hint: Use list comprehension and concatenation. To check if `a` is of type `str`, use `type(a) is str`. To check if `a` is not of type `str`, use `type(a) is not str`.

72: Hint: `x % n` returns the remainder of `x` divided by `n`.

73: Hint: `type(x) is int` is the syntax for testing if `x` is of type `int`.

17. Write a function `mySoS()` that accepts a list of floats as the only argument and returns the sum of squares of the numbers (assume that the argument is indeed a list of floats – no need to test if the condition is met).

```
def mySoS(ns):
    res = 0
    ## Your code here
    return res
```

Verify that the function behaves as expected:

```
assert(mySoS([1.0,2.0,3.0]) == 14.0)
assert(mySoS([-2.5,1.3,13.4]) == 187.5)
```

18. What is the result of the following code?

```
def mystery(func, n):
    return [func(i) for i in range(n)]

print(mystery(lambda x: (2*x+1)**2, 5))
```

Rewrite the function using an anonymous function (single line).

19. Complete the definition of the `Python` function `myRep()` with arguments `x`, `y`, and `n` (where `x` and `y` can be assumed to be strings and `n` can be assumed to be a nonnegative integer) that returns the string `x+y` repeated `n` times.

```
def myRep(x, y, n):
    res = ''
    # Your code here

    return res
```

Verify that the function behaves as expected:

```
assert(myRep('a','b',3) == 'ababab')
assert(myRep('Python','C',0) == '')
```

20. Complete the definition of the `Python` function `posOfi()` with argument `s` and returns a list of indices at which `s` contains the letter 'i'.[74]

74: Hint: use the enumerate function.

```
def posOfi(s):
    # Your code here

    return None
```

Verify that the function behaves as expected:

```
print(posOfi("Mississipi"))
print(posOfi("Harry Potter"))
```

21. Complete the following `Python` function which takes a string consisting of a paragraph of sentences ending with a period and returns a list of all the sentences, with leading and trailing spaces stripped. You may assume that every period ends a proper sentence and there are no sentences not ending in a period.

```
def sentences(p):
    # Your code here
    return None
```

Verify that the function behaves as expected:

```
p = 'The essence of Python.  One can sense. But not learn. '
print(sentences(p))
```

22. What effect do the methods `upper()`, `lower()`, and `title()` have on non-alphabetical characters?
23. Complete the following function which takes a list of full names as argument an returns a list of names that are not properly capitalized. For example, for the argument `['John Doe', 'JANE Kelly', 'nicole dunn', 'David Huang']`, the function returns `['JANE Kelly, 'nicole Dunn']`.

```
def badNames(names):
    # Your code here
    return None
```

24. Complete the following function which takes a list `l` of strings as argument and returns a list consisting of the strings in `l` not containing the symbol `-`. For example, given the argument `['Hi', 'Good-bye', 'Ciao', 'Twenty-one']`, the function should return `['Hi', 'Ciao']`.

```
def filterList(l):
    # Your code here
    return None
```

25. Complete the following function which takes a list of pairs as argument and returns a dictionary with the first components as keys and the second components as the corresponding values. For example, given the argument `[(1,'a'),(2,'b')]`, the function returns `{1: 'a', 2: 'b'}`.

```
def pairListToDict(pairs):
    # Your code here
    return None
```

26. Complete the following function which takes a dictionary as argument and removes all the key-value pairs that do not have values of type `str`. For example, calling the function with the dictionary `{'one': 1, 'two': 'Two', 'three': 3}` will change the dictionary to `{'two': 'Two'}`.

```
def filter(d):
    # Your code here
    return
```

27. Complete the following code so that `sq` is a 1D `numpy` array of the squares of the first 100 positive integers. Use list comprehension.

```
sq = np.array([...])
```

28. Obtain a `NumPy` array from the array `sq` in the section by applying the function $\sqrt{x}+1$ to each entry `x` in `sq`.[75]
29. Complete the following definition of `myFunc()` which takes a positive integer argument `n` and a positive real number `d` and generates an array of `n` random values drawn from the standard normal distribution and returns the number of values whose absolute values are less than or equal to `d`. You may assume that `n` is a positive integer and `d` is a non-negative float when `myFunc()` is called.[76]

75: Hint: use broadcasting and `np.sqrt()`.

76: Hint: use `numpy.random.randn()` for generating the random array.

```
def myFunc(n, d):
    # Your code here
    return 0
```

Verify that the function behaves as expected:

```
np.random.seed(5900)
assert(myFunc(10000,1) == 6848)
assert(myFunc(100000,2) == 95490)
```

30. Obtain the `iris` data set through `seaborn` and generate some summary statistics.
31. Write code to change the labels in the data frame `crashes` from `Cnnn` to `Incident nnn` and turn that column into an index column. Commit these changes to `crashes`.
32. Extract a data frame from `df` consisting only of the columns `speeding` and `alcohol` for which the `speeding` values are at least 3.0 and the `alcohol` values are at most 4.5.
33. There is a powerful way to filter rows involving complex boolean expressions via the `query()` method. For instance,

```
df.query("ins_losses > 160 & ins_premium < 900 & abbr == 'CA'")
```

```
   case  speeding  alcohol  ...  ins_premium  ins_losses  abbr
4  C005       4.2     3.36  ...       878.41      165.63    CA

[1 rows x 8 columns]
```

Extract a data frame from `df` *via* `query()` consisting of records for which `alcohol` is at most 4.0 and `abbr` is neither `CA` nor `LA`.

34. Obtain a data frame `df4` by changing the column name of `Student ID` in the data frame `gpa` to `ID`. Then create `df5` by merging `df4` and `df` using `pd.merge(df4, df, on='ID')` and summarize the resulting data frame.
35. Perform an outer join with `df4` from the previous exercise and `dfB`.

36. Drop the observations in the original `gpa` data frame for which the only `NaN` values are found in the `GPA` column.
37. Replace the `NaN` in the original `gpa`'s `Year` column with the string `Unknown`.
38. Modify `markToGrade` so that a mark between 80 to 100 (incusive) is converted to an `A`, a mark at least 70 but less than 80 is converted to a `B`, a mark at least 60 but less than 70 is converted to a `C`, a mark at least 50 but less than 60 is converted to a `D`, and a mark below 50 is converted to an `F`.

```
def markToGrade(x):
    res = 'F'
    # Your code here
    return res
```

Add a Grade column to `df3` containing the converted grades.

39. Obtain the mean for each of the `Year` groups in the `calc` data frame.
40. Obtain the mean, standard deviation, and median for each of the `Year` groups in the `calc` data frame, using `agg()`.
41. Produce a summary of the `calc` data frame giving the `Grade` mean and standard deviation, and the `GPA` median, grouped by `Years`.
42. Complete the definition of a function that returns `Satisfactory` if the average of the array `x` is at least 65.0 and `Unsatisfactory` otherwise.

```
def groupStatus(arr):
    res = ''
    # Your code here
    return res
```

Determine the group status in the `calc` dataset by both `Sex` and `Year`, for the `Grade` variable.

43. Write a function that produces the pivot table displaying the number of students with a passing grade by `Sex` and `Year`.[77]
44. Carry out the remaining exercises in both `R` and `Python`. There is no need to do the exercises in any particular order. Take the time to design pseudo-code and think about what the code does before jumping directly into the programming. You may choose to carry out each of the exercises separately, or to write a single program that carries out all of the individual exercises. You will find much of the base code you need in the chapter's examples, but you may need to tweak and add to this code to carry out the exercises. Do not hesitate to look for information and inspiration on the Internet and in the documentation.

    a) Create three variables and assign numerical values to each of these variables. Then write one or more statements that carry out the following types of operations using these variables: addition, subtraction, multiplication, division, raising to a power.
    b) Create three variables and assign string values to each of these variables. Write a statement that joins the three strings into a single string. Write some code that prints the string. Write

77: Hint: if `arr` is a `NumPy` array, then `arr >= 50.0` gives an array of the same length such that element `i` is `True` if and only if `a[i] >= 50.0`.

some code that tests to see if a substring of your choice is contained within the larger string.

c) Create three variables and assign lists to each of these variables. Join the three lists into a new list containing three distinct sub-lists (a list of three lists). Create a list from this list without sub-lists (all original list elements are part of a single larger list). Create a fourth list by splitting this resulting list in half and assigning the second half of the list to a new variable. Extract the last item of this list (it can either stay in the original list or be removed from it) and assign this element to a variable.

d) Write a statement that contains at least three nested blocks. Use at least three of the following control flow options: if, if else, while, for, break, continue (`Python` only), next, switch.

e) Write a function that takes three arguments as input and returns one value. Call the function with arguments of your choosing.

f) Execute the relevant command that shows a list of the packages (for `R`) or modules (for `Python`) that are currently installed in your environment. Use the available documentation to determine what some of these do. Write some code that uses functions and objects supplied by these packages.

g) Print to the standard output three sentences of your choosing, on three separate lines, using a single statement of code.

h) Locate a comma separated values (`.csv`) file stored on your computer or online. Read this file into the notebook and store the results in one or more variables.

i) Create a new file and write four lines in `.csv` format to this file. In a separate statement, write four more lines to this existing file, without overwriting the original file.

j) Write enough code to generate at least five different error messages. Copy these error messages into a text document, and write a short note under each explaining the meaning of the error message, and how the code was fixed.

k) Using a language of your choice, write a function that, when passed a dataset, reports 5 interesting pieces of information about the dataset. Load a dataset and run the function on this dataset.

l) Using a language of your choice, write two functions. The output of the first function should work as the input to the second function. The first function should read in a dataset and generate a subset of the dataset based on some chosen criteria. The second function should read in a dataset and provide summary data of some type for each column in the dataset. Load a dataset and run both functions on the dataset.

m) Write a program that sorts a list of numbers, without using the in-built sorting functions.

n) Write a program that sorts a list of character strings, without using the in-built sorting functions.

45. Consider a database consisting of two tables, as shown below.

Professors

| NAME | SALARY | FACULTY_ID | AGE | FACULTY |
|---|---|---|---|---|
| Adam Smith | 60000 | 1 | 67 | Economics |
| Paige Ryans | 70000 | 2 | 48 | Physics |
| Alex Doe | 55000 | 3 | 37 | |

Courses

| COURSE | FULL | SEMESTER | FACULTY_ID |
|---|---|---|---|
| ECON101 | True | Fall | 1 |
| PHYS201 | False | Winter | 2 |
| ECON401 | False | Winter | 1 |
| ... | ... | ... | ... |

a) What is the primary key for each table?
b) What are the foreign keys for each table?
c) What are the `NULL` values?
d) What is the relation between these tables?
e) What type of data does each field support?
f) What constraints might we expect each field to have?
g) What would happen if we tried to mix datatypes without enforcing constraints?

## Chapter References


[1] P. Boily, S. Davies, and J. Schellinck. *The Practice of Data Visualization*. Data Action Lab, 2023.

[2] R. Duursma, J. Powell, and G. Stone. *A Learning Guide to R*. Scribd, 2017.

[3] K. Eliason. *Difference Between Object-oriented Programming and Procedural Programming Languages*. 2013.

[4] T. Herman. *How to Write Software With Mathematical Perfection*.

[5] N.J. Horton and K. Kleinman. *Using R and RStudio for Data Management, Statistical Analysis, and Graphics, 2nd Edition*. Taylor & Francis, 2015.

[6] R. Kabacoff. *R in Action*. Second. Manning, 2015.

[7] J. Kazil and K. Jarmul. *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*. O'Reilly, 2016.

[8] J.H. Maindonald. *Using R for Data Analysis and Graphics Introduction, Code and Commentary*. CRAN, 2004.

[9] W. McKinney. *Python for Data Analysis : agile tools for real-world data*. Sebastopol, CA: O'Reilly, 2013.

[10] *MySQL 8.0 Reference Manual*.

[11] R.D. Peng. *R Programming for Data Science*. Lulu.com, 2012.

[12] AV Content Team. *A Complete Tutorial to learn Data Science in R from Scratch*. 2016.

[13] *SQL Tutorial*, *Tutorials Point*.

[14] J. VanderPlas. *Python Data Science Handbook : essential tools for working with data*. Sebastopol, CA: O'Reilly, Inc, 2016.

[15] *SQL Tutorial*, *W3 Schools*.

[16] H. Wickham and G. Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly, Jan. 2017.

[17] Y. Xie, C. Dervieux, and E. Riederer. *R Markdown Cookbook*. ISBN 9780367563837. Boca Raton, Florida: Chapman and Hall/CRC, 2020.

# Multivariate Calculus for Data Analysis

# 2

by **Fabrizio Donzelli**, with contributions from **Patrick Boily**

---

This chapter contains an essential introduction to multivariable calculus. The goal is to provide the readers interested in statistics and/or data science with some basic mathematical tools that are at the base of the algorithms and the mathematical models of statistical analysis. Theoretical details, such as rigorous proofs and definitions, will be kept at the minimal level.

A more detailed and complete introduction to multivariable calculus is found at the YouTube channel Calc with Fab and in [4, 3, 1].


2.1 Points, Vectors, Coordinates 107
One Dimension . . . . . 108
Two and Three Dimensions 108
More Dimensions . . . . . 108
2.2 Functions . . . . . 109
2.3 Graphical Representation . 111
One Variable . . . . . 111
Two Variables . . . . . 111
Three or More Variables . . 114
Scalars and Vector Fields . . 115
2.4 Derivatives . . . . . 116
Difference Quotients . . . . 116
Rules of Differentiation . . 117
Partial Derivatives . . . . . 118
Gradients . . . . . 121
Directional Derivatives . . . 122
2.5 Optimization . . . . . 125
Critical Points . . . . . 125
Local vs. Global . . . . . 127
Local Extrema . . . . . 127
Global Extrema . . . . . 130
Lagrange Multipliers . . . . 132
2.6 Riemann Integrals . . . . . 135
Local Densities, Total Sums 136
One Variable . . . . . 137
Fundamental Theorem . . . 137
Finding Antiderivatives . . 138
Several Variables . . . . . 139
Applications to Statistics . . 140
2.7 Exercises . . . . . 143
Chapter References . . . . . 146


## 2.1 Points, Vectors, Coordinates, Dimensions

We denote by $\mathbb{R}^n$ the **$n$-dimensional (real) space**. A point $P$ in $\mathbb{R}^n$ is located using the **orthogonal Cartesian coordinates** $(x_1, x_2, \cdots, x_n)$.*

This notation may be adapted according to the context. For instance, we will often denote a **specified point** in $\mathbf{R}^n$ by $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, in contrast with the notation $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ which we reserve for a **generic point**. The number $n$ of coordinates is the **dimension** of $\mathbb{R}^n$.

Given two sets $A$ and $B$ (for examples, two regions in $\mathbb{R}^n$) we write $A \subseteq B$ if $A$ **is a subset of** $B$ (that is, $A$ is contained in $B$: every element of $A$ is also in $B$, but the converse is not necessarily true). Let $P = (a_1, \cdots, a_n)$ be a point in $\mathbb{R}^n$, and $D \subseteq \mathbb{R}^n$. We write $P \in \mathbb{R}^n$ if the point **belongs** to the set $D$, **otherwise** we write $P \notin \mathbb{R}^n$.

The real line $\mathbb{R}$ contains **intervals**:

- **closed** $[a, b]$, the set of all $x$ such that $a \leq x \leq b$;
- **open** $(a, b)$, the set of all $x$ such that $a < x < b$;
- **"clopens"** $(a, b]$ $(a < x \leq b)$ and $[a, b)$ $(a \leq x < b)$, and
- **unbounded** $(a, +\infty)$, $(-\infty, a)$, $(-\infty, +\infty)$.

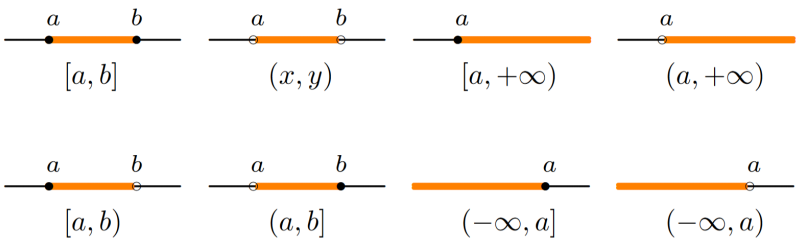



**Figure 2.1:** Intervals on the real line $\mathbb{R}$.

* We assume some familiarity with most of the following notions, but we suggest reading this short section before moving on to the rest of the chapter, as a refresher.

### 2.1.1 One Dimension

The (real) one-dimensional space is denoted by $\mathbb{R}$; it is represented by a **line**, oriented **from left to right** along the directon along which values increase. It is common to denote the position of the points along $\mathbb{R}$ by $x$, but one can choose another name for the variable.[1]

1: We often use $t$ when the real line represents the passage of time, for instance.

The point with coordinate $x = 0$ is known as the **origin** of the line. Positive values of $x$ are located to the **right** of the origin, negative values to the **left**, as in Figure 2.2.


**Figure 2.2:** The real line $\mathbb{R}$, with origin and direction.

### 2.1.2 Two and Three Dimensions

The (real) plane $\mathbb{R}^2$ is two-dimensional; we give it (Cartesian) coordinates $(x, y)$, as shown in Figure 2.3.[2] The four **plane sectors** formed by the coordinate axes (red lines) are the plane's **quadrants**, labeled with Roman numerals in counterclockwise order.

2: As was the case in one-dimensional space, the notation of the coordinates may change according to the context: $(x_1, x_2)$ is also used, for instance, but so are polar coordinates $(r, \theta)$.

For $\mathbb{R}^3$, we typically use the (Cartesian) coordinates $(x, y, z)$ or $(x_1, x_2, x_3)$.[3]

3: Other options: spherical coordinates, cylindrical coordinates.

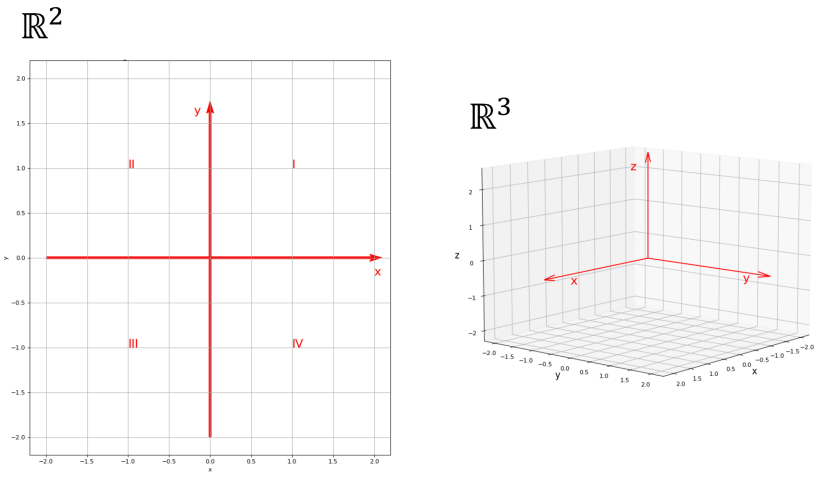



**Figure 2.3:** The real plane $\mathbb{R}^2$, with origin and quadrants (left); the real space $\mathbb{R}^3$ (right).

In general, we do not display the coordinate axes.[4]

4: Unless we do!

### 2.1.3 More Dimensions

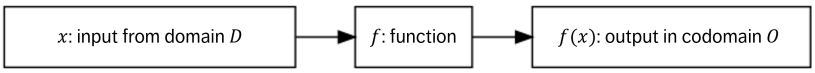
We define the $n$-dimensional (real) space $\mathbb{R}^n$ as the space described by Cartesian coordinates $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. The point $\mathbb{0} = (0, 0, \cdots, 0)$ is the **origin** of $\mathbb{R}^n$, and it is the point of **common intersection** of the $n$ coordinate axes.

In principle, $\mathbb{R}^n$ is not a vector space, but it can be treated as such and so we can perform vector algebra operation with elements of $\mathbb{R}^n$ (see Chapter 3, *Overview of Linear Algebra*).

## 2.2 Functions

Functions are the basic objects of calculus, and are the building blocks of mathematical modelling. Functions are in a general sense **input-output machines**, in the sense of the following general definition, which applies beyond calculus.

If $D$ is a set of input values $O$ is the set of output values, then a **function** $f : D \to O$ is a rule that assigns **to each input** element $x \in I$, a **unique output** value, which we denote by $f(x)$. The notation of the function, the input and output set can vary, as usual, according to the context. Once $f$ has been specified, we refer to $D$ as the **domain** of $f$ and to $O$ as its **codomain**.




If $f : D \to O$ is a function, the set $f(D) = \{f(x) \mid x \in D\} \subseteq O$ is called the **range** (or the **image**) of $f$.

#### Examples

1. Let $P$ be the collection of patients in a COVID emergency hospital, and $O = \{p(\text{ositive}), n(\text{egative})\}$ be the set of possible test responses. We construct the "COVID-TEST" function $T : P \to O$ as follows: If $x \in P$,

$$T(x) = \begin{cases} p, & \text{if patient } x \text{ tests positive} \\ n, & \text{if patient } x \text{ tests negative} \end{cases}$$

   In this example the output values are **categorical**, since they classify the patients into a discrete set of (fixed) classes.[5]

2. Let $S$ denote a sphere of arbitrary radius. A point on $S$ can be located using two coordinates: its **longitude** and its **latitude**.[6] We can then define the temperature function $T : S \to \mathbb{R}$ by

$$T(\text{longitude}, \text{latitude}) = \text{temperature at the point.}$$

   The temperature function is usually assumed to be **continuous**.[7]

3. **Probability theory** is naturally expressed in the language of multivariate calculus (see Chapter 6). For instance, the **density function** of the **multivariate normal distribution** in 2 uncorrelated variables of expectation **0** is a function $f_{\sigma_1,\sigma_2} : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$f_{\sigma_1,\sigma_2}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x^2 + y^2}{2}\right).$$

   The probability that a randomly selected point $P = (x, y)$ from this distribution falls in $\Omega \subseteq \mathbf{R}^2$ is an integral:

$$\iint_\Omega f_{\sigma_1,\sigma_2}(x, y)\, dA.$$

   We will discuss such notions further in Section 2.6, 6.3, and 6.4.

5: In statistics, it is often convenient to represent categorical variables with **numeric** values. For example, we can assign $f(x) = 1$ if the patient $x$ has a positive test, $f(x) = 0$ if their test is negative.

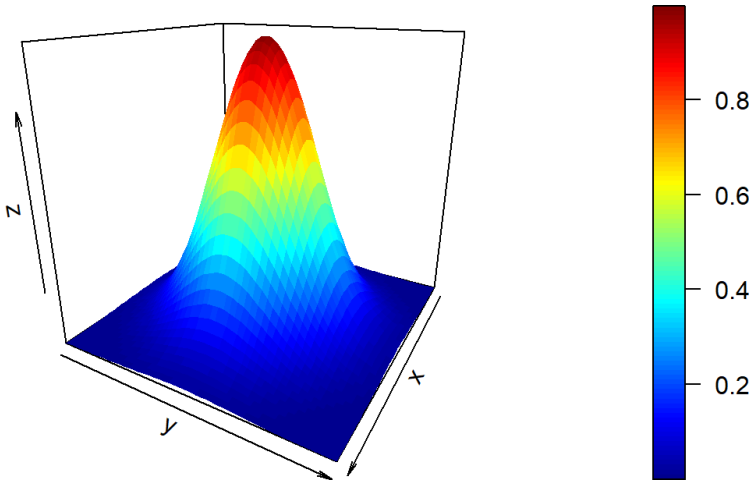
6: Assuming that a special point and great circle through that point have been identified.

7: We will not be discussing this concept except in an intuitive manner: a continuous function is one in which there are no "jumps". An interesting corollary is that if we model the temperature on the Earth in that manner, we can show that at any given moment there are at least two antipodal points which have exactly the same temperature.

4. The following block of R code provides a display of the 3D surface $z = \exp(-x^2 - y^2)$ over $\{(x, y) \in \mathbb{R}^2 \mid -2 \leq x, y \leq 2\}$.

**3D plotting in R**

```
library(plot3D) # for 3D plotting

M <- mesh(seq(-2, 2, length.out = 50),
          seq(-2, 2, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- exp(-x^2-y^2)
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```




**Note:** the domain of a function is part of the recipe, it is not automatically defined by the function itself. However, in calculus, when we use the word **domain**, we usually mean the **largest set** $D_f$ to which the function could be applied. For any $x$ in $D_f$, there is a **unique output** $f(x)$.[8]

8: That is not necessarily the case in the general framework of **multivalued functions**, which, while quite interesting from a geometrical perspective, are outside the scope of this document.

### Examples

1. What is the (largest possible) domain $D_f$ of the function defined by $f(x, y) = \frac{1}{x+y}$? We cannot divide by zero, so the denominator $x + y$ can never be zero when we apply the function $f(x, y)$; $D_f$ therefore consists of all pairs $(x, y)$ expect for those satisfying the equation $x + y = 0$, whose solution set is the line $y = -x$. Thus,

$$D_f = \{(x, y) \in \mathbb{R}^2 \mid x + y \neq 0\};$$

   in other words, the domain consists of the region above the line $y = -x$ and the region below the line $y = -x$.
2. What is the domain $D_f$ of $f(x, y, z, w) = \ln(w) + x + y + z$? Recall that the (real) logarithm is defined only for positive input values. Hence the domain is $D_f = \{(x, y, z, w) \in \mathbb{R}^4 \mid w > 0\}$.

## 2.3 Graphical Representation of Functions

Human eyes (and brains) have a difficult time parsing large data files directly; we typically rely on **graphical representations** to make sense of data (see Chapter 18 and [2] for a *lot* more information on the topic).

Graphical representations are useful in calculus as well; we review a few standard ways of providing these for functions of several variables.

### 2.3.1 One Variable: Sketch the Graph

Let $f : (a, b) \to \mathbb{R}$ be a function of one variable $x$. The **graph** of $f$ is the curve of equation $y = f(x)$; a point in the graph is given by coordinates $(x, f(x))$, for $x \in (a, b)$.

**Example** Sketch the graph of the function $f : [0, \infty) \to \mathbb{R}$ defined by

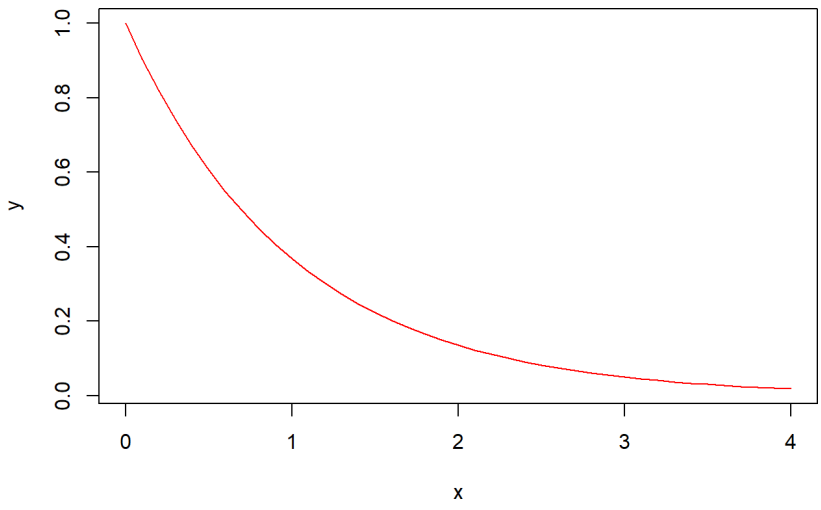
$$f(x) = e^{-x} \text{ for } x \geq 0.$$

Does the point $(1, 2)$ belong to the graph of $f$?[9]

9: This is essentially an example of the **exponential distribution**.

Note that the domain is restricted to the half-real line $x \geq 0$; since the exponent is negative, $e^{-x}$ decays to 0 as $x \to \infty$ (quite rapidly in fact).

```
x <- seq(0,4,0.1)
y <- exp(-x)
plot(x, y, type='l', col = rainbow(25), lty=1)
```




To answer the last question, we evaluate $f(1)$; it is equal to $e^{-1} \neq 2$, and so the point is not on the graph.

### 2.3.2 Two Variables: Graphs or Level Curves

For function of two variables, there are two convenient ways to provide a graphical representation.

### The Graph of a Function

Let $f : D \to \mathbb{R}$ be a function of two variables $x, y$, where $D \subseteq \mathbb{R}^2$. The **graph** of $f$ is the **surface** of equation $z = f(x, y)$.

A point on the graph is given by coordinates $(x, y, f(x, y))$, where $(x, y) \in D$. We can interpret the graph as a **hilly region**, in which case $(x, y)$ are the coordinates of the position with reference to $xy-$plane, and $z$ is the altitude.

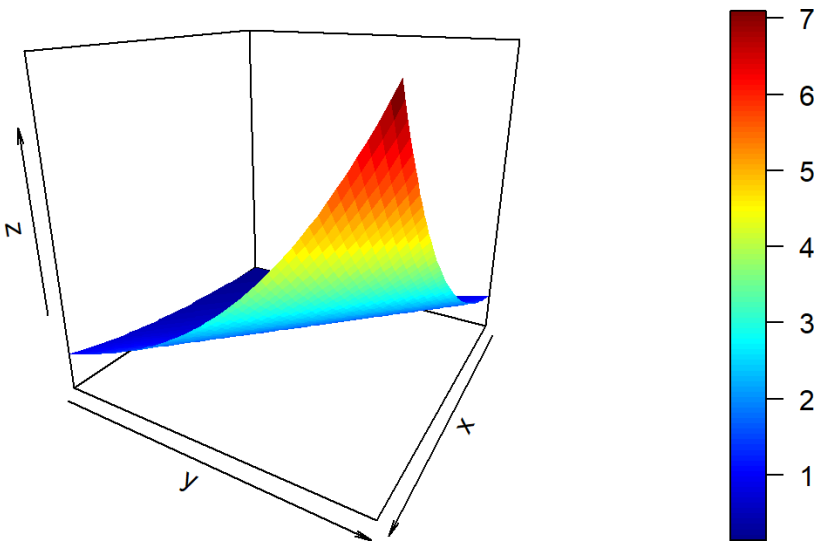
**Example** Sketch the graph of the function $f : D \to \mathbb{R}$ defined by

$$f(x, y) = e^{x+y}, \text{ for } -1 \leq x \leq 1, -1 \leq y \leq 1.$$

Interpret the graph.

We can recycle the code from one of the previous examples.

```
library(plot3D)

M <- mesh(seq(-1, 1, length.out = 50),
          seq(-1, 1, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- exp(x+y)
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```




### Level (Contour) Curves

Let $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$. Depending on the nature of $f$, the graph may be difficult to read (or to plot). An alternative may be to sketch the **level** (or contour) **curves**.

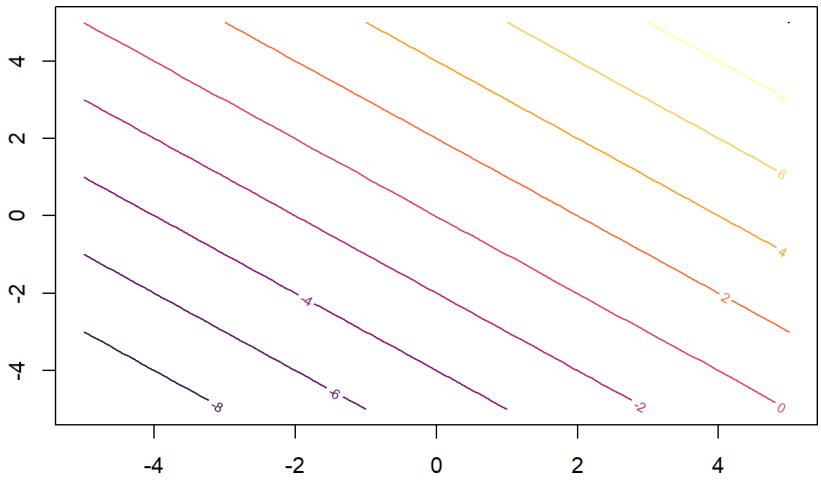
Let $c$ be a value in the range of $f$, which is to say, a **possible output value** of $f$. Generically, the equation $f(x,y) = c$ is a **curve** in the $xy-$plane, a **level curve** (or **contour curve**) of $f$, which consists of **all (and only)** the points $(x,y) \in D$ where the function takes the value $c$.

**Example** Plot a few level curves of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = x + y$.

For any fixed value $c$, the equation $x+y = c$ can be rewritten as $y = -x+c$. The level curves of $f$ are thus all the lines in the $xy-$plane with slope $-1$. Along each line of equation $y = -x + c$, the value of $f$ is given by the $y-$intercept.

Here is a sample code for plotting the level curves of $f$; the numbers displayed on top of the curves are the values $c$ taken by the function along the curves displayed.

```
x <- seq(-5,5,length.out=50)
y <- seq(-5,5,length.out=50)
z <- outer(x,y,"+")
cols <- hcl.colors(10, "Inferno") #color palette
contour(x,y,z,col=cols)
```



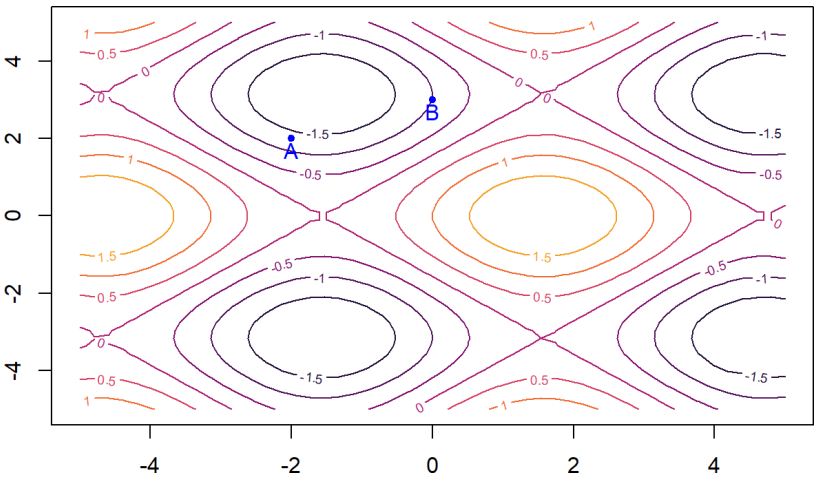

We can use level curves to estimate the values of a function in a certain region of the domain.

**Example** Given the following level curves of $f(x,y) = \sin(x) + \cos(y)$, estimate the value of $f$ at $A$ and $B$.

**Level curves in R**

```
x <- seq(-5,5,length.out=50)
y <- seq(-5,5,length.out=50)
z <- outer(sin(x),cos(y),"+")

cols <- hcl.colors(10, "Inferno") #color palette
```

```
contour(x,y,z, col=cols)
points(-2,2,col='blue',pch=20)
points(0,3,col='blue',pch=20)
points(-2,1.7,col='blue',pch="A")
points(0,2.7,col='blue',pch="B")
```




The point $A$ is located between the level curves $f(x, y) = -1$ and $f(x, y) = -1.5$. Since it is slightly closer to the second curve, we can estimate $f(A) \approx -1.3$.

The point $B$ seems to sit exactly along the level curve $f(x, y) = -1$, hence $f(B) \approx -1$.[10]

10: Of course, we can double check this estimate by finding the coordinates of $A$ and $B$, and computing $f(A)$ and $f(B)$.

**Example** Level curves may **degenerate** to **lower dimensional regions**, or, even "worse", be empty when $c$ is not in the range of $f$.

As an illustration, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2$:

- for $c > 0$, the level curve $x^2 + y^2 = c$ is the circle of center $(0, 0)$ and radius $\sqrt{c}$;
- the level curve $x^2 + y^2 = 0$ degenerates to the point $(0, 0)$, the only point whose coordinates solve the equation $x^2 + y^2 = 0$;
- for $c < 0$, the level curve $x^2 + y^2 = c$ does not exist, since $x^2 + y^2 \geq 0$ for all real values of $x$ and $y$.

### 2.3.3 Three or More Variables

The more variables we have, the more challenging it can be to provide graphical representations of a function.

However both graphs and level sets can be defined, in purely mathematical terms, over an arbitrary number of variables, without needing to be visualized.

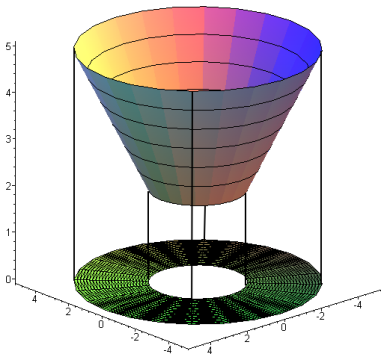
### The Graph of a Function

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function of $n$ variables $\mathbf{x} = (x_1, \cdots, x_n)$. The **graph** of $f$ is the $n-$dimensional **hypersurface** in $\mathbb{R}^{n+1}$ defined by the equation $w = f(\mathbf{x}) = f(x_1, \cdots, x_n)$, for $\mathbf{x} \in D$. A point on the graph is therefore identified by coordinates

$$(\mathbf{x}, f(\mathbf{x})) = (x_1, x_2, \cdots, x_n, f(x_1, x_2, \cdots, x_n)),$$

with $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in D$. We can interpret $f$ as a way of bending and stretching the domain $D$ into a new region embedded in $\mathbb{R}^{n+1}$.[11]

11: We illustrate this for $n = 2$ below:



The cone is a distortion in $\mathbb{R}^3$ of the ring in $\mathbf{R}^2$.

### Level (Contour) Sets

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and let $c$ be a value in the range of $f$. Generically, the equation $f(\mathbf{x}) = f(x_1, \ldots, x_n) = c$ is an $n-1$ dimensional region (**hypersurface**) in $D$, called a **level set** (or **contour set**) of $f$, which consists of **all (and only)** the points $\mathbf{x} = (x_1, \ldots, x_n) \in D$ where the function takes the value $c$.

Level sets may **degenerate** to **lower dimensional regions** $< n-1$, or be empty when $c$ is not in the range of $f$.

**Example** Describe the level sets of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = x^2 + y^2 + z^2$. Are there "degenerate" level sets?

In $\mathbb{R}^3$, the equation of the 2D sphere of radius $R > 0$ and centre at the origin $\mathbf{0} = (0, 0, 0)$ is $x^2 + y^2 + z^2 = R^2$. Thus, the level sets of the functions consists of spheres all centered at the origin.

If $R = 0$, the equation $x^2 + y^2 + z^2 = 0$ is satisfied only for the zero dimensional set $\{(x, y, z) \mid x = y = z = 0\}$; this level set is degenerate.

## 2.3.4 Scalar-Valued Functions and Vector Fields

Let $D \subseteq \mathbb{R}^n$ be a $n$-dimensional domain. A **real valued function** $f : D \to \mathbb{R}$ will be called a **function** (or a **scalar field**), in contrast with a **vector valued function** $\mathbf{F} : D \to \mathbb{R}^n$, which we call a **vector field**.

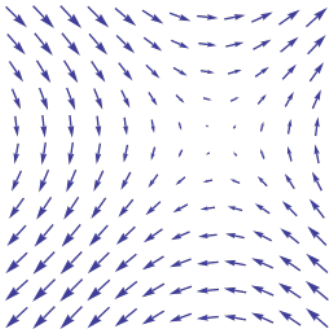


**Figure 2.4:** An illustration of the 2D vector field $\mathbf{F}(x, y) = (\sin y, \sin x)$ [author unknown].

Vector fields play a crucial role in vector calculus and its applications to physics and geometry, but this is out of scope for our purposes. We refer again the reader to [4].

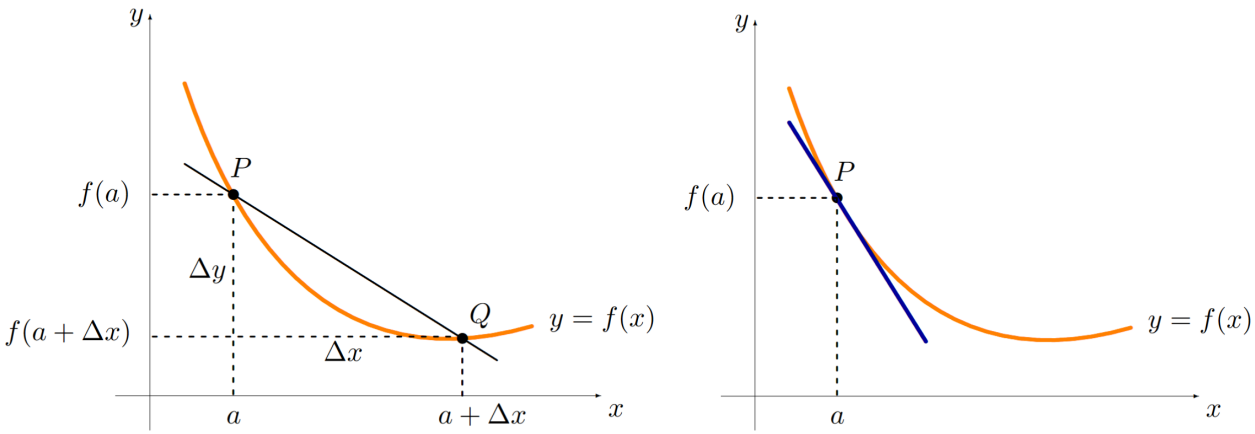



**Figure 2.5:** Difference quotient and slope of the tangent to $y = f(x)$ at $P(a, f(a))$.

## 2.4 Derivatives

After an introduction to functions, the next step is to define the **derivative**, which provides a unified way of measuring the rate of change of a function with respect to its variables.

### 2.4.1 Limit of Difference Quotients

Let $f : (c, d) \longrightarrow \mathbb{R}$ be a function of one variable $x$ and $x = a \in D_f = (c, d)$. The **derivative** of $f(x)$ at $x = a$ is denoted by $f'(a)$ and is defined as the limit (if it exists) of the difference quotients

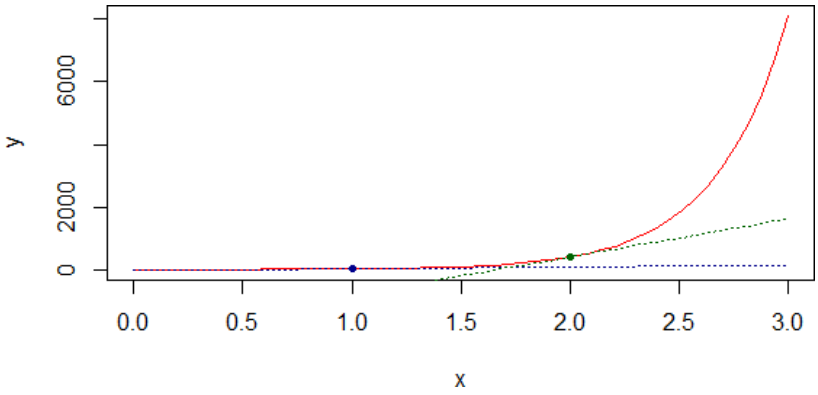
$$f'(a) = \lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}.$$

The **number** $f'(a)$ is a measure of the rate of change of $f$ at $x = a$. Geometrically, the value $f'(a)$ is the slope of the tangent line to the graph of $f$ at the point $(a, f(a))$.

In general, the value of the derivative of $f$ depends on $x$; we therefore define the **derivative function** $f' : (c, d) \longrightarrow \mathbf{R}$, which also carries the meaning of **slope function**.

**Example** Consider the exponential function $f$ defined by $f(x) = e^{3x}$ on $\mathbb{R}$, whose graph is represented by the red curve below.

```
x <- seq(0, 3, length.out=50)
y <- exp(3*x)

plot(x, y, type='l', col=rainbow(25), lty=1)
lines(x, 3*exp(3)*x-3*exp(3)+exp(3),
      col='darkblue', lty=3)
points(1, exp(3), pch=20, col='darkblue')
lines(x, 3*exp(6)*x-3*2*exp(6)+exp(6),
      col='darkgreen', lty=3)
points(2, exp(6), pch=20, col='darkgreen')
```




The graph also shows two tangent lines. The slope of each tangent line is the **rate of change** of $f$ at $x$. By comparing the slopes of the two tangent lines, we observe that the rate of change at $x = 2$ is much larger than the rate of change at $x = 1$, in accordance with the fact that the exponential function grows quite quickly.

The process of calculating the derivative of $f$ is sometimes referred as **differentiation**. The derivative is denoted in two ways:

$$\frac{df(x)}{dx} \quad \text{or} \quad f'(x),$$

it is up to the reader which one (if not both) to use.

### 2.4.2 Rules of Differentiation

But there is no need to use the definition *via* the limit of differential quotients to compute the derivative of a function. The set of **differentiation rules** are recalled here for readers' convenience.[12]

12: A detailed discussion about differentiation can be found in [6, 3].

In the following list, $x$ denotes the variable, while $a$ and $n$ are constants.

1. For a **constant function** $f$, $f'(x) = 0$
2. **Power rule:** $(x^n)' = nx^{n-1}$
3. **Exponentials:** $(e^{ax})' = ae^{ax}$
4. **Logarithms:** $(\ln(x))' = \frac{1}{x}$
5. **Product rule:** $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$
6. **Quotient rule:** $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$
7. **Chain rule:** $f(g(x))' = f'(g(x))g'(x)$

The chain rule, for instance, is important for understanding the construction of the **backpropagation** algorithm of neural network models (see Chapter 31 and [5], say).

**Example** Using the rules, compute the derivative of $f(x) = e^{-x^2}$.[13] What is the value of the rate of change of $f(x)$ at $x = 2$?

13: We will stop using the convoluted phrasing "the function $f : A \to B$ defined by $f(x) = \ldots$" and substitute instead "the function $f(x) = \ldots$" when the context allows it.

From the exponentials derivative rule and the chain rule, we obtain:

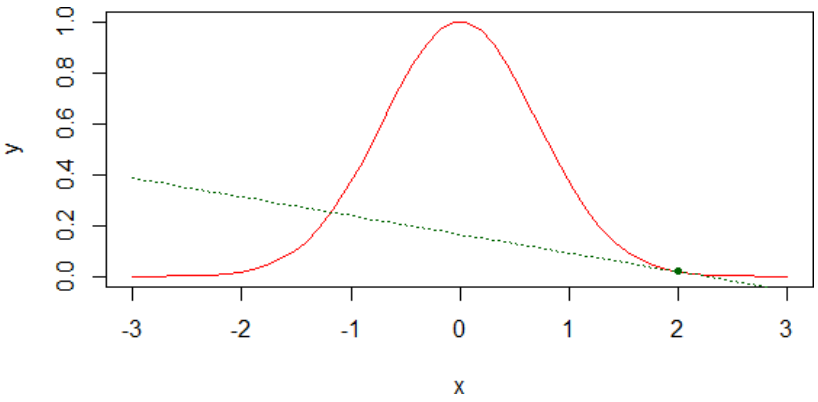
$$f'(x) = (e^{-x^2})' = e^{-x^2}(-x^2)' = -2xe^{-x^2}$$

At $x = 2$, the rate of change of $f(x)$ is

$$f'(2) = -2 \times 2 \times e^{-2^2} = -0.073$$

The **slope** (or **rate of change**) at $x = 2$ is negative, as expected by inspecting the shape of the bell curve representing the curve $y = e^{-x^2}$. Its value is "small", which is also expected since the function decays to zero quite rapidly.

```
x = seq(-3, 3, length.out=50)
y = exp(-x^2)
plot(x, y, type='l', col = rainbow(25), lty=1)
lines(x, -2*2*exp(-2**2)*(x-2)+exp(-2**2), col='darkgreen',
      lty=3)
points(2, exp(-2**2), pch=20, col='darkgreen')
```




### 2.4.3 Partial Derivatives

How do we expand this definition to functions of several variables? In this case, we are interested in defining and computing the rate of change with respect to any of the variables. This is done via **partial derivatives** which, computationally speaking, are a straightforward generalization of the notion of derivative of a function of one variable.

#### Partial Derivatives of Order 1

Let $f(x_1, \cdots, x_n)$, and pick any variable $x_k$, for some $k \in \{1, \cdots, n\}$, with respect to which we want to compute the rate of change of $f$. We can use the one-variable differentiation rules from Section 2.4.2 by treating the remaining variables as constant.

The **partial derivative of order one** of $f$ with respect to the variable $x_k$, denoted in two alternative ways as follows:

$$\lim_{\Delta x \to 0} \frac{f(x_1, \ldots, x_k + \Delta x, \ldots, x_n) - f(x_1, \ldots, x_k, \ldots, x_n)}{\Delta x} = \frac{\partial f}{\partial x_k} = f_{x_k}.$$

**Example** Compute the 3 partial derivatives of $f(x, y, z) = x^2y + 3xz$.

We have 3 variables, and we compute the corresponding partial derivative for each of them:

$$f_x(x, y, z) = \frac{\partial(x^2y + 3xz)}{\partial x} = 2xy + 3z$$
$$f_y(x, y, z) = \frac{\partial(x^2y + 3xz)}{\partial y} = x^2$$
$$f_z(x, y, z) = \frac{\partial(x^2y + 3xz)}{\partial z} = 3x$$

**Tangent Plane**

If $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $x = a$, the equation of the unique **tangent line to the graph** $y = f(x)$ at $P(a, f(a))$ is

$$y = f'(a)(x - a) + f(a).$$

More generally, if $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x} = \mathbf{a}$, there are infinitely many tangent lines to its graph $w = f(\mathbf{x})$ at $P(\mathbf{a}, f(\mathbf{a}))$. All of these lines lie in the same unique **tangent hyperplane**.

When $n = 2$, we have a **tangent plane** to $z = f(x, y)$ at $P(a, b, f(a, b))$; it is the plane that rests on the surface, touching it only at the point of tangency, as illustrated in the figure below.[14]

14: Near the point of tangency, the surface resembles the tangent plane: this is partly why that we've long believed the Earth to be flat!

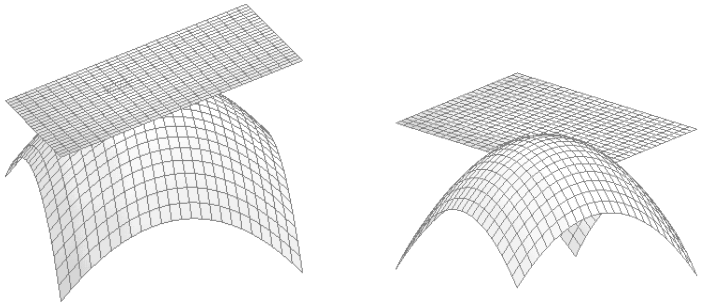


**Figure 2.6:** Tangent placn to $z = -x^2 + y^2$ at $(0, 1, 1)$, seen from two different angles.

When such a plane exists, as do the partial derivatives, the surface is said to be **differentiable** at the point in question.

If $z = f(x, y)$ is a differentiable surface $P(a, b, f(a, b))$, the equation of the tangent plane to the surface at point $P$ is

$$z = f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b).$$

**Example** Find the tangent plane to $z = \sqrt{x - y}$ at $P(2, 1, 1)$

First, we verify that $P$ is indeed on the surface. Since $a = 2$ and $b = 1$, we simply need to check that $\sqrt{a - b} = \sqrt{2 - 1} = 1$, which is indeed the case.

Next we compute the partial derivatives

$$f_x(x,y) = \frac{1}{2\sqrt{x-y}} \quad \text{and} \quad f_y(x,y) = -\frac{1}{2\sqrt{xx-y}}.$$

Thus

$$f_x(a,b) = f_x(2,1) = \frac{1}{2\sqrt{2-1}} = \frac{1}{2} \text{ and } f_y(a,b) = f_y(2,1) = -\frac{1}{2\sqrt{2-1}} = -\frac{1}{2},$$

so the equation of the tangent plane is

$$\begin{aligned} z &= f(2,1) + f_x(2,1)(x-2) + f_y(2,1)(y-1) \\ &= 1 + \frac{1}{2}\cdot(x-2) - \frac{1}{2}(y-1) = \frac{1}{2}(1+x-y). \end{aligned}$$

When the partial derivatives do not exist at a particular point on the surface, then either there is no tangent plane or it is **not unique**.

For example, the partial derivatives of $f(x,y) = 2 - \sqrt{x^2+y^2}$ are not defined when $(x,y) = (0,0)$ (which is in the domain of $f$); graphically, this translates into more than one tangent plane at the vertex of the cone $z = 2 - \sqrt{x^2+y^2}$, as shown below.

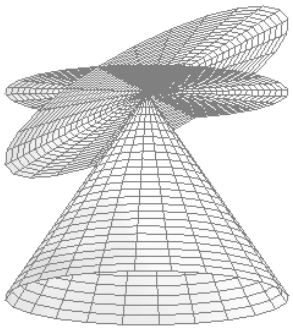


**Figure 2.7:** Two tangent planes at the vertex of the cone $z = 2 - \sqrt{x^2+y^2}$.

### Partial Derivatives of Order 2

15: For example, in optimization.

In calculus problems,[15] it is convenient to have at hand the **partial derivatives of order two**. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and pick any two variables $x_h, x_k$, for $k, h \in \{1, 2, \cdots n\}$.

The **partial derivative of order two** with respect to $x_h$ and $x_k$ (in that order) is the function

$$f_{x_h x_k}(x_1, \cdots, x_n) = \frac{\partial^2 f(x_1, \cdots, x_n)}{\partial x_k \partial x_h}$$

obtained by first computing the partial derivative with respect to $x_h$, and then the partial derivative of that partial derivative with respect to $x_k$.

But what if, when computing a partial derivative of order two, we mistakenly change the order of differentiation with respect to the two chosen variables?

It turns out that for sufficiently regular functions the order does not matter, thanks to Clairaut's Theorem, which is explained in Figure 2.8; "higher order" means that we can keep differentiating $f$,[16] obtaining partial derivatives of order $3, 4, \ldots$ and so on.

16: When the function is differentiable, it needs to be added.

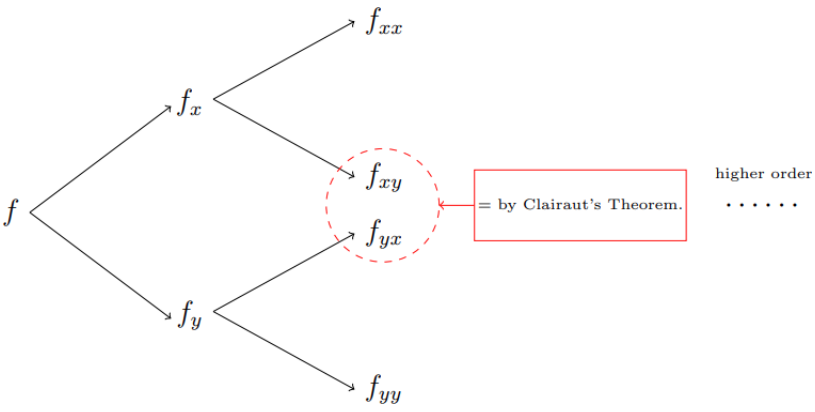



**Figure 2.8:** Illustration of Clairaut's theorem in 2 variables.

Clairaut's Theorem applies to the "standard functions" that we introduce in calculus courses, obtained by combining polynomials, rational funtions, trigonometric functions, exponentials and logarithmig functions, analytic functions (power series), etc.

**Example** Consider such a standard function of 3 variables $(x, y, z)$. In theory, $f$ has 9 partial derivatives of order 2:

$$f_{xx}, f_{xy}, f_{xz}, f_{yx}, f_{yy}, f_{yz}, f_{zx}, f_{zy}, f_{zz}$$

But thanks to Clairaut's Theorem, we have:

$$\begin{aligned} f_{xy} &= f_{yx} \\ f_{xz} &= f_{zx} \\ f_{yz} &= f_{zy} \end{aligned}$$

We only need to compute 6 partial derivatives of order 2 to obtain them all!

### 2.4.4 Gradients

From the point of view of data analysis, the most important vector fields are the gradients of multivariate functions $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$.

The **gradient** $\nabla f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is defined by:[17]

17: Pronounced "nabla".

$$\nabla f(x_1, \cdots, x_n) = \langle f_{x_1}(x_1, \cdots, x_n), \cdots, f_{x_n}(x_1, \cdots, x_n). \rangle$$

The $\langle \ldots \rangle$ notation is used to distinguish vector fields (and vectors) from points in $\mathbf{R}^n$, which are denoted using $(\ldots)$.[18]

18: The gradient is not only a way to collect the first order partial derivatives of a function into a vector, but it carries important geometrical information about the function, as we shall soon see.

**Example** We can easily ompute the gradient of $f(x, y, z) = x^2y + z$, and evaluate it at $(-1, 1, 2)$.

Indeed,

$$\nabla f(x, y, z) = \langle 2xy, x^2, 1\rangle.$$

At $(-1, 1, 2)$, the gradient becomes a 3-dimensional **vector**:

$$\nabla f(-1, 1, 2) = \langle 2 \cdot (-1) \cdot 1, (-1)^2, 1\rangle = \langle -2, 1, 1\rangle.$$

### Gradient and Level Sets

There is a crucial property linking the gradient of a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ and its level sets: wherever $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the gradient is **perpendicular** to the level sets of $f$.

More precisely, given a point $\mathbf{a} = (a_1, \cdots, a_n) \in D$, if $\nabla f(\mathbf{a}) \neq \mathbf{0} = (0, \ldots, 0)$, then $\nabla f(\mathbb{a}) \perp L_{\mathbf{a}}$, where $L_{\mathbf{a}}$ is the level set of $f$ through $\mathbf{a}$. In $\mathbb{R}^2$, we can visualize this property quite easily.

**Example** Consider the function $f(x, y) = x^2 + y^2$, whose level curves are concentric circles. The gradient vector field is represented by the vectors in Figure 2.9. Since $\nabla f)(x, y) = \langle 2x, 2y\rangle$, the gradient is a radial vector field,[19] and the orthogonality is a simple consequence of Euclidean geometry.[20]

19: The vectors point along the radii of the level circles.

20: A radius meets its circle orthogonally [3].

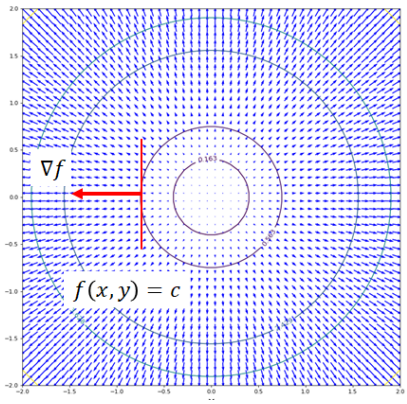



**Figure 2.9:** The gradient $\nabla f = \langle 2x, 2y\rangle$ is perpendicular to the level sets $x^2 + y^2 = c$, as is illustrated with $(x, y) = (-1, 0)$.

## 2.4.5 Directional Derivatives

In studying a function whose domain $D$ is a region of $n-$dimensional space $\mathbb{R}^n$, we usually choose $n$ **preferred** pairwise orthogonal directions, corresponding to the $n$ cartesian coordinates $(x_1, \cdots, x_n)$. Those directions are given by the **canonical basis vectors**

$$\mathbf{e}_1 = \langle 1, 0, \cdots, 0\rangle$$
$$\vdots$$
$$\mathbf{e}_n = \langle 0, 0, \cdots, 1\rangle$$

Note that each canonical basis vector is of length 1. In $\mathbb{R}^3$ we also denote the canonical basis by $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\} = \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$.

The **rate of change** of $f$ **along the direction** $\mathbf{e}_k$ is the partial derivative $f_{x_k}$. We can also use **any** direction $\mathbf{u}$ with unit length. We can find the appropriate formula using "minimally intuitive'' reasoning.[21]

21: To quote Dr. De Oliveira.

The vector $\mathbf{u}$ is a linear combination of the basis elements:

$$\mathbf{u} = c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n.$$

As we have discussed, the rate of change of $f$ along $\mathbf{e}_k$ is $f_{x_k}$. If $\mathbf{u}$ is of length 1, we can interpret the linear combination above as a **signed weighted average** of the canonical basis vectors $\mathbf{e}_k$; consequently, it is reasonable to define the rate of change of $f$ along $\mathbf{u}$ as the signed weighted average of the partial derivatives $f_{x_k}$, with the same coefficients $c_k$.[22]

22: The proof that this indeed the right approach to take is an easy consequence of the chain rule.

### Link With the Gradient

Given a unit vector

$$\mathbf{u} = c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n,$$

the **directional derivative** of $f$ along $\mathbf{u}$ is

$$D_{\mathbf{u}}f(x_1, \cdots, x_n) = c_1 f_{x_1}(x_1, \cdots, x_n) + \cdots + c_n f_{x_n}(x_1, \cdots, x_n).$$

Using the **dot product** of vectors, we can also write

$$D_{\mathbf{u}}f(x_1, \cdots, x_n) = \nabla f(x_1, \ldots, x_n) \cdot \mathbf{u}.$$

**Example** What is the directional derivative of $f(x, y) = \cos(xy) + y$ along the unit vector $\mathbf{u} = \frac{1}{\sqrt{2}}\langle 1, 1\rangle$ at the point $(1, 1)$?

We start computing the gradient of $f$:

$$\nabla f(x, y) = \langle -y\sin(xy), -x\sin(xy) + 1\rangle.$$

The directional derivative as a function (that is, for arbitrary $x, y$) is

$$\begin{aligned} D_{\mathbf{u}}f(x, y) &= \nabla f(x, y) \cdot \mathbf{u} = \langle -y\sin(xy), -x\sin(xy) + 1\rangle \cdot \frac{1}{\sqrt{2}}\langle 1, 1\rangle \\ &= -\frac{1}{\sqrt{2}}y\sin(xy) + \frac{1}{\sqrt{2}}(-x\sin(xy) + 1). \end{aligned}$$

At $x = 1, y = 1$ we obtain

$$D_{\mathbf{u}}f(1, 1) = -\frac{1}{\sqrt{2}}\sin(1) + \frac{1}{\sqrt{2}}(-1\sin(1) + 1) = -\sqrt{2}\sin(1) + \frac{1}{\sqrt{2}}$$

### Minimum and Maximum Rate of Change

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{a} = (a_1, \cdots, a_n) \in D$ with $\nabla f(\mathbf{a}) \neq \mathbf{0}$. The **maximum** rate of change of $f$ at $\mathbf{a}$ occurs along the direction of the

gradient,

$$\frac{\nabla f(\mathbf{a})}{||\nabla f(\mathbf{a})||},$$

while the **minimum** rate of change of $f$ at $\mathbf{a}$ occurs along the opposite direction.

To understand this last statement let us reason in the case of a function of two variables whose graph $z = f(x, y)$ is a surface. In order to climb or go down the hill along the steepest way, we move perpendicularly to the contour line of the hill located at a certain height. The orthogonal direction is given by the gradient.[23]

23: This property is crucial in understanding the **gradient descent** algorithm that searches for the minimum values of a function (the cost function). See Chapter 31, *A Deep Learning Launchpad*.

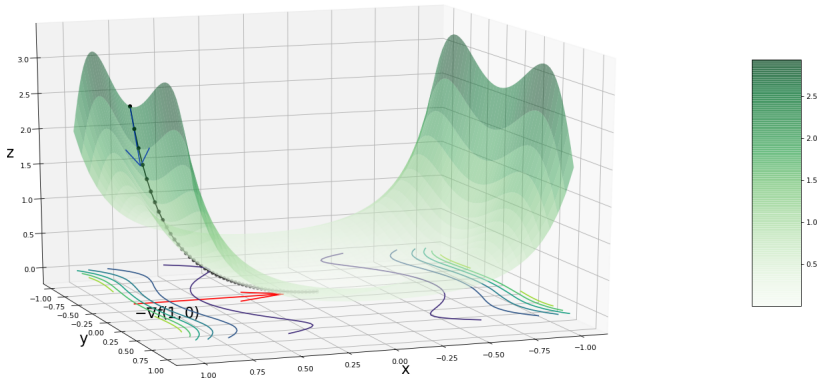



**Figure 2.10:** Gradient descent search for the minimum of $z = (x^2+y^2)\exp(x^4-y^4)$ [5].

**Example** What is the maximum rate of change of $f(x, y) = x^2 + y^2$ at $(1, 1)$?

We start with the calculation of the gradient

$$\nabla f(x, y) = \langle 2x, 2y \rangle.$$

At $(x, y) = (1, 1)$, the gradient is

$$\nabla f(1, 1) = \langle 2, 2 \rangle,$$

the unit vector corresponding to the direction of maximum rate of change is thus

$$\mathbf{u} = \nabla f(1, 1)||\nabla f(1, 1)||\frac{1}{\sqrt{2}}\langle 1, 1 \rangle.$$

The value of the maximum rate of change is thus given by:

$$D_{\mathbf{u}}f = \nabla f(1, 1) \cdot \mathbf{u} = \nabla f(1, 1) = \langle 2, 2 \rangle \cdot \frac{1}{\sqrt{2}}\langle 1, 1 \rangle = 2\sqrt{2}.$$

For a general $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{a} \in D$ such that $\nabla f(\mathbf{a}) \neq \mathbf{0}$, the value of the **maximum rate of change** of $f$ at $\mathbf{a}$ is $||\nabla f(\mathbf{a})||$; conversely, the **minimum rate of change** of $f$ at $\mathbf{a}$ is $-||\nabla f(\mathbf{a})||$.

## 2.5 Optimization

Optimization problems arise in many areas of sciences and mathematics.

1. In **regression analysis**, we minimize a "cost function" in order to find the parameters that best fit the available data (see Chapter 8);
2. in **machine learning**, we use algorithms to adjust the learning parameters, again by minimizing a cost function (see Chapters 19, 20, 21, and 31);
3. in **general relativity**, objects move along **geodesics**, which are the trajectories of minimal length, and
4. in **geometry**, the shortest path joining two points on a sphere is the great circle passing through the points.[24]

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. The goal of **optimization** is to find where $f$ reaches its **maximum** and **minimum** values, and to determine these values as well.[25]

24: These are crucial to navigation, especially when it comes to determining the fastest and cheapest air routes between two cities.

25: We provide a more in-depth look at optimization in Chapter 5.

**Example** In **linear regression**, we construct a linear model, in which a dependent variable (the **response**) is predicted by the independent variables (**predictors**) by means of a linear function.

Consider the case when we have only one independent variable, denoted by $x$. The goal is to find the linear relation that best determines the value of the response $y$ as a function of $x$: $y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon$ is the **error component** of the model.[26] The regression goal is to determine the **optimal** model parameters $\beta_0$ and $\beta_1$. But what does optimal mean in this context?

Let $(x_k, y_k), k = 1, \ldots, N$, be the observed/available data. In the **ordinary least squares** framework, the best estimate of the true parameters $\beta_0, \beta_1$ (assuming that the linear model was appropriate in the first place) are the values minimizing the **residual sum of squares**:

$$Q(\beta_0, \beta_1) = \sum_{k=1}^{N} (\beta_0 + \beta_1 x_k - y_k)^2.$$

26: In practice, the relation between $x$ and $y$ is unlikely to be exact, and the error component (which relies of distribution parameters) is part and parcel of the problem. We will discuss this in much more detail in Chapter 8.

In the rest of this section, we will review a few of the standard concepts and methods for solving optimization problems, which come in two flavours:

1. **analytical methods**, which are based on differential calculus – they yield exact solutions, but fail in practice when the underlying model is too complicated,[27] and
2. **numerical methods** which provide approximate solutions when that is the case.[28]

27: See Chapter 5 for more information.

28: See Chapter 4 for more information.

### 2.5.1 Critical Points

The properties of gradient mentioned above require that the gradient not be zero at the point of interest. But observations where the gradient is zero

are also important. These "equilibrium" points are location candidates for finding function's **extrema** (max/min).

Throughout, let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{a} = (a_1, \ldots, a_n) \in D$. The latter is a **critical point** of $f$ if

$$\nabla f(\mathbf{a}) = \mathbf{0} \quad \text{or} \quad \nabla f(\mathbf{a}) \text{ does not exist.}$$

The latter situation occurs at the cone's apex in Figure 2.7, for instance.

In term of equations, this means that $\mathbf{x} = (x_1, \ldots, x_n) = (x_1, \ldots, x_n) = \mathbf{a}$ is a solution of the system

$$\begin{aligned} f_{x_1}(a_1, \cdots, a_n) &= 0 \\ &\vdots \\ f_{x_n}(a_1, \cdots, a_n) &= 0. \end{aligned}$$

In general situations, it is typically somewhat difficult to find the critical points of a function, for two reasons:

1. the system of equations encoded in $\nabla f = \mathbf{0}$ is often **non-linear**, and so we can not use linear algebra methods to solve it;
2. but even when the system is linear, if the number of variables is large, it may be time consuming to use the Gauss-Jordan algorithm to obtain solution(s).[29]

29: See Chapter 3 for details.

We thus often have to rely on **numerical solvers**: the good news is that most programming languages come with libraries that do the work behind the scenes. But it remains important to have a basic understanding of the underlying mathematics, if we want to make conscientious use of such libraries.

**Example** Find the critical points of $f(x, y) = \sin(xy)$. Plot the graph and the contour curves of $f$ as a solution.

We start by computing the gradient of $f$:

$$\nabla f(x, y) = \langle y \cos(xy), x \cos(xy) \rangle.$$

Next, we solve the system $\nabla f = \mathbf{0}$, which consists of the following equations:

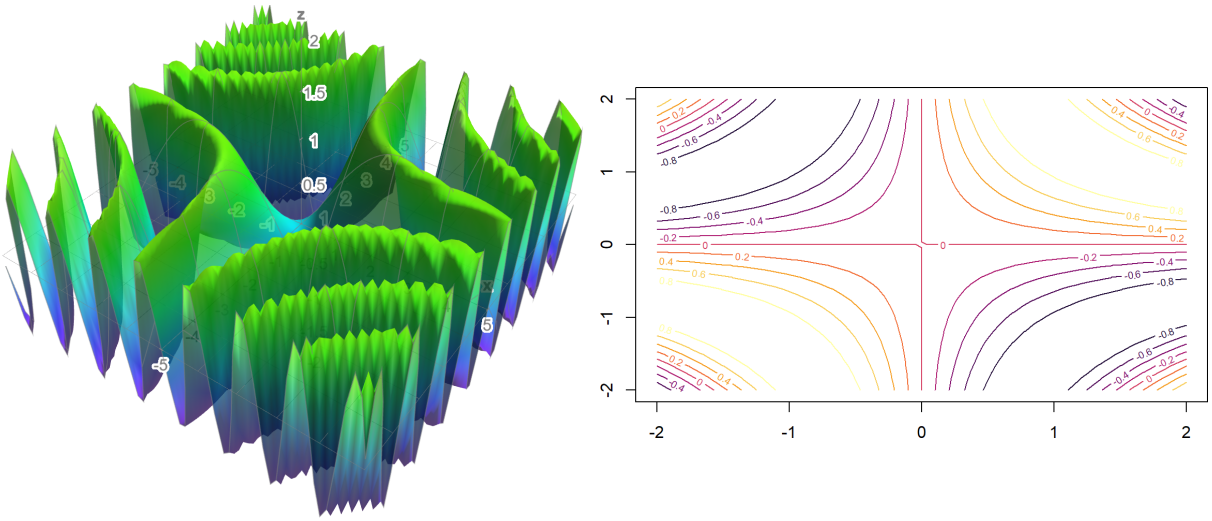
$$y \cos(xy) = 0 \quad \text{and} \quad x \cos(xy) = 0.$$

The first of these has two possible solutions: $y = 0$ or $\cos(xy) = 0$.

Substituting $y = 0$ in the second equation yields $x \cos(0) = x = 0$, which implies that $x = 0$ as well. Thus, $P = (0, 0)$ is a critical point of $f$.

If $\cos(xy) = 0$, then $xy = \frac{\pi}{2} + n\pi$, which automatically satisfies the second equation. We have thus found an infinite collection of critical points of $f$, namely all the points located along the the **hyperbolas** $xy = \frac{\pi}{2} + n\pi$. If we let $xy = t$, we see in fact that the graph of $f$ looks like a "distorted cosine wave" drawn along each hyperbola.

```
# graph
library(plot3D)
M <- mesh(seq(-2, 2, length.out = 50),
          seq(-2, 2, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- sin(x*y)
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)

# contour lines
x <- seq(-2,2,length.out=50)
y <- seq(-2,2,length.out=50)
z <- sin(outer(x,y,"*"))
cols <- hcl.colors(10, "Inferno") #color palette
contour(x,y,z, col=cols)
```



### 2.5.2 Local vs. Global

The extreme values of a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ fall into two main categories: **local** and **global**. In general, a **local** property is a property that is satisfied (detected) on a small subregion of the domain $D$; a **global** property is one that is satisfied everywhere in the domain.

Thus local extrema are extreme values in a sub-region of the domain $D$, global extrema are extreme values along the entire domain.

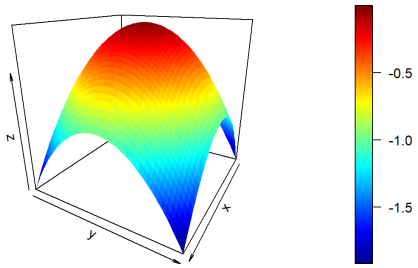
### 2.5.3 Local Extrema

We now discuss how to find the local extrema of multivariate functions using differential calculus.[30] Locally, the 3 standard shapes that we encounter at a critical point $\mathbf{x} = \mathbf{a} \in D$ where $\nabla f(\mathbf{a}) = \mathbf{0}$ resemble the following.

1. **Local maximum**

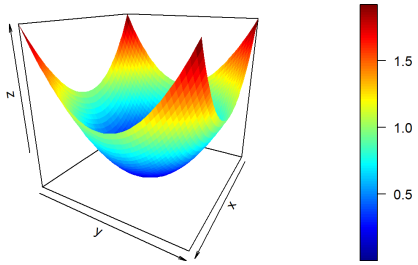
30: In order to keep things simple from a geometrical perspective, we will restrict our efforts to function $f$ of two variables, but the concepts generalize to higher $n$. In this case, the graph is the surface $z = f(x, y)$, which can be interpreted as a hilly region over the domain $D$ of $f$.

```
M <- mesh(seq(-1, 1, length.out = 50),
          seq(-1, 1, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- -x**2-y**2
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```



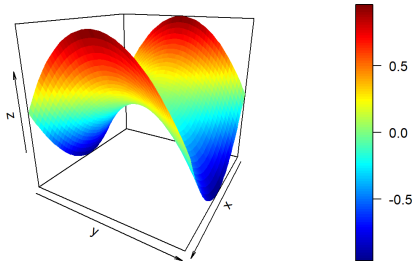

2. **Local mimimum**

```
z <- x**2+y**2
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```




3. **Saddle point** ("hybrid": max on one direction, min on the other one)

```
z <- x**2-y**2
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```




**Definitions**

We say that $f$ has a **local minimum** at $\mathbf{a} = (a_1, \cdots, a_n)$ if $f(\mathbf{a}) \leq f(\mathbf{x})$ for all $\mathbf{x}$ in a small $n$-dimensional region of $D$ centered at $\mathbf{a}$. In contrast, $f$ has

a **local maximum** at $\mathbf{a}$) if $f(\mathbf{a}) \geq f(\mathbf{x})$ for all $\mathbf{x}$ in a small $n$-dimensional region of $D$ centered at $\mathbf{a}$.

### Critical Points and Local Extrema

It is the following result (presented without proof) that justifies the importance of critical points in the optimization context.

**Theorem** If $f$ has a local extremum at $\mathbf{x} = \mathbf{a}$, then $\mathbf{x} = \mathbf{a}$ is a critical point of $f$.

The only candidates for **local** extrema are thus critical points.[31] The first step in the search of local extrema therefore consists in solving the system $\nabla f = \mathbf{0}$.

31: That is not necessarily the case for

Once that is done, we need to determine which critical points are local maxima and which are local minima. Thankfully, the **second derivative test** of introductory calculus can be generalized to any finite dimension $n$, as we shall see shortly.

### The Hessian Matrix

We have already introduced the gradient of $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, a vector field which provides **first-order** information about $f$. Second derivatives are collected into the **Hessian** matrix:

$$H(f)(\mathbf{x}) = \begin{bmatrix} f_{x_1x_1}(\mathbf{x}) & \cdots & f_{x_1x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ f_{x_nx_1}(\mathbf{x}) & \cdots & f_{x_nx_n}(\mathbf{x}). \end{bmatrix}$$

The Hessian matrix is symmetric (according to Clairaut's Theorem): a linear algebra result states that real symmetric matrix have real **eigenvalues**.[32]

32: We will discuss these notions in detail in Chapter 3.

Each eigenvalue $\lambda$ of $H(f)(\mathbf{a})$ is associated to an **eigenvector** $\mathbf{v} \in \mathbb{R}^n$; the sign of the eigenvalue provides information about the local behaviour of $f$ at $\mathbf{x} = \mathbf{a}$, along the direction determined by $\mathbf{v}$.

### Second Derivative Test

Suppose $\mathbf{a} \in D$ is a critical point of $f$ and let

$$H(f)(\mathbf{a}) = \begin{bmatrix} f_{x_1x_1}(\mathbf{a}) & \cdots & f_{x_1x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ f_{x_nx_1}(\mathbf{a}) & \cdots & f_{x_nx_n}(\mathbf{a}). \end{bmatrix}$$

be the Hessian matrix of $f$ at $\mathbf{a}$. If **all** eigenvalues of $H(f)(\mathbf{a})$ are **negative**, then $f$ has a **local maximum** at $\mathbf{x} = \mathbf{a}$; if **all** eigenvalues of $H(f)(\mathbf{a})$ are **positive**, then $f$ has a **local minimum** at $\mathbf{x} = \mathbf{a}$; if some are positive and some are negative, then $f$ has a **saddle point** at $\mathbf{x} = \mathbf{a}$.[33]

33: What happens if some of the eigenvalues are 0?

If $f : D \subseteq \mathbb{R} \to \mathbb{R}$, this is simply the second derivative test in $\mathbb{R}$: let $a$ be a critical point of $f$ with $f'(a) = 0$:

- if $f''(a) < 0$, then $f$ has a local maximum at $x = a$;
- if $f''(a) > 0$, then $f$ has a local maximum at $x = a$, and
- if $f''(a) = 0$, we can not use the second derivative to determine the nature of the critical point.[34]

34: It may be a local maximum (such as $a = 0$ for $f(x) = -x^4$), a local minimum (such as $a = 0$ for $f(x) = x^4$, or an inflection point (such as $a = 0$ for $f(x) = x^3$. Which it is depends on the function in question.

**Example** Find and classify the critical points of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = x^2 + y^2 + xz$.

We start by computing the gradient of $f$:

$$\nabla f(x, y, z) = \langle 2x + z, 2y, x \rangle.$$

The system $\nabla f = \mathbf{0}$ has a unique solution, $x = y = z = 0$; the only critical point of $f$ is thus located at $\mathbf{0} = (0, 0, 0)$.

The Hessian matrix $H(f)(\mathbf{x})$ is constant since $f$ was quadratic. In particular,

$$H(f)(\mathbf{0}) = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

We can compute the eigenvalues and the corresponding eigenvectors of $H(f)(\mathbf{0})$ **algebraically** (see Chapter 3), but we can also solve the eigenvalue/eigenvectors problem numerically with two lines of code in R:

```
H = matrix(c(2, 0, 1, 0, 2, 0, 1, 0, 0), 3, 3)
print(eigen(H))
```

$$\lambda_1 = 2.4 \quad \mathbf{v}_1 = \langle 0.9, 0, 0.4 \rangle$$
$$\lambda_2 = 2 \quad \mathbf{v}_2 = \langle 0, -1, 0 \rangle$$
$$\lambda_3 = -0.4 \quad \mathbf{v}_3 = \langle 0.4, 0, -0.9 \rangle$$

Two of the eigenvalues are positive, the other one is negative; the critical point $\mathbf{0} = (0, 0, 0)$ is a saddle point of $f$.

Geometrically, along the plane spanned by the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$,[35] which corresponds to the positive eigenvalues $\lambda_1$ and $\lambda_2$ of $H(f)(\mathbf{0})$, $f$ behaves like a function of two variables with a **local minimum**; along the line spanned by the vector $\mathbf{v}_3$ associated with the negative eigenvalue $\lambda_3$, $f$ behaves like a function of one variable with a **local maximum**.

35: These concepts are discussed in Chapter 3.

### 2.5.4 Global Extrema

When we attempt of minimizing the cost function in a machine learning algorithm, we hope to find the **smallest possible cost**, which will correspond to the parameters associated with the "best learning". In mathematical terms we are looking for the **global minimum** of the cost function, which does not necessarily occur at a local minimum – indeed,

it is conceivable that the global minimum is reached on the boundary of the domain.

In other types of problems, it could be the **global maximum** that is of interest.

#### Definitions

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. We say that $f$ reaches its **global minimum** at $\mathbf{a} \in D$ if $f(\mathbf{a}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in D$; the value $f(\mathbf{a})$ is the global minimum value of $f$. For the **global maximum**, we replace "$\leq$" by "$\geq$".

Note that global extrema do not necessarily exist: $f : (0, \infty) \to \mathbb{R}$, $x \mapsto \frac{1}{x}$ has neither a global maximum nor a global minimum.

#### Closed and Bounded Domains

A subset $D \subseteq \mathbb{R}^n$ is **bounded** if it can be contained in an $n-$ball of finite radius; formally, it there exists $M > 0$ such that

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \leq M$$

for all $\mathbf{x} \in D$.

It is **closed** if it contains it boundary. This is perhaps more difficult to grasp than it looks. An alternative definition (in $\mathbb{R}^n$) is that $D$ is closed if every $\mathbf{x} \notin D$ is contained in an $n-$ball centered at $\mathbf{x}$ which lies entirely outside of $D$.

**Example** The disk $D \subseteq \mathbb{R}^n$ defined by the inequality $x^2 + y^2 < 1$ is a bounded domain (use $M = 1$, but it not closed – its boundary, which consists of the circle $x^2 + y^2 = 1$, is not contained in $D$. The **closure** of $D$ is $x^2 + y^2 \leq 1$.

#### Extreme Value Theorem

If $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is **continuous** (roughly speaking, if it has no jump or break) over a closed and bounded domain, then $f$ admits a global maximum and a global minimum on $D$.

The EVT is not useful from a computational point of view, but it gives some conditions that guarantee that the problems of searching for global extrema makes sense.

**Example** Let $D$ be the open disk as in the previous example, and denote its closure by $\overline{D}$. Consider the function $f(x, y) = x^2 + y^2$ on $D$: the globabl minimum of $f$ is 0, clearly attained at $x = y = 0$. However there is no global maximum, since the maximum value is "pushed" to the boundary circle, which is not part of the domain.

If we take the same function but extend it to the closed domain $\overline{D}$, then $f$ does reach its maximum value of 1, at infinitely many points along the boundary circle.

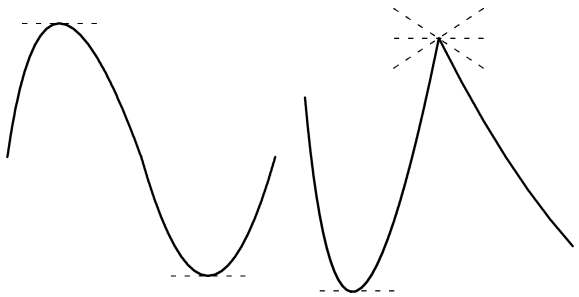


**Figure 2.11:** Critical points for continuous functions of a single real variable.

### 2.5.5 Lagrange Multipliers

We have already discussed the link between optimization and the derivative when it comes to finding local extrema. Is there a link for global optimization?

Recall that a differentiable function $f : [a, b] \to \mathbb{R}$ has a **critical point** at $x^* \in (a, b)$ if either $f'(x^*) = 0$ or $f'(x^*)$ is undefined (see Figure 2.11).

If additionally $f$ is continuous, then the optimal solution of the problem

$$\begin{array}{rl} \max & f(x) \\ \text{s.t.} & x \leq b \\ & x \geq a \\ & x \in \mathbb{R} \end{array}$$

is found at one (or possibly, many) of the following **feasible solutions**: $x = a$, $x = b$, or $x = x^*$ where $x^*$ is a critical point of $f$ in $(a, b)$.

This can be extended fairly easily to multi-dimensional domains, with the following result.

**Theorem** Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a continuous function, where $A$ is a closed subset of $\mathbb{R}^n$. Then $f$ reaches its maximum (resp.minimum) value either at a critical point of $f$ in $A^\circ$, the **interior** of $A$, or somewhere on $\partial A$, the **boundary** of $A$.

**Example** Consider a company that sells gadgets and gizmos. If the company's monthly profits are expressed (in 1000$ dollars) according to

$$f(x, y) = 81 + 16xy - x^4 - y^4,$$

where $x$ and $y$ represent, respectively, the number of gadgets and gizmos sold monthly (in 10,000s of units), and if the company can produce up to 30,000 units of both gadgets and gizmos monthly, what is the optimal number of each items that the company must sell in order to maximize its monthly profits? The monthly profit function is shown in Figure 2.12.

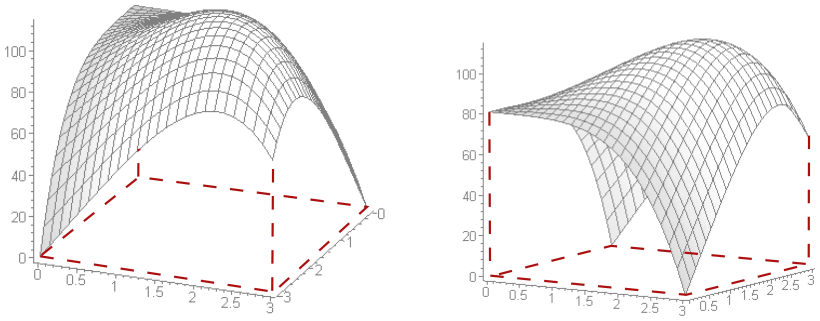


**Figure 2.12:** Monthly profit function for the gadgets and gizmos example.

Since $f$ is continuous, the maximum value is reached at a critical value in

$$A^\circ = (0,3) \times (0,3)$$

or somewhere on the boundary

$$\partial A = \{(x,y) \in [0,3]^2 \mid x = 0 \text{ or } x = 3 \text{ or } y = 0 \text{ or } y = 3\}.$$

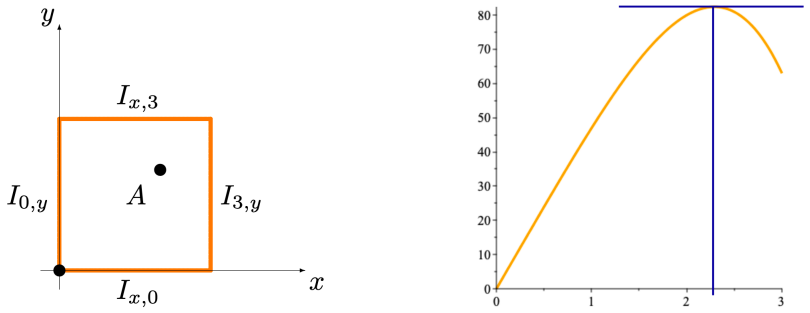



**Figure 2.13:** Boundary of the domain (left); profile for $g_3$ and $h_3$ (right) in the gadgets and gizmos example.

But $f$ is smooth; the gradient $\nabla f(x,y)$ is thus always defined, and the only critical points are those for which $\nabla f(x,y) = (16y - 4x^3, 16x - 4y^3) = (0,0)$. At such a point, $4x = y^3$, which, upon substitution in $f_x$ yields

$$0 = 16y - \frac{1}{16}y^9 = \frac{1}{16}y(256 - y^8) = \frac{1}{16}y(y-2)(y+2)(y^2+4)(y^4+16),$$

which is to say $y = -2, 0, 2$.

Only $y = 2$ can potentially yield a critical point in $A^\circ$, however. When $y = 2$, we must have $x = \frac{1}{4}2^3 = 2$: the only critical point of $f$ in $A^\circ$ is thus $(x^*, y^*) = (2,2)$, and the monthly profit function value at that point is

$$f(x^*, y^*) = 81 + 16(2)(2) - 2^4 - 2^4 = 113.$$

On the boundary $\partial A$, the objective function reduces to one of:

$$\begin{aligned}
f(0,y) &= g_0(y) = 81 - y^4, \quad \text{on } 0 \le y \le 3 \\
f(3,y) &= g_3(y) = 48y - y^4, \quad \text{on } 0 \le y \le 3 \\
f(x,0) &= h_0(x) = 81 - x^4, \quad \text{on } 0 \le x \le 3 \\
f(x,3) &= h_3(x) = 48x - x^4, \quad \text{on } 0 \le x \le 3
\end{aligned}$$

These are easy to optimize, being continuous functions of a single real variable; $g_0$ and $h_0$ are maximized at the origin, with the objective

function taking the value 81 there, while $g_3$ and $h_3$ are maximized at $12^{1/3}$, with the objective function taking the value $\approx 82.42$ there (see Figure 2.13).

Combining all this information, we conclude that the company will maximize its monthly profits at 113,000$ if it sells 20,000 units of both gadgets and gizmos.

---

While the approach we just presented works in this case, there are many instances for which it can be substantially more difficult to find the optimal value on $\partial A$.

The method of **Lagrange multipliers** can simplify the computations, to some extent. Consider the problem

$$\left|\begin{array}{lll} \min/\max & f(\mathbf{x}) & \\ \text{s.t.} & g_i(\mathbf{x}) \leq a_i & i = 1, \dots, m \\ & \mathbf{x} \in \mathcal{D}, & \end{array}\right.$$

where $f, g_i$ are continuous and differentiable on the (closed) region $A$ described by the constraints $g_i \leq a_i$, $i = 1, \dots, m$.[36] If the problem is **feasible** and **bounded**,[37] then the optimal value is reached either at a critical point of $f$ in $A^\circ$ or at a point $\mathbf{x} \in \partial A$ for which

$$\nabla f(\mathbf{x}) = \lambda_1 \nabla g_1(\mathbf{x}) + \cdots + \lambda_m \nabla g_m(\mathbf{x}),$$

where $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ are the **Lagrange multipliers** of the problem.

36: Strictly speaking, differentiability is not required on the entirety of $A$.

37: See Chapter 5.

**Example** Consider a factory that produces various types of deluxe pickle jars. The monthly number of jars $Q$ of a specific kind of pickled radish that can be produced at the factory is given by $Q(K, L) = 900K^{0.6}L^{0.4}$, where $K$ is the number of dedicated canning machines, and $L$ is the monthly number of employee-hours spent on the pickled radish.

The pay rate for the employees is 100$/hour (the pickles are extra deluxe, apparently); the monthly maintenance cost for each canning machine is 200$.

If the factory owners want to maintain monthly production at 36,000 jars of pickled radish, what combination of number of canning machines and employee-hour will minimize the total production costs? The optimization problem is

$$\left|\begin{array}{lll} \min & f(K, L) = 200K + 100L & \\ \text{s.t.} & K^{0.6}L^{0.4} = 40; & K, L \geq 0. \end{array}\right.$$

The **objective function** is linear and so has no critical point. The feasability region $A$ can be described by the constraints $g_1(K, L) = K^{0.6}L^{0.4} \leq 40$ and $g_2(K, L) = -K^{0.6}L^{0.4} \leq -40$. Points of interest on the boundary $\partial A$ are obtained by solving the Lagrange equation

$$(200, 100) = \lambda \left(0.6 \left(\tfrac{L}{K}\right)^{0.4}, 0.4 \left(\tfrac{K}{L}\right)^{0.6}\right),$$

since $\nabla g_1 = -\nabla g_2$, with $K^{0.6}L^{0.4} = 40$.

Numerically, there is only one solution, namely

$$(K_*, L_*, \lambda) \approx (35.65, 47.54, 297.10).$$

The objective function at that point takes on the value

$$f(K_*, L_*) \approx 200(35.65) + 100(47.54) \approx 11884.02,$$

and this value must either be the maximum or the minimum of the objective function subject to the constraints of the problem. But we know, that the point $(K_1, L_1) = (1, 40^{2.5})$ belongs to $\partial A$;[38] since

$$f(K_1, L_1) = 200(1) + 100(40^{2.5}) > f(K_*, L_*),$$

then $(K_*, L_*)$ is indeed the minimal solution of the problem, and the minimal value of the objective function subject to the constraints is $\approx 11,884.02\$$.

38: As $1^{0.6}(40^{2.5})^{0.4} = 40$).

In practice, the value for $K$ has to be an integer,[39] , so we might pick:

- a **sub-optimal** $K'_* = 36$ canning machines, which yields
- a **sub-optimal** $L'_* \approx 46.84$ employee-hours,
- which together yield a **sub-optimal** monthly operating cost of

$$f(K'_*, L'_*) \approx 200(36) + 100(46.84) \approx 11884.85.$$

This departure from optimality would nevertheless be quite likely to be acceptable to the factory owners.

39: Unless we consider using a different number of canning machines at various times during the month.

---

Given how straightforward the method is, it might seem that there is no real need to say anything else – why would anybody ever use something other than Lagrange multipliers to solve optimization problems?

One of the issues is that when the number of constraints is too high relative to the dimension $n$ of $A$,[40] then **there may not be a finite number of candidate** solutions on $\partial A$, which makes this approach useless.

40: Which is usually the case in real-life situations.

Another difficulty that might arise is that the system of equations

$$\nabla f(\mathbf{x}) = \lambda_1 \nabla g_1(\mathbf{x}) + \cdots + \lambda_m \nabla g_m(\mathbf{x})$$

could be **ill-conditioned**, or **highly non-linear**, and numerical solutions could be hard to obtain. We will discuss this further in Chapters 4 and 5.

## 2.6 Riemann Integrals

Integration, as we will see, is the reverse process of differentiation. We start with a review of basic integration rules and methods, starting with one-variable methods which can then be generalized to multiple Riemann integrals in many variables.

### 2.6.1 Motivation: Local Densities vs. Total Quantities

The following argument, motivated by statistics, is one of many possible ways of introducing the concept of Riemann integrals.

In general, the (multi-variable) **Riemann integral**

$$\int_D f(x_1, \cdots, x_n) dV$$

is the continuous version of the infinite series

$$\sum_{k_1, \cdots, k_n = 1}^{\infty} f_{k_1, \cdots, k_n} \Delta V.$$

This realization is at the centre of all approaches to Riemann integration.

Consider a real random variable $x$ with **probability density function** $f(x)$.[41] Let $x_0$ be an arbitrary value of $x$. The **probability** that $x$ takes a value in the interval $[x_0, x_0 + \Delta x]$ of length (size) $\Delta x$ (which is usually quite small) is approximately

41: See Chapter 6 for details.

$$f(x_0)\Delta x.$$

Assume that $[a, b]$ is a finite interval. We compute the probability that $x$ belongs to the (large) interval $[a, b]$ by using **Riemann sums approximations**.

First, we sub-divide the interval $[a, b]$ into $N$ **sub-intervals** of equal length $\Delta x = \frac{b-a}{N}$: if we label the endpoints of each sub-interval as

$$x_0 = a, x_1 = x_0 + \Delta x, \cdots, x_{N-1} = x_0 + (N-1)\Delta x, x_N = b,$$

then the sub-interval $I_k$ can be written as

$$I_k = [x_{k-1}, x_k].$$

If $\Delta x$ is sufficiently small, then we can say that, since the probability of finding $x$ within $I_k$ is approximately $f(x_{k-1})\Delta x$, then the probability of finding $x$ in $[a, b]$ is approximated by the sum of those "local" (**infinitesimal**) probabilities:

$$P(x \in [a, b]) \approx \sum_{k=1}^{N} f(x_{k-1})\Delta x.$$

At this point, we may be nonplussed to realize that this formula is only going to yield an **estimate** (or an **approximation**) of the exact value of the probability.

But the theory of Riemann integrals shows that as we increase the number $N$ of sub-intervals $I_k$,[42] the estimated value **converges** (gets closer and closer) to the exact value, and in the limiting case $N \to \infty$, we obtain

42: And therefore sending $\Delta x \to 0$.

$$P(x \in [a, b]) = \lim_{N \to \infty} \sum_{k=1}^{N} f(x_{k-1})\Delta x.$$

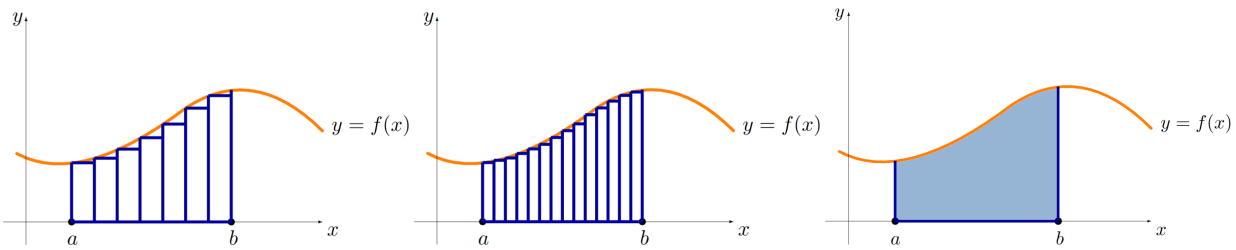



**Figure 2.14:** Graphical illustration of the Riemann integral $\int_a^b f(x)dx$: approximations with left-most sample points and $N = 7$ (left); $N = 14$ (middle); Riemann integral (right).

### 2.6.2 One Variable

Using the same reasoning, we define the Riemann integral for any continuous function $f : [a, b] \to \mathbb{R}$ by

$$\int_a^b f(x)\, dx = \lim_{N\to\infty} \sum_{k=1}^{N} f(x_{k-1})\Delta x,$$

where $N$ is the number of sub-interval $I_k$ of length $\Delta x = (b - a)/N$ and $x_k$ is a sample point in $I_k$ (the centre of the interval, say).

Different choices of sample points lead to **different versions** of the Riemann sum approximation. In the limiting case $N \to \infty$, however, all approximations converge to the same value, which is the {**Riemann integral of** $f$ **over** $[a, b]$; the process is illustrated in Figure 2.14.

### 2.6.3 Fundamental Theorem of Calculus

As is the case with derivatives, the calculation of Riemann integrals can (in principle) be performed without going through the process of Riemann sum approximations.

For a continuous function $f : [a, b] \to \mathbb{R}$, there is a function $F : [a, b] \to \mathbb{R}$ (the **antiderivative** or **indefinite integral** of $f$), which satisfies $F'(x) = f(x)$ and which we denote by

$$F(x) = \int f(x)dx,$$

The antiderivative is **unique** up to an additive constant $c$:

$$(F(x) + c)' = f(x).$$

The **Fundamental Theorem of Calculus** states that, for any antiderivative $F$ of $f$, then

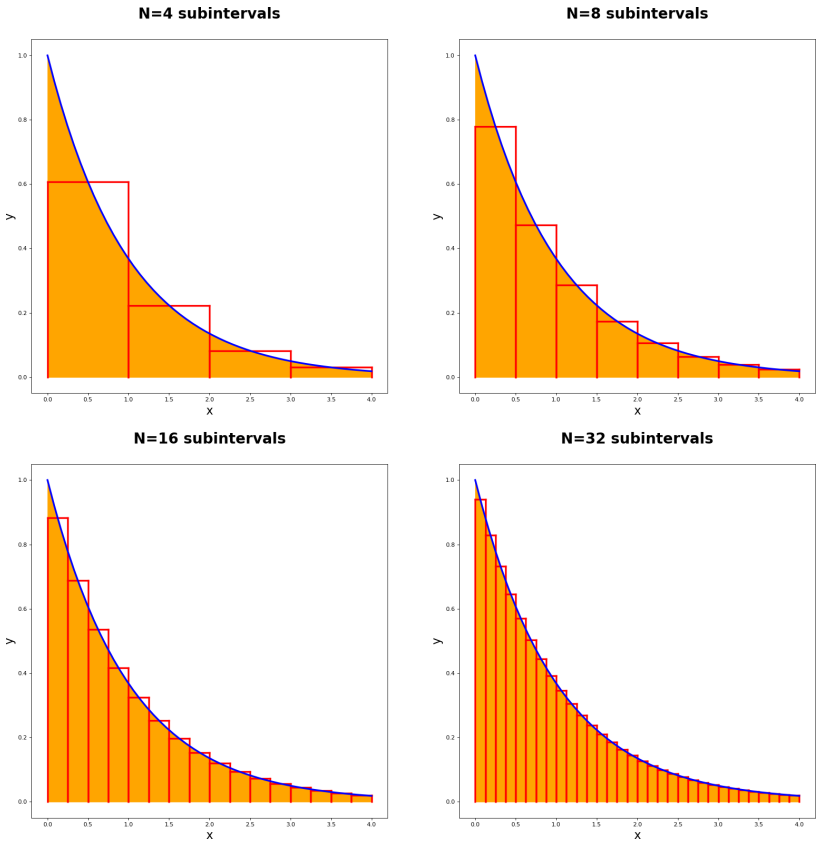
$$\int_a^b f(x)dx = F(b) - F(a) = \Big[F(x)\Big]_a^b.$$

Note that we also denote the difference $F(b) - F(a)$ by $\Big[F(x)\Big]_a^b$.

**Example** Here are the Riemann sum approximations with 4 different sub-interval sub-divisions, for the integral

$$\int_0^4 e^{-x}dx = \Big[-e^{-x}\Big]_0^4 = 1 - e^{-4}.$$

For any of the approximations, the area of each vertical rectangle is $f(\overline{x})\Delta x$, where $\overline{x}$ is the midpoint of the small interval.




The antiderivative $F$ of a continuous function $f$ always exists. However, if the analytic expression of the function is too complicated, it may not be possible to find the antiderivative $F$ of $f$.[43] What to do, then? We have no choice but to proceed with numerical integration.[44]

43: It still exists, however.

44: There are several approaches used to compute a Riemann integrals numerically. In the previous example, we used the midpoint approximation; there are other ways of approximating the integral (left-most point, right-most point, Simpson rule, Gaussian quadratures, Monte Carlo, etc.). We will discuss these in Chapter 4.

45: There are methods, but typically harder to use or understand: how do we select the right $u$-substitution? Or the $u\,dv$ term in integration by parts?

### 2.6.4 Finding Antiderivatives

Computing derivatives is usually easy, since it is (almost) a one-directional, no-choice algorithm: **follow the rules** and all is good to go.

When we find an antiderivative, we are "climbing back" to the source, and that can actually be much harder.[45]

Here are some basic rules for finding antiderivatives. For more advanced techniques, we let the reader look into the literature [6, 3].

1. **Linearity:**

$$\int (af(x) + bg(x))\,dx = a\int f(x)\,dx + b\int g(x)\,dx$$

2. **Power rule:**

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C, \text{ for } n \neq -1$$

3. **Power rule special case:**

$$\int \frac{dx}{x} = \ln|x| + C$$

4. **Exponentials:**

$$\int e^{ax}\,dx = \frac{e^{ax}}{a} + C$$

5. **Integration by parts:**

$$\int f'(x)g(x)\,dx = f(x)g(x) - \int f(x)g'(x)\,dx$$

6. **Integration by substitution:**

$$\int f(x)\,dx = \int f(x(u))\frac{dx}{du}\,du$$

Note that integration by substitution is a sort of inverse of the chain rule, and integration by parts the same for the product rule.

### 2.6.5 Several Variables

We are now ready to introduce **multiple integrals**, that is Riemann integrals of a function defined over a domain of arbitrary dimension.

Let $D \subset \mathbb{R}^n$ and $f$ be a **density function** on $D$, such as a **probability density function** for the configuration $(x_1, \cdots, x_n)$ of $n$ random variables. Let $\mathbf{a} \in D$. If we pick a point $\mathbf{x}$ at random the probability that we find it in a region centered at $\mathbf{x} = \mathbf{a}$ of $n-$volume $\Delta V$ is approximated by

$$f(\mathbf{a})\Delta x_1 \cdots \Delta x_n.$$

Let $S \subset D$ be a subregion of the whole **sample space domain** $D$. The probability $p(S)$ of finding $\mathbf{x} \in S$ is approximated as follows. Subdivide $S$ into $N$ small sample regions $S_k$ $(k = 1, \cdots, N)$, each of volume $\Delta V$. Pick, for each $k$, a sample point $P_k$ in $S_k$. According to the formula above,

we have

$$p(\mathbf{x} \in S) \approx \sum_{k=1}^{N} f(P_k)\Delta V$$

The exact value is obtained in the limiting case $N \to \infty$. This is the multivariate **Riemann integral** construction. If we use **Cartesian coordinates** $(x_1, \cdots, x_n)$, the volume is

$$\Delta V = \Delta x_1 \cdots \Delta x_n,$$

and so

$$p(\mathbf{x} \in S) = \int_S f(S)\, dV = \lim_{N\to\infty} \left( \sum_{k=1}^{N} f(P_k)\Delta x_1 \cdots \Delta x_n \right).$$

The Riemann sum approximation is used to define the Riemann integral for an arbitrary **continuous** function, not necessarily one carrying the meaning of probability.

The **double integral** ($n = 2$ variables) is often denoted by $\iint$, the **triple integral** ($n = 3$ variables) by $\iiint$. If the dimension of the integral is not important (for example, if we are interested in general properties of Riemann integrals) we simply use the symbol $\int$.

### 2.6.6 Applications to Statistics

Let $f$ be a probability density function of $n$ independent continuous random variables, on a domain $D \subset \mathbb{R}^n$. Let $g(x_1, \cdots, x_n)$ be an arbitrary random variable.[46]

46: We can assume that is a continuous function.

The **average value** of $g$ is the integral

$$E\{g\} = \int_D g(x_1, \cdots, x_n) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

The **variance** of $g$ is the integral

$$\sigma^2 = E\left\{(g - E\{g\})^2\right\} = \int_D (g(x_1, \cdots, x_n) - E\{g\})^2 f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

The **standard deviation** of $g$ is the integral

$$\sigma = \sqrt{\int_D (g(x_1, \cdots, x_n) - E\{g\})^2 f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n}.$$

The **covariance** between two random variables $g$ and $h$ is

$$\sigma\{g, h\} = \int_D (g(x_1, \cdots, x_n) - E\{g\})(h(x_1, \cdots, x_n) - E\{h\}) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

**Computing Riemann Integrals in Several Variables**

Several methods can be used to calculate the Riemann integral of a function of several variables. In Cartesian coordinates, we can deduce a formula starting, once again, from the "infinitesimal" point of view.

For simplicity, we can consider a 2D domain $D \subset \mathbb{R}^2$ defined by the inequalities

$$D : a \leq x \leq b, \quad c(x) \leq y \leq d(x).$$

Let $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$ be continuous. In order to compute the integral

$$\iint_D f(x, y)\, dydx,$$

we can proceed by iterating the integration process, one iteration per variable, as follows.

First, for each value of $x \in [a, b]$, we can integrate $\int f(x, y)\, dy$ along the vertical direction. Since $y$ satisfies the bounds $c(x) \leq y \leq d(x)$ for each $x \in [a, b]$, we start by computing the integral along the vertical strips of width $dx$:

$$\int_{c(x)}^{d(x)} f(x, y)dy.$$

Next, we integrate the contributions of each individual strip, by integrating over the remaining variable $x$. We therefore obtain a formula for computing a double integral in Cartesian coordinates, integrating first by vertical strips:

$$\int_D f dA = \int_a^b \left( \int_{c(x)}^{d(x)} f(x, y)dy \right).$$

Note that the role of the variables can be interchanged; refer to [4] for more details.

In general, if a domain $D \subset \mathbb{R}^n$ is described by Cartesian coordinate inequalities $(x_1, \cdots, x_n)$, such as:

$$\begin{aligned}
&a_1 \leq x_1 \leq b_1 \\
&a_2(x_1) \leq x_2 \leq b_2(x_1) \\
&\cdots \\
&a_n(x_1, x_2, \cdots, x_{n-1}) \leq x_n \leq b_n(x_1, x_2, \cdots, x_{n-1})
\end{aligned}$$

then the $n-$ integral of $f$ over $D$ can be computed by the **iterated integral**

$$\int_D f dV = \int_{a_1}^{b_1} \int_{a_2(x_1)}^{b_1(x_1)} \cdots \int_{a_n(x_1,x_2,\cdots,x_{n-1})}^{b_n(x_1,x_2,\cdots,x_{n-1})} f(x_1, x_2, \cdots, x_n)dx_n dx_{n-1} \cdots dx_1.$$

The idea is to integrate one variable per time, using the one-variable rules of integration. As is the case for integration in $\mathbb{R}$, there is a **change of variables** (substitution) formula for integrals in several variables.

We can then derive formulas for double integrals in **polar** coordinates, or triple integrals in **cylindrical** or **spherical** coordinates.[47]

47: Again, refer to [4] for more details.

Let $D \subset \mathbb{R}^n$ and $f : D \to \mathbb{R}$. The Riemann integral of $f$ over $D$, defined as the limit of Riemann sums, is denoted by

$$\int_D f \, dV.$$

The symbol $dV$ denotes the **infinitesimal $n-$dimensional volume element**, and the **infinitesimal quantity** $f \, dV$ represents the infinitesimal portion of $f$ contained in the infinitesimal region of measure $dV$. The **total** ("grand sum") is obtained by integrating $f \, dV$ over the full domain.

The expression of the volume element depends of the choice of coordinates. In Cartesian coordinates, the volume is as expressed above:

$$dV = dx_1 \cdots dx_n.$$

Thus, if $f \equiv 1$, $\int_D f \, dV$ represents the $n-$**volume** of $D$. For other types of coordinate systems, and the corresponding integration formulas, we once again refer to [4].

**Example** Let $E$ be the **solid** region located above the triangle of the $xy-$plane defined by the inequalities $|x| \leq 1, 0 \leq y \leq 1-x$, and below the surface $z = x^2+y^2$. Compute the triple integral of $f(x, y, z) = x$ over $E$.

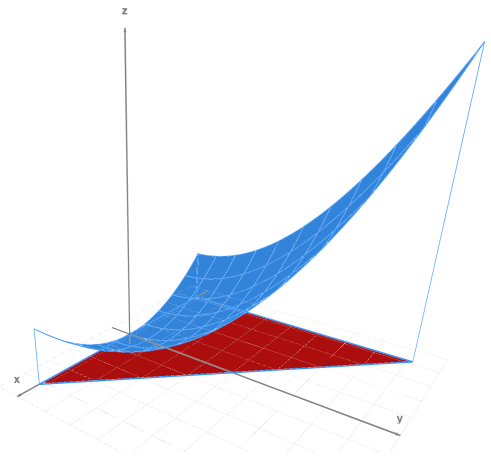
The bounds of the triangle define the region of the $xy-$plane:

$$-1 \leq x \leq 1, \quad 0 \leq 1 \leq 1 - x.$$

The solid is therefore described by the inequalities

$$-1 \leq x \leq 1, \quad 0 \leq y \leq 1 - x, \quad 0 \leq z \leq x^2 + y^2,$$

as shown below.



Therefore, the triple integral is:

$$\begin{aligned}
\iiint_E f\,dV &= \int_{-1}^{1}\int_{0}^{1-x}\int_{0}^{x^2+y^2} x\,dz\,dy\,dx \\
&= \int_{-1}^{1}\int_{0}^{1-x}\left[xz\right]_{z=0}^{z=x^2+y^2} = \int_{-1}^{1}\int_{0}^{1-x}(x^3+xy^2)\,dy\,dx \\
&= \int_{-1}^{1}\left[x^3y + x\frac{y^3}{3}\right]_{y=0}^{y=1-x} dx = \int_{-1}^{1}\left(x^3(1-x) + x\frac{(1-x)^3}{3}\right) dx \\
&= \int_{-1}^{1}\left(\frac{-4x^4}{3} + 2x^3 - x^2 + \frac{x}{3}\right) dx \\
&= \left[-\frac{4x^5}{15} + \frac{2x^4}{4} - \frac{x^3}{3} + \frac{x^2}{6}\right]_{-1}^{1} = -\frac{6}{5}.
\end{aligned}$$

## 2.7 Exercises

1. The price at which an item sells is given by $P(d,s) = k\frac{d^2}{s+10}$, where $k$ is a constant, and $s$ and $d$ are the product supply and demand, respectively.
   a) For what value(s) of $d$ is $P(d,90) = 100k$?
   b) For what value(s) of $s$ is $P(10,s) = 10k$?
   c) If $d = 9$ and $s = 10$, how does $P$ change when $d$ goes from 9 to 11?
   d) If $d = 9$ and $s = 10$, how does $P$ change when $s$ goes from 10 to 8?
   e) Compute and interpret $P(6,3)$.
   f) Compute and interpret $P_d(6,3)$.
   g) Compute and interpret $P_s(6,3)$.
2. Find the largest possible domain (in $\mathbb{R}^2$) and the range (in $\mathbb{R}$) of the following functions.
   a) $f(x,y) = x^2 + 2xy + y^2$.
   b) $f(x,y) = \ln(x-y)$.
   c) $f(x,y) = \frac{1}{(y-2)\ln x}$.
   d) $f(x,y,z) = \frac{xy}{1-z}$.
   e) $f(x,y,z) = \sqrt{36 - x^2 - 4y^2}$.
   f) $f(x,y,z) = \frac{x^2z^2}{(y-2)^2}$.
   g) $f(x,y) = \sqrt{x+y}$.
   h) $f(x,y) = \sqrt{4 - x^2 - y^2}$.
   i) $f(x,y) = \frac{1}{4-x^2-y^2}$.
   j) $f(x,y) = \frac{1}{e^{x^2+y^2}}$.
3. Find the equation of the tangent plane to the surface $z = f(x,y)$ at the given point.
   a) $f(x,y) = x^4 + y^4 - 4xy + 1, (0,0)$.
   b) $f(x,y) = x^2 + y^2 + 4x - 6y, (1,0)$.
   c) $f(x,y) = 2x^3 + xy^2 + 5x^2 + y^2, (0,1)$.

d) $f(x,y) = x^2 + y^2 + x^2y + 4$, $(1,2)$.
e) $f(x,y) = y\sqrt{x} - y^2 - x + 6y$, $(1,-1)$.
f) $f(x,y) = xy - 2x - y$, $(2,3)$.
g) $f(x,y) = xy(1-x-y)$, $(-3,2)$.
h) $f(x,y) = x^2 + y^2 + \frac{1}{x^2y^2}$, $(-1,0)$.
i) $f(x,y) = x^3 + y^3 + 4xy$, $(0,-2)$.
j) $f(x,y) = \frac{1}{xy}$, $(1,-1)$.
k) $f(x,y) = \ln(x^2+y^2)$, $(1,0)$.
l) $f(x,y) = x^y$, $(2,2)$.
m) $f(x,y) = (x+y)e^x$, $(0,2)$.
n) $f(x,y) = \frac{x+y}{x-y}$, $(2,-1)$.
o) $f(x,y) = y\ln(x+2)e^{\sqrt{y}}$, $(-1,4)$.
p) $f(x,y) = xye^{1/y}$, $(-1,1)$.

4. Classify the critical points of the following functions.

a) $f(x,y) = x^4 + y^4 - 4xy + 1$.
b) $f(x,y) = x^2 + y^2 + 4x - 6y$.
c) $f(x,y) = 2x^3 + xy^2 + 5x^2 + y^2$.
d) $f(x,y) = x^2 + y^2 + x^2y + 4$.
e) $f(x,y) = y\sqrt{x} - y^2 - x + 6y$.
f) $f(x,y) = xy - 2x - y$.
g) $f(x,y) = xy(1-x-y)$.
h) $f(x,y) = x^2 + y^2 + \frac{1}{x^2y^2}$
i) $f(x,y) = x^3 + y^3 + 4xy$.

5. Compute the 2nd order partial derivatives of the following functions.

a) $f(x,y) = \frac{1}{\sqrt{x^2+y^2}}$.
b) $f(x,y,z) = xyz$.
c) $f(x,y,z) = \ln\left(\frac{x+y}{x+z}\right)$.
d) $f(x,y) = \frac{x^2+y^2}{1+x}$.
e) $f(x,y,z) = \sqrt{1+x+y-2z}$.
f) $f(x,y,z) = x^2yz^3 + xy^2\sqrt{z}$.
g) $f(x,y) = \frac{xy^2}{\sqrt{x+3}}$.
h) $f(x,y,z) = xz\sqrt{y}$.
i) $f(x,y,z) = x^3\ln(zx)yz^2e^{yx}$.
j) $f(x,y) = xy\sqrt{x^2+7}$.
k) $f(x,y,z) = \frac{1}{xyz}$.
l) $f(x,y,z) = \frac{x^3y-z^2}{3x+y+2z}$.

6. Compute $\int_0^2 \int_0^x e^{x^2}\,dy\,dx$ by first sketching the area of integration.
7. Compute $\int_0^3 \int_{y^2}^9 y\sin(x^2)\,dx\,dy$.
8. What is the volume of the solid bounded by the planes $z = x+2y+4$ and $z = 2x+y$, above the triangle in the $xy$ plane with vertices $A(1,0,0)$, $B(2,1,0)$ and $C(0,1,0)$?
9. Compute $\int_W h\,dV$, where $h(x,y,z) = ax+by+cz$, $W = \{(x,y,z) : 0 \le x \le 1, 0 \le y \le 1, 0 \le z \le 2\}$.
10. Sketch the region of integration $W$ of the triple integral

$$\int_0^1 \int_0^{2-x} \int_0^3 f(x,y,z)\,dz\,dy\,dx.$$

11. What is the volume of the solid defined by the intersection of the two cylinders $x^2 + z^2 = 1$ and $y^2 + z^2 = 1$?
12. Compute $\int_0^{\sqrt{2}} \int_0^{\sqrt{4-y^2}} xy \, dx \, dy$.
13. Compute $\int_W \sin(x^2 + y^2) \, dV$, where $W$ is the cylinder centered about the $z$ axis from $z = -1$ to $z = 3$ with radius 1.
14. Compute

$$\int_0^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{1-x^2-z^2}}^{\sqrt{1-x^2-z^2}} (x^2 + y^2 + z^2)^{-1/2} \, dy \, dz \, dx.$$

15. Compute

$$\int_0^1 \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (x^2 + y^2)^{-1/2} \, dy \, dx \, dz.$$

16. What is the volume of the solid $Q$ directly above the region bounded by $0 \le x \le 1, 1 \le y \le 2$ in the $xy-$plane and below the plane $z = 4 - x - y$?
17. Compute $\int_0^1 \int_{\sqrt{x}}^1 e^{y^3} \, dy \, dx$.
18. Sketch the solid bounded by the the surfaces $z = 0$, $y = 0$, $z = a - x + y$ and $y = a - \frac{1}{a}x^2$, where $a$ is a positive constant. What is the volume of that solid?
19. Evaluate $\int_0^{\ln 2} \int_0^{\ln 5} e^{2x-y} \, dx \, dy$.
20. Evaluate $\int_0^1 \int_0^1 \frac{xy}{\sqrt{x^2+y^2+1}} \, dx \, dy$.
21. Let $D = \{(x, y) : 1 \le y \le e, y^2 \le x \le y^4\}$. Compute $\iint_D \frac{1}{x} \, dA$.
22. What is the volume of the solid lying under the paraboloid $z = x^2 + y^2$ and above the domain bounded by $y = x^2$ and $x = y^2$?
23. Let $R$ be the disk of radius 5, centered at the origin. Evaluate $\iint_R x \, dA$.
24. What is the volume of the solid lying under the cone $z = \sqrt{x^2 + y^2}$ and above the ring $4 \le x^2 + y^2 \le 25$ located in the $xy-$plane?
25. Evaluate $\int_0^3 \int_0^{\sqrt{9-x^2}} \int_0^x yz \, dy \, dz \, dx$.
26. Compute $\iiint_E e^x \, dV$, where

$$E = \{(x, y, z) : 0 \le y \le 1, 0 \le x \le y, 0 \le z \le x + y\}.$$

27. Compute $\iiint_E xz \, dV$, where $E$ is the pyramid with vertices $(0, 0, 0)$, $(0, 1, 0)$, $(1, 1, 0)$ and $(0, 1, 1)$.
28. Let $W$ be a three-dimensional solid. Its volume can be computed by the following iterated integral:

$$V(W) = \int_0^{2\pi} \int_0^2 \int_0^{4-r^2} r \, dz \, dr \, d\theta.$$

Find $W$ and $V(W)$.
29. Compute $\iiint_B (x^2+y^2+z^2) \, dV$, where $B$ is the unit ball $x^2 + y^2 + z^2 \le 1$.
30. Evaluate

$$\int_0^3 \int_0^{\sqrt{9-y^2}} \int_{\sqrt{x^2+y^2}}^{\sqrt{18-x^2-y^2}} (x^2 + y^2 + z^2) \, dz \, dx \, dy.$$

31. Evaluate the integral $\iint_D x^2 y \, dx \, dy$ where $D$ is the region bounded by the curves $y = x^2$ and $x = y^2$ in the first quadrant.
32. Compute the volume of the solid bounded by the cone $z = \sqrt{x^2 + y^2}$ and the sphere of radius $a > 0$ whose center is located at the origin.
33. Compute the volume of the solid bounded by the paraboloïds $z = 10 - x^2 - y^2$ and $z = 2(x^2 + y^2 - 1)$.
34. Compute the area of the planar region bounded by $y = x^2$, $y = 2x^2$, $x = y^2$, and $x = 3y^2$.
35. Find the volume of the solid bounded by the interior of the sphere $x^2 + y^2 + z^2 = a^2$ and the interior of the cylinder $x^2 + y^2 = a^2$, $a > 0$.
36. Find the volume of the solid bounded by the interior of each of the cylinders $x^2 + y^2 = a^2$, $x^2 + z^2 = a^2$ and $y^2 + z^2 = a^2$, $a > 0$.
37. Find the volume of the solid bounded by the interior of the cone $z^2 = x^2 + y^2$ lying above the paraboloïd $z = 6 - x^2 - y^2$.
38. Find the volume of the solid bounded by the plane $z = 3x + 4y$ lying below the paraboloïd $z = x^2 + y^2$.
39. Let $S$ be the sphere of radius $a > 0$ centered at $(0, 0, a)$. Show that $\iiint_S z^2 \, dx \, dy \, dz = \frac{8}{5}\pi a^5$.
40. Compute $\iiint_{\mathbb{R}^3} e^{-(x^2+y^2+z^2)} \, dx \, dy \, dz$.

## Chapter References


[1] P. Boily. *Analysis and Topology Study Aids* . Data Action Lab.

[2] P. Boily, S. Davies, and J. Schellinck. *The Practice of Data Visualization* . Data Action Lab, 2023.

[3] P. Boily and R. Hart. *Le calcul dans la joie* . 2nd ed. 2020.

[4] F. Donzelli. *Multivariable Calculus*. Kendall Hunt, 2022.

[5] M. Nielsen. *Neural Networks and Deep Learning* . 2019.

[6] J. Stewart. *Calculus, Early Transcendentals*. Cengage Learning, 2012.

# Overview of Linear Algebra

# 3

by **Fabrizio Donzelli**, with contributions from **Patrick Boily**

---

This chapter contains an essential introduction to linear algebra. The goal is to provide the readers interested in statistics and/or data science with some basic mathematical tools that are at the base of the algorithms and the mathematical models of statistical analysis. Theoretical details, such as rigorous proofs and definitions, will be kept at a minimal level.

A more detailed introduction to linear algebra can be found in [3, 2].


3.1 **Vector Spaces** . . . . . 147
**Practical Definition** . . . . . 147
**Linear Combinations** . . . . 149
**Bases and Dimension** . . . . 150
**Vector Subspaces** . . . . . 151
**Spanning Sets** . . . . . 153
**Dot Product** . . . . . 153
**Cross Product in $\mathbb{R}^3$** . . . . . 154
3.2 **Linear Transformations** . . 155
3.3 **Matrix Algebra** . . . . . 157
**Matrix Operations** . . . . . 157
**Square Matrices** . . . . . 159
**Determinants** . . . . . 160
3.4 **Linear Systems** . . . . . 162
**Gauss-Jordan Elimination** . 165
**Linear Systems & Matrices** 167
3.5 **Matrix Diagonalization** . . 168
**Eigenvalues & Eigenvectors** 168
**Similar Matrices** . . . . . 174
**Diagonalization** . . . . . 175
3.6 **Exercises** . . . . . 177
**Chapter References** . . . . . 180


## 3.1 Vector Spaces

At its most fundamental level, linear algebra deals with **vector spaces** and **linear transformation** between these.

Linear transformation are represented by **matrices**; a good portion of this chapter will be therefore dedicated to matrix algebra.[1]

### 3.1.1 Practical Definition

While there is a formal definition of vector spaces (see [2], for instance), we will eschew it in these notes. Instead, we use a "recipe" that contains all that we will need.

In the context of linear algebra, the set $\mathbb{R}^n$ is the **$n$-dimensional vector space**, consisting of **$n$-dimensional vectors**.[2]

Here are the key defining properties of these vectors:

- a $n$-dimensional vector $\mathbf{v}$ is a collection of $n$ numbers: $\mathbf{v} = \langle v_1, \cdots, v_n \rangle$, where the numbers $v_k$ are the **components** of the vector;[3]
- vectors belonging to the same vector space can be added, while remaining a part of that vector space: the **vector sum** of $\mathbf{v} = \langle v_1, v_2, \cdots, v_n \rangle$ and $\mathbf{w} = \langle w_1, w_2, \cdots, w_n \rangle$, is

$$\mathbf{v} + \mathbf{w} = \langle v_1 + w_1, v_2 + w_2, \cdots, v_n + w_n \rangle;$$

- in vector algebra, simple numbers are **scalars** – the **multiplication of a vector by a scalar** is defined in the "obvious way": if $c$ is a scalar, and $\mathbf{v} = \langle v_1, v_2, \cdots, v_n \rangle$ is a vector, then

$$c\mathbf{v} = \langle cv_1, cv_2, \cdots, cv_n \rangle;$$

- the zero $n$-dimensional vector is denoted by $\mathbf{0} = \langle 0, 0, \cdots, 0 \rangle$.

1: Note that the order in which the material covered by a first year university linear algebra course could be different than the order presented here – it is common for texts of this nature to start with linear systems before moving to vector spaces; this is not how we will approach the presentation, in no small part because the language of vectors is very useful, not only in mathematics, but also in coding. A mastery of this language makes mathematical modeling more accessible, in general.

2: This definition is not ideal since it implicitly assumes that the vector are expressed with respect to the **standard basis**

$$\begin{aligned} \mathbf{e}_1 &= \langle 1, 0, \cdots, 0 \rangle, \\ \mathbf{e}_2 &= \langle 0, 1, \cdots, 0 \rangle, \\ &\vdots \\ \mathbf{e}_n &= \langle 0, 0, \cdots, 1 \rangle. \end{aligned}$$

3: In the other chapters, we will use $(v_1, \ldots, v_n)$ when the context is clear.

**Example** An aircraft is flying from Ottawa to Milan. The direction and its speed are determined by three values that change over time: latitude $x(t)$, longitude $y(t)$, and altitude $z(t)$. Hence, the velocity of the aircraft is modeled using a 3-dimensional vector $\mathbf{v}(t) = \langle x'(t), y'(t), z'(t) \rangle \in \mathbb{R}^3$.

Note however that the 3 quantities $x(t)$, $y(t)$, and $z(t)$ are not truly Cartesian in nature, since longitude and latitude are described by angles. Locally, however,[4] this $\mathbb{R}^3$ model is a good approximation, assuming that the Earth is **locally flat**.

4: That is to say, as long as we do not look at long distance trajectories, say.

**Example** A boat is sailing in the Pacific Ocean with a velocity vector $\mathbf{v} = \langle 1, 2 \rangle$. At some point the wind starts blowing with speed $\mathbf{w} = \langle 2, 4 \rangle$, helping the boat to sail faster. What is the estimate of the effective velocity of the boat under the influence of the wind?

We need to add the vectors. Luckily for us, velocities add **linearly**, hence the velocity of the wind-boosted boat is

$$\mathbf{v}_{\text{tot}} = \mathbf{v} + \mathbf{w} = \langle 1, 2 \rangle + \langle 2, 4 \rangle = \langle 3, 6 \rangle.$$

The result is only an approximation of the real situation, since in reality there are dissipation effects that may reduce the speed of the boat.[5] ■

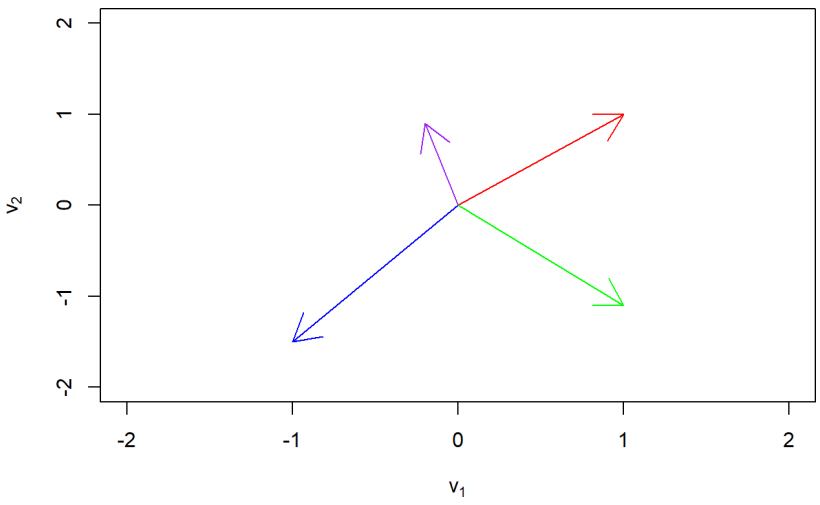
5: But that is a problem for engineers, really, and we will sidestep the challenge simply by ignoring it.

While vectors can be of arbitrary dimension, having a low-dimensional geometric picture helps strengthen vector intuition, which may be otherwise sound too abstract. In practice, vectors in $\mathbb{R}^2$, $\mathbb{R}^3$ are represented by arrows, emanating from the same origin point.

**Example** Here is an example of a representation of 2-dimensional vectors, which include a basic R script that produces the picture.[6]

6: Which can be improved, see Chapter 18 and [1].

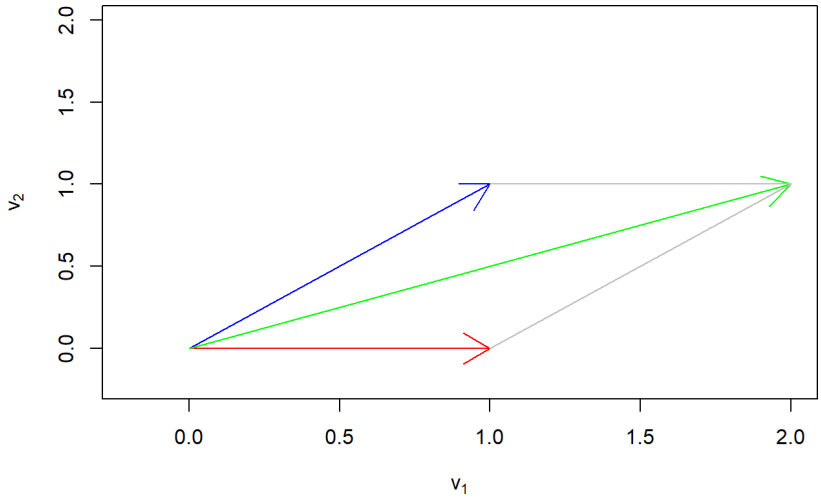
```
plot(NA,xlim=c(-2,2), ylim=c(-2,2),
    xlab = expression(list(v[1])),
    ylab=expression(list(v[2])))
arrows(0,0,1,1, col="red"); arrows(0,0,-1,-1.5, col="blue")
arrows(0,0,1,-1.1, col="green"); arrows(0,0,-0.2,0.9, col="purple")
```



In principle, arrows exist in arbitrary dimensions, but they are difficult to visualize. As we can always represent a vector as an arrow, the next rule applies no matter the dimension $n$.

**Parallelogram rule**: the sum of two vectors $\mathbf{v}$ and $\mathbf{w}$ is the diagonal of the parallelogram generated by $\mathbf{v}$ and $\mathbf{w}$, emanating from the origin:

```
plot(NA,xlim=c(-0.2,2), ylim=c(-0.22,2),
    xlab = expression(list(v[1])),
    ylab=expression(list(v[2])))
arrows(0,0,1,0, col="red")
arrows(0,0,1,1, col="blue")
arrows(0,0,2,1, col="green")
segments(1,1,2,1, col="grey")
segments(1,0,2,1, col="grey")
```




### 3.1.2 Linear Combinations

Given a finite collection of $n$-dimensional vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ and scalar coefficients $c_1, c_2, \cdots, c_k$, the vector

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k$$

is called the **linear combination** of the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ with coefficients $c_1, c_2, \cdots, c_k$.

**Example** Show that the vector $\langle 2, 3\rangle$ can be written as a linear combination of $\mathbf{e}_1 = \langle 1, 0\rangle$ and $\mathbf{e}_2 = \langle 0, 1\rangle$.

This problem can be set up and solved using an algorithm that solves a system of linear equations.[7]

7: See Section 3.4.

However, the situation at hand is a simpler matter of applying the definition of linear combination. We see that we can express

$$\langle 2, 3\rangle = \langle 2, 0\rangle + \langle 0, 3\rangle = 2\langle 1, 0\rangle + 3\langle 0, 1\rangle = 2\mathbf{e}_1 + 3\mathbf{e}_2.$$

### 3.1.3 Bases and Dimension

As we mentioned previously, the components of a vector are not defined in a "universal way", but they depend on the choice of a set of "reference vectors", which form a **basis**: a set of vectors which cover once and only once all possible **independent** directions of the vector space.

Let $V$ be a vector space, and let $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k\}$ be a finite list of vectors in $V$. We say that the vectors are **linearly independent** if:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \text{ if and only if } c_1 = c_2 = \cdots = c_k = 0.$$

Otherwise, we say that they are **linearly dependent**.

If we expand the equation above, we see that the condition of linear independence is equivalent to state that the **homogeneous linear system** (see Section 3.4)

$$\begin{pmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k\end{pmatrix} \cdot \begin{pmatrix}c_1 & c_2 & \cdots & c_k\end{pmatrix}^T = \begin{pmatrix}0 & 0 & \cdots & 0\end{pmatrix}^T$$

only has the trivial solution $c_1 = c_2 = \cdots = c_k = 0$.

We can also view linear dependence is as follows. Suppose, for instance, that we have three vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ related by a linear dependence relation. For example, let us assume that

$$\mathbf{v}_1 - \mathbf{v}_2 - \mathbf{v}_3 = \mathbf{0}.$$

Then we can rewrite this expression as

$$\mathbf{v}_1 = \mathbf{v}_2 + \mathbf{v}_3,$$

which provides an intuition for the idea of linear dependence: one (or more) vector in the collection can be reconstructed as a linear combination of the remaining vectors.

A **basis** of a vector space $V$ is a collection of vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ such that: + The vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are linearly independent. + Every vector $v \in V$ can be expressed **in a unique way** as a linear combination of the basis element $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$.

Note that the linear combination expressed from a basis is **unique**,[8] that is the coefficients $c_1, c_2, \cdots, c_n$ of the equation

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

are uniquely determined.

8: Because the basis vectors are linearly independent.

While a vector space $V$ has more than one basis, **all of its bases have the same cardinality**, meaning that all bases have the same number of vectors. This number $n$ is the **dimension** of the vector space.

The vector space $\mathbb{R}^n$ is $n-$dimensional; we usually (but not always) represent vectors with respect to the **standard basis** $\{\mathbf{e}_1, \ldots, \mathbf{e}_2\}$.

The uniqueness of the expression of a vector as a linear combination of basis vectors explains why we can interpret the components of the vector as coordinates.

**Example** Determine if the following 4 vectors form a basis in $\mathbb{R}^4$:

$$\begin{aligned}\mathbf{v}_1 &= \langle 1,0,0,0\rangle\\ \mathbf{v}_2 &= \langle 1,1,1,1\rangle\\ \mathbf{v}_3 &= \langle 1,0,1,-2\rangle\\ \mathbf{v}_4 &= \langle 0,1,0,-1\rangle\end{aligned}$$

We need to solve the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4, = \langle 0,0,0,0\rangle$, which unwraps into:

$$\begin{aligned}c_1 &= 0\\ c_1 + c_2 + c_3 + c_4 &= 0\\ c_1 + c_3 - 2c_4 &= 0\\ c_2 - c_4 &= 0\end{aligned}$$

Is it clear that the only solution is the trivial one $c_1 = c_2 = c_3 = c_4 = 0$? We will discuss how to demonstrate that it is indeed the only solution in Section 3.4.[9]

9: We can also verify linear independence using the properties of determinants (see Section 3.3.3).

**Example** Show with an example that there can be infinitely many bases for a vector space of positive dimension.

For each $\theta \in [0, 2\pi)$, the set

$$B_\theta = \{\langle \cos\theta, \sin\theta\rangle, \langle -\sin\theta, \cos\theta\rangle\}$$

is a basis of $\mathbf{R}^2$. ■

We will not discuss infinite dimensional vector spaces (that's a topic for advanced courses), but we provide one such example, for curiosity's sake.

**Example** The space

$$\mathbb{P}[x] = \{a_0 + a_1 x + \cdots + a_k x^k \mid a_i \in \mathbb{R}, k \in \mathbb{N}\}$$

of all polynomials in one variable $x$ over the reals is an infinite dimensional vector space; the vectors are polynomials. For all $n \in \mathbb{N}$, the monomials $1, x, x^2, \cdots, x^n$ are linearly independent for all $n$, so there are infinitely many linearly independent vectors in $\mathbb{P}[x]$.[10]

10: This example is interesting not just because it deals with an infinite-dimensional vector space, but also because it shows that the notion of vector space applies beyond the intuitive geometric notion of arrows represented in vector components.

### 3.1.4 Vector Subspaces

The space $W = \mathbb{R}^2$ consists of vectors of the form $\langle x, y\rangle$. The space $V = \mathbb{R}^3$ consists of vectors of the form $\langle x, y, z\rangle$. We can interpret $W$ as a smaller vector space contained in $V$, from which it inherits the operations of sum and multiplication by scalar.

**Example** Show that a linear combination of 2-dimensional vectors of the form $\langle x, y, 0\rangle$ has the same form (i.e., the third component remains zero).

This is a classic problem that looks hard the first time we learn linear algebra, but in fact the solution consists a simple check. Take two arbitrary vectors $\mathbf{v}_1 = \langle x_1, y_1, 0\rangle$ and $\mathbf{v}_2 = \langle x_2, y_2, 0\rangle$. Then, for arbitrary scalars $a, b$, the linear combination of them has the expression

$$a\mathbf{v}_1 + b\mathbf{v}_2 = a\langle x_1, y_1, 0\rangle + b\langle x_2, y_2, 0\rangle = \langle ax_1 + bx_2, ay_1 + by_2, 0\rangle,$$

of the form $\langle x, y, 0\rangle$, if we let $x = ax_1 + bx_2$ and $y = ay_1 + by_2$. ■

Let $V$ be a vector space, and $W \subset V$, a subset of $V$: we say that $W$ is a **vector subspace** (subspace in short) of $V$, denoted $W < V$, if $W$ is a vector space itself (which inherits the operations from the bigger space $V$ in which it is contained).

In particular, if $W < V$, and $\mathbf{v}, \mathbf{w} \in W$ and $a, b \in \mathbb{R}$, then:

- $\mathbf{0} \in W$, and
- $a\mathbf{v} + b\mathbf{w} \in W$.

Note that, by definition, $V$ is a subspace of itself.

The result of the previous example can be recast as $\mathbb{R}^2$ being a vector subspace of $\mathbb{R}^3$.

**Example** Let $V$ be a vector space. What is the "largest" subspace of $V$? What is the "smallest" subspace of $V$?

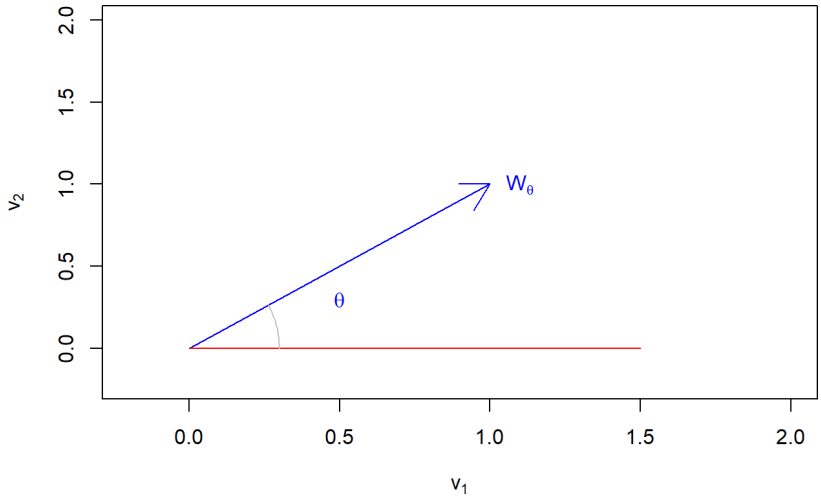
As $V \subseteq V$ is itself a subspace of $V$, it is also the largest subspace of $V$. The smallest subspace of $V$ is the zero-dimensional vector space $\{\mathbf{0}\}$, which consists solely of the zero vector. ■

Let $V$ be a vector space of dimension $n$. Then, it should be intuitive that if $W$ is a subspace of $V$, then $\dim(W) \leq \dim(V)$.

The zero space from the previous example is the only zero-dimensional vector subspace; the space $V$ itself is the only subspace of maximal dimension $n$. There are infintely many "intermediate dimension" (proper) subspaces as soon $\dim V \leq 2$.

**Example** Let $W_\theta = \{a\langle \cos\theta, \sin\theta\rangle \mid a \in \mathbb{R}\} < \mathbb{R}^2$, $\theta \in [0, 2\pi)$. For each angle value $\theta$, the vector $\langle \cos\theta, \sin\theta\rangle$ gives a different direction, hence $W_{\theta_1} = W_{\theta_2}$ if and only if $\theta_1 = \theta_2$.

```
library(plotrix)
plot(NA,xlim=c(-0.2,2), ylim=c(-0.22,2),
     xlab = expression(list(v[1])),
     ylab=expression(list(v[2])))
arrows(0,0,1,1, col="blue")
segments(0,0,1.5,0, col="red")
draw.arc(0,0,0.3,0,0.5,col="grey")
text(0.5,0.3,expression(theta),col="blue")
text(1.1,1,expression(list(W[theta])),col="blue")
```




### 3.1.5 Spanning Sets

How do we "create" subspaces? As long as we do not worry too much about "clean production", we take a finite set of vectors of a given vector space $V$, and consider **all possible linear combination** of such vectors.

Let $V$ a vector space and $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N \in V$. The **spanning set**

$$\text{Span}\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N\} = \{a_1\mathbf{v}_1 + \cdots + a_N\mathbf{v}_N \mid a_i \in \mathbb{R}\} < V.$$

**Example** Let $V$ be a vector space and $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N \in V$. Then $\mathbf{v} \in \text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ if and only if $\mathbf{v} = c_1\mathbf{v}_1 + \cdots c_n\mathbf{v}_N$, for some coefficients $c_1, c_2, \cdots, c_N \in \mathbb{R}$.

This is a "trivial" statement – we simply translated the condition "belonging to span" into the equation "$\mathbf{v}$ is a linear combination of the spanning vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N$".[11]

11: Being trivial, it can still cause confusion at the beginning; but it is crucial to learn how to translate math-related sentences into formulas or equations.

The problem with the definition of the spanning set of a collection of vectors is that it says nothing about the **dimension of the vector space**.

**Example** Let $V = \mathbb{R}^2$. We can write $V = \text{Span}\{\langle 1, 0\rangle, \langle 0, 1\rangle\}$, which makes sense since the two vectors form a basis of $V$. However, we can also generate the entire vector space with three vectors, so that the number of vectors is not linked to the dimension: $V = \text{Span}\{\langle 1, 0\rangle, \langle 0, 1\rangle, \langle 1, 1\rangle\}$.

### 3.1.6 Dot Product

The dot product of two vectors is a scalar quantity that in some sense measure how much of their components two vectors share. The **dot** (or scalar) **product** of two $n$-dimensional vectors $\mathbf{v} = \langle v_1, v_2, \cdots, v_n\rangle$ and $\mathbf{w} = \langle w_1, w_2, \cdots, w_n\rangle$:

$$\mathbf{v} \cdot \mathbf{w} = v_1w_1 + v_2w_2 + \cdots + v_nw_n.$$

From the dot product, we can define the Euclidean **length** (or **norm**) of a vector $\mathbf{v}$:

$$\|\mathbf{v}\| = \sqrt{\mathbf{v}\cdot\mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

Two vectors $\mathbf{v}$ and $\mathbf{w}$ are **orthogonal** if and only if $\mathbf{v}\cdot\mathbf{w} = 0$. In general:

$$\mathbf{v}\cdot\mathbf{w} = \|\mathbf{v}\|\|\mathbf{w}\|\cos(\theta),$$

where $\theta$ is the angle formed by the vector $\mathbf{v}$ and $\mathbf{w}$.[12]

12: In fact, this is how we define the angle between two vectors when the geometrical interpretation is unavailable to us.

Two non-zero vectors $\mathbf{v}, \mathbf{w}$ create two angles, $\theta$ and $2\pi - \theta$: does the dot product depends on the choice between the two angles?

No, because for all angles $\theta$ we have:

$$\cos(2\pi - \theta) = \cos(\theta).$$

**Example** Find the (smallest) angle $\theta$ formed by the vectors $\mathbf{v} = \langle 1, 2\rangle$ and $\mathbf{w} = \langle -1, 1\rangle$.

It's a one line calculation:

$$\theta = \arccos\left(\frac{\mathbf{v}\cdot\mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}\right) = \arccos\left(\frac{-1+2}{\sqrt{1+4}\sqrt{1+1}}\right) = \arccos\left(\frac{1}{\sqrt{10}}\right) = 1.25 \text{ radians}.$$

**Example** Let $t$ be a real parameter Find the vectors of the form $\langle 1, t\rangle$ and with length equal to 5.

The general vector $\langle 1, t\rangle$ has length

$$\|\langle 1, t\rangle\| = \sqrt{1^2 + t^2} = \sqrt{1 + t^2}.$$

We look for the values of $t$ such that

$$\|\langle 1, t\rangle\| = \sqrt{1 + t^2} = 5,$$

which are found by solving the **quadratic equation**:

$$\sqrt{1+t^2} = 5 \implies 1 + t^2 = 25 \implies t^2 = 24 \implies t = \pm\sqrt{24} = \pm 2\sqrt{6}.$$

As expected, there are two vectors $\langle 1, t\rangle$ of length 5: $\langle 1, \pm 2\sqrt{6}\rangle$.

### 3.1.7 Cross Product in $\mathbb{R}^3$

The dot product is also called scalar product, since it outputs a scalar from two given vectors. The cross (or vector) product, which will define below, produces a new vector out of two input vectors.

Given two 3-dimensional vectors $\mathbf{v} = \langle v_1, v_2, v_3, \rangle$ and $\mathbf{w} = \langle w_1, w_2, w_3, \rangle$, the **cross** (or vector) **product** formula can be symbolically represented

with the help of a determinant:

$$\mathbf{v} \times \mathbf{w} = \det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} = \det \begin{pmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}$$
$$= \langle v_2 w_3 - v_3 w_2, -(v_1 w_3 - v_3 w_1), v_1 w_2 - v_2 w_1 \rangle.$$

Note that we left the formula without multiplying out negative sign in front of the second entry, in order to remind the reader that the determinant is an alternating sum.[13]

13: See Section 3.3.3.

Whereas the dot product can be extended to vector space of all dimensions, the cross product is only defined on $\mathbb{R}^3$.

## 3.2 Linear Transformations and Matrices

A **matrix** of size $m \times n$ is a collection of $m \times n$ numbers aligned along $m$ rows and $n$ columns:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

We refer to matrices of size $n \times n$ as **square matrices** of size $n$.

Let $V$ and $W$ be two vector spaces (of arbitrary dimension, possibly infinite-dimensional). A **linear map** $T : V \to W$ is a function that preserves linear combinations of vectors:

$$T(a\mathbf{v} + b\mathbf{w}) = aT(\mathbf{v}) + bT(\mathbf{w}), \text{ for all } a, b \in \mathbb{R} \text{ and for all } \mathbf{v}, \mathbf{w} \in V.$$

Given a basis $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ of $V$ and a basis $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$ of $W$, we can construct the **matrix elements** $t_{i,j}$ of the matrix representing the linear transformation $T$ with respect to the given bases. In fact, there are coefficients $T_{ij}$ such that

$$T(\mathbf{v}_i) = \sum_{j=1}^{m} t_{i,j} \mathbf{w}_j$$

We will use the **convention** that a matrix is given with respect to the **standard basis**.

A linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ is represented by **matrix-vector multiplication**. We write the vectors of $\mathbb{R}^n$ and $\mathbb{R}^m$ as **column vectors**:

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} \in \mathbb{R}^m.$$

The vector-matrix multiplication defines the linear map $T(\mathbf{v}) = \mathbf{w}$ (relative to bases of $\mathbb{R}^n$ and $\mathbb{R}^m$):

$$\begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} = \begin{pmatrix} t_{1,1} & \cdots & t_{1,n} \\ \vdots & \ddots & \vdots \\ t_{m,1} & \cdots & t_{m,n} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} t_{1,1}v_1 + \cdots + t_{1,n}v_n \\ \vdots \\ t_{m,1}v_1 + \cdots + t_{m,n}v_n \end{pmatrix}.$$

Linear maps can be composed in the same way as regular functions, assuming that the range of the second is in the domain of the first.

If $T : \mathbb{R}^n \to \mathbb{R}^m$ and $S : \mathbb{R}^m \to \mathbb{R}^p$ are two linear maps, then the **composition** of $S$ and $T$ (the order is important) is the linear map $S \circ T : \mathbb{R}^n \to \mathbb{R}^p$ defined by

$$(S \circ T)(\mathbf{v}) = S(T(\mathbf{v})) = (ST)\mathbf{v}.$$

If the maps $S$ and $T$ are represented by the matrices S=

$$\begin{pmatrix} s_{1,1} & \cdots & s_{1,m} \\ \vdots & \ddots & \vdots \\ s_{p,1} & \cdots & s_{p,m} \end{pmatrix} \text{ and } T = \begin{pmatrix} t_{1,1} & \cdots & t_{1,n} \\ \vdots & \ddots & \vdots \\ t_{m,1} & \cdots & t_{m,n} \end{pmatrix},$$

then the composite map corresponds to the matrix obtained by **matrix multiplication** (or **matrix product**)

$$ST = \begin{pmatrix} s_{1,1} & \cdots & s_{1,m} \\ \vdots & \ddots & \vdots \\ s_{p,1} & \cdots & s_{p,m} \end{pmatrix} \begin{pmatrix} t_{1,1} & \cdots & t_{1,n} \\ \vdots & \ddots & \vdots \\ t_{m,1} & \cdots & t_{m,n} \end{pmatrix} = \begin{pmatrix} s_{1,1}t_{1,1} + \cdots + s_{1,m}t_{m,1} & \cdots & s_{1,1}t_{1,n} + \cdots + s_{1,m}t_{m,n} \\ \vdots & \ddots & \vdots \\ s_{p,1}t_{1,1} + \cdots + s_{p,m}t_{m,1} & \cdots & s_{p,1}t_{1,n} + \cdots + s_{p,m}t_{m,n} \end{pmatrix}$$

Note that the formula of matrix multiplication can be more easily understood using dot products:

$$(st)_{ij} = (\text{row } i \text{ of S}) \cdot (\text{column } j \text{ of T}).$$

**Example** For any angle value in radians, measured counterclockwise with respect to reference to the positive $x$-axis, the matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

rotates vectors in the $xy$-plane around the origin by an angle $\theta$. For instance, we can rotate the vector $\langle 1, 0\rangle$ by $\frac{\pi}{4}$ (45 degrees counterclockwise):

$$R_{\pi/4}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}.$$

The new vector has the same length as the original one, in agreement with the fact that rotations do not change lengths, and it forms an angle of 45 degrees with respect to the positive $x$-axis.

For a fixed $\theta$, the rotation is a linear map:

$$R_\theta(a\mathbf{v} + b\mathbf{w}) = aR_\theta(\mathbf{v}) + bR_\theta(\mathbf{w}) \qquad \text{(prove it!)}.$$

## 3.3 Matrix Algebra

We have already introduced matrix multiplication as a way to define the composition of two compatible linear maps. In this section we collect the all essential rules of matrix algebra. We start with operations that make sense for all matrices, and then specialize to operations that are defined only for square matrices. For convenience we report again the definition of matrix multiplication.

### 3.3.1 Matrix Operations

#### Matrix Multiplication

Formally, let $A \in \mathbb{M}_{m,n}$ (i.e., $A$ is a $m \times n$ matrix) and $B \in \mathbb{M}_{n,p}$ (i.e., $B$ is a $n \times p$ matrix). Then the **matrix product** of $A$ by $B$ is the matrix $AB \in \mathbb{M}_{m,p}$ (i.e., $AB$ is of size $m \times p$), where the entries $(ab)_{i,j}$ are

$$(ab)_{i,j} = \text{ row } i \text{ of } A \cdot \text{ column } j \text{ of } B.$$

Unlike multiplication between scalars, the product of matrices is not generally commutative – assuming that both $AB$ and $BA$ exist, it is not always the case that $AB = BA$.

**Example** If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$, then

$$AB = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix} \neq \begin{pmatrix} 23 & 34 \\ 31 & 46 \end{pmatrix} = BA.$$

The matrix product $AB$ is only defined when the number of columns of $A$ is equal to the number of rows of $B$:

$$\underbrace{A}_{m\times n}\underbrace{B}_{n\times p} = \underbrace{AB}_{m\times p}.$$

The dot product of two $n$-dimensional vector can also be understood in term of matrix multiplication: if we represent $\mathbf{v}$ as a row vector, and $\mathbf{w}$ as a column vector, then

$$\mathbf{v} \cdot \mathbf{w} = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n.$$

#### Matrix Addition

Given two matrices $A, B \in \mathbb{M}_{m,n}$, their **sum** is the matrix $A + B \in \mathbb{M}_{m,n}$ obtained by adding $A$ and $B$ entry-by-entry, that is

$$(a + b)_{i,j} = a_{i,j} + b_{i,j}.$$

**Example** If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$, then $A + B = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$.

Note that, unlike matrix multiplication, matrix addition is commutative: $A + B = B + A$ for all compatible matrices.

### Multiplication by a Scalar

For any matrix $A \in \mathbb{M}_{m,n}$ and scalar $c \in \mathbb{R}$, the **scalar multiplication** of $A$ by $c$ is the matrix $cA \in \mathbb{M}_{m,n}$ whose entries are the entries of $A$ scaled by the factor $c$, that is:

$$(ca)_{i,j} = ca_{i,j}.$$

**Example** If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $c = -2$, then $cA = \begin{pmatrix} -2 & -4 \\ -6 & -8 \end{pmatrix}$.

### Transpose of a Matrix

The **transpose** of $A \in \mathbb{M}_{m,n}$ is the matrix $A^\top \in \mathbb{M}_{n,m}$ whose columns are the rows of $A$:

$$(a^\top)_{i,j} = a_{j,i}.$$

**Example** If $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, then $A^\top = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$.

The transpose is a linear operation: $(A + B)^\top = A^\top + B^\top$ for all compatible matrices $A, B$. However, it behaves "unexpectedly" with respect to matrix multiplication: $(AB)^\top = B^\top A^\top$ for all compatible matrices $A, B$.[14]

14: While this is not a proof, we see that this formula is at the very least aligned with the compatibility of matrix multiplication: if $A \in \mathbb{M}_{m,n}$ and $B \in \mathbb{M}_{n,p}$, then $AB \in \mathbb{M}_{m,p}$ and $(AB)^\top \in \mathbb{M}_{p,m}$. Since $B^\top \in \mathbb{M}_{p,n}$ and $A^\top \in \mathbb{M}_{n,m}$, we see that $B^\top A^\top$ is always defined, but that $A^\top B^\top$ is only defined when $m = p$.

### Matrix Spaces

The **column space** of a matrix $A = [A_1 \mid \cdots \mid A_n] \in \mathbb{M}_{m,n}$ is the vector subspace of $\mathbb{R}^m$ spanned by the column vectors of $A$:

$$\text{colsp}(A) = \text{Span}\{A_1, \cdots, A_n\}.$$

The **rank** of $A$ is the dimension of $\text{colsp}(A)$. If we interpret $A$ as a linear map (as discussed in Section 3.2), then $\text{colsp}(A)$ is in fact the **image** of this map:

$$\text{Im}(A) = \{A\mathbf{v} \mid \mathbf{v} \in \mathbb{R}^n\} < \mathbb{R}^m.$$

The **nullspace** (or **kernel**) of $A$ is the vector subspace of $\mathbb{R}^m$ that are mapped to the null vector $\mathbf{0}$ by $A$:

$$\text{nullsp}(A) = \ker(A) = \{\mathbf{v} \in \mathbb{R}^m \mid A\mathbf{v} = \mathbf{0}\} \subseteq \mathbb{R}^m.$$

That these two sets are indeed vector subspaces of $\mathbf{R}^m$ is clear:

- $\mathbf{0} \in \text{Im}(A), \ker(A)$ since $A\mathbf{0} = \mathbf{0}$;[15]

15: The null vector pulls double-duty here.

- if $\mathbf{v}, \mathbf{w} \in \ker(A)$, $a, b \in \mathbb{R}$, then $a\mathbf{v} + b\mathbf{w} \in \ker(A)$ since

$$A(a\mathbf{v} + b\mathbf{w}) = aA\mathbf{v} + bA\mathbf{w} = a\mathbf{0} + b\mathbf{0} = \mathbf{0};$$

- if $\mathbf{v}, \mathbf{w} \in \mathrm{Im}(A)$, $a, b \in \mathbb{R}$, then $a\mathbf{v} + b\mathbf{w} \in \mathrm{Im}(A)$ since there exists $\mathbf{u}, \mathbf{z} \in \mathbb{R}^n$ such that $A\mathbf{u} = \mathbf{v}$ and $A\mathbf{z} = \mathbf{w}$, and so

$$a\mathbf{v} + b\mathbf{w} = aA\mathbf{u} + bA\mathbf{z} = A(a\mathbf{u} + b\mathbf{z}).$$

In particular, neither of these spaces is empty since they always contain at least $\mathbf{0}$.

**Rank-Nullity Theorem**

Let $A \in \mathbb{M}_{m,n}$; then

$$\dim(\ker(A)) + \dim(\mathrm{Im}(A)) = m.$$

This theorem is a basic (and very useful) result of linear algebra, with counterparts in other sectors of algebra (such as group theory).

### 3.3.2 Square Matrices

The **identity matrix** of size $n$ is the square matrix, denoted by $\mathbf{I}_n$:

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

The **diagonal** of a square matrix $A$ is the list of elements $A_{ii}$ (that is, the values along the diagonal).

A square matrix is said to be a **diagonal matrix** if the non-diagonal entries are all zero.

A square matrix $A$ is said to be **symmetric** if $A = A^\top$. In fact, the entries are symmetric with respect to the diagonal of the matrix.

A square matrix $A$ of size $n$ is said to be **invertible** (or non-singular) if there exists a matrix, denoted by $A^{-1}$, such that $AA^{-1} = A^{-1}A = \mathbf{I}_n$. The matrix $A^{-1}$ is called the **inverse** of $A$. Note that the inverse of $A^{-1}$ is $A$ (in other words, $(A^{-1})^{-1} = A$).

If $A$ is invertible, then

$$(A^{-1})^\top = (A^\top)^{-1}.$$

If $A$ and $B$ are both invertible (and have the same size), then

$$(AB)^{-1} = B^{-1}A^{-1}.$$

We will discuss a way to compute the inverse of a non-singular matrix in Section 3.4.2.

### 3.3.3 Determinants

There is an important numerical value that can be associated to any square matrix $A$, its **determinant** $\det(A)$.

When we work with large-sized matrices, we rely on a computer program to compute the determinant. However, we need to know what it is and how to compute it for small size examples.

The purely algebraic definition of the determinant makes use of the language of multilinear algebra, which will not discuss here; instead, we proceed with a computational definition.

- For a scalar $a \in \mathbb{R} = \mathbb{M}_{1,1}$, $\det(a) = a$.
- For $A \in \mathbb{M}_{2,2}$,

$$\det(A) = \det\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}.$$

- For $A \in \mathbb{M}_{n,n}$, consider the first row, which consists of the elements $[a_{1,1}, a_{2,1}, \cdots, a_{1,n}]$. Let $M_{1,k}$ be the square matrix of size $n-1$ obtained by removing from $A$ the row and column passing through $a_{1,k}$. Then the determinant of $A$ is the **alternating sum**:

$$\det(A) = \det(M_{1,1}) - \det(M_{1,2}) + \cdots + (-1)^{n+1} \det(M_{1,n})$$

  The quantities $\det(M_{i,j})$ are called the **minors** of the matrix.

In fact, we can pick any row or any column and apply the alternating sum formula as above. However, we need to be careful about the sign in front of the minor $\det(M_{i,j})$, which is called the **cofactor** $C_{i,j}$:

$$C_{i,j} = (-1)^{i+j} \det(M_{i,j}).$$

For more details about the general formula, we refer to [3].

#### Properties

The determinant determines important properties of a square matrix.

- The determinant of a **diagonal** matrix is the product of its diagonal entries.
- The determinant behaves nicely when it comes to matrix multiplication and inversion (assuming $A$ and $B$ are both square and of the same size):

$$\det(AB) = \det(A)\det(B),$$

  and, if $A$ is invertible, then

$$\det(A^{-1}) = \det(A)^{-1},$$

- The determinant is **invariant under transposition**:

$$\det(A) = \det(A^\top).$$

- Let $A$ be a square matrix and let $A[R_i \leftrightarrow R_j]$ (resp. $A[C_i \leftrightarrow C_j]$) be the matrix obtained by interchanging row $i$ with row $j$ (resp. column $i$ with column $j$). Then

$$\det(A[C_i \leftrightarrow C_j]) = -\det(A)$$
$$\det(A[R_i \leftrightarrow R_j]) = -\det(A)$$

- More generally, if we perform an odd number of permutations of rows (columns), the determinant changes sign; if we perform an even number of permutations of rows (columns), the determinant stays the same.

Let $A$ be a square matrix, of size $n$. Then the following conditions are equivalent:

1. $\det(A) \neq 0$;
2. $A$ is invertible;
3. the $n$ column vectors of $A$ are linearly independent, hence they form a basis of $\mathbb{R}^n$;
4. the $n$ row vectors of $A$ are linearly independent, hence they form a basis of $\mathbb{R}^n$;
5. the rank of $A$ is $n$ (maximal rank);
6. the nullspace (kernel) of $A$ consists only of the zero vector $\mathbf{0}$.

**Examples** Determine if the following matrices are invertible or not, without computing the inverse.

1. $A = \begin{pmatrix} 2 & 3 \\ -1 & -3 \end{pmatrix}$ is invertible, since $\det A = 2(-3) - 3(-1) = -2 \neq 0$.
2. $B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 3 & 2 & 1 \\ 1 & 2 & 3 & 4 \\ -1 & 1 & -1 & 1 \end{pmatrix}$ is not invertible, since the first and third rows are equal (and so they are linearly dependent).
3. $C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 2 & 1 \\ 2 & 3 & 5 & 5 \\ -1 & 1 & -1 & 1 \end{pmatrix}$ is not invertible, as we can see either by computing that $\det C = 0$, or by observing that $R_1 + R_2 = R_3$.
4. $D = \begin{pmatrix} 1 & 42 & 0.12 & 4 \\ 0 & 1 & -2 & 21 \\ 1.2 & 23 & 0.5 & 5 \\ -2.2 & 1 & 0 & -0.55 \end{pmatrix}$ is invertible as can be seen in the following R code.

```
D <- rbind(c(1,42,0.12,4),c(0,1,-2,21),c(1.2,23,0.5,5),c(-2.2,1,0,-0.55))
det(D)
```

```
[1] -1336.74
```

5. Suppose that $A$ and $B$ are square matrices of the same size, and that $\det(A) = 3$, $\det(B) = -5$; then

$$\det(A^{-1}B^3A) = \frac{1}{\det(A)} \cdot (\det(B)^3) \cdot \det(A) = (\det(B))^3 = (-5)^3 = -125.$$

There is a closed-form formula for finding the inverse of a square matrix of arbitrary size. Computing the inverse can be very time consuming, and, when the matrices are very large (thousands of entries), we typically consider numerical methods.

But it is convenient to at least remember how to find the inverse of a $2\times 2$ matrix.[16]

16: For inversion of matrices of arbitrary size, we refer to [3]. We mention in passing that the general formula for $A^{-1}$ contains a factor $\frac{1}{\det A}$, re-descovering the fact zero-determinant matrices can not be inverted.

For a $2\times 2$ matrix $A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$, say, the inverse (when it exits) is

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{pmatrix}$$

The formula of the inverse starts with $\frac{1}{\det A}$. If the determinant of $A$ is non-zero, but **close to** zero, we could have issues with the finite precision arithmetic.

We will discuss a row-reduction method to compute the inverse of a general non-singular matrix in the next section.

**Example** Let $A$ and $B$ be the following matrices:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & -2 \\ 8 & 1 \end{pmatrix}$$

Solve the equation $AX = B$ for $X$, where $X \in \mathbb{M}_{2,2}$. We see that

$$AX = B \Rightarrow X = A^{-1}B,$$

but is $A$ invertible? A quick check using the determinant confirms that it is since $\det(A) = 1\cdot 1 - 2\cdot 3 = -5 \neq 0$. Using the formula of the inverse of a $2\times 2$ matrix we obtain:

$$A^{-1} = \frac{1}{1\cdot 1 - 2\cdot 3}\begin{pmatrix} 1 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{3}{5} & -\frac{1}{5} \end{pmatrix}.$$

Finally $X$, the solution of the equation, is

$$X = A^{-1}B = \begin{pmatrix} -\frac{1}{5} & \frac{2}{5} \\ \frac{3}{5} & -\frac{1}{5} \end{pmatrix}\begin{pmatrix} 0 & -2 \\ 8 & 1 \end{pmatrix} = \begin{pmatrix} \frac{16}{5} & \frac{4}{5} \\ -\frac{8}{5} & -\frac{7}{5} \end{pmatrix}$$

## 3.4 Linear Systems

A big motivation for developing the machinery of linear algebra is to find systematic methods for solving **systems of linear equations**, which we can call, in short, **linear systems**. A linear system in $n$ unknowns $x_1, x_2, \cdots, x_n$ and $m$ equations is a system of $m$ linear equations

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n &= b_2 \\ &\vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n &= b_m \end{aligned}$$

Collecting the coefficients of the left hand side of the system into a $m \times n$ matrix, and the coefficients of the right hand side into a $m$ dimensional column vector, we obtain the **matrix-vector form of the linear system**, $A\mathbf{x} = \mathbf{b}$:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

We say that a system of $m$ equations and $n$ variables has size $m \times n$.

If $\mathbf{b} = \mathbf{0}$, the system is called **homogeneous**.

**Example** Let $A$ and $B$ be the following matrices:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & -2 \\ 8 & 1 \end{pmatrix}$$

We have shown how to solve the equation $AX = B$ for $X$, where $X \in \mathbb{M}_{2,2}$. Expand this equation to show that is equivalent to a linear system. Write the linear system in matrix vector form $A\mathbf{x} = \mathbf{b}$.[17]

17: This will not be the same $A$ as in the statement.

The 4 unknowns are the entries of the matrix $X = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$. Then

$$AX = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} x & y \\ z & w \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ 8 & 1 \end{pmatrix}.$$

Expanding the product $AX$ gives the equation

$$AX = \begin{pmatrix} x + 2z & y + 2w \\ 3x + z & 3y + w \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ 8 & 1 \end{pmatrix}.$$

Equating the 4 components gives us a system of 4 equations in 4 unknowns:

$$\begin{aligned} x + 2z &= 0 \\ y + 2w &= -2 \\ 3x + z &= 8 \\ 3y + w &= 1. \end{aligned}$$

In matrix vector form the system is of the form $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{M}_{4,4}$, whose entries are specified in the equation below. The right-hand side is the vector of 4 constant entries, and the unknown vector has component $x, y, z, w$. The system is therefore

$$\begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 3 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ -2 \\ 8 \\ 1 \end{pmatrix}.$$

Rearranging the entries of a matrix in order to obtain a new matrix of different size is a common procedure in coding. Programs like R or

`Python` come with predefined functions that do the rezising for us (but we need to know how they operate!)

The solution set of an arbitray (non-linear) system of equations in $n$ variables is a region of $\mathbb{R}^n$. We learn that such regions are recognized to be objects of euclidean geometry: as we learn in pre-calculus, for example, the solutions of the equation $x^2 + y^2 = 1$ are the points of the circle of radius 1 and centre at the origin of the Cartesian plane.

For a linear system, we will expect the solution set to have some "linearity". More precisely:

- The solution set of a homogeneous linear system $A\mathbf{x} = \mathbf{0}$, in the $n$ unknowns $\mathbf{x} = (x_1, x_2, \cdots, x_n)^\top$, is a vector subspace of $\mathbb{R}^n$.
- The solution set of a linear system $A\mathbf{x} = \mathbf{b}$ is a "vector space shifted away from the origin" of $\mathbb{R}^n$. More precisely, let $\mathbf{x}_p$ be any solution of the system (we call it a particular solution). Then any solution of the system is of the form $\mathbf{x}_0 + \mathbf{x}_p$, where $\mathbf{x}_0$ is a solution of the associated homogeneous linear system $A\mathbf{x} = \mathbf{0}$.

**Example** Let us illustrate the last two points with a simple example. Notice that this example is not meant to propose an algorithm to solve a linear system, but rather to explain the geometrical aspect of the solution set of a linear system. Consider the linear system consisting of one equations in two variables:

$$x + y = 2.$$

It is not homogeneous, since the left hand side coefficient of the equation is not zero. Since there are two variables but only one equation, we expect the general form of the solution of this system to have one free parameter (or free variable), that can be arbitrary chosen. If we use $t$ as the name for the parameter, we write

$$\begin{aligned} x &= t \in \mathbb{R} \\ y &= 2 - t. \end{aligned}$$

Note in particular that with $t = 0$ we obtain the particular solution

$$\mathbf{x}_p = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

which we will use in a second.

The associated homogeneous linear system is

$$x + y = 0.$$

The solution set of the homogeneous system is the line $y = -x$, which in parametric form becomes

$$\begin{aligned} x &= t \\ y &= -t. \end{aligned}$$

Geometrically, the general solution of the non-homogeneous system is obtained by shifting the line $y = -x$ by the vector $(0, 2)^\top$. If we let

$$\mathbf{x}_0(t) = \begin{pmatrix} t \\ -t \end{pmatrix},$$

we see that the general solution is of the form $\mathbf{x}(t) = \mathbf{x}_p + \mathbf{x}_0(t)$.

**Example** Which of the following equation is linear? Why is it important to identify if an equation (or a system of equation) is linear?

a) $x + y - z = 4$ b) $x^2 - y + z = 4$ c) $4x + 4y - z - 4 = 0$

The system a) and c) are linear, while the $x^2$ term in b) makes that one non-linear. It is important to know what are the properties of linear systems: the linear algebra algorithms, such as Gauss-Jordan elimination, do not apply to non-linear systems.

### 3.4.1 Gauss-Jordan Elimination

In introductory linear algebra courses we often start by learning linear systems and how to the **Gauss-Jordan elimination** algorithm. We will not discuss the details of the method in this chapter and we refer to [3] for more details.

The idea of the elimination algorithm is to transform the matrix associated with a linear system into a simpler one. The common approach is to transform the original matrix to a **row echelon form**, or even better the **row reduced echelon form**. Reading the solution of a matrix in echelon form then is quite easy.

The principles behind the Gauss-Jordan elimination are the following. We say that two linear systems are **equivalent** if they have the same solution set.

- Given a linear system, an equivalent system is obtained by adding to one equation a multiple of another one. In term of the matrix associated to the linear system, this amounts to adding to a row a multiple of another one.
- Given a linear system, an equivalent system is obtained by rescaling an equation by a non-zero factor. In term of the matrix associated to the linear system, this amounts to multiplying the row vector corresponding to the equation by a scalar.

We can therefore proceed and start eliminating as much variables as we can, trying to obtain a matrix from which reading the solution is a simple procedure. Let us see an example.

**Example** We solve the following $2 \times 3$ linear system:

$$\begin{aligned} x - y - 2z &= 0 \\ 3x + 2y + z &= 2 \end{aligned}$$

We start by writing the **augmented matrix**

$$\begin{pmatrix} 1 & -1 & 2 & 0 \\ 3 & 2 & 1 & 2 \end{pmatrix},$$

which includes the right hand side of the system in the last column. We proceed with the row reduction in order to reduce the system to an equivalent one that is easier to solve.

We denote by $R_k$ the row number $k$ of the matrix (in this example, $k = 1, 2$). Assume $a \neq 0, b \in \mathbb{R}$; $R_k \to aR_k + bR_j$ denotes the operation of replacing $R_k$ with the linear combination $aR_k + bR_j$.[18] Then

18: $a \neq 0$ is crucial.

$$\begin{pmatrix} 1 & -1 & 2 & 0 \\ 3 & 2 & 1 & 2 \end{pmatrix} \xrightarrow{R_2 \to R_2 - 3R_1} \begin{pmatrix} 1 & -1 & 2 & 0 \\ 0 & 5 & -5 & 2 \end{pmatrix} \xrightarrow{R_2 \to \frac{R_2}{5}} \begin{pmatrix} 1 & -1 & 2 & 0 \\ 0 & 1 & -1 & \frac{2}{5} \end{pmatrix} \xrightarrow{R_1 \to R_1 + R_2|} \begin{pmatrix} 1 & 0 & 1 & \frac{2}{5} \\ 0 & 1 & -1 & \frac{2}{5}. \end{pmatrix}$$

The column in position $j$ corresponds to the variable in position $j$.

With the help of row reduction, the original linear system has been transformed into the equivalent system:

$$\begin{aligned} x + z &= \frac{2}{5} \\ y - z &= \frac{2}{5}, \end{aligned}$$

Selecting $z$ as a **free variable**, we re-write it as:

$$\begin{aligned} x &= \frac{2}{5} - z \\ y &= \frac{2}{5} + z, \end{aligned}$$

We see that $(x, y)$ depends on the value of $z$. The solution set of the linear system is therefore one-dimensional: geometrically, it is the line parametrized by the two equations above, with $z$ being the free parameter. Note that the line does not pass through the origin, in agreement with the fact that the system is not homogeneous. ■

The solution set of a system of homogeneous linear equations is a vector space. The dimension of this vector space coincides with the number of free variables. In particular, if there are no free variables then either the solution is **unique** or the system is **inconsistent** – it does not have solutions.

**Example** Find an example of a linear system with a) no solutions, and b) an example of a linear system whose solution set has 3 free variables out of a total of 5.

To find an example of a) is very easy: write an equation "$\cdots = 1$", then add another equation, obtained by changing the constant to the right hand side, "$\cdots = 2$". Let us take the following example:

$$\begin{aligned} 3x + y - z + w &= 1 \\ 3x + y - z + w &= 2. \end{aligned}$$

It should be clear that no solution can exist, since $1 \neq 2$. Proceeding with row reduction, we can see it algorithmically: we replace $R_2 \to R_2 - R_1$ and we obtain the system:

$$\begin{aligned} 3x + y - z + w &= 1 \\ 0 &= 1, \end{aligned}$$

which is inconsistent.

As for b), we can produce an example of a matrix that gives 3 free variables, if treated as the augmented matrix of a linear system:

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The examples of this section have shown that, if $A$ is a matrix associated with the linear system $A\mathbf{x} = \mathbf{b}$, then:

- the rows of the matrix corresponds to the system's equations, the column to its variables;
- interchanging two rows of the matrix swaps the corresponding equations in the linear system; interchanging two columns swaps the corresponding variables.

**Example** The system

$$\begin{aligned} 3x - y + z &= 0 \\ x + y + z &= 3 \end{aligned}$$

corresponds to the matrix

$$\begin{pmatrix} 3 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

If we switch $y$ and $z$,[19] we obtain

$$\begin{aligned} 3x + z - y &= 0 \\ x + z + y &= 3, \end{aligned}$$

which corresponds to the matrix

$$\begin{pmatrix} 3 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix}.$$

19: Which should not be done unless absolutely necessary, to be honest, but nevermind that for now.

### 3.4.2 Linear Systems and Matrices

Row reduction can be used to invert non-singular matrices. Let $A \in \mathbb{M}_{n,n}$ be such that $\det(A) \neq 0$. Construct the augmented matrix $(A \mid \mathbf{I}_n)$ and row reduce it using only the 3 following allowable operations:

- $R_j \to R_j + bR_k$, $j \neq k$;
- $R_j \to aR_j$, $a \neq 0$;
- $R_j \leftrightarrow R_k$, $j \neq k$.

The process leads to

$$(A \mid \mathbf{I}_n) \xrightarrow{\text{RREF}} (\mathbf{I}_n \mid A^{-1}).$$

**Example** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}$. We have seen that $\det(A) = -5$ and so that $A$ is invertible. We reduce the augmented matrix:

$$\begin{pmatrix} 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \to R_2 - 3R_1} \begin{pmatrix} 1 & 2 & 1 & 0 \\ 0 & -5 & -3 & 1 \end{pmatrix} \xrightarrow{R_2 \to -\frac{R_2}{5}} \begin{pmatrix} 1 & 2 & 1 & 0 \\ 0 & 1 & \frac{3}{5} & -\frac{1}{5} \end{pmatrix} \xrightarrow{R_1 \to R_1 - 2R_2|} \begin{pmatrix} 1 & 0 & -\frac{1}{5} & \frac{2}{5} \\ 0 & 1 & \frac{3}{5} & -\frac{1}{5}; \end{pmatrix}$$

so

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}^{-1} = \frac{1}{5} \begin{pmatrix} -1 & 2 \\ 3 & -1 \end{pmatrix}.$$

## 3.5 Matrix Diagonalization

Through a series of specific transformations, some matrices can be brought into diagonal form. This seemingly inconspicuous property has far-reaching consequences.

### 3.5.1 Eigenvalues and Eigenvectors

A matrix is **diagonal** if its non-zero entries can only be found along the diagonal.[20] Diagonal matrices are very simple: in associated linear systems, the variables involved are "decoupled", and solving the system amounts to solving a collection of linear equations in one variable. In fact, for the diagonal matrix $A$ with diagonal entries denoted, **in order**, by $\lambda_1, \lambda_2, \ldots, \lambda_n$, the linear system $A\mathbf{x} = \mathbf{b}$ is

$$\begin{aligned} \lambda_1 x_1 &= b_1 \\ \lambda_2 x_2 &= b_2 \\ &\vdots \\ \lambda_n x_n &= b_n. \end{aligned}$$

Note that if $\lambda_j = 0$ for some index $j$, the system has solution only if $b_j = 0$, and the variable $x_j$ corresponds to a subspace belonging to $\ker(A)$.

But matrices are not "absolute objects", in the sense that the values of the entries of a matrix depend on the choice of a basis of the vector space where the matrix operates as a linear map. Can we change the coordinates so that a given matrix, with respect to this new coordinate system, is diagonal?[21]

The first step in answering this question requires the introduction of **eigenvalues** and **eigenvectors**.

20: Note that the diagonal entries themselves could be zero.

21: The answer to this question is: "not always, but we can still do partial diagonalization".

- Let $A$ be a square matrix of size $n$. Let $\mathbf{v} \neq \mathbf{0} \in \mathbb{R}^n$. We say that $\mathbf{v}$ is an **eigenvector** of $A$ if

$$A\mathbf{v} = \lambda\mathbf{v}$$

  for some scalar $\lambda \in \mathbb{C}$. The number $\lambda$ is said to be the **eigenvalue** of $A$ associated to the eigenvector $\mathbf{v}$.
- If $\mathbf{v} \neq \mathbf{0} \in \mathbb{R}^n$ is an eigenvector of $A$ associated with eigenvalue $\lambda$, then so is $c\mathbf{v}$, $c \neq 0$. Indeed, if $A\mathbf{v} = \lambda\mathbf{v}$, then

$$A(c\mathbf{v}) = cA\mathbf{v} = c\lambda\mathbf{v} = \lambda(c\mathbf{v}).$$

By definition, the zero vector $\mathbf{0}$ cannot be an eigenvector. Also, note that for a given eigenvector, only one eigenvalue is associated to it.[22]

22: But eigenvalues/eigenvectors can be complex, even if the matrix only has real entries.

What happens when we apply a matrix to one of its eigenvector? A eigenvector spans a one dimensional vector space (a line), and **along this line** the matrix acts like a scalar, rescaling $\mathbf{v}$ by $\lambda$.

The goal of diagonalization is to transform the matrix to a form which is as close as possible to a diagonal; the best form would be a diagonal matrix, as we can see in the next exercise.

**Example** Let $A$ be a diagonal matrix. Show that the eigenvalues of $A$ are the diagonal values. What are the eigenvectors of $A$?

The matrix is of the form

$$A = \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{pmatrix}.$$

For the vector

$$\mathbf{e}_k = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^\top,$$

it is easy to verify that

$$A\mathbf{e}_k = a_{k,k}\mathbf{e}_k.$$

Hence $\mathbf{e}_k$ is the eigenvector with eigenvalue $\lambda_k = a_{k,k}$. ■

An eigenvector,[23] can come from only one eigenvalue. That is in fact almost obvious. Suppose that an eigenvector $\mathbf{v}$ of a matrix $A$ satisfies the eigenvector equation with two different eigenvalues, which we call $\lambda$ and $\mu$, which is to say that

23: Or the 1-dimensional eigenspace spanned by it.

$$A\mathbf{v} = \lambda\mathbf{v} \quad \text{and} \quad A\mathbf{v} = \mu\mathbf{v}.$$

Since the two left-hand sides of the equations above are the same, it follows that $\lambda\mathbf{v} = \mu\mathbf{v}$. Since $\mathbf{v}$, being an eigenvector, is non-zero by definition, this last equation implies that $\lambda = \mu$.

**Example** Can two linearly independent eigenvector have the same eigenvalue? If you believe that this is true (which it is), prove it by finding an example of a a matrix which has the same eigenvalue for more than one independent eigenvector.

The zero matrix can be used, but let us take a non-trivial example. Fix any $\lambda \neq 0 \in \mathbb{R}$ and consider the matrix

$$\begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The eigenvalue $\lambda$ is associated to two linearly independent eigenvectors, $\mathbf{i} = (1,0,0)^\top$ and $\mathbf{j} = (0,1,0)^\top$;[24] the eigenvalue 0 is associated to the eigenvector $\mathbf{k} = (0,0,1)^\top$.[25] ■

24: These are not the only two linearly independent eigenvectors, however.

25: In particular, $\mathbf{k}$ spans $\ker(A)$.

But what do eigenvectors represent, geometrically?[26]

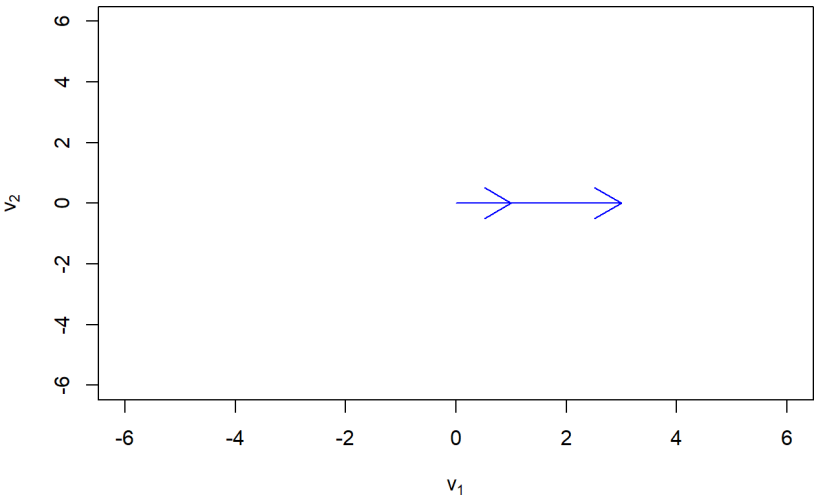
26: It is important to note that while we have illustrated the eigenconcepts with arrows in $\mathbb{R}^n$, any linear mapping of a vector space to another could have eigenvectors; in some cases eigenvectors are functions, not geometrical vectors.

**Example** Let

$$A = \begin{pmatrix} 3 & 2 \\ 0 & 1 \end{pmatrix}.$$

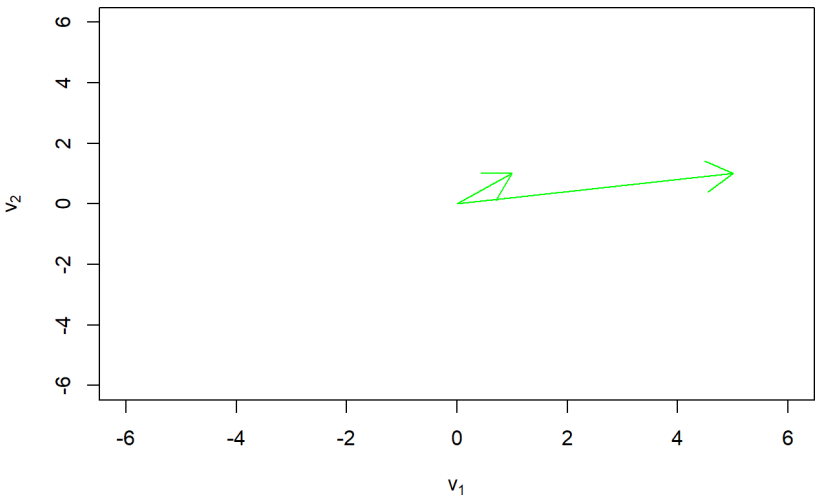
We can show that $\mathbf{v} = (1,0)^\top$ is an eigeventor of $A$, with eigenvalue 3, since $A\mathbf{v} = 3\mathbf{v}$. Applying $A$ to $\mathbf{v}$ stretches it by a factor of 3, as seen below.

```
plot(NA,xlim=c(-6,6), ylim=c(-6,6),
     xlab = expression(list(v[1])),
     ylab=expression(list(v[2])))
arrows(0,0,1,0, col="blue")
arrows(0,0,3,0, col="blue")
```




But the vector $\mathbf{w} = (1,1)^\top$ is not an eigenvector of $A$ since $A\mathbf{w} = (5,1)^\top \neq \lambda(1,1)^\top$, no matter the value of $\lambda$. Applying $A$ to $\mathbf{w}$ does not only dilate it, it also **rotates** it.

```
plot(NA,xlim=c(-6,6), ylim=c(-6,6),
    xlab = expression(list(v[1])),
    ylab=expression(list(v[2])))
arrows(0,0,1,1, col="green")
arrows(0,0,5,1, col="green")
```




The previous examples are easy because the involved matrices are diagonal; finding the eigenvalues and eigenvectors of a general matrix will help us transform it to a form that is closer to a diagonal.

The recipe for finding the eigenvalues and eigenvector of a matrix $A$ starts with constructing a polynomial equation, known as the **characteristic equation**, such that its roots are the eigenvalues of $A$.[27]

27: The characteristic equation is a direct consequence of the properties of determinant from Section 3.3.3.

Suppose that $\lambda$ is an eigenvalue of $A$ (the exact value does not matter): by definition, there is a non-zero eigenvector $\mathbf{v}$ such that $A\mathbf{v} - \lambda v = \mathbf{0}$, which can be re-written as

$$(A - \lambda \mathbf{I}_n)\mathbf{v} = \mathbf{0},$$

where $\mathbf{I}_n$ is the identity matrix with the same size as $A$.

The matrix $A - \lambda \mathbf{I}_n$ has therefore a non-zero nullspace, since it contains the nonzero vector $\mathbf{v}$. It follows that $A - \lambda \mathbf{I}_n$ is not invertible which means that its determinant is zero.

Hence, the eigenvalue $\lambda$ is a solution of the **characteristic equation**

$$\det(A - \lambda \mathbf{I}_n) = 0.$$

The expression $\det(A - \lambda \mathbf{I}_n)$ is a polynomial in the variable $\lambda$, called the **characteristic polynomial of** $A$. The **degree** of the characteristic polynomial (its highest exponent in $\lambda$) is the size $n$ of the $A$.

This works for all sizes $n$, but it is typically easier to find the eigenvalues when $2 \leq n \leq 4$, due to the insolvability of the quintic; for $n \geq 5$, we have to use numerical methods (see Chapter 4).

**Example** Write the characteristic polynomial of the matrix

$$A = \begin{pmatrix} 1 & 4 \\ 1 & 2 \end{pmatrix},$$

and find its eigenvalues.

We need to apply the definition of the characteristic polynomial, expand the determinant, and simplify. The eigenvalues will be the roots of a quadratic equation, since $A$ is of size 2.

$$\begin{aligned} \det(A - \lambda \mathbf{I}_2) &= \det\left(\begin{pmatrix} 1 & 4 \\ 1 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\ &= \det\left(\begin{pmatrix} 1 & 4 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right) \\ &= \det\begin{pmatrix} 1-\lambda & 4 \\ 1 & 2-\lambda \end{pmatrix} \\ &= (1-\lambda)(2-\lambda) - 4 = \lambda^2 - 3\lambda - 2. \end{aligned}$$

The eigenvalues of $A$ are thus the solutions of the equation

$$\lambda^2 - 3\lambda - 2 = 0,$$

namely

$$\lambda_{1,2} = \frac{3 \pm \sqrt{17}}{2}.$$

In this example, both eigenvalues are real. ■

Let $A$ be a square matrix, of any size, and suppose that $\mathbf{v}$ and $\mathbf{w}$ are two eigenvectors of $A$. Is their sum an eigenvector? What about a linear combination of them?

In general the sum is not an eigenvector. However, if $\mathbf{v}$ and $\mathbf{w}$ are associated **with the same eigenvalue** $\lambda$, then their sum is another eigenvector of $A$ with the same eigenvalue, as the following calculations demonstrates:

$$A(\mathbf{v} + \mathbf{w}) = A\mathbf{v} + A\mathbf{w} = \lambda\mathbf{v} + \lambda\mathbf{w} = \lambda(\mathbf{v} + \mathbf{w}).$$

The sum $\mathbf{v} + \mathbf{w}$ is a linear combination; it should not be too difficult to show that a non-trivial linear combination $a\mathbf{v} + b\mathbf{w}$, $a, b \neq 0$ is not an eigenvector of $A$, unless $\mathbf{v}$ and $\mathbf{w}$ share their associated eigenvalue.

After we obtain the eigenvalues of $A$ from the characteristic equation, the next step is to find the corresponding eigenvectors.

As before, we let $A \in \mathbb{M}_{n,n}$ and $\lambda$ be an eigenvalue of $A$. The vector subspace of $\mathbb{R}^n$ spanned by all eigevenctors with this eigenvalue is **eigenspace** $E_\lambda$. The dimension $E_\lambda$, as a vector subspace of $\mathbb{R}^n$, is the **geometric multiplicity** of the associated eigenvalue $\lambda$.

The eigenspace corresponding to an eigenvalue is obtained by solving the homogeneous linear system $(A - \lambda \mathbf{I}_n)\mathbf{v} = \mathbf{0}$, where the unknown are the components of the eigenvector $\mathbf{v}$.

**Example** What are the eigenvectors of the matrix $A$ from the previous example?

We already know the eigenvalues of $A$:

$$\lambda_{1,2} = \frac{3 \pm \sqrt{17}}{2}.$$

To find $\mathbf{v}_1$, the eigenvector of $A$ associated to $\lambda_1$, we must solve the system

$$\begin{pmatrix} 1-\lambda_1 & 4 \\ 1 & 2-\lambda_1 \end{pmatrix} \begin{pmatrix} v_{1,1} \\ v_{1,2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Expanding the system gives two equations in the unknowns $x, y$ (the components of the eigenvector $\mathbf{v}_1$ ).

$$\left(1 - \frac{3+\sqrt{17}}{2}\right) v_{1,1} + 4v_{1,2} = 0$$
$$v_{1,1} + \left(2 - \frac{3+\sqrt{17}}{2}\right) v_{1,2} = 0.$$

We expect this system to have a free variable, since the eigenspace has to be one dimensional.[28]

28: Why is that the case?

We can either get the solution *via* the Gauss-Jordan elimination algorithm or, we can solve directly by substitution since $n$ is quite small. Proceeding with the second option, we solve both equations for $v_{1,2}$, and the second equation collapses into the first:

$$v_{1,2} = \frac{1+\sqrt{17}}{8} v_{1,1}.$$

As expected, we found a one dimensional eigenspace, parametrized by $v_{1,1}$. We can exhibit a basis for $E_{\lambda_1}$ by selecting any non-zero eigenvector in this space; setting $v_{1,1} = 1$, we find

$$E_{\lambda_1} = \text{Span}\{\mathbf{v}_1\} = \text{Span}\left\{\begin{pmatrix} 1 \\ \frac{1+\sqrt{17}}{8}. \end{pmatrix}\right\}$$

Similar computations, which we let the reader perform, yield

$$E_{\lambda_2} = \text{Span}\{\mathbf{v}_2\} = \text{Span}\left\{\begin{pmatrix} 1 \\ \frac{1-\sqrt{17}}{8}. \end{pmatrix}\right\}$$

The **multiplicity** of an eigenvalue is linked to the number of times it appears as a solution of the characteristic equation. We can count properly the number of eigenvalues and eigenvector making use of this concept.

- An eigenvalue is a solution of the characteristic equation $\det(A-\lambda I)$: the multiplicity of the solution is called the **algebraic multiplicity** of the eigenvalue.
- It can be shown that the **geometric multiplicity**, i.e., the dimension of the associated eigenspace $E_\lambda$, is smaller than or equal to the **algebraic multiplicity** (defined above).

### 3.5.2 Similar Matrices

Eigenvectors define subspaces along which the matrix acts by scalar multiplication.[29] Once we have the eigenvectors, we apply a similarity transformation to transform our matrix to a "more diagonal one".

29: That is to say, by stretching or dilation.

Before proceeding, we need to define similarity of matrices: two square matrices $A$ and $B$ of the same size are said to be **similar** if there is an invertible matrix $P$ such that

$$B = P^{-1}AP.$$

The transformation $A \longrightarrow B = P^{-1}AP$ is a **similarity transformation**.

**Example** Similarity is an **equivalence relation**, which means that it satisfies the 3 following properties:

1. **reflexivity** – $A$ is similar to itself;
2. **symmetry** – $A$ is similar to $B$ if and only if $B$ is similar to $A$;
3. **transitivity** – if $A$ is similar to $B$ and $B$ is similar to $C$, then $A$ is similar to $C$.

This exercise is more "theoretical" than our usual fare, but the proof is easy and it will help us familiarize ourselves with the algebra of matrices.

1. Let $P = \mathbf{I}$, the identity matrix of the same size of $A$: then

$$P^{-1}AP = \mathbf{I}^{-1}A\mathbf{I} = \mathbf{I}A\mathbf{I} = A.$$

2. Let $B = P^{-1}AP$ be the similarity relation. Then we can multiply both of its sides to the left by $P$ and to the right by $P^{-1}$:

$$PBP^{-1} = PP^{-1}APP^{-1} = (PP^{-1})A(PP^{-1}) = \mathbf{I}A\mathbf{I} = A.$$

   If we let $Q = P^{-1}$, we therefore obtain the similarity relation:

$$A = Q^{-1}BQ.$$

3. Let $B = P^{-1}AP$ and $C = Q^{-1}BQ$ be the hypothetical similarity relations. Substituting the second into the first yields:

$$C = Q^{-1}BQ = Q^{-1}(P^{-1}AP)Q = (Q^{-1}P^{-1})A(PQ) = (PQ)^{-1}A(PQ).$$

   Hence $C$ is similar to $A$.

It is important to respect the properties of matrix multiplication: for numbers (scalars), the similarity relation reduces directly to equality since $p^{-1}bp = p^{-1}pb = b$ for any number.

For matrices the similarity relation is not trivial, since the matrix product is not commutative... but it does satisfy the other "standard properties" of numbers.

In the proof of the second property above, for instance, we made use of the associative property of matrix multiplication.

### 3.5.3 Diagonalization

Now that we have defined the concept of similarity between matrices, we can conclude our discussion about eigenvalues and eigenvectors with the last step: the diagonalization of a matrix.

We say that a square matrix $A$ is **diagonalizable** if it is similar to a diagonal matrix. That is, there exists an invertible matrix $P$ such that

$$D = P^{-1}AP$$

is a **diagonal matrix**.

As discussed previously, a square matrix $A \in \mathbb{R}^n$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Thus, the matrix $A$ is diagonalizable if and only if there exists a basis of $\mathbb{R}^n$ of eigenvectors of $A$, with respect to which the linear map is represented by a diagonal matrix.

The diagonal values of $D$ are in fact the eigenvalues of $A$, as we will explain in detail soon.

Once the matrix is diagonal, it is "easy to use": a linear system associated to a diagonal matrix of size $n$, for example, is equivalent to $n$ linear equations in one variable. The difficult part is to find the eigenvalues and eigenvectors, since we need to solve equations.[30]

30: Thankfully, we have already discussed how to do this.

Suppose that we found the matrix is diagonalizable, then what is the relation with the **eigenvalue problem**?

Let $A$ be a **square** matrix of size $n$. Suppose that $A$ is diagonalizable. Then $A$ has $n$ (possibly repeated) eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ with corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$. Denote by $D$ the diagonal matrix of the eigenvalues,

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

and denote by $P$ the matrix whose columns are the eigenvectors $A$ (**presented in the same order as the eigenvalues!**):

$$P = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}.$$

The **diagonalization** of $A$ is given by the **similarity transformation**:

$$D = P^{-1}AP.$$

An easy consequence of all this (which we will not prove) is that all **symmetric matrices** are diagonalizable. Moreover, if such a matrix only has real entries, then all of its eigenvalues are real.

**Example** Show that the equation $D = P^{-1}AP$ is equivalent to the equation $A = PDP^{-1}$

We multiply the two sides by $P^{-1}$ from the left, $P$ from the right:

$$P^{-1}AP = P^{-1}(PDP^{-1})P = (P^{-1}P)D(P^{-1}P) = IDI = D.$$

**Example** Prove that the matrix $A$ below is diagonalizable. Diagonalize it. How are the eigenvalues related to the determinant?

$$A = \begin{pmatrix} 2 & 3 & 0.4 & 1 \\ 3 & -1.3 & 0.6 & 17 \\ 0.4 & 0.6 & 0.1 & -23 \\ 1 & 17 & -23 & 0 \end{pmatrix}$$

The matrix is symmetric, hence it is diagonalizable. We expect 4 real eigenvalues (some of which could be duplicates).

We could try to solve the problem by hand, but it would most likely be rather time-consuming. We use R to speed up the process.

```
D <- rbind(c(2,3,0.4,1),c(3,-1.3,0.6,17),
           c(0.4,0.6,0.1,-23),c(1,317,-23,0))
eigen(D)
det(D)
prod(eigen(D)$values)
```

```
eigen() decomposition
$values
[1] -77.8741054  76.0897048   2.9324699  -0.3480693

$vectors
             [,1]        [,2]        [,3]        [,4]
[1,] -0.005237695 -0.01940525  0.46752893 -0.26108024
[2,] -0.210057716 -0.20443758  0.06120210  0.07064722
[3,]  0.278113608  0.28194691  0.87637211  0.96259458
[4,]  0.937283918 -0.93719510 -0.09819847  0.01605501

[1] 6048.09

[1] 6048.09
```

The output of the first two lines of codes produces the set of eigenvectors and eigenvalues. In particular, $A$ is transformed to the diagonal matrix $D$ *via* the eigenvector matrix $P$. The third line computes the determinant of the matrix, which we see is the same as the product of the eigenvalues of $A$, as shown by the fourth line of code.[31]

31: This will always be the case.

### Invariance of the Determinant

The value of the determinant is respected by similarity transformation: if $A$ and $B$ are similar matrices, then $\det(A) = \det(B)$. We can use this fact to prove that the determinant of a diagonalizable matrix is the product of its eigenvalues.

To prove the fist part, we use the property that the determinant respects the product and inverses: $\det(P^{-1}AP) = \det(P)^{-1}\det(A)\det(P) = \det(A)$. From here, the second part is clear, since for a diagonal matrix the determinant is the product of the diagonal entries.

But we must be careful: **not every square matrix is diagonalizable**!

**Example** For any $t \neq 0$, the matrix $T = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ only has one eigenvector $(0,1)^\top$, with eigenvalue $t$. The similarity recipe presented above can thus not be applied.

While the matrix is not diagonalizable, we can still construct its **Jordan normal form**, which is a more general version of a diagonal matrix [2].[32]

32: This is a topic for an advanced linear algebra course; we will not address it here.

## 3.6 Exercises

1. The augmented matrix $[A|B]$ of a system has 15 rows and 18 columns. Assume $\text{rank}(A) = 12$ and $\text{rank}([A|B]) = 13$. Which of the following statements is necessarily true?
   a) The system is inconsistent.
   b) The system has more than one solution, expressed with one parameter.
   c) The system has more than one solution, expressed with two parameters.
   d) The system has a unique solution.
   e) The system has more than one solution, expressed with three parameters.
   f) The system has more than one solution, expressed with four parameters.
2. Find all values of $b$ for which the following system is consistent:
$$\begin{aligned} x + y - z &= 2 \\ x + 2y + z &= 3 \\ x - 3z &= 2b - 1 \end{aligned}$$
3. Find all the values of $h$ for which the following vectors are linearly independent:
$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \\ 1 \\ h \end{pmatrix}.$$
4. Which of the following sets are subspaces of $\mathbb{R}^2$?
$$\begin{aligned} S &= \{(x,y) \in \mathbb{R}^2 \mid 2x - y = 1\} \\ T &= \text{Span}\{(-1,1),(2,-1)\} \\ U &= \{(x,y) \in \mathbb{R}^2 \mid y = x^2\} \\ V &= \{(x,y) \in \mathbb{R}^2 \mid x - 3y = 0\} \end{aligned}$$

5. $A$ is a $3 \times 3$ matrix. Suppose that $\det(A) = 3$. What is $\det(2A^\top A)$? (Hint: $A^\top$ is the transposed of $A$.)
6. Let $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix}$. Which of the following statements is true?
   a) $AB = \begin{pmatrix} -1 & 1 & 2 \\ 5 & 4 & 5 \end{pmatrix}$
   b) $BA = \begin{pmatrix} -1 & 1 & 2 \\ 5 & 4 & 5 \end{pmatrix}$
   c) $BA = \begin{pmatrix} 3 & 3 & 4 \\ 3 & 0 & -1 \end{pmatrix}$
   d) $AB = \begin{pmatrix} 3 & 3 & 4 \\ 3 & 0 & -1 \end{pmatrix}$
   e) $BA = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$
   f) $AB = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$
7. What is the determinant of $\begin{pmatrix} 0 & 0 & 0 & 5 & 0 \\ 2 & 0 & 3 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 7 & 0 & 0 & 0 \end{pmatrix}$?
8. Let $a, b, c, d, e, f$ be the constants and $x, y$ be the unknowns of the system

$$\begin{aligned} ax + by &= e \\ cx + dy &= f. \end{aligned}$$

   a) What condition(s) on $a, b, c, d, e, f$ are needed in order for the system to have a unique solution?
   b) What condition(s) on $a, b, c, d, e, f$ are needed in order for the system to have infinitely many solutions?
   c) What condition(s) on $a, b, c, d, e, f$ are needed in order for the system to have no solution?
9. Let $B = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}$. Find all $2 \times 2$ matrices $A$ that satisfy $AB = BA$. (Hint: write $A = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$, and compute $AB$ and $BA$. Then, solve the system of 4 equations in 4 unknowns that arises from $AB = BA$.)
10. Consider the matrix $A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}$.
    a) Find the eigenvalues of $A$.
    b) For each eigenvalue of $A$, find the corresponding eigenspace of $A$, and state its dimension.
11. Consider the matrix $A = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & -5 & 1 \\ 1 & -1 & 0 & 5 & 1 \\ 2 & -2 & -1 & 9 & 0 \end{pmatrix}$, whose re-

duced row echelon form is

$$\tilde{A} = \begin{bmatrix} 1 & -1 & 0 & 5 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

a) Find the column space of $A$? (Hint: find the columns of $A$ that are necessary to express the column space of $A$.)
b) Are the columns of $A$ linearly independent?
c) What is the dimension of the column space of $A$?
d) Find a basis for the nullspace of $A$.
e) Does the system $Ax = 0$ have a unique solution?

12. Find all values of $x$ for which $\det\begin{pmatrix} 1 & x & x \\ -x & -2 & -x \\ x & -x & -3 \end{pmatrix} = 0$.
13. Let $V$ be a vector space and let $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$. Which of the following statements are true?

13.. If $\{\mathbf{u}, \mathbf{v}\}$ is linearly independent, so is $\{\mathbf{u}, \mathbf{v}, \mathbf{u} + \mathbf{v}\}$.
13.. If $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is linearly independent, so is $\{\mathbf{u}, \mathbf{v}\}$.
13.. If $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is linearly dependent, so is $\{\mathbf{u}, \mathbf{v}\}$.
13.. If $\{\mathbf{u}, \mathbf{v}\}$ is linearly independent, so is $\{\mathbf{u}, \mathbf{u} + \mathbf{v}\}$.

14. Which of the following statements are true?

a) The set $\{(x, x-1, y) \in \mathbb{R}^3 \mid x, y \in \mathbb{R}\}$ is a subspace of $\mathbb{R}^3$.
b) The set $\{p(x) \in \mathbb{P}_4 \mid p(2) = 0\}$ is a subspace of $\mathbb{P}_4$.
c) The set $\{A \in \mathbb{M}_{2,2} \mid A^2 = A\}$ is not a subspace of $\mathbb{M}_{2,2}$.

15. Let $\{\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{z}\}$ be a set of linearly independent vectors. Which of the following sets of vectors are linearly dependent?

a) $\{\mathbf{u} + \mathbf{v}, \mathbf{v} + \mathbf{w}, \mathbf{w} + \mathbf{u}\}$
b) $\{\mathbf{u}, \mathbf{u} + \mathbf{z}, \mathbf{v}, \mathbf{v} + \mathbf{w}\}$
c) $\{\mathbf{u} - \mathbf{v}, \mathbf{v} - \mathbf{w}, \mathbf{w} - \mathbf{z}, \mathbf{z} - \mathbf{u}\}$
d) $\{\mathbf{u}, \mathbf{u} + \mathbf{z}, \mathbf{z}\}$

16. If $\det\begin{pmatrix} 3 & -1 & x \\ 2 & 6 & y \\ -5 & 4 & z \end{pmatrix} = ax + by + cz$, what is the value of $c$?
17. Let $A, B, C$ be square $n \times n$ matrices with $\det(A) = 1$, $\det(B) = 4$ and $\det(C) = -3$. What is the value of $\det(A^2 B C^\top B^{-1})$?
18. For each of the following subspaces, exhibit a basis and find the dimension.

a) $\{(x, y, z, w) | x - y + z - w = 0\}$
b) $\{A \in \mathbb{M}_{2,2} \mid A^\top = -A\}$

19. Let $A = \begin{pmatrix} 2 & -1 & 0 \\ -3 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$.

a) Find $c_A(\lambda)$, the characteristic polynomial of $A$.
b) Use your answer in (a) to determine the eigenvalues of $A$.
c) Find a basis for two of the eigenspaces of $A$.

20. Let $U$ and $W$ be subspaces of $V$. Define

$$U \cup W = \{\mathbf{v} \in V \mid \mathbf{v} \in U \text{ or } \mathbf{v} \in W\}$$
$$U \cap W = \{\mathbf{v} \in V \mid \mathbf{v} \in U \text{ and } \mathbf{v} \in W\}.$$

    a) Show that $U \cap W$ is a subspace of $V$.
    b) Is $U \cup W$ necessarily a subspace of $V$? Explain.

21. The *trace* of a matrix $A$, denoted by $\text{tr}(A)$, is the sum of the elements on the diagonal of $A$. Thus, $\text{tr}\begin{pmatrix} x & y \\ z & w \end{pmatrix} = x + w$.

    a) Show that $\text{tr} : \mathbb{M}_{2,2} \to \mathbb{R}$ is linear, that is, show that

$$\text{tr}\left[a\begin{pmatrix} x_1 & y_1 \\ z_1 & w_1 \end{pmatrix} + b\begin{pmatrix} x_2 & y_2 \\ z_2 & w_2 \end{pmatrix}\right] = a\,\text{tr}\begin{pmatrix} x_1 & y_1 \\ z_1 & w_1 \end{pmatrix} + b\,\text{tr}\begin{pmatrix} x_2 & y_2 \\ z_2 & w_2 \end{pmatrix}$$

    for all $a, b, x_i, y_i, z_i, w_i \in \mathbb{R}$.
    b) Let $x \in \mathbb{R}$. Find a matrix $A \in \mathbb{M}_{2,2}$ such that $\text{tr}(A) = x$.
    c) Using the Rank-Nullity Theorem and the result from part b), can you deduce the value of $\dim(\ker(\text{tr}))$?

22. Let $A = \begin{pmatrix} 2 & -1 & 0 \\ -3 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$. Find $\text{rowsp}(A)$ (the space spanned by the rows of $A$), $\text{colsp}(A)$ and $\text{nullsp}(A)$.

23. Find (if possible) conditions on $a$, $b$ and $c$ such that the system

$$x + ay = 0, \quad y + bz = 0, \quad z + cx = 0.$$

    has:

    a) no solution.
    b) one solution. What is the solution in this case?
    c) infinitely many solutions. What are the solutions in this case?

24. Amongst the following vectors, which one is a linear combination of $(1, 0, 0)$ and $(0, 1, 1)$?

$$(1, 2, 3), \quad (1, 0, 1), \quad (0, 0, 1), \quad (1, 1, 1), \quad (0, 1, 0), \quad (3, 2, 1).$$

25. Let $T : \mathbb{R}^2 \to \mathbb{R}$ be a linear transformation. If $T(1, 2) = 3$ and $T(1, 0) = -1$, what is $T(1, 1)$?

26. Amongst

$$U = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}; \quad V = \{(x, y) \in \mathbb{R}^2 \mid x + y \leq 0\};$$
$$W = \{(x, y) \in \mathbb{R}^2 \mid x = 2y\},$$

    which sets are subspaces of $\mathbb{R}^2$?

## Chapter References

[1] P. Boily, S. Davies, and J. Schellinck. *The Practice of Data Visualization*. Data Action Lab, 2023.

[2] W.K. Nicholson. *Linear Algebra with Applications, 3rd Edition*. PWS Publishing Company, 1994.

[3] G. Strang. *Introduction to Linear Algebra*. Wellesley, 2016.

# Basics of Numerical Methods 4

by **Patrick Boily** (inspired by **Diane Guignard**)


4.1 Basic Concepts . . . . . 181
Round-Off Error . . . . . 182
4.2 Equations With 1 Variable . 185
Bisection Method . . . . . 185
Golden Ratio Method . . . 191
Fixed Point Method . . . . . 193
Newton's Method . . . . . 203
Secant Method . . . . . 207
4.3 Systems of Equations . . . . 208
Linear Systems . . . . . 208
Non-Linear Systems . . . . . 221
4.4 Exercises . . . . . 223
Chapter References . . . . . 226


In today's digital age, it's hard to envision a world devoid of data and computers. Yet, the principles of "data science" predate our modern era of digital computation.

Take, for instance, Johannes Kepler's remarkable 16th-century computations. Before the invention of calculus, he analyzed the orbit of Mars based on Tycho Brahe's observations. This monumental effort culminated in the *Laws of Planetary Motion* [6]. Fast forward to the 20th century, where human computers at the *Jet Propulsion Laboratory* painstakingly calculated the number of rockets needed for space missions. These computations often spanned over a week, filling six to eight notebooks with data and intricate formulas [4]. Such endeavours underscore the invaluable contributions of data-based calculations to our scientific legacy.

Modern technology allows us to retrace and even surpass the feats of our predecessors in a mere fraction of their original time. With advancements in quantum computing, big data processing, and artificial intelligence on the horizon, it seems our computational potential knows no bounds – at least from a technical perspective.[1]

This chapter provides an **overview** of the foundational concepts and techniques at the heart of data science: the often-hidden **mathematics underlying data calculations** and data processing. Substantially more details are available in [1, 3].[2]

1: From a sociological and ethical viewpoint, however, the landscape is potentially more complex.

2: Some of the required topological concepts can also be found in [2].

## 4.1 Basic Concepts

In **scientific computing**, we typically navigate from a **physical problem** (observed phenomenon) to a **computed solution** (algorithm solution) *via* a **mathematical problem** (model) and/or a **numerical problem**, as illustrated in Figure 4.1.

If $u$ is the real solution of the problem and $\hat{u}$ the computed solution, we are often interested in the **computational error**, for obvious reasons: the smaller it is, the more confident we are in exhibiting $\hat{u}$ as a solution.

There are two types of such errors:

- **absolute error:** $|u - \hat{u}|$;
- **relative error:** $\frac{|u-\hat{u}|}{|u|}$.

**Sources of Error** In practice, it is nearly always the case that the computational error is not 0, i.e., that $u \neq \hat{u}$.

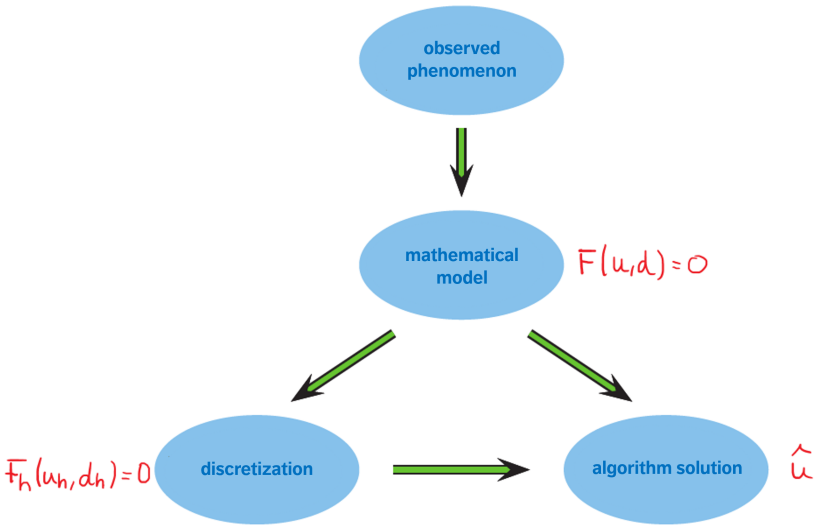



**Figure 4.1:** Schematics of scientific computing (modified from [1]).

That might prove to be the case due to:

- errors in the mathematical model;
- errors in the input data (e.g., due to measurements);
- **approximation errors**, such as **discretization errors** (in interpolation, differentiation, integration, ...) and **convergence errors** (in iterative methods), and/or
- **round-off errors** due to finite machine precision.

**Assessing Numerical Algorithms** In theory, there may be multiple ways of solving a problem numerically. In practice, we usually favour algorithms that are:

- **accurate**;
- **efficient** (in terms of CPU runtime, storage requirements, rate of convergence, etc.), as well as
- **robust/reliable/stable** (roughly speaking, computations do not magnify approximation errors).

### 4.1.1 Round-Off Error

In a computer, a real number $x$ is stored using a **floating point representation**:

$$\mathrm{fl}(x) = (-1)^s \cdot (1.d_1 d_2 \cdots d_t) \cdot 2^e,$$

where

- $s \in \{0, 1\}$ determines the **sign** of $x$, which is positive if $s = 0$, and negative if $s = 1$;
- $f = d_1 d_2 \ldots d_t$ is the **mantissa** (or fraction) of $x$ in base 2, with $d_i \in \{0, 1\}$, $1 \leq i \leq t$, and
- $e$ is the **exponent**, with $L \leq e \leq U$ for some $L, U$.

For instance, the floating point representation of $-6.5$ is

$$\mathrm{fl}(-6.5) = (-1)^1 \cdot (1.101) \cdot 2^2 \quad \Longrightarrow \quad -\left(1 + \frac{1}{2} + \frac{1}{2^3}\right) \cdot 2^2 = -6.5.$$

It is not too difficult to show that the following bound applies on the relative (rounding) error:

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \leq 2^{-(t+1)}.$$

**Single vs. Double Precision** Different operational systems/computational software use different values of $s$, $e$, and $f$.

| | $s$ | $e$ | $f$ | $L$ | $U$ |
|---|---|---|---|---|---|
| **single** (32 bits) | 1 bit | 8 bits | 23 bits | $-126$ | 127 |
| **double** (64 bits) | 1 bit | 11 bits | 52 bits | $-1022$ | 1023 |

In double precision, for instance, we represent numbers as follows:

$$(-1)^s \cdot \left(1 + \sum_{i=1}^{52} \frac{d_i}{2^i}\right) \cdot 2^e \quad \text{with} \quad L = -1022 \leq e \leq 1023 = U.$$

- The **smallest positive number** that can be represented has $s = 0$, $d_i = 0$, and $e = L \implies x_{\min} = 2^{-1022}$;
- the **largest positive number** has $s = 0$, $d_i = 1$, and $e = U \implies x_{\max} = (2 - 2^{-52})2^{1023}$.

We can recover these values (and other parameters) in R.

```
.Machine
```

```
$double.eps
[1] 2.220446e-16

$double.neg.eps
[1] 1.110223e-16

$double.xmin
[1] 2.225074e-308

$double.xmax
[1] 1.797693e+308

$double.base
[1] 2

$double.digits
[1] 53

$double.rounding
[1] 5

$double.guard
[1] 0

$double.ulp.digits
[1] -52

$double.neg.ulp.digits
[1] -53

$double.exponent
[1] 11

$double.min.exp
[1] -1022

$double.max.exp
[1] 1024

$integer.max
[1] 2147483647

$sizeof.long
[1] 4

$sizeof.longlong
[1] 8

$sizeof.longdouble
[1] 16

$sizeof.pointer
[1] 8

$longdouble.eps
[1] 1.084202e-19

$longdouble.neg.eps
[1] 5.421011e-20

$longdouble.digits
[1] 64

$longdouble.rounding
[1] 5

$longdouble.guard
[1] 0

$longdouble.ulp.digits
[1] -63

$longdouble.neg.ulp.digits
[1] -64

$longdouble.exponent
[1] 15

$longdouble.min.exp
[1] -16382

$longdouble.max.exp
[1] 16384
```

Round-off arithmetic can lead to odd behaviour – consider, for instance, the function $f : (0, \infty) \to \mathbf{R}$ defined by

$$f(x) = \frac{(1 + x) - 1}{x}.$$

In theory, we know that $f \equiv 1$ on $(0, \infty)$. In practice, things get messy.

We define the function in R using the following chunk of code.

```
f.test <- function(x){
  ((1+x)-1)/x
}
```

The function evaluates exactly to 1 for $x = 1, 10^{-9}, 10^{-10}$.

```
> f.test(1)
[1] 1
> f.test(0.0000000001)
[1] 1
> f.test(0.00000000001)
[1] 1
```

For smaller values, something strange is happening.

```
> f.test(0.000000000001)
[1] 1.000089
> f.test(0.0000000000001)
[1] 0.9992007
> f.test(0.000000000000001)
[1] 1.110223
> f.test(0.0000000000000001)
[1] 0
```

This phenomenon is know as **cancellation error**. Say we want to compute $f(10^{-16})$. We must first add $10^{-16}$ and 1 – to do so, we first need to **align the exponents**.

$$
\begin{aligned}
1 &= 1.000000000000000 \times 10^{0} \\
10^{-16} &= 1.000000000000000 \times 10^{-16} \\
&= 0.100000000000000 \times 10^{-15} \\
&= 0.010000000000000 \times 10^{-14} \\
&= 0.001000000000000 \times 10^{-13} \\
&= 0.000100000000000 \times 10^{-12} \\
&= 0.000010000000000 \times 10^{-11} \\
&= 0.000001000000000 \times 10^{-10} \\
&= 0.000000100000000 \times 10^{-9} \\
&= 0.000000010000000 \times 10^{-8} \\
&= 0.000000001000000 \times 10^{-7} \\
&= 0.000000000100000 \times 10^{-6} \\
&= 0.000000000010000 \times 10^{-5} \\
&= 0.000000000001000 \times 10^{-4} \\
&= 0.000000000000100 \times 10^{-3} \\
&= 0.000000000000010 \times 10^{-2} \\
&= 0.000000000000001 \times 10^{-1} \\
&= 0.000000000000000 \times 10^{0}
\end{aligned}
$$

From the perspective of double precision arithmetic, $1 + 10^{-16} = 1$! This explains why $f(10^{-16}) = 0$ in R.[3]

3: In R, the only numbers that are represented **exactly** are the integers and negative powers of 2. More information on round-off error (and error propagation) is available in [3].

## 4.2 Solving an Equation in 1 Variable

In this section, we will discuss how to solve an equation of the form

$$f(x) = 0$$

numerically, where $f : [a, b] \to \mathbb{R}$ is a (potentially non-linear) **continuous** function. A real number $x^* \in [a, b]$ for which $f(x^*) = 0$ is a **root** (or a **zero**) of the function $f$; "solving $f$ in $[a, b]$" means finding (at least) one root of $f$ in $[a, b]$.[4]

4: When the context is clear, we will drop "in $[a, b]$" from the conversation.

**Iterative Procedures** In some cases, we may be able to solve $f$ exactly – if $a \neq 0$, for instance, the linear equation $ax + b = 0$ has exactly one zero at $x^* = -b/a$. In practice, we can usually only hope to solve a continuous $f$ **approximately**, assuming a solution even exists.[5]

5: Not every function has a zero: for instance, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2+1$ does not have a root in $\mathbb{R}$.

In general, we must use an **iterative procedure** in order to zoom in on a root. Given an initial guess $x_0$, we generate a sequence of **iterates** $x_1, x_2, x_3, \ldots$ which (hopefully) converges to a root $x^*$ of $f$.

In order to exhibit a candidate $x^*$, we must stop the iterative process after a **finite number of iterations** $n$, according to a prescribed **stopping criterion** such as:

- $|x_n - x_{n-1}| \leq \text{tol}$;
- $|x_n - x_{n-1}|/|x_n| \leq \text{tol}$, provided $x_n \neq 0$, or
- $|f(x_n)| \leq \text{tol}$,

where tol is the algorithm's **prescribed tolerance**. We can avoid infinite loops by also prescribing a **maximum number of iterations** $N_{\max}$.

### 4.2.1 Bisection Method

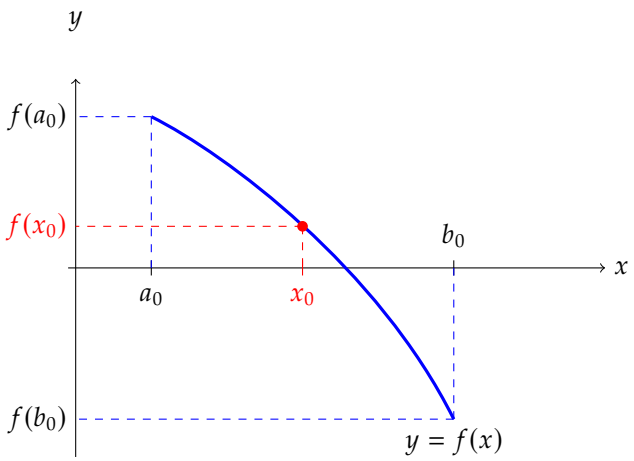
This method is based on the **intermediate value theorem**: if $f \in C([a, b])$ and $f(a)f(b) \leq 0$, then there exists $x^* \in [a, b]$ such that $f(x^*) = 0$.

Let $a_0 = a$, $b_0 = b$ and $x_0 = (a_0 + b_0)/2$. There are three possibilities:

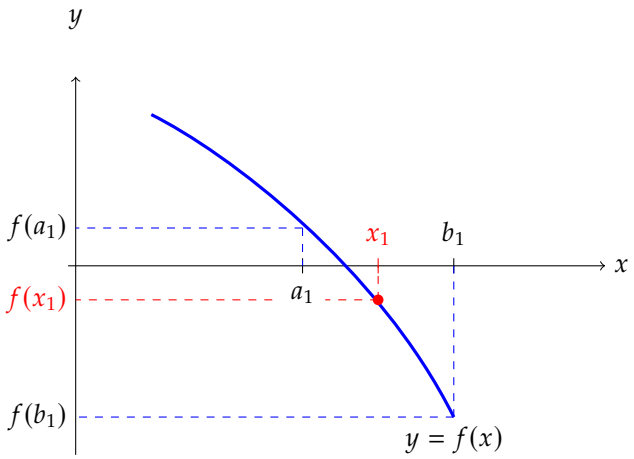
1. $f(x_0) = 0$, in which case $x^* = x_0$ is a root and we are done;
2. $f(a_0)f(x_0) < 0$, in which case $f$ has a root in $[a, x_0]$ and we set $a_1 = a_0$, $b_1 = x_0$, or
3. $f(b_0)f(x_0) < 0$, in which case $f$ has a root in $[x_0, b]$ and we set $a_1 = x_0$, $b_1 = b_0$.

In the latter two cases, we also set $x_1 = (a_1 + b_1)/2$; the **bisection method** re-iterates this process to generate a sequence $\{x_0, x_1, x_2, \ldots\}$, which converges to a root $x^* \in [a, b]$ of $f$.

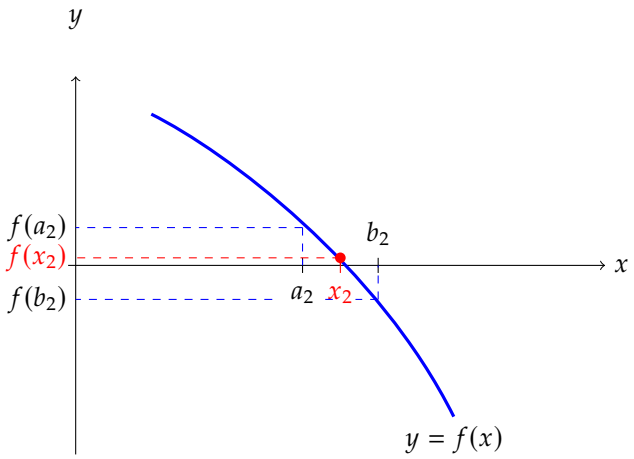
**Illustration of the Method** Let $f : [a, b] \to \mathbb{R}$ be the continuous function whose graph is displayed on the next page. Let $a_0 = a$, $b_0 = b$ and $x_0 = (a_0 + b_0)/2$; clearly, $f(a_0)f(b_0) < 0$.




We find ourselves in the third case, since $f(b_0)f(x_0) < 0$; as such $f$ has a root in $[x_0, b_0]$. In the next iteration, we set $a_1 = x_0$, $b_1 = b_0$, and $x_1 = (a_1 + b_1)/2$.




We find ourselves in the second case, since $f(a_1)f(x_1) < 0$; as such $f$ has a root in $[a_1, x_1]$. In the next iteration, we set $a_2 = a_1$, $b_2 = x_1$, and $x_2 = (a_2 + b_2)/2$, and so on.



Assume that we would like to use the bisection method to find an approximation $x_n$ of a root $x^*$ satisfying

$$|x_n - x^*| \leq \text{tol}$$

for a given tolerance tol $> 0$. How large $n$ should be? We can answer this question using the following result.

**Theorem:** let $f \in C([a,b])$ be such that $f(a)f(b) < 0$. The sequence $\{x_k\}$ generated by the bisection method approximates a root $x^*$ of $f$ with

$$|x_k - x^*| \leq \frac{b-a}{2^{k+1}}, \qquad k \geq 0.$$

**Proof:** we go through the procedure as illustrated previously; at step $k$, we have $x^* \in [a_k, b_k]$ and $x_k = (a_k + b_k)/2$. Moreover, $b_k - a_k = \frac{(b-a)}{2^k}$ as we have divided $[a,b]$ in two $k$ times at that point, and so

$$|x_k - x^*| \leq \frac{1}{2}(b_k - a_k) = \frac{b-a}{2^{k+1}},$$

which completes the proof. ■

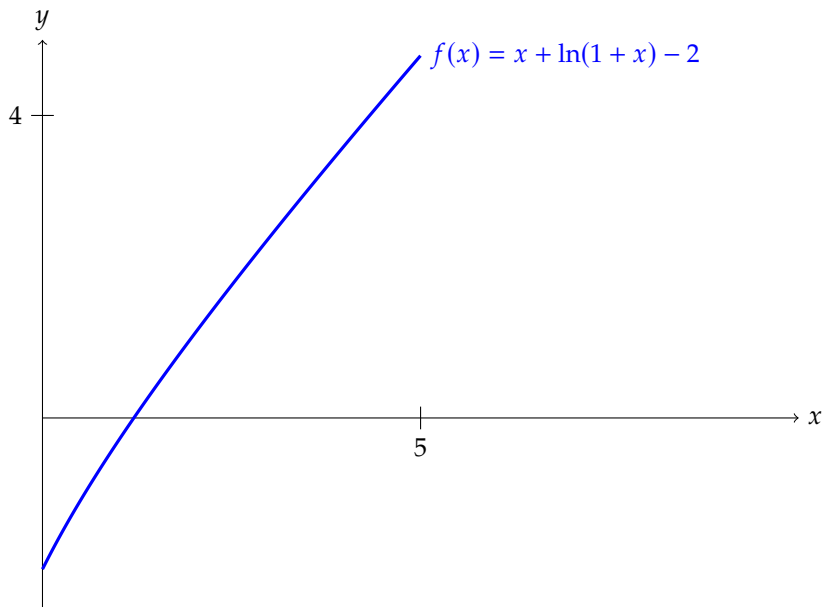
We can guarantee the desired absolute error tolerance if

$$|x_n - x^*| \leq \frac{b-a}{2^{n+1}} < \text{tol},$$

which is to say

$$2^{n+1} \geq \frac{b-a}{\text{tol}} \implies n \geq \log_2\left(\frac{b-a}{\text{tol}}\right) - 1.$$

**Example:** consider the function $f(x) = x + \ln(1+x) - 2$, $x \in [0,5]$, whose graph is given below.



We can guarantee that the bisection iterate $x_n$ is within tol $= 10^{-4}$ of $x^*$ when $n \geq \log_2(5 \cdot 10^4) - 1 = 14.60964$, which is to say when $n \geq 15$.

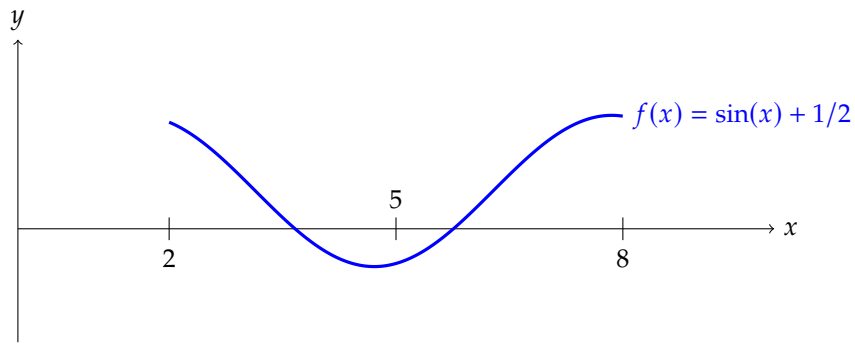
**Algorithm:** bisection method

**Input:** continuous $f$; $a, b$ with $f(a)f(b) < 0$; tol $> 0$
**Output:** approximation $p$ of $x^*$, $n$

**Initialization:** $a_0 = a$, $b_0 = b$, $x_0 = \frac{a_0+b_0}{2}$, $n = \lceil \log_2 \left( \frac{b-a}{\text{tol}} \right) - 1 \rceil$;
**For** $k = 0, 1, 2, \ldots, n-1$ **do**
    **If** $f(x_k) = 0$ **then**
        $p = x_k$, $n = k$;
    **Stop**
    **If** $f(a_k)f(x_k) < 0$ **then**
        $a_{k+1} = a_k$, $b_{k+1} = x_k$;
    **Else**
        $a_{k+1} = x_k$, $b_{k+1} = b_k$;
    **End**
    $x_{k+1} = \frac{a_{k+1}+b_{k+1}}{2}$;
**End**
$p = x_n$.

**Comments** On the positive side, the bisection method always converges when $f$ has a different sign at $a$ and $b$, and we have precise control over the error; on the negative side, the convergence is quite slow (the upper bound on the error only halves with each step), and the method fails to be of use if $f$ does not change sign near a root $x^*$.

**Example** Throughout this section, we will attempt to find roots of the test function $f(x) = \sin(x) + 1/2$ over the interval $[2, 8]$.




Graphically, we see that there are two roots: $x^* \in (2, 5)$ and $x_* \in (5, 8)$. The function is implemented in R as follows.

```
f.test <- function(x){ sin(x)+1/2 }
```

Can the bisection method find $f$'s roots? Here is an implementation of the method in R.

**Bisection method**

```
bisection <- function(f, a, b, tol) {
  # initialization
  k <- 0          # 0th iteration
  x <- (a + b)/2  # first iterate (root approximation)
  x_vec <- c(x)

  # max number of iterations for absolute error control
  n <- ceiling(log2((b - a) / tol) - 1)

  # Bisection method
  while (k < n) {

    if (f(x) == 0) {
      break
    } else {
      k <- k + 1

      if (f(a) * f(x) < 0) {
        b <- x
      } else {
        a <- x
      }
    }

    x <- (a + b) / 2
    x_vec <- c(x_vec, x)
  }

  return(list(x=x, k=k, x_vec=x_vec))
}
```

Note that we have not included input checks to the code: we must have $a < b$, tol $> 0$, $f(a)f(b) < 0$.

We look for $x^*$ in the interval $[2, 5]$, with a tolerance of $0.00005$.

```
bisection(f.test, 2, 5, 0.00005)
```

```
$x
[1] 3.665207

$k
[1] 15

$x_vec
 [1] 3.500000 4.250000 3.875000 3.687500 3.593750 3.640625 3.664062
 [8] 3.675781 3.669922 3.666992 3.665527 3.664795 3.665161 3.665344
[15] 3.665253 3.665207
```

What about $x_*$ in the interval $[5, 8]$, with the same tolerance?

```
bisection(f.test, 5, 8, 0.00005)
```

```
$x
[1] 5.759567

$k
[1] 15

$x_vec
 [1] 6.500000 5.750000 6.125000 5.937500 5.843750 5.796875 5.773438
 [8] 5.761719 5.755859 5.758789 5.760254 5.759521 5.759888 5.759705
[15] 5.759613 5.759567
```

The object `x_vec` lists the iterates $x_0$ to $x_{15}$: the convergence rate is indeed rather slow.

We can verify that the final iterates are quite close to $x^*$ and $x_*$.

```
f.test(3.665207)
f.test(5.759567)
```

```
[1] -1.348466e-05
[1] -1.691475e-05
```

Note however that we manually have to separate the problem into two sub-problems in order to capture both roots. If we were to try to find the roots of the test function over a longer interval containing both $x^*$ and $x_*$, such as [-10,10],[6] the algorithm would find at most one root.

6: We should first verify that $f(-10)f(10) < 0$.

```
bisection(f.test, -10, 10, 0.00005)
```

```
$x
[1] 3.665199

$k
[1] 18

$x_vec
 [1] 0.000000 5.000000 2.500000 3.750000 3.125000 3.437500 3.593750
 [8] 3.671875 3.632812 3.652344 3.662109 3.666992 3.664551 3.665771
[15] 3.665161 3.665466 3.665314 3.665237 3.665199
```

This highlight an important feature of numerical methods in the context of finding roots of a function: they are more useful when we already have a fairly good idea about the location of its roots.

Without the assumption check, the code will still run and might even converge to a root... but not necessarily so. How does the code respond for the test function over $[2, 8]$? Over $[2, 3]$?

### 4.2.2 Golden Ratio Method

We can also "solve" a continuous function $f : [a, b] \to \mathbb{R}$ by finding a value $x^*$ that **maximizes** $f$ over $[a, b]$ and/or a value $x_*$ that **minimizes** $f$ over $[a, b]$.[7]

7: Admittedly, the word "solve" does some heavy lifting here.

In this new context, the **Golden ratio method** plays an analogous role for **unimodal** continuous functions to that played by the bisection method in the original context.

This method is based on the **max/min theorem**: if $f \in C([a, b])$, then there exist $x^*, x_* \in [a, b]$ such that $f(x^*) \geq f(x) \geq f(x_*)$ for all $x \in [a, b]$.

Say we are seeking the minimal value. If $a = b$, then $x^* = x_* = a = b$, so assume that $a < b$. Let $\varphi = (1 + \sqrt{5})/2$, and set $a_0 = a$ and $b_0 = b$.

1. Set $c = b_0 - (b_0 - a_0)/\varphi$ and $d = a_0 + (b_0 - a_0)/\varphi$. We have

$$\begin{aligned}
\varphi < 2 &\implies \frac{b_0 - a_0}{2} < \frac{b_0 - a_0}{\varphi} \implies b_0 - a_0 < 2\left(\frac{b_0 - a_0}{\varphi}\right) \\
&\implies c = b_0 - \frac{b_0 - a_0}{\varphi} < a_0 + \frac{b_0 - a_0}{\varphi} = d, \\
1 < \varphi &\implies \frac{b_0 - a_0}{\varphi} < b_0 - a_0 \implies a_0 < b_0 - \frac{b_0 - a_0}{\varphi} \text{ and} \\
&\qquad\qquad a_0 + \frac{b_0 - a_0}{\varphi} < b_0,
\end{aligned}$$

   and so $[c, d] \subsetneq [a_0, b_0]$.
2. If $f(c) < f(d)$, set $a_1 = a_0$ and $b_1 = d$.
3. Otherwise, set $a_1 = c$ and $b_1 = b_0$.

The algorithm iterates with this new sub-interval $[a_1, b_1]$, to produce a sequence of nested intervals

$$[a_0, b_0] \subsetneq [a_1, b_1] \subsetneq \cdots [a_k, b_k] \subseteq \cdots$$

That the sequence of sub-intervals converges to the minimizer $x_*$ is guaranteed by the **nested interval theorem** since

$$\lim_{k\to\infty} (b_k - a_k) = \lim_{k\to\infty} \left(\frac{b_0 - a_0}{\varphi^{k+1}}\right) = 0.$$

We can guarantee a desired absolute error tolerance tol after $n$ iterations if

$$b_n - a_n = \frac{b_0 - a_0}{\varphi^{n+1}} \leq \text{tol},$$

which is to say

$$\varphi^{n+1} \geq \frac{b_0 - a_0}{\text{tol}} \implies n \geq \log_\varphi \left(\frac{b_0 - a_0}{\text{tol}}\right) - 1.$$

We learn in introductory calculus classes that a differentiable function reaches its max/min at a point where the derivative is 0 or at a point of the domain where the derivative does not exist.[8]

8: So we could use the bisection method on $f'$ instead, say.

The Golden Ratio method does not require knowledge of the derivative, however!

We implement the method (without checks) as follows.

**Golden Ratio method**

```
golden.min <- function(f, a, b, tol) {
  # initialization
  phi = (1 + sqrt(5))/2
  k <- 0 # 0th iteration
  c <- b - (b - a)/phi
  d <- a + (b - a)/phi

  a_vec <- c(a) # first iterate (lower endpoint)
  b_vec <- c(b) # first iterate (upper endpoint)

  # max number of iterations for absolute error control
  n <- ceiling(log((b - a) / tol) / log(phi) - 1)

  # Golden Ratio method
  while (k < n) {
      k <- k + 1
      if (f(c) < f(d)) {
        b <- d
      } else {
        a <- c
      }

    c <- b - (b - a)/phi
    d <- a + (b - a)/phi
    a_vec <- c(a_vec, a)
    b_vec <- c(b_vec, b)
    }

  # point estimate for minimizer
  x = (a + b)/2
  fx = f(x)

  return(list(fx=fx, x=x, k=k, a_vec=a_vec, b_vec=b_vec))
}
```

**Example** In the test function from the previous section, we see that the minimum occurs somewhere in $[4.5, 5]$.

```
golden.min(f.test,2,8,0.00005)
```

```
$fx
[1] -0.5

$x
[1] 4.712396

$k
[1] 24
```

```
$a_vec
 [1] 2.000000 2.000000 3.416408 4.291796 4.291796 4.291796 4.498447
 [8] 4.626165 4.626165 4.674948 4.674948 4.693582 4.705098 4.705098
[15] 4.709497 4.709497 4.711177 4.711177 4.711819 4.712216 4.712216
[22] 4.712216 4.712309 4.712367 4.712367

$b_vec
 [1] 8.000000 5.708204 5.708204 5.708204 5.167184 4.832816 4.832816
 [8] 4.832816 4.753882 4.753882 4.723732 4.723732 4.723732 4.716615
[15] 4.716615 4.713896 4.713896 4.712858 4.712858 4.712858 4.712612
[22] 4.712461 4.712461 4.712461 4.712425
```

From theoretical considerations, we already know that the minimal value of $f(x) = \sin(x) + 1/2$ is indeed $-1/2$.

### 4.2.3 Fixed Point Iteration Method

Both of the previous algorithms **converge slowly**, in the sense that while they do converge, they typically require an unreasonably large number of iterations to do so.

A root-finding problem $f(x) = 0$ can be transformed into an equivalent **fixed point problem** $g(x) = x$. For instance, if

$$g(x) = x - 2f(x) \quad \text{or} \quad g(x) = x + f^2(x),$$

then $f(x^*) = 0$ if and only if $g(x^*) = x^*$.[9] An input $x^*$ for which $g(x^*) = x^*$ is called a **fixed point** of $g$.

9: There are infinitely many different formulations for $g$, as we will see, but not all choices are suitable.

The following theorem gives sufficient conditions under which a function $g : [a, b] \to \mathbb{R}$ has a unique fixed point in $[a, b]$.

**Fixed Point Theorem:**

1. if $g \in C([a, b])$ and $g(x) \in [a, b]$ for all $x \in [a, b]$, then $g$ has a fixed point in $[a, b]$;
2. if $g'$ exists on $(a, b)$ and if there exists $0 < \rho < 1$ such that

$$|g'(x)| \leq \rho, \quad \forall x \in (a, b),$$

then $g$ has a unique fixed point in $[a, b]$.

**Proof:** define $\lambda : [a, b] \to \mathbb{R}$ by $\lambda(x) = g(x) - x$. Since $g(a) \geq a$, then $\lambda(a) = g(a) - a \geq a - a = 0$. Since $g(b) \leq b$, $\lambda(b) = g(b) - b \leq b - b = 0$. But $g$ is continuous; according to the the intermediate value theorem, there is thus a $p \in [a, b]$ such that $\lambda(p) = 0$, which is to say $g(p) = p$.

Now suppose $p^*, p_* \in [a, b]$ are two fixed points of $g$; then

$$|p^* - p_*| = |g(p^*) - g(p_*)|.$$

According to the mean value theorem,[10] if $g$ is differentiable, there is a $c$ between $p_*$ and $p^*$ such that

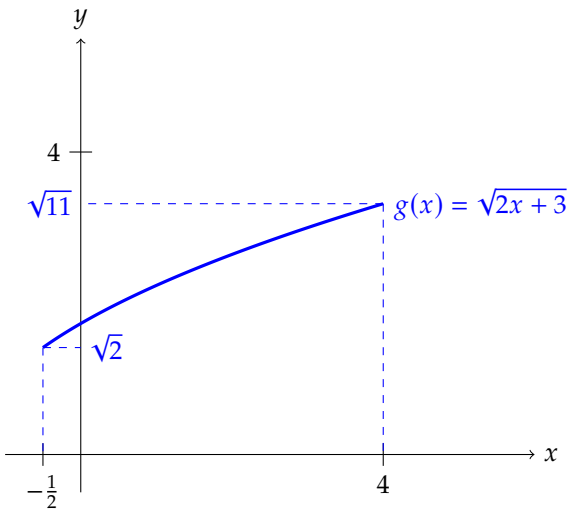
10: See [2] for details.

$$|p^* - p_*| = |g(p^*) - g(p_*)| = |g'(c)| \cdot |p^* - p_*| \leq \rho |p^* - p_*| < |p^* - p_*|.$$

This can only happen if $p^* = p_*$, and so the fixed point is unique. ■

**Example** Consider the equation $f(x) = x^2 - 2x - 3 = 0$, $x \in \left[-\frac{1}{2}, 4\right]$, and the equivalent fixed point equation $x = g(x) = \sqrt{2x+3}$. Show that $g$ has a unique fixed point, and so that $f$ has a unique root, in $\left[-\frac{1}{2}, 4\right]$.

**Solution:** any fixed point of $g$ satisfies $x = \sqrt{2x+3} \implies x^2 - 2x - 3 = 0$, and thus is a root of $f$. Over the interval $\left[-\frac{1}{2}, 4\right]$, $g$ is continuous and increasing, as shown below.



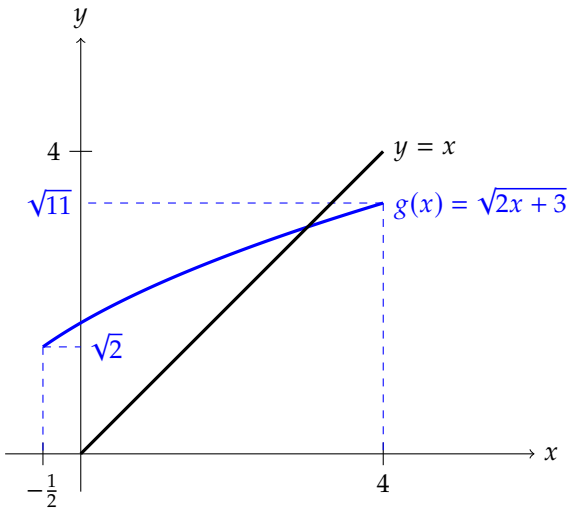

Thus, for any $x \in \left[-\frac{1}{2}, 4\right]$, we have:

$$-\tfrac{1}{2} \le \sqrt{2} \le g\left(-\tfrac{1}{2}\right) \le g(x) \le g(4) \le \sqrt{11} \le 4 \implies g\left(\left[-\tfrac{1}{2}, 4\right]\right) \subseteq \left[-\tfrac{1}{2}, 4\right].$$

Since $g'(x) = \frac{1}{\sqrt{2x+3}}$, then we also have:

$$|g'(x)| \le \frac{1}{\sqrt{2}} < 1, \quad \text{over} \left[-\frac{1}{2}, 4\right].$$

As the assumptions of the theorem are satisfied, $g$ admits a unique fixed point over $\left[-\frac{1}{2}, 4\right]$.



For a given continuous function $g$ on $[a, b]$ and initial iterate $x_0$, the **fixed point iteration** process reads as:

$$x_k = g(x_{k-1}), \quad k \geq 1.$$

If $\{x_k\}$ converges to some $x_* \in [a, b]$, then $x_*$ is a fixed point of $g$;[11] indeed,

$$x_* = \lim_{k\to\infty} x_k = \lim_{k\to\infty} g(x_{k-1}) = g\Big(\lim_{k\to\infty} x_{k-1}\Big) = g(x_*).$$

11: Note that the fixed point is not necessarily unique.

**Illustration of the Fixed Point Procedure** Consider the problem of solving the equation

$$f(x) = x + \ln(1 + x) - 2 = 0, \quad x \in [0, 5],$$

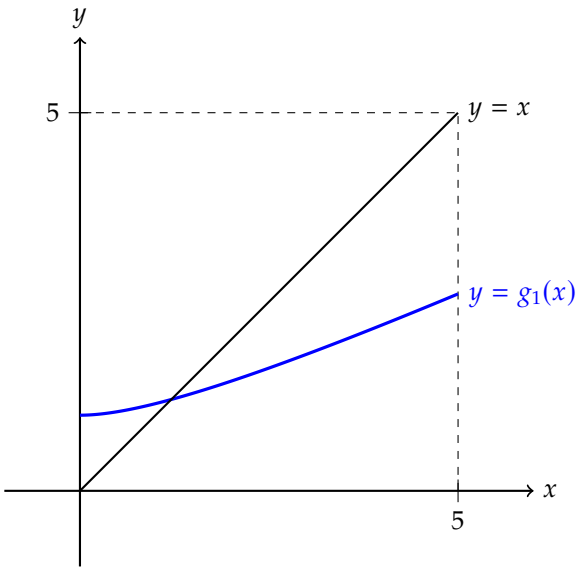
and the three equivalent fixed point equations:

1. $x = g_1(x) = x - \frac{1}{2}[x + \ln(1 + x) - 2]$
2. $x = g_2(x) = 2 - \ln(1 + x)$
3. $x = g_3(x) = e^{2-x} - 1$

We provide a detailed illustration of how the method works on $g_1$; for $g_2$ and $g_3$, we only show the final picture.[12]

12: Is it clear that all the fixed point problems are equivalent to the root-finding problem?

First, we plot the graphs of $y = g_1(x)$ and $y = x$; any intersection of the two curves over the domain $[a, b](= [0, 5])$ must satisfy $g_1(x) = x$ and so is a fixed point of $g_1$ over the domain.



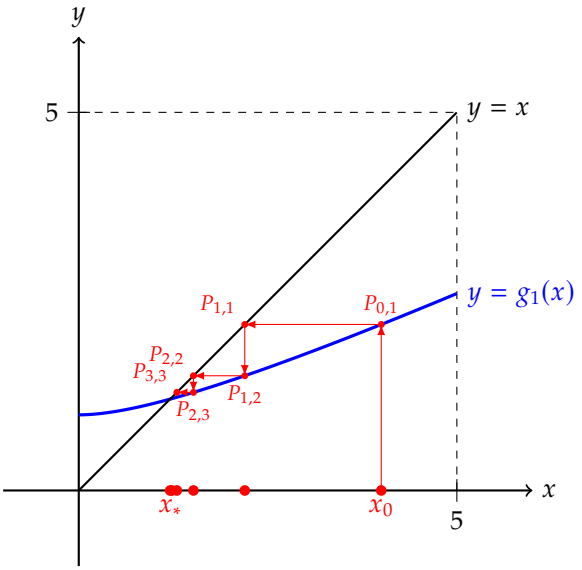

Graphically, we see that there is one such fixed point. How does the procedure find it?

We need an $x_0$ in the domain; we select $x_0 = 4$, for no particular reason, and we obtain:

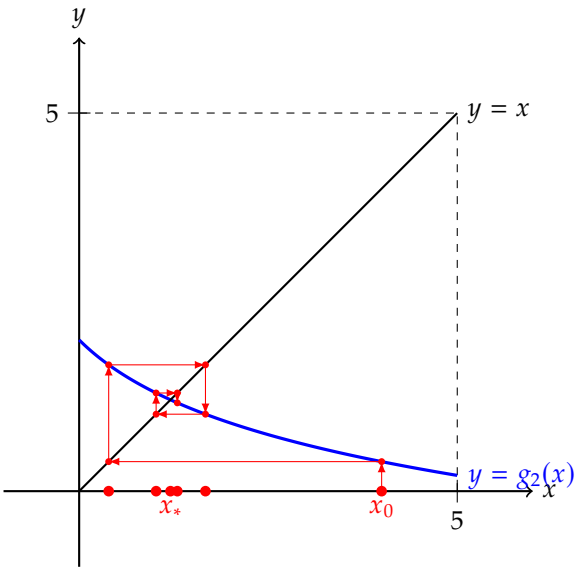
$$x_1 = g_1(x_0) = g_1(4) = 4 - \tfrac{1}{2}[4 + \ln(1 + 4) - 2] \approx 2.195281;$$
$$x_2 = g_1(x_1) = g_1(2.195281) \approx 1.516803;$$
$$x_3 = g_1(x_2) = g_1(1.516803) \approx 1.296907, \quad \text{etc.}$$

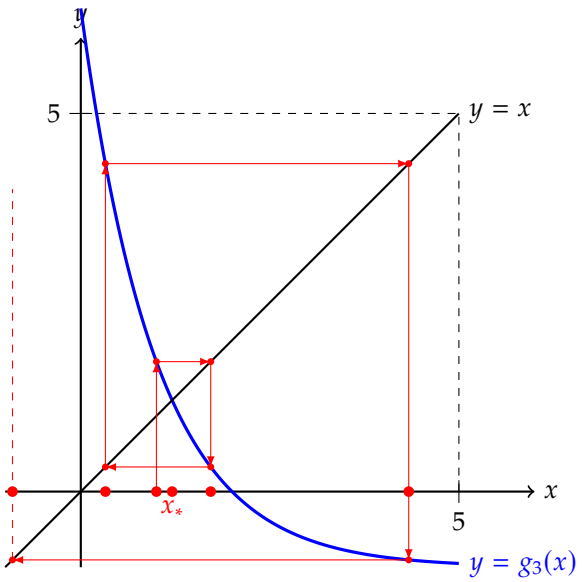
For $k \geq 1$, each $x_k$ plays two roles: it is the $y-$coordinate of a point on the curve $y = g_1(x)$, which becomes the $x-$coordinate of a point on the curve $y = x$, which is then fed back into $g_1$, and so on.

Graphically, this is represented as a rectangular "curve" which converges to the point $(x_*, x_*) \approx (1.20794, 1.20794)$ in a manner resembling a staircase; the label $P_{i,j}$ represents the point with coordinates $(x_i, x_j)$.




With $g_2$ and $g_3$, the fixed point iterations instead take on the following forms.




We see that the method converges for $g_1$ and $g_2$, both to the same fixed point $x_*$, but not for $g_3$, even though $x_*$ is a fixed point for the latter. Note that $|g_i'(x_*)| < 1$ for $i = 1, 2$, while $|g_3'(x_*)| > 1$.

**Convergence** So when can we be sure that fixed point iteration converges to a fixed point?

**Fixed Point Theorem (Reprise):** let $g : [a, b] \to \mathbb{R}$ be a function satisfying hypotheses 1. and 2. of the fixed point theorem of page 193. Then for any initial iterate $x_0 \in [a, b]$, the sequence $\{x_k\}$ defined by

$$x_k = g(x_{k-1}), \quad k \geq 1,$$

converges to the unique fixed point $x_*$ of $g$ in $[a, b]$.

**Proof:** the original fixed point theorem shows that $g$ has a unique fixed point $x_*$ in $[a, b]$. Let $x_0 \in [a, b]$; we must show that $x_k \to x_*$ as $k \to \infty$.

On the one hand, we have $x_k - x_* = g(x_{k-1}) - g(x_*)$ for all $k \geq 1$. On the other hand, since $g$ is differentiable over $(a, b)$, the mean value theorem implies that

$$g(x_{k-1}) - g(x_*) = g'(c_k)(x_{k-1} - x_*), \quad \text{for some } c_k \text{ between } x_{k-1} \text{ and } x_*.$$

Thus,

$$|x_k - x_*| = |g(x_{k-1}) - g(x_*)| = |g'(c_k)||x_{k-1} - x_*| \leq \rho|x_{k-1} - x_*|,$$

by hypothesis. We then have, recursively,

$$|x_k - x_*| \leq \rho|x_{k-1} - x_*| \leq \rho^2|x_{k-2} - x_*| \leq \cdots \leq \rho^k|x_0 - x_*| \to 0$$

as $k \to \infty$ since $\rho < 1$, which completes the proof. ■

**Corollary on the Error Estimates:** under the hypotheses of the fixed point theorem, we can show that:

1. $|x_k - x_*| \leq \rho^k \cdot \max\{x_0 - a, b - x_0\}$ for $k \geq 0$;
2. $|x_k - x_*| \leq \frac{\rho^k}{1-\rho} \cdot \max\{x_0 - a, b - x_0\}$ for $k \geq 1$.

Note that the smaller the value $\rho < 1$ is, the faster the sequence converges to the fixed point $x_*$ of $g$.

**Stopping Criterion** Ideally, we would like the fixed point procedure to stop whenever the error satisfies $e_k = |x_k - x_*| < \text{tol}$ for some prescribed tolerance $\text{tol} > 0$. However, the exact fixed point $x_*$ is not known; instead, we can use the following **stopping criterion**:

$$|x_{k+1} - x_k| < \text{tol}.$$

The value $r_k = |x_k - g(x_k)| = |x_k - x_{k+1}|$ is the **residual** of the fixed point procedure at step $k$. Note that $r_k \approx \text{tol}$ does not imply that $e_k \approx \text{tol}$:

$$\begin{aligned} x_k - x_* &= x_k - x_{k+1} + x_{k+1} - x_* \\ &= x_k - x_{k+1} + g(x_k) - g(x_*) \\ &= x_k - x_{k+1} + g'(c_k)(x_k - x_*), \quad \text{for some } c_k \text{ between } x_k \text{ and } x_*, \end{aligned}$$

so that

$$(1 - g'(c_k))(x_k - x_*) = x_k - x_{k+1} \implies e_k = \frac{r_k}{|1 - g'(c_k)|}.$$

If $|g'(c_k)| \ll 1$, then $e_k \approx r_k \approx \text{tol}$; if $g'(c_k) \approx 1$, then it is possible that $e_k \gg \text{tol}$!

The fixed point iteration is summarized in the following algorithm.

**Algorithm:** fixed point iteration

**Input:** $g$ with the appropriate properties on $[a, b]$, $x_0$, tol $> 0$, $N_{\max}$
**Output:** approximation $p$ of a fixed point $x_*$ of $g$, number of iterations $n$

1 **Initialization:** $x_1 = g(x_0)$, $r_0 = |x_0 - x_1|$, $k = 0$;
2 **While** $r_k >$ tol and $k < N_{\max}$ **do**
3 $k = k + 1$;
4 $x_{k+1} = g(x_k)$;
5 $r_k = |x_k - x_{k+1}|$;
6 **End**
7 $p = x_{k+1}$, $n = k + 1$.

Here is an implementation of the method in R.

**Fixed point method**

```
fixed_point <- function(g, x0, tol, Nmax) {
  # initialization
  x_old <- x0
  x <- g(x_old)
  res <- abs(x - x_old)

  k <- 1
  x_vec <- c(x0, x)

  # fixed point iteration
  while (res > tol && k < Nmax) {
    k <- k + 1
    x_old <- x
    x <- g(x_old)
    res <- abs(x - x_old)
    x_vec <- c(x_vec, x)

    # tolerance not reached
    if (k == Nmax && res > tol) {
      cat('Nmax iterations reached without
       satisfying the prescribed tolerance\n')
    }
  }

  return(list(x = x, k = k, x_vec = x_vec))
}
```

**Example** We can find the fixed point $x_*$ of $g(x) = -\cos(x)$ with tol $=$ 0.00005 as follows.

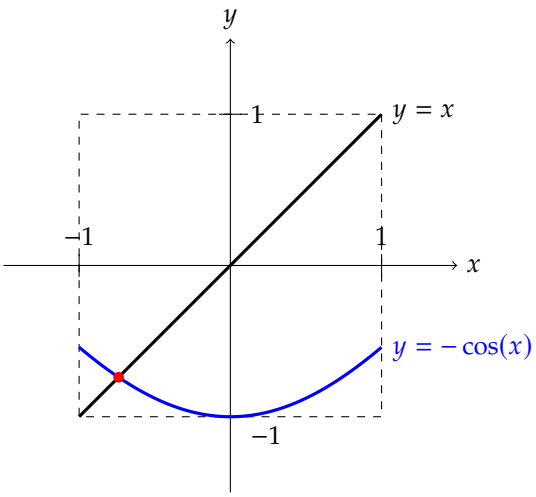
```
g.test <- function(x){ -cos(x) }
fixed_point(g.test, 1, 0.00005, 300)
```

```
$x
[1] -0.7390714

$k
[1] 25

$x_vec
 [1]  1.0000000 -0.5403023 -0.8575532 -0.6542898
 [5] -0.7934804 -0.7013688 -0.7639597 -0.7221024
 [9] -0.7504178 -0.7314040 -0.7442374 -0.7356047
[13] -0.7414251 -0.7375069 -0.7401473 -0.7383692
[17] -0.7395672 -0.7387603 -0.7393039 -0.7389378
[21] -0.7391844 -0.7390183 -0.7391302 -0.7390548
[25] -0.7391056 -0.7390714
```

We can easily verify that the output is at the very least quite near $x_*$, numerically and graphically.

```
g.test(-0.7390714)+0.7390714
```

```
[1] -2.2984e-05
```




**Order of the Method** In the proof of the fixed point theorem (reprise), we saw that

$$|g(x_k) - g(x_*)| = |g'(c_k)||x_k - x_*|$$

for some $c_k$ between $x_k$ and $x^*$.

If $g$, $g'$ are continuous over $[a, b]$ and $\lim_{k\to\infty} x_k = x_*$, where $g(x_*) = x_*$, then we see that

$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} = \lim_{k\to\infty} \frac{|g(x_k) - g(x_*)|}{|x_k - x_*|} = \lim_{k\to\infty} |g'(c_k)| = |g'(x_*)|,$$

since $0 \le |x_* - c_k| \le |x_* - x_k| \to 0$ and $g'$ is continuous.

Thus, $|g'(x_*)|$ provides a measure of the **speed of convergence** of the sequence $\{x_k\}$.

Let $x_k \to x_*$ be such that $x_k \neq x_*$ for all $k$.

1. If
$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} = \lambda \in (0, 1),$$
then $\{x_k\}$ converges **linearly** to $x_*$.

2. If
$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} = 1,$$
then $\{x_k\}$ converges **sublinearly** to $x_*$.

3. If
$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|} = 0,$$
then $\{x_k\}$ converges **superlinearly** to $x_*$.

4. Set $\alpha \geq 1$ an integer; if
$$\lim_{k\to\infty} \frac{|x_{k+1} - x_*|}{|x_k - x_*|^\alpha} = \lambda > 0,$$
then $\{x_k\}$ converges to $x_*$ **with order** $\alpha$; in this case, the value $\lambda$ is known as the **asymptotic error constant**.[13]

13: If $\alpha = 1$, the convergence is linear and we must have $\lambda < 1$; if $\alpha = 2$, the convergence is **quadratic**.

We say that a fixed point iteration $x_k = g(x_{k-1})$ is **of order** $\alpha$ if $\{x_k\}$ converges to a fixed point $x^*$ with order $\alpha$. In that case, when $x_k$ is sufficiently close to $x^*$ then we have

$$|x_{k+1} - x_*| \approx \lambda |x_k - x_*|^\alpha.$$

**Example** Assume that we have two fixed point iterations, one with order $\alpha = 1$ and $\lambda = 0.5$, and the other with order $\alpha = 2$ and $\lambda = 1$. Moreover, suppose that $|x_0 - x^*| = 10^{-1}$. Then we would expect to observe something like the following table.

| $k$ | $\lvert x_k - x_*\rvert$ | $\alpha = 1, \lambda = 0.5$ | $\alpha = 2, \lambda = 1$ |
|---|---|---|---|
| 0 | $\lvert x_0 - x_*\rvert$ | 0.1 | 0.1 |
| 1 | $\lvert x_1 - x_*\rvert$ | 0.05 | 0.01 |
| 2 | $\lvert x_2 - x_*\rvert$ | 0.025 | 0.0001 |
| 3 | $\lvert x_3 - x_*\rvert$ | 0.0125 | 0.00000001 |
| ⋮ | ⋮ | ⋮ | ⋮ |

In both cases, $e_k \to 0$; the convergence is quicker in the second case.[14]

14: In practice, this means that we will not need as many iterations of the fixed point procedure before exiting the 'while' loop in the algorithm.

As mentioned above, the exact fixed point $x_*$ is not known, and so we cannot compute the absolute error $e_k = |x_k - x^*|$ exactly. Instead, we estimate the order $\alpha$ of a fixed point iteration with the help of the residual $r_k = |x_k - x_{k+1}|$ and search for the value of $\alpha$ for which that the ratio $r_{k+1}/r_k^\alpha$ converges to a positive constant.

We already know the relationship between $r_k$ and $e_k$: $r_k = |1 - g'(c_k)|e_k$ for some $c_k$ between $x_k$ and $x_*$; so if $\dfrac{e_{k+1}}{e_k^\alpha} \to \lambda > 0$, then

$$\lim_{k\to\infty} \frac{r_{k+1}}{r_k^\alpha} = \lim_{k\to\infty} \frac{|1 - g'(c_{k+1})|}{|1 - g'(c_k)|^\alpha} \cdot \frac{e_{k+1}}{e_k^\alpha} = \frac{\lambda}{|1 - g'(x_*)|^{\alpha-1}} > 0.$$

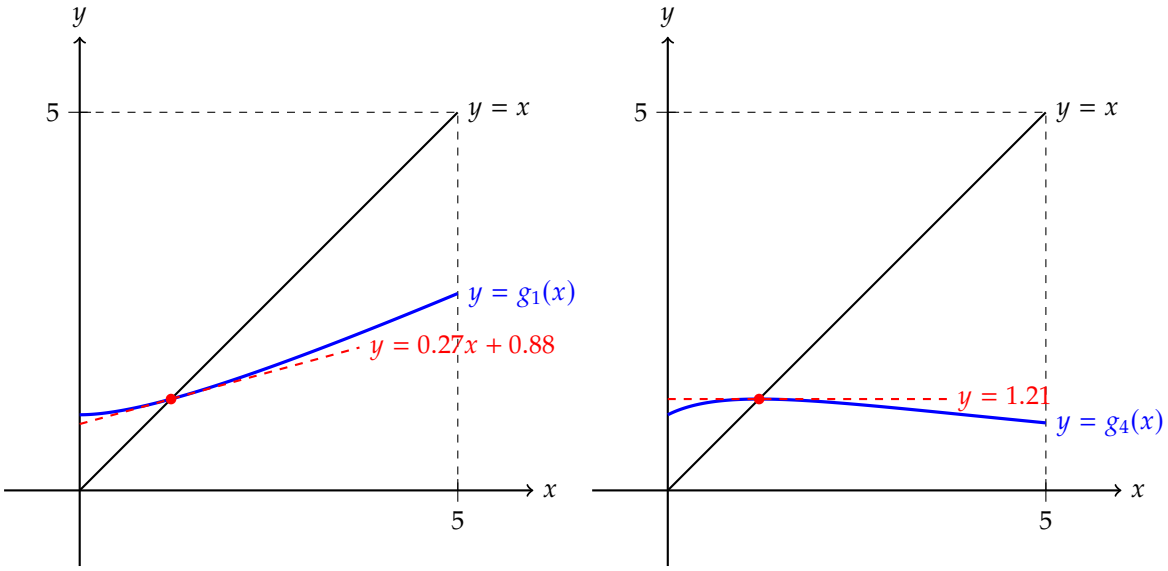
**Example** Consider again the equation

$$f(x) = x + \ln(1 + x) - 2 = 0, \quad x \in [0, 5],$$

and the equivalent fixed point equations $x = g_i(x)$, $i = 1, 4$, with

$$g_1(x) = x - \frac{1}{2}[x + \ln(1 + x) - 2] \quad \text{and} \quad g_4(x) = \frac{3x + 2 - (1 + x)\ln(1 + x)}{2 + x}.$$

The charts are shown below, with their tangent lines at $(x_*, x_*)$.




In both cases, the derivative at the fixed point falls in $(-1, 1)$, so the fixed point procedure converges for every initial iterate $x_0 \in [0, 5]$; note, however, that $|g_4'(x_*)| < |g_1'(x_*)|$, so we expect the convergence to the fixed point to be of higher order for $g_4$ than for $g_1$.

We run the algorithm with $x_0 = 4$ and tol $= 10^{-8}$.

```
g1 <- function(x){x-0.5*(x + log(x+1) - 2)}
x0 = 2
tol = 10^(-8)
Nmax = 1000
fp1 = fixed_point(g1, x0, tol, Nmax)
n1 = length(fp1$x_vec)
```

We compute the residuals, and study the ratios $r_{k+1}/r_k$:

```
res1 = abs(fp1$x_vec[2:n1] - fp1$x_vec[1:(n1-1)] )
res1[2:(n1-1)]/res1[1:(n1-2)]
```

```
[1] 0.3159122 0.2883925 0.2779797 0.2747903
[5] 0.2738879 0.2736387 0.2735703 ...
```

We see that the sequence of ratios seems to converge to $\lambda_1 = 0.2735... > 0$, and so the fixed point convergence is at least linear. For comparison's sake, we also take a look at the ratios $r_{k+1}/r_k^2$:

```
res1[2:(n1-1)]/(res1[1:(n1-2)])^2
```

```
[1]   0.5751114   1.6618933   5.5545402  19.7525620
[5]  71.6462604 261.3517413 954.8594146  ...
```

The sequence of ratios does not seem to converge.

If we repeat the above commands for $g_4$, we find that the fixed point iteration with $g_4$ is of order 2.

```
g4 <- function(x){(3*x+2-(1+x)*log(1+x))/(2+x)}
x0 = 2; tol = 10^(-8); Nmax = 1000;
fp4 = fixed_point(g4, x0, tol, Nmax)
n4 = length(fp4$x_vec)
```

We compute the residuals, and study the ratios $r_{k+1}/r_k$:

```
res4 = abs(fp4$x_vec[2:n4] - fp4$x_vec[1:(n4-1)] )
res4[2:(n4-1)]/res4[1:(n4-2)]
```

```
[1] 3.862611e-02 2.290711e-03 5.146737e-06 ...
```

We see that the sequence of ratios seems to converge to $\lambda_4 = 0$. We take a look at the ratios $r_{k+1}/r_k^2$:

```
res4[2:(n4-1)]/(res4[1:(n4-2)])^2
```

```
[1] 0.04687867 0.07197530 0.07059517 ...
```

These ratios do seem to converge to a non-zero $\lambda_4$, so the convergence is at least of order 2. And for $r_{k+1}/r_k^3$?

```
res4[2:(n4-1)]/(res4[1:(n4-2)])^2
```

```
[1]   0.05689441   2.26150082 968.31804790 ...
```

The sequence of ratios does not seem to converge.

In general, the order of the convergence to a fixed point $x_*$ of $g$ is linked to the order of differentiability of $g$ at $x_*$.

**Theorem:** let $g \in C^{\alpha}([a,b])$, $\alpha \geq 1$ an integer, and let $x^* \in [a,b]$ be a fixed point of $g$, with $x_0$ sufficiently near $x_*$. If

$$0 < |g'(x^*)| < 1,$$

then the fixed point iteration $x_k = g(x_{k-1})$, $k \geq 1$, is only of order 1. If

$$g'(x_*) = g''(x_*) = \cdots = g^{(\alpha-1)}(x_*) = 0 \quad \text{and} \quad g^{(\alpha)}(x_*) \neq 0,$$

then the fixed point iteration is of order $\alpha$.

**Proof:** we only provide an outline for the case $\alpha > 1$. For any $x, x_0 \in [a,b]$, with $x_0$ "sufficiently close" to $x$, we apply **Taylor's theorem** to $g$,[15] around its fixed point $x_* \in [a,b]$, and write

15: See [2] for details.

$$\begin{aligned} g(x) &= g(x_*) + g'(x_*)(x-x_*) + \frac{1}{2}g''(x_*)(x-x_*)^2 + \cdots + \frac{1}{(\alpha-1)!}g^{(\alpha-1)}(x_*)(x-x_*)^{\alpha-1} + \frac{1}{\alpha!}g^{(\alpha)}(c_x)(x-x_*)^{\alpha} \\ &= g(x_*) + \frac{1}{\alpha!}g^{(\alpha)}(c_x)(x-x_*)^{\alpha}, \end{aligned}$$

for some $c_x$ between $x$ and $x_*$.[16] When $x = x_k$, we get

16: The mean value theorem is a special case of Taylor's theorem, with $\alpha = 1$.

$$x_{k+1} - x_* = g(x_k) - g(x_*) = \frac{1}{\alpha!}g^{(\alpha)}(c_x)(x-x_*)^{\alpha},$$

where $c_k$ lies between $x_k$ and $x_*$. Since $x_k \to x_*$, then $c_k \to x_*$ and

$$\lim_{k\to\infty} \frac{|x_{k+1}-x_*|}{|x_k - x_*|^{\alpha}} = \lim_{k\to\infty} \frac{1}{\alpha!}|g^{(\alpha)}(c_k)| = \frac{1}{\alpha!}|g^{(\alpha)}(x_*)|,$$

which is non-zero, by assumption. ■

This explains why some choices of $g$ are better than others; of course, this is of limited applicability as we need to know $x_*$ before we can use this last result to increase the convergence order of the procedure... but if we already know $x_*$, there is no need to improve the speed of convergence.

### 4.2.4 Newton's Method

Newton's method is one of the most frequently-used "fast" method for solving **nonlinear equations**, although in many applications, it is often supplanted by task-specific methods, such as **gradient descent methods**.[17]

17: See Chapters 5 and 31, and Section 4.3.2.

We wish to solve the equation $f(x) = 0$, with $f \in C^2([a,b])$. Assume that $x^* \in [a,b]$ is a root of $f$ and let $x_k \in [a,b]$. According to Taylor's theorem, there is a $c_k$ between $x^*$ and $x_k$ such that

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x) + \frac{1}{2}f''(c_k)(x^* - x_k)^2.$$

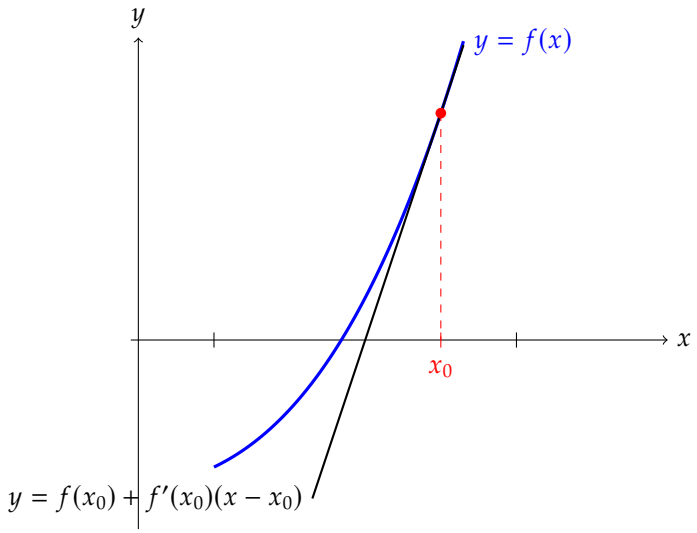
If $x_k$ is near $x^*$, we expect $|x^* - x_k|$ to be small, so that $|x^* - x_k|^2 \ll |x^* - x_k|$. Moreover, if $f'(x_k) \neq 0$, then

$$0 = f(x^*) \approx f(x_k) + f'(x_k)(x^* - x) \implies x^* \approx x_k - \frac{f(x_k)}{f'(x_k)}.$$

Starting from $x_0$, **Newton's method** generates the sequence $\{x_k\}$ defined by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k \geq 0.$$

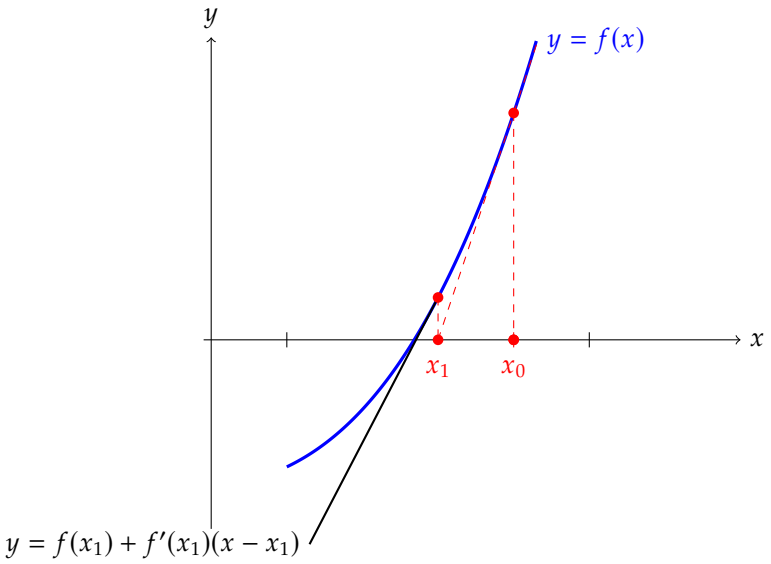
**Illustration of the Method** Let $f : [a, b] \to \mathbb{R}$ be the $C^2$ function whose graph is displayed below, and let $x_0 \in [a, b]$ be near $x^*$. Draw the tangent to $f$ at $x_0$.




The equation of the tangent is $y = f(x_0) + f'(x_0)(x - x_0)$; the intersection of the line with the $x-$axis at

$$0 = f(x_0) + f'(x_0)(x - x_0) \implies x = x_0 - \frac{f(x_0)}{f'(x_0)},$$

which is exactly the first Newton iterate $x_1$. Repeat this procedure starting from $x_1$ to obtain $x_2$, and so on.



**Theorem:** let $f \in C^2([a,b])$. If $x^* \in [a,b]$ is such that $f(x^*) = 0$ and $f'(x^*) \neq 0$, then the sequence $\{x_k\}$ generated by Newton's method converges (at least) quadratically to $x^*$ for any $x_0$ sufficiently near $x^*$.

**Proof:** Newton's method can be recast as a fixed point iteration for the function defined by $g(x) = x - \frac{f(x)}{f'(x)}$. At $x = x^*$,

$$g(x^*) = x^* - \frac{f(x^*)}{f'(x^*)} = x^* - \frac{0}{f'(x^*)} = x^*,$$

so $x^*$ is a fixed point of $g$. But

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2} \implies g'(x^*) = \frac{f(x^*)f''(x^*)}{[f'(x^*)]^2} = 0,$$

so the order of convergence is at least $\alpha = 2$ according to the last theorem of Section 4.2.3. ■

Newton's method may not converge if $x_0$ is too removed from $x^*$, or if the iterations gets caught in a cycle.

**Remark:** if $f'(x^*) = 0$, then Newton's method may still converge with order 1. For instance, $f(x) = x^2$ vanishes at $x^* = 0$ and $f'(x^*) = 0$. We then have

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2}{2x_k} = \frac{1}{2}x_k = \cdots = \left(\frac{1}{2}\right)^k x_0 \to 0 = x^*$$

as $k \to \infty$, and so $x_k \to x^*$. However, for the equivalent fixed point problem $x = g(x) = x/2$, we have $g(x^*) = 0$ and $g'(x^*) = 1/2 \neq 0$, so the convergence is only linear.

Newton's algorithm is summarized in the following algorithm.

**Algorithm:** Newton's method

**Input:** $f$, $f'$, $x_0$, tol $> 0$, $N_{\max}$
**Output:** approximation $p$ of a root $x^*$ of $f$, number of iterations $n$

1 **Initialization:** $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$, $r_0 = |x_1 - x_0|$, $k = 0$;
2 **While** $r_k >$ tol and $k < N_{\max}$ **do**
3 $k = k + 1$;
4 $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$;
5 $r_k = |x_{k+1} - x_k|$;
6 **End**
7 $p = x_{k+1}$, $n = k + 1$.

Here is an implementation of the method in R.

**Newton's method**

```
newton <- function(f, df, x0, tol, Nmax) {
  x_old <- x0
  x <- x_old - f(x_old) / df(x_old)
  res <- abs(x - x_old)
```

```
  k <- 1
  x_vec <- c(x0, x)

  while(res > tol && k < Nmax) {
    k <- k + 1
    x_old <- x
    x <- x_old - f(x_old) / df(x_old)
    res <- abs(x - x_old)

    x_vec <- c(x_vec, x)

    if(k == Nmax && res > tol) {
      cat('Nmax iterations reached without
       satisfying the prescribed tolerance\n')
    }
  }

  return(list(x=x, k=k, x_vec=x_vec))
}
```

**Example** We are looking for roots of the function $f$ defined by $f(x) = x^2 - 4$, whose derivative is $f'$ defined by $f'(x) = 2x$.

```
f <- function(x){x^2 - 4}
df <- function(x){2*x}
```

We initialize the algorithm as follows.

```
x0 <- 1
tol <- 1e-5
Nmax <- 100
```

What does Newton's method find?

```
result <- newton(f, df, x0, tol, Nmax)
print(result$x)
```

```
[1] 2
```

```
print(result$k)
```

```
[1] 5
```

```
print(result$x_vec)
```

```
[1] 1.00000 2.50000 2.05000 2.00061 2.00000 2.00000
```

We know, theoretically, that $f(2) = 0$. But $x^* = 2$ is not the only root of $f$. One of the drawbacks of iterative procedures in the search for roots is that a sequence $\{x_k\}$ converges to one limit (at most).

When we know that there are other roots, we can try playing with the parameters to generate sequences converging to those, but in general that knowledge is not available to us.[18] We can exhibit the other root by using a different $x_0$.

18: That is, in no small part, exactly why we are looking for roots in the first place.

```
x0 <- -1
result <- newton(f, df, x0, tol, Nmax)
print(result$x)
```

```
[1] -2
```

```
print(result$k)
```

```
[1] 5
```

```
print(result$x_vec)
```

```
[1] -1.00000 -2.50000 -2.05000 -2.00061 -2.00000 -2.00000
```

Newton's method, being of order 2, is usually quite fast, but the function's derivative must be known.

### 4.2.5 Secant Method

It might be costly to evaluate $f'$; the **secant method** is a variation of Newton's method where only evaluations of $f$ are needed. The idea is to approximate $f'(x_k)$ by a difference quotient:

$$f'(x_k) = \lim_{x \to x_k} \frac{f(x) - f(x_k)}{x - x_k} \approx \frac{f(x_{k-1}) - f(x_k)}{x_{k-1} - x_k}.$$

The quality of the approximation increases when $x_{k-1}$ is "close" to $x_k$.

Given initial iterates $x_0 \neq x_1 \in [a, b]$ for which $f(x_0) \neq f(x_1)$, the sequence generated by the secant method is similar to the Newton sequence, but substituting $f'(x_k)$ by $\frac{f(x_{k-1}) - f(x_k)}{x_{k-1} - x_k}$:

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}, \quad k \geq 1.$$

Graphically, we obtain $x_2$ as the intersection of the $x-$axis with the line joining the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$.

## 4.3 Systems of Equations

In practice, data problems often give rise to systems of $m$ equations in $n$ unknowns (as opposed to 1 equation in 1 variable). The nature of these systems (**linear** vs. **non-linear**) affects the choice of solution method.[19]

19: Methods derived specifically for linear systems are not easily applicable to non-linear systems, but methods for non-linear systems are usually applicable to linear systems as well.

### 4.3.1 Linear Systems

In simple linear regression, for instance, we are trying to find the coefficients $\beta_0$ and $\beta_1$ that "best" fit the data $\{(X_i, Y_i)\}$ in the least square sense: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \ldots, n$.

In Chapter 8, we see that the estimators $b_0, b_1$ are the solutions of

$$n\overline{Y} = n\beta_0 + n\overline{X}\beta_1, \qquad S_{xy} + n\overline{XY} = n\overline{X}\beta_0 + (S_{xx} + n\overline{X}^2)\beta_1.$$

This is a linear system of two equations in two unknowns, which we can re-write in matrix form as $A\boldsymbol{\beta} = \mathbf{c}$. If $A$ is invertible, the estimated solution vector is $A^{-1}\mathbf{c}$.[20]

20: See Chapter 3 for details.

Consider the linear system $A\mathbf{x} = \mathbf{b}$, where $A$ is an $m \times n$ matrix, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. If $m = n$ and $A$ is **invertible**, the system has a unique solution, $\mathbf{x} = A^{-1}\mathbf{b}$.

In practice, we rarely solve the linear system by explicitly computing $A^{-1}$, especially if $n$ is large.[21]

21: With a computer capable of teraflop speeds, it would take roughly $10^{141}$ years to compute the inverse of an $100 \times 100$ matrix using cofactors or Cramer's rule!

We will briefly discuss two types of methods for solving $A\mathbf{x} = \mathbf{b}$ that do not involve computing $A^{-1}$: direct methods and iterative methods.

#### Direct Methods

In theory, a **direct** method finds the exact solution in a finite number of steps; in practice, the solution is "polluted" by round-off error.

**Gaussian Elimination and Backward Substitution** A linear system may be easy to solve when $A$ has an advantageous structure, such as if it is **upper** (or lower) **triangular**:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n-1} & a_{1,n} \\ 0 & a_{2,2} & \cdots & a_{2,n-1} & a_{2,n} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & 0 & 0 & a_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{pmatrix}.$$

From the last row $a_{n,n}x_n = b_n$, we obtain $x_n = b_n/a_{n,n}$, assuming that $a_{n,n} \neq 0$.[22]

22: All diagonal entries of a triangular matrix $A$ must be non-zero if $A$ is invertible.

Then, from the penultimate row, we have

$$a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \implies x_{n-1} = \frac{1}{a_{n-1,n-1}}\left(b_{n-1} - a_{n-1,n}x_n\right),$$

and so on until we reach the first row.

The formal procedure for triangular matrices are provided below.

---

**Algorithm:** backward substitution

---

**Input:** $A$ upper triangular, $n \times n$, with $a_{i,i} \neq 0$ for all $1 \leq i \leq n$
**Output:** solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$

1 **For** $i = n, n-1, \ldots, 1$ **do**

2 $$x_i = \frac{1}{a_{i,i}} \left( b_i - \sum_{j=i+1}^{n} a_{i,j} x_j \right)$$

3 **End**

4 $\mathbf{x} = (x_1, \ldots, x_n)^\top$

---

**Algorithm:** forward substitution

---

**Input:** $A$ lower triangular, $n \times n$, with $a_{i,i} \neq 0$ for all $1 \leq i \leq n$
**Output:** solution $\mathbf{x}$ of $A\mathbf{x} = \mathbf{b}$

1 **For** $i = 1, 2, \ldots, n$ **do**

2 $$x_i = \frac{1}{a_{i,i}} \left( b_i - \sum_{j=1}^{i-1} a_{i,j} x_j \right)$$

3 **End**

4 $\mathbf{x} = (x_1, \ldots, x_n)^\top$

---

In general, the matrix $A$ is not triangular, but it can be brought to a triangular form via **Gaussian elimination**.[23]

23: See Section 3.4.1 for more details.

**Example** To find the solution of the linear system

$$\begin{cases} x_1 + x_2 + 3x_4 = 4 \\ 2x_1 + x_2 - x_3 + 3x_4 = 1 \\ 3x_1 - x_2 - x_3 + 2x_4 = -3 \\ -x_1 + 2x_2 + 3x_3 - x_4 = 4 \end{cases}$$

we first form the augmented matrix $[A \mid \mathbf{b}]$ and reduce it to its echelon form to obtain

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right).$$

We can read the solution from the reduced matrix directly, *via* backward substitution:

$$\begin{aligned} x_4 &= 13/13 = 1, \\ x_3 &= \frac{1}{3}(13 - 13 \cdot 1) = 0, \\ x_2 &= \frac{1}{-1}(-7 - (-1) \cdot 0 - (-5) \cdot 1) = 2, \\ x_1 &= \frac{1}{1}(4 - 1 \cdot 2 - 0 \cdot 0 - 3 \cdot 1) = -1. \end{aligned}$$

In order to solve a system of $n$ linear equations in $n$ variables, we can show that we need $\mathcal{O}(n^3)$ operations for Gaussian elimination of $[A \mid \mathbf{b}]$, and $\mathcal{O}(n^2)$ operations for backward/forward substitution.[24]

24: We use the "big O" notation $\mathcal{O}(n^k)$ as shorthand for a number of operations $\leq An^k$ for some constant $A > 0$.

**LU Factorization** If $A$ is invertible, then we can perform Gaussian elimination on it, which also means that it can be factored as

$$A = LU,$$

where $L$ and $U$ are lower and upper square triangular, respectively.[25] In fact, $U$ is the reduced matrix of $A$ (after Gaussian-elimination) and

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \ell_{2,1} & 1 & & 0 \\ \vdots & \ddots & \ddots & \\ \ell_{n,1} & \cdots & \ell_{n,n-1} & 1 \end{pmatrix}.$$

25: This assumes that Gaussian elimination can be conducted on $A$ without having to interchange rows, an assumption that we will make throughout this section.

**Example** In the preceding example, we had

$$A = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} \rightsquigarrow U = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & -5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}.$$

With

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix},$$

we indeed have $LU = A$. □

Let $\mathbf{I}_n$ be the $n \times n$ identity matrix, and $\mathbf{M}_n(i, j)$ be the $n \times n$ zero matrix, except in the position $(i, j)$, where the entry is 1. The three types of elementary row transformations that carry $A$ to $U$ can also be written as a left-product of elementary matrices with $A$:

$$U = E^{(n-1,1)} E^{(n-2,2)} E^{(n-2,1)} \cdots E^{(1,n-1)} \cdots E^{(1,1)} A,$$

where

$$E^{(k,\nu)} = \begin{cases} \mathbf{I_n}[R_i \leftrightarrow R_j], & \nu\text{th operation of step } k \text{ is } R_i \leftrightarrow R_j \\ \mathbf{I_n} + a\mathbf{M}_n(i, j), & \nu\text{th operation of step } k \text{ is } aR_i + R_j \to R_j,\ i > j \\ \mathbf{I_n} + (a-1)\mathbf{M}_n(j, j), & \nu\text{th operation of step } k \text{ is } aR_j \to R_j,\ a \neq 0 \end{cases}$$

Note that $E^{(k,\nu)}$ is always invertible; if no row interchange is required, then $E^{(k,\nu)}$ and $\left[E^{(k,\nu)}\right]^{-1}$ are both lower triangular.

By construction, then

$$A = \left[E^{(1,1)}\right]^{-1} \cdot \left[E^{(1,n-1)}\right]^{-1} \cdots \left[E^{(n-1,1)}\right]^{-1} U = LU,$$

where $L$ is lower diagonal with ones on the diagonal.

Once we have the $LU$ factorization of $A$, the system $A\mathbf{x} = \mathbf{b}$ can be solved using first forward, then backward substitution:

$$A\mathbf{x} = LU\mathbf{x} = L\mathbf{y} = \mathbf{b}, \quad \text{and then} \quad U\mathbf{x} = \mathbf{y}.$$

The LU factorization approach is particularly useful if we need to solve $A\mathbf{x} = \mathbf{b}$ for different $\mathbf{b}$.[26] It can also be used to speed up determinant computations, since

$$\det(A) = \det(LU) = \det(L)\det(U) = \left(\prod_{i=1}^{n} \ell_{i,i}\right)\left(\prod_{i=1}^{n} u_{i,i}\right) = \left(\prod_{i=1}^{n} u_{i,i}\right).$$

26: We only need $\mathcal{O}(n^3)$ steps for the Gaussian elimination of $A$ once, then $\mathcal{O}(n^2)$ steps for the forward and backward substitution in each system.

**Pivoting Strategies** When one of the (eventual) **pivot elements** is zero, Gaussian elimination fails because we need access to row interchanges (also known as **pivoting**).[27] But this strategy should also be used when the pivot elements are small in magnitude, relative to the other (reduced) matrix entries, because Gaussian elimination is prone to round-off error.

27: See [7, 5] for details.

**Example** In exact (symbolic) arithmetic, the matrix form of the linear system

$$\begin{cases} 10^{-20}x_1 + x_2 = 1 \\ x_1 + 2x_2 = 4 \end{cases}$$

reduces to

$$\left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 1 & 2 & 4 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 0 & 2-10^{20} & 4-10^{20} \end{array}\right),$$

*via* the row transformation $R_2 - 10^{20}R_1 \to R_2$. Using backward substitution, we then obtain

$$(2-10^{20})x_2 = 4-10^{20} \implies x_2 = \frac{4-10^{20}}{2-10^{20}},$$
$$10^{-20}x_1 = 1-x_2 \implies x_1 = 10^{20}\left(1-\frac{4-10^{20}}{2-10^{20}}\right) = -\frac{2\times 10^{20}}{2-10^{20}};$$

therefore, $x_2 \approx 1$ and $x_1 \approx 2$.

If we are using double precision,[28] we have $2-10^{20} \mapsto -10^{20}$ and $4-10^{20} \mapsto -10^{20}$, and so

$$\left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 1 & 2 & 4 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 0 & -10^{20} & -10^{20} \end{array}\right),$$

which yields $x_2 = 1$ and $x_1 = 0$. That is problematic!

28: Which is to say, $\approx 16$ significant digits.

If we exchange rows 1 and 2 ($R_1 \leftrightarrow R_2$), we obtain instead

$$\left(\begin{array}{cc|c} 1 & 2 & 4 \\ 10^{-20} & 1 & 1 \end{array}\right) \rightsquigarrow \left(\begin{array}{cc|c} 1 & 2 & 4 \\ 0 & 1-2\times 10^{-20} & 1-4\times 10^{-20} \end{array}\right) \mapsto \left(\begin{array}{cc|c} 1 & 2 & 4 \\ 0 & 1 & 1 \end{array}\right),$$

which yields $x_2 = 1$ and $x_1 = 2$.

The elementary matrices in which row interchange are encoded are not lower triangular; an **invertible** matrix $A$ whose Gaussian elimination requires such a transformation does not have an $LU$ decomposition, but it can be decomposed that way up to a **permutation matrix** $P$:[29]

$$PA = LU.$$

29: A permutation matrix is a matrix whose rows are a permutation of the rows of $\mathbf{I}_n$.

**Algorithm:** $LU$ factorization with partial pivoting

**Input:** $n \times n$ matrix $A = (a_{i,j})$
**Output:** $n \times n$ matrices $L, U, P$ such that $PA = LU$

**Initialization:** $P = \mathbf{I}_n$;
**For** $k = 1, 2, \ldots, n-1$ **do**
    Find smallest $q$ such that $|a_{q,k}| = \max_{k \le i \le n} |a_{i,k}|$;
    Exchange rows $q$ and $k$ in $A$ and $P$;
    **For** $i = k+1, \ldots, n$ **do**
        Set $a_{i,k} = a_{i,k}/a_{k,k}$;
        **For** $j = k+1, \ldots, n$ **do**
            Set $a_{i,j} = a_{i,j} - a_{i,k}a_{k,j}$;
        **End**
    **End**
**End**
$L = \mathbf{I}_n$ + strictly lower triangular$(A)$; $U$ = upper triangular$(A)$; $P$

Once we have $P, L, U$ such that $PA = LU$, then we can solve the system $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ by using

$$A\mathbf{x} = \mathbf{b} \iff PA\mathbf{x} = P\mathbf{b} \iff LU\mathbf{x} = P\mathbf{b},$$

namely, we first solve $L\mathbf{y} = P\mathbf{b}$ using forward substitution, then we solve $U\mathbf{x} = \mathbf{y}$ using backward substitution.

**Example** Algorithm 6 is implemented in `R` *via* the `Matrix` package's function `lu()`. We use it to find the partial pivoting $LU$ decomposition of

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 4 & 6 \end{pmatrix}.$$

We start by loading the matrix.

```
require(Matrix)
A=t(matrix(c(1,2,3,2,4,5,3,4,6),3,3))
```

We can decompose and extract the factors of the $LU$ decomposition as follows:

```
D <- lu(A)
expand(D)$L
```

```
3 x 3 Matrix of class "dtrMatrix" (unitriangular)
     [,1]      [,2]      [,3]
[1,] 1.0000000         .         .
[2,] 0.6666667 1.0000000         .
[3,] 0.3333333 0.5000000 1.0000000
```

```
expand(D)$U
```

```
3 x 3 Matrix of class "dtrMatrix"
     [,1]     [,2]     [,3]
[1,] 3.000000 4.000000 6.000000
[2,]        . 1.333333 1.000000
[3,]        .        . 0.500000
```

```
expand(D)$P
```

```
3 x 3 sparse Matrix of class "pMatrix"

[1,] . . |
[2,] . | .
[3,] | . .
```

Other (mostly similar) factorizations may be better suited to various types of matrices $A$:

- for **symmetric** matrices $A$,[30] we use $A = LDL^\top$, where $D$ is a diagonal matrix;
- for **symmetric positive definite** matrices $A$,[31] we use the **Cholesky decomposition** $A = MM^\top$;
- it may be possible to take advantage of some **sparse matrices**' structure (such as is the case for **banded matrices**) to greatly increase the speed of the LU decomposition with partial pivoting.

30: $A^\top = A$

31: $A$ symmetric and $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

**Matrix Norms** A **vector norm** $\|-\| : \mathbb{R}^n \to \mathbb{R}_0^+$ is a function satisfying the following three conditions:

1. $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$;
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$;
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

The 2−norm $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$ is a common example.[32]

32: We can show that all vector norms on $\mathbb{R}^n$ are equivalent, suggesting that there is no real advantage to selecting one over another, in a general setting (although there may be instances where calculations are simpler in one context over another).

Given a vector norm $\|\cdot\|$ on $\mathbb{R}^n$, we can define the **induced matrix norm** $|||-|||$ on the space of $n \times n$ matrices by

$$|||A||| = \sup_{\mathbf{x} \neq \mathbf{0}} \left\{ \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \right\},$$

where $\mathbf{x}$ ranges over $\mathbb{R}^n$.[33]

33: The properties of vector norms also apply to matrix norms – matrices are the vectors of the space of square matrices, with matrix addition and multiplication by a scalar.

That $|||A||| \geq 0$ is a direct consequence of the definition of the supremum and because $\|A\mathbf{x}\|, \|\mathbf{x}\| \geq 0$ for all $\mathbf{x}$.

If $|||A||| = 0$, then $\|A\mathbf{x}\| \leq 0$ for all $\mathbf{x} \neq \mathbf{0}$; since $\|A\mathbf{x}\| \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$, then $\|A\mathbf{x}\| = 0$ for all $\mathbf{x}$. As $\|A\mathbf{0}\| = 0$, then $\|A\mathbf{x}\| = 0$ for all $\mathbf{x}$. In particular, $\|A\mathbf{e}_k\| = \|A_k\| = 0$ for $1 \leq k \leq n$, so that every column $A_k = \mathbf{0}$; hence $A = \mathbf{O}_{n\times n}$. Conversely, if $A = \mathbf{O}_{n\times n}$, then $\|A\mathbf{x}\| = \|\mathbf{0}\| = 0$ for all $\mathbf{x}$, so that $\|A\mathbf{x}\|/\|\mathbf{x}\| = 0$ for all $\mathbf{x} \neq 0$; hence $|||A||| \leq 0$. Since $|||A||| \geq 0$, we must have $|||A||| = 0$.[34]

34: Properties 2 and 3 are left as exercises.

**Theorem:** let $|||\cdot|||$ be the matrix norm induced by a vector norm $\|\cdot\|$. Then:

1. $\|A\mathbf{x}\| \leq |||A||| \cdot \|\mathbf{x}\|$ for all $A$, $\mathbf{x}$;
2. $|||\mathbf{I}_n||| = 1$;
3. $|||AB||| \leq |||A||| \cdot |||B|||$ for all $A$, $B$.

**Proof:** throughout, let $A$, $B$ be generic $n \times n$ matrices, and $\mathbf{x} \in \mathbb{R}^n$.

1. If $\mathbf{x} = \mathbf{0}$, then the property holds as both sides are 0. Now assume that $\mathbf{x} \neq \mathbf{0}$. By definition,

$$|||A||| \geq \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \quad \text{for all } \mathbf{x} \neq \mathbf{0};$$

   thus $|||A||| \cdot \|\mathbf{x}\| \geq \|A\mathbf{x}\|$ for all $\mathbf{x} \neq \mathbf{0}$.
2. For any $\mathbf{x} \neq \mathbf{0}$, we have $\|\mathbf{I}_n\mathbf{x}\|/\|\mathbf{x}\| = 1$, so

$$|||\mathbf{I}_n||| = \sup_{\mathbf{x}\neq\mathbf{0}} \left\{ \frac{\|\mathbf{I}_n\mathbf{x}\|}{\|\mathbf{x}\|} \right\} = \sup_{\mathbf{x}\neq\mathbf{0}} \{1\} = 1.$$

3. For any $\mathbf{x} \neq \mathbf{0}$, we see that

$$\|AB\mathbf{x}\| \leq |||A||| \cdot \|B\mathbf{x}\| \leq |||A||| \cdot |||B||| \cdot \|\mathbf{x}\|;$$

   hence

$$|||AB||| = \sup_{\mathbf{x}\neq\mathbf{0}} \left\{ \frac{\|AB\mathbf{x}\|}{\|\mathbf{x}\|} \right\} \leq |||A||| \cdot |||B|||,$$

   which completes the proof.[35] ■

35: See [2] for more information on the supremum.

The $\ell_p$ vector norm $\|\cdot\|_p$ on $\mathbb{R}^n$ is trivial to compute: for $p \geq 1$, we have

$$\|\mathbf{x}\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p};$$

for $p = \infty$ we have

$$\|\mathbf{x}\|_\infty = \max_{1\leq k\leq n} |x_k|.$$

It is not as clear how we would compute the corresponding induced matrix norm; we can show that

$$|||A|||_1 = \max_{1\leq j\leq n} \left\{ \sum_{i=1}^{n} |a_{i,j}| \right\};$$

$$|||A|||_\infty = \max_{1\leq i\leq n} \left\{ \sum_{j=1}^{n} |a_{i,j}| \right\};$$

$$|||A|||_2 = \sqrt{\lambda_{\max}(A^\top A)}, \quad \lambda_{\max}(B): \text{ largest eigenvalue of } B.$$

Let $|||\cdot|||$ be an induced matrix norm. The **condition number** of an invertible matrix $A$ under that norm is

$$\kappa(A) = |||A||| \cdot |||A^{-1}|||.$$

Because $AA^{-1} = \mathbf{I}_n$, we have

$$1 = |||\mathbf{I}_n||| = |||AA^{-1}||| \leq |||A||| \cdot |||A^{-1}||| = \kappa(A).$$

When $\kappa(A) \gg 1$, we say that $A$ is **ill-conditioned** under $|||\cdot|||$.

**Estimating Error** In this section, we estimate the **relative error** between the exact solution of $A\mathbf{x} = \mathbf{b}$ and an approximate solution $\hat{\mathbf{x}}$.[36]

36: This estimation is useful not only to quantify the effect of round-off error in **direct methods**, but also to analyze stopping criteria for **iterative methods**.

**Theorem:** let $A$ be an invertible $n \times n$ matrix, and let $\mathbf{0} \neq \mathbf{b} \in \mathbb{R}^n$. Let $\mathbf{x} \in \mathbb{R}^n$ be the exact solution to the system $A\mathbf{x} = \mathbf{b}$. Consider a vector norm $\|\cdot\|$ on $\mathbb{R}^n$ and its induced matrix norm $|||\cdot|||$. For any $\hat{\mathbf{x}} \in \mathbb{R}^n$, we have

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa(A) \cdot \frac{\|\mathbf{b} - A\hat{\mathbf{x}}\|}{\|\mathbf{b}\|}.$$

**Proof:** write

$$\mathbf{b} - A\hat{\mathbf{x}} = A\mathbf{x} - A\hat{\mathbf{x}} = A(\mathbf{x} - \hat{\mathbf{x}}) \implies \mathbf{x} - \hat{\mathbf{x}} = A^{-1}(\mathbf{b} - A\hat{\mathbf{x}});$$

hence

$$\|\mathbf{x} - \hat{\mathbf{x}}\| = \|A^{-1}(\mathbf{b} - A\hat{\mathbf{x}})\| \leq |||A^{-1}||| \cdot \|\mathbf{b} - A\hat{\mathbf{x}}\|.$$

We also have

$$\|\mathbf{b}\| = \|A\mathbf{x}\| \leq |||A||| \cdot \|\mathbf{x}\| \implies \frac{1}{\|\mathbf{x}\|} \leq \frac{1}{\|\mathbf{b}\|}|||A|||.$$

Combining both of these inequalities yields

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq |||A||| \cdot |||A^{-1}||| \cdot \frac{\|\mathbf{b} - A\hat{\mathbf{x}}\|}{\|\mathbf{b}\|};$$

as $\kappa(A) = |||A||| \cdot |||A^{-1}|||$, the proof is complete. ■.

In practice, due to floating point representation, we never really solve the system $A\mathbf{x} = \mathbf{b} \neq \mathbf{0}$;[37] instead, we solve the **perturbed system**

37: If $\mathbf{b} = \mathbf{0}$, the homogeneous system has the exact solution $\mathbf{x} = \mathbf{0}$ and no additional work is needed.

$$(A + \delta A)\hat{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b},$$

where the entries of the $n \times n$ matrix $\delta A$ and $\delta\mathbf{b} \in \mathbb{R}^n$ are of the order of $10^{-16}$ those of $A$ and $\mathbf{b}$, respectively.[38]

38: Assuming double precision.

Let $\mathbf{x} \in \mathbb{R}^n$ be the exact solution of the unperturbed system and $\hat{\mathbf{x}}$ that of the perturbed system. Then

$$\mathbf{b} - A\hat{\mathbf{x}} = \mathbf{b} - (\mathbf{b} + \delta\mathbf{b} - \delta A\hat{\mathbf{x}}) = \delta A\hat{\mathbf{x}} - \delta\mathbf{b},$$

and we deduce from the previous theorem that

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa(A)\frac{\|\mathbf{b} - A\hat{\mathbf{x}}\|}{\|\mathbf{b}\|} \leq \kappa(A) \cdot \frac{\|\delta A\|\|\hat{\mathbf{x}}\| + \|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

If $\|\delta A\| \leq \frac{1}{\|A^{-1}\|}$, then we can re-arrange the last equation and write

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \cdot \frac{\|\delta A\|}{\|A\|}} \cdot \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right).$$

**Example** If the perturbation $\delta A$ is $\mathbf{O}_{n\times n}$, then

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa(A) \cdot \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

For instance, consider the exact and perturbed systems

$$\begin{pmatrix} 1 & 10^{-16} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 10^{-16} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + 10^{-16} \\ 1 \end{pmatrix}.$$

The exact solution of $A\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = (1, 0)^\top$, that of $A\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$ is $(1, 1)^\top$: a tiny perturbation $\delta\mathbf{b}$ has a gigantic effect on the solution. This is due to the fact that $A$ is ill-conditioned. Indeed,

$$A^{-1} = \begin{pmatrix} 0 & 1 \\ 10^{16} & -10^{16} \end{pmatrix}, \quad |||A|||_1 = 2, \quad |||A^{-1}|||_1 = 1 + 10^{16},$$

and so $\kappa_1(A) = 2 + 2 \times 10^{16} \gg 1$.

Since the perturbation $\delta A$ is $\mathbf{O}_{2\times 2}$, we would expect, in the $\ell_1$ vector norm and associated induced matrix norm, to find:

$$\begin{aligned} \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_1}{\|\mathbf{x}\|_1} &= \frac{\|(0, -1)^\top\|}{\|(1, 0)^\top\|} = 1 \leq \kappa_1(A) \cdot \frac{\|\delta\mathbf{b}\|_1}{\|\mathbf{b}\|_1} \\ &= (2 + 2 \times 10^{16}) \cdot \frac{\|(10^{-16}, 0)^\top\|_1}{\|(1, 1)^\top\|_1} = (2 + 2 \times 10^{16}) \cdot \frac{10^{-16}}{2} = 1 + 10^{-16}, \end{aligned}$$

which is indeed the case. □

### Iterative Methods

We can get exact solutions from direct methods, but the process is **computationally expensive** and storage can be prohibitive, especially for **large dense** matrices. In this section, we consider **iterative methods**, which operate in the same spirit as **fixed point iteration**.[39]

39: We will discuss other iterative methods, such as **gradient descent** and its variants in Chapter 31. Other modern approaches include the **generalized minimal residual** and **biconjugate gradient** method, among others.

The problem of solving $A\mathbf{x} = \mathbf{b}$ is equivalent to the problem of solving

$$f(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = \mathbf{0}.$$

We re-write this problem into an equivalent problem

$$\mathbf{x} = g(\mathbf{x}) = T\mathbf{x} + \mathbf{c};$$

given an initial guess $\mathbf{x}_0$, we then compute the iterative sequence

$$\mathbf{x}^{(k+1)} = g(\mathbf{x}^{(k)}) = T\mathbf{x}^{(k)} + \mathbf{c}, \quad k = 0, 1, \ldots$$

The hope is that the sequence converges to the solution $\mathbf{x}^*$ of $A\mathbf{x} = \mathbf{b}$.

**Stationary Iteration** As was the case for functions of one variables, we can come up with multiple formulations for the **fixed point system**.

One general technique is based on a **splitting** of $A$: for an invertible matrix $P$ (the **pre-conditioner**), we can write $A = P - (P - A)$:

$$\begin{aligned} A\mathbf{x} = \mathbf{b} &\iff [P - (P - A)]\mathbf{x} = \mathbf{b} \iff P\mathbf{x} = (P - A)\mathbf{x} + \mathbf{b} \\ &\iff \mathbf{x} = P^{-1}(P - A)\mathbf{x} + P^{-1}\mathbf{b} \iff \mathbf{x} = T\mathbf{x} + \mathbf{c}. \end{aligned}$$

The iterative method obtained with this splitting can be written as

$$P\mathbf{x}^{(k+1)} = (P - A)\mathbf{x}^{(k)} + \mathbf{b},$$

or equivalently, upon setting the **residual** $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$ **at step** $k$:

$$\begin{aligned} P\delta\mathbf{x}^{(k+1)} &= \mathbf{r}^{(k)}, \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k+1)}, \quad k = 0, 1, \ldots \end{aligned}$$

This approach is useful when $P\delta\mathbf{x}^{(k+1)} = \mathbf{r}^{(k)}$ is "**much simpler**" to solve than the original system $A\mathbf{x} = \mathbf{b}$, however. This is the case when $P$ is diagonal (**Jacobi**) or triangular (**Gauss-Seidel**).[40]

40: In both cases, we assume that the diagonal entries of $A$ are non-zero, i.e. $a_{i,i} \neq 0$ for $1 \leq i \leq n$.

**Jacobi Method** In this approach, we use

$$P = \begin{pmatrix} a_{1,1} & & \\ & \ddots & \\ & & a_{n,n} \end{pmatrix} \quad \text{and} \quad P - A = -\begin{pmatrix} 0 & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ a_{n,1} & \cdots & a_{n,n-1} & 0 \end{pmatrix}.$$

The iterative procedure $P\mathbf{x}^{(k+1)} = (P - A)\mathbf{x}^{(k)} + \mathbf{b}$ then reduces to a linear system in which the components of $\mathbf{x}^{(k+1)}$ only depend on the components of $\mathbf{x}^{(k)}$:[41]

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i + a_{i,i}x_i^{(k)} - \sum_{j=1}^{n} a_{i,j}x_j^{(k)}\right), \quad i = 1, \ldots, n.$$

41: They can be computed in **parallel**, which is a non-negligible time saver.

**Gauss-Seidel Method** In this approach, we use

$$P = \begin{pmatrix} a_{1,1} & & \\ \vdots & \ddots & \\ a_{n,1} & \cdots & a_{n,n} \end{pmatrix} \quad \text{and} \quad P - A = -\begin{pmatrix} 0 & a_{1,2} & \cdots & a_{1,n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{pmatrix}.$$

The iterative procedure $P\mathbf{x}^{(k+1)} = (P - A)\mathbf{x}^{(k)} + \mathbf{b}$ then reduces to a linear system which can be solved by forward substitution:

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}\right), \quad i = 1, \ldots, n.$$

**Example** Consider the system $A\mathbf{x} = \mathbf{b}$ with

$$A = \begin{pmatrix} 3 & -1 & 1 \\ 3 & 6 & 2 \\ 3 & 3 & 7 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix}.$$

We use $\mathbf{x}_0 = (1, 1, 1)^\top$ to compute the first iterate for both the Jacobi and the Gauss-Seidel methods.

In the **Jacobi** method, the first iterate $\mathbf{x}^{(1)}$ solves

$$\begin{pmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 7 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \\ -3 & 0 & -2 \\ -3 & -3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ -5 \\ -2 \end{pmatrix},$$

which we can solve directly by substitution:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 1/3 \\ -5/6 \\ -2/7 \end{pmatrix}.$$

In the **Gauss-Seidel** method, the first iterate $\mathbf{x}^{(1)}$ solves

$$\begin{pmatrix} 3 & 0 & 0 \\ 3 & 6 & 0 \\ 3 & 3 & 7 \end{pmatrix} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix},$$

which we solve by forward substitution:

$$\begin{aligned} x_1^{(1)} &= 1/3 \\ 3x_1^{(1)} + 6x_2^{(1)} = -2 &\implies x_2^{(1)} = -1/2 \\ 3x_1^{(1)} + 3x_2^{(1)} + 7x_3^{(1)} = 4 &\implies x_3^{(1)} = 9/14. \end{aligned}$$

**Convergence and Stopping Criterion** We know how to compute iterates in the Jacobi and Gauss-Seidel framework, and, more generally, for an **iteration matrix**

$$T = P^{-1}(P - A).$$

How can we tell if the iteration procedure converges, and if it does, whether it converges to the system's unique solution $\mathbf{x}^*$?

The **error** $\mathbf{e}^{(k+1)}$ **at step** $k+1$ is defined by

$$\mathbf{e}^{(k+1)} = \mathbf{x}^* - \mathbf{x}^{(k+1)}.$$

Recall that $\mathbf{x}^* = T\mathbf{x}^* + \mathbf{c}$. Then

$$\mathbf{e}^{(k+1)} = \mathbf{x}^* - \mathbf{x}^{(k+1)} = T\mathbf{x}^* + \mathbf{c} - T\mathbf{x}^{(k)} - \mathbf{c} = T\left(\mathbf{x}^* - \mathbf{x}^{(k)}\right) = T\mathbf{e}^{(k)}.$$

Thus, for any vector norm $\|\cdot\|$ and induced matrix norm $|||\cdot|||$, we have

$$\|\mathbf{e}^{(k+1)}\| = \|T\mathbf{e}^{(k)}\| \le |||T||| \cdot \|\mathbf{e}^{(k)}\| \le |||T|||^2 \cdot \|\mathbf{e}^{(k-1)}\| \le \cdots \le |||T|||^{k+1} \cdot \|\mathbf{e}^{(0)}\|,$$

and so

$$\lim_{k\to\infty} \|\mathbf{e}^{(k+1)}\| = 0, \quad \text{when } |||T||| < 1.$$

**Theorem:** if $|||T||| < 1$ for an induced matrix norm, then for any $\mathbf{x}^{(0)}$, the sequence $\{\mathbf{x}^{(k)}\}$ converges to the solution of $\mathbf{x}^*$ of $A\mathbf{x} = \mathbf{b}$. Moreover,

$$\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \le |||T|||^k \cdot \|\mathbf{x}^* - \mathbf{x}^{(0)}\|;$$

the smaller $|||T|||$ is, the faster the convergence to $\mathbf{x}^*$.

At what point in the iteration should we stop? Given a prescribed tolerance tol $> 0$, the goal is to stop as soon as

$$\|\mathbf{e}^{(k)}\| = \|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \text{tol} \quad \text{or} \quad \frac{\|\mathbf{x}^* - \mathbf{x}^{(k)}\|}{\|\mathbf{x}^*\|} \leq \text{tol},$$

the latter assuming $\mathbf{b} \neq \mathbf{0}$. Since the error cannot be computed in practice, as it involves the exact solution $\mathbf{x}^*$, we need to use an **error estimate**.

One possibility is to use the **normalized residual** and stop as soon as

$$\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{b}\|} = \frac{\|\mathbf{b} - A\mathbf{x}^{(k)}\|}{\|\mathbf{b}\|} \leq \text{tol}.$$

From a previous theorem, we have

$$\frac{\|\mathbf{x}^* - \mathbf{x}^{(k)}\|}{\|\mathbf{x}^*\|} \leq \kappa(A) \cdot \frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{b}\|} \leq \kappa(A) \cdot \text{tol};$$

when $\kappa(A)$ is reasonably small, the normalized residual is suitable to use in the stopping criterion.

Another possibility is to use the **increment between two iterates**, and stop as soon as

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \text{tol}.$$

In this case, since

$$\begin{aligned}\|\mathbf{x}^* - \mathbf{x}^{(k)}\| &\leq |||T||| \cdot \|\mathbf{x}^* - \mathbf{x}^{(k-1)}\| = |||T||| \cdot \|\mathbf{x}^* - \mathbf{x}^{(k)} + \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \\ &\leq |||T||| \left[\|\mathbf{x}^* - \mathbf{x}^{(k)}\| + \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|\right].\end{aligned}$$

Thus, provided $|||T||| < 1$, we have

$$\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \frac{|||T|||}{1 - |||T|||} \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \frac{|||T|||}{1 - |||T|||} \cdot \text{tol}.$$

The incremental stopping criterion would thus be a good choice if $|||T|||$ is not too close to 1.

**Implementation** The Jacobi (`J`) and Gauss-Seidel (`GS`) methods are implemented in `R` as follows.

**Iterative solver**

```
iterative_solver <- function(A, b, x0, nmax, tol, method){
  # Check for valid method
  if(!(method %in% c('J', 'GS'))) {
    stop("Unknown method...")
  }

  # Construct preconditioner matrix based on the method
  if(method == 'J') {
    P <- diag(diag(A))
  } else if(method == 'GS') {
    P <- matrix(0, ncol=ncol(A), nrow=nrow(A))
```

```
    for(i in 1:nrow(A)) {
      for(j in 1:ncol(A)) {
        if(i >= j) {
          P[i, j] <- A[i, j]
        }
      }
    }
  }

  # initialization
  b_norm <- norm(b, type="2")
  if(b_norm == 0) {
    b_norm <- 1
  }

  x <- x0
  r <- b - A %*% x
  r_norm <- norm(r, type="2")
  iter <- 0

  # Iteration
  while((r_norm/b_norm > tol) && (iter < nmax)){
    incr <- solve(P, r)
    x <- x + incr
    r <- b - A %*% x
    iter <- iter + 1
    r_norm <- norm(r, type="2")
  }

  return(list(x=x, iter=iter))
}
```

**Example** The `pracma` library is required to access `norm()`.

```
library(pracma)
```

For instance, we can solve the $4 \times 4$ system

$$\begin{pmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 15 \\ 10 \\ 10 \\ 10 \end{pmatrix}$$

using the Gauss-Seidel method and the normalized residual stopping criterion, with a tolerance of $10^{-5}$ and $\mathbf{x}_0 = (0, 0, 0, 0)^\top$.

```
A <- matrix(c(4,-1,0,0,-1,4,-1,0, 0,-1,4,-1, 0,0,-1,3), 4, 4)
b <- c(15,10,10,10); x0 <- c(0,0,0,0)
nmax <- 100; tol <- 1e-5; method <- "GS"
result <- iterative_solver(A, b, x0, nmax, tol, method)
```

We can see the solution and number of iterations by calling the two

returned items.

```
result$x
```

```
        [,1]
[1,] 4.999974
[2,] 4.999982
[3,] 4.999991
[4,] 4.999997
```

```
result$iter
```

```
[1] 8
```

This compares very well to the exact solution $\mathbf{x}^* = (5, 5, 5, 5)^\top$.

### 4.3.2 Non-Linear Systems

The direct method does not generalize to non-linear systems of equations, but the fundamental concept of iterative methods does.[42]

42: We will have more to say on the topic in Chapter 31.

**Fixed Point Iteration** The ideas of Section 4.2.3 still apply, but they need to be modified somewhat to generalize to non-linear systems of $n$ equations in $n$ unknowns.

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a sufficiently differentiable function. We are looking for points $\mathbf{x}^* \in \mathbb{R}^n$ that solve $f(\mathbf{x}) = \mathbf{0}$. In the general case, the system could admit any finite number of solution,[43] an infinite countable set of solutions,[44] or an uncountable set of solutions.[45] There is no simple criterion to determine in which class a given system falls.

43: Not necessarily only 0 or 1.

44: Such as for $\sin x = 0$ over $\mathbb{R}$.

45: Such as for $A\mathbf{x} = \mathbb{0}$ when $A$ is not of full rank.

**Example** The system

$$f(\mathbf{x}) = \begin{pmatrix} x_1^3 + 2x_1x_2 \\ x_2 + 2x_1^2x_2 \end{pmatrix} = \begin{pmatrix} 8 \\ 13 \end{pmatrix}$$

is equivalent to

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} (8 - x_1^3)/2x_2 \\ 13 - 2x_1^2x_2 \end{pmatrix} = g(\mathbf{x}).$$

Note that there may be multiple ways to transform the system $f(\mathbf{x}) = \mathbf{0}$ into a fixed point problem $g(\mathbf{x}) = \mathbf{x}$, with $g(D) \subseteq D$.

**General Fixed Point Theorem:** let $g : D \to D$, with $D$ a closed subset of $\mathbb{R}^n$, and $\|\cdot\|$ a vector norm on $\mathbb{R}^n$. If $\exists L < 1$ such that

$$\|g(\mathbf{x}) - g(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in D$, then $g$ admits a unique fixed point $\mathbf{x}^* \in D$ and the sequence $\mathbf{x}^{(k+1)} = g(\mathbf{x}^{(k)})$ converges to $\mathbf{x}^*$ for all $\mathbf{x}^{(0)} \in D$.

When $g$ meets the condition stated in the theorem, we say it is **contractive on** $D$; it is not easy to show directly that this property holds. There is a sufficient condition on the **Jacobian matrix** of $g$ at $\mathbf{x}^*$ (see Chapter 2 and the next section on Newton's method) that guarantees that $g$ is contractive in a neighbourhood of $\mathbf{x}^*$:

$$|||Dg(\mathbf{x}^*)||| < 1,$$

assuming that $g$ is at least $C^1$. In that case, the convergence of the fixed point iterates to $\mathbf{x}^*$ is at least of order 1 (linear).

**Newton's Method** In Section 4.2.5, we saw that there was a way to avoid directly evaluating the derivative $f'$ in Newton's Method (which can be costly) by using the **secant approximation**.

This is a reasonable approach for equations in one variable, but it is less obvious how we would do so in a multi-dimensional case – this is where the work we put on linear systems will pay off.

In order to apply Newton's method to the system

$$f(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix} = \mathbf{0},$$

where $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is at least $C^1$, we need to generalize the iterates $x_k$, the function values $f(x_k)$, and the derivative $f'(x_k)$ to the multi-dimensional case.

The natural way to do this is as follows:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \implies \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - Df(\mathbf{x}^{(k)})^{-1} f(\mathbf{x}^k),$$

for $k \geq 0$, where

$$Df(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

In practice it can be quite costly to invert the matrix not only once, but at every step of the iterative process. We can save time (and increase numerical stability) by re-writing the iteration step as a system of linear equations:

$$Df(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = f(\mathbf{x}^k), \quad \text{for } k \geq 0,$$

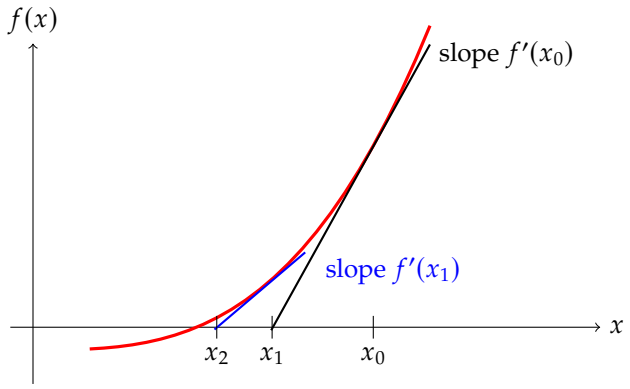
which can be solved using the methods of Section 4.3.1.

Under some regularity conditions on $f$, the sequence $\{\mathbf{x}^{(k)}\}$ converges quadratically to a solution $\mathbf{x}^*$ of $f(\mathbf{x}^*) = \mathbf{0}$.

Potential problems include the poor choice of the starting point $\mathbf{x}_0$ (at a critical point of $f$, $\mathbf{x}_0$ entering a cycle); $f$ not being sufficiently differentiable in a neighbourhood of $\mathbf{x}^*$; $\mathbf{x}^*$ not existing; the derivative of $f$ not being continuous at $\mathbf{x}^*$, etc.

## 4.4 Exercises

1. How must the Golden Ratio method be modified if we are looking for the maximal value of a unimodal continuous function $f$ on $[a, b]$?
2. Is it necessary to use a factor $\varphi$ in the Golden Ratio method or would any other constant $> 1$ do the trick?
3. Implement the secant method in R. Test it on this chapter's example functions.
4. Consider the function defined by $f(x) = x^2 - 2$, which has one positive root $x^* = \sqrt{2}$.




   a) Illustrate Newton's method by performing two steps starting at $x_0$.
   b) Let $\{x_k\}_{k\geq 0}$ be the sequence generated by Newton's method. Write the relationship between $x_{k+1}$ and $x_k$. Then, compute $x_1$ and $x_2$ starting from $x_0 = 2$.
   c) Determine the (exact) order of Newton's method assuming that we start close enough to $x^* = \sqrt{2}$.

5. Let $f(x) = (x+2)(x+1)^2x(x-1)^3(x-2)$. To which zero of $f$ does the bisection method converge when applied on the following intervals?

   a) $[-1.5, 2.5]$ b) $[-0.5, 2.4]$ c) $[-0.5, 3]$ d) $[-3, -0.5]$.

6. Use the bisection method on $[1, 2]$ to find an approximation of $\sqrt{3}$ correct to within $10^{-4}$. Indicate which function $f$ you used and report the values of $x_0$, $x_1$ and $x_2$, the final output and the number of iterations.
7. Let $f(x) = x^2 - 2x - 3$. To find a root of $f$, the following three fixed point method are proposed

   a) $x_k = \dfrac{3}{x_{k-1} - 2}$ b) $x_k = x_{k-1}^2 - x_{k-1} - 3$ c) $x_k = \dfrac{x_{k-1}^2 + 3}{2x_{k-1} - 2}$.

   For each method, compute (if possible) the iterates $x_1$, $x_2$, $x_3$ and $x_4$ starting from $x_0 = 0$. Report the values you obtain in a table. Which methods seem to be appropriate? Among those, which one seems to converge the fastest?
8. Consider the function $g(x) = \frac{1}{3}\sqrt[3]{x+8}$.
   a) Show that $g$ has a unique fixed point in $[0, 1]$.

b) Assuming that we start from $x_0 = \frac{1}{2}$, find a bound for the number of fixed point iterations needed to achieve $10^{-6}$ accuracy.

9. Use Newton's method and the secant method with stopping criterion $|x_{k+1} - x_k| \leq 10^{-5}$ to find solutions for the following problems. For Newton's method, use the midpoint of the given interval for $x_0$ while for the secant method, use the endpoints of the given interval for $x_0$ and $x_1$.
   a) $3x - e^x = 0$ for $1 \leq x \leq 2$;
   b) $2x + 5\cos(x) - e^x = 0$ for $-5 \leq x \leq 0$.
10. Recall that a sequence $\{x_k\}$ that converges to some $x^*$ is said to converge with order $\alpha$ and asymptotic error constant $\lambda$ if
$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^\alpha} = \lambda,$$
where we need $\lambda < 1$ if $\alpha = 1$.
   a) Consider the function $f(x) = 1/x - 1/3, x > 0$, which vanishes at $x^* = 3$. Use Newton's method with stopping criterion $|x_{k+1} - x_k| \leq 10^{-4}$ and $x_0 = 1$ to approximate $x^*$. Determine (numerically) the order $\alpha$ and the asymptotic error constant $\lambda$.
   b) Use the secant method to approximate the root of $f$ defined in a) using $x_0 = 0.5$ and $x_1 = 1.5$. Verify that the order of the method is the golden ratio $\alpha = (1 + \sqrt{5})/2$ and determine the value of $\lambda$.
11. Suppose that $x^*$ is a zero of multiplicity $m \geq 1$ of a function $f$ of class $C^m$, namely
$$f(x^*) = f'(x^*) = f''(x^*) = \ldots = f^{(m-1)}(x^*) = 0 \quad \text{and} \quad f^{(m)}(x^*) \neq 0.$$
   a) Show that Newton's method
$$x_{k+1} = g_1(x_k), \quad k \geq 0, \quad \text{where} \quad g_1(x) = x - \frac{f(x)}{f'(x)},$$
converges only linearly (i.e., with order 1) if $m > 1$.
   b) Consider now the *modified Newton's method*
$$x_{k+1} = g_2(x_k), \quad k \geq 0, \quad \text{where} \quad g_2(x) = x - m\frac{f(x)}{f'(x)}.$$
Show that this method converges at least quadratically (i.e., with order $\geq 2$) for any $m$.

   *Hint: Write $f$ as*
$$f(x) = (x - x^*)^m h(x)$$
*for some (unknown) function $h$ with $h(x^*) \neq 0$ and show that $g_1'(x^*) = 1 - 1/m$ and $g_2'(x^*) = 0$.*
12. Show that the induced matrix norm is indeed a norm (property 1 has already been proved; finish the job with properties 2 and 3).
13. Consider the $n \times n$ matrix $A$ consisting of 1's on the diagonal and in the first column (every other entry being a 0). Compute $\kappa_1(A), \kappa_\infty(A)$, and $\kappa_2(A)$.

14. If $A$ is a $n \times n$ symmetric positive definitive matrix, show that

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

15. Solve the linear system

$$\begin{cases} x_1 - x_2 + 3x_3 = -2 \\ x_1 + x_2 = 5 \\ 3x_1 - 2x_2 + x_3 = 4. \end{cases}$$

using Gaussian elimination in its simplest form (i.e., without pivoting) and backward substitution.

16. Let

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -2 & 2 \\ 2 & 12 & -2 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 6 \\ -10 \end{pmatrix}.$$

a) Compute the $LU$ factorization of $A$, i.e., find a lower triangular matrix $L$ (with ones on the diagonal) and an upper triangular matrix $U$ such that $A = LU$.
b) Solve the system $A\mathbf{x} = \mathbf{b}$ using only forward and backward substitution.

17. Let

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 5 & 3 \\ 4 & 6 & 8 & 0 \\ 3 & 3 & 9 & 8 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -2 \\ 2 \end{pmatrix}.$$

a) Find a lower triangular matrix $L$ (with ones on the diagonal), an upper triangular matrix $U$ and a permutation matrix $P$ such that $PA = LU$.
b) Solve the system $A\mathbf{x} = \mathbf{b}$ using the factorization found in a).
c) Compute the determinant of $A$ using the factorization found in a).

18. a) Prove that for any $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ we have

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2, \quad \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty \quad \text{and} \quad \|\mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_\infty \|\mathbf{x}\|_1. \tag{4.1}$$

b) For each inequality in (4.1), find a vector $\mathbf{x}$ for which equality is attained.
c) Prove that for any matrix $A \in \mathbb{R}^{n\times n}$

$$\frac{1}{\sqrt{n}}|||A|||_\infty \leq |||A|||_2 \leq \sqrt{n}|||A|||_\infty.$$

19. Let

$$\mathbf{x} = (1, -3, 2, -1)^T \quad \text{and} \quad A = \begin{pmatrix} -1 & -1 \\ 2 & -2 \end{pmatrix}.$$

a) Compute $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$.
b) Compute $|||A|||_1$, $|||A|||_2$ and $|||A|||_\infty$.
c) Compute $\kappa_1(A)$, $\kappa_2(A)$ and $\kappa_\infty(A)$.

20. We say that an $n \times n$ matrix $A$ is strictly diagonally dominant by

row if

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad \text{for } i = 1, 2, \ldots, n. \tag{4.2}$$

Prove that if $A$ satisfies (4.2) then the Jacobi method applied to $A\mathbf{x} = \mathbf{b}$, $\mathbf{b} \in \mathbb{R}^n$, converges. *Hint: show that* $\|T_J\|_\infty < 1$, *where* $T_J$ *is the iteration matrix for the Jacobi method.*

21. Consider the system $A\mathbf{x} = \mathbf{b}$ with

$$A = \begin{pmatrix} 2 & -1 & 2 \\ -1 & 1 & 0 \\ 0 & 1 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} -6 \\ 2 \\ -3 \end{pmatrix}.$$

22. Using $\mathbf{x}^{(0)} = (0, 0, 0)^T$ as initial guess:
    a) find (by hand) the first 2 iterations of the Jacobi method;
    b) find (by hand) the first iteration of the Gauss-Seidel method.
23. We consider the Gauss-Seidel method for solving the linear system $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{pmatrix} 1 & \alpha \\ -2 & 1 \end{pmatrix}.$$

24. Determine for which values of $\alpha \in \mathbb{R}$ the method converges for any initial guess $\mathbf{x}^{(0)} \in \mathbb{R}^2$ and any right-hand side $\mathbf{b} \in \mathbb{R}^2$.
25. Implement the fixed point algorithm for systems in `R` and solve the system in Section 4.3.2. Is the function $g$ contractive on some neighbourhood $D$ of the fixed point?
26. Implement Newton's algorithm for systems in `R` and solve the system in Section 4.3.2. What is its Jacobian?

## Chapter References


[1] U.M. Ascher and C. Greif. *A First Course in Numerical Methods*. SIAM, 2011.

[2] P. Boily. *Analysis and Topology Study Aids*. Data Action Lab.

[3] B. Dionne. *Numerical Analysis*. uOttawa, 2023.

[4] B. Holland. 'Human Computers: The Women of NASA'. In: *History* (May 2003).

[5] W.K. Nicholson. *Linear Algebra with Applications, 3rd Edition*. PWS Publishing Company, 1994.

[6] J. Smékal et al. *Data Science in Physics*. Physics and Data Science.

[7] G. Strang. *Introduction to Linear Algebra*. Wellesley, 2016.

# A Survey of Optimization

# 5

by **Patrick Boily** and **Kevin Cheung**

---

Traditionally, optimization has been one of the most-frequently used arrows in the operations researcher's and quantitative analyst's quiver.

From its humble beginning as an offshoot of calculus (see Chapter 2) to its current status as the crown jewel in a variety of industrial contexts (scheduling, financial engineering, transportation networks, rankings, machine learning, etc.), optimization allows users to find the largest output, the smallest wait time, the winning conditions, and so on.

Optimization problems seen in calculus classes are often solved using differential tools. In this whirlwind tour of the optimization landscape, we discuss problems that do not lend themselves to such an approach, providing a quick survey of optimization problems and algorithms, modeling techniques, an software.


**5.1 Beginnings** . . . . . 227
**5.2 Single-Objective Problems** 228
**Feasible/Optimal Solutions** 229
**Unsolvable Problems** . . . . 230
**Possible Tasks** . . . . . 230
**5.3 Problems Types** . . . . . 231
**Classification** . . . . . 231
**Algorithms** . . . . . 232
**5.4 Linear Programming** . . . . 233
**LP Duality** . . . . . 235
**Solving LP Problems** . . . . 237
**5.5 Mixed-Integer LP** . . . . . 238
**Cutting Planes** . . . . . 241
**5.6 Useful Techniques** . . . . . 241
**Activation** . . . . . 242
**Disjunction** . . . . . 242
**Soft Constraints** . . . . . 242
**5.7 Software Solvers** . . . . . 243
**5.8 Data Envelopment Analysis** 244
**Challenges and Pitfalls** . . . 246
**Pros and Cons** . . . . . 247
**DEA Solvers** . . . . . 247
**Case Study: Schools** . . . . . 248
**5.9 Exercises** . . . . . 252
**Chapter References** . . . . . 252


1: And popular techniques for solving them.

## 5.1 Beginnings

We start by looking at some of the most common types of **single-objective optimization problems** that arise in practice.[1] The following toy problems introduce some of the fundamental notions.

1. Let $S$ be the set of all the four-letter English words. What is the maximum number of $\ell$'s a word in $S$ can have?

   There are numerous four-letter words that contain the letter $\ell$ – for example, "line", "long", "tilt", and "full". From this short list alone, we know the maximum number of $\ell$'s is at least 2 and at most 4. As "llll" is not an English word, the maximum number cannot be 4. Can the maximum number be 3? Yes, because "lull" is a four-letter word with three $\ell$'s.

   This example illustrates some fundamental ideas in optimization. In order to say that 3 is the correct answer, we need to

   - search for a word that has three $\ell$'s, and
   - provide an argument that rules out any value higher than 3.

   In this example, the only possible value of $\ell$ higher than 3 is 4, which was easily ruled out. That cannot always be done – if the problem was to find the maximum number of $y$'s, would the same approach work?

2. A pirate lands on an island with a knapsack that can hold 50kg of treasure. She finds a cave with the following items:

| Item | Weight | Value | Value/kg |
|---|---|---|---|
| iron shield | 20kg | \$2800.00 | \$140.00/kg |
| gold chest | 40kg | \$4400.00 | \$110.00/kg |
| brass sceptre | 30kg | \$1200.00 | \$40.00/kg |

Which items can she bring back home in order to maximize her reward without breaking the knapsack?

If the pirate does not take the gold chest, she can take both the iron shield and the brass sceptre for a total value of \$4000. If she takes the gold chest, she cannot take any of the remaining items. However, the value of the gold chest is \$4400, which is larger than the combined value of the iron shield and the brass sceptre. Hence, the pirate should just take the gold chest.

Here, we performed a case analysis and **exhausted all the promising possibilities** to arrive at our answer. Note that a **greedy strategy** that chooses items in descending value per weight would give us the sub-optimal solution of taking the iron shield and brass sceptre.

Even though there are problems for which the greedy approach would return an optimal solution, the second example is not such a problem. The general version of this problem is the classic **binary knapsack problem** and is known to be **NP-hard**.[2]

2: Informally, NP-hard optimization problems are problems for which no algorithm can provide an output in polynomial time – when the problem size is large, the run time explodes.

Many real-world optimization problems are NP-hard. Despite the theoretical difficulty, practitioners often devise methods that return "good-enough solutions" using **approximation methods** and heuristics. There are also ways to obtain **bounds** to gauge the **quality** of the solutions obtained. We will be looking at these issues at a later stage.

## 5.2 Single-Objective Optimization Problems

A typical single-objective optimization problem consists of a **domain set** $\mathcal{D}$, an **objective function** $f : \mathcal{D} \to \mathbb{R}$, and predicates $\mathcal{C}_i$ on $\mathcal{D}$, where $i = 1, \ldots, m$ for some non-negative integer $m$, called **constraints**.

We want to find, if possible, an element $\mathbf{x} \in \mathcal{D}$ such that $\mathcal{C}_i(\mathbf{x})$ holds for $i = 1, \ldots, m$ and the value of $f(\mathbf{x})$ is either as high (in the case of **maximization**) or as low (in the case of **minimization**) as possible. Compactly, **single-objective optimization problems** are written down as:

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathcal{C}_i(\mathbf{x}) \quad i = 1, \ldots, m \\ & \mathbf{x} \in \mathcal{D}, \end{array}$$

in the case of minimizing $f(\mathbf{x})$, or

$$\begin{array}{ll} \max & f(\mathbf{x}) \\ \text{s.t.} & \mathcal{C}_i(\mathbf{x}) \quad i = 1, \ldots, m \\ & \mathbf{x} \in \mathcal{D}, \end{array}$$

in the case of maximizing $f(\mathbf{x})$.

Here, "s.t." is an abbreviation for "subject to." Technically, "min" should be replaced with "inf" (and "max" with "sup") since the minimum value is not necessarily attained. However, we will abuse notation and ignore this subtle distinction.

Some common domain sets include:

- $\mathbb{R}_+^n$ (the set of $n$-tuples of non-negative real numbers)
- $\mathbb{Z}_+^n$ (the set of $n$-tuples of non-negative integers)
- $\{0,1\}^n$ (the set of binary $n$-tuples)

The **Binary Knapsack Problem** (BKP) can be formulated using the notation we have just introduced. Suppose that there are $n$ **items**, with item $i$ having **weight** $w_i$ and **value** $v_i > 0$ for $i = 1, \ldots, n$.

Let $K$ denote the **capacity** of the knapsack. Then the BKP can be formulated as:

$$\begin{array}{ll} \max & \sum_{i=1}^{n} v_i x_i \\ \text{s.t.} & \sum_{i=1}^{n} w_i x_i \leq K \\ & x_i \in \{0,1\} \quad i = 1, \ldots, n. \end{array}$$

Note that there is only one constraint given by the inequality modeling the capacity of the knapsack. For the pirate example discussed previously, the BKP is:

$$\begin{array}{ll} \max & 2800x_1 + 4400x_2 + 1200x_3 \\ \text{s.t.} & 20x_1 + 40x_2 + 30x_3 \leq 50 \\ & x_1, x_2, x_3 \in \{0,1\}. \end{array}$$

### 5.2.1 Feasible and Optimal Solutions

An element $\mathbf{x} \in \mathcal{D}$ satisfying all the constraints (i.e., $\mathcal{C}_i(\mathbf{x})$ holds for all $i = 1, \ldots, m$) is called a **feasible solution** and its **objective function value** is $f(\mathbf{x})$. For a minimization (resp. maximization) problem, a feasible solution $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ (resp. $f(\mathbf{x}^*) \geq f(\mathbf{x})$) for every feasible solution $\mathbf{x}$ is called an **optimal solution**.

The objective function value of an optimal solution, if it exists, is the **optimal value** of the optimization problem. If an optimal value exists, it is by necessity unique, but the problem can have multiple optimal solutions. Consider, for instance, the following example:

$$\begin{array}{ll} \min & x + y \\ \text{s.t.} & x + y \geq 1 \\ & \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2 \end{array}$$

This problem has an optimal solution

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 - t \\ t \end{bmatrix}$$

for every $t \in \mathbb{R}$, but a unique optimal value of 1.

### 5.2.2 Infeasible/Unbounded Problems

It is possible that there exists no element $\mathbf{x} \in \mathcal{D}$ such that $\mathcal{C}_i(\mathbf{x})$ holds for all $i = 1, \ldots, m$. In such a case, the optimization problem is said to be **infeasible**. The following problem, for instance, is infeasible:

$$\begin{array}{ll} \min & x \\ \text{s.t.} & x \leq -1 \\ & x \geq 0 \\ & x \in \mathbb{R} \end{array}$$

Indeed, any solution $x$ must be simultaneously non-negative and smaller than $-1$, which is patently impossible. An optimization problem that is not infeasible can still fail to have an optimal solution, however.

For instance, the problem

$$\begin{array}{ll} \max & x \\ \text{s.t.} & x \in \mathbb{R} \end{array}$$

is not infeasible, but the max/sup does not exist since the objective function can take on values larger than any candidate maximum. Such a problem is said to be **unbounded**.

On the other hand, the problem

$$\begin{array}{ll} \min & e^{-x} \\ \text{s.t.} & x \in \mathbb{R}, \end{array}$$

has a positive objective function value for every feasible solution. Even though the objective function value approaches 0 as $x \to \infty$, there is no feasible solution with an objective function value of 0. Note that this problem is **not** unbounded as the objective function value is bounded below by 0.

### 5.2.3 Possible Tasks

Given an optimization problem, the most natural task is to find an optimal solution (provided that one exists) and to demonstrate that it is optimal.

However, depending on the context of the problem, one might be instead tasked to find:

- a feasible solution (or show that none exists);
- a local optimum;
- a good bound on the optimal value;
- all global solutions;
- a "good" (but **not necessarily optimal**) solution, quickly;
- a "good" solution that is **robust to small changes** in problem data, and/or
- the $N$ best solutions.

In many contexts, the last three tasks are often more important than finding optimal solutions. For example, if the problem data comes from

measurements or forecasts, one needs to have a solution that is still feasible when deviations are taken into account.

Additionally, producing multiple "good" solutions could allow decision makers to choose a solution that has desirable properties (such as political or traditional requirements) but that is not represented by, or difficult to represent with, problem constraints.

## 5.3 Classification of Optimization Problems and Types of Algorithms

The computational difficulty of optimization problems, then, depends on the properties of the domain set, constraints, and the objective function.

### 5.3.1 Classification

Problems without constraints are said to be **unconstrained**. For example, least-squares minimization in statistics can be formulated as an unconstrained problem, and so can

$$\begin{array}{ll} \min & x^2 - 3x \\ \text{s.t.} & x \in \mathbb{R} \end{array}$$

Problems with linear constraints $g_i$ (i.e., linear inequalities or equalities) and a linear objective function $f$ form an important class of problems in **linear programming**.

Linear programming problems are by far the **easiest** to solve in the sense that efficient algorithms exist both in theory and in practice. Linear programming is also the backbone for solving more complex models [2].

**Convex problems** are problems with a **convex** domain set, which is to say a set $\mathcal{D}$ such that

$$t\mathbf{x}_1 + (1-t)\mathbf{x}_2 \in \mathcal{D}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$ and for all $t \in [0,1]$, and convex constraints $g_i$ and function $f$, which is to say,

$$h(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq th(\mathbf{x}_1) + (1-t)h(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$, and for all $t \in [0,1]$, $h \in \{f, g_i\}$.

Convex optimization problems have the property that **every local optimum is also a global optimum**. Such a property permits the development of effective algorithms that could also work well in practice. Linear programming is a special case of convex optimization.

**Nonconvex problems** (such as problems involving integer variables and/or nonlinear constraints that are not convex) are the hardest problems to solve. In general, nonconvex problems are **NP-hard**. Such problems often arise in scheduling and engineering applications.

In the rest of the chapter, we will primarily focus on linear programming and nonconvex problems whose linear constraints $g_i$ and objective function $f$ are linear, but with domain set $\mathcal{D} \subseteq \mathbb{R}^k \times \mathbb{Z}_+^{n-k}$.

These problems cover a large number of applications in operations research, which are often discrete in nature. We will not discuss optimization problems that arise in statistical learning and engineering applications that are modeled as nonconvex continuous models since they require different sets of techniques and methods – more information is available in [1], and in Chapters 4 and 31.

### 5.3.2 Algorithms

We omit the specific algorithmic details of various optimization methods,[3] as consultants and analytsts are usually expected to use **off-the-shelf** solvers for the various tasks, but it could prove insightful for analysts to know of the various types of algorithms or methods that exist for solving optimization problems.

3: Which would be better left for a graduate course on the subject anyway.

Algorithms fall into three families: **heuristics**, **exact**, and **approximate**.

**Heuristics** These are normally quick to execute but do not provide guarantees of optimality. For example, the **greedy heuristic** for the knapsack problem is very quick but does not always return an optimal solution.[4]

4: In fact, no guarantee exists for the "validity" of a solution in that case.

Other heuristics methods include **ant colony**, **particle swarm**, and **evolutionary algorithms**, just to name a few. There are also heuristics that are stochastic in nature and have proof of convergence to an optimal solution. **Simulated annealing** and **multiple random starts** are such heuristics.

Unfortunately, there is no guarantee on the **running time to reach optimality** and there is no way to **identify when one has reached an optimum point**.

**Exact Methods** Some approaches return a global optimum after a finite run time.

However, most exact methods can only guarantee that constraints are approximately satisfied (though the potential violations fall below some pre-specified tolerance). It is therefore possible for the **returned solutions to be infeasible** for the actual problem.

There also exist exact methods that fully control the error. When using such a method, an optimum is usually given as a **box guaranteed to contain an optimal solution** rather than a single element.

Returning boxes rather than single elements are helpful in cases, for example, where the optimum cannot be expressed exactly as a vector of floating point numbers.

Such exact methods are used mostly in academic research and in areas such as medicine and avionics where the tolerance for errors is practically zero.

**Approximate Methods** Some algorithms eventually zoom in on sub-optimal solutions, while providing a guarantee: this solution is at most $\varepsilon$ away from the optimal solution, say.

In other words, approximate methods also provide a proof of **solution quality**.

## 5.4 Linear Programming

**Linear programming** (LP) was developed independently by G.B. Dantzig and L. Kantorovich in the first half of the 20$^{\text{th}}$ century to solve resource planning problems.

Even though linear programming is insufficient for many modern-day applications in operations research, it was used extensively in economic and military contexts in the early days.

To motivate some key ideas in linear programming, we begin with an example.

**Example:** A roadside stand sells lemonade and lemon juice. Each unit of lemonade requires 1 lemon and 2 litres of water to prepare, and each unit of lemon juice requires 3 lemons and 1 litre of water to prepare. Each unit of lemonade gives a profit of 3$ dollars upon selling, while each unit of lemon juice gives a profit of 2$ dollars.

With 6 lemons and 4 litres of water available, how many units of lemonade and lemon juice should be prepared in order to maximize profit?

If we let $x$ and $y$ denote the number of units of lemonade and lemon juice, respectively, to prepare, then the profit is the **objective function**, given by $(3x + 2y)$\$.

Note that a number of constraints must be satisfied by $x$ and $y$:

- $x$ and $y$ should be **non-negative**;
- the number of lemons needed to make $x$ units of lemonade and $y$ units of lemon juice is $x + 3y$ and cannot exceed 6;
- the number of litres of water needed to make $x$ units of lemonade and $y$ units of lemon juice is $2x + y$ and cannot exceed 4;

Hence, to determine the maximum profit, we need to maximize $3x + 2y$ subject to $x$ and $y$ satisfying the constraints $x + 3y \leq 6$, $2x + y \leq 4$, $x \geq 0$, and $y \geq 0$.

A more compact way to write the problem is as follows:

$$
\begin{array}{lrcrcl}
\max & 3x & + & 2y & & \\
\text{s.t.} & x & + & 3y & \leq & 6 \\
 & 2x & + & y & \leq & 4 \\
 & x & & & \geq & 0 \\
 & & & y & \geq & 0. \\
 & x & , & y & \in & \mathbb{R}.
\end{array}
$$

It is customary to omit the specification of the domain set in linear programming since the variables always take on real numbers. Hence, we can simply write

$$\begin{array}{llrcrcl} \max & 3x & + & 2y & & \\ \text{s.t.} & x & + & 3y & \leq & 6 \\ & 2x & + & y & \leq & 4 \\ & x & & & \geq & 0 \\ & & & y & \geq & 0. \end{array}$$

We can solve the above maximization problem graphically, as follows. We first sketch the set of $[x, y]^\top$ satisfying the constraints, called the **feasible region**, on the $(x, y)-$plane.

We then take the objective function $3x + 2y$ and turn it into the equation of a line $3x + 2y = c$ where $c$ is a parameter. Note that as the value of $c$ increases, the line defined by the equation $3x + 2y = c$ moves in the direction of the normal vector $[3, 2]^\top$. We call this direction the **direction of improvement**. Determining the maximum value of the objective function, called the optimal value, subject to the contraints amounts to finding the maximum value of $c$ so that the line defined by the equation $3x + 2y = c$ still intersects the feasible region.

Figure 5.1 shows the (objective function) lines with $c = 0, 4, 6.8$.

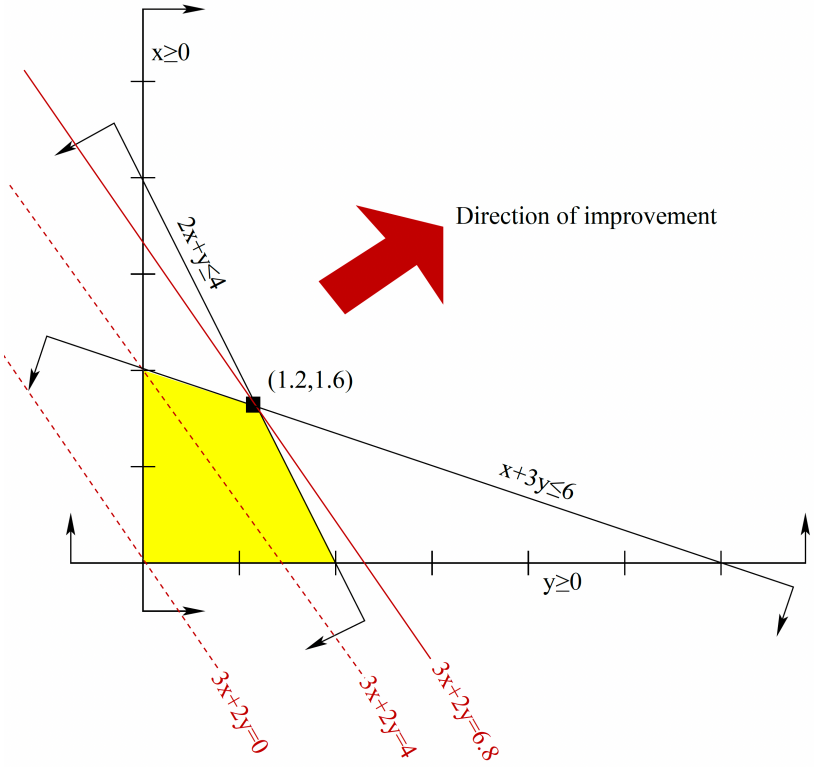



**Figure 5.1:** Graphical solution for the lemonade and lemon juice optimization problem; the feasible region is shown in yellow, and level curves of the objective function in red.

We can see that if $c$ is greater than 6.8, the line defined by $3x + 2y = c$ will not intersect the feasible region. Hence, the profit cannot exceed 6.8 dollars.

As the line $3x + 2y = 6.8$ does intersect the feasible region, 6.8 is the maximum value for the objective function. Note that there is only one

point in the feasible region that intersects the line $3x + 2y = 6.8$, namely $[x^*, y^*]^\top = [1.2, 1.6]^\top$. In other words, to maximize profit, we want to prepare 1.2 units of lemonade and 1.6 units of lemon juice.

This solution method can hardly be regarded as rigorous because we relied on a picture to conclude that $3x + 2y \leq 6.8$ for all $[x, y]^\top$ satisfying the constraints. But we can also obtain this result **algebraically**.

Note that multiplying both sides of the constraint $x + 3y \leq 6$ by 0.2 yields

$$0.2x + 0.6y \leq 1.2,$$

and multiplying both sides of the constraint $2x + y \leq 4$ by 1.4 yields

$$2.8x + 1.4y \leq 5.6.$$

Hence, any $[x, y]^\top$ that satisfies both

$$x + 3y \leq 6 \quad \text{and} \quad 2x + y \leq 4$$

must also satisfy

$$(0.2x + 0.6y) + (2.8x + 1.4y) \leq 1.2 + 5.6,$$

which simplifies to $3x + 2y \leq 6.8$, as desired.

It is always possible to find an algebraic proof like the one above for linear programming problems, which adds to their appeal. To describe the full result, it is convenient to call on **duality**, a central notion in mathematical optimization.

### 5.4.1 Linear Programming Duality

Let $P$ denote following linear programming problem:

$$\begin{array}{ll} \min & \mathbf{c}^\mathbf{T}\mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x} \geq \mathbf{b} \end{array}$$

where $\mathbf{c} \in \mathbb{R}^n$ $\mathbf{b} \in \mathbb{R}^m$ $\mathbf{A} \in \mathbb{R}^{m \times n}$ (inequality on $m-$tuples is applied component-wise.)

Then for every $\mathbf{y} \in \mathbb{R}^m_+$ (that is, all components of $\mathbf{y}$ are non-negative), the inferred inequality $\mathbf{y}^\mathbf{T}\mathbf{A}\mathbf{x} \geq \mathbf{y}^\mathbf{T}\mathbf{b}$ is valid for all $\mathbf{x}$ satisfying $\mathbf{A}\mathbf{x} \geq \mathbf{b}$.

Furthermore, if $\mathbf{y}^\mathbf{T}\mathbf{A} = \mathbf{c}^\mathbf{T}$, the inferred inequality becomes $\mathbf{c}^\mathbf{T}\mathbf{x} \geq \mathbf{y}^\mathbf{T}\mathbf{b}$, making $\mathbf{y}^\mathbf{T}\mathbf{b}$ a lower bound on the optimal value of $P$. To obtain the largest possible bound, we can solve

$$\begin{array}{ll} \max & \mathbf{y}^\mathbf{T}\mathbf{b} \\ \text{s.t.} & \mathbf{y}^\mathbf{T}\mathbf{A} = \mathbf{c}^\mathbf{T} \\ & \mathbf{y} \geq \mathbf{0}. \end{array}$$

This problem is called the **dual problem** of $P$, and $P$ is called the **primal problem**. A remarkable result relating $P$ and its dual $P'$ is the **Duality Theorem for Linear Programming**: if $P$ has an optimal solution, then so does its dual problem $P'$, and the optimal values of the two problems are the same.

A **weaker result** follows easily from the discussion above: the objective function value of a feasible solution to the dual problem $P'$ is a lower bound on the objective function value of a feasible solution to $P$. This result is known as **weak duality**. Despite the fact that it is a simple result, its significance in practice cannot be overlooked because it provides a way to gauge the quality of a feasible solution to $P$.

For example, suppose we have at hand a feasible solution to $P$ with objective function value 3 and a feasible solution to the dual problem $P'$ with objective function value 2. Then we know that the objective function value of our current solution to $P$ is within 1.5 times the actual optimal value since the optimal value cannot be less than 2.

In general, a linear programming problem can have a more complicated form. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$. Let $\mathbf{a}^{(i)^\mathbf{T}}$ denote the $i$th row of $\mathbf{A}$, $\mathbf{A}_j$ denote the $j$th column of $\mathbf{A}$, and $P$ denote the minimization problem, with variables in the tuple $\mathbf{x} = [x_1, \cdots, x_n]^\top$, given as follows:

- the objective function to be minimized is $\mathbf{c}^\mathbf{T}\mathbf{x}$;
- the constraints are $\mathbf{a}^{(i)^\mathbf{T}}\mathbf{x} \sqcup_i b_i$, where $\sqcup_i$ is $\leq$, $\geq$, or $=$ for $i = 1, \ldots, m$, and
- for each $j \in \{1, \ldots, n\}$, $x_j$ is constrained to be non-negative, non-positive, or **free**.

Then the **dual problem** $P'$ is defined to be the maximization problem, with variables in the tuple $\mathbf{y} = [y_1, \cdots, y_m]^\top$ given as follows:

- the objective function to be maximized is $\mathbf{y}^\mathbf{T}\mathbf{b}$;
- for $j = 1, \ldots, n$, the $j$th constraint is

$$\begin{cases} \mathbf{y}^\mathbf{T}\mathbf{A}_j \leq c_j & \text{if } x_j \text{ is constrained to be non-negative} \\ \mathbf{y}^\mathbf{T}\mathbf{A}_j \geq c_j & \text{if } x_j \text{ is constrained to be nonpositive} \\ \mathbf{y}^\mathbf{T}\mathbf{A}_j = c_j & \text{if } x_j \text{ is free.} \end{cases}$$

- and for each $i \in \{1, \ldots, m\}$, $y_i$ is constrained to be non-negative if $\sqcup_i$ is $\geq$; $y_i$ is constrained to be non-positive if $\sqcup_i$ is $\leq$; $y_i$ is free if $\sqcup_i$ is $=$.

The following table can help remember the correspondences:

| Primal (min) | Dual (max) |
|---|---|
| $\geq$ constraint | $\geq 0$ variable |
| $\leq$ constraint | $\leq 0$ variable |
| $=$ constraint | free variable |
| $\geq 0$ variable | $\geq$ constraint |
| $\leq 0$ variable | $\leq$ constraint |
| free variable | $=$ constraint |

Below is an example of a **primal-dual pair** of problems based on the above definition.

Consider the primal problem:

$$\left|\begin{array}{rrrrrrrl} \min & x_1 & - & 2x_2 & + & 3x_3 & & \\ \text{s.t.} & -x_1 & & & + & 4x_3 & = & 5 \\ & 2x_1 & + & 3x_2 & - & 5x_3 & \geq & 6 \\ & & & 7x_2 & & & \leq & 8 \\ & x_1 & & & & & \geq & 0 \\ & & & x_2 & & & & \text{free} \\ & & & & & x_3 & \leq & 0. \end{array}\right.$$

Here, $\mathbf{A} = \begin{bmatrix} -1 & 0 & 4 \\ 2 & 3 & -5 \\ 0 & 7 & 0 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 5 \\ 6 \\ 8 \end{bmatrix}$, and $\mathbf{c} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$.

Since the primal problem has three constraints, the dual problem has three variables:

- the first constraint in the primal is an equation, the corresponding variable in the dual is free;
- the second constraint in the primal is a $\geq$-inequality, the corresponding variable in the dual is non-negative;
- the third constraint in the primal is a $\leq$-inequality, the corresponding variable in the dual is non-positive.

Since the primal problem has three variables, the dual problem has three constraints:

- the first variable in the primal is non-negative, the corresponding constraint in the dual is a $\leq$-inequality;
- the second variable in the primal is free, the corresponding constraint in the dual is an equation;
- the third variable in the primal is non-positive, the corresponding constraint in the dual is a $\geq$-inequality.

Hence, the dual problem is:

$$\left|\begin{array}{rrrrrrrl} \max & 5y_1 & + & 6y_2 & + & 8y_3 & & \\ \text{s.t.} & -y_1 & + & 2y_2 & & & \leq & 1 \\ & & & 3y_2 & + & 7y_3 & = & -2 \\ & 4y_1 & - & 5y_2 & & & \geq & 3 \\ & y_1 & & & & & & \text{free} \\ & & & y_2 & & & \geq & 0 \\ & & & & & y_3 & \leq & 0. \end{array}\right.$$

In some references, the primal problem is always a **maximization problem** – in that case, what we have considered to be a primal problem is their dual problem and *vice-versa*.[5]

5: Note that the **Duality Theorem for Linear Programming** remains true for the more general definition of the primal-dual pair of linear programming problems.

### 5.4.2 Methods for Solving LP Problems

There are currently two families of methods used by modern-day linear programming solvers: **simplex methods** and **interior-point methods**.

We will not get into the technical details of these methods, except to say that the algorithms in either family are iterative, that there is no

known simplex method that runs in polynomial time, but efficient polynomial-time interior-point methods abound in practice. We might wonder why anyone would still use simplex methods, given that they are not polynomial-time methods: simply put, simplex methods are in general more **memory-efficient** than interior-point methods, and they tend to return solutions that have few nonzero entries.

More concretely, suppose that we want to solve the following problem:

$$\begin{array}{ll} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{array}$$

For ease of exposition, we assume that $\mathbf{A}$ has full row rank. Then, each iteration of a simplex method maintains a current solution $\mathbf{x}$ that is **basic**, in the sense that the columns of $\mathbf{A}$ corresponding to the nonzero entries of $\mathbf{x}$ are linearly independent. In contrast, interior-point methods will maintain $\mathbf{x} > \mathbf{0}$ throughout (whence the name "interior point").

When we use an off-the-shelf linear programming solver, the choice of method is usually not too important since solvers have good default settings. Simplex methods are typically used in settings when a problem needs to be resolved after minor changes in the problem data or in problems with additional integrality constraints discussed in the next section.

## 5.5 Mixed-Integer Linear Programming

While the simplicity of linear programming (and duality) make it an appealing tool, its modeling power is insufficient in many real-life applications (for example, there is no simple linear programming formulation of the BKP).

Fortunately, allowing the domain set to restrict one or more variables to integer values drastically extends the modeling power. The price we pay is that there is no guarantee that the problems can be solved in polynomial time.

**Example:** Recall the lemonade and lemon juice problem introduced in the previous section: there is a unique optimal solution at $[x, y]^\top = [1.2, 1.6]^\top$ for a profit of 6.8.

But this solution requires the preparation of **fractional units** of lemonade and lemon juice. What if the number of prepared units needs to be integers?

The solution is to add integrality constraints:

$$\begin{array}{llrcrcl} \max & 3x & + & 2y & & & \\ \text{s.t.} & x & + & 3y & \leq & 6 \\ & 2x & + & y & \leq & 4 \\ & x & & & \geq & 0 \\ & & & y & \geq & 0 \\ & x & , & y & \in & \mathbb{Z}. \end{array}$$

This problem is no longer a linear programming problem; rather, it is an **integer linear programming problem**. Note that we can solve this problem via a case analysis. The second and third inequalities tell us that the possible values for $x$ are 0, 1, and 2.

- If $x = 0$, the first inequality gives $3y \leq 6$, implying that $y \leq 2$. Since we are maximizing $3x + 2y$, we want $y$ to be as large as possible; $[x, y]^\top = [0, 2]^\top$ satisfies all the constraints with an objective function value of 4.
- If $x = 1$, the first inequality gives $3y \leq 5$, implying that $y \leq 1$. Note that $[x, y]^\top = [1, 1]^\top$ satisfies all the constraints with an objective function value of 5.
- If $x = 2$, the second inequality gives $y \leq 0$. Note that $[x, y]^\top = [2, 0]^\top$ satisfies all the constraints with an objective function value of 6.

Thus, $[x^*, y^*]^\top = [2, 0]^\top$ is an **optimal solution**. How does this compare to the solution of the LP problem of the previous section, both in terms of location of the solution and value of the objective function?

A **mixed-integer linear programming problem** (MILP) is a problem of minimizing or maximizing a linear function subject to finitely many linear constraints such that the number of variables are finite, with at least one of them required to take on integer values.

If all the variables are required to take on integer values, the problem is called a **pure integer linear programming problem** or simply an **integer linear programming problem**. Normally, we assume the problem data to be rational numbers to rule out pathological cases.

Many solution methods for solving MILPs have been devised and some of them first solve the **linear programming relaxation** of the original problem, which is the problem obtained from the original problem by dropping all the integrality requirements on the variables.

For instance, if $P_M$ denotes the following MILP:

$$
\begin{array}{llrcrcrcl}
\min & & x_1 & & & + & x_3 & & \\
\text{s.t.} & -x_1 & + & x_2 & + & x_3 & \geq & 1 \\
& -x_1 & - & x_2 & + & 2x_3 & \geq & 0 \\
& -x_1 & + & 5x_2 & - & x_3 & = & 3 \\
& x_1 & , & x_2 & , & x_3 & \geq & 0 \\
& & & & & x_3 & \in & \mathbb{Z}.
\end{array}
$$

then the linear programming relaxation $P_1$ of $P_M$ is:

$$
\begin{array}{llrcrcrcl}
\min & & x_1 & & & + & x_3 & & \\
\text{s.t.} & -x_1 & + & x_2 & + & x_3 & \geq & 1 \\
& -x_1 & - & x_2 & + & 2x_3 & \geq & 0 \\
& -x_1 & + & 5x_2 & - & x_3 & = & 3 \\
& x_1 & , & x_2 & , & x_3 & \geq & 0.
\end{array}
$$

Observe that the optimal value of $P_1$ is a lower bound for the optimal value of $P_M$ since the feasible region of $P_1$ contains all the feasible solutions to $P_M$, thus making it possible to find a feasible solution to $P_1$

with objective function value which is better than the optimal value of $P_M$.

Hence, if an optimal solution to the **LP relaxation** happens to be a feasible solution to the original problem, then it is also an optimal solution to the original problem. Otherwise, there is an integer variable having a nonintegral value $v$.

What we then do is to create two new sub-problems as follows:

- one requiring the variable to be at most the greatest integer less than $v$,
- the other requiring the variable to be at least the smallest integer greater than $v$.

This is the basic idea behind the **branch-and-bound method**. We now illustrate these ideas on $P_M$. Solving the linear programming relaxation $P_1$, we find that $\mathbf{x}' = \frac{1}{3}[0, 2, 1]^\top$ is an optimal solution to $P_1$. Note that $\mathbf{x}'$ is not a feasible solution to $P_M$ because $x_3'$ is not an integer.

We now create two sub-problems $P_2$ and $P_3$. $P_2$ is obtained from $P_1$ by adding the constraint $x_3 \leq \lfloor x_3' \rfloor$,[6] and $P_3$ is obtained from $P_1$ by adding the constraint $x_3 \geq \lceil x_3' \rceil$.

6: $\lfloor a \rfloor$ denotes the **floor of** $a$ and $\lceil a \rceil$ denotes the **ceiling of** $a$.

Hence, $P_2$ is the problem

$$
\begin{array}{llrcrcrcl}
\min & & x_1 & & & + & x_3 & & \\
\text{s.t.} & -x_1 & + & x_2 & + & x_3 & \geq & 1 \\
& -x_1 & - & x_2 & + & 2x_3 & \geq & 0 \\
& -x_1 & + & 5x_2 & - & x_3 & = & 3 \\
& & & & & x_3 & \leq & 0 \\
& x_1 & , & x_2 & , & x_3 & \geq & 0,
\end{array}
$$

and $P_3$ is the problem

$$
\begin{array}{llrcrcrcl}
\min & & x_1 & & & + & x_3 & & \\
\text{s.t.} & -x_1 & + & x_2 & + & x_3 & \geq & 1 \\
& -x_1 & - & x_2 & + & 2x_3 & \geq & 0 \\
& -x_1 & + & 5x_2 & - & x_3 & = & 3 \\
& & & & & x_3 & \geq & 1 \\
& x_1 & , & x_2 & , & x_3 & \geq & 0.
\end{array}
$$

Note that any feasible solution to $P_M$ must be a feasible solution to either $P_2$ or $P_3$. Using the help of a solver, one can see that $P_2$ is infeasible. The problem $P_3$, however, has an optimal solution at $\mathbf{x}^* = \frac{1}{5}[0, 4, 5]^\top$, which is also feasible for $P_M$. Hence, $\mathbf{x}^*$ is an optimal solution of $P_M$.

In many instances, there are multiple choices for the variable on which to branch, and for which sub-problem to solve next. These choices can have an impact on the **total computation time**. But there are no hard-and-fast rules (at the moment) to determine the best branching path. This in area of ongoing research.

### 5.5.1 Cutting Planes

Difficult MILP problems often cannot be solved by branch-and-bound methods alone. A technique that is typically employed in solvers is to add valid inequalities to strengthen the linear programming relaxation.

Such inequalities, known as **cutting planes**, are known to be satisfied by all the feasible solutions to the original problem but not by all the feasible solutions to the initial linear programming relaxation.

**Example:** consider the following PILP problem:

$$\begin{array}{llrcrcl} \min & 3x & + & 2y & & \\ \text{s.t.} & 2x & + & y & \geq & 1 \\ & x & + & 2y & \geq & 4 \\ & x & , & y & \in & \mathbb{Z}. \end{array}$$

An optimal solution to the linear programming relaxation is given by

$$[x^+, y^+]^\top = \frac{1}{3}[-2, 7]^\top.$$

Note that adding the inequalities $2x + y \geq 1$ and $x + 2y \geq 4$ yields $3x + 3y \geq 5$, or equivalently,

$$x + y \geq \frac{5}{3}.$$

Since $x + y$ is an integer for every feasible solution $[x, y]^\top$, $x + y \geq 2$ is a valid inequality for the original problem, but is violated by $[x^+, y^+]^\top$. Hence, $x + y \geq 2$ is a cutting plane.

Adding this to the linear programming relaxation, we have

$$\begin{array}{llrcrcl} \min & 3x & + & 2y & & \\ \text{s.t.} & 2x & + & y & \geq & 1 \\ & x & + & 2y & \geq & 4 \\ & x & + & y & \geq & 2. \end{array}$$

which, upon solving, yields $[x^*, y^*]^\top = [-1, 3]^\top$ as an optimal solution.

Since all the entries are integers, this is also an optimal solution to the original problem. In this example, adding a single cutting plane solved the problem. In practice, one often needs to add numerous cutting planes and then continue with branch-and-bound to solve nontrivial MILP problems.

Many methods for generating cutting planes exist – the problem of generating effective cutting planes efficiently is still an active area of research [4].

## 5.6 Useful Modeling Techniques

So far, we have discussed the kinds of optimization problems that can be solved and certain methods available for solving them. Practical success, however, depends upon the effective **translation** and **formulation** of a

problem description into a mathematical programming problem, which often turns out to be as much an art as it is a science.

We will not be discussing formulation techniques in detail (see [7] for a deep dive into the topic) – instead, we highlight modeling techniques that often arise in business applications, which our examples have not covered so far.

### 5.6.1 Activation

Sometimes, we may want to set a binary variable $y$ to 1 whenever some other variable $x$ is positive. Assuming that $x$ is bounded above by $M$, the inequality

$$x \leq My$$

will model the condition. Note that if there is no valid upper bound on $x$, the condition cannot be modeled using a linear constraint.

### 5.6.2 Disjunction

Sometimes, we want $\mathbf{x}$ to satisfy at least one of a list of inequalities; that is,

$$\mathbf{a}^{(1)^\mathbf{T}}\mathbf{x} \geq b_1 \vee \mathbf{a}^{(2)^\mathbf{T}}\mathbf{x} \geq b_2 \vee \cdots \vee \mathbf{a}^{(k)^\mathbf{T}}\mathbf{x} \geq b_k.$$

To formulate such a **disjunction** using linear constraints, we assume that, for $i = 1, \ldots, k$, there is a lower bound $M_i$ on $\mathbf{a}^{i^\mathbf{T}}\mathbf{x}$ for all $\mathbf{x} \in \mathcal{D}$. Note that such bounds automatically exist when $\mathcal{D}$ is a bounded set, which is often the case in applications.

The disjunction can now be formulated as the following system where $y_i$ is a new 0-1 variable for $i = 1, \ldots, k$:

$$\left|\begin{array}{l} \mathbf{a}^{(1)^\mathbf{T}}\mathbf{x} \geq b_1 y_1 + M_1(1 - y_1) \\ \mathbf{a}^{(2)^\mathbf{T}}\mathbf{x} \geq b_2 y_2 + M_2(1 - y_2) \\ \qquad \vdots \\ \mathbf{a}^{(k)^\mathbf{T}}\mathbf{x} \geq b_k y_k + M_k(1 - y_k) \\ y_1 + \cdots + y_k \geq 1. \end{array}\right.$$

Note that $\mathbf{a}^{i^\mathbf{T}}\mathbf{x} \geq b_i y_i + M_i(1 - y_i)$ reduces to $\mathbf{a}^{i^\mathbf{T}}\mathbf{x} \geq b_i$ when $y_i = 1$, and to $\mathbf{a}^{i^\mathbf{T}}\mathbf{x} \geq M_i$ when $y_i = 0$, which holds for all $\mathbf{x} \in \mathcal{D}$.

Therefore, $y_i$ is an activation for the $i^{\text{th}}$ constraint, and at least one is activated because of the constraint

$$y_1 + \cdots + y_k \geq 1.$$

### 5.6.3 Soft Constraints

Sometimes, we may be willing to pay a price in exchange for specific constraints to be violated (perhaps they represent "nice-to-have" conditions instead of "must-be-met" conditions). Such constraints are referred to as **soft constraints**.

There are situations in which having soft constraints is advisable, say when enforcing all constraints results into an infeasible problem, but a solution is nonetheless needed.

We illustrate the idea on a modified BKP. As usual, there are $n$ items and item $i$ has weight $w_i$ and value $v_i > 0$ for $i = 1, \ldots, n$. The capacity of the knapsack is denoted by $K$. Suppose that we prefer not to take more than $N$ items, but that the **preference** is not an actual constraint.

We assign a penalty for its violation and use the following formulation:

$$\begin{array}{lll} \max & \sum_{i=1}^{n} v_i x_i - py & \\ \text{s.t.} & \sum_{i=1}^{n} w_i x_i \leq K & \\ & \sum_{i=1}^{n} x_i - y \leq N & \\ & x_i \in \{0,1\} & i = 1, \ldots, n \\ & y \geq 0. & \end{array}$$

Here, $p$ is a non-negative number of our choosing. As we are maximizing

$$\sum_{i=1}^{n} v_i x_i - py,$$

$y$ is pushed towards 0 when $p$ is "large". Therefore, the problem will be biased towards solutions that try to violate $x_1 + \cdots + x_n \leq N$ as little as possible.

Experimentation is required to determine What value to select for $p$; the general rule is that if violation is costly in practice, we should set $p$ to be (relatively) high; otherwise, we set it to a moderate value relative to the coefficients of the variables in the objective function value.

Note that when $p = 0$ , the constraint $x_1 + \cdots + x_n \leq N$ has no effect because $y$ can take on any positive value without incurring a penalty.

## 5.7 Software Solvers

A wide variety of solvers exist for all kinds of optimization problems. The NEOS Server is a free online service that hosts many solvers and is a great resource for experimenting with different solvers on **small** problems.

For **large** or **computationally challenging** problems, it is advisable to use a solver installed on a dedicated private machine/server. Commercial solvers can also prove useful:

- IBM ILOG Cplex ;
- Gurobi , or
- FICO Xpress Optimization .

There are popular open-source solvers as well, although they are not as powerful as the commercial tools:

- CBC ;

- GLPLK
- SCIP (requires a commercial licence for consulting work);
- JuliaOpt, to name a few.

We mention in passing that learning how to use of any of these solvers effectively requires a significant time investment. In addition, it is common to build optimization models using a modeling system such as GAMS and LINDO, or a modeling language such as AMPL, ZIMPL, or JuMP.

Note that in the data science and machine learning context, more straightforward methods like **gradient descent**, **stochastic gradient descent** and **Newton's method** are usually sufficient for most applications.

## 5.8 Data Envelopment Analysis

**Operations research** (OR) is a mish-mash of various mathematical methods used to solve complex industrial problems, especially optimization problems, which are being tackled in management and other non-industrial contexts.

**Data Envelopment Analysis** (DEA), based on linear programming, is used to measure the relative performance of units in an organization such as a government department, a school, a company, etc. Typically, a unit's **efficiency** is defined as the quotient of its **outputs**[7] by its **inputs**.[8]

7: Activities of the organization such as service levels or number of deliveries.

8: The resources supporting the organization's operations, such as wages or value of the in-store stock.

In an organization with only one type of input and one type of output, the comparison is simple. For instance, a fictional organization could have the simple input/out data in the table below:

| Unit | Input | Output | Efficiency |
|---|---|---|---|
| A | 10 | 10 | 100% |
| B | 10 | 20 | 200% |
| C | 5 | 15 | 300% |
| D | 15 | 10 | 67% |

However, if there are more than one input or output, the comparisons are less obvious: in the table below, is unit $A$ more efficient than unit $B$?

| Unit | Input 1 | Input 2 | Output 1 | Output 2 |
|---|---|---|---|---|
| A | 10 | 5 | 10 | 20 |
| B | 10 | 15 | 20 | 5 |
| C | 5 | 15 | 15 | 15 |
| D | 15 | 5 | 10 | 20 |

Unit $A$ has fewer total inputs than unit $B$ (as well as fewer outputs of type 1, but it has a substantially more outputs of type 2. Without a system in place to measure relative efficiency, comparison between (potentially incommensurate) units is unlikely to be fruitful.

The **relative efficiency** of unit $k$ is defined by

$$\mathrm{RE}_k = \frac{\sum_j w_{k,j} O_{k,j}}{\sum_i v_{k,i} I_{k,i}},$$

where

- $\{O_{k,j} \mid j = 1, \ldots, n\}$ represent the $n$ **outputs** from unit $k$,
- $\{I_{k,i} \mid i = 1, \ldots, m\}$ represent the $m$ inputs from unit $k$,
- $\{w_{k,j} \mid j = 1, \ldots, n\}$ and $\{v_{k,i} \mid i = 1, \ldots, m\}$ are the **associated unit weights**.

For a specific unit $k$, the DEA model maximizes the **weighted sum of outputs** for a **fixed weighted sum of inputs** (usually set to 100), subject to the weighted sum of outputs of every unit being at most equal to the weighted sum of its inputs when using the DEA weights of unit $k$.

In other words, the optimal set of weights for a given unit could not give another unit a relative efficiency greater than 1.

This is equivalent to solving the following linear program for each unit $k_0$:

$$\left|\begin{array}{ll} \max & \sum_{j=1}^{n} w_{k_0,j} O_{k_0,j} \\ \text{s.t.} & \sum_{i=1}^{m} v_{k_0,i} I_{k_0,i} = 100 \\ & \sum_{j=1}^{n} w_{k_0,j} O_{\ell,j} - \sum_{i=1}^{m} v_{k_0,i} I_{\ell,i} \leq 0, \quad 1 \leq \ell \leq K \\ & (w_{k_0,j}, v_{k_0,i}) \geq \boldsymbol{\varepsilon}, \quad 1 \leq j \leq n,\ 1 \leq i \leq m \end{array}\right.$$

where $\boldsymbol{\varepsilon} \geq 0$ is a parameter vector to be modified by the user.

If we define $\mathbf{w}_\ell$, $\mathbf{v}_\ell$, $\mathbf{O}_\ell$ and $\mathbf{I}_\ell$ as the vectors of output weights, input weights, outputs and inputs, respectively, for unit $\ell$, while $\mathbf{O}$ and $\mathbf{I}$ represent the row matrix of outputs and the row matrix of inputs for all the units, then the linear problem can be re-written simply as

$$\left|\begin{array}{ll} \max & \mathbf{w}_{k_0}^\top \mathbf{O}_{k_0} \\ \text{s.t.} & \mathbf{v}_{k_0}^\top \mathbf{I}_{k_0} = 100 \\ & \mathbf{w}_{k_0}^\top \mathbf{O} - \mathbf{v}_{k_0}^\top \mathbf{I} \leq \mathbf{0} \\ & -(\mathbf{w}_{k_0}, \mathbf{v}_{k_0}) \leq -\boldsymbol{\varepsilon} \end{array}\right.$$

This problem can be solved by the method of **Lagrange multipliers** (see Section 2.5.5) or by using dedicated **numerical solvers** (see previous Section 5.7).

With the data from the example above,the DEA program for unit $A$, for instance, becomes

$$\left|\begin{array}{lrcrcrcrcl} \max & 10w_{A,1} & + & 20w_{A,2} & & & & & & \\ \text{s.t.} & & & & & 10v_{A,1} & + & 5v_{A,2} & = & 100 \\ & 10w_{A,1} & + & 20w_{A,2} & - & 10v_{A,1} & - & 5w_{A,2} & \leq & 0 \\ & 20w_{A,1} & + & 5w_{A,2} & - & 10v_{A,1} & - & 15w_{A,2} & \leq & 0 \\ & 15w_{A,1} & + & 15w_{A,2} & - & 5v_{A,1} & - & 15w_{A,2} & \leq & 0 \\ & 10w_{A,1} & + & 20w_{A,2} & - & 15v_{A,1} & - & 5w_{A,2} & \leq & 0 \\ & w_{A,1} & , & w_{A,2} & , & v_{A,1} & , & v_{A,2} & \geq & \varepsilon \end{array}\right.$$

### 5.8.1 Challenges and Pitfalls

By allowing non-universal (unit-specific) weights, DEA allows each unit to present itself in the **best possible light**, which could potentially lead most units to be deemed efficient. This issue is mitigated to some extent when the number of units $K$ is greater than the product of the number of outputs by the number of inputs $n \cdot m$.

When the number of units is small, a lack of differentiation among units is uninformative since all units could benefit from the best-case scenario described above. When there is differentiation, however, it can be quite telling: units with low DEA relative efficiency have achieved a low score **even when given a chance to put their best foot forward**.

Another concern is that a unit could artificially seem efficient by completely eliminating unfavourable outputs or inputs (i.e.if the associated input/output weights are 0). Constraining the weights to take values in some fixed range can help avoid this issue.

In the example that was discussed above, when we set $\varepsilon = 0$, all units have a relative efficiency of 100. If we set $\varepsilon = 2$, however, the relative efficiency for each unit is

$$\text{RE}_A = 100, \quad \text{RE}_B = 67.7, \quad \text{RE}_C = 100, \quad \text{and} \quad \text{RE}_D = 90.$$

Evidently, insisting that **all** the factors be considered may affect the results.

External factors can easily be added to the model as either inputs or outputs. Available resources are classified as inputs; activity levels or performance measures are classified as outputs.

When units can also be assessed according to some other measure (such as profitability, average rate of success for a task, or environmental cleanliness, say), it can be tempting to solely use the second metric to rank the units.

The combination of **efficiency** and **profitability** (or of any two measures, really) can however offer insights and suggestions:

**Flagships** are units who score high on both measures and that can provide examples of good operating practices (as long as it is recognized that they are also likely beneficiaries of favourable conditions).

**Sleepers** score low on efficiency but high on the other measure, which is probably more a consequence of favourable conditions than good management; as such, they become candidates for efficiency drives.

**Dogs** score high on efficiency but low on the other measure, which indicates good management but unfavourable conditions. In extreme case, these units are candidates for closures, their staff members could be re-assigned to other units.

**Question Marks** are units who score low on both measures; they are subject to unfavourable conditions, but this could also be a consequence of bad management. Attempts should be made to increase the efficiency of these units so that they become Sleepers or Flagships.

Finally, note that in any reasonable application, the linear program to be solved (or its dual) can be fairly **complicated** and sophisticated software can be required to obtain a solution. That is emblematic of industrial optimization problems.

### 5.8.2 Advantages and Disadvantages

The main **benefits** of DEAs are that:

- there is no need to explicitly specify a mathematical form for the production function;
- they have been proven to be useful in uncovering relationships that remain hidden from other methodologies;
- they are capable of handling multiple inputs and outputs;
- they can be used with any input-output measurements, and
- the sources of inefficiency can be analysed and quantified for every evaluated unit.

On the other hand, there are also **disadvantages** to using DEAs:

- the results are known to be sensitive to the selection of inputs and outputs;
- it is impossible to test for the best specification, and
- the number of efficient units on the frontier tends to increase with the number of inputs and output variables.

As is the case for all applications of quantitative methods to real-world problems, DEAs will ultimately prove useless **unless users understand how they function and how to interpret their results**.

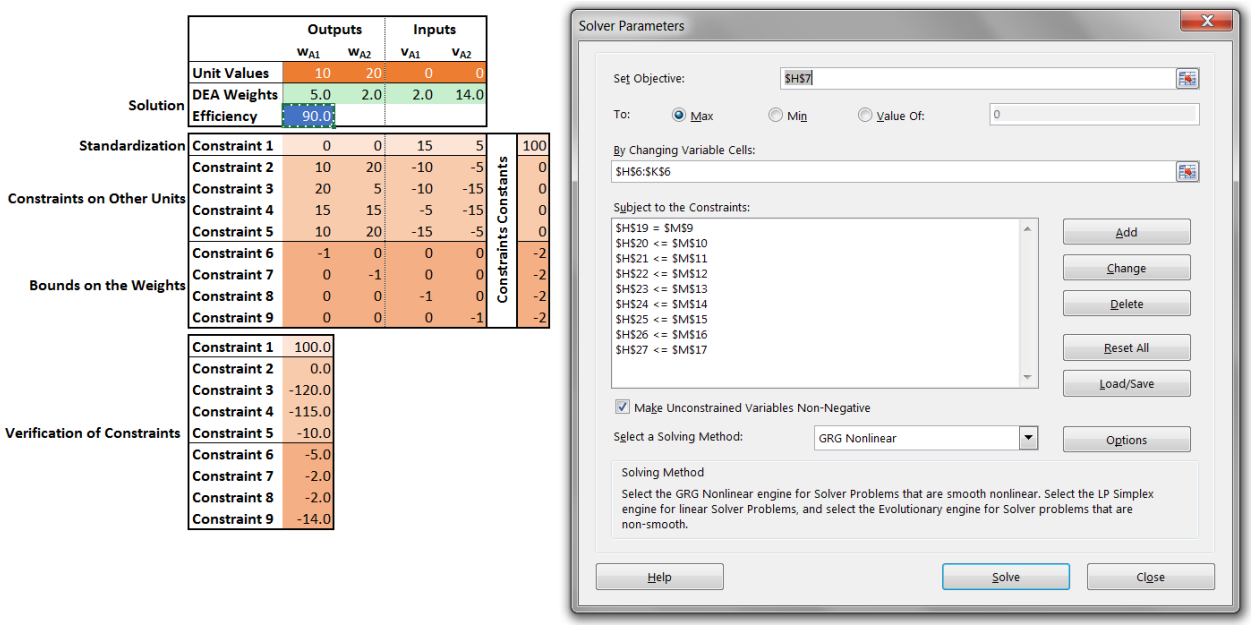
### 5.8.3 SAS, Excel, and R DEA Solvers

For small problems, the numerical cost of solving the problem is not too onerous. Consequently, such problems can typically be solved without having to purchase a commercial solver.

As an illustration, consider the problem of finding the relative efficiency of unit $D$ in the example arising from the data presented above (using a minimal weight threshold of $\varepsilon = 2$, say). Thus, we are looking for the solution to

$$
\begin{array}{lrcrcrcrcl}
\max & 10w_{D,1} & + & 20w_{D,2} & & & & & & \\
\text{s.t.} & & & & & 15v_{D,1} & + & 5v_{D,2} & = & 100 \\
& 10w_{D,1} & + & 20w_{D,2} & - & 10v_{D,1} & - & 5w_{D,2} & \leq & 0 \\
& 20w_{D,1} & + & 5w_{D,2} & - & 10v_{D,1} & - & 15w_{D,2} & \leq & 0 \\
& 15w_{D,1} & + & 15w_{D,2} & - & 5v_{D,1} & - & 15w_{D,2} & \leq & 0 \\
& 10w_{D,1} & + & 20w_{D,2} & - & 15v_{D,1} & - & 5w_{D,2} & \leq & 0 \\
& w_{D,1} & , & w_{D,2} & , & v_{D,1} & , & v_{D,2} & \geq & 2
\end{array}
$$

This is a small problem, and Excel's numerical solver can be used to yield a relative efficiency of 90% (see Figure 5.2 for an illustration).



| | | Outputs | | Inputs | |
|---|---|---|---|---|---|
| | | $w_{A1}$ | $w_{A2}$ | $v_{A1}$ | $v_{A2}$ |
| Solution | Unit Values | 10 | 20 | 0 | 0 |
| | DEA Weights | 5.0 | 2.0 | 2.0 | 14.0 |
| | Efficiency | 90.0 | | | |
| Standardization | Constraint 1 | 0 | 0 | 15 | 5 |
| Constraints on Other Units | Constraint 2 | 10 | 20 | -10 | -5 |
| | Constraint 3 | 20 | 5 | -10 | -15 |
| | Constraint 4 | 15 | 15 | -5 | -15 |
| | Constraint 5 | 10 | 20 | -15 | -5 |
| Bounds on the Weights | Constraint 6 | -1 | 0 | 0 | 0 |
| | Constraint 7 | 0 | -1 | 0 | 0 |
| | Constraint 8 | 0 | 0 | -1 | 0 |
| | Constraint 9 | 0 | 0 | 0 | -1 |

| Constraints Constants |
|---|
| 100 |
| 0 |
| 0 |
| 0 |
| 0 |
| -2 |
| -2 |
| -2 |
| -2 |

| Verification of Constraints | |
|---|---|
| Constraint 1 | 100.0 |
| Constraint 2 | 0.0 |
| Constraint 3 | -120.0 |
| Constraint 4 | -115.0 |
| Constraint 5 | -10.0 |
| Constraint 6 | -5.0 |
| Constraint 7 | -2.0 |
| Constraint 8 | -2.0 |
| Constraint 9 | -14.0 |



**Figure 5.2:** Excel's numerical solver for unit $D$ in the simple DEA problem.

There are a number of non-technical issues with the solver, including the fact that a different worksheet has to be created for every single unit. With larger datasets, this approach may not be practical.

SAS's `proc optmodel`, available in version 9.2+ as part of the OR(R) suite, can also be used; but some additional work has to be done to automate the descriptions of the programs to be solved. R's `rDEA` and `deaR` packages provide other options.

### 5.8.4 Case Study: Barcelona Schools

In this section, we present an illustration of a **resource utlization model** which uses a DEA-like approach.[9]

9: Other optimization case studies can be found in [3].

- **Title:** On centralized resource utilization and its re-allocation by using DEA [6]
- **Authors:** Cecilio Mar-Molinero, Diego Prior, Maria-Manuela Segovia, Fabiola Portillo
- **Date:** 2012
- **Methods:** Data envelopment analysis, simulations

**Abstract** The standard DEA model allows different **Decision-Making Units** (DMUs) to set their own priorities for the inputs and outputs that form part of the efficiency assessment. In the case of a centralized organization with many outlets, such as an education authority that is responsible for many schools, it may be more sensible to operate in the most efficient way, but under a common set of priorities for all DMUs. The centralized resource allocation model does just this; the optimal resource reallocation is found for Spanish public schools and it is shown that the most desirable operating unit is a by-product of the estimation.

**Data** The data consists of 54 secondary public schools in Barcelona during the year 2008, each with three **discretionary inputs** (teaching hours per week, $x_1$; specialized teaching hours per week, $x_2$; capital investments in the last decade, $x_3$), one **non-discretionary input** (total number of students present at the beginning of the academic year, $X$) and two **outputs** (number of students passing their final assessment, $y_1$, and number of students continuing their studies at the end of the academic year, $y_2$).

A subset of the data is shown in Table 5.5.

| School # | $y_1$ | $y_2$ | $x_{1d}$ | $x_{2d}$ | $x_{3d}$ | $X_{1nd}$ |
|---|---|---|---|---|---|---|
| 1 | 260.65 | 378.00 | 44.00 | 3.00 | 8 | 384.00 |
| 2 | 195.18 | 213.00 | 32.01 | 3.00 | 18 | 225.00 |
| 3 | 242.75 | 429.70 | 56.98 | 4.00 | 84 | 446.00 |
| 4 | 283.02 | 350.00 | 49.50 | 3.00 | 39 | 356.00 |
| 5 | 376.76 | 650.80 | 77.50 | 5.50 | 61 | 657.00 |
| 6 | 252.19 | 429.00 | 49.40 | 2.00 | 56 | 440.00 |
| 7 | 225.50 | 247.34 | 33.15 | 1.50 | 43 | 248.00 |
| 8 | 363.85 | 364.34 | 45.90 | 2.00 | 36 | 381.00 |
| 9 | 261.87 | 272.00 | 44.37 | 2.00 | 24 | 288.00 |
| 10 | 235.40 | 251.00 | 35.49 | 1.50 | 51 | 259.00 |
| 11 | 198.63 | 223.34 | 42.00 | 1.50 | 46 | 227.00 |
| 12 | 159.78 | 248.00 | 36.96 | 2.00 | 2 | 250.00 |
| 13 | 98.09 | 193.00 | 35.20 | 1.50 | 55 | 203.00 |
| 14 | 214.92 | 219.00 | 33.60 | 1.50 | 32 | 229.00 |
| 15 | 136.07 | 269.20 | 33.80 | 1.50 | 54 | 271.00 |
| 16 | 214.68 | 346.00 | 54.39 | 2.00 | 33 | 347.00 |
| 17 | 117.12 | 196.00 | 29.00 | 1.50 | 7 | 212.00 |
| 18 | 261.89 | 334.00 | 42.40 | 2.00 | 47 | 339.00 |

**Table 5.5:** Sample from the Barcelona public school dataset used with the radial and simplified models.

**Challenges** A first challenge is that the machinery of DEA cannot directly be brought to bear on the problem since the models under consideration are at best DEA-like. Another challenge is that the number of unknowns to be estimated in the original model is quadratic in the number of units. Consequently, the original model must be simplified to avoid difficulties when the number of units is large. Fortunately, the proposed simplifications can be interpreted logically in the context of re-allocation of resources.

Finally, there are situations where a solution to the simplified problem can be obtained even when the constraints on the total number of units is relaxed, allowing for the possibility of reaching the similar output levels with fewer inputs, in effect advocating for the closure of some units.

While this is a technically-correct solution, it might could prove to be an **unadvisable** one for a variety of non-technical reasons: closing schools is not usually a politically and/or societally palatable strategy. This latter factor should also be incorporated in the decision-making process.

**Project Summary and Results** In the standard DEA model, each unit sets its own priorities, and is evaluated using unit-specific weights. In a **de-centralized environment**, the standard approach is reasonable, but under a central authority where a common set of priorities needs to be met by all units (such as the branches of a bank, or recycling collection vehicles in a city), that approach needs to be modified.

In a school setting, school board administrators may wish to evaluate teachers in a similar manner independently of the school at which they work. Centralized assessment imposes a common set of weights. For weakly centralized management, it is a further assumption that any input excess of inefficient units can be re-allocated among the efficient units, but only as long as this does not contravene the built-in inflexibility of the system, which may make re-allocation rather difficult.

Strongly centralized management, on the other hand, allow for re-allocation of the majority of inputs and outputs among all the units (inefficient or efficient) with the aim of optimizing the performance of the entire system. The original **radial** model of Lozano and Villa [5] is not, strictly speaking, a data envelopment model:

$$\min \theta \text{ (objective)}$$

$$\text{s.t. } \sum_{r=1}^{54}\sum_{j=1}^{54} \lambda_{j,r} x_{i,j} - \theta \sum_{j=1}^{54} x_{i,j} \le 0, \quad \text{for } i = 1,2,3$$

(discretionary inputs)

$$\sum_{r=1}^{54}\sum_{j=1}^{54} \lambda_{j,r} X_j - \sum_{j=1}^{54} X_j \le 0,$$

(non-discretionary input)

$$\sum_{r=1}^{54} y_{kr} - \sum_{r=1}^{54}\sum_{j=1}^{54} \lambda_{j,r} y_{k,j} \le 0, \quad \text{for } k = 1,2$$

(outputs)

$$\sum_{j=1}^{54} \lambda_{j,r} = 54, \quad \text{for } r = 1,\ldots,54$$

$$-\lambda_{j,r} \le 0, \quad \text{for } j,r = 1,\ldots,54, \quad \theta \text{ free}$$

Indeed, this model is not asking every unit to select the weights that make it look as good as possible when comparing itself to the remaining units under the same assessment; rather, it is asking for the system as a whole to find the weights that present it in the best possible light possible, then it assesses the performance of the units separately, using the optimal system weights.

This conceptual shift leads to proposed closures.The main drawback of the radial model is the large number of weights to estimate. A simplification is proposed: if some of the units can be cloned, or equivalently, if some of the units can be closed and their resources re-allocated to other units, then the radial model becomes substantially simpler, and the number of weights to estimate is linear in the number of units (as opposed to quadratic).

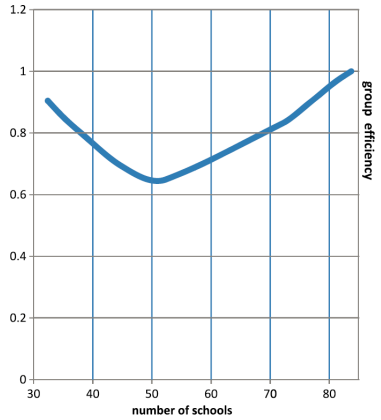



**Figure 5.3:** Results of the re-allocation process in the Barcelona public school dataset.

The new problem is DEA-like:

$$\begin{aligned}
&\min \theta \text{ (objective)}\\
&\text{s.t. } \sum_{j=1}^{54} \lambda_j x_{i,j} - \theta \sum_{j=1}^{54} x_{i,j} \leq 0, \quad \text{for } i = 1,2,3\\
&\text{(discretionary inputs)}\\
&\sum_{j=1}^{54} \lambda_j X_j - \sum_{j=1}^{54} X_j \leq 0\\
&\text{(non-discretionary inputs)}\\
&\sum_{r=1}^{54} y_k - \sum_{j=1}^{54} \lambda_j y_{k,j} \leq 0, \quad \text{for } k = 1,2\\
&\text{(outputs)}\\
&\sum_{j=1}^{54} \lambda_j = 54\\
&-\lambda_j \leq 0, \quad \text{for } j = 1,\ldots,54, \quad \theta \text{ free}
\end{aligned}$$

The numerical solution to the radial model shows a group efficiency of 66%, meaning that the outputs of the system could be produced while reducing the discretionary inputs by $\theta = 34\%$. The simplified model reaches the same group efficiency by cloning units 25 (24.26 times), 26 (20.02 times), 36 (4.71 times), 17 (2.69 times), and 44 (1.70 times).

The re-allocation of inputs and outputs among the 54 schools would produce the aforementioned reduction of the 34% in discretionary inputs.

A simulation experiment shows the effect of dropping the constraint on the number of units: the group efficiency obtained by solving the simplified system for various values of $n$ from 32 to 81 is seen in Figure 5.3.

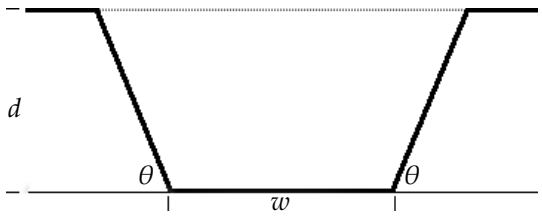
Sure enough, the original solution is good, appearing near the minimum, which reaches $\theta = 0.64$ at $n = 50.36$. This group efficiency corresponds to cloning units 25 (23.96 times), 26 (17.62 times), and 29 (7.87 times). Obviously, schools (and their resources) cannot be cloned, so what are we to make of this result?[10]

10: It could be argued that unit 25 and 26, for instance, are ideal schools under the common priorities imposed by the system: should new schools have to be built, attempts could be made to emulate the stars. Of course, in practice, other factors could come into play.

## 5.9 Exercises

Some of the questions in this section may need to be solved by a combination of the techniques provided in Chapters 2, 4, and 5.

1. Find the extrema of the function defined by $f(x) = x - \sin(x)$ over the interval $[-2, 12]$.
2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = A - (x^2 + Bx + y^2 + Cy)$, where $A, B, C$ are constants. What values must they take so that $f$ admits a maximum value of 15 when $(x, y) = (-2, 1)$? What if it is a minimal value of 15 when $(x, y) = (-2, 1)$?
3. Consider a factory that produces various types of deluxe pickle jars. The monthly number of jars $Q$ of a specific kind of pickled radish that can be produced at the factory is given by $Q(K, L) = 1000K^{0.21}L^{0.79}$, where $K$ is the number of dedicated canning machines, and $L$ is the monthly number of employee-hours spent on the pickled radish. The pay rate for the employees is 22\$/hour; the monthly maintenance cost for each canning machine is 300\$. If the factory owners want to maintain monthly production at 40,000 jars of pickled radish, what combination of number of canning machines and employee-hour will minimize the total production costs?
4. The distance $d$ at which a projectile can be fired depends on the temperature $t$ and the humidity level $h$, according to
$$E(t, h) = 12,000 - t^2 - 2ht - 2h^2 + 200t + 260h,$$
where $t$ is measured in °F and $0 \leq h \leq 100$. Under what atmospheric conditions should we fire the projectile to maximize the distance it travels? To minimize it?
5. The area of the vertical sections of an irrigation canal is 50 square feet. The average flow of liquid in the canal is inversely proportional to the perimeter of the trapezoid, excluding the length length of the dotted segment, which we will denote by $p$. In order to maximise the flow, we must then minimize $p$. Determine the depth $d$, base $w$ and angle $\theta$ that maximizes the flow.




6. Find the extrema of $f(x, y) = x^2 - y$, subject to $x^2 - y^2 = 1$.
7. Find the extrema of $f(x, y, z) = 2\sqrt{x} + y + 4 \ln z$ subject to $x^2 + y + z^2 = 16$.
8. Solve the linear program: $\arg\min\{0.5x_1 + x_2 \mid x_1 + x_2 \geq 1, x_1 + 0.5x_2 \geq 1, x_1, x_2 \geq 0\}$.

## Chapter References

[1] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[2] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. 1st. Athena Scientific, 1997.

[3] P. Boily and J. Schellinck. *Introduction to Quantitative Consulting*. Quadrangle/Data Action Lab, 2025.

[4] G. Cornuéjols. 'Valid inequalities for mixed integer linear programs'. In: *Math. Program.* 112.1 (2008), pp. 3–44.

[5] S. Lozano and G. Villa. 'Centralized Resource Allocation Using Data Envelopment Analysis'. In: *Journal of Productivity Analysis* 22.1 (July 2004), pp. 143–161.

[6] C. Mar-Molinero et al. 'On centralized resource utilization and its reallocation by using DEA'. In: *Ann. Oper. Res.* 221.1 (2014), pp. 273–283.

[7] H. P. Williams. 'Model Building in Linear and Integer Programming'. In: *Computational Mathematical Programming*. Ed. by Klaus Schittkowski. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985, pp. 25–53.

# Probability and Applications

# 6

by **Patrick Boily**; inspired by **Rafal Kulik**

---

Data analysis is sometimes presented in a "point-and-click manner", with tutorials often bypassing foundations in probability and statistics to focus on software use and specific datasets. While modern analysts do not always need to fully understand the theory underpinning the methods that they use, understanding some of the basic concepts can only lead to long-term benefits.

In this chapter, we introduce some of the crucial probabilistic notions that will help analysts get the most out of their data.


6.1 Basic Notions . . . . . 253
Sample Spaces and Events . 253
Counting Techniques . . . . 254
Ordered Samples . . . . . 255
Unordered Samples . . . . . 257
Probability of an Event . . . 257
Conditionality Probability 260
Bayes' Theorem . . . . . 266
6.2 Discrete Distributions . . . 272
Random Variables . . . . . 272
Expectation . . . . . 275
Binomial R.V. . . . . 277
Geometric R.V. . . . . 282
Negative Binomial R.V. . . . 282
Poisson R.V. . . . . 283
Other Discrete R.V. . . . . 288
6.3 Continuous Distributions . 288
Continuous R.V. . . . . 288
Expectation . . . . . 294
Normal R.V. . . . . 296
Exponential R.V. . . . . 301
Gamma R.V. . . . . 304
Binomial Approximations . 305
Other Continuous R.V. . . . 307
6.4 Joint Distributions . . . . . 307
6.5 CLT/Sampling Distributions 313
Sampling Distributions . . 313
Central Limit Theorem . . . 316
Sampling Distributions II . 323
6.6 Exercises . . . . . 327
Chapter References . . . . . 336


## 6.1 Basic Notions

**Probability theory** is the mathematical discipline relating to the numerical description of the likelihood of an event.

### 6.1.1 Sample Spaces and Events

Throughout, we will deal with **random experiments** (e.g., measurements of speed/ weight, number and duration of phone calls, etc.).

For any "experiment," the **sample space** is defined as the set of all its **possible outcomes**, often denoted by the symbol $\mathcal{S}$. A sample space can be **discrete** or **continuous**.

An **event** is a collection of outcomes from the sample space $\mathcal{S}$. Events will be denoted by $A$, $B$, $E_1$, $E_2$, etc.

#### Examples

- Toss a fair coin – the corresponding (discrete) sample space is $\mathcal{S} = \{\text{Head}, \text{Tail}\}$.
- Roll a die – the corresponding (discrete) sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, with various events represented by
  - rolling an even number: $\{2, 4, 6\}$;
  - rolling a prime number: $\{2, 3, 5\}$.
- Suppose we measure the weight (in grams) of a chemical sample – the (continuous) sample space can be represented by $\mathcal{S} = (0, \infty)$, the positive half line, and various events by subsets of $\mathcal{S}$, such as
  - sample is less than 1.5 grams: $(0, 1.5)$;
  - sample exceeds 5 grams: $(5, \infty)$.

For any events $A, B \subseteq \mathcal{S}$:

- the **union** $A \cup B$ of $A$ and $B$ are all outcomes in $\mathcal{S}$ contained in either $A$ or $B$;
- the **intersection** $A \cap B$ of $A$ and $B$ are all outcomes in $\mathcal{S}$ contained in both $A$ and $B$;
- the **complement** $A^c$ of $A$ (sometimes denoted $\overline{A}$ or $-A$) is the set of all outcomes in $\mathcal{S}$ that are **not** in $A$.

If $A$ and $B$ have no outcomes in common, they are **mutually exclusive**; which is denoted by $A \cap B = \varnothing$ (the empty set). In particular, $A$ and $A^c$ are always mutually exclusive.[1]

1: Events can be represented graphically using Venn diagrams – mutually exclusive events are those which do not have a common intersection.

#### Examples

- Roll a die and let $A = \{2, 3, 5\}$ (a prime number) and $B = \{3, 6\}$ (multiples of 3). Then $A \cup B = \{2, 3, 5, 6\}$, $A \cap B = \{3\}$ and $A^c = \{1, 4, 6\}$.
- 100 plastic samples are analyzed for scratch and shock resistance.

| | | **shock resistance** | |
|---|---|---|---|
| | | high | low |
| **scratch** | high | 70 | 4 |
| **resistance** | low | 1 | 25 |

  If $A$ is the event that a sample has high shock resistance and $B$ is the event that a sample has high scratch residence, then $A \cap B$ consists of 70 samples.

### 6.1.2 Counting Techniques

A **two-stage procedure** can be modeled as having $k$ bags, with $m_1$ items in the first bag, ..., $m_k$ items in $k$-th bag.

The **first stage** consists of picking a bag, and the **second stage** consists of drawing an item out of that bag. This is equivalent to picking one of the $m_1 + \cdots + m_k$ total items.

If all the bags have the same number of items, $m_1 = \cdots = m_k = n$, then there are $kn$ items in total, and this is the **total number of ways** the two-stage procedure can occur.

#### Examples

- How many ways are there to first roll a die and then draw a card from a (shuffled) 52−card pack?

  **Answer:** there are 6 ways the first step can turn out, and for each of these (the stages are independent, in fact) there are 52 ways to draw the card. Thus there are $6 \times 52 = 312$ ways this can turn out.

- How many ways are there to draw two tickets numbered 1 to 100 from a bag, the first with the right hand and the second with the left hand?

  **Answer:** There are 100 ways to pick the first number; for *each of these* there are 99 ways to pick the second number. Thus, the task has $100 \times 99 = 9900$ possible outputs.

#### Multi-Stage Procedures

A $k$**-stage process** is a process for which:

- there are $n_1$ possibilities at stage 1;
- regardless of the 1st outcome there are $n_2$ possibilities at stage 2,
- ...
- regardless of the previous outcomes, there are $n_k$ choices at stage $k$.

There are thus $n_1 \times n_2 \cdots \times n_k$ **total ways** the process can turn out.

### 6.1.3 Ordered Samples

Suppose we have a bag of $n$ billiard balls numbered $1, \ldots, n$. We can draw an **ordered sample** of size $r$ by picking balls from the bag:

- **with replacement**, or
- **without replacement**.

With how many different collection of $r$ balls can we end up in each of those cases (each is an $r$-stage procedure)?

**Key Notion:** all the object (balls) can be differentiated (using numbers, colours, etc.)

#### Sampling With Replacement (Order Important)

If we replace each ball into the bag after it is picked, then every draw is the same (there are $n$ ways it can turn out). According to our earlier result, there are

$$\underbrace{n \times n \times \cdots \times n}_{r \text{ stages}} = n^r$$

ways to select an ordered sample of size $r$ **with replacement** from a set with $n$ objects $\{1, 2, \ldots, n\}$.

#### Sampling Without Replacement (Order Important)

If we **do not** replace each ball into the bag after it is drawn, then the choices for the second draw depend on the result of the first draw, and there are only $n - 1$ possible outcomes.

Whatever the first two draws were, there are $n - 2$ ways to draw the third ball, and so on.

Thus there are

$$\underbrace{n \times (n-1) \times \cdots \times (n-r+1)}_{r \text{ stages}} = {}_nP_r \quad \text{(common symbol)}$$

ways to select an ordered sample of size $r \leq n$ **without replacement** from a set of $n$ objects $\{1, 2, \ldots, n\}$.

### Factorial Notation

For a positive integer $n$, write

$$n! = n(n-1)(n-2)\cdots 1.$$

There are two possibilities:

- when $r = n$, ${}_nP_r = n!$, and the ordered selection (without replacement) is called a **permutation**;
- when $r < n$, we can write

$$\begin{aligned} {}_nP_r &= \frac{n(n-1)\cdots(n-r+1)}{} \frac{(n-r)\cdots 1}{(n-r)\cdots 1} \\ &= \frac{n!}{(n-r)!} = n \times \cdots \times (n-r+1). \end{aligned}$$

By convention, we set $0! = 1$, so that

$${}_nP_r = \frac{n!}{(n-r)!}, \quad \text{for all } r \leq n.$$

**Examples:**

- In how many different ways can 6 balls be drawn *in order* without replacement from a bag of balls numbered 1 to 49?

  **Answer:** We compute

  $${}_{49}P_6 = 49 \times 48 \times 47 \times 46 \times 45 \times 44 = 10,068,347,520.$$

  This is the number of ways the actual drawing of the balls can occur for Lotto 6/49 in real-time (balls drawn one by one).

- How many 6-digits PIN codes can you create from the set of digits $\{0, 1, \ldots, 9\}$?

  **Answer:** If the digits may be repeated, we see that

  $$10 \times 10 \times 10 \times 10 \times 10 \times 10 = 10^6 = 1,000,000.$$

  If the digits may not be repeated, we have instead

  $${}_{10}P_6 = 10 \times 9 \times 8 \times 7 \times 6 \times 5 = 151,200.$$

### 6.1.4 Unordered Samples

Suppose that we **cannot** distinguish between different ordered samples; when we look up the Lotto 6/49 results in the newspaper, for instance, we have no way of knowing the order in which the balls were drawn:

$$1 - 2 - 3 - 4 - 5 - 6$$

could mean that the first drawn ball was ball # 1, the second drawn ball was ball # 2, etc., but it could also mean that the first ball drawn was ball # 4, the second one, ball # 3, etc., or **any combination** of the first 6 balls.

Denote the (as yet unknown) number of unordered samples of size $r$ from a set of size $n$ by ${}_nC_r$. We can derive the expression for ${}_nC_r$ by noting that the following two processes are equivalent:

- take an **ordered** sample of size $r$ (there are ${}_nP_r$ ways to do this);
- take an **unordered** sample of size $r$ (there are ${}_nC_r$ ways to do this) **and then** rearrange (permute) the objects in the sample (there are $r!$ ways to do this).

Thus

$${}_nP_r = {}_nC_r \times r! \implies {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!\,r!} = \binom{n}{r};$$

these are known as **binomial coefficients**, read as "$n$-choose-$r$".

**Example** In how many ways can the "Lotto 6/49 draw" be reported in the newspaper (if they are always reported in increasing order)?

This number is the same as the number of *unordered samples* of size 6 (different re-orderings of same 6 numbers are indistinguishable), so

$$\begin{aligned} {}_{49}C_6 &= \binom{49}{6} = \frac{49 \times 48 \times 47 \times 46 \times 45 \times 44}{6 \times 5 \times 4 \times 3 \times 2 \times 1} \\ &= \frac{10,068,347,520}{720} = 13,983,816\,. \end{aligned}$$ ■

There is a variety of binomial coefficient identities, such as

$$\begin{aligned} \binom{n}{k} &= \binom{n}{n-k}, \quad \text{for all } 0 \le k \le n, \\ \sum_{k=0}^{n} \binom{n}{k} &= 2^n, \quad \text{for all } 0 \le n, \\ \binom{n+1}{k+1} &= \binom{n}{k} + \binom{n}{k+1}, \quad \text{for all } 0 \le k \le n-1 \\ \sum_{j=k}^{n} \binom{j}{k} &= \binom{n+1}{k+1}, \quad \text{for all } 0 \le n, \text{ etc.} \end{aligned}$$

### 6.1.5 Probability of an Event

For situations where we have a random experiment which has exactly $N$ possible **mutually exclusive**, **equally likely** outcomes, we can assign

a probability to an event $A$ by counting the number of outcomes that correspond to $A$ – its **relative frequency**. If that count is $a$, then

$$P(A) = \frac{a}{N}.$$

The probability of each individual outcome is thus $1/N$.

#### Examples

- Toss a fair coin – the sample space is $\mathcal{S} = \{\text{Head}, \text{Tail}\}$, i.e., $N = 2$. The probability of observing a Head on a toss is thus $\frac{1}{2}$.
- Throw a fair six sided die. There are $N = 6$ possible outcomes. The sample space is
$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}.$$
If $A$ corresponds to observing a multiple of 3, then $A = \{3, 6\}$ and $a = 2$, so that
$$\text{Prob(number is a multiple of 3)} = P(A) = \frac{2}{6} = \frac{1}{3}.$$
- The probabilities of seeing an even/odd number are:
$$\text{Prob}\{\text{even}\} = P(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2};$$
$$\text{Prob}\{\text{prime}\} = P(\{2, 3, 5\}) = 1 - P(\{1, 4, 6\}) = \frac{1}{2}.$$
- In a group of 1000 people it is known that 545 have high blood pressure. 1 person is selected randomly. What is the probability that this person has high blood pressure?

  **Answer:** the relative frequency of people with high blood pressure is 0.545.

This approach to probability is called the **frequentist interpretation**. It is based on the idea that the theoretical probability of an event is given by the behaviour of the empirical (observed) relative frequency of the event over long-run repeatable and independent experiments.[2]

2: Such as when $N \to \infty$.

This is the classical definition, and the one used in these notes, but there are competing interpretations which may be more appropriate depending on the context; chiefly, the **Bayesian interpretation** (see [2] and Chapter 25 for details) and the **propensity interpretation**.[3]

3: Introducing causality as a mechanism.

#### Axioms of Probability

The modern definition of probability is **axiomatic** (according to Kolmogorov's seminal work [@KOL]).

The **probability of an event** $A \subseteq \mathcal{S}$ is a numerical value satisfying the following properties:

1. for any event $A$, $1 \geq P(A) \geq 0$;
2. for the complete sample space $\mathcal{S}$, $P(\mathcal{S}) = 1$;
3. for the empty event $\varnothing$, $P(\varnothing) = 0$, and

4. for two **mutually exclusive** events $A$ and $B$, the probability that $A$ or $B$ occurs is $P(A \cup B) = P(A) + P(B)$.

Since $\mathcal{S} = A \cup A^c$, and $A$ and $A^c$ are mutually exclusive, then

$$1 \overset{\mathbf{A2}}{=} P(\mathcal{S}) = P(A \cup A^c) \overset{\mathbf{A4}}{=} P(A) + P(A^c)$$
$$\Longrightarrow P(A^c) = 1 - P(A).$$

**Examples**

- Throw a single six sided die and record the number that is shown. Let $A$ and $B$ be the events that the number is a multiple of or smaller than 3, respectively. Then $A = \{3, 6\}$, $B = \{1, 2\}$ and $A$ and $B$ are mutually exclusive since $A \cap B = \varnothing$. Then

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) = \frac{2}{6} + \frac{2}{6} = \frac{2}{3}.$$

- An urn contains 4 white balls, 3 red balls and 1 black ball. Draw one ball, and denote the following events by $W = \{\text{the ball is white}\}$, $R = \{\text{the ball is red}\}$ and $B = \{\text{the ball is black}\}$. Then

$$P(W) = 1/2, \quad P(R) = 3/8, \quad P(B) = 1/8,$$

and $P(W \text{ or } R) = 7/8$.

**General Addition Rule**

This useful rule is a direct consquence of the axioms of probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

**Example** An electronic gadget consists of two components, $A$ and $B$. We know from experience that $P(A \text{ fails}) = 0.2$, $P(B \text{ fails}) = 0.3$ and $P(\text{both } A \text{ and } B \text{ fail}) = 0.15$. Find $P(\text{at least one of } A \text{ and } B \text{ fails})$ and $P(\text{neither } A \text{ nor } B \text{ fails})$.

Write $A$ for "$A$ fails" and similarly for $B$. Then we are looking to compute

$$\begin{aligned} P(\text{at least one fails}) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) = 0.35\,; \\ P(\text{neither fail}) &= 1 - P(\text{at least one fails}) = 0.65\,. \end{aligned}$$

If $A$, $B$ are mutually exclusive, $P(A \cap B) = P(\varnothing) = 0$ and

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B).$$

With three events, the addition rule expands as follows:

$$\begin{aligned} P(A \cup B \cup C) = &P(A) + P(B) + P(C) \\ &- P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &+ P(A \cap B \cap C). \end{aligned}$$

### 6.1.6 Conditional Probability and Independent Events

Any two events $A$ and $B$ satisfying

$$P(A \cap B) = P(A) \times P(B)$$

are said to be **independent**.[4] When events are not independent, we say that they are **dependent** or **conditional**.

4: This is a purely mathematical definition, but it agrees with the intuitive notion of independence in simple examples.

Mutual exclusivity and independence are unrelated concepts. The only way for events $A$ and $B$ to be mutually exclusive **and** independent is for either $A$ or $B$ (or both) to be a non-event (the empty event):

$$\begin{aligned} \varnothing = P(A \cap B) = P(A) \times P(B) &\implies P(A) = 0 \text{ or } P(B) = 0 \\ &\implies A = \varnothing \text{ or } B = \varnothing. \end{aligned}$$

#### Examples

- Flip a **fair** coin twice – the 4 possible outcomes are all equally likely: $\mathcal{S} = \{HH, HT, TH, TT\}$. Let

  $$A = \{HH\} \cup \{HT\}$$

  denote "head on first flip", $B = \{HH\} \cup \{TH\}$ "head on second flip". Note that $A \cup B \neq \mathcal{S}$ and $A \cap B = \{HH\}$. By the general addition rule,

  $$\begin{aligned} P(A) &= P(\{HH\}) + P(\{HT\}) - P(\{HH\} \cap \{HT\}) \\ &= \frac{1}{4} + \frac{1}{4} - P(\varnothing) = \frac{1}{2} - 0 = \frac{1}{2}. \end{aligned}$$

  Similarly, $P(B) = P(\{HH\}) + P(\{TH\}) = \frac{1}{2}$, and so $P(A)P(B) = \frac{1}{4}$. But $P(A \cap B) = P(\{HH\})$ is also $\frac{1}{4}$, so $A$ and $B$ are independent.

- A card is drawn from a regular well-shuffled 52-card North American deck. Let $A$ be the event that it is an ace and $D$ be the event that it is a diamond. These two events are independent. Indeed, there are 4 aces

  $$P(A) = \frac{4}{52} = \frac{1}{13}$$

  and 13 diamonds

  $$P(D) = \frac{13}{52} = \frac{1}{4}$$

  in such a deck, so that

  $$P(A)P(D) = \frac{1}{13} \times \frac{1}{4} = \frac{1}{52},$$

  and exactly 1 ace of diamonds in the deck, so that $P(A \cap D)$ is also $\frac{1}{52}$.

- A six-sided die numbered $1-6$ is loaded in such a way that the probability of rolling each value is *proportional* to that value. Find $P(3)$.

  Let $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ be the value showing after a single toss; for some proportional constant $v$, we have $P(k) = kv$, for $k \in \mathcal{S}$. By

Axiom **A2**, $P(\mathcal{S}) = P(1) + \cdots + P(6) = 1$, so that

$$1 = \sum_{k=1}^{6} P(k) = \sum_{k=1}^{6} kv = v \sum_{k=1}^{6} k = v\frac{(6+1)(6)}{2} = 21v\,.$$

Hence $v = 1/21$ and $P(3) = 3v = 3/21 = 1/7$.

- Now the die is rolled twice, the second toss *independent* of the first. Find $P(3_1, 3_2)$.

  The experiment is such that $P(3_1) = 1/7$ and $P(3_2) = 1/7$, as seen in the previous example. Since the die tosses are independent,[5] then

  $$P(3_1 \cap 3_2) = P(3_1)P(3_2) = 1/49\,.$$

- Is a 2-engine plane more likely to be forced down than a 3-engine plane?

  This question is easier to answer if we assume that **engines fail independently** (this is no doubt convenient, but the jury is still out as to whether it is realistic). In what follows, let $p$ be the probability that an engine fails.[6]
  The next step is to decide what type engine failure will force a plane down:[7]

  - A 2-engine plane will be forced down if both engines fail – the probability is $p^2$;
  - A 3-engine plane will be forced down if any pair of engines fail, or if all 3 fail.
    * **Pair:** the probability that exactly 1 pair of engines will fail independently (i.e., two engines fail and one does not) is

      $$p \times p \times (1-p).$$

      The order in which the engines fail does not matter: there are ${}_3C_2 = \frac{3!}{2!1!} = 3$ ways in which a pair of engines can fail: for 3 engines A, B, C, these are AB, AC, BC.
    * **All 3:** the probability of all three engines failing independently is $p^3$.

    The probability $\geq 2$ engines failing is thus

    $$P(2+ \text{ engines fail}) = 3p^2(1-p) + p^3 = 3p^2 - 2p^3.$$

  Basically it's safer to use a 2-engine plane than a 3-engine plane: the 3-engine plane will be forced down more often, assuming it needs 2 engines to fly.

  This "makes sense": the 2-engine plane need 50% of its engines working, while the 3-engine plane needs 66% (see Figure 6.1 to get a sense of what the probabilities are for $0 \leq p \leq 1$).

- (Taken from [3]) Air traffic control is a safety-related activity – each piece of equipment is designed to the highest safety standards and in many cases duplicate equipment is provided so that if one item fails another takes over.

5: Is it clear what is meant by "independent tosses' '?

6: What are some realistic values of $p$?

7: There is nothing to that effect in the problem statement, so we have to make another set of assumptions.

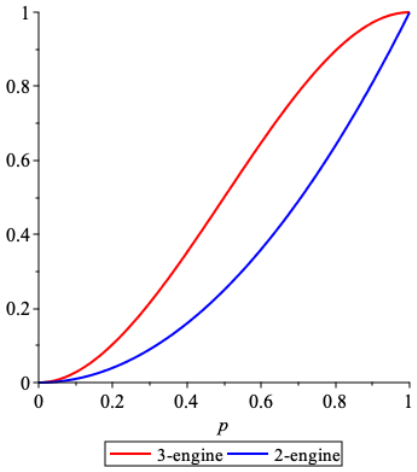



**Figure 6.1:** Failure probability for the 2-engine and 3-engine planes.

A new system is to be provided passing information from Heathrow Airport to Terminal Control at West Drayton. As part of the system design a decision has to be made as to whether it is necessary to provide duplication.

The new system takes data from the *Ground Movements Radar* (GMR) at Heathrow, combines this with data from the *National Airspace System NAS*, and sends the output to a display at *Terminal Control* (a conceptual model is shown in Figure 6.2).

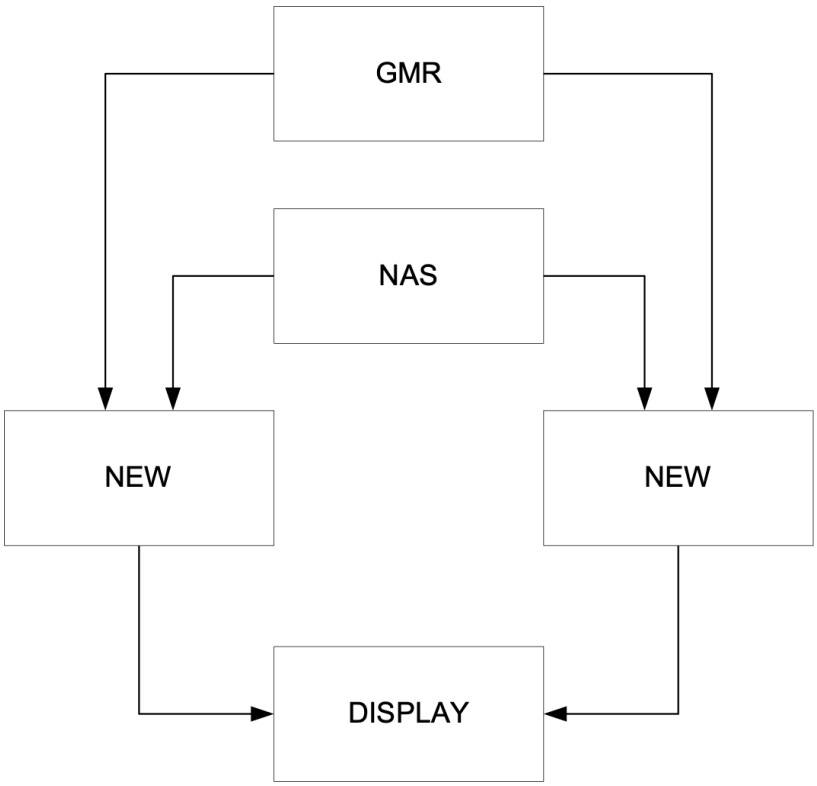



**Figure 6.2:** Conceptual model of air traffic control security system.

For all existing systems, records of failure are kept and an experimental probability of failure is calculated annually using the previous 4 years.

The reliability of a system is defined as $R = 1 - P$, where $P = P(\text{failure})$. We assume that $R_{\text{GMR}} = R_{\text{NAS}} = 0.9999$,[8] and that the components' failure probabilities are independent.

8: That is to say, 1 failure in 10,000 hours.

If a single module is used, the reliability of the **single thread design** (STD) is

$$R_{\text{STD}} = R_{\text{GMR}} \times R_{\text{NEW}} \times R_{\text{NAS}}.$$

If the module is duplicated, the reliability of this **dual thread design** (DTD) is

$$R_{\text{DTD}} = R_{\text{GMR}} \times (1 - (1 - R_{\text{NEW}})^2) \times R_{\text{NAS}}.$$

Duplicating the module causes an improvement in reliability of

$$\rho = \frac{R_{\text{DTD}}}{R_{\text{STD}}} = \frac{(1 - (1 - R_{\text{NEW}})^2)}{R_{\text{NEW}}} \times 100\% .$$

For the module, no historical data is available. Instead, we work out the improvement achieved by using the dual thread design for various values of $R_{\text{NEW}}$.

| $R_{\text{NEW}}$ | 0.1 | 0.2 | 0.5 | 0.75 | 0.99 | 0.999 | 0.9999 | 0.99999 |
|---|---|---|---|---|---|---|---|---|
| $\rho$ (%) | 190 | 180 | 150 | 125 | 101 | 100.1 | 100.01 | 100.001 |

If the module is **very unreliable** (i.e., $R_{\text{NEW}}$ is small), then there is a significant benefit in using the dual thread design ($\rho$ is large).[9] If the new module is **as reliable as** GMR and NAS, that is, if

9: But why would we install a module which we know to be unreliable in the first place?

$$R_{\text{GMR}} = R_{\text{NEW}} = R_{\text{NAS}} = 0.9999,$$

then the single thread design has a combined reliability of 0.9997 (i.e., 3 failures in 10,000 hours), whereas the dual thread design has a combined reliability of 0.9998 (i.e., 2 failures in 10,000 hours).

If the probability of failure is independent for each component, we could conclude from this that the reliability gain from a dual thread design probably does not justify the extra cost.

In the last two examples, we had to make **additional assumptions** in order to answer the questions – this is often the case in practice.

### Conditional Probability

The **conditional probability** of an event $B$ given that another event $A$ has occurred is defined by

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} .$$

Note that this definition only makes sense when "$A$ can happen" i.e., $P(A) > 0$. If $P(A)P(B) > 0$, then

$$P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) = P(B \cap A);$$

$A$ and $B$ are thus independent if $P(B \mid A) = P(B)$ and $P(A \mid B) = P(A)$.

**Examples**

- From a group of 100 people, 1 is selected. What is the probability that this person has high blood pressure (HBP)?

  If we know nothing else about the population, this is an **(unconditional) probability**, namely

  $$P(\text{HBP}) = \frac{\#\text{ individuals with HBP in the population}}{100}.$$

- If instead we first filter out all people with low cholesterol level, and then select 1 person. What is the probability that this person has HBP?

  We are looking for the **conditional probability**

  $$P(\text{HBP} \mid \text{high cholesterol});$$

  the probability of selecting a person with HBP, given high cholesterol levels, presumably different from $P(\text{HBP} \mid \text{low cholesterol})$.

- A sample of 249 individuals is taken and each person is classified by blood type and tuberculosis (TB) status.

| | **O** | **A** | **B** | **AB** | **Total** |
|---|---|---|---|---|---|
| TB | 34 | 37 | 31 | 11 | 113 |
| no TB | 55 | 50 | 24 | 7 | 136 |
| Total | 89 | 87 | 55 | 18 | 249 |

  The (unconditional) probability that a random individual has TB is $P(\text{TB}) = \frac{\#\text{TB}}{249} = \frac{113}{249} = 0.454$. Among those individuals with type **B** blood, the (conditional) probability of having TB is

  $$P(\text{TB} \mid \text{type } \mathbf{B}) = \frac{P(\text{TB} \cap \text{type } \mathbf{B})}{P(\text{type } \mathbf{B})} = \frac{31}{55} = \frac{31/249}{55/249} = 0.564.$$

- A family has two children (not twins). What is the probability that the youngest child is a girl given that at least one of the children is a girl? Assume that boys and girls are equally likely to be born.

  Let $A$ and $B$ be the events that the youngest child is a girl and that at least one child is a girl, respectively:

  $$A = \{\text{GG}, \text{BG}\} \quad \text{and} \quad B = \{\text{GG}, \text{BG}, \text{GB}\},$$

  $A \cap B = A$. Then $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{2/4}{3/4} = \frac{2}{3}$ (and not $\frac{1}{2}$, as might naively be believed).

  Incidentally, $P(A \cap B) = P(A) \neq P(A) \times P(B)$, which means that $A$ and $B$ are **not** independent events.

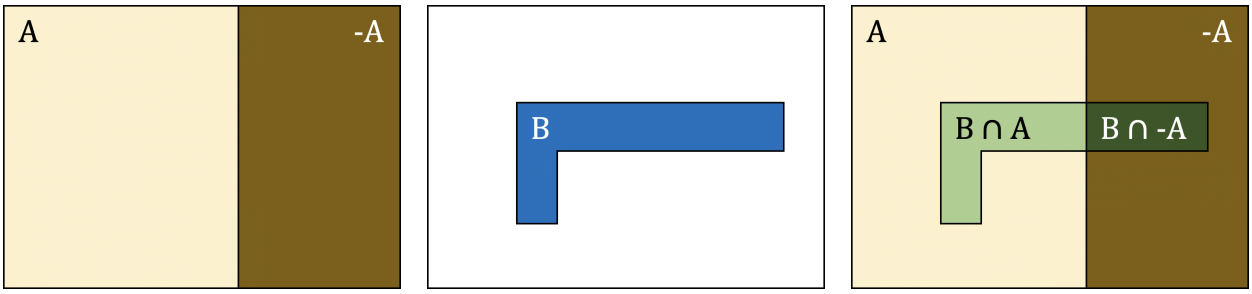



**Figure 6.3:** Decomposition of *B via A*.

**Law of Total Probability**

Let $A$ and $B$ be two events. From set theory, we have

$$B = (A \cap B) \cup (\overline{A} \cap B),$$

as illustrated in Figure 6.3. Note that $A \cap B$ and $\overline{A} \cap B$ are mutually exclusive, so that, according to Axiom **A4**, we have

$$P(B) = P(A \cap B) + P(\overline{A} \cap B).$$

Now, assuming that $\varnothing \neq A \neq \mathcal{S}$, we have

$$P(A \cap B) = P(B \mid A)P(A) \quad \text{and} \quad P(\overline{A} \cap B) = P(B \mid \overline{A})P(\overline{A}),$$

so that

$$P(B) = P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A}).$$

This generalizes as follows: if $A_1, ...A_k$ are **mutually exclusive** and **exhaustive** (i.e., $A\_i \cap A\_j = \varnothing$ for all $i \neq j$ and $A_1 \cup .... \cup A_k = \mathcal{S}$), then for any event $B$

$$P(B) = \sum_{j=1}^{k} P(B \mid A_j)P(A_j) = P(B \mid A_1)P(A_1) + ... + P(B \mid A_k)P(A_k).$$

**Example** With the **Law of Total Probability** (the rule above), compute $P(\text{TB})$ using the data from one of the previous example.

The blood types $\{\mathbf{O}, \mathbf{A}, \mathbf{B}, \mathbf{AB}\}$ form a mutually exclusive partition of the population, with

$$P(\mathbf{O}) = \frac{89}{249},\ P(\mathbf{A}) = \frac{87}{249},\ P(\mathbf{B}) = \frac{55}{249},\ P(\mathbf{AB}) = \frac{18}{249}.$$

It is easy to see that $P(\mathbf{O}) + P(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{AB}) = 1$. Furthermore,

$$P(\text{TB} \mid \mathbf{O}) = \frac{P(\text{TB} \cap \mathbf{O})}{P(\mathbf{O})} = \frac{34}{89},\ P(\text{TB} \mid \mathbf{A}) = \frac{P(\text{TB} \cap \mathbf{A})}{P(\mathbf{A})} = \frac{37}{87},$$
$$P(\text{TB} \mid \mathbf{B}) = \frac{P(\text{TB} \cap \mathbf{B})}{P(\mathbf{B})} = \frac{31}{55},\ P(\text{TB} \mid \mathbf{AB}) = \frac{P(\text{TB} \cap \mathbf{AB})}{P(\mathbf{AB})} = \frac{11}{18}.$$

According to the law of total probability,

$$P(\text{TB}) = P(\text{TB} \mid \mathbf{O})P(\mathbf{O}) + P(\text{TB} \mid \mathbf{A})P(\mathbf{A}) + P(\text{TB} \mid \mathbf{B})P(\mathbf{B}) + P(\text{TB} \mid \mathbf{AB})P(\mathbf{AB}),$$

so that

$$P(\text{TB}) = \frac{34}{89}\cdot\frac{89}{249} + \frac{37}{87}\cdot\frac{87}{249} + \frac{31}{55}\cdot\frac{55}{249} + \frac{11}{18}\cdot\frac{18}{249}$$
$$= \frac{34+37+31+11}{249} = \frac{113}{249} = 0.454,$$

which matches the previous obtained result.

### 6.1.7 Bayes' Theorem

After an experiment generates an outcome, we are often interested in the probability that a certain condition was present given an outcome.[10]

10: Or that a particular hypothesis was valid, say.

We have noted before that if $P(A)P(B) > 0$, then

$$P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) = P(B \cap A);$$

this can be re-written as **Bayes' Theorem**:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}.$$

Bayes' Theorem is a powerful tool in probability analysis, but it is a simple corollary of the rules of probability.

#### Central Data Analysis Question

Given everything that was known prior to the experiment, does the observed data support the hypothesis? The **problem** is that this is usually impossible to compute directly. Bayes' Theorem offers a **possible solution**:

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{data})}$$
$$\propto P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis}),$$

in which the terms on the right might be easier to compute than the term on the left.

#### Bayesian Vernacular

In Bayes' Theorem:

- $P(\text{hypothesis})$ is the **prior** – the probability of the hypothesis being true prior to the experiment;
- $P(\text{hypothesis} \mid \text{data})$ is the **posterior** – the probability of the hypothesis being true once the experimental data is taken into account;
- $P(\text{data} \mid \text{hypothesis})$ is the **likelihood** – the probability of the experimental data being observed assuming that the hypothesis is true.

The theorem is often presented as posterior $\propto$ likelihood $\times$ prior, which is to say, **beliefs should be updated in the presence of new information**.

**Formulations**

If $A, B$ are events for which $P(A)P(B) > 0$, then Bayes' Theorem can be re-written, using the law of total probability, as

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})},$$

or, in the general case where $A_1, \ldots A_k$ are **mutually exclusive** and **exhaustive** events, then for any event $B$ and for each $1 \leq i \leq k$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_k)P(A_k)}.$$

**Examples**

- In 1999, Sinnas sold three car models in North America: Sarten (S), Minima (M), and Papader (PA). Of the vehicles sold that year, 50% were S, 30% were M and 20% were PA; 12% of the S, 15% of the M, and 25% of the PA had a particular defect $D$.

  1. If you own a 1999 Sinnas, what is the probability that it has the defect?

     In the language of conditional probability,

     $$P(\mathrm{S}) = 0.5,\ P(\mathrm{M}) = 0.3,\ P(\mathrm{Pa}) = 0.2,$$
     $$P(D \mid \mathrm{S}) = 0.12,\ P(D \mid \mathrm{M}) = 0.15,\ P(D \mid \mathrm{PA}) = 0.25,$$

     so that

     $$\begin{aligned} P(D) &= P(D \mid \mathrm{S}) \times P(\mathrm{S}) + P(D \mid \mathrm{M}) \times P(\mathrm{M}) + P(D \mid \mathrm{Pa}) \times P(\mathrm{Pa}) \\ &= 0.12 \cdot 0.5 + 0.15 \cdot 0.3 + 0.25 \cdot 0.2 \\ &= 0.155 = 15.5\%. \end{aligned}$$

  2. If a 1999 Sinnas has defect $D$, what model is it likely to be?

     In the first part we computed the total probability $P(D)$; in this part, we compare the posterior probabilities $P(\mathrm{M} \mid D)$, $P(\mathrm{S} \mid D)$, and $P(\mathrm{Pa} \mid D)$ (and not the priors!), computed using Bayes' Theorem:

     $$\begin{aligned} P(\mathrm{S} \mid D) &= \tfrac{P(D|\mathrm{S})P(\mathrm{S})}{P(D)} = \tfrac{0.12 \times 0.5}{0.155} \approx 38.7\% \\ P(\mathrm{M} \mid D) &= \tfrac{P(D|\mathrm{M})P(\mathrm{M})}{P(D)} = \tfrac{0.15 \times 0.3}{0.155} \approx 29.0\% \\ P(\mathrm{Pa} \mid D) &= \tfrac{P(D|\mathrm{Pa})P(\mathrm{Pa})}{P(D)} = \tfrac{0.25 \times 0.2}{0.155} \approx 32.3\% \end{aligned}$$

     Even though Sartens are least likely to have the defect $D$, their overall prevalence in the population carries more weight.

- Suppose that a test for a particular disease has a very high success rate. If a patient:

  1. has the disease, the test is 'positive' with probability 0.99;
  2. does not have the disease, the test reports a 'negative' with prob 0.95.

Assume that only 0.1% of the population has the disease. What is the probability that a patient who tests positive does not have the disease?

Let $D$ be the event that the patient has the disease, and $A$ be the event that the test is positive. The probability of a true positive is

$$P(D \mid A) = \frac{P(A \mid D)P(D)}{P(A \mid D)P(D) + P(A \mid D^c)P(D^c)} = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} \approx 0.019.$$

The probability of a false positive is thus $1 - 0.019 \approx 0.981$. Despite the apparent high accuracy of the test, the incidence of the disease is so low (1 in a 1000) that the vast majority of patients who test positive (98 in 100) do not have the disease.

The 2 in 100 who are true positives still represent 20 times the proportion of positives found in the population (before the outcome of the test is known).[11]

11: It is important to remember that when dealing with probabilities, **both** the likelihood and the prevalence have to be taken into account.

- [**Monty Hall Problem**] On a game show, you are given the choice of three doors. Behind one of the doors is a prize; behind the others, dirty and smelly rubbish bins (as is skillfully rendered in Figure 6.4).

  You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, behind which is a bin. She then says to you, "Do you want to switch from door No. 1 to No. 2?"

  Is it to your advantage to do so?

**Figure 6.4:** The Monty Hall set-up (personal file, ... but that was probably obvious from the artistic quality ).

  In what follows, let and be the events that switching to another door is a successful strategy and that the prize is behind the original door, respectively.

  – Let's first assume that the host opens no door. What is the probability that switching to another door in this scenario would prove to be a successful strategy?

    If the prize is behind the original door, switching would succeed 0% of the time: $P(\mathrm{S} \mid \mathrm{D}) = 0$.[12] If the prize is not behind the original door, switching would succeed 50% of the time: $P(\mathrm{S} \mid \mathrm{D}^c) = 1/2$.[13] Thus,

12: Note that the prior is $P(\mathrm{D}) = 1/3$.

13: Note that the prior is $P(\mathrm{D}^c) = 2/3$.

$$P(\mathrm{S}) = P(\mathrm{S} \mid \mathrm{D})P(\mathrm{D}) + P(\mathrm{S} \mid \mathrm{D}^c)P(\mathrm{D}^c)$$
$$= 0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} \approx 33\%.$$

– Now let's assume that the host opens one of the other two doors to show a rubbish bin. What is the probability that switching to another door in this scenario would prove to be a successful strategy?

  If the prize is behind the original door, switching would succeed 0% of the time: $P(\mathrm{S} \mid \mathrm{D}) = 0$.[14] If the prize is not behind the original door, switching would succeed 100% of the time: $P(\mathrm{S} \mid \mathrm{D}^c) = 1$.[15] Thus,

  $$P(\mathrm{S}) = P(\mathrm{S} \mid \mathrm{D})P(\mathrm{D}) + P(\mathrm{S} \mid \mathrm{D}^c)P(\mathrm{D}^c)$$
  $$= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} \approx 67\%.$$

  If no door is opened, switching is not a winning strategy, resulting in success only 33% of the time. If a door is opened, however, switching becomes the winning strategy, resulting in success 67% of the time.

14: Note that the prior is $P(\mathrm{D}) = 1/3$.

15: Note that the prior is $P(\mathrm{D}^c) = 2/3$.

The Monty Hall problem has attracted a lot of attention over the years due to its counter-intuitive result, but there is no paradox when we understand conditional probabilities.

Perhaps it would be easier to see what happens in practice: if we could pit two players against one another (one who never switches and one who always does so) in a series of Monty Hall games, which one would come out on top in the long run?

We start by setting a number of games $N$ (not too small, or we won't be able to observe long-run behaviour) and a replicability seed (so that we may all obtain the same results).

```
N=500
set.seed(1234)
```

Next, for each of game, we will place the prize behind one of the 3 doors: $A$, $B$, or $C$.

```
locations = sample(c("A","B","C"), N, replace = TRUE)
```

We verify that the prize gets placed behind each door roughly 33% of the time:

```
table(locations)/N
```

```
locations
    A     B     C
0.302 0.344 0.354
```

Le us now obtain a player's first guess for each game – this guess is completely independent of the actual prize location:

```
player.guesses = sample(c("A","B","C"), N, replace = TRUE)
```

Finally, we create a data frame telling the analyst where the prize actually is, and what door the player has selected as their original guess.

```
games = data.frame(locations, player.guesses)
head(games)
```

```
  locations player.guesses
1         B              B
2         B              B
3         A              B
4         C              C
5         A              C
6         A              A
```

In this example (that is, with the data generated above), how often had the player guessed correctly, before a door was opened and they were given a chance to switch?

```
table(games$locations==games$player.guesses)
```

```
FALSE  TRUE
  333   167
```

This should not come as a surprise.

We now initialize the process to find out which door the host opens. For each game, the host opens a door which is not the one selected by the player, nor the one behind which the prize is found.

```
games$open.door <- NA

for(j in 1:N){
  games$open.door[j] <- sample(setdiff(c("A","B","C"),
    union(games$locations[j],games$player.guesses[j])), 1)
}

head(games)
```

```
  locations player.guesses open.door
1         B              B         C
2         B              B         C
3         A              B         C
4         C              C         A
5         A              C         B
6         A              A         B
```

The `union()` call enumerates the doors that the host cannot open; the `setdiff()` call finds the complement of the doors that the host cannot open (i.e.: the doors that she can open), and the `sample()` call picks one of those doors.

If the player never switches, they win whenever they had originally guessed the location of the prize correctly:

```
games$no.switch.win <- games$player.guess==games$locations
```

We find which door the player would have selected if they always switched (the door that is neither the location of the prize nor the one they had originally selected):

```
games$switch.door <- NA

for(j in 1:N){
  games$switch.door[j] <- sample(setdiff(c("A","B","C"),
    union(games$open.door[j],games$player.guesses[j])), 1)
}
```

If the player always switches, they win whenever their switched guess is where the prize is located:

```
games$switch.win <- games$switch.door==games$locations
head(games)
```

```
  locations player.guesses open.door no.switch.win switch.door switch.win
1         B              B         C           TRUE           A      FALSE
2         B              B         C           TRUE           A      FALSE
3         A              B         C          FALSE           A       TRUE
4         C              C         A           TRUE           B      FALSE
5         A              C         B          FALSE           A       TRUE
6         A              A         B           TRUE           C      FALSE
```

The chances of winning by not switching are thus:

```
table(games$no.switch.win)/N
```

```
FALSE  TRUE
0.666 0.334
```

while the chances of winning by switching are:

```
table(games$switch.win)/N
```

```
FALSE  TRUE
0.334 0.666
```

Pretty wild, eh? Numerical simulations show, beyond the shadow of a doubt, that switching IS the better strategy.

## 6.2 Discrete Distributions

In the next sections, we discuss how some of the probability computations can be made easier with the use of **(theoretical) distributions**.[16]

16: Note that the principles of probability theory introduced in the previous section remain valid in all cases.

### 6.2.1 Random Variables and Distributions

Recall that, for any random "experiment", the set of all possible outcomes is denoted by $\mathcal{S}$. A **random variable** (r.v.) is a function $X : \mathcal{S} \to \mathbb{R}$, which is to say, it is a rule that associates a (real) number to every outcome of the experiment; $\mathcal{S}$ is the **domain** of the r.v. $X$ and $X(\mathcal{S}) \subseteq \mathbb{R}$ is its **range**.

A **probability distribution function** (p.d.f.) is a function $f : \mathbb{R} \to \mathbb{R}$ which specifies the probabilities of the values in the range $X(\mathcal{S})$. When $\mathcal{S}$ is **discrete**,[17] we say that $X$ is a **discrete r.v.** and the p.d.f. is called a **probability mass function** (p.m.f.).

17: For the purpose of these notes, a discrete set is one in which all points are **isolated**: $\mathbb{N}$ and finite sets are discrete, but $\mathbb{Q}$ and $\mathbb{R}$ are not.

#### Notation

Throughout, we use the following notation:

- capital roman letters ($X$, $Y$, etc.) denote r.v., and
- corresponding lower case roman letters ($x$, $y$, etc.) denote *generic values taken* by the r.v.

A discrete r.v. can be used to **define events** – if $X$ takes values $X(\mathcal{S}) = \{x_i\}$, then we can define the events $A_i = \{s \in \mathcal{S} : X(s) = x_i\}$ :

- the p.m.f. of $X$ is $f(x) = P(\{s \in \mathcal{S} : X(s) = x\}) := P(X = x)$;
- its **cumulative distribution function** (c.d.f.) is $F(x) = P(X \leq x)$.

#### Properties

If $X$ is a discrete random variable with p.m.f. $f(x)$ and c.d.f. $F(x)$, then

- $0 < f(x) \leq 1$ for all $x \in X(\mathcal{S})$; $\sum_{s\in\mathcal{S}} f(X(s)) = \sum_{x\in X(\mathcal{S})} f(x) = 1$;
- for any event $A \subseteq \mathcal{S}$, $P(X \in A) = \sum_{x\in A} f(x)$;
- for any $a, b \in \mathbb{R}$,

$$P(a < X) = 1 - P(X \leq a) = 1 - F(a)$$
$$P(X < b) = P(X \leq b) - P(X = b) = F(b) - f(b)$$

- for any $a, b \in \mathbb{R}$,

$$P(a \leq X) = 1 - P(X < a) = 1 - (P(X \leq a) - P(X = a)) = 1 - F(a) + f(a).$$

We can use these results to compute the probability of a **discrete** r.v. $X$ falling in various intervals:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a);$$
$$P(a \leq X \leq b) = P(a < X \leq b) + P(X = a) = F(b) - F(a) + f(a);$$
$$P(a < X < b) = P(a < X \leq b) - P(X = b) = F(b) - F(a) - f(b);$$
$$P(a \leq X < b) = P(a \leq X \leq b) - P(X = b) = F(b) - F(a) + f(a) - f(b).$$

**Examples**

- Flip a fair coin – the outcome space is $\mathcal{S} = \{\text{Head}, \text{Tail}\}$. Let $X : S \to \mathbb{R}$ be defined by $X(\text{Head}) = 1$ and $X(\text{Tail}) = 0$. Then $X$ is a discrete random variable.[18]

  If the coin is fair, the p.m.f. of $X$ is $f : \mathbb{R} \to \mathbb{R}$, where

  $$f(0) = P(X = 0) = 1/2,\ f(1) = P(X = 1) = 1/2,$$
  $$f(x) = 0 \text{ for all other } x.$$

- Roll a fair die – the outcome space is $\mathcal{S} = \{1, \ldots, 6\}$. Let $X : \mathcal{S} \to \mathbb{R}$ be defined by $X(i) = i$ for $i = 1, \ldots, 6$. Then $X$ is a discrete r.v.

  If the die is fair, the p.m.f. of $X$ is $f : \mathbb{R} \to \mathbb{R}$, where

  $$f(i) = P(X = i) = 1/6, \text{ for } i = 1, \ldots, 6,$$
  $$f(x) = 0 \text{ for all other } x.$$

- For the random variable $X$ from the previous example, the c.d.f. is $F : \mathbb{R} \to \mathbb{R}$, where

  $$F(x) = P(X \le x) = \begin{cases} 0 & \text{if } x < 1 \\ i/6 & \text{if } i \le x < i + 1,\ i = 1, \ldots, 6 \\ 1 & \text{if } x \ge 6 \end{cases}$$

- For the same random variable, we can compute the probability $P(3 \le X \le 5)$ directly:

  $$\begin{aligned} P(3 \le X \le 5) &= P(X = 3) + P(X = 4) + P(X = 5) \\ &= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}, \end{aligned}$$

  or we can use the c.d.f.:

  $$P(3 \le X \le 5) = F(5) - F(3) + f(3) = \tfrac{5}{6} - \tfrac{3}{6} + \tfrac{1}{6} = \tfrac{1}{2}.$$

- The number of calls received over a specific time period, $X$, is a discrete random variable, with potential values $0, 1, 2, \ldots$.

- Consider a 5−card poker hand consisting of cards selected at random from a 52−card deck. Find the probability distribution of $X$, where $X$ indicates the number of red cards (♦ and ♥) in the hand.

  In all, there are $\binom{52}{5}$ ways to select poker hands. By construction, $X$ can take on values $x = 0, 1, 2, 3, 4, 5$.

  If $X = 0$, then none of the 5 cards in the hands are ♦ or ♥, and all of the 5 cards in the hands are ♠ or ♣. There are thus $\binom{26}{0} \cdot \binom{26}{5}$ 5−card hands that only contain black cards, and

  $$P(X = 0) = \frac{\binom{26}{0} \cdot \binom{26}{5}}{\binom{52}{5}}.$$

  In general, if $X = x, x = 0, 1, 2, 3, 4, 5$, there are $\binom{26}{x}$ ways of having $x$ ♦ or ♥ in the hand, and $\binom{26}{5-x}$ ways of having $5 - x$ ♠ and ♣ in the

18: As a convenience, we write $X = 1$ and $X = 0$.

hand, so that

$$f(x) = P(X = x) = \begin{cases} \frac{\binom{26}{x}\cdot\binom{26}{5-x}}{\binom{52}{5}}, & x = 0, 1, 2, 3, 4, 5; \\ 0 \text{ otherwise} \end{cases}$$

- Find the c.d.f. of a discrete r.v. $X$ with p.m.f. $f(x) = 0.1x$ if $x = 1, 2, 3, 4$ and $f(x) = 0$ otherwise.

  $f(x)$ is indeed a p.m.f. as $0 < f(x) \leq 1$ for all $x$ and

$$\sum_{x=1}^{4} 0.1x = 0.1(1 + 2 + 3 + 4) = 0.1\frac{4(5)}{2} = 1.$$

  Computing $F(x) = P(X \leq x)$ yields

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.1 & \text{if } 1 \leq x < 2 \\ 0.3 & \text{if } 2 \leq x < 3 \\ 0.6 & \text{if } 3 \leq x < 4 \\ 1 & \text{if } x \geq 4 \end{cases}$$

  The p.m.f. and the c.m.f. for this r.v. are shown in Figure 6.5.

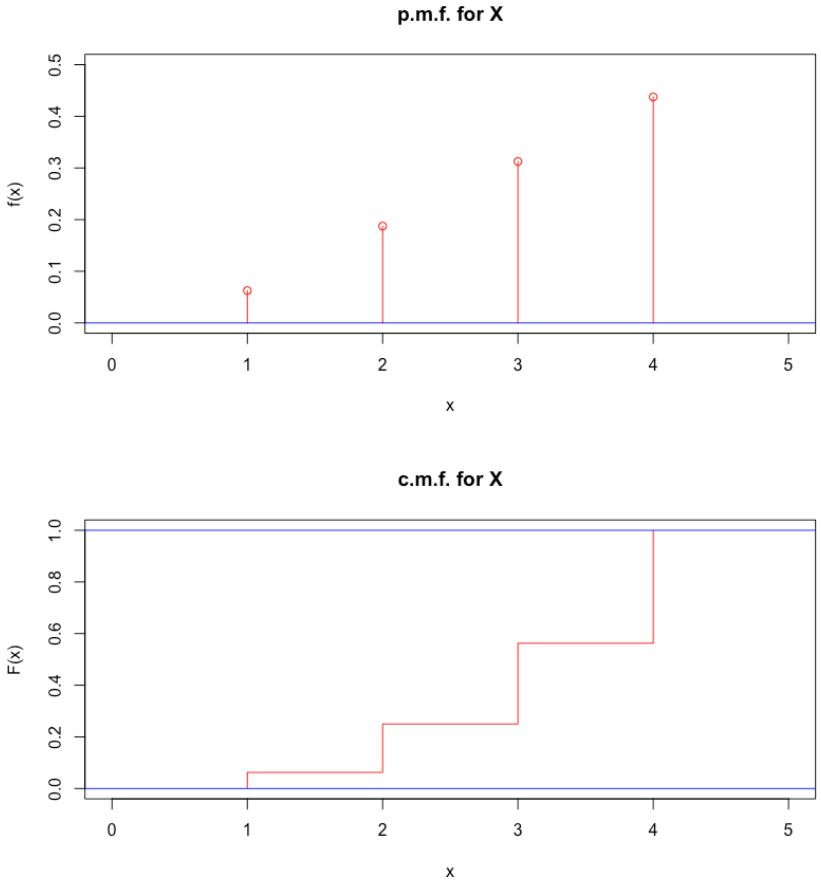



**Figure 6.5:** P.m.f. and c.m.f. for the discrete r.v. $X$ defined in the last example.

### 6.2.2 Expectation of a Discrete Random Variable

The **expectation** of a discrete random variable $X$ is

$$\mathrm{E}[X] = \sum_x x \cdot P(X = x) = \sum_x x f(x)\,,$$

where the sum extends over all values of $x$ taken by $X$.[19] The definition can be extended to a general function of $X$:

$$\mathrm{E}[u(X)] = \sum_x u(x) P(X = x) = \sum_x u(x) f(x).$$

As an important example, note that

$$\mathrm{E}[X^2] = \sum_x x^2 P(X = x) = \sum_x x^2 f(x).$$

19: The expectation of a random variable is simply the average value that it takes, over all possible values.

#### Examples

- What is the expectation on the roll $Z$ of 6−sided die?

  If the die is fair, then

  $$\mathrm{E}[Z] = \sum_{z=1}^{6} z \cdot P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} z = \frac{1}{6} \cdot \frac{6(7)}{2} = 3.5.$$

- For each 1$ bet in a gambling game, a player can win 3$ with probability $\frac{1}{3}$ and lose 1$ with probability $\frac{2}{3}$. Let $X$ be the net gain/loss from the game. Find the expected value of the game.

  $X$ takes on the value 2$ for a win and −2$ for a loss.[20] The expected value of $X$ is thus

  $$\mathrm{E}[X] = 2 \cdot \frac{1}{3} + (-2) \cdot \frac{2}{3} = -\frac{2}{3}.$$

20: That is, win/loss = outcome − bet.

- If $Z$ is the number showing on a roll of a fair 6−sided die, find $\mathrm{E}[Z^2]$ and $\mathrm{E}[(Z - 3.5)^2]$.

$$\mathrm{E}[Z^2] = \sum_z z^2 P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} z^2 = \frac{1}{6}(1^2 + \cdots + 6^2) = \frac{91}{6}$$

$$\begin{aligned}\mathrm{E}[(Z - 3.5)^2] &= \sum_{z=1}^{6} (z - 3.5)^2 \times P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} (z - 3.5)^2 \\ &= \frac{(1 - 3.5)^2 + \cdots + (6 - 3.5)^2}{6} = \frac{35}{12}.\end{aligned}$$

#### Mean and Variance

We can interpret the expectation as the average or the **mean** of $X$, which we often denote by $\mu = \mu_X$. For instance, in the example of the fair die,

$$\mu_Z = \mathrm{E}[Z] = 3.5$$

Note that in the final example, we could have written

$$\mathrm{E}[(Z-3.5)^2] = \mathrm{E}[(Z-\mathrm{E}[Z])^2].$$

This is an important quantity associated to a random variable $X$, its **variance** $\mathrm{Var}[X]$.

The variance of a discrete random variable $X$ is the **expected squared difference from the mean**:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X-\mu_X)^2] = \sum_x (x-\mu_X)^2 P(X=x)\\ &= \sum_x \left(x^2 - 2x\mu_X + \mu_X^2\right) f(x)\\ &= \sum_x x^2 f(x) - 2\mu_X \sum_x x f(x) + \mu_X^2 \sum_x f(x)\\ &= \mathrm{E}[X^2] - 2\mu_X\mu_X + \mu_X^2 \cdot 1\\ &= \mathrm{E}[X^2] - \mu_X^2.\end{aligned}$$

This is also sometimes written as $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}^2[X]$.

### Standard Deviation

The **standard deviation** of a discrete random variable $X$ is defined directly from the variance:

$$\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}\,.$$

The mean is a measure of **centrality** and it gives an idea as to where the **bulk** of a distribution is located; the variance and standard deviation provide information about the **spread** – distributions with higher variance/SD are **more spread out about the average**.

**Example** Let $X$ and $Y$ be random variables with the following p.d.f.

| $x$ | $P(X=x)$ | $y$ | $P(Y=y)$ |
|---|---|---|---|
| −2 | 1/5 | −4 | 1/5 |
| −1 | 1/5 | −2 | 1/5 |
| 0 | 1/5 | 0 | 1/5 |
| 1 | 1/5 | 2 | 1/5 |
| 2 | 1/5 | 4 | 1/5 |

We have $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ and

$$2 = \mathrm{Var}[X] < \mathrm{Var}[Y] = 8,$$

meaning that we expect both distributions to be centered at 0, but $Y$ should be more **spread-out** than $X$ (because its variance is greater, see Figure 6.6).

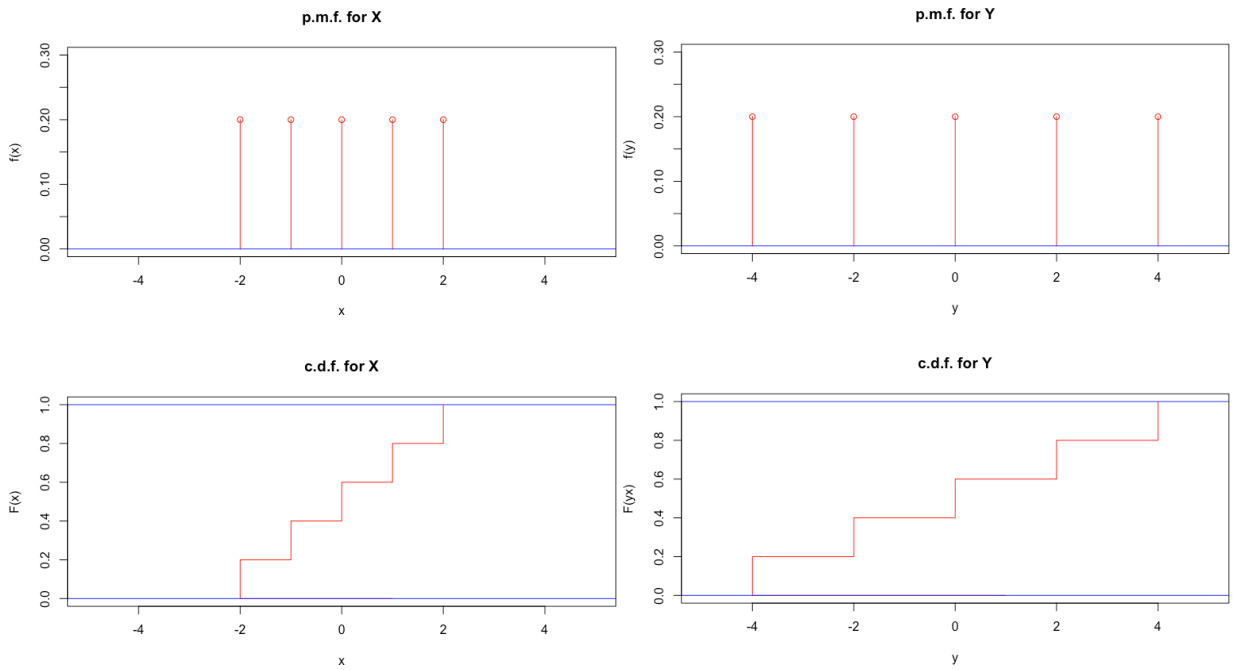



**Figure 6.6:** R.v. $X$ (left) and $Y$ (right) for two uniform distributions, as defined in the example.

#### Properties

Let $X, Y$ be random variables and $a \in \mathbb{R}$. Then

- $\mathrm{E}[aX] = a\mathrm{E}[X]$;
- $\mathrm{E}[X + a] = \mathrm{E}[X] + a$;
- $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$;
- in general, $\mathrm{E}[XY] \neq \mathrm{E}[X]\mathrm{E}[Y]$;
- $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$, $\mathrm{SD}[aX] = |a|\mathrm{SD}[X]$;
- $\mathrm{Var}[X + a] = \mathrm{Var}[X]$, $\mathrm{SD}[X + a] = \mathrm{SD}[X]$.

### 6.2.3 Binomial Distributions

Recall that the number of unordered samples of size $r$ from a set of size $n$ is

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}. $$

#### Examples

- $2! \times 4! = (1 \times 2) \times (1 \times 2 \times 3 \times 4) = 48$, but $(2 \times 4)! = 8! = 40320$.
- $\binom{5}{1} = \frac{5!}{1!\times 4!} = \frac{1\times 2\times 3\times 4\times 5}{1\times(1\times 2\times 3\times 4)} = \frac{5}{1} = 5$.
- In general: $\binom{n}{1} = n$ and $\binom{n}{0} = 1$.
- $\binom{6}{2} = \frac{6!}{2!\times 4!} = \frac{4!\times 5\times 6}{2!\times 4!} = \frac{5\times 6}{2} = 15$.
- $\binom{27}{22} = \frac{27!}{22!\times 5!} = \frac{22!\times 23\times 24\times 25\times 26\times 27}{5!\times 22!} = \frac{23\times 24\times 25\times 26\times 27}{120}$.

#### Binomial Experiments

A **Bernoulli trial** is a random experiment with two possible outcomes, “success" and”failure". Let $p$ denote the probability of a success.

A **binomial experiment** consists of $n$ repeated *independent* Bernoulli trials, each with the same probability of success, $p$, such as:

- female/male births (perahps not truly independent, but often treated as such);
- satisfactory/defective items on a production line;
- sampling with replacement with two types of item,
- etc.

### Probability Mass Function

In a binomial experiment of $n$ independent events, each with probability of success $p$, the number of successes $X$ is a discrete random variable that follows a **binomial distribution** with parameters $(n, p)$:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for } x = 0, 1, 2, \ldots, n.$$

This is often abbreviated to "$X \sim \mathcal{B}(n, p)$".

If $X \sim \mathcal{B}(1, p)$, then $P(X = 0) = 1 - p$ and $P(X = 1) = p$, so

$$\mathrm{E}[X] = (1-p) \cdot 0 + p \cdot 1 = p\,.$$

### Expectation and Variance

If $X \sim \mathcal{B}(n, p)$, it can be shown that

$$\mathrm{E}[X] = \sum_{x=0}^{n} xP(X = x) = np,$$

and

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - np)^2\right] = \sum_{x=0}^{n} (x - np)^2 \cdot P(X = x) = np(1-p)$$

(we will eventually see an easier way to derive these formulas by interpreting $X$ as a sum of discrete random variables).

Recognizing that certain situations can be modeled *via* a distribution whose p.m.f. and c.d.f. are already known can simplify computations.

#### Examples

- Suppose that water samples taken in some well-defined region have a 10% probability of being polluted. If 12 samples are selected independently, then it is reasonable to model the number $X$ of polluted samples as $\mathcal{B}(12, 0.1)$.

  Find

  1. $\mathrm{E}[X]$ and $\mathrm{Var}[X]$;
  2. $P(X = 3)$;
  3. $P(X \leq 3)$.

1. If $X \sim \mathcal{B}(n, p)$, then

$$\mathrm{E}[X] = np \quad \text{and} \quad \mathrm{Var}[X] = np(1-p).$$

With $n = 12$ and $p = 0.1$, we obtain

$$\mathrm{E}[X] = 12 \times 0.1 = 1.2;$$
$$\mathrm{Var}[X] = 12 \times 0.1 \times 0.9 = 1.08\,.$$

2. By definition,

$$P(X = 3) = \binom{12}{3}(0.1)^3(0.9)^9 \approx 0.0852.$$

3. By definition,

$$P(X \le 3) = \sum_{x=0}^{3} P(X = x)$$
$$= \sum_{x=0}^{3} \binom{12}{x}(0.1)^x(0.9)^{12-x}.$$

This sum can be computed directly, however, for $X \sim \mathcal{B}(12, 0.1)$, $P(X \le 3)$ can also be read directly from tabulated values (as in Figure 6.7):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 0 | 0.2824 | 0.0687 | 0.0138 | 0.0022 | 0.0002 | 0.0000 | | | |
| | 1 | 0.6590 | 0.2749 | 0.0850 | 0.0196 | 0.0032 | 0.0003 | 0.0000 | | |
| | 2 | 0.8891 | 0.5583 | 0.2528 | 0.0834 | 0.0193 | 0.0028 | 0.0002 | | |
| | 3 | 0.9744 | 0.7946 | 0.4925 | 0.2253 | 0.0730 | 0.0153 | 0.0017 | 0.0000 | |
| | 4 | 0.9957 | 0.9274 | 0.7237 | 0.4382 | 0.1938 | 0.0573 | 0.0095 | 0.0006 | |
| | 5 | 0.9995 | 0.9806 | 0.8822 | 0.6652 | 0.3872 | 0.1582 | 0.0386 | 0.0039 | 0.0000 |
| | 6 | 0.9999 | 0.9961 | 0.9614 | 0.8418 | 0.6128 | 0.3348 | 0.1178 | 0.0194 | 0.0005 |
| | 7 | 1.0000 | 0.9994 | 0.9905 | 0.9427 | 0.8062 | 0.5618 | 0.2763 | 0.0726 | 0.0043 |
| | 8 | | 0.9999 | 0.9983 | 0.9847 | 0.9270 | 0.7747 | 0.5075 | 0.2054 | 0.0256 |
| | 9 | | 1.0000 | 0.9998 | 0.9972 | 0.9807 | 0.9166 | 0.7472 | 0.4417 | 0.1109 |
| | 10 | | | 1.0000 | 0.9997 | 0.9968 | 0.9804 | 0.9150 | 0.7251 | 0.3410 |
| | 11 | | | | 1.0000 | 0.9998 | 0.9978 | 0.9862 | 0.9313 | 0.7176 |
| | 12 | | | | | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

**Figure 6.7:** Tabulated c.d.f. values for the binomial distribution with $n = 12$ [source unknown].

The appropriate value $\approx 0.9744$ can be found in the group corresponding to $n = 12$, in the row corresponding to $x = 3$, and in the column corresponding to $p = 0.1$. The table can also be used to compute

$$P(X = 3) = P(X \le 3) - P(X \le 2) = 0.9744 - 0.8891 \approx 0.0853.$$

- An airline sells 101 tickets for a flight with 100 seats. Each passenger with a ticket is known to have a probability $p = 0.97$ of showing up for their flight. What is the probability of 101 passengers showing up (and the airline being caught overbooking)? Make appropriate

assumptions. What if the airline sells 125 tickets?

Let $X$ be the number of passengers that show up. We want to compute $P(X > 100)$.

21: No families or late bus?

If all passengers show up independently of one another,[21] we can model $X \sim \mathcal{B}(101, 0.97)$ and

$$\begin{aligned} P(X > 100) &= P(X = 101) \\ &= \binom{101}{101}(0.97)^{101}(0.03)^0 \approx 0.046. \end{aligned}$$

If the airline sells $n = 125$ tickets, we can model the situation with the binomial distribution $\mathcal{B}(125, 0.97)$, so that

$$\begin{aligned} P(X > 100) &= 1 - P(X \leq 100) \\ &= 1 - \sum_{x=0}^{100} \binom{125}{x}(0.97)^x(0.03)^{125-x}. \end{aligned}$$

22: Do these results match your intuition?

This sum is harder to compute directly, but is very nearly 1 (try it with R, say).[22]

We can evaluate related probabilities in R *via* the base functions `rbinom()`, `dbinom()`, etc., whose parameters are `n`, `size`, and `prob`.

We can draw an observation $X$ from a binomial distribution $\mathcal{B}(11, 0.2)$ in R as follows:

```
rbinom(1, size=11, prob=0.2)
```

```
[1] 5
```

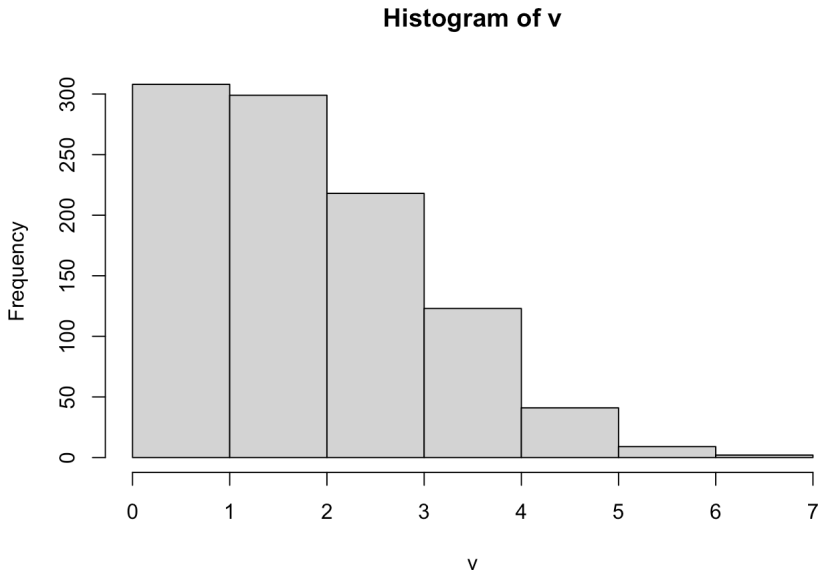
We could also replicate the process 1000 times (and extract the empirical expectation and variance):

```
v<- rbinom(1000,size=11, prob=0.2)
mean(v)
var(v)
```

```
[1] 2.236
[1] 1.794098
```

The histogram of the sample is shown below.

```
brks = min(v):max(v)
hist(v, breaks = brks)
```

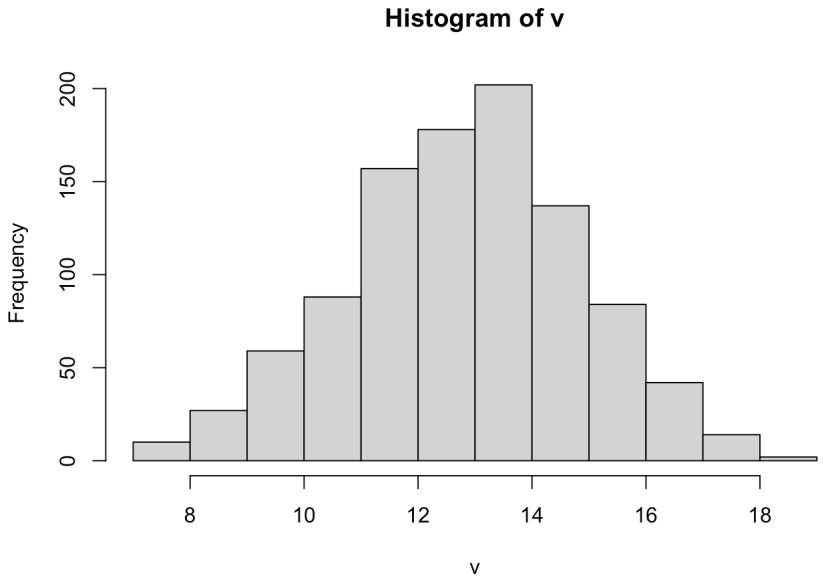



If we change the parameters of the distribution ($\mathcal{B}(19.0.7)$), we get a different looking histogram (and a different expectation and variance).

```
v<- rbinom(1000,size=19, prob=0.7)
mean(v)
var(v)
```

```
[1] 13.308
[1] 4.253389
```

```
brks = min(v):max(v)
hist(v, breaks = brks)
```



### 6.2.4 Geometric Distributions

Now consider a sequence of Bernoulli trials, with probability $p$ of success at each step. Let the **geometric** random variable $X$ denote the number of steps before the first success occurs. Its p.m.f. is given by

$$f(x) = P(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots$$

and we denote it by $X \sim \text{Geo}(p)$. For this r.v., we have

$$\text{E}[X] = \frac{1}{p} \quad \text{and} \quad \text{Var}[X] = \frac{1-p}{p^2}.$$

**Examples**

- A fair 6−sided die is thrown until it shows a 6. What is the probability that 5 throws are required?

  If 5 throws are required, we have to compute $P(X = 5)$, where $X \sim \text{Geo}(1/6)$:

  $$P(X = 5) = (1-p)^{5-1}p = (5/6)^4(1/6) \approx 0.0804.$$

- In the example above, how many throws would you expect to need?

  It's fairly simple: $\text{E}[X] = \frac{1}{1/6} = 6$.[23]

23: Understand, however, that this **does not mean** that we obtain get a 6 every 6 throws.

### 6.2.5 Negative Binomial Distributions

Consider now a sequence of Bernoulli trials, with probability $p$ of success at each step. Let the **negative binomial** random variable $X$ denote the number of steps before the $r$th success occurs. Its p.m.f. is given by

$$f(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r, \quad x = r, r+1, \ldots$$

and we denote it by $X \sim \text{NegBin}(p, r)$. For this r.v., we have

$$\text{E}[X] = \frac{r}{p} \quad \text{and} \quad \text{Var}[X] = \frac{r(1-p)}{p^2}.$$

**Examples**

- A fair 6−sided die is thrown until it three 6's are rolled. What is the probability that 5 throws are required?

  If 5 throws are required, we have to compute $P(X = 5)$, where $X \sim \text{NegBin}(1/6, 3)$:

  $$P(X = 5) = \binom{5-1}{3-1}(1-p)^{5-3}p^3 = \binom{4}{2}(5/6)^2(1/6)^3 \approx 0.0193.$$

- In the example above, how many throws would you expect to need?

  This one is also fairly simple: $\text{E}[X] = \frac{3}{1/6} = 18$.

### 6.2.6 Poisson Distributions

Let us say we are counting the number of "changes" that occur in a continuous interval of time or space.[24]

We have a **Poisson process** with rate $\lambda$, denoted by $\mathcal{P}(\lambda)$, if:

1. the number of changes occurring in non-overlapping intervals are **independent**;
2. the probability of exactly one change in a short interval of length $h$ is approximately $\lambda h$, and
3. The probability of 2+ changes in a sufficiently short interval is essentially 0.

Assume that an experiment satisfies the above properties. Let $X$ be the number of changes in a **unit interval**.[25] What is $P(X = x)$, for $x = 0, 1, \ldots$? We get to the answer by first partition the unit interval into $n$ disjoint sub-intervals of length $1/n$. Then,

1. by the second condition, the probability of one change occurring in one of the sub-intervals is approximately $\lambda/n$;
2. by the third condition, the probability of 2+ changes is $\approx 0$, and
3. by the first condition, we have a sequence of $n$ Bernoulli trials with probability $p = \lambda/n$.

Therefore,

$$\begin{aligned} f(x) = P(X = x) &\approx \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!}\cdot \underbrace{\frac{n!}{(n-x)!}\cdot\frac{1}{n^x}}_{\text{term 1}}\cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\text{term 2}}\cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\text{term 3}}. \end{aligned}$$

Letting $n \to \infty$, we obtain

$$\begin{aligned} P(X = x) &= \lim_{n\to\infty} \frac{\lambda^x}{x!}\cdot \underbrace{\frac{n!}{(n-x)!}\cdot\frac{1}{n^x}}_{\text{term 1}}\cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\text{term 2}}\cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\text{term 3}} \\ &= \frac{\lambda^x}{x!}\cdot 1\cdot \exp(-\lambda)\cdot 1 = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \ldots \end{aligned}$$

Let $X \sim \mathcal{P}(\lambda)$. Then it can be shown that

$$\mathrm{E}[X] = \lambda \quad \text{and} \quad \mathrm{Var}[X] = \lambda;$$

the mean and the variance of a Poisson random variable are **identical**!

We can compute related probabilities in `R` *via* the base functions `rpois()`, `dpois()`, etc., with required parameters `n` and `lambda`. We start by drawing a sample of size 1 from $\mathcal{P}(13)$, say, in `R` as follows:[26]

```
rpois(1,lambda=13)
```

```
[1] 18
```

24: Such as # of defects on a production line over a 1 hr period, # of customers that arrive at a teller over a 15 min interval, etc.

25: This could be 1 day, or 15 minutes, or 10 years, etc.

26: No seed has been specified, so it is conceivable that your results would be different.

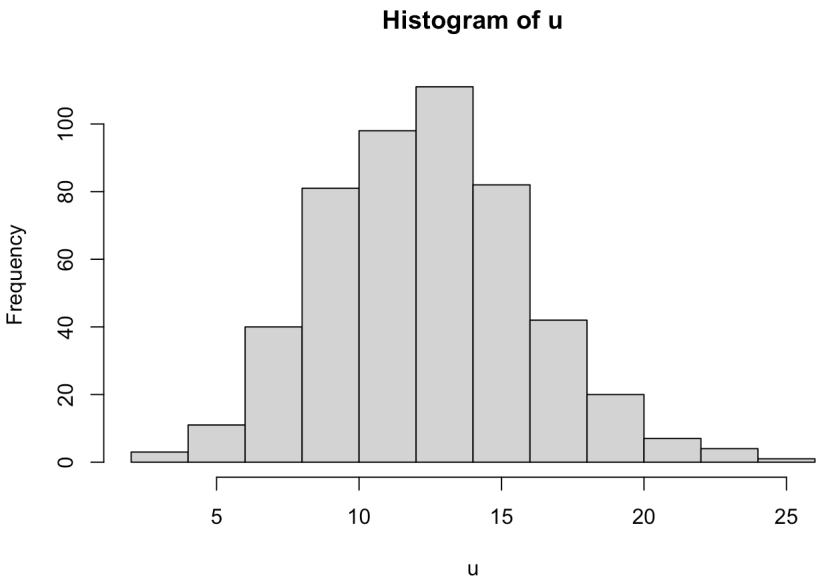
Next, we sample independently 500 times; this yields an empirical expectation and variance.

```
u<-rpois(500,lambda=13)
head(u)
mean(u)
var(u)
```

```
[1] 13 12 14 12 18  9
[1] 12.874
[1] 12.92798
```

The sample's histogram is shown below.

```
hist(u)
```




**Examples**

- A traffic flow is typically modeled by a Poisson distribution. It is known that the traffic flowing through an intersection is 6 cars/minute, on average. What is the probability of no cars entering the intersection in a 30 second period?

  Note that 6 cars/min = 3 cars/30 sec. Thus $\lambda = 3$, and we need to compute

  $$P(X = 0) = \frac{3^0 e^{-3}}{0!} = \frac{e^{-3}}{1} \approx 0.0498.$$

- A hospital needs to schedule night shifts in the maternity ward. It is known that there are 3000 deliveries per year; if these happened randomly round the clock,[27] we would expect 1000 deliveries between the hours of midnight and 8.00 a.m., a time when much of the staff is off-duty.

27: Is this a reasonable assumption?

It is thus important to ensure that the night shift is sufficiently staffed to allow the maternity ward to cope with the workload on any particular night, or at least, on a high proportion of nights.

The average number of deliveries per night

$$\lambda = 1000/365.25 \approx 2.74.$$

If the daily number $X$ of night deliveries follows a Poisson process $\mathcal{P}(\lambda)$, we can compute the probability of delivering $x = 0, 1, 2, \ldots$ babies on each night.

For a Poisson distribution, the p.m.f. values $f(x)$ are obtained *via* `dpois()` in R.[28]

We start by setup the Poisson distribution parameters and the distribution's range.[29]

```
lambda = 2.74
x=0:10
```

The p.m.f. and c.d.f. are shown below:

```
pmf=dpois(x,lambda)
cdf=ppois(x,lambda)
data.frame(x,pmf,cdf)
```

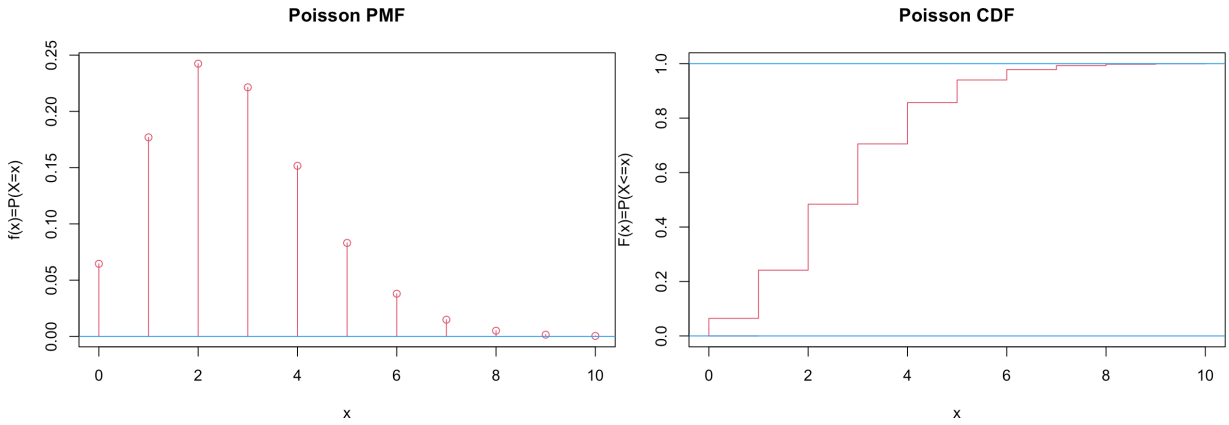
```
x   pmf         cdf
0   0.0645703  0.0645703
1   0.1769228  0.2414931
2   0.2423842  0.4838773
3   0.2213775  0.7052548
4   0.1516436  0.8568984
5   0.0831007  0.9399991
6   0.0379493  0.9779484
7   0.0148544  0.9928029
8   0.0050876  0.9978905
9   0.0015489  0.9994394
10  0.0004244  0.9998638
```

Here are the p.m.f. and c.d.f. plots:

```
plot(x,pmf, type="h", col=2, main="Poisson PMF",
     xlab="x", ylab="f(x)=P(X=x)")
points(x,pmf, col=2)
abline(h=0, col=4)

plot(c(1,x),c(0,cdf), type="s", col=2,
     main="Poisson CDF",
     xlab="x", ylab="F(x)=P(X<=x)")
abline(h=0:1, col=4)
```

28: For a general distribution, replace the `r` in the `rxxxxx(...)` random number generators by `d`: `dxxxxx(...)`.

29: In theory, it goes to infinity, but we have got to stop somewhere in practice.



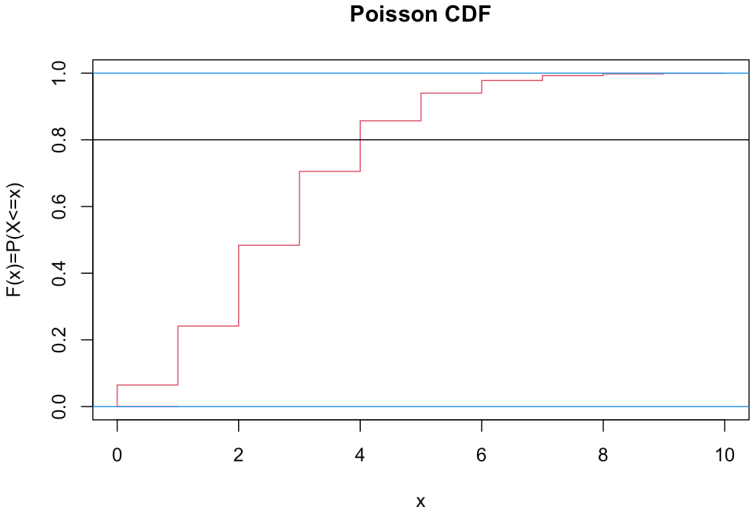

- If the maternity ward wants to prepare for the greatest possible traffic on 80% of the nights, how many deliveries should be expected?

We seek an $x$ for which

$$P(X \leq x-1) \leq 0.80 \leq P(X \leq x).$$

Let's plot the height $F(x) = 0.8$ on the c.d.f.:

```
plot(c(1,x),c(0,cdf), type="s", col=2,
    main="Poisson CDF", xlab="x", ylab="F(x)=P(X<=x)")
abline(h=0:1, col=4)
abline(h=0.8, col=1)
```




The $y = 0.8$ line crosses the CMF at $x = 4$; let's evaluate $F(3) = P(X \leq 3)$ and $F(4) = P(X \leq 4)$ to confirm that $F(3) \leq 0.8 \leq F(4)$.

```
ppois(3,lambda)
ppois(4,lambda)
```

```
[1] 0.7052548
[1] 0.8568984
```

Thus, if the hospital prepares for 4 deliveries a night, they will be ready for the worst on at least 80% of the nights.[30]

30: Note that this is different than asking how many deliveries are expected nightly (namely, $\mathrm{E}[X] = 2.74$).

- On how many nights in the year would 5 or more deliveries be expected?

  We need to evaluate

  $$365.25 \cdot P(X \geq 5) = 365.25(1 - P(X \leq 4)).$$

```
365.25*(1-ppois(4,2.74))
```

```
[1] 52.26785
```

  Thus, on roughly 14% of the nights.

- Over the course of one year, what is the greatest number of deliveries expected on any night?

  We are looking for the largest value of $x$ s.t. $365.25 \cdot P(X = x) \geq 1$.[31] The expected number of nights with each number of deliveries can be computed using:

31: If $365.25 \cdot P(X = x) < 1$, then the probability of that number of deliveries is too low to expect that we would ever see it during the year.

```
nights=c()
for(j in 0:10){
  nights[j+1]=365.25*dpois(j,lambda)
}
rbind(0:10,nights)
```

```
           [,1]     [,2]     [,3]     [,4]
       0.00000  1.00000  2.00000  3.00000
nights 23.58432 64.62103 88.53082 80.85815

           [,5]     [,6]     [,7]     [,8]
       4.00000  5.00000  6.00000 7.000000
nights 55.38783 30.35253 13.86099 5.425587

           [,9]    [,10]      [,11]
      8.000000 9.000000 10.0000000
nights 1.858264 0.565738  0.1550122
```

  The largest index is:

```
max(which(nights>1))-1
```

```
[1] 8
```

  Indeed, for larger values of $x$, we have $365.25 \cdot P(X = x) < 1$.

```
365.25*dpois(8,lambda)
365.25*dpois(9,lambda)
```

```
[1] 1.858264
[1] 0.565738
```

### 6.2.7 Other Discrete Distributions

There are numerous commonly-used discrete distributions [5]:

- the **Rademacher** distribution, which takes values 1 and −1, each with probability 1/2;
- the **beta binomial** distribution, which describes the number of successes in a series of independent Bernoulli experiments with heterogeneity in the success probability;
- the **discrete uniform** distribution, where all elements of a finite set are equally likely (balanced coin, unbiased die, first card of a well-shuffled deck, etc.);
- the **hypergeometric** distribution, which describes the number of successes in the first $m$ of a series of $n$ consecutive Bernoulli experiments, if the total number of successes is known;
- the **Poisson binomial** distribution, which describes the number of successes in a series of independent Bernoulli experiments with different success probabilities;
- **Benford's Law**, which describes the frequency of the first digit of many naturally occurring data.
- **Zipf's Law**, which describes the frequency of words in the English language;
- the **beta negative binomial** distribution, which describes the number of failures needed to obtain $r$ successes in a sequence of independent Bernoulli experiments;
- etc.

## 6.3 Continuous Distributions

How do we approach probabilities where there there are **uncountably infinitely many possible outcomes**, such as one might encounter if $X$ represents the height of an individual in the population, for instance (e.g., the outcomes reside in a continuous interval)? What is the probability that a randomly selected person is about 6 feet tall, say?

### 6.3.1 Continuous Random Variables

In the discrete case, the probability mass function $f_X(x) = P(X = x)$ was the main object of interest. In the continuous case, the analogous role is played by the **probability density function** (p.d.f.), still denoted by $f_X(x)$, but there is a major difference with discrete r.v.:

$$f_X(x) \neq P(X = x).$$

The **(cumulative) distribution function** (c.d.f.) of any such random variable $X$ is also still defined by

$$F_X(x) = P(X \leq x)\,,$$

viewed as a function of a real variable $x$; however $P(X \leq x)$ is not simply computed by adding a few terms of the form $P(X = x_i)$.

Note as well that

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x\to+\infty} F_X(x) = 1.$$

We can describe the **distribution** of the random variable $X$ *via* the following relationship between $f_X(x)$ and $F_X(x)$:[32]

32: In the continuous case, probability is simply an application of calculus!

$$f_X(x) = \frac{d}{dx} F_X(x).$$

**Area Under the Curve**

For any $a < b$, we have

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\},$$

so that

$$P(X \leq a) + P(a < X \leq b) = P(X \leq b)$$

and thus

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx$$

**Probability Density Function**

The **probability density function** (p.d.f.) of a continuous random variable $X$ is an **integrable** function $f_X : X(\mathcal{S}) \to \mathbb{R}$ such that:

- $f_X(x) > 0$ for all $x \in X(\mathcal{S})$ and $\lim_{x\to\pm\infty} f_X(x) = 0$;
- $\int_{\mathcal{S}} f_X(x)\,dx = 1$;
- for any event $A = (a, b) = \{X \mid a < X < b\}$,

$$P(A) = P((a, b)) = \int_a^b f_X(x)\,dx,$$

and the **cumulative distribution function** (c.d.f.) $F_X$ is given by

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

Unlike discrete distributions, the **endpoints** do not affect the probability computations for continuous distributions: for any $a, b$,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b),$$

all taking the value

$$F_X(b) - F_X(a) = \int_a^b f(x)\,dx.$$

Furthermore, for any $x$,

$$P(x > X) = 1 - P(X \leq x) = 1 - F_X(x) = 1 - \int_{-\infty}^{x} f_X(t)\,dt;$$

and for any $a$,

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)\,dx = 0.$$

That last result explains why it is pointless to speak of the probability of a random variable taking on a specific value in the continuous case; rather, we are interested in **ranges** of values.

### Examples

- Assume that $X$ has the following p.d.f.:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x/2 & \text{if } 0 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases}$$

Note that $\int_0^2 f(x)\,dx = 1$. The corresponding c.d.f. is given by:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$
$$= \begin{cases} 0 & \text{if } x < 0 \\ 1/2 \cdot \int_0^x t\,dt = x^2/4 & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

The p.d.f. and the c.d.f. for this r.v. are shown in Figure 6.8.

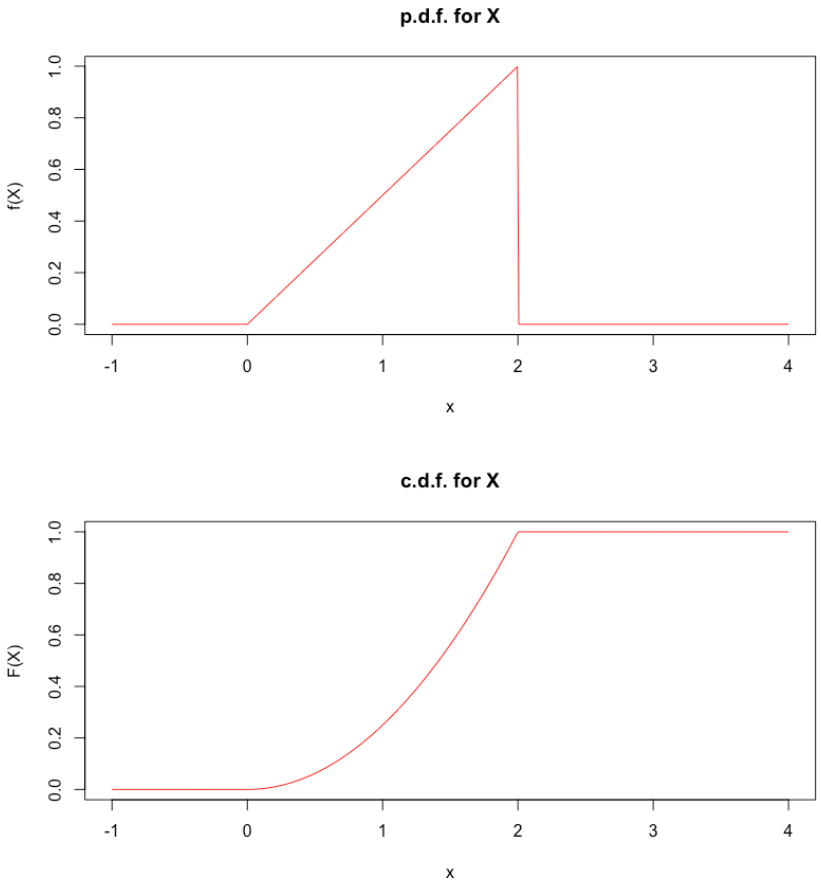



**Figure 6.8:** P.d.f. and c.d.f. for the continuous r.v. $X$ defined above.

- What is the probability of the event $A = \{X \mid 0.5 < X < 1.5\}$ if $X$ is the r.v. above?

  We need to evaluate

$$P(A) = P(0.5 < X < 1.5) = F_X(1.5) - F_X(0.5)$$
$$= \frac{(1.5)^2}{4} - \frac{(0.5)^2}{4} = \frac{1}{2}.$$

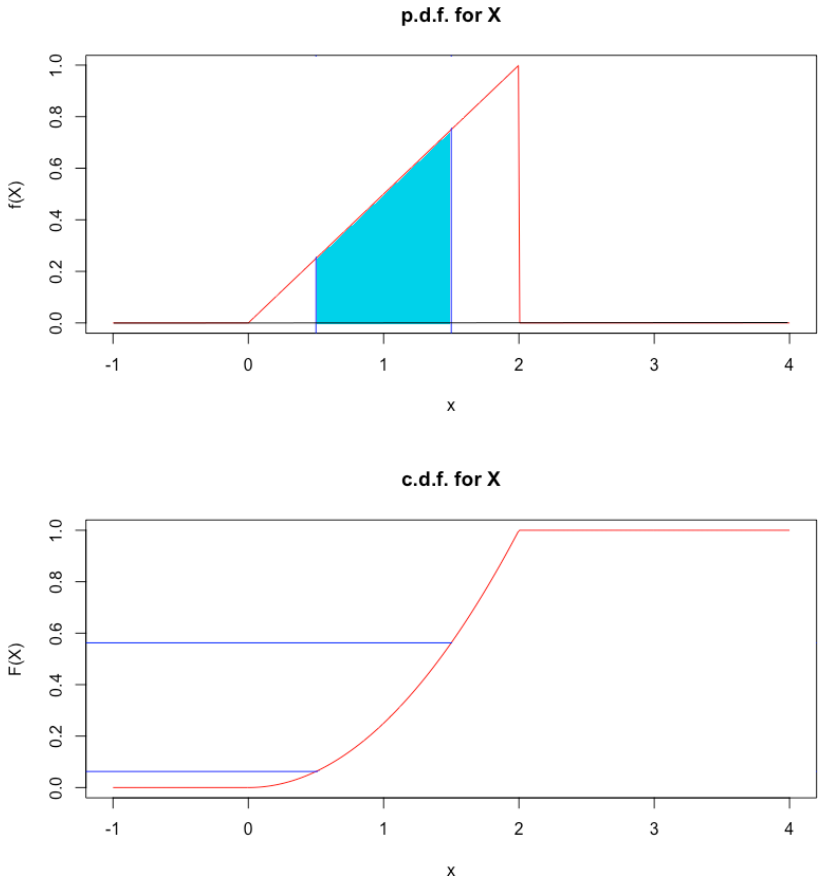



**Figure 6.9:** P.d.f. and c.d.f. for the continuous r.v. $X$ defined above, with event $A$.

- What is the probability of the event $B = \{X \mid X = 1\}$?

  We need to evaluate

$$P(B) = P(X = 1) = P(1 \leq X \leq 1) = F_X(1) - F_X(1) = 0.$$

  This is not unexpected: even though $f_X(1) = 0.5 \neq 0$, $P(X = 1) = 0$, as we saw earlier.

- Assume that, for $\lambda > 0$, $X$ has the following p.d.f.:

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

  Verify that $f_X$ is a p.d.f. for all $\lambda > 0$, and compute the probability that $X > 10.2$.

That $f_X$ is a p.d.f. is obvious; the only work goes into showing that

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} \lambda \exp(-\lambda x)\,dx = \lim_{b\to\infty} \int_0^b \lambda \exp(-\lambda x)\,dx$$
$$= \lim_{b\to\infty} \lambda \left[\frac{\exp(-\lambda x)}{-\lambda}\right]_0^b = \lim_{b\to\infty} \left[-\exp(-\lambda x)\right]_0^b$$
$$= \lim_{b\to\infty} \left[-\exp(-\lambda b) + \exp(0)\right] = 1.$$

The corresponding c.d.f. is given by:

$$F_X(x;\lambda) = P_\lambda(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 0 & \text{if } x < 0 \\ \lambda \int_0^x \exp(-\lambda t)\,dt & \text{if } x \ge 0 \end{cases}$$
$$= \begin{cases} 0 & \text{if } x < 0 \\ \left[-\exp(-\lambda t)\right]_0^x & \text{if } x \ge 0 \end{cases} = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \exp(-\lambda x) & \text{if } x \ge 0 \end{cases}$$

Then

$$P_\lambda(X > 10.2) = 1 - F_X(10.2;\lambda) = 1 - [1 - \exp(-10.2\lambda)] = \exp(-10.2\lambda)$$

is a function of the **distribution parameter** $\lambda$ itself:

| $\lambda$ | 0.002 | 0.02 | 0.2 | 2 | 20 | 200 |
|---|---|---|---|---|---|---|
| $P_\lambda(X > 10.2)$ | 0.9798 | 0.8155 | 0.13 | $1.38 \times 10^{-9}$ | $2.54 \times 10^{-89}$ | $\approx 0$ |

For $\lambda = 0.2$, for instance, the p.d.f. and c.d.f. are:

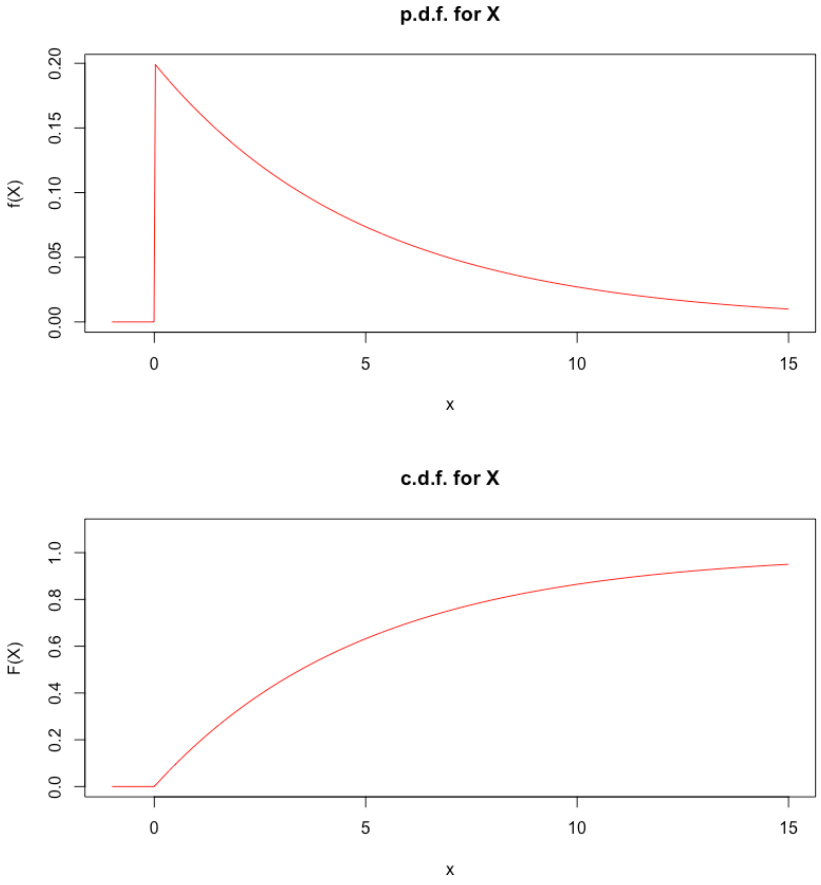



**Figure 6.10:** P.d.f. and c.d.f. for the r.v. $X$ with $\lambda = 0.2$.

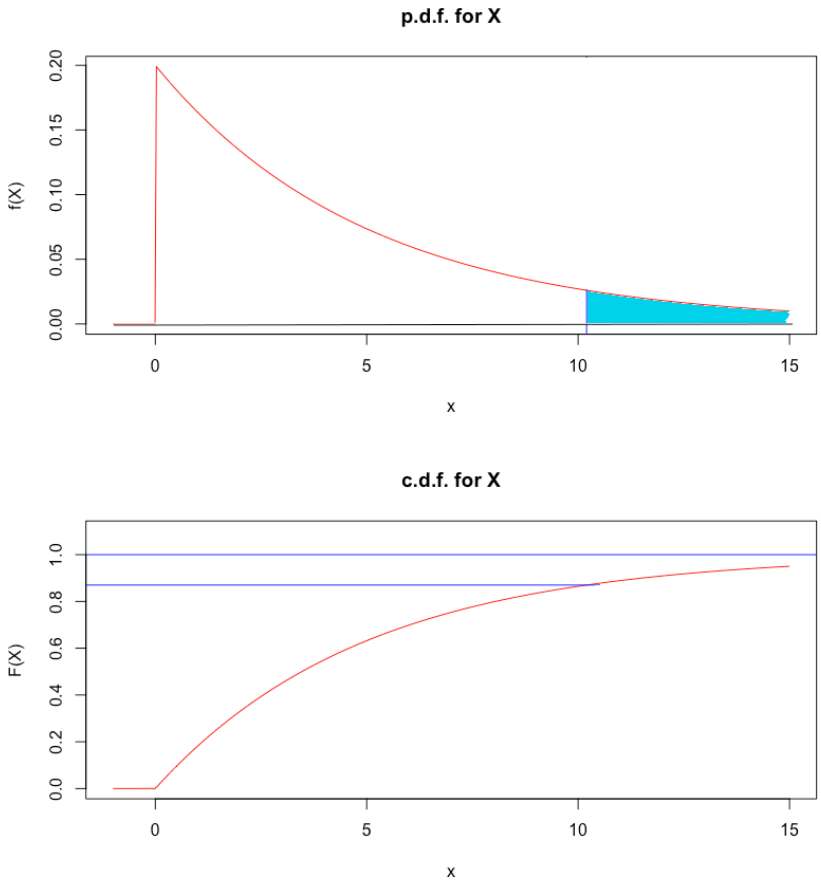



**Figure 6.11:** Probability of $X > 10.2$ (in blue), for $X$ with $\lambda = 0.2$.

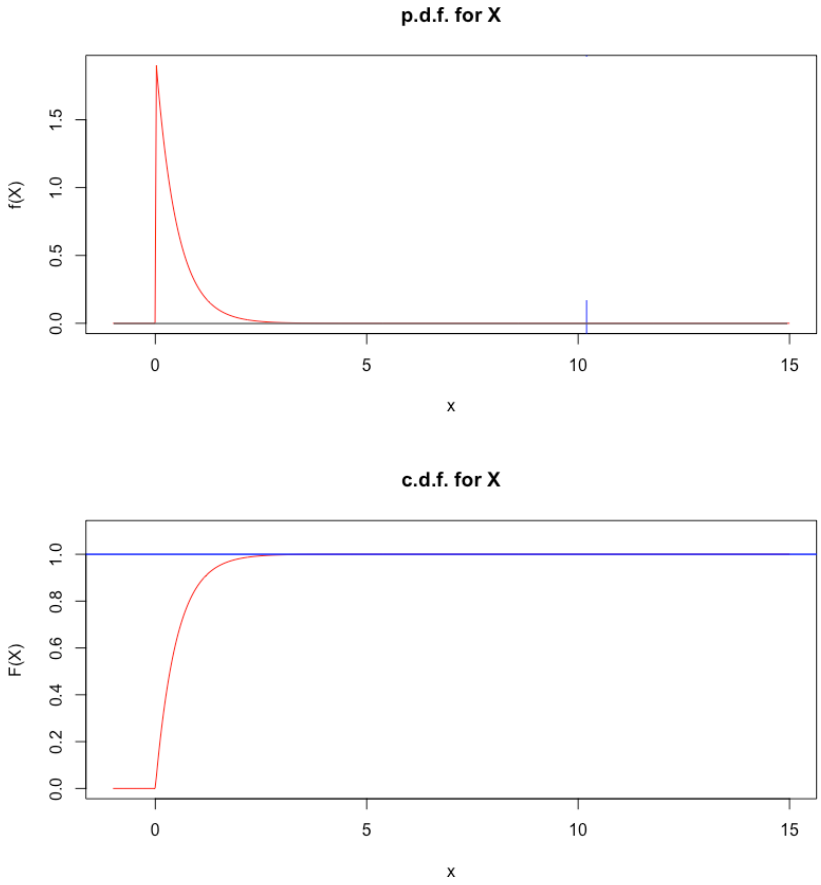



**Figure 6.12:** Probability of $X > 10.2$, for $X$ with $\lambda = 2$; the probability is so small ($1.38 \times 10^{-9}$) that it cannot even be made out in the p.d.f. (blue area).

Note that in all cases, the **shape** of the p.d.f. and the c.d.f. are the same, although the spike when $\lambda = 2$ is much higher than that when $\lambda = 0.2$ – why must that be the case?.[33]

33: This is not a general property of distributions, however, but a property of this specific family of distributions.

### 6.3.2 Expectation of a Continuous Random Variables

For a continuous random variable $X$ with p.d.f. $f_X(x)$, the **expectation** of $X$ is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx\,.$$

For any function $h(X)$, we can also define

$$\mathrm{E}\left[h(X)\right] = \int_{-\infty}^{\infty} h(x) f_X(x)\, dx\,.$$

#### Examples

- Find $\mathrm{E}[X]$ and $\mathrm{E}[X^2]$ in the first example, above.
  we need to evaluate

$$\begin{aligned}\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(X)\, dx = \int_0^2 x f_X(x)\, dx \\ &= \int_0^2 \frac{x^2}{2}\, dx = \left[\frac{x^3}{6}\right]_{x=0}^{x=2} = \frac{4}{3}; \\ \mathrm{E}[X^2] &= \int_0^2 \frac{x^3}{2}\, dx = 2.\end{aligned}$$

- Note that the **expectation need not exist**. Compute the expectation of the random variable $X$ with p.d.f.

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

  let's verify that $f_X(x)$ is indeed a p.d.f.:

$$\begin{aligned}\int_{-\infty}^{\infty} f_X(x)\, dx &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\, dx \\ &= \frac{1}{\pi}[\arctan(x)]_{-\infty}^{\infty} = \frac{1}{\pi}\left[\frac{\pi}{2} + \frac{\pi}{2}\right] = 1.\end{aligned}$$

  We can also easily see that

$$\begin{aligned}F_X(x) = P(X \le x) &= \int_{-\infty}^{x} f_X(t)\, dt \\ &= \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{1+t^2}\, dt = \frac{1}{\pi}\arctan(x) + \frac{1}{2},\end{aligned}$$

  so that $P(X \le 3) = \frac{1}{\pi}\arctan(3) + \frac{1}{2}$, say (see Figure 6.13).
  The expectation of $X$ is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\, dx.$$

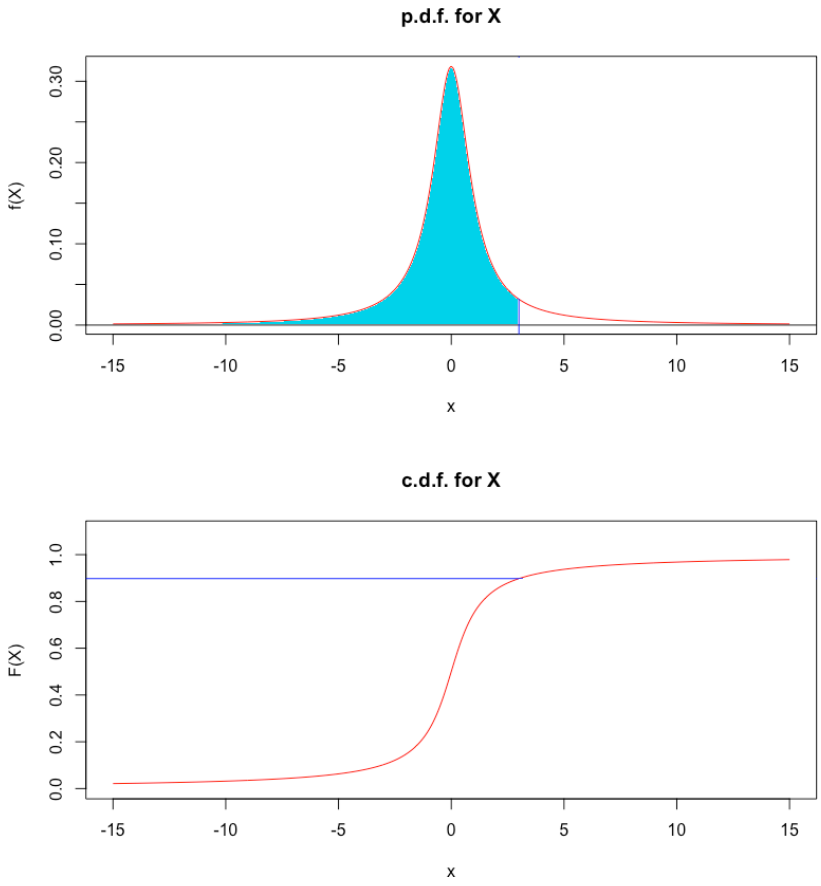



**Figure 6.13:** P.d.f. and c.d.f. for the Cauchy distribution, with area under the curve $F(3)$.

If this improper integral exists, then it needs to be equal **both** to

$$\underbrace{\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}\,dx + \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx}_{\text{candidate 1}}$$

and to the Cauchy principal value

$$\underbrace{\lim_{a\to\infty} \int_{-a}^{a} \frac{x}{\pi(1+x^2)}\,dx}_{\text{candidate 2}}.$$

But it is straightforward to find an antiderivative of $\frac{x}{\pi(1+x^2)}$. Set $u = 1 + x^2$. Then $du = 2x\,dx$ and $x\,dx = \frac{du}{2}$, and we obtain

$$\int \frac{x}{\pi(1+x^2)}\,dx = \frac{1}{2\pi}\int u\,du = \frac{1}{2\pi}\ln|u| = \frac{1}{2\pi}\ln(1+x^2).$$

Then the candidate 2 integral reduces to

$$\lim_{a\to\infty}\left[\frac{\ln(1+x^2)}{2\pi}\right]_{-a}^{a} = \lim_{a\to\infty}\left[\frac{\ln(1+a^2)}{2\pi} - \frac{\ln(1+(-a)^2)}{2\pi}\right] = \lim_{a\to\infty} 0 = 0;$$

while the candidate 1 integral reduces to

$$\left[\frac{\ln(1+x^2)}{2\pi}\right]_{-\infty}^{0} + \left[\frac{\ln(1+x^2)}{2\pi}\right]_{0}^{\infty} = 0 - (\infty) + \infty - 0 = \infty - \infty$$

which is **undefined**. Thus $\mathrm{E}[X]$ cannot not exist, as it would have to be both equal to 0 and be undefined simultaneously.[34]

34: Actually, this is not quite true: the integral for candidate 1 is undetermined of the form $\infty - \infty$; usually, when we reach this point in calculus, we have to use some other approach, such as de l'Hôpital's rule, to reduce the expression to a determinate form. The real reason why the mean does not exist is because the value of the integral for candidate 1 depends on how we approach $-\infty$ and $\infty$ for each of the constituents. For instance, if the integral exists, we should also have

$$\int_{-\infty}^{\infty} x f_X(x)\,dx = \lim_{a\to\infty} \int_{-a}^{2a} x f_X(x)\,dx.$$

In the Cauchy case, that second integral can be shown to take on the value $\ln 2/\pi$, which is different from the principal value 0; hence, the integral does not exist, which is to say, the mean of the Cauchy r.v. does not exist.

### Mean and Variance

Similarly to the discrete case, the **mean** of $X$ is defined to be $\mathrm{E}[X]$, and the **variance** and **standard deviation** of $X$ are, as before,

$$\mathrm{Var}[X] \stackrel{\text{def}}{=} \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}[X^2] - \mathrm{E}^2[X]\,,$$
$$\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}\,.$$

As in the discrete case, if $X, Y$ are continuous random variables, and $a, b \in \mathbb{R}$, then

$$\mathrm{E}[aY + bX] = a\mathrm{E}[Y] + b\mathrm{E}[X]$$
$$\mathrm{Var}[a + bX] = b^2\mathrm{Var}[X]$$
$$\mathrm{SD}[a + bX] = |b|\mathrm{SD}[X]$$

The interpretations of the mean as a measure of **centrality** and of the variance as a measure of **dispersion** still apply in the continuous case.

For the time being, however, we cannot easily compute the variance of a sum $X + Y$, unless $X$ and $Y$ are **independent** random variables:

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

## 6.3.3 Normal Distributions

A **very** important example of a continuous distribution is that provided by the special probability distribution function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,.$$

The corresponding cumulative distribution function is denoted by

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \phi(t)\,dt\,.$$

A random variable $Z$ with this c.d.f. is said to have a **standard normal distribution**, denoted by $Z \sim \mathcal{N}(0, 1)$.

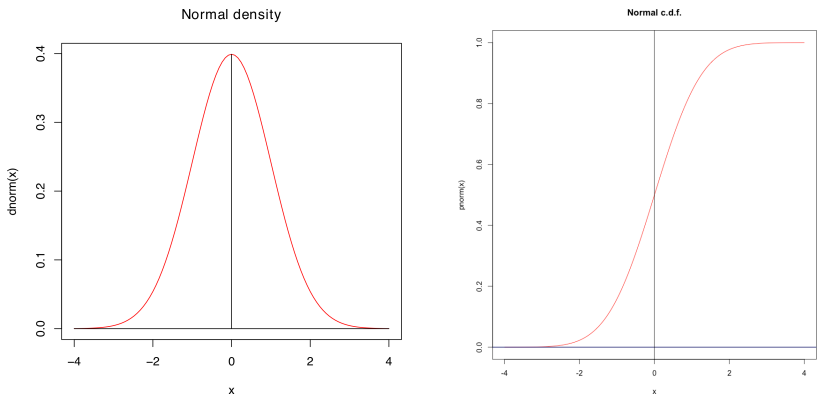



**Figure 6.14:** P.d.f. and c.d.f. for the standard normal distribution.

### Standard Normal Random Variable

The expectation and variance of $Z \sim \mathcal{N}(0,1)$ are

$$\begin{aligned}
\mathrm{E}[Z] &= \int_{-\infty}^{\infty} z\,\phi(z)\,dz = \int_{-\infty}^{\infty} z\,\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,dz = 0,\\
\mathrm{Var}[Z] &= \int_{-\infty}^{\infty} z^2\,\phi(z)\,dz = 1,\\
\mathrm{SD}[Z] &= \sqrt{\mathrm{Var}[Z]} = \sqrt{1} = 1.
\end{aligned}$$

Other quantities of interest include:

$$\begin{aligned}
&\Phi(0) = P(Z \le 0) = \frac{1}{2}, \quad \Phi(-\infty) = 0, \quad \Phi(\infty) = 1,\\
&\Phi(1) = P(Z \le 1) \approx 0.8413, \quad \text{etc.}
\end{aligned}$$

### Normal Random Variables

Let $\sigma > 0$ and $\mu \in \mathbb{R}$. If $Z \sim \mathcal{N}(0,1)$ and $X = \mu + \sigma Z$, then

$$\frac{X-\mu}{\sigma} = Z \sim \mathcal{N}(0,1).$$

Thus, the c.d.f. of $X$ is given by

$$F_X(x) = P(X \le x) = P(\mu + \sigma Z \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right);$$

its p.d.f. must then be

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}\,\phi\left(\frac{x-\mu}{\sigma}\right).$$

Any random variable $X$ with this c.d.f./p.d.f. satisfies

$$\begin{aligned}
\mathrm{E}[X] &= \mu + \sigma\mathrm{E}[Z] = \mu,\\
\mathrm{Var}[X] &= \sigma^2\mathrm{Var}[Z] = \sigma^2,\\
\mathrm{SD}[X] &= \sigma
\end{aligned}$$

and is said to be **normal with mean $\mu$ and variance $\sigma^2$**, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$. As it happens, every general normal $X$ can be obtained by a linear transformation of the standard normal $Z$.

Traditionally, probability computations for normal distributions are done with tables which compile values of the standard normal distribution c.d.f., such as the one found in [4] or at ztable.net. With the advent of freely-available **statistical software**, the need for tabulated values had decreased.[35]

In `R`, the standard normal c.d.f. $F_Z(z) = P(Z \le z)$ can be computed with the function `pnorm(z)` – for instance, `pnorm(0)=0.5`.[36]

#### Examples

- Let $Z$ represent the standard normal random variable. Then:

35: Although it would still be a good idea to learn how to read and use them.

36: In the examples that follow, whenever $P(Z \le a)$ is evaluated for some $a$, the value is found either by consulting a table or using `pnorm`.

1. $P(Z \leq 0.5) = 0.6915$
2. $P(Z < -0.3) = 0.3821$
3. $P(Z > 0.5) = 1 - P(Z \leq 0.5) = 1 - 0.6915 = 0.3085$
4. $P(0.1 < Z < 0.3) = P(Z < 0.3) - P(Z < 0.1) = 0.0781$
5. $P(-1.2 < Z < 0.3) = P(Z < 0.3) - P(Z < -1.2) = 0.5028$

- Suppose that the waiting time (in minutes) in a coffee shop at 9am is normally distributed with mean 5 and standard deviation 0.5.[37] What is the probability that the waiting time for a customer is at most 6 minutes?

37: In theory, this cannot be the true model as this would imply that some of the wait times could be negative, but it may nevertheless be an acceptable assumption in practice.

Let $X$ denote the waiting time. Then $X \sim \mathcal{N}(5, 0.5^2)$ and the **standardised random variable** is a standard normal:

$$Z = \frac{X - 5}{0.5} \sim \mathcal{N}(0, 1)\,.$$

The desired probability is

$$\begin{aligned} P(X \leq 6) &= P\left(\frac{X-5}{0.5} \leq \frac{6-5}{0.5}\right) \\ &= P\left(Z \leq \frac{6-5}{0.5}\right) = \Phi\left(\frac{6-5}{0.5}\right) \\ &= \Phi(2) = P(Z \leq 2) \approx 0.9772. \end{aligned}$$

- Suppose that bottles of beer are filled in such a way that the actual volume of the liquid content (in mL) varies randomly according to a normal distribution with $\mu = 376.1$ and $\sigma = 0.4$.[38] What is the probability that the volume in any randomly selected bottle is less than 375mL?

38: The statement from the previous sidenote applies here as well – we will assume that this is understood from this point onward.

Let $X$ denote the volume of the liquid in the bottle. Then

$$X \sim \mathcal{N}(376.1, 0.4^2) \implies Z = \frac{X - 376.1}{0.4} \sim \mathcal{N}(0, 1)\,.$$

The desired probability is thus

$$\begin{aligned} P(X < 375) &= P\left(\frac{X - 376.1}{0.4} < \frac{375 - 376.1}{0.4}\right) \\ &= P\left(Z < \frac{-1.1}{0.4}\right) \\ &= P(Z \leq -2.75) = \Phi(-2.75) \approx 0.003\,. \end{aligned}$$

- If $Z \sim \mathcal{N}(0, 1)$, for which values $a$, $b$ and $c$ do:
  1. $P(Z \leq a) = 0.95$?

     From the table (or R) we see that

     $$P(Z \leq 1.64) \approx 0.9495,\ P(Z \leq 1.65) \approx 0.9505\,.$$

     Clearly we must have $1.64 < a < 1.65$; a linear interpolation provides a decent guess at $a \approx 1.645$.

     This level of precision is usually not necessary – it is often suf-

ficient to simply present the interval estimate: $a \in (1.64, 1.65)$

2. $P(|Z| \leq b) = P(-b \leq Z \leq b) = 0.99$?

   Note that

   $$P(-b \leq Z \leq b) = P(Z \leq b) - P(Z < -b)$$

   However the p.d.f. $\phi(z)$ is **symmetric** about $z = 0$, which means that

   $$P(Z < -b) = P(Z > b) = 1 - P(Z \leq b),$$

   and so that

   $$\begin{aligned} P(-b \leq Z \leq b) &= P(Z \leq b) - [1 - P(Z \leq b)] \\ &= 2P(Z \leq b) - 1 \end{aligned}$$

   In the question, $P(-b \leq Z \leq b) = 0.99$, so that

   $$2P(Z \leq b) - 1 = 0.99 \implies P(Z \leq b) = \frac{1 + 0.99}{2} = 0.995\,.$$

   Consulting the table we see that

   $$P(Z \leq 2.57) \approx 0.9949,\ P(Z \leq 2.58) \approx 0.9951;$$

   a linear interpolation suggests that $b \approx 2.575$.

3. $P(|Z| \geq c) = 0.01$?

   Note that $\{|Z| \geq c\} = \{|Z| < c\}^c$, so we need to find $c$ such that

   $$P(|Z| < c) = 1 - P(|Z| \geq c) = 0.99.$$

   But this is equivalent to

   $$P(-c < Z < c) = P(-c \leq Z \leq c) = 0.99$$

   as $|x| < y \Leftrightarrow -y < x < y$, and $P(Z = c) = 0$ for all $c$. This problem was solved in part b); set $c \approx 2.575$.

Normally distributed numbers can be generated by `rnorm()` in `R`, which accepts three parameters: `n`, `mean`, and `sd`. The default parameter values are `mean=0` and `sd=1`.

We can draw a single number from $\mathcal{N}(0, 1)$ as follows:[39]

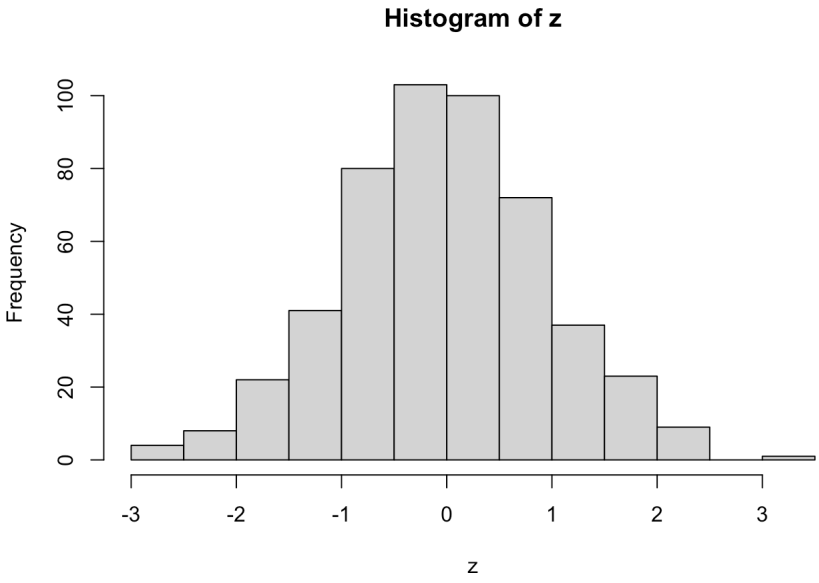
39: Note: no seed is provided, so results may vary.

```
rnorm(1)
```

```
[1] -0.2351372
```

We can generate a histogram of a sample of size 500, say, from $\mathcal{N}(0, 1)$ as follows:

```
z<-rnorm(500)
hist(z)
```




A histogram with 20 bins is shown below:

```
brks = seq(min(z),max(z),(max(z)-min(z))/20)
hist(z, breaks = brks)
```


For normal distributions with mean $\mu$ and standard deviation $\sigma$, we need to modify the call to `rnorm()`.

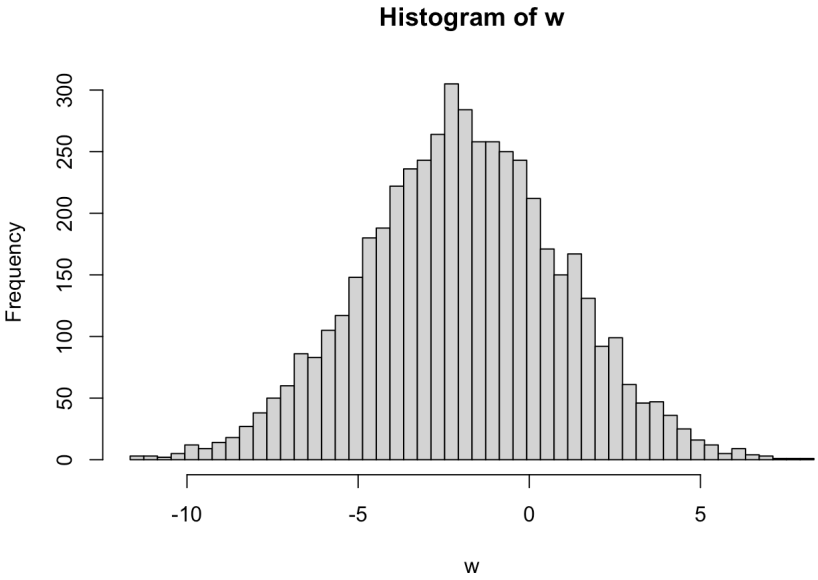
For instance, we can draw 5000 observations from $\mathcal{N}(-2, 3^2)$ using the following code:

```
w<-rnorm(5000, sd=3, mean=-2)
mean(w)
sd(w)
```

```
[1] -1.943782
[1] 2.920071
```

A histogram with 50 bins is displayed below:

```
brks = seq(min(w),max(w),(max(w)-min(w))/50)
hist(w, breaks = brks)
```




### 6.3.4 Exponential Distributions

Assume that cars arrive according to a **Poisson process with rate** $\lambda$, that is, the number of cars arriving within a fixed unit time period is a Poisson random variable with parameter $\lambda$.

Over a period of time $x$, we would then expect the number of arrivals $N$ to follow a Poisson process with parameter $\lambda x$. Let $X$ be the wait time to the first car arrival. Then

$$P(X > x) = 1 - P(X \leq x) = P(N = 0) = \exp(-\lambda x).$$

We say that $X$ follows an **exponential distribution** $\text{Exp}(\lambda)$:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } 0 \leq x \end{cases} \quad \text{and} \quad f_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \lambda e^{-\lambda x} & \text{for } 0 \leq x \end{cases}$$

Note that $f_X(x) = F_X'(x)$ for all $x$.

If $X \sim \text{Exp}(4)$, then $P(X < 0.5) = F_X(0.5) = 1 - e^{-4(0.5)} \approx 0.865$ is the area of the shaded region in Figure 6.15.

#### Properties

If $X \sim \text{Exp}(\lambda)$, then:

- $\mu = \text{E}[X] = 1/\lambda$, since

$$\mu = \int_0^\infty x\lambda e^{-\lambda x}\, dx = \left[-\frac{\lambda x + 1}{\lambda} e^{-\lambda x}\right]_0^\infty = \left[0 + \frac{\lambda(0) + 1}{\lambda} e^{-0}\right] = \frac{1}{\lambda};$$

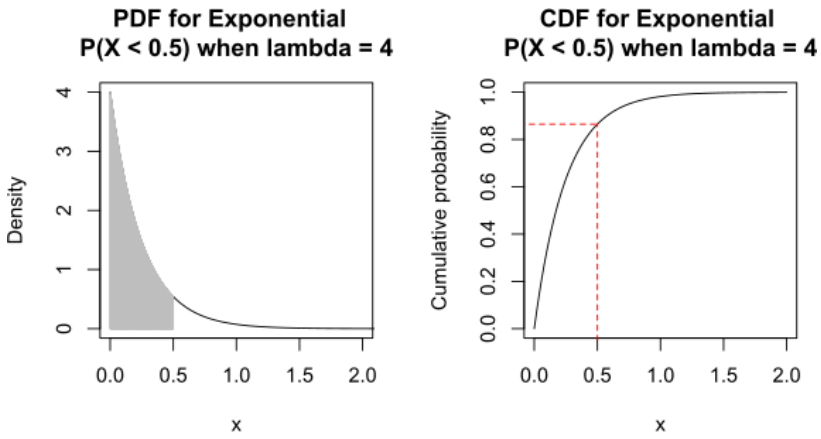



**Figure 6.15:** P.d.f. and c.d.f. for the exponential distribution. with parameter *lambda* = 4 [source unknown].

- $\sigma^2 = \text{Var}[X] = 1/\lambda^2$, since

$$\sigma^2 = \int_0^\infty (x - \text{E}[X])^2 \, \lambda e^{-\lambda x} \, dx = \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x} \, dx$$
$$= \left[-\frac{\lambda^2 x^2 + 1}{\lambda^2} e^{-\lambda x}\right]_0^\infty = \left[0 + \frac{\lambda^2 (0)^2 + 1}{\lambda^2} e^{-0}\right] = \frac{1}{\lambda^2};$$

- and $P(X > s + t \mid X > t) = P(X > s)$, for all $s, t > 0$, since

$$P(X > s + t \mid X > t) = \frac{P(X > s + t \text{ and } X > t)}{P(X > t)}$$
$$= \frac{P(X > s + t)}{P(X > t)} = \frac{1 - F_X(s + t)}{1 - F_X(t)}$$
$$= \frac{\exp(-\lambda(s + t))}{\exp(-\lambda t)}$$
$$= \exp(-\lambda s) = P(X > s).$$

Among continuous r.v., only exponential distributions satisfy this **memoryless** property; geometric distributions are the only memoryless discrete r.v., which makes, In a sense, Exp($\lambda$) the continuous counterpart of Geo($p$).

**Example** The lifetime of a certain type of light bulb follows an exponential distribution whose mean is 100 hours (i.e. $\lambda = 1/100$).

- What is the probability that a light bulb will last at least 100 hours?

  Since $X \sim \text{Exp}(1/100)$, we have

  $$P(X > 100) = 1 - P(X \leq 100) = \exp(-100/100) \approx 0.37.$$

- Given that a light bulb has already been burning for 100 hours, what is the probability that it will last at least 100 hours more?

  We seek $P(X > 200 \mid X > 100)$. By the memory-less property,

  $$P(X > 200 \mid X > 100) = P(X > 200 - 100) = P(X > 100) \approx 0.37.$$

- The manufacturer wants to guarantee that their light bulbs will last at least $t$ hours. What should $t$ be in order to ensure that 90% of

the light bulbs will last longer than $t$ hours?

We need to find $t$ such that $P(X > t) = 0.9$. In other words, we are looking for $t$ such that

$$0.9 = P(X > t) = 1 - P(X \leq t) = 1 - F_X(t) = e^{-0.01t},$$

that is,

$$\ln 0.9 = -0.01t \implies t = -100 \ln 0.9 \approx 10.5 \text{ hours}.$$

Exponentially distributed numbers are generated by `rexp()` in `R`, with required parameters `n` and `rate`.

We can draw from Exp(100) as follows:[40]

40: This is the last time we mention that these are seedless (pseudo-)random numbers.

```
rexp(1,100)
```

```
[1] 0.0009430804
```

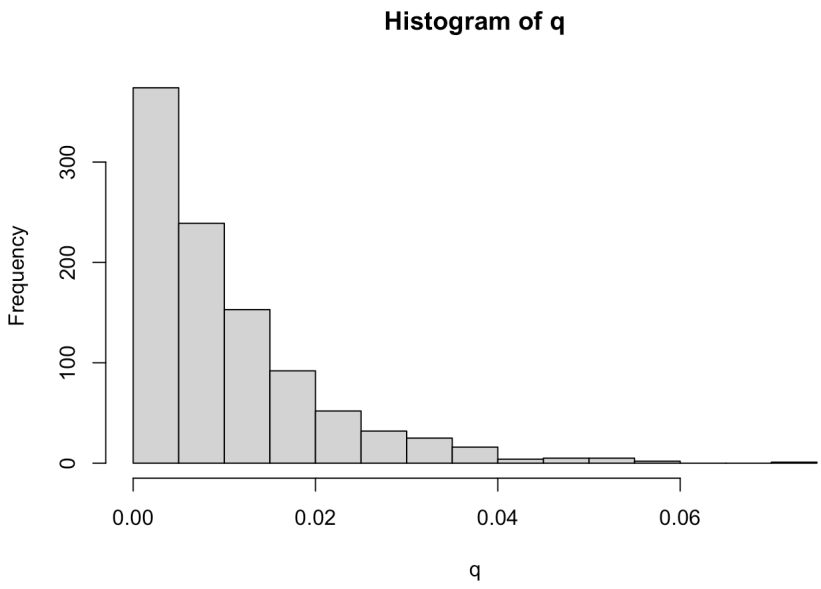
If we repeat the process 1000 times, the empirical mean and variance are:

```
q<-rexp(1000,100)
mean(q)
var(q)
```

```
[1] 0.01029523
[1] 0.000102973
```

The histogram is displayed below:

```
hist(q)
```



### 6.3.5 Gamma Distributions

Assume that cars arrive according to a Poisson process with rate $\lambda$. Recall that if $X$ is the time to the first car arrival, then $X \sim \text{Exp}(\lambda)$.

If $Y$ is the wait time to the $r$th arrival, then $Y$ follows a **Gamma distribution** with parameters $\lambda, r$, denoted $Y \sim \Gamma(\lambda, r)$, for which the p.d.f. is

$$f_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ \frac{y^{r-1}}{06(r-1)!}\lambda^r e^{-\lambda y} & \text{for } y \geq 0 \end{cases}$$

The c.d.f. $F_Y(y)$ exists – it is the area under $f_Y$ from 0 to $y$ – but it cannot be expressed with elementary functions.

We can also show that

$$\mu = \text{E}[Y] = \frac{r}{\lambda} \quad \text{and} \quad \sigma^2 = \text{Var}[Y] = \frac{r}{\lambda^2}.$$

#### Examples

- Suppose that an average of 30 customers per hour arrive at a shop in accordance with a Poisson process, that is to say, $\lambda = 1/2$ customers arrive on average every minute. What is the probability that the shopkeeper will wait more than 5 minutes before both of the first two customers arrive?

  Let $Y$ denote the wait time in minutes until the second customer arrives. Then $Y \sim \Gamma(1/2, 2)$ and

  $$P(Y > 5) = \int_5^\infty \frac{y^{2-1}}{(2-1)!}(1/2)^2 e^{-y/2}\, dy = \int_5^\infty \frac{ye^{-y/2}}{4}\, dy$$
  $$= \frac{1}{4}\left[-2ye^{-y/2} - 4e^{-y/2}\right]_5^\infty = \frac{7}{2}e^{-5/2} \approx 0.287.$$

- Telephone calls arrive at a switchboard at a mean rate of $\lambda = 2$ per minute, according to a Poisson process. Let $Y$ be the waiting time until the 5th call arrives. What is the p.d.f., the mean, and the variance of $Y$?

  We have

  $$f_Y(y) = \frac{2^5 y^4}{4!}e^{-2y}, \text{ for } 0 \leq y < \infty,$$
  $$\text{E}[Y] = \frac{5}{2}, \quad \text{Var}[Y] = \frac{5}{4}.$$

The Gamma distribution can be extended to cases where $r > 0$ is not an integer by replacing $(r-1)!$ by

$$\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}\, dt.$$

The exponential and the $\chi^2$ distributions (we will discuss the latter later) are special cases of the Gamma distribution: $\text{Exp}(\lambda) = \Gamma(\lambda, 1)$ and $\chi^2(r) = \Gamma(1/2, r)$.

Gamma distributed numbers are generated by `rgamma()`, with required parameters `n`, `shape`, and `scale`.

We can draw from a $\Gamma(2,3)$ distribution, for example, using:

```
rgamma(1,shape=2,scale=1/3)
```

```
[1] 2.249483
```

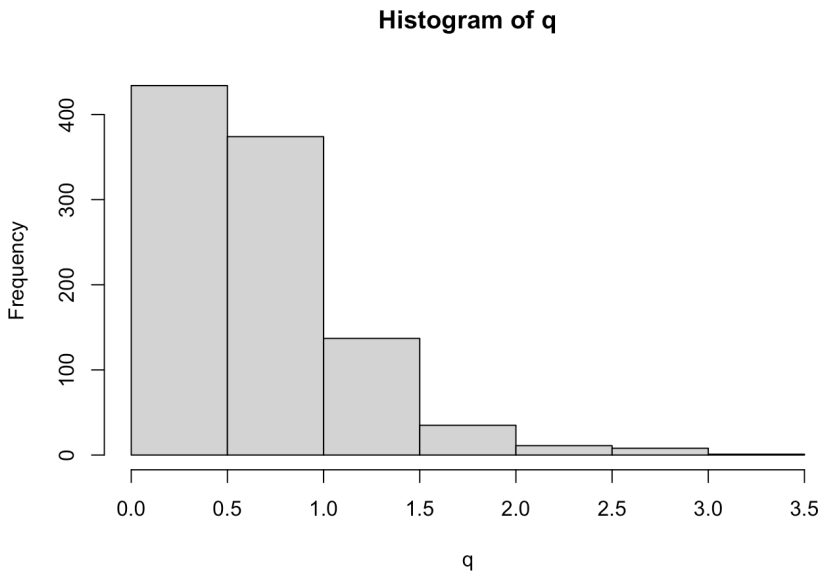
This can be repeated 1000 times, say, and we get the empirical mean and variance:

```
q<-rgamma(1000,shape=2, scale=1/3)
mean(q)
var(q)
```

```
[1] 0.6663675
[1] 0.2205931
```

The corresponding histogram is displayed below:

```
hist(q)
```




### 6.3.6 Approximation of the Binomial Distribution

If $X \sim \mathcal{B}(n,p)$ then we may interpret $X$ as a sum of **independent and identically distributed** random variables

$$X = I_1 + I_2 + \cdots + I_n \quad \text{where each} \quad I_i \sim \mathcal{B}(1,p)\,.$$

Thus, according to the **Central Limit Theorem**,[41] for large $n$ we have

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1)\,;$$

for large $n$ if $X \overset{\text{exact}}{\sim} \mathcal{B}(n,p)$ then $X \overset{\text{approx}}{\sim} \mathcal{N}(np, np(1-p))$.

41: We will have more to say on this crucial topic in Section 6.5.

### Normal Approximation with Continuity Correction

When $X \sim \mathcal{B}(n,p)$, we know that $\mathrm{E}[X] = np$ and $\mathrm{Var}[X] = np(1-p)$. If $n$ is large, we may approximate $X$ by a normal random variable in the following way:

$$P(X \le x) = P(X < x + 0.5) = P\left(Z < \frac{x - np + 0.5}{\sqrt{np(1-p)}}\right)$$

and

$$P(X \ge x) = P(X > x - 0.5) = P\left(Z > \frac{x - np - 0.5}{\sqrt{np(1-p)}}\right).$$

The **continuity correction terms** are the corresponding $\pm 0.5$ in the expressions – they are required.

**Example** Suppose $X \sim \mathcal{B}(36, 0.5)$. Provide a normal approximation to the probability $P(X \le 12)$.[42]

42: The binomial probabilities are not typically available in textbooks (or online) for $n = 36$, although they could be computed directly in R, such as with `pbinom(12,26,0.5)=0.0326`.

The expectation and the variance of a binomial r.v. are known:

$$\mathrm{E}[X] = 36(0.5) = 18 \quad \text{and} \quad \mathrm{Var}[X] = 36(0.5)(1-0.5) = 9,$$

and so

$$P(X \le 12) = P\left(\frac{X - 18}{3} \le \frac{12 - 18 + 0.5}{3}\right) \overset{\text{norm.approx'n}}{\approx} \Phi(-1.83) \overset{\text{table}}{\approx} 0.033\,.$$

### Computing Binomial Probabilities

There are thus at least four ways of computing (or approximating) binomial probabilities:

- using the **exact formula** – if $X \sim \mathcal{B}(n,p)$, then we have $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for each $x = 0, 1, \ldots, n$;
- using **tables** – if $n \le 15$ and $p$ is one of $0.1, \ldots, 0.9$, then the corresponding c.d.f. can be found in many textbooks (we must first express the desired probability in terms of the c.d.f. $P(X \le x)$), such as in
$$P(X < 3) = P(X \le 2);$$
$$P(X = 7) = P(X \le 7) - P(X \le 6)\,;$$
$$P(X > 7) = 1 - P(X \le 7);$$
$$P(X \ge 5) = 1 - P(X \le 4), \text{ etc.}$$
- using **statistical software** (`pbinom()` in R, say), and
- using the **normal approximation** when $np$ and $n(1-p)$ are both $\ge 5$:
$$P(X \le x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right);$$
$$P(X \ge x) \approx 1 - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

### 6.3.7 Other Continuous Distributions

Some other common continuous distributions are listed in [5]:

- the **Beta** distribution, a family of 2-parameter distributions with one mode and which is useful to estimate success probabilities (special cases: uniform, arcsine, PERT distributions);
- the **logit-normal** distribution on $(0, 1)$, which is used to model proportions;
- the **Kumaraswamy** distribution, which is used in simulations in lieu of the Beta distribution (as it has a closed form c.d.f.);
- the **triangular** distribution, which is typically used as a subjective description of a population for which there is only limited sample data (it is based on a knowledge of the minimum and maximum and a guess of the mode);
- the **chi-squared** distribution, which is the sum of the squares of $n$ independent normal random variables, is used in goodness-of-fit tests in statistics;
- the $F-$distribution, which is the ratio of two chi-squared random variables, used in the analysis of variance;
- the **Erlang** distribution is the distribution of the sum of $k$ independent and identically distributed exponential random variables, and it is used in queueing models (it is a special case of the Gammma distribution);
- the **Pareto** distribution, which is used to describe financial data and critical behavior;
- **Student's $T$ statistic**, which arise when estimating the mean of a normally-distributed population in situations where the sample size is small and the population's standard deviation is unknown;
- the **logistic** distribution, whose cumulative distribution function is the logistic function;
- the **log-normal** distribution, which describing variables that are the product of many small independent positive variables;
- etc.

## 6.4 Joint Distributions

Let $X, Y$ be two continuous random variables. The **joint probability distribution function** (joint p.d.f.) of $X, Y$ is a function $f(x, y)$ satisfying:

1. $f(x, y) \geq 0$, for all $x$, $y$;
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$, and
3. $P(A) = \iint_A f(x, y)\, dxdy$, where $A \subseteq \mathbb{R}^2$.

For a discrete variable, the properties are the same, except that we replace integrals by sums, and we add a property to the effect that $f(x, y) \leq 1$ for all $x, y$.

Property 3 implies that $P(A)$ is the **volume** of the solid over the region $A$ in the $xy$ plane bounded by the surface $z = f(x, y)$.

**Examples**

- Roll a pair of unbiased dice. For each of the 36 possible outcomes, let $X$ denote the smaller roll, and $Y$ the larger roll (taken from [1]).

  1. How many outcomes correspond to the event

  $$A = \{(X = 2, Y = 3)\}?$$

  The rolls $(3, 2)$ and $(2, 3)$ both give rise to event $A$.

  2. What is $P(A)$?

  There are 36 possible outcomes, so $P(A) = \frac{2}{36} \approx 0.0556$.

  3. What is the joint p.m.f. of $X, Y$?

  Only one outcome, $(X = a, Y = a)$, gives rise to the event $\{X = Y = a\}$. For every other event $\{X \neq Y\}$, two outcomes do the trick: $(X, Y)$ and $(Y, X)$. The joint p.m.f. is thus

  $$f(x, y) = \begin{cases} 1/36 & 1 \leq x = y \leq 6 \\ 2/36 & 1 \leq x < y \leq 6 \end{cases}$$

  The first property is automatically satisfied, as is the third (by construction). There are only 6 outcomes for which $X = Y$, all the remaining outcomes (of which there are 15) have $X < Y$.

  Thus,

  $$\sum_{x=1}^{6} \sum_{y=x}^{6} f(x, y) = 6 \cdot \frac{1}{36} + 15 \cdot \frac{2}{36} = 1.$$

  4. Compute $P(X = a)$ and $P(Y = b)$, for $a, b = 1, \ldots, 6$.

  For every $a = 1, \ldots, 6$, $\{X = a\}$ corresponds to the following union of events:

  $$\{X = a, Y = a\} \cup \{X = a, Y = a + 1\} \cup \cdots \cup \{X = a, Y = 6\}.$$

  These events are mutually exclusive, so that

  $$\begin{aligned} P(X = a) &= \sum_{y=a}^{6} P(\{X = a, Y = y\}) \\ &= \frac{1}{36} + \sum_{y=a+1}^{6} \frac{2}{36} = \frac{1}{36} + \frac{2(6 - a)}{36}, \quad a = 1, \ldots, 6. \end{aligned}$$

  Similarly, we get

  $$P(Y = b) = \frac{1}{36} + \frac{2(b - 6)}{36}, \; b = 1, \ldots, 6.$$

  These **marginal probabilities** can be found in the margins of the p.m.f.

  5. Compute $P(X = 3 \mid Y > 3)$, $P(Y \leq 3 \mid X \geq 4)$.

The notation suggests how to compute these **conditional probabilities**:

$$P(X = 3 \mid Y > 3) = \frac{P(X = 3 \cap Y > 3)}{P(Y > 3)}$$
$$P(Y = 3 \mid X \geq 4) = \frac{P(Y = 3 \cap X \geq 4)}{P(X \geq 4)}$$

The region corresponding to $P(Y > 3) = \frac{27}{36}$ is shaded in red (see Figure 6.16); the region corresponding to $P(X = 3) = \frac{7}{36}$ is shaded in blue. The region corresponding to

$$P(X = 3 \cap Y > 3) = \frac{6}{36}$$

is the intersection of the regions:

$$P(X = 3 \mid Y > 3) = \frac{6/36}{27/36} = \frac{6}{27} \approx 0.2222.$$

As $P(Y \leq 3 \cap X \geq 4) = 0$, $P(Y \leq 3 \mid X \geq 4) = 0$.

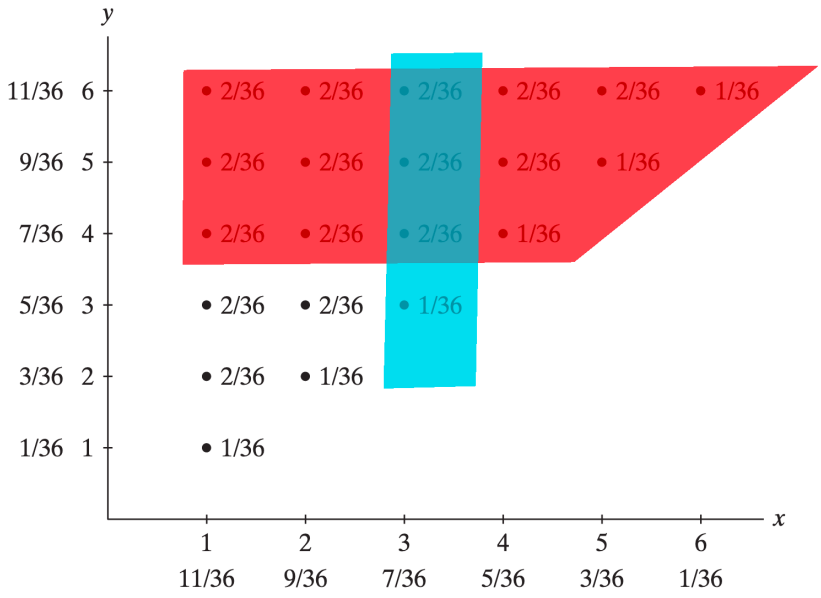



**Figure 6.16:** Conditional and marginal probabilities in the dice example [1].

6. Are $X$ and $Y$ independent?

Why didn't we simply use the multiplicative rule to compute

$$P(X = 3 \cap Y > 3) = P(X = 3)P(Y > 3)?$$

It's because $X$ and $Y$ are **not independent**, that is, it is not always the case that

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all allowable $x, y$. Indeed, $P(X = 1, Y = 1) = \frac{1}{36}$, but

$$P(X = 1)P(Y = 1) = \frac{11}{36} \cdot \frac{1}{36} \neq \frac{1}{36},$$

so $X$ and $Y$ are **dependent**.[43]

43: This is often the case when the domain of the joint p.d.f./p.m.f. is not rectangular.

- There are 8 similar chips in a bowl: three marked $(0,0)$, two marked $(1,0)$, two marked $(0,1)$ and one marked $(1,1)$. A player selects a chip at random and is given the sum of the two coordinates, in dollars (taken from [1]).

  1. What is the joint probability mass function of $X_1$, and $X_2$?

     Let $X_1$ and $X_2$ represent the coordinates; we have

     $$f(x_1, x_2) = \frac{3 - x_1 - x_2}{8}, \quad x_1, x_2 = 0, 1.$$

  2. What is the expected pay-off for this game?

     The pay-off is simply $X_1 + X_2$. The expected pay-off is thus

     $$\begin{aligned} \mathrm{E}[X_1 + X_2] &= \sum_{x_1=0}^{1} \sum_{x_2=1}^{0} (x_1 + x_2) f(x_1, x_2) \\ &= 0 \cdot \frac{3}{8} + 1 \cdot \frac{2}{8} + 1 \cdot \frac{2}{8} + 2 \cdot \frac{1}{8} \\ &= 0.75. \end{aligned}$$

- Let $X$ and $Y$ have joint p.d.f.

  $$f(x, y) = 2, \quad 0 \le y \le x \le 1.$$

  1. What is the support of $f(x, y)$?

     The support is the set $S = \{(x, y) : 0 \le y \le x \le 1\}$, a triangle in the $xy$ plane bounded by the $x-$axis, the line $y = 1$, and the line $y = x$.

     The support is the blue triangle shown in Figure 6.17.

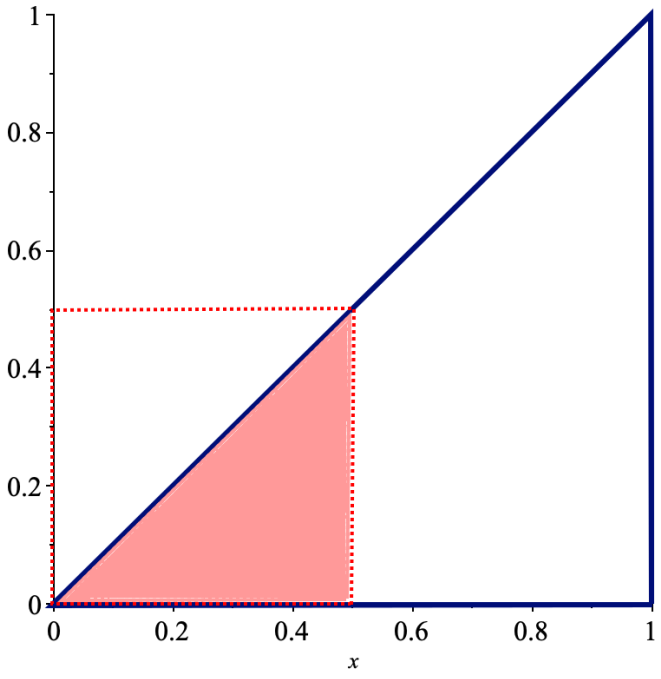



**Figure 6.17:** Support for the joint distribution of $X$ and $Y$ in the above example.

2. What is $P(0 \le X \le 0.5, 0 \le Y \le 0.5)$?

   We need to evaluate the integral over the shaded area:

$$\begin{aligned} P(0 \le X \le 0.5, 0 \le Y \le 0.5) &= P(0 \le X \le 0.5, 0 \le Y \le X) \\ &= \int_0^{0.5} \int_0^x 2\, dy dx = \int_0^{0.5} [2y]_{y=0}^{y=x}\, dx = \int_0^{0.5} 2x\, dx = 1/4. \end{aligned}$$

3. What are the marginal probabilities $P(X = x)$ and $P(Y = y)$?

   For $0 \le x \le 1$, we get

$$P(X = x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{y=0}^{y=x} 2\, dy = [2y]_{y=0}^{y=x} = 2x,$$

   and for $0 \le y \le 1$,

$$P(Y = y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{x=y}^{x=1} 2\, dx = [2x]_{x=y}^{x=1} = 2 - 2y.$$

4. Compute $\mathrm{E}[X]$, $\mathrm{E}[Y]$, $\mathrm{E}[X^2]$, $\mathrm{E}[Y^2]$, and $\mathrm{E}[XY]$.

   We have

$$\begin{aligned} \mathrm{E}[X] &= \iint_S x f(x, y)\, dA = \int_0^1 \int_0^x 2x\, dy dx \\ &= \int_0^1 [2xy]_{y=0}^{y=x}\, dx = \int_0^1 2x^2\, dx = \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}; \\ \mathrm{E}[Y] &= \iint_S y f(x, y)\, dA = \int_0^1 \int_y^1 2y\, dx dy \\ &= \int_0^1 [2xy]_{x=y}^{x=1}\, dy = \int_0^1 (2y - 2y^2)\, dy = \left[y^2 - \frac{2}{3}y^3\right]_0^1 = \frac{1}{3}; \\ \mathrm{E}[X^2] &= \iint_S x^2 f(x, y)\, dA = \int_0^1 \int_0^x 2x^2\, dy dx \\ &= \int_0^1 \left[2x^2 y\right]_{y=0}^{y=x}\, dx = \int_0^1 2x^3\, dx = \left[\frac{1}{2}x^4\right]_0^1 = \frac{1}{2}; \\ \mathrm{E}[Y^2] &= \iint_S y^2 f(x, y)\, dA = \int_0^1 \int_y^1 2y^2\, dx dy \\ &= \int_0^1 \left[2xy^2\right]_{x=y}^{x=1}\, dy = \int_0^1 (2y - 2y^3)\, dy = \left[\frac{2}{3}y^3 - \frac{1}{2}y^4\right]_0^1 = \frac{1}{6}; \\ \mathrm{E}[XY] &= \iint_S xy f(x, y) = \int_0^1 \int_0^x 2xy\, dy dx \\ &= \int_0^2 \left[xy^2\right]_{y=0}^{y=x} = \int_0^1 x^2\, dx = \left[\frac{x^4}{4}\right]_0^1 = \frac{1}{4}. \end{aligned}$$

5. Are $X$ and $Y$ independent?

   They are not, as the **support** of the joint p.d.f. is not rectangular.

The **covariance** of two random variables $X$ and $Y$ can give some indication of how they depend on one another:

$$\text{Cov}(X,Y) = \text{E}[(X - \text{E}[X])(Y - \text{E}[Y])] = \text{E}[XY] - \text{E}[X]\text{E}[Y].$$

When $X = Y$, the covariance reduces to the variance.[44] In the last example, for instance, we have: $\text{Var}[X] = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$, $\text{Var}[X] = \frac{1}{6} - (\frac{1}{3})^2 = \frac{1}{18}$, and $\text{Cov}(X,Y) = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{3} = \frac{1}{36}$.[45]

44: Note that the covariance could be negative, unlike the variance.

45: We will use the covariance again in Chapters 8 and 10.

In R, we can generate a **multivariate joint normal** *via* `MASS`'s `mvrnorm()`, whose required paramters are `n`, a mean vector `mu` and a covariance matrix `Sigma`.

We look at two standard bivariate joint normals.

```
mu1 = c(0,0); mu2 = c(-3,12)
Sigma1 = matrix(c(1,0,0,1),2,2)
Sigma2 = matrix(c(110,15,15,3),2,2)
```

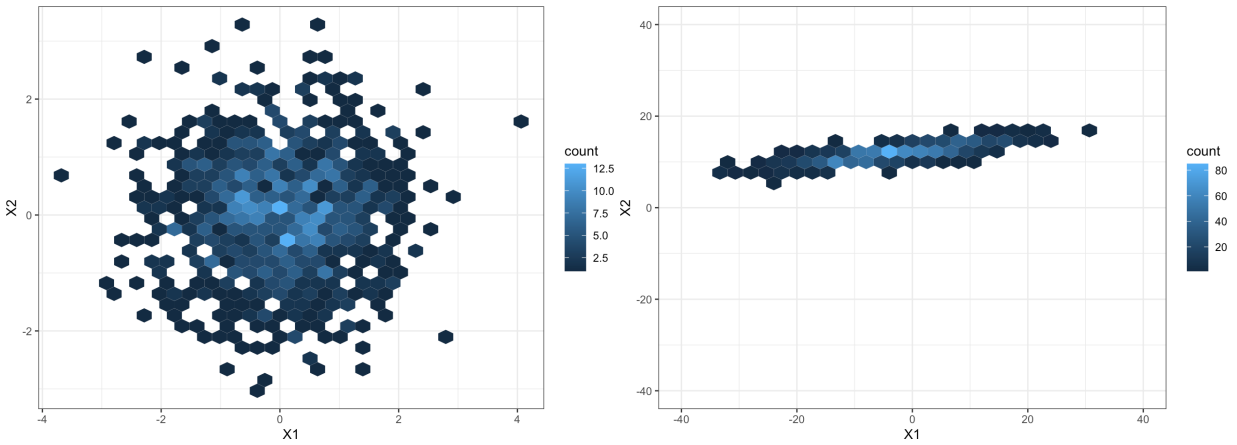
We sample 1000 observations from each joint normal.

```
library(MASS)
a1<-mvrnorm(1000,mu1,Sigma1)
a1<-data.frame(a1)
a2<-mvrnorm(1000,mu2,Sigma2)
a2<-data.frame(a2)
```

What would you expect to see when we plot the data? In the first case, the covariance matrix is the identity (**diagonal**), so we expect the blob to be circular; in the second case, we have a **non-diagonal** covariance matrix, which stretches the blob.[46]

46: The blob will have a "positive" slope since $\text{Cov}(X,Y) = 15 > 0$.

```
library(ggplot2)
library(hexbin)
qplot(X1, X2, data=a1, geom="hex")
qplot(X1, X2, data=a, geom="hex") +
    ylim(-40,40) + xlim(-40,40)
```



## 6.5 Central Limit Theorem and Sampling Distributions

In this section, we introduce one of the fundamental results of probability theory and statistical analysis.

### 6.5.1 Sampling Distributions

A **population** is a set of similar items which of interest in relation to some questions or experiments.

In some situations, it is impossible to observe the entire set of observations that make up a population – perhaps the entire population is too large to query, or some units are out-of-reach.

In these cases, we can only hope to infer the behaviour of the entire population by considering a **sample** (subset) of the population.

Suppose that $X_1, \ldots, X_n$ are $n$ **independent** random variables, each having the same c.d.f. $F$, i.e.they are **identically distributed**. Then, $\{X_1, \ldots, X_n\}$ is a **random sample** of size $n$ from the population, with c.d.f. $F$.

Any function of such a random sample is called a **statistic** of the sample; the probability distribution of a statistic is called a **sampling distribution**.

Recall the linear properties of the expectation and the variance: if $X$ is a random variable and $a, b \in \mathbb{R}$, then

$$\begin{aligned} \mathrm{E}[a + bX] &= a + b\mathrm{E}[X], \\ \mathrm{Var}[a + bX] &= b^2\mathrm{Var}[X], \\ \mathrm{SD}[a + bX] &= |b|\mathrm{SD}[X]. \end{aligned}$$

#### Sum of Independent Random Variables

For any random variables $X$ and $Y$, we have

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

In general,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + 2\mathrm{Cov}(X, Y) + \mathrm{Var}[Y];$$

if **in addition** $X$ and $Y$ are **independent**, then

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

More generally, if $X_1, X_2, \ldots, X_n$ are **independent**, then

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i] \quad \text{and} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

### Independent and Identically Distributed Random Variables

A special case of the above occurs when all of $X_1, \ldots, X_n$ have **exactly the same distribution**. In that case we say they are **independent and identically distributed**, which is traditionally abbreviated to "**iid**".

If $X_1, \ldots, X_n$ are iid, and

$$\mathrm{E}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}[X_i] = \sigma^2 \quad \text{for } i = 1, \ldots, n,$$

then

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu \quad \text{and} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2 .$$

#### Examples

- A random sample of size 100 is taken from a population with mean 50 and variance 0.25. Find the expected value and variance of the **sample total**.

  This problem translates to "if $X_1, \ldots, X_{100}$ are iid with $\mathrm{E}[X_i] = \mu = 50$ and $\mathrm{Var}[X] = \sigma^2 = 0.25$ for $i = 1, \ldots, 100$, find $\mathrm{E}[\tau]$ and $\mathrm{Var}[\tau]$ for

  $$\tau = \sum_{i=1}^{n} X_i ."$$

  According to the iid formulas,

  $$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = 100\mu = 5000, \qquad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = 100\sigma^2 = 25 .$$

- The mean value of potting mix bags weights is 5 kg, with standard deviation 0.2. If a shop assistant carries 4 bags (selected independently from the stock) then what is the expected value and standard deviation of the total weight carried?

  There is an implicit "population" of bag weights. Let $X_1, X_2, X_3, X_4$ be iid with $\mathrm{E}[X_i] = \mu = 5$, $\mathrm{SD}[X_i] = \sigma = 0.2$ and $\mathrm{Var}[X_i] = \sigma^2 = 0.2^2 = 0.04$ for $i = 1, 2, 3, 4$. Let $\tau = X_1 + X_2 + X_3 + X_4$.

  According to the iid formulas,

  $$\mathrm{E}[\tau] = n\mu = 4 \cdot 5 = 20, \qquad \mathrm{Var}[\tau] = n\sigma^2 = 4 \cdot 0.04 = 0.16.$$

  Thus, $\mathrm{SD}[\tau] = \sqrt{0.16} = 0.4$.

### Sample Mean

The **sample mean** is a typical statistic of interest:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i .$$

If $X_1, \ldots, X_n$ are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$ for all $i = 1, \ldots, n$, then

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}(n\mu) = \mu$$
$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}.$$

**Example** A set of scales returns the true weight of the object being weighed plus a random error with mean 0 and standard deviation 0.1 g. Find the standard deviation of the average of 9 measurements of an object.

Suppose the object has true weight $\mu$. The "random error" indicates that each measurement $i = 1, \ldots, 9$ is written as $X_i = \mu + Z_i$ where $\mathrm{E}[Z_i] = 0$ and $\mathrm{SD}[Z_i] = 0.1$ and the $Z_i$'s are iid.

The $X_i$'s are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{SD}[X_i] = \sigma = 0.1$. If we average $X_1, \ldots, X_n$ (with $n = 9$) to get $\overline{X}$, then

$$\mathrm{E}\left[\overline{X}\right] = \mu \quad \text{and} \quad \mathrm{SD}\left[\overline{X}\right] = \tfrac{\sigma}{\sqrt{n}} = \tfrac{0.1}{\sqrt{9}} = \tfrac{1}{30} \approx 0.033.$$

We do not need to know the **actual** distribution of the $X_i$; only $\mu$ and $\sigma^2$ are required to compute $\mathrm{E}[\overline{X}]$ and $\mathrm{Var}[\overline{X}]$.

### Sum of Independent Normal Random Variables

Another interesting case occurs when we have **multiple independent normal** random variables on the same experiment.

Suppose $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ for $i = 1, \ldots, n$, and all the $X_i$ are independent. We already know that

$$\mathrm{E}[\tau] = \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = \mu_1 + \cdots + \mu_n;$$
$$\mathrm{Var}[\tau] = \mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = \sigma_1^2 + \cdots + \sigma_n^2.$$

It turns out that, under these hypotheses, $\tau$ is **also normally distributed**, i.e.

$$\tau = \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mathrm{E}[\tau], \mathrm{Var}[\tau]) = \mathcal{N}\left(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2\right).$$

Thus, if $\{X_1, \ldots, X_n\}$ is a random sample from a normal population **with mean $\mu$ and variance $\sigma^2$**, then $\sum_{i=1}^{n} X_i$ and $\overline{X}$ are also normal, which, combined with the above work, means that

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right) \quad \text{and} \quad \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

**Example** Suppose that the population of students' weights is normal with mean 75 kg and standard deviation 5 kg. If 16 students are picked at random, what is the distribution of the (random) total weight $\tau$? What is the probability that the total weight exceeds 1250 kg?

If $X_1, \ldots, X_{16}$ are iid as $\mathcal{N}(75, 25)$, then the sum $\tau = X_1 + \cdots + X_{16}$ is also normally distributed with

$$\tau = \sum_{i=1}^{16} X_i \sim \mathcal{N}(16 \cdot 75, 16 \cdot 25) = \mathcal{N}(1200, 400), \quad \text{and}$$

$$Z = \frac{\tau - 1200}{\sqrt{400}} \sim \mathcal{N}(0, 1).$$

Thus,

$$P(\tau > 1250) = P\left(\frac{\tau - 1200}{\sqrt{400}} > \frac{1250 - 1200}{20}\right) = P(Z > 2.5) = 1 - P(Z \leq 2.5)$$
$$\approx 1 - 0.9938 = 0.0062\,.$$

### 6.5.2 Central Limit Theorem

Suppose that a professor has been teaching a course for the last 20 years. For every cohort during that period, the mid-term exam grades of all the students have been recorded. Let $X_{i,j}$ be the grade of student $i$ in year $j$. Looking back on the class lists, they find that

$$\mathrm{E}[X_{i,j}] = 56 \quad \text{and} \quad \mathrm{SD}[X_{i,j}] = 11.$$

This year, there are 49 students in the class. What should the professor expect for the class mid-term exam average?

Of course, the professor cannot predict any of the student grades or the class average with absolute certainty, but they could try the following approach:

1. simulate the results of the class of 49 students by generating sample grades $X_{1,1}, \ldots, X_{1,49}$ from a **normal** distribution $\mathcal{N}(65, 15^2)$;
2. compute the sample mean for the sample and record it as $\overline{X}_1$;
3. repeat steps 1-2 $m$ times and compute the standard deviation of the sample means $\overline{X}_1, \ldots, \overline{X}_m$;
4. plot the histogram of the sample means $\overline{X}_1, \ldots, \overline{X}_m$.

What do you think is going to happen?

**Central Limit Theorem:** if $\overline{X}$ is the mean of a random sample of size $n$ taken from a population with mean $\mu$ and finite variance $\sigma^2$, then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

as $n \to \infty$. More precisely, this is a **limiting** result. If we view the **standardization**

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

as functions of $n$, we have, for each $z$,

$$\lim_{n\to\infty} P\left(Z_n \leq z\right) = \Phi(z) \quad \text{and} \quad P\left(Z_n \leq z\right) \approx \Phi(z), \text{ if } n \text{ is large enough,}$$

**whether the original $X_i$'s are normal or not**.

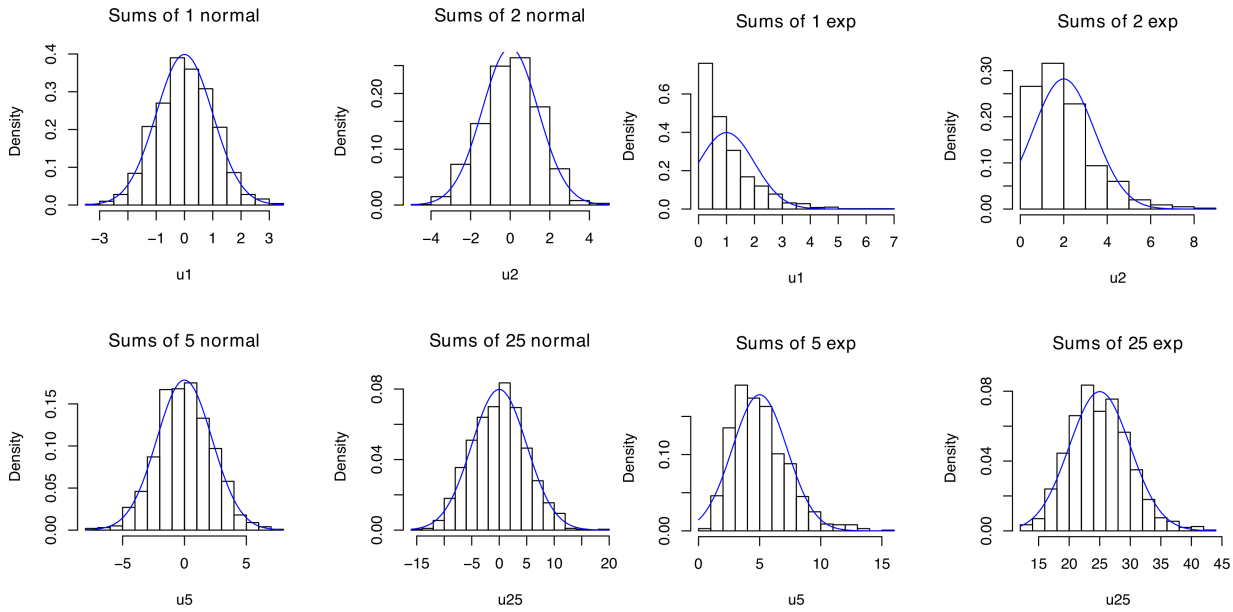



**Figure 6.18:** Illustration of the central limit theorem with a normal underlying distribution and with an exponential underlying distribution [source unknown].

### Examples

- The examination scores in an university course have mean 56 and standard deviation 11. In a class of 49 students, what is the probability that the average mark is below 50? What is the probability that the average mark lies between 50 and 60?

  Let the marks be $X_1, \ldots, X_{49}$ and assume the performances are independent. According to the central limit theorem,

  $$\overline{X} = (X_1 + X_2 + \cdots + X_{49})/49,$$

  with $\mathrm{E}[\overline{X}] = 56$ and $\mathrm{Var}[\overline{X}] = 11^2/49$. We thus have

  $$P(\overline{X} < 50) \approx P\left(Z < \frac{50-56}{11/7}\right) = P(Z < -3.82) = 0.0001$$

  and

  $$P(50 < \overline{X} < 60) \approx P\left(\frac{50-56}{11/7} < Z < \frac{60-56}{11/7}\right)$$
  $$= P(-3.82 < Z < 2.55) = \Phi(2.55) - \Phi(-3.82) = 0.9945.$$

  Note that this says nothing about whether the scores are normally distributed or not, only that the average scores follow an approximate normal distribution.[47]

- Systolic blood pressure readings for pre-menopausal, non-pregnant women aged $35-40$ have mean 122.6 standard deviation 11 mm Hg. An independent sample of 25 women is drawn from this target population and their blood pressure is recorded. What is the probability that the average blood pressure is greater than 125 mm Hg? How would the answer change if the sample size increases to 40?

47: If the scores did arise from a normal distribution, the $\approx$ would be replaced by a $=$.

According to the CLT, $\overline{X} \sim \mathcal{N}(122.6, 121/25)$, approximately. Thus

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{25}}\right) = P(Z > 1.09) = 1 - \Phi(1.09) = 0.14.$$

However, if the sample size is 40, then

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{40}}\right) = 0.08.$$

Increasing the sample size reduces the probability that the average is far from the expectation of each original measurement.

- Suppose that we select a random sample $X_1, \ldots, X_{100}$ from a population with mean 5 and variance 0.01. What is the probability that the difference between the sample mean of the random sample and the mean of the population exceeds 0.027?

  According to the CLT, we know that, approximately, $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has standard normal distribution. The desired probability is thus

$$\begin{aligned} P &= P(|\overline{X} - \mu| \geq 0.027) \\ &= P(\overline{X} - \mu \geq 0.027 \text{ or } \mu - \overline{X} \geq 0.027) \\ &= P\left(\frac{\overline{X} - 5}{0.1/\sqrt{100}} \geq \frac{0.027}{0.1/\sqrt{100}}\right) + P\left(\frac{\overline{X} - 5}{0.1/\sqrt{100}} \leq \frac{-0.027}{0.1/\sqrt{100}}\right) \\ &\approx P\left(Z \geq 2.7\right) + P\left(Z \leq -2.7\right) \\ &= 2P\left(Z \geq 2.7\right) \approx 2(0.0035) = 0.007. \end{aligned}$$

In the next example, we illustrate how to use the CLT with R.

**Example** A large freight elevator can transport a maximum of 9800 lbs. Suppose a load containing 49 boxes must be transported. From experience, the weight of boxes follows a distribution with mean $\mu = 205$ lbs and standard deviation $\sigma = 15$ lbs. Estimate the probability that all 49 boxes can be safely loaded onto the freight elevator and transported.

We are given $n = 49$, $\mu = 205$, and $\sigma = 15$. Let us further assume that the boxes all come from different sources, which is to say, the boxes' weight $x_i$, $i = 1, \ldots, 49$, are independent of one another.

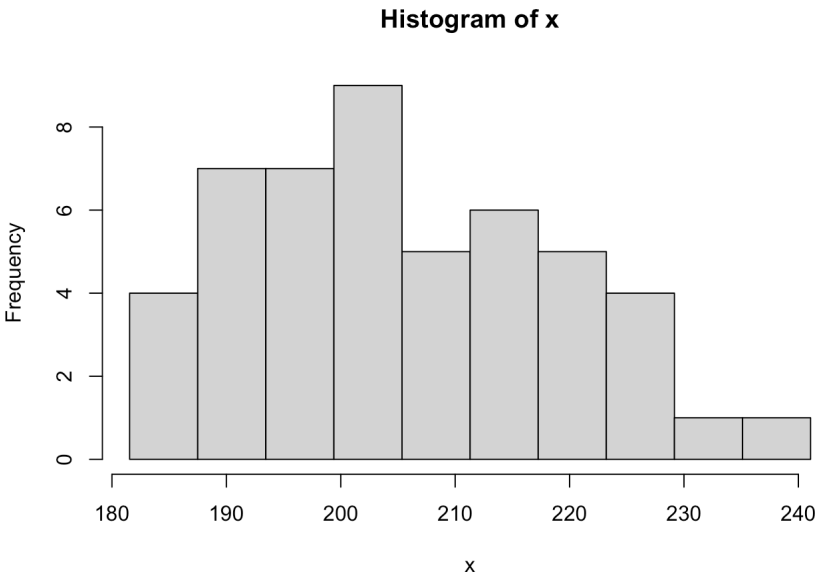
To get a sense of the task's feasibility, we simulate a few scenarios. Note that the problem makes no mention of the type of distribution that the weights follow.

To start, we assume that the weights are normally distributed.

```
set.seed(0) # to ensure replicability
x<-rnorm(49,mean=205,sd=15)
```

The histogram shows a distribution which is roughly normal.

```
brks = seq(min(x),max(x),(max(x)-min(x))/10)
hist(x, breaks = brks)
```




The elevator can transport up to 9800 lbs; the $n = 49$ boxes can be transported if their total weight

$$T = 49w = x_1 + \cdots + x_{49},$$

where $w = \overline{x}$, is less than 9800 lbs. In mathematical terms, we are interested in the value of the probability $P(T < 9800)$.

For the sample x from above, we get:

```
(T<-sum(x))
```

```
[1] 10066.36
```

That specific group of 49 boxes would be too heavy to carry in one trip. But perhaps we were simply unlucky – perhaps another group of boxes would have been light enough. Let us try again, but with a different group of boxes.

```
set.seed(999)
(T=sum(rnorm(49,mean=205,sd=15)))
```

```
[1] 9852.269
```

It's closer, but still no cigar. However, two tries are not enough to establish a trend and to estimate $P(T < 9800)$.

Next, we write a little function to help us find an estimate of the probability. The idea is simple: if we were to try a large number of random combinations of 49 boxes, the proportion of the attempts for which the total weight $T$ falls below 9800 is (hopefully?) going to approximate $P(T < 9800)$.

```
estimate_T.normal <- function(n, T.threshold, mean, sd, num.tries){
  a=0
  for(j in 1:num.tries){
    if(sum(rnorm(n,mean=mean,sd=sd))<T.threshold){
      a=a+1
    }
  }
  estimate_T.normal <- a/num.tries
}
```

What kind of inputs are these meant to be? What does this code do? Note that running this cell will **compile** the function `estimate_T.normal()`, but that it still needs to be called with appropriate inputs to provide an estimate for $P(T < 9800)$.

We try the experiment (`num.tries`) 10, 100, 1000, 10000, 100000, and 1000000 times, with `n=49`, `T.threshold=9800`, `mu=205`, and `sigma=15`.

```
(c(estimate_T.normal(49,9800,205,15,10),
estimate_T.normal(49,9800,205,15,100),
estimate_T.normal(49,9800,205,15,1000),
estimate_T.normal(49,9800,205,15,10000),
estimate_T.normal(49,9800,205,15,100000),
estimate_T.normal(49,9800,205,15,1000000)))
```

```
[1] 0.00000 0.01000 0.00700 0.00990 0.00973 0.00975
```

We cannot say too much from such a simple set up, but it certainly seems as though we should expect success about 1% of the time.

That is a low probability, which suggests that 49 may be too many boxes for the elevator to work correctly, in general, but perhaps that is only the case because we assumed normality. What happens if we used other distributions with the same characteristics, such as $U(179.02, 230.98)$ or $\Lambda(5.32, 0.0054)$?[48]

48: How would we verify that these distributions indeed have the right characteristics? How would we determine the appropriate parameters in the first place?

Let us write new functions `estimate_T.unif()` and `estimate_T.lnormf()` to repeat the previous work with those two distributions.

```
estimate_T.unif <- function(n, T.threshold, min, max, num.tries){
  a=0
  for(j in 1:num.tries){
    if(sum(runif(n,min=min,max=max))<T.threshold){
      a=a+1
    }
  }
estimate_T.unif <- a/num.tries
}

estimate_T.lnorm <- function(n, T.threshold, meanlog, sdlog, num.tries){
  a=0
  for(j in 1:num.tries){
    if(sum(rlnorm(n,meanlog=meanlog,sdlog=sdlog))<T.threshold){
```

```
      a=a+1
    }
  }
estimate_T.lnorm <- a/num.tries
}
```

For the uniform distribution, we obtain:

```
(c(estimate_T.unif(49,9800,179.02,230.98,10),
 estimate_T.unif(49,9800,179.02,230.98,100),
 estimate_T.unif(49,9800,179.02,230.98,1000),
 estimate_T.unif(49,9800,179.02,230.98,10000),
 estimate_T.unif(49,9800,179.02,230.98,100000),
 estimate_T.unif(49,9800,179.02,230.98,1000000)))
```

```
[1] 0.000000 0.010000 0.008000 0.007900 0.010230 0.009613
```

For the log-normal distribution, we obtain:

```
(c(estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),10),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),100),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),1000),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),10000),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),100000),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),1000000)))
```

```
[1] 0.000000 0.000000 0.006000 0.009500 0.009060 0.009184
```

Under all three distributions, it appears as though $P(T < 9800)$ converges to a value near 1%, even though the three distributions are very different. That might be surprising at first glance, but it is really a consequence of the **Central Limit Theorem**.

We are estimating $P(T < 9800) = P(w < 9800/49) = P(w < 200)$, where $w$ is the mean weight of the boxes.

According to the CLT, the distribution of $w$ is approximately normal with mean $\mu = 205$ and variance $\sigma^2/n = 15^2/49$, even if the weights themselves were not normally distributed.

By subtracting the mean of $w$ and dividing by the standard deviation we obtain a new random variable $z$ which is approximately the standard unit normal, i.e.

$$P(w < 200) \approx P\left(z < \frac{200 - 205}{15/7}\right).$$

But

```
(200-205)/(15/7)
```

```
[1] -2.333333
```

Thus, $P(w < 200) \approx P(z < -2.33)$ and we need to find the probability that the standard normal p.d.f. is smaller than $-2.33$.

This can be calculated with the `pnorm()` function:

```
pnorm(-2.33, mean=0, sd=1)
```

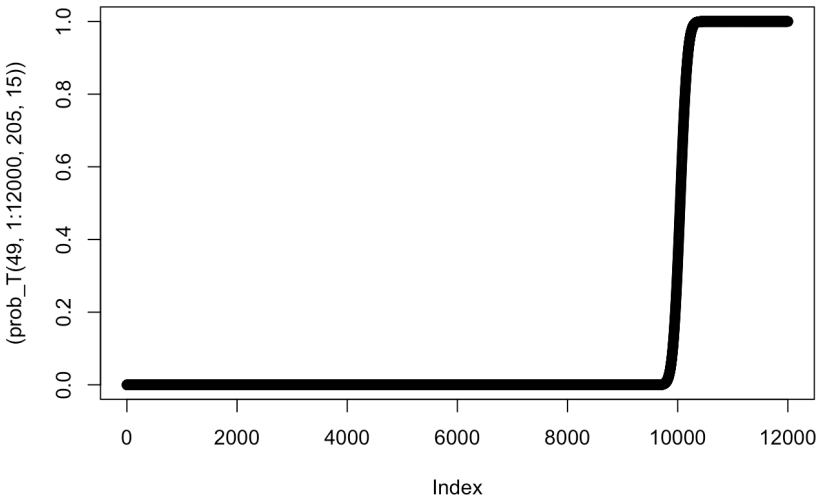
```
[1] 0.009903076
```

Hence, $P(T < 9800) \approx 0.0099$, which means that it is highly unlikely that the 49 boxes can be transported in the elevator all at once.

**Example** What elevator threshold would be required to reach a probability of success of 10%? 50%? 75%?

The following routine approximates the probability in question without resorting to simulating the weights (that is, independently of the underlying distribution of weights) for given `n`, `threshold`, `mean`, and `sd`. Can you figure out what `pnorm()` is doing?

```
prob_T <- function(n,threshold,mean,sd){
  prob_T=pnorm((threshold/n - mean)/(sd/sqrt(n)),0,1)
}

plot((prob_T(49,1:12000,205,15)))
```




We can find the desired thresholds by calling:

```
max(which(prob_T(49,1:12000,205,15)<0.1))
max(which(prob_T(49,1:12000,205,15)<0.5))
max(which(prob_T(49,1:12000,205,15)<0.75))
```

```
[1] 9910
[1] 10044
[1] 10115
```

### 6.5.3 Sampling Distributions (Reprise)

We now revisit sampling distributions in a some specific contexts.

#### Difference Between Two Means

Statisticians are often interested in the difference between various populations; a result akin to the CLT provides guidance in that area.

**Theorem:** let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean $\mu_1$ and variance $\sigma_1^2$, and $\{Y_1, \ldots, Y_m\}$ be another random sample, independent of $X$, from a population with mean $\mu_2$ and variance $\sigma_2^2$.

If $\overline{X}$ and $\overline{Y}$ are the respective sample means, then

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

has standard normal distribution $\mathcal{N}(0,1)$ as $n, m \to \infty$.[49]

[49]: Like the CLT, this is a **limiting** result.

**Example** Two different machines are used to fill cereal boxes on an assembly line. The critical measurement influenced by these machines is the weight of the product in the boxes.

The variances of these weights is identical, $\sigma^2 = 1$. Each machine produces a sample of 36 boxes, and the weights are recorded. What is the probability that the difference between the respective averages is less than 0.2, assuming that the true means are identical?

We have $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2 = 1$, $n = m = 36$. The desired probability is

$$\begin{aligned}
&P\left(|\overline{X} - \overline{Y}| < 0.2\right) = P\left(-0.2 < \overline{X} - \overline{Y} < 0.2\right) \\
&= P\left(\frac{-0.2 - 0}{\sqrt{1/36 + 1/36}} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{1/36 + 1/36}} < \frac{0.2 - 0}{\sqrt{1/36 + 1/36}}\right) \\
&= P(-0.8485 < Z < 0.8485) \\
&\approx \Phi(0.8485) - \Phi(-0.8485) \approx 0.6.
\end{aligned}$$

#### Sample Variance $S^2$

When the underlying variance is unknown (which is usually the case in practice), it must be approximated by the sample variance.

**Theorem:** let $\{X_1, \ldots, X_n\}$ be a random sample taken from a normal population with mean $\sigma^2$, and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

be the sample variance. The statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

follows a **chi-squared distribution with $\nu = n - 1$ degrees of freedom** (d.f.), where $\chi^2(\nu) = \Gamma(1/2, \nu)$.

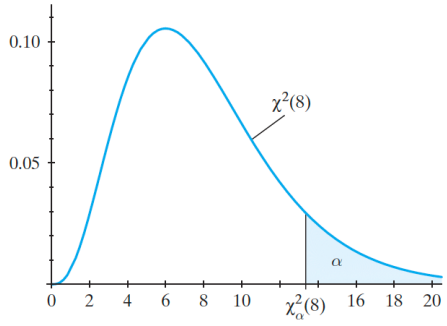



**Figure 6.19:** Chi-squared distribution with 8 degrees of freedom [1].

**Notation:** for $0 < \alpha < 1$ and $\nu \in \mathbb{N}^*$, $\chi^2_\alpha(\nu)$ is the **critical value** for which

$$P(\chi^2 > \chi^2_\alpha(\nu)) = \alpha\,,$$

where $\chi^2 \sim \chi^2(\nu)$ follows a chi-squared distribution with $\nu$ degrees of freedom.

The values of $\chi^2_\alpha(\nu)$ can be found in various textbook tables, or by using `R` or specialized online calculators.

For instance, when $\nu = 8$ and $\alpha = 0.95$, we compute $\chi^2_{0.95}(8)$ *via*

```
qchisq(0.95, df=8,lower.tail = FALSE)
```

```
[1] 2.732637
```

Thus $P(\chi^2 > 2.732) = 0.95$, where $\chi^2 \sim \chi^2(8)$, i.e., $\chi^2$ has a chi-squared distribution with $\nu = 8$ degrees of freedom.

In other words, 95% of the area under the curve of the probability density function of $\chi^2(8)$ is found to the right of 2.732.

### Sample Mean With Unknown Population Variance

Suppose that $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2(\nu)$. If $Z$ and $V$ are independent, then the distribution of the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

is a **Student $t-$distribution with $\nu$ degrees of freedom**, which we denote by $T \sim t(\nu)$.[50]

**Theorem:** let $X_1, \ldots, X_n$ be independent normal random variables with mean $\mu$ and standard deviation $\sigma$. Let $\overline{X}$ and $S^2$ be the sample mean and sample variance, respectively. Then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

follows a **Student $t-$distribution with $\nu = n - 1$ degrees of freedom**.

50: The probability density function of $t(\nu)$ is

$$f(x) = \frac{\Gamma(\nu/2 + 1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)(1 + x^2/\nu)^{\nu/2+1/2}}.$$

Using the same notation as with the chi-squared distribution, let $t_\alpha(\nu)$ represent the **critical $t$-value** above which we find an area under the p.d.f. of $t(\nu)$ equal to $\alpha$ , i.e.

$$P(T > t_\alpha(\nu)) = \alpha \,,$$

where $T \sim t(\nu)$.

For all $\nu$, the Student $t$-distribution is a symmetric distribution around zero, so we have $t_{1-\alpha}(\nu) = -t_\alpha(\nu)$. The critical values can be found in tables, or by using the R function `qt()`.

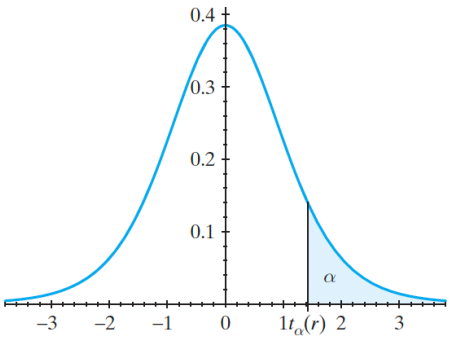



**Figure 6.20:** Student $t-$distribution with $r$ degrees of freedom [1].

If $T \sim t(\nu)$, then for any $0 < \alpha < 1$, we have

$$\begin{aligned} P\left(|T| < t_{\alpha/2}(\nu)\right) &= P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right) \\ &= P\left(T < t_{\alpha/2}(\nu)\right) - P\left(T < -t_{\alpha/2}(\nu)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > -t_{\alpha/2}(\nu)\right)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > t_{1-\alpha/2}(\nu)\right)\right) \\ &= 1 - \alpha/2 - (1 - (1 - \alpha/2)) = 1 - \alpha. \end{aligned}$$

Consequently,

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha \,.$$

We can show that $t(\nu) \to \mathcal{N}(0,1)$ as $\nu \to \infty$; intuitively, this makes sense because the estimate $S$ gets better at estimating $\sigma$ when $n$ increases.

**Example** In R, we can see that when $T \sim t(8)$,

```
qt(0.025, df=8, lower.tail=FALSE)
```

```
[1] 2.306004
```

Thus, $P\left(T > 2.306\right) = 0.025$, which implies

$$P\left(T < -2.306\right) = 0.025$$

, so $t_{0.025}(8) = 2.306$ and

$$\begin{aligned} P(|T| \leq 2.306) &= P(-2.306 \leq T \leq 2.306) \\ &= 1 - P(T < -2.306) - P(T > 2.306) \\ &= 1 - 2P(T < -2.306) = 0.95\,. \end{aligned}$$

The Student $t-$distribution will be useful when the time comes to compute confidence intervals and to do hypothesis testing (see Chapter 7).

### $F-$Distributions

Let $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$. If $U$ and $V$ are independent, then the random variable

$$F = \frac{U/\nu_1}{V/\nu_2}$$

follows an **$F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom**, which we denote by $F \sim F(\nu_1, \nu_2)$.

The probability density function of $F(\nu_1, \nu_2)$ is

$$f(x) = \frac{\Gamma(\nu_1/2 + \nu_2/2)(\nu_1/\nu_2)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1 + x\nu_1/\nu_2)^{\nu_1/2+\nu_2/2}}, \quad x \geq 0.$$

**Theorem:** if $S_1^2$ and $S_2^2$ are the sample variances of independent random samples of size $n$ and $m$, respectively, taken from normal populations with variances $\sigma_1^2$ and $\sigma_2^2$, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

follows an **$F$-distribution with $\nu_1 = n-1$, $\nu_2 = m-1$ d.f.**

**Notation:** for $0 < \alpha < 1$ and $\nu_1, \nu_2 \in \mathbb{N}^*$, $f_\alpha(\nu_1, \nu_2)$ is the **critical value** for which $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$ where $F \sim F(\nu_1, \nu_2)$. Critical values can be found in tables, or by using the `R` function `qf()`.

It can be shown that

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)};$$

for instance, since

```
qf(0.95, df1=6, df2=10, lower.tail=FALSE)
```

```
[1] 0.2463077
```

Thus,

$$f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246\,.$$

These distributions play a role in linear regression and ANOVA models (see Chapters 8 and 11).

## 6.6 Exercises

1. Two events each have probability 0.2 of occurring and are independent. What is the probability that neither occur?
2. Two events each have probability 0.2 and are mutually exclusive. What is the probability that neither occur?
3. A smoke-detector system has two parts, $A$ and $B$. If smoke occurs then the item $A$ detects it with probability 0.95, the item $B$ detects it with probability 0.98 whereas both of them detect it with probability 0.94. What is the probability that the smoke is undetected?
4. Let $A_1, A_2, A_3$ denote the events that the field goal is made by player $1, 2, 3$, respectively. Assume independence and $P(A_1) = 0.5$, $P(A_2) = 0.7$, $P(A_3) = 0.6$. Compute the probability that exactly 1 player is successful.
5. In a group of 16 candidates, 7 are chemists and 9 are physicists. In how many ways can one choose a group of 5 candidates with 2 chemists and 3 physicists?
6. A theorem of combinatorics states that the number of permutations of $n$ objects in which $n_1$ are alike of kind 1, $n_2$ are alike of kind 2, ..., and $n_r$ are alike of kind $r$ (that is, $n = n_1 + n_2 + \cdots + n_r$) is

$$\frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_r!}.$$

   Find the number of different words that can be formed by rearranging the letters in the following words.

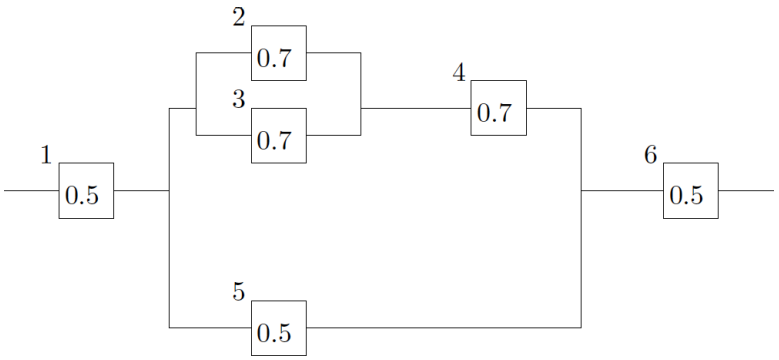
   6.. FRIDGE
   6.. HHTTTT
   6.. LLEWELLYN
   6.. KITCHISSIPPI

7. A class consists of 490 engineering and 510 science students. The students are divided according to their marks:

| | Passed | Failed |
|---|---|---|
| Eng. | 430 | 60 |
| Sci. | 410 | 100 |

   If one person is selected randomly, what is the probability that they failed if they were an engineering student?
8. A company which produces a particular drug has two factories, $A$ and $B$. 70% of the drugs are made in factory $A$, 30% in factory $B$. If 95% of the drugs produced by factory $A$ meet standards while only 75% of those produced by factory $B$ do so, what is the probability that a random dose meets standards?
9. A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease; in 436 cases the test result was positive. The test was also given to a random sample of 500 patients without the disease; only in 5 cases was the result positive. In Canada 11.3% of the population aged 65+ have Alzheimer's disease. Find the probability that a person has the disease given that their test was positive.
10. Twelve items are independently sampled from a production line. If the probability that any given item is defective is 0.1, what is the probability of at most two defectives in the sample?
11. A student can solve 6 problems from a list of 10. For an exam 8 questions are selected at random from the list. What is the probability that the student will solve exactly 5 problems?

12. Consider the following system with six components. We say that it is functional if there exists a path of functional components from left to right. The probability of each component functions is shown. Assume that the components function or fail independently. What is the probability that the system operates?



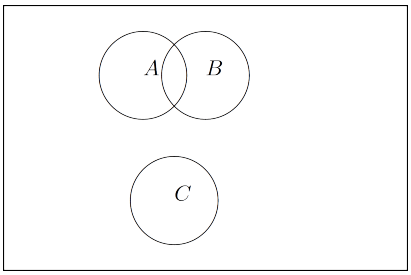

13. Pieces of aluminum are classified according to the finishing of the surface and according to the finishing of edge. The results from 85 samples are summarized as follows:

| | **Edge** | |
|---|---|---|
| **Surface** | excellent | good |
| excellent | 60 | 5 |
| good | 16 | 4 |

Let $A$ denote the event that a selected piece has an "excellent" surface, and let $B$ denote the event that a selected piece has an "excellent" edge. If samples are elected randomly, determine the following probabilities:

13.. $P(A)$ b) $P(B)$ c) $P(A^c)$ d) $P(A \cap B)$ e) $P(A \cup B)$ f) $P(A^c \cup B)$

14. Three events are shown in the Venn diagram below.




Shade the region corresponding to the following events:

a) $A^c$ $(A \cap B) \cup (A \cap B^c)$
b) $(A \cap B) \cup C$
c) $(B \cup C)^c$
d) $(A \cap B)^c \cup C$

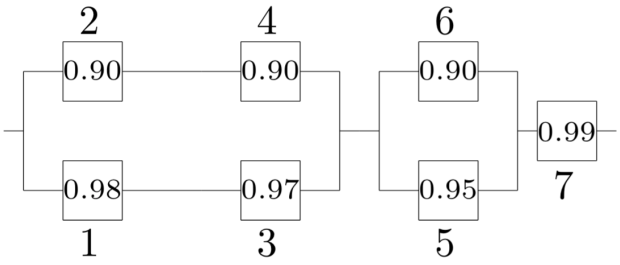
15. If $P(A) = 0.1$, $P(B) = 0.3$, $P(C) = 0.3$, and events $A, B, C$ are mutually exclusive, determine the following probabilities:

a) $P(A \cup B \cup C)$
b) $P(A \cap B \cap C)$
c) $P(A \cap B)$
d) $P((A \cup B) \cap C)$
e) $P(A^c \cap B^c \cap C^c)$
f) $P[(A \cup B \cup C)^c]$

f) The probability that an electrical switch, which is kept in dryness, fails during the guarantee period, is 1%. If the switch is humid, the failure probability is 8%. Assume that 90% of switches are kept in dry conditions, whereas remaining 10% are kept in humid conditions.

a) What is the probability that the switch fails during the guarantee period?
b) If the switch failed during the guarantee period, what is the probability that it was kept in humid conditions?

b) The following system operates only if there is a path of functional device from left to the right. The probability that each device functions is as shown. What is the probability that the circuit operates?




Assume independence.

b) An inspector working for a manufacturing company has a 95% chance of correctly identifying defective items and 2% chance of incorrectly classifying a good item as defective. The company has evidence that 1% of the items it produces are nonconforming (defective).

a) What is the probability that an item selected for inspection is classified as defective?
b) If a random item is classified as non defective, what is the probability that it is indeed good?

b) Consider an ordinary 52-card North American playing deck (4 suits, 13 cards in each suit).

a) How many different 5−card poker hands can be drawn from the deck?
b) How many different 13−card bridge hands can be drawn from the deck?
c) What is the probability of an all-spade 5−card poker hand?
d) What is the probability of a flush (5−cards from the same suit)?
e) What is the probability that a 5−card poker hand contains exactly 3 Kings and 2 Queens?
f) What is the probability that a 5−card poker hand contains exactly 2 Kings, 2 Queens, and 1 Jack?

f) Students on a boat send messages back to shore by arranging seven coloured flags on a vertical flagpole.

a) If they have 4 orange flags and 3 blue flags, how many messages can they send?
b) If they have 7 flags of different colours, how many messages can they send?
c) If they have 3 purple flags, 2 red flags, and 4 yellow flags, how many messages can they send?

c) The Stanley Cup Finals of hockey or the NBA Finals in basketball continue until either the representative team form the Western Conference or from the Eastern Conference wins 4 games. How many different orders are possible ($WWEEEE$ means that the Eastern team won in 6 games) if the series goes

a) 4 games?
b) 5 games?
c) 6 games?
d) 7 games?

d) Consider an ordinary 52-card North American playing deck (4 suits, 13 cards in each suit), from which cards are drawn at random and without replacement, until 3 spades are drawn.

a) What is the probability that there are 2 spades in the first 5 draws?
b) What is the probability that a spade is drawn on the 6th draw given that there were 2 spades in the first 5 draws?
c) What is the probability that 6 cards need to be drawn in order to obtain 3 spades?
d) All the cards are placed back into the deck, and the deck is shuffled. 4 cards are then drawn from. What is the probability of having drawn a spade, a heart, a diamond, and a club, in that order?

d) A student has 5 blue marbles and 4 white marbles in his left pocket, and 4 blue marbles and 5 white marbles in his right pocket. If they transfer one marble at random from their left pocket to his right pocket, what is the probability of them then drawing a blue marble from their right pocket?

d) An insurance company sells a number of different policies; among these, 60% are for cars, 40% are for homes, and 20% are for both. Let $A_1, A_2, A_3, A_4$ represent people with only a car policy, only a home policy, both, or neither, respectively. Let $B$ represent the event that a policyholder renews at least one of the car or home policies.

a) Compute $P(A_1)$, $P(A_2)$, $P(A_3)$, and $P(A_4)$.
b) Assume $P(B \mid A_1) = 0.6$, $P(B \mid A_2) = 0.7$, $P(B \mid A_3) = 0.8$. Given that a client selected at random has a car or a home policy, what is the probability that they will renew one of these policies?

b) An urn contains four balls numbered 1 through 4. The balls are selected one at a time, without replacement. A match occurs if ball $m$ is the $m$th ball selected. Let the event $A_i$ denote a match on the $i$th draw, $i = 1, 2, 3, 4$.

a) Compute $P(A_i)$, $i = 1, 2, 3, 4$.
b) Compute $P(A_i \cap A_j)$, $i, j = 1, 2, 3, 4$, $i \neq j$.
c) Compute $P(A_i \cap A_j \cap A_k)$, $i, j, k = 1, 2, 3, 4$, $i \neq j, i \neq k, j \neq k$.
d) What is the probability of at least 1 match?

d) The probability that a company's workforce has at least one accident in a given month is $(0.01)k$, where $k$ is the number of days in the month. Assume that the numbers of monthly accidents are independent. If the company's year starts on January 1, what is the probability that the first accident occurs in April?

d) A Pap smear is a screening procedure used to detect cervical cancer. Let $T^-$ and $T^+$ represent the events that the test is negative and positive, respectively, and let $C$ represent the event that the person tested has cancer. The false negative rate for this test when the patient has the cancer is 16%; the false positive test for this test when the patient does not have cancer is 19%. In North America, the rate of incidence for this cancer is roughly 8 out of 100,000 women. Based on these numbers, is a Pap smear an effective procedure? What factors influence your conclusion?

d) Of three different fair dice, one each is given to Elowyn, Llewellyn, and Gwynneth. They each roll it. Let $E = \{$Elowyn rolls a 1 or a 2$\}$, $LL = \{$Llewellyn rolls a 3 or a 4$\}$, and $G = \{$Gwynneth rolls a 5 or a 6$\}$.

a) What are the probabilities of each of $E$, $LL$, and $G$ occurring?
b) What are the probabilities of any two of $E$, $LL$, and $G$ occurring simultaneously?
c) What is the probability of all three of the events occurring simultaneously?
d) What is the probability of at least one of $E$, $LL$, or $G$ occurring?

d) Over the course of two baseball seasons, player $A$ obtained 126 hits in 500 at-bats in Season 1, and 90 hits in 300 at-bats in Season 2; player $B$, on the other hand, obtained 75 hits in 300 at-bats in Season 1, and 145 hits in 500 at-bats in Season 2. A player's batting average is the number of hits they obtain divided by the number of at-bats.

a) Which player has the best batting average in Season 1? In Season 2?
b) Which player has the best batting average over the 2-year period?
c) Can you explain what is happening here?

c) A stranger comes to you and shows you what appears to be a normal coin, with two distinct sides: Heads ($H$) and Tails ($T$). They flip the coin 4 times and record the following sequence of tosses: $HHHH$.

a) What is the probability of obtaining this specific sequence of tosses? What assumptions do you make along the way in order to compute the probability? What is the probability that the next toss will be a $T$.
b) The stranger offers you a bet: they will toss the coin another time; if the toss is $T$, they give you 100$, but if it is $H$, you give them 10$. Would you accept the bet (if you are not morally opposed to gambling)?
c) Now the stranger tosses the coin 60 times and records $60 \times H$ in a row: $H \cdots H$. They offer you the same bet. Do you accept it?
d) What if they offered 1000$ instead? 1000000$?

d) An experiment consists in selecting a bowl, and then drawing a ball from that bowl. Bowl $B_1$ contains two red balls and four white balls; bowl $B_2$ contains one red ball and two white balls; and bowl $B_3$ contains five red balls and four white balls. The probabilities for selecting the bowls are not uniform: $P(B_1) = 1/3$, $P(B_2) = 1/6$, and $P(B_3) = 1/2$, respectively.

   a) What is the probability of drawing a red ball $P(R)$?
   b) If the experiment is conducted and a red ball is drawn, what is the probability that the ball was drawn from bowl $B_1$? $B_2$? $B_3$?

b) Two companies $A$ and $B$ consider making an offer for road construction. Company $A$ submits a proposal. The probability that $B$ submits a proposal is $1/3$. If $B$ does not submit the proposal, the probability that $A$ gets the job is $3/5$. If $B$ submits the proposal, the probability that $A$ gets the job is $1/3$. What is the probability that $A$ will get the job?

b) In a box of 50 fuses there are 8 defective ones. We choose 5 fuses randomly (without replacement). What is the probability that all 5 fuses are not defective?

b) The sample space of a random experiment is $\{a, b, c, d, e, f\}$ and each outcome is equally likely. A random variable is defined as follows

| outcome | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ |
|---|---|---|---|---|---|---|
| $X$ | 0 | 0 | 1.5 | 1.5 | 2 | 3 |

Determine the probability mass function of $X$. Determine the following probabilities:

   a) $P(X = 1.5)$
   b) $P(0.5 < X < 2.7)$
   c) $P(X > 3)$
   d) $P(0 \leq X < 2)$
   e) $P(X = 0 \text{ or } 2)$

e) Determine the mean and the variance of the random variable defined in the previous question.

e) $X$ has **uniform distribution** on a set of values $\{X_1, \ldots, X_k\}$ if

$$P(X = X_i) = \frac{1}{k}, \qquad i = 1, \ldots, k.$$

The thickness measurements of a coating process are **uniformly distributed** with values 0.15, 0.16, 0.17, 0.18, 0.19. Determine the mean and variance of the thickness measurements. Is this result compatible with a uniform distribution?

e) Samples of rejuvenated mitochondria are mutated in 1% of cases. Suppose 15 samples are studied and that they can be considered to be independent (from a mutation standpoint). Determine the following probabilities:

   a) no samples are mutated;
   b) at most one sample is mutated, and
   c) more than half the samples are mutated.

c) Samples of 20 parts from a metal punching process are selected every hour. Typically, 1% of the parts require re-work. Let $X$ denote the number of parts in the sample that require re-work. A process problem is suspected if $X$ exceeds its mean by more than three standard deviations.

   a) What is the probability that there is a process problem?
   b) If the re-work percentage increases to 4%, what is the probability that $X$ exceeds 1?
   c) If the re-work percentage increases to 4%, what is the probability that $X$ exceeds 1 in at least one of the next five sampling hours?

c) In a clinical study, volunteers are tested for a gene that has been found to increase the risk for a particular disease. The probability that the person carries a gene is 0.1.

   a) What is the probability that 4 or more people will have to be tested in order to detect 1 person with the gene?
   b) How many people are expected to be tested in order to detect 1 person with the gene?
   c) How many people are expected to be tested in order to detect 2 people with the gene?

c) The number of failures of a testing instrument from contaminated particles on the product is a Poisson random variable with a mean of 0.02 failure per hour.

   a) What is the probability that the instrument does not fail in an 8−hour shift?
   b) What is the probability of at least 1 failure in a 24−hour day?

b) Use R to generate a sample from a binomial distribution and from a Poisson distribution (select parameters as you wish). Use R to compute the sample means and sample variances. Compare these values to population means and population variances.

b) A container of 100 light bulbs contains 5 bad bulbs. We draw 10 bulbs without replacement. Find the probability of drawing at least 1 defective bulb.

b) Let $X$ be a discrete random variable with range $\{0, 1, 2\}$ and probability mass function (p.m.f.) given by $f(0) = 0.5$, $f(1) = 0.3$, and $f(2) = 0.2$. What are the expected value and variance of $X$?

b) A factory employs several thousand workers, of whom 30% are not from an English-speaking background. If 15 members of the union executive committee were chosen from the workers at random, evaluate the probability that exactly 3 members of the committee are not from an English-speaking background.

b) Assuming the context of the previous questions, what is the probability that a majority of the committee members do not come from an English-speaking background?

b) In a video game, a player is confronted with a series of opponents and has an 80% probability of defeating each one. Success with any opponent (that is, defeating the opponent) is independent of previous encounters. The player continues until defeated. What is the probability that the player encounters at least three opponents?

b) Assuming the context of the previous question, how many encounters is the player expected to have?

b) From past experience it is known that 3% of accounts in a large accounting company are in error. The probability that exactly 5 accounts are audited before an account in error is found, is:

b) A receptionist receives on average 2 phone calls per minute. Assume that the number of calls can be modeled using a Poisson random variable. What is the probability that he does not receive a call within a 3−minute interval?

b) Roll a 4−sided die twice, and let $X$ equal the larger of the two outcomes if they are different and the common value if they are the same. Find the p.m.f. and the c.d.f. of $X$.

b) Compute the mean and the variance of $X$ as defined in the previous question, as well as $\mathrm{E}[X(5 - X)]$.

b) A basketball player is successful in 80% of her (independent) free throw attempts. Let $X$ be the minimum number of attempts in order to succeed 10 times. Find the p.m.f. of $X$ and the probability that $X = 12$.

b) Let $X$ be the minimum number of independent trials (each with probability of success $p$) that are needed to observe $r$ successes. The p.m.f. of $X$ is

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-1}, \quad x = r, r+1, \ldots$$

The mean and variance of $X$ are

$$\mathrm{E}[X] = \frac{r}{p} \quad \text{and} \quad \mathrm{Var}[X] = \frac{r(1-p)}{p^2}.$$

Compute the mean minimum number of independent free throw attempts required to observe 10 successful free throws if the probability of success at the free thrown line is 80%. What about the standard deviation of $X$?

b) If $n \geq 20$ and $p \leq 0.05$, it can be shown that the binomial distribution with $n$ trials and an independent probability of success $p$ can be approximated by a Poisson distribution with parameter $\lambda = np$:

$$\frac{(np)^x e^{-np}}{x!} \approx \binom{n}{x} p^x (1-p)^{n-x}.$$

A manufacturer of light bulbs knows that 2% of its bulbs are defective. What is the probability that a box of 100 bulbs contains exactly at most 3 defective bulbs? Use the Poisson approximation to estimate the probability.

b) Consider a discrete random variable $X$ which has a uniform distribution over the first positive $m$ integers, i.e.

$$f(x) = P(X = x) = \frac{1}{m}, \quad x = 1, \ldots, m,$$

and $f(x) = 0$ otherwise. Compute the mean and the variance of $X$. For what values of $m$ is $\mathrm{E}[X] > \mathrm{Var}[X]$?

b) Assume that arrivals of small aircrafts at an airport can be modeled by a Poisson random variable with an average of 1 aircraft per hour.

   a) What is the probability that more than 3 aircrafts arrive within an hour?
   b) Consider 15 consecutive and disjoint 1−hour intervals. What is the probability that in none of these intervals we have more than 3 aircraft arrivals?
   c) What is the probability that exactly 3 aircrafts arrive within 2 hours?

c) In a group of ten students, each student has a probability of 0.7 of passing the exam. What is the probability that exactly 7 of them will pass an exam?

c) A company's warranty states that the probability that a new swimming pool requires some repairs within the 1st year is 20%. What is the probability, that the sixth sold pool is the first one which requires some repairs within the 1st year?

c) Consider the following R output:

```
> pbinom(16,100,0.25)
[1] 0.02111062
> pbinom(30,100,0.25)
[1] 0.8962128
> pbinom(32,100,0.25)
[1] 0.9554037
> pbinom(15,100,0.25)
[1] 0.01108327
> pbinom(17,100,0.25)
[1] 0.03762626
> pbinom(31,100,0.25)
[1] 0.9306511
```

Let $X \sim \mathcal{B}(n, p)$ with $n = 100$ and $p = 0.25$. Using the R output above, calculate $P(16 \leq X \leq 31)$.

c) Consider a random variable $X$ with probability density function given by

$$f(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ 0.75(1 - x^2) & \text{if } -1 \leq x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$

What is the expected value and the standard deviation of $X$?

c) A random variable $X$ has a cumulative distribution function (c.d.f.)

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x/2 & \text{if } 0 < x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

What is the mean value of $X$?

c) Let $X$ be a random variable with p.d.f. $f(x) = \frac{3}{2}x^2$ for $-1 \leq x \leq 1$, and $f(x) = 0$ otherwise. Find $P(X^2 \leq 0.25)$.

c) In the inspection of tin plate produced by a continuous electrolytic process, 0.2 imperfections are spotted per minute, on average. Find the probability of spotting at least 2 imperfections in 5 minutes. Assume that we can model the occurrences of imperfections as a Poisson process.

c) If $X \sim \mathcal{N}(0, 4)$, find $P(|X| \geq 2.2)$.

c) If $X \sim \mathcal{N}(10, 1)$, what value of $k$ yields $P(X \leq k) = 0.701944$?

c) The time it takes a supercomputer to perform a task is normally distributed with mean 10 milliseconds and standard deviation 4 milliseconds. What is the probability that it takes more than 18.2 milliseconds to perform the task? (use the normal table or R).

c) Let $X$ be a random variable. What is the value of $b$ (where $b$ is not a function of $X$) which minimizes $\mathrm{E}[(X - b)^2]$?

c) The time to reaction to a visual signal follows a normal distribution with mean 0.5 seconds and standard deviation 0.035 seconds.

   a) What is the probability that time to react exceeds 1 second?
   b) What is the probability that time to react is between 0.4 and 0.5 seconds?
   c) What is the time to reaction that is exceeded with probability of 0.9?

c) Refer to the situation described in the aircraft question above.

   a) What is the length of the interval such that the probability of having no arrival within this interval is 0.1?
   b) What is the probability that one has to wait at least 3 hours for the arrival of 3 aircrafts?
   c) What is the mean and variance of the waiting time for 3 aircrafts?

c) Assume that $X$ is normally distributed with mean 10 and standard deviation 3. In each case, find the value $x$ such that:

   a) $P(X > x) = 0.5$
   b) $P(X > x) = 0.95$
   c) $P(x < X < 10) = 0.2$
   d) $P(-x < X - 10 < x) = 0.95$
   e) $P(-x < X - 10 < x) = 0.99$

e) Let $X \sim \mathrm{Exp}(\lambda)$ with mean 10. Find $P(X > 30 \mid X > 10)$.

e) Consider a random variable $X$ with the following probability density function:

$$f(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ \frac{3}{4}(1 - x^2) & \text{if } -1 < x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$

What is the value of $P(X \leq 0.5)$?

e) A receptionist receives on average 2 phone calls per minute. If the number of calls follows a Poisson process, what is the probability that the waiting time for call will be greater than 1 minute?

e) A company manufactures hockey pucks. It is known that their weight is normally distributed with mean 1 and standard deviation 0.05. The pucks used by the NHL must weigh between 0.9 and 1.1. What is the probability that a randomly chosen puck can be used by NHL?

e) Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ for the dice example above. Are $X$ and $Y$ independent?
e) Find $\text{Var}[X_1]$, $\text{Var}[X_2]$, and $\text{Cov}(X_1, X_2)$ for the chip example above. Are $X_1$ and $X_2$ independent?
e) Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.m.f.

$$f(x, y) = \frac{x + y}{21}, \quad x = 1, 2, 3, \quad y = 1, 2.$$

e) Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.m.f.

$$f(x, y) = \frac{xy^2}{30}, \quad x = 1, 2, 3, \quad y = 1, 2.$$

Are $X$ and $Y$ independent?
e) Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.m.f.

$$f(x, y) = \frac{xy^2}{13}, \quad (x, y) = (1, 1), (1, 2), (2, 2)$$

Are $X$ and $Y$ independent?
e) Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.d.f.

$$f(x, y) = \frac{3}{2}x^2(1 - |y|), \quad -1 < x < 1, \quad -1 < y < 1.$$

Are $X$ and $Y$ independent?
e) Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ follow

$$f(x, y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}, \quad -\infty < x < \infty, \quad -\infty < y < \infty.$$

e) Suppose that samples of size $n = 25$ are selected at random from a normal population with mean 100 and standard deviation 10. What is the probability that sample mean falls in the interval

$$(\mu_{\overline{X}} - 1.8\sigma_{\overline{X}}, \mu_{\overline{X}} + 1.0\sigma_{\overline{X}})?$$

e) The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean $\mu = 8.2$ minutes and standard deviation $\sigma = 1.5$ minutes. Suppose that a random sample of $n = 49$ customers is taken. Compute the approximate probability that the average waiting time for these customers is:

a) Less than 10 min.
b) Between 5 and 10 min.
c) Less than 6 min.

c) A random sample of size $n_1 = 16$ is selected from a normal population with a mean of 75 and standard deviation of 8. A second random sample of size $n_2 = 9$ is taken independently from another normal population with mean 70 and standard deviation of 12. Let $\overline{X}_1$ and $\overline{X}_2$ be the two sample means. Find

a) The probability that $\overline{X}_1 - \overline{X}_2$ exceeds 4.
b) The probability that $3.5 < \overline{X}_1 - \overline{X}_2 < 5.5$.

b) Using R, illustrate the central limit theorem by generating $M = 300$ samples of size $n = 30$ from:

a) a normal random variable with mean 10 and variance 0.75;
b) a binomial random variable with 3 trials and probability of success 0.3

Repeat the same procedure for samples of size $n = 200$. What do you observe?
b) Suppose that the weight in pounds of a North American adult can be represented by a normal random variable with mean 150 lbs and variance 900 $\text{lbs}^2$. An elevator containing a sign "Maximum 12 people'' can safely carry 2000 lbs. What is the probability that 12 North American adults will not overload the elevator?

b) Let $X_1, \cdots, X_{50}$ be an independent random sample from a Poisson distribution with mean 1. Set $Y = X_1 + \cdots + X_{50}$. Find an approximation of the probability $P(48 \leq Y \leq 52)$.

b) A new type of electronic flash for cameras will last an average of 5000 hours with a standard deviation of 500 hours. A quality control engineer intends to select a random sample of 100 of these flashes and use them until they fail. What is the probability that the mean life time of the sample of 100 flashes will be less than 4928 hours?

b) Assume that random variables $\{X_1, \ldots, X_8\}$ follow a normal distribution with mean 2 and variance 24. Independently, assume that random variables $\{Y_1, \ldots, X_{16}\}$ follow a normal distribution with mean 1 and variance 16. Let $\overline{X}$ and $\overline{Y}$ be the corresponding sample means. What is $P(\overline{X} + \overline{Y} > 4)$?

b) Suppose that $X_1 \sim \mathcal{N}(3, 4)$ and $X_2 \sim \mathcal{N}(3, 45)$. Given that $X_1$ and $X_2$ are independent random variables, what is a good approximation of $P(X_1 + X_2 > 9.5)$?

b) Consider a sample $\{X_1, \ldots, X_{10}\}$ from a normal population $X_i \sim \mathcal{N}(4, 9)$. Denote by $\overline{X}$ and $S^2$ the sample mean and the sample variance, respectively. Find $c$ such that

$$P\left(\frac{\overline{X} - 4}{S/\sqrt{10}} \leq c\right) = 0.99.$$

## Chapter References

[1] R.V. Hogg and E.A Tanis. *Probability and Statistical Inference*. 7th. Pearson/Prentice Hall, 2006.

[2] E.T. Jaynes. *Probability Theory: the Logic of Science*. Cambridge Press, 2003.

[3] Mathematical Association, UK. *An Aeroplane's Guide to A Level Maths*.

[4] R.E. Walpole et al. *Probability and Statistics for Engineers and Scientists*. 8th. Pearson Education, 2007.

[5] Wikipedia. *List of Probability Distributions* . 2021.

# Introductory Statistical Analysis 7

by **Patrick Boily**, with contributions from **Shintaro Hagiwara**

---

Loosely speaking, a **statistic** is any function of a sample from the distribution of a random variable; statistics aim to extract information from an observed sample to summarize the essential features of a dataset.

In this chapter, we introduce basic statistics, and we show how probability theory can be used to build **confidence intervals** and conduct **hypothesis tests**, two of the fundamental tasks of statistical analysis. We also discuss various variance decompositions and multivariate statistics. This review of statistical methods is (by necessity) quite brief; further details can be found in [3, 5, 6, 7, 8, 9, 10, 11, 12].[1]


7.1 Introduction . . . . . 337
7.2 Descriptive Statistics . . . . 337
Data Descriptions . . . . . 338
Outliers . . . . . 343
Visual Summaries . . . . . 343
Coefficient of Correlation . 346
7.3 Estimation . . . . . 349
Standard Error . . . . . 349
C.I. for $\mu$ With $\sigma$ Known . . 351
Confidence Level . . . . . 356
Sample Size . . . . . 358
C.I. for $\mu$ With $\sigma$ Unknown 359
C.I. for a Proportion . . . . . 362
7.4 Hypothesis Testing . . . . . 363
Generalities . . . . . 367
Critical Regions . . . . . 369
Test for a Mean . . . . . 372
Test for a Proportion . . . . 378
Two-Sample Tests . . . . . 379
Difference of 2 Proportions 383
Hypothesis Testing with R . 384
7.5 Additional Topics . . . . . 389
Analysis of Variance . . . . 389
Analysis of Covariance . . . 394
Multivariate Statistics . . . 397
Goodness-of-Fit Test . . . . 401
7.6 Exercises . . . . . 402
Chapter References . . . . . 408


1: A fair number of the examples and exercises we provide in the chapter also come from those references.

## 7.1 Introduction

In general, statistics can be divided into two categories based on their purposes: **descriptive statistics** and **inferential statistics**.

**Descriptive statistics** can be extended to summarize **multivariate** behaviours, *via* sample correlations, contingency tables, scatter plots, etc. They not only provide an easily understandable **overview** of the dataset; they also give analysts a chance to study the collected sample and investigate two important questions:

- is the sample compatible with their understanding of the situation?
- is the sample representative of the underlying population?

**Inferential statistics**, on the other hand, facilitate the process of inference (**induction**) to the general population from which the sample is drawn.

## 7.2 Descriptive Statistics

As its name implies, **descriptive statistics** aim to describe the data; examples include:

- **sample size** (overall and/or subgroups);
- demographic breakdowns of participants;
- measures of **central tendency** (e.g., mean, median, mode, etc.);
- measures of **variability** (e.g., sample variance, minimum, maximum, interquartile range, etc.);
- higher distribution **moments** (skew, kurtosis, etc.);
- **non-parametric** measures (various quantiles);
- **derived** measures (correlation coefficients), etc.

They can be presented as a **single number**, in a **summary table**, or even in **graphical representations** (e.g., histogram, pie chart, etc.).

### 7.2.1 Data Descriptions

Studies and experiments give rise to **statistical units**. These units are typically described with **variables** (and measurements), which are either **qualitative** (categorical) or **quantitative** (numerical).

Categorical variables take values (**levels**) from a finite set of pre-determined **categories** (or classes); numerical variables from a (potentially infinite) set of **quantities**.

#### Examples

1. Age is a **numerical** variable, measured in years, although is is often reported to the nearest year integer, or in an age range of years, in which case it is an **ordinal** variable (mixture of qualitative or quantitative).
2. Typical numerical variables include distance in $m$, volume in $m^3$, etc.
3. Disease diagnosis is a **categorical** variable with (at least) 2 categories (positive/negative).
4. Compliance with a standard is a categorical variable: there could be 2 levels (compliant/non-compliant) or more (compliance, minor non-compliance issues, major non-compliance issues).
5. **Count** variables are numerical variables.

In a first pass, a variable can be described along (at least) 2 dimensions: its **centrality** and its **spread**:[2]

2: The **skew** and the **kurtosis** are also sometimes used.

- **centrality** measures include the **median**, the **mean**, and, less frequently, the **mode**;
- spread (or **dispersion**) measures include the **standard deviation** (sd), the **quartiles**, the **inter-quartile range** (IQR), and, less frequently, the **range**.

The median, range, and quartiles are all easily calculated from an **ordered** list of the data.

#### Sample Median

The **median** $\text{med}(x_1, \ldots, x_n)$ of a sample of size $n$ is a numerical value which splits the ordered data into 2 equal subsets: half the observations fall **below** the median, and half **above** it:

- if $n$ is **odd**, then the **position** of the median (or its **rank**) is $(n+1)/2$ – the median observation is the $\frac{n+1}{2}^{\text{th}}$ ordered observation;
- if $n$ is **even**, then the median is the average of the $\frac{n}{2}^{\text{th}}$ and the $(\frac{n}{2}+1)^{\text{th}}$ ordered observations.

The procedure is simple: order the data, and follow the even/odd rules **to the letter**.

#### Examples

1. $\text{med}(4,6,1,3,7) = \text{med}(1,3,4,6,7) = x_{(5+1)/2} = x_3 = 4$. There are 2 observations below 4 $\{1,3\}$, and 2 observations above 4 $\{6,7\}$.
2. $\text{med}(1,3,4,6,7,23) = \frac{x_{6/2}+x_{6/2+1}}{2} = \frac{x_3+x_4}{2} = \frac{4+6}{2} = 5$. There are 3 observations below 5 $\{1,3,4\}$, and 3 observations above 4 $\{6,7,23\}$.
3. $\text{med}(1,3,3,6,7) = x_{(5+1)/2} = x_3 = 3$. There seems to be only 1 observation below 3 $\{1\}$, but 2 observations above 3 $\{6,7\}$.

Note that there is ambiguity in the definition of the median: **above** and **below** should be interpreted as **after** and **before**, respectively, inclusive of the median value. In the last example above, for instance, there are 2 observations ($x_1 = 1, x_2 = 3$) before the median observation ($x_3 = 3$), and 2 after the median ($x_4 = 6, x_5 = 7$).

### Sample Mean

The **mean** of a sample is simply the arithmetic average of its observations. For observations $x_1, \ldots, x_n$, the sample mean is

$$\text{AM}(x_1, \ldots, x_n) = \overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

Other means exist, such as the **harmonic** mean and the **geometric** mean:

$$\text{HM}(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$
$$\text{GM}(x_1, \ldots, x_n) = \sqrt[n]{x_1 \cdots x_n}.$$

All of these measures attempt to find an "average" of the observations.

#### Examples

1. $\text{AM}(4,6,1,3,7) = \frac{4+6+1+3+7}{5} = \frac{21}{5} = 4.2 \approx 4 = \text{med}(4,6,1,3,7)$.
2. $\text{AM}(1,3,4,6,7,23) = \frac{1+3+4+6+7+23}{6} = \frac{44}{6} \approx 7.3$, which is not nearly as close to $\text{med}(1,3,4,6,7,23) = 5$.
3. $\text{HM}(4,6,1,3,7) = \frac{5}{\frac{1}{4}+\frac{1}{6}+\frac{1}{1}+\frac{1}{3}+\frac{1}{7}} = \frac{5}{53/28} = \frac{140}{53} \approx 2.64$.
4. $\text{GM}(4,6,1,3,7) = \sqrt[5]{4 \cdot 6 \cdot 1 \cdot 3 \cdot 7} \approx \sqrt[5]{(504)} \approx 3.47$.

It can be shown that if $x = (x_1, \ldots, x_n)$ and $x_i > 0$ for all $i$, then

$$\min(x) \leq \text{HM}(x) \leq \text{GM}(x) \leq \text{AM}(x) \leq \max(x).$$

There is no need to decide on a single centrality measure when reporting on the data; in practice, we may use as many of them as we want to.

But there are situations where the mean (or the median) could prove to be a better choice. On the one hand, the use of the mean is **theoretically supported** by the **Central Limit Theorem** (CLT; see Section 6.5.2.

When the data distribution is roughly **symmetric**, then the median and the mean will be near one another. If the data distribution is **skewed** then the mean is pulled toward the long tail and as a result gives a distorted view of the centre (see Figure 7.1).

Consequently, medians are generally used for house prices, incomes, etc., as the median is **robust** against outliers and incorrect readings (whereas the mean is not).

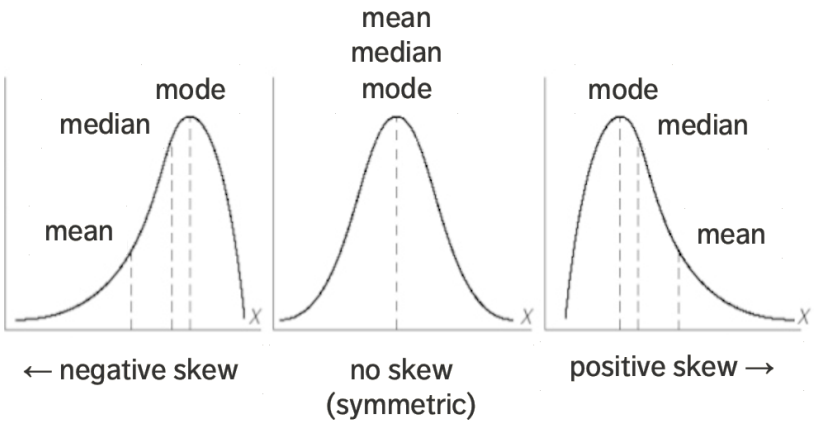



**Figure 7.1:** Mean, median, and mode in various skewness scenarios. [modified from unknown source]

### Standard Deviation

While the mean, the median, and the mode provide an idea as to where some of the distribution's "mass" is located, the **standard deviation** provides some notion of its spread. The higher the standard deviation, the further away from the mean the variable values are likely to fall (see Figure 7.2). We will have more to say on this topic.

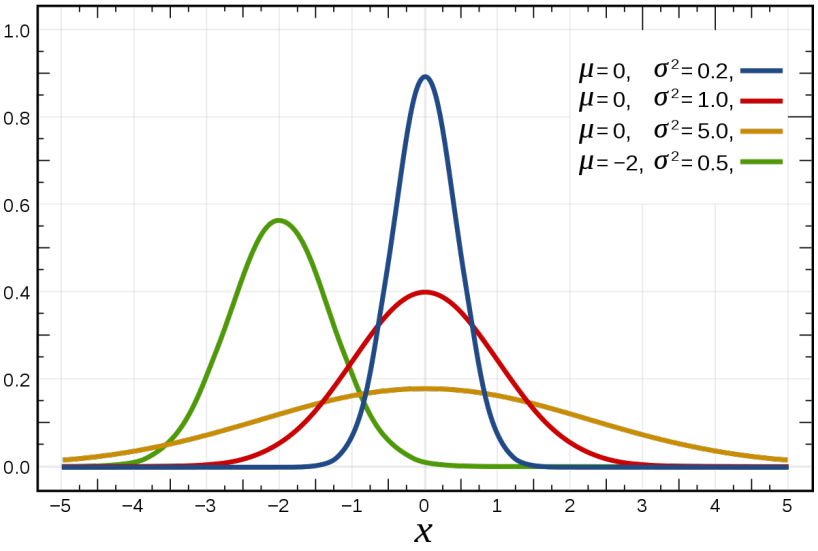



**Figure 7.2:** Normal distributions, with various means and standard deviations. [Wikipedia]

### Quantiles

Another way to provide information about the spread of the data is *via* **centiles**, **deciles**, and/or **quartiles**.

The **lower quartile** $Q_1(x_1, \ldots, x_n)$ of a sample of size $n$, or $Q_1$, is a numerical value which splits the ordered data into 2 unequal subsets: 25% of the observations fall below $Q_1$ **and** 75% of the observations fall above $Q_1$.

Similarly, the **upper quartile** $Q_3$ splits the ordered data into 75% of the observations below $Q_3$, **and** 25% of the observations above $Q_3$.

The median can be interpreted as the **middle quartile** $Q_2$, of the sample, the minimum as $Q_0$, and the maximum as $Q_4$: the vector $(Q_0, Q_1, Q_2, Q_3, Q_4)$ is the **5-pt summary** of the data.

**Centiles** $p_i$, $i = 0, \ldots, 100$ and **deciles** $d_j$, $j = 0, \ldots, 10$ run through different splitting percentages

$$p_{25} = Q_1, p_{75} = Q_3, d_5 = Q_2, \text{ etc.}$$

They are found as with the media: **sort** the sample observations $\{x_1, x_2, \ldots, x_n\}$ in an **increasing order** as

$$y_1 \leq y_2 \leq \ldots \leq y_n.$$

The smallest $y_1$ has **rank** 1 and the largest $y_n$ has **rank** $n$.

Any value that falls between the observations of ranks:

- $\lfloor \frac{n}{4} \rfloor$ and $\lfloor \frac{n}{4} \rfloor + 1$ is a **lower quartile** $Q_1$;
- $\lfloor \frac{3n}{4} \rfloor$ and $\lfloor \frac{3n}{4} \rfloor + 1$ is an **upper quartile** $Q_3$;
- $\lfloor \frac{in}{100} \rfloor$ and $\lfloor \frac{in}{100} \rfloor + 1$ is a **centile** $p_i$, for $i = 1, \ldots, 99$;
- $\lfloor \frac{jn}{10} \rfloor$ and $\lfloor \frac{jn}{10} \rfloor + 1$ is a **decile** $d_j$, for $j = 1, \ldots, 9$.

In practice, we compute the $m-$**quantile of order** $k$ for the data, where $k = 1, \ldots, m-1$ by averaging the observations of rank

$$\left\lfloor \frac{kn}{m} \right\rfloor \quad \text{and} \quad \left\lfloor \frac{kn}{m} \right\rfloor + 1;$$

other protocols exist, such as the use of **weighted averages** (where the weights are determined by rank $k$ of the $m-$quantile of interest).

#### Examples

1. $Q_1(1,3,4,6,7) = \frac{1}{2}\left(y_{\lfloor 5/4 \rfloor} + y_{\lfloor 5/4 \rfloor + 1}\right) = \frac{1}{2}\left(y_1 + y_2\right) = \frac{1}{2}(1+3) = 2.$
2. $d_7(1,3,4,6,7,23) = \frac{1}{2}\left(y_{\lfloor 7(6)/10 \rfloor} + y_{\lfloor 7(6)/10 \rfloor + 1}\right) = \frac{1}{2}\left(y_4 + y_5\right) = \frac{1}{2}(6+7) = 13/2.$
3. $Q_1(1,3,4,6,7,23) = \frac{1}{2}\left(y_{\lfloor 6/4 \rfloor} + y_{\lfloor 6/4 \rfloor + 1}\right) = \frac{1}{2}\left(y_1 + y_2\right) = \frac{1}{2}(1+3) = 2.$
4. $Q_3(1,3,4,6,7,23) = \frac{1}{2}\left(y_{\lfloor 3(6)/4 \rfloor} + y_{\lfloor 3(6)/4 \rfloor + 1}\right) = \frac{1}{2}\left(y_4 + y_5\right) = \frac{1}{2}(6+7) = 6.5.$

5. Consider the following midterm grades:

```
grades<-c(
  80,73,83,60,49,96,87,87,60,53,66,83,32,80,66,90,72,55,76,46,48,69,45,48,77,52,59,97,
  76,89,73,73,48,59,55,76,87,55,80,90,83,66,80,97,80,55,94,73,49,32,76,57,42,94,80,90,
  90,62,85,87,97,50,73,77,66,35,66,76,90,73,80,70,73,94,59,52,81,90,55,73,76,90,46,66,
  76,69,76,80,42,66,83,80,46,55,80,76,94,69,57,55,66,46,87,83,49,82,93,47,59,68,65,66,
  69,76,38,99,61,46,73,90,66,100,83,48,97,69,62,80,66,55,28,83,59,48,61,87,72,46,94,48,
  59,69,97,83,80,66,76,25,55,69,76,38,21,87,52,90,62,73,73,89,25,94,27,66,66,76,90,83,
  52,52,83,66,48,62,80,35,59,72,97,69,62,90,48,83,55,58,66,100,82,78,62,73,55,84,83,66,
  49,76,73,54,55,87,50,73,54,52,62,36,87,80,80
  )
```

The quartiles and mean are:

```
summary(grades)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   55.00   70.00   68.74   82.50  100.00
```

### Dispersion Measures

Some of the dispersion measures are fairly simple to compute: the **sample range** is

$$\text{range}(x_1, \ldots, x_n) = \max\{x_i\} - \min\{x_i\};$$

the **inter-quartile range** is $\text{IQR} = Q_3 - Q_1$.

The **sample standard deviation** $s$ and **sample variance** $s^2$ are estimates of the underlying distribution's $\sigma$ and $\sigma^2$. For observations $x_1, \ldots, x_n$,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2);$$

it differs from the (population) standard deviation and the (population) variance in the denominator: $n-1$ is used instead of $n$.[3]

3: In statistical parlance, we say that 1 **degree of freedom** is lost when we use the sample to estimate the sample mean.

### Examples

1. The sample variance of $\{1, 3, 4, 6, 7\}$ is

$$\frac{1}{5-1}\left(\sum_{i=1}^{5} x_i^2 - \frac{1}{5}\left(\sum_{i=1}^{5} x_i\right)^2\right) = \frac{1}{4}\left(111 - \frac{1}{5}(21)^2\right) = 5.7.$$

2. The interquartile range of $\{1, 3, 4, 6, 7, 23\}$ is

$$\begin{aligned}\text{IQR}(1,3,4,6,7,23) &= Q_3(1,3,4,6,7,23) - Q_1(1,3,4,6,7,23)\\ &= 6.5 - 2 = 4.5.\end{aligned}$$

3. We can provide more data descriptions of the `grades` dataset (see above) using `psych`'s `describe()` function.

```
psych::describe(grades)
```

```
   vars   n  mean    sd median trimmed
X1    1 211 68.74 17.37     70   69.43

    mad min max range  skew kurtosis  se
  19.27  21 100    79 -0.37    -0.46 1.2
```

### 7.2.2 Outliers

An **outlier** is an observation that lies outside the overall pattern in a distribution.[4] Let $x$ be an observation in the sample;[5] it is a

- **suspected outlier** if

$$x < Q_1 - 1.5\,\text{IQR} \quad \text{or} \quad x > Q_3 + 1.5\,\text{IQR},$$

- **definite outlier** if

$$x < Q_1 - 3\,\text{IQR} \quad \text{or} \quad x > Q_3 + 3\,\text{IQR}.$$

4: Outlier analysis (and anomaly detection) is its own discipline – an overview is provided in Chapter 26.

5: In theory, this definition only applies to **normally distributed** data, but it is often used as a first pass for outlier analysis even when the data is not normally distributed.

**Example** In the set $\{1, 3, 4, 6, 7, 23\}$, $Q_1 = 2$, $Q_3 = 6.5$, and IQR $= 4.5$. Thus

$$\begin{aligned} Q_1 - 1.5\text{IQR} &= 2 - 1.5(4.5) = -4.75 \\ Q_3 + 1.5\text{IQR} &= 6.5 + 1.5(4.5) = 13.25 \\ Q_1 - 3\text{IQR} &= 2 - 3(4.5) = -11.5 \\ Q_3 + 3\text{IQR} &= 6.5 + 3(4.5) = 20.0 \end{aligned}$$

Since $23 > Q_3 + 3\text{IQR}$ (and $23 > Q_3 + 1.5\text{IQR}$), 23 is both a definite (and a suspected) outlier of $\{1, 3, 4, 6, 7, 23\}$.

### 7.2.3 Visual Summaries

The **boxplot** (also known as the box-and-whisker plot) is a quick and easy way to present a graphical summary of a univariate distribution:

1. draw a box along the observation axis, with endpoints at the lower and upper quartiles $Q_1$ (knees) and $Q_3$ (shoulders), and with a "belt" at the median $Q_2$;
2. draw a line extending from $Q_1$ to the smallest value closer than 1.5IQR to the left of $Q_1$;
3. draw a line extending from $Q_3$ to the largest value closer than 1.5IQR to the right of $Q_3$;
4. any suspected outlier is plotted separately (as in Figure 7.3):

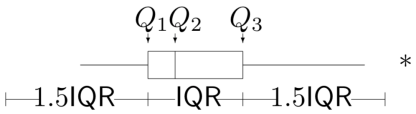



**Figure 7.3:** Boxplot with one (suspected) outlier.

#### Skewness

For **symmetric** distributions, the median and mean are equal, and the quartiles $Q_1$ and $Q_3$ are equidistant from $Q_2$:

- if $Q_3 - Q_2 > Q_2 - Q_1$ then the data distribution is **skewed to the right** (positively skewed);
- if $Q_3 - Q_2 < Q_2 - Q_1$ then the data distribution is **skewed to left** (negatively skewed).

Graphically, if the distance between the shoulders and the belt is larger than the distance between the belt and the knees, then the data is skewed to the right; if it's the opposite, the data is skewed to the left.

In the boxplots below, the data is skewed to the right.

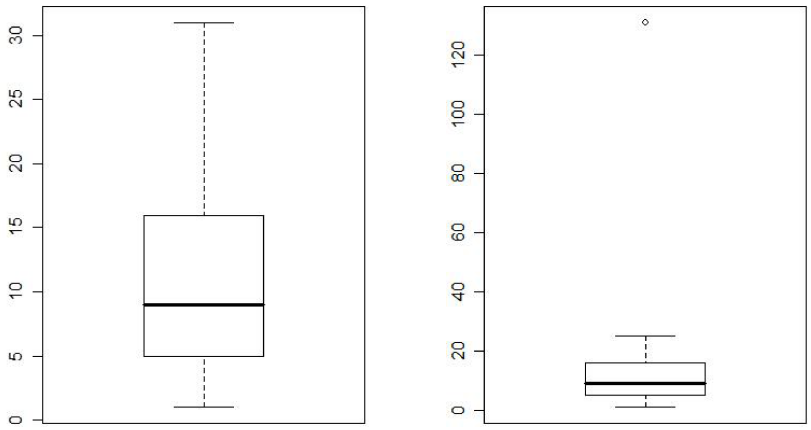



**Figure 7.4:** Boxplot of positively skewed datasets.

### Histograms

Visual information about the distribution of the sample can also be provided *via* **histograms**.

A histogram for the sample $\{x_1, \ldots, x_n\}$ is built according to the following specifications:

- the **range** of the histogram is $r = \max\{x_i\} - \min\{x_i\}$;
- the **number of bins** should approach $k = \sqrt{n}$, where $n$ is the sample size;
- the **bin width** should approach $r/k$, and
- the **frequency of observations** in each bin should be represented by the **bin height**.

### Shapes of Datasets

Boxplots and histograms provide an easy visual impression of the **shape of the data set**, which can eventually suggest a mathematical model for the situation of interest: another way to define skewness is to say that data is **skewed to the right** if the corresponding boxplot or histogram is stretched to the right, and *vice-versa*.

#### Examples

1. Consider the daily number of car accidents in Sydney, Australia, over a 40-day period:
   `6 3 2 24 12 3 7 14 21 9 14 22 15 2 17 10 7 7 31 7`
   `18 6 8 2 3 2 17 7 7 21 13 23 1 11 3 9 4 9 9 25`

   The sorted values are:
   `1 2 2 2 2 3 3 3 3 4 6 6 7 7 7 7 7 7 8 9`
   `9 9 9 10 11 12 13 14 14 15 17 17 18 21 21 22 23 24 25 31`

We can then easily see that

$$\text{min} = y_1 = 1, \quad Q_1 = \frac{1}{2}(y_{10} + y_{11}) = 5, \quad \text{med} = \frac{1}{2}(y_{20} + y_{21}) = 9,$$
$$Q_3 = \frac{1}{2}(y_{30} + y_{31}) = 16, \quad \text{max} = y_{40} = 31.$$

A corresponding histogram and boxplot are shown in Figure 7.5.

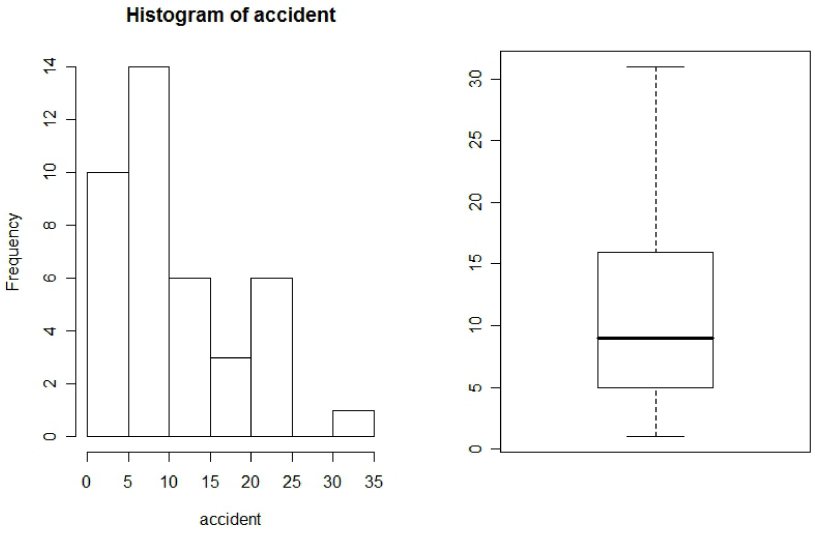
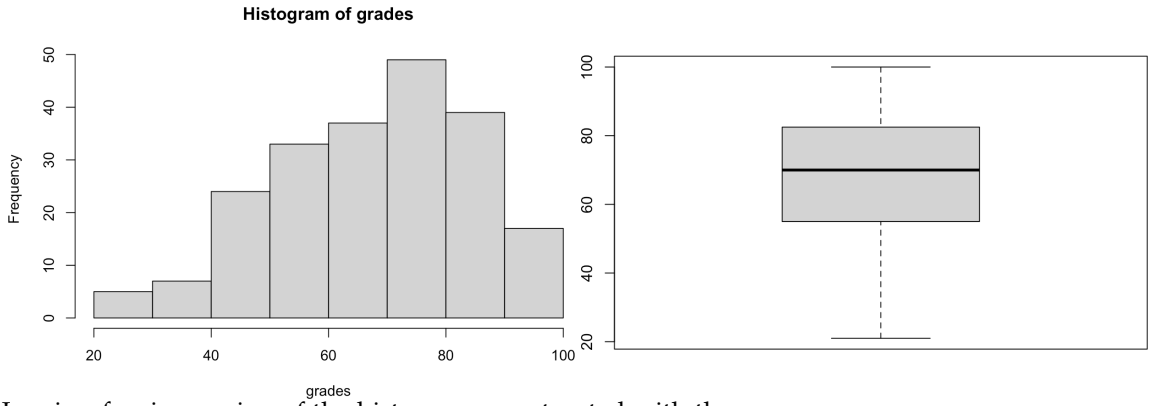



**Figure 7.5:** Histogram and boxplot of the Sydney accident dataset.

2. We can also visualize the `grades` dataset:

```
hist(grades, breaks = seq(20,100,10))
boxplot(grades)
```

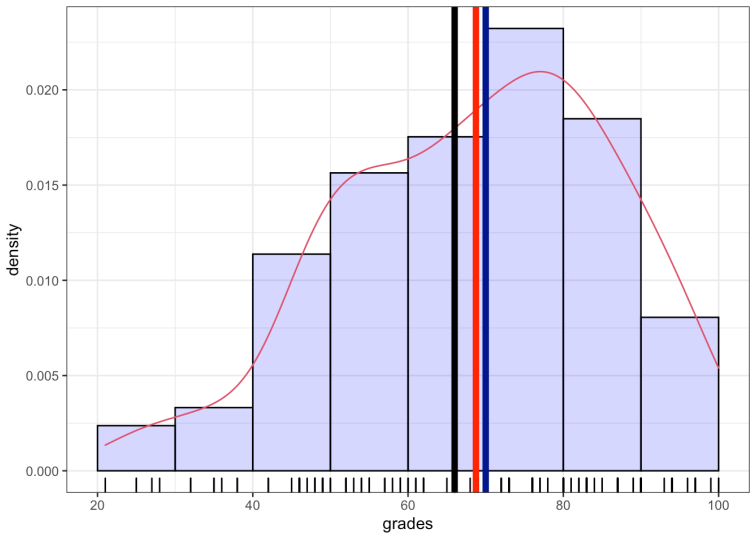



Here is a fancier version of the histogram, constructed with the `ggplot2` package.[6]

6: See Section [1] for details on the use of this R package.

```
# function to find the mode
fun.mode<-function(x){
    as.numeric(names(sort(-table(x)))[1])}

library(ggplot2)
ggplot(data=data.frame(grades), aes(grades)) +
    geom_histogram(aes(y =..density..),    # approximated pdf
        breaks=seq(20, 100, by = 10),              # 8 bins from 20 to 100
        col="black",                               # colour of outline
```

```
    fill="blue",                                # fill colour of bars
    alpha=.2) +                                 # transparency
geom_density(col=2) +                           # colour of pdf curve
geom_rug(aes(grades)) +                  # adding a rug on x-axis
geom_vline(aes(xintercept = mean(grades)),
    col='red',size=2) +                  # vertical line: mean
geom_vline(aes(xintercept = median(grades)),
    col='darkblue',size=2) +             # vertical line: median
geom_vline(aes(xintercept = fun.mode(grades)),
    col='black',size=2)                  # vertical line:  mode
```




What is the shape of this dataset? Is the class in trouble?

### 7.2.4 Coefficient of Correlation

For bivariate (or multivariate) datasets, we can still study each variable separately, as in the previous sections, but we might also be interested in determining how the variables relate to one another.

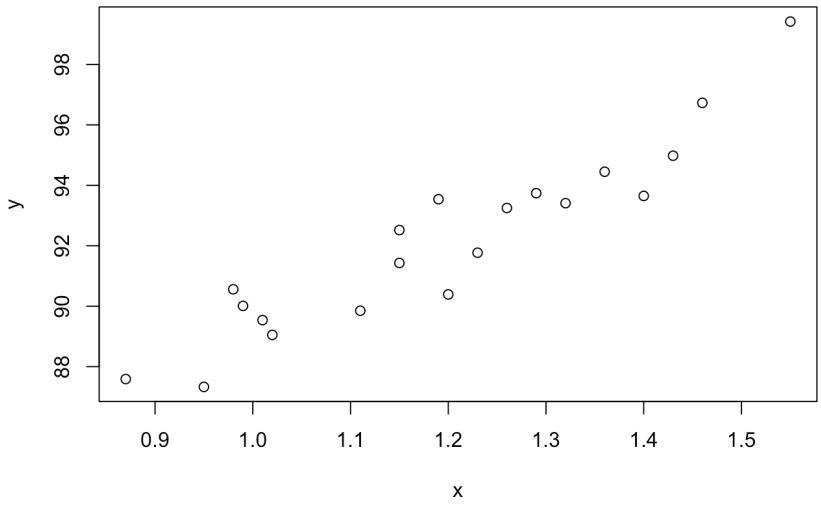
For instance, consider the following data, consisting of $n = 20$ paired measurements $(x_i, y_i)$ of hydrocarbon levels $x$ and pure oxygen levels $y$ in fuels:

```
x = c(
  0.99,1.02,1.15,1.29,1.46,1.36,0.87,1.23,
  1.55,1.40,1.19,1.15,0.98,1.01,1.11,1.20,
  1.26,1.32,1.43,0.95
  )
y = c(
  90.01,89.05,91.43,93.74,96.73,94.45,87.59,91.77,
  99.42,93.65,93.54,92.52,90.56,89.54,89.85,90.39,
  93.25,93.41,94.98,87.33
)

cbind(x,y)
```

```
        x     y           x     y
 [1,] 0.99 90.01  [11,] 1.19 93.54
 [2,] 1.02 89.05  [12,] 1.15 92.52
 [3,] 1.15 91.43  [13,] 0.98 90.56
 [4,] 1.29 93.74  [14,] 1.01 89.54
 [5,] 1.46 96.73  [15,] 1.11 89.85
 [6,] 1.36 94.45  [16,] 1.20 90.39
 [7,] 0.87 87.59  [17,] 1.26 93.25
 [8,] 1.23 91.77  [18,] 1.32 93.41
 [9,] 1.55 99.42  [19,] 1.43 94.98
[10,] 1.40 93.65  [20,] 0.95 87.33
```

Assume that we are interested in measuring the **strength of association** between $x$ and $y$. We can use a graphical display to provide an initial description of the relationship: it appears that the observations lie around a **hidden line**.

```
plot(x,y)
```




For paired data $(x_i, y_i)$, $i = 1, \ldots, n$, the **sample correlation coefficient** of $x$ and $y$ is

$$\rho_{XY} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}.$$

The coefficient $\rho_{XY}$ is defined only if $S_{xx} \neq 0$ and $S_{yy} \neq 0$, i.e. if neither $x_i$ nor $y_i$ are constant.

The variables $x$ and $y$ are **uncorrelated** if $\rho_{XY} = 0$ (or is very small, in practice), and **correlated** if $\rho_{XY} \neq 0$ (or if $|\rho_{XY}|$ is "large", in practice).

**Example** For the data on the previous page, we have

$$S_{xy} \approx 10.18,\ S_{xx} \approx 0.68,\ S_{yy} \approx 173.38,$$

so that

$$\rho_{XY} \approx \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx 0.94.$$

This can also be computed directly in R:

```
(Sxx = sum((x-mean(x))^2))
(Syy = sum((y-mean(y))^2))
(Sxy = sum((x-mean(x))*(y-mean(y))))
(rho = Sxy/sqrt(Sxx*Syy))
```

```
[1] 0.68088
[1] 173.3769
[1] 10.17744
[1] 0.9367154
```

or by using the `cor()` function:

```
cor(x,y)
```

```
[1] 0.9367154
```

**Properties**

- $\rho_{XY}$ is unaffected by changes of scale or origin. Adding constants to $x$ does not change $x - \overline{x}$ (similarly for $y - \overline{y}$) and multiplying $x$ and $y$ by constants changes both the numerator and denominator equally;
- $\rho_{XY}$ is symmetric in $x$ and $y$ (i.e. $\rho_{XY} = \rho_{YX}$) and $-1 \leq \rho_{XY} \leq 1$; if $\rho_{XY} = \pm 1$, then the observations $(x_i, y_i)$ all lie on a straight line with a positive (or negative) slope;
- the sign of $\rho_{XY}$ reflects the trend of the points;
- a high correlation coefficient value $|\rho_{XY}|$ does not necessarily imply a **causal relationship** between the two variables;
- note that $x$ and $y$ can have a very strong **non-linear** relationship without $\rho_{XY}$ reflecting it (see Figure 7.6).

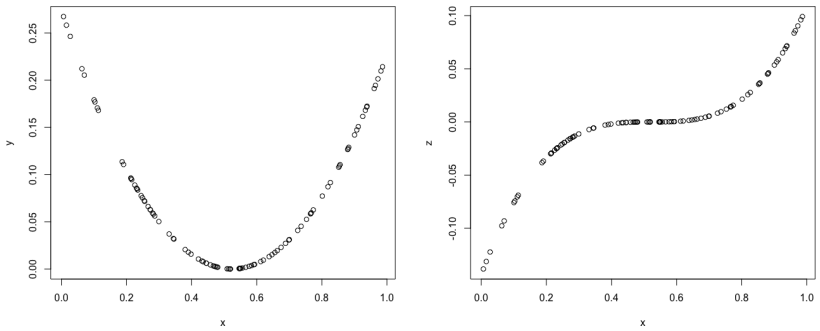



**Figure 7.6:** Examples of strong relationships that are not reflected by the coefficient of correlation.

Human brains are ... not that great at intuiting correlations, even when the relationship has a linear component: in the above figure, how obvious is it that the correlation on the left is $-0.12$, and that the one on the right is 0.93? Beware!

## 7.3 Point and Interval Estimation

One of the goals of **statistical inference** is to draw conclusions about a **population** based on a random sample from the population.

For instance, we might want answers to the following questions.

1. Can we assess the reliability of a product's manufacturing process by randomly selecting a sample of the final product and determining how many of them are compliant according to some quality assessment scheme?
2. Can we determine who will win an election by polling a small sample of respondents?

Specifically, we seek to estimate an unknown **parameter** $\theta$, say, using a single quantity called the **point estimate** $\hat{\theta}$.

This point estimate is obtained *via* a **statistic**, which is simply a function of a random sample.[7]

The probability distribution of the statistic is its **sampling distribution**; as an example, we have discussed the sampling distribution of the **sample mean** in Section 6.5. Describing such sampling distributions is a main focus of statistical research.

7: Common examples of inferential statistics include:

- **sample mean** and **sample median**;
- **sample variance** and **sample standard deviation**;
- **sample quantiles** (median, quartiles, quantiles);
- **test statistics** ($t-$statistics, $\chi^2-$statistics, $f-$statistics, etc.);
- **order statistics** (sample maximum and minimum, sample range, etc.);
- **sample moments** and functions thereof (skewness, kurtosis, etc.);
- etc.

**Example** Consider a process that manufactures gear wheels. Let $X$ be the random variable that records the weight of a randomly selected gear wheel. What is the population mean $\mu_X = \mathrm{E}[X]$?.

In the absence of the p.d.f. $f(x)$, we can estimate $\mu = X$ with the help of a random sample $X_1, \ldots, X_n$ of gear wheel weight measurements, *via* the sample mean statistic:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

which follows approximately a $\mathcal{N}\left(\mu, \sigma^2/n\right)$ distribution, according to the CLT.

### 7.3.1 Estimator (Sampling) Variance and Standard Error

In practice, the point estimator $\hat{\theta}$ varies depending on the choice of the sample $\{X_1, \ldots, X_n\}$.

The **standard error** of a statistic is the **standard deviation of its sampling distribution**.

For instance, if observations $X_1, \ldots, X_n$ come from a a population with **unknown mean** $\mu$ and **known variance** $\sigma^2$, then $\mathrm{Var}(\overline{X}) = \sigma^2/n$ and the **standard error of** $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

If the variance of the original population is **unknown**, then it is estimated by the sample variance $S^2$ and the **estimated standard error of** $\overline{X}$ is

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n}}, \quad \text{where} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

**Examples**

1. A sample of 20 baseball player heights (in inches) is shown below.

```
x=c(74,74,72,72,73,69,69,71,76,71,
    73,73,74,74,69,70,72,73,75,78)
```

What is the standard error of the sample mean $\overline{X}$?

The sampling mean of the heights is

$$\overline{X} = \frac{X_1 + \cdots + X_{20}}{20} = 72.6$$

and the sample variance $S^2$ is

$$S^2 = \frac{1}{20-1}\sum_{i=1}^{20}(X_i - 72.6)^2 \approx 5.6211.$$

The standard error of $\overline{X}$ is thus

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{20}} \approx \sqrt{\frac{5.6211}{20}} \approx 0.5301.$$

The quantities can be computed directly *via* R:[8]

8: Note that `var()` always treats the underlying dataset as a **sample**, not as a **population**.

```
(x.bar = mean(x))
(S2.x = var(x))
(se.x = sqrt(S2.x/length(x)))
```

```
[1] 72.6
[1] 5.621053
[1] 0.530144
```

2. Consider a sample $\{X_1, \ldots, X_{100}\}$ of independent observations selected from a normal population $\mathcal{N}(\mu, \sigma^2)$ where $\sigma = 50$ is known, but $\mu$ is not. What is the best estimate of $\mu$? What is the sampling distribution of that estimate?

The sample mean $\overline{X} = \frac{1}{100}(X_1 + \cdots + X_{100})$ is the best estimate of $\mu_X = \mu_{\overline{X}}$ and the standard error of $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{50}{\sqrt{100}} = 5.$$

Since the observations are sampled independently from a normal population with mean $\mu$ and standard deviation 50, which is to say, $\overline{X} \sim \mathcal{N}(\mu, 5^2) = \mathcal{N}(\mu, 25)$, according to the CLT.

### 7.3.2 Confidence Intervals for $\mu$ When $\sigma$ is Known

Consider a sample $\{x_1, \ldots, x_n\}$ drawn from a **normal population** with **known** variance $\sigma^2$ and **unknown** mean $\mu$. The sample mean

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

is a **point estimate** of $\mu$.[9]

Of course, this estimate is not exact, because $\overline{x}$ is an **observed value** of $\overline{X}$; it is unlikely that the observed value $\overline{x}$ should coincide with $\mu$.

We know that $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, so that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

**The** $68 - 96 - 99.7$ **Rule**

For the standard normal distribution, it can be shown that

$$P(|Z| < 1) \approx 0.683, \quad P(|Z| < 2) \approx 0.955, \quad P(|Z| < 3) \approx 0.997.$$

This says that about 68% of the observations from $\mathcal{N}(0, 1)$ fall within one standard deviation ($\sigma = 1$) from the mean ($\mu = 0$), about 96% within two standard deviations, and about 99.7% within three.

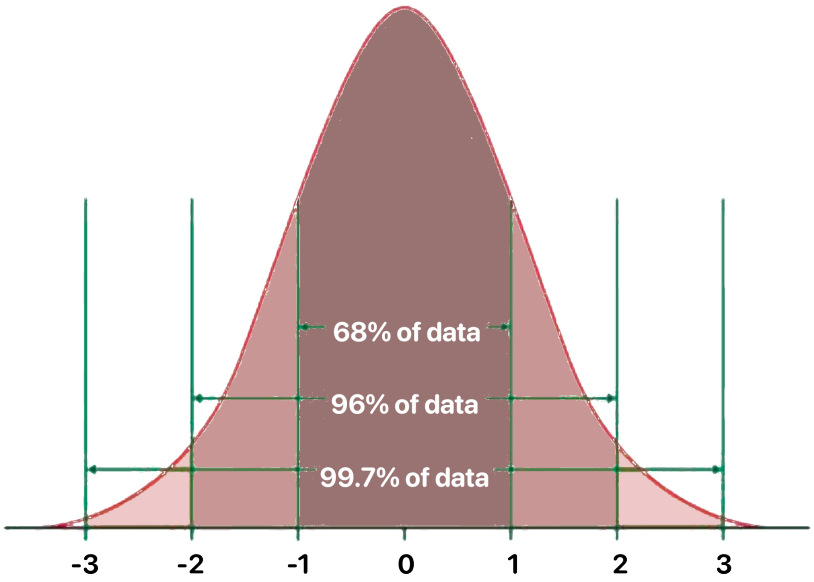



**Figure 7.7:** The 68-96-99.7 rule on the standard normal distribution. [source unknown]

In other words, whenever we observe a sample mean $\overline{X}$ (with sample size $n$) from a normal population with mean $\mu$, we would expect the inequality

$$-k < Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < k$$

9: In general, upper case letters are reserved for a general sample, and lower case letters for a specifically observed sample.

to hold approximately

$$g(k) = \begin{cases} 68.3\% \text{ of the time}, & \text{if } k = 1 \\ 95.5\% \text{ of the time}, & \text{if } k = 2 \\ 99.7\% \text{ of the time}, & \text{if } k = 3 \end{cases}$$

### Confidence Intervals

By re-arranging the terms, we can build a **symmetric** $g(k)$ **confidence interval** (C.I.) **for** $\mu$:

$$\overline{X} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + k\frac{\sigma}{\sqrt{n}} \implies \text{C.I.}(\mu; g(k)) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}}.$$

### Examples

1. Consider a sample $\{X_1, \ldots, X_{64}\}$ from a normal population with known standard deviation $\sigma = 72$. The sample mean is $\overline{X} = 375.2$. Build a symmetric 68.3% confidence interval for $\mu$.

   According to the formula, the symmetric 68.3% confidence interval ($k = 1$) for $\mu$ would be

   $$\text{C.I.}(\mu; 0.683) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}} \equiv 375.2 \pm 1 \cdot \frac{72}{\sqrt{64}},$$

   which is to say

   $$\text{C.I.}(\mu; 0.683) \equiv (375.2 - 9, 375.2 + 9) = (366.2, 384.2).$$

   **VERY IMPORTANT:** this does not say that we are 68.3% sure that the true $\mu$ is between 366.2 and 384.2. What it says is that when a sample of size 64 is taken from a normal population $\mathcal{N}(\mu, 72^2)$ and a symmetric 68.3% confidence interval for $\mu$ is built, $\mu$ will fall between the endpoints of the interval about 68.3% of the time.[10]

2. Build a symmetric 95.5% confidence interval for $\mu$.

   The same formula applies, with $k = 2$:

   $$\text{C.I.}(\mu; 0.955) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}} \equiv 375.2 \pm 2 \cdot \frac{72}{\sqrt{64}},$$

   which is to say

   $$\text{C.I.}(\mu; 0.995) \equiv (375.2 - 18, 375.2 + 18) = (357.2, 393.2).$$

3. Build a symmetric 99.7% confidence interval for $\mu$.

   Again, the same formula applies, with $k = 3$:

   $$\text{C.I.}(\mu; 0.997) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}} \equiv 375.2 \pm 3 \cdot \frac{72}{\sqrt{64}},$$

   which is to say

   $$\text{C.I.}(\mu; 0.995) \equiv (375.2 - 27, 375.2 + 27) = (348.2, 402.2).$$

10: This less than intuitive interpretation of the confidence interval is one of the disadvantages of using the frequentist approach; the analogous concept in Bayesian statistics is called the **credible interval**, which agrees with our naïve expectation of a confidence interval as saying something about how certain we are that the true parameter is in the interval, see [11] and Chapter 25.

Note that the C.I. increases in size with the **confidence level**. The interpretation stays the same, no matter the required confidence level or the parameter of interest.

A 95% C.I. for the mean, for instance, indicates that we would expect 19 out of 20 samples from the same population to produce confidence intervals that contain the true population mean, **on average**.

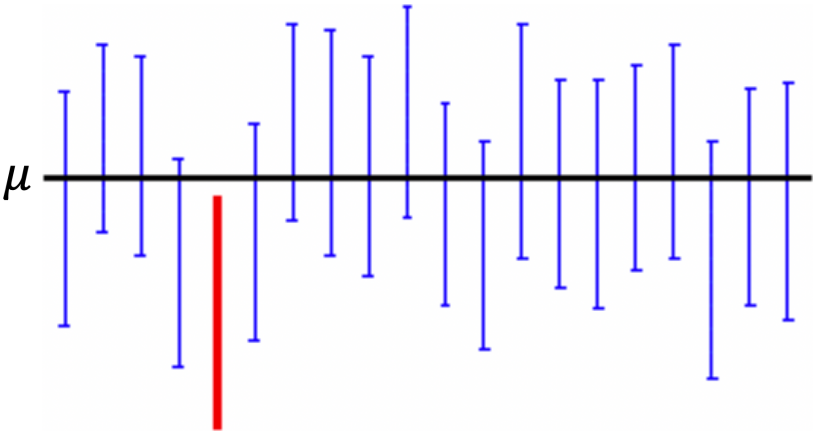



**Figure 7.8:** Frequentist interpretation of confidence intervals: out of 20 experiments, we would expect the true population mean to fall in the confidence interval about 19 times, on average. [source unknown]

**Confidence Interval for $\mu$ when $\sigma$ is Known (Reprise)**

Another approach to C.I. building is to specify the **proportion of the area under $\phi(z)$ of interest**, and then to determine the **critical values** (which is to say, the endpoints of the interval).

Let $\{X_1, \ldots, X_n\}$ be drawn from $\mathcal{N}(\mu, \sigma^2)$. Recall that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

For a **symmetric 95% C.I. for** $\mu$, we need to find $z^* > 0$ such that $P(-z^* < Z < z^*) \approx 0.95$. But the left-hand side of this "equality" can be re-written as

$$\begin{aligned} P(-z^* < Z < z^*) &= \Phi(z^*) - \Phi(-z^*) \\ &= \Phi(z^*) - (1 - \Phi(z^*)) = 2\Phi(z^*) - 1; \end{aligned}$$

we are thus looking for a critical value $z^*$ such that

$$0.95 = 2\Phi(z^*) - 1 \Longrightarrow \Phi(z^*) = \frac{0.95 + 1}{2} = 0.975.$$

From any normal table (or *via* `qnorm(0.975)` in R), we see that $\Phi(1.96) \approx 0.9750$, so that

$$P(-1.96 < Z < 1.96) = P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95.$$

In other words, the inequality

$$-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

holds with probability 0.95, or, equivalently,

$$\text{C.I.}(\mu; 0.95) \equiv \overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

is the **(symmetric) 95% C.I. for $\mu$ when $\sigma$ is known**.

A similar argument shows that

$$\text{C.I.}(\mu; 0.99) \equiv \overline{X} \pm 2.575\frac{\sigma}{\sqrt{n}}$$

is the **(symmetric) 99% C.I. for $\mu$ when $\sigma$ is known**.

#### Examples

1. A sample of size $n = 17$ is selected from a normal population with mean $\mu = -3$ (this is information is unknown to the analysts: this is what they are trying to determine) and standard deviation $\sigma = 2$, which is known.

   The data is shown below:

   ```
   set.seed(0)  # for replicability
   n=17; mu=-3; sigma=2
   (x=rnorm(n,mu,sigma))
   ```

   ```
    [1] -0.4740914 -3.6524667 -0.3404015 -0.4551414 -2.1707171
    [6] -6.0799001 -4.8571341 -3.5894409 -3.0115343  1.8093068
   [11] -1.4728131 -4.5980185 -5.2953140 -3.5789231 -3.5984302
   [16] -3.8230217 -2.4955531
   ```

   Build a 95% confidence interval for $\mu$.

   The sample mean $\overline{x}$ is given by

   ```
   mean(x)
   ```

   ```
   [1] -2.804917
   ```

   The corresponding 95% confidence interval is:

   ```
   lower.bound = mean(x) - 1.96*2/sqrt(17)
   upper.bound = mean(x) + 1.96*2/sqrt(17)
   c(lower.bound,upper.bound)
   ```

   ```
   [1] -3.755657 -1.854178
   ```

   We notice that $\mu = 3$ is indeed found in the confidence interval:

   ```
   lower.bound<mu & mu<upper.bound
   ```

   ```
   [1] TRUE
   ```

2. Repeat the process $M = 1000$ times. How often does $\mu$ fall in the C.I.? We set the seed and the problem parameters.

```
set.seed(0)  # for replicability
n=17; mu=-3; sigma=2; M=1000
```

Next, we initialize the vector which determines if $\mu$ is in the C.I. and the vector which will contain the sample mean for each of the $M = 1000$ repetitions of the experiment:

```
is.mu.in <- c(); sample.means <- c()
```

Finally, we set-up the repetitions: for each sample, we compute the sample mean and the confidence interval bounds, and determine if the true (unknown) value $\mu = 2$ falls in the confidence interval or not.

```
for(j in 1:M){
  x=rnorm(n,mu,sigma)
  sample.means[j] = mean(x)
  lower.bound = sample.means[j] - 1.96*sigma/sqrt(n)
  upper.bound = sample.means[j] + 1.96*sigma/sqrt(n)
  is.mu.in[j] = lower.bound<mu & mu<upper.bound
}
```

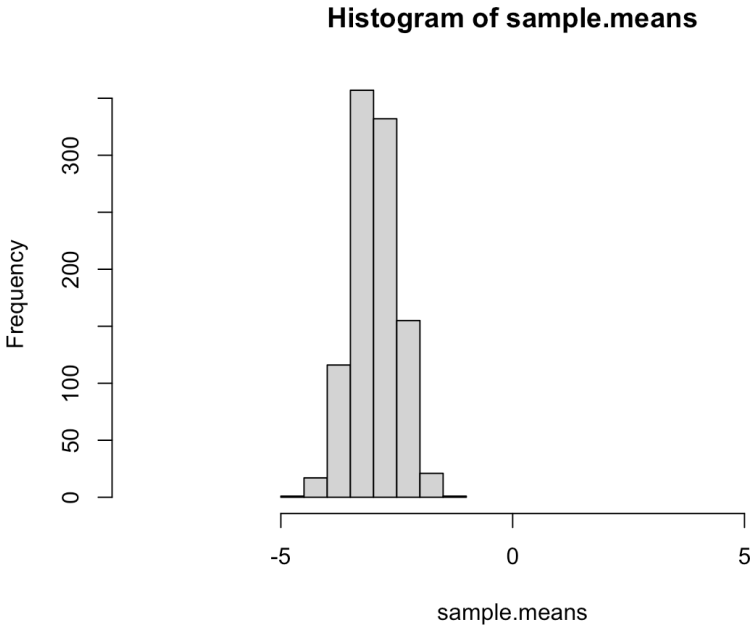
The proportion of the times when it does can thus be obtained *via*:

```
table(is.mu.in)/M
```

```
is.mu.in
FALSE  TRUE
0.055 0.945
```

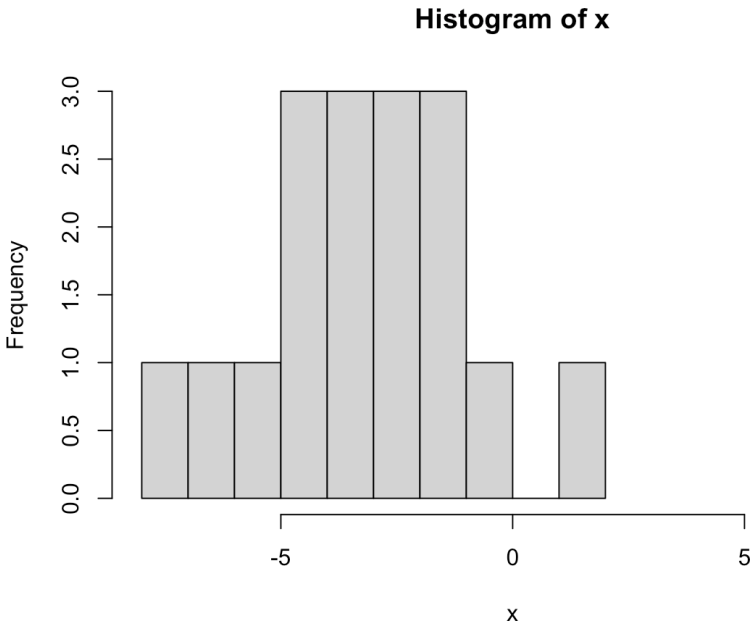
This is indeed very close to 95%. We can also verify the conclusion of the CLT: look at the histogram of the sample means!

```
hist(sample.means, xlim=c(-8,8))
```



This differs markedly from the histogram of the sample values: for instance, the last of the $M = 1000$ samples is distributed as below:

```
hist(x, xlim=c(-8,8))
```




The sample variance is significantly larger than the standard error.

### 7.3.3 Confidence Level

The **confidence level** $1 - \alpha$ is usually expressed in terms of a **small** $\alpha$, so that $\alpha = 0.05$ corresponds to a confidence level of $1 - \alpha = 0.95$.

For $\alpha \in (0, 1)$, the value $z_\alpha$ for which $P(Z > z_\alpha) = \alpha$ is called the $100(1 - \alpha)\%$ **quantiles** of the standard normal distribution. The situation is illustrated in Figure 7.9.

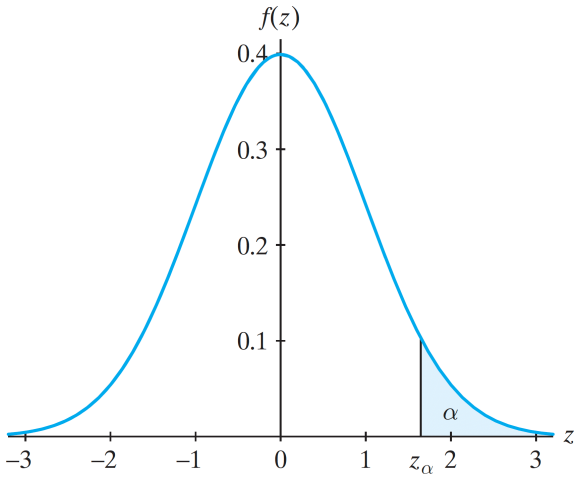



**Figure 7.9:** Quantiles of the standard normal distribution [5].

$$P(Z > z_\alpha) = \alpha$$

$$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$$

For general 2−**sided confidence intervals**,[11] the appropriate quantities are found by solving $P(|Z| > z^*) = \alpha$ for $z^*$. By the properties of $\mathcal{N}(0,1)$,

$$\alpha = P(|Z| > z^*) = 1 - P(-z^* < Z < z^*) = 1 - (2\Phi(z^*) - 1) = 2(1 - \Phi(z^*)),$$

so that

$$\Phi(z^*) = 1 - \alpha/2 \implies z^* = z_{\alpha/2},$$

as illustrated in Figure 7.10.

11: The only ones we will consider in these notes.

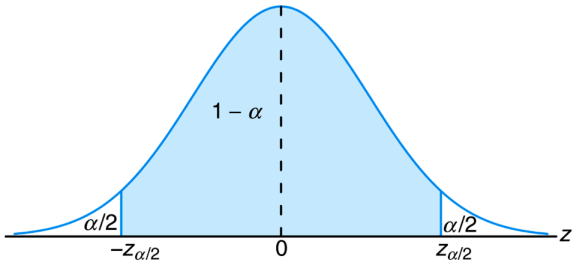



**Figure 7.10:** Two-sided quantiles of the standard normal distribution [5].

The most commonly-used cases are for $\alpha = 0.05$ and $\alpha = 0.01$:

$$P(|Z| > z_{0.025}) = 0.05 \implies z_{0.025} = 1.96$$
$$P(|Z| > z_{0.005}) = 0.01 \implies z_{0.005} = 2.575.$$

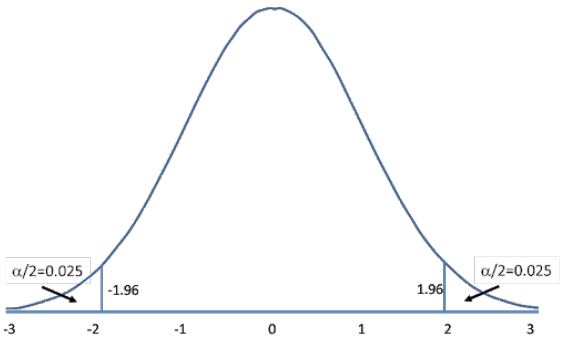



**Figure 7.11:** Two-sided quantiles of the standard normal distribution, for confidence level 0.05.

The symmetric $100(1-\alpha)\%$ C.I. for $\mu$ can thus generally be written as

$$\text{C.I.}(\mu; 1-\alpha)\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

For a given confidence level $\alpha$, **shorter confidence intervals are better** in relation to estimating the mean:

- estimates improve when the sample size $n$ increases;
- estimates improve when $\sigma$ decreases.

For a given sample, if $\alpha_1 > \alpha_2$ then

$$100(1-\alpha_1)\%\text{ C.I.} \subseteq 100(1-\alpha_2)\%\text{ C.I.}$$

For instance, the 95% C.I. built from a sample is always contained in the corresponding 99% C.I.

If the sample comes from a normal population, then the C.I. is **exact**. Otherwise, if $n$ is large, we may use the CLT and get an **approximate** C.I.

**Examples**

- A sample of 9 observations from a normal population with known standard deviation $\sigma = 5$ yields a sample mean $\overline{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean $\mu$.

  The estimate of $\mu$ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)\%$ C.I. is

  $$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

  Thus,

  $$\text{C.I.}(\mu; 0.95) \equiv 19.93 \pm 1.96\frac{5}{\sqrt{9}} = (16.66, 23.20)$$
  $$\text{C.I.}(\mu; 0.99) \equiv 19.93 \pm 2.575\frac{5}{\sqrt{9}} = (15.64, 24.22).$$

- A sample of 25 observations from a normal population with known standard deviation $\sigma = 5$ yields a sample mean $\overline{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean $\mu$.

  The estimate of $\mu$ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)\%$ C.I. are:

  $$\text{C.I.}(\mu; 0.95) \equiv 19.93 \pm 1.96\frac{5}{\sqrt{25}} = (17.97, 21.89)$$
  $$\text{C.I.}(\mu; 0.99) \equiv 19.93 \pm 2.575\frac{5}{\sqrt{25}} = (17.35, 22.51).$$

- A sample of 25 observations from a normal population with known standard deviation $\sigma = 10$ yields a sample mean $\overline{X} = 19.93$. Provide a 95% and a 99% C.I. for the unknown population mean $\mu$.

  The estimate of $\mu$ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)\%$ C.I. are:

  $$\text{C.I.}(\mu; 0.95) \equiv 19.93 \pm 1.96\frac{10}{\sqrt{25}} = (16.01, 23.85)$$
  $$\text{C.I.}(\mu; 0.99) \equiv 19.93 \pm 2.575\frac{10}{\sqrt{25}} = (14.78, 25.08).$$

Note how the confidence intervals are affected by $\alpha$, $n$, and $\sigma$.

### 7.3.4 Sample Size

The **error** $E$ we commit by estimating $\mu$ *via* the sample mean $\overline{X}$ is smaller than $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, with probability $100(1-\alpha)\%$ (in the frequentist interpretation).

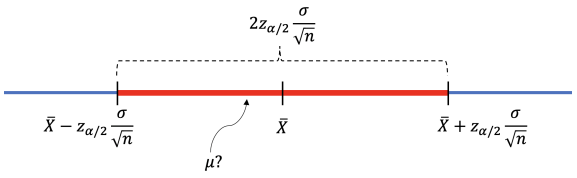



**Figure 7.12:** Estimation error.

At this stage, if we want to **control the error** $E$, the only thing we can really do is control the sample size:[12]

$$E > z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

12: Sampling strategies can also help, but this is a topic for another day (see Chapter 10).

**Examples**

1. A sample $\{X_1, \ldots, X_n\}$ is selected from a normal population with standard deviation $\sigma = 100$. What sample size should be used to insure that the error on the population estimate is at most $E = 10$, at a confidence level $\alpha = 0.05$?

   As long as

   $$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025}\cdot 100}{10}\right)^2 = (19.6)^2 = 384.16,$$

   then the error committed by using $\overline{X}$ to estimate $\mu$ will be at most 10, with 95% probability.

2. Repeat the first example, but with $\sigma = 10$.

   We need

   $$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025}\cdot 10}{10}\right)^2 = (1.96)^2 = 3.8416.$$

3. Repeat the first example, but with $E = 1$.

   We need

   $$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.025}\cdot 100}{1}\right)^2 = (196)^2 = 38416.$$

4. Repeat the first example, but with $\alpha = 0.01$.

   We need

   $$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{z_{0.005}\cdot 100}{10}\right)^2 = (25.75)^2 = 663.0625.$$

The relationship between $\alpha$, $\sigma$, $E$, and $n$ is not always intuitive, but it follows a simple rule.

### 7.3.5 Confidence Intervals for $\mu$ When $\sigma$ is Unknown

So far, we have been in the fortunate situation of sampling from a population with **known** variance $\sigma^2$. What do we do when the population variance is **unknown** (a situation which occurs much more frequently in real world applications)?

The solution is to estimate $\sigma$ using the **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

and the **sample standard deviation** $S = \sqrt{S^2}$; we use $\overline{X}$ instead of $\mu$ since we do not know the value of the latter (that is indeed the parameter whose value we are trying to estimate in the first place).[13]

13: Remember, when $\sigma$ is known (and $n$ is large enough), we already know from the CLT that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is approximately $\mathcal{N}(0,1)$.

If $\sigma$ is unknown, it can be shown that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ follows approximately the **Student $t-$distribution with $n-1$ degrees of freedom**, $t(n-1)$.

Consequently, at a confidence level $\alpha$, we have

$$P\left(-t_{\alpha/2}(n-1) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) \approx 1-\alpha,$$

where $t_{\alpha/2}(n-1)$ is the $100(1-\alpha/2)^{\text{th}}$ quantile of $t(n-1)$. These can be read from pre-compiled tables or computed using the `R` function `qt()`.

Thus,

$$100(1-\alpha)\%\text{C.I. for}\mu \approx \overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}.$$

Equality is reached if the underlying population is normal. For instance, if $\alpha = 0.05$ and $\{X_1, X_2, X_3, X_4, X_5\}$ are samples from a normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$, then $t_{0.025}(5-1) = 2.776$ and

$$P\left(-2.776 < \frac{\overline{X}-\mu}{S/\sqrt{5}} < 2.776\right) = 0.95.$$

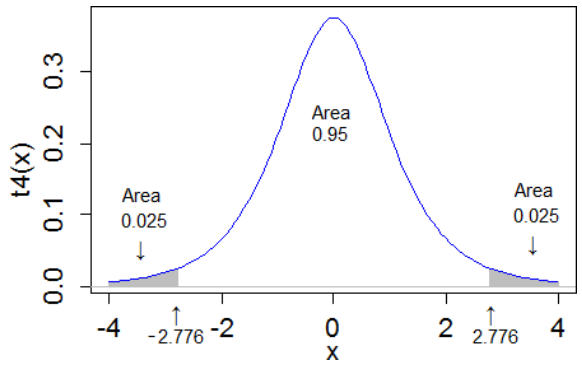



**Figure 7.13:** Critical value for Student distribution with 4 degrees of freedom, at confidence level 0.05. [source unknown]

### Examples

1. For a given year, 9 measurements of ozone concentration are obtained:

   $$3.5, 5.1, 6.6, 6.0, 4.2, 4.4, 5.3, 5.6, 4.4.$$

   Assuming that the measured ozone concentrations follow a normal distribution with variance $\sigma^2 = 1.21$, build a 95% C.I. for the population mean $\mu$. Note that $\overline{X} = 5.01$ and that $S = 0.97$.

We must use the standard normal quantile $z_{\alpha/2} = z_{0.025} = 1.96$ :

$$\overline{X} \pm z_{0.025}\frac{\sigma}{\sqrt{n}} = 5.01 \pm 1.96\frac{\sqrt{1.21}}{\sqrt{9}} = (4.29, 5.73).$$

2. Do the same thing, this time assuming that the true variance of the underlying population is unknown.

We must use the Student quantile $t_{\alpha/2}(n-1) = t_{0.025}(8) = 2.306$:

$$\overline{X} \pm t_{0.025}(n-1)\frac{S}{\sqrt{n}} = 5.01 \pm 2.306\frac{0.97}{\sqrt{9}} = (4.26, 5.76).$$

The quantile value can be obtained from `R` using `qt()`:

```
alpha=0.05
n=9
qt(1-alpha/2,n-1)
```

```
[1] 2.306004
```

3. A sample of size $n = 17$ is selected from a normal population with mean $\mu = -3$ (this is information is unknown to the analysts: this is what they are trying to determine) and unknown standard deviation.

The data is shown below:

```
set.seed(0)  # for replicability
n=17; mu=-3; sigma=2
(x=rnorm(n,mu,sigma))
```

```
 [1] -0.4740914 -3.6524667 -0.3404015 -0.4551414
 [5] -2.1707171 -6.0799001 -4.8571341 -3.5894409
 [9] -3.0115343  1.8093068 -1.4728131 -4.5980185
[13] -5.2953140 -3.5789231 -3.5984302 -3.8230217
[17] -2.4955531
```

Build a 95% confidence interval for $\mu$.

The sample mean $\overline{x}$ is given by

```
mean(x)
```

```
[1] -2.804917
```

The corresponding 95% confidence interval is:

```
lower.bound = mean(x) - qt(1-0.05/2,17-1)*2/sqrt(17)
upper.bound = mean(x) + qt(1-0.05/2,17-1)*2/sqrt(17)
c(lower.bound,upper.bound)
```

```
[1] -3.833222 -1.776612
```

We notice that $\mu = -3$ is indeed found in the confidence interval:

```
lower.bound<mu & mu<upper.bound
```

```
[1] TRUE
```

When the underlying variance is known, the C.I. is **tighter** (smaller), which is only natural as we are more confident about our results when we have more information.

**Note:** what we have seen is that when the underlying distribution is normal, or when it is not normal but the sample size is "large" enough, we can build a C.I. for the population mean, whether the population variance is known or not.

If, however, the underlying population is not normal and the sample size is "small", the approach used in this section cannot guarantee the C.I.'s accuracy.

### 7.3.6 Confidence Intervals for a Proportion

If $X$ is the number of successes in $n$ independent trials, then $X \sim \mathcal{B}(n, p)$, $\mathrm{E}[X] = np$ and $\mathrm{Var}[X] = np(1-p)$, and the point estimator for $p$ is simply $\hat{P} = \frac{X}{n}$.

Since $X$ is a sum of iid random variables, its **standardization**

$$Z = \frac{X - \mu}{\sigma} = \frac{n\hat{P} - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately $\mathcal{N}(0, 1)$, when $n$ is large enough.

Thus, for sufficiently large $n$,

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

Using the construction presented earlier in this section, we conclude that

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

is an **approximate** $100(1-\alpha)\%$ C.I. for $p$. However, this result is not useful in practice because $p$ is unknown, so we use the following approximation instead:

$$\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

#### Examples

1. Two candidates ($A$ and $B$) are running for office. A poll is conducted: 1000 voters are selected randomly and asked for their preference: 52% support $A$, while 48% support their rival, $B$. Provide a 95% C.I. for the support of each candidate.

We use $\alpha = 0.05$ and $\hat{P} = 0.52$. The approximate 95% C.I. for $A$ is thus

$$0.52 \pm 1.96\sqrt{\frac{0.52 \cdot 0.48}{1000}} \approx 0.52 \pm 0.031,$$

while the one for $B$ is $0.48 \pm 0.031$.

2. On the strength of this polling result, a newspaper prints the following headline: "Candidate $A$ Leads Candidate $B$!" Is the headline warranted?

   Although there is a 4–point gap in the poll numbers, the true support for candidate $A$ is in the $48.9\% - 55.1\%$ range, and, the true support for candidate $B$ is in the $44.9\% - 51.1\%$ range, with probability 95% (that is to say, 19 times out of 20).

   Since there is overlap in the confidence intervals, the race is more likely to be a dead heat.

## 7.4 Hypothesis Testing

Consider the following scenario: person A claims they have a fair coin, but for some reason, person B is suspicious of the claim, believing the coin to be biased in favour of tails.

Person B flips the coin 10 times, expecting a low number of heads, which they intend to use as **evidence** against the claim. Let $X = \#$ of Heads.

Suppose $X = 4$. This is less than expected for a binomial random variable $X \sim \mathcal{B}(10, 0.5)$ since $\mathrm{E}[X] = 5$; the results are more in line with a coin for which $P(\text{Head}) = 0.4$.

Does this data constitute evidence against the claim $P(\text{Head}) = 0.5$?

If the coin is fair, then $X \sim \mathcal{B}(10, 0.5)$ and $X = 4$ is still close to $\mathrm{E}[X]$; in fact, $P(X = 4) = 0.205$ (as opposed to $P(X = 5) = 0.246$) so the event $X = 4$ is still quite likely. It would seem that there is no *real* evidence against the claim that the coin is fair.

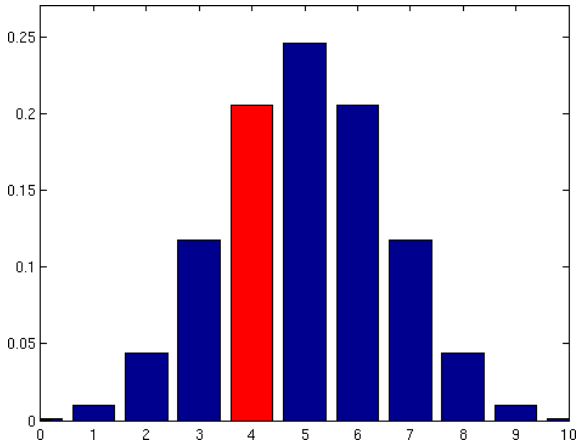



**Figure 7.14:** Binomial distribution for 10 trials, with probability of success 1/2. The probability of exactly 4 successes is highlighted in red.

The way the sentence "*It would seem that there is no evidence against the claim that the coin is fair*" is worded is very important.

We did not reject the claim that $P(\text{Head}) = 0.5$,[14] but this **doesn't mean that, in fact,** $P(\textbf{Head}) = 0.5$. **Not rejecting** (which is not the same as "accepting") a claim is a **weak statement**.

14: Which is to say, that the coin is symmetric.

To see why, let's consider person C, who claims that the coin from the example above has $P(\text{Head}) = 0.3$. Under $X \sim \mathcal{B}(10, 0.3)$, the event $X = 4$ is still quite likely, with $P(X = 4) = 0.22$; we **do not have enough evidence to reject** either $P(\text{Head}) = 0.5$ or $P(\text{Head}) = 0.3$.

However, **rejecting** a claim is a **strong statement**! Let's say that person B convinces person A to flip the coin another 90 times. In the second round of flips, 36 Heads occur, giving a total of 40 Heads out of 100 coin flips.

What can we say now? Does this constitute any evidence against the claim? If so, how much?

Let $Y \sim \mathcal{B}(100, 0.5)$ (i.e.the coin is fair); $Y = 40$ is smaller than what we would expect as $\mathrm{E}[Y] = 50$ if the claim is true, so $Y = 40$ is again more in agreement with $P(\text{Head}) = 0.4$.

But the event $Y = 40$ **does not** lie in the probability mass centre of the distribution as $X = 4$ did; rather, it falls in the **distribution tail** (an area of lower probability).

For $Y \sim \mathcal{B}(100, 0.5)$, $P(Y = 40) = 0.011$.[15] Thus, if the coin is fair, the event $Y = 40$ is quite **unlikely**.

15: Compare this with the previous value $P(X = 4) = 0.205$.

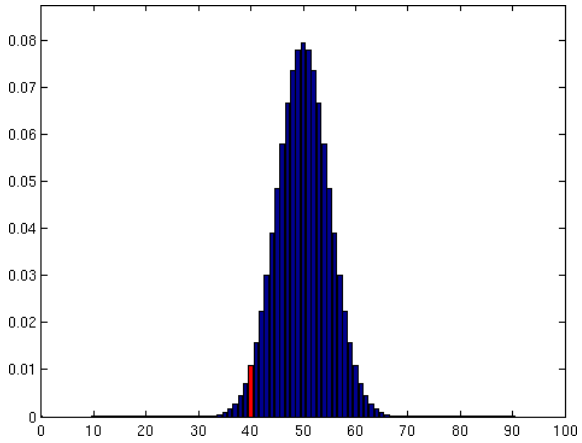



**Figure 7.15:** Binomial distribution for 100 trials, with probability of success 1/2. The probability of exactly 40 successes is highlighted in red.

Values down in the lower tail (or up in the upper tail) provide **some evidence** against the claim. The question is, how much evidence? **How do we quantify it?**

Since values that are "further down the left tail" provide evidence against the claim of a fair coin (in favour of a coin biased against Heads), we will use the actual tail area that goes with the observation: **the smaller the tail area, the greater the evidence against the claim** (and *vice-versa*).

For 4 Heads out of 10 tosses, the evidence is the $p-$**value** $P(X \leq 4)$ , i.e.

$$P(X \leq 4 \mid X \sim \mathcal{B}(10, 0.5)) = 0.377.$$

Thus, if $P(\text{Head}) = 0.5$, the event $X \leq 4$ is still very likely: we would see evidence that extreme (or more) $\approx 38\%$ of the time (simply by chance).

For 40 Heads out of 100 tosses, the evidence is the $p-$**value** $P(Y \leq 40)$,

$$P(Y \leq 40 \mid Y \sim \mathcal{B}(100, 0.5)) = 0.028.$$

Thus, if $P(\text{Head}) = 0.5$, the event $Y \leq 40$ is very unlikely: we would only see evidence that extreme (or more) $\approx 3\%$ of the time. A claim's $p-$value is the **area of the tail** of the distribution's p.d.f. under the assumption that the claim is true:

$$\text{smaller } p-\text{value} \iff \text{more evidence against claim.}$$

### Vocabulary of Hypothesis Testing

A specific language and notation has evolved to describe this approach to "testing hypotheses":

- the "claim" is called the **null hypothesis** and is denoted by $H_0$.
- the "suspicion" is called the **alternative hypothesis** ($H_1$);
- the (random) quantity we use to measure evidence is called a **test statistic** – we need to know its distribution when $H_0$ is true, and
- the $p-$**value** quantifies "the evidence against $H_0$".

Consider the coin tossing situation described previously. The null and alternative hypotheses are

$$H_0 : P(\text{Head}) = 0.5 \quad \text{and} \quad H_1 : P(\text{Head}) < 0.5\,.$$

With $n$ tosses, the test statistic is the number of heads $X$ in $n$ tosses:

- if $n = 10$ and $X = 4$, the $p-$value is

  $$P(X \leq 4 \mid X \sim \mathcal{B}(10, 0.5)) = 0.377,$$

  on the basis of which we would not reject the null hypothesis that the coin was fair;

- if $n = 100$ and $X = 40$, the $p-$value is

  $$P(X \leq 40 \mid X \sim \mathcal{B}(100, 0.5)) = 0.028,$$

  on the basis of which we would reject the null hypothesis that the coin was fair, in favour of the alternative that it was not.

### How Small Does the $p-$Value Need to Be?

We concluded that 37.7% was "not that small", whereas 2.8% was "small enough". How small does a $p-$value need to be before we consider that we have "compelling evidence" against $H_0$?

There is no easy answer to this question.[16] Typically, we look at the probability of making a **type I error**, $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$ :

- if $p-$value $\leq \alpha$, then we **reject** $H_0$ in favour of $H_1$;
- if $p-$value $> \alpha$, then **there is not enough evidence to reject** $H_0$ (**which is not the same as accepting** $H_0$**!**).

16: It depends on many factors, including what penalties we might pay for being wrong.

By convention, we often use $\alpha = 0.01$ or $\alpha = 0.05$.

The use of $p$-values has come under fire recently, as many view them as the root cause of the current **replication crisis**.[17] In this twitter thread K. Carr describes why there is nothing wrong with $p-$values *per se*:

17: The crisis concerns the prevalence of positive findings that are contradicted in subsequent studies [4].

> Don't know what a $p-$VALUE is? Don't know why $p-$VALUES work? Don't know why sometimes $p-$VALUES don't work? **THIS IS THE THREAD FOR YOU!**
>
> **DEFINITION OF A $p-$VALUE:** Assume your theory is false. The $p-$VALUE is the probability of getting an outcome as extreme or even more extreme than what you got in your experiment.
>
> **THE LOGIC OF THE $p-$VALUE:** Assume my theory is false. The probability of getting extreme results should be very small but I got an extreme result in my experiment. Therefore, I conclude that this is strong evidence that my theory is true. That's the logic of the p-value.
>
> **THE $p-$VALUE IS REASONABLE IN THEORY BUT TRICKY IN PRACTICE:** In my opinion, the p-value is just a mathematical version of the way humans think. If we see something that seems unlikely given our beliefs, we often doubt those beliefs. In practice, the p-value can be tricky to use.
>
> **THE $p-$VALUE REQUIRES A GOOD DEFINITION OF WHEN YOUR THEORY IS FALSE:** There are usually an infinite number of ways to define a world where your theory is false. $p-$values often fail when people use overly simplistic mathematical models of the processes that created their data. If the mismatch between their mathematical models of the world and the actual world is too large then the probabilities we compute can become completely disconnected from reality.
>
> **THE $p-$VALUE MAY REQUIRE AN ACCURATE MODEL OF YOU (THE OBSERVER):** The probability of getting the result you got depends on many things. If you sometimes do things like throw out data or repeat measurements then you're part of the system. Your behavior affects the probability of getting your experimental results. Therefore, to be completely realistic, you need to have an ACCURATE model of your own behavior when you gather and analyze data. This is hard and a big part of why the p-value often fails as a tool.
>
> **BY DEFINITION, $p-$VALUES MUST SOMETIMES BE WRONG:** When using $p-$values, we're working off of probabilities. By logic of the p-value itself, even with perfect use, some of your decisions will be wrong. You have to embrace this if you're going to use the $p-$values. Badly defining what it means for your model to be false. Inaccurately modeling the chances of getting your data including your own behaviors. Not treating a p-value as a decision rule that can sometimes be wrong.
>
> These factors all contribute to misuse of the p-value in practice. Hope this cleared some things up for you.
>
> Thanks for coming to my p-value TED talk!

### 7.4.1 Hypothesis Testing in General

A **hypothesis** is a conjecture concerning the value of a population parameter. Hypothesis testing require two **competing** hypotheses:

- a **null hypothesis**, denoted by $H_0$;
- an **alternative hypothesis**, denoted by $H_1$ or $H_A$.

The hypothesis is **tested** by evaluating experimental evidence:

- if the evidence against $H_0$ is **strong enough**, we reject $H_0$ **in favour of** $H_1$, and we say that the evidence against $H_0$ in favour of $H_1$ is **significant**;
- if the evidence against $H_0$ is **not** strong enough, then we fail to reject $H_0$ and we say that the evidence against $H_0$ is **not significant**.

In cases when we fail to reject $H_0$, we **do NOT instead accept** $H_0$; we simply do not have enough evidence to reject $H_0$. We sometimes also say that the evidence is **compatible with** $H_0$.

From a philosophical perspective, the hypotheses should be formulated **prior to the experiment** or the study. The experiment or study is then conducted to evaluate the evidence against the null hypothesis – in order to avoid **data snooping**, it is crucial that we do not formulate $H_1$ after looking at the data.

Scientific hypotheses can be often expressed in terms of whether an effect is found in the data. In this case, we might use the following null hypothesis:

$$H_0 : \text{there is no effect}$$

against the alternative hypothesis:

$$H_1 : \text{there is an effect.}$$

#### Errors in Hypothesis Testing

Two types of errors can be committed when testing $H_0$ against $H_1$:

- if we reject $H_0$ when $H_0$ was in fact true, we have committed a **type I error**;
- if we fail to reject $H_0$ when $H_0$ was in fact is false, we have committed a **type II error**.

| | **Decision:** reject $H_0$ | **Decision:** fail to reject $H_0$ |
|---|---|---|
| **Reality:** $H_0$ is True | Type I Error | No Error |
| **Reality:** $H_0$ is False | No Error | Type II Error |

#### Examples

1. If we conclude that a drug treatment is useful for treating a particular disease, but this is not the case in reality, then we have committed an error of type I.

2. If we cannot conclude that a drug treatment is useful for treating a particular disease, but in reality the treatment is effective, then we have committed an error of type II.

What type of error is worst? It depends on numerous factors.[18]

18: There are other types of errors, but they are not quite of the same nature: when $H_0$ is wrongly rejected, but not for the right (data) reasons, or when $H_0$ is correctly rejected, but $H_1$ is wrongly interpreted; see Wikipedia for more information.

### Power of a Test

The probability of committing a type I error is usually denoted by

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true});$$

that of committing a type II error by

$$\beta = P(\text{ fail to reject } H_0 \mid H_0 \text{ is false}),$$

and that of correctly rejecting $H_0$ by

$$\text{power} = P(\text{reject } H_0 \mid H_0 \text{ is false}) = 1 - \beta.$$

Conventional values of $\alpha$ and $\beta$ are usually 0.05 and 0.2, respectively, although that is not a hard and fast rule.

### Types of Null and Alternative Hypotheses

Let $\mu$ be the population parameter of interest; hypotheses are usually expressed in terms of the values of this parameter (although we could also be testing for other parameters).

The null hypothesis is a **simple hypothesis** of the form:

$$H_0 : \mu = \mu_0,$$

where $\mu_0$ is some candidate value ("simple" means that the parameter is assumed to take on a single value).

The alternative hypothesis $H_1$ is a **composite hypothesis**, i.e. it contains more than one candidate value.

Depending on the context, hypothesis testing takes on one of the following three forms. We test the null hypothesis

$$H_0 : \mu = \mu_0, \quad \text{where } \mu_0 \text{ is a number},$$

against a:

- **two-sided** alternative: $H_1 : \mu \neq \mu_0$;
- **left-sided** alternative: $H_1 : \mu < \mu_0$, or
- **right-sided** alternative: $H_1 : \mu > \mu_0$.

The formulation of the alternative hypothesis depends on the research hypothesis and is determined **prior** to experiment or study.

**Example** Investigators often want to verify if new experimental conditions lead to a change in population parameters.

For instance, an investigator claims that the use of a new type of soil will produce taller plants on average compared to the use of traditional soil. The mean plant height under the use of traditional soil is 20 cm.

1. Formulate the hypotheses to be tested.
2. If another investigator suspects the opposite, that is, that the mean plant height when using the new soil will be smaller than the mean plant height with old soil. What hypotheses should be formulated?
3. A 3rd investigator believes that there will be an effect, but is not sure if the effect with be to produce shorter or taller plants. What hypotheses should be formulated then?

Let $\mu$ represent the mean plant height with the new type of soil. In all three cases, the null hypothesis is $H_0 : \mu = 20$.

The alternative hypothesis depends on the situation:

1. $H_1 : \mu > 20$.
2. $H_1 : \mu < 20$.
3. $H_1 : \mu \neq 20$.

For each $H_1$, the corresponding $p-$values would be computed differently when testing $H_0$ against $H_1$.

### 7.4.2 Test Statistics and Critical Regions

We test a statistical hypothesis we use a **test statistic**. A test statistic is a function of the random sample and the population parameter of interest.

In general, we reject $H_0$ if the value of the test statistic is in the **critical region** or **rejection area** for the test; the critical region is an interval of real numbers.

The critical region is obtained using the definition of errors in hypothesis testing – we select the critical region so that

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$$

is equal to some pre-determined value, such as 0.05 or 0.01.

**Examples** a new curing process developed for a certain type of cement results in a mean compressive strength of 5000 kg/cm$^2$, with a standard deviation of 120 kg/cm$^2$.

We test the hypothesis $H_0 : \mu = 5000$ against the alternative $H_1 : \mu < 5000$ with a random sample of 49 pieces of cement.

Assume that the critical region in this specific instance is $\overline{X} < 4970$, that is, we would reject $H_0$ if $\overline{X} < 4970$.

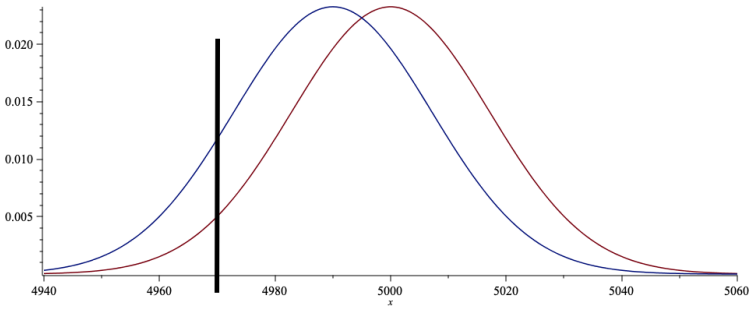
1. Find the probability of committing a type I error when $H_0$ is true.

   By definition, we have

   $$\begin{aligned}\alpha &= P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})\\ &= P(\overline{X} < 4970 \mid \mu = 5000).\end{aligned}$$

   Thus, according to the CLT, we have

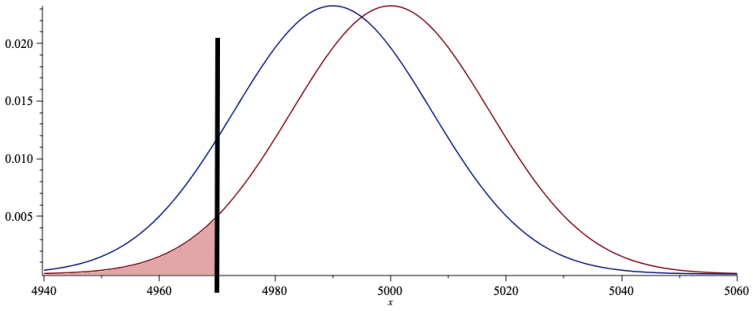
   $$\alpha \approx P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{4970-5000}{120/7}\right) \approx P(Z < -1.75) \approx 0.0401\,.$$




   The sampling distribution of $\overline{X}$ under $H_0$ is shown in **red** in the graph above (and those below): it is a normal distribution with mean = 5000, and standard deviation = 120/7). The sampling distribution of $\overline{X}$ under $H_1$ appears in **blue**: here, a normal distribution with mean = 4990 and standard deviation = 120/7.

   The critical region falls to the left of the vertical **black** line $\overline{X} < 4970$, and the probability of committing a type I error is the area shaded in pale red, below:

   $$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(\overline{X} < 4970 \mid \mu = 5000).$$




   We would thus reject $H_0$ if the observed value of $\overline{X}$ falls to the left of $\overline{X} = 4970$ (in the critical region).

2. Evaluate the probability of committing a type II error if $\mu$ is actually 4990, say (and not 5000, as assumed in $H_0$).

   By definition, we have

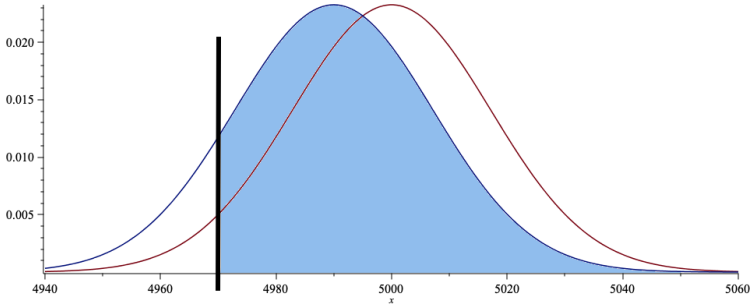
   $$\begin{aligned}\beta &= P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})\\ &= P(\overline{X} > 4970 \mid \mu = 4990).\end{aligned}$$

Thus, according to the CLT, we have

$$\beta = P(\overline{X} > 4970) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{4970 - 4990}{120/7}\right)$$
$$\approx P(Z > -1.17) = 1 - P(Z < -1.17) \approx 0.879\,.$$

The critical region falls to the right of the vertical black line; the probability of committing a type II error is the area in pale blue:

$$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\overline{X} > 4970 \mid \mu = 4990).$$

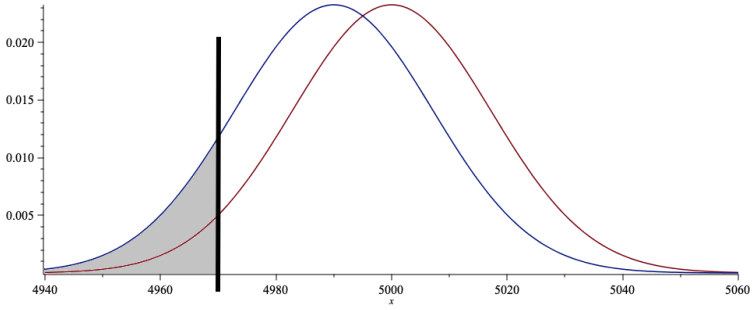



We would thus fail to reject $H_0$ if the observed value of $\overline{X}$ falls to the right of $\overline{X} = 4970$ (outside the critical region).

The power of the test is easily computed as

$$\text{power} = P(\text{reject } H_0 \mid H_0 \text{ is false}) = P(\overline{X} < 4970) = 1 - \beta \approx 0.121,$$

the area shaded in grey below.




3. Evaluate the probability of committing a type II error if $\mu$ is actually 4950, say (and not 5000, as in $H_0$).

By definition, we have

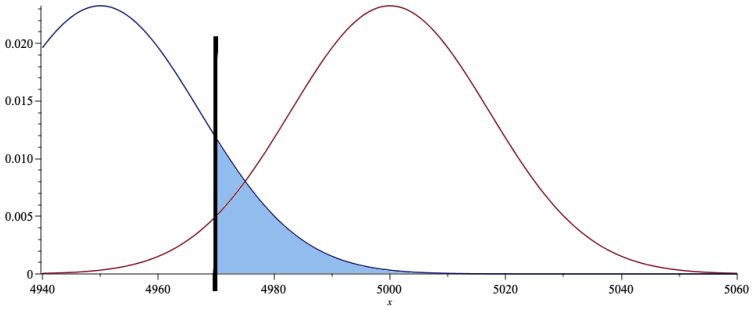
$$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\overline{X} > 4970 | \mu = 4950).$$

Thus, according to the CLT, we have

$$\beta = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{4970 - 4950}{120/7}\right) \approx P(Z > 1.17) \approx 0.121\,.$$

The critical region falls to the right of the vertical black line; the probability of committing a type II error is the area in pale blue:

$$\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\overline{X} > 4970 \mid \mu = 4950).$$




We would thus fail to reject $H_0$ if the observed value of $\overline{X}$ falls to the right of $\overline{X} = 4970$ (outside the critical region).

The probability of making a type II error is much larger in the first case, which means that the threshold $\overline{X} = 4970$ is not ideal in that situation.

### 7.4.3 Test for a Mean

Suppose $X_1, \ldots, X_n$ is a random sample from a population with mean $\mu$ and variance $\sigma^2$, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean:

- if the population is normal, then $\overline{X} \overset{\text{exact}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$;
- if the population is **not** normal, then as long as $n$ is **large enough**, $\overline{X} \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$.

We start by assuming that the population variance $\sigma^2$ is **known**, and that the hypothesis concerns the **unknown** population mean $\mu$.

#### Explanation: Left-Sided Alternative

Consider the unknown population mean $\mu$. Suppose that we test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu < \mu_0,$$

where $\mu_0$ is some candidate value for $\mu$. To evaluate the evidence against $H_0$, we compare $\overline{X}$ to $\mu_0$. Under $H_0$,

$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} \mathcal{N}(0, 1).$$

We say that $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ is the observed value of the $Z$**−test statistic** $Z_0$.

If $z_0 < 0$, we have evidence that $\mu < \mu_0$. However, we only reject $H_0$ in favour of $H_1$ if the evidence is **significant**, which is to say, if

$$z_0 \leq -z_\alpha, \text{ at a level of significance } \alpha.$$

The corresponding $p$**−value** for this test is the probability of observing evidence that is as (or more) extreme than our current evidence in favour of $H_1$, assuming that $H_0$ is true (that is, simply by chance).[19] The **decision rule** for the left-sided test is thus

- if the $p-$value $\leq \alpha$, we **reject** $H_0$ **in favour of** $H_1$;
- if the $p-$value $> \alpha$, we **fail to reject** $H_0$.

19: "Even more extreme", in this case, means further to the left, so that $p$-value $= P(Z \leq z_0) = \Phi(z_0)$, where $z_0$ is the observed value for the $Z$-test statistic.

Formally, the **left-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu < \mu_0;$$

at significance $\alpha$, if $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha$, we reject $H_0$ in favour of $H_1$, as below.

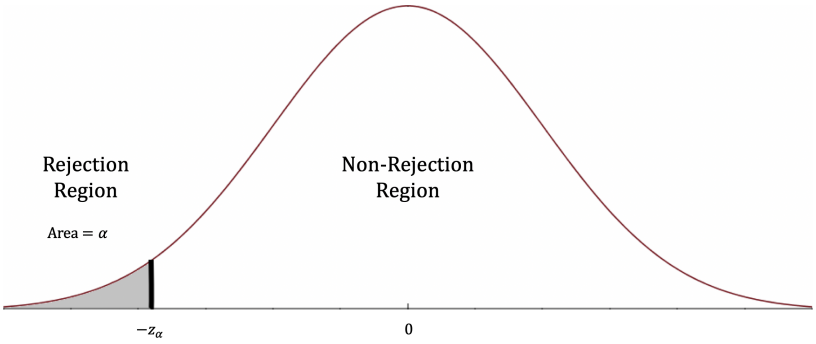



**Figure 7.16:** Critical test region, left-sided test.

An equivalent **right-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu > \mu_0;$$

at significance $\alpha$, if $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$, we reject $H_0$ in favour of $H_1$, as below.

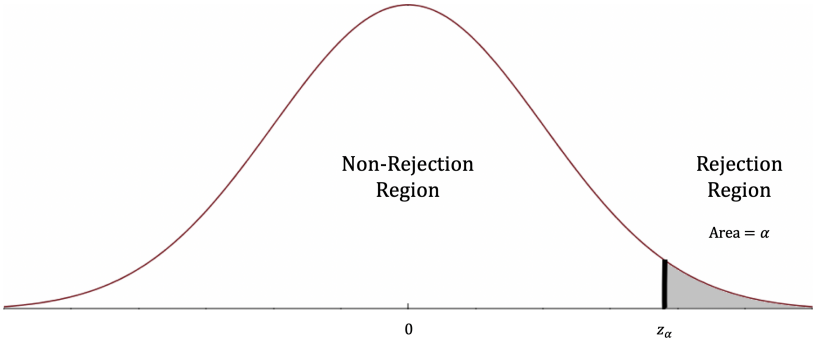



**Figure 7.17:** Critical test region, right-sided test.

The **two-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0;$$

at significance $\alpha$, if $|z_0| = \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2}$, we reject $H_0$ in favour of $H_1$.

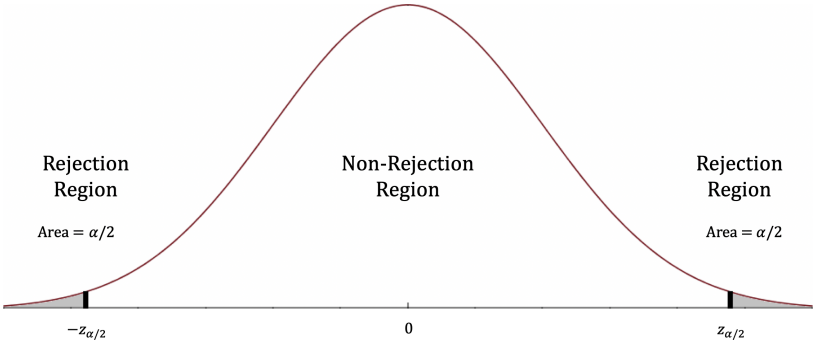



**Figure 7.18:** Critical test region, two-sided test.

The **procedure** to test for $H_0 : \mu = \mu_0$ requires 6 steps.

**Step 1:** set $H_0 : \mu = \mu_0$.

**Step 2:** select an alternative hypothesis $H_1$.[20] Depending on the context, we choose one of these alternatives:

- $H_1 : \mu < \mu_0$ (one-sided test);
- $H_1 : \mu > \mu_0$ (one-sided test);
- $H_1 : \mu \neq \mu_0$ (two-sided test).

20: What we are trying to show using the data.

**Step 3:** choose $\alpha = P(\text{type I error})$, typically $\alpha \in \{0.01, 0.05\}$.

**Step 4:** for the observed sample $\{x_1, \ldots, x_n\}$, compute the observed value of the test statistics $z_0 = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.

**Step 5:** determine the critical region according to:

| Alternative Hypothesis | Critical Region |
|---|---|
| $H_1 : \mu > \mu_0$ | $z_0 > z_\alpha$ |
| $H_1 : \mu < \mu_0$ | $z_0 < -z_\alpha$ |
| $H_1 : \mu \neq \mu_0$ | $\lvert z_0 \rvert > z_{\alpha/2}$ |

where $z_\alpha$ is the critical value satisfying $P(Z > z_\alpha) = \alpha$ , for $Z \sim \mathcal{N}(0, 1)$. The critical values are displayed below for convenience.

| $\alpha$ | $z_\alpha$ | $z_{\alpha/2}$ |
|---|---|---|
| 0.05 | 1.645 | 1.960 |
| 0.01 | 2.327 | 2.576 |

**Step 6:** compute the associated $p-$value according to:

| Alternative Hypothesis | Critical Region |
|---|---|
| $H_1 : \mu > \mu_0$ | $P(Z > z_0)$ |
| $H_1 : \mu < \mu_0$ | $P(Z < z_0)$ |
| $H_1 : \mu \neq \mu_0$ | $2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$ |

**Decision Rule:** as above,

- if the $p-$value $\leq \alpha$, **reject** $H_0$ **in favour of** $H_1$;
- if the $p-$value $> \alpha$, **fail to reject** $H_0$.

A few examples will clarify the procedure.

#### Examples

1. Components are manufactured to have strength normally distributed with mean $\mu = 40$ units and standard deviation $\sigma = 1.2$ units. The manufacturing process has been modified, and an increase in mean strength is claimed (the standard deviation remains the same).

A random sample of $n = 12$ components produced using the modified process had the following strengths:

```
42.5, 39.8, 40.3, 43.1, 39.6, 41.0,
39.9, 42.1, 40.7, 41.6, 42.1, 40.8.
```

Does the data provide strong evidence that the mean strength now exceeds 40 units? Use $\alpha = 0.05$.

We follow the outlined procedure to test for $H_0 : \mu = 40$ against $H_1 : \mu > 40$.

The observed value of the sample mean is $\overline{x} = 41.125$. Hence,

$$\begin{aligned} p\text{−value} &= P(\overline{X} \geq \overline{x}) = P(\overline{X} \geq 41.125) \\ &= P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{41.125 - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= P(Z \geq 3.25) \approx 0.006. \end{aligned}$$

As the $p$−value is smaller than $\alpha$, we reject $H_0$ in favour of $H_1$.

Another way to see this is that if the model '$\mu = 40$' is true, then it is very unlikely that we would observe the event $\{\overline{X} \geq 41.125\}$ entirely by chance, and so the manufacturing process likely has an effect in the claimed direction.

2. A set of scales works properly if the measurements differ from the true weight by a normally distributed random error term with standard deviation $\sigma = 0.007$ grams. Researchers suspect that the scale is systematically adding to the weights.

   To test this hypothesis, $n = 10$ measurements are made on a 1.0g "gold-standard" weight, giving a set of measurements which average out to 1.0038g. Does this provide evidence that the scale adds to the measurement weights? Use $\alpha = 0.05$ and 0.01.

   Let $\mu$ be the weight that the scale would record in the absence of random error terms. We test for $H_0 : \mu = 1.0$ against $H_1 : \mu > 1.0$.

   The observed test statistic is $z_0 = \frac{1.0038-1.0}{0.007/\sqrt{10}} \approx 1.7167$. Since

   $$z_{0.05} = 1.645 < z_0 = 1.7167 \leq z_{0.01} = 2.327,$$

   we reject $H_0$ for $\alpha = 0.05$, but we fail to reject $H_0$ for $\alpha = 0.01$.

   Case closed. Right?

3. In the previous example, assume that we are interested in whether the scale works properly, which means that the investigators think there might be some systematic misreading, but they are not sure in which direction the misreading would occur. Does the sample data provide evidence that the scale is systematically biased? Use $\alpha = 0.05$ and 0.01.

Let $\mu$ be as in the previous example. We test for $H_0 : \mu = 1.0$ against $H_1 : \mu \neq 1.0$.

The test statistic is still $z_0 = 1.7167$; since $|z_0| \leq z_{\alpha/2}$ for both $\alpha = 0.05$ and $\alpha = 0.01$, we fail to reject $H_0$ at either $\alpha = 0.05$ or $\alpha = 0.01$.

Thus, our "reading" of the test statistic depends on what type of alternative hypothesis we have selected (and so, on the overall context).

4. The marks for an "average" class are normally distributed with mean 60 and variance 100. Nine students are selected from the class; their average mark is 55. Is this subgroup "below average"?

   Let $\mu$ be the true mean of the subgroup. We are testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$.

   The observed sample test statistic is

   $$z_0 = \frac{55 - 60}{10/\sqrt{9}} = -1.5.$$

   The corresponding $p-$value is

   $$P(\overline{X} \leq 55) = P(Z \leq -1.5) = 0.07.$$

   Thus there is not enough evidence to reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ or $\alpha = 0.01$.

5. We consider the same set-up as in the previous example, but this time the sample size is $n = 100$, not 9. Is there some evidence to suggest that this subgroup of students is 'below average'?

   Let $\mu$ be as before. We are still testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$, but this time the observed sample test statistic is

   $$z_0 = \frac{55 - 60}{10/\sqrt{100}} = -5.$$

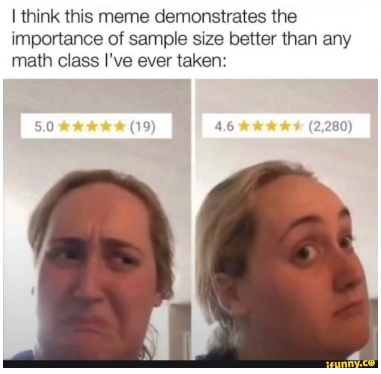
   The corresponding $p-$value is

   $$P(\overline{X} \leq 55) = P(Z \leq -5) \approx 0.00.$$

   Thus we reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ or $\alpha = 0.01$.

The lesson from the last example is that the **sample size plays a role**; in general, an estimate obtained from a larger (representative) sample is more likely to be generalizable to the population as a whole.[21]

21: Or as the iFunny meme has it. . .



### Tests and Confidence Intervals

It is becoming more and more common for analysts to bypass the computation of the $p-$value altogether, in favour of a confidence interval based approach.[22]

22: In order to avoid the controversy surrounding the crisis of replication?

For a given $\alpha$, we reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ if, and only if, $\mu_0$ is **not** in the $100(1-\alpha)\%$ C.I. for $\mu$.

**Example** A manufacturer claims that a type of engine uses 20 gallons of fuel to operate for one hour. It is known from previous studies that this amount is normally distributed with variance $\sigma^2 = 25$ and mean $\mu$.

A sample of size $n = 9$ has been taken and the following value has been observed for the mean amount of fuel per hour: $\overline{X} = 23$. Should we accept the manufacturer's claim? Use $\alpha = 0.05$.

We test for $H_0 : \mu = 20$ against $H_1 : \mu \neq 20$. The observed sample test statistic is

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{23 - 20}{5/\sqrt{9}} = 1.8.$$

For a 2−sided test with $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$. Since $|z_0| \leq z_{0.025}$, $z_0$ is not in the critical region, and we do not reject $H_0$.

The advantage of the **confidence interval** approach is that it allows analysts to test for various claims **simultaneously**. Since we know the variance of the underlying population, an approximate $100(1-\alpha)\%$ C.I. for $\mu$ is given by

$$\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n} = 23 \pm 1.96 \cdot 5/\sqrt{9} = (19.73; 26.26).$$

Based on the data, we would thus not reject the claim that $\mu = 20$, $\mu = 19.74$, $\mu = 26.20$, etc.

### Test for a Mean with Unknown Variance

If the data is normal and $\sigma$ is unknown, we can estimate it *via* the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

As we have seen for confidence intervals, the test statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

follows a **Student's $t-$distribution with $n-1$ df**.

We can follow the same steps as for the test with known variance, with the modified critical regions and $p-$values:

| Alternative Hypothesis | Critical Region |
|---|---|
| $H_1 : \mu > \mu_0$ | $t_0 > t_\alpha(n-1)$ |
| $H_1 : \mu < \mu_0$ | $t_0 < -t_\alpha(n-1)$ |
| $H_1 : \mu \neq \mu_0$ | $\lvert t_0\rvert > t_{\alpha/2}(n-1)$ |

where

$$t_0 = \frac{\overline{x} - \mu_0}{S/\sqrt{n}}$$

and $t_\alpha(n-1)$ is the $t-$value satisfying

$$P(T > t_\alpha(n-1)) = \alpha$$

for $T \sim t(n-1)$. The corresponding $p-$values are given in the table below.

| **Alternative Hypothesis** | $p-$**Value** |
|---|---|
| $H_1 : \mu > \mu_0$ | $P(T > t_0)$ |
| $H_1 : \mu < \mu_0$ | $P(T < t_0)$ |
| $H_1 : \mu \neq \mu_0$ | $2 \cdot \min\{P(T > t_0), P(T < t_0)\}$ |

**Example** Consider the following observations, taken from a normal population with unknown mean $\mu$ and variance:

```
18.0, 17.4, 15.5, 16.8, 19.0, 17.8, 17.4, 15.8,
17.9, 16.3, 16.9, 18.6, 17.7, 16.4, 18.2, 18.7.
```

Conduct a right-side hypothesis test for $H_0 : \mu = 16.6$ vs. $H_1 : \mu > 16.6$, using $\alpha = 0.05$.

The sample size, sample mean, and sample variance are $n = 16$, $\overline{X} = 17.4$ and $S = 1.078$, respectively.

Since the variance $\sigma^2$ is unknown, the observed sample test statistics of interest is

$$t_0 = \frac{\overline{x} - \mu_0}{S/\sqrt{n}} = \frac{17.4 - 16.6}{1.078/4} \approx 2.968,$$

and the corresponding $p-$value is

$$p-\text{value } = P(\overline{X} \geq 17.4) = P(T > 2.968),$$

where $T \sim t(n-1) = t(\nu) = t(15)$.

From the $t-$tables (or by using the `R` function `qt()`), we see that

$$P\left(T(15) \geq 2.947\right) \approx 0.005, \; P\left(T(15) \geq 3.286\right) \approx 0.0025.$$

The $p-$value thus lies in the interval $(0.0025, 0.005)$; in particular, the $p-$value $\leq 0.05$, which is strong evidence against $H_0 : \mu = 16.6$.

### 7.4.4 Test for a Proportion

The principle for proportions is pretty much the same, as we can see in the next example.

**Example** A group of 100 adult American Catholics were asked the following question: "Do you favour allowing women into the priesthood?" 60 of the respondents independently answered 'Yes'; is the evidence strong enough to conclude that more than half of American Catholics favour allowing women to be priests?

Let $X$ be the number of people who answered 'Yes'. We assume that $X \sim \mathcal{B}(100, p)$, where $p$ is the true proportion of American Catholics who favour allowing women to be priests.

We test for $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$. Under $H_0$, $X \sim \mathcal{B}(100, 0.5)$.

The $p-$value that corresponds to the observed sample is

$$\begin{aligned} P(X \geq 60) &= 1 - P(X < 60) = 1 - P(X \leq 59) \\ &\approx 1 - P\left(\frac{X+0.5-np}{\sqrt{np(1-p)}} \leq \frac{59+0.5-50}{\sqrt{25}}\right) \\ &\approx 1 - P(Z \leq 1.9) = 0.0287, \end{aligned}$$

where the $+\mathbf{0.5}$ comes from the correction to the normal approximation of the binomial distribution (see Section 6.3.6 for details).

Thus, we would reject $H_0$ at $\alpha = 0.05$, but not at $\alpha = 0.01$.

### 7.4.5 Two-Sample Tests

Up to this point, we have only tested hypotheses about populations by evaluating the evidence provided by a single sample of observations. **Two-sample tests** allows analysts to compare two populations.[23]

23: These populations are potentially distinct.

#### Paired Test

Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean $\mu_1$ and unknown variance $\sigma^2$; let $X_{2,1}, \ldots, X_{2,n}$ be a random sample from a normal population with unknown mean $\mu_2$ and unknown variance $\sigma^2$, with both populations **not necessarily independent** of one another.[24] We would like to test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

24: It is possible that the 2 samples arise from the same population, or represent two different measurements on the same units, say.

In order to do so, we compute the differences $D_i = X_{1,i} - X_{2,i}$ and consider the $t-$test (as we do not know the variance). The test statistic is

$$T_0 = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t(n-1),$$

where

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \quad \text{and} \quad S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2.$$

**Example** The knowledge of basic statistical concepts for $n = 10$ engineers was measured on a scale from $0 - 100$ *before* and *after* a short course in statistical quality control. The result are as follows:

| **Engineer** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before $X_{1,i}$ | 43 | 82 | 77 | 39 | 51 | 66 | 55 | 61 | 79 | 43 |
| After $X_{2,i}$ | 51 | 84 | 74 | 48 | 53 | 61 | 59 | 75 | 82 | 48 |

Let $\mu_1$ and $\mu_2$ be the mean score before and after the course, respectively.

Assuming the underlying scores are normally distributed, conduct a test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$.

The differences $D_i = X_{1,i} - X_{2,i}$ are:

| **Engineer** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before $X_{1,i}$ | 43 | 82 | 77 | 39 | 51 | 66 | 55 | 61 | 79 | 43 |
| After $X_{2,i}$ | 51 | 84 | 74 | 48 | 53 | 61 | 59 | 75 | 82 | 48 |
| Difference $D_i$ | −8 | −2 | 3 | −9 | −2 | 5 | −4 | −14 | −3 | −5 |

The observed sample mean is $\overline{d} = -3.9$, and the observed sample variance is $s_D^2 = 31.21$.

The test statistic is:

$$T_0 = \frac{\overline{D} - 0}{S_D/\sqrt{n}} \sim t(n-1),$$

with observed value:

$$t_0 = \frac{-3.9}{\sqrt{31.21/10}} \approx -2.21.$$

We compute

$$P(\overline{D} \leq -3.9) = P(T(9) \leq -2.21) = P(T(9) > 2.21).$$

But $t_{0.05}(9) = 1.833 < t_0 = 2.21 < t_{0.01}(9) = 2.821$, so we reject $H_0$ at $\alpha = 0.05$, but not at $\alpha = 0.01$.

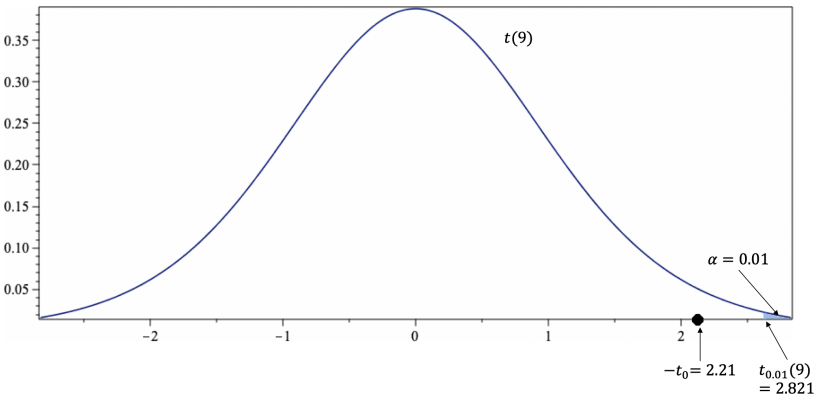



**Figure 7.19:** Critical test regions for the right-sided test, with $n = 10$ observations: confidence levels 0.05 (left) and 0.01 (right).

### Unpaired Test

Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean $\mu_1$ and variance $\sigma_1^2$; let $Y_{2,1}, \ldots, Y_{2,m}$ be a random sample

from a normal population with unknown mean $\mu_2$ and variance $\sigma_2^2$, with both populations **independent** of one another.

We want to test for

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2.$$

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ , $\overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$ . As always, the observed values are denoted by lower case letters: $\overline{x}$, $\overline{y}$.

**When the Variances $\sigma_1^2$ and $\sigma_2^2$ are Known**

We can follow the same steps as for the earlier test, with some modifications:

| Alternative Hypothesis | Critical Region |
|---|---|
| $H_1 : \mu_1 > \mu_2$ | $z_0 > z_\alpha$ |
| $H_1 : \mu_1 < \mu_2$ | $z_0 < -z_\alpha$ |
| $H_1 : \mu_1 \neq \mu_2$ | $\lvert z_0\rvert > z_{\alpha/2}$ |

where

$$z_0 = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}},$$

and $z_\alpha$ satisfies $P(Z > z_\alpha) = \alpha$ , for $Z \sim \mathcal{N}(0, 1)$.

| Alternative Hypothesis | $p-$Value |
|---|---|
| $H_1 : \mu_1 > \mu_2$ | $P(Z > z_0)$ |
| $H_1 : \mu_1 < \mu_2$ | $P(Z < z_0)$ |
| $H_1 : \mu_1 \neq \mu_2$ | $2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$ |

**Example** A sample of $n = 100$ Albertans yields a sample mean income of $\overline{X} = 33,000\$$. A sample of $m = 80$ Ontarians yields $\overline{Y} = 32,000\$$. From previous studies, it is known that the population income standard deviations are, respectively, $\sigma_1 = 5000\$$ in Alberta and $\sigma_2 = 2000\$$ in Ontario. Do Albertans earn more than Ontarians, on average?

We test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$. The observed difference is $\overline{X} - \overline{Y} = 1000$; the observed test statistic is

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} = \frac{1000}{\sqrt{5000^2/100 + 2000^2/80}} = 1.82;$$

the corresponding $p-$value is

$$P\left(\overline{X} - \overline{Y} > 1000\right) = P(Z > 1.82) = 0.035,$$

and so we reject $H_0$ when $\alpha = 0.05$, but not when $\alpha = 0.01$.

**When the Variances $\sigma_1^2$ and $\sigma_2^2$ are Unknown (Small Samples)**

In this case, the modifications are:

| Alternative Hypothesis | Critical Region |
|---|---|
| $H_1 : \mu_1 > \mu_2$ | $t_0 > t_\alpha(n+m-2)$ |
| $H_1 : \mu_1 < \mu_2$ | $t_0 < -t_\alpha(n+m-2)$ |
| $H_1 : \mu_1 \neq \mu_2$ | $\lvert t_0 \rvert > t_{\alpha/2}(n+m-2)$ |

where

$$t_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}} \text{ and } S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

$t_\alpha(n+m-2)$ satisfies $P(T > t_\alpha(n+m-2)) = \alpha$ , and $T \sim t(n+m-2)$.

| Alternative Hypothesis | $p-$Value |
|---|---|
| $H_1 : \mu_1 > \mu_2$ | $P(T > t_0)$ |
| $H_1 : \mu_1 < \mu_2$ | $P(T < t_0)$ |
| $H_1 : \mu_1 \neq \mu_2$ | $2 \cdot \min\{P(T > t_0), P(T < t_0)\}$ |

**Example** A researcher wants to test whether, on average, a new fertilizer yields taller plants. Plants were divided into two groups: a control group treated with an old fertilizer and a study group treated with the new fertilizer. The following data are obtained:

| Sample Size | Sample Mean | Sample Variance |
|---|---|---|
| $n = 8$ | $\overline{X} = 43.14$ | $S_1^2 = 71.65$ |
| $m = 8$ | $\overline{Y} = 47.79$ | $S_2^2 = 52.66$ |

Test for $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 < \mu_2$.

The observed difference is $\overline{X} - \overline{Y} = -4.65$ and the **pooled sampled variance** is

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{7(71.65) + 7(52.66)}{8+8-2} = 62.155 = 7.88^2.$$

The observed test statistic is thus

$$t_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}} = \frac{-4.65}{7.88\sqrt{1/8 + 1/8}} = -1.18;$$

the corresponding $p-$value is

$$\begin{aligned} P\left(\overline{X} - \overline{Y} < -4.65\right) &= P(T(14) < -1.18) \\ &= P(T(14) > 1.18) \in (0.1, 0.25) \end{aligned}$$

(according to the table), and we do not reject $H_0$ when $\alpha = 0.05$, or when $\alpha = 0.01$.

**When the Variances $\sigma_1^2$ and $\sigma_2^2$ are Unknown (Large Samples)**

In this case, the modifications are:

| Alternative Hypothesis | Critical Region |
|---|---|
| $H_1 : \mu_1 > \mu_2$ | $z_0 > z_\alpha$ |
| $H_1 : \mu_1 < \mu_2$ | $z_0 < -z_\alpha$ |
| $H_1 : \mu_1 \neq \mu_2$ | $\lvert z_0 \rvert > z_{\alpha/2}$ |

where

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}},$$

and $z_\alpha$ satisfies $P(Z > z_\alpha) = \alpha$ , for $Z \sim \mathcal{N}(0, 1)$.

| Alternative Hypothesis | $p-$Value |
|---|---|
| $H_1 : \mu_1 > \mu_2$ | $P(Z > z_0)$ |
| $H_1 : \mu_1 < \mu_2$ | $P(Z < z_0)$ |
| $H_1 : \mu_1 \neq \mu_2$ | $2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$ |

**Example** Consider the same set-up as in the previous example, but with larger sample sizes: $n = m = 100$. Now test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$.

The observed difference is (still) $-4.65$. The observed test statistic is

$$z_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}} = \frac{-4.65}{\sqrt{71.65^2/100 + 52.66^2/100}} = -4.17;$$

the corresponding $p-$value is

$$P\left(\overline{X} - \overline{Y} < -4.65\right) = P(Z < -4.17) \approx 0.0000;$$

and we reject $H_0$ when either $\alpha = 0.05$ or $\alpha = 0.01$.

### 7.4.6 Difference of Two Proportions

As always, we can transfer these tests to proportions, using the normal approximation to the binomial distribution.

For instance, to test for $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ in samples of size $n_1$, $n_2$, respectively, we use the **observed sample difference of proportions**

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2}},$$

where $\hat{p}$ is the **pooled proportion**

$$\hat{p} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2.$$

and the $p-$value is, as always, $2 \cdot \min\{P(Z > z_0), P(Z < z_0)\}$.

### 7.4.7 Hypothesis Testing with R

There are built-in functions in R that allow for hypothesis testing.

- We test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ when $\sigma$ is unknown (**two-sided $t-$test**) using:

```
t.test(x,mu=mu.0)
```

- We test for $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ when $\sigma$ is unknown (**right-sided $t-$test**) using:

```
t.test(x,mu=mu.0,alternative="greater")
```

- We test for $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$ when $\sigma$ is unknown (**left-sided $t-$test**) using:

```
t.test(x,mu=mu.0,alternative="less")
```

- We test for $H_0 : \mu_1 = \mu_2$ against$H_1 : \mu_1 \neq \mu_2$ in case of two independent samples, when variances are unknown but equal (**two-sample two-sided $t-$test**) using:

```
t.test(x,y,var.equal=TRUE)
```

- We test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ in case of two independent samples, when variances are unknown but equal (**two-sample right-sided $t-$test**) using:

```
t.test(x,y,var.equal=TRUE,alternative="greater")
```

- We test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ in case of two independent samples, when variances are unknown but equal (**two-sample left-sided $t-$test**) using:

```
t.test(x,y,var.equal=TRUE,alternative="less")
```

For all these tests, we **reject the null hypothesis $H_0$ at significance level $\alpha$** if the $p-$value of the test is **below** $\alpha$.[25]

If the $p-$value of the test is **greater** than the significance level $\alpha$, then we **fail to reject the null hypothesis $H_0$ at significance level $\alpha$**.[26]

Note that the $p-$value for the test will appear in the output, but it can also be computed directly using the appropriate formula. The corresponding 95% confidence intervals also appear in the output.

25: Which means that the probability of wrongly rejecting $H_0$ when $H_0$ is in fact true is below $\alpha$, usually taken to be 0.05 or 0.01).

26: Which, it is worth recalling, is not the same as accepting the null hypothesis.

#### Artificial Examples

1. Let's say that we have a small dataset with $n = 7$ observations:

```
x=c(4,5,4,6,4,4,5)
```

Let $\mu_X$ be the true mean of whatever distribution the sample came from. Is it conceivable that $\mu_X = 5$?

We can test for $H_0 : \mu_X = 5$ against $H_1 : \mu_X \neq 5$ simply by calling:

```
t.test(x,mu=5)
```

```
    One Sample t-test
data:  x
t = -1.4412, df = 6, p-value = 0.1996
alternative hypothesis: true mean is not equal to 5

95 percent confidence interval:
3.843764 5.299093

sample estimates:
mean of x
4.571429
```

All the important information is in the output: the critical $t-$value from Student's $T-$distribution with $n-1=6$ degrees of freedom $t^*=-1.4412$, the probability of wrongly rejecting $H_0$ if it was in fact true ($p-$value $=0.1996$), and the 95% confidence interval $(3.843764, 5.299093)$ for $\mu_X$, whose point estimate is $\overline{x}=4.571429$.

Since the $p-$value is greater than $\alpha=0.05$, we fail to reject the null hypothesis that $\mu_X=5$; there is not enough evidence in the data to categorically state that $\mu_X \neq 5$.[27]

27: Is it problematic that the sample size $n=7$ is small?

2. Let's say that now we have a small dataset with $n=9$ observations:

```
y=c(1,2,1,4,3,2,4,3,2)
```

Let $\mu_Y$ be the true mean of whatever distribution the sample came from. Is it conceivable that $\mu_Y=5$?

We can test for $H_0:\mu_Y=5$ against $H_1:\mu_Y\neq 5$ simply by calling:

```
t.test(y,mu=5)
```

```
    One Sample t-test
data:  y
t = -6.7823, df = 8, p-value = 0.0001403
alternative hypothesis: true mean is not equal to 5

95 percent confidence interval:
1.575551 3.313338

sample estimates:
mean of x
2.444444
```

The $p-$value is 0.0001403, which is substantially smaller than $\alpha=0.05$, and we reject the null hypothesis that the true mean is 5. The test provides no information about what the true mean could be, but the 95% confidence interval $(1.575551, 3.313338)$ does: we would expect $\mu_Y \approx 2.5$.

3. Is it conceivable that $\mu_Y = 2.5$?

   Let's run:

```
t.test(y,mu=2.5)
```

```
    One Sample t-test
data:  y
t = -0.14744, df = 8, p-value = 0.8864
alternative hypothesis: true mean is not equal to 2.5

95 percent confidence interval:
 1.575551 3.313338

sample estimates:
mean of x
 2.444444
```

   With such a large $p$-value, we can definitely accept the null hypothesis, right?[28]

28: Alas, we cannot. All that we can say is that we do not have enough evidence to reject the null hypothesis $H_0 : \mu_Y = 2.5$.

**Teaching Dataset** Suppose that a researcher wants to determine if, as she believes, a new teaching method enables students to understand elementary statistical concepts better than the traditional lectures given in a university setting (based on [9]).

She recruits $N = 80$ second-year students to test her claim. The students are randomly assigned to one of two groups:

- students in group $A$ are given the traditional lectures,
- whereas students in group $B$ are taught using the new teaching method.

After three weeks, a short quiz is administered to the students in order to assess their understanding of statistical concepts.

The results are found in the `teaching.csv` dataset.

```
teaching <- read.csv("teaching.csv", header = TRUE)
colnames(teaching)<-c("ID","Group","Grade")
head(teaching)
```

```
ID  Group  Grade
1  B  75.5
2  B  77.5
3  A  73.5
4  A  75.0
5  B  77.0
6  A  79.0
```

Is there enough evidence to suggest that the new teaching is more effective (as measured by test performance)?

We can summarize the results (sample size, sample mean, sample variance) as follows:

```
library(dplyr)
counts.by.group = aggregate(x = teaching$Grade,
    by = list(teaching$Group), FUN = length)

means.by.group = aggregate(x = teaching$Grade,
    by = list(teaching$Group), FUN = mean)

variances.by.group = aggregate(x = teaching$Grade,
    by = list(teaching$Group), FUN = var)

teaching.summary <- counts.by.group |>
    full_join(means.by.group, by="Group.1" ) |>
    full_join(variances.by.group, by="Group.1" )

colnames(teaching.summary) <- c("Group",
    "Sample Size", "Sample Mean", "Sample Variance")
```

```
Group  Sample Size  Sample Mean  Sample Variance
A  40  75.125  6.650641
B  40  79.000  5.538462
```

If the researcher assumes that both groups have similar background knowledge prior to being taught (which she attempt to enforce by randomising the group assignment), then the effectiveness of the teaching methods may be compared using two hypotheses: the **null hypothesis** $H_0$ and the **alternative** $H_1$.

Let $\mu_i$ represent the true performance of method $i$. Since the researcher wants to claim that the new method is more effective than the traditional ones, it is most appropriate for her to use one-sided hypothesis testing with

$$H_0 : \mu_A \geq \mu_B \quad \text{against} \quad H_1 : \mu_A < \mu_B.$$

The testing procedure is simple:

1. calculate an appropriate **test statistic** under $H_0$;
2. reject $H_0$ in favour of $H_1$ if the test statistic falls in the **critical region** (also called the **rejection region**) of an associated distribution, and
3. fail to reject $H_0$ otherwise.

In this case, she uses a two-sample $t-$test. Assuming that variability in two groups are roughly the same, the test statistic is given by:

$$t_0 = \frac{\overline{y}_B - \overline{y}_A}{S_p \sqrt{\frac{1}{N_A} + \frac{1}{N_B}}},$$

where the pooled variance $S_p^2$ is

$$S_p^2 = \frac{(N_A - 1)S_A^2 + (N_B - 1)S_B^2}{N_A + N_B - 2}.$$

With her data, she obtains the $t-$statistic as follows. First, she identifies the number of observations in each group:

```
(N.A = teaching.summary[1,2])
(N.B = teaching.summary[2,2])
(N=N.A+N.B)
```

```
[1] 40
[1] 40
[1] 80
```

Then, she computes the sample mean score in each group:

```
(y.bar.A = teaching.summary[1,3])
(y.bar.B = teaching.summary[2,3])
```

```
[1] 75.125
[1] 79
```

She computes the sample variance of the scores in each group:

```
(S2.A = teaching.summary[1,4])
(S2.B = teaching.summary[2,4])
```

```
[1] 6.650641
[1] 5.538462
```

She finally computes the sample pooled variance of scores:

```
(S2.P = ((N.A-1)*S2.A+(N.A-1)*S2.B)/(N.A+N.B-2))
```

```
[1] 6.094551
```

From which she obtains the $t-$statistic:

```
(t0 = (y.bar.B - y.bar.A) / sqrt(S2.P*(1/N.A+1/N.B)))
```

```
[1] 7.019656
```

The test statistic value is $t_0 = 7.02$.

In order to reject or fail to reject the null hypothesis, she needs to compare it against the critical value of the Student $T$ distribution with $N - 2 = 78$ degrees of freedom at significance level $\alpha = 0.05$, say.

Set the significance level at 0.05:

```
alpha=0.05
```

Be careful with the `qt()` function – the next call "looks" right, but it will give you a critical value on the wrong side of the distribution's mean:

```
(t.star.wrong = qt(alpha,N-2))
```

```
[1] -1.664625
```

This call, however, gives the correct critical value:

```
(t.star = qt(alpha,N-2, lower.tail=FALSE))
```

```
[1] 1.664625
```

The appropriate critical value is

$$t^* = t_{1-\alpha,N-2} = t_{0.95,78} = 1.665.$$

Since $t_0 > t^*$ at $\alpha = 0.05$, she rejects the null hypothesis $H_0 : \mu_A \geq \mu_B$, which is to say that she has enough evidence to support the claim that the new teaching method is more effective than the traditional methods, at $\alpha = 0.05$.

## 7.5 Additional Topics

We will finish this chapter by introducing and briefly discussing some additional statistical analysis topics (ANOVA, ANCOVA, MANOVA, multivariate statistics, goodness-of-fit tests). Another common application, **linear regression and its variants**, will receive a thorough treatment in subsequent modules.

### 7.5.1 Analysis of Variance

**Analysis of variance** (ANOVA) is a statistical method that partitions a dataset's variability into **explainable variability** (model-based) and **unexplained variability** (error) using various statistical models, to determine whether (multiple) treatment groups have significantly different group means.[29] The **total sample variability** of a feature $y$ in a dataset is defined as

$$\text{SST} = \sum_{k=1}^{N} (y_k - \overline{y})^2,$$

where $\overline{y}$ is the overall mean of the data.

29: We will have more to say on the topic in Chapter 11.

Let us return to the teaching method example of Section 7.4.7.

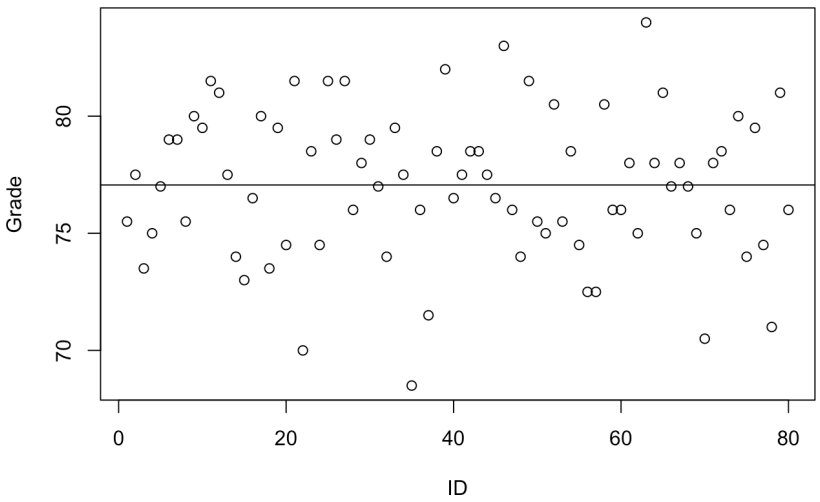
The mean of the grades, for all students, is:

```
(mu = mean(teaching$Grade))
```

```
[1] 77.0625
```

The plot below shows all the students' scores, ordered by participant ID; the overall mean is displayed for comparison.

```
plot(teaching$ID,teaching$Grade, xlab="ID", ylab="Grade")
abline(h = mu)
```




Since the assignment of ID is **arbitrary** (at least, in theory), we do not observe any patterns – if we were to guess someone's score with no knowledge except for their participant ID, then picking the sample mean is as good a guess as any other reasonable guesses.

Statistically speaking, this means that the **null model**

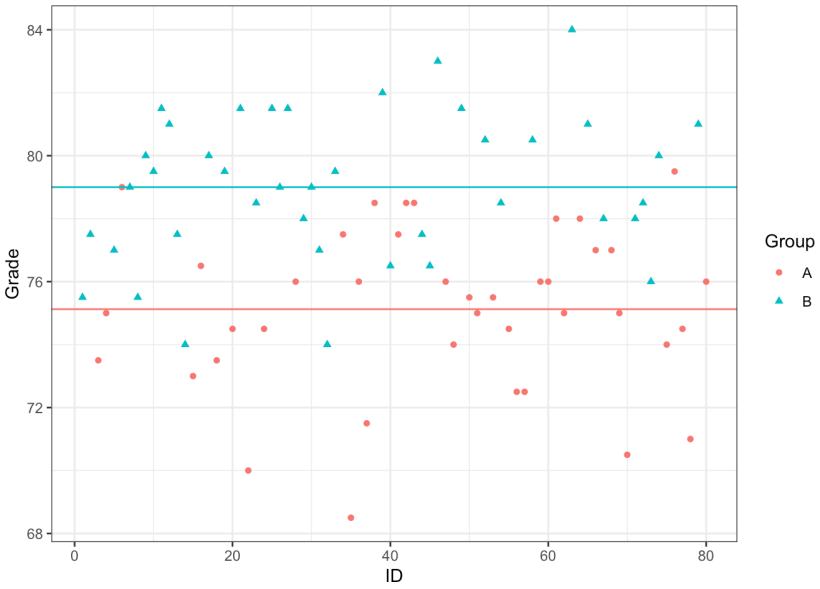
$$y_{i,j} = \mu + \varepsilon_{i,j},$$

where $\mu$ is the **overall mean**, $i = A, B$, and $j = 1, \ldots, 40$, does not explain any of the variability in the student scores (as usual, $\varepsilon_{i,j}$ represents the departure or noise from the model prediction).

But the students DID NOT all receive the same treatment: 40 randomly selected students were assigned to group $A$, and the other 40 to group $B$, and both group were taught using a different method.

When we add this information to the plot, we see that the two study groups show different characteristics in term of their average scores.

```
library(ggplot2)
ggplot(teaching, aes(x=ID,y=Grade,colour=Group,shape=Group)) +
  geom_point() +
  geom_hline(aes(yintercept = y.bar.B),col="#00BFC4") +
  geom_hline(aes(yintercept = y.bar.A),col="#F8766D") + theme_bw()
```




With the group assignment information, we can refine our null model into the **treatment-based model**

$$y_{i,j} = \mu_i + \varepsilon_{i,j},$$

where $\mu_i$, $i = A, B$ represent the group means. Using this model, we can decompose SST into **between-treatment sum of squares** and **error (within-treatment) sum of squares** as

$$\begin{aligned} \text{SST} &= \sum_{i,j}(y_{i,j} - \overline{y})^2 = \sum_{i,j}(y_{i,j} - \overline{y}_i + \overline{y}_i - \overline{y})^2 \\ &= \sum_i N_i(\overline{y}_i - \overline{y})^2 + \sum_{i,j}(y_{i,j} - \overline{y}_i)^2 = \text{SSA} + \text{SSE} \end{aligned}$$

The SSA component looks at the difference between each of the treatment means and the overall mean, which we consider to be **explainable**[30] ; the SSE component, on the other hand, looks at the difference between each observation and its own group mean, and is considered to be **random**.[31]

Thus, SSA/SST $\times$ 100% of the total variability can be explained using a treatment-based model. This ratio is called the **coefficient of determination**, denoted by $R^2$.

Formally, the ANOVA table incorporates a few more items – the table below summarizes all the information that it contains.

| Source | Sum of Squares | df | Mean Square | $F_0$ | p−value |
|---|---|---|---|---|---|
| Treatment | SSA | $p-1$ | MSA = SSA/$(p-1)$ | MSA/MSE | $P(F_0 > F^*)$ |
| Error | SSE | $N-p$ | MSE = SSE/$(N-p)$ | | |
| Total | SST | $N-1$ | | | |

The specific table for the teaching methodology dataset can be obtained directly from the `lm()` function.

30: That is to say, the treatment explains part of the difference in the observed group means.

31: As the spread about the group means is fairly large (relatively-speaking), we suspect that the treatment-based model on its own does not capture all the variability in the data.

```
model.lm <- lm(Grade ~ Group, data = teaching)
SS.Table <- anova(model.lm)
SS.Table
```

| Source | Sum of Squares | df | Mean Square | $F_0$ | p−value |
|---|---|---|---|---|---|
| Treatment | 300.31 | 1 | 300.31 | 49.28 | $7.2 \times 10^{-10}$ *** |
| Error | 475.38 | 78 | 6.095 | | |
| Total | 775.69 | 79 | | | |

The test statistic $F_0$ follows an $F$-distribution with $(\mathrm{df}_{\mathrm{treat}}, \mathrm{df}_{\mathrm{e}}) = (1, 78)$ degrees of freedom. At a significance level of $\alpha = 0.05$, the critical value $F^* = F_{0.95,1,78} = 3.96$ is substantially smaller than the test statistic $F_0 = 49.28$, implying that the two-treatment model is statistically significant.

This, in turn, means that the model recognises a statistically significant difference between the students' scores, based on the teaching methods.

```
(R2 = summary(model.lm)$r.squared)
```

```
[1] 0.3871566
```

The coefficient of determination $R^2$ provides a way to measure the model's **significance**. From the ANOVA table for the teaching example, we compute

$$R^2 = \frac{\mathrm{SSA}}{\mathrm{SST}} = \frac{300.31}{775.69} \approx 0.39,$$

which means that 39% of the total variation in the data can be explained by the two-treatment model.

Is this good enough? That depends on the specifics of the situation (in particular, on the researcher's or the client's needs).

### Diagnostic Checks

As with most statistical procedures, ANOVA relies on certain assumptions for its result to be valid. Recall that the model is given by

$$y_{i,j} = \mu_i + \varepsilon_{i,j}.$$

What assumptions are made?

The main assumption is that the error terms follow independently and identically distributed (iid) normal distributions (i.e., $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$).

Assuming independence, we are required to verify three additional assumptions:

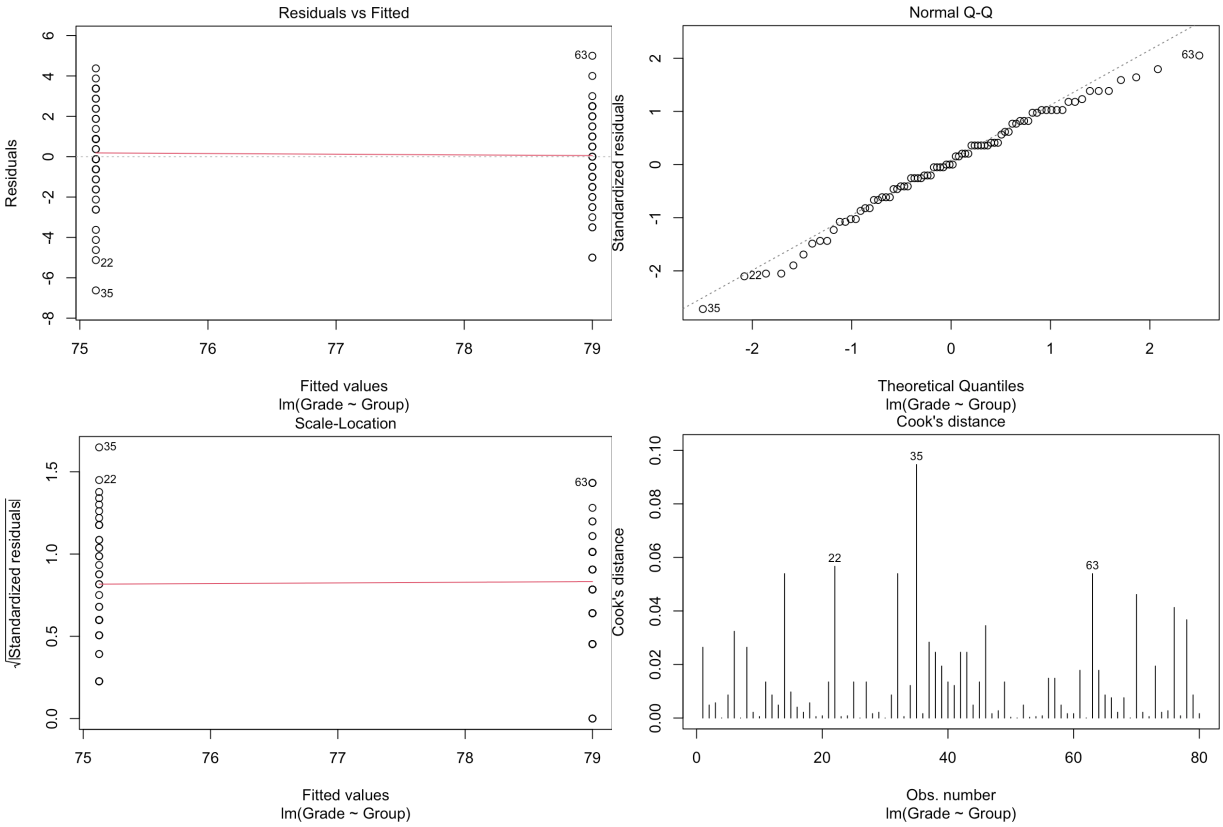
- normality of the error terms;
- constant variance (within treatment groups), and
- equal variances (across treatment groups).

Normality of the errors can be tested visually with the help of a **normal-QQ plot**, which compares the **standardized residuals quantiles** against the **theoretical quantiles** of the standard normal distribution $\mathcal{N}(0,1)$.[32]

32: A straight line indicates normality.

In other words, if the errors are normally distributed with mean 0 and variance $\sigma^2$, we would expect that the 80 standardized residuals $r_{i,j} = \frac{\varepsilon_{i,j}-0}{\sigma}$ should behave as though they had been drawn from $\mathcal{N}(0,1)$.

```
plot(model.lm, which = c(1,2,3,4))
```




The plots above show some departure in the lower tail, however, moderate departure from normality is usually acceptable as long as it is mostly a tail phenomenon.

To test the assumption of constant variance, we can run visual inspection using:

- residuals vs.fitted values, and/or
- residuals vs.order/time.

The standardized residuals in both groups should be approximately distributed according to $\mathcal{N}(0,1)$. The plots also show that variability from the mean in each treatment group is reasonably similar.[33]

More formally, equality of variance is often tested for using **Bartlett's test** (when normality of the residuals is met) or the **modified Levene's test** (when it is not).

33: If a difference is apparent and we cannot conclude that the variances are constant across groups, we need to apply a **variance stabilising transformation**, such as a **logarithmic transformation** or **square-root transformation** before proceeding.

Assuming that we felt the evidence of normal residuals was warranted in the two-treatment model of the teaching dataset, we get a $p-$value of 0.57 for Bartlett's test:

```
(B.T <- bartlett.test(Grade~Group, teaching))
```

```
    Bartlett test of homogeneity of variances

data:  Grade by Group
Bartlett's K-squared = 0.32192, df = 1, p-value = 0.5705
```

Otherwise, we get a $p-$value of 0.76 for Levene's test.

```
(L.T <- lawstat::levene.test(teaching$Grade,
    teaching$Group, location="median",
    correction.method="zero.correction"))
```

```
    Modified robust Brown-Forsythe Levene-type test based
    on the absolute deviations from the median with modified
    structural zero removal method and correction factor

data:  teaching$Grade
Test Statistic = 0.095106, p-value = 0.7586
```

In either case, the $p-$value falls above reasonable significance levels (0.05, say), which means that we cannot reject the null hypothesis of equal variance.

When there are $p > 2$ treatment groups, ANOVA provides a test for

$$H_0 : \mu_1 = \cdots = \mu_p \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j \text{ for at least one } i \neq j.$$

A significant $F_0$ value indicates that **there is at least one group which differs from the others**, but it does not specify which one does.

Specialized methods such as **Scheffe's method** and **Tukey's test** can be used to identify the statistically different treatments.

Finally, while ANOVA can accommodate unequal treatment group sizes, it is recommended to keep those sizes equal across all groups – this makes the test statistic less sensitive to violations of the assumption of equal variances across treatment groups, providing yet another reason to involve the analysts/consultants in the **data collection process**.

### 7.5.2 Analysis of Covariance

In a previous section, we looked at the effectiveness of new teaching method by assigning each group to a specific treatment and comparing the mean test scores. A crucial assumption for that model is that subjects in each group have **similar background knowledge** about statistics prior to the three week lectures.

If this assumption is wrong, however, we may be making incorrect decisions based on the model. Even if each group had similar background knowledge *on average*, there may be large variability from person-to-person, masking the true treatment effect.

### Paired Comparison

One way to avoid such **subject-to-subject variability** is to administer both treatments to each individual, and then compare treatment effects by looking at the **difference in the outcomes**. For instance, if a grocery chain is interested in measuring the effectiveness of two advertising campaigns, it could be reasonable to assume that there is a large variability in total sales, as well as popular items sold, at each store.

It may then be preferable to run both campaigns in each store and analyze the resulting data rather than to split the stores into two groups (in each of which a different advertising campaign is run) and then to compare the mean outcomes in the two groups.

Formally, let $X_{i,1}$ denote the total sales with campaign $A$ and $X_{i,2}$ the total sales with campaign $B$. The quantity of interest is the **difference** $D_i = X_{i,1} - X_{i,2}$ for each store $i = 1, \ldots, N$.

Assuming that the differences $D_i$ follow an iid normal distribution with mean $\delta$ and variance $\sigma_d^2$, then we test for

$$H_0 : \delta = 0 \quad \text{against} \quad H_1 : \delta \neq 0$$

using the test statistic

$$t_0 = \sqrt{N}\frac{\overline{D}}{s_d},$$

which follows a Student's $t$ distribution with $N - 1$ degrees of freedom; thus we reject $H_0$ if the observed test statistic $t_0$ has $p$-value less than the significance level $\alpha/2$.

### ANOVA vs. ANCOVA

ANOVA compares multiple group means and tests whether any of the group means differ from the rest, by breaking down the total variability into a treatment (explainable) variability component and an error (unexplained) variability component, and building a ratio $F_0$ to determine whether or not to reject $H_0$.

**Analysis of covariance** (ANCOVA) introduces **concomitant variables** (or **covariates**) to the ANOVA model, splitting the total variability into 3 components: SSA, $\text{SS}_{\text{con}}$, and SSE, aiming to reduce error variability.

The choice of covariates is thus crucial in running a successful ANCOVA. In order to be useful, a concomitant variable must be related to response variable in some way, otherwise it not only fails to reduce error variability, but it also increases the model complexity:

- in the teaching method example, we could consider administering a pre-study test to measure the **prior knowledge level** of each participant and use this score as a concomitant variable;

- in the advertising campaign example, we could have used the **previous month's sales** as a covariate;
- in medical studies, we could use the **age** and **weight** of subjects, say.

Importantly, concomitant variables should not be affected by treatments. As an example, suppose that the patients in a medical study were asked:

> How strongly do you believe that you were given actual medication rather than a placebo?

If the treatment is indeed effective, then a participant's response to this question could be **markedly different** in the treatment group than in the placebo group.[34]

34: The medication may have strong side-effects which cannot be ignored.

This means that true treatment effect may be masked by concomitant variable due to unequal effects on treatment groups. Note that **qualitative covariates** (such as gender, say) are not part of the ANCOVA framework – indeed, such covariates create new ANOVA treatment groups instead.

When moving from an ANOVA to an ANCOVA model, the error variability is further split into a **pure error** and a **covariate** component, while the **treatment** variability remains unchanged.

### ANCOVA Model and Assumptions

Suppose that we are testing the effect of $p$ treatments, with $N_j$ subjects in each group. Then the ANCOVA model takes the form

$$y_{i,j} = \mu + \tau_j + \gamma(x_{i,j} - \overline{x}) + \varepsilon_{i,j}$$

where

- $y_{i,j}$ is the response of the $i^{\text{th}}$ subject in the $j^{\text{th}}$ treatment group;
- $\mu$ is the overall mean;
- $\tau_j$ is the $j^{\text{th}}$ treatment effect, subject to a constraint

$$\sum_{j=1}^{p} \tau_j = 0;$$

- $\gamma$ is the coefficient for the **covariate effect**;
- $(x_{i,j} - \overline{x})$ is the covariate value of the $i^{\text{th}}$ subject in the $j^{\text{th}}$ treatment group, adjusted by the mean, and
- $\varepsilon_{i,j}$ is the error of $i^{\text{th}}$ subject in the $j^{\text{th}}$ treatment group.

Additionally, four assumptions must be satisfied:

- **independence and normality of residuals** – the residuals follow an $iid$ normal distribution with mean of 0 and variance $\sigma_\varepsilon^2$;
- **homogeneity of residual variances** – the variance of the residuals is uniform across treatment groups;
- **homogeneity of regression slopes** – the regression effect (slope) is uniform across treatment groups, and
- **linearity of regression** – the regression relationship between the response and the covariate is linear.

The first of these assumptions can be tested with the help of a QQ-plot and a scatter-plot of residuals vs.fitted values, while the second may use the Bartlett or the Levene test. The final assumption is not as crucial as the other three assumptions, however. Various remedial methods can be applied should any of these assumptions fail.

The third assumption, however, is **crucial** to the ANCOVA model; it can be tested with the **equal slope test**, which requires an ANCOVA regression with an additional interaction term $x \times \tau$. If the interaction is not significant, the third assumption is satisfied.

In the event that the interaction term is statistically significant, a different approach (e.g. moderated regression analysis, mediation analysis) is required since using the original ANCOVA model is not prescribed.

An in-depth application of an ANCOVA model can be found in [2].

### 7.5.3 Basics of Multivariate Statistics

Up to this point, we have only considered situations where the response is **univariate**. In applications, the situation often calls for **multivariate** responses, where the response variables are thought to have some relationship to one another (e.g., a **correlation structure**).

It remains possible to analyze each response variable independently, but the dependence structure can be exploited to make **joint** (or simultaneous) inferences.

#### Properties of the Multivariate Normal Distribution

The probability density function of a multi-dimensional random vector $\mathbf{X} \in \mathbb{R}^p$ that follows a **multivariate normal distribution** with **mean vector** $\boldsymbol{\mu}$ and **covariance matrix** $\boldsymbol{\Sigma}$, denoted by $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is given by

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right),$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p} \end{bmatrix}.$$

For such an $\mathbf{X}$, the following properties hold:

1. any linear combination of its components are normally distributed;
2. all subsets of components follow a (modified) multivariate normal distribution;
3. a diagonal covariance matrix implies the independence of its components;
4. conditional distributions of components follow a normal distribution, and
5. the quantity $(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ follows a $\chi^2_p$.

These properties make the multivariate normal distribution attractive, from a theoretical point of view (if not always entirely realistic).

For instance:

- using property 1, we can use **contrasts** to test which components are distinct from the others;
- property 5 is the multivariate analogue of the square of a standard normal random variable $Z \sim \mathcal{N}(0,1)$ following a $Z^2 \sim \chi^2_1$ distribution;
- but two univariate normal random variables with zero covariance are not necessarily independent (the joint p.d.f. of two such variables is not necessarily the p.d.f. of a multivariate normal distribution).

### Hypothesis Testing for Mean Vectors

When the sample comes from a univariate normal distribution, we can test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0$$

by using a $t-$statistic. Analogously, if the sample comes from a $p-$variate normal distribution, we can test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{against} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

by using **Hotelling's $\mathbf{T}^2$ test statistic**

$$T^2 = N \cdot (\overline{\mathbf{X}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}),$$

where $\overline{\mathbf{X}}$ denotes the **sample mean**, $\mathbf{S}$ the **sample covariance matrix**, and $N$ the sample size.

Under $H_0$,

$$T^2 \sim \frac{(N-1)p}{(N-p)} F_{p,N-p}.$$

Thus, we do not reject $H_0$ at a significance level of $\alpha$ if

$$N \cdot (\overline{\mathbf{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}_0) \leq \frac{(N-1)p}{(N-p)} F_{p,N-p}(\alpha)$$

and reject it otherwise.

### Confidence Region and Simultaneous Confidence Intervals for Mean Vectors

In the $p-$variate normal distribution, any $\boldsymbol{\mu}$ that satisfies the condition

$$N \cdot (\overline{\mathbf{X}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{(N-1)p}{(N-p)} F_{p,N-p}(\alpha)$$

resides inside a $(1-\alpha)100\%$ **confidence region** (an ellipsoid in this case).

**Simultaneous Bonferroni confidence intervals** with overall error rate $\alpha$ can also be derived, using

$$(\overline{x}_j - \mu_j) \pm t_{N-1}(\alpha/p)\sqrt{\frac{s_{j,j}}{N}} \text{ for } j = 1, \ldots, p.$$

Another approach is to use **Hotelling's $T^2$ simultaneous confidence intervals**, given by

$$(\overline{x}_j - \mu_j) \pm \sqrt{\frac{p(N-1)}{N-p}F_{p,N-p}(\alpha)}\sqrt{\frac{s_{j,j}}{N}} \text{ for } j = 1, \ldots, p.$$

Figure 7.20 shows these regions for a bivariate normal random sample. Note that the Hotelling's $T^2$ simultaneous confidence intervals form a rectangle (in grey) that confines the confidence region, while the Bonferroni confidence intervals (in blue) are slightly narrower.

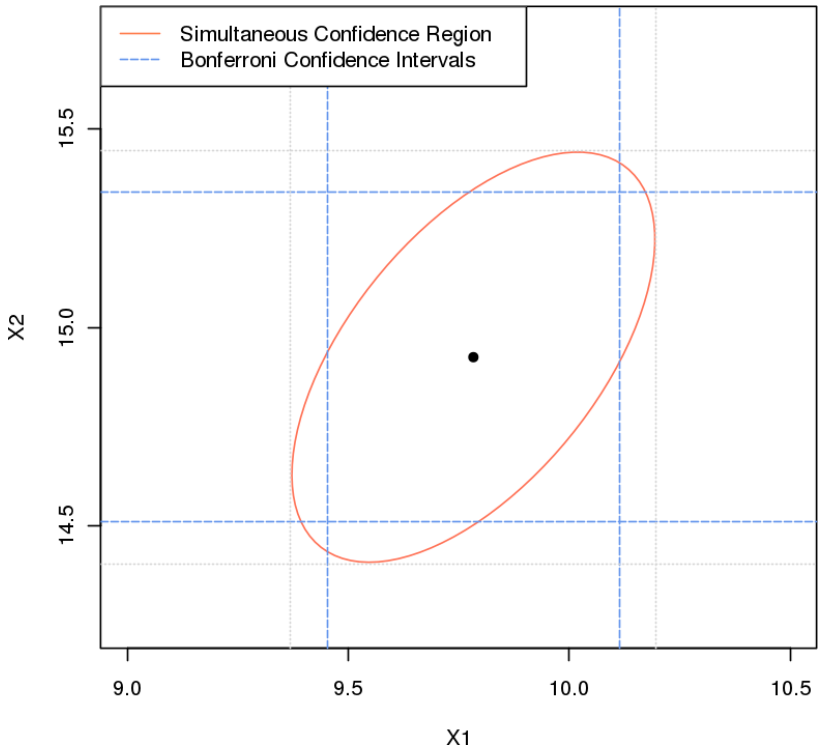



**Figure 7.20:** Confidence region for a bivariate normal random sample (sample not shown).

Given that all the components of the mean vector are correlated (since the covariance matrix is generally non-diagonal), the confidence region should be used if the goal is to study the **plausibility of the mean vector as a whole**, while Bonferroni confidence intervals may be more suitable when **component-wise confidence intervals** are of needed.

### Multivariate Analysis of Variance

ANOVA is often used as a first attempt to determine whether the means from every sub-population are identical.

ANOVA can test means from more than two populations; the **multivariate ANOVA** (MANOVA) is quite simply a multivariate extension of ANOVA which tests whether the mean vectors from all sub-populations are identical.

Assume there are $I$ sub-populations in the population, from each of which $N_i$ $p-$dimensional responses are drawn, for $i = 1, \ldots, I$.

Each observation can be expressed as:

$$\mathbf{X}_{i,j} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_{ij},$$

where $\boldsymbol{\mu}$ is the **overall mean vector**, $\boldsymbol{\tau}_i$ is the $i^{\text{th}}$ **population-specific treatment effect**, and $\boldsymbol{\varepsilon}_{ij}$ is the **random error**, which follows a $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ distribution.

It is important to note that the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be the same for each sub-population, and that

$$\sum_{i=1}^{I} N_i \boldsymbol{\tau}_i = \mathbf{0}$$

to ensure that the estimates are uniquely identifiable.

To test the hypothesis

$$H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_I = \mathbf{0} \quad \text{against} \quad H_1 : \text{some } \boldsymbol{\tau}_i \neq \mathbf{0},$$

we decompose the **total sum of squares and cross-products** $\text{SSP}_{\text{tot}}$ into

$$\text{SSP}_{\text{tot}} = \text{SSP}_{\text{treat}} + \text{SS}_{\text{e}}.$$

Based on this decomposition, we compute the test statistic known as **Wilks' lambda**

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|},$$

where $\mathbf{B}, \mathbf{W}$ are as in the MANOVA table below:

| Source | SSP | df | MSP | $\mathbf{F_0}$ |
|---|---|---|---|---|
| Treatment | $\mathbf{B}$ | $I - 1$ | $\mathbf{B}/(I-1)$ | $\mathbf{W}^{-1}\mathbf{B}$ |
| Error | $\mathbf{W}$ | $\sum_{i=1}^{I} N_i - I$ | $\mathbf{W}/\sum_{i=1}^{I}(N_i - 1)$ | |
| Total | $\mathbf{B} + \mathbf{W}$ | $\sum_{i=1}^{I} N_i - 1$ | $(\mathbf{B} + \mathbf{W})/(\sum_{i=1}^{I} N_i - 1)$ | |

We have

$$\mathbf{B} = \sum_{i=1}^{I} N_i (\mathbf{X}_i - \mathbf{X})(\mathbf{X}_i - \mathbf{X})^\top$$

and

$$\mathbf{W} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \mathbf{X}_i)(\mathbf{X}_{ij} - \mathbf{X}_i)^\top;$$

we reject $H_0$ if $\Lambda^*$ is below some pre-agreed upon threshold, which depends on $p$, $I$, and $N_i$, $i = 1, \ldots, I$.

### 7.5.4 Goodness-of-Fit Test

A (fictitious) 2017 survey asked a sample of $N = 200$ adults between the age of 25 to 35 about their highest educational achievement:

| Year | <HS | HS | CU | CU+ |
|---|---|---|---|---|
| 2017 | 16 | 55 | 83 | 46 |

In a 1997 survey, it was also found that:

| Year | <HS | HS | CU | CU+ |
|---|---|---|---|---|
| 1997 | 13% | 32% | 37% | 18% |

Based on the result of this survey, is there sufficient evidence to believe that educational backgrounds of the population have changed between 1997 and 2007?[35]

We can view the distribution of educational achievements as being **multinomial**. For such a distribution, with parameters $p_1, \cdots, p_k$, the expected frequency in each category is $m_j = Np_j$.

Let $O_j$ denote the observed frequency for the $j^{\text{th}}$ category. If there has been no real change since 1997, we would expect the sum of squared differences between the observed 2017 frequencies and the expected frequencies based on 1997 data to be small.

We can use this information to test the **goodness-of-fit** between the observations and the expected frequencies *via* Pearson's $\chi^2$ test statistic

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - m_{j,0})^2}{m_{j,0}} \sim \chi^2(k-1).$$

In the above example, the hypotheses of interest are

$$H_0 : \mathbf{p} = \mathbf{p}^* = (0.13, 0.32, 0.37, 0.18) \quad \text{vs} \quad H_1 : \mathbf{p} \neq \mathbf{p}^*.$$

The table below summarizes the information under $H_0$.

| Category | $O_j$ | $p_{j,0}$ | $m_{j,0}$ | $(O_j - m_{j,0})^2/m_{j,0}$ |
|---|---|---|---|---|
| 1 | 16 | 0.13 | 26 | 3.846 |
| 2 | 55 | 0.32 | 64 | 1.266 |
| 3 | 83 | 0.37 | 74 | 1.095 |
| 4 | 46 | 0.18 | 36 | 2.778 |
| Total | 200 | 1 | 200 | 7.815 |

Pearson's test statistic is $X^2 = 7.815$, with an associated $p-$value of 0.0295, which implies that there is enough statistical evidence (at the $\alpha = 0.05$ level) to accept that the population's educational achievements have changed over the last 20 years.

35: Since each respondent's educational achievement can only be classified into one of these categories, they are **mutually exclusive**. Furthermore, these categories cover all possibilities on the educational front, so they are also **exhaustive**.

## 7.6 Exercises

1. Consider a sample of $n = 10$ observations displayed in ascending order:

$$15, 16, 18, 18, 20, 20, 21, 22, 23, 75.$$

   a) Compute the sample mean and sample variance.
   b) Find the 5-point summary of the data. Is the distribution skewed?
   c) Are there any likely outliers in the sample? If so, indicate their values.
   d) Build and display the sample's boxplot chart.
   e) Build and display a sample histogram.

2. The daily number of accidents in Sydney over a 40-day period are provided below:

$$6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15, 2, 17, 10, 3, 9, 4, \quad 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17, 7, 7, 21, 13, 23, 1, 11, 9, 9, 25.$$

   a) Compute the sample mean and sample variance.
   b) Find the 5-point summary of the data. Is the distribution skewed?
   c) Are there any likely outliers in the sample? If so, indicate their values.
   d) Build and display the sample's boxplot chart.
   e) Build and display a sample histogram.

3. Repeat the previous question when the "31" is replaced by a "130".
4. The grades in a class are shown below.

$$\begin{gathered}
80, 73, 83, 60, 49, 96, 87, 87, 60, 53, 66, 83, 32, 80, 66 \\
90, 72, 55, 76, 46, 48, 69, 45, 48, 77, 52, 59, 97, 76, 89 \\
73, 73, 48, 59, 55, 76, 87, 55, 80, 90, 83, 66, 80, 97, 80 \\
55, 94, 73, 49, 32, 76, 57, 42, 94, 80, 90, 90, 62, 85, 87 \\
97, 50, 73, 77, 66, 35, 66, 76, 90, 73, 80, 70, 73, 94, 59 \\
52, 81, 90, 55, 73, 76, 90, 46, 66, 76, 69, 76, 80, 42, 66 \\
83, 80, 46, 55, 80, 76, 94, 69, 57, 55, 66, 46, 87, 83, 49 \\
82, 93, 47, 59, 68, 65, 66, 69, 76, 38, 99, 61, 46, 73, 90, \\
66, 100, 83, 48, 97, 69, 62, 80, 66, 55, 28, 83, 59, 48, 61 \\
87, 72, 46, 94, 48, 59, 69, 97, 83, 80, 66, 76, 25, 55, 69 \\
76, 38, 21, 87, 52, 90, 62, 73, 73, 89, 25, 94, 27, 66, 66 \\
76, 90, 83, 52, 52, 83, 66, 48, 62, 80, 35, 59, 72, 97, 69 \\
62, 90, 48, 83, 55, 58, 66, 100, 82, 78, 62, 73, 55, 84, 83 \\
66, 49, 76, 73, 54, 55, 87, 50, 73, 54, 52, 62, 36, 87, 80, 80
\end{gathered}$$

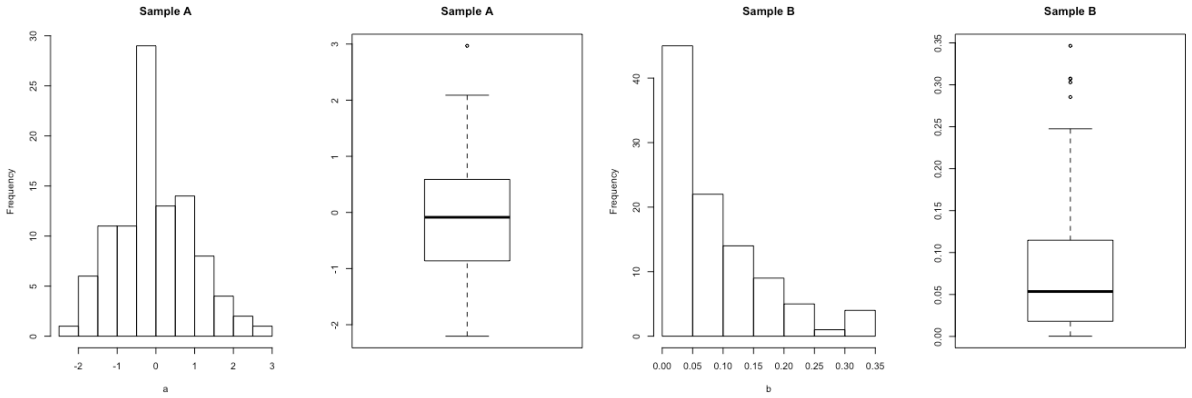
   a) Compute the sample mean and sample variance.
   b) Find the 5-point summary of the data. Is the distribution skewed?
   c) Are there any likely outliers in the sample? If so, indicate their values.
   d) Build and display the sample's boxplot chart.
   e) Build and display a sample histogram.
   f) Based on your analysis, how well did the class do?

5. Consider the following dataset:

$$2.6, 3.7, 0.8, 9.6, 5.8, -0.8, 0.7, 0.6, 4.8, 1.2, 3.3, 5.0, 3.7, 0.1, -3.1, 0.3.$$

   What are the median and the interquartile range of the sample?

f) The following charts show a histogram and a boxplot for two samples, $A$ and $B$. Based on these charts, which of $A$ and/or $B$ (or neither) is likely to arise from a normal population?




f) Consider the following dataset:

$$12, 14, 6, 10, 1, 20, 4, 8.$$

What are its median and its first quartile?

f) A manufacturer of fluoride toothpaste regularly measures the concentration of of fluoride in the toothpaste to make sure that it is within the specifications of 0.85 – 1.10 mg/g. [5]

**Table 6.1-3** Concentrations of fluoride in mg/g in toothpaste

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.98 | 0.92 | 0.89 | 0.90 | 0.94 | 0.99 | 0.86 | 0.85 | 1.06 | 1.01 |
| 1.03 | 0.85 | 0.95 | 0.90 | 1.03 | 0.87 | 1.02 | 0.88 | 0.92 | 0.88 |
| 0.88 | 0.90 | 0.98 | 0.96 | 0.98 | 0.93 | 0.98 | 0.92 | 1.00 | 0.95 |
| 0.88 | 0.90 | 1.01 | 0.98 | 0.85 | 0.91 | 0.95 | 1.01 | 0.88 | 0.89 |
| 0.99 | 0.95 | 0.90 | 0.88 | 0.92 | 0.89 | 0.90 | 0.95 | 0.93 | 0.96 |
| 0.93 | 0.91 | 0.92 | 0.86 | 0.87 | 0.91 | 0.89 | 0.93 | 0.93 | 0.95 |
| 0.92 | 0.88 | 0.87 | 0.98 | 0.98 | 0.91 | 0.93 | 1.00 | 0.90 | 0.93 |
| 0.89 | 0.97 | 0.98 | 0.91 | 0.88 | 0.89 | 1.00 | 0.93 | 0.92 | 0.97 |
| 0.97 | 0.91 | 0.85 | 0.92 | 0.87 | 0.86 | 0.91 | 0.92 | 0.95 | 0.97 |
| 0.88 | 1.05 | 0.91 | 0.89 | 0.92 | 0.94 | 0.90 | 1.00 | 0.90 | 0.93 |

a) Build a relative frequency histogram of the data (a histogram with area = 1).
b) Compute the data's mean $\overline{x}$ and its standard deviation $s_x$.
c) The mean and the variance can also be approximated as follows. Let $u_i$ be the **class mark** for each of the histogram's classes (the midpoint along the rectangles' widths), $n$ be the total number of observations, and $k$ be the number of classes. Then

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{k} f_i u_i \quad \text{and} \quad s_u^2 = \frac{1}{n-1} f_i (u_i - \overline{u})^2.$$

Compute $\overline{u}$ and $s_u$. How do they compare with $\overline{x}$ and $s_x$?
d) Provide a the 5–point summary of the data, as well as the interquartile range IQR.
e) Display this information as a boxplot chart.
f) Compute the **midrange** $\frac{1}{2}(Q_0 + Q_4)$, the **trimean** $\frac{1}{4}(Q_1 + 2Q_2 + Q_3)$, and the **range** $Q_4 - Q_0$ for the fluoride data.

f) The compressive strength of concrete is normally distributed with mean $\mu = 2500$ and standard deviation $\sigma = 50$. A random sample of size 5 is taken. What is the standard error of the sample mean?

f) A new cure has been developed for a certain type of cement that should change its mean compressive strength. It is known that the standard deviation of the compressive strength is 130 kg/cm$^2$ and that we may assume that it follows a normal distribution. 9 chunks of cement have been tested and the observed sample mean is $\overline{X} = 4970$. Find the 95% confidence interval for the mean of the compressive strength.

f) Consider the same set-up as in the previous question, but now 100 chunks of cement have been tested and the observed sample mean is $\overline{X} = 4970$. Find the 95% confidence interval for the mean of the compressive strength.

f) Consider the same set-up as in two questions ago, but now we do not know the standard deviation of the normal distribution. 9 chunks of cement have been tested, and the measurements are

$$5001, 4945, 5008, 5018, 4991, 4990, 4968, 5020, 5003.$$

Find the 95% confidence interval for the mean of the compressive strength.

f) A steel bar is measured with a device which a known precision of $\sigma = 0.5$mm. Suppose we want to estimate the mean measurement with an error of at most 0.2mm at a level of significance $\alpha = 0.05$. What sample size is required? Assume normality.

f) In a random sample of 1000 houses in the city, it is found that 228 are heated by oil. Find a 99% C.I. for the proportion of homes in the city that are heated by oil.

f) Past experience indicates that the breaking strength of yarn used in manufacturing drapery material is normally distributed and that $\sigma = 2$ psi. A random sample of 15 specimens is tested and the average breaking strength is found to be $\overline{x} = 97.5$ psi.

   a) Find a 95% confidence interval on the true mean breaking strength.
   b) Find a 99% confidence interval on the true mean breaking strength.

b) The diameter holes for a cable harness follow a normal distribution with $\sigma = 0.01$ inch. For a sample of size 10, the average diameter is 1.5045 inches.

   a) Find a 99% confidence interval on the mean hole diameter.
   b) Repeat this for $n = 100$.

b) A journal article describes the effect of delamination on the natural frequency of beams made from composite laminates. The observations are as follows:

$$230.66, 233.05, 232.58, 229.48, 232.58, 235.22.$$

Assuming that the population is normal, find a 95% confidence interval on the mean natural frequency.

b) A textile fibre manufacturer is investigating a new drapery yarn, which the company claims has a mean thread elongation of $\mu = 12$ kilograms with standard deviation of $\sigma = 0.5$ kilograms.

   a) What should be the sample size so that with probability 0.95 we will estimate the mean thread elongation with error at most 0.15 kg?
   b) What should be the sample size so that with probability 0.95 we will estimate the mean thread elongation with error at most 0.05 kg?

b) An article in *Computers and Electrical Engineering* considered the speed-up of cellular neural networks (CNN) for a parallel general-purpose computing architecture. Various speed-ups are observed:

$$3.77, 3.35, 4.21, 4.03, 4.03, 4.63, 4.63, 4.13, 4.39, 4.84, 4.26, 4.60.$$

Assume that the population is normally distributed. Find a 99% C.I. for the mean speed-up.

b) An engineer measures the weight of $n = 25$ pieces of steel, which follows a normal distribution with variance 16. The average observed weight for the sample is $\overline{x} = 6$. What is the two-sided 95% C.I. for the mean $\mu$?

b) The brightness of television picture tube can be evaluated by measuring the amount of current required to achieve a particular brightness level. An engineer thinks that one has to use 300 microamps of current to achieve the required brightness level. A sample of size $n = 20$ has been taken to verify the engineer's hypotheses.

   a) Formulate the null and the alternative hypotheses (use a two-sided test alternative).
   b) For the sample of size $n = 20$ we obtain $\overline{x} = 319.2$ and $s = 18.6$. Test the hypotheses from part a) with $\alpha = 5\%$ by computing a critical region. Calculate the $p$-value.
   c) Use the data from part b) to construct a 95% confidence interval for the mean required current.

c) We say that a particular production process is **stable** if it produces at most 2% defective items. Let $p$ be the true proportion of defective items.

   a) We sample $n = 200$ items at random and consider hypotheses testing about $p$. Formulate null and alternative hypotheses.
   b) What is your conclusion of the above test, if one observes 3 defective items out of 200? Note: you have to choose an appropriate confidence level $\alpha$.

b) Ten engineers' knowledge of basic statistical concepts was measured on a scale of $0 - 100$, before and after a short course in statistical quality control. The results are:

| Engineer | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Before $X_{1i}$ | 43 | 82 | 77 | 39 | 51 | 66 | 55 | 61 | 79 | 43 |
| After $X_{2i}$ | 51 | 84 | 74 | 48 | 53 | 61 | 59 | 75 | 82 | 53 |

Let $\mu_1$ and $\mu_2$ be the mean mean score before and after the course. Perform the test $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 < \mu_2$. Use $\alpha = 0.05$.

b) It is claimed that 15% of a certain population is left-handed, but a researcher doubts this claim. They decide to randomly sample 200 people and use the anticipated small number to provide evidence against the claim of 15%. Suppose 22 of the 200 are left-handed. Compute the $p-$value associated with the hypothesis (assuming a binomial distribution), and provide an interpretation.

b) A child psychologist believes that nursery school attendance improves children's social perceptiveness (SP). They use 8 pairs of twins, randomly choosing one to attend nursery school and the other to stay at home, and then obtains scores for all 16. In 6 of the 8 pairs, the twin attending nursery school scored better on the SP test. Compute the $p-$value associated with the hypothesis (assuming a binomial distribution), and provide an interpretation.

b) A certain power supply is stated to provide a constant voltage output of 10kV. Ten measurements are taken and yield the sample mean of 11kV. Formulate a test for this situation. Should it be $1-$sided or $2-$sided? What value of $\alpha$ should you use? What conclusion does the test and the sample yield?

b) A company is currently using titanium alloy rods it purchases from supplier $A$. A new supplier (supplier $B$) approaches the company and offers the same quality (at least according to supplier B's claim) rods at a lower price. The company's decision makers are interested in the offer. At the same time, they want to make sure that the safety of their product is not compromised. They randomly selects ten rods from each of the lots shipped by suppliers $A$ and $B$ and measures the yield strengths of the selected rods. The observed sample mean and sample standard deviation are 651 MPa and 2 MPa for supplier's $A$ rods, respectively, and the same parameters are 657 MPa and 3 MPa for supplier B's rods. Perform the test $H_0 : \mu_A = \mu_B$ against $\mu_A \neq \mu_B$. Use $\alpha = 0.05$. Assume that the variances are equal but unknown.

b) The deflection temperature under load for two different types of plastic pipe is being investigated. Two random samples of 15 pipe specimens are tested, and the deflection temperatures observed are as follows:

   - 206, 188, 205, 187, 194, 193, 207, 185, 189, 213, 192, 210, 194, 178, 205.
   - 177, 197, 206, 201, 180, 176, 185, 200, 197, 192, 198, 188, 189, 203, 192.

Does the data support the claim that the deflection temperature under load for type 1 pipes exceeds that of type 2? Calculate the $p$-value, using $\alpha = 0.05$, and state your conclusion.

b) It is claimed that the breaking strength of yarn used in manufacturing drapery material is normally distributed with mean 97 and $\sigma = 2$ psi. A random sample of nine specimens is tested and the average breaking strength is found to be $\overline{X} = 98$ psi. Formulate a test for this situation. Should it be 1−sided or 2−sided? What value of $\alpha$ should you use? What conclusion does the test and the sample yield?

b) A civil engineer is analyzing the compressive strength of concrete. It is claimed that its mean is 80 and variance is known to be 2. A random sample of size 60 yields the sample mean 59. Formulate a test for this situation. Should it be 1−sided or 2−sided? What value of $\alpha$ should you use? What conclusion does the test and the sample yield?

b) The sugar content of the syrup in canned peaches is claimed to be normally distributed with mean 10 and variance 2. A random sample of $n = 10$ cans yields a sample mean 11. Another random sample of $n = 10$ cans yields a sample mean 9. Formulate a test for this situation. Should it be 1−sided or 2−sided? What value of $\alpha$ should you use? What conclusion does the test and the sample yield?

b) The mean water temperature downstream from a power water plant cooling tower discharge pipe should be no more than 100F. Past experience has indicated that that the standard deviation is 2F. The water temperature is measured on nine randomly chosen days, and the average temperature is found to be 98F. Formulate a test for this situation. Should it be 1−sided or 2−sided? What value of $\alpha$ should you use? What conclusion does the test and the sample yield?

b) We are interested in the mean burning rate of a solid propellant used to power aircrew escape systems. We want to determine whether or not the mean burning rate is 50 cm/second. A sample of 10 specimens is tested and we observe $\overline{X} = 48.5$. Assume normality with $\sigma = 2.5$.

b) Ten individuals have participated in a diet modification program to stimulate weight loss. Their weight both before and after participation in the program is shown below:

| | |
|---|---|
| Before | 195, 213, 247, 201, 187, 210, 215, 246, 294, 310 |
| After | 187, 195, 221, 190, 175, 197, 199, 221, 278, 285 |

Is there evidence to support the claim that this particular diet-modification program is effective in producing mean weight reduction? Use $\alpha = 0.05$. Compute the associated $p$−value.

b) We want to test the hypothesis that the average content of containers of a particular lubricant equals 10L against the two-sided alternative. The contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, 9.5. Find the $p$−value of this two-sided test. Assume that the distribution of contents is normal. Note that if $x_i$ represent the measurements, $\sum_{i=1}^{10} x_i^2 = 1006.79$.

b) An engineer measures the weight of $n = 25$ pieces of steel, which follows a normal distribution with variance 16. The average weight for the sample is $\overline{X} = 6$. They want to test for $H_0 : \mu = 5$ against $H_1 : \mu > 5$. What is the $p$−value for the test?

b) The thickness of a plastic film (in mm) on a substrate material is thought to be influenced by the temperature at which the coating is applied. A completely randomized experiment is carried out. 11 substrates are coated at 125F, resulting in a sample mean coating thickness of $\overline{x}_1 = 103.5$ and a sample standard deviation of $s_1 = 10.2$. Another 11 substrates are coated at 150F, for which $\overline{x}_2 = 99.7$ and $s_2 = 11.7$ are observed. We want to test equality of means against the two-sided alternative. Assume that population variances are unknown but equal. The value of the appropriate test statistics and the decision are (for $\alpha = 0.05$):

b) The following output was produced with `t.test` command in R.

```
One Sample t-test
data:  x
t = 2.0128, df = 99, p-value = 0.02342
alternative hypothesis: true mean is greater than 0
```

Based on this output, which statement is correct?

a) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;
b) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu \neq 0$;
c) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;
d) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu < 0$;
e) The type I error is 0.02342.

e) A pharmaceutical company claims that a drug decreases a blood pressure. A physician doubts this claim. They test 10 patients and records results before and after the drug treatment:

```
Before=c(140,135,122,150,126,138,141,155,128,130)
After=c(135,136,120,148,122,136,140,153,120,128)
```

At the R command prompt, they type:

```
test.t(Before,After,alternative="greater")
```

```
data:  Before and After
t = 0.5499, p-value = 0.2946
alternative hypothesis: true
    difference in means is
    greater than 0
sample estimates: mean of x mean of y
                      136.5     133.8
```

Their assistant claims that the command should instead be:

```
test.t(Before,After,paired=TRUE,alternative="greater")
```

```
data: Before and After t = 3.4825,
    df = 9, p-value = 0.003456
alternative hypothesis: true
    difference in means is
    greater than 0
sample estimates: mean of the differences
                                       2.7
```

Which answer is best?

a) The assistant uses the correct command. There is *not enough* evidence to justify that the new drug decreases blood pressure;
b) The assistant uses the correct command. There is *enough* evidence to justify that the new drug decreases blood pressure for any reasonable choice of $\alpha$;
c) The physician uses the correct command. There is *not enough* evidence to justify that the new drug decreases blood pressure;
d) The physician uses the correct command. There is *enough* evidence to justify that the new drug decreases blood pressure for any reasonable choice of $\alpha$;
e) Nobody is correct, $t-$tests should not be used here.

e) A company claims that the mean deflection of a piece of steel which is 10ft long is equal to 0.012ft. A buyer suspects that it is bigger than 0.012ft. The following data $x_i$ has been collected:

$$0.0132, 0.0138, 0.0108, 0.0126, 0.0136,$$
$$0.0112, 0.0124, 0.0116, 0.0127, 0.0131.$$

Assuming normality and that $\sum_{i=1}^{10} x_i^2 = 0.0016$, what are the $p-$value for the appropriate one-sided test and the corresponding decision?

a) $p \in (0.05, 0.1)$ and reject $H_0$ at $\alpha = 0.05$.
b) $p \in (0.05, 0.1)$ and do not reject $H_0$ at $\alpha = 0.05$.
c) $p \in (0.1, 0.25)$ and reject $H_0$ at $\alpha = 0.05$.
d) $p \in (0.1, 0.25)$ and do not reject $H_0$ at $\alpha = 0.05$.

d) In an effort to compare the durability of two different types of sandpaper, 10 pieces of type $A$ sandpaper and 11 pieces of type $B$ sandpaper were subjected to treatment by a machine which measures abrasive wear. We have the following observations:

$$x_A : 27, 26, 24, 29, 30, 26, 27, 23, 28, 27; \qquad x_B : 24, 23, 22, 27, 24, 21, 24, 25, 24, 23, 20$$

Note that $\sum x_{A,i} = 267$, $\sum x_{B,i} = 257$, $\sum x_{A,i}^2 = 7169$, $\sum x_{B,i}^2 = 6041$. Assuming normality and equality of variances in abrasive wear for $A$ and $B$, we want to test for equality of mean abrasive wear for $A$ and $B$. What is the appropriate $p-$value for this test?

d) The following output was produced with a `t.test` command in `R`.

```
t = 32.9198, df = 999, p-value < 2.2e-16, alternative hypothesis: true mean is not equal to 0
```

Based on this output, which statement is correct?

a) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;
b) If the type I error is 0.05, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu \neq 0$;
c) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu > 0$;
d) If the type I error is 0.01, then we reject $H_0 : \mu = 0$ in favour of $H_1 : \mu < 0$.

d) A medical team wants to test whether a particular drug decreases diastolic blood pressure. Nine people have been tested. The team measured blood pressure before ($X$) and after ($Y$) applying the drug. The corresponding means were $\overline{X} = 91$, $\overline{Y} = 87$. The sample variance of the differences was $S_D^2 = 25$. What is the $p-$value for the appropriate one-sided test?

d) A researcher studies a difference between two programming languages. Twelve experts familiar with both languages were asked to write a code for a particular function using both languages and the time for writing those codes was registered. The observations are as follows.

```
Expert 01 02 03 04 05 06 07 08 09 10 11 12
Lang 1 17 16 21 14 18 24 16 14 21 23 13 18
Lang 2 18 14 19 11 23 21 10 13 19 24 15 29
```

Construct a 95% C.I. for the mean difference between the first and the second language. Do we have any evidence that the average time to write a function is shorter in one of the languages?

d) Consider a proportion of recaptured moths in the light-coloured ($p_1$) and the dark-coloured ($p_2$) populations. Among the $n_1 = 137$ light-coloured moths, $y_1 = 18$ were recaptured; among the $n_2 = 493$ dark-coloured moths, $y_2 = 131$ were recaptured. Is there a significant difference between the proportion of recaptured moths in both populations?

## Chapter References


[1] P. Boily, S. Davies, and J. Schellinck. *The Practice of Data Visualization* . Data Action Lab, 2023.

[2] P. Boily and J. Schellinck. *Introduction to Quantitative Consulting*. Quadrangle/Data Action Lab, 2025.

[3] P. Bruce and A. Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly, 2017.

[4] E.W. Gibson. 'The Role of $p-$Values in Judging the Strength of Evidence and Realistic Replication Expectations'. In: *Statistics in Biopharmaceutical Research* 13.1 (2021), pp. 6–18.

[5] R.V. Hogg and E.A Tanis. *Probability and Statistical Inference*. 7th. Pearson/Prentice Hall, 2006.

[6] M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods*. 2nd. Wiley, 1999.

[7] M.H. Kutner et al. *Applied Linear Statistical Models*. McGraw Hill Irwin, 2004.

[8] A. Reinhart. *Statistics Done Wrong: the Woefully Complete Guide*. No Starch Press, 2015.

[9] M.L. Rizzo. *Statistical Computing with R*. CRC Press, 2007.

[10] H. Sahai and M.I. Ageel. *The Analysis of Variance: Fixed, Random and Mixed Models*. Birkhäuser, 2000.

[11] D.S. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial (2nd ed.)* Oxford Science, 2006.

[12] R.E. Walpole et al. *Probability and Statistics for Engineers and Scientists*. 8th. Pearson Education, 2007.

# Classical Regression Analysis

# 8

by **Patrick Boily** (inspired by **Gilles Lamothe** and **Rafal Kulik**)

---

Regression analysis is quite likely the most frequent application of probability and statistics; it is used extensively in the physical and social sciences, and forms the backbone of statistical learning. No data scientist worthy of the name can be ignorant of this aspect of the discipline.

We use the term "classical" to differentiate the basic process from its myriad variants and modifications, which we discuss further in Chapter 20 (*Regression and Value Estimation*).

Our treatment borrows heavily from a classical reference [7]; other useful resources include [3, 5]. Note that the examples use R, which provides a suite of "natural" tools for regression analysis.


8.1 Preliminaries . . . . . 409
Random Variables . . . . . 409
Multivariate Calculus . . . . 416
Matrix Algebra . . . . . 417
Quadratic Forms . . . . . 417
Optimization . . . . . 419
8.2 Simple Linear Regression . 419
Least Squares Estimation . 421
Inference . . . . . 429
Estimation and Prediction . 437
Significance of Regression . 444
SLR in R . . . . . 446
8.3 Multiple Linear Regression 447
Least Squares Estimation . 448
Inference . . . . . 451
Power of a Test . . . . . 460
Determination Coefficients 461
Diagnostics . . . . . 461
8.4 Extensions of OLS . . . . . 468
Multicollinearity . . . . . 468
Polynomial Regression . . . 471
Interaction Effects . . . . . 474
Categorical Variables . . . . 477
Weighted Least Squares . . 477
Other Extensions . . . . . 480
8.5 OLS and Outliers . . . . . 481
Leverage and Extrapolation 481
Deleted Residuals . . . . . 483
Influential Observations . . 484
Cook's Distance . . . . . 485
8.6 Exercises . . . . . 486
Chapter References . . . . . 490


## 8.1 Preliminaries

Regression analysis is not a very complicated discipline ... assuming that its pre-requisites are mastered well. In this chapter, it will be useful to be familiar with a number of notions relating to:

- **random variables**;
- **multivariate calculus**;
- **linear algebra**;
- **quadratic forms**, and
- **optimization**.

### 8.1.1 Random Variables

A **random experiment** is a **process** (together with its **sample space** $\mathcal{S}$) for which it is impossible to predict the **outcome with certainty**. The **sample space** $\mathcal{S}$ is the set of the random experiment's **possible outcomes**.

A **random variable** $Y$ associated to this process is a function $Y : \mathcal{S} \to \mathbb{R}$. If the set $Y(\mathcal{S}) = \{Y(s) \mid s \in \mathcal{S}\}$ is **countable**, we say that $Y$ is a **discrete random variable**; if it is **uncountable**, we say that $Y$ is a **continuous randcom variable**.

Each r.v. $Y$ has a corresponding **probability function** $f(Y)$, which specifies the probabilities of the values taken by $Y$. $Y_1$ and $Y_2$ are **independent** when their **joint probability function** $f(Y_1, Y_2)$ is the product of the **individual** probability functions $f(Y_1)f(Y_2)$.

**Expectation, Variance, and Covariance** The **expectation operator** $\mathrm{E}\{\cdot\}$ is defined by

$$\mathrm{E}\{Y\} = \begin{cases} \sum_{Y(s)} Y(s) f(Y(s)), & \text{if } Y \text{ is discrete} \\ \int_{\mathbb{R}} Y f(Y)\, dy, & \text{if } Y \text{ is continuous} \end{cases}$$

The expectation $\mathrm{E}\{Y\}$ is the **average value** that we would expect to observe if the experiment is repeated a large number of times. The expectation is sometimes also called the **mean** of $Y$, denoted $\overline{Y}$; it is thus a measure of $Y$'s **centrality**.

The **variance operator** $\sigma^2\{\cdot\}$ is defined by

$$\sigma^2\{Y\} = \mathrm{E}\left\{(Y - \mathrm{E}\{Y\})^2\right\} = \mathrm{E}\left\{Y^2\right\} - \left(\mathrm{E}\{Y\}\right)^2.$$

It is often denoted by $\mathrm{Var}(Y)$. It is a measure of $Y$'s **dispersion** (large variances are associated with r.v. with **heavy dispersion**, and *vice-versa*).

The **covariance operator** $\sigma\{\cdot,\cdot\}$ is defined by

$$\sigma\{Y, W\} = \mathrm{E}\{(Y - \mathrm{E}\{Y\})(W - \mathrm{E}\{W\})\} = \mathrm{E}\{YW\} - \mathrm{E}\{Y\}\,\mathrm{E}\{W\}.$$

It is often denoted by $\mathrm{Cov}(Y, W)$. It is a measure of the **strength of the linear relationship** between two r.v. (large covariance magnitudes are associated with **linearity**, but "large" is a relative concept).

The **standard deviation operator** $\sigma\{\cdot\}$ is defined by

$$\sigma\{Y\} = \sqrt{\sigma^2\{Y\}}.$$

It is always non-negative.

The **correlation operator** $\rho\{\cdot,\cdot\}$ is defined by

$$\rho\{Y, W\} = \frac{\sigma\{Y, W\}}{\sigma\{Y\}\,\sigma\{W\}},$$

assuming that $\sigma\{Y\}\,\sigma\{W\} \neq 0$. When $\rho\{Y, W\} = 0$, we say that the r.v. are **uncorrelated**.

**Operator Properties** Let $Y, Y_i, W$ be random variables, $c, a_i, b_i, c_i \in \mathbb{R}$, $i = 1, \ldots, n$. Then:

- $\mathrm{E}\{\cdot\}$ is **linear** on the space of r.v.: $\mathrm{E}\{aY + b\} = a\mathrm{E}\{Y\} + b$ and

$$\mathrm{E}\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i \mathrm{E}\{Y_i\}$$

- $\sigma^2\{aY + b\} = a^2\sigma^2\{Y\}$ and

$$\sigma^2\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma\{Y_i, Y_j\} = \sum_{i=1}^{n} a_i^2 \sigma^2\{Y_i\} + \sum_{i \neq j} a_i a_j \sigma\{Y_i, Y_j\}$$

- $\sigma\{Y, Y\} = \sigma^2\{Y\}$ and $\sigma\{Y, W\} = \sigma\{W, Y\}$
- $\sigma\{a_1 Y + b_1, a_2 W + b_2\} = a_1 a_2 \sigma\{Y, W\}$

- $\{Y_i\}$ **uncorrelated** $\implies$

$$\sigma\left\{\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i\right\} = \sum_{i=1}^{n} a_i c_i \sigma^2\{Y_i\}$$

- $\sigma\{Y, W\} < 0 \iff$ observations of $Y$ above $\overline{Y}$ tend to accompany corresponding observations of $W$ below $\overline{W}$, and *vice-versa*.
- $\sigma\{Y, W\} > 0 \iff$ observations of $Y$ above $\overline{Y}$ tend to accompany corresponding observations of $W$ above $\overline{W}$, and *vice-versa*.
- $\sigma\{Y, W\} = 0 \implies Y$ and $W$ are **uncorrelated**
- $Y, W$ **independent** $\implies \rho\{Y, W\} = 0$ (uncorrelated)
- $\rho\{Y, W\} = 0 \not\implies Y, W$ **independent**, however
- $|\rho\{Y, W\}| \leq 1$ (consequence of the Cauchy-Schwartz inequality)
- $|\rho\{Y, W\}| = 1 \iff Y = aW + b$ for some $a, b \in \mathbb{R}$,

**Random Vectors** If $Y_1, \ldots, Y_n$ are random variables, then

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

is a **random vector**. The **expectation** of $\mathbf{Y}$ is

$$\mathrm{E}\{\mathbf{Y}\} = \begin{pmatrix} \mathrm{E}\{Y_1\} \\ \vdots \\ \mathrm{E}\{Y_n\} \end{pmatrix}.$$

The components of $\mathbf{Y}$ need not all have identical distributions.

The **variance-covariance matrix** of $\mathbf{Y}$ is the symmetric matrix

$$\sigma^2\{\mathbf{Y}\} = (g_{i,j}), \quad \text{where } g_{i,j} = \begin{cases} \sigma^2\{Y_i\} & i = j \\ \sigma\{Y_i, Y_j\} & i \neq j \end{cases}$$

or

$$\sigma^2\{\mathbf{Y}\} = \begin{pmatrix} \sigma^2\{Y_1\} & \cdots & \sigma\{Y_1, Y_n\} \\ \vdots & \ddots & \vdots \\ \sigma\{Y_1, Y_n\} & \cdots & \sigma^2\{Y_n\} \end{pmatrix}$$

If the components of $\mathbf{Y}$ are **independent** and all have the **same variance** $\sigma^2$, then

$$\sigma^2\{\mathbf{Y}\} = \sigma^2 \mathbf{I}_n.$$

In practice, we usually work with **samples** of the random variables. Let $\{(X_i, Y_i)\}_{i=1}^n$ be observed from the joint distribution of $(X, Y)$:

- the **sample means**

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  are unbiased estimators of $\mathrm{E}\{X\}$ and $\mathrm{E}\{Y\}$, respectively;

- the **sample variances**

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{and} \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

  are unbiased estimators of $\sigma^2\{X\}$ and $\sigma^2\{Y\}$, respectively;
- the **sample variances**

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

  is an unbiased estimator of $\sigma\{X, Y\}$.

**Important Distributions** The **(cumulative) distribution function** (c.d.f.) of any continuous random variable $Y$ is defined by

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(t)\,dt$$

viewed as a function of a real variable $y$.

Alternatively, We can describe the **distribution** of $Y$ *via* the following relationship between $f_Y(y)$ and $F_Y(y)$:

$$f_Y(y) = \frac{d}{dy}F_Y(y).$$

The **probability density function** (p.d.f.) of a continuous random variable $Y$ is function[1] $f_Y : Y(\mathcal{S}) \to \mathbb{R}$ with:

1: Integrable function, that is.

- $f_Y(y) > 0$ for all $y \in Y(\mathcal{S})$
- $\lim_{y\to\pm\infty} f_Y(y) = 0$;
- $\int_{\mathcal{S}} f_Y(y)\,dy = 1$;

For any $a, b$, we have

$$P(a < Y < b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b)$$
$$= F_Y(b) - F_Y(a) = \int_a^b f(y)\,dy.$$

The following distributions all play an important role in the theory of regression analysis (see Section 6.3.3 for more information).

A random variable $Y$ follows a **normal distribution** $\mathcal{N}(\mu, \sigma^2)$ of mean $\mu$ and variance $\sigma^2$ if the c.d.f. of $Y$ is

$$F_Y(y) = P(Y \leq y) = \Phi(y),$$

with

$$f_Y(y) = \Phi'(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right).$$

A random variable $Y$ follows a $\chi^2$ **distribution** $\chi^2(\nu)$ if its p.d.f. is

$$f_Y(y;\nu) = \begin{cases} \dfrac{y^{\frac{\nu}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}, & y > 0; \\ 0, & \text{otherwise.} \end{cases}$$

where $\Gamma(\cdot)$ is the **Gamma function**. If $U_i \sim \chi^2(\nu_i)$, $i = 1, 2$, and $U_1, U_2$ are independent, then

$$U = U_1 + U_2 \sim \chi^2(\nu_1) + \chi^2(\nu_2) = \chi^2(\nu_1 + \nu_2).$$

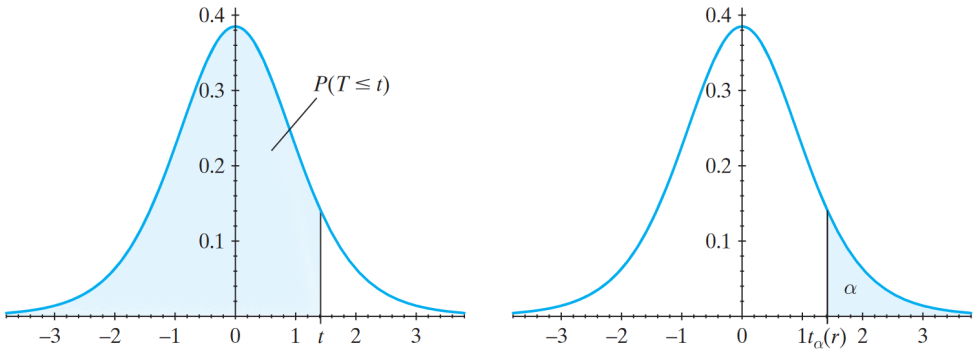
There is an important link between the standard normal distribution and the $\chi^2(1)$ distribution: if $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2(1)$.

If $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(\nu)$, where $Z, U$ are independent, then

$$t = \frac{Z}{\sqrt{U/\nu}} \sim t(\nu)$$

follows a **Student $T$-distribution with $\nu$ degrees of freedom**.

**Table VI** The $t$ Distribution




$$P(T \le t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)(1 + w^2/r)^{(r+1)/2}}\, dw$$

$$P(T \le -t) = 1 - P(T \le t)$$

| | $P(T \le t)$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.60 | 0.75 | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 |
| $r$ | $t_{0.40}(r)$ | $t_{0.25}(r)$ | $t_{0.10}(r)$ | $t_{0.05}(r)$ | $t_{0.025}(r)$ | $t_{0.01}(r)$ | $t_{0.005}(r)$ |
| 1 | 0.325 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 0.289 | 0.816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 0.277 | 0.765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |

**Figure 8.1:** Cumulative distribution function of Student's $T$ distribution, with some critical values for $\nu = 1, 2, 3$ degrees of freedom [6].

If $U_i \sim \chi^2(\nu_i)$, $i = 1, 2$ and $U_1, U_2$ are independent, then

$$F = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1, \nu_2)$$

follows the **Fisher's distribution with $\nu_1$ and $\nu_2$ degrees of freedom**.

In practice, we do not use tables, but rather statistical software (such as R),

**Table VII** The $F$ Distribution

$$P(F \le f) = \int_0^f \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

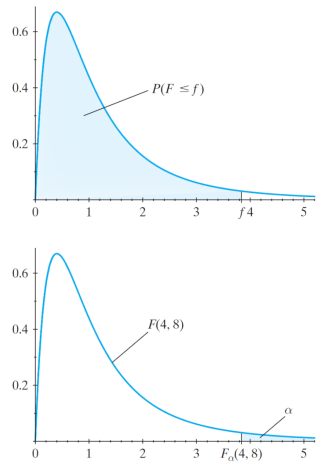






**Table VII** *continued*

$$P(F \le f) = \int_0^f \frac{\Gamma[(r_1+r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

| $\alpha$ | $P(F \le f)$ | Den. d.f. $r_2$ | Numerator Degrees of Freedom, $r_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0.05 | 0.95 | 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 | 241.9 |
| 0.025 | 0.975 | | 647.79 | 799.50 | 864.16 | 899.58 | 921.85 | 937.11 | 948.22 | 956.66 | 963.28 | 968.63 |
| 0.01 | 0.99 | | 4052 | 4999.5 | 5403 | 5625 | 5764 | 5859 | 5928 | 5981 | 6022 | 6056 |
| 0.05 | 0.95 | 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 |
| 0.025 | 0.975 | | 38.51 | 39.00 | 39.17 | 39.25 | 39.30 | 39.33 | 39.36 | 39.37 | 39.39 | 39.40 |
| 0.01 | 0.99 | | 98.50 | 99.00 | 99.17 | 99.25 | 99.30 | 99.33 | 99.36 | 99.37 | 99.39 | 99.40 |
| 0.05 | 0.95 | 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 |
| 0.025 | 0.975 | | 17.44 | 16.04 | 15.44 | 15.10 | 14.88 | 14.73 | 14.62 | 14.54 | 14.47 | 14.42 |
| 0.01 | 0.99 | | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.35 | 27.23 |
| 0.05 | 0.95 | 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 |
| 0.025 | 0.975 | | 12.22 | 10.65 | 9.98 | 9.60 | 9.36 | 9.20 | 9.07 | 8.98 | 8.90 | 8.84 |
| 0.01 | 0.99 | | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 | 14.55 |
| 0.05 | 0.95 | 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 |
| 0.025 | 0.975 | | 10.01 | 8.43 | 7.76 | 7.39 | 7.15 | 6.98 | 6.85 | 6.76 | 6.68 | 6.62 |
| 0.01 | 0.99 | | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.16 | 10.05 |
| 0.05 | 0.95 | 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 |
| 0.025 | 0.975 | | 8.81 | 7.26 | 6.60 | 6.23 | 5.99 | 5.82 | 5.70 | 5.60 | 5.52 | 5.46 |
| 0.01 | 0.99 | | 13.75 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 |
| 0.05 | 0.95 | 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 |
| 0.025 | 0.975 | | 8.07 | 6.54 | 5.89 | 5.52 | 5.29 | 5.12 | 4.99 | 4.90 | 4.82 | 4.76 |
| 0.01 | 0.99 | | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 |

**Figure 8.2:** Cumulative distribution function of Fisher's $F$ distribution, with some critical values [6].

to compute important quantities. The functions `qnorm()`, `qt()`, and `qf()`, for instance, find the quantiles of the corresponding distributions.

```
qnorm(0, mean = 0, sd = 1)
qnorm(1, mean = 0, sd = 1)
qnorm(0.5, mean = 0, sd = 1)
qnorm(0.25, mean = 4, sd = 2)
```

```
[1] -Inf
[1] Inf
[1] 0
[1] 2.65102
```

```
qt(0.95, df = 20)
qf(0.975, df1 = 1, df2 = 19)
```

```
[1] 1.724718
[1] 5.921631
```

The functions `dnorm()`, `dt()`, and `df()` compute the value of the p.d.f. of the corresponding random variables at specified points in their domain.

```
dnorm(0, mean = 0, sd = 1)
dnorm(1, mean = 0, sd = 1)
dnorm(-1, mean = 0, sd = 1)
dnorm(3, mean = 4, sd = 2)
```

```
[1] 0.3989423
```

```
[1] 0.2419707
[1] 0.2419707
[1] 0.1760327
```

```
qf(2, df1 = 1, df2 = 19)
```

```
[1] 0.2844237
```

The functions `pnorm()`, `pt()`, and `pf()` compute the value of the c.d.f. of the corresponding random variables at specified points in their domain.

```
pnorm(0, mean = 0, sd = 1)
pnorm(1, mean = 0, sd = 1)
pnorm(-1, mean = 0, sd = 1)
pnorm(3, mean = 4, sd = 2)
```

```
[1] 0.5
[1] 0.8413447
[1] 0.1586553
[1] 0.3085375
```

```
pt(-1, df = 20)
pf(2, df1 = 1, df2 = 19)
```

```
[1] 0.1646283
[1] 0.8265229
```

Finally, we can generate (pseudo-)random values drawn from the corresponding distribution with `rnorm()`, `rt()`, and `rf()`.

```
set.seed(0) # for replicability
rnorm(10, mean = 0, sd = 1)
```

```
[1]  1.262954285 -0.326233361  1.329799263  1.272429321  0.414641434
[6] -1.539950042 -0.928567035 -0.294720447 -0.005767173  2.404653389
```

```
rt(5, df = 20)
```

```
[1]  0.9000978 -0.9947734 -0.4056054 -0.8546851 -1.3176242
```

```
rf(8, df1 = 1, df2 = 19)
```

```
[1] 1.8583849 1.8137178 0.8621754 0.5502212 1.1415165
[6] 2.4191686 1.8868591 0.6094574
```

**Central Limit Theorems** There are variants on a fundamental result of probability statistics that forms the basis of a fair chunk of applications, not only for regression analysis, but also for sampling theory, the design of experiments, time series analysis, and so on. We present them here without proof.

**Theorem:** let $X_1, \ldots, X_n$ be independent normal random variables with mean $\mu_1, \ldots, \mu_n$ and standard deviations $\sigma_1, \ldots, \sigma_n$. Then

$$X_1 + \cdots + X_n \sim \mathcal{N}(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2).$$

If $\mu_i \equiv \mu$ and $\sigma_i^2 \equiv \sigma$ for $i = 1, \ldots, n$, then $X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$.

**Theorem:** let $X_1, \ldots, X_n$ be independent normal random variables with mean $\mu$ and standard deviation $\sigma$ . Let $\overline{X}$ be the sample mean. Then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

**Theorem:** let $X_1, \ldots, X_n$ be independent random variables with mean $\mu$ and standard deviation $\sigma$ . Let $\overline{X}$ be the sample mean. Then

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to Z \sim \mathcal{N}(0, 1), \quad \text{as } n \to \infty.$$

**Theorem:** let $X_1, \ldots, X_n$ be independent normal random variables with mean $\mu$ and common variance. Let $\overline{X}$ and $s^2$ be the sample mean and the sample variance, respectively. Then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

follows a Student $T$ distribution with $\nu = n - 1$ degrees of freedom.

### 8.1.2 Multivariate Calculus

From a regression analysis's perspective, the main tool of multivariate calculus is the gradient of a multivariate differentiable function.[2]

2: More on the general topic can be found in Chapter 2 and in [2, 1, 4].

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a **differentiable** function. If $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the **derivative** (or **gradient**) of $f$ with respect to $\mathbf{Y}$ is

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \begin{pmatrix} \frac{\partial f(\mathbf{Y})}{\partial Y_1} \\ \vdots \\ \frac{\partial f(\mathbf{Y})}{\partial Y_n} \end{pmatrix}.$$

The gradient is a **linear operator**:

$$\nabla_{\mathbf{Y}}(af + bg)(\mathbf{Y}) = a\nabla_{\mathbf{Y}} f(\mathbf{Y}) + b\nabla_{\mathbf{Y}} g(\mathbf{Y}).$$

The gradient of constant and of linear functions is particular easy to find: if $f(\mathbf{Y}) \equiv a$, then $\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \mathbf{0}$; if $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v}$, then $\nabla_{\mathbf{Y}} f(\mathbf{Y}) = \mathbf{v}$.

### 8.1.3 Matrix Algebra

It turns out that the important concepts of regression analysis are more easily expressed (and ultimately, understandable) in matrix notation.[3]

3: See Chapter 3 and [8] for more information.

Let $A \in M_{m,n}(\mathbb{R})$ and $\mathbf{Y}$ be a random vector. Consider $\mathbf{W} = A\mathbf{Y}$. Then

$$\mathrm{E}\{\mathbf{W}\} = A\mathrm{E}\{\mathbf{Y}\} \quad \text{and} \quad \sigma^2\{\mathbf{W}\} = A\sigma^2\{\mathbf{Y}\}A^\top.$$

Furthermore, if $\mathbf{Y} \sim \mathcal{N}\left(\mathrm{E}\{\mathbf{Y}\}, \sigma^2\{\mathbf{Y}\}\right)$, then

$$\mathbf{W} \sim \mathcal{N}\left(\mathrm{E}\{\mathbf{W}\}, \sigma^2\{\mathbf{W}\}\right) = \mathcal{N}\left(A\mathrm{E}\{\mathbf{Y}\}, A\sigma^2\{\mathbf{Y}\}A^\top\right).$$

If $A \in M_{n,n}(\mathbb{R})$, the **trace** of $A$ is

$$\mathrm{trace}(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

The trace is a **linear operator**: $\mathrm{trace}(kA + B) = k \cdot \mathrm{trace}(A) + \mathrm{trace}(B)$; we also have $\mathrm{trace}(AB) = \mathrm{trace}(BA)$.[4]

4: Assuming, of course, that the matrices are **compatible** with respect to the product.

The **transpose** of a matrix $A$, denoted by $A^\top$, is obtained by interchanging its **rows** and its **columns**, or simply by **reflecting** the matrix along its **primary diagonal**.

**Properties:** if $A \in M_{m,n}(\mathbb{R})$ and $k \in \mathbb{R}$, then

- $(A^\top)^\top = A$
- $k^\top = k$
- $(kA + B)^\top = kA^\top + B^\top$
- $(AB)^\top = B^\top A^\top$

### 8.1.4 Quadratic Forms

A **symmetric quadratic form** in $Y_1, \ldots, Y_n$ is an expression of the form

$$Q_A(\mathbf{Y}) = \mathbf{Y}^\top A\mathbf{Y} = \sum_{i,j=1}^{n} a_{i,j}Y_iY_j,$$

where $A$ is an $n \times n$ **symmetric matrix** ($A^\top = A$). A number of important quantities in regression analysis can be expressed as such forms.

The **degrees of freedom** for a symmetric quadratic form $Q_A(\mathbf{Y})$ can be obtained by computing the **rank** of the associated matrix $A$. For instance, the symmetric matrix associated with the symmetric quadratic form

$$Q_A(\mathbf{Y}) = 4Y_1^2 + 7Y_1Y_2 + 2Y_2^2$$

is

$$A = \begin{pmatrix} 4 & 7/2 \\ 7/2 & 2 \end{pmatrix}.$$

As $\mathrm{rank}(A) = 2$, $Q_A$ has 2 degrees of freedom.

**Theorem:** let $Q_1, \ldots Q_K$ be symmetric quadratic forms of $\mathbf{Y}$ with respective symmetric matrices $A_1, \ldots, A_K$. If $a_i \in \mathbb{R}$ for $i = 1, \ldots, K$, then

$$Q = a_1Q_1 + \cdots + a_KQ_K$$

is a symmetric quadratic form of $\mathbf{Y}$ with symmetric matrix

$$A = a_1 A_1 + \cdots + a_K A_K.$$

For a general $n \times n$ matrix $B$, we have

$$\nabla_{\mathbf{Y}} \left( \mathbf{Y}^\top B \mathbf{Y} \right) = (B^\top + B)\mathbf{Y}.$$

Thus the gradient of a symmetric quadratic form $Q_A(\mathbf{Y})$ is

$$\nabla_{\mathbf{Y}} Q_A(\mathbf{Y}) = 2A\mathbf{Y}.$$

It can be shown that **every** expression of the form $\mathbf{Y}^\top B \mathbf{Y}$ can be associated to a symmetric matrix $A$, even if $B$ is not itself symmetric, so we may as well assume that every such form is symmetric.[5]

The **eigenvalues** of an $n \times n$ matrix $A$ are the roots of the **characteristic polynomial** $p_A(\lambda)$ of $A$: $p_A(\lambda) = \det(A - \lambda \mathbf{I}_n) = 0$.[6] If $\lambda$ is an eigenvalue of $A$, then there exists $\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{v} = \lambda \mathbf{v}$.[7]

Consider a quadratic form $Q_A(\mathbf{Y})$, with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$:

- if $\lambda_i > 0$ for all $i$, we say that $Q_A(\mathbf{Y})$ and $A$ are **positive definite**;
- if $\lambda_i < 0$ for all $i$, we say that $Q_A(\mathbf{Y})$ and $A$ are **negative definite**;
- if $\lambda_i \lambda_j < 0$ for some $i, j$, we say that $Q_A(\mathbf{Y})$ and $A$ are **indefinite**.

**Cochran's Theorem** Let $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Suppose that

$$\mathbf{Y}^\top \mathbf{Y} = Q_1(\mathbf{Y}) + \cdots + Q_K(\mathbf{Y}),$$

with $Q_k$ positive (semi-)definite quadratic forms with $r_k = \operatorname{rank}(A_k)$ degrees of freedom, $k = 1, \ldots, K$. If $r_1 + \cdots + r_K = n$, then $Q_1(\mathbf{Y}), \ldots, Q_K(\mathbf{Y})$ are **independent** random variables and

$$\frac{Q_k(\mathbf{Y})}{\sigma^2} \sim \chi^2(r_k), \quad k = 1, \ldots, K.$$

In particular, if $K = 2$ and $r_1 = r$, then $Q_2(\mathbf{Y})/\sigma^2 \sim \chi^2(n - r)$.

**Important Quadratic Forms** For any positive integer $n$, we define two **special matrices**:

$$\mathbf{J}_n = \mathbf{J} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{1}_{n \times 1} = \mathbf{1}_n = \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Note that $\mathbf{1}_n^\top \mathbf{1}_n = n$ and $\mathbf{1}_n \mathbf{1}_n^\top = \mathbf{J}_n$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ be a random vector. What are the symmetric matrices associated with:

$$Q_A(\mathbf{Y}) = \sum_{i=1}^n Y_i^2, \quad Q_B(\mathbf{Y}) = n\overline{Y}^2, \quad \text{and} \quad Q_C(\mathbf{Y}) = \sum_{i=1}^n (Y_i - \overline{Y})^2?$$

5: The role played by quadratic forms in multi-variable calculus is analogous to the role played by $f(x) = ax^2$ in calculus.

6: There are $n$ such (complex) roots, not all necessarily distinct.

7: If $A$ is symmetric, all of its eigenvalues are **real**.

We re-write the quadratic forms in $\mathbf{Y}$ to obtain:

$$Q_A(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{Y} = \mathbf{Y}^\top \mathbf{I}_n \mathbf{Y} \implies A = \mathbf{I}_n;$$
$$Q_B(\mathbf{Y}) = n\left(\frac{1}{n}\sum_{i=1}^n Y_i\right)^2 = \frac{1}{n}\sum_{i,j=1}^n Y_i Y_j = \frac{1}{n}\mathbf{Y}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{Y} \implies B = \frac{1}{n}\mathbf{J}_n;$$
$$Q_C(\mathbf{Y}) = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 = \mathbf{Y}^\top \mathbf{I}_n \mathbf{Y} - \frac{1}{n}\mathbf{Y}^\top \mathbf{J}_n \mathbf{Y} \implies C = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n.$$

Since $\text{rank}(A) = n$, $\text{rank}(B) = 1$, and $\text{rank}(C) = n-1$, Cochran's Theorem implies that $Q_B(\mathbf{Y})$, and $Q_C(\mathbf{Y})$ are **independent** random variable, and that

$$\frac{Q_A(\mathbf{Y})}{\sigma^2} = \frac{\mathbf{Y}^\top \mathbf{Y}}{\sigma^2} \sim \chi^2(n), \quad \frac{Q_B(\mathbf{Y})}{\sigma^2} = \frac{n\overline{Y}^2}{\sigma^2} \sim \chi^2(1), \quad \frac{Q_C(\mathbf{Y})}{\sigma^2} = \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1).$$

### 8.1.5 Optimization

Let $A$ be a symmetric $n \times n$ matrix, $\mathbf{v} \in \mathbb{R}^n$, $c \in \mathbb{R}$. Consider the function

$$f(\mathbf{Y}) = \frac{1}{2}\mathbf{Y}^\top A \mathbf{Y} - \mathbf{Y}^\top \mathbf{v} + c.$$

Note that $f$ is **differentiable**. The **critical points** of $f$ satisfy

$$\nabla_{\mathbf{Y}} f(\mathbf{Y}) = A\mathbf{Y} - \mathbf{v} = \mathbf{0} \implies A\mathbf{Y} = \mathbf{v}.$$

If $A$ is **invertible** ($\det(A) \neq 0$), there is a **unique** critical point $\mathbf{Y}^* = A^{-1}\mathbf{v}$. If $A$ is **singular** ($\det(A) = 0$), there is **no** critical point if $\mathbf{v} \notin \text{range}(A)$, or there are **infinitely many** critical points if $\mathbf{v} \in \text{range}(A)$.

When $A$ is **invertible**:

- if $A$ is **positive definite**, then $f$ reaches its **global minimum** at $\mathbf{Y}^* = A^{-1}\mathbf{v}$;
- if $A$ is **negative definite**, then $f$ reaches its **global maximum** at $\mathbf{Y}^* = A^{-1}\mathbf{v}$;
- if $A$ is **indefinite** (if $A$ has positive **and** negative eigenvalues), then $\mathbf{Y}^* = A^{-1}\mathbf{v}$ is a **saddle point** for $f$.

If the eigenvalues could be **zero**, we replace "definite" by "semi-definite" throughout.

## 8.2 Simple Linear Regression

We start by considering a simple scenario, with only two **continuous** variables: a **response** $Y$ and a **predictor** $X$.

#### Examples

- $X$: age; $Y$: height
- $X$: age; $Y$: salary
- $X$: income; $Y$: life expectancy
- $X$: number of sunlight hours; $Y$: plant biomass

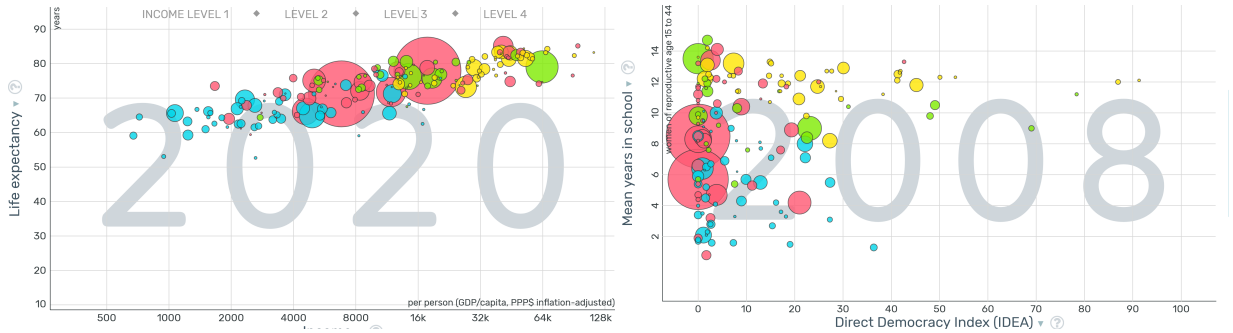



**Figure 8.3:** Response and predictor in the Gapminder data [10, 9]; life expectancy $Y$ against the logarithm of the GDP per capita $X$ (left); mean years in schooling $Y$ against direct democracy index $X$ (right).

We hope that there might be a **functional relationship** $Y = f(X)$ between $X$ and $Y$. In practice (assuming that a relationship even exists), the best that we may be able to achieve is a **statistical relationship**

$$Y = f(X) + \varepsilon,$$

where

- $f(X)$ is the **response function**;
- $\varepsilon$ is the **random error** (or noise).

In **simple linear regression**, we assume that the response function satisfies

$$f(X) = \beta_0 + \beta_1 X.$$

The building blocks of regression analysis are the **observations**:

$$(X_i, Y_i), \quad i = 1, \ldots, n.$$

In an ideal setting, these observations are **(jointly) randomly sampled**, according to some appropriate design.[8]

8: See Chapters 11 and 10.

The **simple linear regression model** (SLRM) is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_0, \beta_1$ are **unknown parameters** (which we want to find) and $\varepsilon_i$ is the **random error on the $i$th observation** (or case).

The SLRM **assumption on the error structure** is that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.[9] Let us unpack the statement: since $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$: we have

9: We use matrix notation to keep the assumption **compact**.

- $\mathrm{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0} \implies \mathrm{E}\{\varepsilon_i\} = 0, \quad i = 1, \ldots, n;$
- $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \mathbf{I}_n \implies \sigma^2\{\varepsilon_i\} = \sigma^2, \quad i = 1, \ldots, n;$
- $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \mathbf{I}_n \implies \sigma\{\varepsilon_i, \varepsilon_j\} = 0, \quad \text{for all } i \neq j.$

This means that the errors $\{\varepsilon_i\}$ are **uncorrelated**, with **mean** 0 and **constant variance**.

In other words, the **dispersion** of observations is **constant** around the regression line.

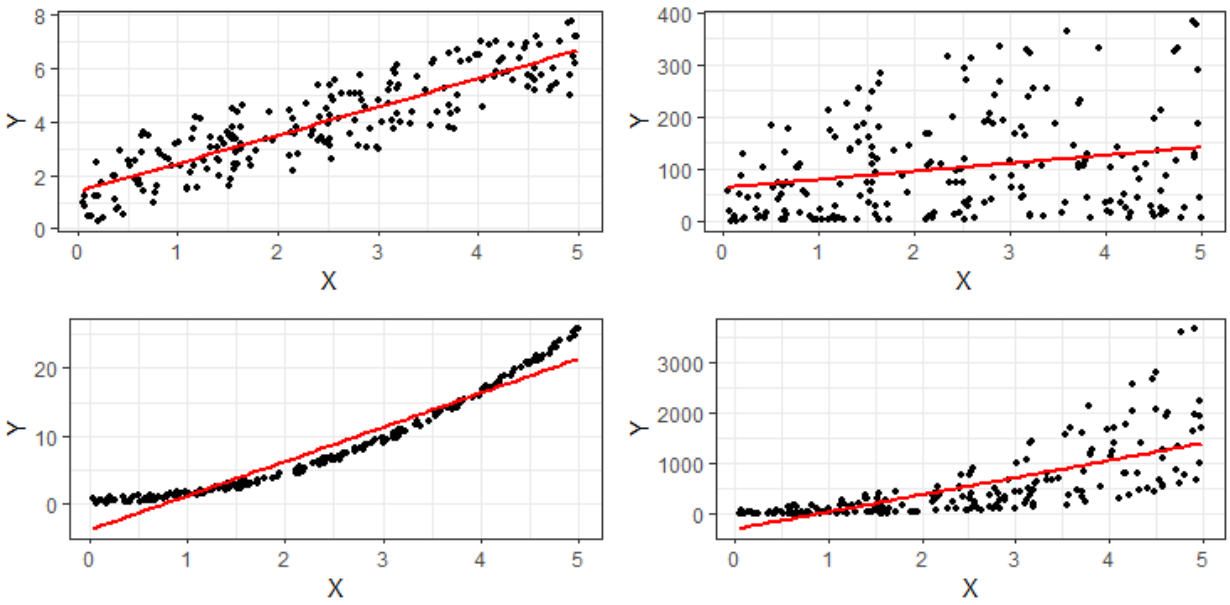



**Figure 8.4:** Illustrations of failed SLRM assumptions: constant, uncorrelated variance (top left); non-constant uncorrelated variance (top right); constant correlated variance (bottom left); non-constant correlated variance (bottom right).

### 8.2.1 Least Squares Estimation

We treat the predictor values $X_i$ as constant, for $i = 1, \ldots, n$.[10] Since $\mathrm{E}\{\varepsilon_i\} = 0$, the **expected** (or mean) **response given** $X_i$ is thus

10: That is, we assume that there is **no measurement error**.

$$\mathrm{E}\{Y_i \mid X_i\} = \mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i \mid X_i\} = \beta_0 + \beta_1 X_i + \mathrm{E}\{\varepsilon_i\} = \beta_0 + \beta_1 X_i.$$

The **deviation at** $X_i$ is the difference between the observed response $Y_i$ and the expected response $\mathrm{E}\{Y_i \mid X_i\}$:

$$e_i = Y_i - \mathrm{E}\{Y_i \mid X_i\};$$

the deviation can be **positive** (if the point lies **above** the line) or **negative** (if it lies **below**).

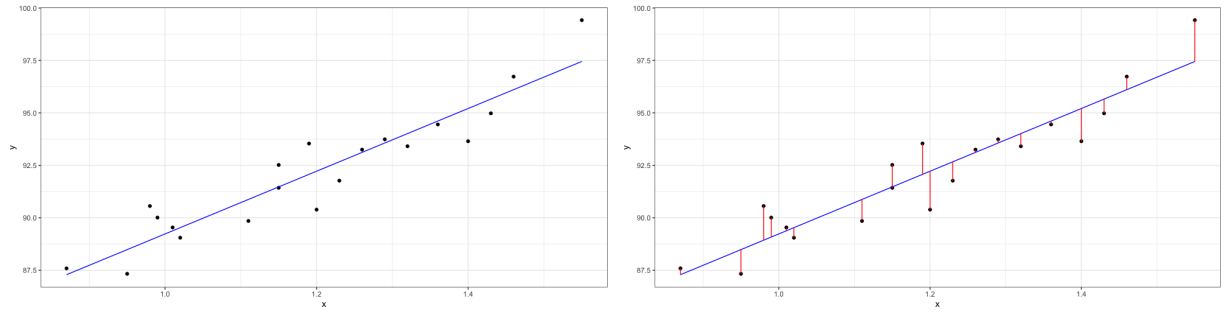


**Figure 8.5:** Line of best fit and deviations (residuals) for a simple dataset.

How do we find **estimators** for $\beta_0$ and $\beta_1$? Incidentally, how do we determine if the fitted line is a **good model for the data**?

Consider the function

$$Q(\boldsymbol{\beta}) = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mathrm{E}\{Y_i|X_i\})^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

If $Q(\boldsymbol{\beta})$ is "small", then the sum of the **squared residuals** is "small", and so we would expect the line $Y = \beta_0 + \beta_1 X$ to be a good fit for the data. The **least-square estimators** of the SLR problem are the pair $\mathbf{b} = (b_0, b_1)$ which minimizes the function $Q$ with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

We must then find the critical points of $Q(\boldsymbol{\beta})$, i.e., solve $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$. Thus, we must solve the following system:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-1) = 0$$
$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-X_i) = 0.$$

This is a linear system of two equations in the two unknowns $\beta_0, \beta_1$, known as the **normal equations**. As seen in Chapter 3, it has either **no solution**, a **unique solution**, or **infinitely many solutions**.[11]

11: From now on, we drop the $\mid X_i$ when we use the $\mathrm{E}\{\cdot \mid X_i\}$.

**Normal Equations** These equations reduce to the following pair:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i, \qquad \sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2.$$

If we use the following shorthand notation:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,\ S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2,\ S_{xy} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}),$$

it is not too difficult to show that

$$\sum_{i=1}^{n} X_i^2 = S_{xx} + n\overline{X}^2 \quad \text{and} \quad \sum_{i=1}^{n} X_i Y_i = S_{xy} + n\overline{XY}.$$

With this notation, the normal equations further reduce to

$$n\overline{Y} = n\beta_0 + n\overline{X}\beta_1, \qquad S_{xy} + n\overline{XY} = n\overline{X}\beta_0 + (S_{xx} + n\overline{X}^2)\beta_1.$$

In matrix form, this can be written as:

$$\begin{bmatrix} 1 & \overline{X} \\ n\overline{X} & S_{xx} + n\overline{X}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix}.$$

A linear system $A\boldsymbol{\beta} = \mathbf{v}$ has a unique solution $\boldsymbol{\beta} = A^{-1}\mathbf{v}$ if the determinant of the coefficient matrix $A$ is non-zero.

In our case, the determinant is

$$S_{xx} + n\overline{X}^2 - n\overline{XX} = S_{xx} > 0 \iff s_X^2 \neq 0.$$

The unique solution is thus

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & \overline{X} \\ n\overline{X} & S_{xx} + n\overline{X}^2 \end{bmatrix}^{-1} \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} S_{xx} + n\overline{X}^2 & -\overline{X} \\ -n\overline{X} & 1 \end{bmatrix} \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix} = \frac{1}{S_{xx}} \begin{bmatrix} (S_{xx} + n\overline{X}^2)\overline{Y} - \overline{X}(S_{xy} + n\overline{XY}) \\ -n\overline{XY} + S_{xy} + n\overline{XY} \end{bmatrix},$$

which reduces to

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = \begin{bmatrix}\overline{Y} - \overline{X}\cdot S_{xy}/S_{xx}\\ S_{xy}/S_{xx}\end{bmatrix}$$

Set $b_0 = \beta_0$ and $b_1 = \beta_1$. Then we may write:

$$b_1 = \frac{S_{xy}}{S_{xx}}\ (\textbf{slope}) \quad \text{and} \quad b_0 = \overline{Y} - b_1\overline{X}\ (\textbf{intercept}).$$

By analogy with $S_{xx}$ (the **total variation of the predictor**), we can also define the **total variation of the response** $S_{yy}$, a quantity that will play an important role in this chapter:[12]

12: And in Chapters 11 and 10.

$$S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2;$$

If the $X_i$ are fixed, $b_0, b_1$ are **linear combinations** of the $Y_i$:

$$b_1 = \frac{1}{S_{xx}}\sum_{i=1}^{n}(X_i - \overline{X})Y_i - \frac{\overline{Y}}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}Y_i,$$

$$b_0 = \sum_{i=1}^{n}\frac{Y_i}{n} - \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{XX}}Y_i\overline{X} = \sum_{i=1}^{n}\left[\frac{1}{n} - \overline{X}\frac{(X_i - \overline{X})}{S_{XX}}\right]Y_i.$$

**Properties of Least Squares Estimators** Both $b_0, b_1$ are **unbiased estimators** of their respective parameters. Indeed,

$$\begin{aligned}
\mathrm{E}\{b_1\} &= \mathrm{E}\left\{\sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}Y_i\right\} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}\mathrm{E}\{Y_i\}\\
&= \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}\mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}(\beta_0 + \beta_1 X_i + \mathrm{E}\{\varepsilon_i\})\\
&= \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}(\beta_0 + \beta_1 X_i) = \frac{\beta_0}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} + \frac{\beta_1}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})X_i}_{=S_{xx}(?)}\\
&= 0 + \beta_1 = \beta_1,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\{b_0\} &= \mathrm{E}\left\{\overline{Y} - b_1\overline{X}\right\} = \mathrm{E}\left\{\overline{Y}\right\} - \mathrm{E}\left\{b_1\overline{X}\right\} = \mathrm{E}\left\{\overline{Y}\right\} - \mathrm{E}\{b_1\}\,\overline{X}\\
&= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}Y_i\right\} - \beta_1\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{Y_i\} - \beta_1\overline{X}\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i\} - \beta_1\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \beta_1\overline{X}\\
&= \frac{\beta_0}{n}\sum_{i=1}^{n}1 + \frac{\beta_1}{n}\sum_{i=1}^{n}X_i - \beta_1\overline{X} = \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} = \beta_0.
\end{aligned}$$

Now is as good a time as any to illustrate these notions with an example.

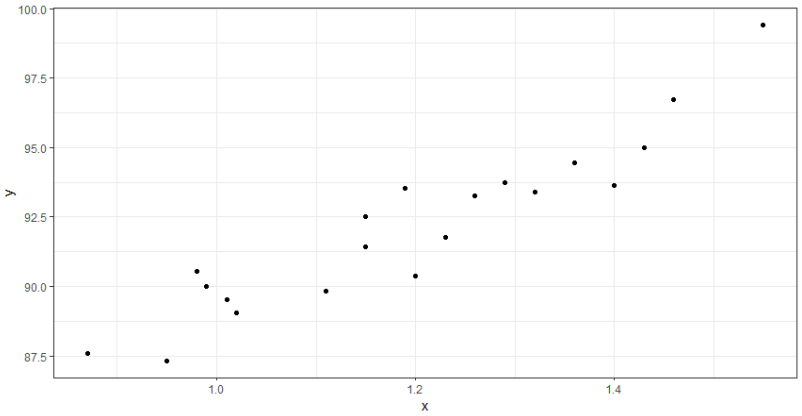
**Fuels Example** Consider the following $n = 20$ paired measurements $(X_i, Y_i)$ of hydrocarbon levels $(X)$ and pure oxygen levels $(Y)$ in fuels:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 0.99 | 1.02 | 1.15 | 1.29 | 1.46 | 1.36 | 0.87 | 1.23 | 1.55 | 1.40 |
| $Y_i$ | 90.01 | 89.05 | 91.43 | 93.74 | 96.73 | 94.45 | 87.59 | 91.77 | 99.42 | 93.65 |
| $i$ | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| $X_i$ | 1.19 | 1.15 | 0.98 | 1.01 | 1.11 | 1.20 | 1.26 | 1.32 | 1.43 | 0.95 |
| $Y_i$ | 93.54 | 92.52 | 90.56 | 89.54 | 89.85 | 90.39 | 93.25 | 93.41 | 94.98 | 87.33 |

Is the simple regression model valid? If so, fit the data to the model.

We start by loading and displaying the data.

```
x = c(0.99, 1.02, 1.15, 1.29, 1.46, 1.36, 0.87, 1.23, 1.55, 1.40,
      1.19, 1.15, 0.98, 1.01, 1.11, 1.20, 1.26, 1.32, 1.43, 0.95)
y = c(90.01, 89.05, 91.43, 93.74, 96.73, 94.45, 87.59, 91.77, 99.42, 93.65,
      93.54, 92.52, 90.56, 89.54, 89.85, 90.39, 93.25, 93.41, 94.98, 87.33)
plot(x,y)
```




Before we go on to compute the basic sums, we should verify visually if the SLR assumptions are met; they appear to be.

```
x.mean = mean(x)
y.mean = mean(y)
Sxy = sum((x-mean(x))*(y-mean(y)))
Sxx = sum((x-mean(x))^2)
Syy = sum((y-mean(y))^2)
```

```
[1] 1.196
[1] 92.1605
[1] 0.68088
[1] 10.17744
[1] 173.3769
```

We compute the least-square estimators:

```
(b1 = Sxy/Sxx)
(b0 = y.mean - b1*x.mean)
```

```
[1] 14.947
[1] 74.283
```

Thus the **regression line** for the data is

$$\hat{Y} = \hat{f}(X) = b_0 + b_1 X = 74.283 + 14.947X,$$

which is displayed in Figure 8.5 (left). Evaluating $\hat{f}$ at $X_i$ yields the $i$**th fitted value** $\hat{Y}_i = \hat{f}(X_i) = b_0 + b_1 X_i$.

**Residuals** The $i$**th regression residual** is $e_i = Y_i - \hat{Y}_i$; the residuals In the fuels dataset are displayed in Figure 8.5 (right).

**Properties of the Residuals**

1. $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$;
2. $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \overline{\hat{Y}}$;
3. $\sum_{i=1}^{n} X_i e_i = 0$;
4. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$;
5. the point $(\overline{X}, \overline{Y})$ lies on the regression line, and
6. $\sum_{i=1}^{n} e_i^2$ is minimal in the OLS sense.

**Proof:**

1. We see that

   $$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = \overline{Y} - b_0 - b_1\overline{X} = 0,$$

   according to the first normal equation.
2. From 1., we have $0 = \overline{e}$. Thus

   $$0 = \overline{e} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y} - \overline{\hat{Y}} \implies \overline{Y} = \overline{\hat{Y}}.$$

3. We see that

   $$\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} X_i Y_i - b_0\sum_{i=1}^{n} X_i - b_1\sum_{i=1}^{n} X_i^2 = 0,$$

   according to the second normal equation.
4. We see that

   $$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n}(b_0 + b_1 X_i)e_i = b_0\sum_{i=1}^{n} e_i + b_1\sum_{i=1}^{n} X_i e_i = 0,$$

   according to 1. and 3.

5. This is automatically true since

$$\hat{f}(\overline{X}) = b_0 + b_1\overline{X} = (\overline{Y} - b_1\overline{X}) + b_1\overline{X} = \overline{Y}.$$

6. For any $\mathbf{b}^* = (b_0^*, b_1^*) \neq \mathbf{b} = (b_0, b_1)$, we must have $Q(\mathbf{b}^*) \geq Q(\mathbf{b})$. Denote the residuals obtained from the line fitted with $\mathbf{b}^*$ by $e_i^*$. Then

$$\sum_{i=1}^{n} e_i^2 = \underbrace{\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2}_{=Q(\mathbf{b})} < \underbrace{\sum_{i=1}^{n}(Y_i - b_0^* - b_1^* X_i)^2}_{=Q(\mathbf{b}^*)} = \sum_{i=1}^{n}(e_i^*)^2.$$

This completes the proof. ■

**Descriptive Statistics and Correlations** The **Pearson sample correlation coefficient** $r$ of 2 variables $X$ and $Y$ is defined by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

This coefficient is such that

1. $-1 \leq r \leq 1$;
2. $|r| = 1 \iff Y_i = b_0 + b_1 X_i$, for all $i = 1, \ldots, n$, and
3. $\text{sgn}(r) = \text{sgn}(b_1)$, so that $r = 0 \iff b_1 = 0$.

If $|r| \approx 1$, then there is a **strong linear association** between $X$ and $Y$. If $|r| \approx 0$, there is **very little linear association** between $X$ and $Y$.[13] Note that we can **decompose** the total deviation as follows:

$$\underbrace{Y_i - \overline{Y}}_{\substack{\text{total deviation} \\ \text{from the mean}}} = \underbrace{(Y_i - \hat{Y}_i)}_{\substack{\text{unexplained deviation} \\ \text{from the mean}}} + \underbrace{(\hat{Y}_i - \overline{Y})}_{\substack{\text{deviation from the mean} \\ \text{explained by regression}}}.$$

This decomposition is shown graphically in Figure 8.6.

13: What can we say when $0 \ll |r| \ll 1$? We will discuss this at later stage.

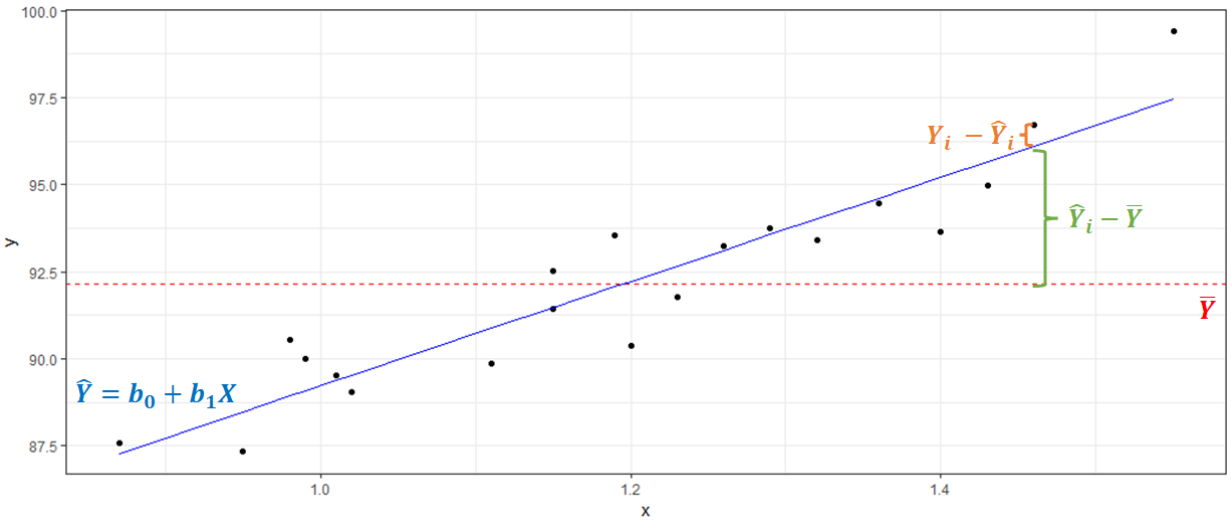



**Figure 8.6:** Illustration of the total deviation decomposition on the fuels dataset.

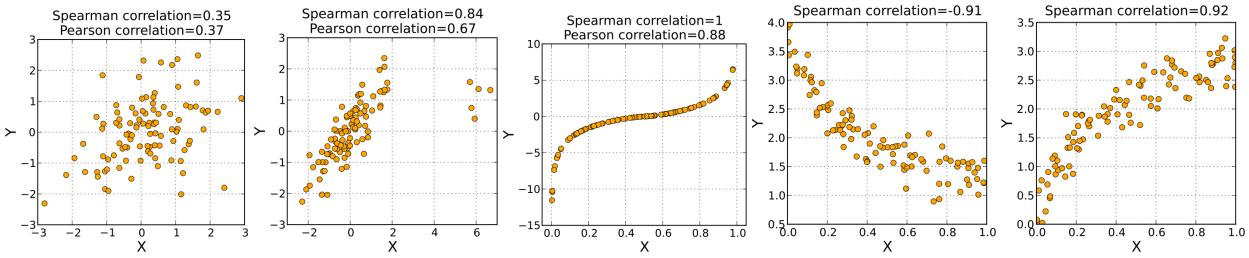



**Figure 8.7:** Illustration of various Spearman correlations (from Wikipedia).

The **Spearman sample correlation coefficient** $r_S$ of 2 variables $X$ and $Y$ is the **Pearson correlation** between the **rank values** $R(X_i)$ and $R(Y_i)$ of $X_i$ and $Y_i$, respectively. This coefficient is such that

1. $-1 \leq r_S \leq 1$;
2. $r_S = 1 \iff$ the relation between $X$ and $Y$ is **monotonic increasing**,
3. $r_S = -1 \iff$ the relation between $X$ and $Y$ is **monotonic decreasing**,
4. if the association between $X$ and $Y$ is **weak**, then $r_S \approx 0$, and
5. $r_S$ is invariant under **order-preserving (monotonic) transformations**.

The computational procedure is simple: for measurements

$$\mathfrak{Z} = \{Z_i \mid i = 1, \ldots, n\},$$

let $R(Z_i)$ be the **rank value** of $Z_i$ in $\mathfrak{Z}$; the smallest value of $Z_i$ has rank 1, the second smallest has rank 2, and so on, until the largest value, which has rank $n$. Ties are dealt with as in the example below:

| $Z_i$ | 0 | 1.5 | 1.5 | −1.5 | 3 | −2 |
|---|---|---|---|---|---|---|
| $R(Z_i)$ | 3 | 4.5 | 4.5 | 2 | 6 | 1 |

Formally, the Spearman correlation is given by

$$r_S = \frac{S_{R(x)R(y)}}{\sqrt{S_{R(x)R(x)} S_{R(y)R(y)}}}.$$

Some examples are shown in Figure 8.7.

**Sums of Squares Decomposition** The total deviation decomposition gives rise to one of the fundamental concepts of regression analysis: **sum of squares (SS) decompositions**.

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})\right]^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}\underbrace{(Y_i - \hat{Y}_i)}_{=e_i}(\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \\
&= \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}_{\text{SSR}} + 2\underbrace{\sum_{i=1}^{n}\hat{Y}_i e_i}_{=0} - 2\overline{Y}\underbrace{\sum_{i=1}^{n} e_i}_{=0}
\end{aligned}$$

This is often written as SST = SSE + SSR, where

- SST is the **total sum of squares**,
- SSE is the **error sum of squares**, and
- SSR is the **regression sum of squares**.

Note that we can write

$$\begin{aligned}\text{SSR} &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(b_0 + b_1X_i - \overline{Y})^2 = \sum_{i=1}^{n}(\overline{Y} - b_1\overline{X} + b_1X_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}\left(b_1(\overline{X} - X_i)\right)^2 = b_1^2\sum_{i=1}^{n}(\overline{X} - X_i)^2 = b_1^2S_{xx}.\end{aligned}$$

As SST = $S_{yy}$ and SSE = $Q(\mathbf{b})$, the decomposition can also be written:

$$S_{yy} = b_1^2S_{xx} + \sum_{i=1}^{n} e_i^2.$$

**Fuels Example** In the fuels dataset, we have

$$S_{xx} = 0.68, \quad S_{xy} = 10.18, \quad S_{yy} = 173.38,$$

so that the sample correlation coefficient is

$$r = \frac{10.18}{\sqrt{0.68}\sqrt{173.38}} \approx 0.94,$$

and the SS decomposition is SST(173.38) = SSR(152.13) + SSE(21.25). We can verify that this is indeed the case with R.

```
cor(x,y, method = "pearson")
cor(x,y, method = "spearman")
```

```
[1] 0.9367154
[1] 0.9236556
```

The values of $r$, $r_S$ are quite close to 1; is this a strong linear association?

**Coefficient of Determination** We can answer the previous question by looking at the quantity

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

also known as the **coefficient of determination**. It is the proportion of variation in the response which can be explained by the fitted line.

When $R^2 \approx 0$, the regression is **not very significant**, whereas when $R^2 \approx 1$, the variables are strongly linearly related.

**Proposition:** $R^2 = r^2$.

**Proof:** we have seen that SSR = $b_1^2S_{xx}$ and SST = $S_{yy}$. Thus

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{S_{xx}}\right)^2\frac{S_{xx}}{S_{yy}} = b_1^2 \cdot \frac{S_{xx}}{S_{yy}} = \frac{\text{SSR}}{\text{SST}} = R^2. \qquad \blacksquare$$

This answers the question relating to the interpretation of $0 \ll |r| \ll 1$: $r^2$ gives a sense of how much variation the regression "explains".

**Fuels Example** In the fuel dataset, we have

$$R^2 = \frac{152.13}{173.98} = 0.8774;$$

thus, about 87.74% of the variation observed in the data can be explained by the fitted line $\hat{Y} = 74.283 + 14.947X$.

This is a **reasonably high** proportion; together with the scatter plot, this suggests that the SRM is likely appropriate in this case. □

But don't get too deeply enamoured of $R^2$ as a figure to validate the regression: the values can be quite large even if the linear association is weak, as can be seen in Figure 8.8.

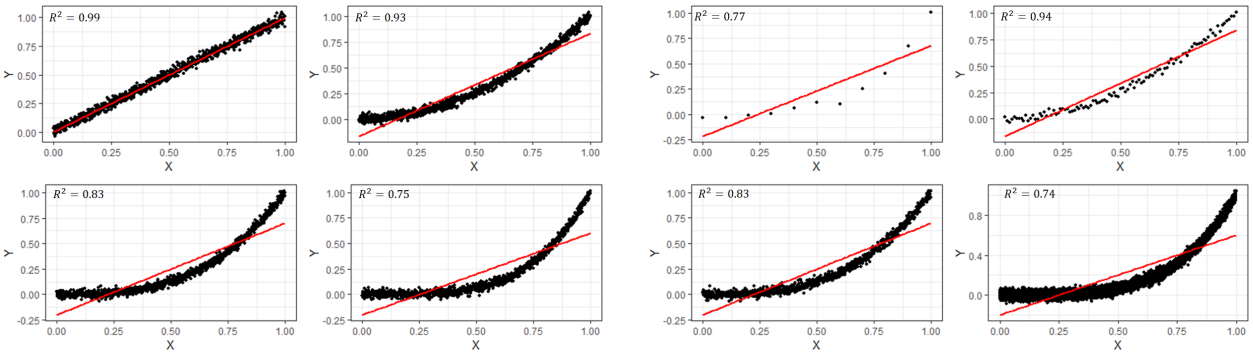



**Figure 8.8:** Various $R^2$ for nonlinear datasets; notice the effect of the number of observations on the coefficient of determination.

## 8.2.2 Inference

In order to test various hypotheses about the regression, we will need an estimation for the **common variance** $\sigma^2$. In the SLR model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

we have independent normal random errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. The probability function of $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ is thus

$$f(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right].$$

The **likelihood function** is

$$L(\beta_0, \beta_1; \sigma^2) = \prod_{i=1}^{n} f(Y_i) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{Q(\beta_0, \beta_1)}{2\sigma^2}\right],$$

where

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

The likelihood $L$ is maximized when $Q$ is minimized with respect to $\beta_0, \beta_1$.

We have already shown that the optimizer occurs at the **maximum likelihood estimator** $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) = (b_0, b_1)$, for which

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{SSE}.$$

Can we also use the data to find an estimator of $\sigma^2$?

Consider the **log-likelihood**

$$\begin{aligned} \ln L(b_0, b_1; \sigma^2) &= \ln \prod_{i=1}^{n} f(Y_i) = \sum_{i=1}^{n} \ln f(Y_i) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} Q(b_0, b_1) \end{aligned}$$

Because the logarithm is a **monotone increasing** function, maximizing $L$ is equivalent to maximizing $\ln L$. But

$$\frac{\partial L}{\partial [\sigma^2]} = -\frac{n}{2} \cdot \frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2} Q(b_0, b_1) = \frac{-1}{2\sigma^2} \left( n - \frac{Q(b_0, b_1)}{\sigma^2} \right).$$

Setting $\frac{\partial L}{\partial [\sigma^2]} = 0$ and solving for $\sigma^2$ yields

$$\widehat{\sigma^2} = \frac{1}{n} Q(b_0, b_1) = \frac{\text{SSE}}{n}.$$

14: It can be shown that $\text{E}\left\{\widehat{\sigma^2}\right\} = \frac{n-2}{n}\sigma^2$.

This estimator is **biased**, however.[14] The **mean squared error**

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

is another estimator of the population variance $\sigma^2$; this one is **unbiased** as

$$\text{E}\{\text{MSE}\} = \text{E}\left\{\frac{\text{SSE}}{n-2}\right\} = \text{E}\left\{\frac{n}{n-2} \cdot \frac{\text{SSE}}{n}\right\} = \frac{n}{n-2} \text{E}\left\{\widehat{\sigma^2}\right\} = \sigma^2.$$

We can think of the variance $\sigma^2$ of a **finite population** of size $n$ as a sum of squares divided by its degrees of freedom $n$:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2.$$

The estimator of the population variance using a **sample** of size $n$ is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2;$$

15: A degree of freedom is lost because we first used the sample to compute the **sample mean** $\overline{Y}$ as an approximation of $\mu$.

a sum of squares divided by its degrees of freedom $n-1$.[15]

Using the same data for two different purposes creates a "link" between $s^2$ and $\overline{Y}$ which did not exist between $\sigma^2$ and $\mu$. The same reasoning explains why it should not come as a surprise that we must divide SSE by $n-2$ to obtain an unbiased estimator of $\sigma^2$: in the error of sum of squares

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2,$$

we must first use the data to estimate 2 quantities, $\beta_0$ and $\beta_1$. Thus, SSE has $n-2$ degrees of freedom, and the unbiased estimator of $\sigma^2$ is

$$\text{MSE} = \frac{\text{SSE}}{n-2}.$$

**Fuels Example** In the fuels dataset with $n = 20$ observations, the **unbiased estimator** of the error variance $\sigma^2$ in the SLR model is computed as below.

```
n = length(x)
SSE = Syy - b1^2*Sxx
(MSE = SSE/(n-2))
```

```
[1] 1.180545
```

Thus $\widehat{\sigma^2} \approx 1.18$. □

In general, if the SLR model is valid we would expect

$$\text{E}\{Y_i\} = \beta_0 + \beta_1 X_i$$

to hold, more or less, for all samples. But the **specific values** for the OLS estimators $b_0, b_1$ depend on the **available data**; with different observations, we would obtain different values for the estimators, and it makes sense to study the **standard error** of $b_0, b_1$:

$$\sigma\{b_k\} = \sqrt{\text{E}\{(b_k - \beta_k)^2\}} = \sqrt{\text{E}\left\{b_k^2\right\} - \beta_k^2}, \quad \text{for } k = 0, 1.$$

**Regression Slope** In theory, we could then

1. collect $M$ independent datasets,
2. repeat the OLS procedure and obtain a slope estimate $b_{1;j}$ of $\beta_1$ for each dataset $j$, and
3. estimate $\sigma\{b_1\}$ by computing the sample standard deviation of $\{b_{1;1}, \ldots, b_{1;M}\}$.

In practice, however, collecting data is often **costly** and we may never have access to more than one set of observations.[16]

16: The use of **resampling methods** (such as the bootstrap or the jackknife, see Chapter 20) is another option, but in the case of OLS estimation, we can use the underlying machinery to obtain standard error estimates from a **single sample**.

As the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent in the SLR model, the response values $Y_1, \ldots, Y_n$ are uncorrelated, with variance $\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$ for $i = 1, \ldots, n$. Since

$$b_1 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{xx}} Y_i, \quad \text{we have } \sigma^2\{b_1\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{xx}}\right)^2 \sigma^2\{Y_i\},$$

so that

$$\sigma^2\{b_1\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{xx}}\right)^2 \sigma^2\{\varepsilon_i\} = \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{\sigma^2}{S_{xx}^2} \cdot S_{xx} = \frac{\sigma^2}{S_{xx}}.$$

Since we do not usually know the actual value of $\sigma^2$, the **estimated standard error of** $b_1$ is:

$$\mathrm{s}\{b_1\} = \sqrt{\frac{\mathrm{MSE}}{S_{xx}}}.$$

**Fuels Example** In the fuels dataset, we have:

```
(s.b1 = sqrt(MSE/Sxx))
```

```
[1] 1.316758
```

and so $\mathrm{s}\{b_1\} \approx 1.317$. □

As $b_1$ is a linear combination of the **independent normal** random variables $\{Y_i\}_{i=1}^n$, it is itself **normal**, by the central limit theorem.[17]

17: See page 416.

Since we already know its expectation and its variance, we know its distribution:

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies \frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1).$$

We now make assumptions that will be justified at a later stage:

$$\frac{\mathrm{SST}}{\sigma^2} \sim \chi^2(n-1), \quad \frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2), \quad \frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(1), \quad b_1, \mathrm{SSE\ indep.}$$

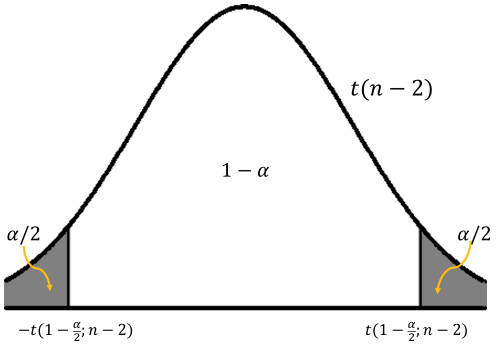
The definition of the Student $t-$distribution (see Section 8.1.1) yields

$$T_1 = \underbrace{\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-2)}_{\nu}} = \frac{b_1 - \beta_1}{\sqrt{\mathrm{MSE}}/\sqrt{S_{xx}}} = \frac{b_1 - \beta_1}{\mathrm{s}\{b_1\}} \sim t(n-2).$$

**Critical Region** Let $\alpha \in (0,1)$. Since $\frac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \sim t(n-2)$, we have

$$\begin{aligned} 1-\alpha &= P\left(-t(1-\tfrac{\alpha}{2}; n-2) \le \tfrac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \le t(1-\tfrac{\alpha}{2}; n-2)\right) \\ &= P\left(b_1 - t(1-\tfrac{\alpha}{2}; n-2)\cdot \mathrm{s}\{b_1\} \le \beta_1 \le b_1 + t(1-\tfrac{\alpha}{2}; n-2)\cdot \mathrm{s}\{b_1\}\right), \end{aligned}$$

as in the image below.



Thus, the $100(1-\alpha)\%$ **confidence interval for** $\beta_1$ is

$$\text{C.I.}(\beta_1; 1-\alpha) \equiv b_1 \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \text{s}\{b_1\}.$$

**Fuels Example** In the fuels dataset, we have

$$b_1 = 14.947, \quad \text{s}\{b_1\} = 1.317.$$

At a **confidence level** of $1-\alpha = 0.95$,[18] the critical value of the Student $t-$distribution with $n-2 = 20-2 = 18$ degrees of freedom is

$$t(1-0.05/2; 20-2) = t(0.975; 18) = 2.101.$$

We can build a 95% confidence interval for $\beta_1$ as follows:

$$\text{C.I.}(\beta_1; 0.95) \equiv 14.947 \pm 2.101(1.317) = [12.17, 17.72].$$

18: Or an **error rate** of $\alpha = 0.05$.

**Regression Intercept** With the same assumptions as with $b_1$, we also have:

$$\begin{aligned}
\sigma^2\{b_0\} &= \sigma^2\left\{\overline{Y} - b_1\overline{X}\right\} = \sigma^2\left\{\frac{1}{n}\sum_{i=1}^n Y_i - \overline{X}\sum_{i=1}^n \frac{(X_i-\overline{X})}{S_{xx}}Y_i\right\} \\
&= \sigma^2\left\{\sum_{i=1}^n\left[\frac{1}{n} - \frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]Y_i\right\} = \sum_{i=1}^n\left[\frac{1}{n} - \frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]^2 \underbrace{\sigma^2\{Y_i\}}_{=\sigma^2} \\
&= \sigma^2\Big[\sum_{i=1}^n \frac{1}{n^2} - \frac{2\overline{X}}{nS_{xx}}\underbrace{\sum_{i=1}^n(X_i-\overline{X})}_{=0} + \frac{\overline{X}^2}{S_{xx}^2}\underbrace{\sum_{i=1}^n(X_i-\overline{X})^2}_{=S_{xx}}\Big].
\end{aligned}$$

Thus,

$$\sigma^2\{b_0\} = \left[\frac{n}{n^2} - 0 + \frac{\overline{X}^2}{S_{xx}^2}S_{XX}\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right],$$

and so the estimated standard error of $b_0$ is:

$$\text{s}\{b_0\} = \sqrt{\text{MSE}}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}.$$

**Fuels Example** In the fuels dataset, we have

$$\text{s}\{b_0\} = \sqrt{1.18}\sqrt{\frac{1}{20} + \frac{(23.92/20)^2}{0.68}} = 1.593. \qquad \square$$

As was the case for $b_1$, $b_0$ follows a normal distribution since it is a linear combination of the **independent normal** random variables $Y_1, \ldots, Y_n$.

As we already know its expectation and its variance, we also know its distribution:

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_xx}\right]\right) \implies \frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_xx}}} \sim \mathcal{N}(0,1).$$

Assuming again that $b_0$ and SSE are independent and that $\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2)$, the definition of the Student $t-$distribution yields that

$$T_0 = \underbrace{\frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-2)}_{\nu}} = \frac{b_0 - \beta_0}{\sqrt{\text{MSE}}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}} = \frac{b_0 - \beta_0}{\text{s}\{b_0\}}$$

follows a $t(n-2)$ distribution.

As is the case with $\beta_1$, the $100(1-\alpha)\%$ **confidence interval for** $\beta_0$ is

$$\text{C.I.}(\beta_0; 1-\alpha) \equiv b_0 \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \text{s}\{b_0\}\,.$$

**Fuels Example** In the fuels dataset, we have $b_0 = 74.283$ and $\text{s}\{b_0\} = 1.593$. At a **confidence level** of $1-\alpha = 0.95$, the critical value of the Student $t-$distribution with $n-2 = 18$ degrees of freedom is $t(0.975; 18) = 2.101$, and we can build a 95% confidence interval for $\beta_0$ as follows:

$$\text{C.I.}(\beta_0; 0.95) \equiv 74.283 \pm 2.101(1.593) = [70.94, 77.63].$$

**Hypothesis Testing** With standard errors, we can **test hypotheses** on the regression parameters.

We try to determine if the true parameters $\beta_0, \beta_1$ take on specific values and whether the line of best fit provides a good description of a bivariate dataset using the following steps:

1. set up a **null** hypothesis $H_0$ and an **alternative** hypothesis $H_1$;
2. compute a **test statistic** (using the studentization);
3. find a **critical region**/$p-$value for the test statistic under $H_0$;
4. **reject** or **fail to reject** $H_0$ based on the critical region/$p-$value.

For instance, we might be interested in testing whether a true parameter value $\beta$ is equal to some **candidate value** $\beta^*$, i.e.

$$H_0 : \beta = \beta^* \text{ against } H_1 : \begin{cases} \beta < \beta^*, & \text{left-tailed test} \\ \beta > \beta^*, & \text{right-tailed test} \\ \beta \neq \beta^*, & \text{two-tailed test} \end{cases}$$

Under $H_0$, we have shown that

$$T_0 = \frac{b - \beta^*}{\text{s}\{b\}} \sim t(n-2).$$

The **critical region** depends on the confidence level $1-\alpha$ and on the **type** of the alternative hypothesis $H_1$.

Let $t^*$ be the observed value of $T_0$; **we reject** $H_0$ **at** $\alpha$ **if** $t^*$ **is in the critical region of the test**.

| Alternative Hypothesis | Rejection Region |
|---|---|
| $H_1 : \beta < \beta^*$ | $t^* < -t(1-\alpha; n-2)$ |
| $H_1 : \beta > \beta^*$ | $t^* > t(1-\alpha; n-2)$ |
| $H_1 : \beta \neq \beta^*$ | $\lvert t^* \rvert > t(1-\alpha/2; n-2)$ |

**Examples** Test the following hypotheses In the fuels dataset.

a) Test for $H_0 : \beta_0 = 75$ against $H_1 : \beta_0 < 75$ at $\alpha = 0.05$.
b) Test for $H_0 : \beta_1 = 10$ against $H_1 : \beta_1 > 10$ at $\alpha = 0.05$.
c) Test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.

We have seen that

$$b_0 = 74.283, \quad \mathrm{s}\{b_0\} = 1.593, \quad b_1 = 14.947, \quad \mathrm{s}\{b_1\} = 1.317.$$

Since the error rate for all tests is $\alpha = 0.05$, we also need to compute the critical values of the Student $t-$distribution with $\nu = 20 - 2 = 18$ degrees of freedom, at confidence levels $1 - \alpha = 0.95$ and $1 - \alpha/2 = 0.975$:

$$t(0.975; 18) = 2.101, \quad \text{and} \quad t(0.95; 18) = 1.734.$$

a) We run a **left-tailed** test for the intercept: the observed test statistic is

$$t_a^* = \frac{b_0 - \beta_0^*}{\mathrm{s}\{b_0\}} = \frac{74.283 - 75}{1.593} = -0.449 \not< -1.734 = -t(0.95; 18),$$

and so we **fail to reject** $H_0$ **at** $\alpha = 0.05$.

b) We run a **right-tailed** test for the slope: the observed test statistic is

$$t_b^* = \frac{b_1 - \beta_1^*}{\mathrm{s}\{b_1\}} = \frac{14.947 - 10}{1.317} = 3.757 > 1.734 = t(0.95; 18),$$

and so we **reject** $H_0$ **in favour of** $H_1$ **at** $\alpha = 0.05$.

c) We run a **two-tailed** test for the slope: the observed test statistic is

$$|t_c^*| = \left|\frac{b_1 - \beta_1^*}{\mathrm{s}\{b_1\}}\right| = \left|\frac{14.947 - 0}{1.317}\right| = 11.351 > 2.101 = t(0.975; 18),$$

and so we **reject** $H_0$ **in favour of** $H_1$ **at** $\alpha = 0.05$.

We will see another test for the slope in Section 8.2.4.

**Mean Response** We can also conduct inferential analysis for the **expected response** at $X = X^*$.[19] We assume that $\mathrm{E}\{Y^*\} = \beta_0 + \beta_1 X^*$.

19: In practice, there could be replicates, say.

The **estimated mean response** at $X = X^*$ is

$$\hat{Y}^* = b_0 + b_1 X^*.$$

The predictor value being **fixed**, $\hat{Y}^*$ is normally distributed with

$$\mathrm{E}\left\{\hat{Y}^*\right\} = \mathrm{E}\{b_0 + b_1 X^*\} = \mathrm{E}\{b_0\} + \mathrm{E}\{b_1\} X^* = \beta_0 + \beta_1 X^*,$$

so that $\hat{Y}^*$ is an **unbiased estimator** of $Y^*$. What is its standard error?

If $b_0, b_1$ were independent, we could simply compute

$$\sigma^2\left\{\hat{Y}^*\right\} = \sigma^2\{b_0\} + (X^*)^2\sigma^2\{b_1\}.$$

But they are **not independent**, as we can see in the following result.

**Theorem:** under the SLR assumptions, $\sigma\left\{\overline{Y}, b_1\right\} = 0$ and

$$\sigma\left\{b_0, b_1\right\} = -\overline{X}\sigma^2\left\{b_1\right\}.$$

**Proof:** throughout, keep in mind that the $Y_i$ are **uncorrelated**. We have

$$\sigma\left\{\overline{Y}, b_1\right\} = \sigma\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n}\frac{(X_j-\overline{X})}{S_{xx}}Y_j\right\} = \sum_{i,j=1}^{n}\frac{1}{n}\cdot\frac{(X_i-\overline{X})}{S_{xx}}\sigma\left\{Y_i, Y_j\right\}.$$

All the terms for which $i \neq j$ have $\sigma\left\{Y_i, Y_j\right\} = 0$, the other ones have $\sigma\left\{Y_i, Y_i\right\} = \sigma^2\left\{Y_i\right\} = \sigma^2$, so

$$\sigma\left\{\overline{Y}, b_1\right\} = \frac{\sigma^2}{nS_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0} = 0.$$

Similarly,

$$\begin{aligned}\sigma\left\{b_0, b_1\right\} &= \sigma\left\{\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]Y_i, \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}Y_i\right\}\\ &= \sum_{i,j=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]\frac{(X_j-\overline{X})}{S_{xx}}\sigma\left\{Y_i, Y_j\right\}\end{aligned}$$

All the terms for which $i \neq j$ have $\sigma\left\{Y_i, Y_j\right\} = 0$, the other ones have $\sigma\left\{Y_i, Y_i\right\} = \sigma^2\left\{Y_i\right\} = \sigma^2$, so

$$\begin{aligned}\sigma\left\{b_0, b_1\right\} &= \sigma^2\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]\frac{(X_i-\overline{X})}{S_{xx}}\\ &= \frac{\sigma^2}{nS_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0} - \frac{\sigma^2\overline{X}}{S_{xx}^2}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})^2}_{S_{xx}}\\ &= -\overline{X}\frac{\sigma^2}{S_{xx}} = -\overline{X}\sigma^2\left\{b_1\right\}.\end{aligned}$$

This completes the proof. ■

We can now determine the standard error of the estimated mean response $Y = \hat{Y}^*$ at $X = X^*$:

$$\begin{aligned}\sigma^2\left\{\hat{Y}^*\right\} &= \sigma^2\left\{b_0 + b_1X^*\right\} = \sigma^2\left\{b_0\right\} + (X^*)^2\sigma^2\left\{b_1\right\} + 2\sigma\left\{b_0, X^*b_1\right\}\\ &= \sigma^2\left[\frac{1}{n}+\frac{\overline{X}^2}{S_{xx}}\right] + \frac{(X^*)^2\sigma^2}{S_{xx}} - 2X^*\overline{X}\frac{\sigma^2}{S_{xx}}\\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\left[(X^*)^2 - 2\overline{X}X^* + \overline{X}^2\right] = \sigma^2\left[\frac{1}{n}+\frac{(X^*-\overline{X})^2}{S_{xx}}\right].\end{aligned}$$

The estimated standard error is thus

$$\mathrm{s}\left\{\hat{Y}^*\right\} = \sqrt{\mathrm{MSE}}\sqrt{\frac{1}{n}+\frac{(X^*-\overline{X})^2}{S_{xx}}}.$$

But there are many ways to skin a cat:

$$\begin{aligned}\sigma^2\left\{\hat{Y}^*\right\} &= \sigma^2\left\{(\overline{Y} - b_1\overline{X}) + b_1X^*\right\} = \sigma^2\left\{\overline{Y} + b_1(X^* - \overline{X})\right\} \\ &= \sigma^2\left\{\overline{Y}\right\} + \sigma^2\left\{b_1(X^* - \overline{X})\right\} + 2(X^* - \overline{X})\sigma\left\{\overline{Y}, b_1\right\} \\ &= \frac{\sigma^2}{n} + (X^* - \overline{X})^2\frac{\sigma^2}{S_{xx}} + 0 = \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right].\end{aligned}$$

Either way, we can show that

$$T^* = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\mathrm{s}\{\hat{Y}^*\}} \sim t(n-2), \quad \text{and so}$$
$$\text{C.I.}(\mathrm{E}\{Y^*\}; 1-\alpha) \equiv \beta_0 + \beta_1X^* \pm t(1-\tfrac{\alpha}{2}; n-2)\cdot \mathrm{s}\{\hat{Y}^*\}.$$

**Fuels Example** In the fuels dataset, the 95% C.I. for $\mathrm{E}\{Y^*\}$ is

$$\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) \equiv 74.28 + 14.95X^* \pm 2.10\sqrt{1.18\left[\frac{1}{20} + \frac{(X^* - 1.12)^2}{0.68}\right]}$$

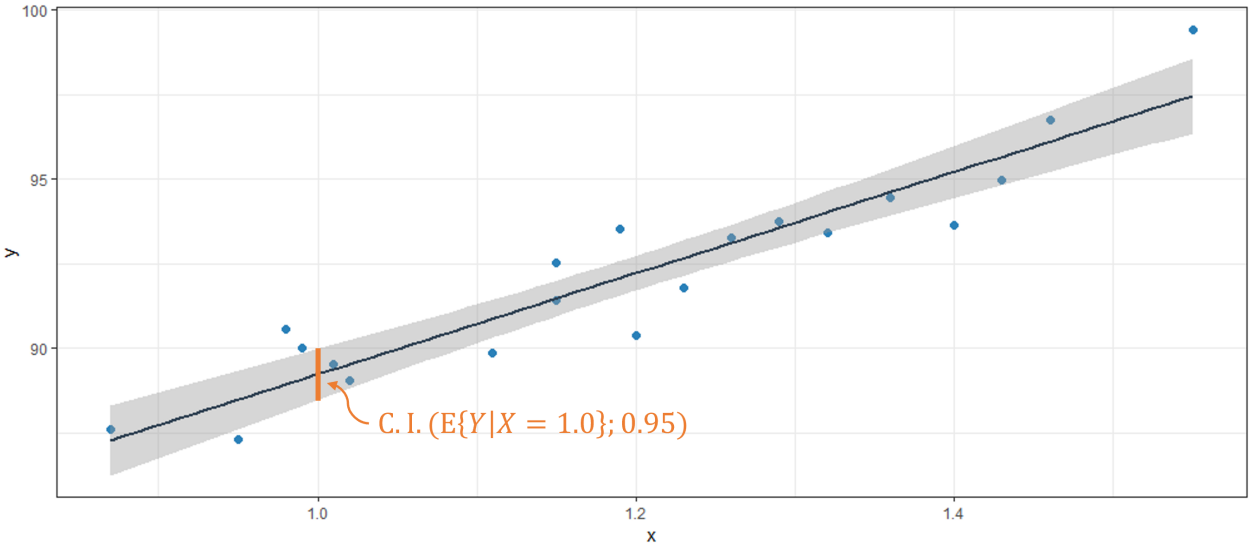



**Figure 8.9:** Confidence interval for the mean response: at $X^* = 1$, the 95% confidence interval for the mean response $\mathrm{E}\{Y^*\}$ is the orange bar.

### 8.2.3 Estimation and Prediction

When we estimate the **expected** (mean) response $\mathrm{E}\{Y^*\}$, we are determining how $(b_0, b_1)$ could **jointly** vary from one sample to the next. As these parameters uniquely determine the line of best fit, finding a confidence interval for the mean response at all $X = X^*$ is equivalent to finding a **confidence band** for the entire line over the predictor domain.[20]

It should come as no surprise that a number of observations fell outside of their respective confidence intervals for the fuels dataset example: we were estimating the **mean response** at a predictor level $X = X^*$, not the **actual** (or new) **responses** at that level.

20: **Warning:** see a bit further down for **joint estimation**.

But what if we wanted to find a range of **likely response values** at $X = X^*$? We use the available data to build **confidence intervals** (C.I.) when we are interested in certain (fixed) population characteristics (parameters) that are unknown to us.

But a new value of the response is not a parameter – it is a **random variable**. We refer to the interval of plausible (likely) values for a new response as a **prediction interval** (P.I.).

In order to determine such a P.I. for the response, we must model the **error** involved in the prediction of the response.[21]

21: Throughout, we assume that the new responses for a predictor level $X = X^*$ are independent of the observed responses, which is to say that the **residuals are uncorrelated**.

**Prediction Intervals** Let $Y_p^*$ represent a **(new) response** at $X = X^*$:

$$Y_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p \quad \text{for some } \varepsilon_p.$$

If the average error is 0, the best prediction for $Y_p^*$ is still the **response on the fitted line at** $X = X^*$:

$$\hat{Y}_p^* = b_0 + b_1 X^*.$$

The **prediction error** at $X = X^*$ is thus

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p - b_0 - b_1 X^*.$$

In the SLR model, the error $\varepsilon_p$ and the estimators $b_0, b_1$ are **normally distributed**. Consequently, so is the prediction error $\text{pred}^*$. We have

$$\text{E}\{\text{pred}^*\} = \underbrace{\text{E}\{\beta_0 + \beta_1 X^* + \varepsilon_p^*\}}_{=\beta_0+\beta_1 X^*} - \underbrace{\text{E}\{b_0 + b_1 X^*\}}_{=\beta_0+\beta_1 X^*} = 0.$$

22: They are not uncorrelated with one another because $\bar{e} = 0$.

Because the residuals are uncorrelated with the responses,[22] we have

$$\begin{aligned}\sigma^2\{\text{pred}^*\} &= \sigma^2\left\{Y_p^*\right\} + \sigma^2\left\{\hat{Y}_p^*\right\} \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right]\end{aligned}$$

Thus

$$\text{pred}^* \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right]\right).$$

The estimated standard error is thus

$$\text{s}\{\text{pred}^*\} = \sqrt{\text{MSE}}\sqrt{1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}}.$$

As before, we can show that

$$T_p^* = \frac{\text{pred}^* - 0}{\text{s}\{\text{pred}^*\}} \sim t(n-2), \quad \text{and so}$$

$$\text{P.I.}(Y_p^*; 1-\alpha) \equiv \beta_0 + \beta_1 X^* \pm t(1 - \tfrac{\alpha}{2}; n-2) \cdot \text{s}\{\text{pred}^*\}.$$

23: Furthermore, these regions are smallest when $X^* = \overline{X}$, and they increase as $|X^* - \overline{X}|$ increases.

Note that $\text{s}\{\hat{Y}^*\} < \text{s}\{\text{pred}^*\}$ so that the C.I. for the mean response at $X^*$ is **contained** in the P.I. for a new response at $X^*$.[23]

**Fuels Example** In the fuels dataset, the 95% P.I. for $Y_p^*$ is

$$\text{P.I.}(Y_p^*; 0.95) \equiv 74.28 + 14.95X^* \pm 2.10\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(X^* - 1.12)^2}{0.68}\right]}.$$

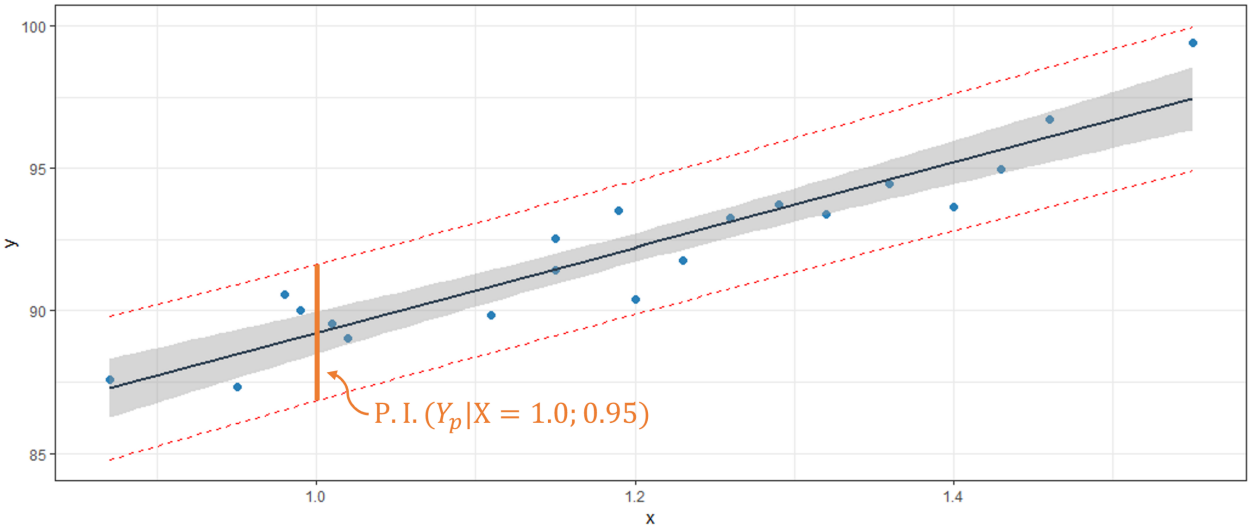



**Figure 8.10:** Prediction interval for a new response: at $X^* = 1$, the 95% prediction interval for a new response $Y_p^*$ is the orange bar.

**Hypothesis Testing** Since the distributions for the estimators of the mean response and for new responses are normal and since we have estimates for their standard errors, we can conduct hypothesis testing as before:

1. identify the **type** of alternative hypothesis $H_1$ (left-tailed, right-tailed, two-tailed),
2. compute the (studentized) **observed test statistic**, and
3. compare to the appropriate **critical value** of the Student $t-$distribution.

**Fuels Example** In the fuels dataset, suppose we would like to test

$$H_0 : \text{E}\{Y^* \mid X^* = 1.2\} = 92.5 \quad \text{against} \quad H_1 : \text{E}\{Y^* \mid X^* = 1.2\} \neq 92.5.$$

Under $H_0$, the test statistic

$$T^* = \frac{\hat{Y}^* - 92.5}{\text{s}\{\hat{Y}^*\}} \sim t(n-2) = t(18).$$

But $\hat{Y}^* = 74.28 + 14.95(1.2) = 92.22$ and

$$\text{s}\{\hat{Y}^*\} = \sqrt{1.18}\sqrt{\frac{1}{20} + \frac{(1.2 - 1.12)^2}{0.68}} = 0.265.$$

The observed value of $T^*$ is thus

$$t^* = \frac{92.22 - 92.5}{0.265} = -1.057.$$

At an error rate of $\alpha = 0.05$, the critical value of the Student $t-$distribution with $n - 2 = 18$ degrees of freedom is $t(0.975; 18) = 2.101$; since $|t^*| \not> t(0.975; 18)$, there is not enough evidence to reject the null hypothesis $H_0$ at a confidence level of 95%.[24]

24: Which is not the same as accepting the null hypothesis $H_0$.

What if we observed a new response $Y_p^* = 80$ for a predictor level $X^* = 1.2$? Is this a reasonable value or should we expect something larger?

At a confidence level of 95%, the prediction interval for the response at the predictor level $X^* = 1.2$ is

$$\begin{aligned}\text{P.I.}(Y_p^*; 0.95) &\equiv \hat{Y}^* \pm t(0.975; 18) \cdot \text{s}\{\text{pred}^*\} \\ &= 74.28 + 14.95(1.2) \pm 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.2 - 1.12)^2}{0.68}\right]} \\ &= 92.22 \pm 2.101(1.061) = [89.99, 94.45].\end{aligned}$$

As $Y_p^* = 80$ is not in the prediction interval, this seems like an unlikely new response for $X^* = 1.2$ (at confidence level 95%).

**Joint Estimations and Predictions** When we use a dataset to estimate the two parameters $\beta_0$ and $\beta_1$ in the SLR model, the **error sum of squares** SSE has $n - 2$ degrees of freedom.

This might seem like an obscure technical point, but there is a practical consequence: the resulting C.I. are necessarily **wider** than those that would be obtained if the sum of squares had more degrees of freedom. For instance, $t(0.975; 18) = 2.101 > t(0.975, 20) = 2.086$.[25]

25: What does this mean for regression analysis? One interpretation is that there is a **penalty** for the simultaneous estimation of parameters: when the same data is used to compute various estimates, it gets **"tired"** (?) and it loses some of its predictive power.

**Bonferroni's Procedure** Say we are interested in the **joint** estimation of $g$ parameters $\theta_1, \ldots, \theta_g$.

For each parameter $\theta_i$, we build $\text{C.I.}(\theta_i) \equiv A_i = \{L_i \leq \theta_i \leq U_i\}$; the **error rate for estimating** $\theta_i$ is $P(\overline{A_i}) = P(\theta_i \notin A_i)$. The **family confidence level** is

$$P(A_1 \cap \cdots \cap A_g) = P(\theta_1 \in A_1, \cdots, \theta_g \in A_g).$$

**Theorem:** for individual error rates $P(\overline{A_i}) = \frac{\alpha}{g}$, we have

$$P(A_1 \cap \cdots \cap A_g) \geq 1 - \alpha.$$

**Proof:** recall that $P(C \cup D) = P(C) + P(D) - P(C \cap D)$. As all probabilities are non-negative, $P(C) + P(D) \geq P(C \cup D)$. This can be extended to unions of $g$ events:

$$P(\overline{A_1} \cup \cdots \cup \overline{A_g}) \leq P(\overline{A_1}) + \cdots + P(\overline{A_g}); \quad \text{or}$$

$$1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g}) \geq 1 - P(\overline{A_1}) - \cdots - P(\overline{A_g}) = 1 - g \cdot \frac{\alpha}{g} = 1 - \alpha.$$

As $P(A_1 \cap \cdots \cap A_g) = 1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g})$, this completes the proof. ■

We can use the **Bonferroni procedure** to provide **joint C.I.** for parameters $\theta_1, \ldots, \theta_g$ at a **family confidence level** of $1-\alpha$:

$$\text{C.I.}_{\text{B}}(\theta_i; 1-\alpha) \equiv \hat{\theta}_i \pm t(1-\tfrac{\alpha/g}{2}; \text{d.f.}) \cdot \text{s}\{\hat{\theta}_i\}, \quad i = 1, \ldots, g.$$

**Joint Estimation of $\beta_0$ and $\beta_1$** At a family confidence level of $1-\alpha$, the joint **Bonferroni** C.I. for $\beta_0$ and $\beta_1$ ($g = 2$) take the form:

$$\text{C.I.}_{\text{B}}(\beta_i; 1-\alpha) \equiv b_i \pm t(1-\tfrac{\alpha}{4}; n-2) \cdot \text{s}\{b_i\}, \quad i = 0, \ldots, 1.$$

At least $100(1-\alpha)\%$ of the times we use this procedure, both $\beta_0$ and $\beta_1$ will fall inside their respective C.I..

**Fuels Example** In the fuels dataset, if we want a family confidence level of $1-\alpha = 0.95$, we need to use $t(1-\frac{0.05}{4}; 20-2) = t(0.9875; 18) = 2.44501$ :

$$\text{C.I.}_{\text{B}}(\boldsymbol{\beta}; 0.95) \equiv \begin{cases} 74.283 \pm 2.445 \cdot 1.593 \equiv [70.39, 78.18] & (\beta_0) \\ 14.947 \pm 2.445 \cdot 1.317 \equiv [11.73, 18.17] & (\beta_1) \end{cases}$$

**Working-Hotelling's Procedure** When we estimate a C.I. for the mean response at $X = X^*$, we express the lower bound and the upper bound of the interval as a function of $X^*$.[26]

If we are only interested in jointly estimating the mean response at a "small" number of levels $X = X_i^*$, $i = 1, \ldots, g$, with a family confidence level $1-\alpha$, we can use the **Bonferroni** procedure:

$$\text{C.I.}_{\text{B}}(\text{E}\left\{Y_i^*\right\}; 1-\alpha) = \hat{Y}_i^* \pm t(1-\tfrac{\alpha/g}{2}; n-2) \cdot \text{s}\{\hat{Y}_i^*\}, \quad i = 1, \ldots, g.$$

If we want to build a $100(1-\alpha)\%$ confidence region for $\text{E}\{Y\} = \beta_0 + \beta_1 X$, the Bonferroni approach would require us to let $g \to \infty$ in the C.I. computations, which is problematic as

$$t(1-\tfrac{\alpha/g}{2}; n-2) \to \infty$$

in that case. Instead, we seek $W > 0$ such that

$$1-\alpha = P\Big(\hat{Y}(X) - W \cdot \text{s}\{\hat{Y}(X)\} \leq \underbrace{\beta_0 + \beta_1 X}_{=\text{E}\{\hat{Y}(X)\}} \leq \hat{Y}(X) + W \cdot \text{s}\{\hat{Y}(X)\}\Big)$$

for all $X$ in the regression domain. This can be achieved if

$$1-\alpha = P\left(\max_X \left\{\left|\frac{\hat{Y}(X) - \text{E}\{\hat{Y}(X)\}}{\text{s}\{\hat{Y}(X)\}}\right|\right\} \leq W\right),$$

or equivalently, if

$$1-\alpha = P\left(\max_X \left\{\frac{(\hat{Y}(X) - \text{E}\{\hat{Y}(X)\})^2}{\text{s}^2\{\hat{Y}(X)\}}\right\} \leq W^2\right).$$

26: It would be tempting to see the union of all these C.I. as a **confidence band** for the mean response at all $X$, i.e., for the **true line of best fit**

$$\text{E}\{Y\} = \beta_0 + \beta_1 X,$$

but that's not how it works.

In order to find the appropriate $W$, we need the distribution of

$$\mathscr{M} = \max_X \left\{ \frac{(\hat{Y}(X) - \mathrm{E}\{\hat{Y}(X)\})^2}{\mathrm{s}^2\{\hat{Y}(X)\}} \right\} = \max_X \left\{ \frac{\left[(b_0 + b_1 X) - (\beta_0 + \beta_1 X)\right]^2}{\mathrm{MSE}\left[\frac{1}{n} + \frac{(X-\overline{X})^2}{S_{xx}}\right]} \right\}.$$

Set $t = X - \overline{X}$; then the quantity can be re-written as:

$$\max_t \left\{ \frac{\left[\overline{Y} - \mathrm{E}\left\{\overline{Y}\right\} + (b_1 - \beta_1)t\right]^2}{\mathrm{MSE}\left[\frac{1}{n} + \frac{t^2}{S_{xx}}\right]} \right\} = \max_t \left\{ \frac{[c_1 + d_1 t]^2}{c_2 + d_2 t^2} \right\} = \max_t \{h(t)\}.$$

Note that $c_2, d_2 > 0$ as $\mathrm{MSE}, S_{xx} > 0$, so $h(t) \geq 0$ for all $t$. This is a continuous rational function of a single variable, with a horizontal asymptote at $h = d_1^2/d_2 \geq 0$; its first derivative is

$$h'(t) = \frac{2\,(c_1 + d_1 t)\,(c_2 d_1 - c_1 d_2 t)}{(c_1 + d_2 t^2)^2}.$$

The critical points are found at $t_1 = -\frac{c_1}{d_1}$ and $t_2 = \frac{c_2 d_1}{c_1 d_2}$. Since

$$h(t_1) = 0 \quad \text{and} \quad h(t_2) = \frac{c_1^2 d_2 + c_2 d_1^2}{c_2 d_2} = \frac{c_1^2}{c_2} + \frac{d_1^2}{d_2} \geq 0,$$

we must have

$$\max_t \{h(t)\} = \frac{c_1^2}{c_2} + \frac{d_1^2}{d_2}.$$

Thus

$$\mathscr{M} = \frac{(\overline{Y} - \mathrm{E}\{\overline{Y}\})^2}{\mathrm{MSE}/n} + \frac{(b_1 - \beta_1)^2}{\mathrm{MSE}/S_{xx}} = \frac{\left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2}{\mathrm{MSE}/\sigma^2}$$

Both of the r.v. in the numerator of $\mathscr{M}$ are independent; we then have

$$\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}, \frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1) \implies \left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2, \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2 \sim \chi^2(1).$$

We can re-write the random variable in the denominator of $\mathscr{M}$ as

$$\mathrm{MSE}/\sigma^2 = \frac{\mathrm{SSE}}{\sigma^2} \Big/ n - 2\,,$$

so that

$$\mathscr{M} = \frac{2 \overbrace{\left[\left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2 + \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2\right]}^{\sim\chi^2(2)} \Big/ 2}{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{\sim\chi^2(n-2)} \Big/ n - 2} \sim 2F(2, n-2).$$

We thus have

$$1-\alpha = P(\mathcal{M} \le W^2) \iff W^2 = 2F(1-\alpha;2,n-2).$$

**Joint Estimation of Mean Responses** At a family confidence level of $1-\alpha$, the joint **Working-Hotelling** C.I. for $\mathrm{E}\left\{Y_i^*\right\}$ at any number of levels $X = X_i^*$ take the form:

$$\text{C.I.}_{\text{WH}}(\mathrm{E}\left\{Y_i^*\right\}; 1-\alpha) = \hat{Y}_i^* \pm \sqrt{2F(1-\alpha;2,n-2)} \cdot \text{s}\{\hat{Y}_i^*\}.$$

We select whichever of the Bonferroni or Working-Hotelling approaches yields the **tighter** C.I..

**Fuels Example** In the fuels dataset, at a family confidence level of 0.95, the required factor is

$$W = \sqrt{2F(0.95;2;18)} = 2.667.$$

The Working-Hotelling confidence band for the line of best fit is shown in pink below; the Bonferroni region for any 20 simultaneous inferences on the mean response also contains the **blue** region.

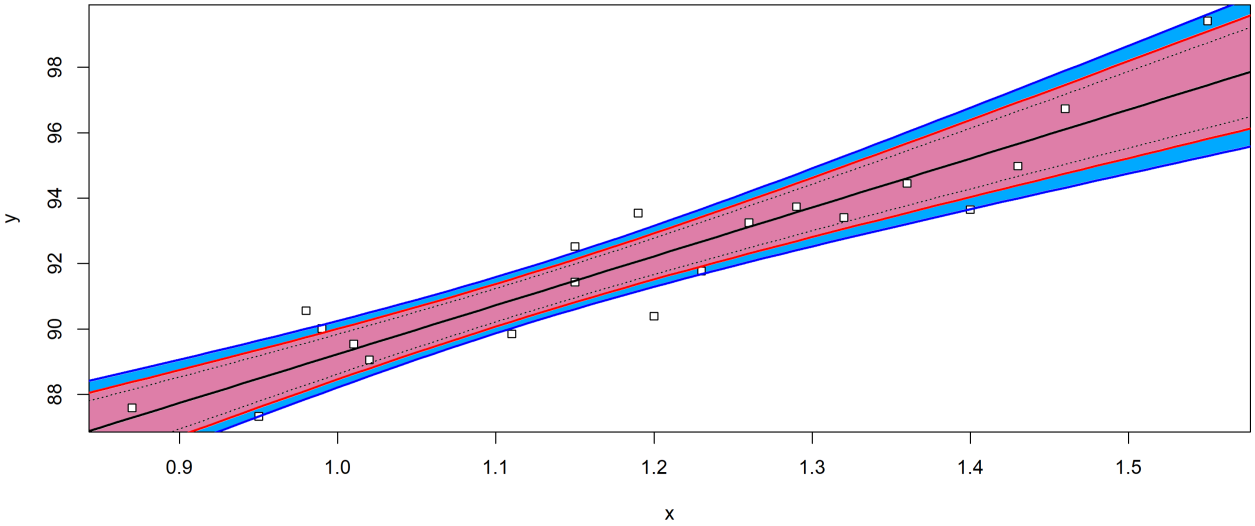



**Figure 8.11:** Joint Working-Hotelling confidence band (pink) and joint Bonferroni region fo 20 simultaneous inferences on the mean response (blue + pink) in the fuels dataset.

**Scheffé's Procedure and Joint Estimation of New Responses** If we want to obtain **joint prediction intervals** at family confidence level $1-\alpha$ for $g$ new responses $Y_{p_i}^*$ at predictor levels $X = X_i^*$, $i = 1,\ldots,g$, we use the approach (among the two below) that leads to "tighter" P.I.:

- if $g$ is "small", the **Bonferroni** prediction intervals are given by

$$\text{P.I.}_{\text{B}}(Y_{p_i}^*; 1-\alpha) \equiv \hat{Y}_{p_i}^* \pm t(1-\tfrac{\alpha/g}{2}; n-2) \cdot \text{s}\{\text{pred}_i^*\}, \quad i = 1,\ldots,g;$$

- if $g$ is "large", the **Scheffé** prediction intervals are

$$\text{P.I.}_{\text{S}}(Y_{p_i}^*; 1-\alpha) \equiv \hat{Y}_{p_i}^* \pm \sqrt{gF(1-\alpha;g,n-2)} \cdot \text{s}\{\text{pred}_i^*\}, \quad i = 1,\ldots,g.$$

### 8.2.4 Significance of Regression

What can we conclude if $\beta_1 = 0$? It could be that:

1. there is **no relationship** between $X$ and $Y$, as in a diffuse cloud of points – knowledge of $X$ explains nothing about the possible values of $Y$;
2. there is a **horizontal relationship** between $X$ and $Y$, so that changes in $X$ do not bring any change in $Y$;
3. there is a **non-linear relationship** between $X$ and $Y$ which is best approximated by a horizontal line.

In each of these cases, we say that regression is **not significant**.

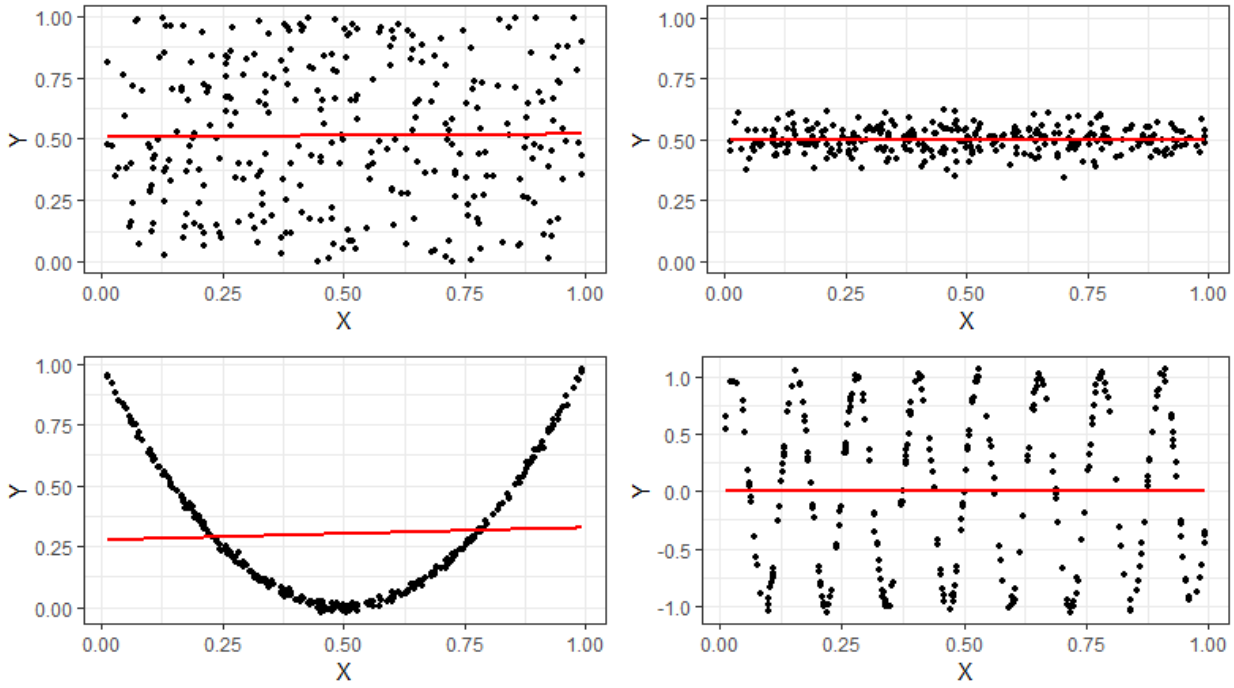



**Figure 8.12:** Examples of non-significant regressions.

This test for **significance of regression** is

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

The underlying assumptions are that:

1. the **simple linear regression model** holds, and
2. the error terms are **independent** and **normal**, with variance $\sigma^2$.

Under these assumptions, we can show that $b_0, b_1$ are **independent of SSE** and that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2).$$

**Analysis of Variance** Whether $H_0$ holds or not, the unbiased estimator for the error variance is

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n-2} \quad \left( \implies \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2) \right).$$

Recall that, in general: SST = SSR + SSE. If $H_0 : \beta_1 = 0$ holds, then $Y_1, \ldots, Y_n$ is an independent random sample drawn from $\mathcal{N}(\beta_0, \sigma^2)$. Our best estimate for $\sigma^2$ is thus

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\text{SST}}{n-1} \quad \Big( \implies \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1) \Big).$$

**Cochran's Theorem** implies that SSE, SSR are **independent**, and that

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2\big((n-1)-(n-2)\big) = \chi^2(1).$$

Thus, if $H_0 : \beta_1 = 0$ holds, the quotient

$$F^* = \frac{\underbrace{\left(\frac{\text{SSR}}{\sigma^2}\right)}_{\chi^2(\nu_1)} \Big/ \underbrace{1}_{\nu_1}}{\underbrace{\left(\frac{\text{SSE}}{\sigma^2}\right)}_{\chi^2(\nu_2)} \Big/ \underbrace{(n-2)}_{\nu_2}} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2)$$

follows a Fisher $F$ distribution with $1, n-2$ degrees of freedom.

It can be shown that $\text{E}\{\text{MSR}\} = \sigma^2 + \beta_1^2 S_{xx}$; if $\beta_1 \neq 0$, we thus have $\text{E}\{\text{MSR}\} > \sigma^2$, which means that large observed values of $F^*$ support $H_1 : \beta_1 \neq 0$.

**Decision Rule:** let $0 < \alpha \ll 1$. If $F^* > F(1-\alpha; 1, n-2)$, then we reject $H_0$ in favour of $H_1$ at level $\alpha$.[27]

27: We have already examined a test for significance of regression in Section 8.2.2. They are linked: when $\beta_1 = 0$, $F^* = (t^*)^2$.

**Fuels Example** In the fuels dataset, we have $n = 20$ and

$$\text{SST} = 173.38, \quad \text{SSR} = 152.13, \quad \text{SSE} = 21.25,$$

so that

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{152.13/1}{21.25/18} = 128.8631 = (11.351)^2;$$

at $\alpha = 0.05$, the critical value is $F(1-0.05; 1, 18) = 4.413873$. Since $F^* > F(0.95; 1, 18)$, we reject $H_0 : \beta_1 = 0$ at $\alpha = 0.05$, in favour of the alternative being that the regression is **significant** ($H_1 : \beta_1 \neq 0$).

**Golden Rule** In general, if SSx is a sum of squares with $n - x$ degrees of freedom, the corresponding **mean sum of squares** is

$$\text{MSx} = \frac{\text{SSx}}{n-x}.$$

Under some specific test assumptions,[28] MSx provides an unbiased estimator for the variance $\sigma^2$ of the error terms. Depending on the situation, Cochran's Theorem can then be used to show that

$$\frac{\text{SSx}}{\sigma^2} \sim \chi^2(n-x).$$

28: Or under general assumptions, depending on the sum of squares in question or the situation.

### 8.2.5 Simple Linear Regression in R

While we can compute quantities associated with the SLR model manually,[29] the `lm()` function in R produces an object from which we can extract most of them.

29: As we have done on numerous occasions earlier in this section.

**Fuels Example** We can easily compute the regression model in R.

```
(model <- lm(y ~ x))
plot(x,y); abline(model) # display points and line
```

```
Coefficients:
(Intercept)            x
      74.28        14.95
```

We can get more information *via* the `summary()` call.

```
summary(model)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.83029 -0.73334  0.04497  0.69969  1.96809

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.283      1.593   46.62  < 2e-16 ***
x             14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared:  0.8774,Adjusted R-squared:  0.8706
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

Other attributes are available, as seen below.

```
attributes(model)
```

```
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

```
attributes(summary(model))
```

```
$names
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

## 8.3 Multiple Linear Regression

The situation is usually more complicated; in particular, in any reasonable dataset we might expect to see $p$ **predictors** $X_k$, $k = 0, \ldots, p-1$.

#### Examples

- $X_1$: age, $X_2$: sex; $Y$: height ($p = 3$)
- $X_1$: age; $X_2$: years of education, $Y$: salary ($p = 3$)
- $X_1$: income; $X_2$: infant mortality; $X_3$: fertility rate, $Y$: life expectancy ($p = 4$)
- etc.

In theory, we hope that there is a **functional relationship** $Y = f(X_0, \ldots, X_{p-1})$ between $X_0(= 1), X_1, \ldots, X_{p-1}$ and $Y$. In practice (assuming that a relationship even exists), the best that we may be able to hope for is a **statistical relationship**

$$Y = f(X_0, X_1, \ldots, X_{p-1}) + \varepsilon,$$

where, as before, $f(X_0, X_1, \ldots, X_{p-1})$ is the **response function**, and $\varepsilon$ is the **random error** (or noise).

In **general linear regression**, we assume that the response function is

$$f(X_0, X_1, \ldots, X_p) = \beta_0 X_0(= 1) + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

The building blocks of regression analysis are the **observations**:

$$(X_{i,0}(= 1), X_{i,1}, \ldots, X_{i,p-1}, Y_i), \quad i = 1, \ldots, n.$$

In an ideal setting, these observations are **(jointly) randomly sampled**, according to some appropriate design.[30]

30: See Chapters 11 and 10.

The **general linear regression** (GLR) model is

$$Y_i = \beta_0 X_{i,0}(= 1) + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_k$, $k = 0, \ldots, p-1$ are **unknown parameters** and $\varepsilon_i$ is the **random error on the $i$th observation** (or case).[31] A GLR model need not necessarily be linear in $X$, but the mean response $\mathrm{E}\{Y\}$ must be **linear in the parameters** $\beta_k$, $k = 0, \ldots, p-1$.

31: Note that a predictor $X_k$ can be a function of other predictors. For instance, the following model is a GLR model:

$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2.$$

In what follows, we write

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p-1} \end{pmatrix},$$

for the **response vector**, the **parameter vector**, and the **design matrix**, respectively.

In the design matrix $\mathbf{X}$, $\mathbf{X}_i$ represents the $i$th case (the $i$th row of $\mathbf{X}$), a single **multiple predictor level**. The columns of the design matrix represent the values taken by the various predictor variables for all cases.

The **multiple linear regression model** is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Note that the SLR model fits into this framework, if we use $p = 2$ with

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix}.$$

### 8.3.1 Least Squares Estimation

We treat the predictor values $X_{i,k}$ as though they were constant, for $i = 1, \ldots, n$, $k = 0, \ldots, p-1$.[32] Since $\mathrm{E}\{\varepsilon_i\} = 0$, the **expected** (or mean) **response conditional on** $X_i$ is thus

32: That is, we assume that there is **no measurement error**.

$$\mathrm{E}\{Y_i \mid \mathbf{X}_i\} = \mathrm{E}\left\{\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \mid \mathbf{X}_i\right\} = \mathbf{X}_i\boldsymbol{\beta} + \mathrm{E}\{\varepsilon_i\} = \mathbf{X}_i\boldsymbol{\beta}.$$

The **deviation at** $\mathbf{X}_i$ is the difference between the observed response $Y_i$ and the expected response $\mathrm{E}\{Y_i \mid \mathbf{X}_i\}$:

$$e_i = Y_i - \mathrm{E}\{Y_i \mid \mathbf{X}_i\};$$

the deviation can be **positive** (if the point lies "**above**" the hyperplane $Y = \mathbf{X}\boldsymbol{\beta}$) or "**negative**" (if it lies **below**).

How do we find **estimators** for $\boldsymbol{\beta}$? Incidentally, how do we determine if the fitted hyperplane is a **good model for the data**?

Consider the function

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mathrm{E}\{Y_i \mid \mathbf{X}_i\})^2 = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2.$$

If $Q(\boldsymbol{\beta})$ is "small", then the sum of the **squared residuals** is "small", and so we would expect the hyperplane $Y = \mathbf{X}\boldsymbol{\beta}$ to be a good fit for the data.

The **least-square estimators** of the GLR problem is the vector $\mathbf{b} \in \mathbb{R}^p$ which minimizes the function $Q$ with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$. We must then find critical points of $Q(\boldsymbol{\beta})$, i.e., solve $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

**Matrix Notation** The OLS regression function is $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$, where $\mathbf{b}$ minimizes

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i\boldsymbol{\beta}\right)^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y}^\top - \boldsymbol{\beta}^\top \mathbf{X}^\top)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Since $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y}$ is a scalar, it is equal to its transpose $\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta}$, and so

$$Q(\boldsymbol{\beta}) = \mathbf{Y}^\top\mathbf{Y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

But $\mathbf{X}^\top\mathbf{X}$ is positive definite, so $Q(\boldsymbol{\beta})$ is minimized at $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

**Normal Equations** The gradient of $Q(\boldsymbol{\beta})$ is

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta},$$

so the critical point $\mathbf{b}$ solves the **normal equations**

$$(\mathbf{X}^\top \mathbf{X})\mathbf{b} = \mathbf{X}^\top \mathbf{Y}.$$

The matrix $\mathbf{X}^\top \mathbf{X}$ is called the **sum of squares and cross products** (SSCP) matrix; when it is invertible, the **unique** solution of the normal equations is

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

also known as the **LS estimates** of the GLR problem.[33]

33: The SSCP matrix is $p \times p$, and so is not usually too costly to invert, no matter the number of observations $n$, although in practice $p$ can be quite large.

For instance, say we have two predictors $X_1, X_2$ and three regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$. If we write $\mathbf{x} = (1, X_1, X_2)$, the **regression function** is

$$\mathrm{E}\{Y\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

If the OLS estimates are

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \left(0.5, -0.1, 2\right)^\top,$$

say, then the **estimated regression function** is

$$\hat{Y} = \mathbf{x}\mathbf{b} = 0.5 - 0.1 X_1 + 2 X_2.$$

**Residuals and Sums of Squares** The **fitted values** for the GLR problem are

$$\begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \underbrace{\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top}_{=\mathbf{H}} \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H}$ is the **hat matrix**.

**Theorem:** $\mathbf{H}$, $\mathbf{I}_n - \mathbf{H}$ are idempotent and symmetric, and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$.
**Proof:** we use the notation $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$. We will first need to show that $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{H}^\top = \mathbf{H}$, $\mathbf{M}^2 = \mathbf{M}$, and $\mathbf{M}^\top = \mathbf{M}$.

That this is the case is obvious:

$$\begin{aligned}
\mathbf{H}^2 &= \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top = \mathbf{X}\mathbf{I}_n \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top = \mathbf{H} \\
\mathbf{H}^\top &= \left(\mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top\right)^\top = \left(\mathbf{X}^\top\right)^\top \left(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)^\top \mathbf{X}^\top = \mathbf{X}\left(\left(\mathbf{X}^\top \mathbf{X}\right)^\top\right)^{-1} \mathbf{X}^\top \\
&= \mathbf{X}^\top \left(\mathbf{X}^\top (\mathbf{X}^\top)^\top\right)^{-1} \mathbf{X}^\top = \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H} \\
\mathbf{M}^2 &= (\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - \mathbf{I}_n\mathbf{H} - \mathbf{H}\mathbf{I}_n + \mathbf{H}^2 = \mathbf{I}_n - 2\mathbf{H} + \mathbf{H} = \mathbf{I}_n - \mathbf{H} = \mathbf{M} \\
\mathbf{M}^\top &= (\mathbf{I}_n - \mathbf{H})^\top = \mathbf{I}_n^\top - \mathbf{H}^\top = \mathbf{I}_n - \mathbf{H} = \mathbf{M}.
\end{aligned}$$

Furthermore,

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{I}_n = \mathbf{0},$$

which completes the proof. ■

The $i$**th residual** is $e_i = Y_i - \hat{Y}_i$. Since $\mathbf{MX} = \mathbf{0}$, the **residual vector** is

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{MY} \\ &= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\boldsymbol{\varepsilon}.\end{aligned}$$

In other words, the residual vector is both a linear transformation of the response vector $\mathbf{Y}$ and of the random error vector $\boldsymbol{\varepsilon}$. Just as in the SLR case (which is a special case of GLR), the residuals have a set of nice properties.

**Theorem:** the design matrix is orthogonal to the residual vector, i.e., $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$ (the columns of $\mathbf{X}$ are orthogonal to $\mathbf{e}$).

**Proof:** from the normal equations, we get

$$\mathbf{X}^\top\mathbf{Xb} = \mathbf{X}^\top\mathbf{Y} \implies \mathbf{X}^\top(\mathbf{Y} - \mathbf{Xb}) = \mathbf{0} \implies \mathbf{X}^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

But $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{e}$, so that $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$. ■

**Theorem:** if the model has an intercept term $\beta_0$, we also have $\mathbf{1}_n^\top\mathbf{e} = 0$, $\bar{e} = \bar{Y} - \bar{\hat{Y}} = 0$, and $\hat{\mathbf{Y}}^\top\mathbf{e} = 0$.

**Proof:** if there is an intercept term, the first column of the design matrix $\mathbf{X}$ is $\mathbf{1}_n$. Thus $\mathbf{1}_n^\top\mathbf{e}$ corresponds to the first entry of $\mathbf{X}^\top\mathbf{e} = \mathbf{0}$, which is to say, 0. This also implies that $\bar{e} = 0$. For the last part, recall that $\hat{\mathbf{Y}} = \mathbf{Xb}$, and so $\hat{\mathbf{Y}}^\top = \mathbf{b}^\top\mathbf{X}^\top$ and $\hat{\mathbf{Y}}^\top\mathbf{e} = \mathbf{b}^\top\mathbf{X}^\top\mathbf{e} = \mathbf{b}^\top\mathbf{0} = 0$. ■

We have already seen that SST is a quadratic form in $\mathbf{Y}$:

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \mathbf{Y}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y};$$

from the definition of the residuals, we see that this also holds for SSE:

$$\begin{aligned}\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^\top\mathbf{e} = (\mathbf{MY})^\top\mathbf{MY} = \mathbf{Y}^\top\mathbf{M}^\top\mathbf{MY} \\ = \mathbf{Y}^\top\mathbf{M}^2\mathbf{Y} = \mathbf{Y}^\top\mathbf{MY} = \mathbf{Y}^\top\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y}.\end{aligned}$$

The sum of squares decomposition can then be re-written as:

$$\text{SSR} = \text{SST} - \text{SSE}.$$

Thus, SSR is also a quadratic form in $\mathbf{Y}$:

$$\begin{aligned}\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 &= \mathbf{Y}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y} - \mathbf{Y}^\top\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y} \\ &= \mathbf{Y}^\top\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n - \mathbf{I}_n + \mathbf{H}\right)\mathbf{Y} = \mathbf{Y}^\top\left(\mathbf{H} - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}.\end{aligned}$$

**Theorem:** $\text{E}\{\text{SSE}\} = (n-p)\sigma^2$ and $\text{rank}(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - p$. Thus, SSE has $n - p$ degrees of freedom.

**Proof:** we have

$$\text{SSE} = \mathbf{e}^\top\mathbf{e} = (\mathbf{M}\boldsymbol{\varepsilon})^\top\mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbf{M}\boldsymbol{\varepsilon} = \sum_{i,j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j = \sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j.$$

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$,

$$\begin{aligned} \mathrm{E}\left\{\varepsilon_i^2\right\} &= \sigma^2\left\{\varepsilon_i\right\} + \left(\mathrm{E}\left\{\varepsilon_i\right\}\right)^2 = \sigma^2 + 0 = \sigma^2, \quad i = 1, \ldots, n, \quad \text{and} \\ \mathrm{E}\left\{\varepsilon_i \varepsilon_j\right\} &= \sigma\left\{\varepsilon_i, \varepsilon_j\right\} + \mathrm{E}\left\{\varepsilon_i\right\} \mathrm{E}\left\{\varepsilon_j\right\} = 0 + 0 = 0, \quad i \neq j. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathrm{E}\left\{\mathrm{SSE}\right\} &= \mathrm{E}\left\{\sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} m_{ii}\varepsilon_i^2\right\} + \mathrm{E}\left\{\sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j\right\} \\ &= \sum_{i=1}^{n} m_{ii}\mathrm{E}\left\{\varepsilon_i^2\right\} + \sum_{i\neq j} m_{ij}\mathrm{E}\left\{\varepsilon_i\varepsilon_j\right\} = \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2 \mathrm{trace}(\mathbf{M}) \\ &= \sigma^2 \mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \sigma^2[\mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H})] = \sigma^2[n - \mathrm{trace}(\mathbf{H})]. \end{aligned}$$

But

$$\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\Big(\underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{A_{n\times p}} \underbrace{\mathbf{X}^\top}_{B_{p\times n}}\Big) = \mathrm{trace}\Big(\underbrace{\mathbf{X}^\top}_{B_{p\times n}} \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{A_{n\times p}}\Big) = \mathrm{trace}(\mathbf{I}_p) = p,$$

whence $\mathrm{E}\left\{\mathrm{SSE}\right\} = (n-p)\sigma^2$. ■

The **mean square error** MSE in the GLR model is

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-p},$$

which is not surprising as we have to estimate the $p$ parameters $\beta_k$, $k = 0, \ldots, p-1$, in order to compute SSE. According to the previous theorem, MSE is an **unbiased estimator of the error variance** $\sigma^2$.

### 8.3.2 Inference, Estimation, and Prediction

Assuming **normality** and **independence** of the random errors, the estimators $b_0, \ldots, b_{p-1}$ are then independent of SSE and

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-p).$$

This information allows us to test for the **significance of regression** using the **overall** $F-$**test**:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \quad \text{against} \quad H_1 : \beta_k \neq 0 \text{ for some } k = 1, \ldots, p-1$$

assuming that the GLR model holds.

**Analysis of Variance** In particular, we have

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \quad i = 1, \ldots, n.$$

Whether $H_0$ holds or not, the unbiased estimator for the error variance is

$$\widehat{\sigma^2} = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-p} \quad \left(\implies \frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-p)\right).$$

If $H_0$ holds, then $Y_1, \ldots, Y_n$ is an independent random sample drawn from $\mathcal{N}(\beta_0, \sigma^2)$. Our best estimate for $\sigma^2$ is thus

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\text{SST}}{n-1} \quad \Big( \implies \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1) \Big).$$

Since SST = SSE + SSR, **Cochran's Theorem** implies that SSE, SSR are **independent**, and that

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2\big((n-1)-(n-p)\big) = \chi^2(p-1).$$

Thus, if $H_0$ holds, the quotient

$$F^* = \frac{\left(\frac{\text{SSR}}{\sigma^2}\right)\Big/(p-1)}{\left(\frac{\text{SSE}}{\sigma^2}\right)\Big/(n-p)} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \frac{\text{MSR}}{\text{MSE}} \sim F(p-1, n-p)$$

follows a Fisher $F$ distribution with $p-1, n-p$ degrees of freedom.

The corresponding **ANOVA** table is

| **Source** | **SS** | **df** | **MS** | **F*** |
|---|---|---|---|---|
| Regression | SSR | $p-1$ | MSR = SSR/$(p-1)$ | MSR/MSE |
| Error | SSE | $n-p$ | MSE = SSE/$(n-p)$ | |
| Total | SST | $n-1$ | | |

The overall $F-$test's **p−value** is

$$P(F(p-1, n-p) > F^*).$$

**Decision Rule:** at confidence level $1-\alpha$, we reject $H_0$ if

$$F^* > F(1-\alpha; p-1, n-p);$$

equivalently, we reject $H_0$ if $P(F(p-1, n-p) > F^*) < \alpha$.

**Toy Example** Consider a dataset with $n = 12$ observations, a response variable $Y$ and $p-1 = 4$ predictors $X_1$, $X_2$, $X_3$, $X_4$. We build a GLR model

$$\begin{aligned} Y_i &= \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, 12 \\ &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \varepsilon_i, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{12}) \end{aligned}$$

The corresponding **ANOVA** table is

| **Source** | **SS** | **df** | **MS** | **F*** |
|---|---|---|---|---|
| Regression | 4957.2 | 4 | 1239.3 | 5.1 |
| Error | 1699.0 | 7 | 242.7 | |
| Total | 6656.2 | 11 | | |

With a $p-$value $= P(F(4,7) > 5.1) = 0.0303$, we **reject** $H_0$ at $\alpha = 0.05$ and conclude that the regression is **significant**.

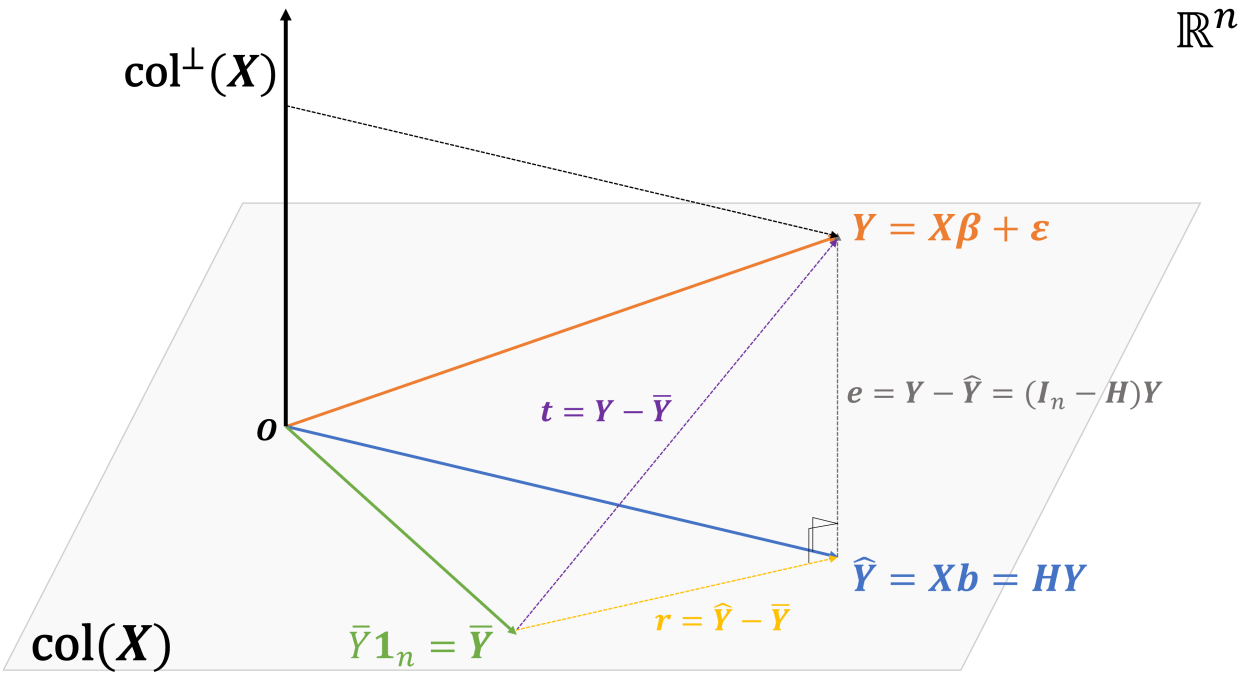



**Figure 8.13:** Geometrical interpretation of multiple linear regression: the sums of squares decomposition is a manifestation of Pythagoras' Theorem (see below).

**Geometrical Interpretation** A number of GLR concepts become easier to understand when viewed through the prism of **geometry** and **vector algebra**. Let

$$\begin{aligned}\mathscr{M}(\mathbf{X}) &= \text{colsp}(\mathbf{X}) = \{\mathbf{X}\gamma \mid \gamma \in \mathbb{R}^p\} < \mathbb{R}^n \\ \mathscr{M}^\perp(\mathbf{X}) &= (\text{colsp}(\mathbf{X}))^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v} \cdot \mathbf{w} = 0,\ \forall \mathbf{w} \in \mathscr{M}(\mathbf{X})\}\end{aligned}$$

The **vector of observations** $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ lies in $\mathbb{R}^n$, while the **fitted vector** $\mathbf{Y} = \mathbf{Xb} = \mathbf{HY}$ lies in $\mathscr{M}(\mathbf{X})$ and

$$\mathbf{e} = \mathbf{Y} - \mathbf{Y} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$$

lies in $\mathscr{M}^\perp(\mathbf{X})$. The hat matrix $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are idempotent (they are the projection matrices on $\mathscr{M}(\mathbf{X})$ and $\mathscr{M}^\perp(\mathbf{X})$) and symmetric.

The OLS estimator $\mathbf{b}$ is such that $\mathbf{Xb}$ is the closest vector to $\mathbf{Y}$ in $\mathscr{M}(\mathbf{X})$:

$$\mathbf{b} = \arg\min_{\gamma \in \mathbb{R}^p} \left\{\|\mathbf{Y} - \mathbf{X}\gamma\|_2^2\right\} = \arg\min_{\gamma \in \mathbb{R}^p} \left\{\|\mathbf{e}\|_2^2\right\} = \arg\min_{\gamma \in \mathbb{R}^p} \{\text{SSE}\}\,.$$

If the GLR model has a constant term $\beta_0$, the mean vector $\overline{\mathbf{Y}} = \overline{Y}\mathbf{1}_n$ lies in $\mathscr{M}(\mathbf{X})$; indeed, for $\gamma^* = (\overline{Y}, 0, \ldots, 0)^\top$, we have $\overline{\mathbf{Y}} = \mathbf{X}\gamma^*$. The triangle $\Delta\mathbf{Y}\mathbf{Y}\overline{\mathbf{Y}}$ is thus a **right angle triangle**, with

$$\mathbf{t} = \mathbf{Y} - \overline{\mathbf{Y}} = (\mathbf{Y} - \mathbf{Y}) + (\mathbf{Y} - \overline{\mathbf{Y}}) = \mathbf{e} + \mathbf{r};$$

Pythagoras' Theorem then gives us

$$\|\mathbf{t}\|_2^2 = \text{SST} = \text{SSE} + \text{SSR} = \|\mathbf{e}\|_2^2 + \|\mathbf{r}\|_2^2.$$

**Model Parameters** As was the case with the SLR model parameters, if $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, then

$$\mathbf{Y} \sim \mathcal{N}(\mathrm{E}\{\mathbf{Y}\}, \sigma^2\{\mathbf{Y}\}) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

If $A$ is any compatible matrix, then

$$A\mathbf{Y} \sim \mathcal{N}(A\mathrm{E}\{\mathbf{Y}\}, A\sigma^2\{\mathbf{Y}\}A^\top) = \mathcal{N}(A\mathbf{X}\boldsymbol{\beta}, \sigma^2 AA^\top).$$

From the normal equations, the OLS estimates for the GLR model are given by a **linear transformation** of the response vector $\mathbf{Y}$:

$$\mathbf{b} = \underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{p\times n}\mathbf{Y} = A\mathbf{Y}.$$

In particular,

$$\mathrm{E}\{\mathbf{b}\} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}\{\mathbf{Y}\} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

so that $\mathbf{b}$ provides **unbiased estimators** of $\boldsymbol{\beta}$. Furthemore,

$$\begin{aligned}\sigma^2\{\mathbf{b}\} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\{\mathbf{Y}\}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}_n\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

Thus,

$$\mathbf{b} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

The **estimated variance-covariance matrix** for the estimators $\mathbf{b}$ is thus

$$\mathrm{s}^2\{\mathbf{b}\} = \mathrm{MSE}\cdot(\mathbf{X}^\top\mathbf{X})^{-1}, \quad \text{and} \quad \mathrm{s}\{\mathbf{b}\} = \sqrt{\mathrm{MSE}}\sqrt{\mathrm{diag}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]}.$$

For each $k = 0, \ldots, p-1$, the **studentization** of $b_k$ is

$$T_k = \frac{b_k - \beta_k}{\sqrt{\mathrm{MSE}}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{k,k}}} = \underbrace{\frac{b_k - \beta_k}{\sigma\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{k,k}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-p)}_{=\nu}} \sim t(n-p),$$

where $(\mathbf{X}^\top\mathbf{X})^{-1}_{k,k}$ represents the $k+1$ entry in $\mathrm{diag}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]$.

For a specific $k \in \{0, \ldots, p-1\}$, the $100(1-\alpha)\%$ C.I. for $\beta_k$ is

$$\text{C.I.}(\beta_k; 0.95) \equiv b_k \pm t\left(1 - \tfrac{\alpha}{2}; n-p\right)\cdot \mathrm{s}\{b_k\}.$$

The corresponding hypothesis tests for

$$H_0 : \beta_k = \beta_k^* \qquad \text{against} \quad H_1 : \begin{cases} \beta_k < \beta_k^* & \text{left-tailed test} \\ \beta_k > \beta_k^* & \text{right-tailed test} \\ \beta_k \neq \beta_k^* & \text{two-tailed test} \end{cases}$$

Under $H_0$, the computed test statistic

$$T_k = \frac{b_k - \beta_k^*}{\mathrm{s}\{b_k\}} \sim t(n-p).$$

The **critical region** for the test depends on the **confidence level** $1-\alpha$ and on the **type** of the alternative hypothesis $H_1$. Let $t^*$ be the observed value of $T_k$. **We reject** $H_0$ **if** $t^*$ **is in the critical region**.

| Alternative Hypothesis | Rejection Region |
|---|---|
| $H_1 : \beta_k < \beta_k^*$ | $t^* < -t(1-\alpha; n-p)$ |
| $H_1 : \beta_k > \beta_k^*$ | $t^* > t(1-\alpha; n-p)$ |
| $H_1 : \beta_k \neq \beta_k^*$ | $\lvert t^* \rvert > t(1-\alpha/2; n-p)$ |

**Toy Example** Consider the situation with $n = 12$ observations and $p-1 = 4$ predictors as described previously. We build the GLR model $\hat{\mathbf{Y}} = \mathbf{Xb}$ and obtain the following results:

| Predictor | Estimate | SE | t |
|---|---|---|---|
| Intercept | -102.71 | 207.86 | -0.49 |
| $X_1$ | 0.61 | 0.37 | 1.64 |
| $X_2$ | 8.92 | 5.3 | 1.68 |
| $X_3$ | 1.44 | 2.39 | 0.60 |
| $X_4$ | 0.01 | 0.77 | 0.02 |

Recall that $n-p = 7$; the 95% C.I. for $\beta_2$ is thus

$$\text{C.I.}(\beta_2; 0.95) \equiv 8.92 \pm t(0.975; 7) \cdot 5.3 = 8.92 \pm 2.365 \cdot 5.3 = [-3.6, 21.5].$$

We could also test for $H_0 : \beta_3 = 2$ against $H_1 : \beta_3 \neq 2$, say: under $H_0$,

$$T_3^* = \frac{b_3 - 2}{\text{s}\{b_3\}} \sim t(7).$$

The observed statistic is

$$t^* = \frac{1.44 - 2}{2.39} = -0.23;$$

we would reject $H_0$ at confidence level $1-\alpha = 0.95$ if

$$\lvert t^* \rvert > t(0.975; 7) = 2.365;$$

as $-0.23 \not> 2.365$, we cannot conclude that $\beta_3 \neq 2$.[34]

34: While we can build a C.I. for $\beta_2$ and test a hypothesis about $\beta_3$, each at the $1-\alpha = 0.95$ confidence level, we cannot do so **jointly**.

**Mean Response** We can also conduct inferential analysis for the **expected response** at $\mathbf{X}^* = (1, X_1^*, \ldots, X_{p-1}^*)$ in the model's **scope**. In the GLR model, we assume that

$$\text{E}\{Y^*\} = \mathbf{X}^*\boldsymbol{\beta} = \beta_0 + \beta_1 X_1^* + \cdots + \beta_{p-1} X_{p-1}^*.$$

The **estimated mean response** at $\mathbf{X}^*$ is

$$\hat{Y}^* = \mathbf{X}^*\mathbf{b} = b_0 + b_1 X_1^* + \cdots + b_{p-1} X_{p-1}^*.$$

The predictor values are **fixed**, thus $\hat{Y}^*$ is normally distributed with

$$\text{E}\{\hat{Y}^*\} = \text{E}\{\mathbf{X}^*\mathbf{b}\} = \mathbf{X}^*\text{E}\{\mathbf{b}\} = \mathbf{X}^*\boldsymbol{\beta},$$

so that $\hat{Y}^*$ is an **unbiased estimator** of $\text{E}\{Y^*\}$.

Furthermore,

$$\sigma^2\{\hat{Y}^*\} = \mathbf{X}^*\sigma^2\{\mathbf{b}\}(\mathbf{X}^*)^\top = \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top,$$

so that

$$s^2\{\hat{Y}^*\} = \text{MSE}\cdot\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = \mathbf{X}^*s^2\{\mathbf{b}\}(\mathbf{X}^*)^\top.$$

The **estimated standard error** is thus

$$s\{\hat{Y}^*\} = \sqrt{\mathbf{X}^*s^2\{\mathbf{b}\}(\mathbf{X}^*)^\top}.$$

Since

$$\hat{Y}^* = \mathbf{X}^*\mathbf{b} = \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

is a **linear transformation** of $\mathbf{Y}$, and since

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),$$

then

$$\hat{Y}^* \sim \mathcal{N}\left(\mathrm{E}\{\hat{Y}^*\}, \sigma^2\{\hat{Y}^*\}\right) = \mathcal{N}\left(\mathbf{X}^*\boldsymbol{\beta}, \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right).$$

Thus

$$Z = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\sigma\{\hat{Y}^*\}} = \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}} \sim \mathcal{N}(0,1).$$

The **studentization** of $\hat{Y}^*$ is then

$$\begin{aligned} T &= \underbrace{\frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-p)}_{=\nu}} \\ &= \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sqrt{\text{MSE}}\sqrt{\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}} \sim t(n-p). \end{aligned}$$

For a specific predictor level $\mathbf{X}^*$, the $100(1-\alpha)\%$ C.I. for $\mathrm{E}\{Y^*\}$ is

$$\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) \equiv \hat{Y}^* \pm t\left(1 - \tfrac{\alpha}{2}; n-p\right)\cdot s\{\hat{Y}^*\}.$$

The corresponding hypothesis tests for

$$H_0 : \mathrm{E}\{Y^*\} = \gamma \quad \text{against} \quad H_1 : \begin{cases} \mathrm{E}\{Y^*\} < \gamma & \text{left-tailed test} \\ \mathrm{E}\{Y^*\} > \gamma & \text{right-tailed test} \\ \mathrm{E}\{Y^*\} \neq \gamma & \text{two-tailed test} \end{cases}$$

Under $H_0$, the computed test statistic

$$T = \frac{\hat{Y}^* - \gamma}{s\{\hat{Y}^*\}} \sim t(n-p).$$

The **critical region** for the test depends on the **confidence level** $1-\alpha$ and on the **type** of the alternative hypothesis $H_1$. Let $t^*$ be the observed value of $T$. **We reject $H_0$ if $t^*$ is in the critical region**.

| Alternative Hypothesis | Rejection Region |
|---|---|
| $H_1 : \mathrm{E}\{Y^*\} < \gamma$ | $t^* < -t(1-\alpha; n-p)$ |
| $H_1 : \mathrm{E}\{Y^*\} > \gamma$ | $t^* > t(1-\alpha; n-p)$ |
| $H_1 : \mathrm{E}\{Y^*\} \neq \gamma$ | $\lvert t^* \rvert > t(1-\alpha/2; n-p)$ |

**Toy Example** Consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously. We would like to predict the expected response at

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{in the model's } \textbf{scope}.$$

Thus

$$\begin{aligned}\hat{Y}^* &= \mathbf{X}^*\mathbf{b} \\ &= -102.71 + 0.61(11.10) + 8.92(20.74) + 1.44(6.61) + 0.01(182.38) \\ &= 100.40.\end{aligned}$$

Recall that MSE = 242.71. Using the data, we computed

$$\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = 1.42,$$

so that

$$\mathrm{s}\{\hat{Y}^*\} = \sqrt{242.71}\sqrt{1.42} = 22.12.$$

Since $n - p = 7$; the 95% C.I. for $\mathrm{E}\{Y^*\}$ is

$$\begin{aligned}\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) &\equiv 100.40 \pm t(0.975; 7) \cdot 22.12 \\ &= 100.40 \pm 2.365 \cdot 22.12 = [48.09, 152.71].\end{aligned}$$

We could also test for $H_0 : \mathrm{E}\{Y^*\} = 150$ against $H_1 : \mathrm{E}\{Y^*\} < 150$, say: under $H_0$,

$$T^* = \frac{\hat{Y}^* - 150}{\mathrm{s}\{\hat{Y}^*\}} \sim t(7).$$

The observed statistic is

$$t^* = \frac{100.40 - 150}{22.12} = -2.24.$$

We would reject $H_0$ at confidence level $1 - \alpha = 0.95$ if

$$t^* < -t(0.95; 7) = -1.89;$$

as $-2.24 < -1.89$, the evidence is strong enough to **reject**

$$H_0 : \mathrm{E}\{Y^*\} = 150 \quad \text{in favour of} \quad H_1 : \mathrm{E}\{Y^*\} < 150.$$

Note, however, that the two-sided 95% C.I. for $\mathrm{E}\{Y^*\}$ contains 150, so we **cannot reject**

$$H_0 : \mathrm{E}\{Y^*\} = 150 \quad \text{in favour of} \quad H_1 : \mathrm{E}\{Y^*\} \neq 150$$

at confidence level $1 - \alpha = 95\%$. As before, we cannot conduct **joint inferences** about various predictor levels $\mathbf{X}^*$ without modifications.

**Prediction Intervals** Let $Y_p^*$ represent a **(new) response** at $\mathbf{X}^*$, so that

$$Y_p^* = \mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p \quad \text{for some } \varepsilon_p.$$

If the average error is 0, the best prediction for $Y_p^*$ is still the **fitted response at $\mathbf{X}^*$** :

$$\hat{Y}_p^* = \mathbf{X}^*\mathbf{b}.$$

The **prediction error** at $\mathbf{X}^*$ is thus

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p - \mathbf{X}^*\mathbf{b}.$$

In the GLR model, the error $\varepsilon_p$ and the estimators $\mathbf{b}$ are **normally distributed**. Consequently, so is the prediction error $\text{pred}^*$. Note that

$$\text{E}\{\text{pred}^*\} = \underbrace{\text{E}\left\{\mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p^*\right\}}_{=\mathbf{X}^*\boldsymbol{\beta}} - \underbrace{\text{E}\{\mathbf{X}^*\mathbf{b}\}}_{=\mathbf{X}^*\boldsymbol{\beta}} = 0.$$

Because the residuals are uncorrelated with the response, we also have

$$\begin{aligned}\sigma^2\{\text{pred}^*\} &= \sigma^2\left\{Y_p^*\right\} + \sigma^2\left\{\hat{Y}_p^*\right\} \\ &= \sigma^2 + \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = \sigma^2\left[1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right].\end{aligned}$$

Thus $\text{pred}^* \sim \mathcal{N}\left(0, \sigma^2\left[1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right]\right)$ and the estimated standard error is

$$\text{s}\{\text{pred}^*\} = \sqrt{\text{MSE}}\sqrt{1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}.$$

As before, we can show that

$$T_p^* = \frac{\text{pred}^* - 0}{\text{s}\{\text{pred}^*\}} \sim t(n-p), \quad \text{and so}$$
$$\text{P.I.}(Y_p^*; 1-\alpha) \equiv \mathbf{X}^*\mathbf{b} \pm t(1-\tfrac{\alpha}{2}; n-p)\cdot \text{s}\{\text{pred}^*\}.$$

Note that $\text{s}\{\hat{Y}^*\} < \text{s}\{\text{pred}^*\}$ so that the C.I. for the mean response is always **contained** in the P.I. for new responses.

**Toy Example** Consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously. We would like to predict the new responses at

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{in the model's } \textbf{scope}.$$

We have already seen that $\hat{Y}^* = \mathbf{X}^*\mathbf{b} = 100.40$. Recall that MSE = 242.71 and

$$\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = 1.42,$$

so that

$$\text{s}\{\text{pred}^*\} = \sqrt{242.71}\sqrt{1 + 1.42} = 37.70.$$

Since $n - p = 7$, the 95% P.I. for $Y^*$ is

$$\begin{aligned}\text{P.I.}(Y^*; 0.95) &\equiv 100.40 \pm t(0.975; 7)\cdot 37.70 \\ &= 100.40 \pm 2.365 \cdot 37.70 = [11.24, 189.56].\end{aligned}$$

**Joint Estimation and Prediction** At a family confidence level of $1-\alpha$:

- the **Bonferroni** procedure can be used to jointly estimate $g$ model parameters $\beta_{k_\ell}$, $g$ mean responses $\mathrm{E}\left\{Y_\ell^*\right\}$, or $g$ new responses $Y_\ell^*$, for $\ell = 1, \ldots, g$;
- the **Working-Hostelling** procedure can be used to jointly estimate $g$ mean responses $\mathrm{E}\left\{Y_\ell^*\right\}$, for $\ell = 1, \ldots, g$;
- the **Scheffé** procedure can be used to jointly predict $g$ new responses $Y_\ell^*$, for $\ell = 1, \ldots, g$.

The process is identical to the SLR approach; depending on the task at hand, we pick the appropriate procedure that yields the **smallest interval**.

The sole difference lies in the composition of the **factors** that accompany the estimated standard errors in the construction of the **joint** confidence/prediction intervals at **family confidence level** $1-\alpha$:

- $t(1-\frac{\alpha/g}{2}; n-p)$ for the Bonferroni procedure;
- $\sqrt{pF(1-\alpha; p, n-p)}$ for the Working-Hotelling procedure, and
- $\sqrt{gF(1-\alpha; g, n-p)}$ for the Scheffé procedure.

**Toy Example** We can provide joint confidence intervals for the **model parameters** in the preceding example at family confidence level $1-\alpha = 0.95$, using $n-p=7$ and $g=5$. The **Bonferroni** factor is

$$t\left(1-\tfrac{0.05/5}{2}; 7\right) = t(0.995; 7) = 3.50;$$

the joint confidence intervals are:

$$\text{C.I.}_{\text{B}}(\beta_k; 0.95) \equiv b_k \pm 3.50 \cdot \text{s}\left\{b_k\right\}.$$

| **Parameter** | $b_k$ | $\text{C.I.}_{\text{B}}(\beta_k; 0.95)$ |
|---|---|---|
| $\beta_0$ | -102.71 | [-830.22, 624.80] |
| $\beta_1$ | 0.61 | [-0.685 , 1.905] |
| $\beta_2$ | 8.92 | [ -9.63 ,27.47] |
| $\beta_3$ | 1.44 | [-6.925 , 9.805] |
| $\beta_4$ | 0.01 | [-2.685 , 2.705] |

Individually, **none of the parameters** are significant at the family confidence level $1-\alpha = 0.95$ (all the confidence intervals contain 0), but the regression **as a whole** is significant (see overall $F$-test example).

Similarly, the **Working-Hotelling** joint confidence intervals for the estimated mean $\mathrm{E}\left\{Y_\ell^*\right\}$ at a variety of predictor levels $\mathbf{X}_\ell^*$, $\ell = 1, \ldots, g$ (family confidence level $1-\alpha = 0.95$) are

$$\begin{aligned}\text{C.I.}_{\text{WH}}(\mathrm{E}\left\{Y_\ell^*\right\}; 0.95) &\equiv \hat{Y}_\ell^* \pm \sqrt{5F(0.95; 5, 7)} \cdot \text{s}\{\hat{Y}_\ell^*\} \\ &= \mathbf{X}_\ell^*\mathbf{b} \pm 4.46 \sqrt{\underbrace{242.71}_{=\text{MSE}}} \sqrt{\mathbf{X}_\ell^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}_\ell^*)^\top}\end{aligned}$$

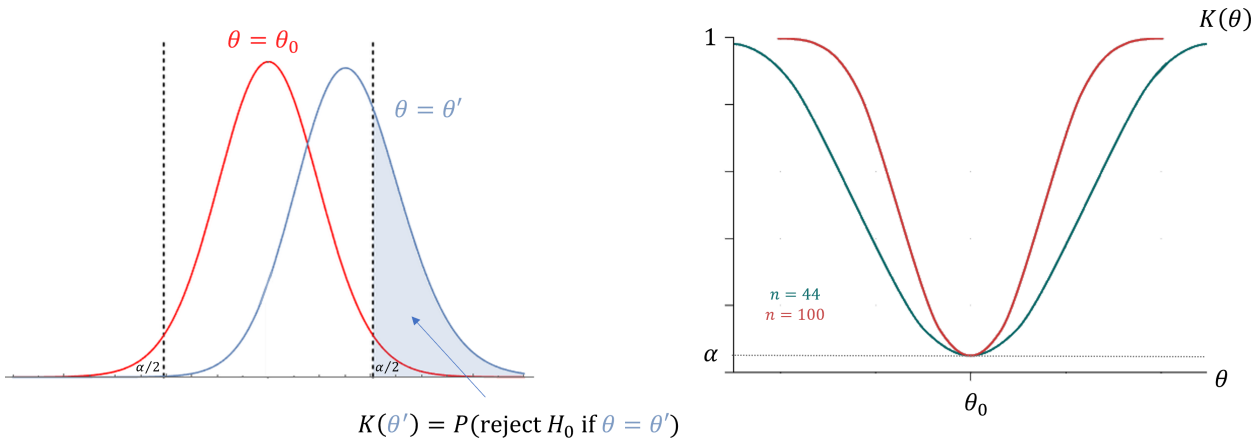



**Figure 8.14:** Power function (right) and error of type I (left).

### 8.3.3 Power of a Test

When we do hypothesis testing, we can make two types of errors.

- **Type I Error:** rejecting a valid $H_0$
- **Type II Error:** failing to reject $H_0$ when $H_1$ is valid.[35]

The **level of significance** $\alpha$ is used to control the risk of making an error of type I; type II errors are harder to control, in general.

Suppose we are testing (2−sided test) for

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Let $\alpha$ be the probability of making an error of type I.

The **power function**

$$K(\theta') = P(\text{reject } H_0 \text{ if } \theta = \theta')$$

is such that $K(\theta_0) = \alpha$.

If $\theta \neq \theta_0$, $t^* = \frac{\hat{\theta} - \theta_0}{\text{s}\{\hat{\theta}\}} \sim t(\nu)$ with **non-centrality parameter**

$$\delta = \frac{|\theta - \theta_0|}{\sigma\{\hat{\theta}\}} \approx \frac{|\theta - \theta_0|}{\text{s}\{\hat{\theta}\}},$$

where $\theta$ is the true value and $\theta_0$ is the value under $H_0$. The **power of the test** is the probability of rejecting $H_0$ if $\theta = \theta'$:

$$K(\theta') = P(|t^*| > t(1 - \alpha/2; \nu); \delta).$$

To control the power, we can either increase $n$ or decrease $S_{xx}$ (as we can see in Figure 8.14).

We will revisit these notions in Chapter 11.

35: There are other types of error, such as "correctly rejecting $H_0$ for the wrong reason", "giving the right answer to the wrong problem ", "choosing the wrong problem representation", "deliberately selecting the wrong questions for intensive and skilled investigation", "incorrectly interpreting a correctly rejected $H_0$" and so on, but that is outside the scope of this chapter. See wikipedia.org/wiki/Type_III_error for details.

### 8.3.4 Coefficients of Determination

The **coefficient of multiple determination** of a GLR model is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

the proportion of the variation in $Y$ which is explained by the regression. If the GLR model incorporates an intercept term ($\beta_0 \neq 0$), then

$$R^2 = r^2_{Y\hat{Y}} = \frac{(s_{Y\hat{Y}})^2}{s_Y s_{\hat{Y}}};$$

this is not the case without an intercept term. When the number of parameters $p$ increases, so does $R^2$; however, the degrees of freedom, $n - p$ decrease. This typically means that the estimates are less precise. We can adjust $R^2$ to take this loss into account.

The **adjusted coefficient of multiple determination** of a GLR model is

$$R^2_a = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p} \cdot \frac{\text{SSE}}{\text{SST}} \quad \text{(which could be } < 0\text{)}.$$

**Toy Example** In the case we have been carrying around for a while, we had

$$\text{SST} = 6656.2, \quad \text{SSE} = 1699.0, \quad n - p = 7, \quad n - 1 = 11,$$

so that

$$R^2 = 1 - \frac{1699.0}{6656.2} = 0.745 \quad \text{and} \quad R^2_a = 1 - \frac{11}{7} \cdot \frac{1699.0}{6656.2} = 0.599.$$

### 8.3.5 Diagnostics and Remedial Measures

We have seen that there are **four** GLR assumptions:

- **linearity** – $\text{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$;
- **variance constancy (homoscedasticity)** – $\sigma^2\{\varepsilon_i\} = \sigma^2, i = 1, \ldots, n$;
- **independence** – $\varepsilon_1, \ldots, \varepsilon_n$ are independent,[36] and
- **normality** – $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$.

36: **Uncorrelated** is in fact sufficient.

We have combined these assumptions in the simpler vector form

$$Y \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

These assumptions must be met before we can trust the GLR model.[37]

37: In theory, at least. In practice, the model may prove useful even if they are not met, but that must be established **on a case-by-case basis**.

Recall that we have the following results on the **residuals**:

1. $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, or $e_i = Y_i - \hat{Y}_i$, for $i = 1, \ldots, n$;
2. if $\beta_0 \neq 0$, $\overline{\mathbf{e}} = 0$, and
3. $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}_n - \mathbf{H})$, so that $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$, for $i = 1, \ldots, n$, and $\sigma\{e_i, e_j\} = \sigma\{e_j, e_i\} = -h_{ij}\sigma^2$ for $i \neq j = 1, \ldots, n$.

The **standard error** is $\text{s}^2\{e_i\} = \text{MSE}(1 - h_{ii})$ and the **internal studentization** is $r_i = \frac{e_i - \overline{e}}{\text{s}\{e_i\}} \sim t(n-p)$, for $i = 1, \ldots, n$.

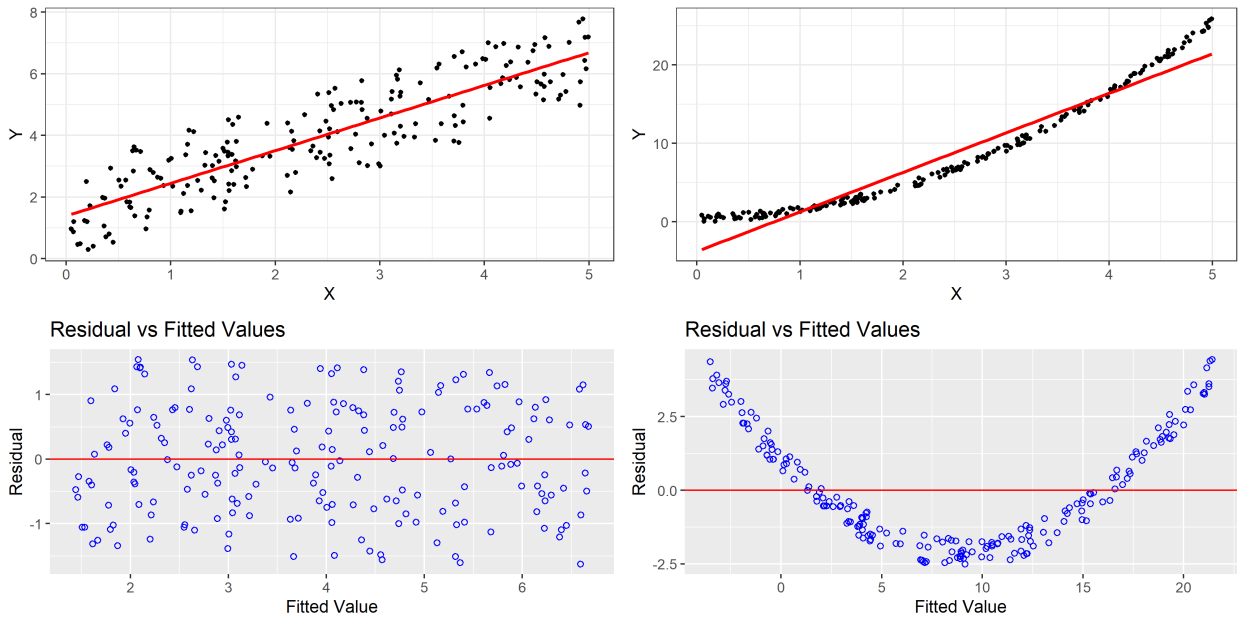



**Figure 8.15:** Illustrations of non-linearity using residuals and fitted values: linear case (left) and non-linear case (trend).

**Linearity** We plot the residuals $e_i$ against the prediction $\hat{Y}_i$: if the linearity assumption is warranted, the points should appear **randomly scattered about** 0.

The **absence** of a trend suggests that the relationship between $X_1, \ldots, X_p$ and $Y$ is indeed linear, the **presence** of a trend provides evidence against the linearity assumption, as we see in Figure 8.15.

There are also formal tests, such as the test for **lack of fit**:[38]

$$\begin{cases} H_0 : \mathrm{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_{p-1}\mathbf{x}_{p-1} \\ H_1 : H_0 \text{ is false} \end{cases}$$

38: The Ramsay RESET test is another such test, which we will not discuss, but which would be useful to know.

Let $\mathbf{W}^1 = (X_1^1, \ldots, X_{p-1}^1), \ldots, \mathbf{W}^c = (X_1^c, \ldots, X_{p-1}^c)$, be the $c$ **distinct** predictor levels.[39]

39: The $j$th level has $n_j$ observations $Y_{i,j}$.

Assume that $\mathrm{E}\{Y\}$ has a **functional dependency on** $X_1, \ldots, X_{p-1}$, and that the residuals are **independent** and follow a **normal distribution** $\mathcal{N}(0, \sigma^2)$, and that **at least one** of the $p-1$ predictor levels $X_k$ has **replicates**. Denote the **average observation** over the $j$th level by $\overline{Y}_j$, and write

$$\mathrm{SST}_j = \sum_i^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2.$$

The corresponding ANOVA table is

| source | SS | df | MS | $F^*$ |
|---|---|---|---|---|
| Regression | SSR | $p-1$ | SSR$/(p-1)$ | MSLF/MSPE |
| Error | SSE | $n-p$ | SSE$/(n-p)$ | |
| Lack of fit | SSLF | $c-p$ | SSLF$/(c-p)$ | |
| Pure Error | SSPE | $n-c$ | SSPE$/(n-c)$ | |
| Total | SST | $n-1$ | | |

Recall that SST = SSE + SSR. We further partition SSE = SSPE + SSLF, where

$$\text{SSPE} = \sum_{j=1}^{c} \text{SST}_j$$

so that

$$\frac{\text{SSPE}}{\sigma^2} \sim \chi^2\left(\sum_{j=1}^{c}(n_j - 1)\right) = \chi^2(n-c).$$

Thus, according to **Cochran's Theorem**, when $H_0$ holds, we have

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p), \quad \frac{\text{SSLF}}{\sigma^2} \sim \chi^2(c-p),$$

and

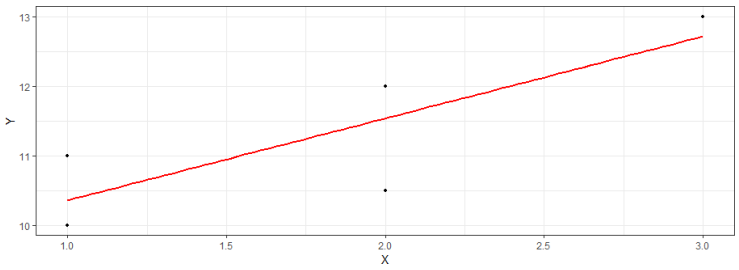
$$F^* = \frac{\left(\frac{\text{SSLF}}{\sigma^2}\right)\Big/(c-p)}{\left(\frac{\text{SSPE}}{\sigma^2}\right)\Big/(n-c)} \sim F(c-p, n-c).$$

**Decision Rule:** If $F^* > F(1-\alpha; c-p, n-c)$, we reject $H_0$ at a significance level of $\alpha$.

**Example** Consider a dataset with the following $(X, Y)$ observations

$$(1,10), (1,11), (2,10.5), (2,12), (3,13).$$

Is the linear model $\text{E}\{Y\} = \beta_0 + \beta_1 X$ warranted? We have $n = 5$, $p = 2$, and $c = 3$. The OLS framework yields $Y = 9.18 + 1.18X$, and the scatterplot is shown below.




Visually, it does seem that the line would be a good model, but it is difficult to say with certainty since there are so few points in the chart. We use the formal test for lack of fitness: we have

$$\text{SST} = S_{yy} = 5.8, \quad \text{SSR} = b_1^2 S_{xx} = 3.8829, \quad \text{SSE} = \text{SST} - \text{SSR} = 1.91071,$$
$$\text{SSPE} = \text{SST}_1 + \text{SST}_2 + \text{SST}_3 = 0.5 + 1.125 + 0 = 1.625,$$
$$\text{SSLF} = \text{SSE} - \text{SSPE} = 1.91071 - 1.625 = 0.28751,$$
$$\text{MSLF} = \frac{\text{SSLF}}{c-p} = \frac{0.28571}{3-2} = 0.28571, \quad \text{MSPE} = \frac{\text{SSPE}}{n-c} = \frac{1.625}{5-3} = 0.8125,$$

so that

$$F^* = \frac{\text{MSLF}}{\text{MSPE}} = \frac{0.28571}{0.8125} = 0.3516.$$

Since the critical value of the $F(3-2, 5-3) = F(1,2)$ distribution at $\alpha = 0.05$ is 18.5 , we **do not reject** the hypothesis of linearity.

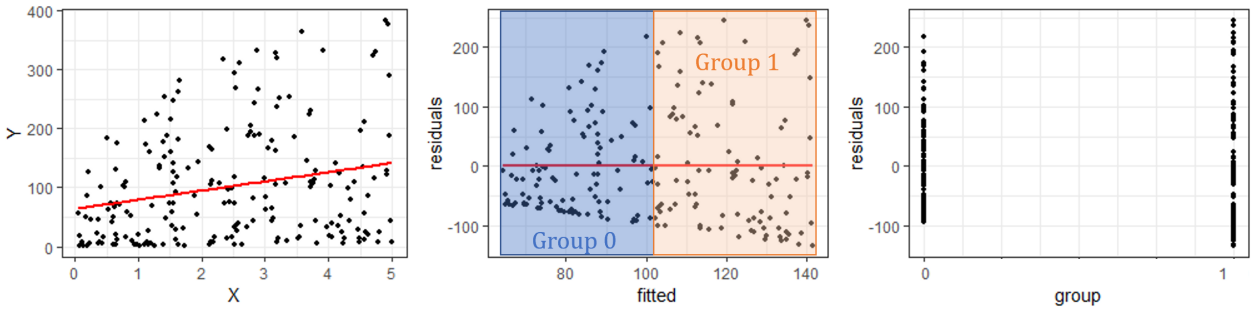



**Figure 8.16:** Illustration of the Brown-Forsythe test: original data and linear model (left), residuals against fitted values (middle), and deviations of residuals by group (right).

**Homoscedasticity** We can use residual plots to determine whether the condition of homoscedasticity is met or not. But there are **formal tests** as well, such as the **Brown-Forsythe** test, which is robust against departures from normality.[40]

40: Anotehr useful alternative is the **Breusch-Pagan** test, which requires normality of the residuals. It is worth looking up.

Let us take a look at the latter. Select a threshold $a \in \mathbb{R}$ and **partition** the residuals into 2 groups:

$$\text{Group 0: } \hat{Y} \leq a \text{ (the } e_{i,0}\text{'s)} \quad \text{vs.} \quad \text{Group 1: } \hat{Y} > a \text{ (the } e_{i,1}\text{'s).}$$

We pick $a$ so that $|\text{Group 0}| = n_0 \approx n_1 = |\text{Group 1}|$. Let $\tilde{e}_j$ be the **median residual of group** $j$ and let $d_{ij} = |e_{ij} - \tilde{e}_j|$ be the **absolute deviation of the $i$th residual in group $j$ from** $\tilde{e}_j$, for $j = 0, 1$.[41]

41: We use this framework rather than using the **mean** and the **square deviation** because of sensitivity to outliers – it is this choice that makes the test robust against departures from the normality assumption.

Set $\overline{d}_j = \dfrac{1}{n_j}\displaystyle\sum_i^{n_j} d_{ij}$, $j = 0, 1$. In order to test for

$$\begin{cases} H_0 : \overline{d}_0 = \overline{d}_1 & \text{(the variance is constant)} \\ H_1 : \overline{d}_0 \neq \overline{d}_1 & \text{(the variance is \textbf{not} constant)} \end{cases}$$

we compute the test statistic

$$t^*_{\text{BF}} = \frac{\overline{d}_0 - \overline{d}_1}{s_p\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

where

$$s_p^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n_0}(d_{i,0} - \overline{d}_0)^2 + \sum_{i=1}^{n_1}(d_{i,1} - \overline{d}_1)^2\right] = \frac{(n_0-1)s_0^2 + (n_1-1)s_1^2}{n_0 + n_1 - 2}$$

is the **pooled variance**. When $H_0$ holds, $t^*_{\text{BF}} \sim t(n_0 + n_1 - 2) = t(n-2)$.

**Decision Rule:** If $|t^*_{\text{BF}}| > t(1-\alpha/2; n-2)$, we reject $H_0$ at $\alpha$.

**Example** In the data displayed in Figure 8.16, the median fitted value is $a = 101.5096$. Visually, the constant variance assumption does not seem to be met.

We divide the datasets into two groups, based on whether the fitted value falls below $a$ (Group 0, in blue) or not (Group 1, in orange); there are $n_0 = n_1 = 100$ observations in each group.

The group median residuals are $\tilde{e}_0 = -15.6$, $\tilde{e}_1 = -22.9$. The mean and variance of the absolute deviations of the residuals to the median in each group are $\overline{d}_0 = 59.1$, $s^2{}_0 = 2197.745$, and $\overline{d}_1 = 86.3$, $s_0^1 = 4783.501$, respectively, which yield the pooled variance $s_p^2 = 3490.623$.

The BF test statistic is $t^*_{\text{BF}} = -3.21$; since

$$|t^*_{\text{BF}}| = 3.21 > \ t(0.975; 198) = 1.97,$$

we **reject** $H_0$ (equal variance) at significance level $\alpha = 0.05$.

**Independence** Independence of the error terms can be gauged visually by plotting the **residuals** $e_i$ against the **fitted values** $\hat{Y}_i$.

If the errors are **independent**, the correlation between these should be small ($|\rho| \approx 0$ ); if a pattern or a trend emerges, then they are likely **dependent**. The residuals vs. fitted values chart of the previous example shows a **slight** pattern, for instance, but the correlation is so **small** ($\rho = -6 \times 10^{-18}$) that we can reasonably treat them as **independent**.[42]

Other tests may be appropriate, depending on the nature of the data and model.[43]

42: The general linear regression assumption is that the **errors** are independent, but we only ever work with the **residuals**, which are definitely **not independent** ($\overline{e} = 0$).

43: For instance, the **Durbin-Watson** test for auto-correlation in the residuals of time series models (see Chapter 9).

**Normality** If the error terms are $\mathcal{N}(0, \sigma^2)$, we expect the residuals to also be $\mathcal{N}(0, \sigma^2)$. Thus, if the histogram of the **studentized residuals**

$$r_i = \frac{e_i}{\text{s}\{e_i\}} = \frac{Y_i - \hat{Y}_i}{\sqrt{\text{MSE}}\sqrt{1 - h_{ii}}}$$

is not symmetrical, then they do not follow a standard normal distribution $\mathcal{N}(0, 1)$ and the error terms are unlikely to be normal.

If the histogram is symmetrical, we build the **normal probability** plot from the **studentized residuals**.[44] For each $i = 1, \ldots, n$, we construct the following table:

44: Also known as **quantile-quantile** plot, or $qq-$plot.

| $i$ | **studentized residual** | **rank** | **percentile** | $z-$**quantile** |
|---|---|---|---|---|
| 1 | $r_1$ | $k_1$ | $p_1$ | $z_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $r_i$ | $k_i$ | $p_i$ | $z_i$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $r_n$ | $k_n$ | $p_n$ | $z_n$ |

The **rank** $k_i$ is given in **increasing** order (ties use the average rank); the **approximate percentile** is

$$p_i = \frac{k_i - 0.375}{n + 0.25}, \quad \text{(blom plotting position);}$$

the **quantile** is $z_i = \Phi^{-1}(p_i)$, where $\Phi(z) = P(Z \leq z)$, $Z \sim \mathcal{N}(0, 1)$.

Next, we plot the studentized residuals $r_i$ against the quantiles $z_i$ – the points should fall randomly about the "**normal**" line, with no systematic trend away from it. If not, the errors are unlikely to be normal.

Finally, we compute the **correlation** $\rho$ between $r_i$ and $z_i$, $i = 1, \ldots, n$. In order to test for

$$\begin{cases} H_0: \text{ error terms are normally distributed} \\ H_1: \ H_0 \text{ is false} \end{cases}$$

we find the critical value $\rho_\alpha$ of the normal **probability plot correlation coefficient** (PPCC) for sample size $n$ at a significance level $\alpha$.[45]

45: Such as could be found here.

**Decision Rule:** If $\rho < \rho_\alpha$, we reject $H_0$ at significance level $\alpha$.

**Example** Consider a dataset with the following $(X, Y)$ observations

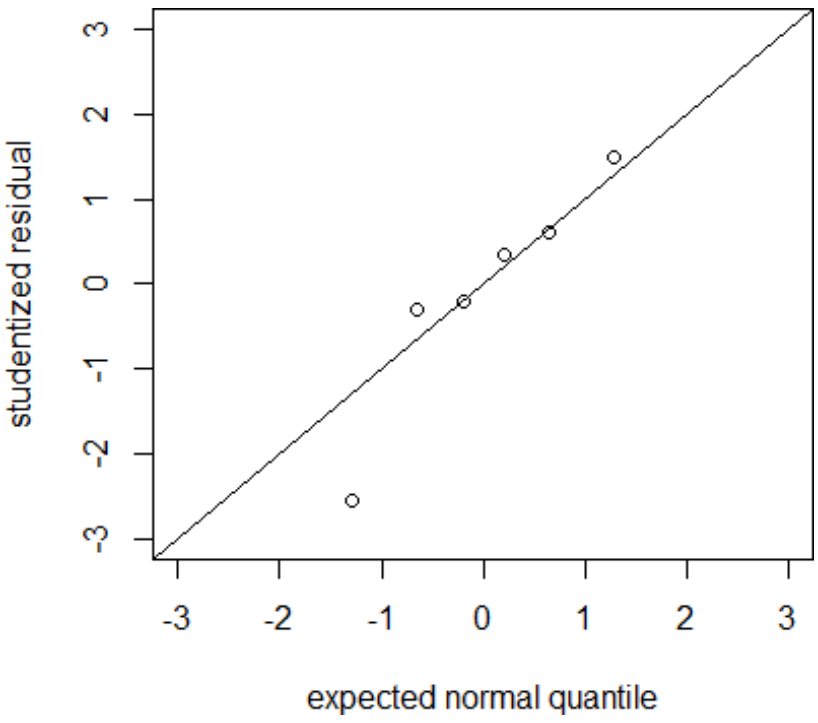
$$(1, 7.4), (1, 8.0), (2, 7.0), (2, 10.4), (3, 19.1), (4, 20.3).$$

Assume a linear model $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$. Is the normality assumption of the error terms warranted?

The linear model is $\mathrm{E}\{Y\} = 1.802 + 4.722X$; the table is

| $x$ | $y$ | **studentized residual** | **rank** | $p$ | $z-$**quantile** |
|---|---|---|---|---|---|
| 1 | 7.4 | 0.35 | 4 | 0.58 | 0.20 |
| 1 | 8.0 | 0.60 | 5 | 0.74 | 0.64 |
| 2 | 7.0 | -2.57 | 1 | 0.10 | -1.28 |
| 2 | 10.4 | -0.29 | 2 | 0.26 | -0.64 |
| 3 | 19.1 | 1.48 | 6 | 0.90 | 1.28 |
| 4 | 20.3 | -0.21 | 3 | 0.42 | -0.20 |

The $qq-$plot is shown below.



The correlation between the studentized residuals and the $z-$quantile is $\rho = 0.939$. At a significance level $\alpha = 0.05$, the critical value of the correlation in the PPCC table with $n = 6$ is 0.888, so we do not reject the normality assumption.[46]

46: Which, as we never tire of pointing out, is not the same as accepting $H_0$.

**Remedial Measures** **Transformations on** $X$ are used when the data exhibits a **monotone non-linear trend** with **variance constancy**; if the trend is increasing and concave down, we might try $X' = \ln X$ or $X' = \sqrt{X}$; if the trend is increasing and concave up, we might try $X' = e^X$ or $X' = X^2$; if it is decreasing and concave up, we might try $X' = \frac{1}{X}$ or $X' = e^{-X}$; if it is decreasing and concave down, we might try $X' = e^{-X^2}$.

**Transformations on** $Y$ are used when the data exhibits **monotone non-linear trend** with **NO variance constancy**, but it is often hard to determine from the scatter plots which transformation on $Y$ is best. The **Box-Cox** transformation helps us find a power $\lambda$ which will be appropriate for the regression model

$$Y_i^{(\lambda)} = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{X}_i$ is the $i$th row of $\mathbf{X}$. Set

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

We pick the $\lambda$ that minimizes the SSE$(\lambda)$ resulting from the regressions.

**Weighted Least Squares** are used if the data exhibits a **linear trend** with **no variance constancy**. An alternative would be to first use a transformation on $Y$ to control the **variance**, and then a transformation on $X$ to control the **linearity** that may have been destroyed by the first transformation.[47]

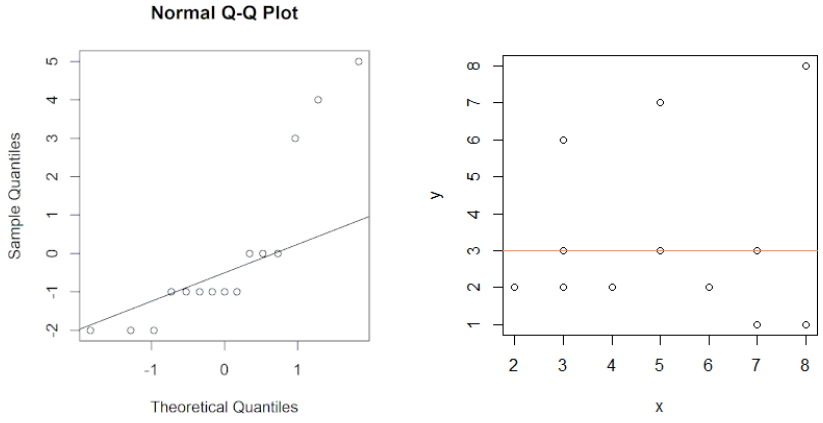
47: We will discuss this further in Section 8.4.5.

**Example** Consider the following dataset

$$(7,1), (7,1), (8,1), (3,2), (2,2), (4,2), (4,2), (6,2),$$
$$(6,2), (7,3), (5,3), (3,3), (3,6), (5,7), (8,8).$$[48]

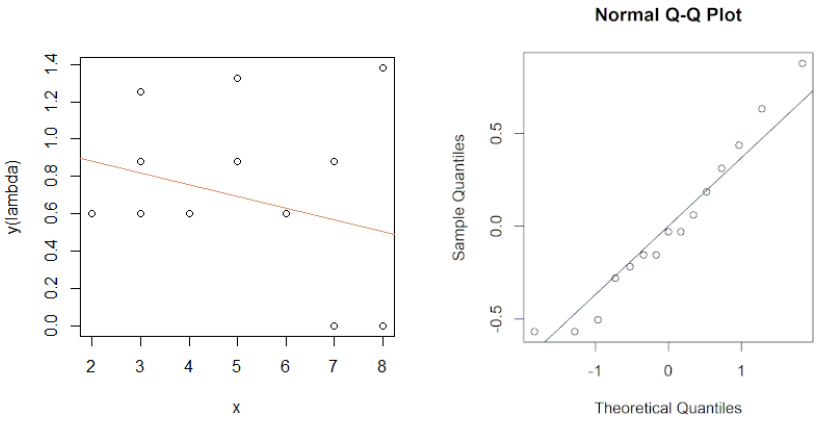
48: This example was found online, at a location that we cannot remember, unfortunately.

The scatterplot, regression line, and normal $qq-$plot are shown below.



The $qq-$plot shows that the error terms are unlikely to be normal, and so the regression model is not valid. The variance is not constant, so we use the Box-Cox transformation on $Y$: the optimal $\lambda$ is -0.42.

The scatterplot, regression line, and normal $qq-$plot on the transformed data are shown below.




**IMPORTANT:** the linear model on the original data is $\mathrm{E}\{Y\} = 3 + 0 \cdot X$. The linear model on the transformed data is

$$\mathrm{E}\left\{Y^{(-0.42)}\right\} = 1.00564 - 0.06264X$$

$\Longrightarrow$

$$\begin{aligned}\mathrm{E}\{Y\} &= \left([\lambda\beta_0 + 1] + \lambda\beta_1 X\right)^{1/\lambda} \\ &= \left([-0.42(1.00564) + 1] + 0.42 \cdot 0.06264X\right)^{1/(-0.42)} \\ &= \frac{1}{(0.5776 + 0.0263X)^{2.380}}\end{aligned}$$

which is **NOT** a straight line in the $xy-$plane.

## 8.4 Extensions of the OLS Model

We have seen that we can fairly easily extend simple linear regression to multiple linear regression with minimal disruption, simply by using the appropriate matrix notation. In practice, the multiple linear regression assumptions are rarely met; we have also presented ways in which we can identify departures from the assumptions, and how we can remedy this situation.

In this chapter, we will discuss more sophisticated extensions of linear regression, extensions that get closer to real-life applications.

### 8.4.1 Multicollinearity

The multiple linear regression **normal equations** are

$$(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{Y}.$$

When $\mathbf{X}^\top\mathbf{X}$ is **invertible**, the solution $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ is **unique**. If one of the variables is a non-trivial linear combination of other variables

$$X_k = \alpha_{j_1} X_{j_1} + \cdots + \alpha_{j_\ell} X_{j_\ell},$$

then $\text{rank}(\mathbf{X}^\top) = \text{rank}(\mathbf{X}^\top\mathbf{X}) < p$ and so $\mathbf{X}^\top\mathbf{X}$ is **singular** (not invertible), and the solution is not **unique** (the system in **under-determined**).

**Example** Consider the design matrix and vector response

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 3 \\ 1 & 3 & 3 & 6 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix}.$$

Find the OLS model $\text{E}\{Y \mid (X_1, X_2, X_3)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

We compute the constituents of the normal equations

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 3 & 5 & 6 & 11 \\ 5 & 11 & 12 & 23 \\ 6 & 12 & 14 & 26 \\ 11 & 23 & 26 & 49 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} 5 \\ 13 \\ 14 \\ 27 \end{pmatrix}.$$

The row echelon form of $[\mathbf{X}^\top\mathbf{X} \mid \mathbf{X}^\top\mathbf{Y}]$ is

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & -2 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

meaning that $\mathbf{b} = (-2, 1 - s, 1 - s, s)$ is an OLS solution for all $s \in \mathbb{R}$. More problematically, we cannot compute the corresponding variance-covariance matrix $\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$. □

In practice, it is quite rate that a predictor is an **exact** linear combination of other predictors; when it is almost so, however, the design matrix may be nearly **singular** (**ill-conditioned**),[49] leading to **uncertainty** in the parameter vector $\mathbf{b}$ that solves the normal equations.[50]

In multiple linear regression, the **variance inflation factor** for $\beta_k$ is

$$\text{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p,$$

where $R_k^2$ is the coefficient of multiple determination obtained when $X_k$ is regressed on the other $p - 2$ predictor variables in the model.[51]

Note that if $X_k$ is **very nearly** a linear combination of the other predictors, then $R_k^2 \approx 1$, yielding a **large** $\text{VIF}_k$, which influence the least-squares estimates. In practice, $\max_k \text{VIF}_k > 10$ implies that there are likely crucial problems with multicollinearity.

Remedial measures include **centering the data**, **ridge regression**, and **principal component regression**.[52]

49: See Chapter 4.

50: This is also the main cause of the "**wrong coefficient sign**" problem, when a coefficient takes on the opposite sign of what is expected based on a first-principle understanding of the situation.

51: Strictly speaking, this is not quite the definition of the variance inflation factor, but it will do for the purpose of these notes.

52: The latter two of these are discussed in Chapter 20.

**Example** Consider the following dataset

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|
| 1 | 1 | 2.063 | 1 | 2.995 |
| 2 | 1 | 3.184 | 1 | 3.773 |
| 1 | 1 | 2.131 | 2 | 2.846 |
| 2 | 1 | 2.867 | 2 | 3.963 |
| 1 | 2 | 3.104 | 1 | 5.291 |
| 2 | 2 | 3.876 | 1 | 6.070 |
| 1 | 2 | 2.999 | 2 | 5.034 |
| 2 | 2 | 3.865 | 2 | 6.014 |

Compare the linear models

$$\mathrm{E}\{Y \mid (X_1, X_2, X_3)\} \quad \text{and} \quad \mathrm{E}\{Y \mid (X_1, X_2, X_4)\}.$$

We start by loading the data in R.

```
X1 = c(1,2,1,2,1,2,1,2); X2 = c(1,1,1,1,2,2,2,2)
X4 = c(1,1,2,2,1,1,2,2)
X3 = c(2.06, 3.18, 2.13, 2.87, 3.10, 3.88, 2.99, 3.87)
Y  = c(2.99, 3.77, 2.85, 3.96, 5.29, 6.07, 5.03, 6.01)
data = data.frame(X1,X2,X3,X4,Y)
```

We build and summarize the two models.

```
summary(lm(Y ~ X1 + X2 + X3, data=data))
summary(lm(Y ~ X1 + X2 + X4, data=data))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08738    0.25633  -0.341   0.7503
X1           1.15410    0.43564   2.649   0.0570 .
X2           2.45576    0.44809   5.481   0.0054 **
X3          -0.27536    0.48844  -0.564   0.6030

Residual standard error: 0.1237 on 4 degrees of freedom
Multiple R-squared:  0.9947,Adjusted R-squared:  0.9907
F-statistic: 248.9 on 3 and 4 DF,  p-value: 5.313e-05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08200    0.22295  -0.368 0.731659
X1           0.91350    0.08427  10.841 0.000411 ***
X2           2.20800    0.08427  26.203 1.26e-05 ***
X4          -0.06800    0.08427  -0.807 0.464935
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1192 on 4 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9913
F-statistic: 268.2 on 3 and 4 DF,  p-value: 4.579e-05
```

The estimated parameters $b_0$, $b_1$, and $b_2$ are **quite similar** in both models, but the standard errors are **starkingly different**; the confidence intervals in the second model are **much tighter** for $\beta_1$ and $\beta_2$ than they are in the first model.

Why is this? Note that $\text{VIF}_1 \approx \text{VIF}_2 \approx \text{VIF}_4 \approx 1$ in the second model,[53] , whereas $\text{VIF}_1 \approx \text{VIF}_2 \approx \text{VIF}_3 \approx 25$ in the first model. This should not come as a surprise, as $X_3$ is very nearly a linear combination of $X_1$ and $X_2$:

$$\|X_3 - X_1 - X_2\|_2^2 \approx 0.324,$$

whereas $\|X_1\|_2^2 \approx 4.47$, $\|X_2\|_2^2 \approx 4.47$, and $\|X_3\|_2^2 \approx 8.70$.

53: The predictors are linearly independent.

### 8.4.2 Polynomial Regression

In a dataset with a predictor $X$ and a response $Y$, both numerical, if the relationship between $X$ and $Y$ is **not linear**, we may consider transforming the data so that the relationship between $X'$ and $Y'$ is **so**, fitting a **linear OLS** model to these new variables, and inverting the results to obtain a relationship between the original $X$ and $Y$.

Another approach is to create a sequence of predictors

$$X_1 = X, \; X_2 = X^2, \quad \ldots, \quad X_k = X^k$$

and to treat the entire situation as a multiple linear regression model

$$\mathrm{E}\{Y \mid (X_1, \ldots, X_k)\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \beta_0 + \beta_1 X + \cdots + \beta_k X^k.$$

**Example** Fit the following data

| $X$ | 1 | 1 | 2 | 4 | 3 | 6 |
|---|---|---|---|---|---|---|
| $Y$ | 0.8 | 1.3 | 4.1 | 15.3 | 8.8 | 36 |

We can fit a linear model to the data as follows.

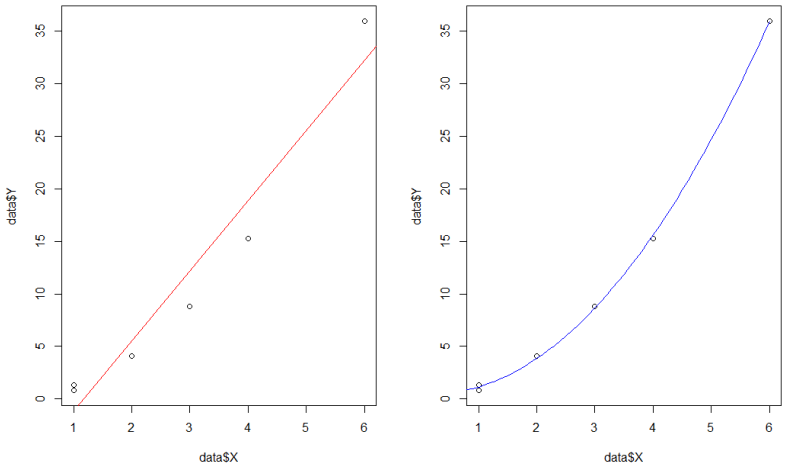
```
X = c(1,1,2,4,3,6)
Y = c(0.8,1.3,4.1,15.3,8.8,36)
summary(lm(Y ~ X))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.913      2.734  -2.895  0.04435 *
X              6.693      0.818   8.182  0.00122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.55 on 4 degrees of freedom
Multiple R-squared:  0.9436, Adjusted R-squared:  0.9295
F-statistic: 66.94 on 1 and 4 DF,  p-value: 0.001215
```

The fit seems decent ($R_a^2 = 0.9295$), but a plot of the data suggests that something is astray: visually, the quadratic fit seems better ($R_a^2 = 0.9994$).

```
X2 = X^2
Y = c(0.8,1.3,4.1,15.3,8.8,36)
summary(lm(Y ~ X + X2))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.56635    0.47768   1.186 0.321128
X           -0.49591    0.34935  -1.420 0.250809
X2           1.06466    0.05046  21.101 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```




One thing we notice is that of the three coefficients, only the quadratic $b_2$ is significant at $\alpha = 0.05$, even though the fit seemed **quite tight**, visually. Part of the problem is that although the relationship between $X$ and $X^2$ is **not linear**, the predictors are still **correlated**, leading to a fairly high VIF term:

$$\text{VIF}_1 = \frac{1}{1 - R_1^2} = \frac{1}{1 - 0.9510685} = 20.43673. \qquad \square$$

This is typical of polynomial regression: the suggested remedial measure is to use **centered predictors** $x_i = X_i - \overline{X}$.

**Example** The quadratic fit of the previous example could also be written as:

$$\begin{aligned} \mathrm{E}\{Y\} &= \gamma_0 + \gamma_1(X - \overline{X}) + \gamma_2(X - \overline{X})^2 \\ &= \left\{\gamma_0 - \gamma_1\overline{X} + \gamma_2\overline{X}^2\right\} + \left\{\gamma_1 - 2\gamma_2\overline{X}\right\} X + \gamma_2 X^2 = \beta_0' + \beta_1' X + \beta_2' X^2 \end{aligned}$$

but now **all** coefficients are significant at $\alpha = 0.05$.

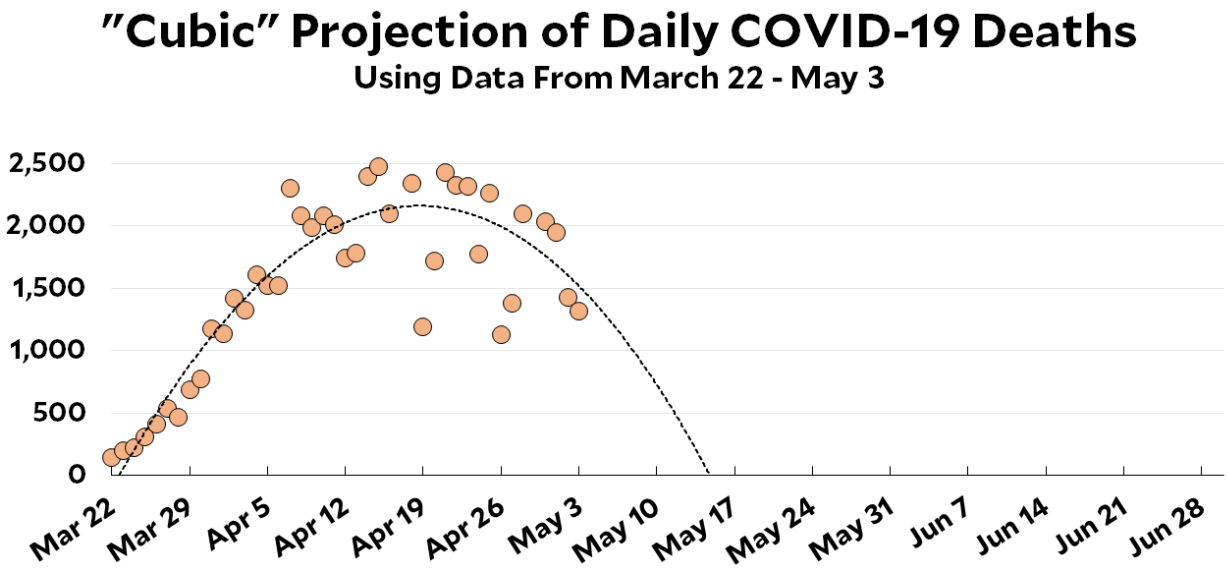



**Figure 8.17:** The White House projections for COVID-19 deaths used a cubic polynomial regression certainly fit the available data (March 22-May 3, 2020); the predicted end of the pandemic by May 16, 2020 did not survive the test of time, however, as no epidemiological domain expertise was brought to bear on the problem, with dire consequences of the United States [author unknown].

```
Xm = X - mean(X)
X2m = Xm^2
summary(lm(Y ~ Xm + X2m))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.70814    0.20935   36.82 4.41e-05 ***
Xm           5.53718    0.09472   58.46 1.10e-05 ***
X2m          1.06466    0.05046   21.10 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```

Note that the centered $\text{VIF}_1$ is much lower at $(1-0.3344)^2 \approx 1.5$.

```
summary(lm(X2m ~ Xm))
```

```
Residual standard error: 3.323 on 4 degrees of freedom
Multiple R-squared:  0.3344,Adjusted R-squared:  0.168
F-statistic: 2.009 on 1 and 4 DF,  p-value: 0.2293
```

The rest of the ordinary least square machinery easily carries over. □

Graphically and/or mathematically, polynomial regression can prove quite powerful and convenient to use. But convenience is not always a sufficient reason to use a regression model.[54]

54: For a modern example, consider the White House prediction in the early days of the COVID-19 pandemic (see Figure 8.17).

### 8.4.3 Interaction Effects

We have seen that we can extend simple linear regression in $X$ to include higher power terms.[55]

55: After centering the data to minimize the effects of multicolinearity.

There is nothing to stop us from doing so with any number of predictors $X_1, \ldots, X_p$, leading to an **additive model**

$$E\{Y\} = f_1(X_1) + \cdots + f_p(X_p),$$

where the $f_i$ are **polynomial functions** in 1 variable.[56] In what follows, we assume that $p = 2$ to keep things simple.

56: This could be modified to any linear function of the regression coefficients $\beta_{i,j}$.

We can refine the model with an **interaction term** $f_3(X_1, X_2) = \beta_3 X_1 X_2$. In keeping with the **hierarchical principle**, we might consider the model

$$\begin{aligned} \mathrm{E}\{Y\} &= f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) \\ &= \beta_0 + \beta_{1,1} X_1 + \beta_{2,1} X_2 + \beta_{1,2} X_1^2 + \beta_3 X_1 X_2 + \beta_{2,2} X_2^2, \end{aligned}$$

although there could also be good reasons to consider something like

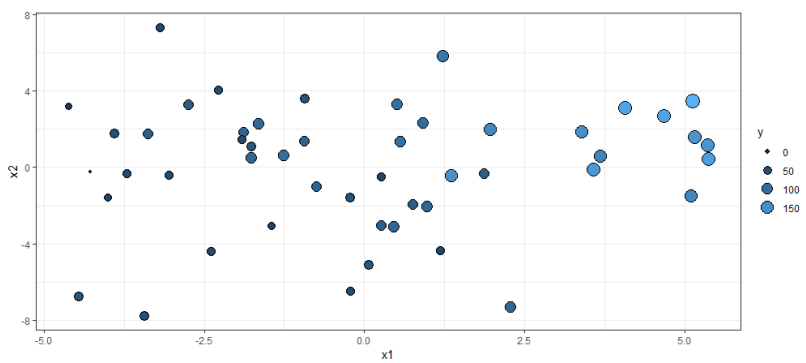
$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_1 X_2.$$

In the latter case, if we assume that $\beta_1 \beta_2 > 0$, then if $\beta_1 \beta_3 > 0$, we have a **reinforcement interaction**; if $\beta_1 \beta_3 < 0$, we have an **interference interaction**.

**Example** We consider a dataset of $n = 50$ observations with 2 centered predictors $X_1, X_2$ and a response $Y$.[57]

57: We do not specify a seed, so the results may vary slightly from one run to the next.

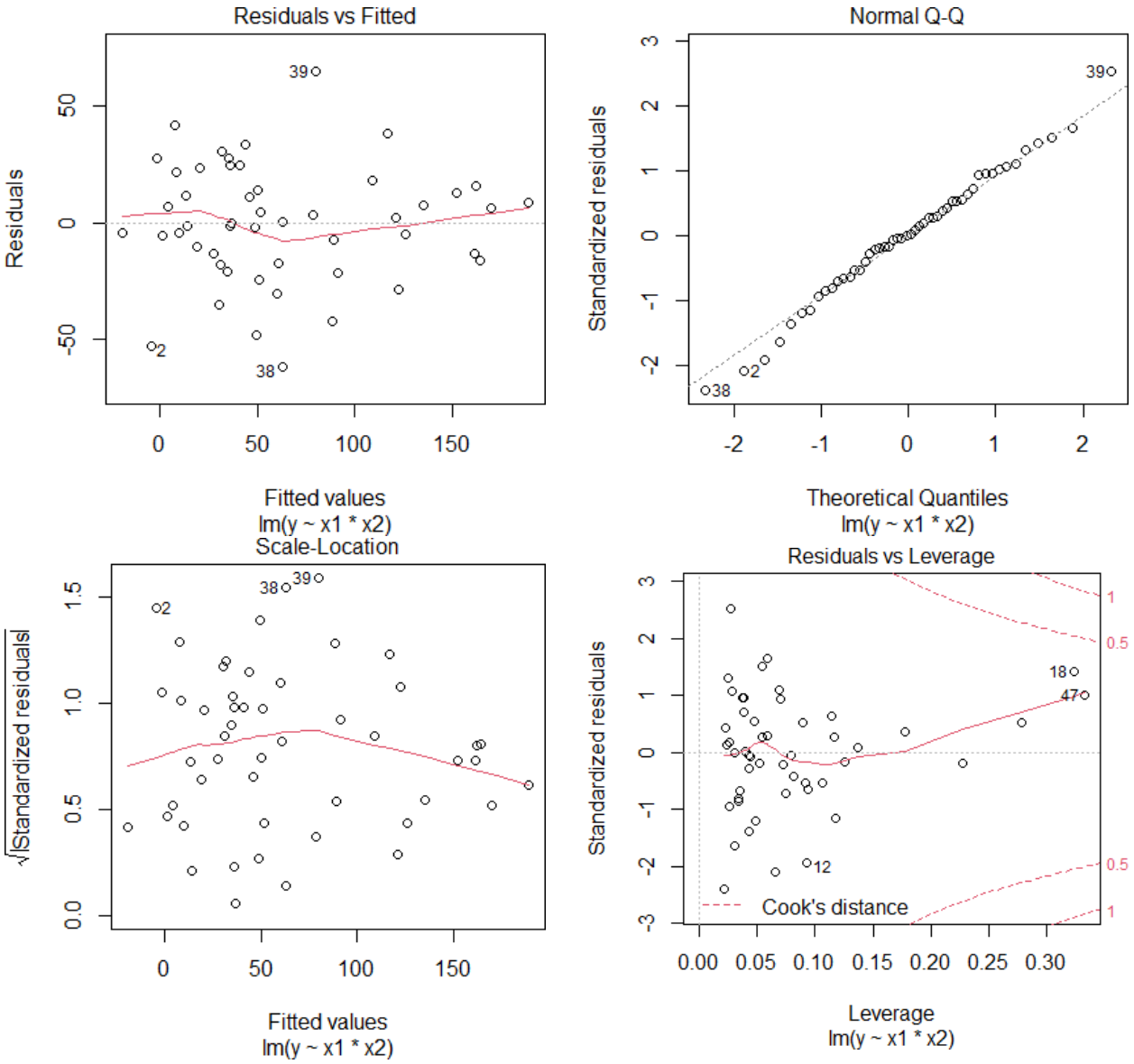
```
x1 <- runif(50, 0, 10); x2 <- rnorm(50, 10, 3)
modmat <- model.matrix(~x1*x2, data.frame(x1=x1, x2=x2))
coeff <- c(1, 2, -1, 1.5)
y <- rnorm(50, mean = modmat %*% coeff, sd = 25)
dat <- data.frame(y = y, x1 = x1, x2 = x2)
dat2 = dat
dat2[,c(2:3)] <- scale(dat[,c(2:3)], scale=FALSE)

library(ggplot2)
ggplot(dat2,aes(x=x1,y=x2,fill=y,size=y)) + theme_bw() +
    geom_point(pch=21) + theme_bw()
```



We compute the fit for the reduced and the full interaction models. The former exhibits reinforcement interaction ($\beta_1\beta_3 > 0$).

```
summary(lm(y ~ x1 * x2, data=dat2))
plot(lm(y ~ x1 * x2, data=dat2))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.7494     3.7043  16.669  < 2e-16 ***
x1           15.6463     1.3017  12.020 8.55e-16 ***
x2            5.1396     1.2010   4.279 9.40e-05 ***
x1:x2         1.6886     0.4379   3.856 0.000356 ***

Residual standard error: 26.06 on 46 degrees of freedom
Multiple R-squared:  0.8166,Adjusted R-squared:  0.8047
F-statistic: 68.28 on 3 and 46 DF,  p-value: < 2.2e-16
```

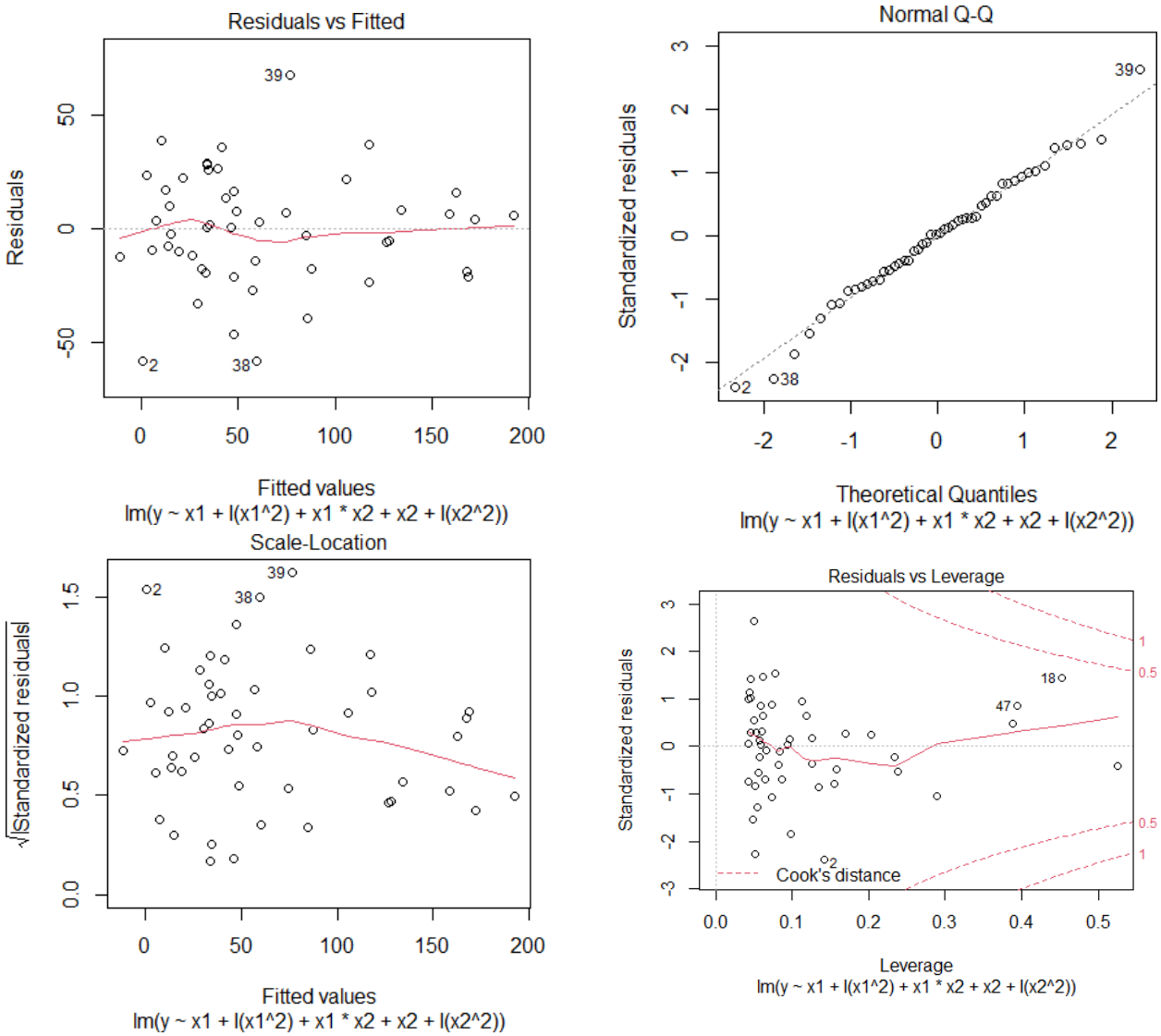


The summary indicates that the reduced interaction linear model is appropriate, which is supported by the diagnostic plots. But what about the full model? The pure quadratic terms are not significant, which suggests that the reduced model is likely a better choice.[58]

58: Although not necessarily so.

```
summary(lm(y ~ x1+I(x1^2)+x1*x2+x2+I(x2^2), data=dat2))
plot(lm(y ~ x1+I(x1^2)+x1*x2+x2+I(x2^2), data=dat2))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.25684    5.94511   9.799 1.24e-12 ***
x1          15.36026    1.38371  11.101 2.42e-14 ***
I(x1^2)      0.41459    0.46486   0.892 0.377316
x2           4.91100    1.31831   3.725 0.000553 ***
I(x2^2)      0.01042    0.26562   0.039 0.968891
x1:x2        1.56368    0.46519   3.361 0.001613 **

Residual standard error: 26.4 on 44 degrees of freedom
Multiple R-squared:  0.8199,Adjusted R-squared:  0.7994
F-statistic: 40.06 on 5 and 44 DF,  p-value: 2.654e-15
```



### 8.4.4 ANOVA/ANCOVA for Categorical Variables

We can also include categorical variables within the OLS framework. Suppose there are $K$ treatments (levels) for predictor $X$.

In the **dummy variable** encoding, we set

$$X_j = \begin{cases} 1 & \text{treatment } j \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, K-1$. The ANOVA/OLS model is then

$$Y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{i,j} + \varepsilon_i \quad \text{and} \quad E\{Y\} = \begin{cases} \beta_0 & \text{treatment } K \\ \beta_0 + \beta_j & \text{treatment } j \end{cases}$$

In the **treatment effect** encoding, we set

$$X_j = \begin{cases} 1 & \text{treatment } j \\ -1 & \text{treatment } K \\ 0 & \text{else} \end{cases}$$

for $j = 1, \ldots, K-1$. The ANOVA/OLS model is as in the dummy encoding case and

$$E\{Y\} = \begin{cases} \beta_0 - (\beta_1 + \cdots + \beta_{K-1}) & \text{treatment } K \\ \beta_0 + \beta_j & \text{treatment } j \end{cases}$$

We will have more to say on the topic in Chapter 11.

### 8.4.5 Weighted Least Squares

We have seen that the OLS regression model $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ requires **constant variance**. When that assumption is not met – but in a "monotonic" manner, such as $\sigma^2\{\varepsilon_i\} = \sigma^2 x_i$, say – various data transformations on the predictors $X$ may be appropriate.

What do we do when the linearity assumption is valid, but the variance $\sigma_i$ does not change in a **systematic** manner?

One way to approach the problem is *via* **weighted least squares** (WLS), which does not require all observations to be **treated equally**, that is to say, to be given the **same weight**.

Let $w_i \geq 0$ be the weight of observation $i$ and write $Z_i = \sqrt{w_i} Y_i$. Define the **weight matrix** as $\mathbf{W} = \text{diag}(w_1, \ldots, w_n)$.

The **WLS problem** is to find the coefficient vector $\boldsymbol{\beta}$ which **minimizes** the weighted sum of squared errors

$$\begin{aligned} \text{SSE}_w = Q_w(\boldsymbol{\beta}) &= \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \\ &= \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\hat{\mathbf{Y}}\|_2^2 = \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^\top \mathbf{W}\mathbf{Y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{Y} - \mathbf{Y}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

But $\nabla_\beta Q_w(\boldsymbol{\beta}) = -2\mathbf{X}^\top\mathbf{W}\mathbf{Y} + 2\mathbf{X}^\top\mathbf{W}\mathbf{X}\boldsymbol{\beta}$, so the WLS estimator $\mathbf{b}_W$ of $\boldsymbol{\beta}$ is

$$\nabla_\beta Q_w(\boldsymbol{\beta}) = \mathbf{0} \implies \mathbf{b}_W = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{Y}.$$

The entire OLS machinery can then be used in the WLS context simply by replacing $\mathbf{Y}$ by $\sqrt{\mathbf{W}}\mathbf{Y}$ and $\mathbf{X}$ by $\sqrt{\mathbf{W}}\mathbf{X}$ throughout.

**Example** Consider a dataset with $n = 11$ observations:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.82 | 1.09 | 1.22 | 1.24 | 1.29 | 1.30 | 1.36 | 1.38 | 1.39 | 1.40 | 1.55 |
| $y$ | 1.47 | 1.33 | 1.32 | 1.30 | 1.35 | 1.34 | 1.38 | 1.52 | 1.40 | 1.44 | 1.58 |

We build the OLS model, a WLS model where the first observation has twice the weight of the other observations, and a OLS model without the first observation.[59]

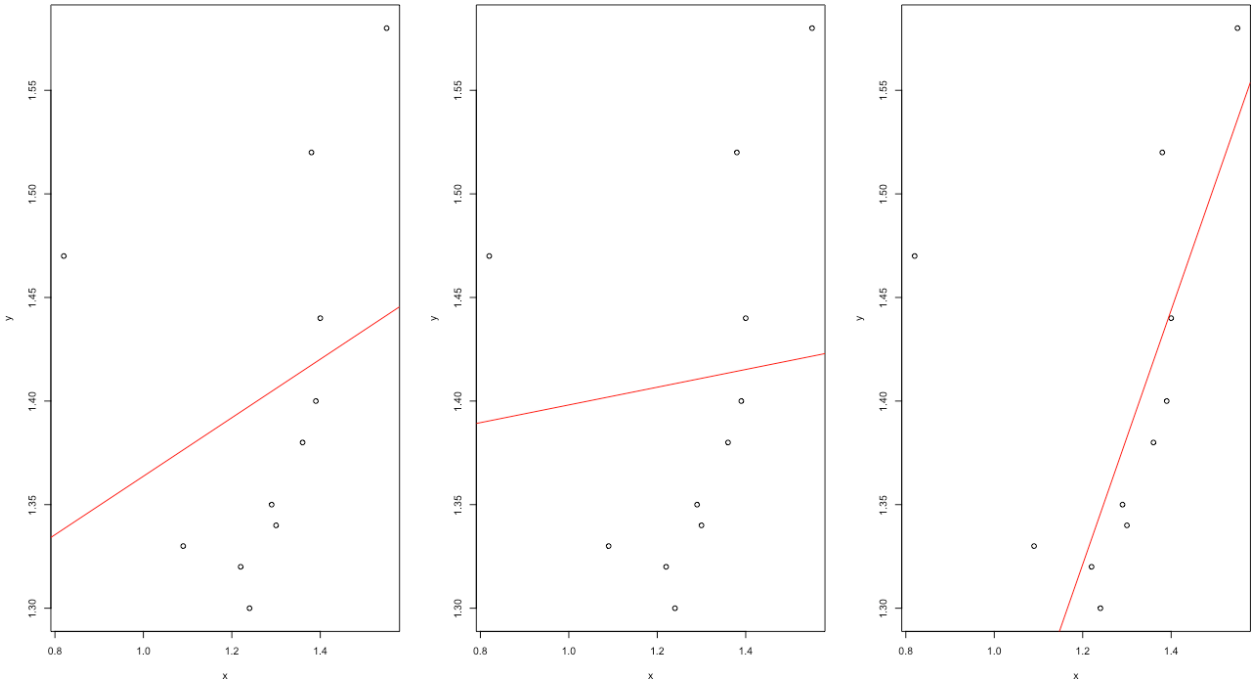
59: Which is equivalent to a WLS model with $w_1 = 0$ and $w_i = 1$ for $i > 1$.

```
x <- c(0.82,1.09,1.22,1.24,1.29,1.30,1.36,1.38,1.39,1.40,1.55)
y <- c(1.47,1.33,1.32,1.30,1.35,1.34,1.38,1.52,1.40,1.44,1.58)
mod.1 <- lm(y ~ x)
summary(mod.1)
mod.2 <- lm(y ~ x, weights = c(2,1,1,1,1,1,1,1,1,1,1))
summary(mod.2)
mod.3 <- lm(y ~ x, weights = c(0,1,1,1,1,1,1,1,1,1,1))
summary(mod.3)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2225     0.1920   6.366  0.00013 ***
x             0.1412     0.1489   0.948  0.36782

Residual standard error: 0.09047 on 9 degrees of freedom
Multiple R-squared:  0.09081, Adjusted R-squared:  -0.01021
F-statistic: 0.899 on 1 and 9 DF,  p-value: 0.3678

------------------------------------------------------------

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3553     0.1624   8.344 1.58e-05 ***
x             0.0428     0.1292   0.331    0.748

Residual standard error: 0.09669 on 9 degrees of freedom
Multiple R-squared:  0.01204, Adjusted R-squared:  -0.09773
F-statistic: 0.1097 on 1 and 9 DF,  p-value: 0.748

------------------------------------------------------------

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5848     0.1916   3.052   0.0158 *
x             0.6136     0.1444   4.250   0.0028 **

Residual standard error: 0.05402 on 8 degrees of freedom
Multiple R-squared:  0.693, Adjusted R-squared:  0.6546
F-statistic: 18.06 on 1 and 8 DF,  p-value: 0.002801
```

The OLS model is $\hat{y} = 1.223 + 0.1412x$ (left in the chart below), the WLS model with $w_1 = 2$ and $w_i = 1$, $i = 2, \ldots, 11$ is $\hat{y} = 1.3553 + 0.0428x$ (middle), and the OLS/WLS without the first observation is $\hat{y} = 0.5848 + 0.6136x$ (right). The plots are shown below.

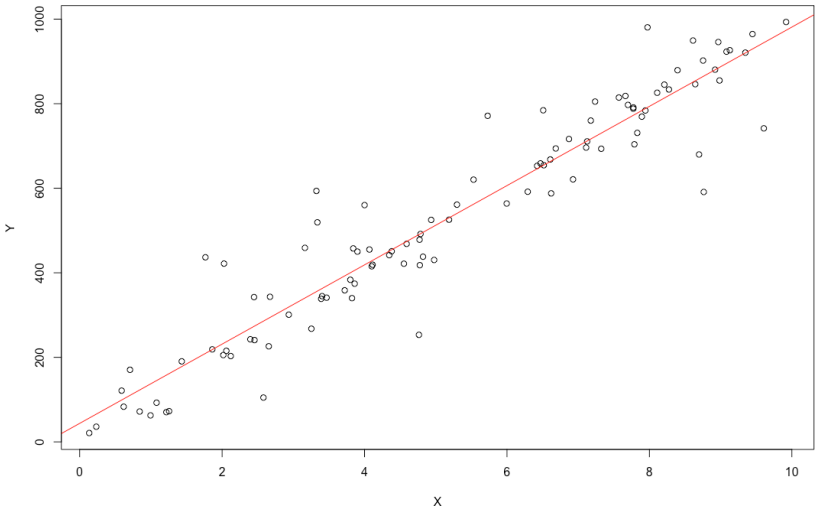
```
par(mfrow=c(1,3))
plot(x,y); abline(mod.1, col="red")
plot(x,y); abline(mod.2, col="red")
plot(x,y); abline(mod.3, col="red")
```




We can use WLS to deal with an error variance which is not constant. Consider the underlying model

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\{\boldsymbol{\varepsilon}\}\right), \quad \text{where} \quad \sigma^2\{\varepsilon_i\} = \sigma_i^2 \not\equiv \sigma^2,$$

such as may be found in the image below:



The procedure goes as in the OLS case, with some slight modifications:

1. if the $\sigma_i^2$ are known, we use the weights $w_i = \frac{1}{\sigma_i^2} \geq 0$;
2. if the $\sigma_i^2$ are unknown:
   a) we use OLS and find the residuals $e_i$;[60]
   b) depending on the choice made above, regress either $e_i^2$ or $|e_i|$ on $X_1, \ldots, X_{p-1}$ to obtain fitted values $\hat{v}_i$ or $\hat{s}_i$, which are point estimate of $\sigma_i^2$ or $\sigma_i$, respectively;
   c) depending on the choice made above, use WLS with $w_i = \frac{1}{\hat{v}_i}$ or $w_i = \frac{1}{\hat{s}_i^2}$ and compute $\text{SSE}_w$ and $\text{MSE}_w = \frac{\text{SSE}_w}{n-p}$. If $\text{MSE}_w \approx 1$, the scaling is **appropriate**; otherwise, repeat steps a) to c), starting with the current **WLS residuals**.

60: $e_i^2$ is an estimate of $\sigma_i^2$ when there are no $Y$-outliers, $|e_i|$ is an estimate of $\sigma_i$ when there are some.

**Example** The number of defective items $Y$ produced by a machine is known to be linearly related to the speed setting $X$ of the machine:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon_i \text{ indép.}$$

An analyst regresses the squared residuals $e_i^2 = (\hat{Y}_i - Y_i)^2$ on the speed setting $X_i$ and obtains the following $n = 12$ fitted values:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{v}_i$ | 68.7 | 317.4 | 193 | 317.4 | 68.7 | 193 | 193 | 317.4 | 68.7 | 317.4 | 68.7 | 193 |

Using weighted OLS with $w_i = \frac{1}{\hat{v}_i}$, her residuals are $e_i^w = \hat{Y}_i^w - Y_i$:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e_i$ | −3.6 | 5.6 | −13.5 | −16.4 | −9.6 | 7.5 | −10.5 | 26.6 | 14.4 | −17.4 | −1.6 | 18.5 |

Is her use of these weights appropriate?

We have

$$\text{SSE}_w = \sum_{i=1}^{12} w_i e_i^2 = \sum_{i=1}^{12} \frac{1}{\hat{v}_i} e_i^2 = 12.2953,$$

a sum of squares with $n - p = 12 - 2 = 10$ degrees of freedom, so that

$$\text{MSE}_w = \frac{\text{SSE}_w}{n-p} = \frac{12.2953}{10} = 1.22953.$$

Since $\text{MSE}_w \approx 1$, we have evidence that the weights are **appropriate** and that the initial $\hat{v}_i$ provide reasonable approximations of $\sigma_i^2$.

### 8.4.6 Other Extensions

The OLS assumptions are **convenient** from a mathematical perspective, but they are not always met in practice. One way out of this conundrum is to use **remedial measures** to transform the data into **compliant inputs**.

Another approach is to **extend/expand the assumptions** and to work out the corresponding mathematical formalism:

- **generalized linear models (GLM)** implement responses with **non-normal** conditional distributions (see Section 20.2.3);
- **classifiers**, such as logistic regression, decision trees, support vector machines, naïve Bayes methods, neural networks, etc., extend regression to **categorical responses** (see Chapter 21);
- **non-linear methods**, such as splines, generalized additive models (GAM), nearest neighbour methods, kernel smoothing methods, etc., are used for responses that are **not linear combinations of the predictors** (see Chapter 20);
- **tree-based methods** and **ensemble learning methods**, such as bagging, random forests, and boosting, are used to simplify the modeling of **predictor interactions** (see Chapter 21);
- **regularization methods**, such as ridge regression, the LASSO, and elastic nets, facilitate the process of **model selection** and **feature selection** (see Section 20).

**Model Selection** With reasonable real-world datasets and situations, we can often build tens (if not hundreds) of models related to a specific scenario.[61] When most of these models are "aligned" with one another, that is, when they yield similar results, picking the simplest model is a good approach.

61: Not necessarily models of the linear regression variety.

But in practice, we can also reach a point of **diminishing returns** – including more variables in the model might not yield better predictive power, due to the **curse of dimensionality**.

The problem of **model selection** is not easy to solve; we tackle it in earnest in Section 20.4 and in Chapter 23.

## 8.5 Outliers and Influential Observations

When we are working with a single predictor, we can usually tell quite quickly if a prediction or a response is unusual, in some sense.

If a predictor value is much smaller/much larger than the other predictor values, we might be hesitant to use the regression model to fit the value because no similar values were used to "train" the model. When $p > 1$, finding the anomalous observations (predictors and/or responses) is not as obvious.

We introduce a small number of methods to do so in this section; there are plenty more, which we will discuss in detail in Chapter 26.

### 8.5.1 Leverage and Hidden Extrapolation

Consider a dataset with two predictors $X_1$, $X_2$, as shown in Figure 8.18. Regression models are typically only useful when we are working within the **model scope**; if regression is an attempt to **interpolate** the data, then we must avoid situations where we are **extrapolating** from the data.

The problem is that we cannot always easily tell if a predictor $\mathbf{X}_h$ is in the model scope or not; in the previous image, each component of $\mathbf{X}_h$ is in

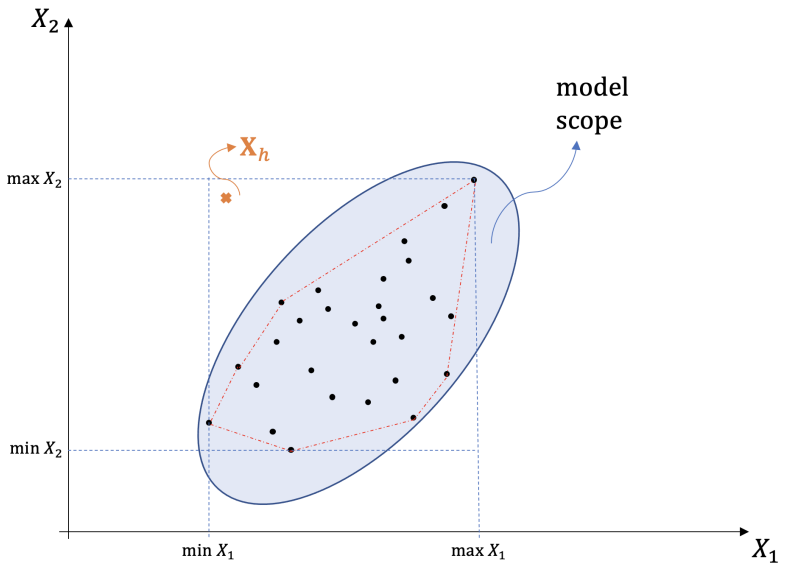



**Figure 8.18:** Model scope in two-dimensional predictor space (in blue); the predictor level $\mathbf{X}_h$ is out-of-scope.

the range of the predictors used to build the model, but $\mathbf{X}_h$ as a whole is **not**. When $p$ is large, this **visual** approach fails.

The **leverage of the $i$th case** is:

$$h_{ii} = \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i^\top, \quad \mathbf{X}_i \text{ is the } i\text{th row of } \mathbf{X};$$

in other words, $h_{ii}$ is the $i$th diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. The **leverage** determines if a predictor level $\mathbf{X}_h$ is in the **model scope**: if

$$\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h > \max\{h_{ii} \mid i = 1, \ldots, n\},$$

$\mathbf{X}_h$ is **outside the scope** and $\hat{Y}_h = \mathbf{X}_h\mathbf{b}$ contains a **hidden extrapolation**.

Note that $0 \leq h_{ii} \leq 1$, for $i = 1, \ldots, n$. Indeed, since:

1. $\mathbf{0} \leq \sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\{\mathbf{HY}\} = \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}^\top = \sigma^2\mathbf{H} \implies h_{ii} \geq 0$ for all $i$
2. $\mathbf{0} \leq \sigma^2\{\mathbf{e}\} = \sigma^2\{(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\} = \sigma^2(\mathbf{I}_n - \mathbf{H}) \implies 1 - h_{ii} \geq 0$ for all $i$

Generally-speaking, the surface of $\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h = c$ is an ellipsoid centred around

$$\overline{\mathbf{X}} = (1, \overline{X}_1, \ldots, \overline{X}_p).$$

The larger $c$, the larger the "distance" to $\overline{\mathbf{X}}$.

An $X-$**outlier** is an observation which is **atypical** with respect to the **predictor levels**.

We note that

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\text{trace}(\mathbf{H}) = \frac{p}{n} \quad (p \leq n);$$

1. if $h_{ii} \leq 0.2$, then the leverage of the $i$th case is **low** (very near $\overline{\mathbf{X}}$);
2. if $0.2 < h_{ii} < 0.5$, then the leverage is **moderate**;
3. if $h_{ii} \geq 0.5$, then the leverage is **high** (potential $X-$outlier);
4. when $n$ is large, if $h_{ii} > 3\overline{h} = \frac{3p}{n}$, then the $i$th case is an $X-$**outlier**.

**Example** We wish to fit the multiple linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

to a dataset with $n$ observations, with

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 1.17991 & -0.00731 & 0.00073 \\ -0.00731 & 0.00008 & -0.00012 \\ 0.00073 & -0.00012 & 0.00046 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 220 \\ 36768 \\ 9965 \end{pmatrix}$$

What are the point estimates for the regression coefficients $\boldsymbol{\beta}$? We would like to predict the value of $Y_h$ when $X_1 = 200$ and $X_2 = 50$, i.e., at the point $\mathbf{X}_h = (1, 200, 50)^\top$. What is the leverage of $\mathbf{X}_h$? Is this case of hidden extrapolation? If not, what is the predicted value $Y_h$?

The OLS estimates of the regression coefficients are

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -1.91943 \\ 0.13744 \\ 0.33234 \end{pmatrix}.$$

The leverage of $\mathbf{X}_h$ is

$$\mathbf{X}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_h = 0.27891;$$

it is small enough to suggest that we are not in a hidden extrapolation situation (although $n$ is unknown, so we cannot compare it against $\frac{3p}{n}$).

The predicted response at $\mathbf{X}_h$ is thus $\hat{Y}_h = \mathbf{X}_h^\top \mathbf{b} = 42.18557$.

## 8.5.2 Deleted Studentized Residuals

While $X-$outliers can be determined without reference to a **regression surface** $\hat{Y}(\mathbf{x}) = \mathbf{xb}$, we can also look for observations whose response values are **unexpectedly distant** from $\hat{Y}(\mathbf{x})$.

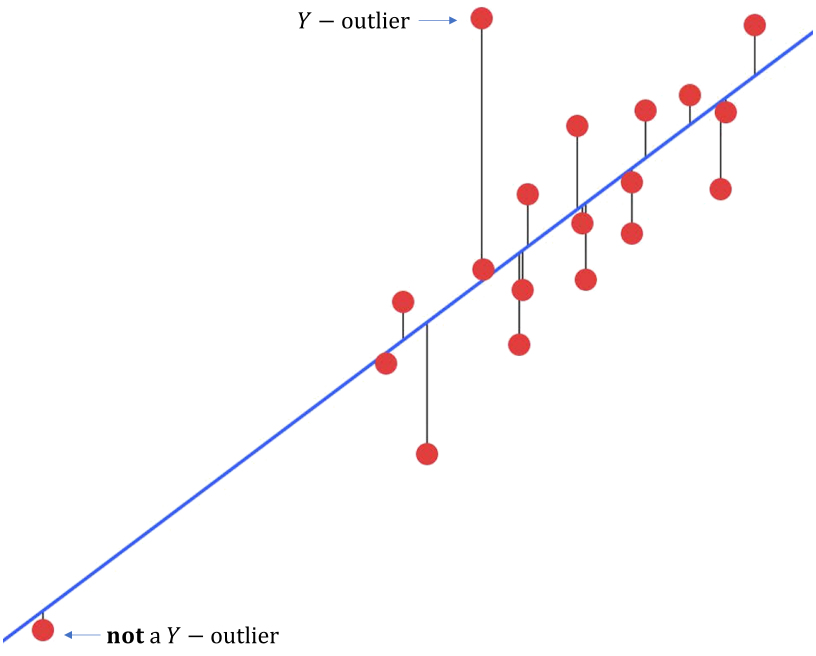



**Figure 8.19:** $X-$outlier and $Y-$outlier in an artificial dataset.

A $Y-$**outlier** is an observation which yields a **large** regression residual.

If the **(internal) studentized residual** is large enough,

$$|r_i| = \left|\frac{e_i}{\text{s}\{e_i\}}\right| = \left|\frac{e_i}{\sqrt{\text{MSE}}\sqrt{1-h_{ii}}}\right| \geq 3,$$

say, then the $i$th point is a $Y-$outlier.

Another approach is to **delete** the $i$th case from the model and refit

$$\mathbf{b}_{(i)} = \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_{(i)}^\top \mathbf{Y}_{(i)},$$

yielding an expected value for the $i$th case, $\hat{Y}_{i(i)}$.

For $i = 1, \ldots, n$, the **deleted residual** is $d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}$ and the **external studentization** is

$$t_i = \frac{d_i}{\text{s}\{d_i\}} = e_i \sqrt{\frac{n-p-1}{\text{SSE}(1-h_{ii}) - e_i^2}} \sim t(n-p-1),$$

where

$$\text{s}^2\{d_i\} = \text{MSE}_{(i)} \left[1 + \mathbf{X}_i \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_i^\top\right].$$

**Decision Rule:** if $|t_i| > t(1-\frac{\alpha/n}{2}; n-p-1)$, then the $i$th case is a $Y-$outlier at significance level $\alpha$.

Note that it is possible for an observation to be an $X-$outlier without being an $Y-$outlier, and *vice-versa* (see previous chart).

### 8.5.3 Influential Observations

In the regression context, we may also be interested in determining which observations are **influential** – observations whose absence from (or presence in) the data significantly change the **nature of the fit** (qualitatively).

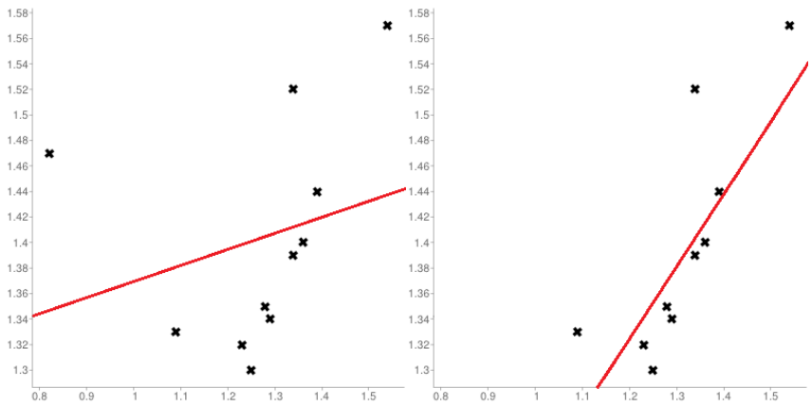



**Figure 8.20:** Influential observation in a dataset; the nature of the regression line changes drastically when the left-most observation is removed from the data.

Influential observations need not be outliers (but they may be!), and *vice-versa*.

For the $i$th case, $\text{DFFITS}_i$ is a measure of the **influence** of the $i$th case on the $\hat{Y}$ in a neighbourhood of $\mathbf{X}_i$. The **difference from the fitted value** is

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

For small and moderately-sized samples, if $|\text{DFFITS}_i| > 2$, then the $i$th case is **likely influential**. For larger samples, if $|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$, then the $i$th case is **influential**.

A similar measure can be determined to see if case $i$ has a lot of influence on the value of the **fitted parameter** $b_k$:

$$\text{DFBETAS}_i^k = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_{(i)} \left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]_{k,k}}}.$$

### 8.5.4 Cook's Distance

We can also use **Cook's distance** to measure observation $i$'s influence:

$$D_i = \frac{1}{p \cdot \text{MSE}} \sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2 = \frac{e_i^2}{p \cdot \text{MSE}} \left[\frac{h_{ii}}{(1 - h_{ii})^2}\right] \sim F(p, n - p).$$

**Decision Rule:**

- if $D_i < F(0.2; p; n - p)$, then the $i$th case **has little influence**;
- if $D_i > F(0.5; p; n - p)$, then the $i$th case **is very influential**.

Regressions based on OLS framework are convenient, but they are not **robust** against outliers and influential observations (median, absolute value).

**Example** Let

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 4 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 2.1 \\ 24.2 \\ 29.5 \\ 27.6 \\ 30.5 \\ 27.5 \end{pmatrix}.$$

Find the data's $X-$outliers, $Y-$outliers, and influential observations.

Since $n = 6$, the sample is small. The OLS estimates are

$$\mathbf{b} = \begin{pmatrix} -7.3 \\ 5.51 \\ 5.70 \end{pmatrix},$$

from which

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b} = (-1.8, 3.2, -2.7, 1.28, -1.32, 1.37)^\top.$$

The external residuals are $(-18.47, 2.40, -1.99, 0.41, -0.5, 0.57)^\top$. Since

$$t\left(1-\frac{\alpha/n}{2}; n-p-1\right) = t\left(1-\frac{0.1/6}{2}; 6-3-1\right) = 7.65,$$

**only the first case** is a $Y-$outlier at $\alpha = 0.1$; conservatively, when $|t_i|$ is large, we should further study the influence of case $i$, so we will be sure to look into case 1 in detail.[62]

62: Note the Bonferroni correction term.

For $X-$outliers, we seek cases with leverages above 0.5:

$$\mathbf{h} = (0.87, 0.45, 0.58, 0.19, 0.41, 0.48)^\top.$$

Cases $1, 3$ are **high** leverage points, suggesting that they are potential $X-$outliers, whereas cases $2, 5, 6$ have **moderate** leverages (but are unlikely to be $X-$outliers, lest $5/6$ observations be so).

The **differences in fitted values** are

$$\text{DFFITS} = (-48.7, 2.29, -2.33, 0.2, -0.42, 0.54)^\top,$$

suggesting that only the first 3 cases are influential. The **Cook distances** are $\mathbf{D} = (6.9, 0.67, 0.91, 0.02, 0.08, 0.13)^\top$; since $D_1$ is the only distance larger than than $F(0.5; p, n-p) = 1$, only the **first** case is likely to be influential.

## 8.6 Exercises

1. a) Let $U_i \sim \chi^2(r_i)$ be independent random variables with $r_1 = 5$, $r_2 = 10$. Set
$$X = \frac{U_1/r_1}{U_2/r_2}.$$
Using R, find $s$ and $t$ such that
$$P(X \leq s) = 0.95 \quad \text{and} \quad P(X \leq t) = 0.99.$$
$P(V \leq w) = 0.95$.
2. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$, and $a \in \mathbb{R}$. Define $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v} + a$. Find the gradient of $f$ with respect to $\mathbf{Y}$. Write a function in R that computes $f(\mathbf{Y})$ given $\mathbf{v}, a$. Evaluate the function at $\mathbf{Y} = (1, 0, -1)$, for $\mathbf{v} = (1, 2, -3)$ and $a = -2$.[63]
3. Let $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$, $\boldsymbol{\mu} = (1, 0, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\mathbf{W} = A\mathbf{Y}$. What distribution does the random vector $\mathbf{W}$ follow? Draw a sample of size 100 for this random vector with R and plot them in a graph. You may use the function `mvrnorm()` from the `MASS` package to help along (but you do not have to).
4. Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_4)$ and set $\overline{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$. Using R, draw 1000 observations (and plot a histrogram) from:
   a) $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$
   b) $4\overline{Y}^2$
   c) $(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2$

63: We write vectors either as columns or as rows, in a more or less arbitrary way. It is up to you to determine which one makes the dimensions compatible.

5. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(\mathbf{Y}) = Y_1^2 + \tfrac{1}{2}Y_2^2 + \tfrac{1}{2}Y_3^2 - Y_1Y_2 + Y_1 + 2Y_2 - 3Y_3 - 2.$$

Using `R`, find the critical point(s) of $f$. If it is unique, does it give rise to a global maximum of $f$? A global minimum? A saddle point?

6. Consider the dataset `Autos.xlsx`. The predictor variable is `VKM.q` ($X$, the average daily distance driven, in km); the response variable is `CC.q` ($Y$, the average daily fuel consumption, in L). Use `R` to:
   a) display the scatterplot of $Y$ versus $X$;
   b) determine the number of observations $n$ in the dataset;
   c) compute the quantities $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_iY_i$, $\sum Y_i^2$;
   d) find the normal equations of the line of best fit;
   e) find the coefficients of the line of best fit (without using `lm()`), and
   f) overlay the line of best fit onto the scatterplot.
7. Use the `R` function `lm()` to obtain the coefficients of the line of best fit and the residuals from exercise 6. Show (by calculating the required quantities directly) that the first 5 properties of residuals are satisfied.
8. Using `R`, compute the Pearson and Spearman correlation coefficients between the predictor and the response in exercise 6. Is there a strong or weak linear association between these two variables? Use the correlation values and diagrams to justify your answer.
9. Using `R`, find the decomposition into sums of squares for the regression in exercise 6.
10. (continuation of the previous question) Using `R`, randomly draw $n$ pairs of observations from the data set. Determine the least squares line of best fit $L_n$ and calculate its coefficient of determination $R_n^2$. Repeat for $n = 10, 50, 100, 500$ and for all observations. Is there anything interesting to report? If so, how is it explained?
11. Using `R`, plot the residuals corresponding to the ls line of best fit when using all observations in the set. Visually, do the SLR assumptions on the error terms appear to be satisfied? Give a visual approximation of $\sigma^2$. Then compute the estimator $\widehat{\sigma}^2$. Compare.
12. Using `R`, compute directly the 95% and the 99% confidence interval of the slope of the regression line.
13. Before even doing the calculations with `R`, do you think we should be able to determine whether the confidence interval for the intercept of the regression line is smaller or larger than the corresponding interval for the slope? If so, why would this be the case? Determine directly the 95% and the 99% confidence interval of the intercept.
14. (continuation of the previous question) Using the fit from the previous questions:
   a) Test for $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 > 0$.
   b) Test for $H_0 : \beta_1 = 10$ vs. $H_1 : \beta_1 \neq 10$.
   c) Test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

   Justify and explain your answers.
15. (continuation of the previous question)
   a) Using the formulas, calculate the covariance $\sigma\{b_0, b_1\}$.
   b) Randomly select a sample of 50 pairs of observations from `Autos.xlsx` (with or without remplacement, as desired). Compute the regression parameters $(b_0^{(1)}, b_1^{(1)})$ corresponding to the sample. Repeat the procedure 300 times, to produce 300 pairs $(b_0^{(j)}, b_1^{(j)})$. Display all pairs in a scatter plot.
   c) Comment on the results. Are they consistent with what you obtained in a)?
16. Determine the 95% confidence interval of the expected response $\mathrm{E}\{Y\}$ when the predictor is $X = X^*$. What is the specific interval when $X^* = 27$? Calculate the mean of the responses $\{Y^*\}$ when $X^* = 27$ in the data. Does this mean fall within the confidence interval? Repeat the exercise for $X^* = 5$. Test $H_0 : \mathrm{E}\{Y^* \mid X^* = 5\} = 0$ vs. $H_1 : \mathrm{E}\{Y^* \mid X^* = 5\} > 0$ at confidence level $\alpha = 0.05$.
17. Determine the 95% prediction interval for a new response $Y_p^*$ when the predictor is $X = X^*$. What is the specific interval when $X^* = 27$? What proportion of the responses $Y_p^*$ fall within the prediction interval when $X^* = 27$? Repeat the exercise for $X^* = 5$. Are the results compatible with the notion of prediction interval? Is the observation (5.25) probable (at $\alpha = 0.05$)?

18. (continuation of the previous question)
    a) Perform a 95% joint estimate of the parameters $\beta_0$ and $\beta_1$ Compare with the results of question 16.
    b) Find the joint 95% Working-Hostelling confidence band for the mean response $E\{Y\}$ when $X = X^*$. Superimpose the line of best fit and the band on the scatterplot of the observations.
    c) Find a joint 95% confidence band for the prediction of $g = 20$ new responses $Y_k^*$ at $X = X_k^*$, $k = 1, \ldots, 20$. Superimpose the line of best fit and the band on the scatterplot of the observations.
19. (continuation of the previous question) Perform an analysis of variance to determine if the regression is significant or not.
20. (continuation of the previous question) Express the SLR $Y_i = beta_0 + beta_1 X_i + varepsilon_i$ using matrix notation. With `R`, determine the OLS solution directly (without using `lm()` or the sums $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_i Y_i$, $\sum Y_i^2$).
21. Consider the dataset `Autos.xlsx`. This time around, we are only interested in the VPAS vehicles. The predictor variables are `VKM.q` ($X_1$, the average daily distance driven, in km) and `Age` ($X_2$, the age of the vehicle, in years); the response variable is `CC.q` ($Y$, the average daily fuel consumption, in L). Use `R` to:
    a) determine the design matrix $\mathbf{X}$ of the SLR model;
    b) compute the fitted values of the response $\mathbf{Y}$ if $\boldsymbol{\beta} = (1, 5, 1)$;
    c) compute the residual sum of squares if $\boldsymbol{\beta} = (1, 5, 1)$.
22. (continuation of the previous question) Determine directly the least squares estimator $\mathbf{b}$ of the SLR problem, using matrix manipulations in `R`. Find the estimated regression function of the response $Y$. Compute the residual sum of squares in the case $\boldsymbol{\beta} = \mathbf{b}$. Is this value consistent with the result obtained in part c) of the previous question?
23. (continuation of the previous question) Using only matrix manipulations in `R`, determine the vector of residuals in the SLR problem, as well as SST, SSE, and SSR. Verify that SST = SSR + SSE. What is the mean square error of the SLR model?
24. (continuation of the previous question) Assuming the SLR model is valid, test whether the regression is significant using the global $F$ test – use `R` as you see fit (but use it!).
25. (continuation of the previous question) Find the estimated variance-covariance matrix $s^2\{\mathbf{b}\}$ for the OLS estimator $\mathbf{b}$. At a confidence level of 95%, test for
    a) $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$;
    b) $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 < 0$.
26. (continuation of the previous question) We want to predict the mean response $E\{Y^*\}$ when $\mathbf{X}^* = (20, 5)$. What is the fitted value $\hat{Y}^*$ in this case? Compute a 95% C.I. for the sought quantity.
27. (continuation of the previous question) We want to predict the new response $Y_p^*$ when $\mathbf{X}^* = (20, 5)$. Compute a 95% P.I. for $Y_p^*$.
28. (continuation of the previous question)
    a) Give joint 95% C.I. for the regression parameters $\beta_0, \beta_1, \beta_2$.
    b) Give joint 95% C.I. for the expected mean value $E\{Y_\ell^*\}$ using the Working-Hotelling procedure for $\mathbf{X}_1^* = (50, 10)$, $\mathbf{X}_2^* = (20, 5)$, $\mathbf{X}_3^* = (200, 8)$.
29. (continuation of the previous question) Is the multiple linear regression model preferable to the two simple linear regression models for the same subset of `Autos.xlsx` (using $X_1$ or $X_2$, but not both)? Support your answer.
30. (continuation of the previous question) Compute the multiple coefficient of determination and the adjusted multiple coefficient of determination directly (without using `lm()`). What do these values tell you about the quality of the fit?
31. (continuation of the previous question) Is the linearity assumption reasonable? Justify your answer.
32. (continuation of the previous question) Is the assumption of constant variance reasonable? Justify your answer.
33. (continuation of the previous question) Is the assumption of independence of the error terms reasonable? Justify your answer.
34. (continuation of the previous question) Is the assumption of normality of the error terms reasonable? Justify your answer.

35. (continuation of the previous question) Overall, do you believe that the multiple linear regression model is appropriate? Justify your answer.
36. (continuation of the previous question) Use appropriate corrective measures to improve the multiple regression results.
37. (continuation of the previous question) Are the predictors in the data set multicollinear? Justify your answer.
38. (continuation of the previous question) For this question, we drop the variable `Age` from the dataset. Fit the response to a cubic regression centered on the predictor $x_1 = X_1 - \overline{X}_1$, by adding one variable at a time, to obtain $E\{Y \mid x_1\} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$. Using $\alpha = 0.05$, test for $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$.
39. (continuation of the previous question) For this question, we re-introduce the variable `Age` to the data. Build a polynomial model of degree 2 in $X_1$ and $X_2$ that includes an interaction term (the full model) and a model that is only of degree 1 in $X_1$ and $X_2$, but still contains an interaction term (the reduced model). Determine the coefficients in both cases. Which of the two models is better?
40. Consider the dataset `Autos.xlsx`. The predictor variable is `Type` ($X$, vehicle type); the response is `CC.q` ($Y$, average daily fuel consumption, in L). Using a dummy variable encoding, find the regression model of $Y$ as a function of $X$. Is this a good model? Justify your answer.
41. Use the data set provided in the example for Section 4.5.
    a) Find and plot the solution of the WLS problem with $w_i = x_i^2$.
    b) Find the solution of the WLS problem with the procedure described in the chapter. Plot the results.
    c) Which of the two options gives the best fit? Justify your answer.
42. Consider the dataset `Autos.xlsx`. The predictor variables are `VKM.q` ($X_1$, average daily distance, in km), `Age` ($X_2$, vehicle age in years), and `Rural` ($X_3$, 0 for urban vehicle, 1 for rural vehicle); the response is `CC.q` ($Y$, average daily fuel consumption, in L). Use the best subset approach with Mallow's $C_p$ criterion to select the best model.
43. Repeat the previous question, with the adjusted coefficient of determination $R_a^2$.
44. Repeat the previous question, with the backward stepwise selection method and with Mallow's $C_p$ criterion.
45. Repeat the previous question, with the backward stepwise selection method and with the adjusted coefficient of determination $R_a^2$.
46. Repeat the previous question, with the forward stepwise selection method and with Mallow's $C_p$ criterion.
47. Repeat the previous question, with the forward stepwise selection method and with the adjusted coefficient of determination $R_a^2$.
48. Consider the dataset `Autos.xlsx`. The predictor variables are `VKM.q` ($X_1$, average daily distance, in km) and `Age` ($X_2$, vehicle age in years), and `Rural` ($X_3$, 0 for urban vehicle, 1 for rural vehicle; the response is still `CC.q` ($Y$, average daily fuel consumption, in L). Find the $X-$outliers in the dataset.
49. (continuation of the previous question) Consider the MLR model $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$. Find the $Y-$outliers in the dataset.

## Chapter References


[1] P. Boily. *Analysis and Topology Study Aids* . Data Action Lab.

[2] P. Boily and R. Hart. *Le calcul dans la joie* . 2nd ed. 2020.

[3] G.E.P. Box. 'Use and Abuse of Regression'. In: *Journal of Technometrics* 8.4 (Nov. 1966), pp. 625–629.

[4] F. Donzelli. *Multivariable Calculus*. Kendall Hunt, 2022.

[5] F.E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing, 2015.

[6] R.V. Hogg and E.A Tanis. *Probability and Statistical Inference*. 7th. Pearson/Prentice Hall, 2006.

[7] M.H. Kutner et al. *Applied Linear Statistical Models*. McGraw Hill Irwin, 2004.

[8] W.K. Nicholson. *Linear Algebra with Applications , 3rd Edition*. PWS Publishing Company, 1994.

[9] H. Rosling. *The Health and Wealth of Nations* . Gapminder Foundation, 2012.

[10] H. Rosling, O. Rosling, and A.R. Rönnlund. *Factfulness: Ten Reasons We're Wrong About The World - And Why Things Are Better Than You Think* . Hodder & Stoughton, 2018.

# Time Series and Forecasting

# 9

by **Patrick Boily**, inspired by **Rafal Kulik**

---

Many traditional statistical methods assume that observations are independently and identically distributed, which is unlikely to happen in real life. At best, this assumption may be sufficiently accurate to allow for some predictive power; at worst, it can lead to wildly inaccurate insights and predictions.

A time series is a sequence of values, measured at regular intervals over time. The motivation of time series analysis lies in the assumption that what happened in the past has an influence on what will happen in the future. Typically, time series are used for **trend analysis** and for **forecasting** future values when there are good reasons to suspect the existence of cycles in the data.* Generally speaking, the forecast horizon is the length of the prediction period: predictions at shorter horizons tend to be more reliable and accurate than predictions at longer horizons.

Ideally, the reporting periods used in time series analysis should be identical (e.g. daily, monthly, quarterly or yearly), the measurements should be taken over discrete (exclusive), consecutive periods, and the concepts and the measurement approach should be consistent over time. Detection of periodicity should be done by graphical representation of the data (and the frequency of data collection) using logic (e.g., is there an expectation of hourly, weekly, monthly, quarterly, and/or x-year cycles). More information is available in [2, 1, 5, 3, 4].

## 9.1 Introduction

Various time series analysis methods and tests are found in applications and in the literature, including:

- **auto-regressive** models (AR),
- **smoothing and filtering** models (such as moving averages (MA) and exponential smoothing (ES)),
- **detrending** models (such as ARMA, finite differences, etc.),
- **seasonal decomposition** models (such as X11, X12, X13, and ARIMA models), and
- **linear** and **non-linear forecasting** models (suc has Holt's Method, Winter's Method, GARCH models, etc.).

We start by providing examples and some of the basic concepts of the discipline.

* For instance, a time series analysis could be used to predict the number of passengers going through Canadian airports at various points in the future. Or an economist might be interested in forecasting the stock market, using time series analysis.


9.1 Introduction . . . . . 491
Simple Examples . . . . . 492
Pre-Processing . . . . . 493
Stationary Models . . . . . 504
Partial Autocorrelation . . . 508
9.2 Estimating Parameters . . . 510
Sample Statistics . . . . . 510
Examples . . . . . 511
9.3 ARMA Models . . . . . 516
Linear Processes . . . . . 516
ARMA in General . . . . . 518
Stationarity and Causality . 519
Linear Representation . . . 521
ACVF . . . . . 523
PACF . . . . . 526
9.4 Forecasting . . . . . 530
Yule-Walker Procedure . . . 530
Durbin-Levinson Algorithm 533
Forecast Limits . . . . . 535
Example . . . . . 535
9.5 ARMA Estimation . . . . . 543
Mean: I.I.D. Case . . . . . 543
Mean: Time Series . . . . . 544
Yule-Walker Estimators . . 545
Example . . . . . 548
9.6 Diagnostic Tests . . . . . 551
Ljung-Box Test . . . . . 552
Example . . . . . 553
9.7 MLE Estimation . . . . . 556
I.I.D. Random Variables . . 556
Time Series Model . . . . . 558
Order Selection . . . . . 560
Examples . . . . . 560
9.8 Nonlinear Time Series . . . 575
ARCH Model . . . . . 575
GARCH Model . . . . . 576
Example . . . . . 577
9.9 Miscellenea . . . . . 580
Seasonality . . . . . 581
Asymptotic Normality . . . 584
9.10 Exercises . . . . . 590
Chapter References . . . . . 598

### 9.1.1 Simple Examples

**White Noise** Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance 1. Sometimes such a sequence is called a **white noise.** A sample white noise path consisting of 100 steps, with independent $Z_t \sim \mathcal{N}(0,1)$, is provided by the R code below.[1]

```
z = rnorm(100);
plot.ts(z)
```

**Random Walk** Let $\{Z_t\}$ be a sequence of i.i.d.[2] random variables with mean 0 and variance $\sigma_Z^2$. Define $X_t = \sum_{i=1}^{\top} Z_i$, $t = 1, 2, \ldots$. A sample random walk of 100 steps, with independent $Z_t \sim \mathcal{N}(0,1)$, is provided by the R code below (see Figure 9.1 for the output).

```
z = rnorm(100);
x = cumsum(z);
plot.ts(x)
```

**Model with Trend** A linear or polynomial trend can sometimes be found in time series models. Consider, for instance, the time series

$$X_t = 1 + 2t + Z_t, \quad t = 1, 2, \ldots,$$

where $\{Z_t\}$ is a sequence of i.i.d. random variables. The linear trend is $m_t = 1 + 2t$. A 100-step realization of this model, with independent $Z_t \sim \mathcal{N}(0,1)$, is provided by the R code below (see Figure 9.1).

**Linear trend**

```
trend = 1+2*seq(1:100);
z = rnorm(100,0,10);
x = z+trend;
plot.ts(x)
```

For economics data, we may want to take into account an exponential inflation trend. If the interest rate $r$ is assumed to be fixed, the nominal price $X_t$ is actually the real (deflated) price $P_t$ with respect to inflation:

$$X_t = P_t e^{rt}, \quad t = 1, 2, \ldots.$$

This phenomenon is illustrated in the quarterly earnings of Johnson & Johnson share (1960–80), as shown in Figure 9.1.

**Exponential trend**

```
require(stats);
x = JohnsonJohnson;
plot.ts(x)
```

1: The output is shown in Figure 9.1. Note that the specific realization of the time series depends on the seed used to generate the pseudo-random numbers in R. In the absence of a `set.seed(...)` command, the realization will change after every call; with the command, the realization will be the same after every call. This comment should be kept in mind at all times when producing examples.

2: Independent, identically distributed

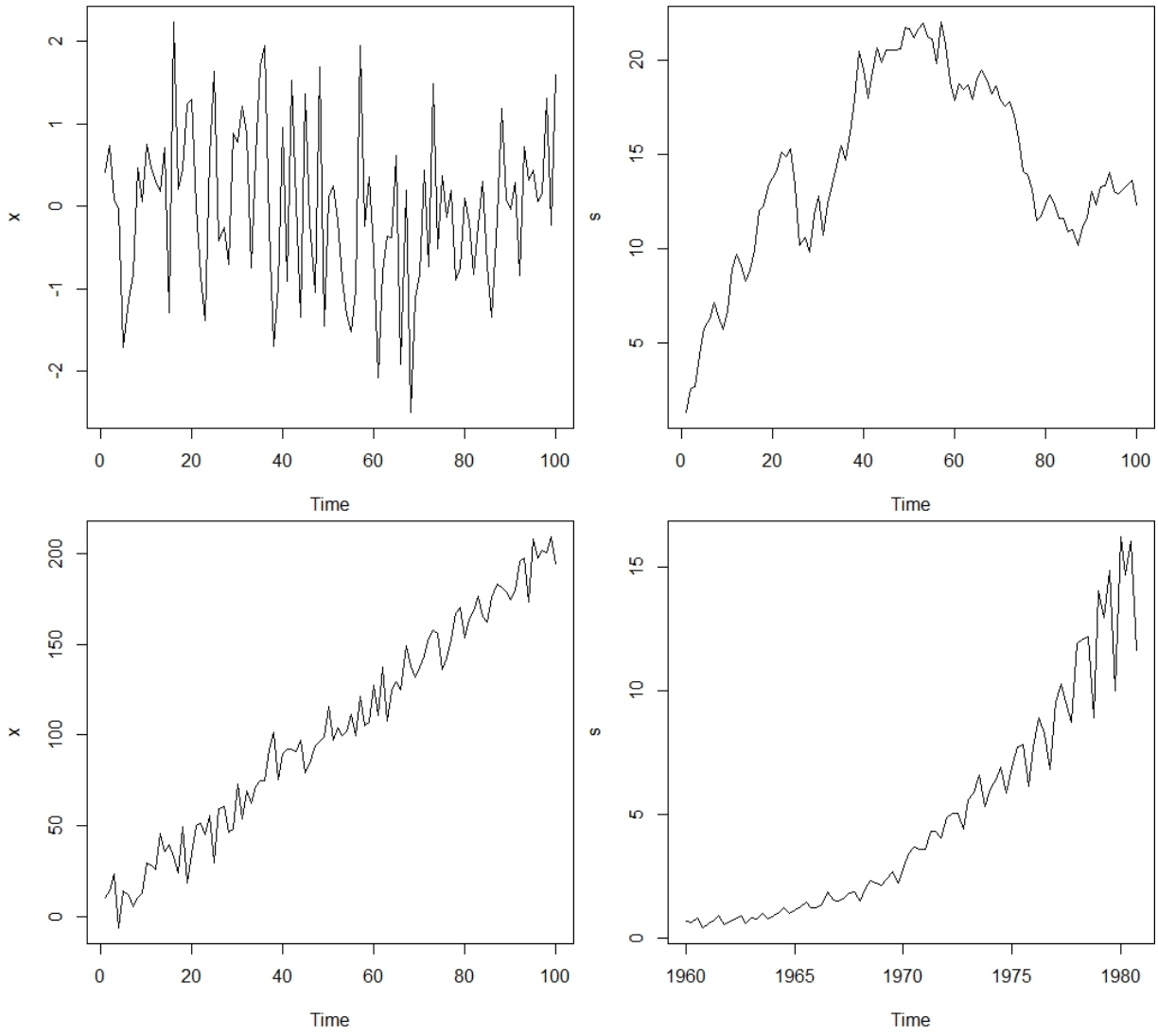



**Figure 9.1:** Simple time series: white noise (top left), random walk (top right), linear trend (bottom left), exponential trend (bottom right).

### 9.1.2 Pre-Processing

**Component decomposition** is central to time series analysis. Displaying the components of a time series is also helpful in understanding the data. Each of the components represents a category of patterns.

Generally speaking, there are three common components of time series: **trend**, **seasonality**, and **irregular**. We briefly discuss other potential components, but for the sake of simplicity, only the first two of these will be discussed in this chapter:

- the **trend** component describes the overall "changing direction" of the data, either increase or decrease or flat, which is a long-term effect and not necessarily linear;[3]
- the **seasonal** component reveals the seasonal effect on a series of data, such as that passengers in the airport will increase during summer vacation season;[4]
- The **irregular** (anomalous) component is a short-term effect, which can vary considerably from period to period, and includes measurement errors, unseasonal change, etc. – once the trend, seasonal,

3: For example, in the linear trend time series model of Figure 9.1, the bottom left graph shows the trend going up, and so we expect $X_t$ to increase with $t$.

4: If the monthly deaths of lung disease in London, UK, shows peaks occurring at the beginning of each year, say, then we conclude that winter is a harsher time for such deaths than summer is, in general.

and cyclical effects are removed, we use the residual of the time series to identify the irregular contributions;
- **cyclical** components usually lasts at least two years – note that, in general, the exact length of an ongoing cycle cannot be predicted;[5]
- **other** components may include calendar effect (trading day, leap year, etc.), government policies, strike actions, exceptional events, inclement weather, etc.

5: For example, the global financial crisis in 2008 lasted about 5 years. The difference between seasonal and cyclical is that the former displays the change over a fixed time period.

**Decomposition Models** Traditionally, decomposition follows one of three models: **multiplicative**, **additive**, and **pseudo-additive**.

The **additive** approach assumes that:

1. the seasonal component $S_t$ and the irregular component $I_t$ are independent of the trend behaviour $m_t$;
2. the seasonal component $S_t$ remains stable from year to year; and
3. the seasonal fluctuations are such that $\sum_{j=1}^{n} S_{t+j} = 0$.[6]

6: For daily series, $n = 365$; for monthly series, $n = 12$; for quarterly series, $n = 4$, and so on.

Mathematically, the model is expressed as:

$$X_t = m_t + S_t + I_t.$$

All components share the same dimensions and units. After seasonality adjustment, the seasonality adjusted series is:

$$SA_t = X_t - S_t = m_t + I_t.$$

The **multiplicative** approach assumes that:

1. the magnitude of the seasonal spikes/troughs increases when the trend increases (and vice versa);
2. the trend $m_t$ has the same dimensions as the original series $X_t$, and the seasonal component $S_t$ and the irregular component $I_t$ are dimensionless and centered around 1;
3. the seasonal fluctuations are such that $\sum_{j=1}^{n} S_{t+j} = 0$, and
4. the original series $X_t$ does not contain zero values.

Mathematically, the model is expressed as:

$$X_t = m_t \times S_t \times I_t.$$

All components share the same units. After seasonality adjustments, the seasonality adjusted series is

$$SA_t = \frac{X_t}{S_t} = m_t \times I_t$$

To transform a multiplicative model into an additive model, we could take a logarithmic transformation, such as:

$$\log X_t = \log m_t + \log S_t + \log I_t,$$

assuming that none of the component values are non-positive.

The **pseudo-additive** approach assumes that some of the values of the original series $X_t$ are 0 (or very close to 0) and that:

1. the seasonal component $S_t$ and the irregular component $I_t$ are both dependent on the trend level $m_t$, but independent of each other, and
2. the trend $m_t$ has the same dimensions as the original series $X_t$, and the seasonal component $S_t$ and the irregular component $I_t$ are dimensionless and centered around 1.

Mathematically, the model is expressed as:

$$X_t = m_t + m_t \times (S_t - 1) + m_t \times (I_t - 1) = m_t \times (S_t + I_t - 1).$$

All components share the same units. After seasonality adjustment, the seasonality adjusted series is:

$$SA_t = X_t - m_t \times (S_t - 1) - m_t \times (D_t - 1) = m_t \times I_t$$

The **choice** of a model is driven by data behaviour and assumptions. The analyst needs to plot the time series graph and test a range of models, selecting the one which stabilized the seasonal component.

The simplest way to determine whether to use multiplicative or additive decomposition, is by graphing the time series. If the size of the seasonal variation increases/decreases over time, multiplicative decomposition should be used (such as in the last chart of Figure 9.1).

On the other hand, if the seasonal variation seems to be constant over time, an additive model should be used (bottom left, Figure 9.1).[7]

7: A pseudo-additive model should be used when the data exhibits the characteristics of the multiplicative series, but with some $X_t$ values near zero.

**Illustration** We illustrate the process of decomposition with an arbitrary time series recording the monthly number of hours for a variable called CV, whose values are shown in the Figure 9.2.

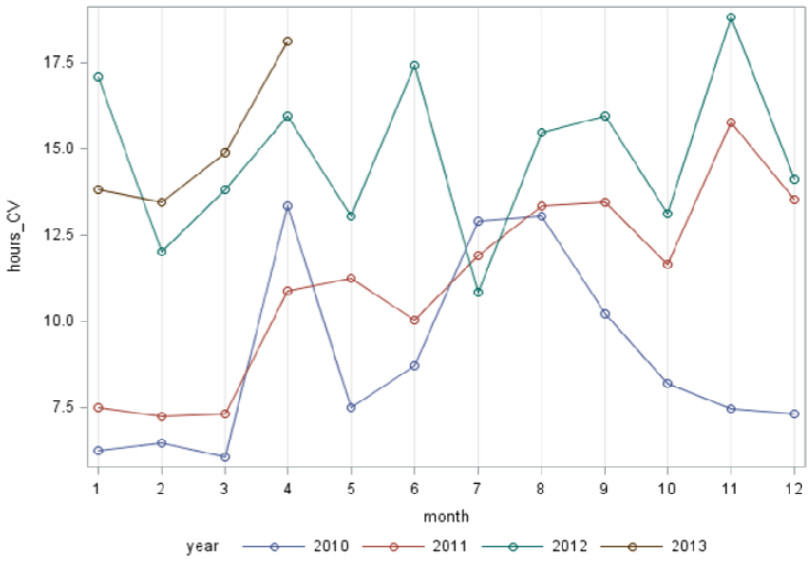



**Figure 9.2:** Time series; CV by year.

The continuous plot, Figure 9.3 ,shows that the size of the peaks and troughs does not seem to follow changing trends: the additive model is thus selected. The SAS procedure X12 agrees with that assessment, and further suggests no data transformation.

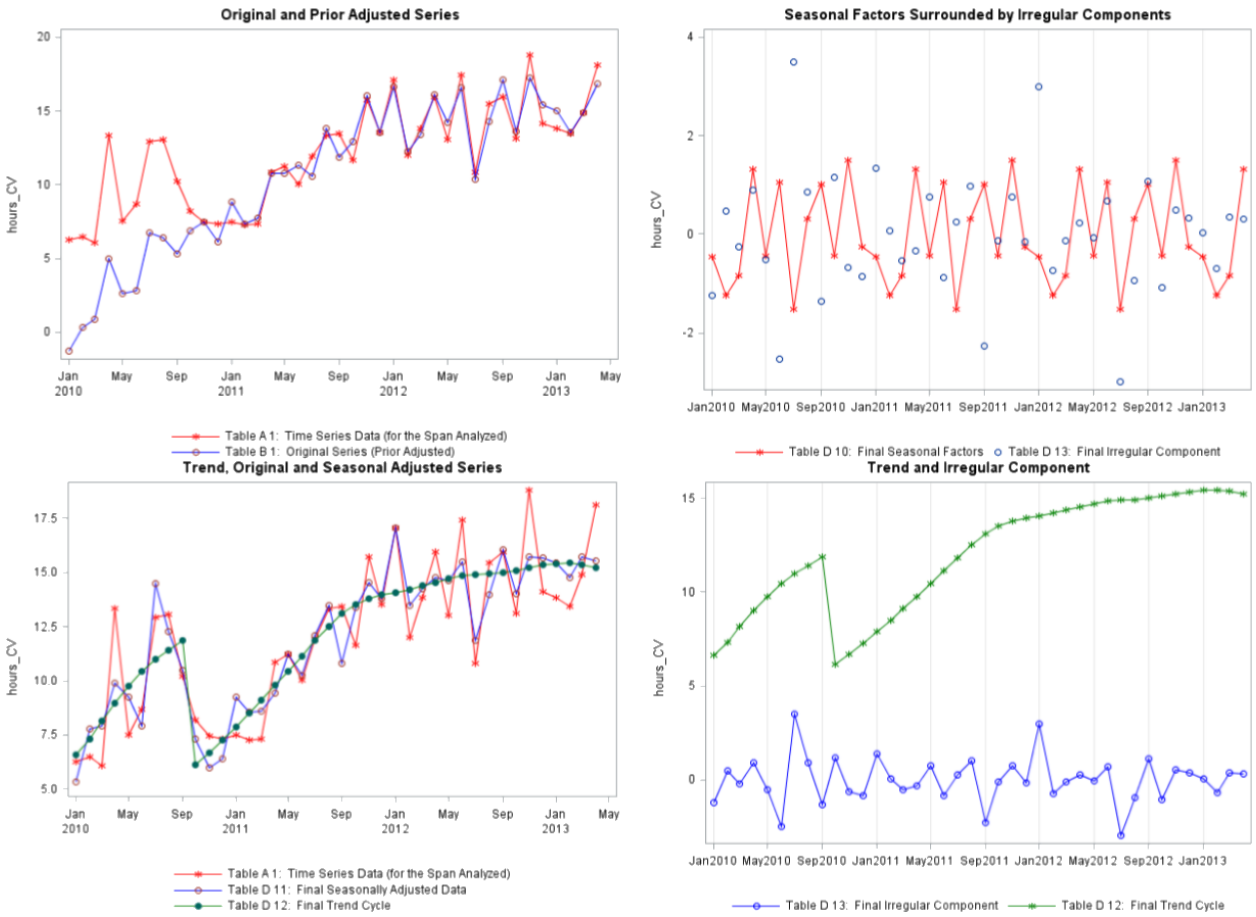
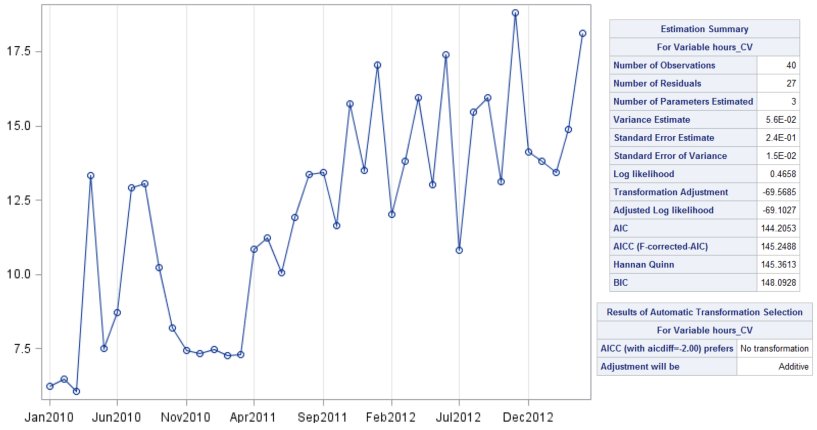



**Figure 9.4:** Diagnostic plots (top row) and adjusted plots (bottom row). Note that the analysis of a time series starts with estimation of the effects of festivals and trading days. These pre-calculated estimates are then used for prior adjustment of the series. The prior adjusted original series is subsequently analyzed using the seasonal adjustment.



| Estimation Summary | |
|---|---|
| For Variable hours_CV | |
| Number of Observations | 40 |
| Number of Residuals | 27 |
| Number of Parameters Estimated | 3 |
| Variance Estimate | 5.6E-02 |
| Standard Error Estimate | 2.4E-01 |
| Standard Error of Variance | 1.5E-02 |
| Log likelihood | 0.4658 |
| Transformation Adjustment | -69.5685 |
| Adjusted Log likelihood | -69.1027 |
| AIC | 144.2053 |
| AICC (F-corrected AIC) | 145.2488 |
| Hannan Quinn | 145.3613 |
| BIC | 148.0928 |

| Results of Automatic Transformation Selection | |
|---|---|
| For Variable hours_CV | |
| AICC (with aicdiff=-2.00) prefers | No transformation |
| Adjustment will be | Additive |

**Figure 9.3:** Continuous CV; estimation summary.

The diagnostic plots are shown in Figure 9.4: the 2010 CV series is prior-adjusted from the beginning until OCT2010 after the detection of a level shift. The SI (Seasonal-Irregular) chart shows that there are more than one irregular component which exhibits volatility. The adjusted series is shown at the bottom of Figure 9.4 (the trend and irregular components are shown separately for readability).

**Roll-Back** In this chapter, however, we will focus on time series whose **structure** can be broken down into three additive components,

$$X_t = m_t + Y_t + S_t,$$

where:

- $m_t$ is the trend;
- $S_t$ is the seasonal component;
- $Y_t$ is the **stationary** component (to be defined shortly).

In order to analyse time series, we first need to eliminate both the trend and the seasonal component.[8] We present a few ways to accomplish this, assuming that there is no seasonal component, i.e. $S_t \equiv 0$.

8: Collectively, these are known as the **non-stationarities** of the time series.

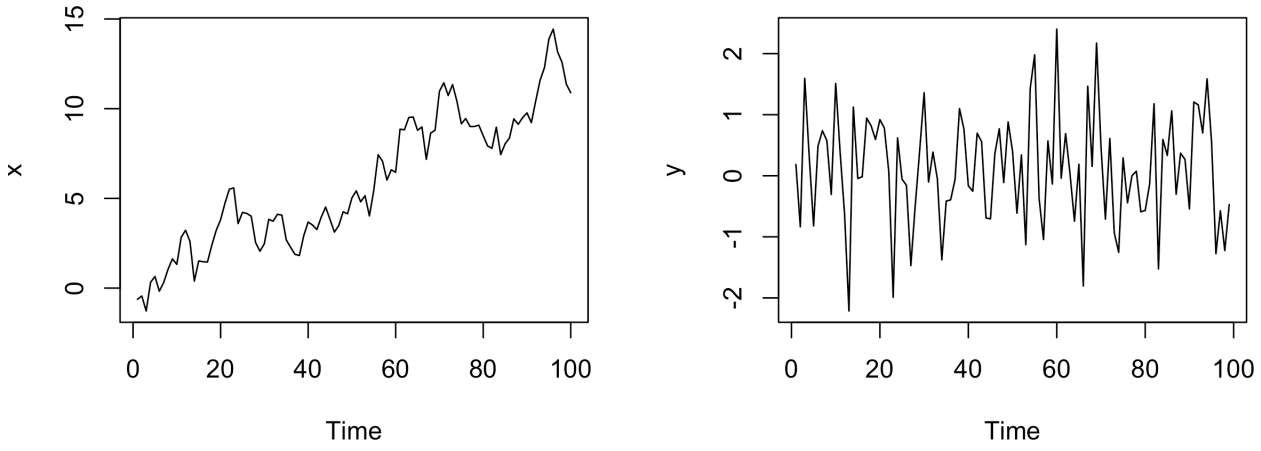
**Differencing** For the time series $\{X_t, t = 1, \ldots, n\}$, we may calculate

$$\nabla X_t = X_t - X_{t-1}, \quad t = 2, \ldots, n.$$

Depending on the nature of the trend in the original time series, the differenced time series may exhibit no trend.

**Differencing a random walk**

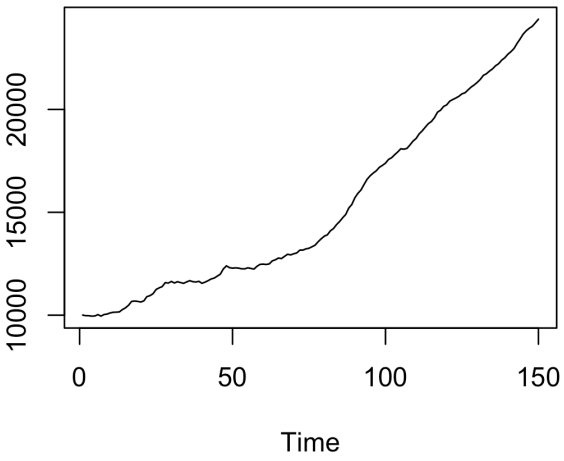
```
set.seed(1)
z = rnorm(100)
x = cumsum(z)
y = diff(x)
par(mfrow=c(1,2))
plot.ts(x)
plot.ts(y)
```

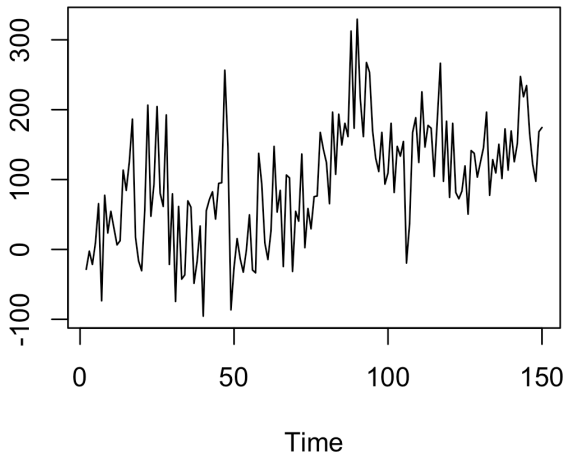



In a sense, differencing a time series is akin to **differentiating** a function $f : \mathbb{R} \to \mathbb{R}$; if the underlying trend is roughly linear, we expect the differenced time series to have **white noise** characteristics.[9]

9: Which is to say, that the trend is horizontal.

But if the underlying trend is not linear, differencing only once might not detrend the original series, as can be seen below, where the trend has a clear (positive) slope.

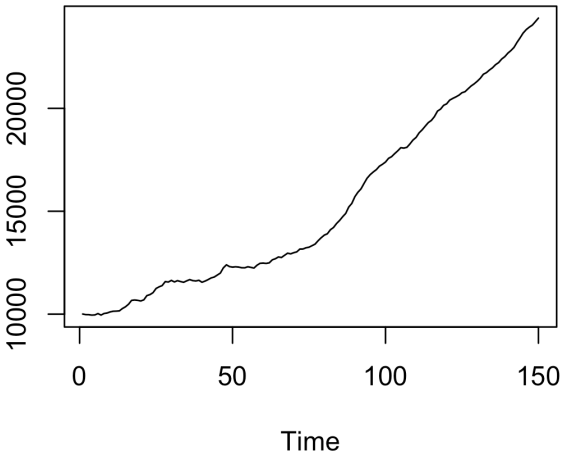


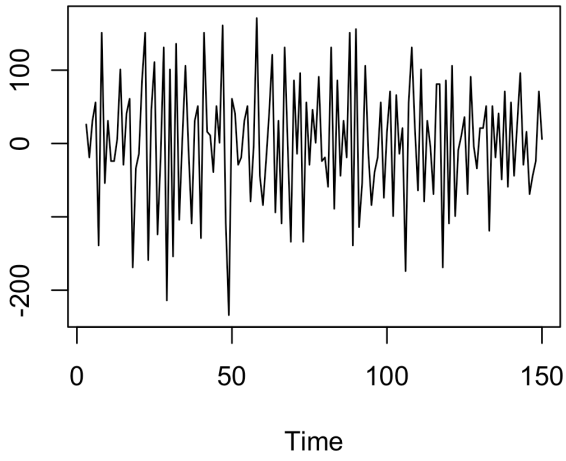




Given that the original time series trend is concave up, differencing a second time could be a good strategy:[10]

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = \nabla X_t - \nabla X_{t-1} = X_t - 2X_{t-1} + X_{t-2},\ t = 3, \ldots, n.$$

10: Since, by analogy, the second derivative of a quadratic function is the zero function.




**Polynomial Fitting** When a linear trend is clearly visible ($m_t = a + bt$), then we can estimate the parameters $a, b$ by minimizing

$$\sum_{t=1}^{n} (X_t - a - bt)^2.$$

This is a simple regression problem (see Chapter 8), where the independent variable is time $t$ and the dependent variable is the time series itself. Consequently, the trend is estimated by

$$\widehat{m}_t = \widehat{a} + \widehat{b}t,$$

where $\widehat{a}$ and $\widehat{b}$ are the **least squares estimators** of $a$ and $b$, respectively. In this case, the **detrended time series** is

$$\widehat{Y}_t = X_t - \widehat{m}_t,\ \ t = 1, \ldots, n.$$

If the trend $m_t$ would be better described by another polynomial, the process is similar; note however that it is not in general easy to justify using a non-linear polynomial trend.

As an example, consider the following time series, whose trend is linear by construction.

```
set.seed(11)
n=89; a=4; b=10;
Time=c(1:n);
X = a + b*Time + 20*rnorm(n)
```

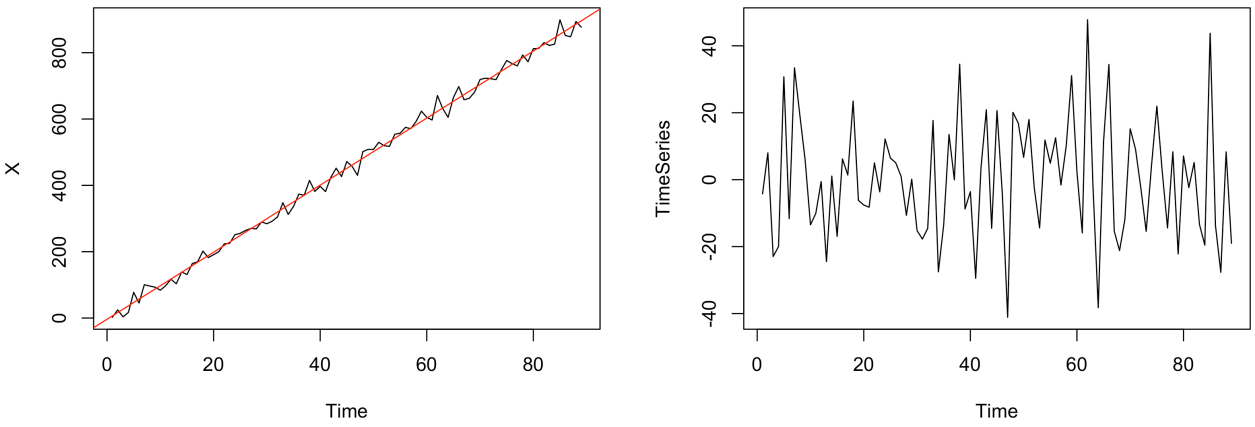
We can find the least squares estimates as follows:

```
estimation = lm(X~Time);
a.est = estimation$coefficients[1]; # Estimated intercept
b.est = estimation$coefficients[2]; # Estimated slope
c(a.est,b.est)
```

```
[1] -3.695528  10.10823
```

We plot the time series with its linear trend and compute the stationary part by removing the linear trend.

```
Fitted.Lin.Trend=a.est+b.est*Time;
TimeSeries=X-Fitted.Lin.Trend;

par(mfrow=c(1,2))
plot.ts(X)
abline(a=a.est,b=b.est, col="red", lwd=1);
plot.ts(TimeSeries);
```




**Exponential Smoothing** Let $\alpha \in (0,1)$. We can estimate the trend *via*:

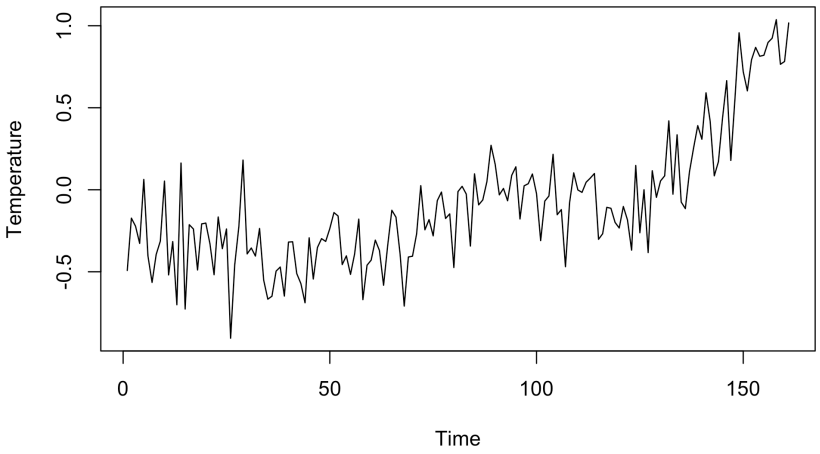
$$\widehat{m}_1 = X_1, \qquad \widehat{m}_t = \alpha X_t + (1-\alpha)\widehat{m}_{t-1}, \;\; t = 2, \ldots, n.$$

In other words, at any time $t$, we assign weights $\alpha$ and $1-\alpha$ to the current observation and the preceding smoothed data. The detrended time series is

$$\widehat{Y}_t = X_t - \widehat{m}_t, \;\; t = 1, \ldots, n.$$

Let us take a look at an example.

```
Temperature = c(-0.492, -0.173, -0.222, -0.327,  0.063,
        -0.403, -0.565, -0.394, -0.313,  0.053, -0.519,
        -0.316, -0.701,  0.163, -0.727, -0.213, -0.239,
        -0.489, -0.208, -0.203, -0.329, -0.518, -0.166,
        -0.359, -0.239, -0.905, -0.456, -0.223,  0.181,
        -0.391, -0.355, -0.404, -0.236, -0.551, -0.667,
        -0.649, -0.496, -0.471, -0.648, -0.319, -0.317,
        -0.511, -0.572, -0.689, -0.293, -0.544, -0.352,
        -0.298, -0.315, -0.236, -0.139, -0.160, -0.456,
        -0.403, -0.516, -0.391, -0.179, -0.670, -0.460,
        -0.429, -0.307, -0.370, -0.582, -0.339, -0.125,
        -0.167, -0.393, -0.709, -0.410, -0.405, -0.268,
         0.025, -0.244, -0.182, -0.281, -0.066, -0.014,
        -0.175, -0.147, -0.474, -0.011,  0.021, -0.026,
        -0.343,  0.097, -0.092, -0.062,  0.050,  0.271,
         0.155, -0.031,  0.008, -0.067,  0.088,  0.140,
        -0.178,  0.024,  0.037,  0.096, -0.024, -0.310,
        -0.069, -0.038,  0.216, -0.152, -0.121, -0.469,
        -0.078,  0.103, -0.001, -0.016,  0.046,  0.071,
         0.099, -0.302, -0.268, -0.107, -0.113, -0.199,
        -0.233, -0.102, -0.184, -0.368,  0.148, -0.262,
         0.000, -0.383,  0.116, -0.046,  0.054,  0.085,
         0.420, -0.027,  0.335, -0.075, -0.115,  0.110,
         0.256,  0.391,  0.308,  0.591,  0.418,  0.085,
         0.171,  0.438,  0.665,  0.179,  0.555,  0.957,
         0.720,  0.603,  0.792,  0.868,  0.814,  0.820,
         0.898,  0.924,  1.037,  0.765,  0.782,  1.017)
plot.ts(Temperature)
```



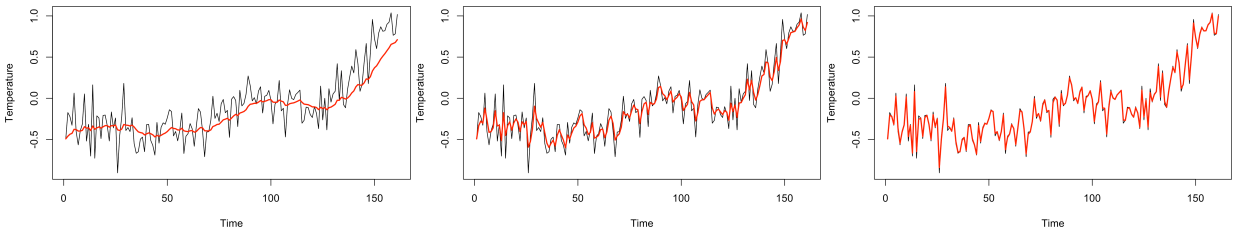

This times series is not stationary, so we need to remove its trend. Exponential smoothing is implemented in the following R function.

```
ExpSmooth <- function(x,alpha){
  # x: data
  # alpha: smoothing parameter
  n = length(x)
  Data = c(rep(0,n))
  Data[1] = x[1]
```

```
  for(i in 2:n){
    Data[i] = alpha*x[i] + (1-alpha)*Data[i-1]
  }
  out <- Data
}
```

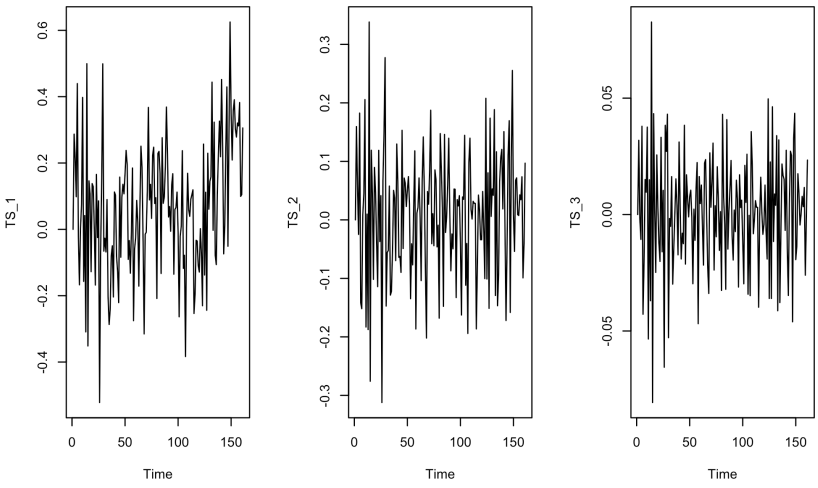
What effect does the parameter $\alpha$ have on the outcome? In general, the smaller $\alpha$ is, the smoother the trend is; here, we try $\alpha = 0.1, 0.5, 0.9$.

```
plot.ts(Temperature)
MySmoothedTS1 = ExpSmooth(Temperature,0.1)
points(MySmoothedTS1,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS2 = ExpSmooth(Temperature,0.5)
points(MySmoothedTS2,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS3 = ExpSmooth(Temperature,0.9)
points(MySmoothedTS3,col="red",type="l", lwd=2)
```




Using $\alpha = 0.1$ (left) indeed achieves the smoothest trend; $\alpha = 0.9$ (right) shows barely any smoothing. Detrending the series, we obtain:

```
TS_1 = Temperature-MySmoothedTS1
TS_2 = Temperature-MySmoothedTS2
TS_3 = Temperature-MySmoothedTS3

par(mfrow=c(1,3))
plot.ts(TS_1); plot.ts(TS_2); plot.ts(TS_3)
```



The outcome of the procedure is a time series (in this example, either `TS_1`, `TS_2`, or `TS_3`), which we hope can be treated as stationary.[11] Of course, different smoothing parameters $\alpha$ lead to different stationary time series – experience will inform the choice of $\alpha$. The main thrust is that the exponential smoothing should not follow the data too closely while preserving the trend and the trend-removed dependence structure.

11: These are the time series that will be analysed using the methods we discuss in this chapter.

**Moving Average Smoothing** Another detrending approach requires us to pick a **window size** $q$ (a positive integer). Then the trend is estimated *via*

$$\widehat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t+j}, \quad q+1 \le t \le n-q.$$

The detrended time series is

$$\widehat{Y}_t = X_t - \widehat{m}_t, \quad t = q+1, \ldots, n-q.$$

Why does this method work? By assumption, we have $X_t = m_t + Y_t$. We assume further that $\mathrm{E}[Y_t] = 0$.[12] Then

12: If this is not the case, the non-zero mean can be always incorporated into the trend.

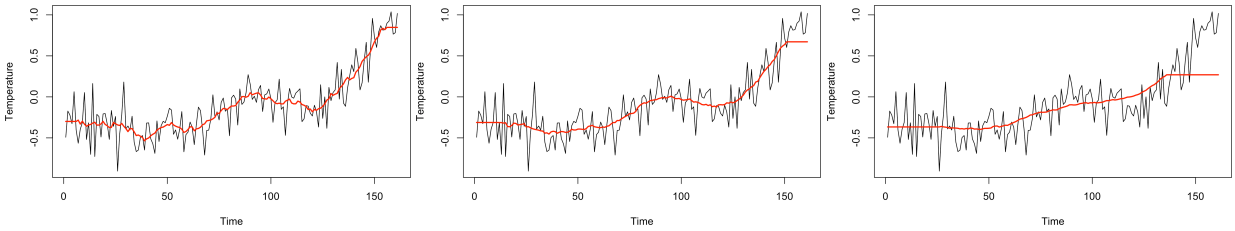
$$(2q+1)^{-1} \sum_{j=-q}^{q} X_{t+j} = (2q+1)^{-1} \sum_{j=-q}^{q} m_{t+j} + (2q+1)^{-1} \sum_{j=-q}^{q} Y_{t+j}.$$

If the trend is linear ($m_t = a + bt$) Then

$$(2q+1)^{-1} \sum_{j=-q}^{q} m_{t+j} = (2q+1)^{-1} \sum_{j=-q}^{q} \{a + b(t+j)\} = a + bt.$$

We apply this approach to the `Temperature` data from the previous method, using $q = 5, 10, 25$.

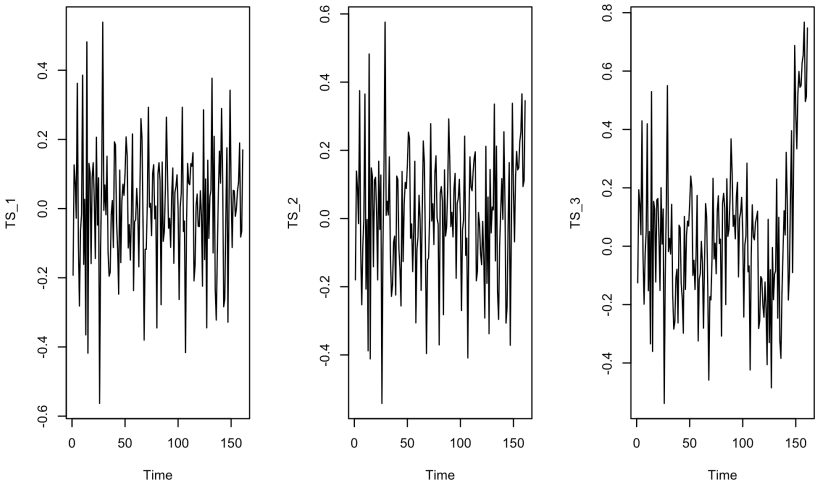
```
MASmooth<-function(x,Q){
  # x: data set
  # Q: MA window size
  n = length(x)
  Smooth = c(rep(0,n))
  for(i in Q+1:(n-Q)){Smooth[i] = mean(x[(i-Q):(i+Q)])}
  for(i in 1:Q){Smooth[i] = Smooth[Q+1]}
  for(i in (n-Q+1):n){Smooth[i] = Smooth[(n-Q)]}
  out <- Smooth }

plot.ts(Temperature)
MySmoothedTS1 = MASmooth(Temperature,5)
points(MySmoothedTS1,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS2 = MASmooth(Temperature,10)
points(MySmoothedTS2,col="red",type="l", lwd=2)

plot.ts(Temperature)
MySmoothedTS3 = MASmooth(Temperature,25)
points(MySmoothedTS3,col="red",type="l", lwd=2)
```




Note the flattening of the trend at the extremities.

The detrended time series are displayed below.

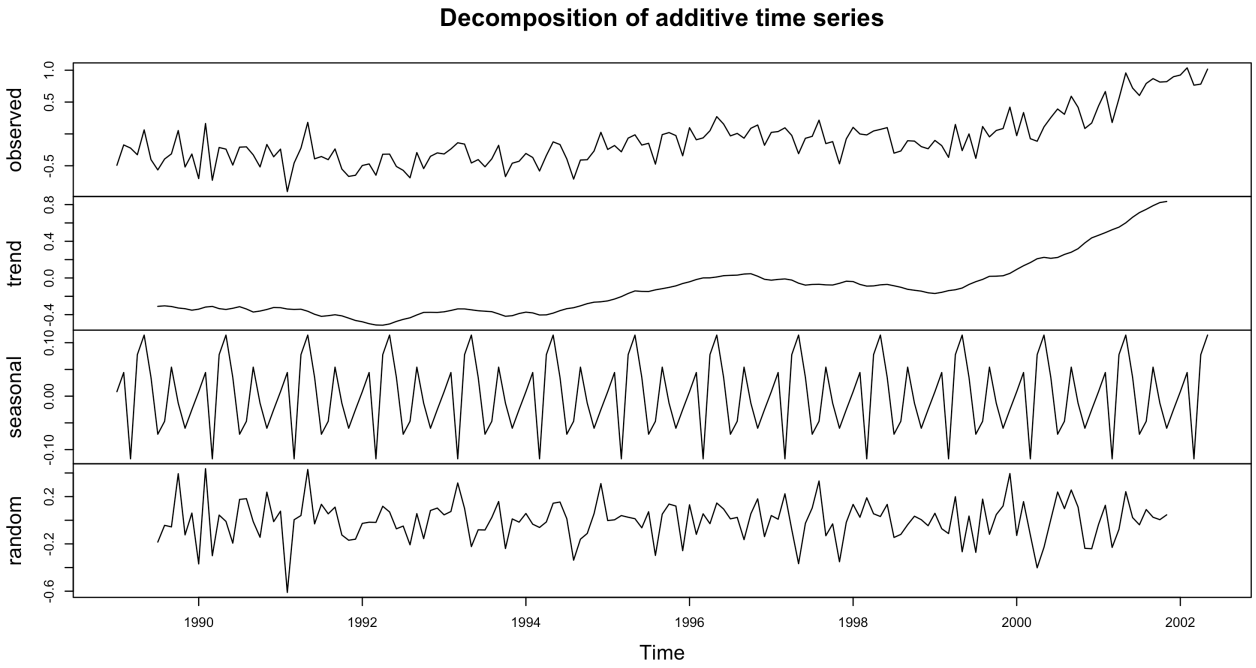
```
TS_1 = Temperature-MySmoothedTS1
TS_2 = Temperature-MySmoothedTS2
TS_3 = Temperature-MySmoothedTS2

par(mfrow=c(1,3))
plot.ts(TS_1)
plot.ts(TS_2)
plot.ts(TS_3)
```




**Built-In Decomposer** Most statistical analysis tools have built-in functions that can decompose time series according to some model.

For instance, if the temperature data is a monthly time series, starting in 1989 (and assuming that there is a seasonal component $S_t$), then `tseries`'s `decompose()` function can extract the stationary component (named `random` in this implementation) using an additive model and a moving average approach.

```
library(tseries)
Temperature.ts <- ts(Temperature, start=1989, freq=12)
plot(decompose(Temperature.ts))
```




The components can be isolated by calling:

- `decompose(Temperature.ts)$trend`,
- `decompose(Temperature.ts)$seasonal`, and
- `decompose(Temperature.ts)$random`.

### 9.1.3 Stationary Models, Autocovariance, and Autocorrelation

13: Throughout this chapter, **time series** are sequences $\{X_t \mid t = t_0, \ldots\}$ of random variables.

We now introduce the fundamental notions of time series analysis. [13]

**Definitions and Properties**

Let $\{X_t\}$ be a time series with $\mathrm{E}[X_t^2] < \infty$ for each $t$.

The expectation $\mu_X(t) = \mathrm{E}[X_t]$ is a function of $t$, the **mean function**. The **(auto)covariance function** of the time series is defined as

$$\gamma_X(t,s) = \mathrm{Cov}(X_t, X_s) = \mathrm{E}[X_s X_t] - \mathrm{E}[X_s]\mathrm{E}[X_t].$$

14: When the context is clear, we will denote the mean function and the autocovariance function simply by $\mu$ and $\gamma$, respectively.

Note that $\gamma_X(t,t) = \mathrm{Var}(X_t)$.[14]

From our perspective, the most important properties of the covariance are that it is:

- **symmetric**
$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X);$$
- **multilinear**
$$\mathrm{Cov}\left(\sum_{k=1}^{K} a_k X_k, \sum_{\ell=1}^{L} b_\ell Y_\ell\right) = \sum_{k=1}^{K}\sum_{\ell=1}^{L} a_k b_\ell \mathrm{Cov}(X_k, Y_\ell),$$
- and $\mathrm{Cov}(X,a) = 0$ for all $a \in \mathbb{R}$.

**Cauchy's Inequality:** if $X, Y$ are r.v., then

$$\mathrm{Cov}(X,Y))^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y).$$

**Proof:** we may assume that $\mathrm{E}[X] = \mathrm{E}[Y] = 0$.[15] Define the function

$$g(t) = \mathrm{E}[(X+tY)^2] = t^2\mathrm{Var}(Y) + 2t\mathrm{Cov}(X,Y) + \mathrm{Var}(X), \quad t \in \mathbb{R}.$$

By construction, $g(t) \geq 0$ for all $t$. Since it is quadratic in $t$, it has at most one root, which is to say that its discriminant is non-positive. In other words

$$\Delta = 4(\mathrm{Cov}(X,Y))^2 - 4\mathrm{Var}(X)\mathrm{Var}(Y) \leq 0,$$

which implies the result. ■

15: Otherwise, set $X' = X - \mathrm{E}[X]$ and $Y' = Y - \mathrm{E}[Y]$ and work with $X', Y'$ instead of $X, Y$. This can be done since the covariance and the variance are invariant under translation by a constant (see properties above).

A time series $\{X_t\}$ is **(weakly) stationary** if

- $\mu_X(t) \equiv \mu_X$, and
- $\gamma_X(t,s) = f_X(t-s)$ for some function $f_X$.

In particular, for such a time series, we must have $\sigma^2\{()\, X_t) \equiv \sigma_X^2$ and

$$\begin{aligned}
\mathrm{Cov}(X_t, X_{t+1}) &= \gamma_X(t, t+1) = f_X(t+1-t) = f_X(1)\\
\mathrm{Cov}(X_{t+1}, X_{t+2}) &= \gamma_X(t+1, t+2) = f_X(t+2-(t+1)) = f_X(1)\\
&\vdots\\
\mathrm{Cov}(X_{t+k}, X_{t+k+1}) &= \gamma_X(t+k, t+k+1) = f_X(1), \quad k \geq 0.
\end{aligned}$$

**Lemma:** assume that $\{X_t\}$ is a (weakly) stationary time series. Then the covariance function $\gamma_X(t,s)$ is a **non-negative definite function**.[16] **Proof:** we have

$$\begin{aligned}
0 \leq \mathrm{Var}\left(\sum_{j=1}^n a_j X_j\right) &= \mathrm{Cov}\left(\sum_{i=1}^n a_j X_j, \sum_{j=1}^n a_j X_j\right)\\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma_X(i,j).
\end{aligned}$$

This completes the proof. ■

16: For all non-negative integers $n$ and all real numbers $a_1, \ldots, a_n$ we have

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma_X(i,j) \geq 0.$$

Under the same hypothesis as above, then $\gamma_X(t,s) = f_X(h)$, $h = t-s$; for simplicity's sake, we often write $\gamma_X(t-s)$ or $\gamma_X(h)$ for the covariance.[17]

17: When the context is un-ambiguous.

The **(auto)correlation function** (ACF) of $\{X_t \mid t = 1, \ldots, n\}$ is given by:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \frac{\mathrm{Cov}(X_1, X_{h+1})}{\mathrm{Var}(X_1)}.$$

Note that $\rho_X(0) = 1$.

### Examples and Illustrations

**White Noise** Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance 1.

Then $\mu_Z(t) = \mathrm{E}[Z_t] = 0$ and $\gamma_Z(t,t) = f_Z(0) = \mathrm{Var}(Z_t) = 1$ for all $t$, while $\gamma_Z(t,s) = f_Z(h) = 0$ for all $t \neq s \implies h \neq 0$. Since $\gamma_Z$ only depends on $h = t - s$ and $\mu_Z \equiv 0$, $\{Z_t\}$ is (weakly) stationary.

**Random Walk** Let $\{Z_t\}$ be a sequence of i.i.d. random variables with mean 0 and variance $\sigma_Z^2$. Define $S_t = \sum_{i=1}^{\top} Z_i$. Then $\mathrm{E}[S_t] = 0$, and

$$\begin{aligned}\gamma_S(t,t+h) &= \mathrm{Cov}(S_t, S_{t+h}) = \mathrm{Cov}(S_t, S_t + Z_{t+1} + \cdots + Z_{t+h}) \\ &= \mathrm{Cov}(S_t, S_t) + \mathrm{Cov}(S_t, Z_{t+1} + \cdots + Z_{t+h}) \\ &= \mathrm{Cov}(S_t, S_t) + \mathrm{Cov}(Z_1 + \cdots + Z_t, Z_{t+1} + \cdots + Z_{t+h}) \\ &= \mathrm{Cov}(S_t, S_t) + \sum_{i=1}^{\top} \sum_{j=1}^{h} \mathrm{Cov}(Z_i, Z_{t+j}) = \mathrm{Cov}(S_t, S_t) + 0 = \mathrm{Var}(S_t).\end{aligned}$$

Since

$$\mathrm{Var}(S_t) = \mathrm{Var}(Z_1 + \cdots + Z_t) = \mathrm{Var}(Z_1) + \cdots \mathrm{Var}(Z_t) = \sigma_Z^2 + \cdots + \sigma_Z^2 = t\sigma_Z^2,$$

the autocovariance function depends on $t$ (and not on $h = t - s$), and the sequence is not (weakly) stationary.

**Model with Trend** We revisti the model $X_t = 1 + 2t + Z_t$, $t = 1, 2, \ldots$, where $\{Z_t\}$ is a sequence of i.i.d. random variables with mean $\mu_Z = \mathrm{E}[Z_t]$. Then

$$\mathrm{E}[X_t] = \mathrm{E}[1 + 2t + Z_t] = 1 + 2t + \mu_Z.$$

The mean function depends on $t$; the model is not (weakly) stationary.

**"Multiplicative" Model** Let $\{Z_t\}$ be i.i.d. with mean 0 and variance $\sigma_Z^2$. Define

$$X_t = Z_t Z_{t-1} Z_{t-2}, \quad t \geq 3.$$

Because $\mathrm{E}[Z_t] = 0$, we have

$$\sigma_Z^2 = \mathrm{Var}(Z_t) = \mathrm{E}[Z_t^2] - \mathrm{E}^2[Z_t] = \mathrm{E}[Z_t^2].$$

Since the $Z_t$ are independent of one another, we have

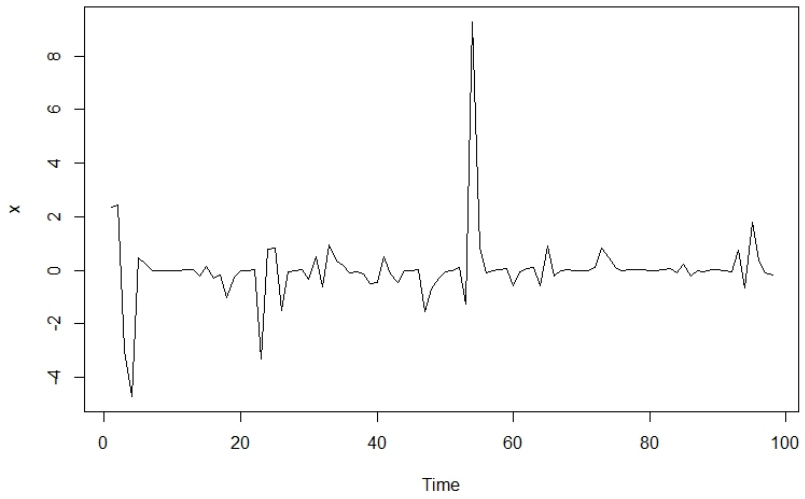
$$\begin{aligned}\mathrm{E}[X_t] &= \mathrm{E}[Z_t Z_{t-1} Z_{t-2}] = \mathrm{E}[Z_t]\mathrm{E}[Z_{t-1}]\mathrm{E}[Z_{t-2}] = 0, \quad \text{and} \\ \mathrm{Var}(X_t) &= \mathrm{E}[X_t^2] = \mathrm{E}[Z_t^2 Z_{t-1}^2 Z_{t-2}^2] = \mathrm{E}[Z_t^2]\mathrm{E}[Z_{t-1}^2]\mathrm{E}[Z_{t-2}^2] = \sigma_Z^6\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}(X_t, X_{t+1}) &= \mathrm{E}[X_t X_{t+1}] - \mathrm{E}[X_t]\mathrm{E}[X_{t+1}] \\ &= \mathrm{E}[\{Z_t Z_{t-1} Z_{t-2}\}\{Z_{t+1} Z_t Z_{t-1}\}] - 0 \\ &= \mathrm{E}[Z_{t+1}]\mathrm{E}[Z_t^2]\mathrm{E}[Z_{t-1}^2]\mathrm{E}[Z_{t-2}] = 0.\end{aligned}$$

Similarly, we have $\mathrm{Cov}(X_t, X_s) = 0$ for $t \neq s$; the model is thus (weakly) stationary.

```
z=rnorm(100)
n=length(z)
zt=z[3:n]
zt1=z[2:(n-1)]
zt2=z[1:(n-2)]
x=zt*zt1*zt2
plot.ts(x)
```




**MA(1)** Let $\{Z_t\}$ be a sequence of independent random variables with $\mu_Z \equiv 0$ and variance $\sigma_Z^2 = \mathrm{Var}(Z)$, and $\theta \in \mathbb{R}$. The MA(1) model is:

$$X_t = Z_t + \theta Z_{t-1}, \quad t \geq 2.$$

We see that $\mathrm{E}[X_t] = \mathrm{E}[Z_t + \theta Z_{t-1}] = \mathrm{E}[Z_t] + \theta \mathrm{E}[Z_{t-1}]$, and that

$$\begin{aligned}\mathrm{Var}(X_t) &= \mathrm{E}[X_t^2] = \mathrm{E}[\{Z_t + \theta Z_{t-1}\}^2] \\ &= \mathrm{E}[Z_t^2] + \theta^2 \mathrm{E}[Z_{t-1}^2] + 2\theta \underbrace{\mathrm{E}[Z_t Z_{t-1}]}_{=0} = \sigma_Z^2 + \theta^2 \sigma_Z^2 = \sigma_Z^2(1+\theta^2).\end{aligned}$$

Thus the autocovariance of MA(1) is

$$\gamma_X(t, t+h) = \gamma_X(h) = \begin{cases} \sigma_Z^2(1+\theta^2) & h = 0; \\ \sigma_Z^2 \theta & h = \pm 1; \\ 0 & |h| > 1 \end{cases}.$$

Note that $\gamma_X(t, t+h) = \gamma_X(h)$ depends only on $h$ and so a MA(1) time series is (weakly) stationary. Furthermore,

$$\rho_X(t, t+h) = \rho_X(h) = \begin{cases} 1 & h = 0; \\ \theta/(1+\theta^2) & h = \pm 1; \\ 0 & |h| > 1 \end{cases}.$$

The ACF then also only depends on $h$:

$$\rho_X(t, t+h) = \rho_X(h).$$

The set $\mathcal{T}_n$ of stationary time series of length $n$ is a vector "subspace" over $\mathbb{R}$ of the set of all independent time series.[18]

18: For a generous definition of subspace.

Indeed,

1. $\{0_t \mid t = 1, \dots, n\} \in \mathcal{T}_n$;
2. if $\{X_t \mid t = 1, \dots, n\} \in \mathcal{T}_n, \lambda \in \mathbb{R}$, then $\{\lambda X_t \mid t = 1, \dots, n\} \in \mathcal{T}_n$;
3. if $\{X_t \mid t = 1, \dots, n\}, \{Y_t \mid t = 1, \dots, n\} \in \mathcal{T}_n$ are **independent** time series, then $\{W_t = X_t + Y_t \mid t = 1, \dots, n\} \in \mathcal{T}_n$.

We only prove the third of these statements (the other two are left as exercises).

Let $\{X_t\}, \{Y_t\} \in \mathcal{T}_n$ be independent time series, with means $\mu_X$, $\mu_Y$ and autocovariance functions $\gamma_X$ and $\gamma_Y$, respectively. Set $W_t = X_t + Y_t$. Then

$$\mu_W(t) = \mathrm{E}[W_t] = \mathrm{E}[X_t + Y_t] = \mathrm{E}[X_t] + \mathrm{E}[Y_t] = \mu_X + \mu_Y (:= \mu_W)$$

and

$$\begin{aligned}
\gamma_W(t, t+h) &= \mathrm{E}[W_t W_{t+h}] - \mathrm{E}[W_t]\mathrm{E}[W_{t+h}] \\
&= \mathrm{E}[(X_t + Y_t)(X_{t+h} + Y_{t+h})] - \mu_W^2 \\
&= \mathrm{E}[X_t X_{t+h}] + \mathrm{E}[Y_t Y_{t+h}] + \underbrace{\mathrm{E}[X_t Y_{t+h}]}_{=\mathrm{E}[X_t]\mathrm{E}[Y_t]} + \underbrace{\mathrm{E}[Y_t X_{t+h}]}_{=\mathrm{E}[X_t]\mathrm{E}[Y_t]} - \mu_W^2 \\
&= \gamma_X(h) + \mu_X^2 + \gamma_Y(h) + \mu_Y^2 + \mu_X\mu_Y + \mu_X\mu_Y - (\mu_X + \mu_Y)^2 \\
&= \gamma_X(h) + \gamma_Y(h).
\end{aligned}$$

That is to say, $\{W_t\} \in \mathcal{T}_n$. ■

### 9.1.4 Partial Autocorrelation (PACF)

Let $\{X_t\} \in \mathcal{T}_n$ with $\mu_X = 0$. The **partial (auto)covariance** between $X_t$ and $X_{t+k}$ is the covariance between $X_t$ and $X_{t+k}$, where we "condition out" the intermediate time series $X_{t+1}, \dots, X_{t+k-1}$.

Assume that the random variables $X_1$ and $X_3$ from the stationary time series have the following relationship:

$$X_1 = \beta_{1,3} X_3 + Z,$$

where $\mu_Z = 0$, and $Z$ is independent of both $X_1, X_3$. Then

$$\begin{aligned}
&X_1 X_3 = \beta_{1,3} X_3^2 + Z X_3 \implies \mathrm{E}[X_1 X_3] = \beta_{1,3}\mathrm{E}[X_3^2] + \mathrm{E}[Z X_3] \\
&\implies \gamma_X(2) = \beta_{1,3}\gamma_X(0) + \mathrm{E}[Z]\mathrm{E}[X_3] \implies \gamma_X(2) = \beta_{1,3}\gamma_X(0),
\end{aligned}$$

and so

$$\beta_{1,3} = \frac{\gamma_X(2)}{\gamma_X(0)} = \rho_X(2).$$

If $Z \sim \mathcal{N}(0, \sigma_Z^2)$, we recognize $\beta_{1,3}$ as the **OLS regression parameter** when regressing $X_1$ against $X_3$.[19] Similarly, if we further assume that

$$X_2 = \beta_{2,3} X_3 + V,$$

19: Strictly speaking, if $Z$ is not normal, the OLS qualifier does not apply but the rest of the argument still works.

where $V \sim \mathcal{N}(0, \sigma_V^2)$ is independent of both $X_2, X_3$, then the OLS regression parameter when regressing $X_2$ against $X_3$ is

$$\beta_{2,3} = \frac{\gamma_X(1)}{\gamma_X(0)} = \rho_X(1).$$

The **partial (auto)correlation** (PACF) between $X_1$ and $X_2$ is the correlation between $X_1$ and $X_2$, removing the effect of $X_3$:

$$\rho_{1,2;3} = \text{Corr}(X_1 - \beta_{1,3}X_3, X_2 - \beta_{2,3}X_3).$$

Hence,

$$\rho_{1,2;3} = \frac{\text{Cov}(X_1 - \beta_{1,3}X_3, X_2 - \beta_{2,3}X_3)}{\sqrt{\text{Var}(X_1 - \beta_{1,3}X_3)}\sqrt{\text{Var}(X_2 - \beta_{2,3}X_3)}}.$$

But we have

$$\begin{aligned}
&\text{Cov}(X_1 - \beta_{1,3}X_3, X_2 - \beta_{2,3}X_3) \\
&= \text{Cov}(X_1, X_2) + \text{Cov}(\beta_{1,3}X_3, \beta_{2,3}X_3) - \text{Cov}(X_1, \beta_{2,3}X_3) - \text{Cov}(\beta_{1,3}X_3, X_2) \\
&= \gamma_X(1) + \beta_{1,3}\beta_{2,3}\text{Cov}(X_3, X_3) - \beta_{2,3}\text{Cov}(X_1, X_3) - \beta_{1,3}\text{Cov}(X_3, X_2) \\
&= \gamma_X(1) + \beta_{1,3}\beta_{2,3}\gamma_X(0) - \beta_{2,3}\gamma_X(2) - \beta_{1,3}\gamma_X(1) \\
&= \gamma_X(1) + \rho_X(2)\rho_X(1)\gamma_X(0) - \rho_X(1)\gamma_X(2) - \rho_X(2)\gamma_X(1) \\
&= \gamma_X(1) + \rho_X(2)\gamma_X(1) - \rho_X(1)\gamma_X(2) - \rho_X(2)\gamma_X(1) \\
&= \gamma_X(1) + \rho_X(2)\gamma_X(1) - \tfrac{\gamma_X(1)}{\gamma_X(0)}\gamma_X(2) - \rho_X(2)\gamma_X(1) \\
&= \gamma_X(1) + [\rho_X(2)\gamma_X(1) - \gamma_X(1)\rho_X(2)] - \rho_X(2)\gamma_X(1) = \gamma_X(1)(1 - \rho_X(2)).
\end{aligned}$$

We also have:

$$\text{Var}(X_1 - \beta_{1,3}X_3) = \gamma_X(0)\left(1 - \rho_X^2(2)\right) \quad \text{and} \quad \text{Var}(X_2 - \beta_{2,3}X_3) = \gamma_X(0)\left(1 - \rho_X^2(1)\right).$$

Thus, the partial correlation is

$$\begin{aligned}
\rho_{1,2;3} &= \frac{\gamma_X(1)(1 - \rho_X(2))}{\gamma_X(0)\sqrt{\left(1 - \rho_X^2(2)\right)\left(1 - \rho_X^2(1)\right)}} = \frac{\rho_X(1) - \rho_X(1)\rho_X(2)}{\sqrt{\left(1 - \rho_X^2(2)\right)\left(1 - \rho_X^2(1)\right)}} \\
&= \frac{\text{Corr}(X_1, X_2) - \text{Corr}(X_2, X_3) \cdot \text{Corr}(X_1, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_3)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}}.
\end{aligned}$$

**Note:** $\gamma_X(1)$, $\text{Cov}(X_1, X_2)$, and $\text{Cov}(X_2, X_3)$ are interchangeable because the time series $\{X_t\}$ is stationary; thus we have $\text{Corr}(X_1, X_2) = \text{Corr}(X_2, X_3)$.

Similarly, the partial (auto)correlation between $X_1$ and $X_3$ is the correlation between $X_1$ and $X_3$, removing the effect of $X_2$:

$$\rho_{1,3;2} = \frac{\text{Corr}(X_1, X_3) - \text{Corr}(X_1, X_2) \cdot \text{Corr}(X_2, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_2)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}}.$$

**The PACF** Given a time series $\{X_t\}$, the partial autocorrelation at lag $h$, denoted $\alpha_X(h)$,[20] is the autocorrelation between $X_t$ and $X_{t+h}$, removing the linear dependence of $X_t$ on $X_{t+1}, \ldots, X_{t+h-1}$; the function $\alpha_X$ is called the **partial autocorrelation function** (PACF).

20: Or $\alpha(h)$ if the context is clear.

Note that:

1. $\alpha(1) = \rho_X(1)$,
2. $\alpha(2) = \rho_{1,3;2}$,
3. $\alpha(3) = \rho_{1,4;2,3}$,
4. and so on.

A non-negligible aspect of the discipline involves computing the PACF for different models; we anticipate the task by providing some some calculations for a special case: the MA(1) model.

**MA(1)** Let $\{Z_t\}$ be a sequence of independent random variables with $\mu_Z \equiv 0$ and variance $\sigma_Z^2 = \text{Var}(Z)$, and $\theta \in \mathbb{R}$. The MA(1) model is $X_t = Z_t + \theta Z_{t-1}$, $t \geq 2$. We have seen that

$$\rho_X(h) = \begin{cases} 1 & h = 0; \\ \theta/(1+\theta^2) & h = \pm 1; \\ 0 & |h| > 1 \end{cases}.$$

Thus,

$$\begin{aligned} \alpha(2) &= \frac{\text{Corr}(X_1, X_3) - \text{Corr}(X_1, X_2) \cdot \text{Corr}(X_2, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_2)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}} \\ &= \frac{\rho_X(2) - \rho_X^2(1)}{\sqrt{1 - \rho_X^2(1)}\sqrt{1 - \rho_X^2(1)}} \\ &= \frac{0 - \dfrac{\theta^2}{(1+\theta^2)^2}}{1 - \dfrac{\theta^2}{(1+\theta^2)^2}} = \frac{-\theta^2}{1 + \theta^2 + \theta^4}. \end{aligned}$$

## 9.2 Estimating Model Parameters

In practice, we typically work with one of the time series' **realizations**, that is to say, the true $\mu(\cdot)$, $\gamma(\cdot)$ and $\alpha(\cdot)$ are not available to us.

### 9.2.1 Sample Statistics

As is usually the case, in statistical analysis, we can use the data at our disposal in order to estimate the model's parameters. As always, assume that $\{X_t\} \in \mathcal{T}_n$ is stationary.

**Sample Mean** The mean $\mu = \mu_X \equiv \text{E}[X_t]$ can be estimated by the **sample mean**:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

**Sample Variance** The variance $\sigma_X^2 \equiv \text{Var}(X_t) = \text{E}[(X_t - \mu)^2]$ can be estimated by the **sample variance**:

$$\widehat{\sigma}_X^2 = \widehat{\gamma}_X(0) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

**Sample (Auto)Covariance** The covariance $\gamma_X(h) = \text{E}[(X_t - \mu)(X_{t+h} - \mu)]$ (ACVF) can be estimated by the **sample (auto)covariance**:

$$\widehat{\gamma}_X(h) = \frac{1}{n-1}\sum_{t=1}^{n-h}(X_t - \overline{X})(X_{t+h} - \overline{X}).$$

**Sample (Auto)Correlation** The (auto)correlation $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$ is estimated by the **sample autocorrelation** (**sample ACF**):

$$\widehat{\rho}_X(h) = \frac{\widehat{\gamma}_X(h)}{\widehat{\gamma}_X(0)}.$$

**Sample PACF** The PACF is estimated by the **sample PACF**; for instance, since

$$\alpha(2) = \frac{\rho_X(2) - \rho_X^2(1)}{\sqrt{1-\rho_X^2(1)}\sqrt{1-\rho_X^2(1)}} = \frac{\rho_X(2) - \rho_X^2(1)}{1-\rho_X^2(1)},$$

then

$$\widehat{\alpha}(2) = \frac{\widehat{\rho}_X(2) - \widehat{\rho}_X^2(1)}{1-\widehat{\rho}_X^2(1)}.$$

### 9.2.2 Examples

**White Noise** Recall that white noise $\{Z_t\}$ is a sequence of independent random variables with mean 0 and variance 1. Then $\gamma_X(0) = \rho_X(0) = 1$ and $\gamma_X(h) = \rho_X(h) = 0$ for $h \neq 0$.

We prepare a realization of the white noise time series.

```
set.seed(1)
z = rnorm(100)
n = length(z)
(muz = mean(z))
gamma0 = sum((z-muz)^2)/(n-1)
var(z)
```

```
[1] 0.1088874
[1] 0.8067621
```

We see that the sample mean and the sample variance are near 0 and 1, respectively. We can exhibit the sample ACF using the `acf()` function.

```
zt = z[2:n]; zt1 = z[1:(n-1)]
(corr = acf(z))
```

```
autocorrelations of series 'z', by lag

     0      1      2      3      4      5      6      7      8
 1.000 -0.004 -0.027 -0.107 -0.113 -0.093 -0.125  0.065  0.043
     9     10     11     12     13     14     15     16     17
 0.026  0.025 -0.032 -0.042  0.053 -0.038 -0.022 -0.140  0.063
    18     19     20
-0.023 -0.084 -0.112
```

For instance, we can extract $\widehat{\rho}(1)$ using the following call:

```
corr$acf[2]
```

```
[1] -0.003651251
```

But we can also compute it directly:

```
gamma1 = sum((zt1-muz)*(zt-muz))/(n-1)
(rho1 = gamma1/gamma0)
```

```
[1] -0.003651251
```

The sample PACF can be obtained *via* the `pacf()` function.

```
(partial.corr = pacf(z))
```

```
Partial autocorrelations of series 'z', by lag

     1      2      3      4      5      6      7      8      9
-0.004 -0.027 -0.108 -0.116 -0.105 -0.153  0.023 -0.002 -0.025
    10     11     12     13     14     15     16     17     18
-0.005 -0.046 -0.052  0.069 -0.042 -0.039 -0.157  0.035 -0.053
    19     20
-0.121 -0.190
```

For instance, we can extract $\widehat{\alpha}(2)$ using the following call:

```
partial.corr$acf[2]
```

```
[1] -0.02703468
```

But we can also compute it directly:

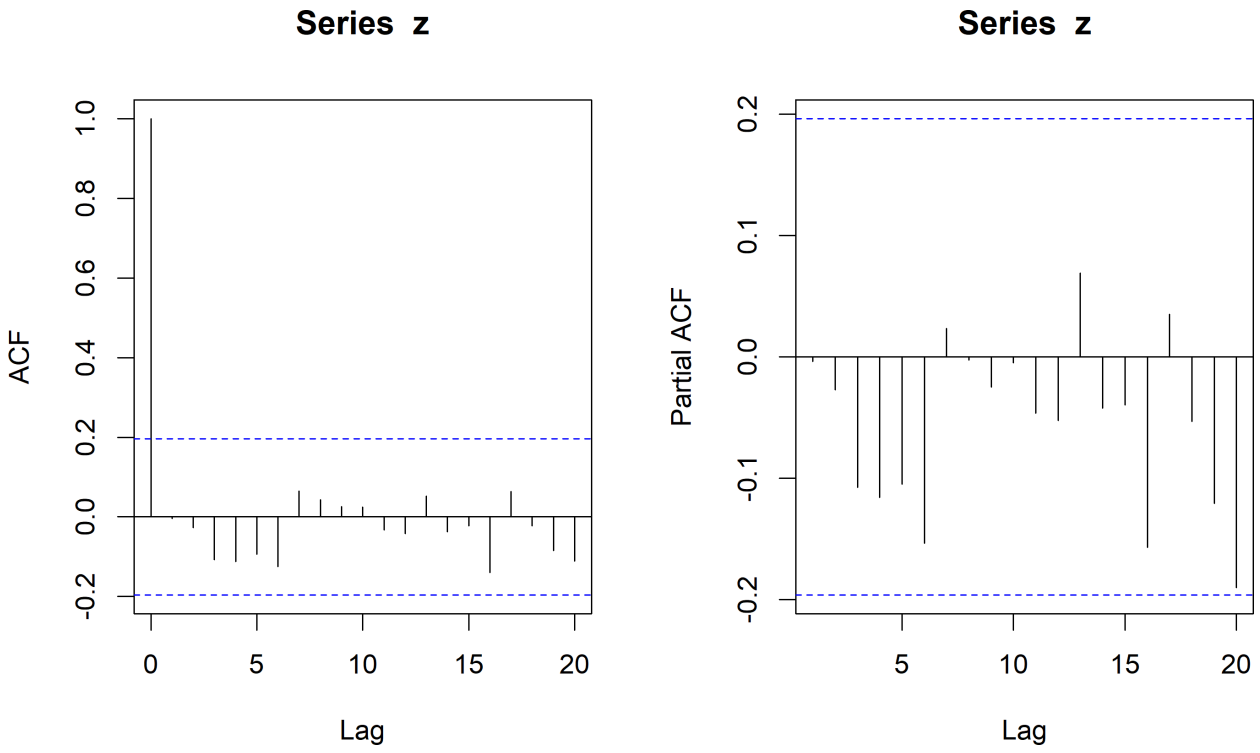
```
(alpha2 = (corr$acf[3]-(corr$acf[2])^2)/(1-(corr$acf[2])^2))
```

```
[1] -0.02703468
```

Finally, we plot the sample ACF and sample PACF of the white noise time series against the lag $h$.[21]

21: The dotted blue lines in the ACF and PACF chart indicate the thresholds beyond which the recorded values can be seen as statistically different rom zero. These lines are located at a height of $\pm\frac{1.96}{\sqrt{n}}$ (see Section 9.6 for an explanation).

```
par(mfrow=c(1,2))
acf(z); pacf(z)
```

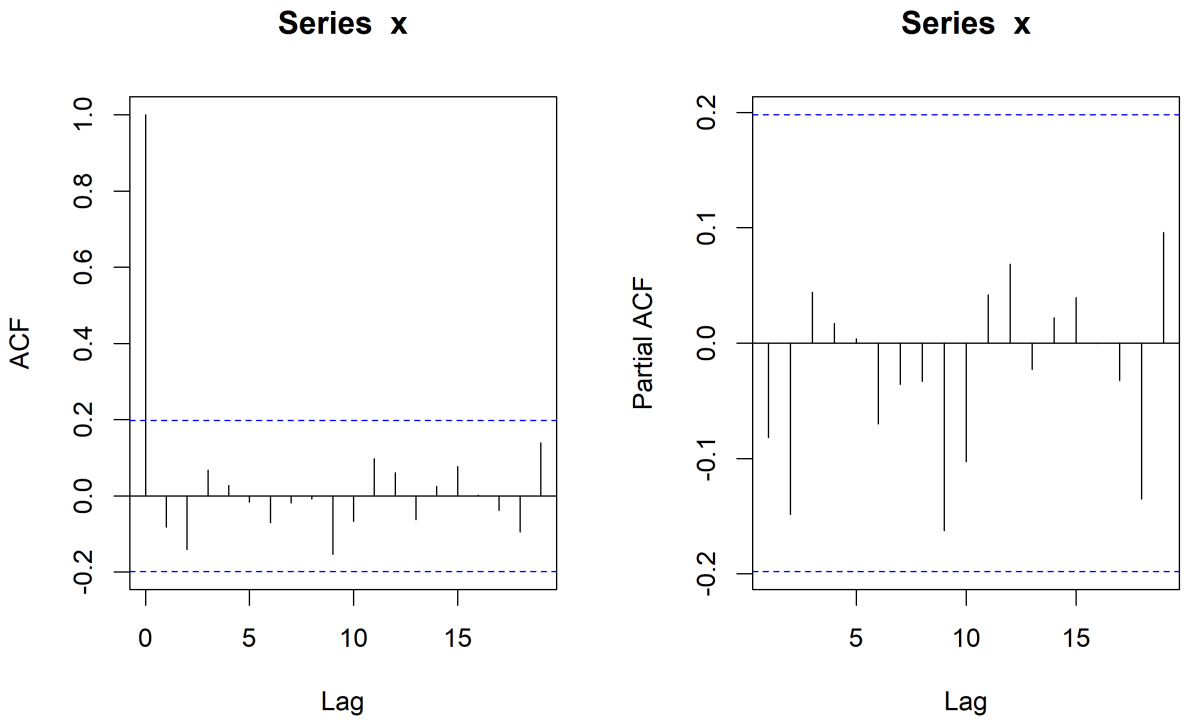



**"Multiplicative" Model** Let $\{Z_t\}$ be i.i.d. with mean 0 and variance $\sigma_Z^2$. Define

$$X_t = Z_t Z_{t-1} Z_{t-2}, \quad t \geq 3.$$

We prepare a realization of this time series, assuming that $Z_t \sim \mathcal{N}(0,1)$, and display its sample ACF and sample PACF.

```
set.seed(2)
z = rnorm(100)
n = length(z)
zt = z[3:n]; zt1 = z[2:(n-1)]; zt2 = z[1:(n-2)];
x = zt*zt1*zt2
par(mfrow=c(1,2))
acf(x)
pacf(x)
```




Are the results fundamentally different than those of the white noise time series?[22]

22: Keeping in mind that we are working with (potentially) different realizations of the respective time series.

**MA(1)** Recall MA(1) model

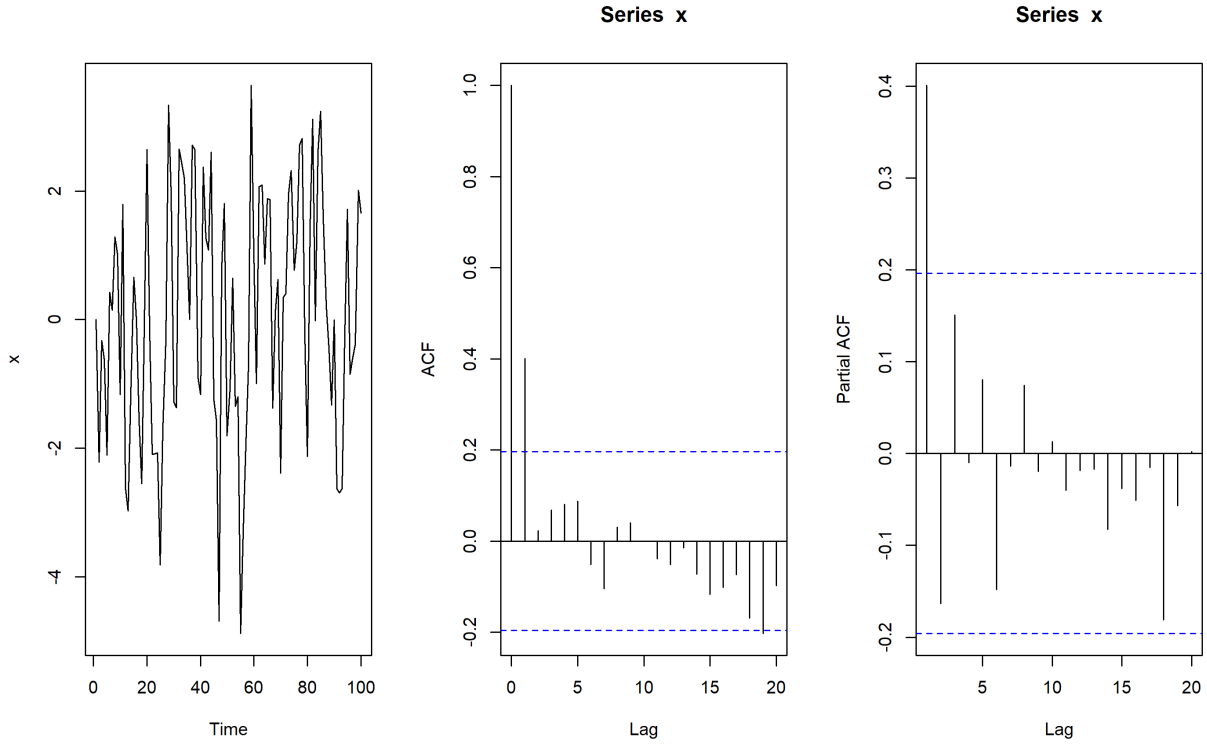
$$X_t = Z_t + \theta Z_{t-1},$$

We have derived the ACF of this model previously: $\rho_X(0) = 1$, $rho_X(1) = \theta/(1+\theta^2)$, and $rho_X(h) = 0$ for $h > 1$. We prepare a realization of MA(1) as follows:

```
set.seed(3)
z = rnorm(100,0,1)
n = length(z)
x = rep(0,n)
theta = 2
for(i in 2:n){
    x[i] = z[i] + theta*z[i-1]
}
```

Theoretically, the only non-zero values of the ACF are at $h = 0$ and $h = 1$; is that also going to be the case in the sample ACF?

```
par(mfrow=c(1,3))
plot.ts(x)
corr = acf(x)
pacf(x)
```




It is not exactly so, obviously, but $\widehat{\rho}_X(0)$ and $\widehat{\rho}_X(1)$ are substantially larger than the remaining $\widehat{\rho}_X(h)$.

The theoretical value of $\rho_X(1)$ can be computed exactly:

```
(rho1 = theta/(1+theta^2))
```

```
[1] 0.4
```

How does that compare to the sample estimate $\widehat{\rho}_X(1)$?

```
corr[1]
```

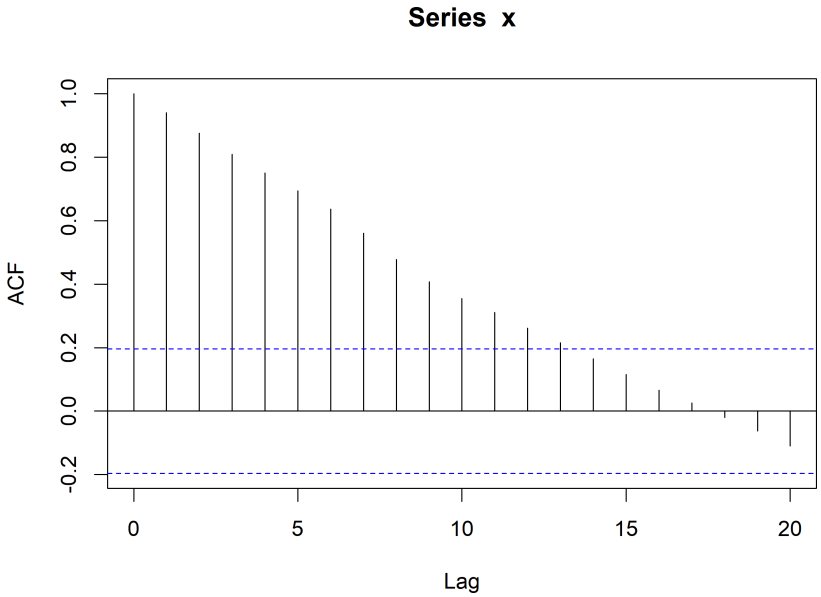
```
autocorrelations of series 'x', by lag

    1
0.401
```

Pretty darn close, we'd say.

**Random Walk** Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance $\sigma_Z^2$, and set $X_t = \sum_{i=1}^{\top} Z_i$.

We prepare a realization of a random walk and display its sample ACF.

```
set.seed(4)
z=rnorm(100)
x=cumsum(z)
acf(x)
```




Well, that is certainly rather different than the other sample ACF we have studied so far... but perhaps it should not come as a surprise when we remember that random walks are **not** stationary.

---

Time series analysis, then, requires first that the time series be decomposed into its

- **stationary** (random) and
- **non-stationary** components (trend, level shifts, seasonality, etc.).

Next, we try to identify the nature of the random component *via* a model (using tools like the sample ACF and the sample PACF).

We will discuss commonly-encountered models in the following sections.

## 9.3 ARMA Models

In this section, we assume that the time series $\{X_t\} \in \mathcal{T}_n$ is stationary. We will discuss the simplest of the non-trivial time series analysis models, the **auto-regressive moving average** model (ARMA).

### 9.3.1 Linear Processes/Moving Averages

Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance $\text{Var}(Z_t) = \text{E}[Z_t^2] = \sigma_Z^2$.[23] Let $\psi_j$, $j \geq 0$, be a sequence of

23: In the rest of this section, the assumptions on $\{Z_t\}$ will be taken for granted.

constants such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Then

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

is called a **linear process** or a **moving average**.[24]

The condition $\sum_{j=0}^{\infty} |\psi_j| < \infty$ ensures that the infinite series converges:

$$\mathrm{E}[|X_t|] \leq \sum_{j=0}^{\infty} |\psi_j| \mathrm{E}[|Z_{t-j}|] = \mathrm{E}[|Z_0|] \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Note that this condition is not necessary, however.[25]

**Lemma:** a linear process is a stationary time series with $\mathrm{E}[X_t] = 0$ and

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}.$$

**Proof:** if we assume that the convergence of the infinite sum of random variables is "uniform", then since $\mathrm{E}[Z_t] \equiv 0$, we have

$$\mathrm{E}[X_t] = \mathrm{E}\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j}\right] = \sum_{j=0}^{\infty} \psi_j \mathrm{E}[Z_{t-j}] = 0;$$

that this is indeed the case is not trivial to show.[26]

We interchange $\sum$ and $\mathrm{E}[\cdot]$ once more,[27] to obtain:

$$\begin{aligned}\gamma_X(h) &= \mathrm{E}[X_t X_{t+h}] - \mathrm{E}[X_t]\mathrm{E}[X_{t+h}] = \mathrm{E}\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j} \sum_{i=0}^{\infty} \psi_i Z_{t+h-i}\right] - 0 \\ &= \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \psi_j \psi_i \mathrm{E}[Z_{t-j} Z_{t+h-i}].\end{aligned}$$

Since the noise variables $Z_t$ are independent, the only terms that contributes to the double sum are those for which $j = i - h$. Hence, the double sum collapses to a single sum:

$$\gamma_X(h) = \sum_{j=0}^{\infty} \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \mathrm{E}[Z_{t-j}^2] = \sum_{j=0}^{\infty} \sum_{j=0}^{\infty} \psi_j \psi_{j+h} (\mu_Z^2 + \sigma_Z^2).$$

As $\mu_Z = 0$, we obtain the desired conclusion. ■

**AR(1)** The auto-regressive model of order 1, AR(1), with parameter $\phi$ takes the form

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

If $|\phi| < 1$, AR(1) is the linear process with $\psi_j = \phi^j$; according to the preceding lemma, we have

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} = \sigma_Z^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \sigma_Z^2 \phi^h \sum_{j=0}^{\infty} (\phi^2)^j = \sigma_Z^2 \frac{\phi^h}{1-\phi^2},$$

24: The terms **causal moving average** or **one-sided moving average** are also used, to indicate that the sum starts at a finite index $j$; a **non-causal linear process** would take the form $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, but we need a bi-directional sequence $\{Z_t \mid t \in \mathbb{Z}\}$ of independent random variables with mean 0 and variance $\sigma_Z^2$ for this to make sense.

25: $\sum_{j=0}^{\infty} |\psi_j| < \infty \implies \sum_{j=0}^{\infty} \psi_j^2 < \infty$.

26: The proof is outside the scope of these notes; we will take it as valid, sight unseen.

27: Again, because of the $L_2-$convergence of the $\psi-$series.

using the formula for the sum of a geometric series.[28]

28: Note that the sum does not converge for $|\phi| \geq 1$.

**MA($q$)** The moving average model of order $q$, MR($q$), with parameter vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q)$ takes the form

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

This is a linear process with $\psi_0 = 1$, $\psi_1 = \theta_1, \ldots, \psi_q = \theta_q$, and $\psi_j = 0$, for all $j > q$;[29] according to the preceding lemma, we have

29: We set $\theta_0 = 1$, by convention.

$$\gamma_X(h) = \begin{cases} \sigma_Z^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+h} = \sigma_Z^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & h = 0, \ldots, q \\ 0 & h > q. \end{cases}$$

### 9.3.2 ARMA in General

In order to define the general ARMA model, we introduce a crucial element of time series analysis.

**Backward Shift Operator** Recall that the difference operator $\nabla$ acts on a time series $\{X_t\}$ according to

$$\nabla X_t = X_t - X_{t-1}, \qquad \text{as long as } X_{t-1} \text{ exists.}$$

The **backward shift operator** $B$ is defined by

$$BX_t = (1 - \nabla)X_t = X_t - (X_t - X_{t-1}) = X_{t-1}.$$

It is easy to show (by induction, say) that $B^k X_t = X_{t-k}$, for all $k$ for which $X_{t-k}$ exists.

**AR(1)** If

$$X_t = \phi X_{t-1} + Z_t,$$

then, by formal manipulations of the expressions, we have

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t = \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t, \\ &= \phi^3 X_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t = \cdots \\ &= \cdots + \phi^4 Z_{t-4} + \phi^3 Z_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t, \end{aligned}$$

which we recognize as the AR(1) process.[30]

30: Convergence still requires $|\phi| < 1$.

Equivalently, if we set $\phi(x) = 1 - \phi x$, then AR(1) rewrites as:

$$X_t - \phi B X_t = Z_t \iff (1 - \phi B)X_t = Z_t \iff \phi(B)X_t = Z_t.$$

**MA(1)** Recall that MA(1) is the linear process

$$X_t = Z_t + \theta Z_{t-1},$$

where the $Z_t$ are as in AR(1) above. If we set $\theta(z) = 1 + \theta z$, then MA(1) rewrites as:

$$X_t = Z_t + \theta B Z_t \iff X_t = (1 + \theta B) Z_t \iff X_t = \theta(B) Z_t.$$

**ARMA(1, 1)** We can use $\phi(x)$ and $\theta(z)$ to define a new model:

$$\phi(B) X_t = \theta(B) Z_t,$$

which upon expansion becomes

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}.$$

This model combines the AR(1) and MA(1) models, which is why we call it an **auto-regressive moving average model of order** $(1, 1)$.

**ARMA($p, q$)** Let $\{Z_t\}$ be a sequence of independent random variables with mean 0 and variance $\text{Var}(Z_t) = \text{E}[Z_t^2] = \sigma_Z^2$. A time series $\{X_t\}$ is an **auto-regressive moving average model of order** $(p, q)$, denoted ARMA$(p, q)$, if it solves the equation

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots \theta_q Z_{t-q}.$$

Equivalently,

$$\phi(B) X_t = \theta(B) Z_t,$$

where

$$\phi(x) = 1 - \phi_1 x - \cdots - \phi_p x^p, \quad \text{and} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

are the **auto-regressive** and **moving average** polynomials, respectively.

The statement "ARMA$(p, q)$ solves the equation" means that we can write $X_t$ as a stationary linear process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

where the coefficients $\psi_j$ depend on the model parameters $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$.

While ARMA models do not need to be causal, we will only be interested in causal models:

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

### 9.3.3 Stationarity and Causality

A **stationary solution** for ARMA$(p, q)$ exists whenever the auto-regressive polynomial

$$\phi(z) = 1 - \phi_1 x - \cdots - \phi_p x^p$$

**has no root on the complex unit circle**, which is to say that of $\phi$'s roots satisfy $|x| \neq 1$.

A causal solution for ARMA$(p, q)$ exists whenever the roots of the auto-regressive polynomial auto-regressive polynomial $\phi(x)$ **all lie outside the complex open unit disk**, which is to say that all of $\phi$'s roots satisfy $|x| > 1$.

#### Examples

1. The auto-regressive polynomial of the AR(1) model
$$X_t - 1.1X_{t-1} = Z_t$$
is $\phi(x) = 1 - 1.1x$; its only root is at $x_0 = 1/1.1$, for which $|x_0| < 1$. Thus we can write $X_t$ as a stationary linear process, but there are no causal solution.
2. The model $X_t - 0.1X_{t-1} = Z_t$ is both stationary and causal.
3. The model $X_t - X_{t-1} = Z_t$ is causal but non-stationary; its auto-regressive polynomial $\phi(x)$ only has a root at $x = 1$.
4. Consider the AR(2) process $X_t - 0.1X_{t-1} - 0.4X_{t-2} = Z_t$. Equivalently, we can write $X_t - 0.1BX_t - 0.4B^2X_t = Z_t$; its auto-regressive polynomial is thus
$$\phi(x) = 1 - 0.1x - 0.4x^2,$$
whose roots are $x_1 \approx 1.46$ and $x_2 \approx -1.71$. Both of these roots have modulus larger than one, so the process is causal and there is a stationary solution.
5. Consider the AR(2) process $(1-B-B^2)X_t = Z_t$. The auto-regressive polynomial is
$$\phi(x) = 1 - x - x^2,$$
whose roots are $x_{1,2} = (-1 \pm i\sqrt{3})/2$. The modulus is 1 and so there are no stationary solution (but the process is causal).
6. Consider the AR(2) process $X_t - 0.1X_{t-1} + 0.4X_{t-2} = Z_t$. The auto-regressive polynomial is
$$\phi(x) = 1 - 0.2x + 0.4x^2,$$
whose only roots are imaginary:
$$x_{1,2} = \frac{0.1 \pm i\sqrt{1.56}}{0.8} = 0.25 \pm 0.1561249500i.$$
Both roots have the same modulus which is $\approx 1.58$; this is larger than 1 so the linear process is stationary and causal.
7. Consider the AR(2) process $X_t - \phi X_{t-1} - \phi X_{t-2} = Z_t$; its auto-regressive polynomial is
$$\phi(x) = 1 - \phi x - \phi x^2,$$
whose roots are
$$x_{1,2}(\phi) = -\frac{\phi \pm \sqrt{\phi^2 + 4\phi}}{2\phi}.$$
Then $\Delta = \phi^2 + 4\phi = \phi(\phi + 4) > 0$ if $\phi < -4$ and $\phi > 0$, so the roots are real when $\phi \notin [-4, 0]$; over $(-4, 0)$, the roots are complex

conjugates, with

$$|x_{1,2}(\phi)| = \left|\frac{1}{2} \pm i\frac{\sqrt{-\phi^2-4\phi}}{2\phi}\right| = \sqrt{\frac{1}{4} + \frac{(-\phi^2-4\phi)}{4\phi^2}} = \sqrt{-\frac{1}{\phi}}.$$

We seek the instances where $|x_{1,2}(\phi)| = 1$.

a) When $\phi \notin [-4, 0]$, $x_{1,2}(\phi) = \pm 1$ if and only if

$$-\frac{\phi \pm \sqrt{\phi^2+4\phi}}{2\phi} = \pm 1 \iff \phi \pm \sqrt{\phi^2+4\phi} = \pm 2\phi$$
$$\iff \phi \pm 2\phi = \pm\sqrt{\phi^2+4\phi},$$

that is, $-\phi = \pm\sqrt{\phi^2+4\phi}$ or $3\phi = \pm\sqrt{\phi^2+4\phi}$. Squaring on both sides yields $\phi^2 = \phi^2+4\phi$ or $9\phi^2 = \phi^2+4\phi$; this becomes $\phi = 0$, which we must reject as it is not in the domain of $x_{1,2}(\phi)$, or $\phi = 1/2$, which is.

b) When $\phi \in (-4, 0)$, $|x_{1,2}(\phi)| = 1$ if and only if $\sqrt{-1/\phi} = 1$, so that $-1/\phi = 1$, or $\phi = -1$.

The situation is summarized in Figure 9.5.

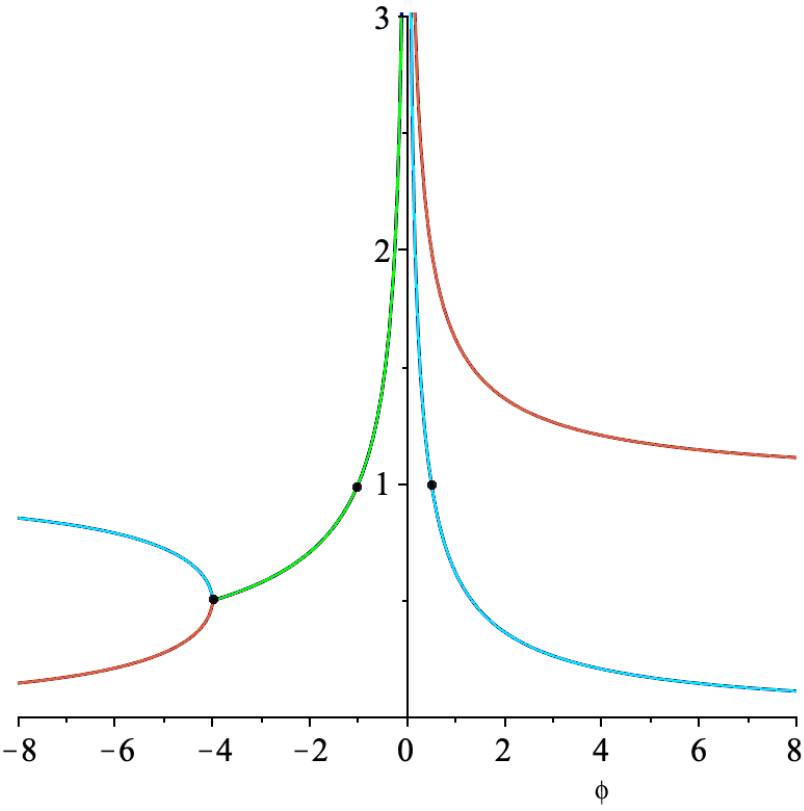



**Figure 9.5:** Modulus of the roots of the quadratic polynomial $\phi(x) = 1-\phi x-\phi x^2$ as a function of $\phi$; the roots are real and distinct when $\phi < -4$ or $\phi > 0$ (red, blue); they are complex conjugates when $-4 < \phi < 0$ (green). The corresponding linear process is causal and stationary when the modulus is larger than or equal to 1 for both roots; by piecewise continuity of the modulii, we see that this is the case for $\phi \in [-2, 0) \cup (0, 1/2]$.

### 9.3.4 Linear Representation

Given an ARMA($p$, $q$) model, how do we **represent** it as a linear process? There is no easy way to do this in the general case, but we will study some basic models.

**MA($q$)** If $p = 0$, then an ARMA$(0, q)$ model is simply an MA$(q)$ model, and its linear representation is trivial:

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

with $\psi_0 = 1$, $\psi_1 = \theta_1, \ldots, \psi_q = \theta_q$, and $\psi_k = 0$ for all $k > q$.

As $q$ is finite,

$$\sum_{j=0}^{\infty} |\psi_j| = 1 + |\theta_1| + \cdots + |\theta_q| < \infty.$$

**AR(1)** The simplest auto-regressive model is obtained by setting $p = 1$ and $q = 0$ in ARMA$(p, q)$:

$$\phi(B)X_t = Z_t,$$

where the auto-regressive polynomial is $\phi(x) = 1 - \phi x$. Define

$$\chi(x) = \frac{1}{\phi(x)} = \frac{1}{1 - \phi x}.$$

This function has a power series expansion:

$$\chi(x) = \frac{1}{1 - \phi x} = \sum_{j=0}^{\infty} \phi^j x^j,$$

which we know converges whenever $|\phi| < 1$. Multiplying the original model on both sides by $\chi(B)$ yields:

$$\chi(B)\phi(B)X_t = \chi(B)Z_t \implies X_t = \chi(B)Z_t,$$

since $\chi(x)\phi(x) = 1$ for all $x$, by construction. Thus, the linear representation of AR(1) is

$$X_t = \chi(B)Z_t = \sum_{j=0}^{\infty} \phi^j B^j Z_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j},$$

a formula we have seen before.

We note that the formal computation above only yields a **causal** linear representation when $|\phi| < 1$.[31]

31: If $|\phi| > 1$, we one can still represent the process linearly, but it is not causal.

**ARMA(1, 1)** What can we say if $p = 1$ and $q = 1$, that is, if

$$\phi(B)X_t = \theta(B)Z_t,$$

where $\phi(x) = 1 - \phi x$ and $\theta(z) = 1 + \theta z$?

We once again define

$$\chi(x) = \frac{1}{\phi(x)} = \frac{1}{1 - \phi x} = \sum_{j=0}^{\infty} \phi^j x^j.$$

Multiplying the original model on both sides by $\chi(B)$ yields:

$$\chi(B)\phi(B)X_t = \chi(B)\theta(B)Z_t, \implies X_t = \chi(B)\theta(B)Z_t,$$

since $\chi(x)\phi(x) = 1$ for all $x$. In other words,

$$X_t = \chi(B)\theta(B)Z_t = \sum_{j=0}^{\infty} \phi^j B^j (1 + \theta B) Z_t = \sum_{j=0}^{\infty} \phi^j B^j Z_t + \theta \sum_{j=0}^{\infty} \phi^j B^{j+1} Z_t$$
$$= \sum_{j=0}^{\infty} \phi^j Z_{t-j} + \theta \sum_{j=0}^{\infty} \phi^j Z_{t-(j+1)}.$$

But we would like $X_t$ to take the form $\sum_{j=0}^{\infty} \psi_j Z_{t-j}$, that is, we want:

$$\sum_{j=0}^{\infty} \psi_j Z_{t-j} = \sum_{j=0}^{\infty} \phi^j Z_{t-j} + \theta \sum_{j=0}^{\infty} \phi^j Z_{t-j-1}.$$

We rewrite this equation as:

$$\psi_0 Z_t + \sum_{j=1}^{\infty} \psi_j Z_{t-j} = \phi^0 Z_t + \sum_{j=1}^{\infty} \phi^j Z_{t-j} + \theta \sum_{j=1}^{\infty} \phi^{j-1}) Z_{t-j}$$
$$= \phi^0 Z_t + \sum_{j=1}^{\infty} (\phi^j + \theta\phi^{j-1}) Z_{t-j}.$$

The linear representation of ARMA(1,1) is thus

$$\psi_0 = 1, \qquad \psi_j = \phi^{j-1}(\phi + \theta), \quad j \geq 1;$$

This formula was obtained under the assumptions that $|\phi| < 1$,[32] and that $\phi + \theta \neq 0$.[33]

32: To insure the convergence of the power series representation of $\chi(x)$.

33: Otherwise, $X_t = Z_t$ for all $t$.

**ARMA$(1, q)$** The procedure for ARMA$(1, q)$ works in much the same way as it did for ARMA$(1, 1)$.

**AR$(p)$** The general procedure for AR$(p)$, $p \geq 2$, is much more involved; we will not discuss it.

### 9.3.5 Autocovariance Function

The simplest ways to obtain the ACVF of an ARMA model either use the model's linear representation or a recursive method.

**MA$(q)$ and AR$(1)$** The linear representation of the MA$(q)$ model is trivial; for AR(1), we use the linear representation from Section 9.3.1. In both cases, we used the Lemma in that section to compute each model's ACVF (see p. 9.3.2).

**ARMA$(1, 1)$** For this special case (and for ARMA$(1,q)$ in general), we also use the linear representation from Section 9.3.4 and the Lemma from Section 9.3.1 to obtain the ACVF.

Specifically, since $\psi_0 = 1$, $\psi_j = \phi^{j-1}(\phi+\theta)$, $j \geq 1$, and $|\phi| < 1$, we have

$$\begin{aligned}\gamma_X(0) &= \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma_Z^2 \psi_0^2 + \sigma_Z^2 \sum_{j=1}^{\infty} \psi_j^2 \\ &= \sigma_Z^2 + \sigma_Z^2 \sum_{j=1}^{\infty} (\phi^{j-1})^2 (\phi+\theta)^2 \\ &= \sigma_Z^2 \left[1 + (\phi+\theta)^2 \sum_{j=1}^{\infty} \phi^{2(j-1)}\right] = \sigma_Z^2 \left[1 + \frac{(\phi+\theta)^2}{1-\phi^2}\right].\end{aligned}$$

Similarly,

$$\begin{aligned}\gamma_X(1) &= \sigma_Z^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} = \sigma_Z^2 \psi_1 + \sigma_Z^2 \sum_{j=1}^{\infty} \psi_j \psi_{j+1} \\ &= \sigma_Z^2 (\phi+\theta) + \sigma_Z^2 \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)\phi^j(\phi+\theta) \\ &= \sigma_Z^2 \left[(\phi+\theta) + \frac{1}{\phi}(\phi+\theta)^2 \sum_{j=1}^{\infty} \phi^{2j}\right] = \sigma_Z^2 \left[(\phi+\theta) + \phi \frac{(\phi+\theta)^2}{1-\phi^2}\right].\end{aligned}$$

For a general $h \geq 1$, note first that

$$\psi_0 \psi_h = \psi_h = \phi^{h-1}(\phi+\theta) = \phi^{h-1}\phi^{1-1}(\phi+\theta) = \phi^{h-1}\psi_1 = \phi^{h-1}\psi_0\psi_1;$$

if $j \geq 1$, we also have

$$\begin{aligned}\psi_j \psi_{j+h} &= \phi^{j-1}(\phi+\theta)\phi^{j+h-1}(\phi+\theta) = \phi^{h-1}\left[\phi^{j-1}(\phi+\theta)\phi^j(\phi+\theta)\right] \\ &= \phi^{h-1}\psi_j\psi_{j+1}.\end{aligned}$$

Thus, $\gamma_X(h) = \phi^{h-1}\gamma_X(1)$ for $h \geq 1$, and so

$$\gamma_X(h) = \begin{cases} \sigma_Z^2 \left[1 + \dfrac{(\phi+\theta)^2}{1-\phi^2}\right] & h = 0, \\ \sigma_Z^2 \phi^{h-1} \left[(\phi+\theta) + \phi \dfrac{(\phi+\theta)^2}{1-\phi^2}\right] & h \geq 1. \end{cases}$$

**AR(1)** We can obtain $\gamma_X(h)$ for AR(1) by setting $\theta = 0$ in the the ACVF for the ARMA(1, 1) model, but we will illustrate a **recursive method** that generalizes to AR($p$) or general ARMA($p, q$) models with $p \geq 2$.

Let $h \in \mathbb{N}$. We start by multiplying the AR(1) equation $X_t = \phi X_{t-1} + Z_t$ by $X_{t-h}$ on both sides and applying the expectation operator to obtain:

$$\mathrm{E}[X_t X_{t-h}] = \phi \mathrm{E}[X_{t-1} X_{t-h}] + \mathrm{E}[Z_t X_{t-h}].$$

By definition, $\gamma_X(h) = \mathrm{E}[X_t X_{t-h}] - \mathrm{E}[X_t]\mathrm{E}[X_{t-h}]$. But $\mathrm{E}[X_t] = 0$ for all $t$ as $\{X_t\}$ is assumed to be stationary; thus $\mathrm{E}[X_t X_{t-h}] = \gamma_X(h)$ and $\mathrm{E}[X_{t-1} X_{t-h}] = \gamma_X(h-1)$.

For all $h \geq 1$ we know that $Z_t$ is independent of $X_{t-h}$, which is most easily seen with the linear representation of AR(1): $X_{t-h} = \sum_{j=0}^{\infty} \phi^j Z_{t-h-j}$.[34]

Thus, $\mathrm{E}[Z_t X_{t-h}] = \mathrm{E}[Z_t]\mathrm{E}[X_{t-h}] = 0$, and the AR(1) equation is equiva-

34: Note that this would not be the case if we had multiplied by $X_{t+h}$ to start with.

lent to the recursive formula:

$$\gamma_X(h) = \mathrm{E}[X_t X_{t-h}] = \phi \mathrm{E}[X_{t-1} X_{t-h}] = \phi \gamma_X(h-1), \quad h \geq 1,$$

or, by induction:

$$\gamma_X(h) = \phi^{h-1} \gamma_X(0), \qquad h \geq 1.$$

We start the recursion by computing $\gamma_X(0) = \mathrm{Var}(X_t) = \sigma_X^2$. We have

$$\mathrm{Var}(X_t) = \phi^2 \mathrm{Var}(X_{t-1}) + \mathrm{Var}(Z_t),$$

again, since $X_{t-1}$ and $Z_t$ are independent.

As $X_t$ is stationary, $\mathrm{Var}(X_t) = \mathrm{Var}(X_{t-1})$ for all $t$ and we have

$$\sigma_X^2 = \phi^2 \sigma_X^2 + \sigma_Z^2.$$

Solving for $\sigma_X^2$ yields:

$$\sigma_X^2 = \frac{\sigma_Z^2}{1 - \phi^2}.$$

Finally

$$\gamma_X(h) = \phi^h \gamma_X(0) = \sigma_Z^2 \frac{\phi^h}{1 - \phi^2},$$

which agrees with the ACVF that was calculated in Section 9.3.1.

**AR(2)** This model's equation is $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. We use a similar approach: we multiply both sides by $X_{t-h}$ and apply the expectation operator to obtain:

$$\mathrm{E}[X_t X_{t-h}] = \phi_1 \mathrm{E}[X_{t-1} X_{t-h}] + \phi_2 \mathrm{E}[X_{t-2} X_{t-h}] + \mathrm{E}[Z_t X_{t-h}].$$

An argument similar to the one presented for AR(1) yields the AR(2) recursion formula:

$$\gamma_X(h) = \phi_1 \gamma_X(h-1) + \phi_2 \gamma_X(h-2), \quad h \geq 2.$$

We start the recursion by computing $\gamma_X(0) = \mathrm{Var}(X_t) = \sigma_X^2$ and $\gamma_X(1)$.

To do so, we multiply the AR(2) equation by $X_{t-1}$ and once again apply the expectation operator to get:

$$\mathrm{E}[X_t X_{t-1}] = \phi_1 \mathrm{E}[X_{t-1}^2] + \phi_2 \mathrm{E}[X_{t-2} X_{t-1}] + \underbrace{\mathrm{E}[Z_t X_{t-1}]}_{=0},$$

so that

$$\gamma_X(1) = \phi_1 \gamma_X(0) + \phi_2 \gamma_X(1) \implies \gamma_X(1) \frac{1 - \phi_2}{\phi_1} = \gamma_X(0).$$

Next, we multiply the AR(2) equation by $X_t$ and apply the expectation operator one last time to get:

$$\mathrm{E}[X_t^2] = \phi_1 \mathrm{E}[X_{t-1} X_t] + \phi_2 \mathrm{E}[X_{t-2} X_t] + \mathrm{E}[Z_t X_t].$$

But $Z_t$ and $X_t$ are **not** independent; in fact,

$$\mathrm{E}[Z_t X_t] = \mathrm{E}[Z_t(\phi_1 X_{t-1}+\phi_2 X_{t-2}+Z_t)] = \phi_1 \mathrm{E}[Z_t X_t]+\phi_2 \mathrm{E}[Z_t X_{t-2}]+\mathrm{E}[Z_t^2] = \sigma_Z^2,$$

and so

$$\gamma_X(0) = \phi_1 \gamma_X(1) + \phi_2 \gamma_X(2) + \sigma_Z^2.$$

However, we know that

$$\gamma_X(2) = \phi_1 \gamma_X(1) + \phi_2 \gamma_X(0)$$

from the AR(2) recursion formula, with $h = 2$; we can substitute this expression into the equation for $\gamma_X(0)$ to obtain:

$$\gamma_X(0) = \phi_1 \gamma_X(1) + \phi_2 \left\{\phi_1 \gamma_X(1) + \phi_2 \gamma_X(0)\right\} + \sigma_Z^2,$$

which yields:

$$\gamma_X(h) = \phi_1 \gamma_X(h-1) + \phi_2 \gamma_X(h-2), \qquad h \geq 2,$$

$$\gamma_X(1) = \sigma_Z^2 \frac{\phi_1}{(1+\phi_2)\left\{(1-\phi_2)^2 - \phi_1^2\right\}},$$

$$\gamma_X(0) = \sigma_Z^2 \frac{1-\phi_2}{(1+\phi_2)\left\{(1-\phi_2)^2 - \phi_1^2\right\}}.$$

We can perform a sanity check, by letting $\phi_2 = 0$, $\phi_1 = \phi$; the last two formulas reduce to $\gamma_X(0)$ and $\gamma_X(1)$ for AR(1).[35]

35: It is easy to see that the recursive formula for the ACVF of AR($p$) takes the form:

$$\gamma_X(h) = \sum_{j=1}^{p} \phi_j \gamma_X(h-j).$$

### 9.3.6 Partial Autocorrelation Function

The **partial autocorrelation** of a time series $\{X_t\}$ **at lag** $h$, denoted by $\alpha(h)$, is the autocorrelation between $X_t$ and $X_{t+h}$, after removing the linear dependence of $X_t$ on $X_{t+1}, \ldots, X_{t+h-1}$.

**MA(1)** We have already calculated $\alpha(2)$ for MA(1); for a general $h \in \mathbb{N}$, it can be shown that the PACF is:

$$\alpha(h) = \frac{-(-\theta)^h}{1+\theta^2+\cdots+\theta^{2h}}.$$

Since the denominator is always positive, we see that MA(1)'s PACF has an oscillating behaviour, but that it tapers to 0 when $h \to \infty$.

**AR(1)** The PACF for the AR(1) model $X_t = \phi X_{t-1} + Z_t$ is such that

$$\alpha(1) = \rho_X(1) = \phi, \quad \alpha(2) = \mathrm{Corr}(X_t, X_{t+2} - \phi X_{t+1}) = \mathrm{Corr}(X_t, Z_{t+2}) = 0.$$

It turns out that this PACF behaviour is typical of AR($p$) models.

**Theorem:** consider a stationary AR($p$) time series. Then

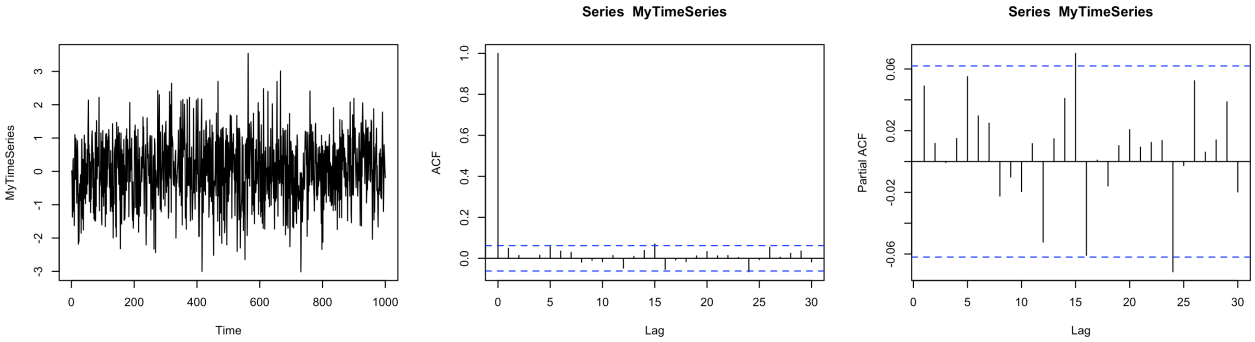
$$\alpha(h) = 0, \quad h = p+1, p+2, \ldots.$$

**Examples** In what follows, we generate a realization of various ARMA($p, q$) models through package `tseries`' `arima()` function, and display the sample ACF and sample PACF plots.[36] Do the graphs have the expected characteristics?

36: The examples will also showcase the syntax of the simulation function.

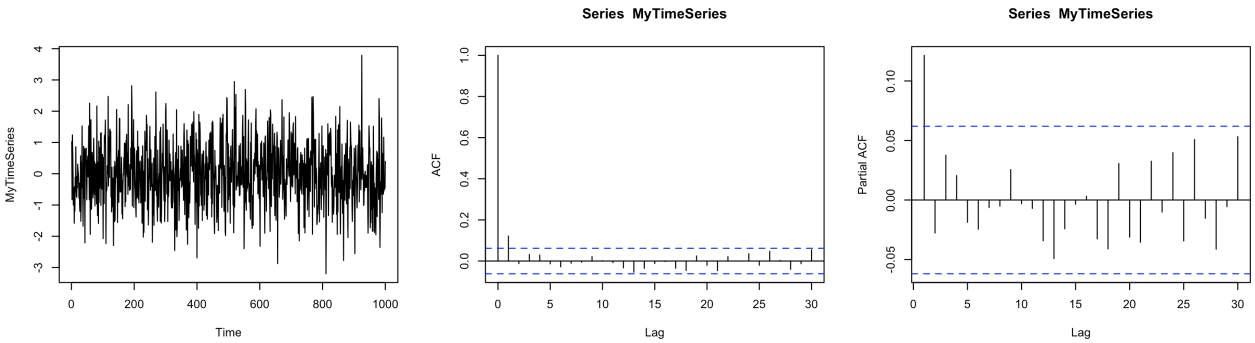
### White Noise

```
library(tseries)
set.seed(10)
MyTimeSeries = arima.sim(model = list(ar = c()),
                         n = 1000,
                         rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```

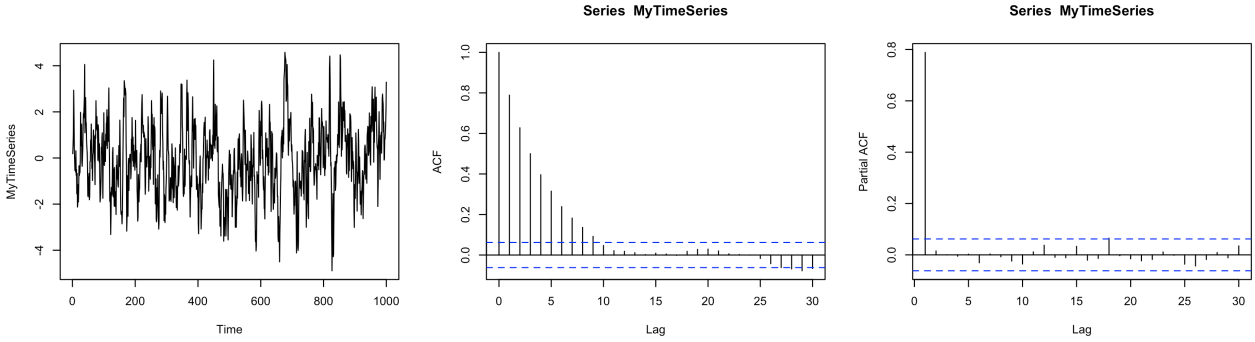



### AR(1)

```
set.seed(11)
MyTimeSeries = arima.sim(model = list(ar = c(0.1)),
                         n = 1000,
                         rand.gen = rnorm);
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



```
set.seed(12)
MyTimeSeries = arima.sim(model = list(ar = c(0.8)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```

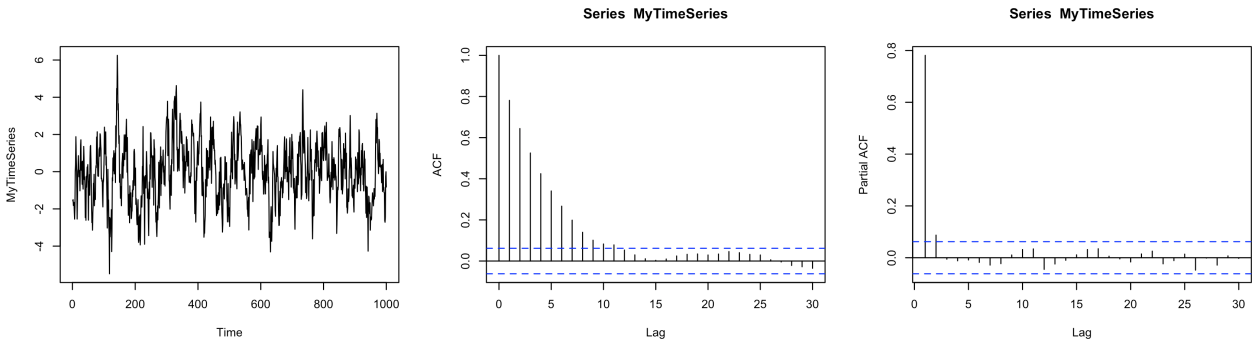



```
set.seed(14)
MyTimeSeries = arima.sim(model = list(ar = c(1.1)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```

```
Error in arima.sim(model = list(ar = c(1.1)), n = 1000, rand.gen = rnorm) :
  'ar' part of model is not stationary
```

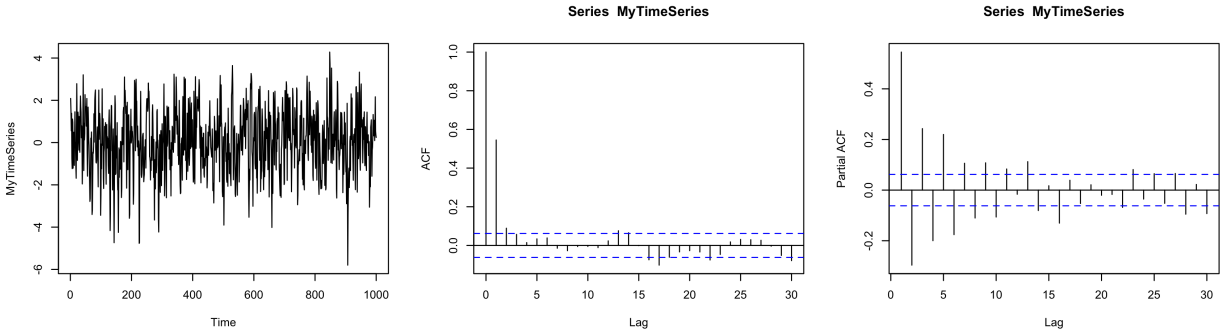
**AR(2)**

```
set.seed(13)
MyTimeSeries = arima.sim(model = list(ar = c(0.7,0.1)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```



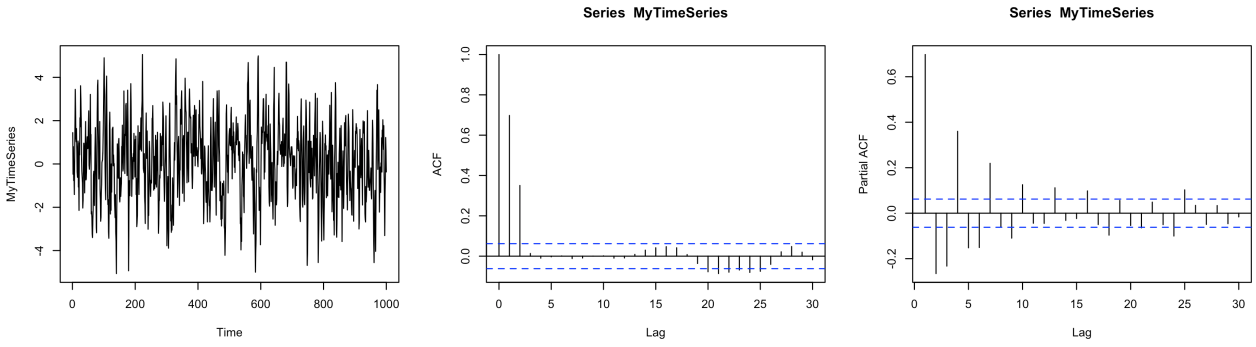
**MA**(1)

```
set.seed(15)
MyTimeSeries = arima.sim(model = list(ma = c(1)),
    n = 1000, rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```




**MA**(2)

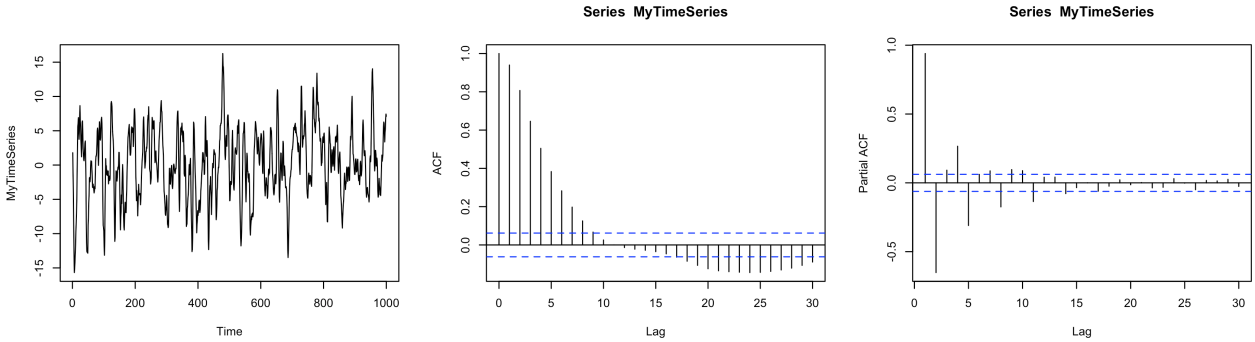
```
set.seed(16)
MyTimeSeries = arima.sim(model = list(ma = c(1,1)),
                         n = 1000,
                         rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
acf(MyTimeSeries)
pacf(MyTimeSeries)
```




**ARMA**$(1, 2)$

```
set.seed(17)
MyTimeSeries = arima.sim(model = list(ar = c(0.8),
                                      ma = c(1,1)),
                         n = 1000,
                         rand.gen = rnorm)
par(mfrow=c(1,3))
plot.ts(MyTimeSeries)
```

```
acf(MyTimeSeries)
pacf(MyTimeSeries)
```




**Summary:**

- for AR($p$) models $\gamma_X(h) \neq 0$ for any $h$.
- for MA($q$) models $\gamma_X(h) = 0$ for any $|h| > q$;
- for AR($p$) models $\alpha(h) = 0$ for any $|h| > p$;
- for MA($q$) models $\alpha_X(h) \neq 0$ for any $h$.

## 9.4 Forecasting with Stationary Time Series

In practice, once of the main objectives of time series analysis is to **predict** (or **forecast**) $X_{n+k}$ for some $k \geq 1$, having observed $\{X_1, \ldots, X_n\}$ from a time series with **known** mean $\mu$ and ACVF $\gamma_X(k)$, $k \geq 0$.

Consider a stationary sequence with mean $\mu = \mathrm{E}[X_t]$ and covariance $\gamma_X(h)$. Denote by $P_n X_{n+k}$ a prediction for $X_{n+k}$, given the $n$ observations $X_1, \ldots, X_n$.

We will restrict ourselves to **linear predictors**, that is to say, predictors of the form:

$$P_n X_{n+k} = a_0 + a_1 X_n + \cdots + a_n X_1 = a_0 + \sum_{i=1}^{n} a_i X_{n+1-i},$$

where $a_0, a_1, \ldots, a_n \in \mathbb{R}$.

As is usually the case in statistical applications, this can be recast as an optimization problem. We seek values $\mathbf{a} = (a_0, \ldots, a_n)$ which minimize the expected **mean squared error** (MSE):

$$\mathrm{E}\left[(X_{n+k} - P_n X_{n+k})^2\right],$$

One challenge is that we cannot minimize $(X_{n+k} - P_n X_{n+k})^2$ directly since, there would be no reason to predict $X_{n+k}$ if we already knew it.[37]

37: While the whole entreprise is reminiscent of OLS regression, there are some important differences, chief among them being that the predictors $X_{n+1-i}$ are typically correlated with one another.

### 9.4.1 Yule-Walker Procedure

Let

$$S(\mathbf{a}) = \mathrm{E}\left[(X_{n+k} - P_n X_{n+k})^2\right] = \mathrm{E}\left[(X_{n+k} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i})^2\right].$$

We minimize $S$ by finding its critical points, i.e. by solving $\nabla S(\mathbf{a}) = \mathbf{0}$.

The partial derivative of $S$ with respect to $a_0$ is

$$\mathrm{E}\Big[2(X_{n+k} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}) \cdot 1\Big] = 2\Big(\mu - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\Big);$$

setting it equal to 0 yields

$$a_0 = \mu\Big(1 - \sum_{i=1}^{n} a_i\Big).$$

If $\{X_t\}$ is assumed to be stationary, then $\mu = 0$, and so $a_0 = 0$.

The partial derivatives with respect to $a_1, \ldots, a_n$ are thus:

$$\mathrm{E}\Big[ -2\Big(X_{n+k} - \sum_{i=1}^{n} a_i X_{n+1-i}\Big) X_{n+1-j}\Big], \qquad j = 1, \ldots, n.$$

Setting each of these to 0 yields:

$$\mathrm{E}[X_{n+k} X_{n+1-j}] - \sum_{i=1}^{n} a_i \mathrm{E}[X_{n+1-i} X_{n+1-j}] = 0, \qquad j = 1, \ldots, n.$$

Since $E[X_t] = \mu = 0$, the above expectations are the covariances of $\{X_t\}$ at lags $n + k - (n + 1 - j) = k - 1 + j$ and $n + 1 - i - (n + 1 - j) = i - j$, and we can thus write the system of equations as:

$$\gamma_X(k - 1 + j) = \sum_{i=1}^{n} a_i \gamma_X(i - j), \qquad j = 1, \ldots, n. \tag{9.1}$$

Define the matrix

$$\Gamma_n = [\gamma_X(|i - j|)]_{i,j=1}^{n}$$

and the column vectors

$$\boldsymbol{\gamma}(n; k) = (\gamma_X(k), \ldots, \gamma_X(k + n - 1))^\top, \qquad \underline{\mathbf{a}}_n = (a_1, \ldots, a_n)^\top.$$

We recognize $\Gamma_n$ as the **variance-covariance matrix** of $(X_1, \ldots, X_n)$, whose diagonal entries are $\gamma_X(0) = \mathrm{Var}(X_t) = \sigma_X^2$.

If $n = 1$, for instance, then $\Gamma_1 = \gamma_X(0)$; if $n = 2$, then

$$\Gamma_2 = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(|-1|) & \gamma_X(0) \end{bmatrix} = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(1) & \gamma_X(0) \end{bmatrix}.$$

We can write the system of $n$ equations in $n$ unknowns from (9.1) in a matrix-vector notation:

$$\Gamma_n \underline{\mathbf{a}}_n = \boldsymbol{\gamma}(n; k),$$

whose solution, assuming that $\Gamma_n$ is invertible, is the **Yule-Walker forecasting formula**:

$$\underline{\mathbf{a}}_n = \Gamma_n^{-1} \boldsymbol{\gamma}(n; k).$$

Note that it is **model-independent**.[38]

38: Well, the formula for $\underline{\mathbf{a}}_n$ is, at any rate. It only really assumes that the time series is stationary. But it does depend on the autocovariances of the time series; with a model, it is usually rather straightforward to compute these. Without a model, we have to use the sample autocovariances.

**MSPE** The above procedure guarantees that the **mean squared prediction error**

$$\mathrm{MSPE}_n(k) = \mathrm{E}\left[(X_{n+k} - \sum_{i=1}^{n} a_i X_{n+1-i})^2\right]$$

is minimized when $\underline{\mathbf{a}}$ is chosen according to the Yule-Walker procedure. Can we calculate the MSPE value?

Recall that the $\mathrm{E}[X_t] \equiv 0$ by stationarity. Thus,

$$\begin{aligned}
&\mathrm{E}\left[(X_{n+k} - \sum_{i=1}^{n} a_i X_{n+1-i})^2\right] \\
&= \mathrm{E}\left[X_{n+k}^2\right] - 2\sum_{i=1}^{n} a_i \mathrm{E}[X_{n+k}X_{n+1-i}] + \mathrm{E}\left[\left(\sum_{i=1}^{n} a_i X_{n+1-i}\right)^2\right] \\
&= \gamma_X(0) - 2\sum_{i=1}^{n} a_i \gamma_X(k+i-1) + \mathrm{E}\left[\sum_{i,j=1}^{n} a_i X_{n+1-i} X_{n+1-j} a_j\right] \\
&= \gamma_X(0) - 2\sum_{i=1}^{n} a_i \gamma_X(k+i-1) + \sum_{i,j=1}^{n} a_i \gamma_X(i-j) a_j \\
&= \gamma_X(0) - 2\underline{\mathbf{a}}_n^\top \boldsymbol{\gamma}(n;k) + \underline{\mathbf{a}}_n^\top \Gamma_n \underline{\mathbf{a}}_n = \gamma_X(0) - \underline{\mathbf{a}}_n^\top \boldsymbol{\gamma}(n;k).
\end{aligned}$$

An important remark is that the MSPE formula depends on $k$; in particular, it is possible that, given a set of observations $X_1, \ldots, X_n$, predictions further in the future (i.e., having a larger $k$) may have a larger prediction error than those nearer $t = n$.[39]

39: Of course, it could also be the other way around – but the point is that we should not expect $\mathrm{MSPE}_n(k)$ to be constant with $k$.

**Example: AR(1)** Consider the auto-regressive model $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and $Z_t$ are i.i.d. with mean 0 and variance $\sigma_Z^2$. We have already seen that $\{X_t\}$ is stationary, and so that $\mu = \mathrm{E}[X_t] \equiv 0$.

Recall that the autocovariances for this model are:

$$\gamma_X(h) = \phi^h \frac{\sigma_Z^2}{1-\phi^2}, \qquad h \geq 0.$$

If we are interested in predicting $X_{n+1}$, then we need:

$$\boldsymbol{\gamma}(n;k) = \boldsymbol{\gamma}(n;1) = (\gamma_X(1), \ldots, \gamma_X(n))^\top = \frac{\sigma_Z^2}{1-\phi^2}(\phi, \ldots, \phi^n)^\top.$$

The Yule-Walker forecasting equation in this case becomes

$$\frac{\sigma_Z^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \cdots & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \frac{\sigma_Z^2}{1-\phi^2}\begin{pmatrix} \phi \\ \vdots \\ \phi^n \end{pmatrix}.$$

We can show that the determinant of of $\Gamma_n$ is

$$\det(\Gamma_n) = (-1)^{n-1}(\phi-1)^{n-1}(\phi+1)^{n-1}\left(\frac{\sigma_Z^2}{1-\phi^2}\right)^n \neq 0$$

since $|\phi| < 1$. There is thus a unique forecasting solution $\underline{\mathbf{a}}_n$.

But

$$\Gamma_n \begin{pmatrix} \phi \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \frac{\sigma_Z^2}{1-\phi^2} \begin{pmatrix} 1 \cdot \phi + 0 \cdot (\ldots) \\ \phi \cdot \phi + 0 \cdot (\ldots) \\ \vdots \\ \phi^{n-1} \cdot \phi + 0 \cdot (\ldots) \end{pmatrix} = \frac{\sigma_Z^2}{1-\phi^2} \begin{pmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{pmatrix},$$

and so $\underline{\mathbf{a}}_n = (\phi, 0, \ldots, 0)^\top$ is the unique Yule-Walker forecast vector for the AR(1) model.

The Yule-Walker prediction for $X_{n+1}$ is thus

$$P_n X_{n+1} = a_1 X_n + a_2 X_{n-1} + \cdots + a_n X_1 = \phi X_n,$$

while the MSPE is

$$\begin{aligned} \text{MSPE}_n(1) &= \gamma_X(0) - \underline{\mathbf{a}}_n^\top \boldsymbol{\gamma}(n;1) = \gamma_X(0) - \phi\gamma_X(1) - 0 \cdot \gamma_X(2) - \cdots - 0 \cdot \gamma_X(n) \\ &= \frac{\sigma_Z^2}{1-\phi^2} - \phi^2 \frac{\sigma_Z^2}{1-\phi^2} = \sigma_Z^2. \end{aligned}$$

Note, however, that these formulas cannot yet be used in a practical setting since they involve the unknown parameters $\phi$ and $\sigma_Z^2$.

### 9.4.2 Durbin-Levinson Algorithm

In the AR(1) prediction example, we were lucky that the solution $\underline{\mathbf{a}}_n$ was provided *in extremis*; there is a way to find the best linear predictor without having to compute the inverse of $\Gamma_n$. But it comes at a price: the approach only allows **one-step** prediction to $P_n X_{n+1}$.

We assume that $\mu = \mathrm{E}[X_t] \equiv 0$ and $a_0 = 0$, as in the Yule-Walker procedure.

We re-write the linear predictor as

$$P_n X_{n+1} = \phi_{n,1} X_n + \cdots + \phi_{n,n} X_1.$$

That is, $a_1 = \phi_{n,1}, \ldots, a_n = \phi_{n,n}$.

- If $n = 1$, we seek to find $P_1 X_2 = \phi_{1,1} X_1$ which minimizes

  $$\mathrm{E}\left[(X_2 - P_1 X_2)^2\right] = \mathrm{E}\left[(X_2 - \phi_{1,1} X_1)^2\right].$$

  We differentiate with respect to $\phi_{1,1}$ and set equal to 0 to find the critical point:

  $$\mathrm{E}\left[2(X_2 - \phi_{1,1} X_1)(-X_1)\right] = 0 \implies \mathrm{E}[X_1 X_2] = \phi_{1,1} \mathrm{E}[X_1^2],$$

  which is to say that

  $$\phi_{1,1} = \frac{\gamma_X(1)}{\gamma_X(0)} = \rho_X(1).$$

- If $n = 2$, we seek to find $\phi_{2,1}$ and $\phi_{2,2}$ in

  $$P_2 X_3 = \phi_{2,1} X_2 + \phi_{2,2} X_2.$$

As in the Yule-Walker procedure we minimize

$$\mathrm{E}[(X_3 - \phi_{2,1}X_2 - \phi_{2,2}X_1)^2].$$

Taking derivatives with respect to $\phi_{2,1}$ and $\phi_{2,2}$ leads to:

$$\begin{aligned}\mathrm{E}[-2X_2(X_3 - \phi_{2,1}X_2 - \phi_{2,2}X_1)] = 0\\ \mathrm{E}[-2X_1(X_3 - \phi_{2,1}X_2 - \phi_{2,2}X_1)] = 0;\end{aligned}$$

equivalently, since the mixed expectations are covariances and the squared ones are variances, this can be written as:

$$\begin{aligned}\gamma_X(1) - \phi_{2,1}\gamma_X(0) - \phi_{2,2}\gamma_X(1) = 0\\ \gamma_X(2) - \phi_{2,1}\gamma_X(1) - \phi_{2,2}\gamma_X(0) = 0.\end{aligned}$$

We divide both equations by $\gamma_X(0)$ and re-organize the terms to obtain:

$$\begin{aligned}\phi_{2,1} = \rho_X(1) - \phi_{2,2}\rho_X(1) = \rho_X(1) - \phi_{2,2}\phi_{1,1}, \quad \text{by step } n = 1;\\ 0 = \rho_X(2) - \phi_{2,1}\rho_X(1) - \phi_{2,2}\end{aligned}$$

Solving for $\phi_{2,1}$ and $\phi_{2,2}$, we arrive at

$$\begin{aligned}\phi_{2,2} &= \frac{\rho_X(2) - \phi_{1,1}\rho_X(1)}{1 - \phi_{1,1}\rho_X(1)},\\ \phi_{2,1} &= \rho_X(1) - \phi_{2,2}\phi_{1,1}.\end{aligned}$$

We use either $\phi_{1,1}$ or $\rho_X(1)$, solely based on convenience (since they are equal). In the last system of equations, the coefficients $\phi_{2,2}$ and $\phi_{2,1}$ are computed using sample autocorrelations, as well as $\phi_{1,1}$ (from the step $n = 1$).

This recursive procedure can be extended for a general $n$.

**Durbin-Levinson Algorithm** The coefficients $\phi_{n,1}, \ldots, \phi_{n,n}$ in the best linear prediction $P_n X_{n+1}$ can be computed recursively as:

$$\phi_{n,n} = \left[\gamma_X(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma_X(n-j)\right] v_{n-1}^{-1};$$

$$\begin{pmatrix}\phi_{n,1}\\ \vdots\\ \phi_{n,n-1}\end{pmatrix} = \begin{pmatrix}\phi_{n-1,1}\\ \vdots\\ \phi_{n-1,n-1}\end{pmatrix} - \phi_{n,n}\begin{pmatrix}\phi_{n-1,n-1}\\ \vdots\\ \phi_{n-1,1}\end{pmatrix},$$

and

$$v_n = v_{n-1}[1 - \phi_{n,n}^2], \qquad v_0 = \gamma_X(0), \qquad \phi_{1,1} = \rho_X(1).$$

Note that the Durbin-Levinson algorithm and the Yule-Walker procedure lead to the same results for $P_n X_{n+1}$; indeed, in both cases we compute the coefficents of the linear prediction $P_n X_{n+1}$ using the mean squared error criterion, the difference being that we approach the problem from two different angles.

**AR(1)** Consider the auto-regressive model $X_t = \phi X_{t-1} + Z_t$, where $Z_t$ are i.i.d. with mean 0 and variance $\sigma_Z^2$.

We know the ACVF and ACF of $\{X_t\}$ are

$$\gamma_X(h) = \phi^h \frac{\sigma_Z^2}{1-\phi^2}, \quad \text{and} \quad \rho_X(h) = \gamma_X(h)/\gamma_X(0) = \phi^h.$$

Using the Durbin-Levinson algorithm, we find the linear coefficients and predictors as follows:

$$\begin{aligned} \phi_{1,1} = \phi, \qquad & P_1 X_2 = \phi X_1; \\ \phi_{2,1} = \phi, \phi_{2,1} = 0, \qquad & P_2 X_3 = \phi X_2; \\ & \vdots \quad \vdots \\ \phi_{n,1} = \phi, \phi_{n,2} = \cdots = \phi_{n,n} = 0, \qquad & P_n X_{n+1} = \phi X_n. \end{aligned}$$

**Partial Autocovariance function (PACF)** As a by-product of the Durbin-Levinson algorithm, we obtain the PACF *via*:

$$\alpha(0) = 1; \qquad \alpha(h) = \phi_{h,h}, \quad h \geq 1.$$

### 9.4.3 Forecast Limits and Prediction Intervals

We obtained **model-independent** formulas for (linearly) predicted time series values in the preceding sections, depending solely on the sample autocovariances.[40] Discussions of **accuracy**, however, require model assumptions.

40: Although we can use a model if one is available.

Let $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ be a causal linear process with $\mathrm{E}[Z_t] = 0$ and $\mathrm{Var}(Z_t) = \sigma_Z^2$, and $k \geq 1$ an integer.

It can be shown that the mean squared prediction error at $P_n X_{n+k}$ is:

$$\mathrm{MSPE}_n(k) = \mathrm{E}[(X_{n+k} - P_n X_{n+k})^2] = \sigma_Z^2 \sum_{j=0}^{k-1} \psi_j^2.$$

The theoretical forecast limits of the $100(1-\alpha)\%$ **prediction interval** are thus:

$$P_n X_{n+k} \pm z_{\alpha/2} \sqrt{\mathrm{MSPE}_n(k)} = P_n X_{n+k} \pm z_{\alpha/2} \sigma_Z \sqrt{\sum_{j=0}^{k-1} \psi_j^2},$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution.[41] Note that MSPE (and so the coefficients $\psi_j$) are **model-dependent**: no model, no prediction interval!

41: You know the one: if $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$.

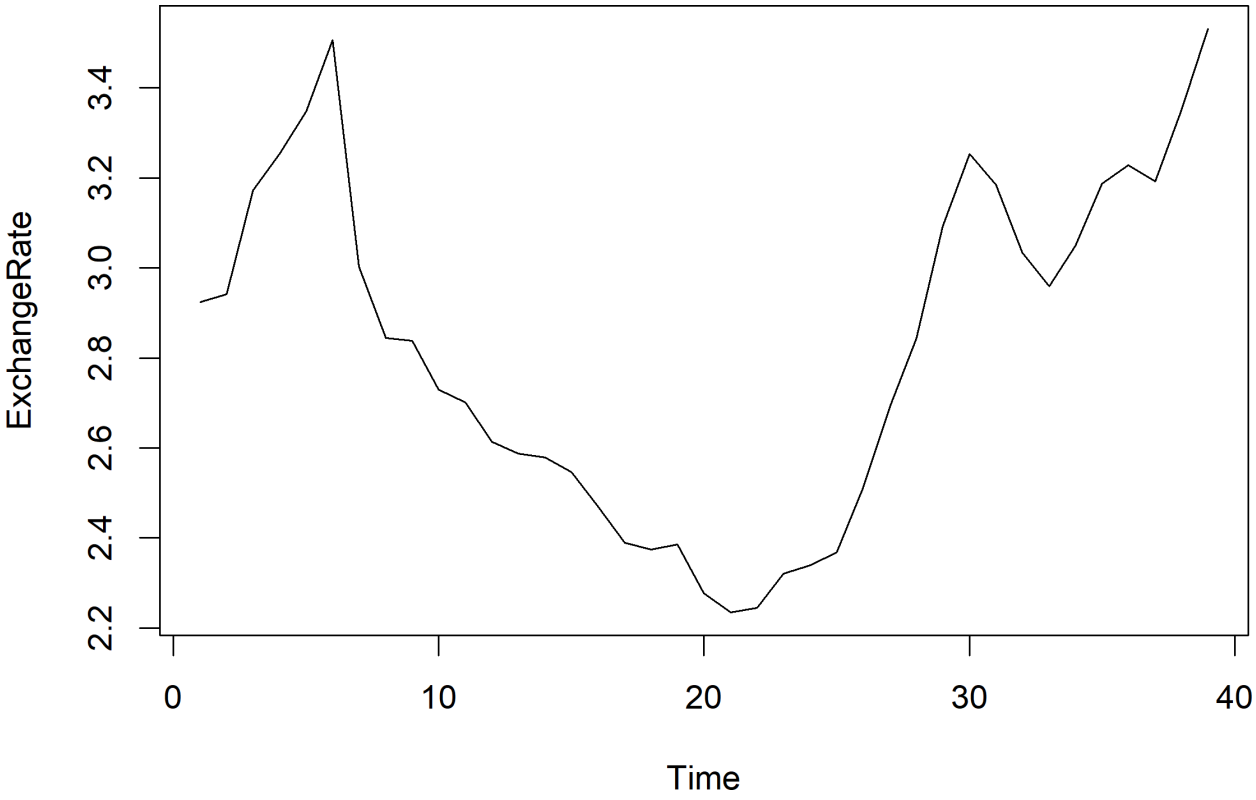
### 9.4.4 Example: Currency Conversion Data

We illustrate the notions presented in this section with an example, using the quarterly mean exchange rate between British pounds (UK) and New Zealand dollar (NZD), from Jan 1991 to Mar 2000 (prepared by Darrin Speegler).

```
ExchangeRate = c(2.9243,2.9422,3.1719,3.2542,3.3479,
                 3.5066,3.0027,2.8440,2.8378,2.7301,
                 2.7008,2.6138,2.5874,2.5787,2.5470,
                 2.4701,2.3895,2.3705,2.3859,2.2766,
                 2.2351,2.2450,2.3208,2.3390,2.3687,
                 2.5120,2.6917,2.8435,3.0922,3.2528,
                 3.1852,3.0340,2.9593,3.0498,3.1869,
                 3.2286,3.1925,3.3522,3.5310)
```

The time series plot tells a better story.

```
plot.ts(ExchangeRate)
```



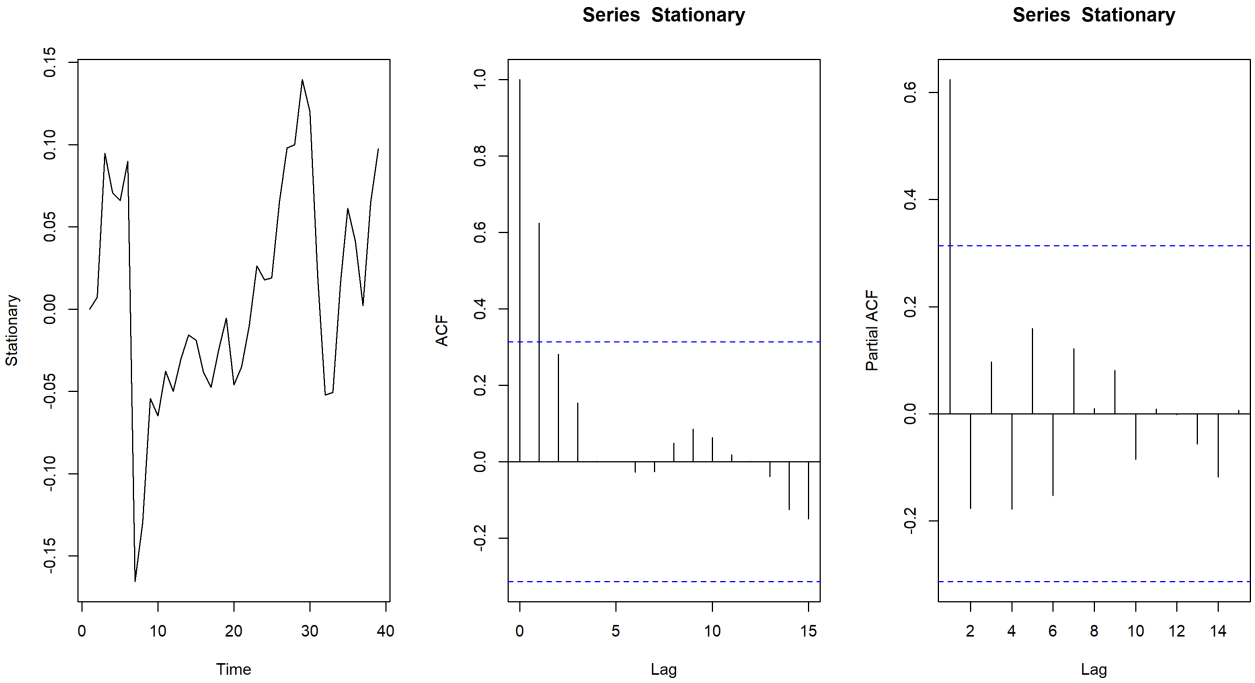

The model is clearly not stationary.

We detrend the data *via* the exponential smoother `ExpSmooth` of Section 9.1.2, with $\alpha = 0.6$.

```
alpha = 0.6
ExchangeRate.smoothed <- ExpSmooth(ExchangeRate,alpha)
Stationary = ExchangeRate - ExchangeRate.smoothed
```

The ACF and PACF of the stationary components are found below.

```
par(mfrow=c(1,3))
plot.ts(Stationary)
acf(Stationary)
pacf(Stationary)
```




The detrended time series looks like AR(1).[42] We centre the time series and use the Yule-Walker method to verify that this is indeed an appropriate model – we will be discussing this further in Section 9.5.3.

42: Does it? How could you tell?

```
MyTimeSeries = Stationary
n = length(MyTimeSeries)
mean = mean(MyTimeSeries)
MyTimeSeries.centered = MyTimeSeries-mean(MyTimeSeries)
(fit.ar <- ar(MyTimeSeries.centered,method="yule-walker"))
```

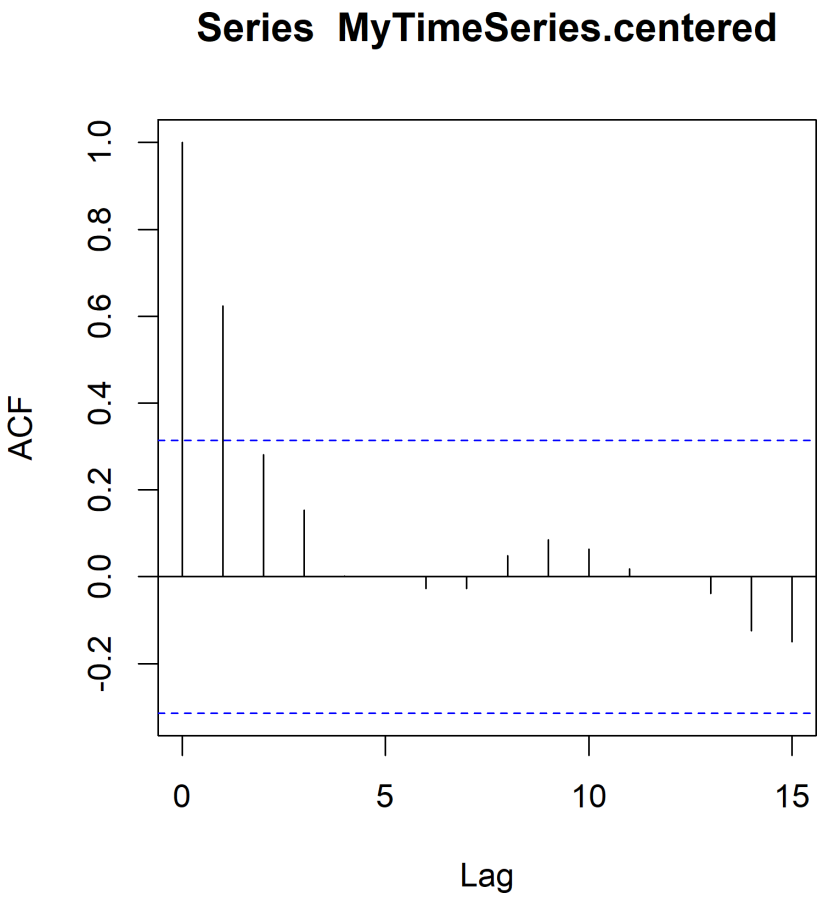
```
Coefficients:
     1
0.6241

Order selected 1  sigma^2 estimated as  0.002842
```

The Yule-Walker estimates of the selected AR(1) model are $\hat{\phi} = 0.6241$, $\sigma_X^2 = 0.002842$, respectively.

We can verify the Yule-Walker output by comparing with the ACF.

```
par(mfrow=c(1,1))
(ACF <- acf(MyTimeSeries.centered))
```

```
Autocorrelations of series 'MyTimeSeries.centered', by lag

     0      1      2      3      4      5      6      7      8
 1.000  0.624  0.281  0.154  0.001  0.000 -0.027 -0.027  0.048
     9     10     11     12     13     14     15
 0.085  0.063  0.018  0.001 -0.039 -0.125 -0.149
```




The second entry is indeed 0.624, the estimator of $\phi$, which can also be accessed as follows.

```
phi = acf(MyTimeSeries.centered)$acf[2]
```

The sample variance of the centered data is:

```
(v = var(MyTimeSeries.centered))
```

```
[1] 0.004532399
```

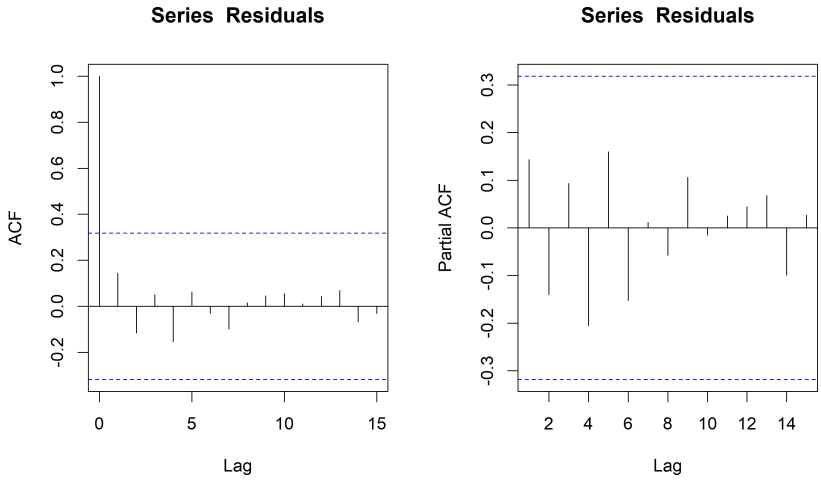
The estimator of $\sigma_Z^2$ is:

```
v-phi^2*v
```

```
[1] 0.002767046
```

How can we tell if the AR(1) fit is appropriate? We can compute the "residuals" of the $X_t - \hat{\phi} X_{t-1}$ and compare it to $Z_t$, which is to say an i.i.d. random variable with mean 0 and variance $\sigma_Z^2$. What do the residual time series ACF and PACF look like?

```
Residuals <- MyTimeSeries.centered[2:n] -
            phi*MyTimeSeries.centered[1:(n-1)]
par(mfrow=c(1,2))
acf(Residuals)
pacf(Residuals)
```




It certainly seems as though there is little dependence left in the residuals time series. We can apply the Ljung-Box test (which we will discuss in Section 9.6).

```
Box.test(Residuals,type="Ljung",lag=1,fitdf=1)
```

```
Box-Ljung test

data:  Residuals
X-squared = 30.799, df = 1, p-value = 2.862e-08
```

The outcome is compatible with the notion that the residuals are i.i.d. random variables.

We can also extract the residuals directly.

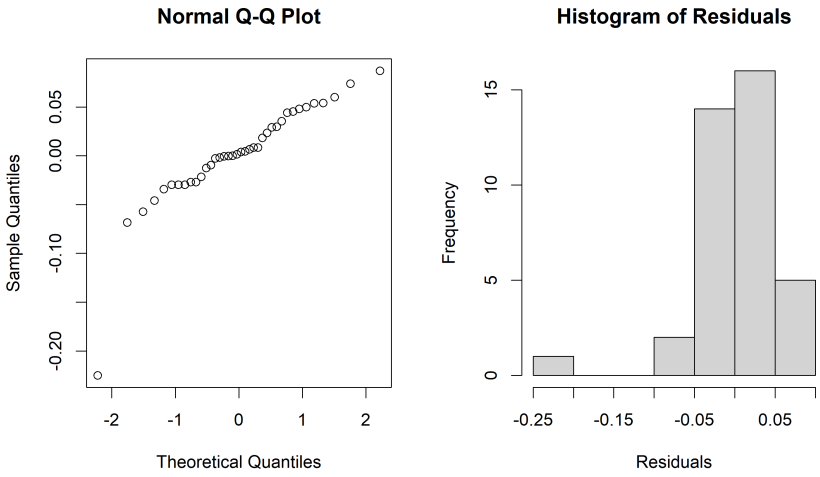
```
fit.ar$resid;
```

```
 [1]            NA  3.887364e-03  8.700283e-02  8.415571e-03
 [5]  1.833739e-02  4.546024e-02 -2.249571e-01 -2.963355e-02
 [9]  2.332064e-02 -3.416757e-02 -4.645337e-04 -2.963498e-02
[13] -2.659019e-03  8.326911e-05 -1.243842e-02 -2.978541e-02
[17] -2.692054e-02  1.589073e-03  6.651807e-03 -4.574392e-02
[21] -9.575653e-03  8.526158e-03  2.929546e-02 -1.888002e-03
[25]  4.617796e-03  4.978927e-02  5.405892e-02  3.551993e-02
[29]  7.382926e-02  2.972301e-02 -5.720633e-02 -6.845055e-02
[33] -2.147845e-02  4.429304e-02  4.800134e-02 -3.084927e-04
[37] -2.693691e-02  6.015361e-02  5.375058e-02
```

Note that this produces one "NA", as the first residual corresponds to $X_1 - \hat{\phi} X_0$, but $X_0$ does not exist in the original stationary time series.

The normality of the residuals (as well as their mean) can be visually assessed as follows.

```
par(mfrow=c(1,2))
qqnorm(Residuals);
hist(Residuals)
```




There are some off-the-beaten-track values, but for the most part, the data is compatible with the idea of the residuals being normally distributed, with mean 0 and variance $\widehat{\sigma}_Z^2$.

We can predict the next value of `MyTimeSeries`, and get the MSPE and its prediction interval as follows.

```
(prediction.next <- mean*(1-phi) + phi*MyTimeSeries[n])
(MSPE = (v-phi^2*v))
```

```
[1] 0.06405712
[1] 0.002767046
```

MSPE can also be obtained by typing `fit.ar$var.pred` at the prompt.

```
alpha=0.05
quantile = qnorm(1-alpha/2)
c(prediction.next - quantile*sqrt(MSPE),
  prediction.next + quantile*sqrt(MSPE))
```

```
[1] -0.03904232  0.16715655
```

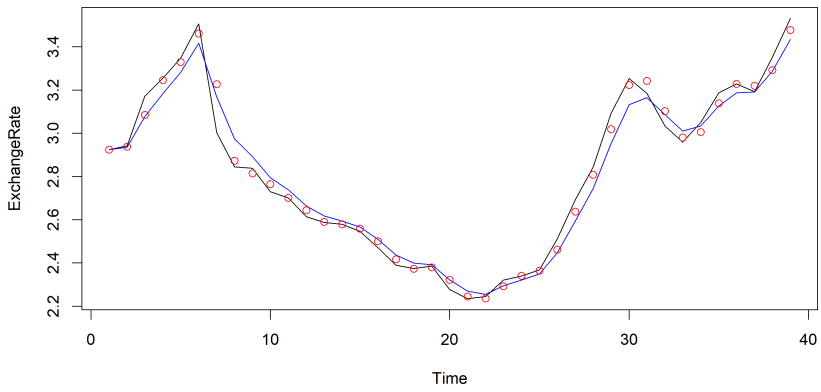
But to make a prediction in the original data, we need to take the last value in the smoothed time series and add the prediction for the stationary component; this serves as the prediction of the next observation for the original time series.

```
(Prediction.Exchange.Rate.next <-
    ExchangeRate.smoothed[n] + prediction.next)
```

```
[1] 3.497661
```

We can also determine the quality of the model fit by "predicting" past values of the original time series using the same process as above (black: original; blue: smoothed model: red: predictions).

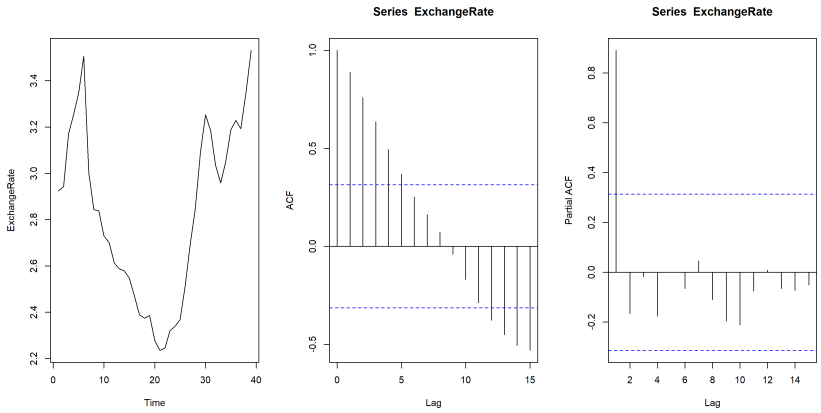
```
prediction <- mean*(1-phi) + phi*(MyTimeSeries)
prediction <- c(MyTimeSeries[1],prediction[1:n-1])
Prediction.Exchange.Rate <- ExchangeRate.smoothed +
        prediction[1:n]

par(mfrow=c(1,1))
plot.ts(ExchangeRate)
points(ExchangeRate.smoothed,type="l",col="blue")
points(Prediction.Exchange.Rate,type="p",col="red")
(Squared.Error =
    sum((Prediction.Exchange.Rate - ExchangeRate)^2))
```




```
[1] 0.1020082
```

What happens if we ignore the non-stationary behaviour and work on the original data itself instead of the stationary component? The Yule-Walker method says the data follows an AR(1) model, but with different $\widehat{\phi}$ and $\widetilde{\sigma}^2_X$ values.

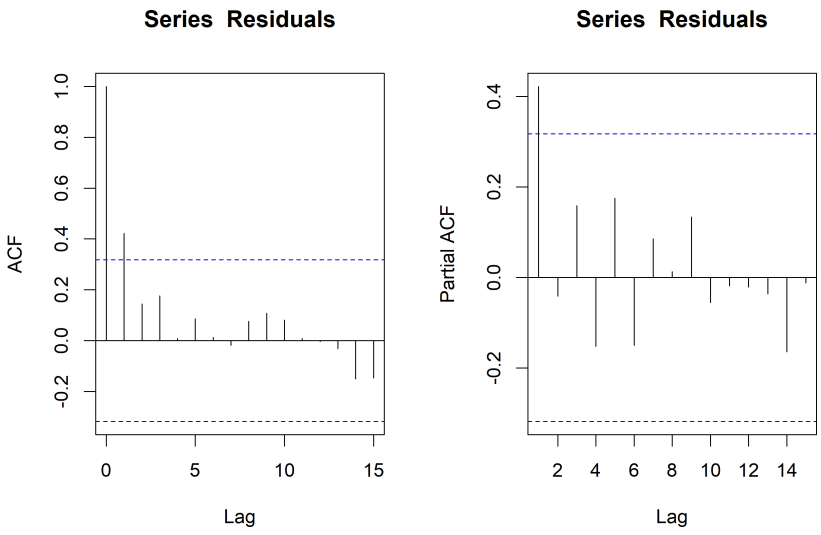
```
par(mfrow=c(1,3))
plot.ts(ExchangeRate)
acf(ExchangeRate)
pacf(ExchangeRate)

mean = mean(ExchangeRate)
ExchangeRate.centered = ExchangeRate - mean(ExchangeRate);
(fit.ar <- ar(ExchangeRate.centered,method="yule-walker"))
```




```
Coefficients:
     1
0.8903

Order selected 1  sigma^2 estimated as  0.03125
```

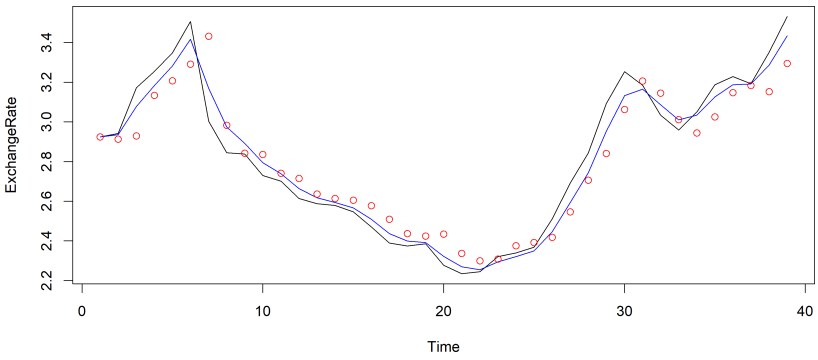
This fit's residuals do not appear to form an i.i.d. sequence.

```
phi = fit.ar$ar
Residuals <- ExchangeRate.centered[2:n] -
    phi*ExchangeRate.centered[1:(n-1)]

par(mfrow=c(1,2))
acf(Residuals)
pacf(Residuals)
```




Note, in particular, the large value of $\widehat{\rho}_X(1) \approx 0.5$. The fitted AR(1) model is the best of the AR models for the data, but it is unlikely to be correct. Nothing is stopping us from predicting new values on the (false) assumption that it was correct, unfortunately.

```
prediction <- mean*(1-phi) + phi*ExchangeRate
prediction <- c(ExchangeRate[1],prediction[1:n-1])
Prediction.Exchange.Rate.Wrong <- prediction[1:n]

par(mfrow=c(1,1))
plot.ts(ExchangeRate)
points(ExchangeRate.smoothed,type="l",col="blue")
points(Prediction.Exchange.Rate.Wrong,type="p",col="red")
(Squared.Error.Wrong =
    sum((Prediction.Exchange.Rate.Wrong-ExchangeRate)^2))
```




```
[1] 0.7490375
```

The predictions are clearly not as accurate as they were in our first attempt at analyzing the data – the squared error is seven times larger now than it was then.[43]

43: This example highlights the importance of **understanding** the process; it is not sufficient to know how to produce new predictions from a time series data – we also need to know not to apply the procedure when the time series is not stationary, or when the model is a poor fit to the data.

## 9.5 Estimation of ARMA Models

Let's assume that we have observations $\{X_1, \ldots, X_n\}$ from a time series and that we have also identified that a model ARMA$(p, q)$ from which they could conceivably arise. How can we best estimate the parameters $\phi_1, \ldots, \phi_p$ and/or $\theta_1, \ldots, \theta_q$?

### 9.5.1 Mean: I.I.D. Case

Assume first that $X_1, \ldots, X_n$ are i.i.d. In practice, the mean of such a sequence is not typically 0. We estimate $\mu \equiv \mathrm{E}[X_t]$ by the **method of moments**, using the sample mean $\overline{X}$:

$$\mathrm{E}[\overline{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} n\mu = \mu.$$

Using the **independence** of the $X_t$, we have:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{\gamma_X(0)}{n}.$$

This computation leads to the **Central Limit Theorem**.

**Lemma:** assume that $X_1, \ldots, X_n$ are i.i.d. with mean $\mu$ and variance $\gamma_X(0)$. Then

$$\sqrt{n}\left\{\frac{\overline{X}-\mu}{\sqrt{\gamma_X(0)}}\right\} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

that is

$$\lim_{n\to\infty} P\left(\sqrt{n}\left\{\frac{\overline{X}-\mu}{\sqrt{\gamma_X(0)}}\right\} \le x\right) = \Phi(x),$$

where $\Phi$ is the standard normal **cumulative distribution function**.

This allows us to construct a **95% confidence interval** for the mean $\mu$:

$$\text{C.I.}(\mu; 0.95) \equiv \left(\overline{X} - 1.96\frac{\sqrt{\gamma_X(0)}}{\sqrt{n}}, \overline{X} + 1.96\frac{\sqrt{\gamma_X(0)}}{\sqrt{n}}\right).$$

This confidence interval involves the unknown $\gamma_X(0)$, which can be estimated with the sample variance.

### 9.5.2 Mean: Time Series

When the time series $\{X_1, \ldots, X_n\}$ does not consist of i.i.d. random variables but arises from a stationary time series, the estimate for $\mu$ remains valid, but the variance computation has to be modified.

Instead, we have

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Cov}\left(\overline{X}, \overline{X}\right) = \mathrm{Cov}\left(\frac{X_1+\cdots+X_n}{n}, \frac{X_1+\cdots+X_n}{n}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j) \\
&= \frac{1}{n^2}\sum_{h=-(n-1)}^{n-1}(n-|h|)\gamma_X(h) = \frac{1}{n^2}\sum_{h=-n}^{n}(n-|h|)\gamma_X(h) \\
&= \frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|).
\end{aligned}$$

As an illustration, assume that $n = 3$. Then

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j) &= 3\gamma_X(0) + 2\gamma_X(1) + 2\gamma_X(-1) + \gamma_X(2) + \gamma_X(-2) \\
&= \sum_{h=-2}^{2}(3-|h|)\gamma_X(h).
\end{aligned}$$

Assume now that $\gamma_X(|h|) \to 0$ as $|h| \to \infty$. Then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|) = \lim_{n\to\infty}\frac{1}{n}\sum_{h=-n}^{n}\gamma_X(|h|) = 0,$$

and

$$\lim_{n\to\infty} n\mathrm{Var}(\overline{X}) = \lim_{n\to\infty} n\frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|)$$
$$= \lim_{n\to\infty}\sum_{h=-n}^{n}\gamma_X(|h|) = \sum_{h=-\infty}^{\infty}\gamma_X(|h|) = \gamma_X(0) + 2\sum_{h=1}^{\infty}\gamma_X(h)$$

as long as $\{X_t\}$ is **short-range dependent** ($\sum_{-\infty}^{\infty}|\gamma_X(|h|)| < \infty$).

This computation is one of the main steps to establish the Central Limit Theorem in the general case.

**Lemma:** assume that $X_1, \ldots, X_n$ is a stationary short-range dependent time series with mean $\mu$, variance $\gamma_X(0)$, and covariance function $\gamma_X(h)$. Then

$$\sqrt{n}\left\{\frac{\overline{X}-\mu}{v}\right\} \xrightarrow{\mathrm{d}} N(0,1),$$

that is

$$\lim_{n\to\infty} P\left(\sqrt{n}\left\{\frac{\overline{X}-\mu}{v}\right\} \le x\right) = \Phi(x),$$

where $\Phi$ is as above, and

$$v^2 = \gamma_X(0) + 2\sum_{h=1}^{\infty}\gamma_X(h).$$

This allows us to construct a **95% confidence interval for the mean** $\mu$:

$$\text{C.I.}(\mu; 0.95) \equiv \left(\overline{X} - 1.96\frac{v}{\sqrt{n}}, \overline{X} + 1.96\frac{v}{\sqrt{n}}\right).$$

This confidence interval involves the unknown $v$.

**Example** Recall that the AR(1) model is $X_t = \phi X_{t-1} + Z_t$, with the usual assumptions on $Z_t$.[44] Then $\gamma_X(h) = \sigma_Z^2\frac{\phi^h}{1-\phi^2}$, and so

$$v^2 = \gamma_X(0) + 2\sum_{h=1}^{\infty}\gamma_X(h) = \sigma_Z^2\frac{1}{1-\phi^2} + 2\sigma_Z^2\frac{1}{1-\phi^2}\frac{\phi}{1-\phi} = \sigma_Z^2\frac{1}{(1-\phi)^2}.$$

44: In order to obtain the linear representation of the model, we need to have $\mu = 0$. If the data is not centered ($\mu \neq 0$), consider instead the shifted model

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t.$$

The stationary solution will then be

$$X_t = \mu + \sum_{j=0}^{\infty}\phi^j Z_{t-j}.$$

### 9.5.3 Yule-Walker Estimators

The method we present now has similarities with Yule-Walker forecasting; it works quite well for AR($p$) models.

Assume a stationary and causal AR(1) model: $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$, $\mathrm{E}[Z_t] \equiv 0$, and $\mathrm{Var}(Z_t) \equiv \sigma_Z^2$. Multiply both sides of the equation, once by $X_{t-1}$ and another time by $X_t$, to get

$$X_t X_{t-1} = \phi X_{t-1}X_{t-1} + Z_t X_{t-1},$$
$$X_t^2 = \phi X_t X_{t-1} + X_t Z_t.$$

We apply the expectation operator on both of these new equations (recall that $\mathrm{E}[X_t] = 0$ and that $X_{t-1}$ is independent of $Z_t$ because the time series is causal) to obtain:

$$\begin{aligned}\gamma_X(1) &= \phi\gamma_X(0) + 0,\\ \gamma_X(0) &= \phi\gamma_X(1) + \mathrm{E}[X_t Z_t].\end{aligned}$$

That last term evaluates to

$$\mathrm{E}[X_t Z_t] = \mathrm{E}[(\phi X_{t-1} + Z_t)Z_t] = \phi\mathrm{E}[X_{t-1}Z_t] + \mathrm{E}[Z_t^2] = \sigma_Z^2.$$

Hence, the system reduces to:

$$\begin{aligned}\gamma_X(0)\phi &= \gamma_X(1)\\ \sigma_Z^2 &= \gamma_X(0) - \phi\gamma_X(1).\end{aligned}$$

Now, consider $p = 2$: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. Multiply both sides of that equation, once each by by $X_{t-2}$, $X_{t-1}$, and $X_t$, to obtain:

$$\begin{aligned}X_t X_{t-2} - \phi_1 X_{t-1}X_{t-2} - \phi_2 X_{t-2}^2 &= Z_t X_{t-2},\\ X_t X_{t-1} - \phi_1 X_{t-1}^2 - \phi_2 X_{t-2}X_{t-1} &= Z_t X_{t-1},\\ X_t^2 - \phi_1 X_{t-1}X_t - \phi_2 X_{t-1}X_t &= X_t Z_t.\end{aligned}$$

We once again apply the expectation operator on each of these new equations to obtain:

$$\begin{aligned}\gamma_X(1) - \phi_1\gamma_X(1) - \phi_2\gamma_X(0) &= 0\\ \gamma_X(1) - \phi_1\gamma_X(0) - \phi_2\gamma_X(1) &= 0\\ \gamma_X(0) - \phi_1\gamma_X(1) - \phi_2\gamma_X(2) &= \sigma_Z^2.\end{aligned}$$

As in section 9.4.1 we consider the variance-covariance matrix

$$\Gamma_p = [\gamma_X(i-j)]_{i,j=1}^p,$$

and the vectors

$$\boldsymbol{\phi}_p = \left(\phi_1, \ldots, \phi_p\right)^\top \quad \text{and} \quad \boldsymbol{\gamma}(p;1) = \left(\gamma_X(1), \ldots, \gamma_X(p)\right)^\top.$$

For $p = 1$, $\Gamma_1 = \gamma_X(0)$; for $p = 2$,

$$\Gamma_2 = \begin{pmatrix}\gamma_X(0) & \gamma_X(1)\\ \gamma_X(-1) & \gamma_X(0)\end{pmatrix} = \begin{pmatrix}\gamma_X(0) & \gamma_X(1)\\ \gamma_X(1) & \gamma_X(0)\end{pmatrix}.$$

We can thus re-write the AR(1) and AR(2) systems above as:

$$\Gamma_p\boldsymbol{\phi}_p = \boldsymbol{\gamma}(p;1), \qquad \sigma_Z^2 = \gamma_X(0) - \boldsymbol{\phi}_p^\top\boldsymbol{\gamma}(p;1).$$

Equivalently, we obtain the **Yule-Walker equations**

$$\boldsymbol{\phi}_p = \Gamma_p^{-1}\boldsymbol{\gamma}(p;1), \qquad \sigma_Z^2 = \gamma_X(0) - \boldsymbol{\phi}_p^\top\boldsymbol{\gamma}(p;1),$$

which are very similar to the Yule-Walker forecast equations.[45] It is not hard to see that the equations hold for a general AR($p$).

45: Note that they do involve unknown autocovariances.

We can combine them with the method of moments,[46] to obtain the **Yule-Walker estimators**:

$$\widehat{\boldsymbol{\phi}}_p = \widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}(p;1), \qquad \widehat{\sigma}_Z^2 = \widehat{\gamma}_X(0) - \widehat{\boldsymbol{\phi}}_p^\intercal \widehat{\boldsymbol{\gamma}}(p;1),$$

where $\widehat{\Gamma}_p$ and $\widehat{\boldsymbol{\gamma}}(p;1)$ are obtained by substituting $\gamma_X$ by $\widehat{\gamma}_X$.

**Theorem:** for a large-enough sample size $n$, the Yule-Walker estimators are approximately normal, with

$$\widehat{\boldsymbol{\phi}}_p \sim \mathcal{N}\left(\boldsymbol{\phi}_p, \frac{1}{n}\sigma_Z^2\Gamma_p^{-1}\right).$$

In particular, for $p = 1$,

$$\widehat{\phi} \sim \mathcal{N}\left(\phi, \frac{1}{n}\sigma_Z^2\gamma_X^{-1}(0)\right).$$

That is, $\text{Var}(\widehat{\phi}) \sim \frac{1}{n}\sigma_Z^2\gamma_X^{-1}(0)$.

**Confidence interval for AR(1)** The **theoretical** confidence interval for the parameter $\phi$ of AR(1) is

$$\text{C.I.}_\alpha(\phi) \equiv \widehat{\phi} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sigma_Z\sqrt{\gamma_X^{-1}(0)},$$

where $z_{\alpha/2}$ is the standard normal quantile. Since $\sigma_Z^2$ and $\gamma_X(0)$ are unknown, we replace them with estimators to obtain the **empirical** (practical) confidence interval

$$\text{C.I.}_\alpha(\phi) \approx \widehat{\phi} \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\widehat{\sigma}_Z\sqrt{\widehat{\gamma}_X^{-1}(0)},$$

where

$$\widehat{\phi} = \frac{\widehat{\gamma}_X(1)}{\widehat{\gamma}_X(0)} \quad \text{and} \quad \widehat{\sigma}_Z^2 = \widehat{\gamma}_X(0) - \widehat{\phi}\widehat{\gamma}_X(1).$$

**Confidence interval for AR(2)** The limiting variance-covariance matrix for the Yule-Walker estimators $\widehat{\phi}_1, \widehat{\phi}_2$ is

$$\sigma_Z^2\Gamma_2^{-1} = \begin{bmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix}.$$

Indeed, we have

$$\Gamma_2 = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(1) & \gamma_X(0) \end{bmatrix} \implies \Gamma_2^{-1} = \frac{1}{\gamma_X^2(0) - \gamma_X^2(1)}\begin{bmatrix} \gamma_X(0) & -\gamma_X(1) \\ -\gamma_X(1) & \gamma_X(0) \end{bmatrix}.$$

Previously, we saw that

$$\gamma_X(1) = \sigma_Z^2\frac{\phi_1}{(1+\phi_2)\left\{(1-\phi_2)^2 - \phi_1^2\right\}} \quad \text{and} \quad \gamma_X(0) = \sigma_Z^2\frac{1-\phi_2}{(1+\phi_2)\left\{(1-\phi_2)^2 - \phi_1^2\right\}}.$$

Substituting these in the expression for $\Gamma_2^{-1}$ yields the desired result.

46: We simply replace the mean with the sample mean and the autocovariances with the sample autocovariances.

In particular, $\text{Var}(\widehat{\phi}_1) \sim \frac{1}{n}(1-\phi_2^2)$ and $\text{Var}(\widehat{\phi}_2) \sim \frac{1}{n}(1-\phi_2^2)$. Consequently,

$$\text{C.I.}_\alpha(\phi_1) \equiv \widehat{\phi}_1 \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{1-\widehat{\phi}_2^2} \quad \text{and} \quad \text{C.I.}_\alpha(\phi_2) \equiv \widehat{\phi}_2 \pm z_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{1-\widehat{\phi}_2^2},$$

where $\widehat{\phi}_1$ and $\widehat{\phi}_2$ are obtained from the Yule-Walker estimators.

### 9.5.4 Example

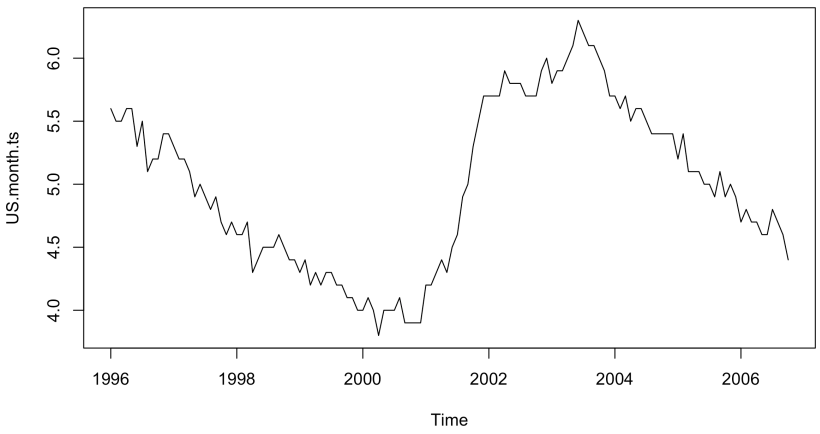
We illustrate this last concept with a simple example.

**US Unemployment Data** The United States' monthly unemployment rate starting with January 1996 is collected in `USunemp.txt` [3].

```
US.month <- c(5.6,5.5,5.5,5.6,5.6,5.3,5.5,5.1,5.2,5.2,
              5.4,5.4,5.3,5.2,5.2,5.1,4.9,5.0,4.9,4.8,
              4.9,4.7,4.6,4.7,4.6,4.6,4.7,4.3,4.4,4.5,
              4.5,4.5,4.6,4.5,4.4,4.4,4.3,4.4,4.2,4.3,
              4.2,4.3,4.3,4.2,4.2,4.1,4.1,4.0,4.0,4.1,
              4.0,3.8,4.0,4.0,4.0,4.1,3.9,3.9,3.9,3.9,
              4.2,4.2,4.3,4.4,4.3,4.5,4.6,4.9,5.0,5.3,
              5.5,5.7,5.7,5.7,5.7,5.9,5.8,5.8,5.8,5.7,
              5.7,5.7,5.9,6.0,5.8,5.9,5.9,6.0,6.1,6.3,
              6.2,6.1,6.1,6.0,5.9,5.7,5.7,5.6,5.7,5.5,
              5.6,5.6,5.5,5.4,5.4,5.4,5.4,5.4,5.2,5.4,
              5.1,5.1,5.1,5.0,5.0,4.9,5.1,4.9,5.0,4.9,
              4.7,4.8,4.7,4.7,4.6,4.6,4.8,4.7,4.6,4.4)
```

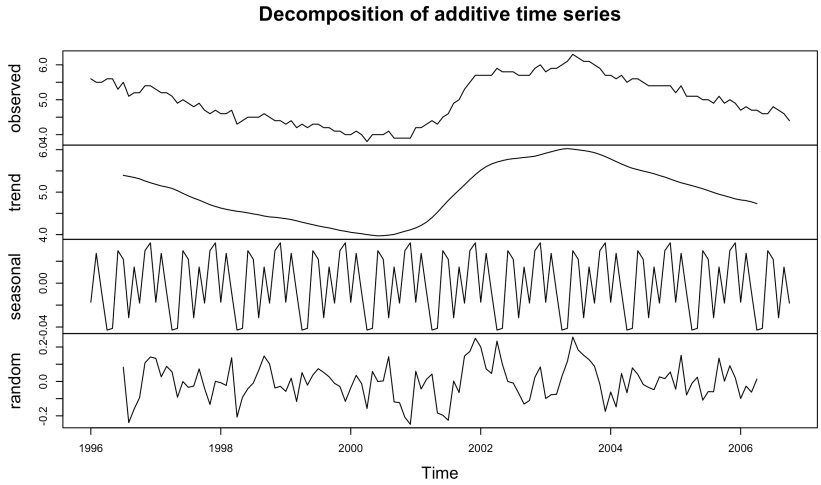
We put the data in a `ts` object and plot the data.

```
US.month.ts <- ts(US.month,start=c(1996,1), freq=12)
plot.ts(US.month.ts)
```




The time series is clearly not stationary, so we decompose it.

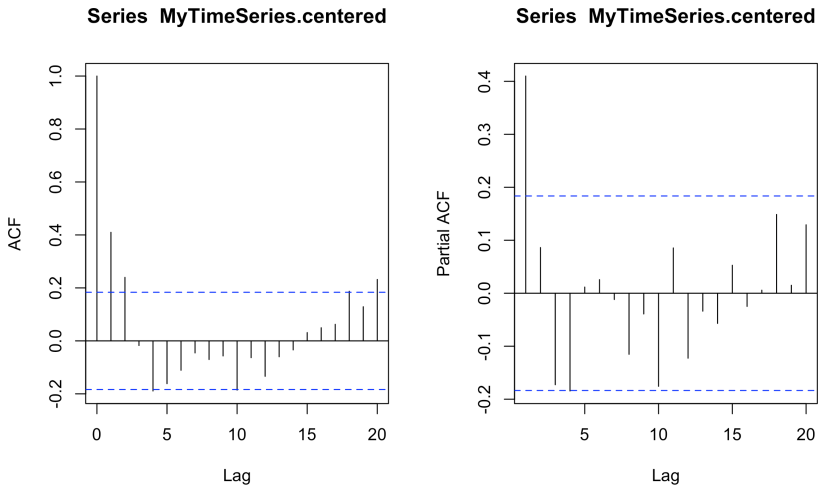
```
plot(decompose(US.month.ts))
```




We recover the stationary part from this decomposition and analyse it as below.[47]

47: The default smoother for the `decompose()` function is a moving average.

```
Stationary <- decompose(US.month.ts)$random
MyTimeSeries = Stationary[7:120]
mean = mean(MyTimeSeries);
MyTimeSeries.centered = MyTimeSeries-mean
par(mfrow=c(1,2))
acf(MyTimeSeries.centered)
pacf(MyTimeSeries.centered)
```




The ACVF/ACF has non-zero values at various lags $h$ (outside the band); the PACF has all zero values for $h > 1$ (inside the band); the eye test suggests an AR(1) model.

But a formal test (using the Yule-Walker) method suggests instead that the order of the model is more likely to be $p = 4$.

```
n = length(MyTimeSeries)
fit.ar <- ar(MyTimeSeries.centered,method="yule-walker")
fit.ar$order
```

```
[1] 4
```

The Yule-Walker estimates for the coefficients $\phi_1, \phi_2, \phi_3, \phi_4$ and for the random component variance $\sigma_Z^2$ are given by:

```
fit.ar$ar
fit.ar$var.pred
```

```
[1] 0.3576   0.1788  -0.1008  -0.1845
[1] 0.009106
```

We compute the limiting variance covariance matrix $\sigma_Z^2 \Gamma_4^{-1}$ as follows.

```
rho = acf(MyTimeSeries.centered)$acf
gamma.0 = var(MyTimeSeries.centered)
sigma.2.Z = fit.ar$var.pred
gamma.h = rho * gamma.0
Gamma.4 = matrix(c(gamma.h[1],gamma.h[2],gamma.h[3],gamma.h[4],
         gamma.h[2],gamma.h[1],gamma.h[2],gamma.h[3],
         gamma.h[3],gamma.h[2],gamma.h[1],gamma.h[2],
         gamma.h[4],gamma.h[3],gamma.h[2],gamma.h[1]),4,4)
Gamma.4.inv = solve(Gamma.4)
(limit.V_CV = sigma.2.Z*Gamma.4.inv)
```

```
           [,1]       [,2]       [,3]       [,4]
[1,]  1.0014029 -0.3900340 -0.1511252  0.1729498
[2,] -0.3900340  1.1234465 -0.3050721 -0.1511252
[3,] -0.1511252 -0.3050721  1.1234465 -0.3900340
[4,]  0.1729498 -0.1511252 -0.3900340  1.0014029
```

Note that we can obtain the matrix directly from the `fit.ar` object.

```
(n-1)*fit.ar$asy.var.coef
```

Finally, we simply apply the formulas to obtain approximate 95% confidence intervals on the AR(4) coefficients.

```
rbind(fit.ar$ar - 1.96/sqrt(n)*sqrt(diag(limit.V_CV)),
      fit.ar$ar + 1.96/sqrt(n)*sqrt(diag(limit.V_CV)))
```

```
          [,1]        [,2]        [,3]          [,4]
[1,] 0.1739213 -0.01581274 -0.29541378 -0.3682125028
[2,] 0.5413204  0.37333082  0.09372978 -0.0008134319
```

## 9.6 Diagnostic Tests

Assume that an AR(1) model $X_t = \phi X_{t-1} + Z_t$ is fit to the data, i.e., we estimate $\phi$ by $\widehat{\phi}$ and $\sigma_Z^2$ by $\widehat{\sigma}_Z^2$. We can now compute the time series of **residuals**

$$\widehat{Z}_t = X_t - \widehat{\phi} X_{t-1}.$$

Note that $\widehat{Z}_t \neq Z_t$, in general, but we would expect them to be near one another if the fit is good. As such, the properties of $Z_t$ should be similar to those of $\widehat{Z}_t$.

It is important to ensure that the model is an **adequate fit** to the data – in particular, the residuals should not exhibit significant autocorrelations at lags $|h| \geq 1$.

If the random variables $\widehat{Z}_t$ are i.i.d., then the correlations $\rho_X(|h|) = 0$ at any lag $h \neq 0$ zero. However, the sample correlations are typically not zero, since there usually are random fluctuations in the data. In general, for large $n$, the sample correlation at any lag is normally distributed with mean zero and variance $1/n$. This provides a 95% confidence interval for the sample autororrelations: $\pm 1.96/\sqrt{n}$.[48]

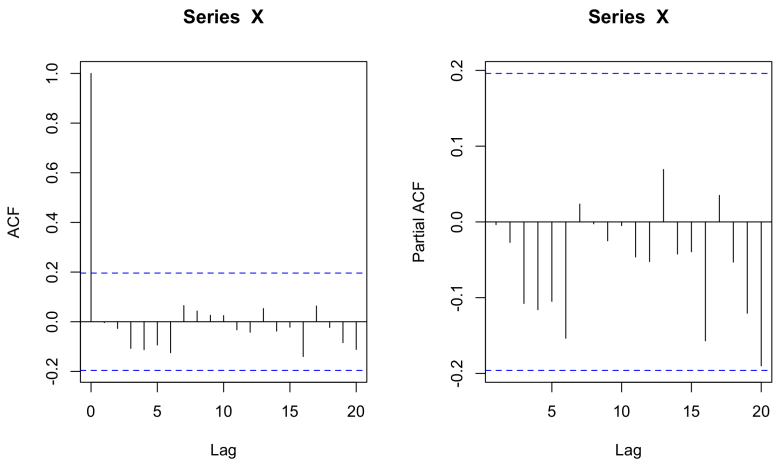
48: This corresponds to the blue lines seen on the ACF plot. Whenever the sample ACF is within the confidence intervals, the rule-of-thumb is to treat the corresponding auto-correlation as zero.

**White Noise** The 95% threshold for a white noise time series with $\mu = 0$ and $\sigma^2 = 1$, with $n = 100$ observations is computed below.

```
n = 100
set.seed(1)
X = rnorm(n)
(threshold = 1.96/sqrt(n))
```

```
[1] 0.196
```

```
par(mfrow=c(1,2))
acf(X)
pacf(X)
```



### 9.6.1 Ljung-Box Test

That is not the only approach, however. Let $h$ be a positive integer (the **lag**) and define

$$Q_h = n \sum_{j=1}^{h} \frac{\widehat{\gamma}_X(j)}{\widehat{\gamma}_X(0)}.$$

Under the null hypothesis that the residuals are i.i.d. , the statistic $Q_h$ has a $\chi^2$ distribution with $h$ degrees of freedom. A large value of $Q_h$ suggests that the sample autocorrelations are too large for the data to arise from the draw of an i.i.d. sequence. We would therefore reject the i.i.d. hypothesis at confidence level $\alpha$ if $Q_h > \chi^2_{1-\alpha}(h)$.

**White Noise** We can conduct the Ljung-Box test on the white noise time series from the previous section, with $h = 2$, say.

```
Box.test(X,type="Ljung",lag=2,fitdf=0)
```

```
        Box-Ljung test

data:  X
X-squared = 0.077367, df = 2, p-value = 0.9621
```

Thus, we conclude that the data is compatible with $X$ being i.i.d., at confidence level $\alpha = 0.05$

**AR(1) Model** This time, we simulate an auto-regressive model (so the time series not i.i.d.) and repeat the procedure.

```
set.seed(1)
MyTimeSeries = arima.sim(model=list(ar=c(0.8)),
                         n=1000,rand.gen=rnorm)
Box.test(MyTimeSeries,type="Ljung",lag=2,fitdf=0)
```

```
        Box-Ljung test

data:  MyTimeSeries
X-squared = 904.66, df = 2, p-value < 2.2e-16
```

We see that the i.i.d. assumption is correctly rejected.

The Ljung-Box test is applied to the residuals. The parameter `fitdf` is the number of the parameters that need to be estimated. In an ARMA$(p, q)$, model, it is $p + q$.[49]

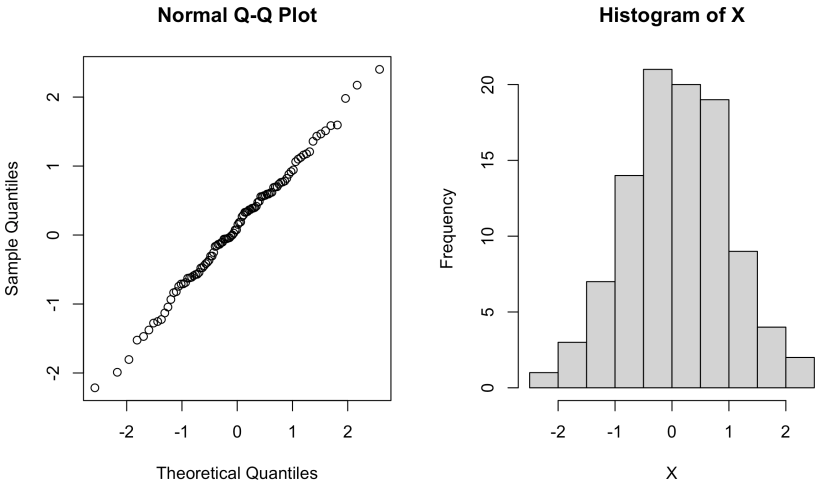
When we reject the hypothesis that the residuals are i.i.d., we are claiming that the fitted ARMA$(p, q)$ model is incorrect.[50] If the test results are compatible with the null hypothesis, we must also verify that the residuals are normally distributed, however, either by plotting a Q-Q plot or a histogram.

In the first example, the time series $\{X_t\}$ is normally distributed.

49: Be careful! Here, we are testing whether the sequence `MyTimeSeries`, which we know to be AR(1), could be white noise (i.i.d.), which is why we use `fitdf=0`. That is, we are assuming that it is a time series of residuals that arose naturally, not as a result of having fit an ARMA$(p, q)$ model to the data. The `lag` parameter represents the positive value $h$.

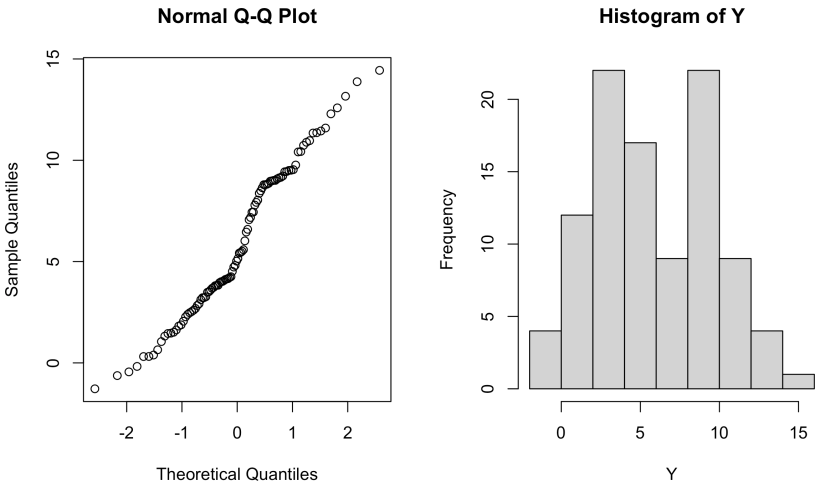
50: We must thus remove $p + q$ degrees of freedom from $h$, since we had to estimate $p + q$ parameters from the data before obtaining the residual time series.

```
par(mfrow=c(1,2))
qqnorm(X)
hist(X)
```




In the second case, the time series $\{Y_t\}$ is a random walk, and it is not normally distributed.

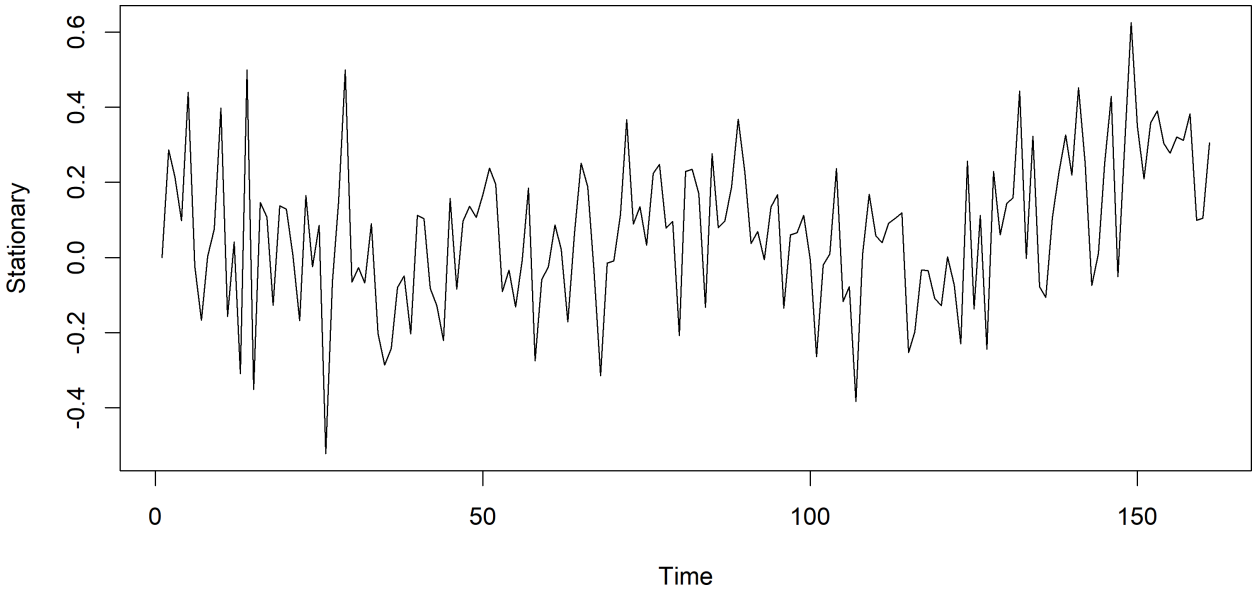
```
Y = cumsum(X)
par(mfrow=c(1,2))
qqnorm(Y)
hist(Y)
```




### 9.6.2 Example: Temperature

We consider the temperature data from page 499; it is clearly not stationary, so we conduct exponential smoothing on it, with smoothing parameter 0.1, yielding the time series `MySmoothedTS1`, which is then centered.

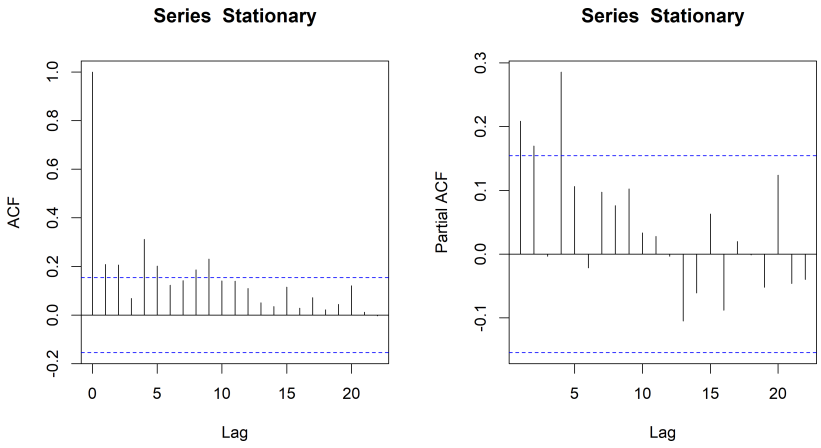
```
Stationary = Temperature - MySmoothedTS1
plot.ts(Stationary, type="l")
```




51: Although there is a bit of growth near the end.

This time series certainly appears stationary.[51] Could it arise from an ARMA($p, q$) model? We plot its ACF and PACF.

```
par(mfrow=c(1,2))
acf(Stationary); pacf(Stationary)
```




52: Be sure to understand why!

AR(4) seems like a reasonable model;[52] Yule-Walker agrees.

```
(fit.ar.yw <- ar(Stationary,method="yule-walker"))
```

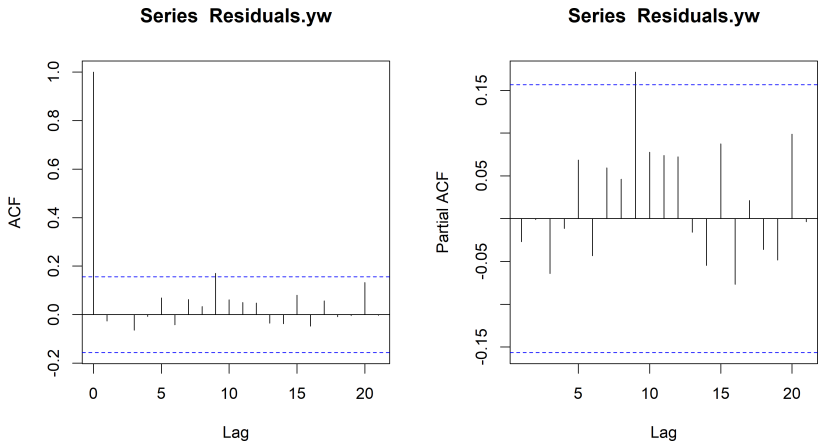
```
Coefficients:
      1        2        3        4
 0.1745   0.1218  -0.0529   0.2855

Order selected 4  sigma^2 estimated as  0.03412
```

We compute the residuals for which $\widehat{Z}_t = \widehat{\phi}(B)X_t$ is defined.

```
phi.yw = fit.ar.yw$ar
n = length(Stationary)
Residuals.yw <- fit.ar.yw$resid
Residuals.yw = na.omit(Residuals.yw)
```

The ACF and PACF of the obtained residuals are as follows.

```
par(mfrow=c(1,2))
acf(Residuals.yw)
pacf(Residuals.yw)
```




There is no dependence left in the residuals (although you can argue that there is a significant lag at 9); the fit seems appropriate.

We can conduct the Box-Ljung test with $h = 5 > 4 = p + q$, say.

```
Box.test(Residuals.yw,type="Ljung",lag=5,fitdf=4)
```

```
Box-Ljung test

data:  Residuals.yw
X-squared = 1.5724, df = 1, p-value = 0.2099
```

The $p-$value is small, but not that small... does the value of $h$ matter? What if we used $h = 4$, instead?

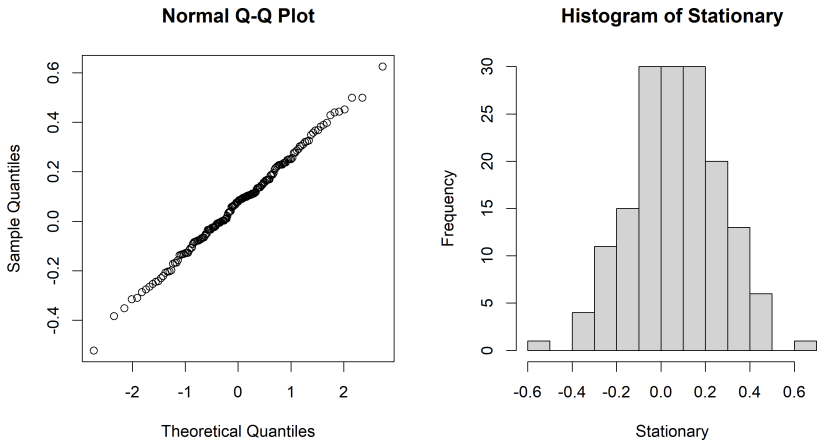
```
Box.test(Residuals.yw,type="Ljung",lag=4,fitdf=4)
```

```
Box-Ljung test

data:  Residuals.yw
X-squared = 0.79007, df = 0, p-value < 2.2e-16
```

The $p-$value is indeed much smaller than 0.05, but it's not clear how the test implementation handles the case where $h = p + q$.

Either way, we should study the normality of the residuals visually.

```
par(mfrow=c(1,2))
qqnorm(Stationary)
hist(Stationary)
```




So, what do you think? We will return this example in the next section.

One thing to note is that the Box-Ljung test is not unanimously favoured by practitioners: see the Breusch-Godfrey test for an alternative.

## 9.7 Maximum Likelihood Estimation

We start with a brief refresher on the topic.

### 9.7.1 I.I.D. Random Variables

Assume that the random variables $X_1, \ldots, X_n$ are i.i.d. with a known probability density function $f_X(x;\theta)$. The objective of **maximum likelihood estimation** (MLE) is to find the parameter $\theta$ that best fits the observed data, in the MLE sense.[53]

53: This does not have to be a univariate problem; we might be interested in the parameter vector $\boldsymbol{\theta}$, depending on the context. The principle is the same, but we will be working with $\nabla_{\boldsymbol{\theta}}$ instead of the derivative $\frac{d}{d\theta}$.

The **likelihood function** is

$$L(\theta) = L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} f_X(X_i;\theta).$$

The **log-likelihood function** is $\ell(\theta) = \log L(\theta) = \ln L(\theta)$. The **maximum likelihood estimator** $\widehat{\theta}_{\text{MLE}}$ is a parameter value (often unique, for commonly-used $f$, but it also depends on the observed data) satisfying

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta).$$

**Example: Exponential Distribution** Assume that $X_1, \ldots, X_n$ is a random sample from an exponential distribution. Recall that $X \sim \text{Exp}(\beta)$, $\theta = \beta > 0$ if

$$f_X(x;\beta) = \begin{cases} \beta^{-1}\exp(-x/\beta), & x > 0; \\ 0, & x \leq 0 \end{cases}$$

The likelihood function is:

$$L(\beta) = \beta^{-n}\prod_{i=1}^{n}\exp(-X_i/\beta) = \beta^{-n}\exp\left(-\beta^{-1}\sum_{i=1}^{n}X_i\right),$$

and the log-likelihood is:

$$\ell(\beta) = -n\log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n}X_i.$$

To optimize $\ell$, we must find its critical points with respect to $\beta$. There is only one such point, since

$$\frac{\partial\ell(\beta)}{\partial\beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n}X_i = 0 \implies \widehat{\beta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X}.$$

Technically, this only tells us that $\overline{X}$ is a critical point of $\ell(\beta)$, not necessarily that it is a maximizer. But

$$\left.\frac{\partial^2\ell(\beta)}{\partial\beta^2}\right|_{\beta=\overline{X}} = -n\overline{X}^2 < 0,$$

so $\widehat{\beta}_{\text{MLE}} = \overline{X}$ is indeed a global maximizer, according to the second derivative test.

The sample mean is not only the MLE estimator for the Exponential distribution, however.

**Example: Normal Distribution** Assume that $Z_1, \ldots, Z_n$ is a i.i.d. sample from a normal distribution with mean $\mu$ and variance $\sigma_Z^2$. The likelihood function is

$$L(\mu, \sigma_Z) = \frac{1}{(\sqrt{2\pi})^n\sigma_Z^n}\exp\left(-\frac{1}{2\sigma_Z^2}\sum_{i=1}^{n}(Z_i-\mu)^2\right),$$

and the log-likelihood is:

$$\ell(\mu, \sigma_Z) = -\frac{n}{2}\log(2\pi) - n\log\sigma_Z - \frac{1}{2\sigma_Z^2}\sum_{i=1}^{n}(Z_i-\mu)^2.$$

We proceed as above, differentiating with respect to $\mu$ to find the critical points:

$$\frac{\partial\ell(\mu,\sigma_Z)}{\partial\mu} = -\frac{1}{\sigma_Z^2}\sum_{i=1}^{n}(Z_i-\mu) = 0 \implies \widehat{\mu}_{\text{MLE}} = \overline{Z} = \frac{1}{n}\sum_{i=1}^{n}Z_i.$$

Substituting $\widehat{\mu}_{\text{MLE}} = \overline{Z}$ in $L$, differentiating with respect to $\sigma_Z$, setting to 0 and solving yields

$$\widehat{\sigma}^2_{Z;\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \widehat{\mu}_{\text{MLE}})^2,$$

demonstrating that the MLE estimators are not always **unbiased**.

### 9.7.2 Time Series Model

We now assume that $X_1, \ldots, X_n$ are observation from a stationary time series. Let $f_n(x_1, \ldots, x_n)$ be their joint density.[54] We further assume that the time series is **Gaussian** and **centered**.[55]

54: No longer in the product form.

55: This is an important assumption – we need to verify that it applies to the data of interest.

We introduce the following notation:

$$\mathbf{X}_n = (X_1, \ldots, X_n)^\top, \qquad \widehat{\mathbf{X}}_n = (\widehat{X}_1, \ldots, \widehat{X}_n)^\top, \qquad \mathbf{U}_n = (U_1, \ldots, U_n)^\top,$$

where $U_i = X_i - \widehat{X}_i$, $i = 1, \ldots, n$, are the **innovations**. Recall that $\Gamma_n = \text{E}[\mathbf{X}_n^\top \mathbf{X}_n] = [\gamma_X(i-j)]_{i,j=1}^n$ is the **variance-covariance matrix** of $\mathbf{X}_n$ (see Section 9.4.1).

The likelihood (the joint density of $X_1, \ldots, X_n$) is

$$L = \frac{1}{(2\pi)^{n/2}} \frac{1}{\det(\Gamma_n)^{1/2}} \exp\left(-\frac{1}{2}\mathbf{X}_n^\top \Gamma_n^{-1} \mathbf{X}_n\right),$$

where $\det(\Gamma_n)$ is the determinant. Note that the ACVF (and hence, also the covariance matrix $\Gamma_n$) depends on model parameters.

For example, if the model is AR(1), then $\gamma_X(h) = \sigma_Z^2 \phi^h/(1-\phi^2)$. Thus, its variance-covariance matrix and the log-likelihood depend on the model parameters $\sigma_Z, \phi$, so that we can write $L(\sigma_Z, \phi)$.

In this particular case, the MLE estimators are obtained by maximizing $L(\sigma_Z, \phi)$ with respect to $\sigma_Z, \phi$. In the general case, there are no no explicit formulas to do so and everything must be conducted numerically (see Chapter 4).

It turns out that

$$\mathbf{X}_n^\top \Gamma_n^{-1} \mathbf{X}_n = \mathbf{U}_n^\top \mathbf{D}^{-1} \mathbf{U}_n,$$

where $\mathbf{D} = \text{diag}(v_0, \ldots, v_{n-1})$, for $v_i = \text{E}\left[\left(X_{i+1} - \widehat{X}_{i+1}\right)^2\right]$.[56] Thus, we have

56: See Section 9.9.3 for more details.

$$\mathbf{X}_n^\top \Gamma_n^{-1} \mathbf{X}_n = \sum_{i=1}^{n}(X_i - \widehat{X}_i)^2/v_{i-1}.$$

Furthermore, $\det(\Gamma_n) = v_0 \cdots v_{n-1}$, and so the likelihood function takes the form

$$L = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{v_0 \cdots v_{n-1}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \widehat{X}_i)^2/v_{i-1}\right).$$

The form of $L$ above can be used as long as we have formulas for $\widehat{X}_i$, even if those do not arise from the innovation algorithm.

**Theorem:** the MLE estimator $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ is **asymptotically normal**,[57] with mean $\boldsymbol{\theta}$ and variance $n^{-1}\text{V}(\boldsymbol{\theta})$, where $\text{V}(\boldsymbol{\theta})$ is a covariance matrix.

57: See Section 9.9.2.

If the data arises from an ARMA$(p, q)$ process, we would use the innovation algorithm to express $\widehat{X}_i$ in terms of the coefficients $\theta_1, \ldots, \theta_q$, and then plug them into the likelihood function

$$L(\boldsymbol{\theta}) = L(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_Z^2),$$

which can be maximized using the MLE approach, as above.

**AR(1)** Consider the auto-regressive model $X_t = \phi X_{t-1} + Z_t$, where $Z_t$ are i.i.d. normal random variables with mean 0 and variance $\sigma_Z^2$, starting with $t = 1$. Then $\widehat{X}_{i+1} = \phi X_i$ and $v_i = \text{E}[(X_{i+1} - \widehat{X}_{i+1})^2] = \sigma_Z^2$ for all $i = 1, \ldots, n-1$. The likelihood function is thus:

$$L = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma_Z^n} \exp\left( -\frac{1}{2} \sum_{i=2}^{n} (X_i - \phi X_{i-1})^2 / \sigma_Z^2 \right).$$

Ignoring the constant term $\frac{1}{(2\pi)^{n/2}}$, the log-likelihood is

$$\ell = -n \log \sigma_Z - \frac{1}{2\sigma_Z^2} \sum_{i=2}^{n} (X_i - \phi X_{i-1})^2.$$

Hence,

$$\widehat{\phi}_{\text{MLE}} = \frac{\sum_{i=2}^{n} X_{i-1} X_i}{\sum_{i=2}^{n} X_{i-1}^2}$$

and

$$\widehat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=2}^{n} (X_i - \widehat{\phi}_{\text{MLE}} X_{i-1})^2.$$

If $\sigma_Z$ is known, then we do not need to use the MLE estimator; we have $\theta = \phi$ and $\text{V}(\theta)$ becomes

$$\text{V}(\theta) = \text{V}(\phi) = \sigma_Z^2 (1 - \phi^2).$$

We note that the MLE estimator of $\phi$ (as well as its asymptotic variance) are the same as those obtained by the Yule-Walker procedure.

**AR($p$)** In general, for AR(p) models, the Yule-Walker estimator and MLE of $(\phi_1, \ldots, \phi_p)$ also agree; in both cases the asymptotic variance is

$$\text{V}(\phi_1, \ldots, \phi_p) = \sigma_Z^2 \Gamma_p^{-1}.$$

For AR(2) we have seen that

$$\text{V}(\phi_1, \phi_2) = \begin{pmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & 1 - \phi_2^2 \end{pmatrix}.$$

However, the MLE and Yule-Walker estimators of variance $\sigma_Z^2$ do not need to agree in general!

**AR($p$) Models (Revisited)** For simplicity's sake, consider the AR(1) model $X_t = \phi X_{t-1} + Z_t$, where $Z_t$ are i.i.d. normal with mean 0 and variance $\sigma_Z^2$.

We assume that $\mu = 0$; then,

$$L = \frac{1}{(\sqrt{2\pi})^n \sigma_Z^n} \exp\left(-\frac{1}{2\sigma_Z^2} \sum_{i=1}^{n} Z_i^2\right),$$

Since $Z_t = X_t - \phi X_{t-1}$, this transforms to

$$L(\phi, \sigma_Z) = \frac{1}{(\sqrt{2\pi})^n \sigma_Z^n} \exp\left(-\frac{1}{2\sigma_Z^2} \sum_{i=2}^{n} (X_i - \phi X_{i-1})^2\right).$$

The likelihood function now depends explicitly on $\phi$ and $\sigma_Z$, and we can continue as we did in the previous section (without having to use innovations).

This approach works for arbitrary AR($p$) models, but not for MA($q$) or general ARMA($p, q$) models.

### 9.7.3 Order Selection

We have discussed a visual criterion to identify a time series follows a AR($p$) or MA($q$) model, as well as a formal approach (Yule-Walker). Another classical approach to ARMA ($p, q$) order selection is provided by the **Akaike information criteria** (AIC) method.

We consider several ARMA($p, q$) models, all depending on parameter vectors $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$. The `ar()` function in `R`, for instance, has $q = 0$ and tries $p = 1, \ldots, 12$.

For each model we calculate the following expression:

$$\text{AIC} = 2 \log L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_Z) - 2(p + q + 1) \frac{n}{n - p - q - 2}.$$

When $q = 0$ (i.e., when we consider AR($p$) models), this reduces to:

$$\text{AIC} = 2 \log L(\boldsymbol{\phi}, \sigma_Z) - 2(p + 1) \frac{n}{n - p - 2}.$$

The AIC method chooses a model with a high likelihood but penalizes models with too many parameters (i.e., if $p$ and $q$ are too large).[58] Another function, `arima()`, computes AIC as follows:

$$\text{AIC} = -2 \log L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_Z) + 2(p + q + k + 1),$$

where $k$ is the number of additional parameters to estimate (in our case, $k = 1$ since we estimate $\sigma_Z$ and there is no seasonality); the optimal model is the one that **minimizes** that version of AIC.

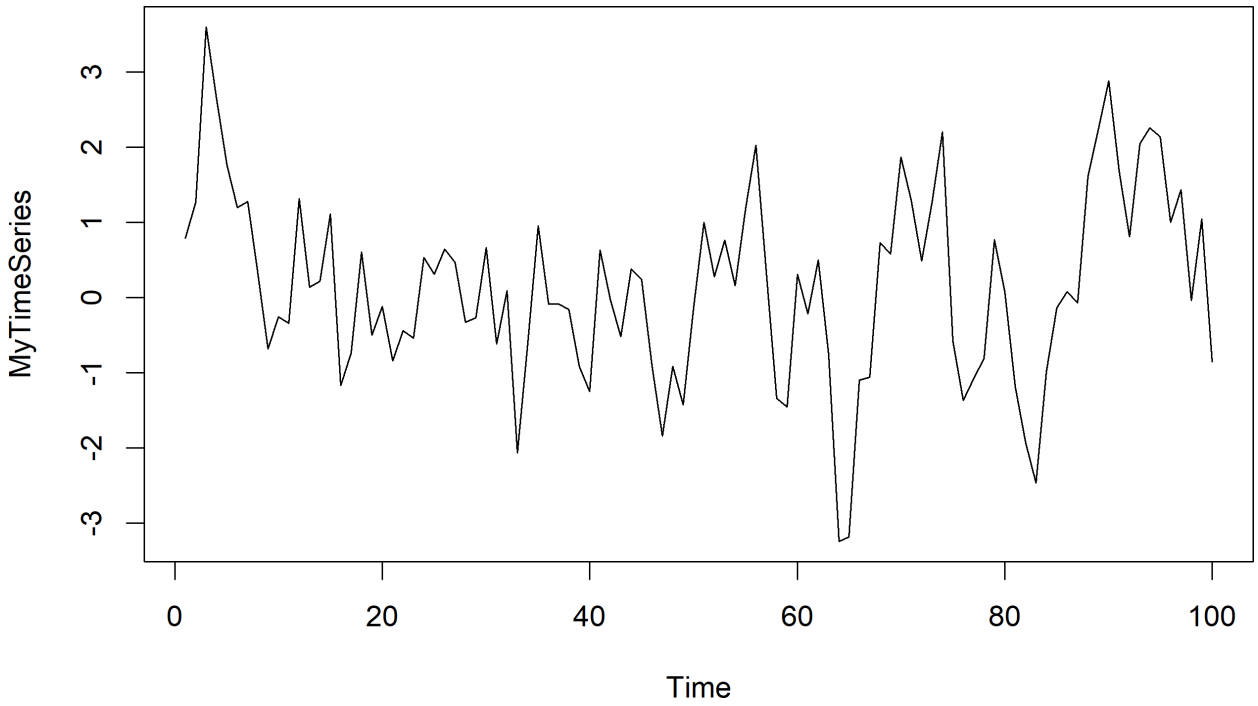
58: Note that maximizing AIC is equivalent to minimizing −AIC.

### 9.7.4 Examples

We consider three examples: an artificial time series, a Lake Huron time series, and a continuation of the Temperature example.

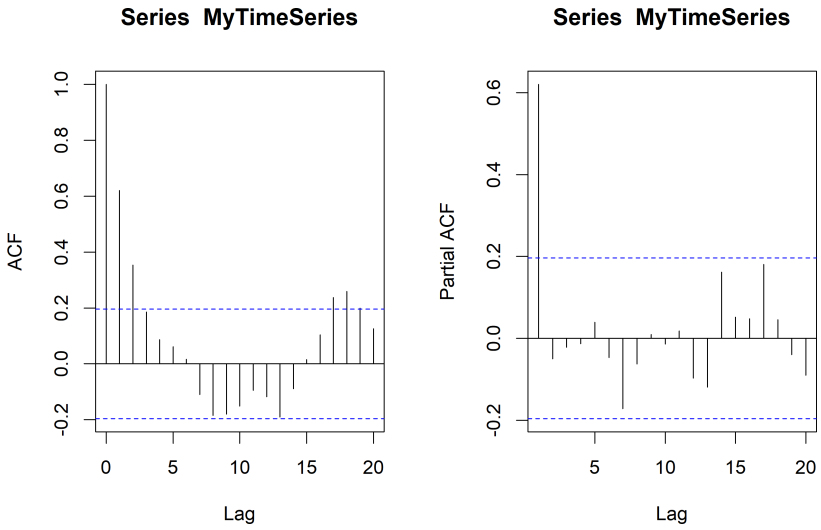
**Example: Artificial Data** This artificial time series appears to be stationary (more or less).

```
MyTimeSeries <- c(0.793,  1.270,  3.600,  2.649,  1.767,
                  1.198,  1.278,  0.347, -0.683, -0.255,
                 -0.338,  1.316,  0.142,  0.218,  1.118,
                 -1.170, -0.731,  0.609, -0.498, -0.118,
                 -0.839, -0.439, -0.537,  0.537,  0.314,
                  0.647,  0.470, -0.323, -0.264,  0.670,
                 -0.616,  0.092, -2.062, -0.603,  0.958,
                 -0.084, -0.083, -0.156, -0.914, -1.250,
                  0.634, -0.031, -0.519,  0.383,  0.241,
                 -0.903, -1.838, -0.912, -1.422, -0.134,
                  1.004,  0.282,  0.766,  0.164,  1.180,
                  2.030,  0.341, -1.337, -1.452,  0.313,
                 -0.212,  0.500, -0.762, -3.239, -3.179,
                 -1.094, -1.055,  0.735,  0.582,  1.869,
                  1.295,  0.492,  1.272,  2.210, -0.574,
                 -1.363, -1.076, -0.809,  0.774,  0.082,
                 -1.180, -1.925, -2.463, -0.983, -0.135,
                  0.081, -0.071,  1.612,  2.241,  2.884,
                  1.686,  0.811,  2.046,  2.260,  2.142,
                  1.003,  1.435, -0.039,  1.049, -0.855)

plot.ts(MyTimeSeries)
```



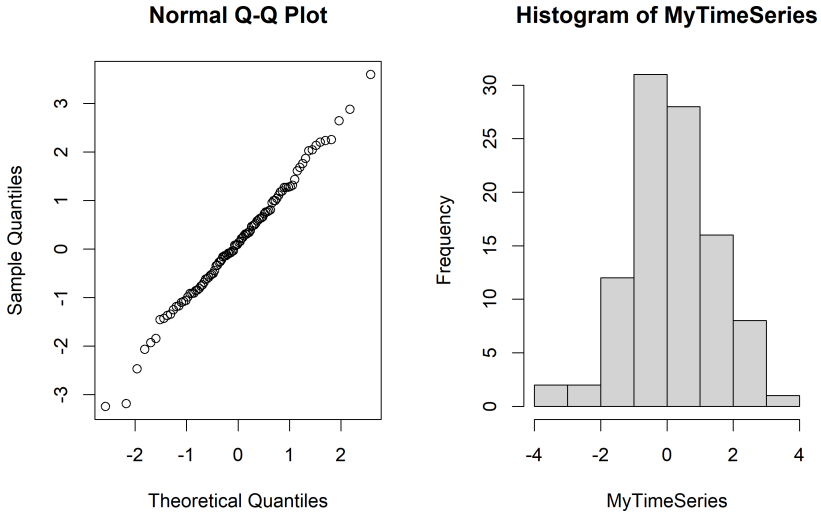

The ACF and PACF displays suggest that the data could arise from an AR(1) process.

```
par(mfrow=c(1,2))
acf(MyTimeSeries)
pacf(MyTimeSeries)
```




We draw your attention to the structure of the ACF; a continuous stretch of positive values, followed by a continuous stretch of negative values, followed by a continuous stretch of positive values (and so on?). This could be indicative of a seasonality effect in the data (see Section 9.9.1). Are the values of the time series normally distributed?

```
par(mfrow=c(1,2))
qqnorm(MyTimeSeries)
hist(MyTimeSeries)
```




We perform model estimation using two approaches: Yule-Walker and MLE.[59]

59: We do not need normality for the former, but we do need it for the latter, which is why we took the time to verify that the time series values *could* be normally distributed.

```
(fit.ar.yw <- ar(MyTimeSeries,method="yule-walker"))
```

```
Coefficients:
     1
0.6201

Order selected 1  sigma^2 estimated as  0.9707
```

```
(fit.ar.mle <- ar(MyTimeSeries,method="mle"))
```

```
Coefficients:
     1
0.6197

Order selected 1  sigma^2 estimated as  0.9458
```

In both cases, the selected model is AR(1), but the estimated parameters are slightly different. However, the estimates of the autoregressive parameter $\phi$ should be be the same, regardless of the method used. What is going on?

The difference comes from the fact that the R implementation of the MLE approach uses a fairly complicated optimization algorithm, leading to numerical discrepancies – the differences are not significant, to be honest, which is comforting.

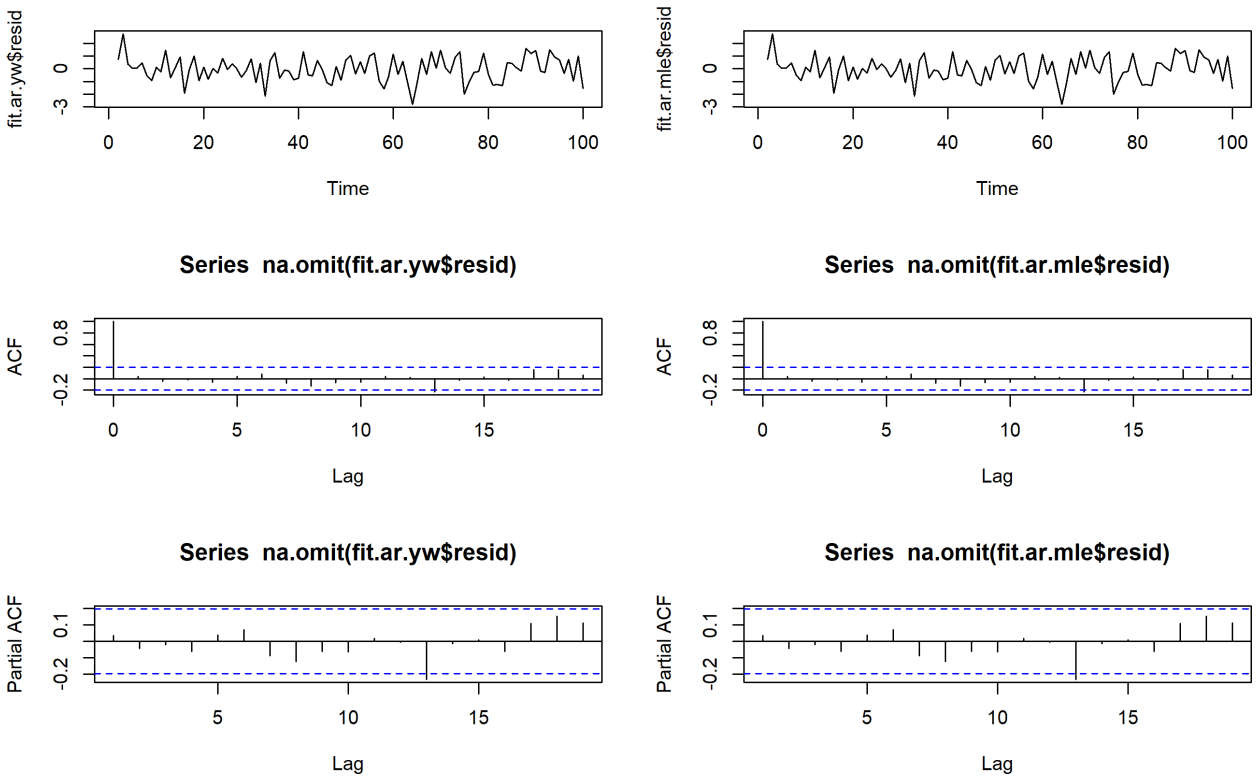
Note, however, that the estimates for $\sigma_Z^2$ are different, as they should be, since one is unbiased (Yule-Walker), whereas the other is biased (MLE).

The order and the coefficient value can be extracted using the following code – the displays are suppressed as they can be read above.

```
fit.ar.yw$order
fit.ar.mle$order
fit.ar.yw$ar
fit.ar.mle$ar
```

In order to assess the fit, we can take a look at the residuals.

```
par(mfrow=c(3,2))
plot.ts(fit.ar.yw$resid)
plot.ts(fit.ar.mle$resid)
acf(na.omit(fit.ar.yw$resid))
acf(na.omit(fit.ar.mle$resid))
pacf(na.omit(fit.ar.yw$resid))
pacf(na.omit(fit.ar.mle$resid))
```




In both cases, the residuals certainly look like they could arise from i.i.d. processes.

What would the prediction for the next value of the time series be, in both cases?

```
predict(fit.ar.yw)
```

```
$pred
Time Series:
Start = 101
End = 101
Frequency = 1
[1] -0.4737512
```

```
$se
Time Series:
Start = 101
End = 101
Frequency = 1
[1] 0.9852252
```

```
predict(fit.ar.mle)
```

```
$pred
Time Series:
Start = 101
End = 101
Frequency = 1
[1] -0.4754649
```

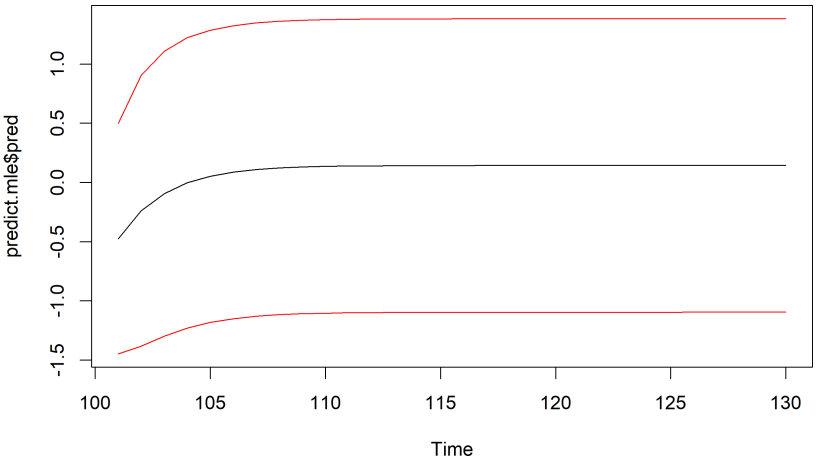
```
$se
Time Series:
Start = 101
End = 101
Frequency = 1
[1] 0.9725163
```

The different predictions values stem from the fact that $\phi_{\text{YW}}$ is slightly different from $\phi_{\text{MLE}}$.

Both models seem appropriate – which one should we choose? We select the MLE model, for no particular reason. We forecast the next 30 iterations of the model; the **confidence bands** with confidence bands obtained as

$$\text{prediction} \pm \text{standard error of prediction}.$$

```
predict.mle <- predict(fit.ar.mle,n.ahead=30)

par(mfrow=c(1,1))
y.max = max(predict.mle$pred+predict.mle$se)
y.min = min(predict.mle$pred-predict.mle$se)
plot.ts(predict.mle$pred,ylim=c(y.min,y.max))
lines(predict.mle$pred-predict.mle$se,col="red")
lines(predict.mle$pred+predict.mle$se,col="red")
```




Note that these prediction bounds are quite **wide** – the moral of this story is that **long-term forecasts are a fool's errand**, more often than not. Tread with care.

In both estimation methods, the order of the AR model is selected according to AIC (with the maximal order controlled by `order.max`).

```
fit.ar.mle$aic
```

```
       0        1        2        3        4        5
46.636914  0.000000  1.621425  3.598935  5.537506  7.360361
       6        7        8        9       10       11
 8.973913  7.460411  8.709559 10.705111 12.469661 14.417006
      12
14.506713
```

Sure enough, the lowest value (AIC minus a constant) is for AR(1).[60]

We can also use the more general `arima()` function (but we need to specify the order).

60: How this value is computed depends on the implementation.

```
(fit.arma <- arima(MyTimeSeries, order=c(1,0,0)))
```

```
Coefficients:
         ar1  intercept
      0.6197     0.1430
s.e.  0.0777     0.2517

sigma^2 estimated as 0.9458:  log likelihood = -139.35,  aic = 284.7
```

The results are readily seen to be identical to those of MLE (suggesting a reason to select MLE over YW, perhaps).

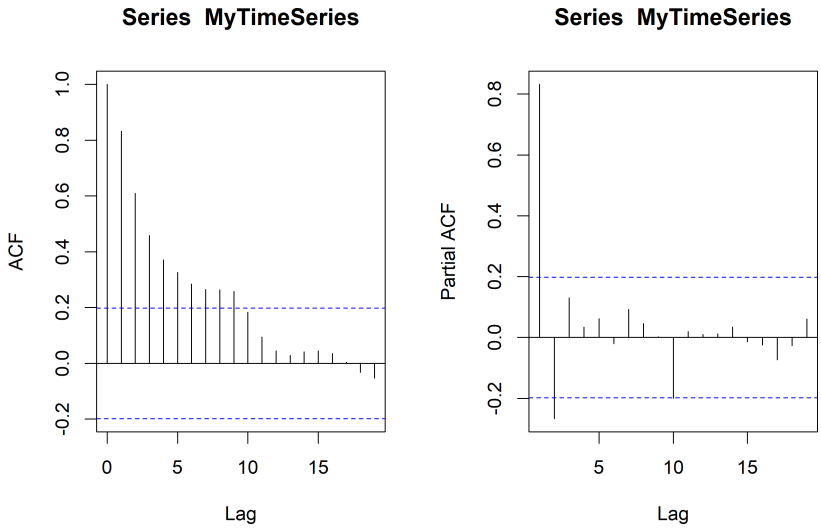
**Example: Lake Huron** We now conduct a similar analysis with the built-in Lake Huron dataset. We start by loading and plotting the data.

```
MyTimeSeries = LakeHuron
plot.ts(MyTimeSeries)
```


There is a downward trend in the first half of the data (from 1875 to 1925), but it seems almost accidental – if a few of these points were lower, the trend would probably appear to be horizontal. We will treat the time series as stationary, with the caveat that it might make sense to analyze the de-trended time series instead.

We can achieve a first pass at the order by looking at the ACF and PACF graphs.

```
par(mfrow=c(1,2))
acf(MyTimeSeries)
pacf(MyTimeSeries)
```

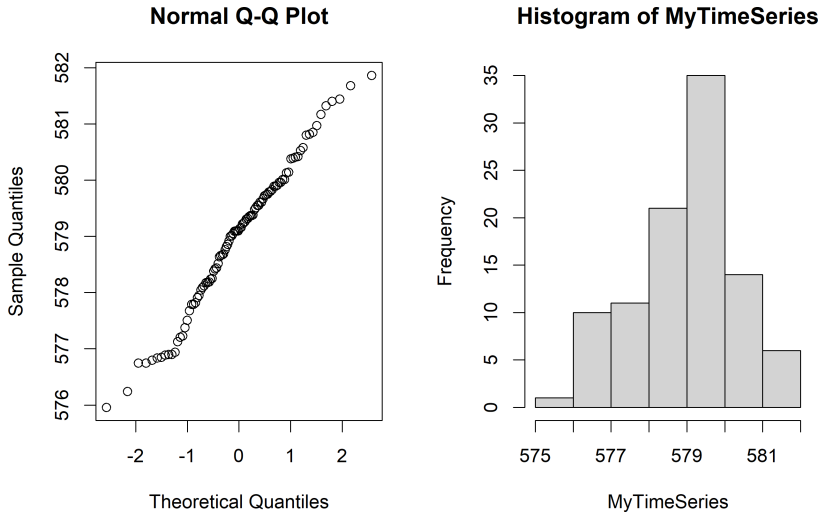



These plots suggest an AR(2) model, or potentially an ARMA(1, 1) model.[61]

61: The ACF and PACF of an ARMA model both converge to 0, but the order $(p, q)$ is not usually obvious... there is a lot of guess-and-check involved in the process.

The time series appears to take on normally distributed values, as can be seen below.

```
par(mfrow=c(1,2))
qqnorm(MyTimeSeries)
hist(MyTimeSeries)
```




We start by assuming that the data is best fit by an auto-regressive model; what would its order and coefficient estimates be?

Using the Yule-Walker approach, we get the following.

```
(fit.ar.yw <- ar(MyTimeSeries,method="yule-walker"))
```

```
Coefficients:
      1        2
 1.0538  -0.2668

Order selected 2  sigma^2 estimated as  0.5075
```

The MLE approach instead yields the following.

```
(fit.ar.mle <- ar(MyTimeSeries,method="mle"))
```

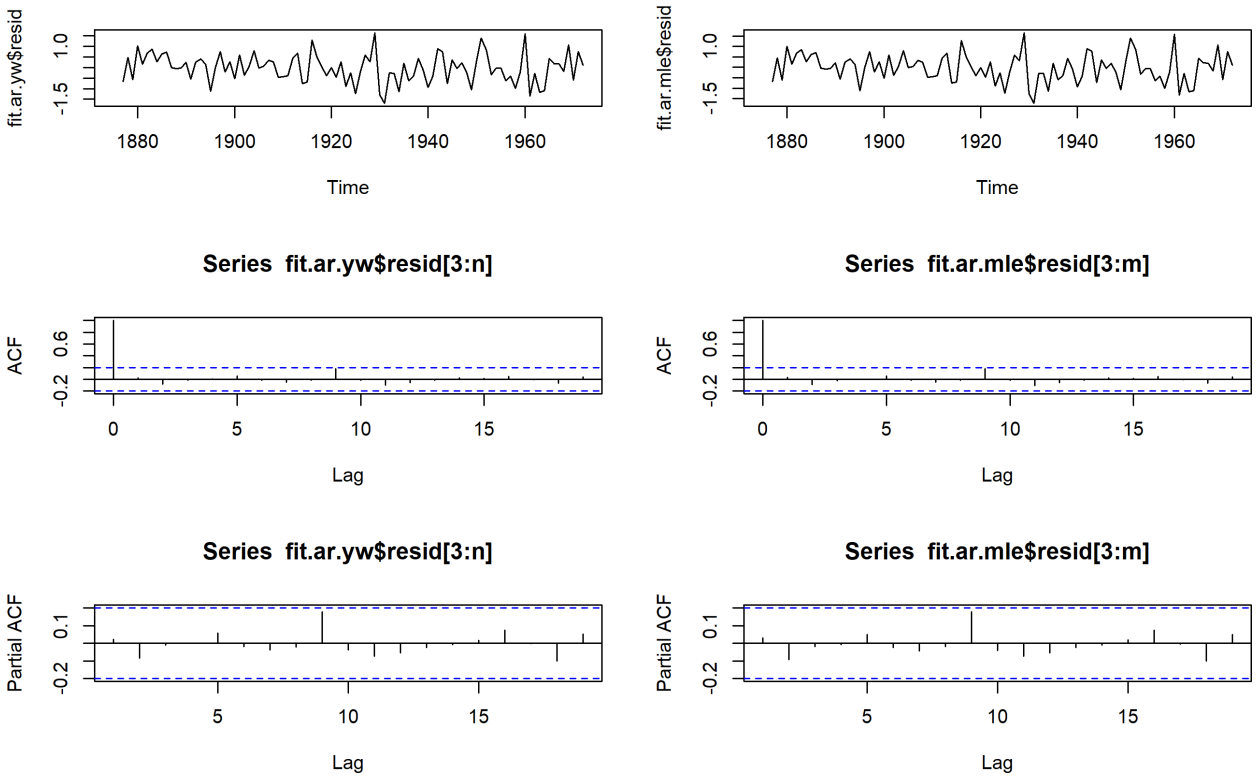
```
Coefficients:
      1        2
 1.0437  -0.2496

Order selected 2  sigma^2 estimated as  0.4788
```

Both of them suggest an AR(2) model, which agrees with our visual determination of the order.[62]

62: The $\phi_i$ should be identical in both approaches, but we have already discussed that the discrepancies are due to the choice of numerical algorithms in the implementations.

Are either of the fits good? We take a look at the residuals.

```
par(mfrow=c(3,2))
plot.ts(fit.ar.yw$resid); plot.ts(fit.ar.mle$resid)
n=length(fit.ar.yw$resid); m=length(fit.ar.mle$resid)
acf(fit.ar.yw$resid[3:n]); acf(fit.ar.mle$resid[3:m])
pacf(fit.ar.yw$resid[3:n]); pacf(fit.ar.mle$resid[3:m])
```



The residual plots look as we would expect if the data arose from either of the two AR(2) processes (i.e., there does not appear to be dependences in the residuals). So which model should be chosen? We could pick the one with smallest AIC, or selecting the model that best "predicts" past values of the data (as done in Section 9.4.4 with the currency exchange rate data). We select the MLE model for the purpose of illustration.

In order to investigate ARMA(1, 1) as a model for the data, we use the `arima()` function. We will re-fit the MLE AR(2) model in this framework, to gain access to the same set of attributes for both models.

```
(fit.arma.1 <- arima(MyTimeSeries, order=c(2,0,0)))
```

```
Coefficients:
         ar1      ar2  intercept
      1.0436  -0.2495   579.0473
s.e.  0.0983   0.1008     0.3319

sigma^2 estimated as 0.4788:  log likelihood = -103.63,  aic = 215.27
```

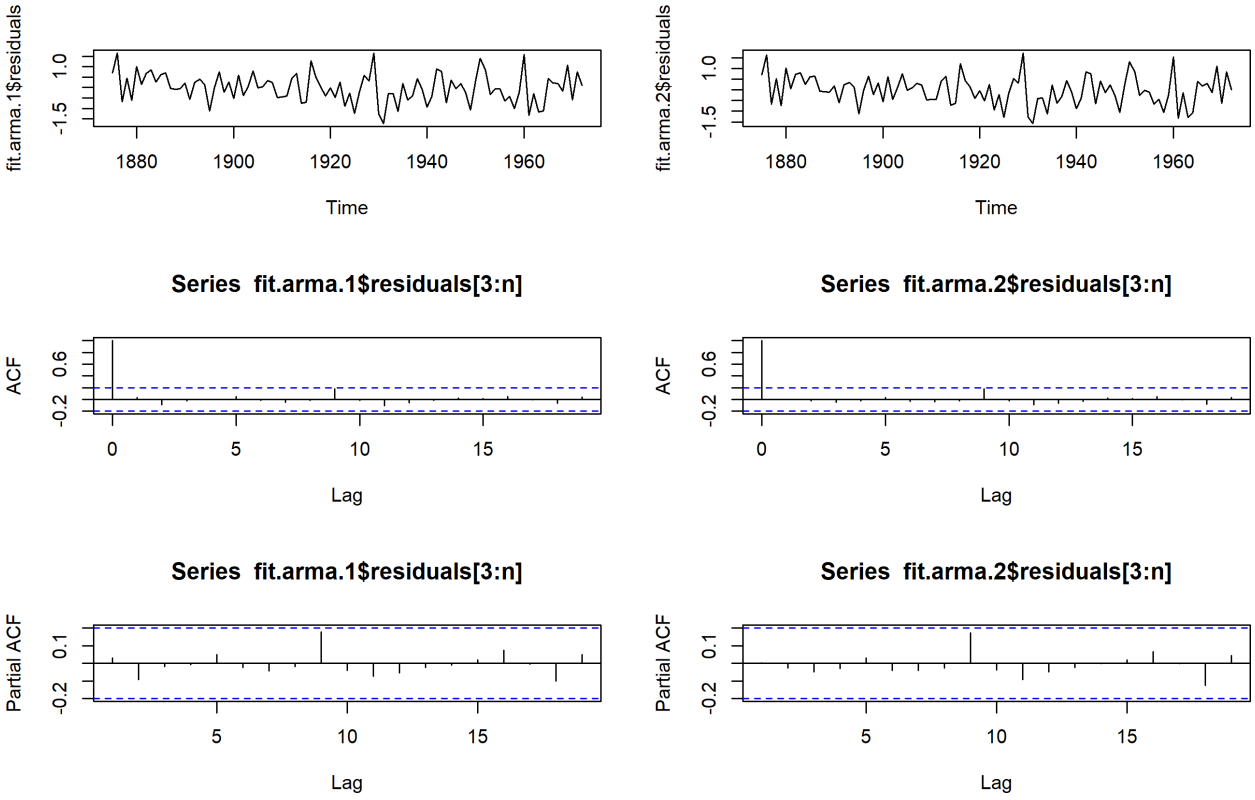
```
(fit.arma.2 <- arima(MyTimeSeries, order=c(1,0,1)))
```

```
Coefficients:
         ar1     ma1  intercept
      0.7449  0.3206   579.0555
s.e.  0.0777  0.1135     0.3501

sigma^2 estimated as 0.4749:  log likelihood = -103.25,  aic = 214.49
```

The intercept term represents the expectation $\mu = \mathrm{E}[X_t]$ of the time series. An important take-away is that there is no obvious relationship between the $\phi_1$ of the AR(2) model and the $\phi_1$ of the ARMA(1, 1) model.

What do the residuals look like?

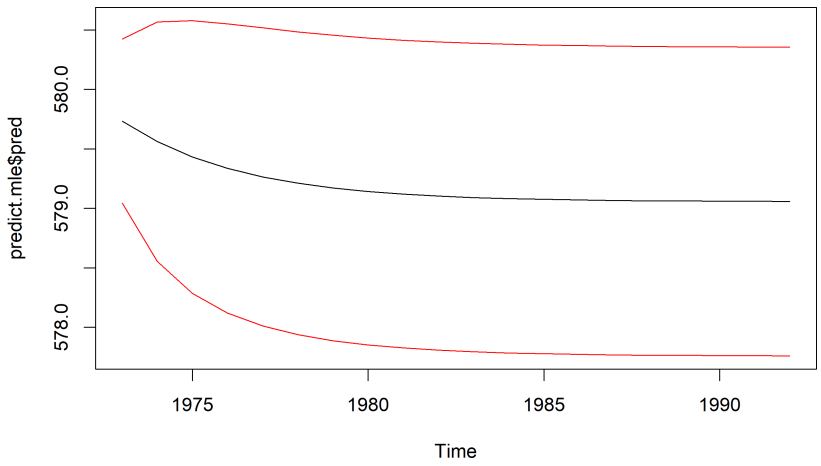
```
par(mfrow=c(3,2))
plot.ts(fit.arma.1$residuals)
plot.ts(fit.arma.2$residuals)

n = length(fit.arma.1$residuals)
m = length(fit.arma.2$residuals)

acf(fit.arma.1$residuals[3:n])
acf(fit.arma.2$residuals[3:n])
pacf(fit.arma.1$residuals[3:n])
pacf(fit.arma.2$residuals[3:n])
```




63: The AIC can be read off of the outputs above, but they can also be extracted directly with `fit.arma.1$aic` and `fit.arma.2$aic`.

Both AR(2) and ARMA(1, 1) are acceptable; we select the latter since it has the smallest AIC.[63] We can predict the next 20 time steps.

```
par(mfrow=c(1,1))
predict.mle <- predict(fit.arma.2,n.ahead=20)
y.max = max(predict.mle$pred+predict.mle$se)
y.min = min(predict.mle$pred-predict.mle$se)
plot.ts(predict.mle$pred,ylim=c(y.min,y.max))
lines(predict.mle$pred-predict.mle$se,col="red")
lines(predict.mle$pred+predict.mle$se,col="red")
```




64: When seasonality is taken into account, we might expect to see some up-and-down motion in the predictions.

Note that the predictions are not as "jagged" as the original time series.[64]

**Example: Temperature (cont.)** We consider the temperature data from pages 499 and 553; using the Yule-Walker procedure, we found that the centered stationary part of the exponentially smoothed time series (`Stationary`) was decently approximated by an AR(4) process.

We now approach the same time series via the MLE procedure. The chart on page 556 indicates that the normality assumption is reasonable. We can thus safely apply the procedure.

```
(fit.ar.mle <- ar(Stationary,method="mle"))
```

```
Coefficients:
      1        2        3        4        5
 0.1427   0.1290  -0.0682   0.2716   0.1187

Order selected 5  sigma^2 estimated as  0.03241
```

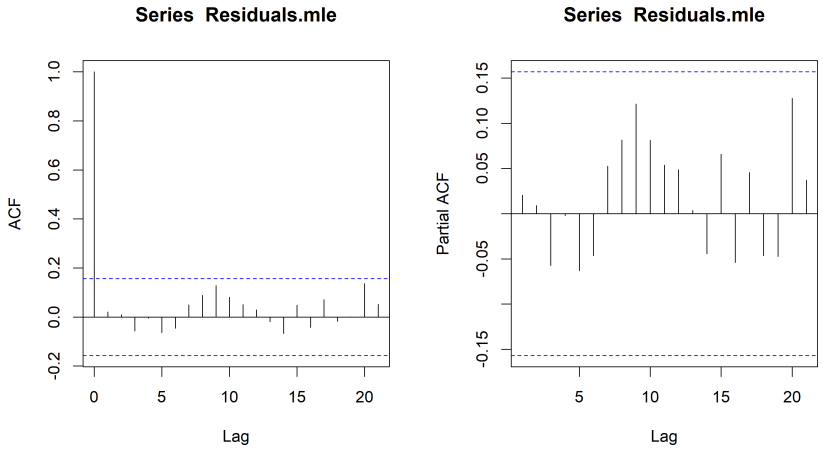
The MLE procedure selected a different order – but there is nothing wrong with that! Note that we could recover this model with the `arima` function (which also displays the standard errors for the AR coefficients).

```
arima(Stationary,order=c(5,0,0),method="ML")
```

```
Coefficients:
         ar1     ar2      ar3     ar4     ar5  intercept
      0.1427  0.1290  -0.0682  0.2716  0.1187     0.0743
s.e.  0.0786  0.0761   0.0766  0.0763  0.0798     0.0342

sigma^2 estimated as 0.03241:  log likelihood = 47.34,  aic = -80.69
```

Is the MLE fit appropriate? Do the residuals appear to be white noise?

```
Residuals.mle = fit.ar.mle$resid
Residuals.mle = na.omit(Residuals.mle)
par(mfrow=c(1,2))
acf(Residuals.mle)
pacf(Residuals.mle)
```



Yes-ish. Close enough is good enough, certainly. We accept the fit. But now we have two competing models. Which one should we choose? We can check the quality of the prediction, for instance.

```
(Squared.Error.yw = mean((Residuals.yw)^2))
```

```
[1] 0.0332902
```

```
(Squared.Error.mle = mean((Residuals.mle)^2))
```

```
[1] 0.03214238
```

The MLE approach yields a lower total error, so we might as well select the MLE model.

But why was AR(5) selected by the MLE procedure? We can compare with the AR(4) MLE model and calculate the respective AIC.

```
(fit.mle.4 <- arima(Stationary,order=c(4,0,0),method="ML"))
```

```
Coefficients:
         ar1     ar2      ar3     ar4  intercept
      0.1782  0.1196  -0.0541  0.2918     0.0715
s.e.  0.0754  0.0764   0.0765  0.0757     0.0303

sigma^2 estimated as 0.03287:  log likelihood = 46.25,  aic = -80.49
```

```
(fit.mle.5 <- arima(Stationary,order=c(5,0,0),method="ML"))
```

```
Coefficients:
         ar1     ar2      ar3     ar4     ar5  intercept
      0.1427  0.1290  -0.0682  0.2716  0.1187     0.0743
s.e.  0.0786  0.0761   0.0766  0.0763  0.0798     0.0342

sigma^2 estimated as 0.03241:  log likelihood = 47.34,  aic = -80.69
```

Note the values of log-likelihod and AIC.[65]

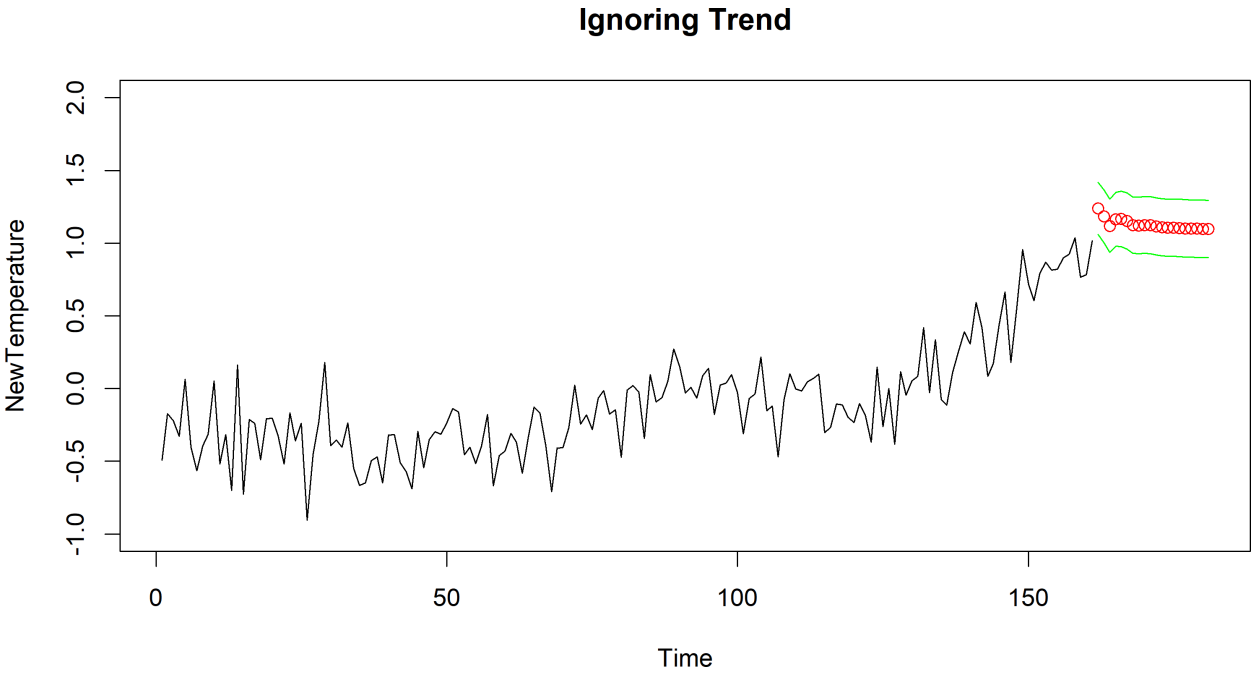
65: It would be important to make sure that you can recover the AIC values from the log-likelihood values, with the formula.

We can use the MLE model to predict the next 20 observations

```
k = 20
prediction = predict(fit.ar.mle,n.ahead=k)$pred
error = predict(fit.ar.mle,n.ahead=k)$se
```

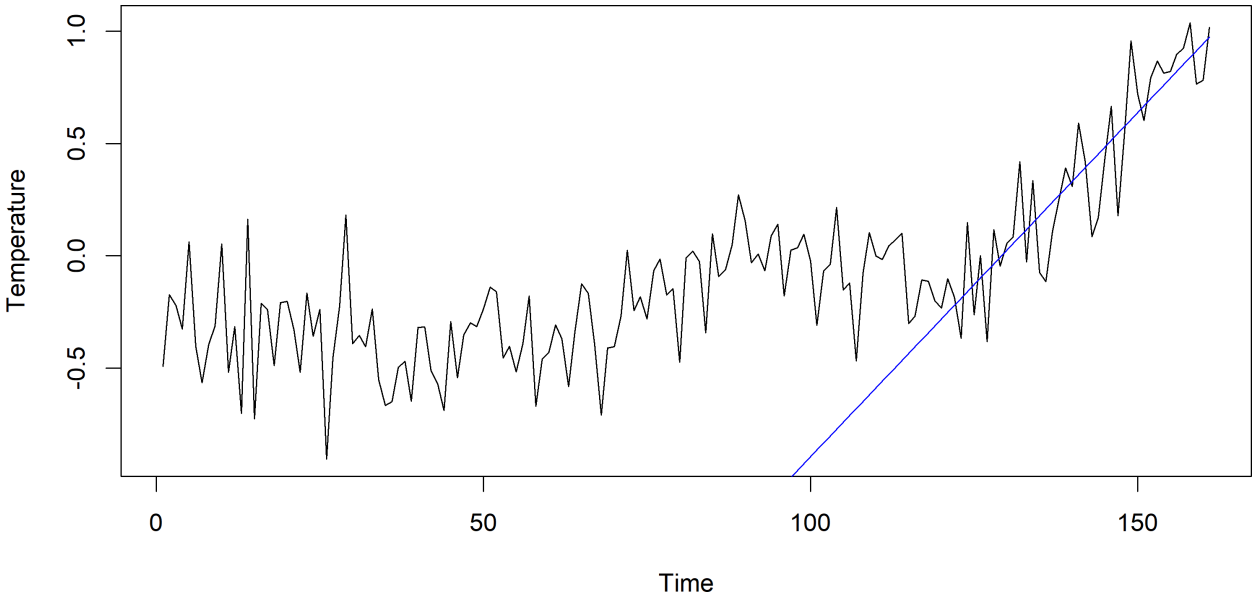
In order to transform these `Stationary` predictions into values in the original time series, we have to add them to the Temperature data. In the next code chunk, we will ignore the trend in the original data.

```
n=length(Temperature)
k = 20
prediction.1 = prediction+Temperature[n]
prediction.1.upper = prediction.1 + error
prediction.1.lower = prediction.1 - error

dummy.ts = c(rep(NA,k))
NewTemperature = c(Temperature,dummy.ts)
dummy.pred=c(rep(NA,n))
PredictedStationary = c(dummy.pred,prediction.1)
PredictionUpperLimit = c(dummy.pred,prediction.1.upper)
PredictionLowerLimit = c(dummy.pred,prediction.1.lower)

par(mfrow=c(1,1))
plot.ts(NewTemperature,ylim=c(-1,2),main="Ignoring Trend")
points(PredictedStationary,col="red",type="p")
points(PredictionUpperLimit,col="green",type="l")
points(PredictionLowerLimit,col="green",type="l")
```




Something about this is definitely not right. The problem is that we ignored the trend in the original data, but starting in year 120 (or thereabouts), the time series follows a linear trend (more or less). We fit a linear trend to this part of the data.

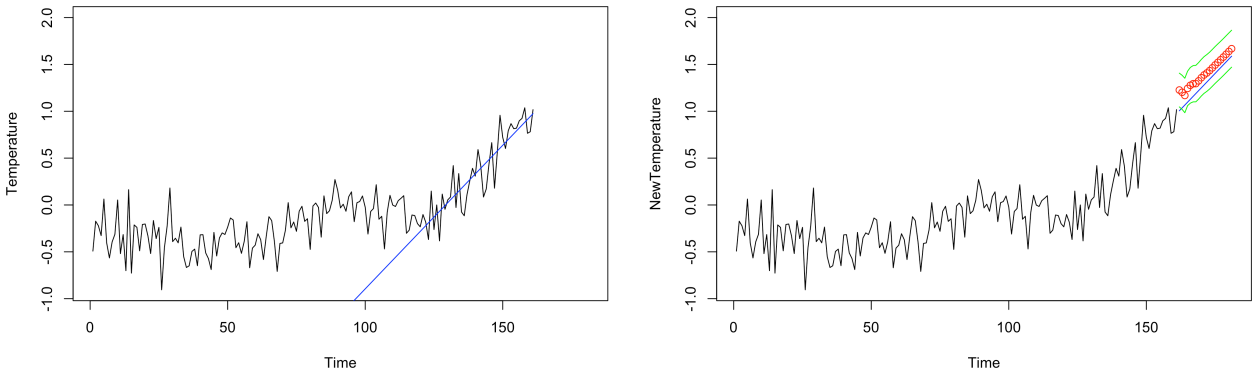
```
n = length(Temperature)
Time = seq(1,n,by=1)
lin.reg = lm(Temperature[120:n]~Time[120:n])
Lin.Trend = lin.reg[[1]][1] + lin.reg[[1]][2]*Time

plot.ts(Temperature)
points(Lin.Trend,col="blue",type="l")
```




66: The code for the prediction limits is not included – how could they be produced?

The next step is to extend the linear trend and the predictions.[66]

```
k = 20
dummy.ts = c(rep(NA,k))
NewTemperature = c(Temperature,dummy.ts)
dummy.trend = c(rep(NA,n)); Time = seq(1,n+k,by=1)
Extended.Trend = lin.reg[[1]][1] + lin.reg[[1]][2]*Time
Trend = c(dummy.trend,Extended.Trend[(n+1):(n+k)])
y.max = 2; y.min = min(Temperature)

par(mfrow=c(1,2))
plot.ts(Temperature,xlim=c(1,n+k),ylim=c(y.min,y.max))
points(Lin.Trend,col="blue",type="l")
plot.ts(NewTemperature,xlim=c(1,n+k),ylim=c(y.min,y.max))
points(Trend,col="blue",type="l")
Prediction.stationary = c(dummy.trend,prediction)
PredictedStationary = Trend+Prediction.stationary
points(PredictedStationary,col="red",type="p")
```



## 9.8 Nonlinear Time Series

The log-returns of financial data typically have the following properties:

- they are **uncorrelated**;
- their **squares** are correlated;
- they are **not normally distributed**.

Such features cannot be modelled by ARMA models.

### 9.8.1 ARCH model

A time series $\{X_t \mid t = 1, \dots, n\}$ is **autoregressive conditionally heteroscedastic of order** $p$, denoted ARCH($p$) if

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2,$$

where $Z_t$ are i.i.d. with mean 0 and variance 1, $\alpha_0 > 0$, $\alpha_i \geq 0$ for all $i$. We note explicitly that the values of $\sigma_t$ depend on the past values of the sequence $\{X_t\}$: $X_{t-1}, X_{t-2}, \dots$.

If $p = 1$, then

$$\begin{aligned} X_t^2 &= \sigma_t^2 Z_t^2 = \left(\alpha_0 + \alpha_1 X_{t-1}^2\right) Z_t^2 = \left(\alpha_0 + \alpha_1 \sigma_{t-1}^2 Z_{t-1}^2\right) Z_t^2 \\ &= \alpha_0 Z_t^2 + \alpha_1 Z_t^2 Z_{t-1}^2 \sigma_{t-1}^2. \end{aligned}$$

We can continue on this way by replacing $\sigma_{t-1}^2$ by its formulation, and so on. Consequently, we see that $X_t^2$ depends only on $Z_t, Z_{t-1}, Z_{t-2}, \dots$.[67] This is valid for all ARCH models, not only ARCH(1).

67: As a further consequence, $Z_t$ and $X_{t-1}$ are independent.

For a general $p$, we have

$$\mathrm{E}[X_t \mid X_{t-1}, \dots, X_{t-p}] = \mathrm{E}[\sigma_t Z_t \mid X_{t-1}, \dots, X_{t-p}] = \sigma_t \mathrm{E}[Z_t \mid X_{t-1}, \dots, X_{t-p}] = 0$$

and

$$\begin{aligned} \mathrm{Var}(X_t \mid X_{t-1}, \dots, X_{t-p}) &= \mathrm{E}[X_t^2 \mid X_{t-1}, \dots, X_{t-p}] = \mathrm{E}[\sigma_t^2 Z_t^2 \mid X_{t-1}, \dots, X_{t-p}] \\ &= \sigma_t^2 \mathrm{E}[Z_t^2 \mid X_{t-1}, \dots, X_{t-p}] = \sigma_t^2 \mathrm{E}[Z_t^2] = \sigma_t^2. \end{aligned}$$

The "conditionally heteroscedastic" in ARCH refers to this last equation.

The series $\{\sigma_t^2 \mid t \geq 1\}$ is the **volatility** of the time series; ARCH($p$) is an example of a **stochastic volatility process**.

**Proposition:** the ARCH(1) process is stationary if and only if $\alpha_1 < 1$. A stationary solution is given by

$$X_t^2 = \alpha_0 \sum_{i=0}^{\infty} \alpha_1^i \prod_{j=0}^{i} Z_{t-j}^2.$$

In an ARCH(1) model, we have

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2.$$

We can estimate the model parameters using the maximum likelihood principle. Consider the joint density

$$f_{(X_0,\dots,X_n)}(x_0,\dots,x_n) = f_{X_0}(x_0)\prod_{i=1}^{n} f_{X_t|X_{t-1}}(x_t \mid x_{t-1})$$

where

$$f_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{1}{\sigma_t} g(x_t/\sigma_t)\ ,$$

with $\sigma_t^2 = \alpha_0 + \alpha_1 x_{t-1}^2$ and $g$ is the density of $Z_0$ (which is typically a normal or Student $T$ distribution).

Let

$$\begin{aligned} F_{X_t|X_{t-1}}(x_t \mid x_{t-1}) &= P(X_t \le x_t \mid X_{t-1} = x_{t-1}) \\ &= P(\sigma_t Z_t \le x_t \mid X_{t-1} = x_{t-1}) \\ &= P(Z_t \le x_t/\sigma_t \mid X_{t-1} = x_{t-1}) \\ &= P(Z_t \le x_t/\sigma_t) = G(x_t/\sigma_t), \end{aligned}$$

where $G$ is the cumulative distribution function of $Z$: $G(z) = P(Z \le z)$.

We can show that

$$f_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{d}{dx_t} F_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{1}{\sigma_t} g(x_t/\sigma_t)\ .$$

Indeed, we start with the conditional distribution:

$$\begin{aligned} F_{X_t|X_{t-1}}(x_t \mid x_{t-1}) &= P(X_t \le x_t \mid X_{t-1} = x_{t-1}) = P(\sigma_t Z_t \le x_t \mid X_{t-1} = x_{t-1}) \\ &= P(Z_t \le x_t/\sigma_t \mid X_{t-1} = x_{t-1}) \\ &= P(Z_t \le x_t/\sqrt{\alpha_0 + \alpha_1 x_{t-1}^2} \mid X_{t-1} = x_{t-1}) \\ &= F_Z(Z_t \le x_t/\sqrt{\alpha_0 + \alpha_1 x_{t-1}^2}) = F_Z(x_t/\sigma_t) \end{aligned}$$

and the density is

$$\frac{d}{dx_t} F_Z(x_t/\sigma_t) = \frac{1}{\sigma_t} f_Z(x/\sigma_t) = \frac{1}{\sigma_t} g(x/\sigma_t),$$

keeping in mind that $\sigma_t^2 = \alpha_0 + \alpha_1 x_{t-1}^2$.

Thus, the likelihood function has the form

$$L(\alpha_0, \alpha_1) = \prod_{t=1}^{n} \frac{1}{\sigma_t} g(X_t/\sigma_t)$$

and

$$(\alpha_0, \alpha_1) = \arg \max_{\alpha_0>0, 0<\alpha_1<1} L(\alpha_0, \alpha_1),$$

where the optimization problem is solved numerically (see Section 4).

### 9.8.2 GARCH Model

A time series $\{X_t \mid t = 1,\dots,n\}$ is a **generalized autoregressive conditionally heteroscedastic** model of **order** $(p, q)$, denoted GARCH$(p, q)$ if

the variance $\sigma_t^2$ is modeled using past squared observations $X_{t-i}^2$ **and** past variances $\sigma_{t-j}^2$:

$$X_t = \sigma_t Z_t, \quad \text{Var}(X_t \mid X_{t-1}, \ldots, X_{t-p}) = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

where $Z_t$ are i.i.d. with mean 0 and variance 1, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for all $i, j$.

On the topic of identifying an ARCH/GARCH model in practice, [5] has this to say:

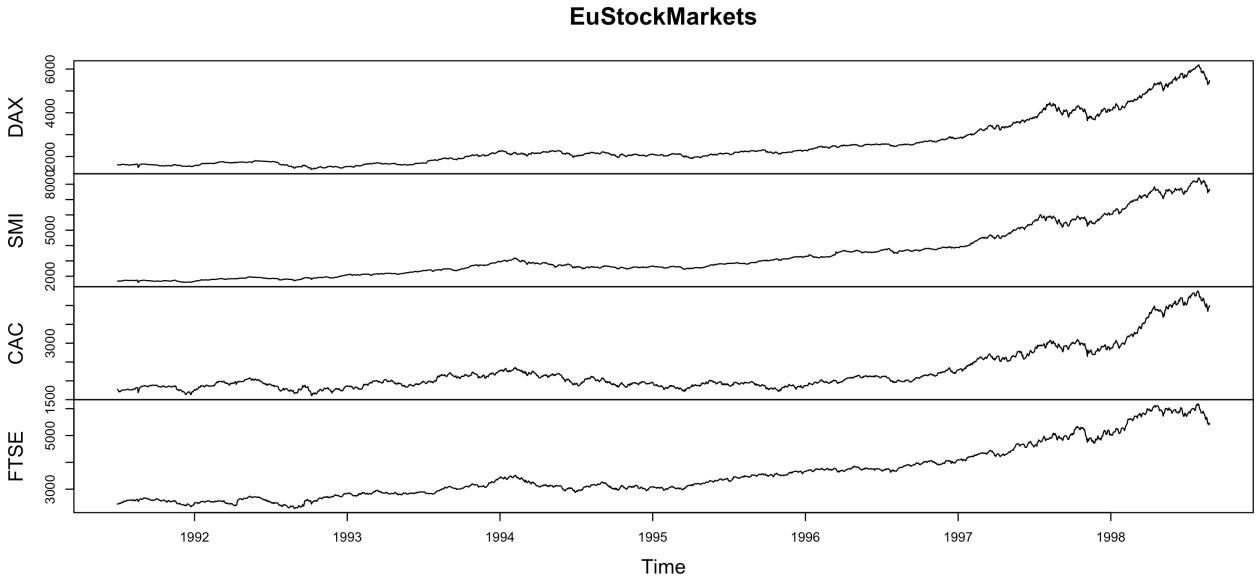
> The best identification tool may be a time series plot of the series. It's usually easy to spot periods of increased variation sprinkled through the series. It can be fruitful to look at the ACF and PACF of both $X_t$ and $X_t^2$. For instance, if $X_t$ appears to be white noise and $X_t^2$ appears to be AR(1), then an ARCH(1) model for the variance is suggested. If the PACF of $X_t^2$ suggests AR($p$), then ARCH($p$) may work. GARCH models may be suggested by an ARMA-type look to the ACF and PACF of $X_t^2$. [...] You might have to experiment with various ARCH and GARCH structures after spotting the need in the time series plot of the series.

### 9.8.3 Example: Stock Returns

We consider the daily closing price of Germany's DAX stock index, from 1991 to 1998;[68] the dataset is pre-built in R.

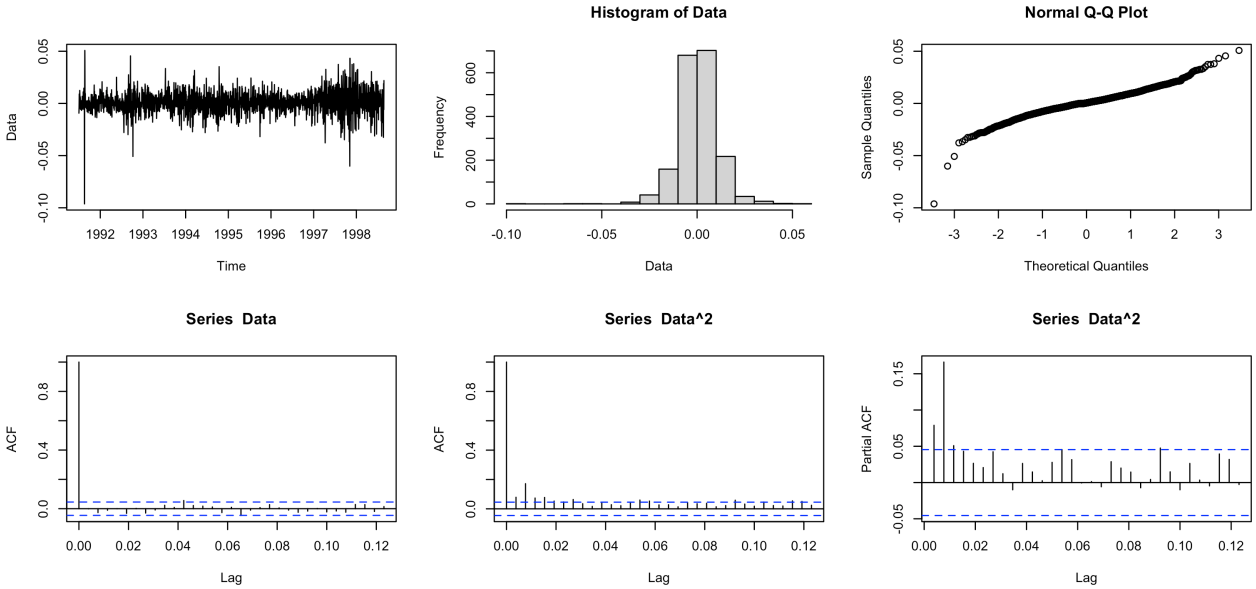
68: The observations are recorded on business days, and are also available for 3 other indices: SMI, CAC, UK FTSE.

```
library(tseries)
plot(EuStockMarkets)
```




We differentiate the log-returns of the DAX index to obtain a time series which appears to be stationary, but which is not normally distributed. We display the ACF of the data, as well as the ACF and PACF of the square of the data.

```
Data <- diff(log(EuStockMarkets))[,"DAX"]
par(mfrow=c(2,3))
plot.ts(Data)
hist(Data); qqnorm(Data);
acf(Data); acf(Data^2); pacf(Data^2)
```




It is reasonable to fit to model the data as an ARCH(1) process.

```
fit.ARCH1 <- garch(Data,order=c(0,1))
```

```
***** ESTIMATION WITH ANALYTICAL GRADIENT *****

     I     INITIAL X(I)        D(I)
     1     1.008019e-04     1.000e+00
     2     5.000000e-02     1.000e+00

    IT   NF      F         RELDF    PRELDF    RELDX   STPPAR   D*STEP   NPRELDF
     0    1 -7.582e+03
     1    8 -7.582e+03  7.08e-06  1.27e-05  1.0e-05  9.4e+10  1.0e-06  5.95e+05
     2    9 -7.582e+03  9.60e-08  9.77e-08  1.0e-05  2.0e+00  1.0e-06  7.31e-01
     3   18 -7.584e+03  2.66e-04  4.85e-04  2.6e-01  2.0e+00  3.5e-02  7.31e-01
     4   19 -7.584e+03  1.47e-05  1.13e-05  4.4e-02  0.0e+00  7.9e-03  1.13e-05
     5   20 -7.584e+03  1.81e-06  1.67e-06  2.0e-02  0.0e+00  3.8e-03  1.67e-06
     6   21 -7.584e+03  1.51e-08  1.46e-08  1.9e-03  0.0e+00  3.6e-04  1.46e-08
     7   22 -7.584e+03  1.47e-11  1.47e-11  6.3e-05  0.0e+00  1.2e-05  1.47e-11

 ***** RELATIVE FUNCTION CONVERGENCE *****

 FUNCTION    -7.584131e+03   RELDX        6.254e-05
 FUNC. EVALS      22         GRAD. EVALS       8
 PRELDF       1.471e-11      NPRELDF      1.471e-11

     I      FINAL X(I)        D(I)          G(I)
     1    9.611161e-05     1.000e+00    -9.229e-01
     2    9.703263e-02     1.000e+00    -7.850e-05
```

The resulting GARCH object has the following attributes.

```
attributes(fit.ARCH1)
```

```
$names
 [1] "order"         "coef"          "n.likeli"      "n.used"
 [5] "residuals"     "fitted.values" "series"        "frequency"
 [9] "call"          "vcov"

$class
[1] "garch"
```

The estimated coefficient values of $\alpha_0$ and $\alpha_1$ are obtained as below.

```
(Coefficients <- fit.ARCH1$coef)
alpha0=Coefficients[1]; alpha1=Coefficients[2]
```

```
          a0           a1
9.611161e-05 9.703263e-02
```

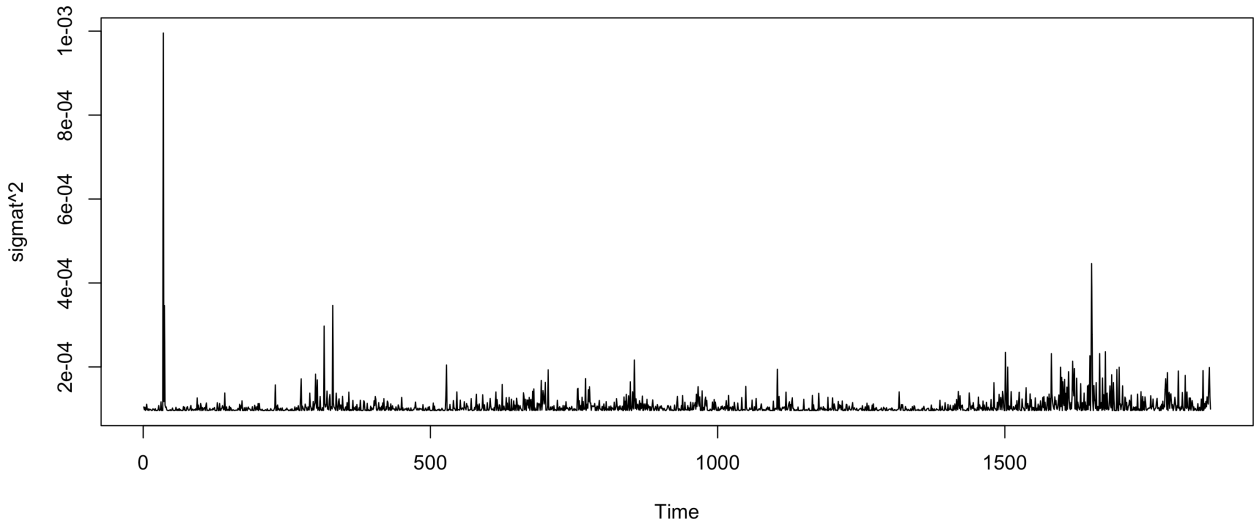
We can view the fitted values as past prediction of $\sigma_t$, the first 10 of which are as below.

```
past.prediction = fit.ARCH1$fitted.values
past.prediction[1:10]
```

```
[1]          NA 0.010225064 0.009899957 0.010196954 0.009819289 0.009911300
[7] 0.010540232 0.009966482 0.009844322 0.010001275
```

We can plot the time series $\{\sigma_t^2\}$:

```
n = length(Data)
sigmat = past.prediction[2:n]
par(mfrow=c(1,1))
plot.ts(sigmat^2)
```



The prediction of the next observation in the sequence can be obtained directly.

```
n1 = length(sigmat)
sqrt(alpha0+alpha1*sigmat[n1]^2)
```

```
        a0
0.01028445
```

It is easy to extract the residuals, the first 10 of which are:

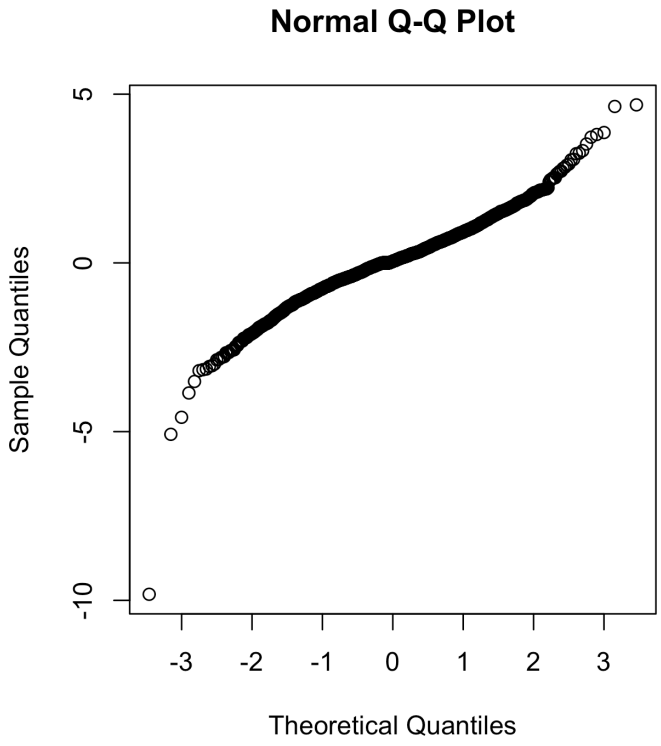
```
residuals <- fit.ARCH1$residuals
residuals[1:10]
```

```
[1]         NA -0.4324838  0.9094782 -0.1743871
[5] -0.4762781  1.2538257  0.5464646
[8] -0.2879321  0.6451482  0.1183917
```

69: Remember that normality of the residuals ($Z_t$) is not the same as normality of the data ($X_t$).

We can see that the residuals are normally distributed, roughly.[69]

```
residuals = residuals[2:n]
qqnorm(residuals)
```




## 9.9 Miscellenous Topics

We will finish this chapter by briefly discussing three additional topics: **seasonality**, **asymptotic normality**, and **innovations**.

### 9.9.1 Seasonality

In the study of time series data, **seasonality** – a repeating pattern that occurs at regular intervals – is an important concept. For instance, we might expect a time series of the average monthly temperature in a specific location to show regularity from one year to the next. Or, assuming that an employee's salary is deposited twice monthly directly into their bank account and that expenses come out on a monthly basis form the same account, we would expect the time series of end-of-day balances in the account to follow a regular monthly pattern.

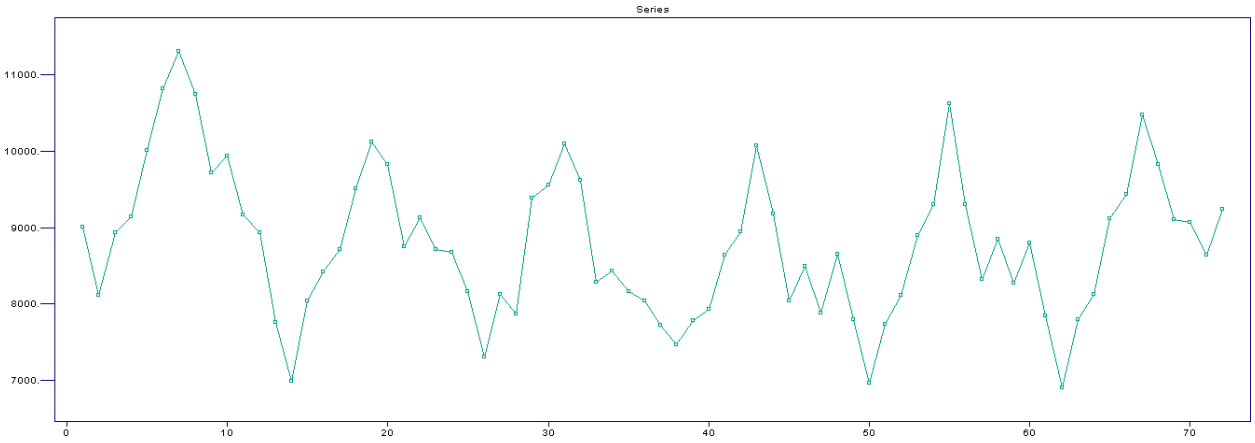
**Differencing** is a simple way to correct for a seasonal component: if we have identified such a component with a period of $T$ time steps,[70] then we can remove it on $X_t$ by subtracting from it the value $X_{t-T}$, yielding a time series

$$Y_t = \nabla_T X_t = X_t - X_{t-T}, t > T.$$

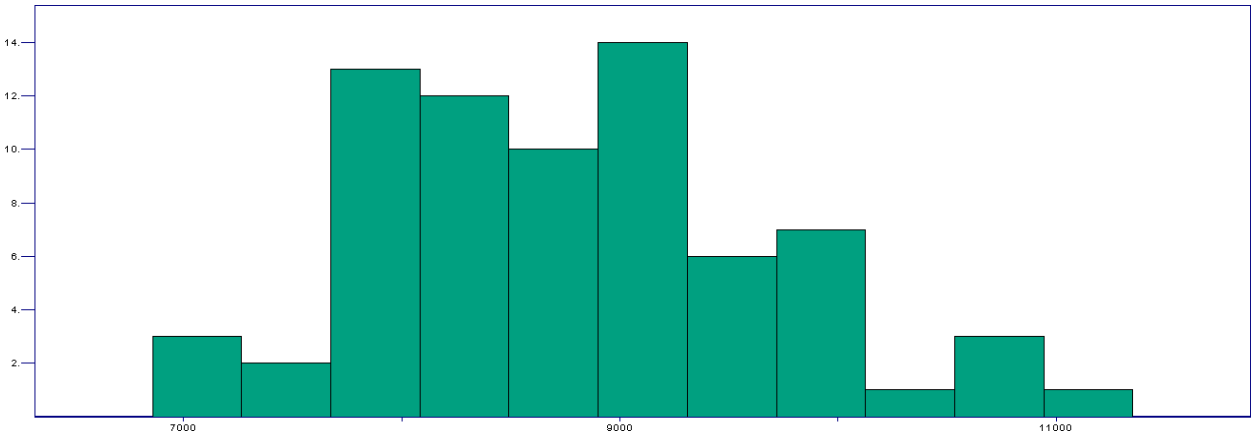
70: By searching for regularities in the ACVF, through a Fourier analysis of the data, or using domain expertise.

We have seen some examples of seasonal decomposition when we were using the `decompose()` function to de-trend the data and obtain the stationary (random) component for analysis (see page 504, for instance).

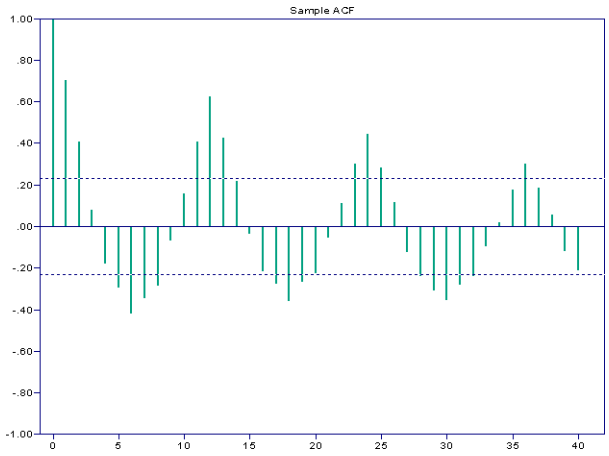
**Example: Accidental Deaths** The monthly accidental deaths figures (`USAccDeaths`) in the US from January 1973 ($t = 1$) to December 1978 ($t = 72$) are plotted below.



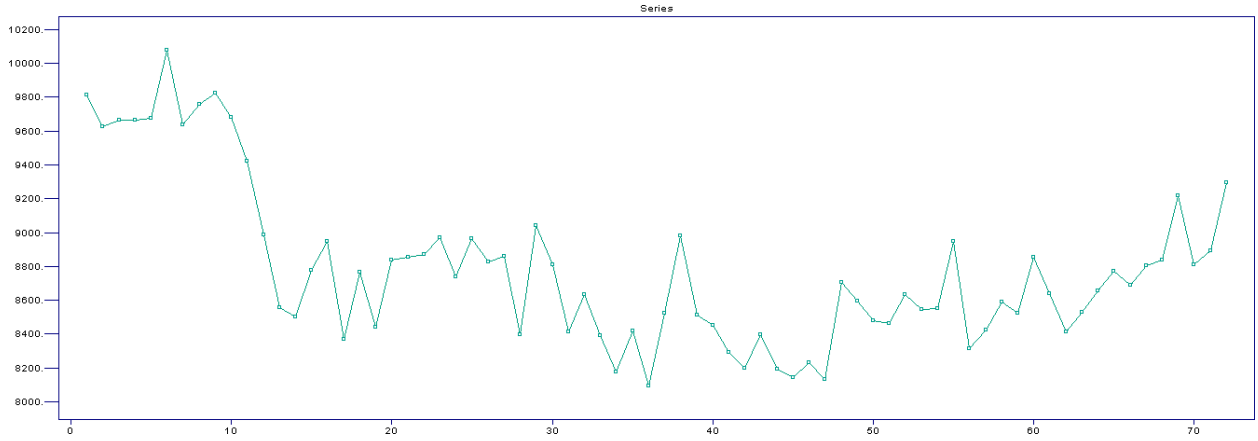

A histogram of the data is also provided.



The sample autocorrelation function also shows a seasonal trend with period $T = 12$.

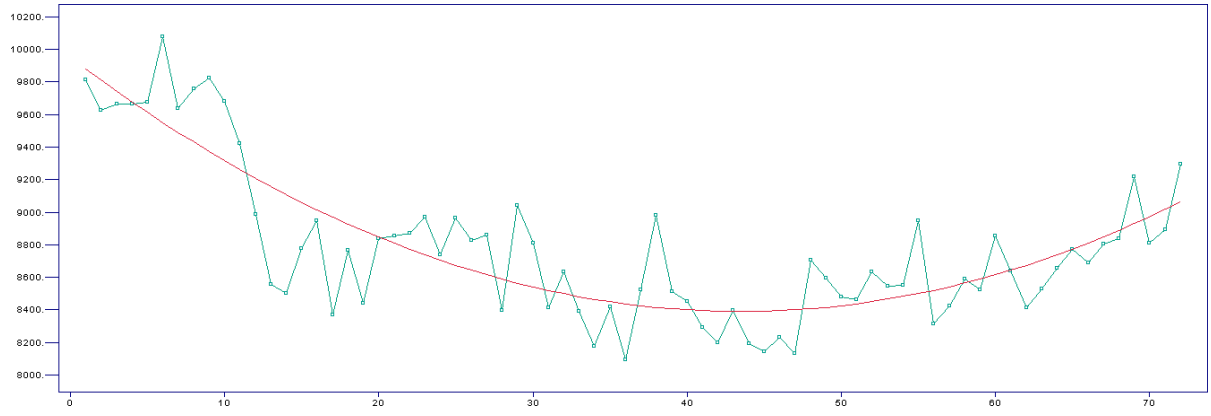



The deseasonalized deaths data is shown below.




This graphs suggests the presence of an additional quadratic component:

$$x_t = \underbrace{m_t}_{\text{local trend}} + \underbrace{s_t}_{\text{seasonal trend}} + \underbrace{Z_t}_{\text{noise}}, \quad m_t = a_0 + a_1 t + a_2 t^2.$$



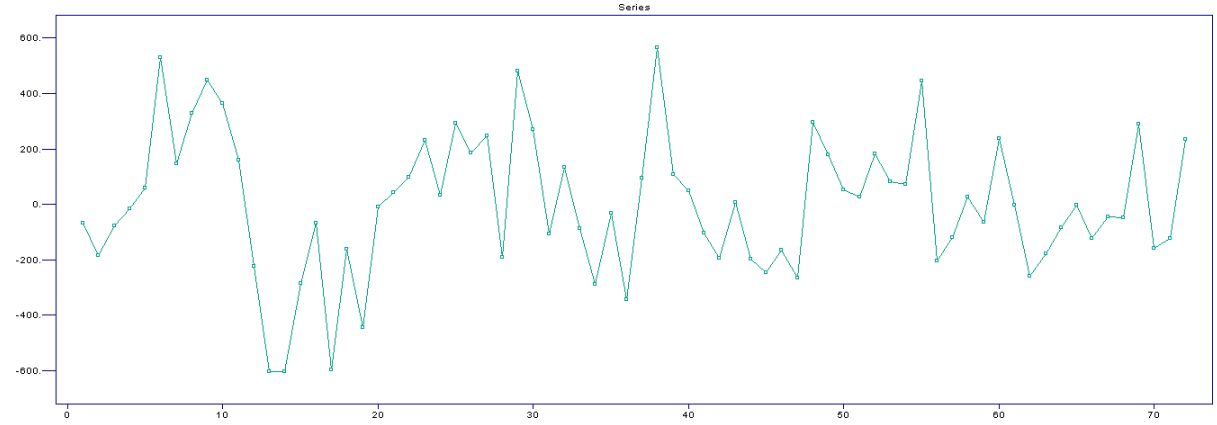

We estimate the local trend as

$$\hat{m}_t = 9951.822 - 71.817t + 0.826t^2, \quad 1 \le t \le 72.$$

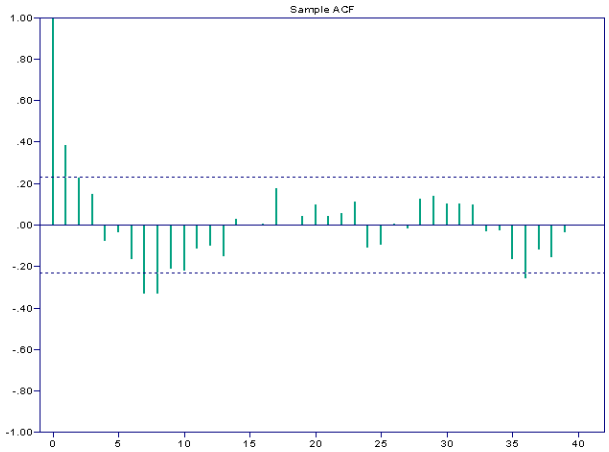
The estimated residuals (the stationary signal)

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad 1 \leq t \leq 72$$

is shown below.




The residuals do appear to be dependent, as there are long stretches of residuals with the same sign. Furthermore, 10% of the autocorrelations are outside the bounds $\pm 1.96/\sqrt{72}$, which is also an indication that we should reject the i.i.d. hypothesis.



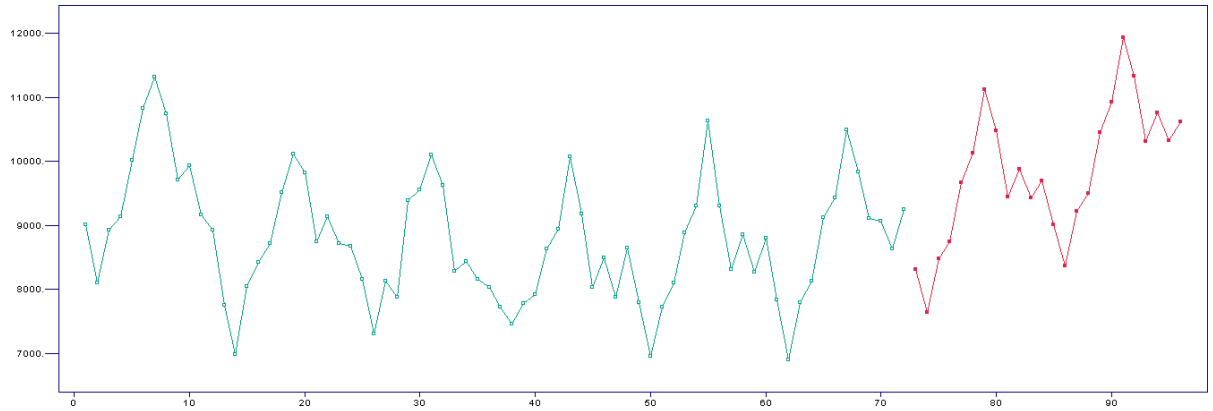

The results of the randomness tests for residuals are:

```
Ljung - Box statistic = 55.384 Chi-Square ( 20 ), p-value = .00004
Order of Min AICC YW Model for Residuals = 1
```

The sample value of the Ljung-Box statistic $Q_{\text{LB}}$ with lag $h = 20$ is 51.84. Since the corresponding $p-$value is $0.00004 < 0.05$ we reject the i.i.d. hypothesis at a level of 0.05. The minimum-AICC Yule-Walker auto-regressive model for the data is of order 1 ($\neq 0$), which supports the evidence provided by the sample ACF and the Ljung-Box statistic against the i.i.d. hypothesis.

We forecast data for the years 1979 and 1980 (using an ARMA model) and display the prediction in red below.[71]

71: The order is not provided.




Note the "jaggedness" of the predictions.

### 9.9.2 Asymptotic Normality

Asymptotic normality is an important concept in time series analysis for several reasons, some of which are outlined below.

- **Statistical Inference:** Asymptotic normality allows for the application of standard statistical tests (like $t-$tests and $z-$tests) for hypothesis testing and confidence interval construction. This simplifies the analysis by using familiar and well-understood techniques.
- **Large Sample Approximation:** Time series data often involve a large number of observations. The Central Limit Theorem suggests that the sampling distribution of many statistics will be approximately normal in large samples, making the results generalizable.
- **Parameter Estimation:** In many time series models, parameter estimates are often obtained through methods like Maximum Likelihood Estimation (MLE) or Ordinary Least Squares (OLS). Asymptotic normality of these estimators provides a basis for conducting inference about the parameters.
- **Model Validation:** When fitting models to time series data, it is important to know under what conditions the model will produce reliable forecasts. Knowing that a model's estimators are asymptotically normal helps in understanding its long-term behaviour.
- **Comparison of Models:** Asymptotic normality provides a common ground for comparing different models. This is especially useful in model selection criteria, like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), where the likelihood function plays a crucial role.
- **Robustness:** Models that possess asymptotically normal properties are often more robust to minor deviations from assumptions, like non-normality of errors in small samples.
- **Simplicity and Computation:** When the statistics of interest are asymptotically normal, it simplifies both the theoretical and computational aspects of the analysis. This allows for easier interpretation and faster computation, which is crucial in real-world applications where time and computational resources may be limited.

The **score function** of a probability density $f(x;\theta)$ is:

$$s(x;\theta) = \frac{\partial \log f(x;\theta)}{\partial \theta} = \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta}.$$

The **Fisher information** of the time series $\{X_t \mid t = 1, \ldots, n\}$ is:

$$I_n(\theta) = \text{Var}\left( \sum_{i=1}^{n} s(X_i;\theta) \right).$$

If the random variables are i.i.d., then the Fisher information collapses to

$$I_n(\theta) = n\text{Var}(s(X_1;\theta)) = nI_1(\theta) = nI(\theta).$$

**Lemma:** the score function satisfies $\text{E}[s(X;\theta)] = 0$.

**Proof:** we used the definition of the expectation to obtain:

$$\begin{aligned}
\text{E}[s(X;\theta)] &= \int s(X;\theta) f(x;\theta) dx = \int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta) dx \\
&= \int \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta} f(x;\theta) dx = \int \frac{\partial f(x;\theta)}{\partial \theta} dx \\
&= \frac{\partial}{\partial \theta} \int f(x;\theta) dx = \frac{\partial}{\partial \theta}[1] = 0. \qquad \blacksquare
\end{aligned}$$

In the proof, we assumed that we could interchange integration and differentiation.[72] Using the above lemma, we then find:

72: This holds for most reasonable density functions $f(x;\theta)$.

$$I(\theta) = \text{Var}(s(X;\theta)) = \text{E}[s^2(X;\theta)].$$

**Lemma:** we have

$$I(\theta) = \text{E}[s^2(X;\theta)] = -\text{E}\left[ \frac{\partial s(X;\theta)}{\partial \theta} \right] = -\text{E}\left[ \frac{\partial^2 \log f(X;\theta)}{\partial \theta^2} \right].$$

**Proof**: first, we note that:

$$\text{E}\left[ s^2(X;\theta) \right] = \int \left( \frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2 f(x;\theta) dx = \int \frac{1}{f^2(x;\theta)} \left( \frac{\partial f(x;\theta)}{\partial \theta} \right)^2 f(x;\theta) dx = \int \frac{1}{f(x;\theta)} \left( \frac{\partial f(x;\theta)}{\partial \theta} \right)^2 dx.$$

Next, we see that:

$$\begin{aligned}
-\text{E}\left[ \frac{\partial s(X;\theta)}{\partial \theta} \right] &= -\int \frac{\partial s(X;\theta)}{\partial \theta} f(x;\theta) dx = -\int \frac{1}{f^2(x;\theta)} \left( \frac{\partial^2 f(x;\theta)}{\partial \theta^2} f(x;\theta) - \left( \frac{\partial f(x;\theta)}{\partial \theta} \right)^2 \right) f(x;\theta) dx \\
&= -\int \left( \frac{\partial^2 f(x;\theta)}{\partial \theta^2} \right) dx + \int \frac{1}{f(x;\theta)} \left( \frac{\partial f(x;\theta)}{\partial \theta} \right)^2 dx = -\frac{\partial^2}{\partial \theta^2} \underbrace{\int f(x;\theta) dx}_{=1} + \text{E}[s^2(X;\theta)] = \text{E}[s^2(X;\theta)].
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
\text{E}\left[ \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} \right] &= \int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) dx = \int \frac{\partial}{\partial \theta} \left( \frac{\partial \log f(x;\theta)}{\partial \theta} \right) f(x;\theta) dx \\
&= \int \frac{\partial}{\partial \theta} \left( \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta} \right) f(x;\theta) dx = \int \frac{\partial s(x;\theta)}{\partial \theta} f(x;\theta) dx = \text{E}\left[ \frac{\partial s(X;\theta)}{\partial \theta} \right]. \qquad \blacksquare
\end{aligned}$$

**Example** Consider $X \sim \text{Exp}(\beta)$. The density function of $X$ is

$$f(x;\beta) = \frac{1}{\beta} e^{-x/\beta},$$

with $\text{E}[X] = \beta$, so that $\log f(x;\beta) = -\log(\beta) - x\beta$. The score function of $X$ is thus

$$s(x;\beta) = -\frac{1}{\beta} + \frac{1}{\beta^2} x$$

and its derivative (w.r.t. $\beta$) is

$$-\frac{\partial s(x;\beta)}{\partial \beta} = -\frac{1}{\beta^2} + \frac{2}{\beta^3} x.$$

Hence,

$$I(\beta) = \text{E}\left[-\frac{\partial s(x;\beta)}{\partial \beta}\right] = -\frac{1}{\beta^2} + \frac{2}{\beta^3}\text{E}[X] = -\frac{1}{\beta^2} + \frac{2}{\beta^3}\beta = \frac{1}{\beta^2}.$$

Thus,

$$I_n(\beta) = \frac{n}{\beta^2}.$$

Note that for $\overline{X}_n = (X_1 + \ldots + X_n)/n$, we have

$$\text{Var}(\overline{X}_n) = \frac{\text{Var}(X)}{n} = \frac{\beta^2}{n},$$

so that $\text{Var}(\overline{X}_n) = I_n^{-1}(\beta)$.

This can be generalized to other distributions.

**Theorem:** under appropriate regularity conditions, we have

$$\frac{\widehat{\theta}_{\text{MLE}} - \theta}{\text{Var}\left(\sqrt{\widehat{\theta}_{\text{MLE}}}\right)} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where

$$\text{Var}\left(\sqrt{\widehat{\theta}_{\text{MLE}}}\right) = I_n^{-1}(\theta).$$

**Proof:** the MLE estimator, $\widehat{\theta}_{\text{MLE}}$, solves

$$\frac{\partial}{\partial \theta} \ell(\widehat{\theta}_{\text{MLE}}) = 0,$$

where $\ell$ is the log-likelihood. We apply Taylor's theorem to $\ell$ around $\theta = \widehat{\theta}_{\text{MLE}}$ to obtain

$$\ell(\theta) + \left(\widehat{\theta}_{\text{MLE}} - \theta\right) \frac{\partial^2}{\partial \theta^2} \ell(\theta) \approx 0.$$

Rearranging the terms, we get:

$$\sqrt{n}\left(\widehat{\theta}_{\text{MLE}} - \theta\right) = \frac{\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \ell(\theta)}{-\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ell(\theta)}.$$

Next, we show that the numerator converges to a normal distribution, whereas the denominator converges in probability to a constant.

Recall that

$$\ell(\theta) = \log f(X_1;\theta) + \cdots + \log f(X_n)(\theta)$$

and so

$$\frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta) = \sum_{i=1}^{n} s(X_i;\theta).$$

We have already shown that $\mathrm{E}[s(X_i;\theta)] = 0$. Hence, the numerator can be written as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i,$$

where $Y_i = s(X_i;\theta)$ are i.i.d. with mean 0 and variance

$$\mathrm{Var}(s(X_i;\theta)) = \mathrm{E}[s^2(X_i,\theta)] = I(\theta).$$

Thus, we have

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathrm{E}[s^2(X_1,\theta)]\right) = \mathcal{N}\left(0, I(\theta)\right).$$

Similarly, the numerator can be written as

$$\frac{1}{n}\sum_{i=1}^{n} U_i,$$

where

$$U_i = \frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta), \qquad i = 1,\ldots,n$$

are i.i.d. random variables. From the previous Lemma, we can write

$$\mathrm{E}[U_i] = \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta)\right] = -I(\theta).$$

The **Law of Large Numbers**[73] then yields

$$-\frac{1}{n}\sum_{i=1}^{n} U_i \to I(\theta),$$

from which we conclude the result. ■

**Example: Exponential Distribution (continued)** Applying the theorem on the , we have

$$\sqrt{n}(\overline{X}_n - \beta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \beta^2\right).$$

### 9.9.3 Innovations

We now provide some of the details that allowed us to use innovations in Section 9.7.2. The goal is to try to determine a "good" prediction for the $n+1$th observation in the time series, which we denote by $P_n X_{n+1}$.

A by-product of the **innovation algorithm** is that we will also "predict" $X_1,\ldots,X_n$.[74]

73: To wit: if the $X_i$ are i.i.d. with finite mean $\mu$, then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i = \mu.$$

There are two versions of this, the **weak law** and the **strong law**, depending on the type of convergence, but that falls outside the scope of these course notes, as does **convergence in distribution**, which basically states that the corresponding cumulative distribution functions $F_n$ converge pointwise to a cumulative distribution function $F$.

74: Of course, we do not need to predict these values since they have already been observed in practice, but we can use the **innovations**, i.e., the differences between the observed values $X_i$ and the "predicted" values $\widehat{X}_i$ for model choice and estimation purposes.

As in Section 9.7.2, we define

$$\widehat{X}_{i+1} = P_i X_{i+1} = a_{i1} X_i + \cdots + a_{ii} X_1, \qquad i = 0, \ldots, n;$$

which is to say that $\widehat{X}_{n+1}$ is the **predicted value** for $X_{n+1}$, whereas $\widehat{X}_1, \ldots, \widehat{X}_n$ are the **"predicted" values** for $X_1, \ldots, X_n$.

We also define the column vectors

$$\mathbf{X}_n = (X_1, \ldots, X_n)^\top, \qquad \widehat{\mathbf{X}}_n = (\widehat{X}_1, \ldots, \widehat{X}_n)^\top, \qquad \mathbf{U}_n = (U_1, \ldots, U_n)^\top,$$

where $U_i = X_i - \widehat{X}_i$, $i = 1, \ldots, n$, are the *innovations* of the time series; a "good" prediction is such that these errors are small. As we have no data before $n = 1$ on which to base the prediction, we opt for $\widehat{X}_1 = \mathrm{E}[X_1] = 0$.[75]

75: Remember, we are assuming that $\{X_t\}$ is a stationary time series.

Omitting $\widehat{X}_{n+1}$, we re-write the predictions, individually, as

$$\begin{aligned}
i = 0: &\quad \widehat{X}_1 = 0, \\
i = 1: &\quad \widehat{X}_2 = a_{1,1} X_1, \\
i = 2: &\quad \widehat{X}_3 = a_{2,1} X_2 + a_{2,2} X_1, \\
i = 3: &\quad \widehat{X}_4 = a_{3,1} X_3 + a_{3,2} X_2 + a_{3,1} X_1, \\
&\quad \vdots \\
i = n-2: &\quad \widehat{X}_n = a_{n-1,1} X_{n-1} + \cdots + a_{n-1,n-1} X_1,
\end{aligned}$$

or, simultaneously, as

$$\widehat{\mathbf{X}}_n = \mathbf{A}^* \mathbf{X}_n,$$

where

$$\mathbf{A}^* = \begin{pmatrix}
0 & 0 & 0 & \cdots & 0 \\
a_{1,1} & 0 & 0 & \cdots & 0 \\
a_{2,2} & a_{2,1} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n-1,n-1} & a_{n-1,n-2} & \cdots & a_{n-1,1} & 0
\end{pmatrix}.$$

Note that the matrix is lower diagonal.

We write

$$\mathbf{U}_n = \mathbf{X}_n - \widehat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{A}^* \mathbf{X}_n = \mathbf{A} \mathbf{X}_n,$$

where $\mathbf{A} = \mathbf{I}_n - \mathbf{A}^*$. This matrix is invertible since $\det(\mathbf{A}) = 1 \neq 0$.

Let $\mathbf{C} = \mathbf{A}^{-1}$ and $\mathbf{B} = \mathbf{C} - \mathbf{I}_n$; then we can write

$$\mathbf{X}_n = \mathbf{C} \mathbf{U}_n, \quad \text{and} \quad \widehat{\mathbf{X}}_n = (\mathbf{C} - \mathbf{I}_n)\, \mathbf{U}_n = \mathbf{B} \mathbf{U}_n,$$

representing the "predicted" values in terms of the innovations $\mathbf{U}_n$ and the lower diagonal matrix $\mathbf{B}$ (indeed, $\mathbf{C}$ must be lower diagonal, as is $\mathbf{I}_n$, so that $\mathbf{B} = \mathbf{C} - \mathbf{I}_n$ is also lower diagonal).

We can write the second of these equations as

$$\widehat{\mathbf{X}}_n = (\mathbf{C} - \mathbf{I}_n)\,\mathbf{U}_n = \begin{pmatrix} \widehat{X}_1 \\ \widehat{X}_2 \\ \widehat{X}_3 \\ \vdots \\ \widehat{X}_n \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 0 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{pmatrix},$$

and the first as

$$\mathbf{X}_n = \mathbf{C}\mathbf{U}_n = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_n \end{pmatrix}.$$

Note that the coefficients $\theta_{k,j}$ have nothing to do with the Durbin-Levinson algorithm (see Section 9.4.2).

From the above matrix equation, we have, for instance,

$$\begin{aligned} \widehat{X}_1 &= 0, \\ \widehat{X}_2 &= \theta_{1,1}(X_1 - \widehat{X}_1), \\ \widehat{X}_3 &= \theta_{2,1}(X_2 - \widehat{X}_2) + \theta_{2,2}(X_1 - \widehat{X}_1). \end{aligned}$$

The prediction of $X_3$ is then based on the first and the second innovations $X_1 - \widehat{X}_1$ and $X_2 - \widehat{X}_2$.

In general, for a MA($q$) model, we can write

$$\widehat{X}_{i+1} = \begin{cases} 0 & i = 0 \\ \displaystyle\sum_{j=1}^{i} \theta_{i,j}(X_{i+1-j} - \widehat{X}_{i+1-j}) & i \geq 1 \end{cases}.$$

For an ARMA($p, q$) model, we have instead

$$\widehat{X}_{i+1} = \begin{cases} 0 & i = 0 \\ \phi_1 X_i + \cdots + \phi_p X_{i+1-p} + \displaystyle\sum_{j=1}^{i} \theta_{i,j}(X_{i+1-j} - \widehat{X}_{i+1-j}) & i \geq 1 \end{cases}.$$

The only thing left is to determine how to evaluate the coefficients $\theta_{i,j}$; this is the subject of the next theorem.

**Innovation Algorithm:** assume that $\{X_t\}$ is a stationary time series with mean 0. Let $v_i = \mathrm{E}[(X_{i+1} - \widehat{X}_{i+1})^2]$, $i \geq 0$, and $v_0 = \mathrm{E}[X_1^2] = \gamma_X(0)$.

Then

$$\theta_{n,n-i} = v_i^{-1}\left(\gamma_X(n-i) - \sum_{j=0}^{i-1} \theta_{i,i-j}\theta_{n,n-j}v_j\right), \qquad 0 \leq i < n,$$

$$v_n = \gamma_X(n-1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j.$$

**Example** Consider the MA(1) model $X_t = Z_t + \theta Z_{t-1}$, where $\mathrm{E}[Z_t] = 0$ and $\mathrm{Var}(Z_t) = \sigma_Z^2$. Recall that $\gamma_X(0) = \sigma_Z^2(1+\theta^2)$, $\gamma_X(1) = \theta\sigma_Z^2$ and $\gamma_X(h) = 0$, $h > 1$.

We have:

- $n = 1$
  - $i = 0$: $v_0 = \gamma_X(0) = \sigma_Z^2(1+\theta^2)$, $\theta_{1,1} = v_0^{-1}\gamma_X(1) = \rho_X(1)$ and $v_1 = \gamma_X(0) - \theta_{1,1}^2 v_0$
- $n = 2$
  - $i = 0$: $\theta_{2,2} = v_0^{-1}\gamma_X(2) = 0$
  - $i = 1$: $\theta_{2,1} = v_1^{-1}\gamma_X(1)$ and $v_2 = v_n = [1 + \theta^2 - v_1^{-1}\theta^2\sigma_Z^2]\sigma_Z^2$
- general $n$
  - $i = 0, \ldots, n-2$: $\theta_{n,j} = 0, 2 \le j \le n$,
  - $i = n-1$: $\theta_{n,1} = v_{n-1}^{-1}\gamma_X(1)$ and $v_n = [1 + \theta^2 - v_{n-1}^{-1}\theta^2\sigma_Z^2]\sigma_Z^2$

**Important Property** The innovations $U_1, \ldots, U_n$ are **uncorrelated**: we have $\mathrm{Cov}(U_i, U_j) = 0$ for $i \neq j$.[76] Remembering that the sequence is centered, we have:

76: This is not trivial to show.

$$\Gamma_n = \mathrm{E}\left[\mathbf{X}_n\mathbf{X}_n^\top\right] = \mathrm{E}[\mathbf{C}\mathbf{U}_n\mathbf{U}_n^\top\mathbf{C}^\top] = \mathbf{C}\mathrm{E}[\mathbf{U}_n\mathbf{U}_n^\top]\mathbf{C}^\top = \mathbf{C}\mathbf{D}\mathbf{C}^\top$$

where $\mathbf{D}$ is the diagonal matrix with entries $v_0, \ldots, v_{n-1}$, where the values $v_i = \mathrm{E}[U_i^2] = \mathrm{E}[(X_i - \widehat{X}_i^2)]$ are the same quantities as those in the innovation algorithm.

## 9.10 Exercises

1. Show that the set $\mathcal{T}_n$ of stationary time series of length $n$ is a vector subspace (over $\mathbb{R}$) of the set of all time series.
2. Let $\{Z_t\}$ be independent normalrandom variables with mean 0 and variance $\sigma_Z^2$. Let $a, b, c$ be constants. Which of the following processes are stationary? Evaluate the mean and the autocovariance functions.
   a) $X_t = Z_t\cos(at) + Z_{t-1}\sin(at)$.
   b) $X_t = a + bZ_t + cZ_{t-2}$.
   c) $X_t = Z_tZ_{t-2}$.
3. Let $\{Z_t\}$ be a sequence of independent normal random variables with mean 0 and variance $\sigma_Z^2 = 1$. Consider the sequence
   $$X_t = Z_t + (Z_{t-1}^2 - 1), t = 1, 2, \ldots.$$
   a) Show that $\mathrm{E}[X_t] = 0$.
   b) Show that $\mathrm{E}[X_tX_{t+h}] = 0$ for $h \neq 0$.
4. Let $\{Z_t\}$ be independent random variables with mean 0 and variance $\sigma_Z^2$. Let $\{Y_t\}$ be a stationary sequence with a covariance function $\gamma_Y(h)$. Assume that the sequences $\{Z_t\}$ and $\{Y_t\}$ are independent from each other. Define $X_t = Y_tZ_t$. Verify that $\mathrm{Cov}(X_t, X_{t+h}) = 0$ for $h \ge 1$.

5. Show that the PACF between $X_1$ and $X_3$ when removing the effect of $X_2$ is:

$$\rho_{1,3;2} = \frac{\text{Corr}(X_1, X_3) - \text{Corr}(X_1, X_2) \cdot \text{Corr}(X_2, X_3)}{\sqrt{\left(1 - \text{Corr}^2(X_1, X_2)\right)\left(1 - \text{Corr}^2(X_2, X_3)\right)}}.$$

6. Let $\{Z_t\}$ be independent random variables with mean 0 and variance $\sigma_Z^2$. Consider the model $X_t = Z_t + Z_{t-1}$. Evaluate $\alpha(1)$ and $\alpha(2)$.
7. Let $\{Z_t\}$ be independent random variables with mean 0 and variance $\sigma_Z^2$. Determine if the following processes are stationary and causal.
   a) $X_t + 0.2X_{t-1} + 0.48X_{t-2} = Z_t$.
   b) $X_t + 1.6X_{t-1} = Z_t - 0.42Z_{t-1} + 0.04Z_{t-2}$.
8. Derive a linear representation of the general ARMA$(1, 2)$ model.
9. Derive a linear representation of the general ARMA$(1, q)$ model.
10. Derive a linear representation of the AR(2) model $X_t = \phi X_{t-2} + Z_t$.
11. Use the linear representation of ARMA$(1, 1)$ to compute its covariance function.
12. Use the recursive method to compute the covariance function of the general AR(2) model.
13. This is an exercise about simulating time series.
    a) Generate ARMA$(p, q)$ sequence $X_t$. You have to choose $p, q$ as well as the required parameters. Make sure that the chosen parameters imply existence of a stationary solution.
    b) Identify the model using ACF and PACF. Include graphs of ACF and PACF (2 graphs).
    c) Add a linear or a polynomial trend $m_t$. The new sequence is $Y_t = m_t + X_t$.
    d) Estimate $m_t$ using all three methods:
       - parametric method;
       - exponential smoothing;
       - moving average smoothing with your chosen $Q$.
    e) For each of the three methods, plot $Y_t$ and the estimated trend $\widehat{m}_t$ on the same graphs (3 graphs).
    f) For each of the three methods, compute $\widehat{X}_t = Y_t - \widehat{m}_t$. Plot residuals (that is $\widehat{X}_t$) (3 graphs).
    g) Analyze $\widehat{X}_t$ using ACF and PACF. Graph ACF and PACF for all three methods (6 graphs). Identify the most likely ARMA model for the data. Compare with your identification in b).
14. Download a data set from this page or use your own data set.
    a) Remove the trend using any of the methods, if needed, to obtain a stationary time series. State the chosen $\widehat{m}_t$.
    b) Plot the original sequence together with the estimated trend.
    c) Plot the stationary part, then its ACF and PACF. Comment on the results when it comes to the choice of a model.
15. Assume that $Z_t$ are i.i.d random variables with mean 0 and variance $\sigma_Z^2$.
    a) Apply the Yule-Walker procedure to obtain $P_n X_{n+2}$ (two step prediction) for AR(1) model $X_t = \phi X_{t-1} + Z_t$, $|\phi| < 1$. Compute the corresponding $\text{MSPE}_n(2)$. Can you guess a general formula for $P_n X_{n+k}$?
    b) Apply the Yule-Walker procedure to obtain $P_n X_{n+1}$ for AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. Compute the corresponding $\text{MSPE}_n(1)$.
16. Consider the ARMA$(1, 1)$ model $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$, $|\phi| < 1$, $\theta \in \mathbb{R}$, where $Z_t$ are i.i.d. random variables with mean 0 and variance $\sigma_Z^2$. The goal is to find the best linear predictor $P_n X_{n+1}$ of $X_{n+1}$ based on $X_1, \ldots, X_n$.
    a) Let $n = 1$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n; 1)$ to obtain $a_1$ in $P_1 X_2 = a_1 X_1$.
    b) Let $n = 2$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n; 1)$ to obtain coefficients $a_1, a_2$ in $P_2 X_3 = a_1 X_2 + a_2 X_1$.

    Hint: We have the following formulas for the covariance function:

$$\gamma_X(0) = \sigma_Z^2 \left[1 + \frac{(\phi + \theta)^2}{1 - \phi^2}\right], \quad X(1) = \sigma_Z^2 \left[(\phi + \theta) + \frac{(\phi + \theta)^2 \phi}{1 - \phi^2}\right], \quad \gamma_X(h) = \phi^{h-1} \gamma_X(1), \qquad h \geq 2.$$

17. Consider the MA(1) model $X_t = Z_t + \theta Z_{t-1}$, $\theta \in \mathbb{R}$, where $Z_t$ are i.i.d. random variables with mean 0 and variance $\sigma_Z^2$. The goal is to find the best linear predictor $P_n X_{n+1}$ of $X_{n+1}$ based on $X_1, \ldots, X_n$.
    a) Let $n = 1$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n;1)$ to conclude that
    $$P_1 X_2 = \frac{\gamma_X(1)}{\gamma_X(0)} X_1 = \frac{\theta}{1+\theta^2} X_1.$$
    b) Let $n = 2$. Use the formula $\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}(n;1)$ to obtain coefficients $a_1, a_2$ in $P_2 X_3 = a_1 X_2 + a_2 X_1$.
    c) Let $n = 2$. Apply the Durbin-Levinson algorithm to get $P_2 X_3 = \phi_{2,1} X_2 + \phi_{2,2} X_1$.
18. Consider a stationary ARMA(1, 1) model
    $$(X_t - \mu) = \phi(X_{t-1} - \mu) + Z_t + \theta Z_{t-1}.$$
    Evaluate $\sum_{h=-\infty}^{\infty} \gamma_X(h)$.
19. Assume that $Z_t$ are i.i.d random variables with mean 0 and variance $\sigma_Z^2$. Consider the AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$.
    a) Derive confidence intervals for $\widehat{\phi}_1$ and $\widehat{\phi}_2$.
    b) Assume that $n = 100$, $\widehat{\gamma}_X(0) = 3$, $\widehat{\gamma}_X(1) = 1.5$, $\widehat{\gamma}_X(2) = 0.5$. Use a) to get the confidence intervals.
20. In this question we develop Yule-Walker estimators for the AR(1) and ARMA(1, 1) models and study their numerical performance. Recall that the Yule-Walker estimator for the AR(1) model is
    $$\widehat{\phi} = \frac{\widehat{\gamma}_X(1)}{\widehat{\gamma}_X(0)} = \widehat{\rho}_X(1), \quad \widehat{\sigma}_Z^2 = \widehat{\gamma}_X(0) - \widehat{\phi}\widehat{\gamma}_X(1) = \widehat{\gamma}_X(0) - \widehat{\rho}_X(1)^2 \widehat{\gamma}_X(0).$$
    a) Numerical experiment for AR(1):
        i. Load the file `Data-AR.txt` into `R`. This is a data set generated from a AR(1) model with $\phi = 0.8$.
        ii. Type `var(Data)` to obtain $\widehat{\gamma}_X(0)$.
        iii. Type `ACF<-acf(Data)`. Then type `ACF`. You will get $\widehat{\rho}_X(h)$, the estimators of $\rho_X(h)$. The second entry is $\widehat{\rho}_X(1) = \widehat{\phi}$.
        iv. Write the final values for $\widehat{\phi}$ and $\widehat{\sigma}_Z^2$.
        v. Compare the estimated $\widehat{\phi}$ with the true $\phi$.
    b) Consider the ARMA(1, 1) model $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$, $|\phi| < 1$; the sequence $X_t$ is causal. Apply the Yule-Walker procedure to obtain the estimators for $\phi$, $\theta$ and $\sigma_Z^2 = \mathrm{Var}(Z_t)$.
    c) Numerical experiment for ARMA(1, 1):
        i. Load the file `Data-ARMA.txt` into `R`. This is a data set generated from a ARMA(1, 1) model with $\phi = 0.8$ and $\theta = 1$.
        ii. Identify the values of $\widehat{\phi}$, $\widehat{\theta}$, and $\widehat{\sigma}_Z^2$.
        iii. Compare the estimated $\widehat{\phi}$ with the true $\phi$. Which estimate is more accurate: ARMA(1, 1) or AR(1)?
21. a) One hundred observations from AR(1) yield the following sample statistics:
    $$\overline{x} = 0, \quad \widehat{\gamma}_X(0) = 1.1, \quad \widehat{\rho}_X(1) = 0.42.$$
        i. Find the Yule-Walker estimators of $\phi$ and $\sigma_Z^2$.
        ii. Write the confidence interval for $\phi$.
        iii. If $X_{100} = 1.5$, what is the predicted value of $X_{101}$? What is the squared error of this prediction?
    b) Two hundred observation from AR(2) yields the following sample statistics:
    $$\overline{x} = 3.82, \quad \widehat{\gamma}_X(0) = 1.15, \quad \widehat{\rho}_X(1) = 0.427, \widehat{\rho}_2 = 0.475.$$
        i. Find the Yule-Walker estimators of $\phi_1$, $\phi_2$ and $\sigma_Z^2$.
        ii. Is the estimated model causal?.
        iii. If $X_{100} = 3.84$ and $X_{99} = 3.26$, what is the predicted value of $X_{101}$?

22. Consider the general AR(1) model. Derive the MLE for $\phi$ and $\sigma_Z^2$.
23. We have already fitted an AR(4) model to US unemployment data, and estimated the parameters using the Yule-Walker procedure.
    a) Calculate the residuals, and plot their ACF and PACF. Is the chosen AR(4) model appropriate?
    b) Predict the next observation in the time series.
    c) Backcast the past observations and verify the quality of the "prediction" by plotting the original values and the "predicted" values on the same graph. Compute the squared error of that prediction.
    d) Now, pretend that the model is AR(1). Estimate the model's parameters. Repeat b)-d). State conclusions.
24. Use the Lake Huron data for this question (an in-built dataset in `R`).
    a) Type the following code at the prompt.

```
My.TS <- LakeHuron
help(LakeHuron)
mean = mean(My.TS)
My.Centered.TS <- My.TS - mean(My.TS)
```

    b) Fit an AR(2) model to the data using the Yule-Walker estimator. Obtain $\widehat{\phi}_1$, $\widehat{\phi}_2$, $\widehat{\sigma}_Z^2$.

```
fit.ar <- ar(My.Centered.TS, method="yule-walker")
```

    c) Verify that the command `ar()` leads to the correct Yule-Walker estimator.
        i. At the prompt, type the following code.

```
ACF <- acf(LakeHuron)
var(LakeHuron)
```

        Read off $\widehat{\rho}_X(1)$ and $\widehat{\rho}_X(2)$ and $\widehat{\gamma}_X(0)$. Using theis information, compute $\widehat{\gamma}_X(1)$, $\widehat{\gamma}_X(2)$.
        ii. Create a vector $(\widehat{\gamma}_X(1), \widehat{\gamma}_X(2))$ and call it `gamma.vector`.
        iii. Create a matrix $\widehat{\Gamma}_2$ and call it `Gamma.matrix`.
        iv. Compute $\hat{\Gamma}_2^{-1} * \gamma_{X,2}$ by typing in

```
solve(Gamma.matrix)%*%gamma.vector
```

        Compare the results with those of part b).
25. When $p \geq 2$, it can be rather difficult to identify the right $p$ from the data. Start by loading `BadData.txt` into the `R` variable `X`.
    a) Based on the ACF and PACF of the data, argue that an AR(3) model can be reasonably chosen.
    b) Type the following code at the prompt.

```
(fit.ar <- ar(X,method="mle"))
```

    What order does `ar()` select? Denote this order by `p`.
    c) Using `p` from the step above, type the following code at the prompt.

```
(fit.arima  <- arima(X,order=c(3,0,0)))
(fit.arima1 <- arima(X,order=c(p,0,0)))
```

    Why did MLE select `p` and not 3?
26. Derive the formulas for the spectral density of MA(1) and ARMA(1, 1).

27. Assume that $(X_1, X_2)$ is a vector of dependent normal random variables with mean 0 and variance $\sigma^2$ each. Assume that the covariance matrix is given by

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

In other words, $\rho$ is the correlation between $X_1$ and $X_2$. Assuming that $\sigma$ is known, find the maximum likelihood estimator of $\rho$.

28. Let $\{Z_t\}$ be an i.i.d. sequence of normal random variables with mean 0 and variance $\sigma_Z^2 = 1$. Define

$$X_t = \begin{cases} Z_t, & t \text{ even}, \\ (Z_{t-1}^2 - 1)/\sqrt{2}, & t \text{ odd}. \end{cases}$$

Find $\mathrm{E}[X_t]$, $\gamma_X(t, t+1)$ and $\gamma_X(t, t+2)$.

29. Consider the sequence

$$X_t = Z_t Z_{t-1} + 0.5 Z_{t-1},$$

where $Z_t$ are i.i.d random variables with mean 0 and variance $\sigma_Z^2$.

a) Show that $\mathrm{E}[X_t] = 0$ for all $t$.
b) Compute $\gamma_X(t, t+h) = \mathrm{E}[X_t X_{t+h}]$ for $h = 0, 1, 2$.
c) Is the sequence $X_t$ stationary? Why?

30. Assume that $A$ and $B$ are random variables with mean 0 and variance $\sigma^2$. Assume also that $\mathrm{Cov}(A, B) = 0$. Let $\omega \in \mathbb{R}$ and define

$$X_t = A\cos(at) + B\sin(bt), \quad a, b \neq 0.$$

Is $\{X_t\}$ stationary?

31. Consider the ARMA(2, 1) model given by

$$X_t - 0.75X_{t-1} + 0.5625X_{t-2} = Z_t + 2.25Z_{t-1}.$$

Is this process causal? Is this process stationary?

32. Consider the linear process given by

$$X_t = \sum_{j=0}^{\infty} (\phi^j + \phi^{j+1}) Z_{t-j},$$

where $|\phi| < 1$ and $Z_t$ is an i.i.d sequence with mean 0 and variance $\sigma_Z^2$. Write the formula for $\gamma_X(h)$, $h \geq 0$.

33. Consider the ARMA(1, 2) model

$$X_t - \phi X_{t-1} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2},$$

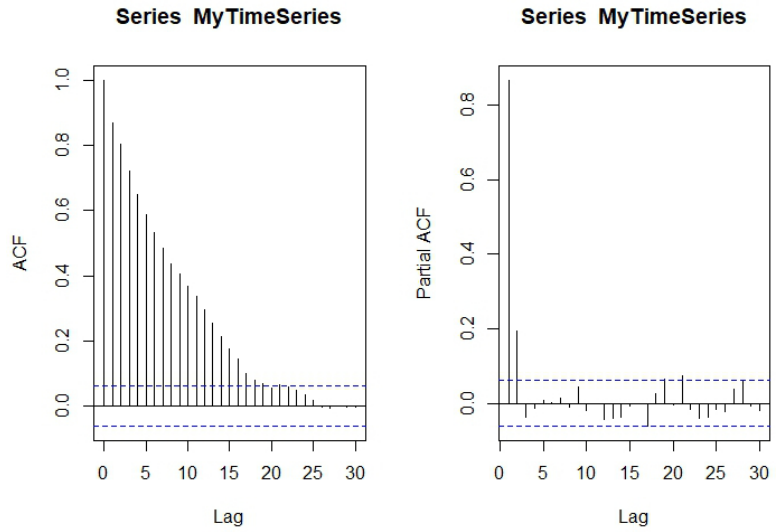
where $|\phi| < 1$, $\theta_1, \theta_2 \in \mathbb{R}$, and $Z_t$ is an i.i.d sequence with mean 0 and variance $\sigma_Z^2$. Derive the linear representation for $X_t$, i.e. find the coefficients $\psi_j$ in $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$.

34. Consider a stationary AR(3) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} = Z_t$. Use the recursive method to conclude

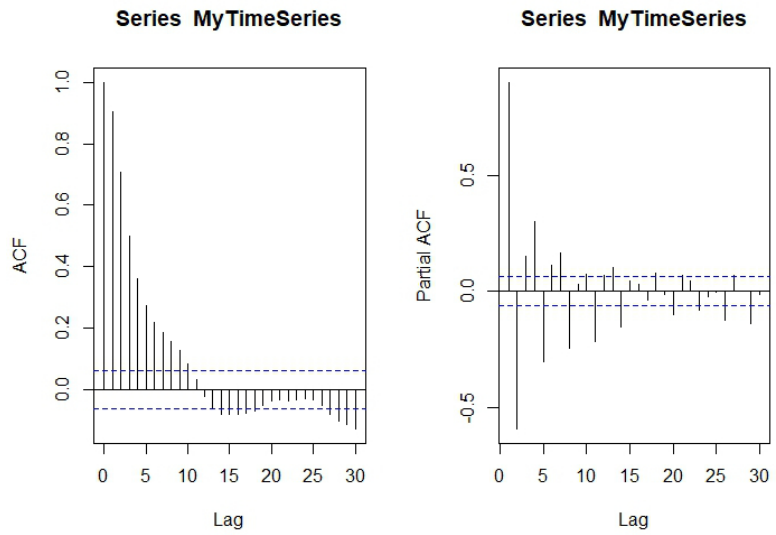
$$\gamma_X(h) = \phi_1 \gamma_X(h-1) + \phi_1 \gamma_X(h-2) + \phi_1 \gamma_X(h-3), \quad h \geq 3.$$

35. Derive the linear representation of a stationary AR(2) model $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$.
36. Write the non-causal linear representation of an AR(1) $X_t = \phi X_{t-1} + Z_t$ with $\phi > 1$.
37. Obtain the coefficients $\phi_{1,1}$, $\phi_{2,2}$, $\phi_{3,3}$ for the AR(1) model. Compare with the Yule-Walker procedure.
38. Obtain the coefficients $\phi_{1,1}$, $\phi_{2,2}$, $\phi_{2,1}$ for the AR(2) model.
39. If $\{X_t\}$ and $\{Y_t\}$ are two uncorrelated stationary processes, show that $\{X_t + Y_t\}$ is a stationary process. What is its ACVF?

40. Identify the ARMA model based on the ACF and PACF below.




41. Identify the ARMA model based on the ACF and PACF below.




42. Consider the AR(1) model $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and the random variables $Z_t$ are i.i.d. with mean 0 and variance $\sigma_Z^2$. Prove that $\phi_{n,n} = 0$ for all $n \geq 2$ (recall that $\phi_{n,n}$ = partial autocovariance at lag $n$).

43. a) Let $X$ and $Y$ be random variables with $\mathrm{E}[Y^2] < \infty$. Show that $\mathrm{E}[Y \mid X]$ minimizes

$$\mathrm{MSE} = \mathrm{E}\left([Y - g(X)]^2\right)$$

over all functions $g$ such that $\mathrm{E}\left([g(X)]^2\right) < \infty$.

b) Generalize to $X_1, \ldots, X_n$ to show that $\mathrm{E}[X_{n+1} | X_1, \ldots, X_n]$ minimizes

$$\mathrm{MSE} = \mathrm{E}\left([X_{n+1} - g(X_1, \ldots, X_n)]^2\right)$$

over all functions $g$ such that $\mathrm{E}\left([g(X_1, \ldots, X_n)]^2\right) < \infty$.

c) If $X_1, X_2, \ldots$ are i.i.d. with $\mathrm{E}[X_i^2] < \infty$ and $\mathrm{E}[X_i] = \mu$ for all $i$, where $\mu$ is known, what is the minimum mean square predictor of $X_{n+1}$ in terms of $X_1, \ldots, X_n$?

d) If $X_1, \ldots, X_n$ are i.i.d. with $\mathrm{E}[X_i^2] < \infty$ and $\mathrm{E}[X_i] = \mu$ for all $i$, where $\mu$ is unknown, show that the best linear unbiased estimator (BLUE) of $\mu$ is $\overline{X}$.

44. Let $\{Z_t\}$ be i.i.d. with $Z_t \sim N(0,1)$ and define

$$X_t = \begin{cases} Z_t & \text{if } t \text{ is even} \\ \frac{Z_{t-1}^2 - 1}{\sqrt{2}} & \text{if } t \text{ is odd} \end{cases}$$

a) Show that $\{X_t\}$ is WN$(0,1)$ but not i.i.d. $(0,1)$ noise.
b) Find $\mathrm{E}[X_{n+1}|X_1,\ldots,X_n]$ for $n$ even and for $n$ odd and compare the results.

45. Consider the time series

$$X_t = \underbrace{m_t}_{\text{local trend}} + \underbrace{Z_t}_{\text{noise}}$$

and the simple moving average filter with weights $a_j = (2q+1)^{-1}$ for $-q \le j \le q$.

a) If $m_t = c_0 + c_1 t$ show that $\sum_{j=-q}^{q} a_j m_{t-j} = m_t$.
b) If $\{Z_t\}_{t\in\mathbb{Z}}$ are i.i.d. with mean 0 and variance $\sigma^2$, show that the moving average

$$A_t = \sum_{j=-q}^{q} a_j Z_{t-j}$$

is small in the sense that $\mathrm{E}[A_t] = 0$ and $\mathrm{Var}(A_t^2) = \frac{\sigma^2}{2q+1}$.

46. Compute the ACF of the model $X_t - 0.6X_{t-1} = Z_t + 1.2Z_{t-1}$, where $Z_t$ is WN$(0,\sigma^2)$.
47. Let $X_t$ denote a non-causal AR(1) process $X_t = \phi X_{t-1} + Z_t$ where $\{Z_t\} \sim$ WN$(0,\sigma^2)$ and $|\phi| > 1$.

a) Denote $W_t = X_t - \frac{1}{\phi}X_{t-1}$. Show that $\{W_t\} \sim$ WN$(0,\sigma_w^2)$ and express $\sigma_w^2$ in terms of $\sigma^2$ and $\phi$.
b) Show that $Y_t = \frac{1}{\phi}Y_{t-1} + W_t$ is causal and has the same ACVF as $X_t$ above.
c) Find the causal form of $X_t = 1.2X_{t-1} + Z_t$ where $\{Z_t\} \sim$ WN$(0,1)$.

48. Let $\{Y_t\}$ be the AR(1) plus white noise time series defined by $Y_t = X_t + W_t$ where $\{W_t\} \sim$ WN$(0,\sigma_w^2)$, $\{X_t\}$ is the AR(1) process $X_t - \phi X_{t-1} = Z_t$, $|\phi| < 1$, $\{Z_t\} \sim$ WN$(0,\sigma_z^2)$, $\mathrm{E}[X_s Z_t] = 0$ for all $s < t$ and $\mathrm{E}[W_s Z_t] = 0$ for all $s,t$.

a) Show that $\{Y_t\}$ is stationary and find its ACVF.
b) Show that the time series $U_t = Y_t - \phi Y_{t-1}$ is 1−correlated and hence is an MA(1) process.
c) Conclude from b) that $\{Y_t\}$ is an ARMA$(1,1)$ process and express the three parameters of this model in terms of $\phi$, $\sigma_w^2$ and $\sigma_z^2$.

49. Let $\{X_t\}$ be an AR$(p)$ process defined by

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

where $\{Z_t\} \sim$ WN$(0,\sigma^2)$ and $\mathrm{E}[X_s Z_t] = 0$ for all $s < t$.

a) Show that for $n > p$, the best linear predictor $P_n X_{n+1}$ is $\phi_1 X_n + \cdots + \phi_p X_{n-p}$.
b) Compute the mean square error of this forecast.

50. Let $\{X_t\}$ be an MA(1) process defined by $X_t = Z_t - \theta Z_{t-1}, \quad t \in \mathbb{Z}$ where $\{Z_t\} \sim$ WN$(0,\sigma^2)$ and $|\theta| < 1$.

a) Show that the best linear predictor $\tilde{P}_n X_{n+1}$ based on $\{X_j | j \le n\}$ is

$$\tilde{P}_n X_{n+1} = -\sum_{j=1}^{\infty} \theta^j X_{n+1-j}.$$

b) Find the mean square error of $\tilde{P}_n X_{n+1}$.

51. In the innovations algorithm, show that for each $n \ge 2$, the innovation $X_n - \hat{X}_n$ is uncorrelated with $X_1,\ldots,X_{n-1}$. Conclude also that the innovation $X_n - \hat{X}_n$ is uncorrelated with the innovations $X_1 - \hat{X}_1,\ldots,X_{n-1} - \hat{X}_{n-1}$.

52. Let $X_1, X_2, X_4, X_5$ be observations from an MA(1) process defined by $X_t = Z_t - \theta Z_{t-1}$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.
    a) Find the best linear estimate of the missing value $X_3$ in terms of $X_1, X_2$.
    b) Find the best linear estimate of the missing value $X_3$ in terms of $X_4, X_5$.
    c) Find the best linear estimate of the missing value $X_3$ in terms of $X_1, X_2, X_4, X_5$.
    d) Compute the mean squared error of the previous estimates. Which one of them is the best estimate for $X_3$.
53. Let $\{X_t\}$ be an AR($p$) process defined by $X_t = \phi X_{t-1} + Z_t$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.
    a) Show that $\sqrt{n}\frac{\hat{\rho}(1)-\rho(1)}{\sqrt{1-(\rho(1))^2}}$ has asymptotically standard normal distribution $N(0,1)$.
    b) If $n = 100$ and $\hat{\rho}(1) = 0.64$, build an approximate 95% confidence interval for $\phi$.
54. Let $\{X_t\}$ be an AR(1) process defined by $X_t = \phi X_{t-1} + Z_t$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ with the usual hypotheses. For $h = 1, 2, \ldots$, compute the $h-$step ahead forecast $P_n X_{n+h} = \hat{X}_n(h)$ in terms of $\{1, X_n, \ldots, X_1\}$ and find its mean square error.
55. Suppose that $\{X_t\}$ is a non-causal and non-invertible ARMA(1, 1) process satisfying $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, with $|\phi|, |\theta| > 1$. Define $\tilde{\phi}(B) = 1 - \frac{B}{\phi}$ and $\tilde{\theta}(B) = 1 + \frac{B}{\theta}$ and let $W_t = \tilde{\theta}^{-1}(B)\tilde{\phi}(B)X_t$.
    a) Show that $\{W_t\}$ has a constant spectral density function.
    b) Conclude that $\{W_t\} \sim \text{WN}(0, \sigma_w^2)$. Give an explicit formula for $\sigma_w^2$ in terms of $\sigma^2$, $\theta$ and $\phi$.
    c) Deduce that $\tilde{\phi}(B)X_t = \tilde{\theta}(B)W_t$, so that $\{X_t\}$ is a causal and invertible ARMA(1, 1) process relative to the white noise $\{W_t\}$ (see [1] for definition).
56. Let $\{X_t\}$ be the MA(1) process defined by $X_t = Z_t + \theta Z_{t-1}$ where $|\theta| < 1$ and $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. The best linear predictor of $X_{n+1}$ based on $X_1, \ldots, X_n$ is

$$\hat{X}_{n+1} = \phi_{n,1}X_n + \cdots + \phi_{n,n}X_1,$$

where $\phi_n = (\phi_{n,1}, \ldots, \phi_{n,n})^\top$ satisfies $R_n\phi_n = \rho_n$; $\rho_n = (\rho(1), \ldots, \rho(n))^\top$. Show that

$$\phi_{n,n-j} = (1 + \theta^2 + \cdots + \theta^{2j})(-\theta)^{-j}\phi_{n,n} \quad \text{for } 1 \le j < n$$

and conclude that the PACF of the process is

$$\phi_{n,n} = -\frac{(-\theta)^n}{1 + \theta^2 + \cdots + \theta^{2n}}.$$

57. Let $\{X_t\}$ be a causal ARMA(1, 1) process of the form $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Consider the innovation algorithm

$$\hat{X}_{n+1} = \phi X_n + \theta_{n,1}(X_n - \hat{X}_n)$$

for this process. It can be shown that the innovation algorithm coefficients $\theta_{n,1}$ can be found recursively as follows:

$$r_0 = \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2}, \quad \theta_{n,1} = \frac{\theta}{r_{n-1}}, \quad r_n = 1 + \theta^2\left(1 - \frac{1}{r_{n-1}}\right).$$

    a) With the notation $y_n = \frac{r_n}{r_n - 1}$, show that

$$y_n = \theta^{-2}y_{n-1} + 1, \quad n \ge 1.$$

    b) Deduce that

$$y_n = \theta^{-2n}y_0 + \sum_{j=1}^{n} \theta^{-2(j-1)} := A(n).$$

    Determine $r_n$ and $\theta_{n,1}$ for all $n \ge 1$.
    c) Evaluate the limits of $r_n$ and $\theta_{n,1}$ for $|\theta| < 1$ as $n \to \infty$.

58. a) Compute and plot the spectral density of the stationary series $\{X_t\}$ satisfying

$$X_t - 0.99X_{t-3} = Z_t, \quad \{Z_t\} \sim \text{WN}(0,1).$$

b) Does the spectral density suggest that the sample paths of $\{X_t\}$ will exhibit approximately oscillatory behaviour? If so, then with what period?
c) Simulate and plot a realization of $X_1, \ldots, X_{60}$. Does the graph of the realization support the conclusion in part b)?
d) Compute the spectral density of the filtered process

$$Y_t = \frac{1}{3}(X_{t-1} + X_t + X_{t+1})$$

and compare the numerical values of the spectral densities of $\{X_t\}$ and $\{Y_t\}$ at frequency $\lambda = \frac{2\pi}{3}$ radians per unit time. What effect would you expect the filter to have on the oscillations of $\{X_t\}$?
e) Apply the filter of part d) to the realization of part c). Comment on the result.

59. Consider the sunspot numbers $\{X_t, t = 1, \ldots, 100\}$, filed as SUNSPOTS.TSM.
a) Compute the sample autocovariances $\hat{\gamma}(0)$, $\hat{\gamma}(1)$, $\hat{\gamma}(2)$ and $\hat{\gamma}(3)$.
b) Use these values to find the Yule-Walker estimates of $\phi_1$, $\phi_2$ and $\sigma^2$ in the AR(2) model

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2)$$

for the mean corrected series $Y_t = X_t - \overline{X}_t$.
c) Assuming that the data really are a realization of an AR(2) process, find 95% C.I. for $\hat{\phi}_1$ and $\hat{\phi}_2$.
d) Use the Durbin-Levinson algorithm to compute the sample PACF $\hat{\phi}_{1,1}$, $\hat{\phi}_{2,2}$ and $\hat{\phi}_{3,3}$ of the sunspot series. Is the value of $\hat{\phi}_{3,3}$ compatible with the assumption that the data are generated from an AR(2) process? Use significance level $\alpha = 0.05$.

60. Use the ARMA Process Gaussian Likelihood formula to prove that if $\{X_t\}$ is an AR($p$) process with the equation $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, then for $n > p$, the likelihood function can be written as

$$L(\phi, \sigma^2) = (2\pi\sigma^2)^{-n/2}(\det(G_p))^{-1/2} \exp\left\{-\tfrac{1}{2\sigma^2}\left[\mathbf{X}_p^\top G_p^{-1}\mathbf{X}_p + \textstyle\sum_{t=p+1}^{n} Z_t^2\right]\right\},$$

where $\mathbf{X}_p = (X_1, \ldots, X_p)^\top$, $\phi = (\phi_1, \ldots, \phi_p)^\top$ and $G_p = \sigma^{-2}\Gamma_p = \sigma^{-2}E(\mathbf{X}_p\mathbf{X}_p^\top)$.

61. If $\{Y_t\}$ is a zero-mean causal ARMA process and $X_0$ is uncorrelated with $Y_t$ for all $t$, show that the best linear predictor of $Y_{n+1}$ in terms of $1, X_0, Y_1, \ldots, Y_n$ is the same as the best linear predictor of $Y_{n+1}$ in terms of $1, Y_1, \ldots, Y_n$.

62. Suppose that $\{Z_t\}$ is a causal stationary AR($p$) process with $\text{E}[Z_t^4] < \infty$, and $Z_t = \sqrt{h_t}e_t$ where $\{e_t\} \sim$ i.i.d. $(0,1)$ and

$$h_t = \alpha_0 + \alpha_1 Z_{t-1}^2 + \cdots + \alpha_p Z_{t-p}^2, \quad \sum_{j=1}^{p} \alpha_j < 1.$$

a) Show that $\text{E}[Z_t^2 | Z_{t-1}^2, Z_{t-2}^2, \ldots] = h_t$.
b) Show that $\{Z_t^2\}$ is an AR($p$) process. Identify its parameters.

## Chapter References


[1] A. Aue. *Time Series Analysis*. LibreTexts, 2021.

[2] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer New York, 2006.

[3] P.S.P. Cowpertwait and A.V. Metcalfe. *Introductory Time Series with R*. Use R! Springer New York, 2009.

[4] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[5] Department of Statistics. *Applied Time Series Analysis*. PennState's College of Science.

# Survey Sampling Methods

# 10

by **Patrick Boily** (inspired by **Patrick Farrell**)

---

Simply put, data analysis requires data. In pedagogical settings, we take for granted that the data at our disposal is "perfect" (or "ideal"): it either consists of the totality of potentially available data, or it is a representative subset thereof. In practice, either of these can be difficult to achieve; it can prove costly (and sometimes impractical) to collect data from which we can infer population trends and characteristics.

While web scraping (and automated methods) are sometimes used to facilitate the data collection process (see Chapter 16, *Web Scraping and Automatic Data Collection*), the samples that they provide often fail to be representative enough to be of use in practice.

In this chapter, we discuss the principles that underlie statistical sampling methods, and show how to obtain estimates for various sampling plans.

## 10.1 Background

> To call in the statistician after the experiment is done may be no more than asking them to perform a post-mortem examination: at best, they may be able to say what the experiment died of. [R.A. Fisher, Presidential Address to the *First Indian Statistical Congress*, 1938]

Data analysis tools and techniques work in conjunction with collected data. The type of data that needs to be collected to carry out such analyses, as well as the priority placed on the collection of quality data relative to other demands, will dictate the choice of data collection strategies.

The manner in which the resulting outputs of these analyses are used for decision support will, in turn, influence appropriate data presentation strategies and system functionality, which is an important access of the analytical process. Although analysts should always endeavour to work with **representative** and **unbiased data**, there will be times when the available data is flawed and not easily repaired.

Analysts have a professional responsibility to explore the data, looking for potential fatal flaws **prior** to the analysis and to inform their client and stakeholders of any findings that could **halt**, **skew**, or simply **hinder** the analytical process or its applicability to the situation at hand.[1]


10.1 Background . . . . . 599
Sampling Generalities . . . 602
Survey Frames . . . . . 604
Fundamental Concepts . . . 604
Data Collection Basics . . . 607
Sampling Types . . . . . 607
10.2 Questionnaire Design . . . 610
Basic Concepts . . . . . 610
Question Types . . . . . 611
Design Considerations . . . 612
Question Order . . . . . 613
10.3 Simple Random Sampling . 615
Basic Notions . . . . . 619
Estimators and C.I. . . . . 622
Sample Size . . . . . 635
10.4 Stratified Sampling . . . . . 637
Estimators and C.I. . . . . 644
Sample Size and Allocation 654
Comparison: SRS and STS 661
10.5 Auxiliary Information . . . 663
Ratio Estimation . . . . . 663
Regression Estimation . . . 674
Difference Estimation . . . . 681
Comparisons . . . . . 684
10.6 Cluster Sampling . . . . . 688
Estimators and C.I. . . . . 688
Sample Size . . . . . 704
Comparison: SRS and CLS 706
10.7 Special Topics . . . . . 707
Systematic Sampling . . . . 707
Sampling with PPS . . . . . 713
Multi-Stage Sampling . . . 716
Multi-Phase Sampling . . . 720
Miscellaneous . . . . . 722
10.8 Exercises . . . . . 730
Chapter References . . . . . 732


1: Unless some clause has specifically been put in the contract/agreement to allow a graceful exit at this point, consultants will have to proceed with the analysis, flaws and all. It is **EXTREMELY IMPORTANT** that one does not simply sweep these flaws under the carpet. Address them repeatedly in meetings with the clients, and make sure that the analysis results that are presented or reported on include an appropriate *caveat*.

### Formulating the Problem

The **objectives** drive all other aspects of quantitative analysis. With a **question** (or questions) in mind, an investigator can start the process that leads to **model selection**.

With potential models in tow, the next step is to consider:

- what **variates** (fields, variables) are needed,
- the **number** of observations required to achieve a pre-determined **precision**, and
- how to best go about **collecting**, **storing** and **accessing** the data.

Another important aspect of the problem is to determine whether the questions are being asked of the data in and of **itself**, or whether the data is used as a **stand-in for a larger population**. In the later case, there are other technical issues to incorporate into the analysis in order to be able to obtain generalizable results.

Questions do more than just drive the other aspects of data analysis – they also drive the development of quantitative methods. They come in all flavours and their variability and breadth make attempts to answer them challenging: no single approach can work for all of them, or even for a majority of them, which leads to the discovery of better methods, which are in turn applicable to new situations, and so on, and so on.

**Not every question is answerable**, of course, but a large proportion of them may be answerable partially or completely; quantitative methods can provide **insights**, **estimates**, and **ranges** for possible answers, and they can point the way towards possible implementations of the solutions.

As an illustration, consider the following questions:

- Is cancer incidence higher for second-hand smokers than it is for smoke-free individuals?
- Using past fatal collision data and economic indicators, can we predict future fatal collision rates given a specific national unemployment rate?
- What effect would moving a central office to a new location have on average employee commuting time?
- Is a clinical agent effective in the treatment against acne?
- Can we predict when border-crossing traffic is likely to be higher than usual, in order to appropriately schedule staff rotations?
- Can personalized offers be provided to past clients to increase the likelihood of them becoming repeat customers?
- Has employee productivity increased since the company introduced mandatory language training?
- Is there a link between early marijuana use and heavy drug use later in life?
- How do selfies from over the world differ in everything from mood to mouth gape to head tilt?

Next steps nearly always requires obtaining relevant data.

### Data Types

Data has **attributes** and **properties**. Fields are classified as **response**, **auxiliary**, **demographic** or **classification** variables; they can be **quantitative** or **qualitative**; **categorical**, **ordinal** or **continuous**; **text-based** or **numerical**.

Furthermore, data is **collected** through experiments, interviews, censuses, surveys, sensors, scraped from the Internet, etc. Collection methods are not always sophisticated, but new technologies usually improves the process in many ways (while introducing new issues and challenges): modern data collection can occur over **one pass**, in **batches**, or **continuously**.

How does one decide which data collection method to use?

The type of question to answer obviously has an effect, as do the required precision, cost and timeliness. Statistics Canada's *Survey Methods and Practices* [10] provides a wealth of information on probabilistic sampling and questionnaire design, which remain relevant in this day of big (and real-time) data.

The importance of this step cannot be overstated: without a **well-designed plan** to collect meaningful data, and without safeguards to identify flaws (and possible fixes) as the data comes in, subsequent steps are likely to prove a waste of time and resources.

As an illustration of the potential effect that data collection can have on the final analysis results, contrast the two following "ways" to collect similar data.

> The Government of Québec has made public its proposal to negotiate a new agreement with the rest of Canada, based on the equality of nations; this agreement would enable Québec to acquire the exclusive power to make its laws, levy its taxes and establish relations abroad – in other words, sovereignty – and at the same time to maintain with Canada an economic association including a common currency; any change in political status resulting from these negotiations will only be implemented with popular approval through another referendum; on these terms, do you give the Government of Québec the mandate to negotiate the proposed agreement between Québec and Canada? [1980 Québec sovereignty referendum question]
>
> Should Scotland be an independent country? [2014 Scotland independence referendum question]

The end result was the same in both instances (no to independence), but an argument can easily be made that the 2014 Scottish 'No' was a much clearer 'No' than the Québec 'No' of 34 years earlier, in spite of the smaller 2014 victory margin.[2]

2: 55.3%-44.7% in the Scotland referendum, as opposed to 59.6%-40.4% in the Québec referendum.

#### Data Storage and Access

Data **storage** is also strongly linked with the data collection process, in which decisions need to be made to reflect how the data is being collected (one pass, batch, continuously), the volume of data that is being collected, and the type of access and processing that will be required (how fast, how much, by whom).

Stored data may go **stale** (e.g., people move, addresses are no longer accurate, etc.), so it may be necessary to implement regular updating collection procedures.

Until very recently, the story of data analysis has only been written for small datasets: useful collection techniques yielded data that could, for the most part, be stored on personal computers or on small servers.

The advent of "Big Data" has introduced new challenges *vis-à-vis* the collection, capture, access, storage, analysis and visualisation of datasets; some effective solutions have been proposed and implemented, and intriguing new approaches are on the way.[3]

3: Such as DNA storing [8], to name but one (!).

We shall not discuss those challenges in detail in this module, but we urge analysts and consultants alike to be aware of their existence.

### 10.1.1 Survey Sampling Generalities

> The latest survey shows that 3 out of 4 people make up 75% of the world's population. [David Letterman]

While the *World Wide Web* does contain troves of data, web scraping (see Chapter 16) does not address the question of **data validity**: will the extracted data be **useful** as an analytical component? Will it suffice to provide the quantitative answers that clients and stakeholders are seeking?

A **survey** [10] is any activity that collects information about characteristics of interest:

- in an **organized** and **methodical** manner;
- from some or all **units** of a population;
- using **well-defined** concepts, methods, and procedures, and
- compiles such information into a **meaningful** summary form.

A **census** is a survey where information is collected from all units of a population, whereas a **sample survey** uses only a fraction of the units.

#### Sampling Model

When survey sampling is done properly, we may be able to use various statistical methods to make inferences about the **target population** by sampling a (comparatively) small number of units in the **study population**.

The relationship between the various populations (**target**, **study**, **respondent**) and samples (**sample**, **intended**, **achieved**) is illustrated in Figure 10.1.

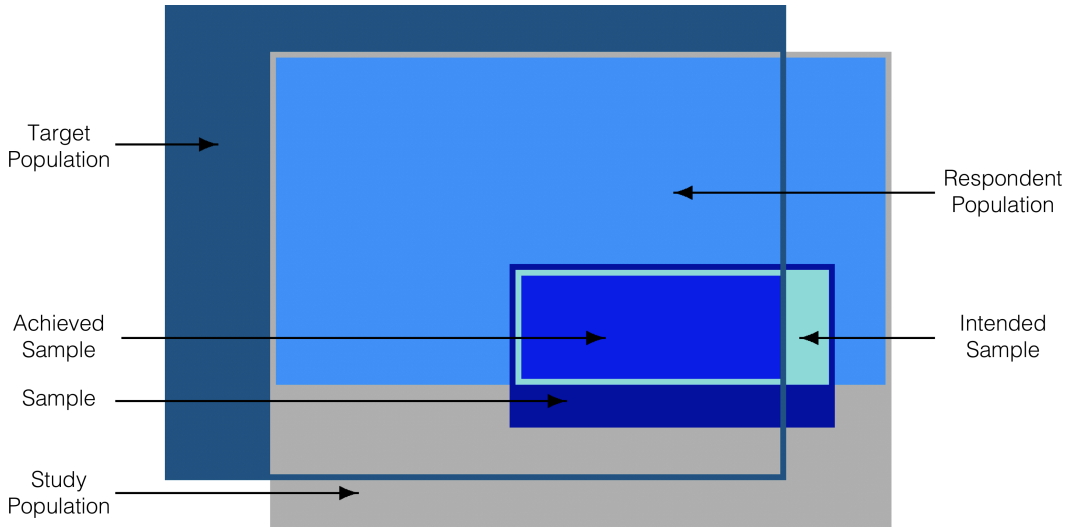



**Figure 10.1:** Various populations and samples in the sampling model.

- **Target population:** population for which we want to obtain information;
- **Study population** (survey population): population covered by the survey (it may be different from the target population, but ideally the two are very similar);[4] conclusions drawn from the survey results only apply to the study population;
- **Respondent population:** units of the study population that would participate in the survey if they were asked to do so; it may be different from the study population if the respondents are not representative of the study population;
- **Survey frame:** provides the means to **identify** and **communicate** with the units in the survey population; it takes the form of a list, which is linked to the population under study;
- **Intended sample:** subset of the study population targeted by the survey;
- **Achieved sample:** subset of the study population whose characteristics were in fact measured.

4: The difference may be due to the **difficulty/high cost** of data collection for some units excluded from the study population.

In general, a survey is preferred to a census if it is **expensive/laborious** to measure the characteristics of interest for each unit, or if the units are **destroyed** by measuring the characteristics.

### Deciding Factors

In some instances, information about the **entire** population is required in order to solve the client's problem, whereas in others it is not necessary. How do we determine which type of survey must be conducted to collect data? The answer depends on multiple factors:

- the type of question that needs to be answered;
- the required precision;
- the cost of surveying a unit;
- the time required to survey a unit;
- size of the population under investigation, and
- the prevalence of the attributes of interest.

Once a choice has been made, each survey typically follows the same **general steps**:

1. statement of objective
2. selection of survey frame
3. sampling design
4. questionnaire design
5. data collection
6. data capture and coding
7. data processing and imputation
8. estimation
9. data analysis
10. dissemination and documentation

The process is not always linear, in that preliminary planning and data collection may guide the implementation (selection of a frame and of a sampling design, questionnaire design), but there is a definite movement from objective to dissemination.[5]

5: Compare with Figure 14.4, Section 14.4.1.

### 10.1.2 Survey Frames

The **frame** provides the means of **identifying** and **contacting** the units of the study population. It is generally costly to create and to maintain (in fact, there are organisations and companies that specialize in building and/or selling such frames).

Useful frames contain:

- identification data,
- contact data,
- classification data,
- maintenance data, and
- linkage data.

The ideal frame must minimize the risk of **undercoverage** or **overcoverage**, as well as the number of **duplications** and **misclassifications** (although some issues that arise can be fixed at the data processing stage).

Unless the selected frame is **relevant** (which is to say, it corresponds, and permits accessibility to, the target population), **accurate** (the information it contains is valid), **timely** (it is up-to-date), and **competitively priced**, the statistical sampling approach is contra-indicated.

### 10.1.3 Fundamental Sampling Concepts

In general, a survey is conducted to **estimate certain attributes of a population** (statistics), such as, for example

- a **mean**;
- a **total**, or
- a **proportion**.

A **population** (either target, study, or respondent) has a finite number $N$ of members, called **units** or **items**. The **response** associated with the $j-$th unit of the population is represented by $u_j$.

Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be a population of size $N < \infty$. If $u_j$ represents a numerical variable,[6] the **mean**, **variance**, and **total** of the **response** in the population are respectively

6: E.g., if $u_j$ is the salary of the $j-$th unit in the population.

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j, \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2, \quad \text{and} \quad \tau = \sum_{j=1}^{N} u_j = N\mu.$$

If $u_j$ represents a **binary variable**,[7] the **proportion** of the **response** in the population is

7: E.g., 1 if the $j-$th unit earns more than \$70$K$ per year, 0 otherwise.

$$p = \frac{1}{N}\sum_{j=1}^{N} u_j.$$

We seek to estimate $\mu$, $\tau$, $\sigma^2$ and/or $p$ using the values of the response variable for the units in the achieved sample $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U}$. The relationship between $\mathcal{Y}$ and $\mathcal{U}$ is simple: in general, $n \ll N$ and $\forall i \in \{1, \ldots, n\}, \exists! j \in \{1, \ldots, N\}$ such that $y_i = u_j$.

The **empirical mean**, **empirical total**, and **empirical variance** are:

$$\overline{y}(, \hat{p}) = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \hat{\tau} = \left(\frac{N}{n}\right)\sum_{i=1}^{n} y_i = N\overline{y}.$$

Let $X_1, \ldots, X_n$ be random variables, $b_1, \ldots, b_n \in \mathbb{R}$, and E, V, and Cov be the **expectation**, **variance** and **covariance** operators. Recall that

$$\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \mathrm{E}(X_i), \quad \mathrm{V}(X_i) = \mathrm{Cov}(X_i, X_i) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2(X_i)$$

$$\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{V}(X_i) + \sum_{1 \le i \ne j}^{n} b_i b_j \mathrm{Cov}(X_i, X_j)$$

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j).$$

The **bias** in an error component is the average of that error component if the survey is repeated many times independently under the same conditions. The **variability** in an error component is the extent to which that component would vary about its average value in this scenario.

The **mean square error** of an error component is a measure of the size of the error component:

$$\begin{aligned}\mathrm{MSE}(\hat{\beta}) &= \mathrm{E}\left((\hat{\beta} - \beta)^2\right) = \mathrm{E}\left((\hat{\beta} - \mathrm{E}(\hat{\beta}) + \mathrm{E}(\hat{\beta}) - \beta)^2\right) \\ &= \mathrm{V}(\hat{\beta}) + \left(\mathrm{E}(\hat{\beta}) - \beta\right)^2 = \mathrm{V}(\hat{\beta}) + \mathrm{Bias}^2(\hat{\beta})\end{aligned}$$

where $\hat{\beta}$ is an estimate of $\beta$. Finally, if the estimate is **unbiased**, then an approximate **95% confidence interval** (95% C.I.) for $\beta$ is given by

$$\hat{\beta} \pm 2\sqrt{\hat{\mathrm{V}}(\hat{\beta})},$$

where $\hat{V}(\hat{\beta})$ is a **sampling design-specific** estimate of $V(\hat{\beta})$.

### Survey Error

One of the strengths of statistical sampling is in its ability to provide estimates of various quantities of interest in the target population, and to provide some control over the **total error** (TE) of the estimates. The TE of an estimate is the amount by which it **differs from the true value** for the target population:

$$\text{Total Error} = \text{Measurement Error} + \text{Sampling Error} + \text{Non-response Error} + \text{Coverage Error},$$

where the:

- **coverage error** is due to differences in the study and target populations;
- **non-response error** is due to differences in the respondent and study populations;
- **sampling error** is due to differences in the achieved sample and the respondent population;
- **measurement error** is due to true value in the achieved sample not being assessed correctly.[8]

8: We sometimes also include the **processing error** in this component, due to the fact that the real value of the characteristic of interest can be affected by the data transformations performed throughout the analysis.

If we let:

- $\overline{x}$ be the computed attribute value in the achieved sample;
- $\overline{x}_{\text{true}}$ be the true attribute value in the achieved sample under perfect measurement;
- $x_{\text{resp}}$ be the attribute value in the respondent population;
- $x_{\text{study}}$ be the attribute value in the study population, and
- $x_{\text{target}}$ be the attribute value in the target population,

then

$$\underbrace{\overline{x} - x_{\text{target}}}_{\text{total error (TE)}} = \underbrace{(\overline{x} - \overline{x}_{\text{true}})}_{\text{meas. \& proc. error}} + \underbrace{(\overline{x}_{\text{true}} - x_{\text{resp}})}_{\text{sampling error}} + \underbrace{(x_{\text{resp}} - x_{\text{study}})}_{\text{non-response error}} + \underbrace{(x_{\text{study}} - x_{\text{target}})}_{\text{coverage error}}.$$

In an ideal scenario, TE = 0. In practice, there are two main contributions to Total Error: **sampling errors** (which are this module's main concern) and **nonsampling errors**, which include every contribution to survey error which is not due to the choice of sampling scheme.

The latter can be controlled, to some extent:

- **coverage error** can be minimized by selecting a high quality, up-to-date survey frame;
- **non-response error** can be minimized by careful choice of the data collection mode and questionnaire design, and by using "call-backs" and "follow-ups";
- **measurement error** can be minimized by careful questionnaire design, pre-testing of the measurement apparatus, and cross-validation of answers.

These suggestions are perhaps less useful than one could hope in modern times: survey frames based on landline telephones are quickly becoming irrelevant in light of an increasingly large and younger population who eschew such phones, for instance, while response rates for surveys that are not mandated by law are surprisingly low.[9]

9: This explains, in part, the impetus towards **automated data collection** and the use of **non-probabilistic sampling** methods.

### 10.1.4 Data Collection Basics

How is data traditionally captured, then? There are **paper-based** approaches, **computer-assisted** approaches, and a suite of other modes.

- **Self-administered questionnaires** are used when the survey requires detailed information to allow the units to consult personal records (which reduces measurement errors), they are useful to measure responses to sensitive issues as they provide an extra layer of privacy, and are typically not as costly as other collection modes, but they tend to be associated with high non-response rate since there is less pressure to respond.
- **Interviewer-assisted questionnaires** use trained interviewers to increase the response rate and overall quality of the data. Face-to-face **personal interviews** achieve the highest response rates, but they are costly (both in training and in salaries). Furthermore, the interviewer may be required to visit any selected respondents many times before contact is established. **Telephone interviews**, on the other hand produce "reasonable" response rates at a reasonable cost and they are safer for the interviewers, but they are limited in length due to respondent phone fatigue. With random dialing, 4-6 minutes of the interviewer's time is spent in out-of-scope numbers for each completed interview.
- **Computer-assisted interviews** combine data collection and data capture, which saves valuable time, but the drawback is that not every sampling unit may have access to a computer/data recorder (although this is becomine less prevalent). All paper-based modes have a computer-assisted equivalent: **computer-assisted self-interview** (CASI), **computer-assisted interview** (CAI), **computer-assisted telephone interview** (CATI), and **computer-assisted personal interview** (CAPI).
- Other approaches include unobtrusive direct observation; diaries to be filled (paper or electronic); omnibus surveys; email, Internet (e.g., Survey Monkey ), social media, etc.

### 10.1.5 Types of Sampling Methods

There is a large variety of methods to select sampling units from the target population.

#### Non-Probabilistic Sampling

Those that use subjective, non-random approaches are called **non-probabilistic sampling** (NPS) methods; these tend to be **quick**, **relatively inexpensive** and **convenient** in that a survey frame is not needed.

NPS methods are ideal for **exploratory analysis** and **survey development**. Unfortunately, they are sometimes used **instead** of probabilistic sampling designs, which is problematic; the associated selection bias makes NPS methods **unsound** when it comes to **inferences**, as they cannot be used to provide **reliable estimates of the sampling error**.[10]

10: The only component of the total error TE on which the analysts has direct control.

Automated data collection often fall squarely in the NPS camp, for instance. While we can still analyse data collected with a NPS approach, we **may not generalize the results** to the target population (except in rare, census-like situations).

NPS methods include:

- **haphazard** sampling, also known as "person on the street" sampling; it assumes that the population is homogeneous, but the selection remains subject to interviewer biases and the availability of units;
- **volunteer** sampling in which the respondents are self-selected; there is a large selection bias since the silent majority does not usually volunteer; this method is often imposed upon analysts due to ethical considerations; it is also used for focus groups or qualitative testing;
- **judgement** sampling is based on the analysts' ideas of the target population composition and behaviour (sometimes using a prior study); the units are selected by population experts, but inaccurate preconceptions can introduce large biases in the study;
- **quota** sampling is very common (and is used in exit polling to this day in spite of the infamous "Dewey Defeats Truman" debacle of 1948 [2]); sampling continues until a specific number of units have been selected for various sub-populations; it is preferable to other NPS methods because of inclusion of sub-populations, but it ignores non-response bias;
- **modified** sampling starts out using probability sampling (more on this later), but turns to quota sampling in its last stage, in part as a reaction to high non-response rates;
- **snowball** sampling asks sampled units to recruit other units among their acquaintances; this NPS approach may help locate hidden populations, but it biased in favour of units with larger social circles and units that are charming enough to convince their acquaintances to participate.


**Figure 10.2:** Dewey vs Truman – the aftermath: Truman victorious!

There are contexts where NPS methods might fit a client's need (and that remains their decision to make, ultimately), but the analyst MUST still inform the client of the drawbacks, and present some probabilistic alternatives.

### Probabilistic Sampling

The inability to make sound inferences in NPS contexts is a monumental strike against their use. While probabilistic sample designs are usually **more difficult and expensive** to set-up (due to the need for a quality survey frame), and take **longer** to complete, they provide **reliable estimates** for the attribute of interest and the sampling error, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

In this chapter, we take a deeper look at the traditional probability sample designs:

- **simple random sampling** (SRS), see Section 10.3;
- **stratified random sampling** (STS), see Section 10.4;
- **systematic random sampling** (SyS), see Section 10.7.1;
- **cluster random sampling** (CLS), see Section 10.6;
- **sampling with probability proportional to size** (PPS), see Section 10.7.2, and
- more advanced designs, see Section 10.7.

In this chapter, the analysis is made easier by assuming that the sampling error dominates the survey error, i.e., that:

- the study population is **representative** of the target population ($x_{\text{study}} \approx x_{\text{target}}$);
- the respondent population and the study population **coincide**, as are the achieved sample and the target sample ($x_{\text{resp}} \approx x_{\text{study}}$), and

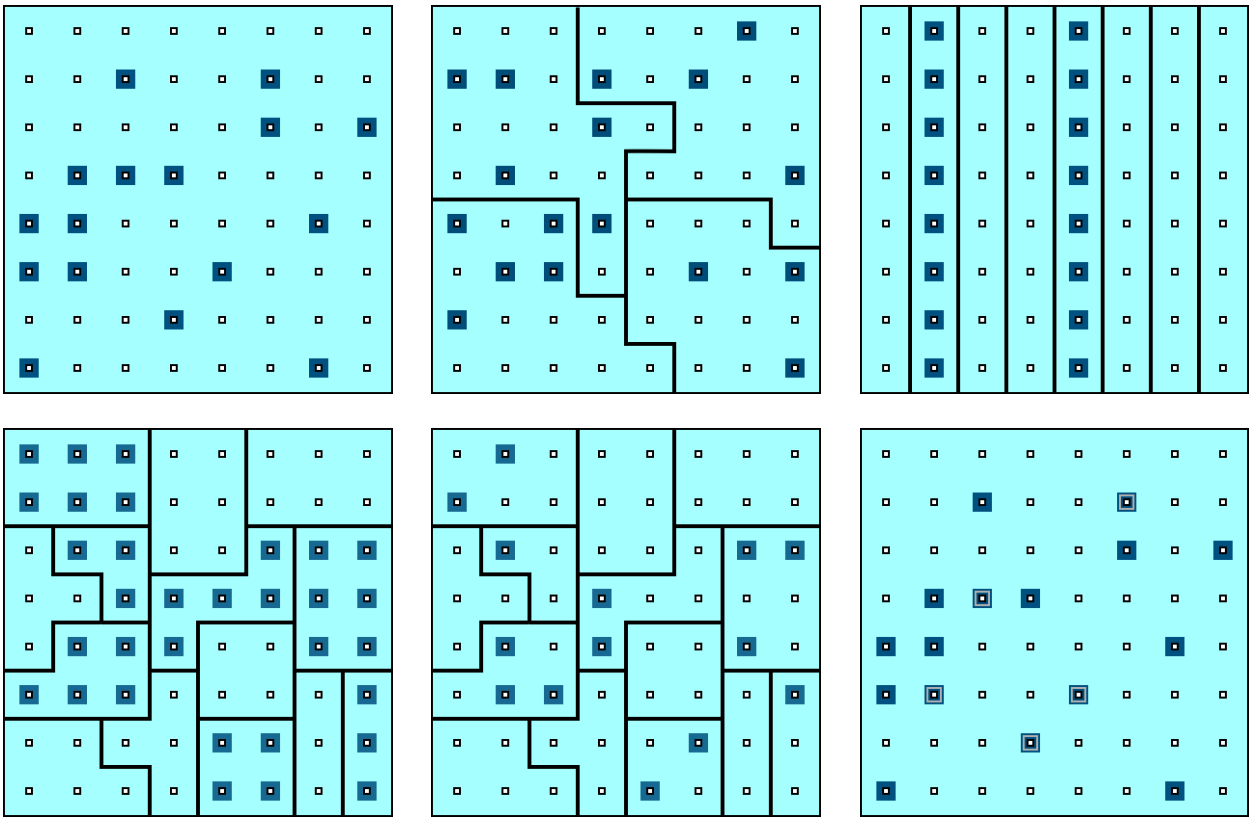


**Figure 10.3:** Schematics of various sampling designs (from left to right, top to bottom): simple random sampling, stratified sampling, systematic sampling, cluster sampling, multi-stage sampling, multi-phase sampling.

- the response is measured without error in the achieved sample ($\overline{x} \approx \overline{x}_{\text{true}}$).

The objective is to **control and evaluate the sampling error** ($\overline{x}_{\text{true}} - \overline{x}_{\text{resp}}$) for various random sampling designs.

## 10.2 Questionnaire Design

> People resist a census, but give them a profile page and they'll spend all day telling you who they are [1].

A **questionnaire** is a series of questions designed to **obtain information on a topic** from respondents. Of course, design principles vary depending on the **subject** and **method of data collection**, but it is considered good practice to test various questionnaires on **random pilot populations** before rolling it out on the study population.

### 10.2.1 Basic Concepts

In general, a questionnaire should:

- be as brief as possible, and free of unnecessary questions;
- be accompanied by clear and concise instructions;
- keep the respondent's interests in mind;
- emphasize confidentiality;

- keep a serious and courteous tone;
- be error-free and attractively presented;
- be clearly and precisely worded;
- be designed so that it can be answered accurately, and
- neatly arranged.

The quality of the collected data depends to a large extent on the quality of the questionnaire – **this is a practical aspect of the discipline on which much more time should be spent than on data analysis**; reputable survey firms employ **specialized teams** for questionnaire design.

There is an added challenge for Government of Canada (GoC) federal departments that are collecting and reporting information about the public and representatives of businesses or other entities, including federal public servants: see Public opinion research in the Government of Canada for details. Some of the information presented in this section will overlap with the POR guidelines, but at other times, our (generic) advice will differ.

When working with the GoC, the POR guidelines must obviously take precedence.[11]

11: Fancy footwork might be required to overcome the challenges presented by the guidelines, but that is par for the course.

### 10.2.2 Question Types

The basic unit of the questionnaire is, of course, the **question**, which comes in two forms:

- **closed** questions, with a fixed number of predetermined, mutually exclusive, and collectively exhaustive answer choices (and which should always include an "Other (please specify)" category to counteract loss of expressiveness), and
- **open** questions, which are used primarily to identify common response choices for use in closed-ended questions in a subsequent questionnaire; any closed-ended question should have been an open-ended question at some point.

In everyday conversation, closed-ended questions are not appropriate:

> Asking open-ended questions is a friendly way to approach others in discussions. Knowing the difference between open and closed questions will be invaluable in your career and social life. How to ask open-ended questions, *WikiHow*

In a survey, it is rather open-ended questions that are not appropriate: closed-ended questions require less **effort** on the part of respondents, and they are generally **easier to quantify**, allowing more questions to be asked in a restricted **amount of time** and for a given **budget**.

For example, compare the two following questions.

**Open question:** What is the most important issue facing Ontario in 2022?

**Closed-ended question:** Which of these is the most important challenge for Ontario in 2022?

- economy and unemployment

- impact of COVID-19
- reconciliation with indigenous communities
- taxes
- budget deficit
- the environment
- organized crime
- gang violence
- racism
- other (please specify)

However, closed-ended questions can also lead to:

- a **loss of an opportunity to test the waters** in order to obtain further clarification;
- **introducing response bias** by presenting alternatives that respondents would never have thought of, and
- a potential **loss of interest** if the choice of answers does not match a respondents' expectations.

Adding open-ended questions to the questionnaire can mitigate these risks. The use of text analysis and natural language processing methods can also help to extract the main meaning or sentiments of an answer to an open-ended question.[12]

12: See Chapters 27 and 32 for details and for limitations of such approaches.

### 10.2.3 Design Considerations

It is well known that the **formulation of questions** can influence the responses of a questionnaire; it is good idea to keep the following **wording considerations** in mind when developing questionnaires:

- Avoid **abbreviations** and **jargon**: "Does your organization use TTWQ practices?"
- Avoid using **complex terms** when **simpler terms** will do: "How many times have you been defenestrated?" vs. "How many times have you been thrown out a window?"
- Ensure that all respondents can answer the questions, by asking **relevant** and **appropriate-level** questions;
- Clarify the **framework**: "What is your annual income?" vs. "What was your total household income from all sources, before taxes and deductions, in 2021?
- Make the question as **accurate** as possible: "How much fuel did your moving company use last year?" (answers received: 2,500 liters, 800 gallons, $13500, more than the previous year, etc.) vs. "How much did your moving company spend on fuel last year?"
- Avoid "**double-barreled**" questions: "Do you plan to leave your car at home and take LRT to work?" vs. "Do you plan to leave your car at home? If so, do you plan to take LRT to work?", and
- Avoid **leading questions**: the always excellent *Yes, Prime Minister* gives a clear-cut example: [13] Sir Humphrey demonstrates that asking leading questions in a particular order can lead a respondent to support the reintroduction of national service:
  - Are you concerned about the number of unemployed youth?
  - Are you concerned about the increase in teenage crime?
  - Do you think there is a lack of discipline in our schools?

13: Which is not nearly as facetious as it appears, in the final analysis.

Yes, Prime Minister | S04xE02 | Leading Questions | *The Ministerial Broadcast*

- Do you think young people would appreciate some leadership?
- Do you think they would respond to a challenge?
- Would you support the re-introduction of national service in the UK?

The first five questions are designed and presented in such a way as to elicit support – the obvious answer to each is "yes". After this pattern of agreement, Sir Humphrey launches the crucial question, framed in such a way that it proposes national service as a supposed solution to all the above problems. In the second part of the exchange, Sir Humphrey demonstrates that another set of leading questions can lead the respondent to oppose the reintroduction of national service:

- Does the danger presented by war worry you?
- Does the arms race worry you?
- Do you think it is dangerous to arm young people and teach them to kill?
- Is it bad to force people to take up arms against their will?
- Would you oppose the reintroduction of national service?

Sir Humphrey's first four questions are deliberately designed to produce agreement. In keeping with the survey design, the fifth question does the same: a person who answers "yes" to each of these questions is necessarily opposed to the reintroduction of national service.[14]

14: Based on an idea by Nagesh Belludi.

### 10.2.4 Question Order

The **order** in which the questions are presented is as important as their wording. Questionnaires should be designed to be **seamless** and **follow a logical process**, from the perspective of the respondents:[15]

1. begin with an **introduction** that provides the title, topic and purpose of the survey;
2. ask for **cooperation** from respondents and explain the importance of the survey and how the results will be used;
3. indicate the degree of **confidentiality** and provide a deadline and contact address;
4. follow up with a series of **easy** and **interesting** questions to build respondent confidence;
5. group similar questions under the same heading;
6. only introduce **sensitive topics** when a relationship of trust is likely to have been established with the respondents;
7. leave some space and/or time for **additional comments**, and
8. **thank** respondents for their participation.

It is worth remembering that without a "sound sampling plan", collected data may be of such poor quality that it is impossible to use it to draw any meaningful conclusions. It is also essential to capture **demographic information** that allows classification of units into **stratas** (STS) or **clusters** (CLS); we will revisit those concepts in subsequent sections.

**Example:** Consider the following video.

15: Questionnaire design is discussed in the following references:

- Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing Nonresponse in Surveys," *Survey Methodology*, v.19, n.1, pp.81-94 [4]
- Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2, pp.125-136 [3]
- *Survey Methods and Practices*, Statistics Canada, catalogue number 12-587-X [10]


**Figure 10.4:** 2021 Census – How do I complete the questionnaire?

**Transcription of the video**

In May, your household will receive a letter to complete the 2021 Census questionnaire. On your letter, you will find a secure access code that allows you to complete the questionnaire online. Once online, you can complete the questionnaire in three easy steps. Simply log on using your secure access code, complete the questionnaire and select "Submit." If you need help or require a paper version, please call the Census Help Line. For more information or to complete the 2021 Census questionnaire, visit census.gc.ca . It's safe, quick and easy.

**Message from the Chief Statistician of Canada**

Thank you for taking a few minutes to participate in the 2021 Census. The information you provide is converted into statistics used by communities, businesses and governments to plan services and make informed decisions about employment, education, health care, market development and more. Your answers are collected under the authority of the Statistics Act and kept strictly confidential. By law, every household must complete a 2021 Census of Population questionnaire. Statistics Canada makes use of existing sources of information such as immigration, income tax and benefits data to ensure the least amount of burden is placed on households. The information that you provide may be used by Statistics Canada for other statistical and research purposes or may be combined with other survey or administrative data sources. Make sure you count yourself into Canada's statistical portrait, and **complete your census questionnaire today.**

Thank you,

Anil Arora
Chief Statistician of Canada

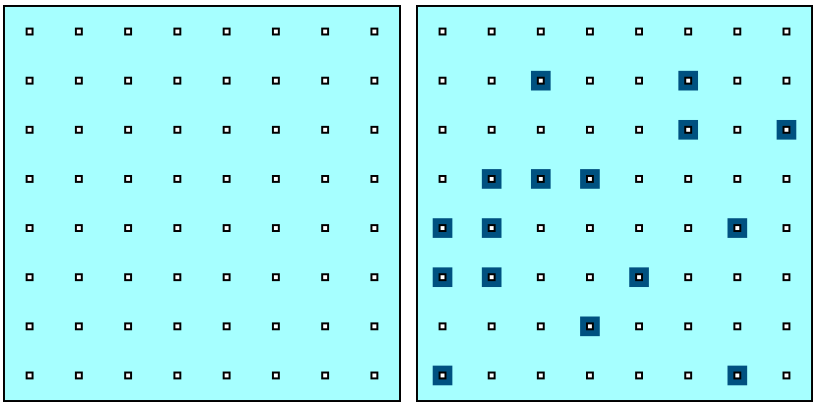


**Figure 10.5:** Schematics of SRS: target population (left) and sample (right).

## 10.3 Simple Random Sampling

Let $\mathcal{U}$ be a population composed of $N$ units, whose responses are

$$\mathcal{U} = \{u_1, \ldots, u_N\}.$$

Suppose we are interested in the **mean** $\mu$ of this target population $\mathcal{U}$, where

$$\mu = \frac{1}{N} \sum_{j=1}^{N} u_j.$$

Since the population is of finite size, it is possible to compute $\mu$ directly... at least, in theory. In practice, we rarely have access to the response values for the entire population $\mathcal{U}$, which leads us to use **sampling methods**.

A **sample** $\mathcal{Y}$ of size $n$ is a subset of the target population $\mathcal{U}$,

$$\mathcal{Y} \subseteq \{y_1, \ldots, y_n\} \subseteq \{u_1, \ldots, u_N\} = \mathcal{U},$$

from which we can approximate $\mu$ using the **sample mean**[16]

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

16: This is not the only estimator of $\mu$.

A **simple random sample** (SRS) of size $n$ is obtained by randomly selecting $n$ units from the target population, **one at a time, without replacement**. In Figure 10.5, a SRS of size $n = 16$ is selected from a population of size $N = 64$.

At each stage of the sampling procedure, all units not yet in the sample have the same probability of being added to the sample. In an SRS, each subset of $n$ units **has the same probability of being selected**.

How do we choose a **random** sample?

This used to be done "by hand", using tables of random numbers. Nowadays, we simply use software (SAS, R, etc.) to obtain **(pseudo-)**random samples.

**Example** What is the average life span, by country, in 2011?

We use the data available in the `Gapminder` dataset.

```
library(tidyverse) # for dplyr, ggplot2
gapminder = read.csv("gapminder_SS.csv",
                     stringsAsFactors=TRUE)
gapminder <- gapminder[,c("country","year","region",
                          "continent","population",
                          "infant_mortality","fertility",
                          "gdp","life_expectancy")]
```

The structure is provided below:

```
str(gapminder)
```

```
'data.frame':   10545 obs. of  9 variables:
 $ country         : Factor w/ 185 levels "Albania","Algeria",..: 1 2 3 4 5 6 7 8 9 ...
 $ year            : int  1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...
 $ region          : Factor w/ 22 levels "Australia and New Zealand",..: 19 11 10 2 ...
 $ continent       : Factor w/ 5 levels "Africa","Americas",..: 4 1 1 2 2 3 2 5 4 3 ...
 $ population      : int  1636054 11124892 5270844 54681 20619075 1867396 54208 ...
 $ infant_mortality: num  115.4 148.2 208 NA 59.9 ...
 $ fertility       : num  6.19 7.65 7.32 4.43 3.11 4.55 4.82 3.45 2.7 5.57 ...
 $ gdp             : num  NA 1.38e+10 NA NA 1.08e+11 ...
 $ life_expectancy : num  62.9 47.5 36 63 65.4 ...
```

A famous chart displays the relationship between 4 of the variables [9]. Our version for 2011 (built with R) can be found in Figure 10.6.

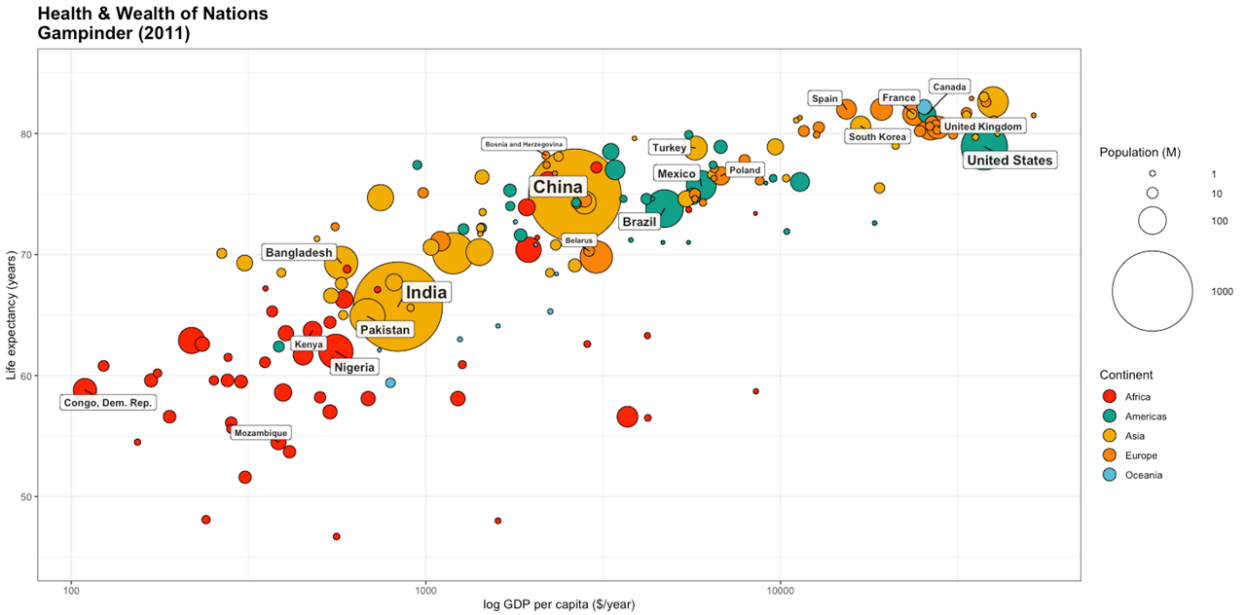



**Figure 10.6:** Health and wealth of nations for the 2011 Gapminder data.

We start by extracting the information of interest.

```
gapminder.SRS <- gapminder |>
  filter(year==2011) |>
  select(life_expectancy)
str(gapminder.SRS)
```

```
'data.frame':   185 obs. of  1 variable:
 $ life_expectancy: num  77.4 76.1 58.1 75.9 76 ...
```

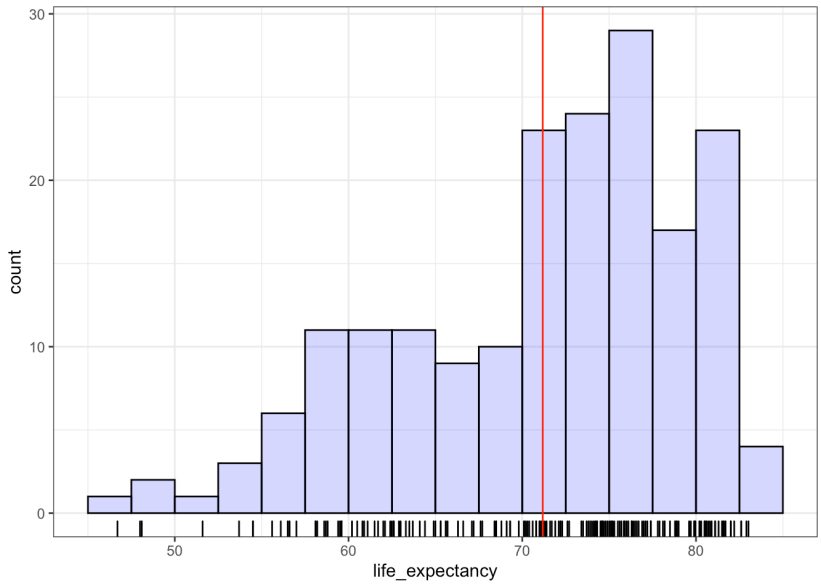
In this specific example, we know the true average life expectancy per country in 2011 (at least, for the $N = 185$ countries in the dataset).

```
mean(gapminder.SRS)
```

```
[1] 71.18
```

The distribution of the population $\mathcal{U} = \{u_1, \ldots, u_{185}\}$ is shown below (with mean in red):

```
ggplot(data=gapminder.SRS, aes(life_expectancy)) +
  geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2,
                 breaks=seq(45, 85, by = 2.5)) +
  geom_vline(xintercept=mean(gapminder.SRS$life_expectancy),
             color="red")
```




We select a random sample of size $n = 10$ from $\mathcal{U}$. The indices are:

```
set.seed(1234) # for replicability
N = dim(gapminder.SRS)[1]
n = 10
(sample.ind = sample(1:N,n, replace=FALSE))
```

```
[1]  28  80 150 101 111 137 133 166 144 132
```

The corresponding sample $\mathcal{Y} = \{y_1, \ldots, y_{10}\}$ is obtained *via*:

```
(gapminder.SRS.n = gapminder.SRS[sample.ind,])
```

```
[1] 67.60 67.70 76.10 79.97 75.70 79.70 70.20 59.60 78.90 78.50
```

Its empirical mean $\overline{y}$ is:

```
(y.bar = mean(gapminder.SRS.n))
```

```
[1] 73.397
```

But a different sample may lead to a different estimate. Case in point, consider the following:

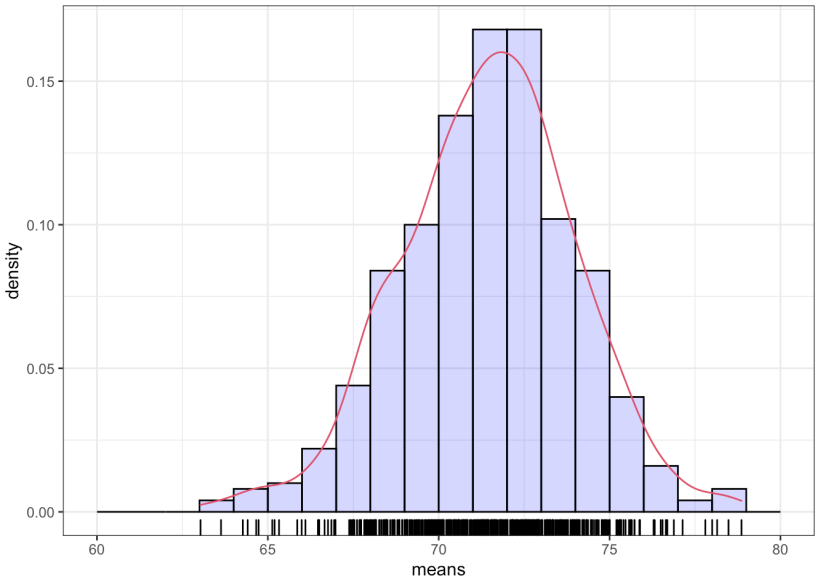
```
set.seed(12345) # replicability
(sample.ind = sample(1:N,n, replace=FALSE))
(gapminder.SRS.n = gapminder.SRS[sample.ind,])
(y.bar = mean(gapminder.SRS.n))
```

```
[1] 142  51 152  58  93  75  96   2  86 180
[1] 71.0 74.3 63.0 81.6 65.0 75.0 46.7 76.1 78.1 74.8
[1] 70.56
```

It is quite reasonable for the two estimates to be different – since each $y_i$ in a SRS is a **random variable**, so is the mean $\overline{y}$.

The **sampling variability** explains how the estimates vary with the sample. For example, if we prepare $m = 500$ samples, each of size $n = 10$, we could obtain the empirical means below:

```
set.seed(12) # for replicability
N=dim(gapminder.SRS)[1]
n=10
m=500
means <- c()
for(k in 1:m){
 means[k] <- mean(gapminder.SRS[sample(1:N,n,
                                       replace=FALSE),])
 }

ggplot(data=data.frame(means), aes(means)) +
    geom_histogram(aes(y =..density..),
                   breaks=seq(60, 80, by = 1),
                   col="black", fill="blue", alpha=.2) +
    geom_density(col=2) + geom_rug(aes(means))
```




There is some variability, of course, but the sample means seem to congregate around the 72 mark:

```
summary(data.frame(means))
```

```
means
  Min. :63.03
1st Qu.:69.83
Median :71.53
  Mean :71.44
3rd Qu.:73.05
  Max. :78.86
```

### 10.3.1 Basic Notions

The **population variance** $\sigma^2$ is a measure of **dispersion**, i.e., the tendency of the response values to deviate from the **population mean** $\mu$:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{j=1}^{N}(u_j-\mu)^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j^2-2u_j\mu+\mu^2)\\ &= \frac{1}{N}\left(\sum_{j=1}^{N}u_j^2-2\mu\sum_{j=1}^{N}u_j+N\mu^2\right) = \frac{1}{N}\left(\sum_{j=1}^{N}u_j^2-2N\mu^2+N\mu^2\right)\\ &= \frac{1}{N}\sum_{j=1}^{N}\left(u_j^2-N\mu^2\right) = \frac{1}{N}\sum_{j=1}^{N}u_j^2-\mu^2\end{aligned}$$

The parameters $\mu$ and $\sigma^2$ can be interpreted in terms of the **expectation** and **variance** of a random variable.

Let $X$ be a discrete random variable whose **probability mass function** (p.m.f.) is $f(x) = P(X = x)$. Thus,

$$\mathrm{E}[X] = \sum_x xf(x), \quad \mathrm{V}[X] = \sum_x (x-\mathrm{E}[X])^2 f(x), \quad \mathrm{SD}[X] = \sqrt{\mathrm{V}[X]}.$$

For a sample of size $n = 1$ from this population, whose value is represented by the random variable $Y_1$, we have $f(u_j) = P(Y_1 = u_j) = \frac{1}{N}$ for $j = 1, \ldots, N$, from which we see that

$$\mathrm{E}[Y_1] = \sum_{j=1}^{N} u_j f(u_j) = \frac{1}{N} \sum_{j=1}^{N} u_j = \mu,$$

and

$$\mathrm{V}[Y_1] = \sum_{j=1}^{N} (u_j - \mu)^2 f(u_j) = \frac{1}{N} \sum_{j=1}^{N} u_j^2 - \mu^2 = \sigma^2, \quad \mathrm{SD}[Y_1] = \sqrt{\mathrm{V}[Y_1]} = \sigma.$$

In general, however, the estimator $\overline{y}$ of the population mean $\mu$ is computed using **more than one observation** – different sample sizes $n$ could yield different values of $\overline{y}$. In order to control the sampling error associated with an SRS, one needs to know the **distribution of** $\overline{Y}$; in particular, $\mathrm{E}[\overline{Y}]$ and $\mathrm{V}[\overline{Y}]$.

If $y_1, \ldots, y_n$ are **independent and identically distributed** (i.i.d.) random variables, the **central limit theorem** (CLT) imposes

$$\overline{Y} \sim_{\text{approx.}} \mathcal{N}(\mu, \sigma^2/n).$$

**Example** Consider a finite population with $N = 4$ elements:

$$u_1 = 2, \quad u_2 = 0, \quad u_3 = 1, \quad u_4 = 5.$$

The population mean and variance are, respectively,

$$\mu = \frac{1}{4}(2 + 0 + 1 + 5) = 2 \quad \text{and} \quad \sigma^2 = \frac{1}{4}(2^2 + 0^2 + 1^2 + 5^2) - 2^2 = \frac{7}{2}.$$

Suppose that draw a SRS of size $n = 3$ without replacement from this population in order to approximate (estimate) the true mean $\mu$. There are $\binom{4}{3} = 4$ such samples:

| **Sample** | **Values** | $\overline{y}$ | $P(\overline{Y} = \overline{y})$ |
|---|---|---|---|
| $u_1, u_2, u_3$ | $2, 0, 1$ | $1$ | $1/4$ |
| $u_1, u_2, u_4$ | $2, 0, 5$ | $7/3$ | $1/4$ |
| $u_1, u_3, u_4$ | $2, 1, 5$ | $8/3$ | $1/4$ |
| $u_2, u_3, u_4$ | $0, 1, 5$ | $2$ | $1/4$ |

Then

$$\mathrm{E}[\overline{Y}] = \sum_{\overline{y}} \overline{y} P(\overline{Y} = \overline{y}) = \tfrac{1}{4}\left(1 + \tfrac{7}{3} + \tfrac{8}{3} + 2\right) = 2 = \mu$$

$$\mathrm{V}[\overline{Y}] = \sum_{\overline{y}} \overline{y}^2 P(\overline{Y} = \overline{y}) - \mathrm{E}^2[\overline{Y}] = \tfrac{1}{4}\left(1^2 + \left(\tfrac{7}{3}\right)^2 + \left(\tfrac{8}{3}\right)^2 + 2^2\right) - 2^2 = \tfrac{7}{18}.$$

This is all great... except that $\mathrm{V}[\overline{Y}] \neq \frac{\sigma^2}{n} = \frac{7}{6}$. What is going on? ■

Here's how we can explain this discrepancy. Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be a finite population of size $N$. A SRS $\mathcal{Y} = \{y_1, \ldots, y_n\}$ of size $n$ is drawn

from $\mathcal{U}$ without replacement. Let $Y_i$ be the random variable which represents the value of the $i-$th unit of the sample, respectively.

All $Y_i$ have **identical distributions**: for any $u_j \in \mathcal{U}$, we have:[17]

17: Be careful not to confuse the unit $u_j$ with its response value $u_j$; we use the same notation by laziness, but they represent different concepts.

$$P(Y_1 = u_j) = \frac{1}{N},$$

$$P(Y_2 = u_j) = \frac{P(Y_2 = u_j \mid Y_1 \neq u_j) \cdot P(Y_1 \neq u_j)}{P(Y_1 \neq u_j \mid Y_2 = u_j)} = \frac{\frac{1}{N-1} \cdot \frac{N-1}{N}}{1} = \frac{1}{N},$$

$$P(Y_3 = u_j) = \frac{P(Y_3 = u_j \mid Y_1, Y_2 \neq u_j) \cdot P(Y_1, Y_2 \neq u_j)}{P(Y_1, Y_2 \neq u_j \mid Y_3 = u_j)} = \frac{\frac{1}{N-2} \cdot \frac{N-2}{N-1} \cdot \frac{N-1}{N}}{1} = \frac{1}{N},$$

and so on:

$$P(Y_i = u_j) = \frac{1}{N}$$

for any $1 \leq i \leq n, 1 \leq j \leq N$, and so $\mathrm{E}[Y_i] = \mu$, $\mathrm{V}[Y_i] = \sigma^2$ for any $i$.

Thus, in the preceding example, we would have

$$\mathrm{E}[Y_1] = \mathrm{E}[Y_2] = \mathrm{E}[Y_3] = \mu = 2 \quad \text{and} \quad \mathrm{V}[Y_1] = \mathrm{V}[Y_2] = \mathrm{V}[Y_3] = \sigma^2 = \frac{7}{2}.$$

But the $\{Y_i\}$ are **not independent** of each other since (for example)

$$\mathrm{E}[\overline{Y}] = \mu = 2, \quad \text{but} \quad \mathrm{V}[\overline{Y}] = \mathrm{V}\left[\tfrac{Y_1+Y_2+Y_3}{3}\right] = \frac{7}{18} \neq \frac{\sigma^2}{3} = \frac{7/2}{3} = \frac{7}{6}.$$

It is in the variance that we observe a difference. The **covariance** between two (discrete) random variables $X_1, X_2$ is a **measure of the strength of association between $X_1$ and $X_2$**. If $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \sigma_i^2 < \infty$ for all $i$, then

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathrm{E}[X_1 X_2] - \mu_1 \mu_2.$$

If $X_1, X_2$ both take values in $\mathcal{U} = \{u_1, \ldots, u_N\}$, then their **joint expectation** is

$$\mathrm{E}[X_1 X_2] = \sum_{j=1}^{N} \sum_{k=1}^{N} u_j u_k P(X_1 = u_j, X_2 = u_k).$$

In the case where $X_1 = Y_i$ and $X_2 = Y_\ell$ (with the interpretation given before) for $1 \leq i \neq \ell \leq n$, we get

$$P(Y_i = u_j, Y_\ell = u_k) = P(Y_i = u_j) P(Y_\ell = u_k \mid Y_i = u_j) = \begin{cases} \frac{1}{N} \cdot \frac{1}{N-1} & \text{if } j \neq k \\ 0 & \text{if } j = k \end{cases}$$

But $\mathrm{E}[Y_i] = \mathrm{E}[Y_\ell] = \mu$, and so

$$\mathrm{Cov}(Y_i, Y_\ell) = \begin{cases} \frac{1}{N(N-1)} \Big[ \sum_{j=1}^{N} \sum_{k=1}^{N} u_j u_k - \underbrace{\sum_{m=1}^{N} u_m^2}_{\text{doublecounting}} \Big] - \mu^2 & \text{if } i \neq \ell \\ \sigma^2 & \text{if } i = \ell \text{ (by convention)} \end{cases}$$

We use the properties $\sum u_\xi = N\mu$ and $\sum u_\xi^2 = N(\mu^2 + \sigma^2)$ to simplify the

expression when $i =\neq \ell$:

$$\begin{aligned}
\mathrm{Cov}(Y_i, Y_\ell) &= \frac{1}{N(N-1)}\Big[\sum_{j=1}^{N}\sum_{k=1}^{N} u_j u_k - \sum_{m=1}^{N} u_m^2 - N(N-1)\mu^2\Big] \\
&= \frac{1}{N(N-1)}\Big[\sum_{j=1}^{N} u_j\Big(\sum_{k=1}^{N} u_k\Big) - N(\sigma^2+\mu^2) - N(N-1)\mu^2\Big] \\
&= \frac{1}{N(N-1)}\Big[N\mu\sum_{j=1}^{N} u_j - N\sigma^2 - N\mu^2 - N^2\mu^2 + N\mu^2\Big] \\
&= \frac{1}{N(N-1)}\Big[N\mu\cdot N\mu - N\sigma^2 - N^2\mu^2\Big] = -\frac{\sigma^2}{N-1}.
\end{aligned}$$

Using the formulas of the previous section, we thus obtain

$$\begin{aligned}
\mathrm{E}[\overline{Y}] &= \mathrm{E}\Big[\frac{Y_1+\cdots+Y_n}{n}\Big] = \frac{1}{n}\mathrm{E}[Y_1+\cdots+Y_n] = \frac{1}{n}\Big(\mathrm{E}[Y_1]+\cdots\mathrm{E}[Y_n]\Big) \\
&= \frac{1}{n}(\underbrace{\mu+\cdots+\mu}_{n \text{ times}}) = \mu, \quad \text{and} \\
\mathrm{V}[\overline{Y}] &= \mathrm{V}\Big[\frac{Y_1+\cdots+Y_n}{n}\Big] = \frac{1}{n^2}\mathrm{V}[Y_1+\cdots+Y_n] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}\mathrm{Cov}(Y_i, Y_\ell) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n}\sigma^2 + 2\sum_{i=1}^{n}\sum_{\ell=i+1}^{n}\mathrm{Cov}(Y_i, Y_\ell)\Big] = \frac{1}{n^2}\Big[n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1}\Big] \\
&= \frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).
\end{aligned}$$

Let's go back to the above example: we have $N = 4$, $n = 3$, $\mu = 2$, and $\sigma^2 = \frac{7}{2}$. According to what we have just found, we indeed get

$$\mathrm{E}[\overline{Y}] = 2 \qquad \text{and} \qquad \mathrm{V}[\overline{Y}] = \frac{7/2}{3}\left(\frac{4-3}{4-1}\right) = \frac{7}{18}.$$

The component $\frac{N-n}{N-1}$ is the **finite population correction factor** (FPCF); it shows up because the population is not infinite. Since the SRS is constructed without replacing the units in the finite population after they have been drawn into the sample, the presence of a unit in the SRS affects the probability that another unit will also be in the SRS – **the random variables $Y_i$ are not independent**.[18]

18: When $N$ is "large" and the ratio $\frac{n}{N}$ is "small", the FPCF $\approx 1$, in which case the situation is very similar to sampling with replacement.

### 10.3.2 Estimators and Confidence Intervals

The estimator $\overline{y}$ is unbiased under SRS. In that case, how do we interpret the sapling variance $\mathrm{V}(\overline{y})$? Quite simply, it provides an idea of the typical distance between the **empirical mean** $\overline{y}$ and the **population mean** $\mu$.

The **mean square error** of $\overline{y}$ under SRS is

$$\mathrm{MSE}(\overline{y}) = \mathrm{V}(\overline{y}) + (\mathrm{E}(\overline{y}) - \mu)^2 = \mathrm{V}(\overline{y}) + 0 = \mathrm{V}(\overline{y}),$$

which is to say that the estimation error is entirely dominated by $\mathrm{V}(\overline{y})$.

When we sample with replacement,[19] the samples $y_1, \ldots, y_n$ are viewed as **independent** from one another. If they are also **indentically distributed**, we then have $\mathrm{E}(y_i) = \mu$ and $\mathrm{V}(y_i) = \sigma^2$, or

$$\mathrm{E}(\overline{y}) = \mu, \quad \text{and} \quad \mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}.$$

When $n \to \infty$, the CLT states that $\overline{y} \sim_{\text{approx.}} \mathcal{N}(\mu, \sigma^2/n)$, whence

$$Z = \frac{\overline{y} - \mu}{\mathrm{SD}(\overline{y})} = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \sim_{\text{approx.}} \mathcal{N}(0, 1).$$

Let $\alpha \in (0, 1)$. Denote the $(1-\alpha)^{\text{th}}$ **quantile of a standard normal random variable** $Z \sim \mathcal{N}(0, 1)$ by $z_\alpha > 0$. According to the frequentist interpretation of probability, we can expect that $\frac{\overline{y}-\mu}{\sigma/\sqrt{n}}$ will fall in the interval $(-z_{\alpha/2}, z_{\alpha/2})$ roughly $100(1-\alpha)\%$ of the time:[20]

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{y} - \mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

The quantity

$$B_\alpha = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = z_{\alpha/2}\mathrm{SD}(\overline{y})$$

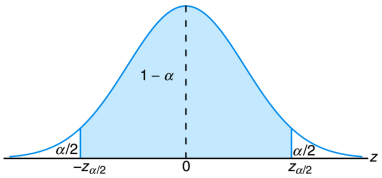
is the **bound on the error of estimation**, and we can build an approximate **95% confidence interval** for the mean $\mu$:

$$\text{C.I.}(\mu; 100(1-\alpha)\%): \quad y \pm B_\alpha = y \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

However, in a SRS scenario, we are **NOT** dealing with i.i.d. random variables. How must this argument be modified when we sample without replacement from a finite population?

19: Which is not a SRS situation.

20: The important quantiles are illustrated below:




### Sampling Context – Gapminder Data

We will illustrate the important concepts of sampling theory with the help of the 2011 Gapminder dataset, as we had done at the start of the section. In addition to average life expectancy, we are also interested in:

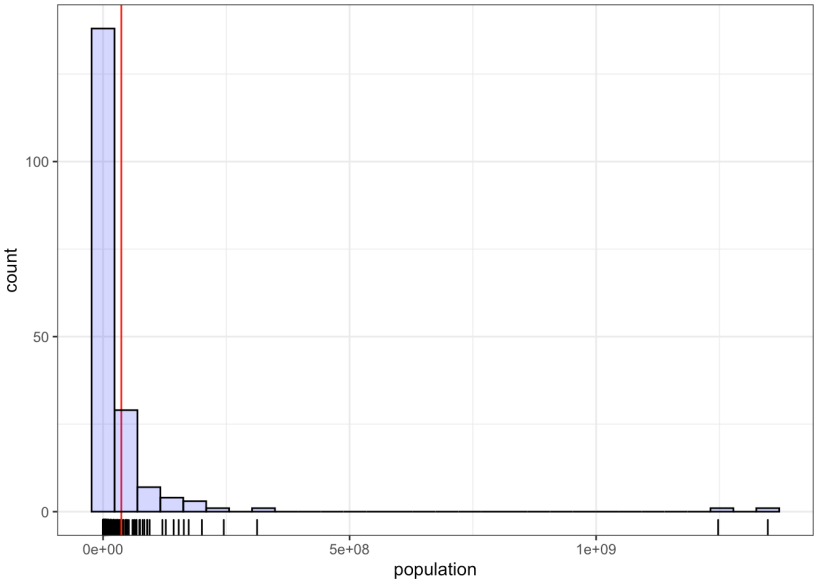
- the **total population** of the planet,
- the **average population** per country, and
- the **proportion** of countries with a population of less than 10M.

The population of 185 countries is available – it ranges from 56,641 to 1,348,174,478, with an average value $\mu = 37,080,426$.

```
gapminder.SRS <- gapminder |>
  filter(year==2011) |> select(life_expectancy,population)
str(gapminder.SRS)
summary(gapmider.SRS)
```

```
'data.frame':   185 obs. of  2 variables:
 $ life_expectancy: num  77.4 76.1 58.1 75.9 76 ...
 $ population     : int  2886010 36717132 21942296 88152 41655616 ...

life_expectancy  population
Min.    46.70       5.644e+04
1st Qu: 65.30       2.064e+06
Median :73.70       7.563e+06
Mean   :71.18       3.708e+07
3rd Qu.:77.40       2.423e+07
Max.   :83.02       1.348e+09
```

```
ggplot(data=gapminder.SRS, aes(population)) +
  geom_rug() +
  geom_vline(xintercept=mean(gapminder.SRS$population),
             color="red") +
  geom_histogram(col="black", fill="blue", alpha=.2)
```

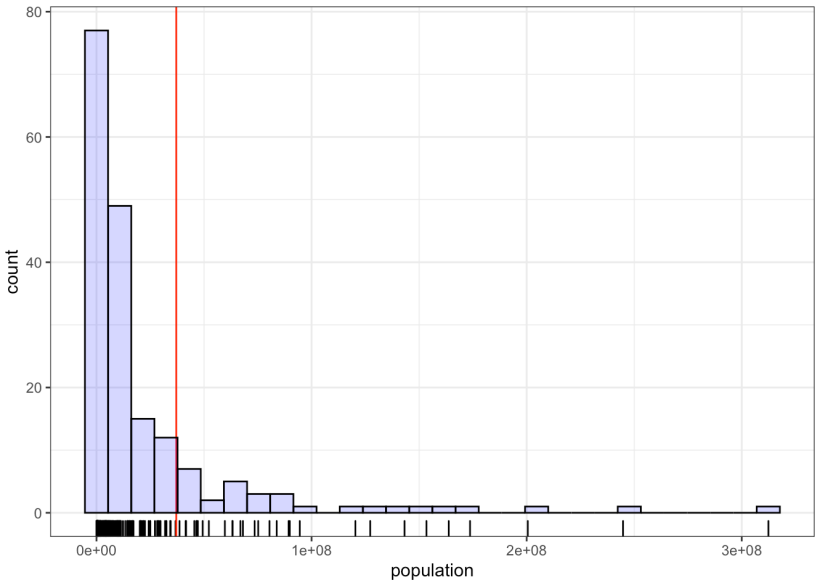



The population distribution by country is **asymmetric**, with a tail that **spreads to the right**, and two outliers (China and India). These observations will sometimes be removed from the data set.

```
gapminder.SRS.2 <- gapminder |>
  filter(year==2011) |>
  select(life_expectancy,population) |>
  filter(population<500000000)
nrow(gapminder.SRS.2)
summary(data.frame(gapminder.SRS.2$population))
```

```
[1] 183

Min.   1st Qu.  Median   Mean      3rd Qu.   Max.
56441  2061342  7355231  23301958  22242334  312390368
```

```
ggplot(data=gapminder.SRS.2, aes(population)) +
  geom_rug() +
  geom_vline(xintercept=mean(gapminder.SRS$population),
             color="red") +
  geom_histogram(col="black", fill="blue", alpha=.2)
```




The associated distribution has the same shape as the one with all countries, but the 183 populations all fall below 312,390,368, with a mean value of $\mu = 23,301,958$.

### Estimating the Mean $\mu$

In an SRS, we have shown that the empirical mean $\overline{y}$ computed from a sample of size $n$ is an **unbiased estimator** of the mean $\mu$ of a population of size $N$ and variance $\sigma^2$. We have also shown that the **sampling variance** of the $\overline{y}$ estimator is

$$\mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

What distribution can we expect $\overline{y}$ to follow? Let's go back to the example of the world population (without China and India). We produce 500 SRS samples of $n = 20$ countries from the list of $N = 183$ countries. For each sample $1 \leq i \leq 500$, we compute the **empirical mean** $\overline{y}_i$:

```
set.seed(12) # replicability
N=dim(gapminder.SRS.2)[1]
n=20
m=500

means <- c()
for(k in 1:m){
    means[k] <- mean(gapminder.SRS.2[sample(1:N,n,
                    replace=FALSE),2])
```

```
  }
```

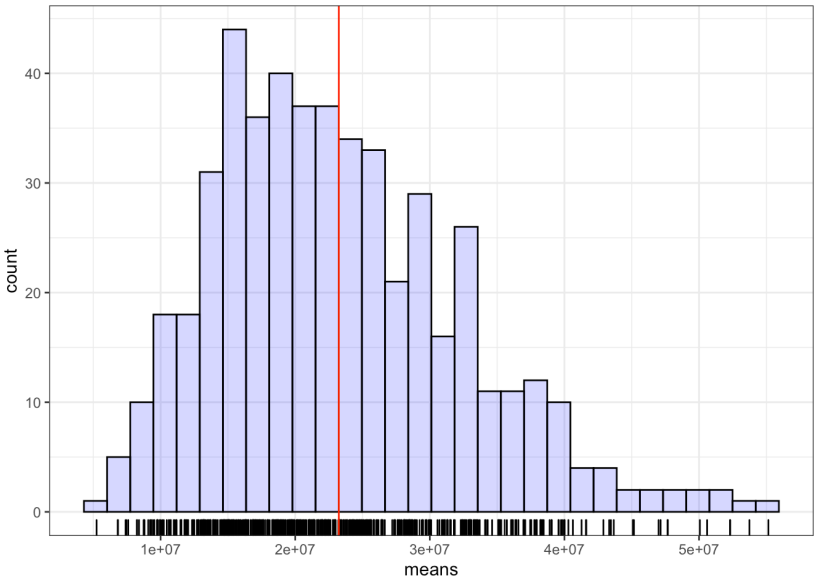
The SRS sample means are listed below:

```
summary(data.frame(means))
```

```
Min.   : 5244486
1st Qu.:16289930
Median :21986525
Mean   :23238867
3rd Qu.:28718720
Max.   :55152022
```

Their distribution (and mean) is:

```
ggplot(data=data.frame(means), aes(means)) +
  geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  geom_vline(xintercept=mean(means), color="red")
```




Although the distribution of empirical means $\overline{y}_i$ is **asymmetric with a tail spreading to the right**, the density curve still resembles that of a **normal distribution**.

**Central Limit Theorem – SRS** Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be a finite population with mean $\mu$ and variance $\sigma^2$, and let $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U}$ be a simple random sample. If $n$ and $N - n$ are both "sufficiently large", then

$$\overline{y} \sim_{\text{approx.}} \mathcal{N}\left(\mathrm{E}(\overline{y}), \mathrm{V}(\overline{y})\right) = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right).$$

In a SRS, the **bound on the error of estimation** and the approximate **95% C.I.** are given by:

$$B_\mu = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \quad \text{and} \quad P(|\overline{y}-\mu| \le B_\mu) \approx P\left(\left|\frac{\overline{y}-\mu}{\mathrm{SD}(\overline{y})}\right| \le 2\right) \approx 0.9544.$$

In practice, the **population variance** $\sigma^2$ is rarely known. We usually approximate it with the **empirical variance**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right], \quad \{y_i\} \text{ i.i.d.}$$

Unfortunately, $s^2$ is a **biased estimator** of $\sigma^2$ when the simple random sample is selected without replacement from a finite population. Indeed,

$$\begin{aligned}
\mathrm{E}(s^2) &= \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2\right] \\
&= \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu+\mu-\overline{y})^2\right] \\
&= \mathrm{E}\left[\frac{1}{n-1}\left[\sum_{i=1}^{n}(y_i-\mu)^2 - n(\overline{y}-\mu)^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathrm{E}\left[(y_i-\mu)^2\right] - n\mathrm{E}\left[(\overline{y}-\mu)^2\right]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma^2 - n\mathrm{V}(\overline{y})\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right] = \frac{\sigma^2}{n-1}\left[n - \frac{N-n}{N-1}\right] \\
&= \frac{\sigma^2}{n-1}\left[\frac{nN-n-N+n}{N-1}\right] = \frac{\sigma^2}{n-1}\cdot\frac{N(n-1)}{N-1} = \frac{N}{N-1}\sigma^2.
\end{aligned}$$

The **unbiased estimator** of $\sigma^2$ in the SRS context is instead

$$\frac{N-1}{N}s^2$$

since

$$\mathrm{E}\left[\frac{N-1}{N}s^2\right] = \frac{N-1}{N}\mathrm{E}(s^2) = \frac{N-1}{N}\cdot\frac{N}{N-1}\sigma^2 = \sigma^2.$$

We can approximate the **sampling variance** by replacing $\sigma^2$ by $\frac{N-1}{N}s^2$ in the expression for $\mathrm{V}(\overline{y})$:

$$\hat{\mathrm{V}}(\overline{y}) = \frac{N-1}{N}\cdot\frac{s^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}\left(1-\frac{n}{N}\right).$$

The **bound on the error of estimation** is thus approxiated by

$$B_\mu \approx \hat{B}_\mu = 2\sqrt{\hat{\mathrm{V}}(\overline{y})} = 2\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)},$$

from which we conclude that

$$\text{C.I.}(\mu; 0.95): \quad \overline{y} \pm 2\sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

is an approximate **95% confidence interval for** $\mu$.

If the population variance $\sigma^2$ is **known**, the FPCF is $\frac{N-n}{N-1}$; if it is **unknwon**, the FPCF in $1 - \frac{n}{N}$. In practice, when the **sampling rate** $\frac{n}{N}$ is below 5%, we can easily drop the FPCF ($1 - \frac{n}{N} \approx 1$) without affecting the resulting quantities too greatly.

**Example** We draw a SRS sample $\mathcal{Y}$ of size $n = 132$ from a finite population $\mathcal{U}$ with $N = 37,444$ units. Let the sample mean and sample standard deviation be $\overline{y} = 111.3$ and $s = 16.35$, respectively. Find an approximate 95% C.I. for the population average $\mu$.

The bound on the error of estimation is roughly

$$\hat{B}_\mu = 2\sqrt{\hat{\mathrm{V}}(\overline{y})} = 2\sqrt{\frac{16.35^2}{132}\left(1 - \frac{132}{37444}\right)} \approx 2.8,$$

which implies that

$$\text{C.I.}(\mu; 0.95) \approx 111.3 \pm 2.8;$$

the outcome is basically the same without the FPCF. ■

**Example** Find an approximate 95% C.I. for the average population per country in 2011 (excluding China and India) with a SRS of size $n = 20$.

We draw such a SRS sample and compute its sample mean $\overline{y}$ and sample variance $s^2$ (the outcomes will of course vary from one sample to another).

```
set.seed(12) # replicability
N = dim(gapminder.SRS.2)[1]
n = 20
SRS = gapminder.SRS.2[sample(1:N,n, replace=FALSE),2]
(y.bar = mean(SRS))
(s.2 = var(SRS))
```

```
[1] 35217143
[1] 5.492071e+15
```

If we do not know the population variance, the bound $\hat{B}_\mu$ and the corresponding approximate 95% C.I. for $\mu$ are given by:

```
(B.hat = 2*sqrt(s.2/n*(1-n/N)))
(IC.hat = c(y.bar-B.hat,y.bar+B.hat))
```

```
[1] 31278890
[1]  3938253 66496034
```

We can compare with the true mean $\mu$:

```
(mu = mean(gapminder.SRS.2[,2]))
```

```
[1] 23301958
```

Sure enough, $\mu$ is in the confidence interval:

```
mu > IC.hat[1] & mu < IC.hat[2]
```

```
[1] TRUE
```

In this case, however, we also knew the population variance $\sigma^2$:

```
(sigma.2 = var(gapminder.SRS.2[,2]))
```

```
[1] 1.885224e+15
```

The bound $B_\mu$ and the corresponding approximate 95% C.I. for $\mu$ are then obtained *via*:

```
(B = 2*sqrt(sigma.2/n*(N-1)/(N-n)))
(IC = c(y.bar-B,y.bar+B))
```

```
[1] 20518160
[1] 14698984 55735303
```

Sure enough, $\mu$ is again in the confidence interval:

```
mu > IC[1] & mu < IC[2]
```

```
[1] TRUE
```

In both cases, the true mean $\mu = 23,301,958$ is contained in the confidence interval. We also notice that the C.I. when the variance $\sigma^2$ is known is contained in the 95% C.I. when the variance is not known.[21] ■

21: Will this always be the case?

In this case, the true mean was in the confidence interval. But it could be that the 95% C.I. constructed from a sample does not contain the mean $\mu$.

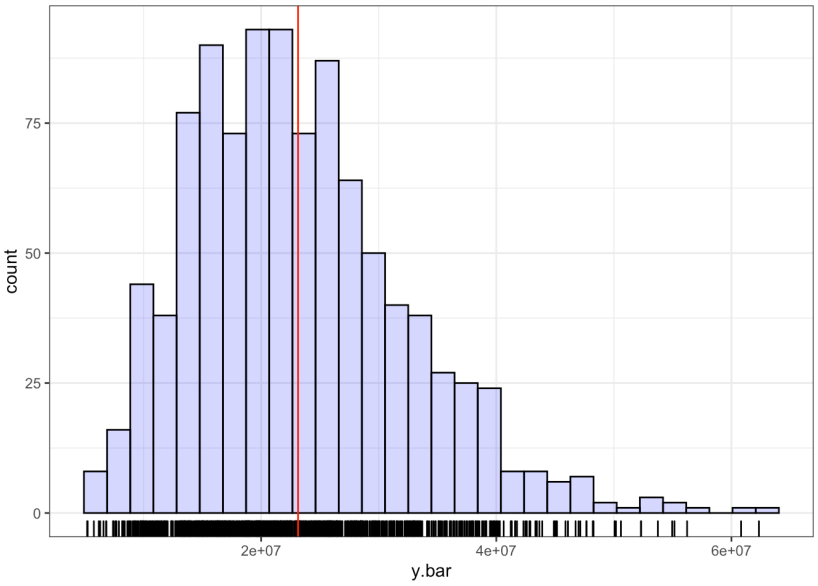
**Example** We repeat this procedure $m = 1000$ times (with different samples each time). If the CLT for SRS applies, how many times would we expect $\mu$ to be in the approximate 95% C.I. built from the simple random samples? Assume that $\sigma^2$ is not known.

```
m = 1000
mu.in.IC = c()
y.bar = c()
for(j in 1:m){
  test = gapminder.SRS.2[sample(1:N,n, replace=FALSE),2]
  s.2 = var(test)
  B.hat = 2*sqrt(s.2/n*(1-n/N))
  y.bar[j] = mean(test)
  mu.in.IC[j] = y.bar[j]-B.hat < mu & mu < y.bar[j]+B.hat
  }
mean(mu.in.IC)
```

```
[1] 0.821
```

This is not the $\approx 95\%$ we expected; but if we increase the sample size, the proportion gets closer to 95% (see Exercises). The long tail of the population distribution for $N = 183$ units probably plays a role – the distribution of the sample meas $\overline{y}$ (with $m = 1000$ samples of size $n = 20$) does not appear to be normal.

```
ggplot(data=data.frame(y.bar), aes(y.bar)) +
  geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  geom_vline(xintercept=mean(y.bar), color="red")
```




**Estimating the Total $\tau$**

Most of the work has been done: since the **total** $\tau$ can be re-written as

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

we can approximate $\tau$ with a SRS through the formula

$$\hat{\tau} = N\overline{y} = \frac{N}{n}\sum_{i=1}^{n} y_i.$$

This estimator is unbiased since its **expectation** is

$$\mathrm{E}(\hat{\tau}) = \mathrm{E}(N\overline{y}) = N \cdot \mathrm{E}(\overline{y}) = N\mu = \tau.$$

Its **sampling variance** is given by

$$\mathrm{V}(\hat{\tau}) = \mathrm{V}(N\overline{y}) = N^2 \cdot \mathrm{V}(\overline{y}) = N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right);$$

the **bound on the estimation error** is thus

$$B_\tau = 2\sqrt{\mathrm{V}(\hat{\tau})} = 2\sqrt{N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} = N \cdot B_\mu.$$

Since we do not usually know the true population variance $\sigma^2$ of $\mathcal{U}$, we provide an approximation by substituting $\sigma^2$ by the sample variance $s^2$, which needs to be multiplied by the "biased" factor $\frac{N-1}{N}$.[22] We can thus provide an **approximation of the sampling variance** using

22: Recall that $s^2$ is a biased estimator of $\sigma^2$ in a SRS.

$$\hat{\mathrm{V}}(\hat{\tau}) = \hat{\mathrm{V}}(N\overline{y}) = N^2 \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right);$$

this yields an **approximate bound on the estimation error** of

$$B_\tau \approx \hat{B}_\tau = 2\sqrt{\hat{\mathrm{V}}(\hat{\tau})} = 2\sqrt{N^2 \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right)} = N \cdot \hat{B}_\mu,$$

and an **approximate 95% C.I. for** $\tau$:

$$\text{C.I.}(\tau; 0.95): \quad \hat{\tau} \pm 2\sqrt{N^2 \cdot \frac{s^2}{n}\left(1 - \frac{n}{N}\right)}.$$

**Example** Consider a sample $\mathcal{Y}$ of size $n = 132$ drawn from a finite population $\mathcal{U}$ of size $N = 37,444$. Suppose the empirical mean and standard deviation of the sample are $\overline{y} = 111.3$ and $s = 16.35$, respectively. Give an approximate 95% C.I. for the total $\tau$ in $\mathcal{U}$.

The approximate bound on the error of estimation

$$\hat{B}_\tau = 2\sqrt{N^2 \cdot \hat{\mathrm{V}}(\overline{y})} = 2\sqrt{37444^2 \cdot \frac{16.35^2}{132}\left(1 - \frac{132}{37444}\right)} \approx 106,383.9643,$$

which yields

$$\text{C.I.}(\tau; 0.95) \approx 37,444 \cdot 111.3 \pm 106,383.9643 = 4,167,517.2 \pm 106,384.0,$$

or simply $(4,061,133.2; 4,273,901.2)$. ■

**Example** Find an approximate 95% C.I. for the population of the planet in 2011 (excluding China and India), using a SRS of size $n = 20$, assuming

that

$$\overline{y} = 27,396,632 \quad \text{and} \quad \text{C.I.}(\mu; 0.95) \equiv (6,755,099; 48,038,164).$$

We have $\hat{B}_\mu \approx 48,038,164 - 27,396,632 = 20,641,532$ and

$$\hat{B}_\tau \approx N\hat{B}_\mu = 183 \cdot 20,641,532 = 3,777,400,356,$$

from which we conclude that

$$\text{C.I.}(\tau; 0.95): \quad N\overline{y} \pm B_\tau = 183(27,396,632) \pm 3,777,400,356,$$

or simply, $\text{C.I.}(\tau; 0.95) :\equiv (1,236,183,300; 8,790,984,012)$.[23] ■

23: The interval is "valid", but it is perhaps too wide to be of practical use. We will discuss ways to improve the prediction in future sections.

**Estimating a Proportion** $p$

In a population $\mathcal{U}$ where $u_j \in \{0,1\}$ represents a **binary response** for all $1 \le j \le N$,[24] the **mean** takes a particular interpretation:

24: For example, $u_j = 1$ when the corresponding unit possesses a certain characteristic, and $u_j = 0$ when it does not.

$$p = \mu = \frac{1}{N}\sum_{j=1}^{N} u_j$$

is the **proportion** of the units possessing the characteristic in question.

This proportion can be estimated with a SRS *via*:

$$\hat{p} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad y_i \in \{0,1\}.$$

It is an unbiased estimator of the proportion since its **expectation** is

$$\text{E}(\hat{p}) = \text{E}(\overline{y}) = \mu = p;$$

its **sampling variance** is

$$\text{V}(\hat{p}) = \text{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

But $U^2 = U$ when $U$ is a binary response, from which we see that

$$\sigma^2 = \text{E}[U^2] - \text{E}^2[U] = \text{E}[U] - \text{E}^2[U] = p - p^2 = p(1-p),$$

and so

$$\text{V}(\hat{p}) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

The **bound on the error of estimation** is thus

$$B_p = 2\sqrt{\text{V}(\hat{p})} = 2\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}.$$

When the population variance $\sigma^2$ is unknown (which is to say, when the true $p$ is unknown, which is usually the case), the **sampling variation approximation** is

$$\hat{\text{V}}(\hat{p}) = \hat{\text{V}}(\overline{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

But recall that $y_i$ only takes on the values 0 and 1, so that $y_i^2 = y_i$ for $1 \le i \le n$, from which we see that

$$s^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\Big) = \frac{n\overline{y} - n\overline{y}^2}{n-1} = \frac{n(\hat{p} - \hat{p}^2)}{n-1} = \frac{n\hat{p}(1-\hat{p})}{n-1},$$

and

$$\hat{\mathrm{V}}(\hat{p}) = \frac{n\hat{p}(1-\hat{p})}{(n-1)n}\Big(1 - \frac{n}{N}\Big) = \frac{\hat{p}(1-\hat{p})}{n-1}\Big(1 - \frac{n}{N}\Big).$$

The **approximate estimation error bound** becomes

$$B_p \approx \hat{B}_p = 2\sqrt{\hat{\mathrm{V}}(\hat{p})} = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\Big(1 - \frac{n}{N}\Big)},$$

with the corresponding approximate 95% C.I. for $p$ being

$$\text{C.I.}(p; 0.95): \quad \hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\Big(1 - \frac{n}{N}\Big)}.$$

**Example** Consider a sample $\mathcal{Y}$ of size $n = 132$ drawn from a finite population $\mathcal{U}$ of size $N = 37,444$. Suppose that 25 of the observations of $\mathcal{Y}$ have a particular characteristic. Find an approximate 95% C.I. for the proportion $p$ of the observations of $\mathcal{U}$ that possess the feature.

In this case, $\hat{p} = 25/132 \approx 0.19$. The required approximate bound is thus

$$\hat{B}_p = 2\sqrt{\hat{\mathrm{V}}(\hat{p})} = 2\sqrt{\frac{0.19(1-0.19)}{132-1}\Big(1 - \frac{132}{37444}\Big)} \approx 0.0684,$$

from which we get

$$\text{C.I.}(p; 0.95) \approx 0.19 \pm 0.0684 \equiv (0.121, 0.258). \quad \blacksquare$$

**Example** Find an approximate 95% C.I. for the proportion of countries for which the 2011 population fell below 10M, using a SRS with sample size $n = 20$.

Let's draw a SRS sample of size $n = 20$ and compute $\hat{p}$ (results will vary from one sample to when the population of a country is smaller than 10M and FALSE otherwise.

```
set.seed(1234) # replicability
N=dim(gapminder.SRS.2)[1]
n=20
thresh.10 <- gapminder.SRS.2[,2] < 10000000
SRS = thresh.10[sample(1:N,n, replace=FALSE)]
```

The proportion of countries with a population smaller than 10M in that sample is:

```
(p.hat = mean(SRS))
```

```
[1] 0.6
```

The true proportion $p$, amongst the $N = 185$ countries, is:

```
(p = mean(thresh.10))
```

```
[1] 0.5737705
```

If we assume that population variance is unknown, the bound $\hat{B}_p$ and the approximate 95% C.I. are given by:

```
(B.p = 2*sqrt(p.hat*(1-p.hat)/(n-1)*(1-n/N)))
(IC = c(p.hat-B.p,p.hat+B.p))
```

```
[1] 0.2121422
[1] 0.3878578 0.8121422
```

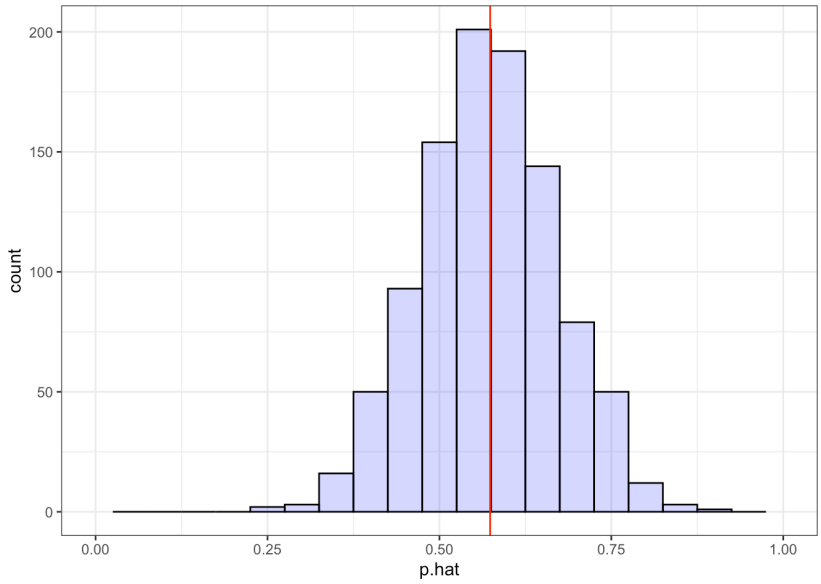
The true proportion $p \approx 0.568$ is indeed in the confidence interval. If we repeat this process $m = 1000$ times, how often is the true proportion found inside the obtained C.I.?

```
m=1000
p.in.IC = c()
p.hat = c()
for(j in 1:m){
  p.hat[j] = mean(thresh.10[sample(1:N,n, replace=FALSE)])
  B.p = 2*sqrt(p.hat[j]*(1-p.hat[j])/(n-1)*(1-n/N))
  p.in.IC[j] = p.hat[j]-B.p < p & p < p.hat[j]+B.p
  }
mean(p.in.IC)
```

```
[1] 0.963
```

Quite close to 95%, you will agree. The distribution of the $m = 1000$ estimates $\hat{p}$ is shown below, with the true proportion (red vertical line).

```
ggplot(data=data.frame(p.hat), aes(p.hat)) +
  geom_histogram(bins=21, col="black", fill="blue",
                 alpha=.2) +
  geom_vline(xintercept=mean(gapminder.SRS.2[,2]<10000000),
             color="red") + xlim(0,1)
```




### 10.3.3 Sample Size

Selecting an appropriate sample size is a challenge, and there is a bit of a chicken-and-egg scenario at play.

Firstly, there is a **practical** problem associated with sampling: since the cost associated with each response can be **costly** (in terms of **time/cost**), we often seek to **minimize the size** of the **realized** sample $\mathcal{Y}$, given a **desired error bound**.

Secondly, the SRS error bound is expressed as

$$B_\xi = 2\sqrt{\mathrm{V}(\hat{\xi})}, \quad \xi \in \{\mu, \tau, p\},$$

but the variance depends on the sample size $|\mathcal{Y}| = n$. We must then express $n$ in terms of the (known) parameters $N$, $\sigma^2$, and $B_\xi$.

#### Mean $\mu$

If we are trying to estimate the mean $\mu$, we have:

$$B_\mu = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \iff \underbrace{\frac{B_\mu^2}{4}}_{=D_\mu} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$\frac{(N-1)D_\mu}{\sigma^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \iff \frac{(N-1)D_\mu + \sigma^2}{\sigma^2} = \frac{N}{n}$$

$$\iff n_\mu = \frac{N\sigma^2}{(N-1)D_\mu + \sigma^2}.$$

However, we can only use this formula is we **know the population variance** $\sigma^2$. We could chose to use the **empirical variance** $s^2$ of the sample $\mathcal{Y}$ as we did when we estimated the sample variance, **but we haven't drawn $\mathcal{Y}$ from $\mathcal{U}$ yet**!

**Stratagies** (to obtain $\sigma^2$):

- use a **preliminary sample** (not necessarily random),
- use the empirical variance obtained in a previous study, or
- for a proportion, use a conservative estimate ($p = 0.5$).

**Example** Consider a finite population $\mathcal{U}$ with size $N = 37,444$. We are interested in the mean $\mu$ of the response variable in $\mathcal{U}$. In a preliminary SRS of size $n = 132$, we computed an (empirical) standard deviation of $s = 16.35$.

Using $\sigma = s$, find the minimal SRS sample size $n_\mu$ required to estimate the mean with a bound on the error of estimation at most $B_\mu = 1.7$.

We can use the formula directly to get

$$D_\mu = \frac{(1.7)^2}{4} \approx 0.73 \implies n_\mu = \frac{37444(16.35)^2}{(37444-1)(0.73)+16.35^2} = 366.39 \approx 367. \quad \blacksquare$$

**Total $\tau$**

If instead, we are seeking to estimate the total $\tau$, we have:

$$\begin{aligned} B_\tau = 2\sqrt{N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} &\iff \underbrace{\frac{B_\tau^2}{4N^2}}_{=D_\tau} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \\ &\iff \frac{(N-1)D_\tau}{\sigma^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &\iff \frac{(N-1)D_\tau + \sigma^2}{\sigma^2} = \frac{N}{n} \\ &\iff n_\tau = \frac{N\sigma^2}{(N-1)D_\tau + \sigma^2}. \end{aligned}$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$. We are interested in the total $\tau$ of the response variable of $\mathcal{U}$. In a preliminary SRS of size $n = 132$, we computed an empirical standard deviation of $s = 16.35$.

Using $\sigma = s$, find the minimal SRS sample size $n_\tau$ required to estimate the total response with a bound on the error of estimation at most $B_\tau = 10000$.

We can use the formula directly to obtain

$$D_\tau = \frac{(10000)^2}{4(37444)^2} \approx 0.018 \implies n_\tau = \frac{37444(16.35)^2}{(37444-1)(0.018)+16.35^2} \approx 10706. \quad \blacksquare$$

**Proportion** $p$

If we are interested in the proportion $p$, we have:

$$B_p = 2\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \iff \underbrace{\frac{B_p^2}{4}}_{=D_p} = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

$$\iff \frac{(N-1)D_p}{p(1-p)} = \frac{N-n}{n} = \frac{N}{n} - 1$$

$$\iff \frac{(N-1)D_p + p(1-p)}{p(1-p)} = \frac{N}{n}$$

$$\iff n_p = \frac{Np(1-p)}{(N-1)D_p + p(1-p)}.$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$. We are interested in the proportion $p$ of units that have a particular feature. In a preliminary SRS of size $n = 132$, we identify 25 observations possessing the feature.

Using the approximation $\sigma^2 = \frac{25}{132} \cdot \frac{107}{132}$ from the preliminary sample, find the minimal SRS sample size $n_p$ required to estimate the true proportion with a bound on the error of estimation of at most $B_p = 0.03$.

We use the formula directly and obtain

$$D_p = \frac{(0.03)^2}{4} \approx 0.0002 \implies n_p = \frac{37444(0.189)(0.811)}{(37444-1)(0.0002) + (0.189)(0.811)} \approx 671. \quad \blacksquare$$

**Example** Consider a situation similar to the previous example. Using the (conservative) approximation $\sigma^2 = (0.5)^2$, find the minimal SRS sample size $n_p$ required to estimate the true proportion with a bound on the error of estimation of at most $B_p = 0.03$.

We use the formula directly and obtain

$$D_p = \frac{(0.03)^2}{4} \approx 0.0002 \implies n_p = \frac{37444(0.5)(0.5)}{(37444-1)(0.0002) + (0.5)(0.5)} \approx 1080. \quad \blacksquare$$

## 10.4 Stratified Random Sampling

The theory we developed in the previous section allows us to determine the distribution of the three **unbiased** estimators $\overline{y}$, $\hat{tau}$, and $p$.

For instance, we have shown that if the size $N$ of a finite population $\mathcal{U} = \{u_1, \ldots, u_N\}$ of expectation $\mu$ and variance $\sigma^2$ and the size $n$ of the SRS $\mathcal{Y}$ from which the estimator $\overline{y}$ is constructed are **sufficiently large**, and if moreover the responses $u_j$ are **i.i.d.** for $1 \leq j \leq N$, then $\overline{y}$ follows **approximately** a normal distribution whose parameters are

$$\mathrm{E}(\overline{y}) = \mu \quad \text{and} \quad \mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

The higher $\sigma^2$ is, the more the repeated SRS $\overline{y}$ values vary.

In practice, the normal approximation is:

- often **acceptable** – see average life expectancy, previous section;
- but **it is not always so**, which can lead to some challenges – cf. the C.I.$(\mu; 0.95)$ for the average population which was in fact only an 80% C.I. for a SRS of size $n = 20$ in the previous section.

In the presence of **outliers** or when $n, N$ are **too small**, the performance of an SRS may leave something to be desired.

**Example** Consider a finite population with $N = 16$ elements:

$$2, 2, 2, 2, 0, 0, 0, 0, 1, 1, 1, 1, 5, 5, 5, 5.$$

The population mean and variance are, respectively:

$$\mu = \frac{1}{16}(4 \cdot 2 + 4 \cdot 0 + 4 \cdot 1 + 4 \cdot 5) = 2;$$
$$\sigma^2 = \frac{1}{16}(4 \cdot 2^2 + 4 \cdot 0^2 + 4 \cdot 1^2 + 4 \cdot 5^2) - 2^2 = \frac{7}{2}.$$

Suppose that we draw an SRS of size $n = 4$ from this population, in order to estimate the mean $\mu$.

From what we discussed in the previous section, the expectation and sampling variance of the estimator $\overline{y}$ are, respectively:

$$\mathrm{E}(\overline{y}) = 2 \quad \text{and} \quad \mathrm{V}(\overline{y}) = \frac{\sqrt{7/2}^2}{4}\left(\frac{16-4}{16-1}\right) = \frac{7}{10}.$$

We could also restrict the sampling structure in the following manner:

1. we start by **separating the population** into 4 segments (the **strata**):

   **strata 1** : 2, 2, 2, 2
   **strata 2** : 0, 0, 0, 0
   **strata 3** : 1, 1, 1, 1
   **strata 4** : 5, 5, 5, 5

2. we then draw a SRS of size $n = 4$ **by selecting one unit per stratum**.

In such a situation (which is **NOT** a SRS$(n = 4, N = 16)$), **each achieved sample** takes the form $\{2, 0, 1, 5\}$: the empirical mean is **always** 2, and so the sampling variance **is null**.

In practice, this artificial situation rarely occurs, but if the units of the population can be grouped into **natural strata**, i.e., **sub-populations** for which:

- the response value is **homogeneous** within each stratum, but
- it is **heterogeneous** from one stratum to another, then

this approach can produce an estimator whose sampling variance **is lower** than that of the SRS estimator; as a bonus, the sample **preserves certain population structures**.

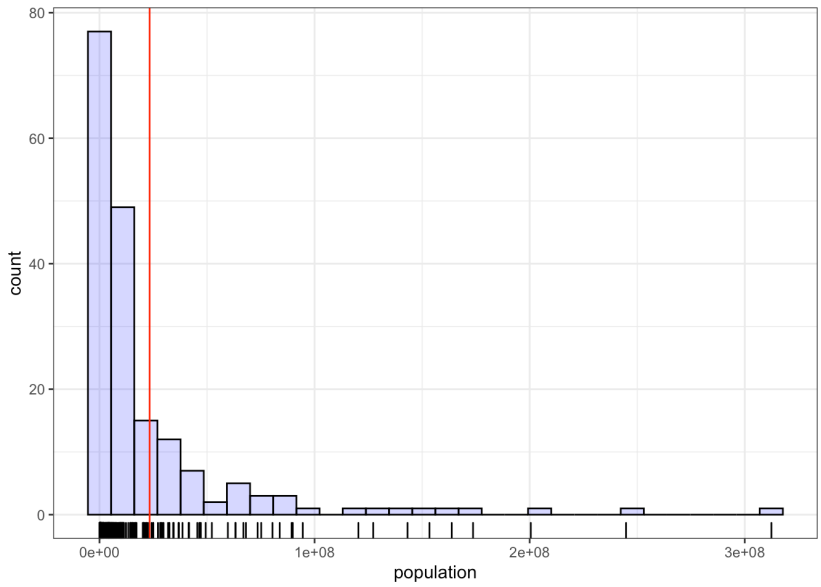
**Example** Find an approximate 95% C.I. for the average population per country (excluding China and India) in 2011. The population distribution in the 2011 Gapminder dataset has the following characteristics:

```
gapminder.STS <- gapminder |>
    filter(year==2011) |> select(population) |>
    filter(population < 1000000000)
summary(gapminder.STS$population)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   56441   2061342   7355231  23301958  22242334 312390368
```

The true average population, by country, is $\mu = 23,301,958$. Recall that the population distribution is asymmetrical:

```
N = nrow(gapminder.STS)
ggplot(data=gapminder.STS, aes(population)) +
    geom_histogram(col="black", fill="blue", alpha=.2) +
    geom_vline(xintercept=mean(gapminder.STS$population),
        color="red") + geom_rug()
```




We use the population strata $[0, 10M), [10M, 25M), [25M, 50M)$, 100M+.

```
gapminder.STS <- gapminder.STS |>
      mutate(strata = ifelse(population<10000000,"S1",
            ifelse(population<25000000,"S2",
            ifelse(population<50000000,"S3",
            ifelse(population<100000000,"S4","S5")))))

gapminder.STS <- gapminder.STS[order(gapminder.STS$population),]

gapminder.STS$strata <- as.factor(gapminder.STS$strata)
```

The number of countries in each stratum is:

```
(strata.N <- tapply(gapminder.STS$population,
                    gapminder.STS$strata, length))
```

```
 S1  S2  S3  S4  S5
105  35  21  13   9
```

For a sample size of $n = 20$, we use approximately $n_i$ countries per stratum $S_i$:

```
strata.N/sum(strata.N)*20
```

```
        S1         S2         S3         S4         S5
11.4754098  3.8251366  2.2950820  1.4207650  0.9836066
```

Some practical considerations might suggest the use of a **different allocation** (more on this later). The distribution of the population by stratum has the following characteristics:

```
tapply(gapminder.STS$population, gapminder.STS$strata,
       summary)
```

```
$S1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  56441  622957 2886010 3386819 5411377 9988846

$S2
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
10027140 11234699 15177280 15682124 20213668 24928503

$S3
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
25016921 29427631 34499905 36211465 41655616 49356692

$S4
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
52237272 63268405 73517002 73841185 83787634 94501233

$S5
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
120365271 143211476 163770669 182154642 200517584 312390368
```

In the first attempt, we draw a SRS from each stratum, using the following sizes: $(n_1, n_2, n_3, n_4, n_5) = (11, 4, 3, 1, 1)$:

```
set.seed(12345) # replicability
n=c(); n[1] = 11; n[2] = 4; n[3] = 3; n[4] = 1; n[5] = 1
ind = list()

# draw a SRS of indices in each of the 5 strata
```

```
ind[[1]] <- sample(1:strata.N[1],n[1])
ind[[2]] <- sum(strata.N[1:1]) + sample(1:strata.N[2],n[2])
ind[[3]] <- sum(strata.N[1:2]) + sample(1:strata.N[3],n[3])
ind[[4]] <- sum(strata.N[1:3]) + sample(1:strata.N[4],n[4])
ind[[5]] <- sum(strata.N[1:4]) + sample(1:strata.N[5],n[5])
```

The average population in the sample is computed as below (this value will change from one STS to another).

```
sample.STS <- gapminder.STS[unique(unlist(ind)),]
mean(sample.STS$population)
```

```
[1] 24378331
```

This naïve approach is not ideal.[25] The estimator

$$\overline{y}_{\text{STS}} = \tfrac{1}{20}(y_1 + \cdots + y_{20})$$

implies that **each observation had the same probability of being chosen**, which is not the case in reality.[26] In our second attempt, the weight of each selected observation depends on the size of the stratum.[27]

25: Despite the relative accuracy of the estimate.

26: Remember, we are not dealing with a SRS situation.

27: We will discuss the theoretical details in the next section.

```
set.seed(123456) # replicability
cumul.n = cumsum(n); cumul.N = cumsum(strata.N)

ind = list()
ind[[1]] <- sample(1:strata.N[1],n[1])
for(j in 2:length(n)){
  ind[[j]] <- cumul.N[j-1] + sample(1:strata.N[j],n[j])
}
sample.STS <- gapminder.STS[unique(unlist(ind)),]
sample.STS = sample.STS[order(sample.STS$population),]

y.bar <- list()
y.bar[[1]] <- mean(sample.STS[1:n[1],c("population")])
for(j in 2:length(n)){
  y.bar[[j]] <- mean(sample.STS[(cumul.n[j-1]+1):cumul.n[j], c("population")])
}

y.bar.STS <- 0
for(j in 1:length(n)){
  y.bar.STS <- y.bar.STS +
            as.numeric(strata.N[j])*y.bar[[j]]
}

y.bar.STS/N
```

```
[1] 22668202
```

The estimate is very close to the actual value of $\mu$, but a lone point estimate does not tell the full story.

We repeat this procedure 500 times, each time using the same size allocation $(n_1, n_2, n_3, n_4, n_5) = (9, 3, 3, 3, 2)$:

```
set.seed(12) # replicability
strata.N <- tapply(gapminder.STS$population,
                   gapminder.STS$strata, length)
cumul.N = cumsum(strata.N)

n=c(); n[1] = 9; n[2] = 3; n[3] = 3; n[4] = 3; n[5] = 2
cumul.n = cumsum(n)

m=500
means <- c()
for(k in 1:m){
    ind = list()
    ind[[1]] <- sample(1:strata.N[1],n[1])
    for(j in 2:length(n)){
      ind[[j]] <- cumul.N[j-1] +
                       sample(1:strata.N[j],n[j])
    }
    ind.STS <-unique(unlist(ind))
    sample.STS <- gapminder.STS[ind.STS,]
    sample.STS = sample.STS[order(sample.STS$population),]

    y.bar <- list()
    y.bar[[1]] <- mean(sample.STS[1:n[1],c("population")])
    for(j in 2:length(n)){
      y.bar[[j]] = mean(sample.STS[(cumul.n[j-1]+1):
                            cumul.n[j],c("population")])
    }

    y.bar.STS <- 0
    for(j in 1:length(n)){
      y.bar.STS <- y.bar.STS +
                    as.numeric(strata.N[j])*y.bar[[j]]
    }

    means[k] <- y.bar.STS/N
}
```

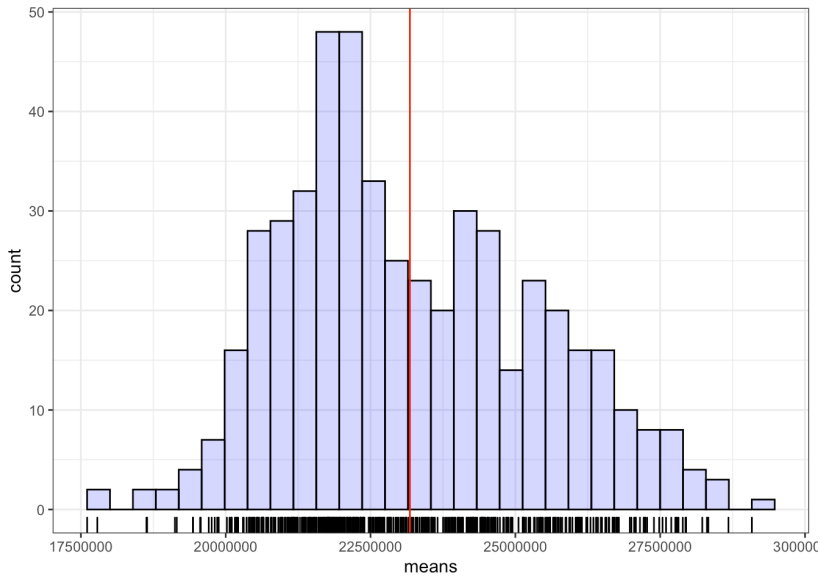
For each sample $1 \leq i \leq 500$, we then compute the **empirical mean** – their distribution has the following characteristics:

```
summary(means)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
17608174 21602380 22735650 23179372 24655297 29082447
```

Finally, we plot the histogram of the STS means (with their mean in red):

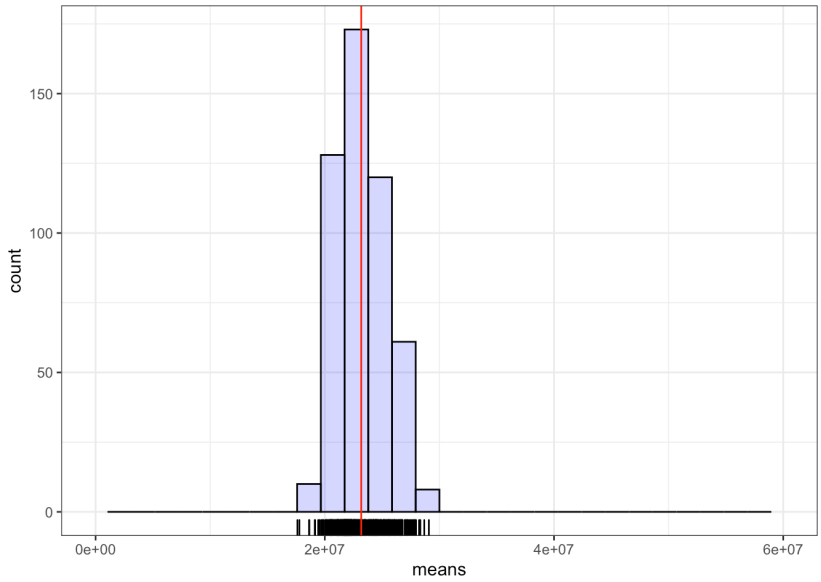
```
ggplot(data=data.frame(means), aes(means)) + geom_rug() +
    geom_histogram(col="black", fill="blue", alpha=.2) +
    geom_vline(xintercept=mean(means), color="red")
```




Not only is the shape of the distribution closer to a normal distribution, compared to the distribution of $\overline{y}$ obtained using SRS, but its variance is also much lower.

As an illustration, ccompare the following image, on the same scale as the corresponding histogram for SRS in Section 10.3.2.

```
ggplot(data=data.frame(means), aes(means)) + geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  xlim(0,60000000) +
  geom_vline(xintercept=mean(means), color="red")
```



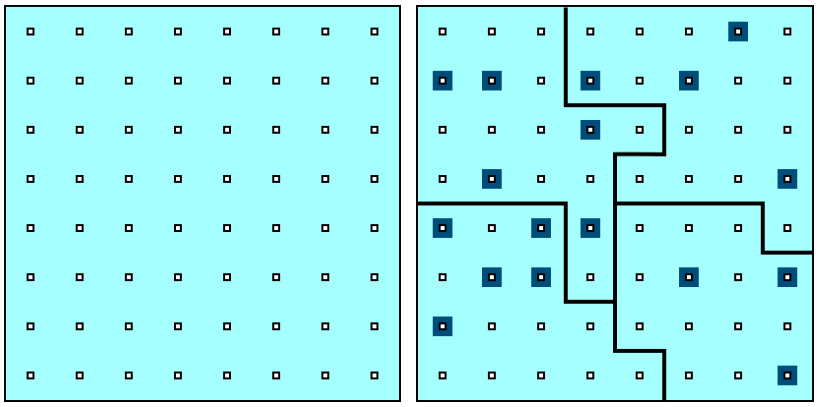


**Figure 10.7:** Schematics of STS: target population (left) and sample (right).

### 10.4.1 Estimators and Confidence Intervals

Assume that we are interested in a finite population $\mathcal{U} = \{u_1, \ldots, u_N\}$, whose expectation is $\mu$ and variance is $\sigma^2$. We cover the population with $M$ disjoint **strata**, containing, respectively, $N_1, \ldots, N_M$ units:

$$\mathcal{U}_1 = \{u_{1,1}, \ldots, u_{1,N_1}\}, \cdots, \mathcal{U}_M = \{u_{M,1}, \ldots, u_{M,N_M}\},$$

with **stratum mean** and **stratum variance**

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{i,j} \quad \text{and} \quad \sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{i,j}^2 - \mu_i^2, \quad 1 \le i \le M.$$

A stratified sample $\mathcal{Y}$ of size $n \le N$ is a subset of the target population $\mathcal{U}$, with$n_1 + \cdots + n_M = n$ and $n_i \le N_i$ for $1 \le i \le M$:

$$\{\underbrace{y_{1,1}, \ldots, y_{1,n_1}}_{\in \text{ strate } \mathcal{U}_1}, \ldots, \underbrace{y_{M,1}, \ldots, y_{M,n_m}}_{\in \text{ strate } \mathcal{U}_M}\} \subseteq \bigcup_{i=1}^{M} \mathcal{U}_i = \mathcal{U}.$$

If every sample $\mathcal{Y}_i = \{y_{i,j} \mid 1 \le j \le n_i\}$ is drawn from the corresponding stratum $\mathcal{U}_i$ *via* a SRS, **independently from one stratum to another**, we obtain a **stratified random sample** (STS) of size $n$. The **sample mean** and the **sample variance**[28] of $\mathcal{Y}_i$ are denoted by $\overline{y}_i$ and $s_i^2$, respectively. In a STS design, each observation in a stratum **has the same probability of being selected**, but it **may differ from one stratum to another**.

28: Which it is important to remember is not the same thing as the "sampling variance" of an estimator.

#### Mean $\mu$

In a STS, the **sample mean** of the observations of the sample $\mathcal{Y}$ falling in the stratum $\mathcal{U}_i$ is an estimator of $\mu_i$ given by

$$\overline{y}_i = \frac{1}{n_i} \sum_{\ell=1}^{n_i} y_{i,\ell}, \quad \text{where } n_i = |\mathcal{U} \cap \mathcal{Y}_i|, \ 1 \le i \le M.$$

The true mean $\mu$ and the **STS estimator** of $\mu$ are thus:

$$\mu = \frac{1}{N} \sum_{j=1}^{N} u_j = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} u_{i,j} = \frac{1}{N} \sum_{i=1}^{M} N_i \mu_i \quad \text{and} \quad \overline{y}_{\text{STS}} = \frac{1}{N} \sum_{i=1}^{M} N_i \overline{y}_i.$$

Since $\mathcal{Y}_i$ is a SRS drawn from $\mathcal{U}_i$, we have:[29]

$$\mathrm{E}(\overline{y}_i) = \mu_i \quad \text{and} \quad \mathrm{V}(\overline{y}_i) = \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right), \quad \text{pour } 1 \leq i \leq M.$$

The **expectation** of the STS estimator is thus:

$$\mathrm{E}\left(\overline{y}_{\mathrm{STS}}\right) = \mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{M} N_i \overline{y}_i\right) = \frac{1}{N}\sum_{i=1}^{M} N_i \mathrm{E}(\overline{y}_i) = \frac{1}{N}\sum_{i=1}^{M} N_i \mu_i = \mu,$$

which is to say that $\overline{y}_{\mathrm{STS}}$ is an **unbiased estimator** of the true mean $\mu$ for a population of size $N$ with variance $\sigma^2$.[30]

The **sampling variance** of the estimator $\overline{y}_{\mathrm{STS}}$ is

$$\begin{aligned}\mathrm{V}\left(\overline{y}_{\mathrm{STS}}\right) &= \mathrm{V}\left(\frac{1}{N}\sum_{i=1}^{M} N_i \overline{y}_i\right) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \mathrm{V}(\overline{y}_i) + \sum_{i \neq i'}^{M} N_i N_{i'} \underbrace{\mathrm{Cov}(\overline{y}_i, \overline{y}_{i'})}_{=0} \\ &= \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \mathrm{V}(\overline{y}_i) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right).\end{aligned}$$

**Central Limit Theorem – STS** If $n, N-n, n_i$, and $N_i - n_i$ are all sufficiently large, for all $i$, then

$$\overline{y}_{\mathrm{STS}} \sim_{\text{approx.}} \mathcal{N}\left(\mathrm{E}(\overline{y}_{\mathrm{STS}}), \mathrm{V}(\overline{y}_{\mathrm{STS}})\right) = \mathcal{N}\left(\mu, \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)\right).$$

In a STS, the **bound on the error of estimation** is

$$B_{\mu;\mathrm{STS}} = 2\sqrt{\mathrm{V}(\overline{y}_{\mathrm{STS}})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)}$$

and the corresponding approximate **95% C.I. for** $\mu$ is

$$\mathrm{C.I.}_{\mathrm{STS}}(\mu; 0.95): \quad \overline{y}_{\mathrm{STS}} \pm B_{\mu;\mathrm{STS}}.$$

In practice, the **population variance** $\sigma^2$ is rarely known,[31] in which case we use the **sample variance**.[32]

In each stratum, the **empirical variance** $s_i^2$ is

$$s_i^2 = \frac{1}{n_i - 1}\sum_{\ell=1}^{n_i}(y_{i,\ell} - \overline{y}_i)^2 = \frac{1}{n_i - 1}\left[\sum_{\ell=1}^{n_i} y_{i,\ell}^2 - n_i \overline{y}_i^2\right], \quad 1 \leq i \leq M.$$

We can then approximate the **sampling variance** in $\mathcal{U}_i$ as we did for a SRS, using

$$\hat{\mathrm{V}}(\overline{y}_i) = \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right), \quad 1 \leq i \leq M.$$

29: For the sake of completeness, the SRS estimator is sometimes denoted by $\overline{y}_{\mathrm{SRS}}$.

30: It is evidently not the one as $\overline{y}_{\mathrm{SRS}}$ is also such an estimator.

31: As is the **variance** $\sigma_i^2$ **in each stratum** $\mathcal{U}_i$, $1 \leq i \leq M$.

32: And the corresponding **finite population correction factor**.

The **sampling variance** of the estimator $\overline{y}_{\text{STS}}$ is thus

$$\hat{\text{V}}(\overline{y}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \text{V}(\overline{y}_i) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right).$$

The **bound of the estimation error** is approximated by

$$B_{\mu;\text{STS}} \approx \hat{B}_{\mu;\text{STS}} = 2\sqrt{\hat{\text{V}}(\overline{y}_{\text{STS}})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)},$$

whence

$$\text{C.I.}_{\text{STS}}(\mu; 0.95): \quad \overline{y}_{\text{STS}} \pm \hat{B}_{\mu;\text{STS}} \equiv \overline{y}_{\text{STS}} \pm 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)}$$

is an **approximate 95% C.I. for $\mu$.**

In practice, when the **stratum sampling rate** $\frac{n_i}{N_i}$ is below 5%, we can drop the FPCF in the corresponding stratum.

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. A STS sample $\mathcal{Y}$ of size $n = 132$ is drawn from $\mathcal{U}$, with $n_1 = 82$ and $n_2 = 50$.

Suppose the empirical mean and standard deviation in $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are $\overline{y}_1 = 120.7$, $\overline{y}_2 = 96.6$, $s_1 = 18.99$, and $s_2 = 14.31$, respectively. Find a 95% C.I. for the mean $\mu$ of $\mathcal{U}$.

The bound on the error of estimation is $\approx \hat{B}_{\mu;\text{STS}} = 2\sqrt{\hat{\text{V}}(\overline{y}_{\text{STS}})}$:

$$2\sqrt{\frac{21123^2}{37444^2}\cdot\frac{18.99^2}{82}\left(1-\frac{82}{21123}\right) + \frac{16321^2}{37444^2}\cdot\frac{14.31^2}{50}\left(1-\frac{50}{16321}\right)} \approx 2.95,$$

so $\text{C.I.}_{\text{STS}}(\mu; 0.95) \approx \left(\frac{21,123(120.7)}{37,444} + \frac{16,321(96.6)}{37,444}\right) \pm 2.95 \equiv (107.25, 113.14)$.

**Example** Find a 95% confidence interval for the average life expectancy by country in 2011 (including India and China), using a STS of size $n = 20$.[33]

33: Stratifying using the country **populations**, as we did earlier in this section.

We can basically re-use the same code:

```
LE.1 <- gapminder |> filter(year==2011) |>
    select(population,life_expectancy)
summary(LE.1)
```

```
  population        life_expectancy
Min.   :5.644e+04   Min.   :46.70
1st Qu.:2.064e+06   1st Qu.:65.30
Median :7.563e+06   Median :73.70
Mean   :3.708e+07   Mean   :71.18
3rd Qu.:2.423e+07   3rd Qu.:77.40
Max.   :1.348e+09   Max.   :83.02
```

The average life expectancy is $\mu = 71.18$. We now prepare the strata according to the population, and we sort the observations from the smallest population to the largest:

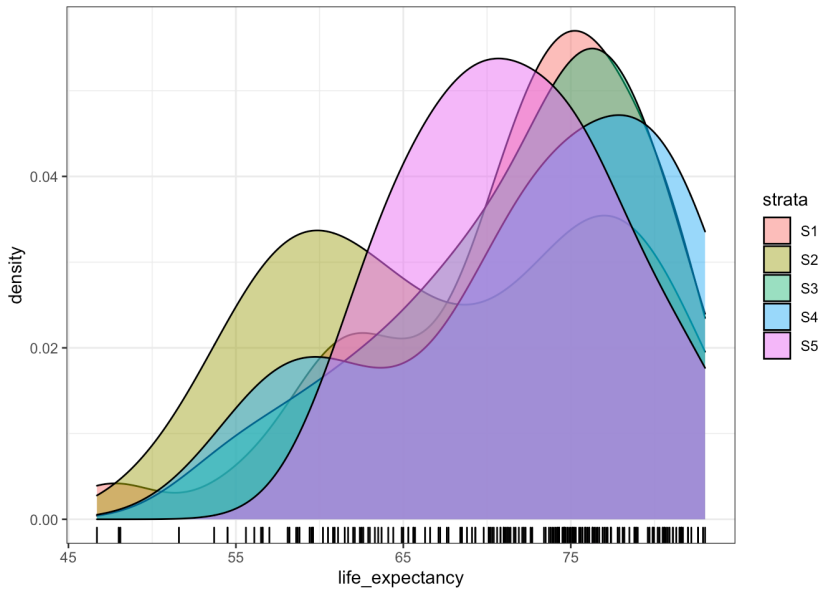
```
LE.1 <- LE.1 |> mutate(strata = ifelse(population<10000000,"S1",
    ifelse(population<25000000,"S2", ifelse(population<50000000,"S3",
    ifelse(population<100000000,"S4","S5")))))
LE.1 <- LE.1[order(LE.1$population),]
LE.1$strata <- as.factor(LE.1$strata)

# number of countries in each stratum
(strata.N <- tapply(LE.1$life_expectancy, LE.1$strata, length))
```

```
 S1  S2  S3  S4  S5
105  35  21  13  11
```

Unfortunately, the life expectancy distributions in each stratum overlap to a great extent: this is not a good sign as it suggests that a country's population is not aligned with its life expectancy.[34]

34: And so that the strata are heterogeneous with respect to life expectancy.

```
ggplot(LE.1,aes(x=life_expectancy,fill=strata)) +
    geom_density(alpha=0.5) + geom_rug()
```




Since there are $N = 185$ observations in the data set, a sample of size $n = 20$, allocated in such a way as to maintain the relative frequencies of the number of observations in each $\mathcal{U}_i$ (this is known as **proportional allocation**), would have the following stratum allocation:

```
N=sum(strata.N)
strata.N/sum(strata.N)*20
```

```
       S1        S2        S3        S4        S5
11.351351  3.783784  2.270270  1.405405  1.189189
```

In practice, we prefer to have at least 2 observations per stratum, so we might use $(n_1, n_2, n_3, n_4, n_5) = (11, 3, 2, 2, 2)$.

```
n=c(11,3,2,2,2)
```

We select a STS sample $\mathcal{Y}$ with these characteristics *via*:

```
set.seed(123456) # replicability
cumul.n = cumsum(n)
cumul.N = cumsum(strata.N)

ind = list()
ind[[1]] <- sample(1:strata.N[1],n[1])
for(j in 2:length(n)){
    ind[[j]] <- cumul.N[j-1] + sample(1:strata.N[j],n[j])
  }

sam.LE.1 <- LE.1[unique(unlist(ind)),]
sam.LE.1 <- sam.LE.1[order(sam.LE.1$population),]
```

Next, we compute the mean $\overline{y}_i$ and the standard deviation $s_i$ in each bucket $\mathcal{Y}_i$, $1 \leq i \leq 5$.

```
y.bar <- list()
std.dev <- list()
y.bar[[1]] <- mean(sam.LE.1[1:n[1],c("life_expectancy")])
std.dev[[1]] <- sd(sam.LE.1[1:n[1],c("life_expectancy")])

for(j in 2:length(n)){
  y.bar[[j]] <- mean(sam.LE.1[(cumul.n[j-1]+1):cumul.n[j],
                     c("life_expectancy")])
  std.dev[[j]] <- sd(sam.LE.1[(cumul.n[j-1]+1):cumul.n[j],
                     c("life_expectancy")])
}

rbind(y.bar,std.dev)
```

```
        [,1]     [,2]     [,3]     [,4]     [,5]
y.bar   70.83636 71.6     67.55    72.15    76.2
std.dev 7.551327 3.774917 18.45549 2.757716 9.050967
```

There is not much variation in the means, but the standard deviation values are all over the place: this is due to small sample sizes in some strata, and overlapping distributions of life expectancy by strata.

As we've already mentioned, **the stratification of countries by population does not align with the estimate of mean life expectancy**. We will continue the STS estimation procedure, for illustration purposes, but in practice, this is the stage at which we would require a different stratification or another sampling plan altogether.

The estimator $\overline{y}_{\text{STS}}$ is:

```
mean.LE.1 <- 0
for(j in 1:length(n)){
    mean.LE.1 <- mean.LE.1 +
        as.numeric(strata.N[j])*y.bar[[j]]
}
(mean.LE.1 <- mean.LE.1/N)
```

```
[1] 71.01902
```

This is fairly close to the true mean $\mu$. The bound on the error of estimation $\hat{B}_{\mu;\text{STS}}$ is:

```
B=0
for(j in 1:length(n)){
    B <- B + as.numeric((strata.N[j]/N)^2*
        std.dev[[j]]^2/n[j]*(1-n[j]/strata.N[j]))
  }
(B <- 2*sqrt(B))
```

```
[1] 3.883388
```

This is quite a large bound, all things considered. The 95% C.I. is thus:

```
c(mean.LE.1 - B, mean.LE.1 + B)
```

```
[1] 67.13563 74.90241
```

Compare with the $\text{C.I.}_{\text{SRS}}(\mu; 0.95)$ obtained previously – the SRS interval was much narrower. This is no doubt due to stratification on the basis of population being a poor choice when dealing with life expectancy.

**Example** Find a 95% confidence interval for the average life expectancy by country in 2011 (including India and China), using a STS of size $n = 20$.[35]

35: This time stratifying the data using the country **life expectations**. In general, we do not stratify with respect to the variable of interest, but with the help of auxiliary variables that are linked to the variable of interest.

We make the appropriate modifications to the code, using the following strata, say:

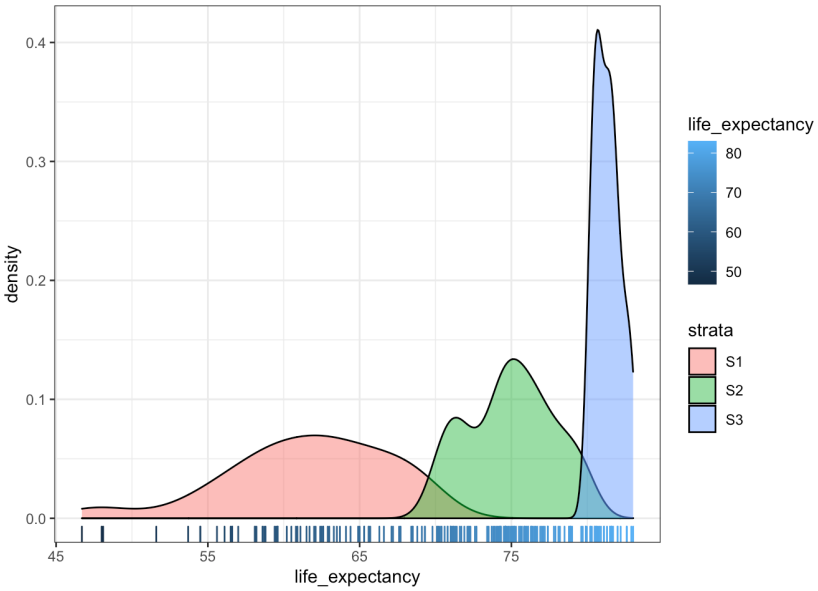
$$\mathcal{U}_1 = \{u_j \mid u_j < 70\}, \quad \mathcal{U}_2 = \{u_j \mid 70 \leq u_j < 80\}, \quad \mathcal{U}_3 = \{u_j \mid u_j \geq 80\}.$$

```
LE.2 <- gapminder |> filter(year==2011) |>  select(life_expectancy)
LE.2 <- LE.2 |> mutate(strata = ifelse(life_expectancy<70,"S1",
                                       ifelse(life_expectancy<80,"S2","S3")))
LE.2 <- LE.2[order(LE.2$life_expectancy),]
LE.2$strata <- as.factor(LE.2$strata)
(strata.N <- tapply(LE.2$life_expectancy, LE.2$strata, length))
```

```
S1 S2 S3
65 93 27
```

By construction, the life expectancy distributions do not overlap from stratum to stratum.

```
ggplot(LE.2,aes(x=life_expectancy,fill=strata)) +
    geom_density(alpha=0.5) +
    geom_rug(aes(color=life_expectancy))
```




Since there are $N = 185$ observations in the data set, with $(N_1, N_2, N_3) = (65, 93, 27)$, a sample of size $n = 20$ could be drawn according to:

```
N=sum(strata.N)
strata.N/sum(strata.N)*20
```

```
      S1        S2        S3
7.027027 10.054054  2.918919
```

We will use $(n_1, n_2, n_3) = (7, 10, 3)$.

```
n=c(7,10,3)
```

The rest of the code runs as in the previous example.

```
cumul.n = cumsum(n)
cumul.N = cumsum(strata.N)

set.seed(123456) # replicability
ind = list()
ind[[1]] <- sample(1:strata.N[1],n[1])
```

```
for(j in 2:length(n)){
    ind[[j]] <- cumul.N[j-1] + sample(1:strata.N[j],n[j])
  }

sam.LE.2 <- LE.2[unique(unlist(ind)),]
sam.LE.2 <- sam.LE.1[order(sam.LE.2$life_expectancy),]

y.bar <- list()
std.dev <- list()
y.bar[[1]] <- mean(sam.LE.2[1:n[1],c("life_expectancy")])
std.dev[[1]] <- sd(sam.LE.2[1:n[1],c("life_expectancy")])

for(j in 2:length(n)){
  y.bar[[j]] <- mean(sam.LE.2[(cumul.n[j-1]+1):cumul.n[j],
                    c("life_expectancy")])
  std.dev[[j]] <- sd(sam.LE.2[(cumul.n[j-1]+1):cumul.n[j],
                    c("life_expectancy")])
}
```

With this sample $\mathcal{Y}$, the strata means and standard deviations are:

```
rbind(y.bar,std.dev)
```

```
        [,1]     [,2]     [,3]
y.bar   71.5     70.27    74.2
std.dev 8.469553 7.721838 7.277362
```

These quantities are more reasonable than with the previous stratification (why?), but they could change from one STS sample to the next. The values for $\overline{y}_{\text{STS}}$ and $\hat{B}_{\mu;\text{STS}}$ are:

```
mean.LE.2 <- 0
for(j in 1:length(n)){
  mean.LE.2 <- mean.LE.2 +
                as.numeric(strata.N[j])*y.bar[[j]]
}
(mean.LE.2 <- mean.LE.2/N)

B=0
for(j in 1:length(n)){
    B <- B + as.numeric((strata.N[j]/N)^2*
               std.dev[[j]]^2/n[j]*(1-n[j]/strata.N[j]))
}

(B <- 2*sqrt(B))
```

```
[1] 71.27573
[1] 3.35133
```

The estimator is quite close to the true value $\mu = 71.18$, but it is when calculating the bound on the error of estimation that the STS approach proves its superiority. In this case, the 95% C.I. for $\mu$ is:

```
c(mean.LE.2 - B, mean.LE.2 + B)
```

```
[1] 67.92440 74.62706
```

These examples show that stratified sampling can improve SRS estimation, **but that this is not always going to be the case**.

**Total $\tau$**

Most of the work has been done: since the **total** $\tau$ can be re-written as

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

we can estimate the total with a STS using:

$$\hat{\tau}_{\text{STS}} = N\overline{y}_{\text{STS}} = \frac{N}{N}\sum_{i=1}^{M} N_i\overline{y}_i = \sum_{i=1}^{M} N_i\overline{y}_i.$$

It is an **unbiased** estimator of the total since its **expectation** is

$$\text{E}(\hat{\tau}_{\text{STS}}) = \text{E}(N\overline{y}_{\text{STS}}) = N \cdot \text{E}(\overline{y}_{\text{STS}}) = N\mu = \tau.$$

Its **sampling variance** is

$$\text{V}(\hat{\tau}_{\text{STS}}) = \text{V}(N\overline{y}_{\text{STS}}) = N^2 \cdot \text{V}(\overline{y}_{\text{STS}}) = \sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right),$$

assuming that we know the variance $\sigma_i^2$ in each strata $\mathcal{U}_i$, $1 \leq i \leq M$, whence the **bound on the error of estimation** is

$$B_{\tau;\text{STS}} = 2\sqrt{\text{V}(\hat{\tau}_{\text{STS}})} = 2\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)} = N \cdot B_{\mu;\text{STS}}.$$

Since the variances $\sigma_i^2$ are usually unknown, we often use the stratum variances $s_i^2$, with correction factors $\frac{N_i - 1}{N_i}$, $1 \leq i \leq M$. The **approximation of the sampling variance** is thus

$$\hat{\text{V}}(\hat{\tau}_{\text{STS}}) = \hat{\text{V}}(N\overline{y}_{\text{STS}}) = \sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right),$$

whence the **bound on the error of estimation** is

$$B_{\tau;\text{STS}} \approx \hat{B}_{\tau;\text{STS}} = 2\sqrt{\hat{\text{V}}(\hat{\tau}_{\text{STS}})} = 2\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)} = N \cdot \hat{B}_{\mu;\text{STS}},$$

and the **approximate 95% C.I. for** $\tau$ is

$$\text{C.I.}_{\text{STS}}(\tau; 0.95): \quad \hat{\tau}_{\text{STS}} \pm \hat{B}_{\tau;\text{STS}}.$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, split into two strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of sizes $N_1 = 21,123$ and $N_2 = 16,321$, respectively. A STS $\mathcal{Y}$ of size $n = 132$ is drawn from $\mathcal{U}$, with $n_1 = 82$ and $n_2 = 50$.

Suppose the empirical mean and standard deviation in $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are $\overline{y}_1 = 120.7$, $\overline{y}_2 = 96.6$, $s_1 = 18.99$, and $s_2 = 14.31$, respectively. Find a 95% C.I. of the total $\tau$ in $\mathcal{U}$.

The bound on the error of estimation is $\approx \hat{B}_{\tau;\text{STS}} = 2\sqrt{\hat{\text{V}}(\hat{\tau}_{\text{STS}})}$:

$$2\sqrt{21123^2 \cdot \frac{18.99^2}{82}\left(1 - \frac{82}{21123}\right) + 16321^2 \cdot \frac{14.31^2}{50}\left(1 - \frac{50}{16321}\right)} \approx 110312.3;$$

$$\text{C.I.}_{\text{STS}}(\tau; 0.95) \approx 21123(120.7) + 16321(96.6) \pm 110312.3 \approx (4015842, 4236467).$$

**Proportion** $p$

If the response $u_{i,\ell} \in \{0, 1\}$ represents the absence or the presence of a certain characteristic for the $\ell$th unit in the $i$th strata $\mathcal{U}_i$, the **mean**

$$p = \mu = \frac{1}{N}\sum_{i=1}^{M}\sum_{\ell=1}^{N_i} u_{i,\ell}$$

is the **proportion** of all units in $\mathcal{U}$ which posess the characteristic. This proportion can be estimated with a STS *via*

$$\hat{p}_{\text{STS}} = \frac{1}{N}\sum_{i=1}^{M} N_i \hat{p}_i, \quad \text{where } \hat{p}_i = \frac{1}{n_i}\sum_{\ell=1}^{n_i} u_{i,\ell},\ 1 \le i \le M.$$

This is an unbiased estimator of $p$ since

$$\text{E}(\hat{p}_{\text{STS}}) = \text{E}(\overline{y}_{\text{STS}}) = \mu = p;$$

its **sampling variance** is:

$$\begin{aligned}\text{V}(\hat{p}_{\text{STS}}) = \text{V}(\overline{y}_{\text{STS}}) &= \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{p_i(1-p_i)}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right),\end{aligned}$$

where $\sigma_i^2 = p_i(1 - p_i)$ is the variance of the response variable $u$ in the stratum $\mathcal{U}_i$.

The **bound on the error of estimation** is

$$B_{p;\text{STS}} = 2\sqrt{\text{V}(\hat{p}_{\text{STS}})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{p_i(1-p_i)}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right)}.$$

Since the proportions $p_i$ are not usually known, the **approximate sampling variance** is used instead:

$$\hat{\text{V}}(\hat{p}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}\left(1 - \frac{n_i}{N_i}\right).$$

The **approximate bound on the error of estimation** is thus

$$B_{p;\text{STS}} \approx \hat{B}_{p;\text{STS}} = 2\sqrt{\hat{\text{V}}(\hat{p}_{\text{STS}})} = \frac{2}{N}\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}\left(1 - \frac{n_i}{N_i}\right)},$$

and the corresponding **approximate 95% C.I. for** $p$ is

$$\text{C.I.}_{\text{STS}}(p; 0.95): \quad \hat{p}_{\text{STS}} \pm \frac{2}{N}\sqrt{\sum_{i=1}^{M} N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}\left(1 - \frac{n_i}{N_i}\right)}.$$

If the sample size in a stratum is too small, we can use the conservative estimate $\hat{p}_i = 0.5$.

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, split into two strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of sizes $N_1 = 21,123$ and $N_2 = 16,321$, respectively. A STS $\mathcal{Y}$ of size $n = 132$ is drawn from $\mathcal{U}$, with $n_1 = 82$ and $n_2 = 50$.

Suppose that $n_1 = 20$ of the observations from $\mathcal{Y}_1$ and $n_2 = 5$ of the observations from $\mathcal{Y}_2$ possess a certain characteristic. Find a 95% C.I. for the proportion $p$ of the units in $\mathcal{U}$ that possess the characteristic.

In this case, $\hat{p}_1 = 20/82 \approx 0.244$ and $\hat{p}_2 = 5/50 = 0.10$, from which we obtain

$$\hat{p}_{\text{STS}} = \frac{21123}{37444}(0.244) + \frac{16321}{37444}(0.10) = 0.181.$$

The bound on the error of estimation is thus

$$\hat{B}_p = \tfrac{2}{37444}\sqrt{21123^2 \tfrac{0.244(1-0.244)}{82-1}\left(1 - \tfrac{82}{21123}\right) + 16321^2 \tfrac{0.1(1-0.1)}{50-1}\left(1 - \tfrac{50}{16321}\right)} \approx 0.0654,$$

from which we conclude that

$$\text{C.I.}(p; 0.95) \approx 0.181 \pm 0.0654 \equiv (0.116, 0.247).$$

### 10.4.2 Sample Size and Allocation

When determining the size of a STS sample $\mathcal{Y}$, we must also consider the problem of **allocating the number of units $n_i$ in each stratum $\mathcal{Y}_i$**. If $|\mathcal{Y}_i| = n_i$, $1 \le i \le M$, then $n = n_1 + \cdots + n_M$. But what are the $n_i$?

In a STS, the sampling variance of the estimator $\overline{y}_{\text{STS}}$ is

$$\text{V}(\overline{y}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right).$$

When $N_i \gg 1$, then $N_i \approx N_i - 1$ and so

$$\text{V}(\overline{y}_{\text{STS}}) \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^{M} N_i \sigma_i^2.$$

Since the sampling variance $\text{V}(\overline{y}_{\text{STS}})$ determines the bound on the error of estimation $\hat{B}_{\mu;\text{STS}}$, we can minimize the bound (and thus the error) **by minimizing the sampling variance**. The quantities $N$, $N_i$, $\sigma_i^2$, are fixed

for $1 \leq i \leq M$; what we minimize against is the sample size $n$ and the allocation $n_i$ in each stratum.

The **total cost of the survey** $\tilde{C}$ can also affect the allocation. The survey budget includes the **overhead cost** (indirect costs) $c_0$ and the **cost per response** $c_i$ in each stratum $\mathcal{U}_i$, $1 \leq i \leq M$. The total cost is thus

$$\tilde{C} = c_0 + \sum_{i=1}^{M} c_i n_i,$$

which must remain below than **available survey budget** $C$. The allocation problem is an optimization problem: we seek to solve

$$\arg_{(n,n_1,\ldots,n_M)} \min \mathrm{V}(\overline{y}_{\mathrm{STS}}), \quad \text{subject to } \tilde{C} \leq C.$$

We use the method of **Lagrange multipliers**. The objective function becomes

$$\begin{aligned} f(n_1,\ldots,n_M,\lambda) &= \mathrm{V}(\overline{y}_{\mathrm{STS}}) + \lambda(\tilde{C} - C) \\ &= \frac{1}{N^2}\sum_{k=1}^{M} N_i^2 \cdot \frac{\sigma_k^2}{n_k} - \frac{1}{N^2}\sum_{k=1}^{M} N_k\sigma_k^2 + \lambda(c_0 + \sum_{k=1}^{M} c_k n_k - C). \end{aligned}$$

Its critical points solve

$$\begin{aligned} 0 = \frac{\partial f(n_1,\ldots,n_M,\lambda)}{\partial n_i} &= \frac{1}{N^2}\sum_{k=1}^{M} N_k^2\sigma_k^2 \frac{\partial(1/n_k)}{\partial n_i} + \lambda \sum_{k=1}^{M} c_k \frac{\partial(n_k)}{\partial n_i} \\ &= -\frac{N_i^2\sigma_i^2}{N^2 n_i^2} + \lambda c_i, \quad 1 \leq i \leq M, \end{aligned}$$

which is to say that

$$n_i = \frac{N_i\sigma_i}{N\sqrt{\lambda}\sqrt{c_i}}, \quad 1 \leq i \leq M.$$

The **strata sampling weights** $w_i$ are

$$w_i = \frac{n_i}{n_1 + \cdots + n_M}, \quad 1 \leq i \leq M.$$

The **general optimal allocation** is thus

$$w_i = \frac{n_i}{n} = \frac{\dfrac{N_i\sigma_i}{N\sqrt{\lambda}\sqrt{c_i}}}{\displaystyle\sum_{k=1}^{M} \frac{N_k\sigma_k}{N\sqrt{\lambda}\sqrt{c_k}}} = \frac{\dfrac{N_i\sigma_i}{\sqrt{c_i}}}{\displaystyle\sum_{k=1}^{M} \frac{N_k\sigma_k}{\sqrt{c_k}}}, \quad 1 \leq i \leq M.$$

Once we have determined the size $n$ of the sample $\mathcal{Y}$, we compute the size of the sample $n_i$ in each $\mathcal{Y}_i$ using $w_i \cdot n$, $1 \leq i \leq M$. Since the product $w_i \cdot n$ is not typically an integer, we allocate $[w_i \cdot n]$ units to each $\mathcal{Y}_i$,[36] and distribute the remaining

$$n - [w_1 \cdot n] - \cdots - [w_M \cdot n]$$

units using "common sense" (while ensuring that $\tilde{C} \leq C$).

36: The **integer part** $[x]$ of $x$ is the largest integer smaller than $x$.

If the cost per response in each stratum is constant, $c_1 = \cdots = c_M$, **Neyman allocation** yields the following stratum sampling weights:

$$w_i = \frac{n_i}{n} = \frac{N_i\sigma_i}{N_1\sigma_1 + \cdots + N_M\sigma_M}, \quad 1 \le i \le M.$$

If moreover the variance is the same in each stratum, $\sigma_1^2 = \cdots = \sigma_M^2$, **proportional allocation** yields the following stratum sampling weights:

$$w_i = \frac{n_i}{n} = \frac{N_i}{N_1 + \cdots + N_M} = \frac{N_i}{N}, \quad 1 \le i \le M.$$

Once the sample size and allocation have been selected, the methods in the previous section can be used to provide confidence intervals for the mean $\mu$, for the total $\tau$, or for a proportion $p$. When the variances are unknown, the usual approximations can be used.

We may at times use allocation schemes that are not necessarily **ideal** from a technical perspective, but which facilitate the preparation of reports or the dissemination of results:

$$w_i = \frac{n_i}{n} = \frac{f(N_i)}{f(N_1) + \cdots + f(N_M)}, \quad 1 \le i \le M, \ f \text{ a random function.}$$

For instance, when studying Canadian populations, we often stratify according to the provinces and use $f(x) = \sqrt{x}$; the proportional allocation and square root allocation sampling weights for the 13 Canadian jurisdictions (based on 2022 population data) are shown below.

| Juridiction | Prop. | Racine | Juridiction | Prop. | Racine |
|---|---|---|---|---|---|
| Ontario | 38.26% | 22.4% | Nouvelle-Ecosse | 2.63% | 5.9% |
| Québec | 23.23% | 17.4% | Nouveau-Brunswick | 2.13% | 5.3% |
| Colombie-Britannique | 13.22% | 13.2% | Terre-Neuve-et-Labrador | 1.48% | 4.4% |
| Alberta | 11.57% | 12.3% | Ile-du-Prince-Edward | 0.41% | 2.3% |
| Manitoba | 3.64% | 6.9% | Territoires-du-Nord-Ouest | 0.12% | 1.2% |
| Saskatchewan | 3.12% | 6.4% | Yukon | 0.10% | 1.2% |
| | | | Nunavut | 0.10% | 1.2% |

**Table 10.2:** Sampling weights for Canadian provinces, under proportional allocation and square root allocation (racine, in French).

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. We seek to estimate the mean $\mu$ of $\mathcal{U}$ using a STS. The survey budget allows for a sample size $n = 132$.

In a preliminary study, we estimated $\sigma_1 \approx 20$ and $\sigma_2 \approx 15$. If the cost of a response in the first stratum is four times that of the cost of a response in the second stratum, find the general optimal allocation. If the response cost per stratum is constant, determine the Neyman and the proportional allocations.

In the general case, we have $c_1 = 4c_2$,

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{4c_2}} = \frac{211230}{\sqrt{c_2}}, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{c_2}} = \frac{244815}{\sqrt{c_2}},$$

and

$$\frac{N_1\sigma_1}{\sqrt{c_1}} + \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{211230}{\sqrt{c_2}} + \frac{244815}{\sqrt{c_2}} = \frac{456045}{\sqrt{c_2}},$$

from which we conclude that

$$n_1 = 132\left(\frac{211230}{456045}\right) = 61.13 \quad \text{and} \quad n_2 = 132\left(\frac{244815}{456045}\right) = 70.87;$$

the general optimal allocation is thus $n_1 = 61$ and $n_2 = 71$.

If the cost for a response is the same in both strata, $c_1 = c_2$, then:

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$

and

$$N_1\sigma_1 + N_2\sigma_2 = 422460 + 244815 = 667275,$$

from which we conclude that

$$n_1 = 132\left(\frac{422460}{667275}\right) = 83.57 \quad \text{and} \quad n_2 = 132\left(\frac{244815}{667275}\right) = 48.43;$$

the Neyman allocation is thus $n_1 = 84$ and $n_2 = 48$.

If we do not trust the study conducted beforehand, and we assume that the variance is constant in each stratum ($\sigma_1 = \sigma_2$), then we have

$$N_1 = 21123, \quad N_2 = 16321, \quad \text{and} \quad N_1 + N_2 = 21123 + 16321 = 37444,$$

from which we conclude that

$$n_1 = 132\left(\frac{21123}{37444}\right) = 74.46 \quad \text{and} \quad n_2 = 132\left(\frac{16321}{37444}\right) = 57.54;$$

the proportional allocation is thus $n_1 = 74$ and $n_2 = 58$. ■

**Sample Size, Given a Bound on the Error of Estimation**

In theory, only **analytical considerations** should influence the sample size. Recall that in a STS of size $n$, the sampling weight corresponding to the stratum $\mathcal{U}_i$ is $w_i = \frac{n_i}{n}$, for $1 \le i \le M$. When we estimate $\mu$ *via* $\overline{y}_{\text{STS}}$, the bound on the error of estimation can be written

$$B_{\mu;\text{STS}} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{w_i \cdot n}\left(\frac{N_i - w_i \cdot n}{N_i - 1}\right)}.$$

We seek to express $n$ in terms of the parameters $N_i$, $\sigma_i$, $w_i$, and $B_{\mu;\text{STS}}$. If $N_i \gg 1$,[37] then $N_i \approx N_i - 1$ and so

37: Which is hopefully the case in practice.

$$\underbrace{\frac{B_{\mu;\text{STS}}^2}{4}}_{=D_{\mu;\text{STS}}} \approx \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{w_i \cdot n}\left(\frac{N_i - w_i \cdot n}{N_i}\right)$$

$$\iff N^2 D_{\mu;\text{STS}} \approx \frac{1}{n}\left\{\sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i}\right\} - \sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i} \cdot \frac{w_i}{N_i}$$

$$\iff \frac{N^2 D_{\mu;\text{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2}{\sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i}} \approx \frac{1}{n} \iff n_{\mu;\text{STS}} \approx \frac{\sum_{i=1}^{M} \frac{N_i^2\sigma_i^2}{w_i}}{N^2 D_{\mu;\text{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2}.$$

Under **general optimal allocation**, the stratum sampling weights are given by

$$w_i = \frac{N_i\sigma_i}{\sqrt{c_i}}\left(\sum_{k=1}^{M}\frac{N_k\sigma_k}{\sqrt{c_k}}\right)^{-1}, \quad 1 \le i \le M,$$

and the sample size is then

$$n_{\mu;\mathrm{STS}} \approx \frac{\left(\sum_{i=1}^{M}\frac{N_i^2\sigma_i^2}{N_i\sigma_i/\sqrt{c_i}}\right) \div \left(\sum_{k=1}^{M}\frac{N_k\sigma_k}{\sqrt{c_k}}\right)^{-1}}{N^2 D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2} = \frac{\left(\sum_{i=1}^{M} N_i\sigma_i\sqrt{c_i}\right)\left(\sum_{i=1}^{M}\frac{N_i\sigma_i}{\sqrt{c_i}}\right)}{N^2 D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2}$$

Under **Neyman allocation**, the stratum sampling weights are given by

$$w_i = N_i\sigma_i\left(\sum_{k=1}^{M} N_k\sigma_k\right)^{-1}, \quad 1 \le i \le M,$$

and the sample size is then

$$n_{\mu;\mathrm{STS}} \approx \frac{\left(\sum_{i=1}^{M}\frac{N_i^2\sigma_i^2}{N_i\sigma_i}\right) \div \left(\sum_{k=1}^{M} N_k\sigma_k\right)^{-1}}{N^2 D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2} = \frac{\left(\sum_{i=1}^{M} N_i\sigma_i\right)^2}{N^2 D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2}$$

In a **proportional allocation** scenario, the stratum sampling weights are given by

$$w_i = N_i\left(\sum_{k=1}^{M} N_k\right)^{-1}, \quad 1 \le i \le M,$$

and the sample size is then

$$n_{\mu;\mathrm{STS}} \approx \frac{\left(\sum_{i=1}^{M}\frac{N_i^2\sigma_i^2}{N_i}\right) \div \left(\sum_{k=1}^{M} N_k\right)^{-1}}{N^2 D_{\mu;\mathrm{STS}} + \sum_{i=1}^{M} N_i\sigma_i^2} = \frac{\sum_{i=1}^{M} N_i\sigma_i^2}{N D_{\mu;\mathrm{STS}} + \frac{1}{N}\sum_{i=1}^{M} N_i\sigma_i^2}$$

When we try to estimate the total $\tau$ using the estimator $\hat{\tau}_{\mathrm{STS}}$, we must substitute

$$D_{\mu;\mathrm{STS}} = \frac{B_{\mu;\mathrm{STS}}^2}{4} \quad \text{by} \quad D_{\tau;\mathrm{STS}} = \frac{B_{\tau;\mathrm{STS}}^2}{4N^2}.$$

When we want to estimate a proportion $p$ using the estimator $\hat{p}_{\mathrm{STS}}$, the bound remains

$$D_{p;\mathrm{STS}} = \frac{B_{p;\mathrm{STS}}^2}{4},$$

but we have to substitute the stratum variances $\sigma_i^2$ by $p_i(1-p_i)$. The proportions $p_i$ can be estimated with the help of a previous study, or, **conservatively**, by using $p_i = 0.5$.

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. We seek to estimate the mean $\mu$ of $\mathcal{U}$ using a STS, with a bound on the error of estimation of $B_{\mu;\text{STS}} = 5$. The response costs by stratum are $c_1 = 400\$$ and $c_2 = 100\$$.

In a preliminary study, we estimated $\sigma_1 \approx 20$ and $\sigma_2 \approx 15$. Determine the sample size and allocation in each of the three scenarios: general optimal allocation, Neyman allocation, and proportional allocation (in the last two cases, use $c_1 = c_2 = 100\$$).

In the general case, we have

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{400}} = 21123, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{100}} = 24481.5,$$
$$N_1\sigma_1\sqrt{c_1} = 21123(20)\sqrt{400} = 8449200, \quad N_2\sigma_2\sqrt{c_2} = 16321(15)\sqrt{100} = 2448150$$
$$N_1\sigma_1^2 = 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225,$$
$$\sum_{i=1}^{2}\frac{N_i\sigma_i}{\sqrt{c_i}} = 45604.5, \quad \sum_{i=1}^{2}N_i\sigma_i\sqrt{c_i} = 10897350, \quad \sum_{i=1}^{2}N_i\sigma_i^2 = 12121425,$$
$$D_{\mu;\text{STS}} = \frac{5^2}{4} = 6.25, \quad n = \frac{(10897350)(45604.5)}{(37444)^2(6.25) + 12121425} = 56.63 \approx 57$$
$$n_1 = 57\left(\frac{21123}{45604.5}\right) = 26.4 \approx 26, \quad n_2 = 57\left(\frac{24481.5}{45604.5}\right) = 30.6 \approx 31.$$

If instead the response cost per stratum is constant ($c_1 = c_2 = 100$), we have:

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$
$$N_1\sigma_1^2 = 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225,$$
$$\sum_{i=1}^{2}N_i\sigma_i = 667275, \quad \sum_{i=1}^{2}N_i\sigma_i^2 = 12121425,$$
$$D_{\mu;\text{STS}} = \frac{5^2}{4} = 6.25, \quad n = \frac{(667275)^2}{(37444)^2(6.25) + 12121425} = 50.74 \approx 51$$
$$n_1 = 51\left(\frac{422460}{667275}\right) = 32.30 \approx 32, \quad n_2 = 51\left(\frac{244815}{667275}\right) = 18.71 \approx 19.$$

It turns out that the exact value of $c_1 = c_2$ does not come into play.

If we look for a proportional allocation, we still have

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$
$$N_1\sigma_1^2 = 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225,$$
$$\sum_{i=1}^{2}N_i\sigma_i = 667275, \quad \sum_{i=1}^{2}N_i\sigma_i^2 = 12121425,$$
$$D_{\mu;\text{STS}} = \frac{5^2}{4} = 6.25, \quad n = \frac{12121425}{37444(6.25) + \frac{12121425}{37444}} = 51.72 \approx 52$$
$$n_1 = 52\left(\frac{21123}{37444}\right) = 29.33 \approx 29, \quad n_2 = 52\left(\frac{16321}{37444}\right) = 22.67 \approx 23.$$

The exact value of $c_1 = c_2$ also does not come into play. ■

### Sample Size, Given a Budget

In practice, however, it is often **budgetary considerations** that play the most important role in sample size selection.

In a STS of size $n$, the stratum sampling weights are $w_i = \frac{n_i}{n}$, for $1 \leq i \leq M$. In this case, we seek to **maximize the size $n$ allowed by the survey budget $C$**:

$$C = c_0 + \sum_{i=1}^{M} c_i n_i = c_0 + n \sum_{i=1}^{M} c_i w_i \implies n = \frac{C - c_0}{\sum_{i=1}^{M} c_i w_i}.$$

In a **general optimal allocation** scenario, we have

$$w_i = \frac{N_i \sigma_i}{\sqrt{c_i}} \left( \sum_{k=1}^{M} \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M,$$

from which we see that

$$c_i w_i = c_i \cdot \frac{N_i \sigma_i}{\sqrt{c_i}} \left( \sum_{k=1}^{M} \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1} = N_i \sigma_i \sqrt{c_i} \left( \sum_{k=1}^{M} \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M;$$

the sample size is then

$$n_{\text{STS}} = (C - c_0) \left( \sum_{i=1}^{M} \frac{N_i \sigma_i}{\sqrt{c_i}} \right) \left( \sum_{i=1}^{M} N_i \sigma_i \sqrt{c_i} \right)^{-1}.$$

In a **Neyman allocation** or **proportional allocation** scenario, the sample weights are

$$w_i = N_i \sigma_i \left( \sum_{k=1}^{M} N_k \sigma_k \right)^{-1} \quad 1 \leq i \leq M,$$

from which we see that

$$c_i w_i = c \cdot N_i \sigma_i \left( \sum_{k=1}^{M} N_k \sigma_k \right)^{-1} \quad 1 \leq i \leq M;$$

the sample size is then

$$n_{\text{STS}} = (C - c_0) \left( \sum_{i=1}^{M} N_i \sigma_i \right) \left( c \sum_{i=1}^{M} N_i \sigma_i \right)^{-1} = \frac{C - c_0}{c}.$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, separated in two disjoint strata $\mathcal{U}_1$ and $\mathcal{U}_2$, of respective sizes $N_1 = 21,123$ and $N_2 = 16,321$. We seek to estimate the mean $\mu$ of $\mathcal{U}$ using a STS. The budget for the study is $C = 20,000\$$, minus $c_0 = 4,000\$$ for overhead costs. The cost of a response in each stratum are $c_1 = 400\$$ and $c_2 = 100\$$, respectively.

In a preliminary study, we estimate $\sigma_1 = 20$ and $\sigma_2 = 15$. Determine the sample size and allocation in each of the three scenarios: general optimal

allocation, Neyman allocation, and proportional allocation (in the last two cases, use $c_1 = c_2 = 100\$$).

In the general case, we have

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{400}} = 21123, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{100}} = 24481.5,$$
$$N_1\sigma_1\sqrt{c_1} = 21123(20)\sqrt{400} = 8449200,$$
$$N_2\sigma_2\sqrt{c_2} = 16321(15)\sqrt{100} = 2448150$$
$$\frac{N_1\sigma_1}{\sqrt{c_1}} + \frac{N_2\sigma_2}{\sqrt{c_2}} = 21123 + 24481.5 = 45604.5,$$
$$N_1\sigma_1\sqrt{c_1} + N_2\sigma_2\sqrt{c_2} = 8449200 + 2448150 = 10897350,$$
$$n = (20000 - 4000)\left(\frac{45604.5}{10897350}\right) = 66.96 \approx 66,$$
$$n_1 = 66\left(\frac{21123}{45604.5}\right) = 30.56 \approx 31, \quad n_2 = 66\left(\frac{24481.5}{45604.5}\right) = 35.43 \approx 35.$$

If the response cost per stratum is constant ($c_1 = c_2 = 100$):

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$
$$N_1\sigma_1 + N_2\sigma_2 = 422460 + 244815 = 667275,$$
$$n = \frac{20000 - 4000}{100} = 160,$$
$$n_1 = 160\left(\frac{422460}{667275}\right) = 101.3 \approx 101, \quad n_2 = 160\left(\frac{244815}{667275}\right) = 58.7 \approx 59.$$

If we also assume that the variances are equal in the 2 strata, the sample size remains $n = 160$, but the proportional allocation yields

$$n_1 = 160\left(\frac{21123}{37444}\right) = 90.25 \approx 90 \quad \text{and} \quad n_2 = 160\left(\frac{16321}{37444}\right) = 69.74 \approx 70. \quad \blacksquare$$

### 10.4.3 Comparison Between SRS and STS

Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ have mean $\mu$ and variance $\sigma^2$.

Using a SRS of size $n$, we can construct the estimator $\overline{y}_{\text{SRS}}$, with sampling variance

$$\text{V}(\overline{y}_{\text{SRS}}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

We have studied the properties of such estimators in section 10.3.

If $\mathcal{U}$ can be split into $M$ strata

$$\mathcal{U}_1 = \{u_{1,1}, \ldots, u_{1,N_1}\}, \cdots, \mathcal{U}_M = \{u_{M,1}, \ldots, u_{M,N_M}\},$$

with mean and variance $\mu_i$ and $\sigma_i^2$, respectively, for $1 \leq i \leq M$.

Using a STS of size $n = (n_1, \ldots, n_M)$, we can construct the estimator $\overline{y}_{\text{STS}}$, with sampling variance

$$\text{V}(\overline{y}_{\text{STS}}) = \frac{1}{N^2}\sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right).$$

Both samples have the same size; is there any way to determine which of the two approaches is preferable **before** computing the confidence intervals? In general, the sample design for which the **sampling variance** of the corresponding estimator is **smallest** is preferred.[38]

38: This corresponds to a tighter (smaller) C.I.

If $N \gg n$ and $N_i \gg n_i$ for all $1 \leq i \leq M$, then $N - n \approx N - 1$ and $N_i - n_1 \approx N_i - 1$ for all $1 \leq i \leq M$. Consequently,

$$\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \approx \frac{\sigma^2}{n} = \frac{1}{nN} \sum_{i=1}^{M} \sum_{j=1}^{N_i} (u_{i,j} - \mu)^2 \quad \text{and} \quad \mathrm{V}(\overline{y}_{\mathrm{STS}}) \approx \frac{1}{N^2} \sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2}{n_i}.$$

In a proportional allocation scenario, $n_i = n \cdot \frac{N_i}{N}$ for all $1 \leq i \leq M$, from which we see that

$$\mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}} \approx \frac{1}{N^2} \sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2 \cdot N}{n N_i} = \frac{1}{nN} \sum_{i=1}^{M} N_i \sigma_i^2.$$

In a Neyman allocation scenario, $n_i = n \cdot \frac{N_i \sigma_i}{N_1 \sigma_1 + \cdots + N_M \sigma_M}$ for all $1 \leq i \leq M$, from which we see that

$$\mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Neyman}} \approx \frac{1}{N^2} \sum_{i=1}^{M} N_i^2 \cdot \frac{\sigma_i^2 \left( \sum_{k=1}^{M} N_k \sigma_k \right)}{n N_i \sigma_i} = \frac{1}{nN^2} \left( \sum_{i=1}^{M} N_i \sigma_i \right)^2.$$

But

$$\begin{aligned}
\mathrm{V}(\overline{y}_{\mathrm{SRS}}) &\approx \frac{1}{nN} \sum_{i=1}^{M} \sum_{j=1}^{N_i} (u_{i,j} - \mu)^2 = \frac{1}{nN} \sum_{i=1}^{M} \sum_{j=1}^{N_i} (u_{i,j} - \mu_i + \mu_i - \mu)^2 \\
&= \frac{1}{nN} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \left\{ (u_{i,j} - \mu_i)^2 + 2(u_{i,j} - \mu_i)(\mu_i - \mu) + (\mu_i - \mu)^2 \right\} \\
&= \frac{1}{nN} \left\{ \sum_{i=1}^{M} \underbrace{\sum_{j=1}^{N_i} (u_{i,j} - \mu_i)^2}_{N_i \sigma_i^2} + 2 \sum_{i=1}^{M} (\mu_i - \mu) \underbrace{\sum_{j=1}^{N_i} (u_{i,j} - \mu_i)}_{N_i \mu_i - N_i \mu_i = 0} + \sum_{i=1}^{M} (\mu_i - \mu)^2 \underbrace{\sum_{j=1}^{N_i} 1}_{N_i} \right\} \\
&= \frac{1}{nN} \left\{ \sum_{i=1}^{M} N_i \sigma_i^2 + \sum_{i=1}^{M} N_i (\mu_i - \mu)^2 \right\} = \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}} + \frac{1}{nN} \sum_{i=1}^{M} N_i (\mu_i - \mu)^2.
\end{aligned}$$

As such,

$$\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \gg \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}}, \quad \text{whenever } \frac{1}{nN} \sum_{i=1}^{M} N_i (\mu_i - \mu)^2 \gg 0;$$

a STS under proportional allocation is substantially preferable to a SRS when **the variance of the stratum means is high**.

Similarly, set

$$\overline{\sigma} = \frac{1}{N} \sum_{i=1}^{M} N_i \sigma_i = \sqrt{n \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Neyman}}}.$$

As such,

$$\begin{aligned} \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Prop}} - \mathrm{V}(\overline{y}_{\mathrm{STS}})_{\mathrm{Neyman}} &= \frac{1}{nN}\sum_{i=1}^{M} N_i\sigma_i^2 - \frac{\overline{\sigma}^2}{n} \\ &= \frac{1}{nN}\left\{\sum_{i=1}^{M} N_i\sigma_i^2 - N\overline{\sigma}^2\right\} \\ &= \frac{1}{nN}\sum_{i=1}^{M} N_i(\sigma_i^2 - 2\sigma_i\overline{\sigma} + \overline{\sigma}^2) \\ &= \frac{1}{nN}\sum_{i=1}^{M} N_i(\sigma_i - \overline{\sigma})^2 \geq 0; \end{aligned}$$

a STS under Neyman allocation is substantially preferable to a STS under proportional allocation **when the variance of the stratum standard deviations is high**.

Combining these, we can conclude that a STS under Neyman allocation is substantially preferable to a SRS when **stratum means and standard deviations vary greatly across strata**.

Since in practice there are other considerations at play (sampling cost, etc.), one may still decide in favor of a SRS or a STS under proportional allocation, especially if the difference in the corresponding variances is (relatively) small.

## 10.5 Using Auxiliary Information

In what follows we present ways to obtain estimates of the mean, the total, or of a proportion with the help of **auxiliary information**. So far, we have only discussed **univariate** SRS and STS estimators. Can we use more than one response per unit to obtain better approximations?

In the 2011 `Gapminder` dataset, there are $N = 168$ countries in 2011 for which the **life expectancy** $Y$ and the (logarithm of the) **gross domestic product per capita** $X$ are available. Suppose it is known that $\mathrm{E}[X] = \mu_X = 7.84$. If we draw a sample $\{(x_1, y_1), \ldots, (x_{10}, y_{10})\} \subseteq \mathcal{U}$ for which the mean of $y_i/x_i$ is 8.67, can we expect that $\mu_Y \approx 8.67\mu_X = 68.00$?[39]

39: See Figure 10.6.

### 10.5.1 Ratio Estimation

Let $\mathcal{U} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ be a finite population of size $N$ for which each unit $u_j$ has 2 observed values: $X_j$ and $Y_j$. The **ratio of the means** $R$ is the ratio of the means (or totals):

$$R = \frac{\sum_{j=1}^{N} Y_j}{\sum_{j=1}^{N} X_j} = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}, \quad \text{as long as } \mu_X, \tau_X \neq 0.$$

We are interested in such quotients when we try to determine the average wage $Y$ as a function of years of schooling $X$ in Canada, for example.

### Ratio Estimator

Let $\mathcal{Y} = \{(x_{i_1}, y_{i_1}), \ldots, (x_{i_n}, y_{i_n})\} \subseteq \mathcal{U}$ a **bivariate simple random sample** of size $n$. We often simplify the notation by writing

$$\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

The **sample ratio of means** $r$ is an estimator of $R$:

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\overline{y}}{\overline{x}} = \frac{\hat{\tau}_Y}{\hat{\tau}_X}, \quad \text{as long as } \overline{x}, \hat{\tau}_X \neq 0.$$

**Warning:** this is a biased estimator!

**Example** Consider a finite bivariate population with $N = 4$ units:

$$u_1 = (1, 2), \quad u_2 = (1, 0), \quad u_3 = (2, 1), \quad u_4 = (4, 5).$$

The population ratio of means $R$ is simply

$$R = \frac{2 + 0 + 1 + 5}{1 + 1 + 2 + 4} = \frac{8}{8} = 1.$$

Suppose that we want to provide an estimate of $R$ by drawing a SRS of size $n = 3$ from $\mathcal{U}$. There are $\binom{4}{3} = 4$ such samples.

| Sample | $y$ Values | $\overline{y}$ | $x$ Values | $\overline{x}$ | $r$ | $P(r)$ |
|---|---|---|---|---|---|---|
| $u_1, u_2, u_3$ | 2, 0, 1 | 1 | 1, 1, 2 | 4/3 | 3/4 | 1/4 |
| $u_1, u_2, u_4$ | 2, 0, 5 | 7/3 | 1, 1, 4 | 2 | 7/6 | 1/4 |
| $u_1, u_3, u_4$ | 2, 1, 5 | 8/3 | 1, 2, 4 | 7/3 | 8/7 | 1/4 |
| $u_2, u_3, u_4$ | 0, 1, 5 | 2 | 1, 2, 3 | 2 | 1 | 1/4 |

We can compute the expectation of the estimator $r$ directly:

$$\mathrm{E}[r] = \sum_r rP(r) = \frac{1}{4}(3/4 + 7/6 + 8/7 + 1) = \frac{341}{336} \approx 1.014881 \neq R = 1. \quad \blacksquare$$

What is the **sampling bias** of $r$ as an estimator of $R$, then?

$$\mathrm{E}[r - R] = \mathrm{E}\left[\frac{\overline{y}}{\overline{x}} - R\right] = \mathrm{E}\left[\frac{1}{\overline{x}}(\overline{y} - R\overline{x})\right] = ??$$

### Ratio Estimator Bias

In this last expression for the sampling bias, only $\overline{x}$ and $\overline{y}$ change when the sample changes: $R$ remains constant. But there is no simple expression allowing us to compute exactly the **expectation of a quotient of random variables**; we must use approximations.

Let $f : [a, b] \to \mathbb{R}$ be $C^2$ over $[a, b]$ (i.e., $f, f', f''$ are all continuous over $[a, b]$). According to Taylor's theorem, for all $c \in (a, b)$, there exists a $\xi$ between $c$ and $z$ such that

$$f(z) = f(c) + f'(c)(z - c) + \frac{f''(\xi)}{2}(z - c)^2.$$

Since $f''$ is continuous over $[a, b]$, $f''$ is bounded on $[a, b]$: $\exists M > 0$ such that $|f''(z)| \leq M$ for all $z \in [a, b]$.

Thus, if $z$ is **sufficiently close** to $c$,

$$|f(c) + f'(c)(z - c)| \gg \frac{M}{2}(z - c)^2 \geq \left|\frac{f''(\xi)}{2}(z - c)^2\right|,$$

from which we conclude that

$$f(z) \approx f(c) + f'(c)(z - c);$$

this is the linear approximation of $f$ at $z = c$. If $f(z) = \frac{1}{z}$, we know that $f'(z) = -\frac{1}{z^2}$. Set $z = \overline{x}$ and $c = \mu_X$.

Since $f$ is $C^2$ over any interval $[a, b]$ with $a > 0$, if $\overline{x}$ is sufficiently close to $\mu_X$, then the liner approximation becomes

$$\frac{1}{\overline{x}} \approx \frac{1}{\mu_X} - \frac{1}{\mu_X^2}(\overline{x} - \mu_X)$$

(the constant approximation would be $\frac{1}{\overline{x}} \approx \frac{1}{\mu_X}$).

But $\mathrm{E}(\overline{x}) = \mu_X$, $\mathrm{E}(\overline{y}) = \mu_Y$ (SRS), and $\mu_Y = R\mu_X$, so that

$$\begin{aligned}
\mathrm{E}[r - R] &= \mathrm{E}\Big[\frac{\overline{y} - R\overline{x}}{\overline{x}}\Big] \approx \mathrm{E}\Big[\Big(\frac{1}{\mu_X} - \frac{1}{\mu_X^2}(\overline{x} - \mu_X)\Big)(\overline{y} - R\overline{x})\Big] \\
&= \mathrm{E}\Big[\frac{1}{\mu_X}(\overline{y} - R\overline{x})\Big] - \mathrm{E}\Big[\frac{1}{\mu_X^2}(\overline{x} - \mu_X)(\overline{y} - R\overline{x})\Big] \\
&= \frac{1}{\mu_X}\Big(\mathrm{E}(\overline{y}) - R \cdot \mathrm{E}(\overline{x})\Big) - \frac{1}{\mu_X^2}\Big(\mathrm{E}\big[\overline{xy} - \mu_X\overline{y} - R\overline{x}^2 - R\mu_X\overline{x}\big]\Big) \\
&= \frac{1}{\mu_X}\underbrace{\Big(\mu_Y - R\mu_X\Big)}_{=0} - \frac{1}{\mu_X^2}\Big(\mathrm{E}(\overline{xy}) - \mu_X\mathrm{E}(\overline{y}) - R\big(\mathrm{E}(\overline{x}^2) - \mu_X\mathrm{E}(\overline{x})\big)\Big) \\
&= -\frac{1}{\mu_X^2}\Big(\mathrm{E}(\overline{xy}) - \mu_X\mu_Y - R\big(\mathrm{E}(\overline{x}^2) - \mu_X^2\big)\Big)
\end{aligned}$$

We further simplify the sampling bias $\mathrm{E}[r - R]$ with the help of $\mathrm{E}(\overline{xy}) = \mu_X\mu_Y + \mathrm{Cov}(\overline{x}, \overline{y})$, and $\mathrm{E}(\overline{x}^2) = \mu_X^2 + \mathrm{V}(\overline{x})$. Thus,

$$\mathrm{E}[r - R] \approx -\frac{1}{\mu_X^2}\big[\mathrm{Cov}(\overline{x}, \overline{y}) - R \cdot \mathrm{V}(\overline{x})\big].$$

In an SRS of size $n$, drawn from a finite population withsize $N$ and variance $sigma^2$, we have already seen that

$$V(\overline{x}) = \frac{\sigma_X^2}{n}\Big(\frac{N - n}{N - 1}\Big) \quad \text{and} \quad V(\overline{y}) = \frac{\sigma_Y^2}{n}\Big(\frac{N - n}{N - 1}\Big).$$

Consider the random variable $Z = X + Y$. The SRS estimator of

$$\mu_Z = \mu_X + \mu_Y$$

is

$$\overline{z} = \overline{x} + \overline{y};$$

its **sampling variance** is

$$\begin{aligned}
\mathrm{V}(\overline{z}) &= \frac{\sigma_Z^2}{n}\left(\frac{N-n}{N-1}\right), \quad \text{where} \\
\sigma_Z^2 &= \frac{1}{N}\sum_{j=1}^{N}(z_j - \mu_Z)^2 = \frac{1}{N}\sum_{j=1}^{N}\left\{(x_j + y_j) - (\mu_X + \mu_Y)\right\}^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu_X)^2 + \frac{2}{N}\sum_{j=1}^{N}(x_j - \mu_X)(y_j - \mu_Y) + \frac{1}{N}\sum_{j=1}^{N}(y_j - \mu_Y)^2 \\
&= \sigma_X^2 + 2\sigma_{XY} + \sigma_Y^2 = \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2,
\end{aligned}$$

where $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$ is the **Pearson correlation coefficient between** $X$ **and** $Y$.

On the one hand,

$$\mathrm{V}(\overline{z}) = \frac{\sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right);$$

on the other,

$$\begin{aligned}
\mathrm{V}(\overline{z}) &= \mathrm{V}(\overline{x} + \overline{y}) = \mathrm{V}(\overline{x}) + 2\mathrm{Cov}(\overline{x}, \overline{y}) + \mathrm{V}(\overline{y}) \\
&= \frac{\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right) + 2\mathrm{Cov}(\overline{x}, \overline{y}) + \frac{\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right);
\end{aligned}$$

we can thus conclude that

$$\mathrm{Cov}(\overline{x}, \overline{y}) = \frac{\rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right).$$

Consequently,

$$\begin{aligned}
\mathrm{E}[r - R] &\approx -\frac{1}{\mu_X^2}\left[\mathrm{Cov}(\overline{x}, \overline{y}) - R \cdot \mathrm{V}(\overline{x})\right] \\
&= -\frac{1}{\mu_X^2}\left[\frac{\rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right) - R\frac{\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right)\right] \\
&= \frac{1}{\mu_X^2} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right)
\end{aligned}$$

But the **systematic error** is not the only way to qualify the magnitude of the error made when using $r$ to estimate $R$: the **mean square error** (MSE) of $r$ is

$$\mathrm{MSE}(r) = \mathrm{E}\left((r - R)^2\right) = \mathrm{V}(r) + \left(\mathrm{E}(r) - R\right)^2.$$

**Ratio Estimator Variability**

We can obtain an approximation of $\mathrm{V}(r)$ using the constant Taylor approximation (of order 0):

$$\frac{1}{\overline{x}} \approx \frac{1}{\mu_X}.$$

Thus,

$$\mathrm{V}(r) = \mathrm{V}(r - R) = \mathrm{V}\Big[\frac{\overline{y}}{\overline{x}} - R\Big] = \mathrm{V}\Big[\frac{\overline{y} - R\overline{x}}{\overline{x}}\Big] \approx \mathrm{V}\Big[\frac{\overline{y} - R\overline{x}}{\mu_X}\Big].$$

Consider the random variable $W = Y - RX$. Since $\mu_Y = R\mu_X$,

$$\mu_W = \mu_Y - R\mu_X = 0.$$

The SRS sample mean of $W$ in $\mathcal{Y}$ is thus

$$\overline{w} = \overline{y} - R\overline{x} \implies \mathrm{V}(r) \approx \mathrm{V}\Big[\frac{\overline{w}}{\mu_X}\Big] = \frac{1}{\mu_X^2}\mathrm{V}(\overline{w}) = \frac{1}{\mu_X^2} \cdot \frac{\sigma_W^2}{n}\Big(\frac{N-n}{N-1}\Big),$$

where

$$\sigma_W^2 = \frac{1}{N}\sum_{j=1}^{N}(W_j - \mu_W)^2 = \frac{1}{N}\sum_{j=1}^{N} W_j^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - RX_j)^2.$$

We thus have

$$\mathrm{V}(r) \approx \frac{1}{\mu_X^2} \cdot \frac{1}{n} \cdot \frac{1}{N}\sum_{j=1}^{N}(Y_j - RX_j)^2\Big(\frac{N-n}{N-1}\Big).$$

The ratio between the systematic error $\mathrm{E}[r-R]$ and the standard deviation of $r$ is then

$$\frac{\mathrm{E}[r-R]}{\mathrm{SD}(r)} \approx \frac{1}{\sqrt{n}} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{\sigma_W}\sqrt{\frac{N-1}{N-n}};$$

when $n, N \to \infty$ (while $N \gg n$), we must have

$$\frac{\mathrm{E}[r-R]}{\mathrm{SD}(r)} \to 0.$$

In other words, although it is impossible to get rid of the bias, the estimation error

$$\mathrm{MSE}(r) = \mathrm{V}(r) + (\mathrm{E}(r) - R)^2$$

is dominated by the variance $\mathrm{V}(r)$ if the sample size $n$ is **sufficiently large**.

**Example** The list of countries for which both life expectancy and (logarithm of) gross domestic product per capita are available in 2011 contains $N = 168$ observations.

```
gapminder.RLD <- gapminder |>  filter(year==2011) |>
    select(life_expectancy,gdp,population)

# we keep only the observations that have both
gapminder.RLD <-  gapminder.RLD[complete.cases(gapminder.RLD),]
gapminder.RLD <- gapminder.RLD |> mutate(lgdppc=log(gdp/population))
(N=nrow(gapminder.RLD))
```

```
[1] 168
```

We draw $m = 500$ SRS samples of $n = 20$, and we compute the estimator $r$ of the ratio $R$ for each of these samples.

```
set.seed(12) # replicability
n=20
m=500

quotients <- c()
for(k in 1:m){
    samp <- gapminder.RLD[sample(1:N,n, replace=FALSE),c("life_expectancy","lgdppc")]
    quotients[k] <- mean(samp$life_expectancy/samp$lgdppc)
  }
```

The average of the 500 estimators is shown below:

```
quotients <- data.frame(quotients)
mean(quotients$quotients)
```

```
[1] 9.238648
```

We already know that $\mu_X = 7.84$. It would be reasonable to expect that $\mu_Y \approx \bar{r}\mu_X$:

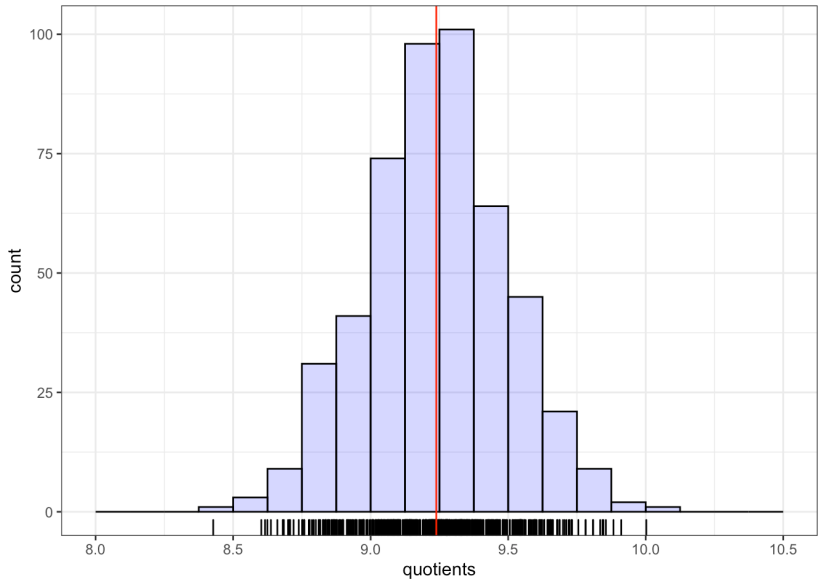
```
mean(gapminder.RLD$lgdppc)*mean(quotients$quotients)
```

```
[1] 72.45559
```

Is this a better approximation than the one we obtained at the beginning of the section: $\mu_Y \approx 68.00$? This question cannot be answered without knowing the **distribution of the estimator** $r$.[40]

40: Keep in mind that it is indeed a random variable since its value depends on the sample $\mathcal{Y}$ selected.

```
ggplot(quotients, aes(quotients)) +
    geom_rug(aes(quotients)) +
    geom_histogram(breaks=seq(8, 10.5, by = .125),
                   col="black", fill="blue", alpha=.2) +
    geom_vline(xintercept=mean(quotients$quotients),
               color="red")
```




```
summary(quotients)
```

```
Min.    1st Qu.  Median   Mean    3rd Qu.  Max.
8.428   9.073    9.246    9.239   9.401    10.002
```

**Ratio Estimator Confidence Intervals**

We can show that the estimator $r$ follows **approximately** a normal distribution $\mathcal{N}(\mathrm{E}(r), \mathrm{V}(r))$, from which we conclude that the **bound on the error of estimation is**

$$B_R \approx \hat{B}_R = 2\sqrt{\hat{\mathrm{V}}(r)} \approx 2\sqrt{\frac{1}{\mu_X^2} \cdot \frac{s_W^2}{n}\left(1 - \frac{n}{N}\right)} \approx 2\sqrt{\frac{1}{\bar{x}^2} \cdot \frac{s_W^2}{n}\left(1 - \frac{n}{N}\right)},$$

where

$$s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - r x_i)^2.$$

Thus

$$\text{C.I.}(R; 0.95): \quad r \pm \hat{B}_R$$

is an **approximate 95% C.I. for** $R$.

Write $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$. We notice that

$$\begin{aligned}
\sigma_W^2 &= \frac{1}{N}\sum_{j=1}^{N} W_j^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - RX_j)^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y + \mu_Y - RX_j) \\
&= \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y + R\mu_X - RX_j)^2 = \frac{1}{N}\sum_{j=1}^{N}[(Y_j - \mu_Y) - R(X_j - \mu_X)]^2 \\
&= \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y)^2 - 2R\frac{1}{N}\sum_{j=1}^{N}(X_j - \mu_X)(Y_j - \mu_Y) + R^2\frac{1}{N}\sum_{j=1}^{N}(X_j - \mu_X)^2 \\
&= \sigma_Y^2 - 2R\mathrm{Cov}(X,Y) + R^2\sigma_X^2 = \sigma_Y^2 - 2R\rho\sigma_X\sigma_Y + R^2\sigma_X^2,
\end{aligned}$$

By analogy, we then have $s_W^2 = s_Y^2 - 2r\hat{\rho}s_X s_Y + r^2 s_X^2$, where

$$s_X^2 = \frac{1}{n-1}\left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right), \quad s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^n y_i^2 - n\overline{y}^2\right),$$
$$s_{XY} = \frac{1}{n-1}\left(\sum_{i=1}^n x_i y_i - n\overline{xy}\right), \quad \text{and} \quad \hat{\rho} = \frac{s_{XY}}{s_X s_Y}.$$

In practice, we can also use the following formula:

$$s_W^2 = \frac{1}{n-1}\left(\sum_{i=1}^n y_i^2 - 2r\sum_{i=1}^n x_i y_i + r^2\sum_{i=1}^n x_i^2\right).$$

**Example** Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of size $n = 132$, drawn from a population of size $N = 37,444$. Find a 95% C.I. for $R$ if

$$\sum_{i=1}^n x_i = 9464.6, \quad \sum_{i=1}^n y_i = 14691.6,$$
$$\sum_{i=1}^n x_i^2 = 686773.2, \quad \sum_{i=1}^n x_i y_i = 1062186, \quad \sum_{i=1}^n y_i^2 = 1670194.$$

With this sample, we have $r = \frac{14691.6}{9464.6} \approx 1.55$, so that

$$s_W^2 = \frac{1670194 - 2(1.55)(1062186) + (1.55)^2(686773.2)}{132-1} \approx 209.2, \quad \text{and}$$
$$\hat{\mathrm{V}}(r) \approx \frac{132^2}{9464.6^2}\frac{209.2}{132}\left(1 - \frac{132}{37444}\right) = 0.0003 \implies \text{C.I.}(R; 0.95) \approx 1.552 \pm 0.035.$$

**Example** Find a 95% C.I. for the ratio of life expectancy by the logarithm of the GDO per capita in 2011 with the help of a SRS of size $n = 20$.

The true ratio is:

```
(R = mean(gapminder.RLD$life_expectancy)/mean(gapminder.RLD$lgdppc))
```

```
[1] 9.046742
```

We draw a sample of size $n = 20$, and we calculate the intermediate sums:

```
N=nrow(gapminder.RLD); n=20
set.seed(123456) # replicability
index = sample(1:N,n, replace=FALSE)
samp = gapminder.RLD[index,c("life_expectancy","lgdppc")]

(sum.xi = sum(samp$lgdppc))
(sum.yi = sum(samp$life_expectancy))
(sum.xi.2 = sum(samp$lgdppc^2))
(sum.yi.2 = sum(samp$life_expectancy^2))
(sum.xiyi = sum(samp$lgdppc*samp$life_expectancy))
```

```
[1] 167.2794
[1] 1450.82
[1] 1430.912
[1] 106117.4
[1] 12245.93
```

Finally, we compute the estimator $r$ and its variance, as well as the desired confidence interval.

```
r = sum.yi/sum.xi
s2.W = 1/(n-1)*(sum.yi.2-2*r*sum.xiyi+r^2*sum.xi.2)
V = n^2/sum.xi^2*(1/n)*s2.W*(1-n/N)
B = 2*sqrt(V)
c(r-B,r+B)
```

```
[1] 8.252515 9.093552
```

We would expect the quotient $R$ to be in the interval $(8.25, 9.09)$ with 95% probability;[41] since $R = 9.046742$, it is indeed the case.[42]

41: According to the frequentist interpretation of confidence intervals.

42: As we have noticed several times, the confidence interval can of course change depending on which sample is drawn from the population.

**Estimation of the Mean and the Total Using the Ratio Estimator**

In practice, we often know $\tau_X$ and/or $\mu_X$. It is possible to use the relation

$$\mu_Y = R\mu_X, \quad \text{where } R = \frac{\mu_Y}{\mu_X}$$

in order to approximate $\mu_Y$ (if $\mu_X$ is unknown, one uses $\mu_X \approx \overline{x}$).

Since $r = \overline{y}/\overline{x}$, the **ratio-based estimator for** $\hat{\mu}_{Y;R}$ is simply:

$$\hat{\mu}_{Y;R} = r \cdot \mu_X.$$

But we have already observed that $r$ is a biased estimator of $R$, so we expect $\hat{\mu}_{Y;R}$ to be a biased estimator of $\mu_Y$, with a normal distribution: $\hat{\mu}_{Y;R} \sim_{\text{approx}} \mathcal{N}(\mathrm{E}(\hat{\mu}_{Y;R}), \mathrm{V}(\hat{\mu}_{Y;R}))$.

It is easy to show

$$\mathrm{E}[\hat{\mu}_{Y;R} - \mu_Y] = \mu_X \mathrm{E}[r - R] \approx \frac{1}{\mu_X} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right)$$

$$\mathrm{V}(\hat{\mu}_{Y;R}) = \mathrm{V}(r \cdot \mu_X) = \mu_X^2 \mathrm{V}(r) \approx \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right).$$

The bound of error on the estimation of $\mu_{Y;R}$ is thus

$$B_{\mu_{Y;R}} \approx \hat{B}_{\mu_{Y;R}} = 2\sqrt{\mathrm{V}(\hat{\mu}_{Y;R})} \approx 2\sqrt{\frac{s_W^2}{n}\left(1 - \frac{n}{N}\right)}, \; s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2,$$

from which we see that $\mathrm{C.I.}_R(\mu_Y; 0.95) \equiv \hat{\mu}_{Y;R} \pm \hat{B}_{\mu_{Y;R}}$ is an **approximate 95% C.I. for** $\mu_Y$.

It is also possible to use the relationship

$$\tau_Y = R\tau_X, \quad \text{where} \quad R = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}$$

to approximate $\tau_Y$ (if $\tau_X$ is unknown, we use $\tau_X \approx N\overline{x}$).

Since $r = \overline{y}/\overline{x}$, the **ratio-based estimator for** $\hat{\tau}_{Y;R}$ is simply:

$$\hat{\tau}_{Y;R} = r \cdot \tau_X.$$

But we have already observed that $r$ is a biased estimator of $R$, so we expect $\hat{\tau}_{Y;R}$ to be a biased estimator of $\tau_Y$, which follows a normal distribution:

$$\hat{\tau}_{Y;R} \sim_{\text{approx}} \mathcal{N}\left(\mathrm{E}(\hat{\tau}_{Y;R}), \mathrm{V}(\hat{\tau}_{Y;R})\right).$$

It is easy to show

$$\mathrm{E}[\hat{\tau}_{Y;R} - \tau_Y] = \tau_X \mathrm{E}[r - R] \approx \frac{N}{\mu_X} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n}\left(\frac{N-n}{N-1}\right)$$

$$\mathrm{V}(\hat{\tau}_{Y;R}) = \mathrm{V}(r \cdot \tau_X) = \tau_X^2 \mathrm{V}(r) = N^2\mu_X^2 \mathrm{V}(r) \approx N^2 \cdot \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right).$$

The **bound of error on the estimation of** $\tau_{Y;R}$ is thus

$$B_{\tau_{Y;R}} \approx \hat{B}_{\tau_{Y;R}} = 2\sqrt{\mathrm{V}(\hat{\tau}_{Y;R})} \approx 2N\sqrt{\frac{s_W^2}{n}\left(1 - \frac{n}{N}\right)}, \; s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2,$$

from which we conclude that $\mathrm{C.I.}_R(\tau_Y; 0.95) \equiv \hat{\tau}_{Y;R} \pm \hat{B}_{\tau_{Y;R}}$ is an **approximate 95% C.I. for** $\tau_Y$.

**Example** Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ o size $n = 132$, drawn from a population of size $N = 37,444$. Find a 95% C.I. for $\mu_Y$ using ratio-based estimation, given that

$$\sum_{i=1}^{n} x_i = 9464.6, \quad \sum_{i=1}^{n} y_i = 14691.6,$$

$$\sum_{i=1}^{n} x_i^2 = 686773.2, \quad \sum_{i=1}^{n} x_i y_i = 1062186, \quad \sum_{i=1}^{n} y_i^2 = 1670194.$$

With this sample, we have $r \approx 1.55$, $s_W^2 \approx 209.2$, $\hat{\mathrm{V}}(r) \approx 0.00031$, and $\mathrm{C.I.}(R; 0.95) \approx 1.552 \pm 0.035$. Moreover, $\overline{x} = 9464.6/132 = 71.70$. Thus

$$\mathrm{C.I.}_R(\mu_Y; 0.95) = \mu_X \cdot \mathrm{C.I.}(R; 0.95) \approx \overline{x} \cdot \mathrm{C.I.}(R; 0.95) \equiv 111.29 \pm 2.51. \quad \blacksquare$$

**Example** Find a 95% C.I. for the average life expectancy by country $\mu_Y$, in 2011, using ratio estimation and the logarithm of the gross domestic product per capita in 2011 ($X$), with a SRS sample of size $n = 20$.

We use the same sample as in the preceding example on the topic. We have already obtained a confidence interval for the ratio:

$$\mathrm{C.I.}(R; 0.95) = (8.25, 9.09).$$

The sample mean of $X$ was $\overline{x} = \frac{167.2794}{20} = 8.364$. The 95% confidence interval for the average life expectancy using ratio estimation is thus

$$\text{C.I.}_R(\mu_Y; 0.95) = \mu_X \cdot \text{C.I.}(R; 0.95) \approx \overline{x} \cdot (8.25, 9.09) = (69.00, 76.03).$$

Recall that the true value is $\mu_Y = 70.95$.

### Sample Size

Just as was the case with SRS and STS, we can determine the required sample size assuming that we have some information about the population distribution.

To give an estimate for $R$, use:

$$\begin{aligned} B_R \approx 2\sqrt{\frac{1}{\mu_X^2} \cdot \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right)} &\iff \underbrace{\frac{B_R^2 \mu_X^2}{4}}_{=D_R} = \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right) \\ &\iff \frac{(N-1)D_R}{\sigma_W^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &\iff \frac{(N-1)D_R + \sigma_W^2}{\sigma_W^2} = \frac{N}{n} \\ &\iff n_R = \frac{N\sigma_W^2}{(N-1)D_R + \sigma_W^2}. \end{aligned}$$

To give an estimate for $\mu_Y$ with ratio estimation, use:

$$B_{\mu_Y;R} \approx 2\sqrt{\frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\mu_Y} = \frac{N\sigma_W^2}{(N-1)D_{\mu_Y} + \sigma_W^2}, \quad D_{\mu_Y} = \frac{B_{\mu_Y;R}^2}{4};$$

for $\tau_Y$, use:

$$B_{\tau_Y;R} \approx 2\sqrt{N^2 \cdot \frac{\sigma_W^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\tau_Y} = \frac{N\sigma_W^2}{(N-1)D_{\tau_Y} + \sigma_W^2}, \quad D_{\tau_Y} = \frac{B_{\tau_Y;R}^2}{4N^2}.$$

Since we do not typically know $\sigma_W^2$, we often use a small preliminary sample and use the empirical variance $s_W^2$ as an estimator of $\sigma_W^2$.

**Example** Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of size $n$, drawn from a population of size $N = 37,444$. Assume that we have $\sigma_W^2 \approx 209.2$ and $\mu_X \approx 71.7$, perhaps from a previous study.

Determine the minimum sample size required to ensure that the bound on the error of estimation of the:

1. ratio $R$ using $r$ is at most 0.025;
2. mean $\mu_Y$ using $\hat{\mu}_{Y;R}$ is at most 5, and
3. total $\tau_Y$ using $\hat{\tau}_{Y;R}$ is at most 25.

We simply use the formulas.

1. since $D_R = \frac{B_R^2 \mu_X^2}{4} = \frac{0.025^2(71.7)^2}{4} \approx 0.8033$, we have

$$n_R = \frac{37444(209.2)}{(37444-1)(0.8033)+209.2} = 258.6453 \implies n_R \geq 259;$$

2. since $D_{\mu_Y} = \frac{B_{\mu_{Y;R}}^2}{4} = \frac{5^2}{4} \approx 6.25$, we have

$$n_{\mu_Y} = \frac{37444(209.2)}{(37444-1)(6.25)+209.2} = 33.443 \implies n_{\mu_Y} \geq 34;$$

3. since $D_{\tau_Y} = \frac{B_{\tau_{Y;R}}^2}{4N^2} = \frac{25^2}{4(37444)} \approx 0.001502243$, we have

$$n_{\tau_Y} = \frac{37444(209.2)}{(37444-1)(0.001502243)+209.2} = 29509.62 \implies n_{\tau_Y} \geq 29510.$$

In this last case, the desired bound $B_{\tau_{Y;R}}$ is probably too tight (the resulting sample size is way too large). ■

### 10.5.2 Regression Estimation

Ratio estimation is a special case of a more general method, **regression estimation**. In the `gapminder.csv` dataset for 2011, we recognize that there is a more or less linear relationship between the **life expectancy** $Y$ and the **logarithm of the GDP per capita** $X$ for $N = 168$ countries.

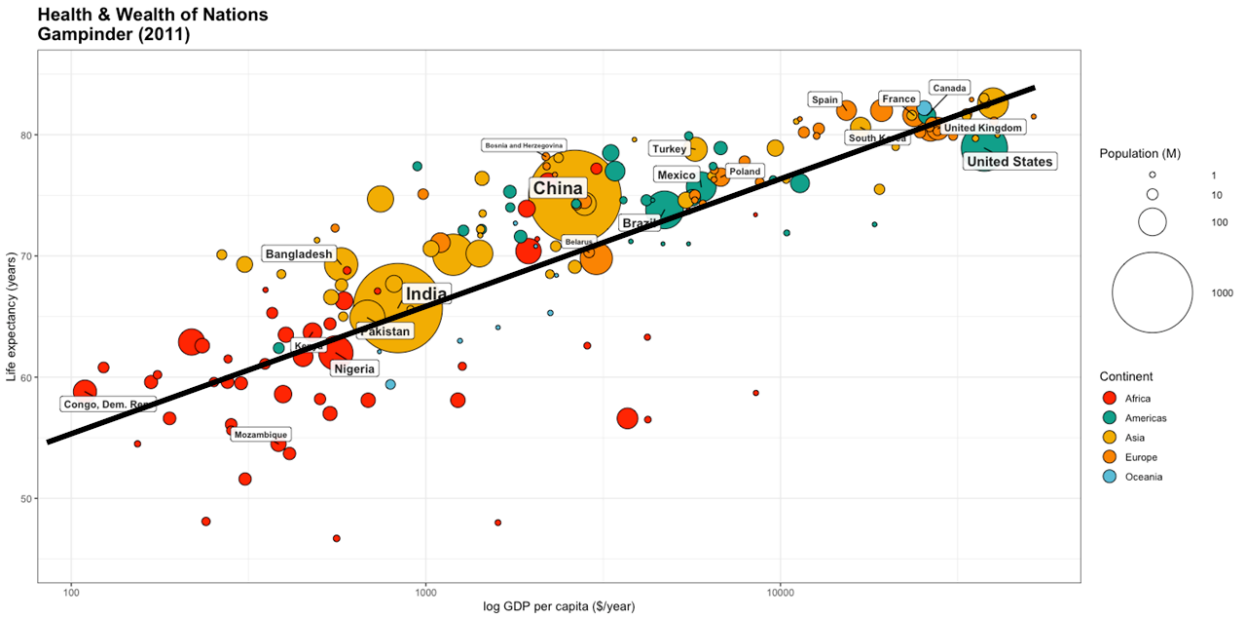



**Figure 10.8:** Health and wealth of nations for the 2011 Gapminder data, with superimposed line of best fit.

When we compute

$$r = \overline{y}/\overline{x}$$

using a SRS of size $n$, we are really assuming that the true relationship between $Y$ and $X$ takes the form $Y = RX \approx rX$, i.e., that it is a straight line of slope $r$ **passing through the origin**. But this last condition does not seem to be met. What to do in this case?

**Regression Estimator**

As above, let $\mathcal{U}$ be a finite bivariate population of size $N$, and $\mathcal{Y} \subseteq \mathcal{U}$ be a finite bivariate random sample of size $n$. We assume that the relationship between $Y$ and $X$ takes the form

$$Y - \mu_Y = \beta(X - \mu_X).$$

If $\mu_X$ is known (as we had assumed was the case for ratio estimation), the **regression estimator** $\hat{\mu}_{Y;L}$ **of** $\mu_Y$ obtained with the SRS $\mathcal{Y}$ is

$$\hat{\mu}_{Y;L} = \overline{y} + \beta(\mu_X - \overline{x}).$$

For now, we treat $\beta$ as an **unknown** constant (since $\mu_Y$ is also unknown). Since $\mathcal{Y}$ is drawn in a SRS context, $\mathrm{E}(\overline{x}) = \mu_X$ and $\mathrm{E}(\overline{y}) = \mu_Y$, so that

$$\mathrm{E}(\hat{\mu}_{Y;L}) = \mathrm{E}(\overline{y}) + \beta(\mu_X - \mathrm{E}(\overline{x})) = \mu_Y + \beta(\mu_X - \mu_X) = \mu_Y.$$

Consider the random variable $W = Y + \beta(\mu_X - X)$. As $\beta$ is constant, we have

$$\mu_W = \mu_Y + \beta(\mu_X - \mu_X) = \mu_Y.$$

The sample mean of $W$ is thus

$$\overline{w} = \overline{y} + \beta(\mu_X - \overline{x}) = \hat{\mu}_{Y;L} \implies \mathrm{V}(\hat{\mu}_{Y;L}) = \mathrm{V}(\overline{w}) = \frac{\sigma^2_{W;L}}{n}\Big(\frac{N-n}{N-1}\Big).$$

But

$$\begin{aligned}\sigma^2_{W;L} &= \frac{1}{N}\sum_{j=1}^{N}(W_j - \mu_W)^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j + \beta(\mu_X - X_j) - \mu_Y)^2 \\ &= \frac{1}{N}\sum_{j=1}^{N}\big\{(Y_j - \mu_Y) - \beta(X_j - \mu_X)\big\}^2 = \sigma^2_Y - 2\beta\rho\sigma_X\sigma_Y + \beta^2\sigma^2_X,\end{aligned}$$

where $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$. Consequently,

$$\mathrm{V}(\hat{\mu}_{Y;L}) = \frac{\sigma^2_Y - 2\beta\rho\sigma_X\sigma_Y + \beta^2\sigma^2_X}{n}\Big(\frac{N-n}{N-1}\Big).$$

In general, for a given systematic error (bias), preference is given to the estimator **with the lowest variance**. The value of $\beta$ which minimizes $\mathrm{V}(\hat{\mu}_{Y;L})$ would then satisfy

$$\frac{\partial \mathrm{V}(\hat{\mu}_{Y;L})}{\partial\beta}(\beta^*) = \frac{1}{n}\Big(\frac{N-n}{N-1}\Big)(-2\rho\sigma_X\sigma_Y + 2\beta^*\sigma^2_X) = 0,$$

which is to say that

$$\beta^* = \rho\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}\cdot\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma^2_X},$$

from which we conclude that

$$\begin{aligned}
\mathrm{V}(\hat{\mu}_{Y;L}) &= \frac{\sigma_Y^2 - 2\beta^*\rho\sigma_X\sigma_Y + (\beta^*)^2\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right) \\
&= \frac{\sigma_Y^2 - 2\rho\frac{\sigma_Y}{\sigma_X}\rho\sigma_X\sigma_Y + (\rho\frac{\sigma_Y}{\sigma_X})^2\sigma_X^2}{n}\left(\frac{N-n}{N-1}\right) \\
&= \frac{\sigma_Y^2 - 2\rho^2\sigma_Y^2 + \rho^2\sigma_Y^2}{n}\left(\frac{N-n}{N-1}\right) \\
&= \frac{\sigma_Y^2(1-\rho^2)}{n}\left(\frac{N-n}{n-1}\right).
\end{aligned}$$

**Regression Estimator Bias**

The task is to determine the coefficients $\alpha, \beta$ that "best describe" the linear relationship between $X$ and $Y$,

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where we assume that $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \sim_{\text{approx.}} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

There are several ways to interpret the phrase "best describe" – the **least squares estimators** $\hat{\alpha}$ and $\hat{\beta}$ are those that minimize the residual sum of squares

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

We solve the system of equations

$$\frac{\partial Q}{\partial \alpha}(a, b) = \sum_{i=1}^{n} -2(y_i - a - bx_i) = 0, \quad \frac{\partial Q}{\partial \beta}(a, b) = \sum_{i=1}^{n} -2x_i(y_i - a - bx_i) = 0,$$

which yields

$$\hat{\alpha} = a = \overline{y} - b\overline{x} \quad \text{and} \quad \hat{\beta} = b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

In practice, it is this $b = \hat{\rho}\frac{s_Y}{s_X}$ that plays the role of the estimator $\beta^*$; note that it varies from one SRS to another. Since $b$ is a random variable,[43] **we cannot conclude that** $\mathrm{E}(b\overline{x}) = \mathrm{E}(b)\mathrm{E}(\overline{x})$, so that

$$\mathrm{E}(\hat{\mu}_{Y;L}) = \mathrm{E}(\overline{y}) + \mu_X\mathrm{E}(b) - \mathrm{E}(b\overline{x}) \neq \mu_Y,$$

in general.

However, if the sample size $n$ is large, it is possible to show that

$$\mathrm{E}[\hat{\mu}_{Y;L} - \mu_Y]$$

is of order $\frac{1}{n}$ (as was the case for the systematic error in ratio estimation); $\hat{\mu}_{Y;L}$ is therefore a **biased estimator** of $\mu_Y$.

43: in the sense that we obtain (potentially) a different slope with every SRS $\mathcal{Y}$.

**Regression Estimator Variability**

The sampling variance of $\hat{\mu}_{Y;L}$ is also of order $\frac{1}{n}$, and so the quotient of the bias $\mathrm{E}[\hat{\mu}_{Y;L} - \mu_Y]$ by the standard deviation of $\hat{\mu}_{Y;L}$ is of order $\frac{1}{\sqrt{n}}$.

Thus, when $n, N \longrightarrow \infty$ (assuming that $N \gg n$), we have

$$\frac{\mathrm{E}[\hat{\mu}_{Y;L} - \mu_Y]}{\mathrm{SD}(\hat{\mu}_{Y;l})} \longrightarrow 0.$$

Although it is impossible to get rid of the bias, the estimation error

$$\mathrm{MSE}(\hat{\mu}_{Y;L}) = \mathrm{V}(\hat{\mu}_{Y;L}) + (\mathrm{E}(\hat{\mu}_{Y;L}) - \mu_Y)^2$$

is dominated byt the variance $\mathrm{V}(\hat{\mu}_{Y;L})$ when $n$ is **sufficiently large**.

**Regression Estimator Confidence Intervals**

The regression estimator $\hat{\mu}_{Y;L}$ follows **approximately** a normal distribution $\mathcal{N}\,(\mathrm{E}(\hat{\mu}_{Y;L}), \mathrm{V}(\hat{\mu}_{Y;L}))$, from which we conclude that the **bound on the error of estimation** is

$$B_L \approx \hat{B}_L = 2\sqrt{\hat{\mathrm{V}}(\hat{\mu}_{Y;L})} \approx 2\sqrt{\frac{s^2_{W;L}}{n}\left(1 - \frac{n}{N}\right)},$$

where $s^2_{W;L}$ is the **regression mean square error**,

$$s^2_{W;L} = \frac{n-1}{n-2}(s^2_Y - b^2 s^2_X) = \frac{n-1}{n-2} \cdot s^2_Y(1 - \hat{\rho}^2).$$

Consequently $\mathrm{C.I.}_L(\mu_Y; 0.95) : \hat{\mu}_{Y;L} \pm \hat{B}_L$ is an **approximate 95% C.I. for** $\mu_Y$.[44]

44: We tackle $\tau_Y$ and $p_Y$ in the usual manner.

**Example** Consider a SRS $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $n = 132$, drawn from population of size $N = 37,444$. In a preceding study, we have shown that $\mu_X \approx 70.3$. Find a 95% C.I. for $\mu_Y$ using regression estimation if

$$\sum_{i=1}^{n} x_i = 9464.6, \quad \sum_{i=1}^{n} y_i = 14691.6,$$
$$\sum_{i=1}^{n} x_i^2 = 686773.2, \quad \sum_{i=1}^{n} x_i y_i = 1062186, \quad \sum_{i=1}^{n} y_i^2 = 1670194.$$

We must evaluate $\overline{x}$, $\overline{y}$, $s^2_X$, $s_{XY}$, $s^2_Y$, and $\hat{\rho}$. But

$$\overline{x} = \frac{9464.6}{132} \approx 71.7, \quad \overline{y} = \frac{14691.6}{132} \approx 111.3,$$
$$s^2_X = \frac{686773.2 - 132(71.7)^2}{132 - 1} \approx 62.2, \quad s^2_Y = \frac{1670194 - 132(111.3)^2}{132 - 1} \approx 267.3$$
$$s_{XY} = \frac{1062186 - 132(71.7)(111.3)}{132 - 1} \approx 67.2, \quad \hat{\rho} = \frac{67.2}{\sqrt{(62.2)(267.3)}} \approx 0.521.$$

The estimator for the regression slope is therefore $b = \hat{\rho}\frac{s_Y}{s_X} = 1.08$. Moreover,

$$s^2_{W;L} = \frac{131}{130} \cdot 267.3 \cdot (1 - 0.521^2) \approx 196.77.$$

Consequently,

$$\hat{\mu}_{Y;L} = 111.3 + 1.08(\underbrace{70.3}_{\mu_X} - 71.7) = 109.8, \quad \text{and}$$

$$\hat{B}_L \approx 2\sqrt{\frac{196.77}{132}\left(1 - \frac{132}{37444}\right)} = 2.43,$$

from which we conclude that

$$\text{C.I.}_L(\mu_Y; 0.95) \equiv 109.8 \pm 2.43.$$

Of course, if the linearity assumption is not valid, we should not expect the bound on the error of estimation using regression estimation to be substantially tighter than the one obtained in a SRS, say.

**Example** Find a 95% C.I. for the average life expectancy by country in 2011 using regression estimation against the logarithm of the GDP per capita, with $n = 20$, assuming that it is known that $\mu_X = 7.84$.

We draw a sample of size $n = 20$ and calculate the required quantities:

```
set.seed(123456) # replicability
N=nrow(gapminder.RLD); n=20
index = sample(1:N,n, replace=FALSE)
samp = gapminder.RLD[index,c("life_expectancy","lgdppc")]
mu.X = mean(gapminder.RLD$lgdppc)
```

The sample means are:

```
(y.bar = mean(samp$life_expectancy))
(x.bar = mean(samp$lgdppc))
```

```
[1] 72.541
[1] 8.363971
```

The intermediate sums and the correlation coefficient are:

```
sum.xi = sum(samp$lgdppc)
sum.yi = sum(samp$life_expectancy)
sum.xi.2 = sum(samp$lgdppc^2)
sum.yi.2 = sum(samp$life_expectancy^2)
sum.xiyi = sum(samp$lgdppc*samp$life_expectancy)

s2.X = (sum.xi.2-n*x.bar^2)/(n-1)
s2.Y = (sum.yi.2-n*y.bar^2)/(n-1)
s.XY = (sum.xiyi-n*x.bar*y.bar)/(n-1)
```

```
(rho = s.XY/sqrt(s2.X*s2.Y))
```

```
[1] 0.667983
```

Next, we evaluate the MSE:

```
(s2.W.L = (n-1)/(n-2)*s2.Y*(1-rho^2))
```

```
[1] 26.8736
```

The bound on the error of estimation is thus:

```
(B = 2*sqrt(s2.W.L/n*(1-n/N)))
```

```
[1] 2.175976
```

and the corresponding 95% C.I. for the mean life expectancy by country is:

```
(hat.mu.Y.L = y.bar + rho*sqrt(s2.Y/s2.X)*(mu.X-x.bar))
c(hat.mu.Y.L-B,hat.mu.Y.L+B)
```

```
[1] 70.71572
[1] 68.53974 72.89170
```

For comparison's sake, the true mean is $\mu_Y = 70.95$.

We can also compute the estimate and the confidence interval directly, with the base `lm()` function.

```
reg.lin = lm(life_expectancy~lgdppc, data=samp)
summary(reg.lin)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-16.2467   0.1592   1.6513   2.6614   5.8812

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.2559     7.7768   5.562  2.8e-05 ***
lgdppc        3.5013     0.9194   3.808  0.00129 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.184 on 18 degrees of freedom
Multiple R-squared:  0.4462,    Adjusted R-squared:  0.4154
F-statistic:  14.5 on 1 and 18 DF,  p-value: 0.001287
```

The required quantities can be extracted as follows:

```
(b = as.numeric(reg.lin$coefficients[2]))
```

```
[1] 3.501336
```

```
(s2.W.L = summary(reg.lin)$sigma^2)
```

```
[1] 26.8736
```

**Sample Size**

If we seek an regression estimate of $\mu_Y$, we use:

$$B_L \approx 2\sqrt{\frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right)} \iff \underbrace{\frac{B_L^2}{4}}_{=D_L} = \frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right) \iff$$

$$\frac{(N-1)D_L}{\sigma^2_{W;L}} = \frac{N-n}{n} = \frac{N}{n} - 1 \iff \frac{(N-1)D_L + \sigma^2_{W;L}}{\sigma^2_{W;L}} = \frac{N}{n},$$

$$\iff n_L = \frac{N\sigma^2_{W;L}}{(N-1)D_L + \sigma^2_{W;L}}.$$

For $\tau_Y$, we use:

$$B_{\tau;L} \approx 2N\sqrt{\frac{\sigma^2_{W;L}}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\tau;L} = \frac{N\sigma^2_{W;L}}{(N-1)D_{\tau;L} + \sigma^2_{W;L}},$$

$$\text{where } D_{\tau;L} = \frac{B^2_{\tau;L}}{4N^2}.$$

Since we do not know usually know $\sigma^2_{W;L}$, we often draw a small preliminary sample on which we compute the sample $s^2_{W;L}$, which is used as an estimator of $\sigma^2_{W;L}$.[45]

45: **Warning:** Even if formal manipulations can still be performed, the estimate may not be valid **if the relationship between the variables $X$ and $Y$ is not linear**.

**Example** Determine the sample size $n$ required to estimate the average life expectancy $\mu_Y$ using regression estimation against the logarithm of GDP per capita in 2011, with a bound of error on the estimation of $B_L = 1$, if $\sigma_{W;L} \approx 5.194$ and $N = 168$.

Using the formula, we have:

$$n_L = \frac{168(5.194)^2}{167(1^2/4) + (5.194)^2} = 65.94498 \implies n_L \geq 66.$$

Since there are good reasons to trust that the relationship between life expectancy and log GNP per capita in 2011 is approximately linear (see Figure 10.8), the regression approach is a strong one.[46] How does it compare with the example that uses ratio estimation?

46: Assuming, of course, that $\mu_X$ is known; otherwise, it is pretty much useless.

### 10.5.3 Difference Estimation

**Difference estimation** is another special case of regression estimation, where the slope $\beta$ is now assumed to be 1.

If $\mu_X$ is known, the **difference estimator** $\hat{\mu}_{Y;D}$ **of** $\mu_Y$ computed from a SRS $\mathcal{Y}$ is

$$\hat{\mu}_{Y;D} = \overline{y} + (\mu_X - \overline{x}).$$

Difference estimation is a good strategy when the relationship between $X$ and $Y$ is approximately **linear** and of **slope** 1,[47] as long as the variance of $Y$ along this line is **constant for all** $X$. Since $\mathcal{Y}$ is drawn according to a SRS, $\mathrm{E}(\overline{x}) = \mu_X$ and $\mathrm{E}(\overline{y}) = \mu_Y$, from which we conclude that

47: Passing or not **through the origin**.

$$\mathrm{E}(\hat{\mu}_{Y;D}) = \mathrm{E}(\overline{y}) + (\mu_X - \mathrm{E}(\overline{x})) = \mu_Y + (\mu_X - \mu_X) = \mu_Y.$$

Consider the random variable $D = Y - X$, whose expectation is

$$\mu_D = \mu_Y - \mu_X.$$

The sample mean of $D$ is thus

$$\overline{d} = \overline{y} - \overline{x} \implies \hat{\mu}_{Y;D} = \mu_X + (\overline{y} - \overline{x}) = \mu_X + \overline{d}.$$

Consequently,

$$\mathrm{V}(\hat{\mu}_{Y;D}) = \mathrm{V}(\mu_X + \overline{d}) = \mathrm{V}(\overline{d}) = \frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right).$$

But

$$\begin{aligned}\sigma_D^2 &= \frac{1}{N}\sum_{j=1}^{N}(D_j - \mu_D)^2 = \frac{1}{N}\sum_{j=1}^{N}\left\{(Y_j - X_j) - (\mu_Y - \mu_X)\right\}^2 \\ &= \frac{1}{N}\sum_{j=1}^{N}\left\{(Y_j - \mu_Y) - (X_j - \mu_X)\right\}^2 = \sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2,\end{aligned}$$

where $\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$. As such,

$$\mathrm{V}(\hat{\mu}_{Y;D}) = \frac{\sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2}{n}\left(\frac{N-n}{N-1}\right).$$

The difference estimator $\hat{\mu}_{Y;D}$ follows **approximately** a normal distribution $\mathcal{N}\left(\mathrm{E}(\hat{\mu}_{Y;D}), \mathrm{V}(\hat{\mu}_{Y;D})\right)$, from which we obtain the **bound on the error of estimation**

$$B_D \approx \hat{B}_D = 2\sqrt{\hat{\mathrm{V}}(\hat{\mu}_{Y;D})} \approx 2\sqrt{\frac{s_D^2}{n}\left(1 - \frac{n}{N}\right)},$$

where

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2 = s_Y^2 - 2\hat{\rho}s_X s_Y + s_X^2,$$

so that C.I.$_D(\mu_Y; 0.95) : \hat{\mu}_{Y;D} \pm \hat{B}_D$ is an **approximate 95% C.I. for** $\mu_Y$.[48]

48: We tackle $\tau_Y$ and $p_Y$ in the usual manner.

**Example** Auditors are often interested in comparing the audited value $Y$ of items with their book value $X$. Suppose that $N = 180$ items in inventory have a book value of $\tau_X = 13,320$. A SRS of $n = 10$ items yields the following data:

| **item** $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Audit** $y_i$ | 9 | 14 | 7 | 29 | 45 | 109 | 40 | 238 | 60 | 170 |
| **Book** $x_i$ | 10 | 12 | 8 | 26 | 47 | 112 | 36 | 240 | 59 | 167 |
| $d_i = y_i - x_i$ | −1 | 2 | −1 | 3 | −2 | −3 | 4 | −2 | 1 | 3 |

Find a 95% C.I. for the mean audit value $\mu_Y$ using difference estimation.

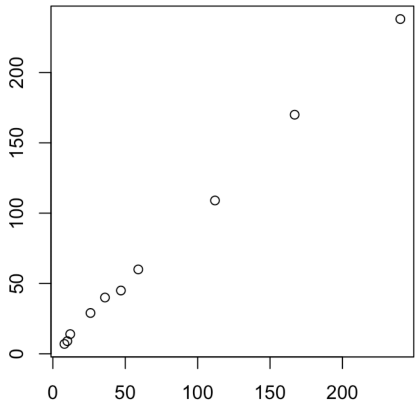



**Figure 10.9:** Scatterplot of $X$ and $Y$.

From the scatterplot, we surmise that the slope of the linear fit of $Y$ against $X$ is approximately 1. We must compute $\overline{d}$ and $s_D^2$:

$$\sum_{i=1}^{10} d_i = 4, \quad \sum_{i=1}^{10} d_i^2 = 58, \implies \overline{d} = \frac{4}{10} \quad \text{and} \quad s_D^2 = \frac{58 - 10(0.4)^2}{10 - 1} = 6.27.$$

Since $\mu_X = \frac{\tau_X}{N} = \frac{13320}{180} = 74$, the difference estimator is

$$\hat{\mu}_{Y;D} = \mu_X + \overline{d} = 74 + 0.4 = 74.4$$

and the bound is

$$\hat{B}_D \approx 2\sqrt{\hat{\mathbb{V}}(\hat{\mu}_D)} = 2\sqrt{\frac{6.27}{10}\left(1 - \frac{10}{180}\right)} = 1.54,$$

from which

$$\text{C.I.}_D(\mu_Y; 0.95) : \quad 74.4 \pm 1.54 \equiv (72.86, 75.94).$$

**Example** Consider a bivariate SRS sample $\mathcal{Y} = \{(x_i, y_i)\}$ of size $n = 132$, drawn from a population of size $N = 37,444$. In a preceding study, we

found that $\mu_X \approx 70.3$. Find a 95% C.I. for $\mu_Y$ using difference estimation, assuming that

$$\sum_{i=1}^{n} x_i = 9464.6, \quad \sum_{i=1}^{n} y_i = 14691.6,$$
$$\sum_{i=1}^{n} x_i^2 = 686773.2, \quad \sum_{i=1}^{n} x_i y_i = 1062186, \quad \sum_{i=1}^{n} y_i^2 = 1670194.$$

In a previous example, we have already computed

$$\overline{x} = 71.7,\ \overline{y} \approx 111.3,\ s_X^2 \approx 62.2,\ s_Y^2 \approx 267.3,\ s_{XY} \approx 67.2.$$

The difference estimator is thus

$$\hat{\mu}_{Y;D} = \overline{y} + (\mu_x - \overline{x}) = 111.3 + (70.3 - 71.7) = 109.9,$$

so that

$$\hat{B}_D \approx 2\sqrt{\frac{267.3 - 2(67.2) + 62.2}{132}\left(1 - \frac{132}{37444}\right)} = 2.427,$$

and

$$\text{C.I.}_D(\mu_Y; 0.95) \equiv 109.9 \pm 2.427.$$

**Example** Find a 95% C.I. for the average life expectancy by country in 2011 $\mu_Y$ using the difference method with the logarithm of GDP per capita per country ($X$), using a sample of size $n = 20$. Assume that $\mu_X = 7.84$ is known.

We draw a sample of size $n = 20$ and compute the various required quantities.

```
set.seed(1234567) # for replicability
N=nrow(gapminder.RLD); n=20
index = sample(1:N,n, replace=FALSE)
samp = gapminder.RLD[index,c("life_expectancy","lgdppc")]
d = samp$life_expectancy - samp$lgdppc
```

```
(mu.X = mean(gapminder.RLD[,"lgdppc"]))
(y.bar = mean(samp$life_expectancy))
(x.bar = mean(samp$lgdppc))
(d.bar = mean(d))
(s2.d = var(d))
```

```
[1] 7.842661
[1] 70.105
[1] 7.577646
[1] 62.52735
[1] 47.69057
```

Note that the regression slope does not seem to be 1 (if that was the case, we would expect $\overline{y}/\overline{x} \approx 1$). Difference estimation is not recommended in

this case, but we will continue the example nonetheless.

The bound on the error of estimation and the difference estimate are computed below, and the confidence interval is:

```
B = 2*sqrt(s2.d/n*(1-n/N))
hat.mu.Y.D = y.bar + (mu.X-x.bar)
c(hat.mu.Y.D-B,hat.mu.Y.D+B)
```

```
[1] 67.47129 73.26874
```

In spite of the difference estimation assumptions not being met, the 95% C.I. for $Y$ does contain the true value, $\mu_Y = 70.95$! A happy coincidence, no more.

#### Sample Size

As with the other methods, we can determine the sample size required to achieve a certain bound on the error of estimation.

In order to estimate $\mu_Y$ and $\tau_Y$ *via* difference estimation, use:

$$B_{\mu;D} \approx 2\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\mu;D} = \frac{N\sigma_D^2}{(N-1)D_{\mu;D} + \sigma_D^2};$$

$$B_{\tau;D} \approx 2N\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\tau;D} = \frac{N\sigma_D^2}{(N-1)D_{\tau;D} + \sigma_D^2},$$

where

$$D_{\mu;D} = \frac{B_{\mu;D}^2}{4} \quad \text{and} \quad D_{\tau;D} = \frac{B_{\tau;D}^2}{4N^2}.$$

As we do not know usually know $\sigma_D^2$, we often draw a small preliminary sample and use the **empirical variance** $s_D^2$ as an estimator of $\sigma_D^2$.

**Warning!** Even if formal manipulations can still be performed, **the estimate may not be valid if the relationship between the variables $X$ and $Y$ is not linear with slope $\approx 1$.**

### 10.5.4 Comparisons

We have already compared the bounds on the error of estimation for SRS, STS (Prop), and STS (Neyman), and discussed contexts in which one might expect a STS to be preferable to an SRS, or a STS (Neyman) preferable to a STS (Prop).

What can be said about ratio, regression, and difference estimation, both compared to SRS and to each other?

**Comparaison Between SRS and the Ratio Method**

In what context can we expect ratio estimation to perform "well"? Obviously, the relationship between $Y$ and $X$ must at least be **linear** and **pass through the origin**, i.e.,

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

It is generally assumed that the observations $\{x_i > 0\}$ are fixed, and that the error terms $\{\varepsilon_i\}$ are independent of each other, with

$$\mathrm{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathrm{V}(\varepsilon_i) = f(x_i)\sigma^2 > 0.$$

The question becomes: what form must $f(x_i)$ take so that the least squares solution $\hat{\beta}$ is **exactly** the estimator $r$ of the ratio $R$?

If we set

$$\underbrace{\frac{y_i}{\sqrt{f(x_i)}}}_{y_i'} = \beta \underbrace{\frac{x_i}{\sqrt{f(x_i)}}}_{x_i'} + \underbrace{\frac{\varepsilon_i}{\sqrt{f(x_i)}}}_{\varepsilon_i'}, \quad i = 1, \ldots, n,$$

we get

$$\mathrm{E}(\varepsilon_i') = \frac{1}{\sqrt{f(x_i)}}\mathrm{E}(\varepsilon) = 0 \quad \text{and} \quad \mathrm{V}(\varepsilon_i') = \frac{1}{f(x_i)}\mathrm{V}(\varepsilon_i') = \frac{f(x_i)\sigma^2}{f(x_i)} = \sigma^2,$$

and the assumptions of the least squares problem are satisfied. The estimator $\beta$ is obtained by minimizing

$$Q(\beta) = \sum_{i=1}^{n}(\varepsilon_i')^2 = \sum_{i=1}^{n}(y_i' - \beta x_i')^2 = \sum_{i=1}^{n}\frac{1}{f(x_i)}(y_i - \beta x_i)^2;$$

since

$$Q'(\beta) = -2\sum_{i=1}^{n}\frac{x_i}{f(x_i)}(y_i - \beta x_i),$$

this is equivalent to solving

$$0 = \sum_{i=1}^{n}\frac{x_i}{f(x_i)}(y_i - \hat{\beta}x_i) \Longleftrightarrow 0 = \sum_{i=1}^{n}\left(\frac{x_i y_i}{f(x_i)} - \hat{\beta}\frac{x_i^2}{f(x_i)}\right) \Longleftrightarrow \hat{\beta} = \frac{\displaystyle\sum_{i=1}^{n}\frac{x_i y_i}{f(x_i)}}{\displaystyle\sum_{i=1}^{n}\frac{x_i^2}{f(x_i)}}.$$

If $\frac{x_i}{f(x_i)} = k > 0$ for all $i = 1, \ldots, n$, the estimator $\hat{\beta}$ becomes

$$\hat{\beta} = \frac{k\displaystyle\sum_{i=1}^{n} y_i}{k\displaystyle\sum_{i=1}^{n} x_i} = \frac{\displaystyle\sum_{i=1}^{n} y_i}{\displaystyle\sum_{i=1}^{n} x_i} = r.$$

Thus, when the variance of $Y$ along the line $Y = \beta X$ is

$$\mathrm{V}(y_i) = \mathrm{V}(\beta x_i + \varepsilon_i) = \mathrm{V}(\varepsilon_i) = x_i\sigma^2$$

(i.e., the variance of $Y$ is **proportional to** $X$), the estimator $r$ of the ratio $R$ is exactly the least squares solution, $\hat{\beta} = r$, and we can expect ratio estimation to produce "good" results.

Of course, one can use the ratio estimation method with a SRS $\mathcal{Y}$ to obtain an estimate $\hat{\mu}_{Y;R}$ of $\mu_Y$ even if $\mathrm{V}(\varepsilon) \neq x\sigma^2$.

We have already determined the variance of this estimator:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{Y;R}) = \mathrm{V}(r\mu_X) = \mu_X^2\mathrm{V}(r) &\approx \frac{1}{n}(\sigma_Y^2 + R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y)\Big(\frac{N-n}{N-1}\Big)\\ &= \underbrace{\frac{\sigma_Y^2}{n}\Big(\frac{N-n}{N-1}\Big)}_{\mathrm{V}(\overline{y}_{\mathrm{SRS}})} + \frac{R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y}{n}\Big(\frac{N-n}{N-1}\Big).\end{aligned}$$

Consequently, $\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \gg \mathrm{V}(\hat{\mu}_{Y;R})$ if and only if $R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \ll 0$, which is to say if

$$\rho \gg \frac{R\sigma_X}{2\sigma_Y} = \frac{\mu_Y\sigma_X}{2\mu_X\sigma_Y} = \frac{1}{2}\cdot\frac{\mathrm{CV}_X}{\mathrm{CV}_Y}.$$

**Comparaison Between SRS and the Regression Method**

We have already determined the variance of the estimator $\hat{\mu}_{Y;L}$ of $\mu_Y$:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{Y;L}) &\approx (1-\rho^2)\frac{\sigma_Y^2}{n}\Big(\frac{N-n}{N-1}\Big) = \underbrace{\frac{\sigma_Y^2}{n}\Big(\frac{N-n}{N-1}\Big)}_{\mathrm{V}(\overline{y}_{\mathrm{SRS}})} - \rho^2\cdot\frac{\sigma_Y^2}{n}\Big(\frac{N-n}{N-1}\Big)\\ &= (1-\rho^2)\mathrm{V}(\overline{y}_{\mathrm{SRS}}).\end{aligned}$$

Consequently, $\mathrm{V}(\hat{\mu}_{Y;L}) \ll \mathrm{V}(\overline{y}_{\mathrm{SRS}})$ when $(1-\rho^2)\mathrm{V}(\overline{y}_{\mathrm{SRS}}) \ll \mathrm{V}(\overline{y}_{\mathrm{SRS}})$, which is to say that

$$1 - \rho^2 \ll 1 \iff 0 \ll |\rho| \leq 1.$$

**Comparaison Between SRS and the Difference Method**

We have already determined the variance of the estimator $\hat{\mu}_{Y;D}$ of $\mu_Y$:

$$\begin{aligned}\mathrm{V}(\hat{\mu}_{Y;D}) &= \frac{\sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2}{n}\Big(\frac{N-n}{N-1}\Big)\\ &= \underbrace{\frac{\sigma_Y^2}{n}\Big(\frac{N-n}{N-1}\Big)}_{\mathrm{V}(\overline{y}_{\mathrm{SRS}})} + \frac{\sigma_X^2 - 2\rho\sigma_X\sigma_Y}{n}\Big(\frac{N-n}{N-1}\Big).\end{aligned}$$

Consequently, $\mathrm{V}(\hat{\mu}_{Y;D}) \ll \mathrm{V}(\overline{y}_{\mathrm{SRS}})$ when $\sigma_X^2 - 2\rho\sigma_X\sigma_Y \ll 0 \iff \sigma_X^2 \ll 2\sigma_{XY}$.

**Comparaison Between the Ratio, Regression, and Difference Methods**

For each of the estimators $\hat{\mu}_{Y;\alpha}$, $\alpha \in \{R, L, D\}$, we have shown that the sampling variance takes the (approximate) form

$$\mathrm{V}(\hat{\mu}_{Y;\alpha}) \approx \mathrm{V}(\overline{y}_{\mathrm{SRS}}) + \frac{A_\alpha}{n}\left(\frac{N-n}{N-1}\right),$$

where

$$A_\alpha = \begin{cases} R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y, & \alpha = R \\ -\rho^2\sigma_Y^2, & \alpha = L \\ \sigma_X^2 - 2\rho\sigma_X\sigma_Y, & \alpha = D \end{cases}$$

In general, $\mathrm{V}(\hat{\mu}_{Y;\alpha}) \ll \mathrm{V}(\hat{\mu}_{Y;\gamma})$ if and only if $A_\alpha \ll A_\gamma$; these are the terms that must be compared to one another.

For instance,

$$\begin{aligned} \mathrm{V}(\hat{\mu}_{Y;R}) \gg \mathrm{V}(\hat{\mu}_{Y;L}) &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \gg -\rho^2\sigma_Y^2 \\ &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y + \rho^2\sigma_Y^2 \gg 0 \\ &\iff (R\sigma_X - \rho\sigma_Y)^2 \gg 0 \iff |R\sigma_X - \rho\sigma_Y| \gg 0 \\ &\iff R \gg \rho\frac{\sigma_Y}{\sigma_X} = \hat{\beta} \quad \text{or} \quad R \ll \hat{\beta} \end{aligned}$$

All things being equal, the regression estimator is preferable to the ratio estimator (according to their bounds on the error of estimation) when **the ratio is quite different from the slope of the regression line**.

Similarly,

$$\begin{aligned} \mathrm{V}(\hat{\mu}_{Y;D}) \gg \mathrm{V}(\hat{\mu}_{Y;L}) &\iff \sigma_X^2 - 2\rho\sigma_X\sigma_Y \gg -\rho^2\sigma_Y^2 \\ &\iff \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \rho^2\sigma_Y^2 \gg 0 \\ &\iff (\sigma_X - \rho\sigma_Y)^2 \gg 0 \iff |\sigma_X - \rho\sigma_Y| \gg 0 \\ &\iff 1 \gg \rho\frac{\sigma_Y}{\sigma_X} = \hat{\beta} \quad \text{or} \quad 1 \ll \hat{\beta}. \end{aligned}$$

All things being equal, the regression estimator is preferable to the difference estimator (according to their bounds on the error of estimation) when **the slope of the regression line takes a value far from** 1.

But the regression estimator is always **at least as good as the other two** since the latter two are special cases of regression estimation.

Finally, we can also compare the estimators by the ratio and by the difference:

$$\begin{aligned} \mathrm{V}(\hat{\mu}_{Y;R}) \gg \mathrm{V}(\hat{\mu}_{Y;D}) &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \gg \sigma_X^2 - 2\rho\sigma_X\sigma_Y \\ &\iff |R| \neq 1 \quad \text{and} \quad \sigma_X^2 \gg \frac{2}{R+1}\sigma_{XY} \end{aligned}$$

and

$$\mathrm{V}(\hat{\mu}_{Y;D}) \gg \mathrm{V}(\hat{\mu}_{Y;R}) \iff |R| \neq 1 \quad \text{and} \quad \sigma_X^2 \ll \frac{2}{R+1}\sigma_{XY}$$

Otherwise, the variances are of relatively similar magnitude.

## 10.6 Cluster Sampling

In practice, collecting sample data can require a tremendous amount of **travel**. Imagine a survey where the residents of the entire country are the **target population**, and a range of demographic and health indicators are measured about the **units**:

- age, height, weight, ethnicity, neighborhood, etc;
- blood pressure, blood cholesterol and mercury levels, body-mass index, etc.

Some of the information can be **self-reported by the units** (age, ethnicity, etc.), but in many cases (body-mass index, mercury levels, etc.), data collection requires the use of **health experts** and **specialized equipment**.

If all the sample units are from the Greater Toronto Area (GTA), say, it may be efficient to move the panel of experts (with all the required equipment in a trailer) from site to site, staying 2 weeks at each site. With about 20 sites in the GTA, data collection would take about a year to complete, but the cost of the survey would be greatly reduced: each night, the interviewers would **go home**; the cost of **moving the equipment** would also be minimized because of the small distances involved.

In a national study, where units could be drawn from several jurisdictions and remote locations, this approach is no longer necessarily recommended as it is potentially very expensive. Instead, one could start by taking a **first sample of geographic areas** (cities, regional municipalities, etc.), and then select a **sub-sample of units** (residents) in each of these areas.

Such a strategy is known as **multi-stage sampling** (M$n$S, see Section 10.7.3). Stratified sampling, for example, is a M2S for which the first level sample is a **census** and the second level sample is a SRS.

As another example, when the first level sample comes from a SRS and the second level sample is a **census** (all units are selected), we speak of **cluster sampling** (CLS).

### 10.6.1 Estimators and Confidence Intervals

As it was the case in the second chapter, we are interested in a finite population $\mathcal{U} = \{u_1, \ldots, u_N\}$ of expectation $\mu$ and variance $\sigma^2$.

Suppose we can cover the population with $M$ disjoint **clusters** containing, respectively, $N_1, \ldots, N_M$ units, so that $N_1 + \cdots + N_M = N$:

$$\mathcal{G}_1 = \{u_{1,1}, \ldots, u_{1,N_1}\}, \ \cdots \ , \ \mathcal{G}_M = \{u_{M,1}, \ldots, u_{M,N_M}\},$$

with cluster **expectation**, **total**, and **variance** given by

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{i,j}, \quad \tau_i = N_i \mu_i, \quad \text{and} \quad \sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{i,j}^2 - \mu_i^2, \quad 1 \leq i \leq M.$$

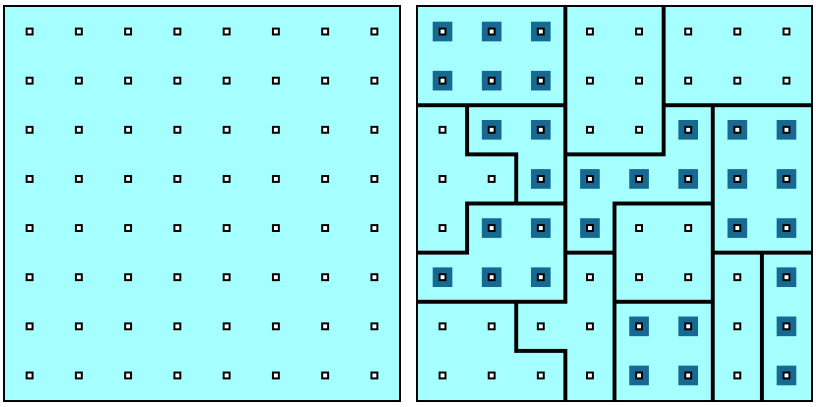


**Figure 10.10:** Schematics of CLS: target population (left) and sample (right).

A **cluster random sample** (CLS) $\mathcal{Y}$ is a subset of the target population $\mathcal{U}$ which is obtained by first drawing a SRS of $m > 1$ clusters, and then selecting all units in the selected clusters:

$$\mathcal{G}_{i_1} \cup \cdots \cup \mathcal{G}_{i_m} = \{\underbrace{y_{i_1,1}, \ldots, y_{i_1,N_{i_1}}}_{\text{cluster } \mathcal{G}_{i_1}}, \ldots, \underbrace{y_{i_m,1}, \ldots, y_{i_m,N_{i_m}}}_{\text{cluster } \mathcal{G}_{i_m}}\} \subseteq \bigcup_{\ell=1}^{M} \mathcal{G}_\ell = \mathcal{U}.$$

When $\mathcal{G}_{i_k}$ belongs to the CLS $\mathcal{Y}$, we denote its **mean**, **total**, and **variance** by $\overline{y}_{i_k}$, $y_{i_k}$, and $s^2_{i_k}$, respectively, for $1 \le k \le m$.

In a CLS design, each observation **has the same probability of being selected**, but the sample size **may change from one CLS to another**, unless the clusters all have the same size in the first place.

### Estimating the Mean $\mu$ for Clusters of Equal Size

Let us assume that all clusters have the same size: $N_1 = \cdots = N_M = n \implies N = Mn$. The **cluster mean** of the sample observations in $\mathcal{Y}$ is an estimator of $\mu$:

$$\overline{y}_C = \frac{1}{mn} \sum_{k=1}^{m} \sum_{j=1}^{n} y_{i_k,j} = \frac{1}{mn} \sum_{k=1}^{m} y_{i_k} = \frac{1}{m} \sum_{k=1}^{m} \overline{y}_{i_k} = \frac{1}{m} \sum_{k=1}^{m} \mu_{i_k}.$$

Therefore, the cluster average is simply the **average of the selected cluster averages**. This is not surprising since

$$\mu = \frac{1}{N} \sum_{\ell=1}^{M} \sum_{j=1}^{n} u_{\ell,j} = \frac{1}{Mn} \sum_{\ell=1}^{M} \sum_{j=1}^{n} u_{\ell,j} = \frac{1}{Mn} \sum_{\ell=1}^{M} \tau_\ell = \frac{1}{M} \sum_{\ell=1}^{M} \mu_\ell.$$

We can easily show that $\overline{y}_C$ is an **unbiased estimator** of $\mu$:

$$\mathrm{E}(\overline{y}_C) = \frac{1}{m} \sum_{k=1}^{m} \mathrm{E}(\mu_{i_k}) = \frac{1}{m} \sum_{k=1}^{m} \mu = \mu.$$

Furthemore, its **sampling variance** is

$$\mathrm{V}(\overline{y}_C) = \frac{\sigma^2_C}{m} \left( \frac{M-m}{M-1} \right), \quad \text{where } \sigma^2_C = \frac{1}{M} \sum_{\ell=1}^{M} (\mu_\ell - \mu)^2,$$

since clusters are drawn using an SRS. Indeed, $\overline{y}_C$ is the mean of a SRS with $m$:

$$\{\mu_{i_1}, \ldots, \mu_{i_m}\} \subseteq \{\mu_1, \ldots, \mu_M\}.$$

**Central Limit Theorem – CLS:** if $m$ and $M - m$ are sufficiently large, then

$$\overline{y}_C \sim_{\text{approx.}} \mathcal{N}\left(\mathrm{E}(\overline{y}_C), \mathrm{V}(\overline{y}_C)\right) = \mathcal{N}\left(\mu, \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right)\right).$$

In a CLS, the **bound on the error of estimation** is thus

$$B_{\mu;C} = 2\sqrt{\mathrm{V}(\overline{y}_C)} = 2\sqrt{\frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right)},$$

and the corresponding **95% C.I. for** $\mu$ is simply

$$\text{C.I.}_C(\mu; 0.95): \quad \overline{y}_C \pm B_{\mu;C}.$$

In practice, the **variance of the cluster means** $\sigma_C^2$ is rarely known – the empirical variance (and the corresponding **correction factor**) is used instead:

$$\hat{\mathrm{V}}(\overline{y}_C) = \frac{s_C^2}{m}\left(1 - \frac{m}{M}\right), \text{ where } s_C^2 = \frac{1}{m-1}\sum_{k=1}^{m}(\overline{y}_{i_k} - \overline{y}_C)^2.$$

The **bound on the error of estimation** is then approximated by

$$B_{\mu;C} \approx \hat{B}_{\mu;C} = 2\sqrt{\hat{\mathrm{V}}(\overline{y}_C)} = 2\sqrt{\frac{s_C^2}{m}\left(1 - \frac{m}{M}\right)},$$

$$\Longrightarrow \text{C.I.}_C(\mu; 0.95): \quad \overline{y}_C \pm \hat{B}_{\mu;C} \equiv \overline{y}_C \pm 2\sqrt{\frac{s_C^2}{m}\left(1 - \frac{m}{M}\right)}.$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$, each of size $n = 851$. We draw a SRS of $m = 6$ clusters. The means of these clusters are:

$$\overline{y}_1 = 120.7, \overline{y}_2 = 75.2, \overline{y}_3 = 116.3, \overline{y}_4 = 111.1, \overline{y}_5 = 116.9, \overline{y}_6 = 96.6.$$

Find a 95% C.I. for the mean $\mu$.

The bound on the error of estimation for $\mu$ is $\approx \hat{B}_{\mu;C} = 2\sqrt{\hat{\mathrm{V}}(\overline{y}_C)}$; we see that

$$\overline{y}_C = \tfrac{1}{6}\Sigma_{k=1}^{6}\overline{y}_k \approx 106.1, \; s_C^2 = \tfrac{1}{6-1}\Sigma_{k=1}^{6}(\overline{y}_k - \overline{y}_C)^2 = \tfrac{69089.6 - 6(106.1)^2}{6-1} \approx 300.8,$$

from which we have

$$\text{C.I.}_C(\mu; 0.95) \approx 106.1 \pm 2\sqrt{\tfrac{300.8}{6}\left(1 - \tfrac{6}{44}\right)} \equiv (93.0, 119.3).$$

**Estimating the Mean $\mu$ for Clusters of Different Sizes**

In practice, the clusters are often all of **different** sizes, so we could write

$$\mu = \frac{\sum_{\ell=1}^{M}\sum_{j=1}^{N_\ell} u_{\ell,j}}{\sum_{\ell=1}^{M} N_\ell} = \frac{\sum_{\ell=1}^{M}\tau_\ell}{\sum_{\ell=1}^{M} N_\ell},$$

where $\tau_\ell$ is the sum of $u_{\ell,j}$ for units in the cluster $\mathcal{G}_\ell$, $1 \le \ell \le M$.[49]

49: If $N_1 = \cdots = N_M = n$, the formulas we will develop will collapse to those seen in the preceding section.

If we still draw $m$ clusters from the population of $M$ clusters using an SRS, the form of $\mu$ suggests the use of the following estimator:

$$\overline{y}_C = \frac{\sum_{k=1}^{m}\sum_{j=1}^{N_{i_k}} y_{i_k,j}}{\sum_{k=1}^{m} N_{i_k}} = \frac{\sum_{k=1}^{m} y_{i_k}}{\sum_{k=1}^{m} N_{i_k}},$$

where we are using the notation of Section 10.5.

If the **average cluster size** is $\overline{N} = \frac{N}{M}$, this is similar to the situation that leads to **ratio estimation of the mean**. By performing the mapping $(\overline{y}_C, \mu, \overline{N}, \tau_\ell, N_\ell) \leftrightsquigarrow (r, R, \mu_X, Y_j, X_j)$, we can therefore conclude that $\overline{y}_C$ is a **biased estimator** of $\mu$, whose **sampling variance** is

$$\mathrm{V}(\overline{y}_C) \approx \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(\frac{M-m}{M-1}\right)\cdot\frac{1}{M}\sum_{\ell=1}^{M}\underbrace{(\tau_\ell - \mu N_\ell)}_{=N_\ell(\mu_\ell-\mu)}{}^2.$$

Consequently, the **bound on the error of estimation** is given by

$$B_{\mu;C} = 2\sqrt{\mathrm{V}(\overline{y}_C)} \approx 2\sqrt{\frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(\frac{M-m}{M-1}\right)\cdot\frac{1}{M}\sum_{\ell=1}^{M}(\tau_\ell - \mu N_\ell)^2}.$$

In practice, we often only have access to the sampled clusters – we must then use the **empirical variance**:

$$\begin{aligned}
\hat{\mathrm{V}}(\overline{y}_C) &\approx \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(1-\frac{m}{M}\right)\cdot\frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_C N_{i_k})^2 \\
&= \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\left(1-\frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y), \quad \text{where} \\
s_Y^2 &= \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{\overline{y}})^2, \quad s_N^2 = \frac{1}{m-1}\sum_{k=1}^{m}(N_{i_k} - \overline{N})^2, \\
\hat{\rho} &= \frac{\sum_{k=1}^{m}(y_{i_k} - \overline{\overline{y}})(N_{i_k} - \overline{N})}{\sqrt{\sum_{k=1}^{m}(y_{i_k} - \overline{\overline{y}})^2 \sum_{k=1}^{m}(N_{i_k} - \overline{N})^2}} \quad , \quad \overline{\overline{y}} = \frac{1}{m}\sum_{k=1}^{m} y_{i_k}.
\end{aligned}$$

Since it is not always possible to determine the average $\overline{N}$ of the clusters in the population $\mathcal{U}$, we often use $\overline{n}$, the **average cluster size in the**

**sample** $\mathcal{Y}$ instead:

$$\overline{n} = \frac{N_{i_1} + \cdots + N_m}{m}.$$

The **bound on the error of estimation** is thus

$$\hat{B}_{\mu;C} \approx 2\sqrt{\frac{1}{\overline{n}^2} \cdot \frac{1}{m}\left(1 - \frac{m}{M}\right)(s_Y^2 + s_N^2 \overline{y}_C^2 - 2\overline{y}_C \hat{\rho} s_N s_Y)}$$

and the **approximate 95% C.I. for** $\mu$ is

$$\text{C.I.}_C(\mu; 0.95): \quad \overline{y}_C \pm \hat{B}_{\mu;C}.$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$. We draw a SRS of $m = 6$ clusters. The means of the observations in these clusters are:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\overline{y}_k$ | 120.7 | 75.2 | 116.3 | 111.1 | 116.9 | 96.6 |
| $N_k$ | 850 | 176 | 1011 | 1001 | 843 | 910 |

Find a 95% C.I. for the mean $\mu$.

The bound on the error of estimation is $\approx \hat{B}_{\mu;C} = 2\sqrt{\hat{\text{V}}(\overline{y}_C)}$; we see that

$$\overline{y}_C = \frac{\sum_{k=1}^6 N_k \overline{y}_k}{\sum_{k=1}^6 N_k} = \frac{531073.3}{4791} \approx 110.8, \quad \overline{n} = \frac{1}{6}\sum_{k=1}^6 N_k = \frac{4791}{6} = 798.5$$

$$\overline{\overline{y}} = \frac{\sum_{k=1}^6 N_k \overline{y}_k}{6} = \frac{531073.3}{6} = 88,512.2,$$

$$s_N^2 = \frac{1}{6-1}\sum_{k=1}^6 (N_k - \overline{n})^2 = 98,146.7$$

$$s_Y^2 = \frac{1}{6-1}\sum_{k=1}^6 (N_k \overline{y}_k - \overline{\overline{y}})^2 = 1,465,229,403.4,$$

$$\hat{\rho} = \frac{\sum_{k=1}^6 (N_k - \overline{n})(N_k \overline{y}_k - \overline{\overline{y}})}{\sqrt{\sum_{k=1}^6 (N_k - \overline{n})^2 \sum_{k=1}^6 (N_k \overline{y}_k - \overline{\overline{y}})^2}} \approx 0.9796$$

$$s_Y^2 + s_N^2 \overline{y}_C^2 - 2\overline{y}_C \hat{\rho} s_N s_Y = 66,814,598.95$$

from which we conclude that

$$\hat{\text{V}}(\overline{y}_C) = \frac{1}{798.5^2} \cdot \frac{1}{6}\left(1 - \frac{6}{44}\right)(66,814,598.95) \approx 15.1$$

and $\text{C.I.}_C(\mu; 0.95) \approx 110.8 \pm 2\sqrt{15.1} \equiv (103.1, 118.6)$.

**Example** Find a 95% C.I. for the average life expectancy by country in 2011 (including India and China), using a CLS of size $m = 8$, assuming that the $N = 185$ countries have been grouped into $M = 22$ **clusters** determined by **geographic regions**.

We re-use the code from the previous sections,[50] The cluster sizes in the population are as follows.

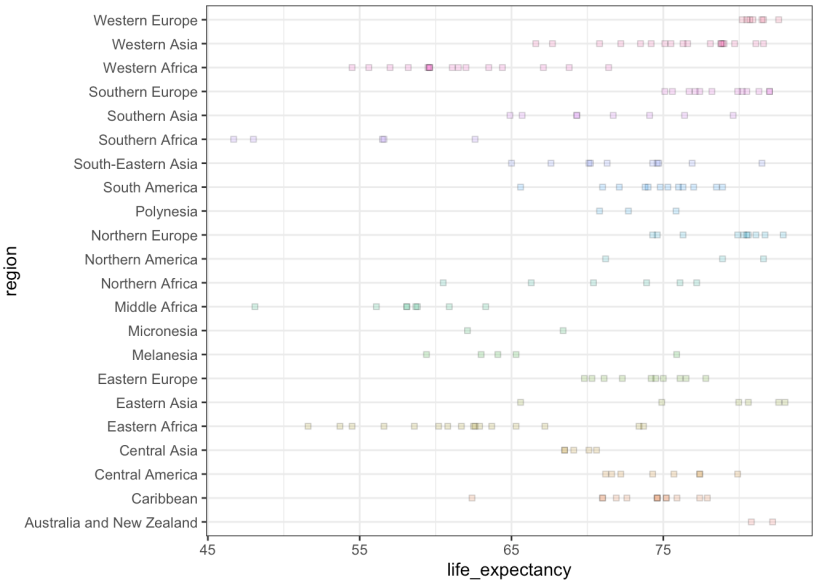
50: With some modifications, in particular with respect to the **clusters** (`region`.

```
gapminder.CLS <- gapminder |>  filter(year==2011) |> select(life_expectancy, region)
summary(gapminder.CLS,22)
```

```
life_expectancy                       region
 Min.   :46.70   Australia and New Zealand: 2
 1st Qu.:65.30   Caribbean                :13
 Median :73.70   Central America          : 8
 Mean   :71.18   Central Asia             : 5
 3rd Qu.:77.40   Eastern Africa           :16
 Max.   :83.02   Eastern Asia             : 6
                 Eastern Europe           :10
                 Melanesia                : 5
                 Micronesia               : 2
                 Middle Africa            : 8
                 Northern Africa          : 6
                 Northern America         : 3
                 Northern Europe          :10
                 Polynesia                : 3
                 South America            :12
                 South-Eastern Asia       :10
                 Southern Africa          : 5
                 Southern Asia            : 8
                 Southern Europe          :12
                 Western Africa           :16
                 Western Asia             :18
                 Western Europe           : 7
```

We note that the average life expectancy is $\mu = 71.18$. We can explore the distribution of life expectancy by cluster using the following code:

```
ggplot(data=gapminder.CLS, aes(x=life_expectancy, y=region, fill=region)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```



We notice a significant variability between some clusters (Southern Africa vs. Southern Europe, for example), but there is still a lot of overlap (which is a good sign). Next, we draw a SRS of $m = 8$ clusters:

```
set.seed(12345) # for replicability
regions=unique(gapminder.CLS[,"region"])
M=length(regions); m=8
(sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1] 14 19 16 11  2 21  6  7
```

We provide a summary of the observations in the sampled clusters:

```
sample.ind = gapminder.CLS$region %in% regions[sample.reg]
gapminder.CLS.n = gapminder.CLS[sample.ind,]
gapminder.CLS.n$region <- as.factor(gapminder.CLS.n$region)
(summ = gapminder.CLS.n |> group_by(region) |>
    summarise(N=n(), y.bar=mean(life_expectancy),
              total.y=sum(life_expectancy)))
```

```
# A tibble: 8 × 4
  region                       N y.barre total.y
  <fct>                    <int>   <dbl>   <dbl>
1 Australia and New Zealand     2    81.5    163
2 Central America               8    75.0    600.
3 Central Asia                  5    69.4    347.
4 Melanesia                     5    65.5    328.
5 Northern Africa               6    70.7    424.
6 Northern America              3    77.2    232.
7 South-Eastern Asia           10    72.6    726.
8 Western Asia                 18    75.8   1364.
```

We can also produce a summary of this summary:

```
(summ.final = summ |>
    summarise(sum.N = sum(N), moy.N = mean(N),
              y.bar.bar = mean(total.y),
              sum.y.bar = sum(total.y)))
```

```
# A tibble: 1 × 4
  sum.N moy.N y.barre.barre sum.y.barre
  <int> <dbl>         <dbl>       <dbl>
1    57  7.12          523.       4184.
```

We can now calculate the cluster estimator:

```
(est.y.bar.G=summ.final$sum.y.bar/summ.final$sum.N)
```

```
[1] 73.40316
```

Next, its sampling variance:

```
s2.Y = var(summ$total.y)
s2.N = var(summ$N)
rho = cor(summ$N,summ$total.y)
V.est.y.G = 1/summ.final$moy.N^2*1/m*(1-m/M)*
    (s2.Y+s2.N*est.y.bar.G^2-
    2*est.y.bar.G*rho*sqrt(s2.N*s2.Y))
```

The bound on the error of estimation and the 95% C.I. for $\mu$ are:

```
B = 2*sqrt(V.est.y.G)
c(est.y.bar.G - B,est.y.bar.G + B)
```

```
[1] 71.35310 75.45321
```

The performance of CLS is generally worse than that of SRS and/or STS – no surprise, given the discussion at the beginning of this section. The nature of the clusters may also play a role (in contrast to STS, CLS is more efficient when the cluster structure is **similar from one cluster to another**), which is not really the case here. We will discuss this further.

#### Estimating the Total $\tau$

Most of the work has already been done: since the **total** $\tau$ can be rewritten as

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

we can estimate the total with a CLS using the formula

$$\hat{\tau}_C = N\overline{y}_C.$$

There are two possibilities: either $N_1 = \cdots = N_M = n$, or the clusters are not all the same size.

If $N_1 = \cdots = N_M = n$, we have an **unbiased** estimator of $\tau$:

$$\mathrm{E}(\hat{\tau}_C) = \mathrm{E}(N\overline{y}_C) = N \cdot \mathrm{E}(\overline{y}_C) = N\mu = \tau,$$

$$\mathrm{V}(\hat{\tau}_C) = N^2 \cdot \mathrm{V}(\overline{y}_C) = N^2 \cdot \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right) \approx N^2 \cdot \hat{\mathrm{V}}(\overline{y}_C) = N^2 \cdot \frac{s_C^2}{m}\left(1 - \frac{m}{M}\right).$$

If the clusters are of different sizes, we have a **biased** estimator of $\tau$, with **sampling variance** given by

$$\begin{aligned}\mathrm{V}(\hat{\tau}_C) = \mathrm{V}(N\overline{y}_C) &= N^2 \cdot \mathrm{V}(\overline{y}_C) \approx N^2 \cdot \hat{\mathrm{V}}(\overline{y}_C) \\ &= \frac{N^2}{\overline{N}^2} \cdot \frac{1}{m}\left(1 - \frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y) \\ &= M^2 \cdot \frac{1}{m}\left(1 - \frac{m}{M}\right)(s_Y^2 + s_N^2\overline{y}_C^2 - 2\overline{y}_C\hat{\rho}s_N s_Y).\end{aligned}$$

The estimator follows an approximate normal distribution

$$\hat{\tau}_C \sim_{\text{approx}} \mathcal{N}\left(\mathrm{E}(\hat{\tau}_C), \hat{\mathrm{V}}(\hat{\tau}_C)\right),$$

as long as the quantities $m$, and $M - m$ are both "large enough".

In both cases, the **bound on the error of estimation** is

$$B_{\tau;C} \approx \hat{B}_{\tau;C} = 2\sqrt{\hat{\mathrm{V}}(\hat{\tau}_C)}$$

and the **95% C.I. for** $\tau$ takes the usual form:

$$\text{C.I.}_C(\tau; 0.95): \quad \hat{\tau}_C \pm \hat{B}_{\tau;C}.$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$, each of size $n = 851$. We draw a SRS of $m = 6$ clusters. The means of the observations in these clusters are:

$$\overline{y}_1 = 120.7,\ \overline{y}_2 = 75.2,\ \overline{y}_3 = 116.3,\ \overline{y}_4 = 111.1,\ \overline{y}_5 = 116.9,\ \overline{y}_6 = 96.6.$$

Find a 95% C.I. for the total $\tau$ in $\mathcal{U}$.

We have previously seen that $\text{C.I.}_C(\mu; 0.95) \equiv (93.0, 119.3)$ for this CLS, with $N_1 = \cdots = N_6 = 851$. Therefore,

$$\text{C.I.}_C(\tau; 0.95) \approx 37444(93.0, 119.3) \equiv (3481307.7, 4466805.3).$$

**Example** Consider a finite population $\mathcal{U}$ of size $N = 37,444$, divided into $M = 44$ clusters $\mathcal{G}_\ell$. We draw a SRS of $m = 6$ clusters. The mean of the observations in these clusters are:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\overline{y}_k$ | 120.7 | 75.2 | 116.3 | 111.1 | 116.9 | 96.6 |
| $N_k$ | 850 | 176 | 1011 | 1001 | 843 | 910 |

Find a 95% C.I. for the total $\tau$ in $\mathcal{U}$.

We have already seen in a previous example that $\text{C.I.}_C(\mu; 0.95) \equiv (103.1, 118.6)$ for this CLS with different cluster sizes. Therefore,

$$\text{C.I.}_C(\tau; 0.95) \approx 37444(103.1, 118.6) \equiv (3860476, 4440858).$$

**WARNING:** how do we do this **if the size $N$ of the population is unknown?** Note that

$$\tau = \sum_{\ell=1}^{M} \tau_\ell = M \cdot \frac{1}{M} \sum_{\ell=1}^{M} \tau_\ell = M\overline{\tau},$$

where $\overline{\tau}$ is **mean of the cluster totals in the population**.

We could then use the estimator

$$M\overline{y}_T = M \cdot \frac{1}{m} \sum_{k=1}^{m} y_{i_k},$$

where $\overline{y}_T$ is the **mean of the $m$ cluster totals in the CLS**.

In that case, we are dealing with a SRS of size $m$, drawn from $M$ cluster totals, i.e., this is an **unbiased** estimator:

$$\begin{aligned}
&\mathrm{E}(M\overline{y}_T) = \tau \\
&\mathrm{V}(M\overline{y}_T) = M^2 \cdot \mathrm{V}(\overline{y}_T) = M^2 \cdot \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right) \\
&\hat{\mathrm{V}}(M\overline{y}_T) \approx M^2 \cdot \hat{\mathrm{V}}(\overline{y}_T) = M^2 \cdot \frac{s_T^2}{m}\left(1-\frac{m}{M}\right),
\end{aligned}$$

where

$$\sigma_T^2 = \frac{1}{M}\sum_{\ell=1}^{M}(\tau_\ell - \overline{\tau})^2 \quad \text{and} \quad s_T^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_T)^2.$$

The estimator follows an approximate normal distribution

$$M\overline{y}_T \sim_{\text{approx}} \mathcal{N}\left(\tau, \hat{\mathrm{V}}(M\overline{y}_T)\right),$$

as long as the quantities $m$, and $M - m$ are both "large enough".

The **bound on the error of estimation** is then

$$B_{\tau;T} \approx \hat{B}_{\tau;T} = 2\sqrt{\hat{\mathrm{V}}(M\overline{y}_T)}$$

and the 95% C.I. for $\tau$ takes the usual form:

$$\text{C.I.}_T(\tau; 0.95): \quad M\overline{y}_T \pm \hat{B}_{\tau;T}.$$

**Example** Consider a finite population $\mathcal{U}$ of unknown size, divided into $M = 44$ clusters $\mathcal{G}_\ell$. We draw a SRS of $m = 6$ clusters. The mean of the observations in these clusters are:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\overline{y}_k$ | 120.7 | 75.2 | 116.3 | 111.1 | 116.9 | 96.6 |
| $N_k$ | 850 | 176 | 1011 | 1001 | 843 | 910 |

Find a 95% C.I. for the total $\tau$ in $\mathcal{U}$.

Since the population size $N$ is unknown, the bound on the error of estimation for $\tau$ is $\approx \hat{B}_{\tau;T} = 2\sqrt{\hat{\mathrm{V}}(M\overline{y}_T)}$; we see that

$$\overline{y}_T = \frac{1}{6}\sum_{k=1}^{6} N_k\overline{y}_k = \frac{531073.3}{6} \approx 88512.2, \quad M\overline{y}_T = 44(88512.2) = 3894537.5$$

and

$$s_T^2 = \frac{1}{6-1}\sum_{k=1}^{6}(N_k\overline{y}_k - \overline{y}_T)^2 = \frac{1}{5}\left(\sum_{k=1}^{6} N_k^2\overline{y}_k^2 - 6\overline{y}_T^2\right) = 1465229403,$$

from which we conclude that

$$\hat{\mathrm{V}}(M\overline{y}_T) = (44)^2 \cdot \frac{1465229403}{6}\left(1 - \frac{6}{44}\right) = 408310593755.73$$

and $\mathrm{C.I.}_T(\tau; 0.95) \approx 3894537.5 \pm 2\sqrt{408310593755.73} \equiv (2616554, 5172521)$.

The estimator is unbiased, but the confidence interval for $\tau$ is much wider than that given by $\mathrm{C.I.}_C(\tau; 0.95) \equiv (3860476, 4440858)$; this is not surprising since we have more information in the latter case (namely, the size $N$ of the population).

**Example** Find a 95% C.I. for the world population in 2011 (excluding China and India), using a CLS of size $m = 8$, drawn from $M = 22$ clusters determined by geographic regions.

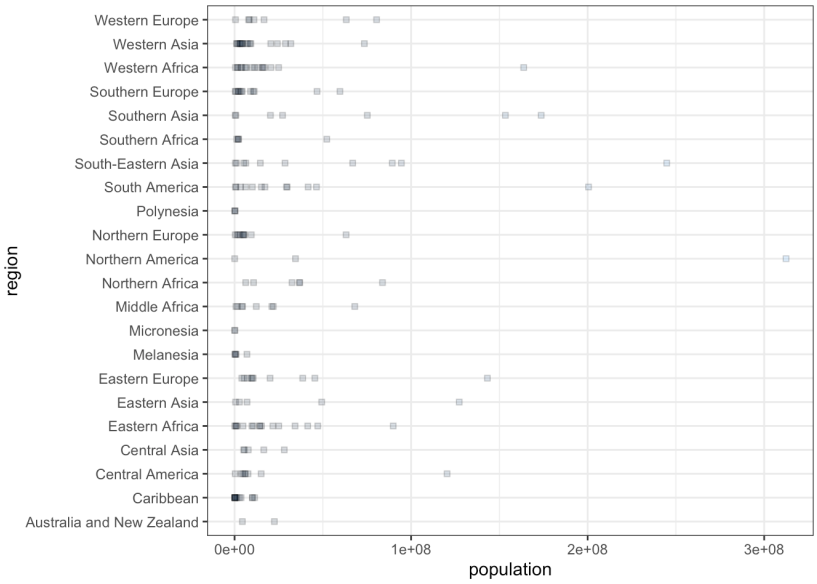
We re-use the code from the previous sections to create the clusters. The true population total is found below:

```
gapminder.CLS.pop <- gapminder |> filter(year==2011) |>
    select(population, region) |>
    filter(population < 500000000)
(sum(gapminder.CLS.pop$population))
```

```
[1] 4264258312
```

We start by studying the distribution of population by region:

```
ggplot(data=gapminder.CLS.pop, aes(x=population, y=region,
    fill=population)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(),
          legend.position="none")
```



The essential statistics are calculated as follows:

```
summ.pop = gapminder.CLS.pop |> group_by(region) |>
    summarise(N=n(), y.pop=mean(population),
              tau.pop=sum(population))
```

Next we draw a SRS of clusters:

```
set.seed(22) # for replicability
regions = unique(gapminder.CLS.pop[,"region"])
M=length(regions); m=8
N=nrow(gapminder.CLS.pop)
(sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1]  6  9 10 12 17  5 11  3
```

The sample is summarized as follows:

```
sample.ind = gapminder.CLS.pop$region %in%
        regions[sample.reg]
gapminder.CLS.T = gapminder.CLS.pop[sample.ind,]
gapminder.CLS.T$region <- as.factor(gapminder.CLS.T$region)
(summ.T = gapminder.CLS.T |> group_by(region) |>
    summarise(N=n(), tau=sum(population)))
```

```
# A tibble: 8 × 3
  region              N       tau
  <fct>           <int>     <int>
1 Central America     8 163510619
2 Eastern Europe     10 294249971
3 Middle Africa       8 134483803
4 Northern Europe    10  99989705
5 South America      12 401182686
6 Southern Asia       7 450825356
7 Western Africa     16 316604189
8 Western Asia       18 237909741
```

If we assume the number of units in the population to be known ($N = 183$), the estimator of the average population per country is:

```
(y.G = sum(summ.T$tau)/sum(summ.T$N))
```

```
[1] 23581529
```

The estimator for the total population (excluding China and India) is:

```
(tau.G = N*y.G)
```

```
[1] 4315419784
```

The bound on the error of estimation and the 95% C.I. for $\tau$ are:

```
s2.G =  1/(m-1)*sum((summ.T$tau-y.G*summ.T$N)^2)
V = M^2*s2.G/m*(1-m/M)
B = 2*sqrt(V)
c(tau.G-B,tau.G+B)
```

```
[1] 2441918142 6188921427
```

If we assume instead that the number of units is unknown, the estimator of the population per cluster is:

```
(y.T = sum(summ.T$tau)/m)
```

```
[1] 262344509
```

The estimator for the total population (excluding China and India) would then be:

```
(tau.T = M*y.T)
```

```
[1] 5771579192
```

The bound on the error of estimation and the 95% C.I. for $\tau$ in that case are computed below:

```
s2.T =  1/(m-1)*sum((summ.T$tau-y.T)^2)
V = M^2*s2.T/m*(1-m/M)
B = 2*sqrt(V)
c(tau.G-B,tau.G+B)
```

```
[1] 2746857662 5883981906
```

The actual value $\tau = 4,264,258,312$ is found within the 95% C.I.[51]

51: But different SRS of clusters might lead to different outcomes.

**Estimating a Proportion $p$**

In a population where $A_{\ell,j} \in \{0,1\}$ represents the absence or presence of a characteristic for the $j$th unit in the $\ell$th cluster, the **mean**

$$p = \frac{1}{N}\sum_{\ell=1}^{M}\sum_{j=1}^{N_\ell} A_{\ell,j} = \frac{\sum_{\ell=1}^{M} A_\ell}{\sum_{\ell=1}^{M} N_\ell}$$

is the **proportion** of the population units possessing the characteristic, where $A_\ell$ is the number of units with the characteristic in the $\ell$th cluster.

If we are still drawing $m$ clusters using a SRS from the $M$ clusters in the population, the form taken by $p$ suggests the use of the following estimator:

$$\hat{p}_C = \frac{\sum_{k=1}^{m}\sum_{j=1}^{N_{i_k}} a_{i_k,j}}{\sum_{k=1}^{m} N_{i_k}} = \frac{\sum_{k=1}^{m} a_{i_k}}{\sum_{k=1}^{m} N_{i_k}},$$

where $a_{i_k}$ is the number of units with the characteristic in the $k$th sampled cluster.

Set $\overline{N} = \frac{N}{M}$. If $N$ is unknown, we use $\overline{N} \approx \overline{n} = \frac{1}{m}(N_{i_1} + \cdots + N_{i_m})$. There are then two possibilities: either $N_1 = \cdots = N_M = n$, or the clusters are not all of the same size. If $N_1 = \cdots = N_M = n$, we have an **unbiased** estimator of $p$:

$$\mathrm{E}(\hat{p}_C) = p, \quad \mathrm{V}(\hat{p}_C) = \frac{1}{n^2} \cdot \frac{\sigma_P^2}{m}\left(\frac{M-m}{M-1}\right) \approx \frac{1}{n^2} \cdot \frac{s_P^2}{m}\left(1 - \frac{m}{M}\right) = \hat{\mathrm{V}}(\hat{p}_C),$$

where

$$\sigma_P^2 = \frac{1}{M}\sum_{\ell=1}^{M}(A_\ell - pN_\ell)^2 \quad \text{and} \quad s_P^2 = \frac{1}{m-1}\sum_{k=1}^{m}(a_{i_k} - \hat{p}_C N_{i_k})^2.$$

If the clusters are of different sizes, we have a **biased** estimator of $p$, whose **sampling variance** is:

$$\mathrm{V}(\hat{p}_C) \approx \frac{1}{\overline{N}^2} \cdot \frac{\sigma_P^2}{m}\left(\frac{M-m}{M-1}\right), \quad \hat{\mathrm{V}}(\hat{p}_C) \approx \frac{1}{\overline{n}^2} \cdot \frac{s_P^2}{m}\left(1 - \frac{m}{M}\right).$$

The estimator follows an approximate normal distribution

$$\hat{p}_C \sim_{\text{approx}} \mathcal{N}\left(\mathrm{E}(\hat{p}_C), \hat{\mathrm{V}}(\hat{p}_C)\right),$$

as long as the quantities $m$, and $M - m$ are both "large enough".

In both cases, the **bound on the error of estimation** is

$$B_{p;C} \approx \hat{B}_{p;C} = 2\sqrt{\hat{\mathrm{V}}(\hat{p}_C)}$$

and the **95% C.I. for** $p$ takes the usual form:

$$\text{C.I.}_C(p; 0.95): \quad \hat{p}_C \pm \hat{B}_{p;C}.$$

**Example** Find a 95% C.I. for the proportion of countries whose life expectancy is above 75 years in 2011, using a CLS with $m = 8$, assuming that the countries are grouped into $M = 22$ clusters determined by geographic regions.

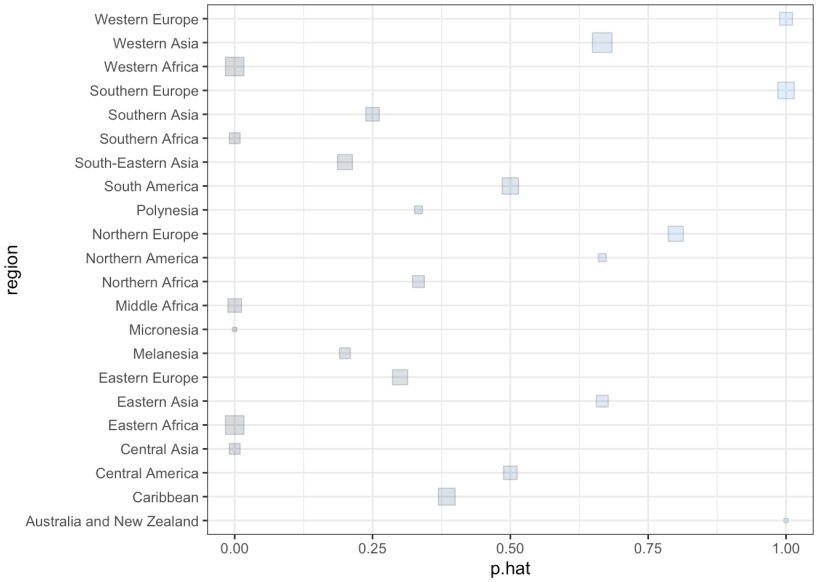
We re-use the code of the previous sections to create the clusters, and we create a new indicator variable for the 75 years life expectancy threshold:

```
gapminder.CLS$life.75 <- ifelse(
        gapminder.CLS$life_expectancy>75,1,0)
gapminder.CLS.75 <- gapminder.CLS |> select(life.75,region)
(mean(gapminder.CLS.75$life.75)) # true proportion
```

```
[1] 0.3945946
```

We begin by examining the proportions in each region:

```
summ.75 = gapminder.CLS.75 |>
    group_by(region) |>
    summarise(N=n(), p.hat=mean(life.75))

ggplot(data=summ.75,aes(x=p.hat, y=region, size=N, fill=p.hat)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```




The proportion of countries with a life expectancy of more than 75 years is found to vary greatly from region to region – this may affect the quality of the estimate.

Next, we draw a SRS of $m - 8$ clusters:

```
set.seed(0) # for replicability
regions = unique(gapminder.CLS[,"region"])
M=length(regions)
m=8
(sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1] 14  4  7  1  2 11 22 18
```

Then, we provide a summary of the proportions by cluster:

```
sample.ind = gapminder.CLS$region %in% regions[sample.reg]
gapminder.CLS.G = gapminder.CLS[sample.ind,]
gapminder.CLS.G$region <- as.factor(gapminder.CLS.G$region)
(summ.75.n = gapminder.CLS.G |>
    group_by(region) |>
    summarise(N=n(), p.hat=mean(life.75)))
```

```
# A tibble: 8 × 3
  region                       N p.hat
  <fct>                    <int> <dbl>
1 Australia and New Zealand     2 1
2 Caribbean                    13 0.385
3 Central America               8 0.5
4 Micronesia                    2 0
5 Northern Africa               6 0.333
6 Northern Europe              10 0.8
7 South-Eastern Asia           10 0.2
8 Southern Europe              12 1
```

We now have enough information to compute the CLS estimator of the proportion:

```
(p.G = sum(summ.75.n$N*summ.75.n$p.hat)/sum(summ.75.n$N))
```

```
[1] 0.5555556
```

Finally, we compute the sampling variance, the margin of error, and the 95% C.I. for $p$ (assuming that the average cluster size is not known):

```
mean.size = sum(summ.75.n$N)/m
s2.p.G =  1/(m-1)*sum((summ.75.n$N*summ.75.n$p.hat-
            p.G*summ.75.n$N)^2)
V = 1/mean.size^2*s2.p.G/m*(1-m/M)
(B = 2*sqrt(V))
c(p.G-B,p.G+B)
```

```
[1] 0.2025966
[1] 0.3529590 0.7581521
```

The actual value $p = 0.394$ is indeed within the 95% confidence interval. We assumed that the average cluster size was unknown; is this also the case if we use the known value $\overline{N} = \frac{185}{22} \approx 8.41$?

The observations of the Gapmider dataset are probably not that suitable for CLS ... at least, not if we use regions as clusters.

### 10.6.2 Sample Size

Depending on whether the clusters are of equal size or not, the variance formulas take different forms; however, they coincide when $N_i = n$ for all $i$; it is only the **nature of the estimator bias** and the **exactness of its sampling variance** that are affected.

Consequently, we will only study the situation where the clusters are assumed to be of different sizes. In what follows, we will use the notations

$$\sigma_E^2 = \frac{1}{M}\sum_{\ell=1}^{M}(\tau_\ell - \mu N_\ell)^2 \quad \text{and} \quad s_E^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_C N_{i_k})^2.$$

**Mean $\mu$**

If we want to estimate $\mu$ with $\overline{y}_C$, we use:

$$\begin{aligned} B_{\mu;C} = 2\sqrt{\frac{1}{\overline{N}^2}\cdot\frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)} &\iff \underbrace{\frac{B_{\mu;C}^2\overline{N}^2}{4}}_{=D_\mu} = \frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right) \\ &\iff \frac{(M-1)D_\mu}{\sigma_E^2} = \frac{M-m}{m} = \frac{M}{m} - 1 \\ &\iff \frac{(M-1)D_\mu + \sigma_E^2}{\sigma_E^2} = \frac{M}{m} \\ &\iff m_{\mu;C} = \frac{M\sigma_E^2}{(M-1)D_\mu + \sigma_E^2}. \end{aligned}$$

Obviously, we can only use this formula **if we know the variance** $\sigma_E^2$ of the cluster totals in the population $\mathcal{U}$. If that is not available, we can use the **empirical variance** $s_E^2$ from a **preliminary sample**, or that from **a prior survey**.[52]

52: If the average size $\overline{N}$ of the clusters of $\mathcal{U}$ is unknown, we use the **empirical average size** $\overline{n} = (N_{i_1} + \cdots + N_{i_m})/m$ from the preliminary sample.

Finally, note that this formula allows us to determine the **number of clusters** $m$ to be drawn from a SRS of clusters in order to obtain some margin of error on the estimate; the sample size may change from one realization to another, depending on the size of the sampled clusters.

**Example** Consider a company that wants a cost inventory for the $N = 625$ items in stock. In practice, it might be tedious to obtain a SRS of these items; however, the items are arranged on $M = 100$ shelves and it is relatively easy to select a SRS of shelves, treating each shelf as a cluster of items. How many shelves would need to be sampled in order to estimate the average value of all items in inventory with a bound on the error of estimation of at most $B_{\mu;C} = 1.25\$$, assuming $\sigma_E^2 \approx 317.53\$$?

Set $D_\mu = \frac{B_{\mu;C}^2\overline{N}^2}{4} = \frac{(1.25)^2(6.25)^2}{4} \approx 15.26$; then

$$m_{\mu;C} = \frac{M\sigma_E^2}{(M-1)D_\mu + \sigma_E^2} = \frac{100(317.53)}{(100-1)(15.26) + 317.53} = 17.4 \approx 18. \quad \blacksquare$$

**Total $\tau$**

If we want to estimate $\tau$ with $N\overline{y}_C$, we use:

$$B_{\tau;C} = 2\sqrt{M^2 \cdot \frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)} \iff \underbrace{\frac{B_{\tau;C}^2}{4M^2}}_{=D_{\tau;C}} = \frac{\sigma_E^2}{m}\left(\frac{M-m}{M-1}\right)$$

$$\iff \frac{(M-1)D_{\tau;C}}{\sigma_E^2} = \frac{M-m}{m} = \frac{M}{m} - 1$$

$$\iff \frac{(M-1)D_{\tau;C} + \sigma_E^2}{\sigma_E^2} = \frac{M}{m}$$

$$\iff m_{\tau;C} = \frac{M\sigma_E^2}{(M-1)D_{\tau;C} + \sigma_E^2}.$$

**Example** Consider a company that wants a cost inventory for the $N = 625$ items in stock. In practice, it might be tedious to obtain a SRS of these items; however, the items are arranged on $M = 100$ shelves and it is relatively easy to select a SRS of shelves, treating each shelf as a cluster of items. How many shelves would need to be sampled in order to estimate the total value of all items in inventory with a bound on the error of estimation of at most $B_{\tau;C} = 600\$$, assuming $\sigma_E^2 \approx 317.53\$$?

Set $D_{\tau;C} = \frac{B_{\tau;C}^2}{4M^2} = \frac{(600)^2}{4(100)^2} = 9$; then

$$m_{\tau;C} = \frac{M\sigma_E^2}{(M-1)D_{\tau;C} + \sigma_E^2} = \frac{100(317.53)}{(100-1)(9) + 317.53} = 26.3 \approx 27. \quad \blacksquare$$

If we want to estimate $\tau$ with $M\overline{y}_T$, we use:

$$B_{\tau;T} = 2\sqrt{M^2 \cdot \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right)} \iff \underbrace{\frac{B_{\tau;T}^2}{4M^2}}_{=D_{\tau}} = \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right)$$

$$\iff \frac{(M-1)D_{\tau}}{\sigma_T^2} = \frac{M-m}{m} = \frac{M}{m} - 1$$

$$\iff \frac{(M-1)D_{\tau} + \sigma_T^2}{\sigma_T^2} = \frac{M}{m}$$

$$\iff m_{\tau;T} = \frac{M\sigma_T^2}{(M-1)D_{\tau} + \sigma_T^2}.$$

**Example** Consider a company that wants a cost inventory for the $N = 625$ items in stock. In practice, it might be tedious to obtain a SRS of these items; however, the items are arranged on $M = 100$ shelves and it is relatively easy to select a SRS of shelves, treating each shelf as a cluster of items. How many shelves would need to be sampled in order to estimate the total value of all items in inventory with a bound on the error of estimation of at most $B_{\tau;T} = 600\$$, assuming $\sigma_T^2 \approx 682.77\$$?

Set $D_{\tau;T} = \frac{B_{\tau;T}^2}{4M^2} = \frac{(600)^2}{4(100)^2} = 9$; then

$$m_{\tau;T} = \frac{M\sigma_T^2}{(M-1)D_{\tau;T} + \sigma_T^2} = \frac{100(682.77)}{(100-1)(9) + 682.77} = 43.4 \approx 44. \quad \blacksquare$$

**Proportion $p$**

If we want to estimate $p$ with $\hat{p}_C$, we use:

$$\begin{aligned} B_{p;C} = 2\sqrt{\frac{1}{\overline{N}^2} \cdot \frac{\sigma_p^2}{m}\left(\frac{M-m}{M-1}\right)} &\iff \underbrace{\frac{B_{p;C}^2 \overline{N}^2}{4}}_{=D_{p;C}} = \frac{\sigma_p^2}{m}\left(\frac{M-m}{M-1}\right) \\ &\iff \frac{(M-1)D_{p;C}}{\sigma_P^2} = \frac{M-m}{m} = \frac{M}{m} - 1 \\ &\iff \frac{(M-1)D_{p;C} + \sigma_P^2}{\sigma_P^2} = \frac{M}{m} \\ &\iff m_{p;C} = \frac{M\sigma_P^2}{(M-1)D_{p;C} + \sigma_P^2}. \end{aligned}$$

### 10.6.3 Comparison Between SRS and CLS

Consider a ClS $\mathcal{Y}$ consisting of $m$ clusters drawn from a population $\mathcal{U}$ of size $N$, distributed in $M$ clusters. Let $\mu$ be the mean and $\sigma^2$ the variance of the population $\mathcal{U}$.

If the clusters are all of size $n$, we can show that

$$\mathrm{V}(\overline{y}_C) \approx \frac{\sigma^2 - \overline{\sigma^2}}{m}\left(1 - \frac{m}{M}\right), \quad \text{where} \quad \overline{\sigma^2} = \frac{1}{M}\sum_{\ell=1}^{M} \sigma_\ell^2,$$

where $\sigma_\ell^2$ is the variance in the $\ell$th cluster.

But we can also consider $\mathcal{Y}$ as having arisen from a SRS with size $mn$. In that case, we have

$$\mathrm{V}(\overline{y}_{\mathrm{SRS}}) = \frac{\sigma^2}{mn}\left(\frac{N-mn}{N-1}\right) \approx \frac{\sigma^2}{mn}\left(1 - \frac{mn}{N}\right) = \frac{\sigma^2}{mn}\left(1 - \frac{mn}{Mn}\right) = \frac{\sigma^2}{mn}\left(1 - \frac{m}{M}\right),$$

from which we conclude that

$$\begin{aligned} \mathrm{V}(\overline{y}_C) - \mathrm{V}(\overline{y}_{\mathrm{SRS}}) &\approx \frac{1}{m}\left(1 - \frac{m}{M}\right)\left(\sigma^2 - \overline{\sigma^2} - \frac{\sigma^2}{n}\right) = \frac{1}{m}\left(1 - \frac{m}{M}\right)\left(\frac{n-1}{n}\sigma^2 - \overline{\sigma^2}\right) \\ &\approx \frac{1}{m}\left(1 - \frac{m}{M}\right)(\sigma^2 - \overline{\sigma^2}), \quad \text{si } n - 1 \approx n. \end{aligned}$$

Consequently, $\mathrm{V}(\overline{y}_C) \gg \mathrm{V}(\overline{y}_{\mathrm{SRS}})$ if and only if $\sigma^2 \gg \overline{\sigma^2}$, which is the case when the **mean of the cluster variances is smaller than the variance in the population**.

The moral of the story is that a ClS is effective if the clusters, regardless of their size, are **as heterogeneous as the population itself**.

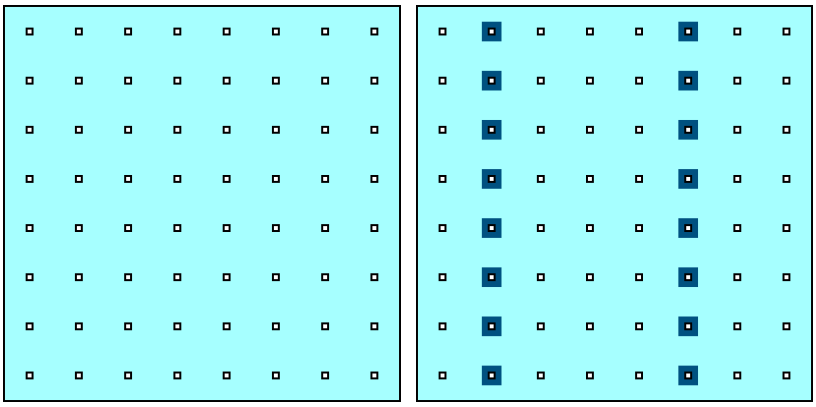


**Figure 10.11:** Schematics of SYS: target population (left) and sample (right).

## 10.7 Special Topics

We complete this introduction to survey sampling by discussing a few additional topics.[53]

53: A few of which could even be called advanced.

### 10.7.1 Systematic Sampling

With the advent of easy-to-access pseudo-random number generators,[54] it is not very arduous to draw a pseudo SRS $\mathcal{Y}$ of size $n$ from a population $\mathcal{U}$ of size $N$ (assuming that we have an appropriate sampling frame, of course).

54: Excel, `R`, SAS, `Python`, etc.

However, it remains possible for the obtained sample to **not be representative** of the population: a SRS of countries that do not include China or India, for example, would not be very useful if we are trying to estimate the average population of the world's countries.

In some cases, a **systematic sampling design** (SYS) can be used to maximize the probability that the random sample $\mathcal{Y}$ represents the population.

Here is how we draw a $1-\text{in}-M$ systematic sample of size $n$ (or $n+1$) from an ordered list of size $N$:

1. determine the integer part $M = \lfloor \frac{N}{n} \rfloor$;
2. randomly select an integer $\gamma$ in $\{1, 2, \ldots, M\}$;
3. the sample $\mathcal{Y}$ then contains the values corresponding to units

$$\underbrace{\gamma, \gamma + M,\ \gamma + 2M,\ \ldots\ , \gamma + (n-1)M}_{n \text{ units}},\ \underbrace{\gamma + nM}_{\text{if } \gamma + nM \leq N}.$$

If the ordering of the units in the sampling frame is fixed, there can only be $M$ different SYS samples of size $n$ (or $n+1$, in some cases).

**Example** The Gapminder dataset contains socio-economic information on $N = 185$ countries in 2011. What are the average life expectancy and population of the world's countries?

We modify the code allowing us to access the data set:

```
gapminder.SYS <- gapminder |> filter(year==2011) |>
    select(country, life_expectancy, population)
N=nrow(gapminder.SYS)
```

There are 185 units in the data set. If we are interested in a SYS of size $n = 20$, say, the integer $M$ is:

```
n=20
(M=floor(N/n))
```

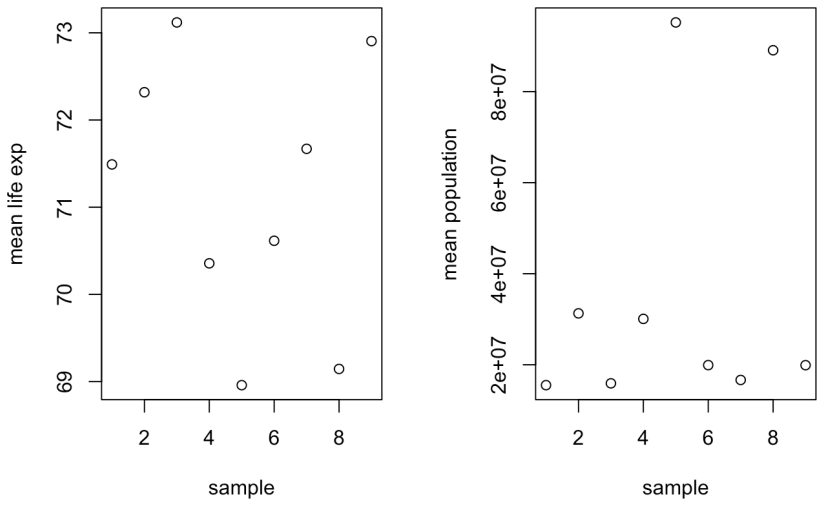
```
[1] 9
```

The vector of observations $0, M, 2M, \ldots, nM$ is therefore:

```
index = M*(0:n)
```

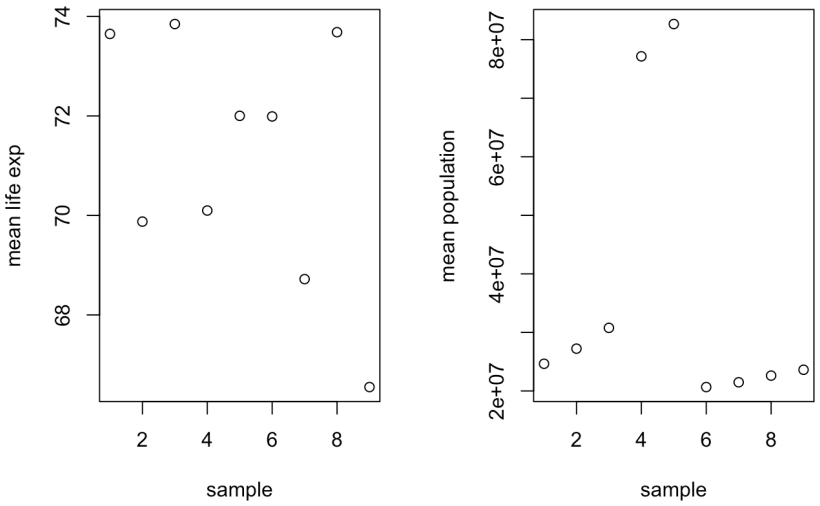
We construct $M = 9$ samples $\mathcal{Y}_i$, $i = 1, \ldots, 9$, assuming that the units appear in alphabetical order (by country name) in the dataset.

```
moy.SYS.life_exp = c() # initialization - life expectation
moy.SYS.pop = c()      # initialization - population

for(j in 1:M){# all SYS of size n or n+1, alpha order
    index.tmp = j + index
    index.tmp <- index.tmp[index.tmp < N+1]  # keeping indices <= N
    sample.sys = gapminder.SYS[index.tmp,2:3]
    moy.SYS.life_exp[j]=mean(sample.sys$life_expectancy)
    moy.SYS.pop[j]=mean(sample.sys$population)
  }

# charts
par(mfrow=c(1,2))
plot(moy.SYS.life_exp, xlab="sample", ylab="mean life exp")
plot(moy.SYS.pop, xlab="sample", ylab="mean population")
```



Could you identify the sample that contains China or India? What if we change the order in which the countries are listed in the dataset?

```
gapminder.SYS <- gapminder.SYS[order(gapminder.SYS$population),]

for(j in 1:M){# all SYS of size n or n+1, population order
    index.tmp = j + index
    index.tmp <- index.tmp[index.tmp < N+1]
    sample.sys = gapminder.SYS[index.tmp,2:3]
    moy.SYS.life_exp[j]=mean(sample.sys$life_expectancy)
    moy.SYS.pop[j]=mean(sample.sys$population)
  }

par(mfrow=c(1,2))
plot(moy.SYS.life_exp, xlab="sample", ylab="mean life exp")
plot(moy.SYS.pop, xlab="sample", ylab="mean population")
```




We obtain similar results when ordering the units in the dataset by life expectancy. ■

In general, if there is a correlation between the **position (rank) of the unit** in the sampling frame and the **value of the variable of interest**, the sampling variance of the SYS estimator will be **lower** than that of the SRS estimator, because the sample is **more likely** to be representative of the population.

If there is no such correlation, the SYS sample is essentially an SRS sample, and the sampling variances are comparable – a SYS is as likely to be **representative** of the population as an SRS.

Finally, if the step $M$ is aligned with the periodicity of the values of the variable of interest, it is the opposite: the sampling variance of a SYS is larger than that of an SRS – a SYS is then **less representative** of the population than an SRS.

Some examples illustrating these situations are shown in Figure 10.12.

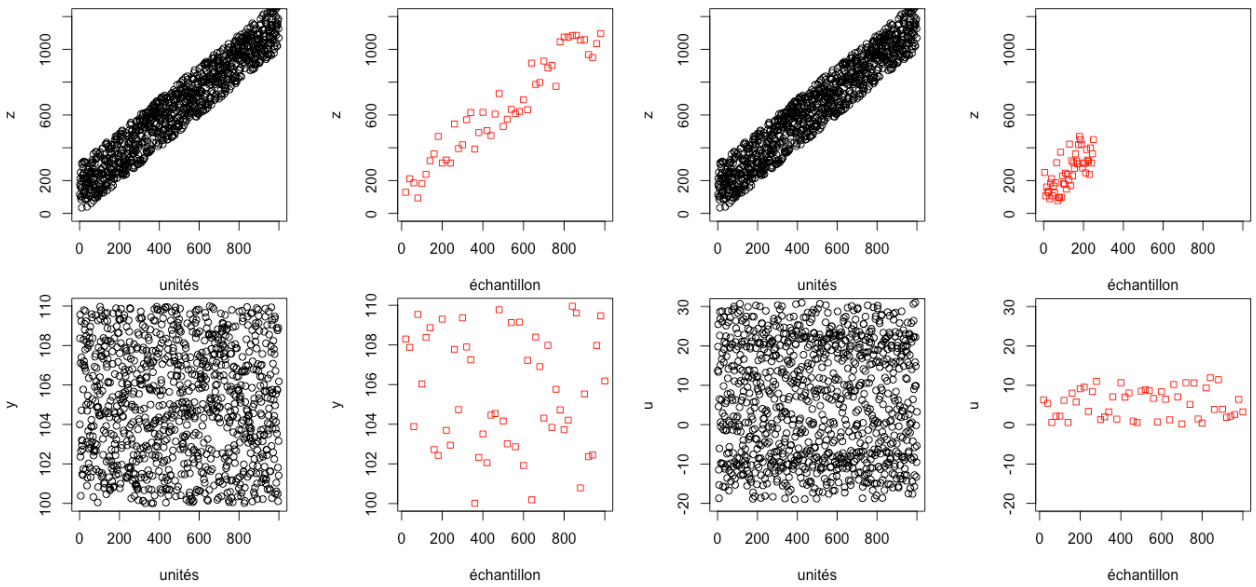



**Figure 10.12:** Various populations and systematic samplings: the order in which the population observations are presented may affect the representativity of the SYS sample.

### SYS as SRS

If the order in which the units are listed in the sampling frame is **random**,[55] we can simply consider that the sample

$$\mathcal{Y}_{\text{SYS}} = \underbrace{\{y_1, y_2, y_3, \ldots, y_{n-1}, y_n\}}_{\{u_\gamma, u_{\gamma+M}, \ldots, u_{\gamma+(n-1)M}\}} \subseteq \mathcal{U}$$

of size $n \approx \frac{N}{M}$ is in fact a SRS of size $n$. In that case, the theory developed in Section 10.3 for SRS remains valid.

55: **Careful!** this is not always easy to demonstrate.

**Estimating the Mean $\mu$** The empirical mean

$$\overline{y}_{\text{SYS}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is an **unbiased** estimator of the true population mean $\mu$, with **bound on the error of estimation**

$$B_{\mu;\text{SYS}} \approx \hat{B}_{\mu;\text{SYS}} = 2\sqrt{\hat{\mathbb{V}}(\overline{y}_{\text{SYS}})} = 2\sqrt{\frac{s^2_{\text{SYS}}}{n}\left(1 - \frac{n}{N}\right)},$$

where

$$s^2_{\text{SYS}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y}_{\text{SYS}})^2;$$

the corresponding **95% C.I. for** $\mu$ is thus

$$\text{C.I.}_{\text{SYS}}(\mu; 0.95): \quad \overline{y}_{\text{SYS}} \pm \hat{B}_{\mu;\text{SYS}}.$$

**Estimating the Total $\tau$** The quantity

$$\hat{\tau}_{\text{SYS}} = N\overline{y}_{\text{SYS}} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

is an **unbiased** estimator of the true population total $\tau$, with **bound on the error of estimation**

$$B_{\tau;\text{SYS}} \approx \hat{B}_{\tau;\text{SYS}} = 2N\sqrt{\hat{\text{V}}(\overline{y}_{\text{SYS}})} = 2N\sqrt{\frac{s^2_{\text{SYS}}}{n}\Big(1 - \frac{n}{N}\Big)};$$

the corresponding **95% C.I. for** $\tau$ is thus

$$\text{C.I.}_{\text{SYS}}(\tau; 0.95): \quad \hat{\tau}_{\text{SYS}} \pm \hat{B}_{\tau;\text{SYS}}.$$

**Estimating the Proportion $p$** If $y_i \in \{0, 1\}$ denotes the absence or presence of a certain characteristic, the quantity

$$\hat{p}_{\text{SYS}} = \overline{y}_{\text{SYS}}$$

is an **unbiased** estimator of the true proportion $p$ of units with the characteristic, with **bound on the error of estimation**

$$B_{p;\text{SYS}} \approx \hat{B}_{p;\text{SYS}} = 2\sqrt{\hat{\text{V}}(\hat{p}_{\text{SYS}})} = 2\sqrt{\frac{\hat{p}_{\text{SYS}}(1 - \hat{p}_{\text{SYS}})}{n-1}\Big(1 - \frac{n}{N}\Big)};$$

the corresponding **95% C.I. for** $p$ is thus

$$\text{C.I.}_{\text{SYS}}(p; 0.95): \quad \hat{p}_{\text{SYS}} \pm \hat{B}_{p;\text{SYS}}.$$

**SYS as CLS**

In practice, SYS is equivalent to a CLS of size $m = 1$, where each cluster is one of the $1-\text{in}-M$ SYS samples.

The quantity

$$\overline{y}_C = \frac{\sum_{k=1}^{m}\sum_{j=1}^{N_{i_k}} y_{i_k,j}}{\sum_{k=1}^{m} N_{i_k}} = \frac{\sum_{k=1}^{m} y_{i_k}}{\sum_{k=1}^{m} N_{i_k}},$$

where we use the CLS notation, is thus a **biased** estimator of the **population mean**, $\mu$.

The **average cluster size** is denoted by $\overline{N} = \frac{N}{M}$; its **sampling variance** is

$$\text{V}(\overline{y}_C) \approx \frac{1}{\overline{N}^2}\cdot\frac{1}{m}\Big(\frac{M-m}{M-1}\Big)\cdot\frac{1}{M}\sum_{\ell=1}^{M}\underbrace{(\tau_\ell - \mu N_\ell)^2}_{=N_\ell(\mu_\ell-\mu)} := \frac{1}{\overline{N}^2}\cdot\frac{\sigma^2_C}{m}\Big(\frac{M-m}{M-1}\Big),$$

and the corresponding **95% C.I. for** $\mu$ is thus

$$\text{C.I.}_{\text{G}}(\mu; 0.95): \quad \overline{y}_C \pm 2\sqrt{\text{V}(\overline{y}_C)}.$$

If the average cluster size $\overline{N}$ is unknown, we simply substitute it by

$$\overline{n} = \frac{1}{n}\sum_{k=1}^{m} N_{i_k}.$$

The estimator of the **total population** $\tau$ is thus either:

- $N\overline{y}_C$, when the number of units $N$ in the population is known, or
- $M\overline{y}_T$, where $\overline{y}_T$ is the **(empirical) mean of the sampled cluster totals**, when only $M$ is known.

Consequently, the sampling variances are

$$\mathrm{V}(N\overline{y}_C) \approx M^2 \cdot \frac{\sigma_C^2}{m}\left(\frac{M-m}{M-1}\right) \quad \text{and} \quad \mathrm{V}(M\overline{y}_T) \approx M^2 \cdot \frac{\sigma_T^2}{m}\left(\frac{M-m}{M-1}\right),$$

where $\sigma_C^2$ and $\sigma_T^2$ are computed as for a CLS. We can then construct the **95% C.I. for** $\tau$ in the usual manner.

Pretty simple, eh?


The sample contains exactly $m = 1$ cluster, so $\overline{n} = n$. The problem doesn't end there – since we don't know $\sigma_C^2$ or $\sigma_T^2$ in general, we would use the empirical variances

$$\hat{\mathrm{V}}(\overline{y}_C) \approx \frac{1}{\overline{N}^2} \cdot \frac{1}{m}\left(1 - \frac{m}{M}\right) \cdot \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_C N_{i_k})^2$$

$$\hat{\mathrm{V}}(M\overline{y}_T) \approx M^2 \cdot \frac{1}{m}\left(1 - \frac{m}{M}\right) \cdot \frac{1}{m-1}\sum_{k=1}^{m}(y_{i_k} - \overline{y}_T)^2.$$

But if $m = 1$, these variances do not exist. How do we get out of this mess? If we cannot treat the SYS as if it were a SRS (for whatever reason), the solution is to **draw additional SYS samples (replicates) and treat it as a CLS**, modifying the value of $M$ as necessary.

### 10.7.2 Sampling with Probability Proportional to Size

In practice, the **size** (whether or not this is a physical characteristic) of the sample units is often quite **variable** – a SRS is not always effective since it does not take into account the **importance that larger population units** may have.

**Additional information on the unit size** can sometimes be used to select a sample that provides a more accurate estimator of the parameters of interest.

One possible way to do this is to assign (potentially) **equal** selection probabilities to different units, based on their size.

**Example** To a certain extent ($\rho = 0.46$), the larger the area of a country, the larger its population. If we are trying to estimate the population of the planet, it might be desirable to adopt a sampling scheme in which the probability of selecting a country is **proportional to its area** – in an SRS, it is very likely that neither **China** nor **India** will be selected, resulting in an underestimate of the total sought. ■

If the variable of interest is (more or less) related to the size of the unit, one can assign a **probability of selection proportional to the size** of the unit (PPS). Note that in a PPS, previously selected units are **replaced** in the population, allowing for the **multiple selection of a single unit**.

#### Selecting a PPS With Replacement

We consider two selection methods for a PPS sample:

- **cumulative totals**, and
- the **Lahiri method**.

In both cases, the PPS sample selection procedure consists of associating with each unit a **range of numbers**,[56] related to the **size of the unit**, and taking the units that correspond to numbers chosen **at random** from the set of numbers associated with the **entire** population of $N$ units.

56: These are often **integers**, but that is not necessary.

In the **method of cumulative totals**, the **size** of the $i-$th unit is denoted by $x_i$, $1 \leq i \leq N$. We associate a **range** to each unit as follows:

| Unit | Range | | |
|---|---|---|---|
| 1 | 1 | to | $x_1$ |
| 2 | $x_1 + 1$ | to | $x_1 + x_2$ |
| 3 | $x_1 + x_2 + 1$ | to | $x_1 + x_2 + x_3$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $N-1$ | $x_1 + \cdots + x_{N-2} + 1$ | to | $x_1 + \cdots + x_{N-2} + x_{N-1}$ |
| $N$ | $x_1 + \cdots + x_{N-1} + 1$ | to | $x_1 + \cdots + x_{N-1} + x_N$ |

Finally, we draw a PPS sample by choosing $n$ integers **at random** between 1 and $X = x_1 + \cdots + x_{N-1} + x_N$ (**with replacement**) and by selecting the units **associated with these integers**.

**Example** In a village, there are 8 orchards, each containing a certain number of apple trees. A sample of $n = 3$ orchards is drawn (with replacement), in proportion to the number of apple trees per orchard.

| ID $i$ | Size $x_i$ | Cumulative Totals | Associated Range |
|---|---|---|---|
| 1 | 50 | 50 | 1 – 50 |
| 2 | 30 | 80 | 51 – 80 |
| 3 | 25 | 105 | 81 – 105 |
| 4 | 40 | 145 | 106 – 145 |
| 5 | 26 | 171 | 146 – 171 |
| 6 | 44 | 215 | 172 – 215 |
| 7 | 20 | 235 | 216 – 235 |
| 8 | 35 | 270 | 236 – 270 |

We choose $n = 3$ integers at random between 1 and 270: 108, 140, and 201, say. The associated units are the 4th, the 4th, and the 6th. ■

In the **Lahiri method**, we still denote the size of a unit by $x_i$, $1 \le i \le N$, but without having **to calculate and report the successive cumulative totals**, which can be tedious to accomplish, even with a computer.

The method consists in selecting a pair of integers $(i, j)$, where $1 \le i \le N$ and $1 \le j \le M = \max\{x_i \mid 1 \le i \le N\}$. If $j \le x_i$, the $i$th unit is added to the sample. Otherwise, the pair $(i, j)$ is rejected.

We continue in this manner until $n$ units have been selected.[57]

57: There are other ways to do this, of course; the important thing is to have a **mechanism for selecting a PPS sample**. We generally prefer sampling without replacement to sampling with replacement, but the latter is a reasonable substitute to the former if $\frac{n}{N}$ is " **sufficiently small**".

### Estimation

Let us revisit the orchard example, where $u_i$ is the yield of all apple trees in the $i$th orchard.

| ID $i$ | # Trees $x_i$ | $\pi_i$ | Yield |
|---|---|---|---|
| 1 | 50 | 50/270 | $u_1 = 2250$ |
| 2 | 30 | 30/270 | $u_2 = 1080$ |
| 3 | 25 | 25/270 | $u_3 = 1300$ |
| 4 | 40 | 40/270 | $u_4 = 1400$ |
| 5 | 26 | 26/270 | $u_5 = 1196$ |
| 6 | 44 | 44/270 | $u_6 = 1716$ |
| 7 | 20 | 20/270 | $u_7 = 820$ |
| 8 | 35 | 35/270 | $u_8 = 1680$ |

We are interested in the **total** apple production of the village, which we know in this case to be $\tau = 11,442$. Since **in principle** an orchard with more apple trees should produce more apples, we draw a PPS sample of $n = 3$ units (with replacement), where the number of apple trees in the orchard is used as the unit size.

In what follows, we illustrate the concepts using the sample

$$y_1 = u_4 = 1400, y_2 = u_4 = 1400, y_3 = u_6 = 1716.$$

If the sample $\mathcal{Y}$, with $|\mathcal{Y}| = n$, is drawn from $\mathcal{U}$ using a PPS, the units $y_1, \ldots, y_n$ are **independent** and distributed according to

| $y_i$ | $u_1$ | $\cdots$ | $u_j$ | $\cdots$ | $u_N$ |
|---|---|---|---|---|---|
| $p(y_i)$ | $\pi_1$ | $\cdots$ | $\pi_j$ | $\cdots$ | $\pi_N$ |

where $0 < \pi_j < 1$ for all $1 \le j \le N$ and $\pi_1 + \cdots + \pi_N = 1$.

For all $1 \le i \le n$, there is a $1 \le j \le N$ such that $y_i = u_j$. Set $w_i = \frac{u_j}{\pi_j}$. The **sampling weights** $w_i$ are also **independent** and distributed according to

$$P(y_i = u_j) = P\Big(w_i = \frac{u_j}{\pi_j}\Big) = \pi_j, \quad 1 \le i \le n,\ 1 \le j \le N.$$

We note that for any $1 \le i \le n$, the **expected weight** is

$$\mathrm{E}(w_i) = \sum_{j=1}^{N} w_j P(w_i = w_j) = \sum_{j=1}^{N} \frac{u_j}{\pi_j} \cdot \pi_j = \sum_{j=1}^{N} u_j = \tau.$$

In other words,

$$\hat{\tau}_{\mathrm{pps}} = \overline{w} = \frac{1}{n} \sum_{i=1}^{n} w_i$$

is an **unbiased estimator of the total** $\tau$. Its **sampling variance** is computed as follows:

$$\mathrm{V}(\hat{\tau}_{\mathrm{pps}}) = \mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) = \underbrace{\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{V}(w_i)}_{\text{ind. des } w_i} = \frac{1}{n^2}\sum_{i=1}^{n}\Big[\sum_{j=1}^{N}(w_j - \tau)^2 P(w_i = w_j)\Big]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{N}\Big(\frac{u_j}{\pi_j} - \tau\Big)^2\pi_j = \frac{1}{n}\sum_{j=1}^{N}\Big(\frac{u_j}{\pi_j} - \tau\Big)^2\pi_j = \frac{1}{n}\sum_{j=1}^{N}\Big(\frac{u_j^2}{\pi^j} - \frac{2\tau u_j}{\pi_j} + \tau^2\Big)\pi_j$$

$$= \frac{1}{n}\Big(\sum_{j=1}^{N}\frac{u_j^2}{\pi_j} - 2\tau\underbrace{\sum_{j=1}^{N} u_j}_{=\tau} + \tau^2\underbrace{\sum_{j=1}^{N}\pi_j}_{=1}\Big) = \frac{1}{n}\Big(\sum_{j=1}^{N}\frac{u_j^2}{\pi_j} - \tau^2\Big).$$

In practice, we do not typically know the true value of $\tau$, so we use the **unbiased estimator**

$$\hat{\mathrm{V}}(\hat{\tau}_{\mathrm{pps}}) = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} w_i^2 - n\hat{\tau}_{\mathrm{pps}}^2\right).$$

**Central Limit Theorem – PPS:** if $n$ and $N - n$ are sufficiently large, then

$$\hat{\tau}_{\mathrm{pps}} \sim_{\mathrm{approx.}} \mathcal{N}\left(\tau, \hat{\mathrm{V}}(\hat{\tau}_{\mathrm{pps}})\right).$$

The **bound on the error of estimation** and the **95% C.I. for** $\tau$ are therefore

$$\hat{B}_{\tau;\mathrm{pps}} = 2\sqrt{\hat{\mathrm{V}}(\hat{\tau}_{\mathrm{pps}})} \quad \text{and} \quad \mathrm{C.I.}_{\mathrm{pps}}(\tau; 0.95) = \hat{\tau}_{\mathrm{pps}} \pm \hat{B}_{\tau;\mathrm{pps}}.$$

**Example** In the orchard dataset, we have

$$\hat{\tau}_{\text{pps}} = \frac{1}{3}\Big[\underbrace{\frac{1400}{40/270}}_{w_1} + \underbrace{\frac{1400}{40/270}}_{w_2} + \underbrace{\frac{1716}{44/270}}_{w_3}\Big] = 9810;$$

$$\hat{\mathrm{V}}(\hat{\tau}_{\text{pps}}) = \frac{1}{3(2)}\Big[\underbrace{\Big(\frac{1400}{40/270}\Big)^2}_{w_1} + \underbrace{\Big(\frac{1400}{40/270}\Big)^2}_{w_2} + \underbrace{\Big(\frac{1716}{44/270}\Big)^2}_{w_3} - 3\cdot\underbrace{9810^2}_{\hat{\tau}^2_{\text{pps}}}\Big]$$
$$= 129,600.$$

Consequently, the 95% C.I. for the total apple yield in the village is

$$\text{C.I.}_{\text{pps}}(\tau; 0.95) = 9810 \pm 2\sqrt{129,600} \equiv (9090, 10530).$$

The actual total yield ($\tau = 11,442$) **does not fall** within the confidence interval – why might this be the case? Is this problematic? ■

In general, $\mathrm{V}(\hat{\tau}_{\text{pps}}) \leq \mathrm{V}(\hat{\tau}_{\text{SRS}})$. In the orchards example, we can show that

$$\mathrm{V}(\hat{\tau}_{\text{SRS}}) \approx 8^2 \cdot \frac{172981.4375}{3}\Big(\frac{8-3}{8-1}\Big) = 2,635,907.619, \quad \text{and}$$
$$\mathrm{V}(\hat{\tau}_{\text{pps}}) \approx \frac{1}{3}\Big[\frac{2250^2}{50/270} + \cdots + \frac{1680^2}{35/270} - 11,442^2\Big] = 723,912.$$

We can also give an estimate of the population average $\mu$ using

$$\hat{\mu}_{\text{pps}} = \frac{\hat{\tau}_{\text{pps}}}{N}, \quad \hat{\mathrm{V}}(\hat{\mu}_{\text{pps}}) = \frac{\hat{\mathrm{V}}(\hat{\tau}_{\text{pps}})}{N^2}, \quad \text{C.I.}_{\text{pps}}(\mu; 0.95) = \frac{\text{C.I.}_{\text{pps}}(\tau; 0.95)}{N}.$$

A lot more can be said on the topic; PPS usually provides a springboard to more sophisticated sampling designs and other theoretical considerations [5, 7, 6].

### 10.7.3 Multi-Stage Sampling

By splitting the sampling process into several stages, one can **reduce costs** and **focus the logistical aspects of sampling on a few focal points**. In **multi-stage sampling** (M*n*S), a sample of large units (**primary units**) is drawn, then sub-units (**secondary units**) are drawn from the large units, and so on.

**Example** Sampling units in a Canadian province could be decomposed into three steps:

1. conduct a sample of municipalities (**primary units**);
2. sample neighbourhoods in the sampled municipalities (**secondary units**), and
3. sample households in the samples neighbourhoods (**tertiary units**).

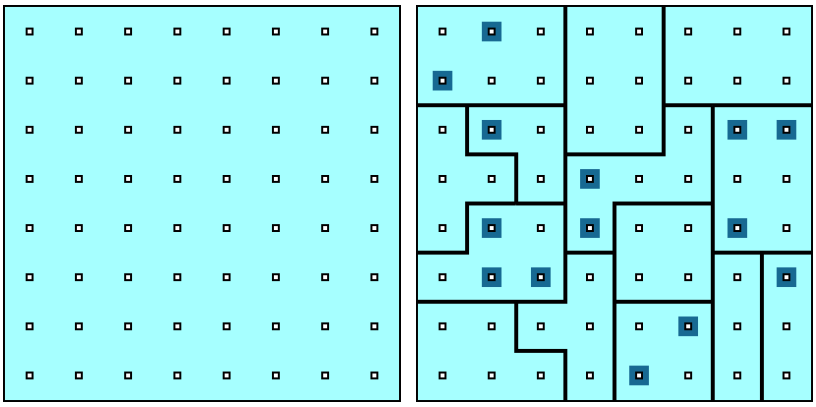


**Figure 10.13:** Schematics of SRS2S: target population (left) and sample (right).

In a M$n$S, the sample is concentrated around several **pivots**: in field studies, for example, this has the advantage of considerably reducing the survey area, which helps to **reduce non-sampling errors**.[58]

58: In addition to reducing operational costs.

Furthermore, detailed information is often available for **groups** of sample units, but not for **individual** units: it is therefore not necessary to obtain a **complete** sampling frame for **all** sample units, but only for those belonging to the primary units selected in the first round, for example.

Any probability sampling method can be used at **each stage**, and they **can change from stage to stage**: e.g., a municipality SRS, a neighborhood SRS, a household SRS, etc.

### Two-Stage Simple Random Sampling

In a 2-stage process, if sampling is conducted using a SRS for both stages, the method is known as **two-stage simple random sampling** (SRS2S).

**Example** The biomass of a plant species in a forest area can be estimated by drawing a SRS of $m = 8$ compartments (primary units) from the $M = 40$ compartments composing the population under study.

For each of these compartments $1 \leq i \leq m$, we then draw a SRS of $n_i$ plots, and measure the biomass in the plot. Estimates of the average or total amount of biomass in the forest area can be calculated using appropriate formulas. ■

### Estimation

Let be a population consisting of $M$ primary units, having $N_\ell$ secondary units in the $\ell$th primary unit. Denote by $u_{i,j}$ the value of the response variable of the $j$th secondary unit in the $i$th primary unit.

The **population mean** is

$$\mu = \frac{\sum_{\ell=1}^{M} \sum_{j=1}^{N_\ell} u_{\ell,j}}{\sum_{\ell=1}^{M} N_\ell}.$$

Suppose we draw a SRS of $m$ primary units, and a SRS of $n_i$ secondary units in the $i$th primary unit. The total sample size is thus $n = n_1 + cdots + n_m$. We obtain an unbiased estimator of $\mu$ from:

$$\overline{y}_{\text{SRS2S}} = \frac{1}{m\overline{N}} \sum_{i=1}^{m} N_i \overline{y}_i = \frac{1}{m\overline{N}} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{k=1}^{n_i} y_{i,k} = \frac{1}{m\overline{N}} \sum_{i=1}^{m} \sum_{k=1}^{n_i} \frac{MN_i}{mn_i} y_{i,k},$$

where

$$\overline{N} = \frac{1}{M} \sum_{\ell=1}^{M} N_\ell \approx \frac{N_1 + \cdots + N_m}{m}.$$

The sampling variance is composed of two components:

- a measure of the variation **between the primary units**, and
- a measure of the variation **within the primary units**.

When $n_i = N_i$ for all $1 \leq i \leq m$, we are dealing with a **CLS** and the variance is only given by the first component (see Section 10.6). In the case where $m = M$, we are dealing with a **STS** and the variance is only given by the second component (see Section 10.4).

When $m \neq M$ and $n_i \neq N_i$ for at least one primary unit $i$, the variance is a combination of these two extremes: in that case, the second component represents **the contribution of sub-sampling** (another name for M$n$S). We use the **law of total variance** to estimate the sampling variance:

$$\begin{aligned}
\mathrm{V}(\overline{y}_{\text{SRS2S}}) &= \mathrm{E}[\mathrm{V}(\overline{y}_{\text{SRS2S}} \mid m)] + \mathrm{V}(\mathrm{E}[\overline{y}_{\text{SRS2S}} \mid m]) \\
&= \frac{1}{\overline{N}^2} \cdot \frac{\sigma_T^2}{m} \left( \frac{M-m}{M-1} \right) + \frac{1}{mM\overline{N}^2} \sum_{i=1}^{m} N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left( \frac{N_i - n_i}{N_i - 1} \right) \\
&\approx \underbrace{\frac{1}{\overline{N}^2} \cdot \frac{s_T^2}{m} \left( 1 - \frac{m}{M} \right)}_{\text{between primary units}} + \underbrace{\frac{1}{mM\overline{N}^2} \sum_{i=1}^{m} N_i^2 \cdot \frac{s_i^2}{n_i} \left( 1 - \frac{n_i}{N_i} \right)}_{\text{within primary units}},
\end{aligned}$$

where

$$s_T^2 = \frac{1}{m-1} \sum_{i=1}^{n} \left( N_i \overline{y}_i - \overline{N} \overline{y}_{\text{SRS2S}} \right)^2, \quad s_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (y_{i,k} - \overline{y}_i)^2.$$

**Example** The biomass of a plant species (kg) is measured in plots of 0.025 ha (secondary units) selected from $m = 8$ compartments (primary units), randomly selected themselves among the $M = 40$ compartments of a forested area. The summary of results is shown in the following table:

| **Comp.** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\overline{y}_i$ | 118 | 107 | 109 | 110 | 120 | 95 | 93 | 90 |
| $s_i^2$ | 436 | 516 | 586 | 456 | 412 | 497 | 755 | 496 |
| $N_i$ | 1760 | 1975 | 1615 | 1785 | 1775 | 2050 | 1680 | 1865 |
| $n_i$ | 9 | 10 | 8 | 9 | 9 | 10 | 8 | 9 |

Find a 95% C.I. for the average biomass per plot and per compartment, and for its total in the forested area.

**Solution:** Since we do not know $\overline{N}$, we approximate it with the mean

$$\overline{N} \approx \frac{1}{8}(1760 + \cdots + 1865) = 1813.125.$$

The totals in the selected primary units are then:

| **Comp.** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $N_i\overline{y}_i(\times 10^5)$ | 2.077 | 2.113 | 1.760 | 1.964 | 2.130 | 1.946 | 1.562 | 1.679 |

The SRS2S estimators of the mean $\mu$, of the mean of the totals in the compartments, and of the total are:

$$\overline{y}_{\text{SRS2S}} = \frac{1}{8(1813.125)}(2.077 + \cdots + 1.679) \times 10^5 = 105.01;$$
$$\overline{N}\overline{y}_{\text{SRS2S}} = 1813.125 \cdot 105.01 = 190,403.75; \quad \tau_{\text{SRS2S}} = M \cdot \overline{N}\overline{y}_{\text{SRS2S}} = 7,616,150.$$

The variance between compartments (primary units) is thus:

$$s_T^2 = \frac{1}{8-1}\sum_{i=1}^{8}(N_i\overline{y}_i - 190,403.75)^2 = 4.55 \times 10^8$$

Finally, we calculate the variance within the compartments:

| **Comp.** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\frac{N_i^2}{\overline{N}^2} \cdot \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right)$ | 48.2 | 51.3 | 72.7 | 50.4 | 45.6 | 49.4 | 93.9 | 54.9 |

The sampling variance is thus

$$\begin{aligned}\hat{\text{V}}(\overline{y}_{\text{SRS2S}}) &= \frac{4.55 \times 10^8}{8(1813.125)^2}\left(1 - \frac{8}{40}\right) + \frac{1}{8(40)}(48.2 + \cdots + 54.9) \\ &= 14.03\end{aligned}$$

The variances of the other two estimators are easily calculated:

$$\hat{\text{V}}(\overline{N}\overline{y}_{\text{SRS2S}}) = \overline{N}^2\hat{\text{V}}(\overline{y}_{\text{SRS2S}}) = (1813.125)^2 \cdot 14.03 = 46,141,324.55;$$
$$\hat{\text{V}}(\tau_{\text{SRS2S}}) = M^2\overline{N}^2\hat{\text{V}}(\overline{y}_{\text{SRS2S}}) = (40)^2 \cdot (1813.125)^2 \cdot 14.03 = 73,826,119,284;$$

the confidence intervals are thus

$$\text{C.I.}_{\text{SRS2S}}(\mu; 0.95): \quad 105.01 \pm 2\sqrt{14.03} \equiv (97.5, 112.5)$$
$$\text{C.I.}_{\text{SRS2S}}(\tfrac{N_0}{M}\mu; 0.95): \quad 190,403.75 \pm 2\sqrt{46,141,324.55} \equiv (176818, 203989.2312),$$
$$\text{C.I.}_{\text{SRS2S}}(\tau; 0.95): \quad 7,616,150 \pm 2\sqrt{73,826,119,284} \equiv (7072730, 8159569)$$

assuming of course that the central limit theorem remains valid in the context of a SRS2S. ■

### 10.7.4 Multi-Phase Sampling

**Multi-stage sampling** (M$n$P) plays a crucial role in many types of surveys, including those conducted by **remote sensing**.

In the first phase, a **selected** number of units are sampled, but only a **small** number of characteristics are captured for each unit. In each successive phase, a larger **number** of features is measured on a smaller **sub-sample** of units.

In this way, the target parameter can be estimated with **more accuracy** and at **lower cost**, by studying the relationship between the features measured in the different sampling phases.

#### Two-Phase Random Sampling

A M$n$P with only two phases is called a **two-phase sampling** (M2P). M2Ps are particularly useful in a situation where enumeration of the **main trait** is expensive (in terms of costs or labor), but in which an **auxiliary trait** correlated to the main trait can easily be observed.

Thus, it is sometimes preferable to draw a **large** SRS in the **first phase** in order to analyze the auxiliary variables, which leads to more accurate estimates of $\tau$ or $\mu$ for that auxiliary variable (at least, that is the hope). In the second phase, a **smaller** sample is drawn, usually a **sub-sample** of the characteristic, and the **auxiliary variable** are measured.

Estimates of the main characteristic are then obtained using the information obtained in the **first phase**, using the **ratio method** or the **regression method**, for instance. The precision of the final estimates can be increased by including **several correlated auxiliary variables**.

**Example** If we want to estimate the total volume of wood $\tau$ in a forest, we could first measure the circumference $c_i$ and height $h_i$ of the trees $i$ in some sample, then the volume $v_{i_k}$ of the trees $i_k$ in a sub-sample. We only need to determine the statistical relationship between $\tau_v$, $\tau_c$, and $\tau_h$ to complete the procedure. ■

The M$n$P sampling method helps to reduce the **cost of enumeration** and increase the **accuracy of estimates**. It can also be used to **stratify** a population: an initial sample is taken based on the auxiliary characteristic, which is used to subdivide the population into strata in which the main characteristic is more or less **homogeneous**.

As long as the two characteristics are **correlated**, accurate estimates of the main characteristic are obtained from a second, relatively small sample.

M2P can also be paired with M2S, for example (or with any other sampling design). If both selection steps are performed with SRS, the method is called **two-phase simple random sampling** (SRS2P).

In the first phase, the population is divided into well-defined sampling units; a SRS $\mathcal{Y}_1$ of size $n_1$ is drawn from these units; the **auxiliary variable** $x$ is measured on all units of $\mathcal{Y}_1$. Next, a sub-SRS $\mathcal{Y}_2$ of size $n_2$ is drawn from $\mathcal{Y}_1$; the **main characteristic** $y$ is measured on all units of $\mathcal{Y}_2$.

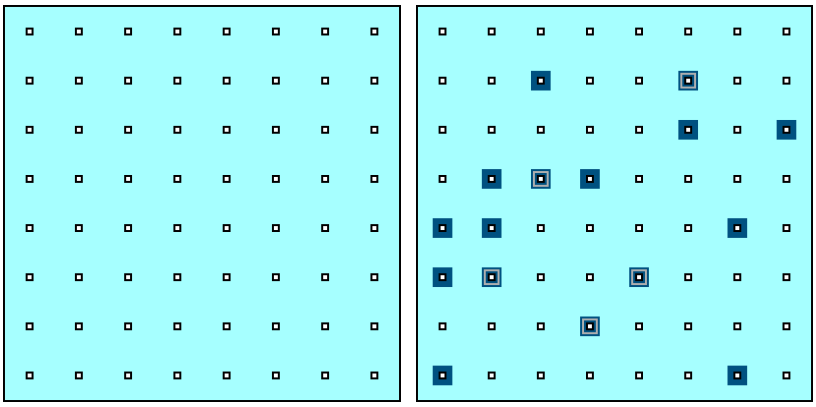


**Figure 10.14:** Schematics of SRS2P: target population (left) and sample (right).

We can evaluate $r_{\mathcal{Y}_2}$ or $b_{\mathcal{Y}_2}$ from the observations in $\mathcal{Y}_2$ (using either the ratio method or the regression method), which yields

$$\begin{aligned}\hat{\mu}_{Y;R;\text{SRS2P}} &= r_{\mathcal{Y}_2} \cdot \overline{x}_{\mathcal{Y}_1} \quad \text{or} \\ \hat{\mu}_{Y;L;\text{SRS2P}} &= \overline{y}_{\mathcal{Y}_2} + b_{\mathcal{Y}_2}(\overline{x}_{\mathcal{Y}_1} - \overline{x}_{\mathcal{Y}_2}).\end{aligned}$$

**Estimation**

Due to the **double sampling**, two terms contribute to sampling variances of the estimators (the first when going from $\mathcal{U}$ to $\mathcal{Y}_1$, and the second from $\mathcal{Y}_1$ to $\mathcal{Y}_2$):

$$\begin{aligned}\hat{\mathrm{V}}(\hat{\mu}_{Y;R;\text{SRS2P}}) &= \frac{1}{n_2}\left(s_Y^2 - 2r_{\mathcal{Y}_2}s_{XY} + (r_{\mathcal{Y}_2})^2 s_X^2\right) + \frac{1}{n_1}\left(2r_{\mathcal{Y}_2}s_{XY} - (r_{\mathcal{Y}_2})^2 s_X^2\right) \\ \hat{\mathrm{V}}(\hat{\mu}_{Y;L;\text{SRS2P}}) &= \frac{1}{n_2}s_{XY;L}^2 + \frac{1}{n_1}\left(s_{XY;L}^2 - s_Y^2\right)\end{aligned}$$

where $s_Y^2$, $s_{XY}$, and $s_X^2$ are the usual quantities (in $\mathcal{Y}_2$), and

$$\begin{aligned}r_{\mathcal{Y}_2} = \frac{\overline{y}_{\mathcal{Y}_2}}{\overline{x}_{\mathcal{Y}_2}},\ b_{\mathcal{Y}_2} = \frac{s_{XY}}{s_X^2}, \quad \text{and} \\ s_{XY;L}^2 = \frac{n_2 - 1}{n_2 - 2} \cdot \left\{s_Y^2 - b_{\mathcal{Y}_2}^2 s_X^2\right\}\end{aligned}.$$

**Example** We are interested in the biomass of any plant in a region, which is divided into plots of 0.025 ha each. First, we measure the number $x$ of groves per unit in a SRS $\mathcal{Y}_1$ of $n_1 = 200$ plots.

Then, the biomass $y$ of the plant in question is calculated in each unit of a sub-SRS $\mathcal{Y}_2$ of $n_2 = 40$ plots:

$$\begin{aligned}&\overline{x}_{\mathcal{Y}_1} = 374.4; \quad \sum_{i=1}^{40} x_i = 15,419; \quad \sum_{i=1}^{40} y_i = 2104; \\ &\sum_{i=1}^{40} x_i^2 = 7,744,481; \quad \sum_{i=1}^{40} x_i y_i = 960,320; \quad \sum_{i=1}^{40} y_i^2 = 125,346.\end{aligned}$$

What would a 95% C.I. for the average biomass per plot look like?

Let us compute the required intermediate quantities:

$$\overline{x}_{\mathcal{Y}_2} = \frac{15419}{40} = 385.5; \quad \overline{y}_{\mathcal{Y}_2} = \frac{2104}{40} = 52.6; \quad r_{\mathcal{Y}_2} = \frac{\overline{y}_{\mathcal{Y}_2}}{\overline{x}_{\mathcal{Y}_2}} = \frac{52.6}{385.5} = 0.14;$$

$$s_X^2 = \frac{1}{39}[7744481 - 40(385.5)^2] \approx 46175; \; s_Y^2 = \frac{1}{39}[125346 - 40(52.6)^2] \approx 376$$

$$s_{XY} = \frac{1}{39}[960320 - 40(385.5)(52.6)] \approx 3827.7; \quad b_{\mathcal{Y}_2} = \frac{s_{XY}}{s_X^2} = \frac{3827.7}{46175.4} \approx 0.08;$$

$$s_{XY;L}^2 = \frac{39}{38}[376.3 - 0.08^2(46175.4)] \approx 82.9;$$

which gives us

$$\hat{\mu}_{Y;R;\text{SRS2P}} = 0.14(374.4) \approx 51.1; \quad \hat{\mu}_{Y;L;\text{SRS2P}} = 52.6 + 0.08(374.4 - 385.5) \approx 51.7$$

and

$$\hat{\text{V}}(\hat{\mu}_{Y;R;\text{SRS2P}}) = \frac{376.3 - 2(0.14)(3827.7) + (0.14)^2 46175.4}{40} + \frac{2(0.14)3827.7 - (0.14)^2 46175.4}{200} \approx 5.67;$$

$$\hat{\text{V}}(\hat{\mu}_{Y;L;\text{SRS2P}}) = \frac{82.9}{40} + \frac{82.9 - 376.3}{200} \approx 3.54;$$

from which we conclude that

$$\text{C.I.}_{R;\text{SRS2P}}(\mu_Y; 0.95) = 51.1 \pm 2\sqrt{5.67} \equiv (46.3, 55.8)$$
$$\text{C.I.}_{L;\text{SRS2P}}(\mu_Y; 0.95) = 51.7 \pm 2\sqrt{3.54} \equiv (47.9, 55.5). \quad \blacksquare$$

### 10.7.5 Miscellaneous

We end the module by briefly discussing a few notions that did not find a natural slot in the previous sections:

- design effects;
- adjusting for non-response;
- estimating the size of a population,
- randomized responses, and
- Bernoulli sampling.

#### Design Effect

The **design effect** compares the estimator for a given sampling design and for a SRS. It is the ratio of the **sampling variance of the estimator under the given sampling design** to the **sampling variance of the estimator under a SRS** (assuming samples of the **same size**).

This value is often applied to compare the **efficiency** of estimators from different sampling designs. If the ratio $< 1$, the sampling design is more efficient than SRS; if it is $> 1$, it is less efficient than SRS.

We directly compared the theoretical variances of several sampling designs in sections 10.4.3, 10.5.4, and 10.6.3, but in practice we compute the design effect using the achieved samples (assuming that they had been drawn under various sampling plans).

Design effects also help to obtain approximate variance estimates for complex **sampling designs**. If a design effect estimate is available from a previous survey (that used the sampling design we will be using for this survey), it can be used to determine the **sample size** required to meet some pre-determined condition(s).

### Adjusting for Nonresponse

Non-response is a problem in **all** surveys. **Total non-response** (when all or almost all data from a sampled unit are missing) occurs when:

- a sample unit **refuses to participate** in the survey;
- we cannot **establish contact with a sample unit**;
- the sampled unit cannot be **found**, or
- the information obtained form the unit is **useless/invalid**.

The simplest way to deal with such non-response is to ignore it; in some **exceptional** circumstances (when the affected observations are not in any way different from those for whom we have valid and complete measurements), proportions or means that are estimated without adjusting for non-response are **more or less the same** as those produced by applying adjustment for non-response.

If one neglects to **compensate** for nonresponding units, however, the **totals are generally underestimated** (e.g., the size of a population, total revenue, or total acres harvested, say).

The most common way to deal with total non-response is to **adjust the base sampling weights** by assuming that the responding units represent both responding and nonresponding units. If the **nonrespondents are equivalent to the respondents** for the characteristics measured in the survey, this is a reasonable approach.

The base weights for nonrespondents are then redistributed among respondents, using a **adjustment factor for nonrespondents** that is multiplied by the base weight, to obtain an adjusted weight.

**Example** If we draw a SRS of size $n = 25$ from a stratum of size $N = 1000$, the **probability of inclusion** of each of these units and the corresponding **basic weight** are

$$\pi = \frac{n}{N} = \frac{25}{1000} = 0.025, \quad w = \frac{1}{\pi} = \frac{1}{0.025} = 40.$$

In other words, each selected unit represents 40 units in the stratum.

If we only get a response from $n_r = 20$ of the $n = 25$ selected units, the **non-response adjustment factor** (NRAF) and the **adjusted weight** (for non-response) become:

$$\text{NRAF} = \frac{n}{n_r} = \frac{25}{20} = 1.25$$
$$w_{\text{nr}} = w \cdot \text{NRAF} = 1.25(40) = 50;$$

each responding unit then represents 50 units in the stratum. This adjusted weighting is what we would end up working with. ■

Of course, the adjusted weight may vary from stratum to stratum, depending on the sample design and the sample size/allocation.

When we want to determine the optimal sample size/allocation across various strata, what we obtain is the **target sample size**.[59] We then have to resort to **inflation** of the sample size to achieve the target.

59: Assuming that the target and study populations coincide.

**Example** The allocation of a StS of size $n = 29$ is found to be $(17, 9, 3)$. In a prior study, the non-response rates by stratum were determined to be $(16.2\%, 20.8\%, 31.2\%)$. Which allocation optimizes the likelihood of achieving the target allocation?

We only need to solve

$$n_1(1 - 0.162) = 17, \quad n_2(1 - 0.208) = 9, \quad n_3(1 - 0.312) = 3,$$

which gives a practical sample allocation of $(n_1, n_2, n_3) = (20.3, 11.3, 4.3) \approx (21, 12, 5)$, and a practical sample size of $n = 38$.

### Estimating a Population Size

How do we proceed if the size $N$ of the population $\mathcal{U}$ is unknown? When the population is large enough, we can always use the approximation $N \approx \infty$ in the sampling variance formulas.

But sometimes it is the parameter $N$ that represents the quantity of interest; as an example, how would we find out the number $N$ of \$5 bill in circulation?

We approach such a problem using the **catch-and-release** method (compare with the approach used in Module 25):

1. we capture $n_1$ bills at random (without replacement) from the population;
2. we mark them and release them back into circulation;
3. at a later time, $n_2$ bills are captured at random (without replacement) from the population;
4. we count the number $X$ of marked bills, $0 < X \leq n_2$.

If we wait long enough (to let the marked bills propagate in the population, say), we obtain

$$\frac{n_1}{N} \approx \frac{X}{n_2}, \quad \text{from which we have } \hat{N} = \frac{n_1 n_2}{X},$$

where $X$ follows a **hypergeometric** distribution with parameters $n_1, N - n_1, n_2$, and probability mass function

$$P(X = x) = \frac{\binom{n_1}{x}\binom{N - n_1}{n_2 - x}}{\binom{N}{n_2}}, \quad 0 \leq x \leq n_2$$

$$\mu_X = \mathrm{E}[X] = n_2 \underbrace{\left(\frac{n_1}{N}\right)}_{p} = n_2 p, \quad \sigma_X^2 = \mathrm{V}[X] = n_2 p(1 - p)\left(\frac{N - n_2}{N - 1}\right).$$

If $\frac{n_2}{N} < 0.05$, we can ignore the FPCF term in the variance:

$$\sigma_X^2 = \mathrm{V}[X] \approx n_2 p(1-p).$$

We can now develop expressions for $\mathrm{E}[\hat{N}]$ and $\mathrm{V}[\hat{N}]$, using a **Taylor series of order 2 near** $X \approx \mu_X = n_2 p$:

$$f(X) \approx f(\mu_X) + f'(\mu_X)(X-\mu_X) + \frac{f''(\mu_X)}{2}(X-\mu_X)^2.$$

Si $\hat{N} = f(X) = \frac{n_1 n_2}{X}$, so that

$$\begin{aligned}\hat{N} &\approx \frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2}(X-\mu_X) + \frac{n_1 n_2}{\mu_X^3}(X-\mu_X)^2 \\ &= \frac{n_1}{p} - \frac{n_1}{n_2 p^2}(X-n_2 p) + \frac{n_1}{n_2^2 p^3}(X-n_2 p)^3.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathrm{E}[\hat{N}] &= \mathrm{E}\left[\frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2}(X-\mu_X) + \frac{n_1 n_2}{\mu_X^3}(X-\mu_X)^2\right] \\ &= \mathrm{E}\left[\frac{n_1 n_2}{\mu_X}\right] - \mathrm{E}\left[\frac{n_1 n_2}{\mu_X^2}(X-\mu_X)\right] + \mathrm{E}\left[\frac{n_1 n_2}{\mu_X^3}(X-\mu_X)^2\right] \\ &= \frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2}(\underbrace{\mathrm{E}[X]}_{\mu_X} - \mu_X) + \frac{n_1 n_2}{\mu_X^3}\mathrm{E}\left[(X-\mu_X)^2\right] \\ &= \frac{n_1 n_2}{\mu_X} + \frac{n_1 n_2}{\mu_X^3}\mathrm{V}[X] \approx \frac{n_1}{p} + \frac{n_1}{n_2^2 p^3}\cdot n_2 p(1-p) = \frac{n_1}{p} + \frac{n_1}{n_2 p^2}(1-p) \\ &= \frac{n_1}{p}\left(1 + \frac{1-p}{n_2 p}\right) = N\left(1 + \frac{1-p}{n_2 p}\right).\end{aligned}$$

Since $\frac{1-p}{n_2 p} > 0$, $\mathrm{E}[\hat{N}] \neq N$, and so $\hat{N}$ is an **asympotically unbiased estimator of** $N$ when the sample size $n_2$ increases.

We can provide an approximation of the variance using a **Taylor series of order 1 near** $X \approx \mu_X = n_2 p$:

$$\hat{N} \approx \frac{n_1 n_2}{\mu_X} - \frac{n_1 n_2}{\mu_X^2}(X-\mu_X) = \frac{n_1}{p}\left(1 - \frac{X-n_2 p}{n_2 p}\right) = \frac{n_1}{p}\left(2 - \frac{X}{n_2 p}\right).$$

Putting all this together, we get

$$\begin{aligned}\mathrm{V}[\hat{N}] &\approx \mathrm{V}\left[\frac{n_1}{p}\left(2 - \frac{X}{n_2 p}\right)\right] = \frac{n_1^2}{p^2}\cdot \mathrm{V}\left[-\frac{X}{n_2 p}\right] = \frac{n_1^2}{n_2^2 p^4}\cdot \mathrm{V}[X] \\ &\approx \frac{n_1^2 n_2 p(1-p)}{n_2^2 p^4} = \frac{n_1^2(1-p)}{n_2 p^3}.\end{aligned}$$

In practice, we do not know the true $p$, so we use

$$\hat{\mathrm{V}}[\hat{N}] = \frac{n_1^2(1-\hat{p})}{n_2 \hat{p}^3}, \quad \text{where } \hat{p} = \frac{X}{n_2}.$$

**Central Limit Theorem – Population Size** $N$**:** if $n_2$ and $N$ are sufficiently large, we have

$$\hat{N} \sim_{\text{approx.}} \mathcal{N}\left(\mathrm{E}[\hat{N}], \hat{\mathrm{V}}[\hat{N}]\right) \approx \mathcal{N}\left(\frac{n_1 n_2}{X}, \frac{n_1^2(1-\hat{p})}{n_2 \hat{p}^3}\right),$$

and the corresponding **95% C.I. for** $N$ is thus

$$\text{C.I.}(N; 0.95): \quad \frac{n_1 n_2}{X} \pm 2\sqrt{\frac{n_1^2(1-\hat{p})}{n_2 \hat{p}^3}}.$$

**Example** Say that $n_1 = 500$ bills were initially captured, marked, and releases; of the $n_2 = 300$ bills recaptured at a later date, $X = 127$ were marked. Give a 95% C.I. for the total number of \$5.

The point estimate is $\hat{N} = \frac{500 \cdot 300}{127} \approx 1181.102$. We also have $\hat{p} = \frac{X}{n_2} = \frac{127}{300} \approx 0.423$, from which we get the bound on the error of estimation

$$2\sqrt{\hat{\mathrm{V}}(\hat{N})} = 2\sqrt{\frac{500^2 \cdot (1-0.42)}{300 \cdot (0.42)^3}} = 159.176,$$

and

$$\text{C.I.}(N; 0.95): \quad 1181.102 \pm 159.176 \equiv (1021.9, 1340.3). \quad \blacksquare$$

### Randomized Response

Let's say we ask students whether they cheated on a test or an assignment during the pandemic. If the answer is "Yes," we can likely conclude that it is the true answer. But since there is a **social cost** associated with such an answer, we can expect that some cheaters will answer "No". What can we do to reduce the measurement error for **sensitive** questions?

**First approach:** with such questions, the skill of the interviewer plays a crucial role – this aspect should not be overlooked.

**Second approach:** the **randomized response** technique requires the use of two questions:

- the **sensitive question**, and
- an **innocent** question,

as well as a **random mechanism with known parameters** (heads or tails, etc.).

Randomized responses work as follows: the respondent flips a coin (without announcing the result to the interviewer), and answers honestly one of the 2 questions:

- "**head**": "Have you ever cheated on a test?";
- "**tail**": "Were you born in January?";

Since the interviewer does not know the outcome of the draw, they do not know whether the respondent is answering the sensitive question or the innocent one. **In theory**, the anonymity provided by the randomized response is freeing (the social cost is **diminished, if not eliminated altogether**) – therefore, we could expect an honest answer, regardless of the question.

**But we have to be careful:** this approach can only be successful if we know the probabilities:

- $\theta$ of **observing a positive response to the innocent question**;
- $\rho$ of **the question being answered actually being the sensitive question**, and
- $\phi$ of **observing a positive response, whatever the question**.

Let $p$ be the **proportion of positive responses to the sensitive question**, which is the quantity of interest. According to the Law of Total Probability, we have

$$\begin{aligned}\phi &= P(\text{positive response}) \\ &= \underbrace{P(\text{positive} \mid \text{sensitive})}_{p} \times \underbrace{P(\text{sensitive})}_{\rho} + \underbrace{P(\text{positive} \mid \text{innocent})}_{\theta} \times \underbrace{P(\text{innocent})}_{1-\rho}, \\ &= p\rho + \theta(1-\rho)\end{aligned}$$

or

$$p = \frac{\phi - \theta(1-\rho)}{\rho}.$$

If $\hat{\phi}$ is the proportion of positive responses in the achieved sample, then the **randomized response estimator** is

$$\hat{p}_{\text{rr}} = \frac{\hat{\phi} - \theta(1-\rho)}{\rho}, \quad \theta, \rho \text{ constants},$$

whose sampling variance is

$$\text{V}(\hat{p}_{\text{rr}}) = \text{V}\left(\frac{\hat{\phi} - \theta(1-\rho)}{\rho}\right) = \text{V}\left(\frac{\hat{\phi}}{\rho}\right) = \frac{1}{\rho^2} \cdot \text{V}(\hat{\phi}).$$

Since $\hat{\phi}$ is a SRS proportion estimator obtained from a sample of size $n$ in a population $\mathcal{U}$ of size $N$, its **sampling variance** is

$$\text{V}(\hat{\phi}) = \frac{\phi(1-\phi)}{n}\left(\frac{N-n}{N-1}\right),$$

from which we conclude that

$$\text{V}(\hat{p}_{\text{rr}}) = \frac{1}{\rho^2} \cdot \frac{\phi(1-\phi)}{n}\left(\frac{N-n}{N-1}\right).$$

As the true value of $\phi$ is typically not known, we instead use the unbiased estimator

$$\hat{\text{V}}(\hat{p}_{\text{rr}}) = \frac{1}{\rho^2} \cdot \frac{\hat{\phi}(1-\hat{\phi})}{n-1}\left(1 - \frac{n}{N}\right),$$

and we build a **95% C.I. for** $p$ *via*

$$\text{C.I.}_{\text{rr}}(p; 0.95): \quad \hat{p}_{\text{rr}} \pm 2\sqrt{\hat{\text{V}}(\hat{p}_{\text{rr}})}.$$

The factor $1/\rho^2$ **penalizes the uncertainty** brought by the randomized response – the higher $\rho$ is, the lower $\hat{\text{V}}(\hat{p}_{\text{rr}})$ is.

There are practical considerations that limit how high $\rho$ can get: if it is **too large**, the anonymity conferred by the approach evaporates, and we risk ruining the study by causing an increase in non-response.

**Example** We seek to determine the incidence of cheating in online courses among students in the Department of Mathematics and Statistics ($N = 442$), using a SRS with $n = 65$. We use the scheme described in this section with $\rho = 1/2$, and observe $\theta = \frac{52}{442}$ and $\hat{\phi} = \frac{21}{65}$. Find a 95% C.I. for the proportion of students who cheated during the pandemic.

We only need compute

$$\hat{p}_{\text{rr}} = \frac{21/65 - 52/442(1 - 1/2)}{1/2} = 0.53$$
$$\hat{\text{V}}(\hat{p}_{\text{rr}}) = \frac{1}{1/2^2} \cdot \frac{21/65(1 - 21/65)}{65 - 1}\left(1 - \frac{65}{442}\right) = 0.012,$$

which yields $\text{C.I.}_{\text{rr}}(p; 0.95) = 0.53 \pm 2\sqrt{0.012} \equiv (0.31, 0.74)$. ■

### Bernoulli Sampling

**Bernoulli sampling** (BS) is a **random** sampling design – we do not know the sample size **before** it is drawn.

Each unit of the population $\mathcal{U} = \{u_1, \ldots, u_N\}$ is assigned the same probability of inclusion in the sample $\mathcal{Y}$: $\pi_j = \pi \in (0, 1)$, for all $j$. We denote the **achieved sample size** by $n_a$.

60: I know, I know.

The BS design[60] consists of performing $N$ independent Bernoulli trials, each with probability of success $\pi$ (where a success means that the unit is **included** in the sample, and a failure means that it **rejected**).

The probability of obtaining a sample $\mathcal{Y}$ of size $n_a$ is then:

$$P(|\mathcal{Y}| = n_a) = \pi^{n_a}(1 - \pi)^{N - n_a}.$$

There are $2^N$ possible samples, with size varying from $n_a = 0$ to $n_a = N$.

The sample size follows a **binomial** distribution $n_a \sim B(N, \pi)$:

$$P(n_a = n) = \binom{N}{n}\pi^n(1 - \pi)^{N-n}, \quad \text{E}[n_a] = N\pi, \quad \text{V}[n_a] = N\pi(1 - \pi).$$

When $N$ is sufficiently large, this distribution is **approximately normal**; the **95% C.I. for** $n$ is thus

$$\text{C.I.}(n_a; 0.95): \quad N\pi \pm 2\sqrt{N\pi(1 - \pi)}.$$

Let $\pi_{j,k}$ be the probability of inclusion of units $u_j$ and $u_k$, $j \neq k$ in the smaple $\mathcal{Y}$. Since the Bernouilli trials are independent of one another,

$$\pi_{j,k} = P(\{u_j, u_k\} \in \mathcal{Y}) = P(u_j \in \mathcal{Y}) \cdot P(u_k \in \mathcal{Y}) = \pi_j \pi_k = \pi^2.$$

The estimator

$$\hat{\tau}_{\text{BS}} = \frac{1}{\pi} \sum_{i=1}^{n_a} y_i$$

is an **unbiased estimator of the total** $\tau$ in $\mathcal{U}$: indeed,

$$\text{E}[\hat{\tau}_{\text{BS}}] = \frac{1}{\pi}\text{E}[n_a \overline{y}] = \frac{\text{E}[n_a]\text{E}[\overline{y}]}{\pi} = \frac{N\pi\mu}{\pi} = N\mu = \tau,$$

as $n_a$ and $\overline{y}$ are independent of each other.

In the same vein, the **sampling variance** of $\hat{\tau}_{\text{BS}}$ is approximately

$$\hat{\text{V}}[\hat{\tau}_{\text{BS}}] = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right) \sum_{i=1}^{n_a} y_i^2.$$

If $N$ and $n_a$ are sufficiently large, the Central Limit Theorem comes into play again, and we build a **95% C.I. for** $\tau$ using

$$\text{C.I.}_{\text{BS}}(\tau; 0.95): \quad \hat{\tau}_{\text{BS}} \pm 2\sqrt{\hat{\text{V}}[\hat{\tau}_{\text{BS}}]}.$$

The corresponding estimators for the mean $\overline{y}_{\text{BS}}$ and the proportion $\hat{p}_{\text{BS}}$ are obtained in the usual manner.

**Example** A teacher has to correct 600 exam papers. For each paper, she rolls a die and only corrects it (at this stage) if it shows a 6.

At the end of the process, she has graded 90 papers, of which 60 have received a passing grade. Find a 95% C.I. for the total number of passes in her class.

Let $y_i = 1$ if the $i$th marked examen received a passing grade, and $y_i = 0$ otherwise. We have $N = 600$, $\pi = 1/6$, $n_a = 90$,

$$\sum_{i=1}^{90} y_i = 60, \quad \sum_{i=1}^{90} y_i^2 = 60, \hat{\tau}_{\text{BS}} = \frac{1}{1/6} \sum_{i=1}^{90} y_i = 6(60) = 360$$

$$\hat{\text{V}}[\hat{\tau}_{\text{BS}}] = \frac{1}{1/6}\left(\frac{1}{1/6} - 1\right) \sum_{i=1}^{90} y_i^2 = 6(5)(60) = 1800.$$

The 95% C.I. is thus $\text{C.I.}_{\text{BS}}(\tau; 0.95) = 360 \pm 2\sqrt{1800} \equiv [277, 443]$. We are not going to lie... it is looking particularly bleak for the students. ■

## 10.8 Exercises

1. You are tasked with estimating the annual salary of data scientists in Canada. Determine the: populations (target, study, respondent); sampling frames; samples (target, achieved); information about units (units, response variable, attributes); sources of error (coverage, non-response, sampling, measurement and processing) and variability (sampling, measurement).
2. We seek to estimate the average daily distance travelled by Ontario cars, as well as their daily fuel consumption. Discuss various approaches to be used. What are some of the issues and challenges that could be encountered?
3. We seek an estimate of the average daily distance travelled in Winter 2012 in Ontario, as are the average daily fuel consumption and the proportion of vehicles not in use. An SRS is selected from the Ontario fleet (size $N = 7,868,359$); the responses are collected in the file `Autos.xlsx`. Discuss issues that may affect the quality of the data. Provide a numerical and visual summary of the data for the sample. Give an approximate 95% C.I. for each population mean sought, with corresponding coefficient of variation.
4. We seek an estimate of the average daily distance travelled in Winter 2012 in Ontario, as are the average daily fuel consumption and the proportion of vehicles not in use. An STS is selected from the Ontario fleet (size $N = 7,868,359$), with information concerning vehicle type and age (the strata); the responses are collected in the file `Autos.xlsx`. Discuss issues that may affect the quality of the data. Provide a numerical and visual summary of the data for the sample. Give an approximate 95% C.I. for each population mean sought, with corresponding coefficient of variation. Conduct the same exercise for each stratum.
5. We seek an estimate of the average daily distance travelled in Winter 2012 in Ontario. An SRS is selected from the Ontario fleet (size $N = 7,868,359$). The responses, as well as the corresponding daily fuel consumption, are collected in the file `Autos.xlsx`. Give an approximate 95% C.I. for the characteristic of interest using quotient, regression, and difference estimation.
6. Could cluster sampling be used to provide estimates of average daily distance travelled, average daily fuel consumption, and proportion of vehicles not in use in Winter 2012 in Ontario? Treat the vehicle type and age information found in `Autos.xlsx` as cluster information.
7. Repeat the previous exercise using multi-phase and multi-stage sampling.
8. Draw $m = 1000$ SRS samples of size $n$ from the $N = 183$ countries (excluding China and India) in the 2011 Gapminder dataset to estimate the average population by country $\mu$. For $n = 30, 60, 90, 120$, what proportion of the $m$ samples yield an approximate 95% C.I. containing $\mu$? Assume that $\sigma^2$ is not known.
9. Find an approximate 95% C.I. for the average life expectancy $\mu$ of the $N = 185$ countries in the 2011 Gapminder dataset using a SRS of size $n = 20$. Is the true average life expectancy in your confidence interval? Repeat this task $m = 1000$ times, with different SRS samples. What proportion of the $m$ samples yield approximate

95% C.I. containing $\mu$? Assume that $\sigma^2$ is not known. Compare with the results of the previous exercise. How do you explain the discrepancy?

10. Find an approximate 95% C.I. for the proportion $p$ of countries whose life expectancy fell below 60 years in the 2011 Gapminder dataset ($N = 185$), using a SRS of size $n = 20$. Is the true proportion in the confidence interval? Repeat this task $m = 1000$ times, with different SRS samples. What proportion of the $m$ samples yield approximate 95% C.I. containing the true $p$? Assume that $\sigma^2$ is not known. Compare with the results of exercises 8 and 9.
11. Find an approximate 95% C.I. for the total population of the planet in the 2011 Gapminder dataset ($N = 185$), using a STS of size $n = 20$. What variable will you use to stratify the data? Repeat this task $m = 1000$ times, with different STS samples. What proportion of the $m$ samples yield approximate 95% C.I. containing the true total $\tau$? Is the distribution of the obtained totals (approximately) normal? How do you explain the shape of this distribution?
12. Find an approximate 95% C.I. for the proportion $p$ of countries whose life expectancy fell below 60 years in the 2011 Gapminder dataset ($N = 185$), using a STS of size $n = 20$. What variable will you use to stratify the data? Is the true proportion in the confidence interval? Repeat this task $m = 1000$ times, with different STS samples. What proportion of the $m$ samples yield approximate 95% C.I. containing the true $p$? Compare with the results of exercise 10.
13. Consider a sample $\mathcal{Y} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from a population of size $N = 37,444$. In a preceding study, we have found that $\sigma^2_{W;L} \approx 188.2$. Find the minimal $n$ which ensures that the bound on the error of (regression) estimation of the mean $\mu_Y$ is at most 5. Do the same for the total $\tau_Y$ and a bound of at most 250.
14. Find a 95% C.I. for the proportion of countries in the 2011 Gapminder dataset ($N = 185$) whose life expectancy is above 75 years, using a CLS with $m = 8$, assuming that the countries are grouped into $M = 22$ clusters determined by geographic regions. Assume further that the average cluster size is known to be $\overline{N} = 8.41$.
15. Consider a CLS $\mathcal{Y}$ consisting of $m$ clusters drawn from a population $\mathcal{U}$ of size $N$, distributed in $M$ clusters. Let $\mu$ be the mean and $\sigma^2$ the variance of the population $\mathcal{U}$. If the clusters are all of size $n$, show that

$$\mathrm{V}(\overline{y}_C) \approx \frac{\sigma^2 - \overline{\sigma^2}}{m}\left(1 - \frac{m}{M}\right), \quad \text{where } \overline{\sigma^2} = \frac{1}{M}\sum_{\ell=1}^{M} \sigma^2_\ell,$$

where $\sigma^2_\ell$ is the variance in the $\ell$th cluster.

## Chapter References


[1] M. Barry. *Lexicon*. Penguin Press, 2013.

[2] D. DeTurck. *Case Study 2: the 1948 Presidential Election* . 2018.

[3] A. Gower. 'Questionnaire Design for Business Surveys'. In: 20.2 (1994), pp. 125–136.

[4] M. Hidiroglou, J. Drew, and G. Gray. 'A Framework for Measuring and Reducing Non-Response in Surveys'. In: 19.1 (1993), pp. 81–94.

[5] R. Latpate et al. *Advanced Sampling Methods*. Springer Nature Singapore, 2021.

[6] S.L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, 1999.

[7] *Méthodes de sondage pour les enquêtes statistiques agricoles*. Rome: FAO. Développement Statistique.

[8] MIT Technology Review. *Storing data in DNA is a lot easier than getting it back out* . 2018.

[9] H. Rosling. *The Health and Wealth of Nations* . Gapminder Foundation, 2012.

[10] *Survey Methods and Practices, Catalogue no.12-587-X*. Statistics Canada.

# 11 The Design of Experiments

by **Patrick Boily** (inspired by **David Haziza**)

---

In the world of data-driven decision-making, it is not enough to simply possess vast datasets (which are often passively collected) and analytical prowess.

The manner in which experiments are **designed**, **conducted**, and **analyzed** can make a huge difference in the validity and reliability of the conclusions that analysts draw.

The design of experiment provides the foundation for sound experimental methodology, enabling scientists and data professionals to meticulously control variables, uncover hidden patterns, and discern causality amidst the complexity of real-world data.*

## 11.1 Basic Notions

At its core, statistics serves as the science of **collecting**, **analyzing**, and **deriving meaningful conclusions** from data.

Data can be obtained through several primary methods, each with its own unique characteristics.

One common approach to data collection involves conducting **sample surveys**. These surveys are often carried out by entities such as National Statistical Offices and polling market firms.†

The main objective of sample surveys is typically to estimate parameters for finite populations. For instance, they may aim to determine the average income within the Canadian population or calculate the unemployment rate.‡

Another method involves **observational studies**, where researchers gather data by observing and recording natural occurrences. These studies provide valuable insights into real-world phenomena but may not always allow for the establishment of causality between variables.

**Experimentation** represents a powerful way to investigate causal relationships. In experiments, researchers manipulate one or more variables and observe the effects on others. This controlled approach helps establish potential **causal networks**,[1] a crucial aspect of scientific inquiry.

These foundational concepts lay the groundwork for our exploration of experimental design.

* More details, examples, and exercises are available in [2, 5], among others.
† Such as *Statistics Canada* or *EKOS Research*, say.
‡ Survey sampling is explored in depth in Chapter 10.


11.1 Basic Notions . . . . . 733
Experiments . . . . . 734
Useful Distributions . . . . 737
11.2 Hypothesis Testing . . . . . 740
Inference on $\mu$ . . . . . 740
Inference on $\mu_1 - \mu_2$ . . . . 745
Inference on $\sigma^2$ . . . . . 751
Inference on $\sigma_1^2/\sigma_2^2$ . . . . . 753
11.3 One-Way Classification . . . 754
Randomized Designs . . . . 754
1−Way Model . . . . . 756
Analysis of Variance . . . . 757
Estimation of Parameters . 761
Unbalanced Designs . . . . 762
Contrasts . . . . . 763
Multiple Comparisons . . . 765
Model Validation . . . . . 773
Power and Sample Size . . . 776
11.4 Random Effects . . . . . 778
Estimation of Parameters . 779
Analysis of Variance . . . . 780
Inference on $\sigma^2$, $\sigma_T^2$, $\mu$ . . . . 782
Power . . . . . 783
11.5 Randomized Block Designs 784
Analysis of Variance . . . . 785
Estimation of Parameters . 789
Multiple Comparisons . . . 789
Power and Sample Size . . . 790
Model Validation . . . . . 790
11.6 Factorial Designs . . . . . 791
2-Way Factorial Experiments 791
Model Validation . . . . . 798
Model Without Interaction 799
Multiple Comparisons . . . 800
$n$-Way Factorial Designs . . 801
11.7 Exercises . . . . . 801
Chapter References . . . . . 802


1: Or **cause-and-effect** connections.

### 11.1.1 Experiments

The essence of **experimental studies** lies in the comparison of **treatments** and their respective **outcomes**. Researchers leverage experiments to address crucial questions, often revolving around topics such as:

- Is a drug a safe and effective cure for a disease? This could involve testing how AZT affects the progression of AIDS.
- What combination of protein and carbohydrate sources provides the optimal nutrition for growing lambs?
- How will long-distance telephone usage patterns change if our company introduces a different rate structure for our customers?
- Can an ice cream manufactured with a new kind of stabilizer match the palatability of our current ice cream?

A fundamental aspect of scientific reasoning involves drawing conclusions from experiments that have been meticulously designed, executed appropriately, and rigorously analyzed. Key elements include the **treatments** and **experimental units** to be employed, the **methodology** for assigning treatments to units, and the measured **responses**.

Note that the environment and observation conditions must be carefully **controlled** and **fixed**.[2]

2: Explanatory variables are under the direct control of the researchers; some are intentionally **altered**, while others are held **constant**.

#### Observational Studies against Experiments

Both observational studies and experiments are typically employed to establish relationships between two or more measured quantities. However, there is a fundamental distinction between observational studies and experiments.

In an observational study, researchers do not actively manipulate or create data; instead, they solely observe the characteristics of pre-existing data. Consequently, an observational study entails the observation of units/individuals and the measurement of variables of interest, **without any attempt to influence their responses**.

Conversely, an experiment involves the **deliberate imposition of specific treatments** on individuals/units to observe their responses. Causal inferences find justification in experiments, where the explanatory variables $x_1, \cdots, x_p$, often referred to as the "possible causes," are directly controlled by the researcher. Such experiments are known as **randomized trials** because the values of the explanatory variables are assigned to experimental units through some random mechanism.

In observational studies, the values of the explanatory variables are **observed** rather than assigned by the researcher, alongside the value of the response. In such studies, causal inferences are **not warranted** because, although efforts can be made to "control" for certain "confounding" factors, it is generally impossible to control for all relevant factors.

What constitutes a **relevant** (or confounding) **factor** in observational studies? It is a factor that both **influences the response variable(s)** and **relates to the explanatory variable(s)** on which the research focuses.

A drawback of observational studies is that the grouping of individuals into "treatments" is **beyond the experimenter's control**, and the mechanism underlying this grouping is often **unknown**.

Consequently, observed differences in responses between treatment groups may be attributable to **hidden mechanisms** rather than to the treatments.

**Example** Consider a dataset from Canada's *Health Care System* comparing the effectiveness of two procedures for treating prostate disease:

1. traditional surgery, or
2. a new method that does not require surgery.

The dataset includes many patients suffering from prostate disease, with their doctors choosing one of the two methods. Initially, the study found that patients treated with the new method were significantly more likely to die within 8 years. H

However, further data analysis revealed that this conclusion was incorrect. Why? What potential confounding variables might be at play?

### Definitions

Some concepts will re-appear time and time again in this chapter, and so we take the time to define them properly.

- **Treatments** represent the different procedures under examination. These could encompass various types or amounts of fertilizer in agronomy or distinct long-distance rate structures in marketing.
- An **experimental unit** refers to the physical entity that can be randomly assigned to a treatment. This unit may be an individual, an animal, a plot of land receiving fertilizer, and so forth, upon which measurements are taken.[3]
- The **dependent** (or response) **variable**, often denoted by $Y$, represents the observed outcome after applying a treatment to an experimental unit.
- **Randomization** involves the use of a known and perfectly controlled probabilistic mechanism to assign treatments to units.
- A **factor** in an experiment is a controlled independent variable, a variable whose levels are determined by the experimenter. Factors combine to create treatments. For instance, the baking treatment for a cake may involve specific time and temperature settings, with each variable varied independently.
- A **level** denotes the intensity setting (or value) of a factor.
- The **effect** is the change in the response caused by a change in a factor.
- A **lurking** (or hidden) **variable** is an uncontrolled variable that falls outside the experimenter's awareness and control, which could influence the experiment's outcome.
- A **cell** refers to the subset of data occurring at the intersection of one level of every treatment.

3: It does not have to be a "physical" entity *per se*, as the data may arise in a simulation context (see Chapter 12).

**Example** In each of five different campuses across the country, we selected 10 students at random to assess their attitudes toward industrial pollution. Each student responded to a specific set of questions, and their responses were aggregated into a total interview score.

| **Campus** | **I** | **II** | **III** | **IV** | **V** |
|---|---|---|---|---|---|
| **Score** | 172 | 248 | 236 | 250 | 241 |

- Experimental unit: a student
- Response variable: total aggregated score
- Factor: campus, with 5 levels
- There are 5 cells in this experiment

**Example** We would like to compare the effects of three different insecticides on a particular variety of string beans. Four plots were prepared, with each plot subdivided into three rows. Each row was planted with 100 seeds and then maintained under the insecticide assigned to the row. The insecticides were randomly assigned to the rows within a plot so that each insecticide appeared in one row in all four plots. The response variable was the number of seedlings that emerged per row.

| | **Plot** | | | |
|---|---|---|---|---|
| **Row** | **I** | **II** | **III** | **IV** |
| **1** | (A) 121 | (A) 73 | (B) 144 | (B) 134 |
| **2** | (B) 128 | (B) 141 | (C) 118 | (A) 85 |
| **3** | (C) 112 | (C) 118 | (A) 109 | (C) 111 |

Of course, we do not need to physically refer to the rows in order; in fact, it might make more sense to represent the experiment using the treatments instead of the location.

| | **Plot** | | | |
|---|---|---|---|---|
| **Insecticide** | **I** | **II** | **III** | **IV** |
| **A** | (1) 121 | (1) 73 | (3) 109 | (2) 85 |
| **B** | (2) 128 | (2) 141 | (1) 144 | (1) 134 |
| **C** | (3) 112 | (3) 118 | (2) 118 | (3) 111 |

- Experimental unit: variety of string beans
- Response variable: number of seedlings
- Factors: plot and insecticide
- Levels of the factors:
  - Plot: four levels (I, II, III, IV)
  - Insecticide: three levels (A, B, C)
- There are $3 \cdot 4 = 12$ cells in this experiment

**Example** We aim to test whether a chemical agent can prevent symptomatic infection from a respiratory diseases. A clinical trial was conducted where patients received either the compound (C) or a placebo (P). The treatment was administered to both men (M) and women (F), each belonging to a specific age group. The information is summarized below.

| | Gender | | | |
|---|---|---|---|---|
| | M | | F | |
| | Drug | | | |
| Age | P | C | P | C |
| 29− | (102)<br>0.31 | (99)<br>0.29 | (105)<br>0.28 | (105)<br>0.30 |
| 30-59 | (117)<br>0.35 | (119)<br>0.31 | (120)<br>0.31 | (119)<br>0.27 |
| 60+ | (89)<br>0.38 | (85)<br>0.41 | (91)<br>0.38 | (90)<br>0.37 |

- Experimental unit: individual on which the infected/non-infected status is measured
- Response variable: 1 = infection, 0 = no infection.
- Factors: gender, drug, and age group
- Levels of the factors:
  - Drug: two levels (compound and placebo)
  - Gender: two levels (male and female)
  - Age group: three levels (29−, 30-60, 60+)
- There are $2 \cdot 2 \cdot 3 = 12$ cells in this experiment

### 11.1.2 Useful Distributions

We have encountered several probabilistic and statistical concepts that arise time and time again in applications.[4] We briefly mention those properties that will be useful in the analysis and design of experiments.

4: See Chapters 6, 7, 8, 9, and 10.

**Sample Mean and Sample Variance** Consider a **random sample**

$$\mathcal{Y} = \{y_1, \ldots, y_n\}$$

drawn from a population with mean $\mu$ and variance $\sigma^2$, where $\mathrm{E}(y_i) = \mu$ and $\mathrm{Var}(y_i) = \sigma^2$ for $i = 1, \ldots, n$.

We assume that the sample observations in $\mathcal{Y}$ are **independent and identically distributed** (i.i.d), indicating that they were generated from the same distribution (or from the same population $\mathcal{U}$).

The **sample mean** and **sample variance** of $\mathcal{Y}$ are given by:

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \overline{y})^2.$$

As a reminder, both the sample mean and the sample variance are **unbiased estimators** of the population mean and the population variance, respectively:

$$\mathrm{E}(\overline{y}) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(y_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu,$$
$$\mathrm{Var}(\overline{y}) = \frac{1}{n^2}\sum_{i=1}^{N}\mathrm{Var}(y_i) = \frac{1}{n^2}\sum_{i=1}^{N}\sigma^2 = \frac{\sigma^2}{n},$$

and

$$\begin{aligned}\mathrm{E}(s^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{N}\mathrm{E}(y_i^2) - n\mathrm{E}(\overline{y}^2)\right] = \frac{1}{n-1}\left[\sum_{i=1}^{N}(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\ &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] = \sigma^2.\end{aligned}$$

**Probability Distributions** The distribution of sample observations is described by a probability distribution. For a continuous variable $Y$, the probability distribution is characterized by a **density function**, denoted as $f(y)$, with the following properties:

$$f(y) \geq 0, \quad P(a \leq Y \leq b) = \int_a^b f(y)dy, \quad \int_{-\infty}^{+\infty} f(y)dy = 1.$$

The mean of a probability distribution, denoted by $\mu$, serves as a measure of **centrality location** and is defined as:

$$\mu = \mathrm{E}(Y) = \int_{-\infty}^{+\infty} yf(y)dy.$$

The variance $\sigma^2$ can be used to quantify the **dispersion** of a variable:

$$\sigma^2 = \mathrm{Var}(Y) = \mathrm{E}[(y-\mu)^2] = \int_{-\infty}^{+\infty}(y-\mu)^2 f(y)dy.$$

**Normal Distributions** If $Y$ follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu$ and variance $\sigma^2$, its probability density function is given by

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), \quad -\infty < y < \infty$$

.

If $y_1, \ldots, y_n$, is a random sample generated from a $\mathcal{N}(\mu, \sigma^2)$, then $\overline{y}$ and $s^2$ are **statistically independent**.

Normal distributions are entirely characterized by their expectation $\mathrm{E}(Y) = \mu$ and variance $\mathrm{Var}(y) = \sigma^2$; any other normal random variable with the same properties must in fact be exactly $Y \sim \mathcal{N}(\mu, \sigma^2)$. We can **standardize** any such random variable:

$$Z = \frac{Y-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

The resulting random variable $Z$ is said to be **standard normal**.

We have discussed normal distributions in detail in Section 6.3.3; the primacy of normal distributions in statistical applications is explained by the following oft-used result.

**Central Limit Theorem:** let $Y_1, \ldots, Y_n$, be $n$ i.i.d. random variables with mean $\mu$ and variance $\sigma^2$. The random variable

$$Z_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

**converges in distribution** to $Z \sim \mathcal{N}(0,1)$, where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$.[5]

5: A sequence $\{X_n\}$ of random variables, with cumulative distribution functions $\{F_1\}$ converges in distribution to a random variable $X$ with cumulative distribution function $F$ if $F_n(x) \to F(x)$ for all $x$ where $F$ is continuous.

**$\chi^2$ Distributions** If $Z_1, \ldots, Z_k \sim \mathcal{N}(0,1)$ are $k$ i.i.d. random variables, then the random variable

$$Y = Z_1^2 + \cdots + Z_k^2$$

follows a $\chi^2_k$ distribution (with $k$ degrees of freedom).[6]

6: We have also used the notation $\chi^2(k)$ in these notes.

The probability density function of such a random variable is

$$f(y) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} y^{k/2-1} e^{-y/2}, \quad y > 0,$$

where $\Gamma$ is the Gamma function.

When $Y \sim \chi^2_k$, we have $\mathrm{E}(Y) = k$ and $\mathrm{Var}(Y) = 2k$.

As the degrees of freedom parameter $k$ increases, the chi-square distribution converges in distribution to a normal distribution with a mean equal to $k$ and a variance equal to $2k$. This convergence is a direct consequence of the Central Limit Theorem.

Now, if we have a random sample $y_1, \ldots, y_n$ generated from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, we can make the following observation:

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}.$$

This implies that we can obtain an unbiased estimator of $\sigma^2$ by dividing the sum of squares by the number of degrees of freedom, which is $n-1$. This unbiased estimator of the population variance will prove useful when introduce **ANOVA tables**.

**Student's $T-$Distributions** If $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_k$ independent, then the distribution of the random variable

$$W = \frac{Z}{\sqrt{Y/k}}$$

is that of a Student $T-$distribution with $k$ degrees of freedom, denoted by $W \sim t_k$.[7]

7: We have also used the notation $t(k)$ in these notes.

The probability density function of the $T-$distribution is :

$$f(w) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)\left(1+\frac{w^2}{k}\right)^{\frac{k+1}{2}}}, \quad -\infty < w < \infty.$$

The $T-$distribution is symmetric, and its expected value is $\mathrm{E}(W) = 0$, while its variance is $\mathrm{Var}(W) = \frac{k}{k-2}$ for $k > 2$. As the degrees of freedom parameter $k$ increases, $W$ converges in distribution to the standard normal distribution $\mathcal{N}(0,1)$.

**Fisher's $F-$Distributions** If $X \sim \chi^2_u$ and $Y \sim \chi^2_v$ are independent, then the distribution of the random variable

$$W = \frac{\frac{X}{u}}{\frac{Y}{v}}$$

is that of a Ficher $F-$distribution with $(u, v)$ degrees of freedom, denoted by $W \sim F_{u,v}$.[8]

8: We have also used the notation $F(u,v)$ in these notes. The order of the degrees of freedom is important: if $W \sim F_{u,v}$, then $\frac{1}{W} \sim F_{v,u}$.

The probability density function of the $F-$distribution is given by:

$$f(w) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \frac{w^{\frac{u}{2}-1}}{\left(1+\frac{u}{v}w\right)^{\frac{u+v}{2}}}, \quad w > 0.$$

The expectation of $W \sim F_{u,v}$ is only defined if $v > 2$; its variance is only defined if $v > 4$. In those cases, we have

$$\mathrm{E}(W) = \frac{u}{v-2} \quad \text{and} \quad \mathrm{Var}(W) = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)};$$

if $v \leq 4$, $W$ does not have a well-defined variance, if $v \leq 2$, it does not have a well-defined expectation. Moreover, if $X \sim t(k)$, then $X^2 \sim F_{1,k}$.

## 11.2 Review of Hypothesis Testing

We have discussed hypothesis testing in detail in Section 7.4 (and in the chapters on applications); we briefly review its important features as it relates to the design of experiment.

### 11.2.1 Inference on the Population Mean

The customary Student $T-$test relies on several key assumptions:

1. a random sample of size $n$ is selected for analysis;
2. the individual observations in this sample are denoted by $y_1, y_2, \ldots, y_n$;
3. these observations are assumed to have been generated from a normal population with a mean parameter $\mu$ and variance $\sigma^2$, expressed as:
$$y_1, y_2, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2).$$

However, what if the underlying population does not follow a normal distribution? The Student $t-$test exhibits robustness in the sense that the distribution of the test statistic remains relatively stable even when the normality assumption is not strictly met. This robustness holds, provided that the sampled population exhibits an **approximately mound-shaped** distribution.

In the context of hypothesis testing: we typically formulate both **null** and **alternative hypotheses** as follows. We pit the

null hypothesis ($H_0$): $\mu = \mu_0$

against the **two-tailed**

alternative hypothesis ($H_1$): $\mu \neq \mu_0$,

or either of the **one-tailed**

alternative hypothesis ($H_1$): $\mu > \mu_0$ (one-tailed test), or
alternative hypothesis ($H_1$): $\mu < \mu_0$ (one-tailed test).

We define the following terms related to hypothesis testing (see Table 11.5 for a summary):

- a **type I error** occurs when we wrongly reject the null hypothesis $H_0$ but it is in fact valid:

$$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true});$$

- a **type II error** occurs when we do not reject the null hypothesis $H_0$ but it should in fact be rejected:

$$\beta = P(\text{Type II error}) = P(\text{do not reject } H_0 \mid H_0 \text{ is false});$$

- the **power of the test** is the probability of correctly rejecting the null hypothesis when it is in fact false:

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

We discuss other types of error in one of the sidenotes of Section 7.4.1.

| | | **Reality** | |
|---|---|---|---|
| | | $H_0$ **is true** | $H_0$ **is false** |
| **Decision** | **Reject** $H_0$ | type I error ($\alpha$) | right decision ($1-\beta$) |
| | **Do not reject** $H_0$ | right decision ($1-\alpha$) | type II error ($\beta$) |

**Table 11.5:** The four possible outcomes for hypothesis testing.

We usually set the **significance level** $\alpha$ of the test, typically chosen as $\alpha = 0.01, 0.05, 0.1$, and aim to construct a test with **high power** $1-\beta$, typically for $\beta = 0.1, 0.2$.

The **test statistic** $t_0$ is calculated as follows:

$$t_0 = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}.$$

If $H_0$ is true,[9] the distribution of $t_0$ follows a $T-$distribution with $n-1$ degrees of freedom ($t_{n-1}$).

9: We also say "under $H_0$".

For a two-tailed test at the level $\alpha$, we **reject** $H_0$ when $|t_0|$ is greater than the **critical value** $t_{\alpha/2;n-1}$.[10] For a one-tailed test, either $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ or $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, we reject $H_0$ based on the sign of $t_0$:

10: See Section 7.3.2 for more information.

- for $H_1 : \mu > \mu_0$, we reject $H_0$ when $t_0 > t_{\alpha;n-1}$;
- for $H_1 : \mu < \mu_0$, we reject $H_0$ when $t_0 < -t_{\alpha;n-1}$.

We can then build an $100(1-\alpha)\%$ **confidence interval** for $\mu$ according to:

$$\overline{y} \pm t_{\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}}$$

The **margin of error** $m$ (sometimes known as the **bound on the error of estimation**, see Chapter 10) is

$$m = t_{\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}}$$

We reject $H_0$ if $\left|\frac{\overline{y}-\mu_0}{s/\sqrt{n}}\right| > t_{\alpha/2;n-1}$, that is, if $\overline{y}$ lies in the **rejection region**

$$\overline{y} \geq \mu_0 + m \quad \text{or} \quad \overline{y} \leq \mu_0 - m.$$

**Inference about $\mu$: Power** The power of a test depends on various factors, including the **specific alternative hypothesis**, the **significance level** $\alpha$, the **variance** $\sigma^2$, and the **sample size** $n$.

We can think of the power as a function

$$\pi(\theta) = P(\text{reject } H_0 : \theta = \theta_0 \mid \text{observed sample}).$$

The **power function** $\pi(\theta)$ obviously depends on the **true value** of the parameter $\theta$, of course, but may also be influenced by the **sample size** and the **rejection rule** or **significance level** of the test. By construction, we must have $\pi(\theta_0) = \alpha$.

We can compute the power of the Student $T-$test with the help of the following random variable: if $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_k$ are independent, the distribution of

$$W = \frac{Z+\delta}{\sqrt{X/k}}$$

is a **non-central $T-$distribution with $k$ degrees of freedom and non-centrality parameter** $\delta$, denoted by $W \sim t_k(\delta)$.[11]

11: When $\delta = 0$, this clearly reduces to the standard Student $T-$distribution.

We take a detailed look at computing the power of the test for a one-tailed test with hypotheses $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

In this case, we reject $H_0$ if $t_0 > t_{\alpha;n-1}$, which is equivalent to

$$\frac{\overline{y}-\mu_0}{s/\sqrt{n}} > t_{\alpha;n-1}.$$

The power function of the test can then be expressed as:

$$\pi(\mu) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} > t_{\alpha;n-1} \,\middle|\, H_0 \text{ is false}\right) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} > t_{\alpha;n-1} \,\middle|\, \mu > \mu_0\right).$$

To compute this probability, we first note that

$$\frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{(\overline{y} - \mu) + (\mu - \mu_0)}{s/\sqrt{n}} = \frac{\sqrt{n}(\overline{y} - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)/\sigma}{s/\sigma}.$$

According to the central limit theorem, $Z = \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

Furthermore, $X = (n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. If $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$,[12] then

12: In practice, we use $\delta \approx \sqrt{n}(\mu - \mu_0)/s$.

$$\frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{Z + \delta}{X/(n-1)}.$$

Under $H_1$, then, we have:

$$W = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}(\sqrt{n}(\mu - \mu_0)/\sigma).$$

**Example** Let $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d., with $s = 10$. We want to test $H_0 : \mu = 60$ against $H_1 : \mu > 60$; assume that we reject $H_0$ if $\overline{y} \geq 62$.

- What is the power of the test when $n = 25$ and the true value of the mean is $\mu = 63$?

  In this case, we have $\delta \approx \sqrt{25}(63 - 60)/10 = 1.5$ and

  $$\pi(63) = P(\overline{y} \geq 62 \mid \mu = 63) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{25}} \,\middle|\, \mu = 63\right) = P(t_{24}(1.5) \geq 1) = 0.6933.$$

  We can compute this in R as follows:

  ```
  1 - pt(q=1, df=24, ncp=1.5)
  ```

  Thus, if $\mu = 63$, the probability of correctly rejecting $H_0$ is $\approx 70\%$.
- Repeat the calculation, but assuming that $n = 100$ instead. In this case, we have $\delta \approx \sqrt{100}(63 - 60)/10 = 3$ and

  $$\pi(63) = P(\text{reject } H_0 \mid \mu = 63) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{100}} \,\middle|\, \mu = 63\right) = P(t_{99}(3) \geq 2) = 0.8401,$$

  which can also be obtained in R as follows:

  ```
  1 - pt(q=2, df=99, ncp=3)
  ```

  We note that, for given values of $\mu$ and $s$, the power of the test increases as the sample size $n$ increases.
- For an arbitrary $n$, we have $\delta \approx \sqrt{n}(63 - 60)/10 = 0.3\sqrt{n}$, and

  $$\begin{aligned}\pi(63) &= P(\text{reject } H_0 \mid \mu = 63) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{n}} \,\middle|\, \mu = 63\right) \\ &= P(t_{n-1}(0.3\sqrt{n}) \geq 0.2\sqrt{n}).\end{aligned}$$

- If the true parameter value is $\mu = 60$, then for an arbitrary sample size $n$, we have $\delta = \sqrt{n}(60 - 60)/10 = 0$ and

$$\pi(60) = P(\text{reject } H_0 \mid \mu = 60) = P\left(\frac{\overline{y} - \mu_0}{s/\sqrt{n}} \geq \frac{62 - 60}{10/\sqrt{n}} \middle| \mu = 60\right)$$
$$= P(t_{n-1} \geq 0.2\sqrt{n}) = \alpha.$$

  Note that $\pi(60)$ corresponds to the probability of a Type I error for a given decision rule and sample size. □

In general, the power of a test increases as:

- the effect $|\mu - \mu_0|$ increases for fixed values of $n$ and $s$;
- the sample size increases for fixed values of $\mu$ and $s$;
- $s$ decreases for fixed values of $\mu$ and $n$.

**Sample Size** When designing an experiment, it is crucial to determine an appropriate sample size. Typically, researchers aim to determine the sample size $n$ that guarantees a high statistical power.[13] To achieve this, the they need to specify the following **key factors**.

13: Often set at $1 - \beta = 0.8$ or 0.9.

1. The desired **power**, which represents the probability of detecting a true effect if it exists;
2. the **significance level** $\alpha$ (the probability of making a Type I error);
3. the **effect size** $|\mu - \mu_0|$, which is chosen to represent a practically meaningful difference between groups or conditions, and
4. an estimate or range for the **population variance** $\sigma^2$.

We illustrate the process *via* a simple example.

**Example** Let $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d. We wish to test the following hypotheses:

$$H_0 : \mu = 100 \quad \text{against} \quad H_1 : \mu > 100.$$

We assume that 20 a plausible value for $\sigma$, and that the level of significance $\alpha$ is 0.05. If an effect $\mu - \mu_0 = 10$ is considered meaningful, what sample size is required to detect such a difference with a power of 0.9?

Given our assumption about $\sigma^2$, the distribution of the test statistic

$$Z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}}$$

is standard normal, $\mathcal{N}(0, 1)$.

In order for $\mu - \mu_0$ to be 10, we must have $\mu = 110$; we can achieve a power of 0.9 as follows:

$$\pi(110) = P(\text{reject } H_0 \mid \mu = 110) = 0.9 \Leftrightarrow P\left(Z \geq z_{0.05} \mid \mu = 110\right) = 0.9$$
$$\Leftrightarrow P\left(\frac{\overline{y} - 100}{20/\sqrt{n}} \geq 1.645 \middle| \mu = 110\right) = 0.9 \Leftrightarrow P\left(\overline{y} \geq 1.645 \cdot \frac{20}{\sqrt{n}} + 100 \middle| \mu = 110\right) = 0$$
$$\Leftrightarrow P\left(\frac{\overline{y} - 110}{20/\sqrt{n}} \geq 1.645 - \frac{10\sqrt{n}}{20}\right) = 0.9.$$

What is the corresponding quantile of the standard normal distribution?

```
qnorm(p=0.9, mean=0, sd=1, lower.tail=FALSE)
```

```
[1] -1.281552
```

Then, we must have

$$1.645 - \frac{10\sqrt{n}}{20} = -1.29,$$

which is to say, $n \approx 35$. □

### 11.2.2 Inference on the Difference of Means

We start with an example borrowed from [4].

**Motivational Example** An experiment was conducted to compare the mean number of tapeworms in the stomachs of sheep that had been treated for worms against the mean number in those that were untreated.

A sample of 14 worms-infected lambs was randomly divided into two groups: 7 were injected with the drug and the remainder were left untreated. After a 6-month period, the lambs were slaughtered and the following worm counts were recorded.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Drug-treated sheep | 18 | 43 | 28 | 50 | 16 | 32 | 13 |
| Untreated sheep | 40 | 54 | 26 | 63 | 21 | 37 | 39 |

How would we test the hypothesis that there is no difference in the mean number of worms between treated and untreated lambs? □

We will return to this example after some important notions.

To test for the **difference of means**, we assume two populations, denoted by I and II, in each of which the distribution of the response variable is taken to be normal.[14]

For Population 1, let $\mu_1$ and $\sigma_1^2$ be the respective **population mean** and **variance**, and analogously, for Population II, $\mu_2$, and $\sigma_2^2$.[15] A key assumption is that the population variances are equal: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Let $y_{1,1}, \ldots, y_{1,n_1}$ be a random sample of size $n_1$ drawn from Population I, with sample mean $\overline{y}_1$, and $y_{2,1}, \ldots, y_{2,n_2}$ be a random sample of size $n_2$ drawn from Population II, with sample mean $\overline{y}_2$. Crucially, these samples are assumed to be **independent**.

Expressed in distributional terms:

$$y_{1,1}, \ldots, y_{1,n_1} \sim \mathcal{N}(\mu_1, \sigma^2), \qquad y_{2,1}, \ldots, y_{2,n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$$

or equivalently:

$$\begin{aligned} y_{1,i} &= \mu_1 + \varepsilon_{1,i}, \quad \varepsilon_{1,i} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n_1, \quad \text{and} \\ y_{2,i} &= \mu_2 + \varepsilon_{2,i}, \quad \varepsilon_{2,i} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n_2. \end{aligned}$$

14: Note that in the motivational example, the response is the worm count, which cannot be normally distributed as negative and fractional values cannot arise. Nevertheless, that assumption may be a good approximation to reality (see Section 6.3.6, for instance).

15: Referring to the motivational example, $\mu_1$ and $\mu_2$ are the true worm count means in the populations of treated and untreated lambs, respectively.

The test's **null** and **alternative** hypotheses are:

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2;$$

the **treatment effect** is denoted by $\mu_1 - \mu_2$.[16]

We require a **test statistic** to determine whether to reject or accept the null hypothesis, $H_0$. Setting the level of the test as $\alpha$,[17] we aim to formulate a test with a substantial power.

The customary $T-$statistic with significance level $\alpha$ is

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{where} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the **pooled estimate of the common variance** $\sigma^2$.

If the null hypothesis $H_0$ holds true, the test statistics $t_0$ follows a $T-$distribution with $n_1 + n_2 - 2$ degrees of freedom, $t_0 \sim t_{n_1+n_2-2}$. The decision to **reject the null hypothesis** at level $\alpha$ is made when

$$|t_0| > t_{\alpha/2;n_1+n_2-2}.$$

In practice, the decision often hinges on the $p$-value. The computation for the $p$-value (in the two-tailed case) is:

$$p-\text{value} = 2P(t_{n_1+n_2-2} > |t_0|);$$

that quantity is smaller then $\alpha$ if and only if the test rejects $H_0$ at level $\alpha$.

**Motivational Example (Cont.)** We compute the required quantities.

```
y.1 <- c(18,43,28,50,16,32,13)
y.2 <- c(40,54,26,63,21,37,39)
(y.bar.1 <- mean(y.1))
(y.bar.2 <- mean(y.2))
(s.2.1 <- var(y.1))
(s.2.2 <- var(y.2))
```

```
[1] 28.57143
[1] 40
[1] 198.619
[1] 215.3333
```

The pooled estimate of the variance is easy to compute.

```
n.1 = length(y.1)
n.2 = length(y.2)
(n.1+n.2-2)
(s.2.p <- ((n.1-1)*s.2.1 + (n.2-1)*s.2.2)/(n.1 + n.2 - 2))
```

16: When the alternative hypothesis is in the form $H_1 : \mu_1 \neq \mu_2$, the test is a **two-tailed test**. If, however, the alternative hypothesis is either $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$, the test becomes a **one-tailed test**.

17: Common values: $\alpha = 0.01, 0.05, 0.1$.

```
[1] 12
[1] 206.9762
```

The test statistic is computed below.

```
(t_0 <- (y.bar.1 - y.bar.2)/sqrt(s.2.p*(1/n.1 + 1/n.2)))
```

```
[1] -1.486161
```

The $p-$value for the two-sided test is thus $2P(t_{12} > |-1.486161|)$.

```
2*pt(q=t_0, df=n.1 + n.2 - 2, lower.tail=TRUE)
```

```
[1] 0.1630303
```

Since the $p-$value is larger than $\alpha = 0.05$, we have insufficient evidence to reject $H_0$, which is to say that the observed data is compatible with the idea that the treatment has no effect. □

We have discussed this before (in Section 7.4, notably), but we will repeat it here for good measure: failure to reject the null hypothesis $H_0$ is not the same as accepting the null hypothesis $H_0$. We cannot **prove** $H_0$, we can only show that the observed data is at least compatible with it.[18]

18: We can **reject** $H_0$, however, which is equivalent to saying that the observed data is not compatible with it.

**Power and Sample Size** We now turn to the sample size determination $n_1$ and $n_2$. In a study, these are usually determined based on the need to offer **sufficient statistical power**.

When $H_0$ is true, the test statistic $t_0$ follows a Student $T-$distribution with $n_1 + n_2 - 2$ degrees of freedom. However, when $H_0$ is false, $t_0$ follows a non-central $T-$distribution with non-centrality parameter

$$\delta = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Suppose we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$.

The power function of the test is then given by

$$\begin{aligned}\pi(\mu_1 - \mu_2) &= P\left(\frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha;n_1+n_2-2} \,\middle|\, H_0 \text{ is false}\right) \\ &= P\left(\frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha;n_1+n_2-2} \,\middle|\, \mu_1 - \mu_2 > 0\right).\end{aligned}$$

The power function increases with $\delta$. Thus, the power **increases** when:

1. $|\mu_1 - \mu_2|$ **increases** – a large difference between the means is easier to detect;

2. $\sigma$ **decreases** – a given difference between $\mu_1$ and $\mu_2$ is easier to detect when the errors $\varepsilon_{\ell,j}$ are small, and/or
3. $n_1$ and/or $n_2$ **increases**.

**Confidence Intervals** We can construct an **approximate** $100(1-\alpha)\%$ **confidence interval for** $\mu_1 - \mu_2$**:**

$$\text{C.I.}(\mu_1 - \mu_2; 1-\alpha) \equiv \overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2;n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

In the previous example, we get a 95% confidence interval for $\mu_1 - \mu_2$ by computing

$$\text{C.I.}(\mu_1 - \mu_2; 0.95) \equiv (28.57 - 40) \pm 2.1788 \cdot 14.39 \sqrt{\frac{1}{7} + \frac{1}{7}} \iff (-28.18, 5.32).$$

Because the interval contains 0, we do not have enough evidence to reject $H_0$ – the data is not incompatible with the notion that $\mu_1 - \mu_2 = 0$.[19] This matches the $p-$test result from the previous section.

19: Which is not the same as saying that we accept $H_1 : \mu_1 - \mu_2 \neq 0$.

**Paired-Difference Test** When the samples are drawn independently from the two populations, we refer to the test as **unpaired**.[20] In a **paired** scenario, the units are not independent:[21] we could imagine selecting $n = 7$ sheep, testing them for tapeworm **before** treating them with a drug, then testing the same sheep for tapeworm **after** the treatment.

20: We often have $n_1 \neq n_2$.

21: In some sense, they are maximally dependent.

If a given specimen is somehow more likely to be afflicted by tapeworm due to genetics or farmer care, we wouldn't be surprised to find a link in its before/after measurements.

**Motivational Example** To compare the wear-and-tear qualities of two types of road paints, A and B, a sample of each is applied to a small area of five randomly selected roads. The roads operate as they normally do, with their specific usage patters, and the number of weeks to some "failure" threshold is recorded for each sample.

These measurements appear in the table below. Do the data present sufficient evidence to indicate a difference in the average wear for the two paint types?

| Road | Paint A | Paint B |
|---|---|---|
| **1** | 9.1 | 8.7 |
| **2** | 11.2 | 10.7 |
| **3** | 9.6 | 9.0 |
| **4** | 8.6 | 8.2 |
| **5** | 8.9 | 8.4 |

If we treated these samples as independent, we would be able be able to answer the question using the pooled variance $s_p^2$, computed with the help of $\overline{y}_A$, $\overline{y}_B$, $s_A$, $s_B$, and $n_A = n_B = 5$.

The **two-sample pooled $T-$test** would conclude that we cannot reject the null hypothesis $H_0 : \mu_A = \mu_B$, which is certainly thought-provoking given that the time to "failure" is systematically longer for Paint A than it is for Paint B. □

We have alluded to this problem at the start of the section: the two-sampled pooled $T-$test **is not the proper statistical test** to use in this case because the two samples are **not independent**.

**Motivational Example (Cont.)** Indeed, the (pair of) measurements Paint A and Paint B for a particular roadway are definitely **related**. The readings have approximately the same magnitude for a road but vary markedly from one road to another. Paint wear-and-tear is largely determined by **traffic volume** and **type**, the **weather**, and the **road surface**, say.

Since each road is likely to have different characteristics on that front, we expect a large amount of variability in the data from one road to another.

In designing the paint wear-and-tear experiment, the experimenters realized that the measurements would vary greatly from road to road. If the paint types (five of type A and five of type B) were randomly assigned to 10 roads, resulting in two independent random samples of size 5, this variability would result in a large standard error and make it difficult to detect a difference in the means.

Instead, they chose to "**pair**" the measurements, comparing the wear-and-tear for Paint A and Paint B on each of the five roads.

| Road | Paint A | Paint B | Difference $d$ |
|---|---|---|---|
| **1** | 9.1 | 8.7 | 0.4 |
| **2** | 11.2 | 10.7 | 0.5 |
| **3** | 9.6 | 9.0 | 0.6 |
| **4** | 8.6 | 8.2 | 0.4 |
| **5** | 8.9 | 8.4 | 0.5 |

This experimental design, sometimes called a **paired-difference** or **matched pairs design**, allows us to eliminate the road-to-road variability by looking at only the five difference measurements shown above. These five differences form a **single random sample** of size $n = 5$. □

For a **paired-difference test** with $n$ samples, we compute $d_i = y_{1,i} - y_{2,i}$ for $i = 1, \ldots, n$. The **null** and the **alternative hypotheses** are:

$$H_0 : \mu_d = 0$$

and

$$H_1 : \mu_d \neq 0 \quad \text{or} \quad H_1 : \mu_d > 0 \quad \text{or} \quad H_1 : \mu_d < 0,$$

while the **test statistic** is:

$$t_0 = \frac{\overline{d} - 0}{s_d/\sqrt{n}}, \tag{11.1}$$

where $\overline{d} = (d_1 + \cdots + d_n)/n$ and

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2.$$

For a two-tailed test at level $\alpha$, we reject $H_0$ when

$$|t_0| > t_{\alpha/2;n-1}.$$

For a one-tailed test $H_0 : \mu_d = 0$ against $H_1 : \mu_d > 0$ (respectively, $H_1 : \mu_d < 0$), we reject $H_0$ when

$$t_0 > t_{\alpha;n-1}; \quad (\text{resp. } t_0 < -t_{\alpha;n-1}).$$

We can build an approximate $100(1-\alpha)\%$ confidence interval for $\mu_d$ using

$$\text{C.I.}(\mu_d; 1-\alpha) \equiv \overline{d} \pm t_{\alpha/2;n-1} \cdot \frac{s_d}{\sqrt{n}}.$$

**Motivational Example (Cont.)** We prepare the data.

```
d <- c(0.4, 0.5, 0.6, 0.4, 0.5)
n = length(d)
```

Simple calculations leads to $\overline{d}$ and $s_d$.

```
(d.bar <- mean(d))
(s.2.d <- var(d))
```

```
[1] 0.48
[1] 0.007
```

The test statistic $t_0$ can be computed easily.

```
(t_0 <- (d.bar - 0)/sqrt(s.2.d/n))
```

```
[1] 12.8285
```

At significance level $\alpha$, the critical value of Student's $T$ distribution with $n - 1 = 4$ degrees of freedom is $t_{\alpha/2;n-1}$, which can compute using either of the following ways in R.

```
alpha = 0.05
(t.crit = qt(p=1 - 0.05/2, df=n-1))
qt(p=0.05/2, df=n-1, lower.tail = FALSE)
```

```
[1] 2.776445
```

Since $12.829 = t_0 > t_{4;0.025} = 2.776$, we reject $H_0$ and we conclude that there is a difference in the mean wear-and-tear for paints A and B.[22]

We build an approximate 95% confidence interval for $\mu_d$ as follows.

22: Note that the observed value $t_0 = 12.829$ is quite large for the Student $T$ distribution with 4 degrees of freedom, and the test result is highly significant.

```
c(d.bar - t.crit*sqrt(s.2.d/n),
  d.bar + t.crit*sqrt(s.2.d/n))
```

```
[1] 0.3761149 0.5838851
```

Note that this interval is much narrower than the interval that would have been obtained using the unpaired data, which indicates that the paired difference design increased the accuracy of the estimate – we have gained valuable information by using this design. □

The paired-difference test or matched pairs design used in the paint wear-and-tear experiment is a special case of an experimental design called a **randomized block design** (see Section 11.5). Importantly, the pairing (or blocking) must occur when the experiment is **planned**, and not after the data are collected.

### 11.2.3 Inference on the Population Variance

In some research situations, the primary interest lies in making inferences concerning **population variances** rather than focusing solely on population means. We begin by considering a test designed for a **single** population variance.

Imagine we have selected a random sample, represented as $y_1, \ldots, y_n$, from a population characterized by a mean of $\mu$ and a variance of $\sigma^2$. An important assumption is that the population from which this sample is drawn is **normally distributed**, i.e. $Y \sim \mathcal{N}(\mu, \sigma^2)$.

The hypothesis test pits

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{against} \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

The analysis uses the **test statistic**

$$\chi_0^2 = (n-1)s^2/\sigma_0^2.$$

Under the assumption that $H_0$ is indeed true, the distribution of $\chi_0^2$ follows a $\chi_{n-1}^2$ distribution.

We reject $H_0$ if $\chi_0^2 > \chi_{\alpha/2;n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2;n-1}^2$, with

$$P(W > \chi_{\alpha/2;n-1}^2) = P(W < \chi_{1-\alpha/2;n-1}^2) = \alpha/2, \quad \text{where } W \sim \chi_{n-1}^2.$$

We build an approximate $100(1-\alpha)\%$ **confidence interval for** $\sigma^2$ *via*:

$$\frac{(n-1)s^2}{\chi_{\alpha/2;n-1}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2;n-1}^2}.$$

**Example** An experimenter believes that the variability of her measuring apparatus has a standard deviation of $\sigma = 2.5$. During an experiment, the measurements recorded were 4.2, 5.3, and 10.3. The question arises: do these observations support or contradict her belief? We test this assertion using a significance level of $\alpha = 0.05$.

Firstly, we define our null and alternative hypotheses as:

$$H_0 : \sigma^2 = 6.25 \quad \text{against} \quad H_1 : \sigma^2 \neq 6.25.$$

We can find the test statistics $\chi_0^2$ as follows.

```
x <- c(4.2, 5.3, 10.3)
n = length(x)
(s.2 = var(x))
```

```
[1] 10.57
```

```
sigma.2 = 6.25
chi.2.0 = (n-1)*s.2/sigma.2
```

```
[1] 3.3824
```

We can compute the critical $\chi_{n-1}^2$ values at $\alpha = 0.05$.

```
alpha = 0.05
(crit.lv = qchisq(p=alpha/2, df=2))
(crit.uv = qchisq(p=1-alpha/2, df=2))
```

```
[1] 0.05063562
[1] 7.377759
```

We reject the null hypothesis $H_0$ if $\chi_0^2 > 7.38$ or $\chi_0^2 < 0.05$. Since the observed value of $\chi_0^2 = 3.3824$ lies between the critical values, we do not reject $H_0$.[23]

23: This indicates that the data does not provide sufficient evidence to dispute the experimenter's initial belief about the variability of her instrument.

She can build an approximate 95% confidence interval for $\sigma^2$ by using the formula.

```
c((n-1)*s.2/crit.uv, (n-1)*s.2/crit.lv)
```

```
[1] 2.865369  417.4927
```

This wide range implies a high level of uncertainty about the true variance, which further underscores the need for more data (or a different testing approach). □

### 11.2.4 Inference on the Ratio of Variances

We now turn our attention to the case of comparing two population variances. Consider two **normal populations**, labeled I and II. Denote the population variances associated with each populations by $\sigma_1^2$ and $\sigma_2^2$.

We draw a random sample of size $n_1$ from Population I:

$$y_{1,1}, \ldots, y_{1,n_1} \sim \mathcal{N}(0, \sigma_1^2)$$

and similarly from Population II:

$$y_{2,1}, \ldots, y_{2,n_2} \sim \mathcal{N}(0, \sigma_2^2).$$

The samples are **unpaired**, and so assumed to be **independent** of one another.

The **hypothesis test** for the variances is framed as:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

The **test statistic** employed for the test is

$$F_0 = s_1^2 / s_2^2.$$

Under the assumption that $H_0$ is true, the distribution of $F_0$ follows an $F$ distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The null hypothesis $H_0$ is **rejected** at significance level $\alpha$ if

$$F_0 > F_{\alpha/2;n_1-1,n_2-1} \quad \text{or} \quad F_0 < F_{1-\alpha/2;n_1-1,n_2-1},$$

with

$$P(W > F_{\alpha/2;n_1-1,n_2-1}) = P(W < F_{1-\alpha/2;n_1-1,n_2-1}) = \alpha/2, \quad \text{where } W \sim F_{n_1-1,n_2-1}.$$

Equivalently, we can express a $100(1-\alpha)\%$ **confidence interval for the ratio** $\sigma_1^2/\sigma_2^2$ *via*:

$$s_1^2/s_2^2 \cdot F_{1-\alpha/2;n_2-1,n_1-1} < \sigma_1^2/\sigma_2^2 < s_1^2/s_2^2 \cdot F_{\alpha/2;n_2-1,n_1-1}.$$

Note the order of the degrees of freedom.[24]

24: We may need to leverage the relationship

$$F_{1-\gamma,\nu_1,\nu_2} = \frac{1}{F_{\gamma,\nu_2,\nu_1}}$$

in the analysis.

**Example** The same experimenter is concerned that the variability of her responses may not be the same when she is using two different experimental procedures.

She conducts a preliminary study with random samples of $n_1 = 11$ and $n_2 = 9$ responses and obtains $s_1^2 = 8.25$ and $s_2^2 = 4.32$, respectively. Do the sample variances present sufficient evidence to indicate that the population variances are unequal?

The null and alternative hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

The test statistic is given by $F_0 = 8.25/4.32 = 1.91$. We reject $H_0$ at level $\alpha = 0.05$ if

$$F_0 > F_{0.025,10,8} = 4.29 \quad \text{or} \quad F_0 < F_{0.975,10,8} = 0.26;$$

therefore, we cannot reject $H_0$ based on the observed data: there is insufficient evidence to indicate a difference in the population variances.[25]

25: Perhaps if we increased the sample sizes?

We can build a 95% confidence interval for $\sigma_1^2/\sigma_2^2$ *via*:

$$\begin{aligned} \text{C.I.}(\sigma_1^2/\sigma_2^2; 0.95) &\equiv (8.25/4.32 \cdot F_{0.975,8,10}, 8.25/4.32 \cdot F_{0.025,8,10}) \\ &\equiv (8.25/4.32 \cdot 0.23, 8.25/4.32 \cdot 3.85) \\ &\equiv (0.44, 7.36). \end{aligned}$$

Because the confidence interval includes 1 (which corresponds to the situation of equal variances), we cannot reject $H_0$ at significance level $\alpha = 0.05$.

## 11.3 One-Way Classification

In the worm/sheep example of Section 11.2.1, we were primarily concerned with comparing the worm counts in treated versus untreated lambs, represented as $\mu_1 - \mu_2$. Within the context of experimental designs, the drug administered (or lack thereof) to the lambs is considered a **factor** with two levels: **treated**, **untreated**.

As we progress through this chapter, our focus shifts to a model where the factor encompasses $a$ levels, thereby giving rise to $a$ treatments. The primary objective is to examine hypothesis testing for **equality among more than two population means**. To achieve this, we leverage a method of data analysis known as the **analysis of variance** (ANOVA).[26]

26: In essence, ANOVA can be perceived as a generalization of the customary $T-$test.

### 11.3.1 Completely Randomized Designs

In experiments where we have $a$ treatments to compare and $N$ units available for the study, a **completely randomized design** offers an efficient approach. To implement such a design:

1. decide on sample sizes $n_1, n_2, \ldots, n_a$ such that $n_1+n_2+\ldots+n_a = N$;
2. randomly allocate $n_1$ units to Treatment 1, $n_2$ units to Treatment 2, and so forth, until $n_a$ units are assigned to Treatment $a$.

In this design, the $N$ experimental units are randomly divided into $a$ groups. Taking the worm/sheep example of Section 11.2.1 as an illustration, the $N = 14$ lambs were divided **at random** into $a = 2$ groups: the treated group and the untreated group.

Alternatively, one could view the completely randomized design as drawing random samples from each of $a$ distinct populations. Each population represents a unique **level** (or treatment) of the **factor** under consideration.

Regardless of the perspective – whether through **random selection** or **random assignment** – completely randomized designs are centered around a **single factor**, which is why they are often referred to as a **one-way classification**.

The next example (modified from [3]) illustrates the basic notation.

**Example** A horticulturist is investigating the phosphorus content of tree leaves from three different varieties of apple trees (A, B and C). Random samples of five leaves from each three varieties are analyzed for phosphorus content. The observations are shown below.

| variety | sample size | phosphorus content | totals | means |
|---|---|---|---|---|
| 1 | 5 | 0.45, 0.50, 0.68, 0.60, 0.57 | 2.80 | 0.560 |
| 2 | 5 | 0.65, 0.70, 0.90, 0.84, 0.79 | 3.88 | 0.776 |
| 3 | 5 | 0.50, 0.70, 0.65, 0.63, 0.56 | 3.04 | 0.608 |

The **response variable** is the phosphorus content, the **factor** (with three levels) is the tree variety.

**Notation**

- $y_{i,j}$ is the $j$**th observation for the** $i$**th factor level** (group, class), $i = 1, \ldots, a$; $j = 1, \ldots, n_i$;
- $n_i$ is the **number of sample observations for the** $i$**th factor level**;
- the **total sample size** is

$$N = \sum_{i=1}^{a} n_i;$$

- $y_{i,\bullet}$ is the **total of the sample observations for the** $i$**th factor level**, so

$$y_{i,\bullet} = \sum_{j=1}^{n_i} y_{i,j};$$

- $y_{\bullet,\bullet}$ is the **grand total of the sample observations**, so

$$y_{\bullet,\bullet} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{i,j};$$

- $\overline{y}_{i,\bullet}$ is the **average of the sample observations for the** $i$**th factor level**, so

$$\overline{y}_{i,\bullet} = y_{i,\bullet}/n_i;$$

- $\overline{y}_{\bullet,\bullet}$ is the **average of all sample observations**, so $\overline{y}_{\bullet,\bullet} = y_{\bullet,\bullet}/N$.

In the example, we have:

- $y_{i,j}$ is the phosphorus content from leave $j$ of variety $i$, $i = 1, 2, 3$; $j = 1, \ldots, 5$;
- $n_1 = n_2 = n_3 \equiv n = 5$;
- $N = n \cdot 3 = 5 \cdot 3 = 15$;
- $y_{1,\bullet} = 2.80$, $y_{2,\bullet} = 3.88$, $y_{3,\bullet} = 3.04$;
- $y_{\bullet,\bullet} = 9.72$;
- $\overline{y}_{1,\bullet} = 0.560$, $\overline{y}_{2,\bullet} = 0.776$, $\overline{y}_{3,\bullet} = 0.608$;
- $\overline{y}_{\bullet,\bullet} = 0.648$.

### 11.3.2 One-Way Classification Model

We consider $a$ populations (**groups**, **treatments**). Initially, we address the scenario of **balanced data**, with $n_i = n = N/a$ observations for each treatment $i$.

The data can be summarized in the following manner:

- from Population 1, we gather the observations $y_{1,1}, \ldots, y_{1,n}$;
- from Population 2, we gather the observations $y_{2,1}, \ldots, y_{2,n}$;
- ...
- from Population $a$, we gather the observations $y_{a,1}, \ldots, y_{a,n}$.

For each treatment $i = 1, \ldots, a$, we assume that the observations

$$y_{i,1}, \ldots, y_{i,n} \sim \mathcal{N}(\mu_i, \sigma^2).$$

Equivalently, we can express the model as

$$y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a; \; j = 1, \ldots, n,$$

with the errors $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ being i.i.d. random variables. We also assume a **common variance** $\sigma^2$ for the $a$ populations.[27] The parameters to be **estimated** include $\mu_1, \ldots, \mu_a$, and $\sigma^2$.

27: See **homoscedasticity**, Chapter 8.

We can deduce that:

$$\mathrm{E}(y_{i,j}) = \mu_i \text{ for the } j\text{th observation in treatment group } i$$

and the variance is given by:

$$\mathrm{Var}(y_{i,j}) = \sigma^2 \quad \text{for all } i, j.$$

**Alternative Reparametrization** We can also recast the problem in a different manner:

$$\mu_i = \mu + (\mu_i - \mu) \equiv \mu + \tau_i,$$

where $\tau_i = \mu_i - \mu$ for all $i = 1, \ldots, a$. Here, $\tau_i$ represents the $i$th **treatment effect** (or treatment effect).

Given this, the **one-way classification model** can be expressed as:

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a; \; j = 1, \ldots, n,$$

where $\mu$ stands for the global (or common) mean applicable to all observations, and the error term $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. This yields an expectation of:

$$\mathrm{E}(y_{i,j}) = \mu + \tau_i.$$

The original model has $a$ parameters, specifically: $\mu_1, \ldots, \mu_a$. However, the new model presents $a + 1$ parameters: $\mu, \tau_1, \ldots, \tau_a$. This makes the model **over-parametrized**.

Addressing this, we set the constraint:

$$\sum_{i=1}^{a} \tau_i = 0.$$

It's clear that the both the **original** model and the **reparametrized** model are equivalent, provided we adhere to the constraint. This constraint enables us to express:

$$\begin{aligned} \mu_1 &= \mu + \tau_1, \\ &\vdots \\ \mu_{a-1} &= \mu + \tau_{a-1}, \\ \mu_a &= \mu - (\tau_1 + \cdots + \tau_{a-1}), \end{aligned}$$

reducing the parameter count to $a$ parameters: $\mu, \tau_1, \ldots, \tau_{a-1}$.

**Overview** Most often, the main objective in ANOVA is to determine if there are differences between the $a$ populations (or treatments). A pertinent question arises: why do we need a new procedure to compare population means when Student's $T-$test is available?

Consider an instance with $a = 3$ population means: $\mu_1, \mu_2$, and $\mu_3$. We could hypothetically test each of the **three pairs** of hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_0 : \mu_1 = \mu_3, \quad \text{and} \quad H_0 : \mu_2 = \mu_3$$

against the appropriate alternatives to identify where the differences (if any) are located.

But each test we conduct is prone to **errors** – consequently, the more tests we perform, the greater the likelihood that at least one of our conclusions will be erroneous.[28]

28: We will delve deeper into this subject at a later date.

ANOVA offers a **singular, comprehensive test** to evaluate the equality of the $a$ population means. Once we discern if a genuine difference exists among the means, we can then use a designated procedure to pinpoint the origins of these differences.

The hypothesis tests pits

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{against} \quad H_1 : \mu_i \neq \mu_j, \text{ for at least one pair } (i, j),$$

or, in an equivalent form:

$$H_0 : \tau_1 = \cdots = \tau_{a-1} = 0 \quad \text{against} \quad H_1 : \text{at least one } \tau_i \neq 0.$$

### 11.3.3 Analysis of Variance

In the analysis of variance, we focus on **partitioning the total sum of squares**, starting with the **basic decomposition**

$$y_{i,j} - \overline{y}_{\bullet,\bullet} = (\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}) + (y_{i,j} - \overline{y}_{i,\bullet}).$$

Each component of the decomposition is interpreted as follows:

- $y_{i,j} - \overline{y}_{\bullet,\bullet}$ is the **total deviation** component;
- $\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}$ is the **deviation of the estimated factor level mean around the overall mean**, and
- $y_{i,j} - \overline{y}_{i,\bullet}$ is the **deviation around the estimated factor level mean**.

We can show (see Exercises) that the sums of squares decomposition for this scenario is:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{\bullet,\bullet})^2 = n\sum_{i=1}^{a}(\overline{y}_{i,\bullet}-\overline{y}_{\bullet,\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{i,\bullet})^2,$$

or

$$\text{SST} = \text{SSA} + \text{SSE},$$

where:

- SST is the **total sum of squares**;
- SSA is the **treatment (or factor) sum of squares**, and
- SSE is the **error sum of squares**.

Given that the total sum of squares SST is **fixed**, an **increase** in SSA corresponds to a **decrease** in SSE and *vice versa*.

If all the observations within a given factor level are identical across all factor levels, then SSE = 0 and SST = SSA. Conversely, if all the estimated factor levels $\overline{y}_{i,\bullet}$ are equal, then SSA = 0 and SST = SSE.

Each sum of squares in the decomposition is associated to a **degree of freedom** (df):

- SST $\rightsquigarrow N-1$
- SSA $\rightsquigarrow a-1$
- SSE $\rightsquigarrow a(n-1) = N-a$

The decomposition's "structure" applies to the degrees of freedome:

$$N-1 = (a-1) + a(n-1) = a-1+N-a.$$

**Variance Considerations** The $i$th treatment **sample variance** is:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{i,\bullet})^2, \quad i=1,\ldots,a;$$

we know that $\text{E}(s_i^2) = \sigma^2$ and $(n-1)s_i^2/\sigma^2 \sim \chi^2_{n-1}$ for all $i = 1,\ldots,a$.

Thus, we can express SSE as

$$\text{SSE} = \sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{i,j}-\overline{y}_{i,\bullet})^2\right] = \sum_{i=1}^{a}(n-1)s_i^2 = (n-1)\sum_{i=1}^{a}s_i^2,$$

and using the typical argument related to the trace of quadratic forms, we can show that

$$\text{SSE}/\sigma^2 \sim \chi^2_{N-a}.$$

**Theorem:** The random variable

$$\text{MSE} = \frac{\text{SSE}}{N-a} = \frac{n-1}{N-a}\sum_{i=1}^{a}s_i^2 = \frac{1}{a}\sum_{i=1}^{a}s_i^2$$

is an unbiased estimator of $\sigma^2$.[29]

So, what exactly does SSA estimate?

29: This holds true regardless of whether the factor level means $\mu_i$ are equal or not. Intuitively, this is reasonable: the variability of observations within each factor level is not influenced by the magnitude of the estimated factor level means when the populations are normal.

**Theorem:** the expectation of SSA is:

$$\begin{aligned}\mathrm{E(SSA)} &= \frac{1}{N}\sum_{i=1}^{a}\left\{n\sigma^2 + [n(\mu+\tau_i)]^2\right\} - \frac{1}{an}[an\sigma^2 + (an\mu)^2] \\ &= (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2.\end{aligned}$$

If we denote the mean square due to the factor $A$ (commonly known as the **treatment mean square**) as $\mathrm{MSA} = \mathrm{SSA}/(a-1)$, then:

$$\mathrm{E(MSA)} = \sigma^2 + \frac{n}{a-1}\sum_{i=1}^{a}\tau_i^2.$$

In situations where all the factor level means are the same ($\mu_i \equiv \mu$), then we have $\tau_i^2 = (\mu_i - \mu)^2 \equiv 0$ and $\mathrm{E(MSA)} = \sigma^2$. Consequently, both MSE and MSA offer unbiased estimates of $\sigma^2$. However, when the $\mu_i$'s differ, MSA tends to be larger than MSE on average.

It can be shown (although it is beyond the scope of these notes) that:

- $\mathrm{SSA}/\sigma^2$ follows a **non-central $\chi^2$ distribution**:

$$\mathrm{SSA}/\sigma^2 \sim \chi^2_{a-1}\left(n\sum_{i=1}^{a}\tau_i^2/\sigma^2\right);$$

- the random variables SSE and SSA are **independent**.

**$F-$Test for the Equality of Treatment Means** How can we tell if the treatment means are identical?

The $F-$test pits

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{against} \quad H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i,j).$$

The test statistic to be used is

$$F_0 = \frac{\mathrm{MSA}}{\mathrm{MSE}}.$$

Large values of $F_0$ support $H_1$ since MSA will tend to exceed MSE when $H_1$ holds.[30] On the other hand, values of $F_0$ near 1 tend to support $H_0$ since both MSE and MSA have the same expected value when $H_0$ holds.[31] Hence, the appropriate test is an **upper-tail one**.

When $H_0$ holds, $\mathrm{SSE}/\sigma^2$ and $\mathrm{SSA}/\sigma^2$ are independent $\chi^2$ variables. Therefore, under $H_0$,

$$F_0 = \frac{\mathrm{SSA}/(a-1)}{\mathrm{SSE}/(N-a)} \sim F_{a-1,N-a}.$$

When $H_1$ holds, that is, when the $\mu_i$'s are not all equal, $F_0$ does not follow the customary $F$ distribution.[32]

It is thus reasonable to reject $H_0$ if we observe large values of $F_0$. Formally, we reject $H_0$ at significance level $\alpha$ if

$$F_0 > F_{\alpha;a-1,N-a}.$$

30: We have seen above that the ratio of the expected values, $\frac{\mathrm{E(MSA)}}{\mathrm{E(MSE)}}$, is greater than 1 under $H_1$.

31: Indeed, under $H_0$,

$$\frac{\mathrm{E(MSA)}}{\mathrm{E(MSE)}} = 1.$$

32: It follows instead a more complicated **non-central $F$ distribution**.

We can construct an **ANOVA table** for the $F-$test for equality of treatment means in the one-way classification scenario, based on the test statistic $F_0 = \text{MSA}/\text{MSE}$ (see Table 11.11).

| Source | SS | df | MS | $F_0$ |
|---|---|---|---|---|
| **Treatment** | SSA | $a-1$ | MSA | $F_0 = \text{MSA}/\text{MSE}$ |
| **Error** | SSE | $N-a$ | MSE | |
| **Total** | SST | $N-1$ | | |

**Table 11.11:** ANOVA table for the equality of the treatment means $\mu_i$ in the one-way classification scenario.

From a computational perspective, the following equivalent formulas are sometimes used, since they are easier to handle when we do not use software:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{i,j}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \quad \text{SSA} = \frac{1}{N}\sum_{i=1}^{a} y_{i,\bullet}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \quad \text{SSE} = \text{SST} - \text{SSA}.$$

**Example** The ANOVA table for the phosphorus dataset of the previous section can be obtained as follows in R.

First we load the data.

```
variety <- c(1,2,3)
sample.size <- c(5,5,5)
content.1 <- c(0.45, 0.50, 0.68, 0.60, 0.57)
content.2 <- c(0.65, 0.70, 0.90, 0.84, 0.79)
content.3 <- c(0.50, 0.70, 0.65, 0.63, 0.56)
content <- rbind(content.1, content.2, content.3)
data <- data.frame(cbind(sample.size, content))
rownames(data) <- variety
colnames(data) <- c("sample.size", "leaf.1", "leaf.2",
                    "leaf.3", "leaf.4", "leaf.5")
data$totals <- rowSums(content)
data$means <- data$totals/data$sample.size
data
```

```
  sample.size leaf.1 leaf.2 leaf.3 leaf.4 leaf.5 totals means
1           5   0.45    0.5   0.68   0.60   0.57   2.80 0.560
2           5   0.65    0.7   0.90   0.84   0.79   3.88 0.776
3           5   0.50    0.7   0.65   0.63   0.56   3.04 0.608
```

We compute the necessary quantities and place them in the ANOVA table.

```
a = nrow(data)
n = length(content.1)
N = a*n
grand.mean = mean(unlist(data[,c(2:(n+1))]))
SST = sum((data[,c(2:(n+1))]-grand.mean)^2)
SSA = n * sum((data$means-grand.mean)^2)
SSE = SST - SSA
```

```
ANOVA = as.data.frame(cbind(c(SSA,SSE,SST),
                        c(a-1, N-a, N-1),
                        c(SSA/(a-1),SSE/(N-a),0),
                        c((SSA/(a-1))/(SSE/(N-a)),0,0)))
rownames(ANOVA) = c("Treatment", "Error", "Total")
colnames(ANOVA) = c("SS", "df", "MS", "F0")
ANOVA
```

```
              SS df      MS       F0
Treatment 0.12864  2 0.06432 7.892025
Error     0.09780 12 0.00815
Total     0.22644 14
```

At significance level $\alpha = 0.05$, the critical value of $F_{2,12}$ is:

```
alpha=0.05
qf(p=1-alpha, df1 = a-1, df2 = N-a)
```

```
[1] 3.885294
```

Since $7.89 = F_0 > F_{0.05;2,12} = 3.89$, we reject $H_0$ at $\alpha = 0.05$ and we conclude that the mean phosphorus content is unlikely to be the same for all $a = 3$ varieties of trees. □

### 11.3.4 Estimation of Model Parameters

Recall that in the one-way classification model, $a + 1$ parameters require estimation, namely $\mu, \tau_1, \ldots, \tau_a$. We use the **least square estimation principle** to find them based on the observed data.

The sum of squares is defined as

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{i,j} - \mu - \tau_i)^2.$$

We find $\hat{\mu}$ and $\hat{\tau}_i$ that minimize $L$ by differentiating $L$ with respect to $\mu$ and $\tau_i$, $i = 1, \ldots, a$, and setting to 0. This yields the **normal equations**:

$$\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{i,j} - \hat{\mu} - \hat{\tau}_i) = 0 \quad (\mu\text{-equation}),$$
$$\sum_{j=1}^{n} (y_{i,j} - \hat{\mu} - \hat{\tau}_i) = 0 \quad (\tau_i\text{-equation}, i = 1, \ldots, a).$$

The corresponding system of linear equations is:

$$\begin{aligned} N\hat{\mu} + n \sum_{i=1}^{a} \hat{\tau}_i &= y_{\bullet,\bullet}, \\ n\hat{\mu} + n\hat{\tau}_1 &= y_{1,\bullet}, \\ &\vdots \\ n\hat{\mu} + n\hat{\tau}_a &= y_{a,\bullet}. \end{aligned}$$

Given the constraint $\tau_1 + \cdots + \tau_a = 0$, the solution is:

$$\begin{aligned}\hat{\mu} &= \overline{y}_{\bullet,\bullet},\\ \hat{\tau}_i &= \overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet} \quad \text{for } i = 1, \ldots, a.\end{aligned}$$

Thus, the **estimated treatment effect** for the $i$th treatment is

$$\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i,\bullet};$$

the **difference between treatments** $i$ and $j$ is given by

$$\hat{\mu}_i - \hat{\mu}_j = \overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}.$$

Using the **pooled estimate** MSE as an estimator of $\sigma^2$, we can exhibit a $100(1-\alpha)\%$ confidence interval for $\mu_i$ *via*:

$$\overline{y}_{i,\bullet} \pm t_{\alpha/2;N-a}\sqrt{\frac{\text{MSE}}{n}};$$

for $\mu_i - \mu_j$, we have instead:

$$(\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}) \pm t_{\alpha/2;N-a}\sqrt{\frac{2\text{MSE}}{n}}.$$

### 11.3.5 Unbalanced Designs

We could also opt for an **unbalanced design**, in which the number of observations $n_i$ we sample in each treatment group $i$ is not necessarily the same from one group to the other. However, a balanced design has several advantages.[33]

33: First and foremost, the theoretical derivations are simpler to obtain in the balanced case.

In particular, the power of the $F-$test is **larger** with balanced data. Indeed for $a = 2$ (two treatments), we can show that the power of the $F-$test is maximized when $\frac{1}{n} + \frac{1}{N-n}$ is minimized (see Exercises); if $N$ is even and fixed, the minimum is thus achieved when $n = N/2$.

Moreover, the $F$ test is only **robust against unequal variances** when data is balanced. For the case of $a = 2$ treatments, the $F$ test is equivalent to the Student's $T-$test, with $F_0 = t_0^2$.

If we define $\theta$ as the ratio $\sigma_1^2/\sigma_2^2$ and $R$ as the ratio $n_1/n_2$, the Student's $T-$test can be expressed as:

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}\left(\frac{1}{s_p^2}\cdot\frac{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}{\frac{1}{n_1} + \frac{1}{n_2}}\right)^{1/2}.$$

When $n_1, n_2 \to \infty$,

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \to \sigma_2^2\frac{1+R\theta}{1+R}.$$

Consquently, $t_0 \to \mathcal{N}(0, (R+\theta)/(1+R\theta))$, and $(R+\theta)/(1+R\theta) = 1$ when $R = 1$, regardless of $\theta$'s value.

For unbalanced data, the sum of squares formulas must be modified:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{i,j}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \quad \text{SSA} = \sum_{i=1}^{a}\frac{y_{i,\bullet}^2}{n_i} - \frac{y_{\bullet,\bullet}^2}{N}.$$

Finally, when we estimate the model parameters, we solve the **normal equations** subject to the constraint:

$$\sum_{i=1}^{a} n_i\hat{\tau}_i = 0$$

as opposed to the constraint $\sum_{i=1}^{a}\hat{\tau}_i = 0$.[34]

34: If we note that the latter can also be written as $n\tau_1 + \cdots + n\tau_a = 0$ in the balanced case, we see that it is simply a special instance of the unbalanced case. The same comment applies to the modified formula for SSA.

### 11.3.6 Contrasts

The analysis of variance can tell us an indication that not all the treatment groups have the same mean response, but an ANOVA does not, by itself, provide information about **which treatments** are different or in what ways they differ.

To get answers to these questions, we must examine the **treatment means**, or equivalently, the **treatment effects**. We can do so through **contrasts**, which enable us to focus in on **specific** (narrow) features of the data.[35]

35: In fact, a single contrast's focus is so narrow that it may obscure the overall picture.

By using several contrasts, we can move the focus around and explore more features. Intelligent use of contrasts involves choosing the contrasts so that they highlight interesting data features.[36]

36: But that's easier said than done without a solid understanding of the domain under study, which can be improved *via* data exploration, among others (see Chapter 18 for more information).

**Linear Contrasts** Linear combinations of the treatment effects $\mu_i$

$$C = \sum_{i=1}^{a} c_i\mu_i, \quad \text{where } \sum_{i=1}^{a} c_i = 0 \text{ with } c_i \in \mathbb{R}.$$

are called **linear contrasts**; in general, we are interested in testing for

$$H_0 : C = 0 \quad \text{against} \quad H_1 : C \neq 0.$$

When there are $a$ treatment effects, we sometimes identify the linear contrast $C$ with its **signature vector** $(c_1, \ldots, c_a)$.

#### Examples

1. Suppose that we wish to test for

   $$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \mu_1 \neq \mu_2;$$

   we must then work with the linear contrast $(1, -1, 0, \ldots, 0)$.[37]

   37: The linear contrast $(1, -1, 0, \ldots, 0)$ also does the trick, being equivalent to the one in the text when it comes to hypothesis testing.

2. Suppose that we wish to test for

   $$H_0 : \tfrac{1}{2}(\mu_1 + \mu_2) = \tfrac{1}{2}(\mu_3 + \mu_4) \quad \text{against} \quad H_1 : \tfrac{1}{2}(\mu_1 + \mu_2) \neq \tfrac{1}{2}(\mu_3 + \mu_4);$$

   we work with the linear contrast $(1/2, 1/2, -1/2, -1/2, 0, \ldots, 0)$.

To test a contrast hypothesis, we start by estimating $C$ using

$$\widehat{C} = \sum_{i=1}^{a} c_i \overline{y}_{i,\bullet}.$$

Assume a balanced design; if the observations are obtained independently, we have $\text{Cov}(\overline{y}_{i_1,\bullet}, \overline{y}_{i_2,\bullet}) = 0$ if $i_1 = i_2$, and $\text{Var}(\overline{y}_{i,\bullet}) = \sigma^2/n$ for all $i$, so

$$\text{Var}(\widehat{C}) = \sum_{i=1}^{a} c_i^2 \text{Var}(\overline{y}_{i,\bullet}) = \frac{\sigma^2}{n} \sum_{i=1}^{a} c_i^2.$$

We obtain an **estimator** of $\text{Var}(\widehat{C})$ *via*:

$$\widehat{\text{Var}}(\widehat{C}) = \frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2.$$

It follows that the **test statistic** is given by

$$t_0 = \frac{\widehat{C}}{\sqrt{\frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2}}.$$

We can show that $t_0 \sim t_{N-a}$; therefore, we reject $H_0$ at significance level $\alpha$ if $|t_0| > t_{\alpha/2;N-a}$.

Instead of the $T-$test, however, we could use the **equivalent** $F-$test, with test statistic $F_0 = \frac{\text{SSC}}{\text{MSE}}$, which rejects $H_0$ at significance level $\alpha$ if $F_0 > F_{\alpha;1,N-a}$, where

$$\text{SSC} = \left( \sum_{i=1}^{a} c_i \overline{y}_{i,\bullet} \right)^2 \Bigg/ \sum_{i=1}^{a} c_i^2/n.$$

We can build a $100(1-\alpha)\%$ **confidence interval** for $C$ is given by

$$\widehat{C} \pm t_{\alpha/2;N-a} \sqrt{\frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2}.$$

**Example** In the phosphorus dataset, suppose we want to test

$$H_0 : \mu_2 = \frac{\mu_1 + \mu_3}{2} \quad \text{against} \quad H_1 : \mu_2 \neq \frac{\mu_1 + \mu_3}{2}.$$

This is a contrast with $c_1 = -1/2$, $c_2 = 1$ and $c_3 = -1/2$.

The test statistics is given by

$$t_0 = \frac{(-1/2) \cdot 0.560 + 1 \cdot 0.776 + (-1/2) \cdot 0.608}{\sqrt{\left(\frac{0.00815/12}{5}\right) \{(-1/2)^2 + 1^2 + (-1/2)^2\}}} = \frac{0.192}{0.0142741} = 13.45094.$$

Since $|t_0| > t_{0.025,12} = 2.17881$, we reject $H_0$ and we conclude that there is enough evidence to conclude that $\mu_2$ is different from the average of $\mu_1$ and $\mu_3$. □

**Orthogonal Contrasts** Two contrasts with coefficients $\{c_i\}$ and $\{d_i\}$ are **orthogonal** if $c_1d_1+\cdots+c_ad_a = 0$. For instance, the contrasts $-2\mu_1+\mu_2+\mu_3$ and $\mu_3 - \mu_2$ are orthogonal since

$$(-2)(0) + (1)(-1) + (1)(1) = 0.$$

If there are $a$ treatments, we can find a set of $a - 1$ contrasts that are mutually orthogonal, that is, each one is orthogonal to all of the others. With 5 treatments (say), we can define 4 mutually orthogonal contrasts:

$$\begin{array}{lccccc} C_1 = & & & \mu_4 & -\mu_5 \\ C_2 = \mu_1 & & +\mu_3 & -\mu_4 & -\mu_5 \\ C_3 = \mu_1 & & -\mu_3 & & \\ C_4 = \mu_1 & -4\mu_2 & +\mu_3 & +\mu_4 & +\mu_5 \end{array}$$

The important feature of orthogonal contrasts is that they are **independent** (as random variables).[38]

38: An additional useful fact is that they **partition** the treatment sum of squares:

$$\text{SSA} = \sum_{i=1}^{a-1} \text{SSC}_i.$$

In other words, if we compute the sums of squares for a full set of orthogonal contrasts ($a-1$ contrasts for $a$ groups), adding up those $a - 1$ sums of squares yields exactly the treatment sum of squares, which also has $a - 1$ degrees of freedom.

### 11.3.7 Multiple Comparisons

We have discussed **multiple hypothesis testing** in Section 8.2.3; how does it apply to design of experiments?

**Example** Suppose we want to compare four treatments, so $a = 4$. We may want to compare all the pairs

$$\begin{array}{lll} H_0 : \mu_1 = \mu_2, & H_0 : \mu_1 = \mu_3, & H_0 : \mu_1 = \mu_4, \\ H_0 : \mu_2 = \mu_3, & H_0 : \mu_2 = \mu_4, & H_0 : \mu_3 = \mu_4. \end{array}$$

Overall, there we have $k = 6$ possible tests of the form $H_0 : \mu_i = \mu_j$ against some fixed alternative type. □

In general, suppose that we wish to conduct $k$ hypothesis tests. If the level of each individual test is $\alpha$, then the overall error rate is likely to be **(much) larger** than $\alpha$.

As an illustration, suppose that we conduct $k = 2$ tests, each one at significance level 5% .[39] Then, the probability of rejecting at least one of the null hypotheses when they are both true will be higher than 5%.

39: That is, the probability of a Type 1 error is 5% for each test separately.

Indeed, let $E_j$ be the event that we reject the null hypothesis for the $j$th test, $j = 1, 2$. Then,

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &= 0.05 + 0.05 - P(E_1 \cap E_2) = 0.1 - P(E_1 \cap E_2). \end{aligned}$$

As $0 \le P(E_1 \cap E_2) \le 0.5$, the probability of making at least one mistake is now between 5% and 10%.[40]

40: If the events are independent, then $0 < P(E_1 \cap E_2)$, and $P(E_1 \cup E_2) > 0.05$.

We can extend this argument to the general case of $k$ tests. Suppose that the $k$ null hypotheses $H_0$ are true. Once again, let's define $E_j$ as the event that we reject the null hypothesis for the $j$th test, $j = 1, \cdots, k$.

**Boole's inequality** states that

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq \sum_{j=1}^{k} P(E_j) = \sum_{j=1}^{k} \alpha_j,$$

where $\alpha_j$ denotes the **probability of Type 1 error associated with the $j$th test**. If $\alpha_j = \alpha$ then

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq k\alpha.$$

For instance, for $k = 10$, and $\alpha_j = 0.05$ for all $j$, then the best that we can say is that

$$P\left(E_1 \cup \cdots \cup E_{10}\right) \leq 0.5.$$

**Conclusion:** the level of significance of a **family of tests** may differ from that of an **individual test**.

We use **multiple comparison procedures** to conduct multiple inference while **controlling the overall error rate**. The rationale behind these procedures is simple – we seek to ensure a **global significance level** below (or at) $\alpha$. More specifically, we seek a procedure for which the probability of rejecting at least one the null hypotheses when they are all true is not larger than $\alpha$.

Several procedures have been proposed in the literature, including:

- Bonferroni's (1936);
- Tukey's (1949);
- Scheffé's (1959).

**Bonferroni's Procedure** When investigating a particular set of $k$ pairwise comparisons and/or contrasts, it is essential to specify the family of interest **in advance**. The **Bonferroni procedure** is versatile, and applicable whether the $n_i$'s are equal or unequal and irrespective of whether the focus is on pairwise comparisons, contrasts, or a mix of both.

Instead of conducting each of the $k$ tests at the usual $\alpha$ level, we conduct each test at the $\alpha/k$ level. With this adjustment, the probability of making at least one Type I error across all $k$ tests is bounded by $\alpha$:

$$P\left(\bigcup_{j=1}^{k} E_j\right) \leq \sum_{j=1}^{k} \frac{\alpha}{k} = k\left(\frac{\alpha}{k}\right) = \alpha.$$

For example, for an analysis involving 10 tests with an intended overall error rate of $\alpha = 0.05$, the Bonferroni correction would adjust the significance level for each test to $0.05/10 = 0.005$.

This method can also be extended to the construction of **simultaneous confidence intervals**. If we denote by $\text{C.I.}_1, \cdots, \text{C.I.}_k$ the associated confidence intervals, each constructed at a **coverage level** of $1 - \alpha$,[41] then the probability that all $k$ intervals simultaneously contain their true parameter values is bounded above by $1 - \alpha$:

$$P\left(\bigcap_{j=1}^{k} E_j\right) \leq 1 - \alpha.$$

41: That is,

$$P(C_j \in \text{C.I.}_j) = 1 - \alpha, \quad j = 1, \cdots, k,$$

where $C_j$ is the true value of the $j$th parameter or contrast of interest.

However, with Bonferroni's adjustment, if each interval is constructed to have a coverage probability of $1-\alpha/k$, then the **joint coverage probability** is at least $1-\alpha$:

$$P\left(\bigcap_{j=1}^{k} E_j\right) = 1 - P\left(\bigcup_{j=1}^{k} E_j^c\right) \geq 1 - \sum_{j=1}^{k} P(E_j^c) = 1 - \sum_{j=1}^{k} \frac{\alpha}{k} = 1 - \alpha.$$

An undoubted **advantage** of the Bonferroni method lies in its **generality**: it is applicable to a wide range of probability-based inferences across various distributions, not merely confidence intervals within a normal linear model.

But this method is not without its **drawbacks**. Chief among them being that for larger values of $k$, the individual significance level for each test can become **exceedingly stringent**.

With an overall error rate of $\alpha = 5\%$ and $k = 10$, say, the significance level for each test under Bonferroni's method is $1 - \alpha/k = 0.995$. This means each confidence interval might be so wide that its **practical utility** diminishes.[42]

42: In such scenarios, one might consider increasing the overall (joint) error rate, perhaps to 10%, to make the results **more easily interpretable**.

**Tukey's Procedure** The **Tukey multiple comparison procedure** is particularly valuable when our focus is on analyzing the set of **all pairwise comparisons** of **factor level means**. Specifically, when utilizing this method, the primary interest revolves around the tests defined by:

$$H_0 : \mu_i = \mu_j \quad \text{against} \quad H_1 : \mu_i \neq \mu_j.$$

When all sample sizes are balanced, the family confidence coefficient for the Tukey method aligns precisely with $1 - \alpha$, ensuring the family significance level is consistent with $\alpha$. However, for unbalanced data, where sample sizes diverge, the Tukey procedure exhibits a **conservative behaviour**. This results in the family confidence coefficient surpassing $1 - \alpha$, and subsequently, the family significance level falling below $\alpha$.

A key component of the Tukey procedure is the use of the **Studentized range distribution**. Given a set of i.i.d. random variables $y_1, \ldots, y_k \sim \mathcal{N}(\mu, \sigma^2)$, their **range** $R$ is defined as:

$$R = \max\{y_1, \ldots, y_k\} - \min\{y_1, \ldots, y_k\}.$$

If $s^2$ be an estimator of $\sigma^2$ independent of $R$, and assume that $\frac{\upsilon s^2}{\sigma^2} \sim \chi^2_\upsilon$. Then the variable $\frac{R}{s}$ follows a **Studentized range distribution** $q_{k,\upsilon}$. Let $q_{\alpha;k,\upsilon}$ be the critical value for which

$$P\left(\frac{R}{s} > q_{\alpha;k,\upsilon}\right) = \alpha.$$

**Theorem:** suppose we have $a$ means, $\overline{y}_{1,\bullet}, \ldots, \overline{y}_{a,\bullet}$, obtained from $a$ independent normal samples, each of size $n$, with respective means $\mu_1, \ldots, \mu_a$ and a shared variance $\sigma^2$.[43]

43: That is, $\overline{y}_{i,\bullet} \sim \mathcal{N}(\mu_i, \sigma^2/n)$, for all $i$.

We know that MSE is an unbiased estimator of $\sigma^2$ independent of $R$ and

$$\frac{(N-a)\text{MSE}}{\sigma^2} \sim \chi^2_{N-a}.$$

Under these conditions, the simultaneous probability for all pairwise comparisons is:

$$(\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}) - q_{\alpha;a,N-a}\sqrt{\frac{\text{MSE}}{n}} < \mu_i - \mu_j < (\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}) + q_{\alpha;a,N-a}\sqrt{\frac{\text{MSE}}{n}}.$$

The family confidence coefficient $1-\alpha$ pertaining to the multiple pairwise comparisons refers to the **proportion of correct families**, each consisting of all pairwise comparisons, when repeated sets of samples are selected and all pairwise confidence intervals are calculated each time.[44]

44: A family of pairwise comparisons is considered to be correct if **every pairwise comparison** in the family is correct.

This family confidence coefficient implies that, across repeated sampling, all pairwise comparisons in the family will be accurate in $100(1-\alpha)\%$ of the instances.

Transitioning our focus to **simultaneous testing**, the objective is to conduct a comprehensive set of tests that pit

$$H_0 : \mu_i = \mu_j \quad \text{against} \quad H_1 : \mu_i \neq \mu_j$$

for all potential pairwise comparisons. The pivotal test statistic in this context is:

$$q_0 = \frac{\overline{y}_{i,\bullet} - \overline{y}_{j,\bullet}}{\sqrt{\text{MSE}/n}}.$$

We reject $H_0$ at significance level $\alpha$ if $|q_0| \geq q_{\alpha;a,N-a}$.[45]

45: Selected percentiles for the Studentized range distribution can be found in tables, such as on this page. In R, we can use the functions `qtukey()` and `ptukey()`.

We illustrate the procedure with the help of a classical example.[46]

46: See here, for instance.

**Example** In a study of the effectiveness of different rust inhibitors, four brands (A, E, C, D) were tested. Altogether, 40 experimental units were randomly assigned to the four brands, with 10 units assigned to each brand. The results obtained after exposing the experimental units to severe weather conditions are given below.[47]

47: The higher the value, the more effective the rust inhibitor.

| | **Rust Inhibitor Brand** | | | |
|---|---|---|---|---|
| | A | B | C | D |
| | $i=1$ | $i=2$ | $i=3$ | $i=4$ |
| $\overline{y}_{i,\bullet}$ | 43.14 | 89.44 | 67.95 | 40.47 |
| $\overline{y}_{\bullet,\bullet} = 60.25$ | | | | |
| MSE = 6.14 | | | | |

This study is a completely randomized design, where the levels of the single factor correspond to the four rust inhibitor brands. Suppose we are interested in all pairwise comparisons, which we evaluate *via* the Tukey procedure.

The important parameters are loaded below.

```
a = 4; N = 40; n = 10; alpha = 0.05
y.bar.1 = 43.14; y.bar.2 = 89.44
y.bar.3 = 67.95; y.bar.4 = 40.47
y.bar = 60.25; MSE = 6.14
(q.crit = qtukey(alpha, a, N-a, lower.tail = FALSE))
B = q.crit*sqrt(MSE/n)
```

```
[1] 3.808798
```

The 6 confidence intervals (with corresponding test statitsics) are computed as follows.

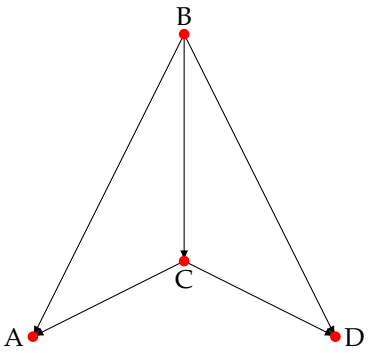
```
ci.2.1 = y.bar.2 - y.bar.1 +B*c(-1,1); q0.2.1 = (y.bar.2 - y.bar.1)/sqrt(MSE/n)
ci.3.1 = y.bar.3 - y.bar.1 +B*c(-1,1); q0.3.1 = (y.bar.3 - y.bar.1)/sqrt(MSE/n)
ci.4.1 = y.bar.4 - y.bar.1 +B*c(-1,1); q0.4.1 = (y.bar.4 - y.bar.1)/sqrt(MSE/n)
ci.3.2 = y.bar.3 - y.bar.2 +B*c(-1,1); q0.3.2 = (y.bar.3 - y.bar.2)/sqrt(MSE/n)
ci.4.2 = y.bar.4 - y.bar.2 +B*c(-1,1); q0.4.2 = (y.bar.4 - y.bar.2)/sqrt(MSE/n)
ci.4.3 = y.bar.4 - y.bar.3 +B*c(-1,1); q0.4.3 = (y.bar.4 - y.bar.3)/sqrt(MSE/n)
```

The simultaneous confidence intervals and tests for pairwise differences are shown in the table below.

| **Confidence Interval** | **Test** | | |
|---|---|---|---|
| | $H_0$ | $H_1$ | $q_0$ |
| $43.3 < \mu_2 - \mu_1 < 49.3$ | $\mu_2 = \mu_1$ | $\mu_2 \neq \mu_1$ | 58.99 |
| $21.8 < \mu_3 - \mu_1 < 27.8$ | $\mu_3 = \mu_1$ | $\mu_3 \neq \mu_1$ | 31.61 |
| $-0.3 < \mu_4 - \mu_1 < 5.7$ | $\mu_1 = \mu_4$ | $\mu_1 \neq \mu_4$ | 3.40 |
| $18.5 < \mu_2 - \mu_3 < 24.5$ | $\mu_2 = \mu_3$ | $\mu_2 \neq \mu_3$ | 27.37 |
| $46.0 < \mu_2 - \mu_4 < 52.0$ | $\mu_2 = \mu_4$ | $\mu_2 \neq \mu_4$ | 62.39 |
| $24.5 < \mu_3 - \mu_4 < 30.5$ | $\mu_3 = \mu_4$ | $\mu_3 \neq \mu_4$ | 35.01 |

Only in the comparison between A and D does the confidence interval include 0. Therefore, there is no clear evidence that either D or A is the better rust inhibitor. For the other pairs, we conclude that there is a difference in performance:

> B $\geq$ A, C $\geq$ A, B $\geq$ C, B $\geq$ D, C $\geq$ D (see the diagram format below).




We obtain the same conclusions if we look at the test statistics, and compare their absolute value to $q_{0.05;4,36} = 3.814$ – except for A and D, all differences are found to be statistically significant. □

**Scheffé's Procedure** The family of interest refers to the set of **all possible contrasts** among the factor level means:

$$C = \sum_{i=1}^{a} c_i \mu_i, \quad \text{where } \sum_{i=1}^{a} c_i = 0, \quad \text{with } c_i \in \mathbb{R}.$$

In essence, the family is comprised of estimates of all possible contrasts $C$ or tests concerning all possible contrasts of the type:

$$H_0 : C = 0 \quad \text{versus} \quad H_1 : C \neq 0;$$

thus, the family consists of **infinitely many statements**.

The confidence level for the **Scheffé procedure** for the entire family is exactly $1 - \alpha$, regardless of whether the design is balanced or unbalanced.

Recall that

$$C = \sum_{i=1}^{a} c_i \mu_i$$

is estimated by

$$\widehat{C} = \sum_{i=1}^{a} c_i \overline{y}_{i,\bullet},$$

and that the variance of this estimate is

$$\widehat{\text{Var}}(\widehat{C}) = \frac{\text{MSE}}{n} \sum_{i=1}^{a} c_i^2.$$

For simultaneous estimation through confidence intervals, the **Scheffé confidence intervals for the family of contrasts** $C$ take the form:

$$\widehat{C} - W\sqrt{\widehat{\text{Var}}(\widehat{C})} < C < \widehat{C} + W\sqrt{\widehat{\text{Var}}(\widehat{C})},$$

where $W^2 = (a-1)F_{\alpha;a-1,N-a}$.[48]

48: See the justification for the Working-Hostelling test in Section 8.2.3 for an indication of how to prove this statement.

If we were to compute the confidence intervals for every conceivable contrast, then we would expect that the entire set of confidence intervals in the family would be accurate in roughly $100(1-\alpha)\%$ of the experimental repetitions. Note that the simultaneous confidence limits differ from those for a single confidence limit solely in terms of the **estimated standard deviation multiple** in front of the square root.

Considering the problem of **simultaneous testing**, we are interested in tests of the form:

$$H_0^C : C = 0 \quad \text{versus} \quad H_1^C : C \neq 0.$$

The corresponding test statistics are

$$F_0 = \frac{\widehat{C}^2}{(a-1)\widehat{\text{Var}}(\widehat{C})},$$

and we reject the specific test $H_0^C$ if $F_0 > F_{\alpha;a-1,N-a}$.[49]

49: Given that applications of the Scheffé procedure never involve all conceivable contrasts, the confidence coefficient for the finite family of statements under consideration will exceed $1 - \alpha$. Thus, $1 - \alpha$ acts as a **guaranteed lower bound**. In a similar vein, the significance level for the finite family of tests will be below $\alpha$.

The following example is found in [1].

**Example** The Kenton Food Company tested four different package designs for a new breakfast cereal. Twenty stores were selected as the experimental units. Each store was randomly assigned one of the package designs, with each package design assigned to five stores. A fire occurred in one store during the study period, so this store was dropped from the study. Hence, one of the designs was tested in only four stores.

The stores were chosen to be comparable in location and sales volume. Other relevant conditions that could affect sales, such as price, amount and location of shelf space, and special promotional efforts, were kept the same for all of the stores in the experiment.

Sales were observed for the study period; the results are recorded below.

| | **Package Design ($i$)** | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | **Total** |
| $n_i$ | 5 | 5 | 4 | 5 | **19** |
| $y_{i,\bullet}$ | 73 | 67 | 78 | 136 | **354** |
| $\overline{y}_{i,\bullet}$ | 14.6 | 13.4 | 19.5 | 27.2 | **18.63** |

This study is a completely randomized unbalanced design with package type as the single, four-level factor.

For what it is worth, the package types had the following characteristics

- Package 1: 3-colour design, with a cartoon character;
- Package 2: 3-colour design, without a cartoon character;
- Package 3: 5-colour design, with a cartoon character;
- Package 4: 5-colour design, without a cartoon character.

The one-way classification ANOVA table for the observed data is:

| **Source** | **SS** | **df** | **MS** | **$F_0$** |
|---|---|---|---|---|
| **Treatment** | 588.2 | 3 | 196.07 | 18.585 |
| **Error** | 158.2 | 15 | 10.55 | |
| **Total** | 746.42 | 8 | | |

We are interested in estimating the following 4 contrasts with family confidence coefficient 0.90:

$$
\begin{aligned}
C_1 &= \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2} \\
C_2 &= \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2} \\
C_3 &= \mu_1 - \mu_2 \\
C_4 &= \mu_3 - \mu_4.
\end{aligned}
$$

We can compute the coefficient $W$ for significance level $\alpha = 0.1$.

```
a=4; alpha=0.1; N=19;
(W = sqrt((a-1)*qf(alpha, df1=a-1, df2=N-a, lower.tail=FALSE)))
```

```
[1] 2.733014
```

We can easily compute the estimated contrasts.

```
y.bar.1 = 14.6; y.bar.2 = 13.4; y.bar.3 = 19.5; y.bar.4 = 27.2
C.hat.1 = (y.bar.1 + y.bar.2)/2 - (y.bar.3 + y.bar.4)/2
C.hat.2 = (y.bar.1 + y.bar.3)/2 - (y.bar.2 + y.bar.4)/2
C.hat.3 = y.bar.1 - y.bar.2
C.hat.4 = y.bar.3 - y.bar.4
```

The designed is unbalanced, so $n$ is not constant. For the purposes of this exercise, we use the average value of the $n_i$ for $n$. Moreover, we can read the value of MSE from the ANOVA table.

```
n = mean(c(5,5,4,5)); MSE = 10.55
```

We now compute the variance of the contrasts.

```
sum.c2.1 = 4*(1/2)^2; sum.c2.2 = 4*(1/2)^2
sum.c2.3 = 2*(1)^2; sum.c2.4 = 2*(1)^2
B.1 = sqrt(MSE/n*sum.c2.1); B.2 = sqrt(MSE/n*sum.c2.2)
B.3 = sqrt(MSE/n*sum.c2.3); B.4 = sqrt(MSE/n*sum.c2.4)
```

We are now able to obtain the joint 90% confidence intervals for the contrasts.

```
C.hat.1 + W*B.1*c(-1,1)
C.hat.2 + W*B.2*c(-1,1)
C.hat.3 + W*B.3*c(-1,1)
C.hat.4 + W*B.4*c(-1,1)
```

```
[1] -13.423064  -5.276936
[1] -7.3230638  0.8230638
[1] -4.560182  6.960182
[1] -13.460182  -1.939818
```

Note that the confidence interval for $C_1$ does not include 0. Hence, if we wished to test $H_0 : C_1 = 0$ versus $H_1 : C_1 \neq 0$ at 90% confidence (among 3 other contrasts), we would reject $H_0$ in favour of $H_1$, namely that the mean sales for the 3-colour and 5-colour designs differ.

The confidence interval provides additional information, however; the mean sales for the 5-colour designs exceed the mean sales for the 3-colour designs, by somewhere between 5.3 and 13.4 cases per store.

Using the other contrasts, the sales manager also concluded that no overall effect of cartoon characters in the package design is indicated by the data, although the use of a cartoon character in the 5-colour designs is associated with lower mean sales than when no cartoon character is used.[50] □

50: Is the link necessarily causal?

**Bonferroni vs. Tukey vs. Scheffé** If **all pairwise comparisons** are of interest, the Tukey procedure is superior to the Bonferroni and Scheffé procedures, leading to narrower confidence intervals. If **not all pairwise comparisons** are to be considered, the Bonferroni procedure may be prove to be a better choice (at times).

The Bonferroni procedure yields **tighter** confidence intervals than Scheffé's when the number of contrasts of interest is **about the same** as (or is **smaller than**) the **number of factor levels**. Indeed, the number of contrasts of interest must exceed the number of factor levels **by a considerable amount** before the Scheffé procedure becomes a better choice.

All three procedures are of the form

$$\text{Estimator} \pm \text{Multiplier} \cdot \text{SE}.$$

The only difference among the three procedures is the **multiplier**. In any given problem, one may then compute the Bonferroni and Scheffé multipliers (and, when appropriate, the Tukey multiplier), and select the smallest option.[51]

51: This is an appropriate choice because the multiplier does not depend on the observed data, only on the structure of the design and the desierd joint signficance level.

### 11.3.8 Model Validation

In our analysis of experimental results, we have primarily compared the average responses across various treatment groups. These comparisons have been conducted using an **overarall ANOVA test** or more targeted procedures based on **contrasts** and **pairwise comparisons**.

The foundation of these methods rests on the assumption that the data follows the model

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \dots, a; j = 1, \dots, n,$$

where $\mu$ symbolizes the global mean applicable to all observations, and $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Note that the designed is assumed to be **balanced**.

The $\tau_i$'s are fixed but unknown coefficients, whereas the $\varepsilon_{i,j}$'s are independent normal random variables with constant but potentially unknown variance $\sigma^2$.

At its core, a model is essentially a **set of assumptions** – but we have done nothing so far to verify if (or ensure that) these assumptions are **reasonable**.

Specifically, we must verify three primary assumptions about the errors:

1. they are independent;
2. they are normally distributed, and
3. they have constant variance.

The model's analytical rigour and the consequent inferences largely depend on the extent to which the errors $\varepsilon_{i,j}$, adhere to these assumptions. Unfortunately, we **never observe the true errors** $\varepsilon_{i,j}$; the most accurate representation we possess for them are the **residuals** $e_{i,j}$, derived from the full model.

**Validation** must then be based on these **observable residuals** rather than the genuine errors. Due to the indirect nature of this process, diagnostics are sometimes **complicated**.

In any practical data set, it's almost inevitable that we encounter violations of **one or more** of these core assumptions. But there is reason for optimism: even in the face of **slight deviations**, the procedures can still yield **reasonable inferences**.

We now delve deeper into **diagnostics** and possible **remedial measures** for scenarios where the model assumptions are not met.

**Residuals** The (unobservable) errors are given by

$$\varepsilon_{i,j} = y_{i,j} - \mu - \tau_i.$$

After the model parameters have been estimated, we can compute the **residuals**

$$e_{i,j} = y_{i,j} - \hat{\mu} - \hat{\tau}_i = y_{i,j} - \hat{y}_{i,j} = y_{i,j} - \overline{y}_{i,\bullet}.$$

These residuals are often referred to as **raw residuals**.

The **error sum of squares** is simply

$$\text{SSE} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} e_i^2,$$

and the mean square error is

$$\text{MSE} = \frac{\text{SSE}}{N-a}, \quad \text{where } N = n_1 + \cdots + n_a.$$

At times, we may also use the **Studentized residuals**

$$d_{i,j} = \frac{e_{i,j}}{\sqrt{\text{MSE}}},$$

which we have discussed in Section 8.3.5.

**Assessing Non-Normality** The $qq-$plot, also known as the **normal probability plot**, is used to determine if the **errors align with a normal distribution**. The assessment is made by comparing the **observed quantiles** of the residuals with the **expected quantiles** from a normal distribution.

A **straight line** is indicative of errors following a normal distribution, albeit slight deviations at the tails are customary (and anticipated).[52] For **non-normal data**, the curvature of the plot provides insights into how the data varies from the normal distribution.

In the context of $qq-$plots, the choice between raw residuals and Studentized residuals is generally inconsequential.

52: See Section 8.3.5 for description and examples.

**Assessing Non-Constant Variance** We look for non-constant variance occurring when the responses within a treatment group all have the same variance $\sigma_i^2$, but the $\sigma_i^2$ differ between different groups.

This can be assessed visually by plotting the residuals, either $e_{i,j}$ or $d_{i,j}$, against the fitted values $\widehat{y}_{i,j}$. With constant variance, the **vertical dispersion** observed within the stripes of this plot remains **fairly consistent**; any **discernible pattern** in the residuals signals **non-constant variance**.

The most common deviations from constant variance are those where the residual variation depends on the mean. Usually we see variances increasing as the mean increases, but other patterns can occur.

**Assessing Independence** Serial dependence, also known as **autocorrelation**, is a common deviation from the assumption of independence in data analysis. This phenomenon emerges when consecutive data points, particularly those in **close temporal proximity**, exhibit excessive similarity (indicating **positive dependence**) or marked dissimilarity (suggesting **negative dependence**). Among these, positive dependence is the more prevalent form.

To visually discern the presence of serial dependence, analysts frequently use an **index plot**, which plots residuals on the vertical axis against their temporal sequence on the horizontal axis. By examining this plot, one can gauge the **degree of dependence**.

A **discernible drift** across the plot, for instance, is indicative of positive dependence. On the other hand, residuals **rapidly alternating** between positive and negative values, all the while centering around zero, typically suggest negative dependence.

**Remedial Measures** **Non-normality** and **non-constant variance** can sometimes be alleviated by transforming the response to a **different scale**:

- **skewness to the right** is often mitigated by employing a square root, logarithm, or other transformation to a power **smaller** than 1;
- in contrast, **skewness to the left** can be lessened by a square, cube, or other transformation to a power **greater** than 1;
- similarly, a prevalent method to address non-constant error variances is through the transformation of the **response variable**.

The **Box-Cox transformation** is particularly well-suited to such a situation, offering a suite of transformations indexed by a parameter $\lambda$:[53]

53: We also discuss it in Section 8.3.5.

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0. \end{cases}$$

The idea is to transform the data over a spectrum of $\lambda$ values, perhaps between $-3$ and 3, and subsequently perform the ANOVA using $Y^(\lambda)$. We compute the sum of squared errors $\text{SSE}(\lambda)$ for every chosen $\lambda$.

Specifically, the optimal $\lambda$ is the one that **maximizes the log-likelihood**

$$-\frac{N}{2}\log[\text{SSE}(\lambda)] + (\lambda - 1)\sum_{i=1}^{a}\sum_{j=1}^{n_i}\log(y_{i,j}).$$

And what can we do about the assumption of **data independence**? Unfortunately, straightforward methodologies to confront data dependence are limited. Advanced analytical techniques like **time series analysis** (see Chapter 9) and **spatial statistics** could be used to model such data, but these are beyond the scope of this chapter.

### 11.3.9 Power and Sample Size

So far, our focus has primarily been on analyzing experimental results. A new focus now emerges as we proceed: how do we determine an appropriate **sample size** for a completely randomized design?

Ideally, the sample size should be **as small as possible**, in order to optimize both time and costs, yet it must also be **sufficiently large** to fulfill the analytical requirements.[54]

54: Making an informed decision on the appropriate sample size requires the analysts to have some knowledge of the system being examined; we will discuss this further in Chapters 13 and 14.

We need two additional distributions to answer the original question:

- if $X_1, \ldots, X_a \sim \mathcal{N}(\mu_i, 1)$ are independent random variables, then $X_1^2 + \cdots + x_a^2$ follows a **non-central $\chi^2$ distributions with $a$ degrees of freedom and non-centrality parameter** $\delta = \mu_1^2 + \cdots + \mu_a^2$, denoted by $a\overline{X} \sim \chi_a^2(\delta)$;[55]
- if $X \sim \chi_n(\eta)$ and $Y \sim \chi_m$, then

$$F = \frac{X/n}{Y/m} \sim F_{n,m}(\eta),$$

  where $F_{n,m}(\eta)$ is the **non-central $F$ distribution with $n$ and $m$ degrees of freedom and non-centrality parameter** $\eta$.

55: This definition is a generalization of the original definition of the (central) $chi_a^2$ distribution.

Recall that the statistic $F_0$ for testing

$$H_0 : \tau_1 = \cdots = \tau_a = 0 \quad \text{against} \quad H_1 : \tau_i \neq 0, \text{ for at least one } i$$

follows a distribution $F_{a-1,N-a}$ when $H_0$ is true. Under the alternative hypothesis $H_1$, this distribution assumption no longer holds.

Instead, the statistic $F_0$ follows a non-central $F_{a-1,N-a}(\delta^2)$, where

$$\delta^2 = n \sum_{i=1}^{a} \tau_i^2 / \sigma^2$$

is the non-centrality parameter.

This parameter essentially measures the extent to which the treatment means deviate from being equal, scaled relative to the variation of $\overline{y}_{i,\bullet}$, which is $\sigma^2/n$.

When computing the **power** for a specific sample size or determining the necessary **sample size** for a desired power, we have to use non-central $F$-distributions.

A potential complication arises from the fact that each value of the non-centrality parameter corresponds to a unique alternative distribution, meaning that there is a **distinct** non-central $F-$distribution for every possible non-centrality parameter value.

**Example** Suppose that $a = 5$ and that the treatment means are

$$\mu_1 = 11,\ \mu_2 = 12,\ \mu_3 = 15,\ \mu_4 = 18,\ \text{and}\ \mu_5 = 19.$$

From previous studies, we know that it is reasonable to expect that $\sigma^2 = 9$. What should $N$ (or $n$) be in a balanced complete design if we use a test with $\alpha = 0.01$, assuming we want a power of at least $1 - \beta = 0.9$? □

In order to answer this question, we need to actually know **ahead of time** what the true individual values of $\mu_1, \ldots, \mu_5$ are, which may prove challenging; we also needed to specify a plausible value (or range of values) for $\sigma^2$.

An alternative approach is to determine the sample size $N$ such that the **largest difference between** treatment means

$$\max\{\mu_i\} - \min\{\mu_i\}$$

is larger than a given value $D$.

If $D = \max\{\mu_i\} - \min\{\mu_i\}$, the non-centrality parameter is minimized when the other means are exactly in the middle of the interval

$$(\min\{\mu_i\}, \max\{\mu_i\}) = (\mu_{i_*}, \mu_{i^*}).$$

In that case, we would have

$$\tau_{i_*} = \mu_{i_*} - \mu = -\frac{D}{2} \quad \text{and} \quad \tau_{i^*} = \mu_{i^*} - \mu = \frac{D}{2},$$

and all other $\tau_i \equiv 0$, from which we conclude

$$\sum_{i=1}^{a} \tau_i^2 = 2(D/2)^2 = D^2/2.$$

It follows that

$$\delta^2_{\min} = nD^2/(2\sigma^2),$$

for a power equal to

$$P\left(F_{a-1,N-a}(\delta^2_{\min}) \geq F_{\alpha;a-1,N-a}\right).$$

**Example** With the data in the statement of the previous example, suppose that we have reason to believe that the largest difference between the treatment means is $D = 8$. Then

$$\delta^2_{\min} = n \cdot 8^2/(2 \cdot 9) = (32/9)n.$$

The power of the test is

$$P\left[F_{4,5(n-1)}((32/9)n) \geq F_{0.01;4,5(n-1)}\right].$$

We try different values of $n$, until we obtain a power which is at least 0.9.

```
for(n in c(2:9)){
delta.2.min = 32/9*n; df1 = 4; df2 = 5*(n-1); alpha = 0.01
crit = qf(0.01, df1=df1, df2=df2, lower.tail=FALSE)
print(c(n,
     pf(crit, df1=df1, df2=df2, ncp=delta.2.min,
     lower.tail=FALSE)))
}
```

```
[1] 2.0000000 0.0704121
[1] 3.0000000 0.2308392
[1] 4.0000000 0.4413316
[1] 5.0000000 0.6376441
[1] 6.0000000 0.7861772
[1] 7.0000000 0.8833954
[1] 8.0000000 0.9405001
[1] 9.0000000 0.9713123
```

With a value of $N = 40$ (i.e., with $n = 8$), the test's power is 94.1%. □

## 11.4 One-Way ANOVA with Random Effects

In the one-way ANOVA model from the previous section

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a;\ j = 1, \ldots, n,$$

where $\mu$ is the common mean to all observations and $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$, the treatment effects $\tau_1, \ldots, \tau_a$ are viewed as **fixed**; this one-way ANOVA model is known as a **fixed-effect model**.

But in some situations, the fixed-effect model is not appropriate; in this section, we consider treatments that are drawn randomly from a population of **potential treatments**, leading to a **random effects model**.

#### Examples: Fixe vs. Random Effects

- A business operates 50 machines that produce cardboard boxes for canned products. To analyze the consistency in the carton's durability, they randomly select ten machines out of the 50 and manufacture 40 boxes from each. They distribute 400 batches of feedstock cardboard randomly among these ten machines. Subsequently, the boxes' strength is assessed. This approach follows a completely randomized design, encompassing ten treatment groups and 400 units.

  In this context, a fixed-effect model is not suitable since the goal is to understand and draw conclusions **about the entire population of machines**, not merely the ten we tested in the experiment – we want to make assertions for the entire population, not just the random subset we examined. Moreover, if the experiment was repeated with a fresh batch of 10 machines, we would most likely end up with a completely distinct group of machines (and so with different observations).

- Imagine a home gardener conducting a small experiment using 24 tomato plants, divided into 4 varieties with 6 plants each. These varieties have piqued the gardener's interest after occasional use over recent summers. Now, the gardener plans to compare these varieties within a 12′ x 8′ garden patch. Each plant is randomly placed in one of the 2′ x 2′ sections. In this scenario, the gardener's focus is solely on these specific four varieties, with no consideration for any other types. The emphasis is strictly on the varieties being tested, and nothing else, so we can use fixed effects.

  Suppose, on the other hand, the 4 tomato varieties were chosen at random from a broader population of tomato types. In this scenario, we'd be dealing with random effects. If the experiment were repeated with a different batch of 4 varieties, it would likely result in a completely distinct group of tomato varieties.

- To determine how proficiently Ontario students can read by the conclusion of first grade, imagine we randomly select 6 schools within the province. From each chosen school, a group of students is randomly picked to undergo a reading assessment. Given that these schools are a random sample from a broader group of interest (all the schools in Ontario), we are operating under a random effect model.

  If our sole focus was on the performance of those specific 6 schools, then a fixed-effect model would have been appropriate. However, that is not the intention in this scenario.

### 11.4.1 Estimation of Model Parameters

The **one-way ANOVA model with random effects** is given by

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}, \quad i = 1, \ldots, a;\ j = 1, \ldots, n,$$

where

- $\mu$ is the global (or common) mean to all observations;
- $\tau_i \sim$ i.i.d. $\mathcal{N}(0, \sigma_T^2)$, $i = 1, \ldots, a$;
- $\varepsilon_{i,j} \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$, $\quad i = 1, \ldots, a,\ j = 1, \ldots, n_i$;
- $\tau_i$ and $\varepsilon_{i,j}$ are independent.

It follows that

$$\begin{aligned}
\mathrm{E}(\tau_i) = 0, &\quad \mathrm{Var}(\tau_i) = \sigma_T^2, \quad \mathrm{Cov}(\tau_i, \tau_{i'}) = 0, i \neq i';\\
\mathrm{E}(\varepsilon_{i,j}) = 0, &\quad \mathrm{Var}(\varepsilon_{i,j}) = \sigma^2, \quad \text{and}\\
\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = 0, &\quad \text{except when } i = i' \text{ and } j = j';\\
\mathrm{Cov}(\tau_i, \varepsilon_{i',j'}) = 0, &\quad \text{for all } i \text{ and } i'.
\end{aligned}$$

Consequently, we have

$$\mathrm{E}(y_{i,j}) = \mathrm{E}(y_{i,j} \mid \tau_i) = \mathrm{E}(\mu + \tau_i + \varepsilon_{i,j} \mid \tau_i) = \mathrm{E}(\mu + \tau_i) = \mu$$

and

$$\mathrm{Var}(y_{i,j}) = \mathrm{Var}(y_{i,j} \mid \tau_i) + \mathrm{VarE}(y_{i,j} \mid \tau_i) = \sigma_T^2 + \sigma^2.$$

Although the $\tau_i$'s and the $\varepsilon_{i,j}$'s are **uncorrelated**, the $y_{i,j}$'s are **correlated**. Indeed, for those in the **same treatment class**, we have

$$\text{Cov}(y_{i,j}, y_{i,j'}) = \text{Cov}(\mu + \tau_i + \varepsilon_{i,j}, \mu + \tau_i + \varepsilon_{i,j'}) = \sigma_T^2, \text{ for } j \neq j',$$

whereas for those in **different treatment classes**, we have

$$\text{Cov}(y_{i,j}, y_{i',j'}) = \text{Cov}(\mu + \tau_i + \varepsilon_{i,j}, \mu + \tau_{i'} + \varepsilon_{i',j'}) = 0, \text{ for } i \neq i'.$$

**Estimation of Parameters** The **intra-class correlation coefficient** is defined as

$$\rho = \frac{\text{Cov}(y_{i,j}, y_{i,j'})}{\sqrt{\text{Var}(y_{i,j})}\sqrt{\text{Var}(y_{i,j'})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}, \quad \text{if } j \neq j'.$$

It is a measure of the correlation between two observations from the same factor level (or class); the parameters $\sigma_T^2$ and $\sigma^2$ are the **variance components**.

In practice, there are 4 family of parameters to be estimated and/or predicted: the common mean $\mu$, the treatment effects $\tau_i$, and the variance components $\sigma_T^2$ and $\sigma^2$.

The common mean and the variance components are **fixed parameters**; these we seek to **estimate**. The treatment effects are random variables, these we seek to **predict**.

### 11.4.2 Analysis of Variance

In the one-way fixed-effects ANOVA model, we considered the overall test of hypothesis $H_0 : \tau_1 = \cdots = \tau_a = 0$. In the context of a random- effects ANOVA model, this hypothesis is nonsensical as the $\tau_i$'s are **random variables**.

Instead, we look to test if the factor (treatment) has an impact on the **variability of the response** $Y$. The null hypothesis is then expressed as $H_0 : \sigma_T^2 = 0$. The alternative stipulates that the factor has an effect on the variability of the response $Y$, which we express as $H_1 : \sigma_T^2 > 0$.

In effect, if $H_0$ is valid, then all the $\tau_i$'s are equal, whereas if $H_1$ holds, then at least two of the $\tau_i$'s differ.

Despite the fact that the fixed-effects model is emphatically not equivalent to the random-effects model, their **analysis of variance** for a single-factor study (one-way classification) is conducted in similar fashions.

We can show (see Exercises) that

$$\text{E(MSE)} = \sigma^2 \quad \text{and} \quad \text{E(MSA)} = \sigma^2 + n\sigma_T^2.$$

It then follows that MSE and MSA have the same expectation $\sigma^2$ if $\sigma_T^2 = 0$. If $\sigma_T^2 > 0$, on the other hand, then E(MSA) > E(MSE) as $n > 0$.

Therefore, we would **reject** $H_0$ **at significance level** $\alpha$ if

$$F_0 = \frac{\text{MSA}}{\text{MSE}} > F_{\alpha;a-1,N-a}.$$

To understand why we compare the observed test statistic $F_0$ to critical values of the $F_{a-1,N-a}$ distribution, we first note that

$$\overline{y}_{i,\bullet} = \frac{1}{N}\sum_{j=1}^{n} y_{i,j} = \mu + \tau_i + \overline{\varepsilon}_{i,\bullet}, \quad \text{where}$$
$$\overline{\varepsilon}_{i,\bullet} = \sum_{j=1}^{n} \frac{\varepsilon_{i,j}}{n} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right),$$

from which it follows that

$$\overline{y}_{i,\bullet} \sim \mathcal{N}\left(\mu, \sigma_T^2 + \frac{\sigma^2}{n}\right), \quad i = 1, \ldots, a.$$

The random variables $\overline{y}_{i,\bullet}$ being i.i.d., we must then have

$$\frac{(a-1)\text{MSA}}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{a-1}.$$

In the context of a balanced design, SSA can be expressed as

$$\begin{aligned}\text{SSA} &= \sum_{i=1}^{a} n_i(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet})^2 = n\sum_{i=1}^{a}(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet})^2 \\ &= n\sum_{i=1}^{a}\left[(\tau_i - \overline{\tau}_\bullet) + (\overline{\varepsilon}_{i,\bullet} - \overline{\varepsilon}_{\bullet,\bullet})\right]^2,\end{aligned}$$

where

$$\overline{\tau}_\bullet = \sum_{i=1}^{a}\frac{\tau_i}{a} \quad \text{and} \quad \overline{\varepsilon}_{\bullet,\bullet} = \sum_{i=1}^{a}\frac{\overline{\varepsilon}_{i,\bullet}}{a}.$$

On the other hand, we have

$$(n-1)s_i^2 = \sum_{j=1}^{n}(y_{i,j} - \overline{y}_{i,\bullet})^2 = \sum_{j=1}^{n}(\varepsilon_{i,j} - \overline{\varepsilon}_{i,\bullet})^2.$$

According to the i.i.d. condition,

$$\frac{(n-1)s_i^2}{\sigma^2} \sim \chi^2_{n-1}$$

independently for all $i = 1, \ldots .a$. As a result, we then have

$$\frac{(N-a)\text{MSE}}{\sigma^2} = (N-a)\sum_{i=1}^{a}\frac{s_i^2}{\sigma^2} = \sum_{i=1}^{a}\frac{(n-1)s_i^2}{\sigma^2} \sim \chi^2_{N-a}.$$

Thus, MSA only depends on $\{\tau_1, \ldots, \tau_a\}$ and $\{\overline{\varepsilon}_{1,\bullet}, \ldots, \overline{\varepsilon}_{a,\bullet}\}$ and MSE only depends on $\left\{s_1^2, \ldots, s_a^2\right\}$. But the sets $\{\tau_1, \ldots, \tau_a\}$ and $\left\{s_1^2, \ldots, s_a^2\right\}$ are independent, as are the sets $\{\overline{\varepsilon}_{1,\bullet}, \ldots, \overline{\varepsilon}_{a,\bullet}\}$ and $\left\{s_1^2, \ldots, s_a^2\right\}$. Therefore, MSA and MSE are independent and so we have, by definition of the $F$ distribution,

$$\frac{\sigma^2}{\sigma^2 + n\sigma_T^2}\frac{\text{MSA}}{\text{MSE}} \sim F_{a-1,N-a}.$$

Under $H_0 : \sigma_T^2 = 0$, this collapses to the decision protocol presented above.

### 11.4.3 Inference on $\sigma^2$, $\sigma_T^2$, and $\mu$

As was the case with the fixed-effects model, we can conduct inference on the model parameters.[56]

56: Assuming, as before, a balanced model.

**Confidence interval for $\sigma^2$ and $\sigma_T^2$** As before, MSE $= \widehat{\sigma^2}$ is an **unbiased estimator** of $\sigma^2$. Since $(N-a)\text{MSE}/\sigma^2 \sim \chi^2_{N-a}$, it follows from that we obtain a $100(1-\alpha)\%$ **confidence interval for** $\sigma^2$ *via*

$$\left[\frac{(N-a)\text{MSE}}{\chi^2_{\alpha/2;N-a}}, \frac{(N-a)\text{MSE}}{\chi^2_{1-\alpha/2;N-a}}\right].$$

But we also have

$$E\left(\frac{\text{MSA}-\text{MSE}}{n}\right) = \frac{\sigma^2}{n} - \frac{\sigma^2+n\sigma_T^2}{n} = \sigma_T^2;$$

consequently, $(\text{MSA}-\text{MSE})/n$ is an **unbiased estimator** of $\sigma_T^2$.

However, nothing precludes this estimator to take on **negative** values, which may occur when MSA $<$ MSE.[57] To overcome this issue, we use the **truncated estimator**

57: This can occur when we are evaluating MSE and MSA from actual data (not their expectations).

$$\hat{\sigma}_T^2 = \begin{cases} (\text{MSA}-\text{MSE})/n, & \text{if MSA} \geq \text{MSE}, \\ 0, & \text{otherwise.} \end{cases}$$

The distribution of $\hat{\sigma}_T^2$ is not simple since it is expressed as the linear combination of two chi-square distributions. As a result, we cannot derive an exact confidence interval for $\sigma_T^2$; we will have to settle for an **approximate confidence interval for** $\sigma_T^2$.

However, we can construct an **exact** confidence interval for the intra-class correlation coefficient $\rho = \sigma_T^2/(\sigma_T^2+\sigma^2)$. Indeed,

$$\begin{aligned}
1-\alpha &= P\left(F_{1-\alpha/2;a-1,N-a} \leq \frac{\sigma^2}{\sigma^2+n\sigma_T^2}\frac{\text{MSA}}{\text{MSE}} \leq F_{\alpha/2;a-1,N-a}\right) \\
&= P\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{\alpha/2;a-1,N-a}}-1\right) \leq \frac{\sigma_T^2}{\sigma^2} \leq \frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{1-\alpha/2;a-1,N-a}}-1\right)\right) \\
&= P\left(g\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{\alpha/2;a-1,N-a}}-1\right)\right) \leq g\left(\frac{\sigma_T^2}{\sigma^2}\right) = \frac{\sigma_T^2}{\sigma_T^2+\sigma^2} \leq g\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{1-\alpha/2;a-1,N-a}}-1\right)\right)\right),
\end{aligned}$$

where $g(t) = t/(t+1)$ is an **increasing** function. Therefore, a $100(1-\alpha)\%$ confidence interval for $\sigma_T^2/(\sigma_T^2+\sigma^2)$ is obtained *via*:

$$\left[\frac{\text{MSA}-F_{\alpha/2;a-1,N-a}\text{MSE}}{\text{MSA}+(n-1)F_{\alpha/2;a-1,N-a}\text{MSE}}, \frac{\text{MSA}-F_{1-\alpha/2;a-1,N-a}\text{MSE}}{\text{MSA}+(n-1)F_{1-\alpha/2;a-1,N-a}\text{MSE}}\right].$$

When $N-a$ is large, the estimator MSE of $\sigma^2$ becomes more precise and we can write $\sigma^2 \approx \text{MSE}$.

It follows that

$$1-\alpha \approx P\left(\frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{\alpha/2;a-1,N-a}}-1\right) \le \frac{\sigma_T^2}{\text{MSE}} \le \frac{1}{n}\left(\frac{\text{MSA}}{\text{MSE}}\frac{1}{F_{1-\alpha/2;a-1,N-a}}-1\right)\right)$$
$$= P\left(\frac{1}{n}\left(\frac{\text{MSA}}{F_{\alpha/2;a-1,N-a}}-\text{MSE}\right) \le \sigma_T^2 \le \frac{1}{n}\left(\frac{\text{MSA}}{F_{1-\alpha/2;a-1,N-a}}-\text{MSE}\right)\right),$$

which provides an approximate $100(1-\alpha)\%$ confidence interval for $\sigma_T^2$.

**Confidence interval for $\mu$** Inferences about the global mean are simpler to obtain. The expression

$$\widehat{\mu} = \overline{y}_{\bullet,\bullet} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{i,j}$$

provides an unbiased estimator of $\mu$. Its variance is given by

$$\text{Var}(\widehat{\mu}) = \frac{n\sigma_T^2+\sigma^2}{N};$$

an unbiased estimator of which is given by

$$\widehat{\text{Var}}(\widehat{\mu}) = \frac{\text{MSA}}{N}.$$

It follows that we can find a $100(1-\alpha)\%$ confidence interval for $\mu$ *via*

$$\overline{y}_{\bullet,\bullet} \pm t_{\alpha/2;a-1}\sqrt{\frac{\text{MSA}}{N}}.$$

### 11.4.4 Power of a Test

In the case of the $F-$test at significance level $\alpha$ for a one-way random-effects model, the power of the test

$$H_0 : \sigma_T^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_T^2 \neq 0$$

is:

$$P(\Delta) = P\left(\frac{\text{MSA}}{\text{MSE}} > F_{\alpha;a-1,N-a} \,\middle|\, \sigma_T^2 \neq 0\right)$$
$$= P\left(\frac{\sigma^2}{\sigma^2+n\sigma_T^2}\frac{\text{MSA}}{\text{MSE}} > \frac{\sigma^2}{\sigma^2+n\sigma_T^2}F_{\alpha;a-1,N-a}\right)$$
$$= P\left(F_{a-1,N-a} > \frac{1}{1+\Delta}F_{\alpha;a-1,N-a}\right),$$

where $\Delta = n\sigma_T^2/\sigma^2$.

## 11.5 Randomized Complete Block Designs

As the variance of the experimental error $\sigma^2$ **increases**, the corresponding confidence intervals get **longer** and the power of tests **decreases**.

All other things being equal, we would thus prefer to conduct experiments with units that are **homogeneous** so that $\sigma^2$ is as small as possible.

We can achieve this through **variance-reduction designs**, which almost exclusively use **blocking**. A **block of units** is a collection of units that are homogeneous in **some sense** – field plots located in the same general area, or units that came from a single supplier, say.

These similarities in the units themselves lead us to anticipate that units **within a block** may have **similar responses**.

When constructing blocks, the goal is to achieve homogeneity of the units within blocks, with the caveat that units in different blocks may be **dissimilar**.[58] The primary purpose of blocking is to remove or isolate the **block-to-block variability**. This helps ensure that this variability does not overshadow or **mask the treatment effects** under consideration.

58: Compare with the notion of stratified random sampling in Section 10.4.

A notable experimental design that makes use of this concept is the **Randomized Complete Block Design** (RCBD). This design is structured for comparing $a$ treatments **across** $b$ blocks. In this setup, treatments are **randomly assigned** to experimental units within a block – each treatment appears **exactly once** in every block. If a RCBD integrates $a$ treatments within each of $b$ blocks, then the **total number of observations** would be $N = ab$.

Randomized block designs can be seen as an **extension** of the paired-difference designs that were discussed in Section 11.2.

#### Examples: Randomized Complete Block Design

- A production supervisor is keen on comparing the mean assembly times of operators using three distinct methods: A, B, and C. Given the anticipated variation in assembly times across different operators, the supervisor employs an RCDB for the comparison.

  Specifically, six assembly-line operators are selected, each representing a block. Each operator is tasked with assembling the item three times, once for every method. The importance of the sequence in which the methods are applied is recognized, as factors like fatigue or heightened dexterity might influence the results. Therefore, every operator is assigned a randomized sequence of the three methods. For instance:

  - Operator 1 first uses method A, proceeds to B, and finishes with C.
  - Operator 2 first uses method A, proceeds to C, and finishes with B.
  - Operator 3 first uses method B, proceeds to A, and finishes with C.
  - Operator 4 first uses method B, proceeds to C, and finishes with A.
  - Operator 5 first uses method C, proceeds to A, and finishes with B.
  - Operator 6 first uses method C, proceeds to B, and finishes with A.

- The credit card industry is engaged in an intense competition for cardholders. Each company designs its unique, intricate reward and fee structure in an attempt to attract customers. Notably, the benefits or costs associated with a credit card can vary significantly depending on the cardholder's monthly spending.

  To investigate this, a consumer watchdog group set out to compare the average rewards or fees of four different credit card companies (A, B, C, D). They used three distinct spending levels as blocks:

  – low spending – \$500 per month,
  – middle spending – \$2,500 per month, and
  – high spending – \$10,000 per month.

  If the rewards are not monetary in nature, the watchdog group has first converted them to a monetary value. The average monthly rewards, as quoted by the credit card companies for cardholders across these spending levels, are presented in the table below.

| **Rewards** | **Credit Card Company** | | | |
|---|---|---|---|---|
| **Spending Level** | **A** ($i = 1$) | **B** ($i = 2$) | **C** ($i = 3$) | **D** ($i = 4$) |
| **Low** ($j = 1$) | 30 | 27 | 34 | 26 |
| **Middle** ($j = 2$) | 68 | 76 | 65 | 67 |
| **High** ($j = 3$) | 304 | 322 | 308 | 296 |

### 11.5.1 Analysis of Variance

In an RCBD, we consider two key factors: **treatments** and **blocks**, both of which play a significant role in influencing the response. Let $y_{i,j}$ represent the response when the $i$th treatment is applied within the $j$th block. The underlying RCBD is described *via*:

$$y_{i,j} = \mu + \tau_i + \beta_j + \varepsilon_{i,j}, \quad i = 1, \ldots, a; \quad j = 1, \ldots, b,$$

where the error terms $\varepsilon_{i,j}$ are independent random variables from a $\mathcal{N}(0, \sigma^2)$ distribution.

In this model, the parameter $\mu$ represents the global effect, while $\tau_i$ denotes the treatment effect for the $i$th treatment level, and $\beta_j$ indicates the effect associated with the $j$th block.[59]

Both treatments and blocks are regarded as **fixed factors**; the expected value of the response can thus be expressed as:

$$\mathrm{E}(y_{i,j}) = \mu + \tau_i + \beta_j.$$

Just as in the one-way (single-factor) fixed-effect experimental design discussed previously, the RCBD model is **over-parameterized**.[60] The primary aim is to test the **uniformity of the treatment means**, effectively examining the presence or absence of an effect for Factor A.

59: We also refer to the treatment as the **first factor** (or Factor A), and to blocking as the **second factor** (or Factor B).

60: We can bypass this problem by enforcing constraints on the treatment and block effects:

$$\sum_{i=1}^{a} \tau_i = 0 \quad \text{and} \quad \sum_{j=1}^{b} \beta_j = 0.$$

Formally, we test for

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0, \quad \text{against} \quad H_1 : \tau_i \neq 0 \quad \text{for at least one } i.$$

The **totals** for the $i$th treatment, the $j$th block, and the overall total of the $N = ab$ observations are given, respectively, by

$$y_{i,\bullet} = \sum_{j=1}^{b} y_{i,j}, \quad i = 1, \ldots, a, \quad y_{\bullet,j} = \sum_{i=1}^{a} y_{i,j}, \quad j = 1, \ldots, b$$

$$y_{\bullet,\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{i,j} = \sum_{i=1}^{a} y_{i,\bullet} = \sum_{j=1}^{b} y_{\bullet,j}.$$

Similarly, we define the various **means**

$$\overline{y}_{i,\bullet} = \frac{y_{i,\bullet}}{b}, \quad \overline{y}_{\bullet,j} = \frac{y_{\bullet,j}}{a}, \quad \text{and} \quad \overline{y}_{\bullet,\bullet} = \frac{y_{\bullet,\bullet}}{N}.$$

**Example (cont.)** In the credit card example from earlier in the section, the totals and means are given in the table below.

| **Rewards** | **Credit Card Company** | | | | | |
|---|---|---|---|---|---|---|
| **Spending Level** | **A** $(i = 1)$ | **B** $(i = 2)$ | **C** $(i = 3)$ | **D** $(i = 4)$ | **Totals** $y_{\bullet,j}$ | **Means** $\overline{y}_{\bullet,j}$ |
| **Low** $(j = 1)$ | 30 | 27 | 34 | 26 | 117 | 29.25 |
| **Middle** $(j = 2)$ | 68 | 76 | 65 | 67 | 276 | 69 |
| **High** $(j = 3)$ | 304 | 322 | 308 | 296 | 1230 | 307.5 |
| **Totals** $y_{i,\bullet}$ | 402 | 425 | 407 | 389 | 1623 | |
| **Means** $\overline{y}_{i,\bullet}$ | 134 | 141.7 | 135.7 | 129.7 | | 135.25 |

The **total sum of square** (SST) can be expressed as the sum of three sums of squares:

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \overline{y}_{\bullet,\bullet})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left[(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}) + (\overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet}) + (y_{i,j} - \overline{y}_{i,\bullet} - \overline{y}_{\bullet,j} + \overline{y}_{\bullet,\bullet})\right]^2 \\ &= b\sum_{i=1}^{a}(\overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet})^2 + a\sum_{j=1}^{b}(\overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \overline{y}_{i,\bullet} - \overline{y}_{\bullet,j} + \overline{y}_{\bullet,\bullet})^2,\end{aligned}$$

or, using the customary symbols (along with the corresponding degrees of freedom):

$$\begin{aligned}\text{SST} &= \text{SSA} + \text{SSB} + \text{SSE} \\ N - 1 &= (a-1) + (b-1) + (a-1)(b-1) = ab - 1\end{aligned}$$

There are equivalent formulas (which are slightly easier to use) for the sums of squares:

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{a}\sum_{j=1}^{b} y_{i,j}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \\
\text{SSA} &= \frac{1}{b}\sum_{i=1}^{a} y_{i,\bullet}^2 - \frac{y_{\bullet,\bullet}^2}{N}, \\
\text{SSB} &= \frac{1}{a}\sum_{j=1}^{b} y_{\bullet,j}^2 - \frac{y_{\bullet,\bullet}^2}{N},
\end{aligned}$$

Finally, SSE is obtained as

$$\text{SSE} = \text{SST} - \text{SSA} - \text{SSB}.$$

It can be shown that

$$\frac{\text{SSA}}{\sigma^2} \sim \chi^2_{a-1}\left(b\sum_{i=1}^{a}\frac{\tau_i^2}{\sigma^2}\right), \quad \frac{\text{SSB}}{\sigma^2} \sim \chi^2_{b-1}\left(a\sum_{j=1}^{b}\frac{\beta_j^2}{\sigma^2}\right),$$

and

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{(a-1)(b-1)}.$$

As has been the case throughout, we can also show that the three sums of squares SSA, SSB, and SSE are **mutually independent**. The corresponding mean squares are obtained in the usual way:

$$\text{MSA} = \frac{\text{SSA}}{a-1}, \quad \text{MSB} = \frac{\text{SSB}}{b-1}, \quad \text{and} \quad \text{MSE} = \frac{\text{SSE}}{(a-1)(b-1)}.$$

We can show (see Exercises) that

$$\begin{aligned}
\text{E(MSA)} &= \sigma^2 + \frac{b}{a-1}\sum_{i=1}^{a}\tau_i^2, \\
\text{E(MSB)} &= \sigma^2 + \frac{a}{b-1}\sum_{j=1}^{b}\beta_j^2, \\
\text{E(MSE)} &= \sigma^2.
\end{aligned}$$

We can test for the absence of a treatment effect (Factor A) by pitting $H_0 : \tau_1 = \cdots = \tau_a = 0$ against $H_1 : \tau_i \neq 0$ for at least one $i$, using the test statistics

$$F_0 = \frac{\text{MSA}}{\text{MSE}},$$

which follows an $F_{a-1,(a-1)(b-1)}$ distribution under $H_0$.

All of this is summarized in Table 11.19.[61]

61: A "large" value of the ratio MSB/MSE implies that blocking was a good strategy.

| **Source** | **SS** | **df** | **MS** | $\mathbf{F_0}$ |
|---|---|---|---|---|
| **Treatment** | SSA | $a-1$ | MSA | $F_0 = \text{MSA/MSE}$ |
| **Block** | SSB | $b-1$ | MSB | |
| **Error** | SSE | $(a-1)(b-1)$ | MSE | |
| **Total** | SST | $N-1$ | | |

**Table 11.19:** ANOVA table for the equality of the treatment means $\tau_i$ in a two-factor randomized complete block design.

**Example (cont.)** In the credit card example from earlier in the section, we have a randomized block design with $b = 3$ spending levels (blocks) and $a = 4$ companies (treatments), so there are $N = ab = 12$ observations.

We start by loading the data.

```
content.1 <- c(30, 27, 34, 26)
content.2 <- c(68, 76, 65, 67)
content.3 <- c(304, 322, 308, 296)
data <- data.frame(rbind(content.1, content.2, content.3))
rownames(data) <- c("Low", "Middle", "High")
colnames(data) <- c("A", "B", "C", "D")
row.totals <- rowSums(data)
row.means <- rowMeans(data)
data <- cbind(data, row.totals, row.means)
col.totals <- colSums(data)
col.means <- colMeans(data)
data <- rbind(data, col.totals, col.means)
rownames(data) <- c("Low", "Middle", "High", "col.totals",
                    "col.means")
data[4,6] <- NA; data[5,5] <- NA
```

```
             A        B        C        D row.totals row.means
Low         30  27.0000  34.0000  26.0000        117     29.25
Middle      68  76.0000  65.0000  67.0000        276     69.00
High       304 322.0000 308.0000 296.0000       1230    307.50
col.totals 402 425.0000 407.0000 389.0000       1623
col.means  134 141.6667 135.6667 129.6667                135.25
```

We compute the necessary quantities and place them in the ANOVA table.

```
a = ncol(content)
b = nrow(content)
N = a*b
grand.mean = data[b+2,a+2]
SST = sum((data[c(1:b),c(1:a)]-grand.mean)^2)
SSA = b * sum((data[b+2,c(1:a)]-grand.mean)^2)
SSB = a * sum((data[c(1:b),a+2]-grand.mean)^2)
SSE = SST - SSA - SSB
ANOVA = as.data.frame(cbind(c(SSA,SSB,SSE,SST),
            c(a-1, b-1, (a-1)*(b-1), N-1),
            c(SSA/(a-1),SSB/(b-1),SSE/((a-1)*(b-1)),0),
            c((SSA/(a-1))/(SSE/((a-1)*(b-1))),(SSB/(b-1))/(SSE/((a-1)*(b-1))),0,0)))
rownames(ANOVA) = c("Treatment", "Block", "Error", "Total")
colnames(ANOVA) = c("SS", "df", "MS", "F0")
ANOVA
```

```
                SS df          MS         F0
Treatment    222.25  3    74.08333    1.84058
Block     181180.50  2 90590.25000 2250.68944
Error        241.50  6    40.25000
Total     181644.25 11
```

At significance level $\alpha = 0.05$, the critical value of $F_{4-1,(4-1)(3-1)} = F_{3,6}$ is given below.

```
qf(0.05, df1 = a-1, df2 = (a-1)*(b-1), lower.tail=FALSE)
```

```
[1] 4.757063
```

We see that $F_0 = \text{MSA/MSE} = 1.84 < F_{0.05;3,6} = 4.76$; therefore, the results do not show a significant difference in the treatment means. That is, there is insufficient evidence to indicate a difference in the credit card companies' monthly rewards.[62]

62: The ratio MSB/MSE is quite large, which suggests that blocking is effective, even if we cannot say that the treatment is so.

### 11.5.2 Estimation of Model Parameters

The RCBD model parameters are the grand mean $\mu$, the treatment effects $\tau_i$, and the blocking effect $\beta_j$, which can be estimated from the data as follows.

We seek to minimize the sum of squares errors:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\varepsilon_{i,j}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \mu - \tau_i - \beta_j)^2.$$

We determine the model values of $\mu$, $\tau_i$ and $\beta_j$ by differentiating the expression above, setting the gradient to 0, and solving for the parameters. In the RCBD context, this leads to:

$$\begin{aligned}
\mu &: \ -2\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{i,j} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0, \\
\tau_i &: \ -2\sum_{j=1}^{b}(y_{i,j} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0, \quad i = 1, \ldots, a, \\
\beta_j &: \ -2\sum_{i=1}^{a}(y_{i,j} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j) = 0, \quad j = 1, \ldots, b.
\end{aligned}$$

After some simplifications, we obtain the following system of linear equations:

$$\begin{aligned}
\mu &: \ N\hat{\mu} = y_{\bullet,\bullet}, \\
\tau_i &: \ b\hat{\mu} + b\hat{\tau}_i = y_{i,\bullet}, \quad i = 1, \ldots, a, \\
\beta_j &: \ a\hat{\mu} + a\hat{\beta}_j = y_{\bullet,j}, \quad j = 1, \ldots, b,
\end{aligned}$$

whose solution is

$$\hat{\mu} = \overline{y}_{\bullet,\bullet}, \quad \hat{\tau}_i = \overline{y}_{i,\bullet} - \overline{y}_{\bullet,\bullet}, \quad \hat{\beta}_j = \overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet}.$$

### 11.5.3 Multiple Comparisons

We can compare two treatments $i$ and $i'$, by looking at the difference of treatments $\tau_i - \tau_{i'}$, which we estimate *via* $\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}$.

The variance of $\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}$ is given by

$$\text{Var}(\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}) = \sigma^2 \cdot \frac{2}{b}.$$

We obtain an $100(1-\alpha)\%$ confidence interval for $\tau_i - \tau_{i'}$ in the usual manner:

$$\tau_i - \tau_{i'} : \ (\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}) \pm t_{\alpha/2;(a-1)(b-1)} \sqrt{\text{MSE}} \sqrt{\frac{2}{b}}.$$

For simultaneous confidence intervals, we must use a modification (as in Section 11.3.7). If we use **Tukey's method**, for instance, the confidence interval with family confidence $100(1-\alpha)\%$ becomes

$$\tau_i - \tau_{i'} : \ (\overline{y}_{i,\bullet} - \overline{y}_{i',\bullet}) \pm q_{\alpha;a,(a-1)(b-1)} \sqrt{\text{MSE}} \sqrt{\frac{1}{b}}.$$

### 11.5.4 Power and Sample Size

Whether or not Factor A has an effect, the distribution of the test statistic $F_0$ is a non-central $F_{a-1,(a-1)(b-1)}(\delta^2)$, with non-centrality parameter

$$\delta^2 = b \sum_{i=1}^{a} \tau_i^2 / \sigma^2.$$

To determine the **sample size**, we can use an approach similar to the one described in Section 11.3.9.

The differences between the treatment effects are $\tau_i - \tau_{i'}$; the largest difference between the treatment averages is thus

$$D = \max\{\tau_i\} - \min\{\tau_i\}.$$

The **minimal** non-centrality parameter is thus

$$\delta^2_{\text{min}} = bD^2/(2\sigma^2),$$

which yields a **test power** of

$$P\left(F_{a-1,(a-1)(b-1)}(\delta^2_{\text{min}}) \geq F_{\alpha;a-1,(a-1)(b-1)}\right).$$

### 11.5.5 Model Validation

As in the previously studied design, three basic assumptions about errors must be checked: **independence**, **normality**, and **homoscedasticity**. As before, we use the residuals to verify whether the assumptions seem reasonable. In the RCBD **predicted responses** are

$$\hat{y}_{i,j} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j = \overline{y}_{i,\bullet} + \overline{y}_{\bullet,j} - \overline{y}_{\bullet,\bullet};$$

their **residuals** are thus

$$e_{i,j} = y_{i,j} - \hat{y}_{i,j} = y_{i,j} - \overline{y}_{i,\bullet} - \overline{y}_{\bullet,j} + \overline{y}_{\bullet,\bullet}.$$

## 11.6 Factorial Designs

In our discussions up to this point, we primarily focused on the foundational problem of understanding how a **single independent factor** influences the **response**. However, it's not uncommon in research to encounter situations where the interest lies in studying the **combined effects** of **multiple independent variables** on a given response. We call such experimental setups, where two or more factors are simultaneously investigated, **factorial designs**.

Consider an example where researchers wish to determine the effect of sleep deprivation on student test performance. If the study only revolves around the sleep factor and the test performance, it is a **simple (one-way) experiment**. But we can add a twist: what if the researchers also wants to know whether the impacts of sleep deprivation vary between high school and university students? This introduces a second factor, school level,[63] into the study, turning it into a factorial design.

63: Which is presumably linked to age.

Factorial designs can vary in their complexity. A frequently encountered type is the $2 \times 2$ factorial design, where **two factors** are being analyzed, and each factor has **two distinct levels**. The numeric representation of a factorial design offers quick insights: the number of digits indicates the **number of factors**, while the value of each number shows **how many levels** the corresponding factor has. For instance, a $4 \times 3$ factorial design consists of two factors, with the first having four levels and the second comprising three levels. Extending this understanding, a $2 \times 2 \times 2$ factorial design would mean the experiment has three factors, each of which having two levels.

### 11.6.1 Two-Way Factorial Experiments

We start by looking into **two-factor designs**. The data from a two-way factorial design can be illustratively showcased using a table, as in Table 11.21.

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | $y_{1,1,1}, \ldots, y_{1,1,n}$ | $y_{1,2,1}, \ldots, y_{1,2,n}$ | $y_{1,3,1}, \ldots, y_{1,3,n}$ |
| $A_2$ | $y_{2,1,1}, \ldots, y_{2,1,n}$ | $y_{2,2,1}, \ldots, y_{2,2,n}$ | $y_{2,3,1}, \ldots, y_{2,3,n}$ |
| $A_3$ | $y_{3,1,1}, \ldots, y_{3,1,n}$ | $y_{3,2,1}, \ldots, y_{3,2,n}$ | $y_{3,3,1}, \ldots, y_{3,3,n}$ |
| $A_4$ | $y_{4,1,1}, \ldots, y_{4,1,n}$ | $y_{4,2,1}, \ldots, y_{4,2,n}$ | $y_{4,3,1}, \ldots, y_{4,3,n}$ |

**Table 11.21:** $4 \times 3$ factorial design treatment structure, with $n$ observations per cell.

In this representation, **rows** align with the levels of one specific factor (designated as Factor A), while columns represent the levels of the second factor (Factor B).

In that design, there are $4 \times 3 = 12$ total treatments. In **balanced** factorial designs, the number of observations $n$ per unique combination of factor levels (which we also call a **cell**) is the same value across **all combinations**. For the current discussion, we assume that the collected data is balanced, $n$ responses to a cell.

Assume that we are working with an $a \times b$ two-way design; there are $N = abn$ observations in total. We refer to the $k$th response in the $(i, j)-$cell by $y_{i,j,k}$.

By similarity to the one-way design, we adopt the following notation:

$$y_{i,\bullet,\bullet} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{i,\bullet,\bullet} = \frac{y_{i,\bullet,\bullet}}{bn};$$

$$y_{\bullet,j,\bullet} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{\bullet,j,\bullet} = \frac{y_{\bullet,j,\bullet}}{an};$$

$$y_{i,j,\bullet} = \sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{i,j,\bullet} = \frac{y_{i,j,\bullet}}{n};$$

$$y_{\bullet,\bullet,\bullet} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}, \quad \overline{y}_{\bullet,\bullet,\bullet} = \frac{y_{\bullet,\bullet,\bullet}}{N}.$$

**Example** We are interested in determining if a medicated agent can help reduce inflammation among athletes. 6000 college-level athletes are assigned to 30 lots of 200 athletes each.

The 30 lots are divided at random into ten groups of three lots each, with each group receiving a different treatment.

A treatment is factorial combination of the medication dosage (Factor A, with two levels), and when the medication is applied (Factor B, with five levels: 1 hour after a game, immediately after the game, during the game, immediately before the game, 1 hour before game).

In each lot, the response is the number of athletes who experience inflammation at some point within a 24-hour period after the game.

| **Cases** | **Application Period** | | | | |
|---|---|---|---|---|---|
| **Dosage** | **1** | **2** | **3** | **4** | **5** |
| **Low** | 10 | 6 | 8 | 12 | 19 |
| | 7 | 18 | 36 | 29 | 46 |
| | 9 | 16 | 19 | 35 | 37 |
| **High** | 3 | 7 | 9 | 10 | 15 |
| | 4 | 4 | 10 | 10 | 26 |
| | 7 | 0 | 4 | 0 | 10 |

The data is summarized below.

| **Cases** | **Application Period** | | | | | |
|---|---|---|---|---|---|---|
| **Dosage** | **1** | **2** | **3** | **4** | **5** | $y_{i,\bullet,\bullet}$ |
| **Low** | 26 | 40 | 63 | 76 | 102 | 307 |
| **High** | 14 | 11 | 23 | 20 | 51 | 119 |
| $y_{\bullet,j,\bullet}$ | 40 | 51 | 86 | 96 | 153 | 426 |

We will discuss how to estimate the two-way factorial design model parameters shortly. □

Typically, we are interested in the **treatment effects** and **interaction effects**. The mathematical representation of a two-way factorial experiment is given by the model:

$$y_{i,j,k} = \mu_{i,j} + \varepsilon_{i,j,k}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n;$$

the subscripts $i$ and $j$ serve as indices for the treatment levels A and B, respectively.

We can re-write the treatment effects as follows:

$$\begin{aligned}\mu_{i,j} &= \overline{\mu}_{\bullet,\bullet} + (\overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,\bullet}) + (\overline{\mu}_{\bullet,j} - \overline{\mu}_{\bullet,\bullet}) + (\mu_{i,j} - \overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,j} + \overline{\mu}_{\bullet,\bullet}) \\ &= \mu + \tau_i + \beta_j + (\tau\beta)_{i,j}\end{aligned}$$

By adopting this perspective, we can reformulate the model as:

$$y_{i,j,k} = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j} + \varepsilon_{i,j,k}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n.$$

As always, we incorporate constraints to avoid an over-parametrized model:

$$\sum_{i=1}^{a} \tau_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} (\tau\beta)_{i,j} = 0, \quad \sum_{j=1}^{b} (\tau\beta)_{i,j} = 0.$$

The **main treatment effects** are represented by $\tau_i$ (Factor A) and $\beta_j$ (Factor B); the **interaction effect** by $(\tau\beta)_{i,j}$. This interaction plays a pivotal role in understanding the experiment's nuances.

The **row effects** tells us how the response changes as we transition from one row to the next, averaged across all columns. In contrast, the **column effect** tells us how the response changes as we move from once column to the next, averaged across all rows.

The **interaction effects** tell us how the change in response depends on columns when moving between rows, or how the change in response depends on rows when moving between columns. An interaction term between Factor A and Factor B means that the change in mean response going from level $i_1$ of Factor A to level $i_2$ of Factor A depends on the level of Factor B under consideration.[64]

64: We cannot simply say that changing the level of Factor A changes the response by a given amount; we may need a different amount of change for each level of Factor B.

**Advantages** Factorial experiments present several advantages.

- When the factors do not interact, factorial experiments are more efficient than one-at-a-time experiments, as the units can be used to assess the (main) effects for both factors. Units in a one-at-a-time experiment can only be used to assess the effects of one factor.
- When the factors interact, factorial experiments can estimate the interaction. One-at-at-time experiments cannot estimate interaction. Use of one-at-a-time experiments in the presence of interaction can lead to serious misunderstanding of how the response varies as a function of the factors.

When there is no interaction, then the **main treatment effects alone** are suffcient to describe the means of the response – such a model is said to be **additive**.

**Estimation of Model Parameters** As before, we seek to minimize the sum of squared residuals

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{i,j,k}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\mu-\tau_i-\beta_j-\gamma_{i,j})^2,$$

where we write $\gamma_{i,j}$ for $(\tau\beta)_{i,j}$ to simplify the notation.

We compute the partial derivatives and set them to 0:

$$\begin{aligned}
\mu &: \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{\bullet,\bullet,\bullet}-N\hat\mu = 0;\\
\tau_i &: \sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{i,\bullet,\bullet}-bn\hat\mu-bn\hat\tau_i = 0;\\
\beta_j &: \sum_{i=1}^{a}\sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{\bullet,j,\bullet}-an\hat\mu-an\hat\beta_j = 0;\\
\gamma_{i,j} &: \sum_{k=1}^{n}(y_{i,j,k}-\hat\mu-\hat\tau_i-\hat\beta_j-\hat\gamma_{i,j}) = y_{i,j,\bullet}-n\hat\mu-n\hat\tau_i-n\hat\beta_j-n\hat\gamma_{i,j} = 0.
\end{aligned}$$

The system's solution is

$$\begin{aligned}
\hat\mu &= \overline{y}_{\bullet,\bullet,\bullet},\\
\hat\tau_i &= \overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,\bullet,\bullet}, \quad i=1,\dots,a,\\
\hat\beta_j &= \overline{y}_{\bullet,j,\bullet}-\overline{y}_{\bullet,\bullet,\bullet}, \quad j=1,\dots,b,\\
\hat\gamma_{i,j} &= \overline{y}_{i,j,\bullet}-\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,j,\bullet}+\overline{y}_{\bullet,\bullet,\bullet}, \quad i=1,\dots,a,\ j=1,\dots,b.
\end{aligned}$$

**Analysis of Variance** The total sum of squares can be decomposed as

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{i,j,k}-\overline{y}_{\bullet,\bullet,\bullet})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left[(\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})+(\overline{y}_{\bullet,j,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})+(\overline{y}_{i,j,\bullet}-\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,j,\bullet}+\overline{y}_{\bullet,\bullet,\bullet})+(y_{i,j,k}-\overline{y}_{i,j,\bullet})\right]^2\\
&= bn\sum_{i=1}^{a}(\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})^2+an\sum_{j=1}^{b}(\overline{y}_{\bullet,j,\bullet}-\overline{y}_{\bullet,\bullet,\bullet})^2+n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{i,j,\bullet}-\overline{y}_{i,\bullet,\bullet}-\overline{y}_{\bullet,j,\bullet}+\overline{y}_{\bullet,\bullet,\bullet})^2+\sum_{i,j,k}(y_{i,j,k}-\overline{y}_{i,j,\bullet})^2,
\end{aligned}$$

which we can re-write simply as

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}.$$

The corresponding ANOVA table is shown below.

| Source | SS | df | MS | F |
|---|---|---|---|---|
| **Treatment A** | SSA | $a-1$ | MSA | $F_A = \text{MSA}/\text{MSE}$ |
| **Treatment B** | SSB | $b-1$ | MSB | $F_B = \text{MSB}/\text{MSE}$ |
| **Interaction AB** | SSAB | $(a-1)(b-1)$ | MSAB | $F_{AB} = \text{MSAB}/\text{MSE}$ |
| **Error** | SSE | $ab(n-1)$ | MSE | |
| **Total** | SST | $N-1$ | | |

**Table 11.25:** ANOVA table for equality of factorial effects and of interaction effects, in a two-way design.

As always, there are equivalent formulas for the sums of squares:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}^2 - \frac{y_{\bullet,\bullet,\bullet}^2}{N}; \quad \text{SSA} = \sum_{i=1}^{a} \frac{y_{i,\bullet,\bullet}^2}{bn} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};$$

$$\text{SSB} = \sum_{j=1}^{b} \frac{y_{\bullet,j,\bullet}^2}{an} - \frac{y_{\bullet,\bullet,\bullet}^2}{N}; \quad \text{SSTR} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{y_{i,j,\bullet}^2}{n} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};$$

$$\text{SSAB} = \text{SSTR} - \text{SSA} - \text{SSB}; \quad \text{SSE} = \text{SST} - \text{SSTR}.$$

**Example** In a comprehensive study aimed at understanding the growth dynamics of plants, we use a 33 factorial design, resulting in a total of 9 distinct treatments. For each treatment combination, we collect information on $n = 4$ replicates, ensuring robustness in the observations.

The response variable of interest is the height of the plants (in cm), all of the same species, after a span of 30 days. We examine two critical factors: the amount of daily sunlight exposure,[65] and the type of fertilizer used.[66] The responses are shown below.

65: With three specific levels (4 hours, 8 hours, and 12 hours).

66: With three unique compositions (Type 1, Type 2, Type 3).

| **Height** (cm) | **Daily Sunlight Exposure** (hours) | | | | | |
|---|---|---|---|---|---|---|
| **Fertilizer** | **12** | | **8** | | **4** | |
| **Type 1** | 14.0 | 19.8 | 12.6 | 13.2 | 1.5 | 8.0 |
| | 14.9 | 13.6 | 9.6 | 12.5 | 4.8 | 5.5 |
| **Type 2** | 14.0 | 14.5 | 4.4 | 3.0 | 3.0 | 6.0 |
| | 8.4 | 17.0 | 9.0 | 6.5 | 9.2 | 4.8 |
| **Type 3** | 13.8 | 11.0 | 17.4 | 12.0 | 9.6 | 10.4 |
| | 16.8 | 16.0 | 15.0 | 13.9 | 8.2 | 6.0 |

The primary objective of the study is not only to tease out the individual and interactive effects of sunlight exposure and fertilizer composition on the plant's growth, but also to pinpoint whether a particular fertilizer type consistently supports optimal growth across sunlight conditions. The data is summarized below.

| **Height** (cm) | **Exposure** (hrs) | | | |
|---|---|---|---|---|
| **Dosage** | **12** | **8** | **4** | $y_{i,\bullet,\bullet}$ |
| **Type 1** | 62.3 | 47.9 | 19.8 | 130.0 |
| **Type 2** | 53.9 | 22.9 | 23.0 | 99.8 |
| **Type 3** | 57.6 | 58.3 | 34.2 | 150.1 |
| $y_{\bullet,j,\bullet}$ | 173.8 | 129.1 | 77.0 | 379.9 |

We can also create this table in R.[67]

67: We will use the `tidyverse` package this time around, just to show it can be done.

```
data = data.frame(
  Fertilizer = as.factor(c(rep("Type 1",4),rep("Type 2",4),rep("Type 3",4))),
  Height_12 = c(14.0, 14.9, 19.8, 13.6, 14.0, 8.4, 14.5, 17.0, 13.8, 16.8, 11.0, 16.0),
  Height_8 = c(12.6, 9.6, 13.2, 12.5, 4.4, 9.0, 3.0, 6.5, 17.4, 15.0, 12.0, 13.9),
  Height_4 = c(1.5, 4.8, 8.0, 5.5, 3.0, 9.2, 6.0, 4.8, 9.6, 8.2, 10.4, 6.0))
```

```
library(tidyverse)
summary.main <- data |> group_by(Fertilizer) |>
    summarise(h12 = sum(Height_12), h8 = sum(Height_8), h4 = sum(Height_4))
totals <- summary.main$h12 + summary.main$h8 + summary.main$h4
summary.big <- data.frame(cbind(summary.main[,c(2:4)], totals))
summary.end <- summary.big |>
    summarise(h12 = sum(h12), h8 = sum(h8), h4 = sum(h4), totals = sum(totals))
summary.data <- rbind(summary.big,summary.end)
rownames(summary.data) <- c("Type 1", "Type 2", "Type 3", "totals")
summary.data
```

```
         h12    h8   h4 totals
Type 1  62.3  47.9 19.8  130.0
Type 2  53.9  22.9 23.0   99.8
Type 3  57.6  58.3 34.2  150.1
totals 173.8 129.1 77.0  379.9
```

We can obtain the ANOVA table as follows.

```
a = nrow(summary.data) - 1
b = ncol(summary.data) - 1
n = nrow(data)/a
N = a*b*n

SST = sum(data[,c(2:(b+1))]^2)-summary.data[4,4]^2/N
SSA = sum(summary.data[b+1,c(1:a)]^2)/(b*n)-summary.data[4,4]^2/N
SSB = sum(summary.data[c(1:b),a+1]^2)/(a*n)-summary.data[4,4]^2/N
SSTR = sum(summary.data[c(1:b),c(1:a)]^2)/n-summary.data[4,4]^2/N
SSAB = SSTR - SSA - SSB
SSE = SST - SSTR
MSA = SSA/(a-1)
MSB = SSB/(b-1)
MSAB = SSAB/((a-1)*(b-1))
MSE = SSE/(a*b*(n-1))

ANOVA = as.data.frame(cbind(c(SSA,SSB,SSAB,SSE,SST),
                            c(a-1, b-1, (a-1)*(b-1), a*b*(n-1), N-1),
                            c(MSA,MSB,MSAB,MSE,0),
                            c(MSA/MSE,MSB/MSE,MSAB/MSE,0,0)))
rownames(ANOVA) = c("Treatment A", "Treatment B", "Interaction AB", "Error", "Total")
colnames(ANOVA) = c("SS", "df", "MS", "F0")
ANOVA
```

```
                      SS df         MS        F0
Treatment A    391.18722  2 195.593611 29.159629
Treatment B    106.83722  2  53.418611  7.963792
Interaction AB  96.13778  4  24.034444  3.583121
Error          181.10750 27   6.707685
Total          775.26972 35
```

**Hypothesis Testing** Before discussing the different hypothesis tests, we need the following results (see Exercises):

$$\mathrm{E(MSA)} = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\tau_i^2; \quad \mathrm{E(MSB)} = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_j^2;$$

$$\mathrm{E(MSAB)} = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{i,j}^2; \quad \mathrm{E(MSE)} = \sigma^2.$$

In general, we may be interested in the following tests:

- presence/absence of interactions between Factor A and Factor B;
- presence/absence of a Factor A effect;
- presence/absence of a Factor B effect.

The hypothesis of **absence of interaction** between Factors A and B can be formulated as

$$H_0^{AB} : \mu_{i,j} - \overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,j} + \overline{\mu}_{\bullet,\bullet} = (\tau\beta)_{i,j} = 0, \quad i = 1,\ldots,a; j = 1,\ldots,b.$$

In the absence of interaction, the **difference between averages** obtained by varying either Factor A or Factor B is the same regardless of the level of the other factor:

$$\begin{aligned}&\mu_{i,j} - \mu_{i,j'} = \mu_{i',j} - \mu_{i',j'}\\ &\mu_{i,j} - \mu_{i',j} = \mu_{i,j'} - \mu_{i',j'}, \quad i,i' = 1,\ldots,a,\ j,j' = 1,\ldots,b.\end{aligned}$$

The **absence of effect for Factor A** can be formulated as

$$H_0^A : \overline{\mu}_{i,\bullet} - \overline{\mu}_{\bullet,\bullet} = \tau_i = 0, \quad i = 1,\ldots,a.$$

In the absence of interaction, we can rewrite the hypothesis as

$$H_0^A : \mu_{i,j} = \mu_{i',j}, \quad i,i' = 1,\ldots,a,\ j = 1,\ldots,b,$$

which corresponds to the intuitive notion of the absence of effect of Factor A.

Similarly, the **absence of effect for Factor B** can be formulated as

$$H_0^B : \overline{\mu}_{\bullet,j} - \overline{\mu}_{\bullet,\bullet} = \beta_j = 0, \quad j = 1,\ldots,b.$$

In the absence of interaction, we can rewrite the hypothesis as

$$H_0^B : \mu_{i,j} = \mu_{i,j'}, \quad i = 1,\ldots,a,\ j,j' = 1,\ldots,b,$$

which corresponds to the intuitive notion of the absence of effect of Factor B.

The hypotheses $H_0^{AB}$, $H_0^A$ and $H_0^B$ use, respectively, the following tests:

$$F_{AB} = \frac{\mathrm{MSAB}}{\mathrm{MSE}} \sim F_{(a-1)(b-1),N-ab}(\delta_{AB}^2), \quad \delta_{AB}^2 = \frac{n}{\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{i,j}^2;$$

$$F_A = \frac{\mathrm{MSA}}{\mathrm{MSE}} \sim F_{a-1,N-ab}(\delta_A^2), \quad \delta_A^2 = \frac{bn}{\sigma^2}\sum_{i=1}^{a}\tau_i^2;$$

$$F_B = \frac{\mathrm{MSB}}{\mathrm{MSE}} \sim F_{b-1,N-ab}(\delta_B^2), \quad \delta_B^2 = \frac{an}{\sigma^2}\sum_{j=1}^{b}\beta_j^2.$$

When $H_0^{AB}$, $H_0^A$, and/or $H_0^B$ hold, we note that $E(\text{MSAB})$, $E(\text{MSA})$, and/or $E(\text{MSB})$ take on the value $\text{MSE} = \sigma^2$, respectively. Thus, **large values** of $F_{AB}$, $F_A$, and/or $F_B$ imply that the observations **do not support** the corresponding null hypotheses.

More generally, When $H_0^{AB}$, $H_0^A$, and/or $H_0^B$ hold, the corresponding test statistic $F_{AB}$, $F_A$, and/or $F_B$ follow a **central $F$-distribution**. Thus, we **reject** $H_0^{AB}$, $H_0^A$, and/or $H_0^B$, respectively, at **significance** $\alpha$ if

$$\begin{aligned} AB &: F_0 > F_{\alpha;(a-1)(b-1),N-ab}; \\ A &: F_0 > F_{\alpha;a-1,N-ab}, \quad \text{and/or} \\ B &: F_0 > F_{\alpha;b-1,N-ab}. \end{aligned}$$

In practice, we start by testing the **absence/presence of interactions**. If the interaction is **not significant**, then we perform the tests corresponding to treatment effects for Factors A and B.[68]

68: In the latter case, the hypotheses $H_0^A$ et $H_0^B$ can easily be interpreted; when the interaction is statistically significant, the interpretation of the treatment effect may be more challenging.

**Example** In the plant growth example, we have $F_{AB} = 3.58$, $F_A = 29.16$, and $F_B = 7.96$. At significance level $\alpha = 0.05$, we find:

```
qf(0.05, df1=(a-1)*(b-1), df2=N-a*b, lower.tail=FALSE)
qf(0.05, df1=a-1, df2=N-a*b, lower.tail=FALSE)
qf(0.05, df1=b-1, df2=N-a*b, lower.tail=FALSE)
```

```
[1] 2.727765
[1] 3.354131
[1] 3.354131
```

Since $3.58 > F_{0.05,4,27} = 2.73$, we reject $H_0^{AB}$ and conclude that the interaction is significant at $\alpha = 0.05$. Also, since $7.96 > F_{0.05,2,27} = 3.35$ and since $29.16 > F_{0.05,2,27} = 3.35$, we reject both $H_0^A$ and $H_0^B$, but it is not as obvious what the means for the data. □

### 11.6.2 Model Validation

The three basic model assumptions are still that the errors are **independent**, **normally distributed**, and have **constant variance**. As we have done before, we would use the residuals in lieu of the errors to validate these assumptions.

In the two-way balanced factorial design, the **predicted values** are given by

$$\hat{y}_{i,j,k} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \widehat{(\tau\beta)}_{i,j} = \overline{y}_{i,j,\bullet};$$

the **model residuals** are thus given by

$$e_{i,j,k} = y_{i,j,k} - \hat{y}_{i,j,k} = y_{i,j,k} - \overline{y}_{i,j,\bullet}.$$

### 11.6.3 Model Without Interaction

In the **absence of interaction**, the model simplifies to

$$y_{i,j,k} = \mu + \tau_i + \beta_j + \varepsilon_{i,j,k},\ i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n.$$

In that case, the estimators of the model parameters are given by

$$\begin{aligned}
\hat{\mu} &= \overline{y}_{\bullet,\bullet,\bullet} \\
\hat{\tau}_i &= \overline{y}_{i,\bullet,\bullet} - \overline{y}_{\bullet,\bullet,\bullet},\ i = 1, \ldots, a \\
\hat{\beta}_j &= \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,\bullet,\bullet},\ j = 1, \ldots, b,
\end{aligned}$$

and the decomposition of the total sum of squares is

$$\begin{aligned}
\text{SST} &= \text{SSA} + \text{SSB} + \text{SSE} \\
N - 1 &= (a-1) + (b-1) + [(a-1)(b-1) + ab(n-1)] \\
&= (a-1) + (b-1) + (N - a - b + 1).
\end{aligned}$$

The treatment sums of squares SSA and SSB are identical to those in the ANOVA model with interaction. The simpler formulas collapse to:

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{i,j,k}^2 - \frac{y_{\bullet,\bullet,\bullet}^2}{N};\quad \text{SSA} = \sum_{i=1}^{a}\frac{y_{i,\bullet,\bullet}^2}{bn} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};$$

$$\text{SSB} = \sum_{j=1}^{b}\frac{y_{\bullet,j,\bullet}^2}{an} - \frac{y_{\bullet,\bullet,\bullet}^2}{N};\quad \text{SSE} = \text{SST} - \text{SSA} - \text{SSB}.$$

The corresponding ANOVA table is given below:

| Source | SS | df | MS | F |
|---|---|---|---|---|
| **Treatment A** | SSA | $a-1$ | MSA | $F_A = \text{MSA/MSE}$ |
| **Treatment B** | SSB | $b-1$ | MSB | $F_B = \text{MSB/MSE}$ |
| **Error** | SSE | $N-a-b+1$ | MSE | |
| **Total** | SST | $N-1$ | | |

**Table 11.29:** ANOVA table for equality of factorial effects, with no interaction effects, in a two-way design.

We test for the null hypotheses

$$H_0^A : \mu_{i,j} = \mu_{i',j} \quad \text{and} \quad H_0^B : \mu_{i,j} = \mu_{i,j'}$$

using the test statistics

$$F_A = \frac{\text{MSA}}{\text{MSE}} \sim F_{a-1,N-a-b+1}(\delta_A^2),\quad \delta_A^2 = \frac{bn}{\sigma^2}\sum_{i=1}^{a}\tau_i^2,$$

$$F_B = \frac{\text{MSB}}{\text{MSE}} \sim F_{b-1,N-a-b+1}(\delta_B^2),\quad \delta_B^2 = \frac{an}{\sigma^2}\sum_{j=1}^{b}\beta_j^2.$$

The analysis of the residuals is based on the following residuals

$$e_{i,j,k} = y_{i,j,k} - \hat{y}_{i,j,k} = y_{i,j,k} - (\hat{\mu} + \hat{\tau}_i + \hat{\beta}_j) = y_{i,j,k} - \overline{y}_{i,\bullet,\bullet} - \overline{y}_{\bullet,j,\bullet} + \overline{y}_{\bullet,\bullet,\bullet}.$$

The rest of the analysis proceeds as before.

### 11.6.4 Multiple Comparisons

As in previous sections, we may want to perform **multiple comparisons**. More often than not, we are interested in constructing **simultaneous confidence intervals** that compare the effects for each factor.

Throughout, recall that we estimate $\sigma^2$ by

$$s^2 = \text{MSE} = \frac{\text{SSE}}{N-ab} = \frac{\text{SSE}}{ab(n-1)}.$$

Suppose that we are interested in all possible pairwise comparisons for treatment A; in that case, there are $K = \binom{a}{2} = a(a-1)/2$ possible pairs to test. For treatment B, there are $L = b(b-1)/2$ possible pairs to test.

We could use the **Bonferroni procedure** to do so; the simultaneous confidence intervals corresponding to Factor $A$ take the form

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm t_{\alpha/(2K),N-ab}\sqrt{\text{MSE}}\sqrt{\frac{2}{bn}},$$

and those for Factor $B$, the form

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm t_{\alpha/(2L),N-ab}\sqrt{\text{MSE}}\sqrt{\frac{2}{an}}.$$

If instead we use **Tukey's method**, the simultaneous confidence intervals corresponding to Factor A are given by

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm q_{\alpha;a,N-ab}\sqrt{\text{MSE}}\sqrt{\frac{1}{bn}},$$

and those for Factor $B$, by

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm q_{\alpha;b,N-ab}\sqrt{\text{MSE}}\sqrt{\frac{1}{an}}.$$

For **Scheffé's approach**, the simultaneous confidence intervals corresponding to Factor A are

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm \sqrt{(a-1)F_{\alpha;a-1,N-ab}}^{1/2}\sqrt{\text{MSE}}\sqrt{\frac{2}{bn}},$$

and those for Factor B,

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm \sqrt{(b-1)F_{\alpha;b-1,N-ab}}^{1/2}\sqrt{\text{MSE}}\sqrt{\frac{2}{an}}.$$

For the two-way balanced factorial model **without interaction**, the simultaneous confidence intervals are similar, except that the number of degrees of freedom in the residual sum of squares SSE is now $N-a-b+1$. In that case, the estimator of $\sigma^2$ is

$$\tilde{s}^2 = \tilde{\text{MSE}} = \frac{\text{SSE}}{N-a-b+1}.$$

For instance, the simultaneous confidence intervals for Factor A obtained

using **Tukey's method** are given by

$$\tau_i - \tau_{i'} : \overline{y}_{i,\bullet,\bullet} - \overline{y}_{i',\bullet,\bullet} \pm q_{\alpha;a,N-a-b+1}\sqrt{\tilde{\text{MSE}}}\sqrt{\frac{1}{bn}},$$

whereas the simultaneous confidence intervals for Factor B obtained *via* **Scheffé's approach**, say, are given by

$$\beta_j - \beta_{j'} : \overline{y}_{\bullet,j,\bullet} - \overline{y}_{\bullet,j',\bullet} \pm \sqrt{(b-1)F_{\alpha;b-1,N-a-b+1}}\sqrt{\tilde{\text{MSE}}}\sqrt{\frac{2}{an}}.$$

### 11.6.5 Factorial Designs with Multiple Factors

The two-way factorial design can be naturally extended to **multiple factors**. For instance, the **three-way factorial design** $a \times b \times c$ is:

$$y_{i,j,k,l} = \mu_{i,j,k} + \varepsilon_{i,j,k,l}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, c;\ l = 1, \ldots, n$$

where

$$\mu_{i,j,k} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{i,j} + (\tau\gamma)_{i,k} + (\beta\gamma)_{j,k} + (\tau\beta\gamma)_{i,j,k}.$$

With three factors, we can explore **second-order interactions** $AB$, $AC$, and $BC$, or the **third-order interaction** $ABC$. Such designs are out of scope for these course notes,[69] more details are available in [2, 5].

69: The ANOVA table for the $a \times b \times c$ design has 9 rows, but is otherwise what one would expect to see.

## 11.7 Exercises

1. Conduct an analysis of the paint example of Section 11.2.1 assuming that the samples are independent (unpaired test). Compare with the results of the paired test on the same data.
2. Recreate the analysis of the apparatus example of Section 11.2.4 using R. What if the sample sizes were $n_1 = 25$ and $n_2 = 30$, instead?
3. Show directly that the decomposition SST = SSA + SSE of one-way classification holds.
4. In a one-way classification model with $a = 2$, show that the power of the $F-$test is maximized when $\frac{1}{n} + \frac{1}{N-n}$ is minimized.
5. Use the least square estimation principles to establish the normal equations, and estimate the parameters in the unbalanced one-way classification model. What are the estimated treatment effects and the estimated difference between treatments in that case? What about their confidence intervals?
6. Compute the ANOVA table for the completely randomized unbalanced design in the Kenton Food Company example.
7. In the one-way random-effects ANOVA model, show that E(MSE) = $\sigma^2$ and E(MSA) = $\sigma^2 + n\sigma_T^2$.
8. In the one-way random-effects ANOVA model, show that

$$\frac{(a-1)\text{MSA}}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{a-1}.$$

9. In a two-factor RCBD, show that

$$\text{E(MSA)} = \sigma^2 + \frac{b}{a-1}\sum_{i=1}^{a}\tau_i^2,$$

$$\text{E(MSB)} = \sigma^2 + \frac{a}{b-1}\sum_{j=1}^{b}\beta_j^2.$$

10. Verify if the RCBD model assumptions are met for the credit card example.
11. Show directly that the total sum of squares in a balanced two-way factorial design breaks down as

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}.$$

12. In the two-way balanced factorial design, show that

$$\text{E(MSA)} = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\tau_i^2;$$

$$\text{E(MSB)} = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_j^2;$$

$$\text{E(MSAB)} = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{i,j}^2;$$

$$\text{E(MSE)} = \sigma^2.$$

13. In the medical agent example (two-way factorial design), is the interaction effect significant at $\alpha = 0.05$? What about the dosage effect? The application period effect?
14. Produce simultaneous confidence intervals at family significance $\alpha$ for treatment effects (Factors A and B) in the medical agent and plant growth examples.

## Chapter References

[1] M.H. Kutner et al. *Applied Linear Statistical Models*. McGraw Hill Irwin, 2004.

[2] D.C. Montgomery. *Design and Analysis of Experiments*. Wiley, 2012.

[3] L. Ott and M. Longnecker. *A First Course in Statistical Methods*. Thomson-Brooks/Cole, 2004.

[4] R.L. Ott and M.T. Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning, 2015.

[5] H. Scheffé. *Analysis of Variance*. London: John Wiley and Sons Inc., 1959.

# 12 Simulations and Modeling

by **Jen Schellinck** and **Patrick Boily**


12.1 Introduction . . . . . 803
Static Models . . . . . 805
Dynamic Models . . . . . 808
Uses, Data, Contrast . . . . . 809
Simulation Types . . . . . 813
12.2 Modeling Strategies . . . . . 815
Information Gathering . . . 815
Conceptual Model . . . . . 816
Building the Model . . . . . 818
Verification and Validation 818
Analysis of Results . . . . . 818
12.3 Practical Considerations . . 820
Computational Complexity 820
Applications . . . . . 820
Software . . . . . 822
12.4 Case Study: NWMO . . . . . 822
12.5 Exercise . . . . . 825
Chapter References . . . . . 825


Modeling plays a central role in a wide range of quantitative endeavours. To thrive as a successful quantitative analyst and consultant, it is crucial to grasp the various types of modeling and models, grasp their similarities and distinctions, and identify suitable applications.

However, due to its pervasive presence across the quantitative spectrum, the significance of modeling is often underestimated and taken for granted, in part because it serves as the foundation of, and is integrated into, numerous techniques.

In reality, quantitative analysts and consultants are inherently modelers. As such, possessing a solid overall understanding of modeling ( beyond mastering specific techniques) and being able to construct models in a broader sense greatly enhances various quantitative undertakings.

## 12.1 Introduction

**Analogical reasoning** is the act of reasoning from one specific occurrence to another specific occurrence, on the basis of similarity. For example,

> [HAND:FINGERS, FOOT:—].

A major benefit of this type of reasoning is that it can reveal new aspects or relationships between objects that have not previously been considered.

Clearly, the choice of objects used in an analogy is important:

> [HAND:FINGERS, ORANGE:—]

likely yields little useful insight, but

> [HAND:FINGERS, PLANT STEM:—]

might be more interesting (see Figure 12.1).

Analogical reasoning is viewed by some as a primary **cognitive strategy**, underlying much of human cognition [7, 6, 4].

Keeping this context in mind, a model is simply an independent entity, or structure, that has useful similarities to another structure of interest, and which allows for analogical reasoning. This structure of interest is referred to as the target of the model.

We can carry out inductive or deductive reasoning on the model and then, via analogical reasoning, transfer our insights about the model over to the target, and in this way learn something about the target. The target

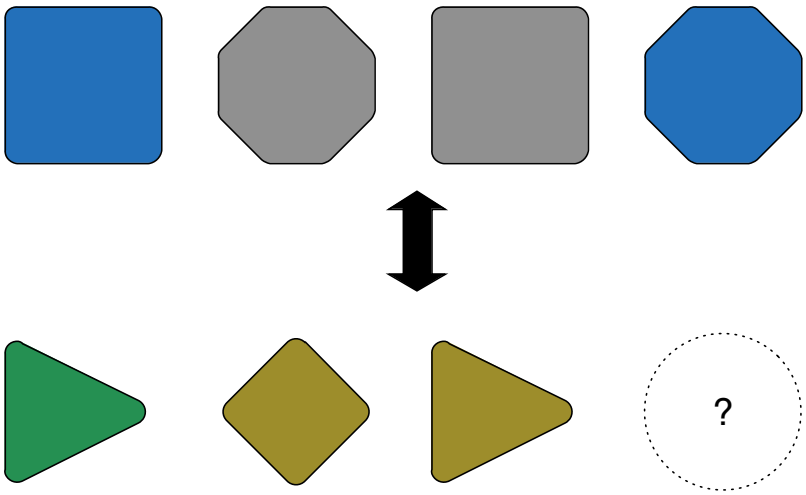



**Figure 12.1:** Can we draw an analogy between the top row of shapes and the bottom row of images? What should the shape and colour of the last image in the bottom row be?

structure might be a single object or a system of objects, or a process being carried out by this system of objects.

Our ability to create a model with *useful similarities* to the target system, and then learn about our chosen target system using this model, can be extremely powerful.

For instance, we can make a very small model of something that is, in reality, very large or very distant – for example, a small scale model of the solar system, made out of wire and styrofoam – and use this small simple model to come up with accurate predictions about this large and distant system.

The solar system model example also showcases the importance of understanding **which parts of the model are usefully similar to the target system** in the context of our intended use of the model. If we try to use our simple solar system model to draw conclusions relating to the relative densities of planets in the solar system, we will be disappointed.

Although there are many different types of models, which we will further discussed later, in general we can say that models have two main functions: **explanation** and **prediction**.

- In some cases, we might have a system whose behaviour we do not fully understand and cannot explain. Models can help us increase our understanding of the mechanisms underlying the behaviours or properties of interest.
- In other cases, regardless of how a type of system is generating a particular behaviour, or came to have a certain property, our interest is not in understanding how this came to be, but rather in predicting the presence (or absence) of that behaviour or property in another system of the same type.

Modelers often try to create **taxonomies** or categorisations of models. These efforts have arguably not been that successful from a conceptually rigorous point of view but, pragmatically, it is still useful to consider the types of models that people commonly use and discuss (see [16] for a useful review and discussion of a variety model and simulation types).

It has been our experienced that clients and stakeholders usually take a dim view of simulations, as though they are somehow less 'valid' or

'real' than other quantitative approaches.[1] This is worth remembering when producing simulation solutions.

1: The reasons for this are varied, and perhaps not entirely unfounded as simulations can easily be used in the wrong way or with the wrong endgame in mind.

### 12.1.1 Static Models

At the heart of simulations lies the concept of a **model**. Models serve as essential tools in understanding **systems**, employing various strategies. Ultimately, their purpose is to enhance the modeler's comprehension of a system, using the term "system" in an axiomatic sense.

#### Conceptual Models

A conceptual model is an **abstraction of a real world system** or process that defines which **elements** of the system or process are of interest in the current context, and how these elements and their **relationships** will be defined for the purposes of **drawing conclusions** about the behaviours or properties of the system.

Arguably, before any other type of model can be generated, a conceptual model must first be created, either implicitly or explicitly.

**Explicit conceptual models** may take the form of diagrams or formalized descriptions of the system. Conceptual models may then be implemented as other types of models (e.g. mathematical, simulation).

**Implicit conceptual models** are often linked with gaps in the understanding of a system – assumptions that go unchallenged and unstated are often less clear and obvious than is originally believed. An engineer may, for instance, state to a consultant that the probability of a certain component failing by time $t$ is 0 without feeling the need to specify that, in the jargon of the discipline, this really means that

$$P(\text{failure by time } t > T) = \varepsilon > 0,$$

for a "sufficiently large" $T$ and a "sufficiently small" $\varepsilon$; the consultant, not knowing the conventions of the field, might mistake this for

$$P(\text{failure by time } t) = 0 \quad \text{for all } t;$$

if not cleared up, the misunderstanding can propagate through the simulation, potentially making it **useless in practice**.

#### Mathematical Models

A mathematical model uses mathematics to **support reasoning** about a real world system. Relationships between objects in the system,[2] are represented by mathematical relationships between variables.

2: Or their properties.

If the relationships within the mathematical model are **sufficiently similar** to relationships between objects in the system of interest, then carrying out **truth-preserving mathematical manipulations** on the model should result in **valid new conclusions** about the system.

Arguments represented by **symbolic logic** also fall under this category. As a result, it could readily be said that all models implemented on computers are a type of mathematical model. That being said, the expression 'mathematical model' typically refers to models that are not necessarily implemented on computers, and which consist of systems of mathematical equations.

Although mathematical models may represent processes and dynamic elements of systems by including time and space as variables, the models themselves are **static**, in the sense that they do not change over time in a manner that is similar to the ways in which the target system itself changes over time.

Mathematical models may still be implemented on computers and methods for **solving the systems of equations** in these models (e.g. symbolic manipulations, numerical analysis) may be carried out using computer algorithms.

Nevertheless, it is important to remember that although both the work performed on a computer and simulations take place within a computational environment, finding solutions to equations through programmatic strategies should not be conflated with the conventional understanding of "simulations". We will elaborate on this topic in the subsequent discussion.

#### Statistical Models

Conceptually, statistics help us represent the world in terms of **populations** and **processes**, which have certain properties that can be themselves be represented using mathematical expressions. **Statistical models** could thus be described as mathematical models motivated by a certain (statistical) conceptualisation of real world processes.

#### To-Scale Physical Models

A to-scale physical model is a model that is constructed from **physical materials**, which are shaped and positioned in such a way as to accurately represent the physical layout, positions, and sizes of elements of the target system, as well as relative to each other (see Figure 12.2 for an example of an architectural model).

#### Data Models

A data model is a conceptual model used to design the structure of data storage. Since data itself represents facts about a system, it is appropriate to first conceptually model the properties and relationships that exist within the system, and which are represented by the data, and then use this conceptual model to create a data storage structure that can be used to efficiently **hold**, **extract**, **edit** and **add** to the stored data (see Figure 12.3 for an example).

**Figure 12.2:** To-scale architectural model of the interior of an office building [5].

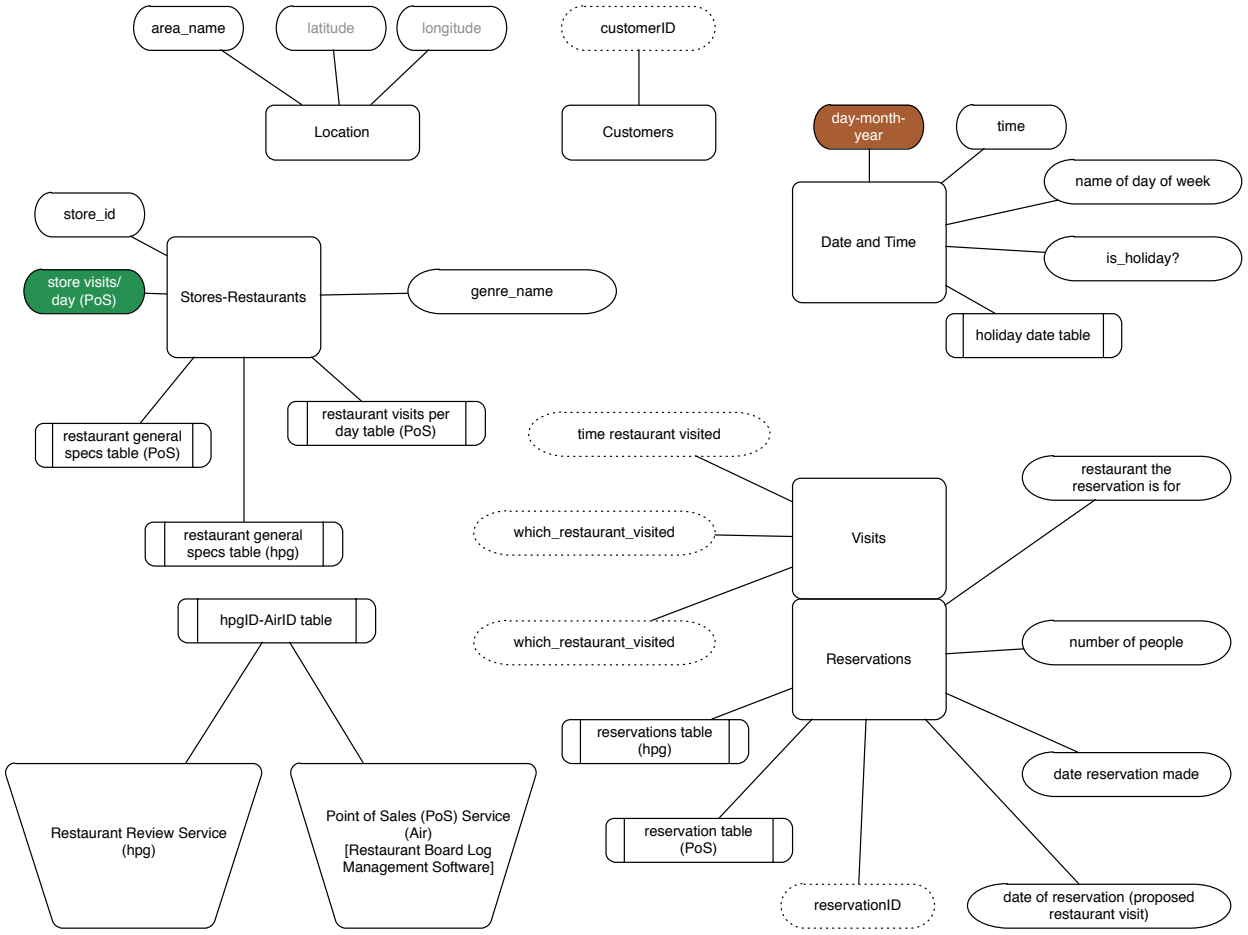



**Figure 12.3:** A preliminary data model of a restaurant reservation system, which can be used to help design an efficient data storage structure, as well as develop data analysis strategies.

### 12.1.2 Dynamic Models

In some situations, only the static aspects of a system are interesting, or the system itself is mostly static.

For instance, if we build a physical model of a house, we expect both the house and the model to be **relatively unchanging** – the measurements of the rooms and the furniture in the house will not change from minute to minute,[3] and the model will not need to change either.

3: Although they could change over years due to remodeling, or even drastically if the house is sold to new owners with a different sense of aesthetics.

We can then use the model to reason about the house:

- if the model couch fits against this wall in the model house, we can reason that the real couch will fit in the real house;
- if eight model chairs cannot simultaneously be placed around the model kitchen table, than eight real chairs cannot be simultaneously be placed around the real kitchen table, etc.

Other systems, however, are more active, or **dynamic**, with processes taking place within the system. When modeling these dynamic system elements, we often talk about **simulation models** or simply **simulations**. Although the term 'simulation' is not precisely defined, it typically indicates that a model is intended to **reflect the behaviour of the target system** – its processes – over time, and also that the **model itself will independently change over time**, when it is run.

The goal is to construct the simulation in such a way that it will change over time in ways that are similar to the manner in which the system itself changes over time. As a result we can use the simulation to predict past, current, and future behaviours of the system.

Historically, simulations have often modeled **individual object-level properties and behaviours**, as well as the mechanisms underlying relevant behaviours, rather than **group-level properties or system outputs**, but this does not have to be the case.

#### Modeling Time and Movement

How do we incorporate time and movement into a model? To return to our styrofoam and wires model of the solar system, if we set it up so that when we turn a crank the planets and moons move realistically around a light bulb in the centre of the model (representing the sun), then we have a dynamic model, or simulation of the solar system. We can simulate what will happen within the actual solar system over time.

As another example, if we wish to know how emergency responders might behave in different plane crash scenarios, we could set up a number of simulated crash scenarios, with a life-size model of a crashed plane, and actors behaving as injured people might. We can then have the emergency responders try out (i.e., simulate) different approaches and strategies to dealing with plane crashes.

The advent of computers greatly facilitated the construction and possible uses of simulations, because it made it possible to simulate dynamic systems **virtually** instead of having to create a dynamic physical model of the system, whose elements could be represented as data structures (and variables within these structures) within computer programs. The

physical interactions between these system elements could then, in turn, be represented by logical rules and mathematical equations operating over these data structures.

These logical rules and mathematical equations pushed computer simulations closer to the domain of mathematical models, relative to physically constructed models. At the same time, computer simulations retained the strategy used by these physical models of determining what would happen to the system by moving the model through its expected behaviours step-by-step, over time.

Rather than mechanically moving the model (or using people and other elements in this capacity) computer models rely on the computer processor to run the program that represents the system, and essentially 'move' (in an electronic sense) the model based on the behaviours the model implements. As discussed earlier, this is a different technique than the one used by mathematical models implemented on computers.

### 12.1.3 Uses, Data, and Contrast with Mathematical Modeling

Simulations are typically used to

- better understand actual real-world phenomena and systems, and
- explore phenomena that don't currently exist but which could exist hypothetically.

Simulations can allow us to both **predict** what our target system will do under particular circumstances, but also **explain** why a system behaves the way it does. However, given that we build simulations using only what is already known (or possibly suspected) to be either currently the case about the system, or at least plausible within the conceptual phase space in which the system resides, you may wonder how a simulation could possibly tell us anything new about the system, and thus, why we would ever bother running simulations.

Human thinking is typically unable to capture all the possible interactions between a system's various parts, and how these parts influence each other in particular circumstances; **merely** thinking through the behaviours of a system which is even slightly complicated is likely lead us to miss implications, and, as a result, **incorrectly predict or explain** the system's behaviour. If, instead, we introduce what we do know into the simulation and allow it to behave based on these rules, behaviours that we would not easily have anticipated can emerge from the process.

Consequently, the notion of **emergence** is crucial in simulations. We can say that simulation behaviours emerge when they are not programmed in the simulation directly, but rather occur as the result of interactions between model components that are themselves programmed into the simulation directly.

The emergent behaviours may occur at different **levels of granularity** of the system. For example, if we create a simulation of people in a city, we might see emergent behaviour with respect to which people most frequently interact with which other people, and we might also see emergent behaviour at the population level, where the average number of people in a given location is equal to a particular value over time.

We can see from this example how emergence allows us both to predict and to explain elements of a system that were not previously amenable to such efforts. We can predict average numbers of people in a particular location, if this information is not available from another source; if it is, we can still use the simulation to explain the origins and underpinnings of this number, by referring to the more granular system components whose interactions lead to the value.

'Emergence' is a concept that has crossed-over into a large number of areas of human endeavour. Don't be surprised to hear clients and end users talk about "emergent phenomena" in contexts where you would not normally expect to hear it.[4] Be sure to **clarify the situation** at an early stage (in the proposal, say) in order to avoid the confusion and headaches that can result when deliverables are handed off.

4: It is quite conceivable that they have a very thorough understanding of what emergence means and what it entails – don't make the classic quantitative consulting mistake of assuming that clients do not understand technical concepts ... you never know what their background and interests are – but, together with terms like 'synergy' or 'big data', it seems to have entered the business lexicon as a trendy but ultimately meaningless term.

### Simulations and Data

All modeling activities rely on the modeler having **accurate** and **relevant** information or data about the target system, which allows for the construction of a model with useful similarities to the target system, which is basically a data collection/information gathering problem. But even then, simulations have a particular relationship with data:

- first and foremost, data is needed in order to properly set simulation parameters – the initial simulation settings that determine how the simulation will run in a particular instance; in the absence of this type of information, although the simulation may generate outputs that could, in principle, have some relevance to the target system in some circumstances, the simulation behaviour is unlikely (or at least, should not be expected) to overlap with target system behaviours of interest within the specific context in which the simulation was generated;
- secondly, simulations have the capacity to generate large amounts of data about the behaviour of the simulation, and by extension, the target system. This data, sometimes referred to as 'synthetic data' or 'simulated data', can be uses as a stand-in for actual data about the system, just as the model is being used as a stand-in for the target system.

When very little is know about reasonable parameters values, a preliminary simulation might first be required in order to produce data which could then be used to set simulation parameters, which, in turn, could be used to produce data for analysis.

It is not too difficult to conceive of multiple links being added to this chain; our advice is to keep the number of such links to a minimum (preferably zero) – in light of the point made in the first item above, it might be preferable to garner information about parameters from **first principles** (or other models).

### Simulations vs. Mathematical Models

The procedural element of computer models, whereby the behaviour of the target system must be, in a sense, mechanically replicated by the

data structures and procedures of the computer program, distinguishes **computer simulations** from mathematical models, which, rather than modeling the temporal, dynamic components of systems by incorporating a temporal, dynamic component directly into the model, instead represent them as variables in mathematical equations that represent components and behaviours of the system.

On this front, the advantage of mathematical models is that **deductive reasoning** (or first principles reasoning) can, in theory, be used to determine the target system behaviour, rather than have to resort to 'running' the model over a range of starting conditions. This is appealing, as mathematical strategies can allow for more definitive and general statements about the system (e.g. "The system will never do the following"; "The system will always do the following", etc.); these types of statements are typically outside the reach of even the most advanced mechanical or programmatic simulations. In practice, however, the underlying complexity of such models limit the usefulness of this approach in most scenarios.

**Example** Consider, for instance, the $n-$**body problem** ($n$BP) of classical mechanics, which consists in predicting the individual trajectories of $n$ celestial bodies bound by gravitational attraction.

Using Newtonian mechanics, the trajectories can be deduced to follow the paths described by the following system of differential equations:

$$m_1 \frac{\mathrm{d}^2\mathbf{q}_1}{\mathrm{d}t^2} = \sum_{j\neq 1} \frac{Gm_1m_j(\mathbf{q}_j - \mathbf{q}_1)}{\|\mathbf{q}_j - \mathbf{q}_1\|^3}, \cdots, m_n \frac{\mathrm{d}^2\mathbf{q}_n}{\mathrm{d}t^2} = \sum_{j\neq n} \frac{Gm_nm_j(\mathbf{q}_j - \mathbf{q}_n)}{\|\mathbf{q}_j - \mathbf{q}_n\|^3},$$

where $m_i$ and $\mathbf{q}_i(t)$ are, respectively, the mass and the trajectory of the $i^{\text{th}}$ celestial body in 3-space, and $G$ is Newton's constant. These equations describe, in principle, the behaviour of stars in a globular cluster, say, or of the Earth-Sun or the Earth-Moon system.

They cannot provide a **complete** description as the range of gravitational attraction is infinite – every 'object' in the Universe influences every 'other' object to some extent, no matter how distant,[5] and other forces/factors may also act on the bodies,[6] but for most practical applications,[7] they are more than sufficient as long as we are willing to ignore relativistic effects.

5: At large distances, the force due to gravity overwhelms the other 3 forces, however.

6: See the precession of Mercury, for instance.

7: If one can consider astronomy a practical discipline.

What do the solutions look like? A typical mathematical approach would be to try to solve the 2BP, and to see if the solution can be generalized to more complex cases.

The **two-body problem** has an exact solution. The **centre of mass** of the two bodies is the vector

$$\mathbf{x}(t) = \frac{m_1\mathbf{q}_1(t) + m_2\mathbf{q}_2(t)}{m_1 + m_2}.$$

In the 'centre-of-mass frame',[8] physical conservation laws show that the trajectories of the two bodies are co-planar and 'orbit' the system's **barycentre**, with an angle $\theta_i(t)$ depending on the **reduced mass** of the system $m_* = \frac{m_1m_2}{m_1+m_2}$ and on the **effective potential** $U(r(t), \ell, m_*)$, where $r(t) = \|\mathbf{q}_2 - \mathbf{q}_1\|$ and $\ell$ is the system's angular momentum.

8: That is, in the frame that moves along with the centre of mass.

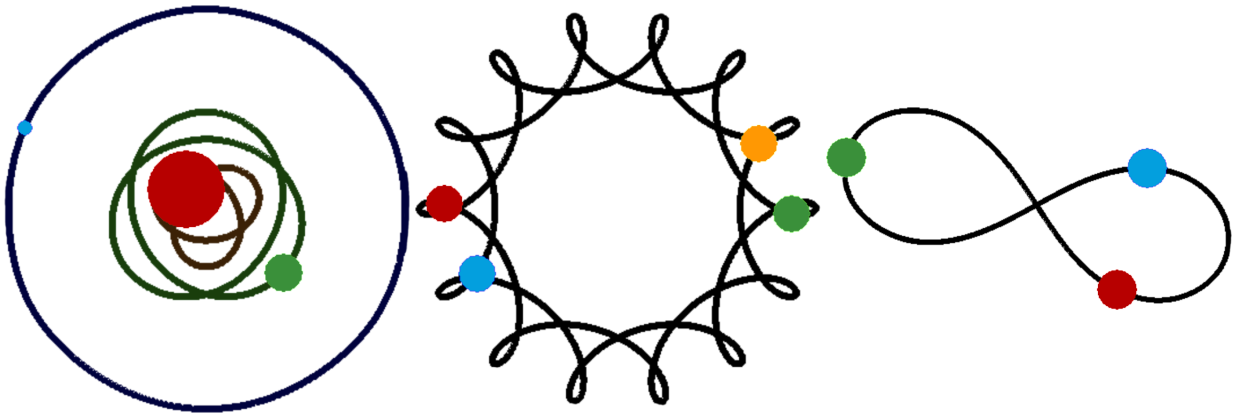


**Figure 12.4:** Possible solutions of the $n-$body problem, based on [14]: planetary 3BD (left), 4BP pairs of bodies orbiting each other (middle), 3−body choreography (right).

Various combinations of parameters lead to various orbits; if the effective potential admits a local minimum, for instance, the orbits will oscillate around the barycentre;[9] if the effective potential does not admit a minimum, then the orbits may **escape to infinity**.[10]

9: Elliptic or precessing elliptic paths, in each Sun-planet system.

10: Hyperbolic or parabolic paths, such as in some Sun-comet systems.

Under some restrictions on the masses and momenta of the bodies, the $n$BP can be shown to have closed-form solutions or theoretically understood approximate solutions (see [8] for a list, and Figure 12.4 for some illustrations), including:

- **Euler's Problem of Two-Fixed Centres** allows for colinear motion in systems where two of the three masses are comparable and fixed;
- the **restricted** 3BP shows the existence of 5 fixed configurations (involving the **Lagrangian points**) which rotate around the system's barycentre in cases where one of the masses is negligible, such as is the case in the Sun-Jupiter-Trojans systems (there are two);
- the **planetary** $n$BP admits quasi-periodic solutions in systems where one of the masses is significantly larger than the other $n-1$ masses, which shows that planets in stable, planar, and nearly circular orbits around a star *can* transition to chaotic orbits, but that these orbits would be bounded by quasiperiodic tori and so would preserve some regularity, and
- $n-$**body choreography** in which all the masses move on the same manifold, without collisions.

The **general** $n-$body problem can be solved analytically using Taylor Series (known as **Sundman**'s series), but the series converge so slowly as to be of no practical use for astronomical results.[11]

11: Which would require at least $10^{8000000}$ terms in the 3BP case, well beyond even what modern computers can produce [1].

By contrast, in order to draw conclusions from a simulation we must first set certain initial conditions and then run the simulation and examine the resulting output. Each simulation run represents only one specific instance in the model space. As a result, it can be difficult, if not downright impossible, to draw general conclusions from the results of one or even multiple simulation runs.[12]

12: To say nothing of exploring the outcome of using different parameter values.

This has lead to criticism over the use of simulations in some milieus, on the basis that simulations should **never** be used if mathematical models can be used instead.

However, the $n$BP illustrates why taking this hard-line position may be inadvisable; clearly, there are circumstances in which it is difficult to create solvable (actionable) mathematical models that represents the target system in ways sufficiently similar to the system in relevant respects in order to for salient and accurate conclusions to be drawn about that system, in which case a simulation might provide greater insight.

It is also possible to create **hybrids** of mathematical and simulation models to allow for increased insight into system behaviours.

If $n$ is relatively small, the $n$BP trajectories can be approximated to a high-level of accuracy by using numerical methods to solve the corresponding system of differential equations.[13] For astronomical bodies that avoid collisions (or near encounters), there are two main technical issues:

- the first one is that the $n$BP problem is **chaotic** for $n > 2$,[14] so that small errors such as can be generated by truncating initial conditions or intermediate calculations may lead to simulated solutions that are wildly divergent from the true paths;
- astronomical simulations typically run over million of years, leading to an accumulation of integration errors; this is problematic as the approximate solutions are only mathematical objects, whereas the actual bodies they represent have to satisfy physical laws (including the various conservation laws); this can be tackled by using analytical methods such as the **variational principle** and **perturbation theory** to produce trajectory manifolds on which to 'project' the integrated approximations.

13: See [11] for an example of planetary system formation).

14: A whimsical take on the effects of such unpredictable behaviour is offered in Liu Cixin's *The Three-Body Problem* [3].

For many bodies, the time complexity is related to the square of the number of bodies, which can make the direct simulation unpractical.

In that case, useful simulations must approximate the essential character of the actual trajectories while reducing the computational complexity. There are many dedicated methods to achieve this goal, including so-called **tree code** and **particle mesh** methods [8].

While these particular issues may not apply to general simulations, the interplay of valid approximation and computational feasibility lies at the core of successful simulations.

### 12.1.4 Simulation Types

We have already alluded to some simulation types; in this section we provide more concrete descriptions of the available modeling avenues.

#### Full-Scale Physical Simulations

Full-scale physical simulations are **life-sized**, **physically realistic** simulations, which make use of structures that already exist to replicate or reproduce target system behaviours.

For example, to simulate boat rescue situations (and then practice responding under various scenarios), the Coast Guard might make use of existing vessels and emergency personnel, and introduce actors playing the part of accident victims, a wave machine to simulate possible environmental conditions, etc.

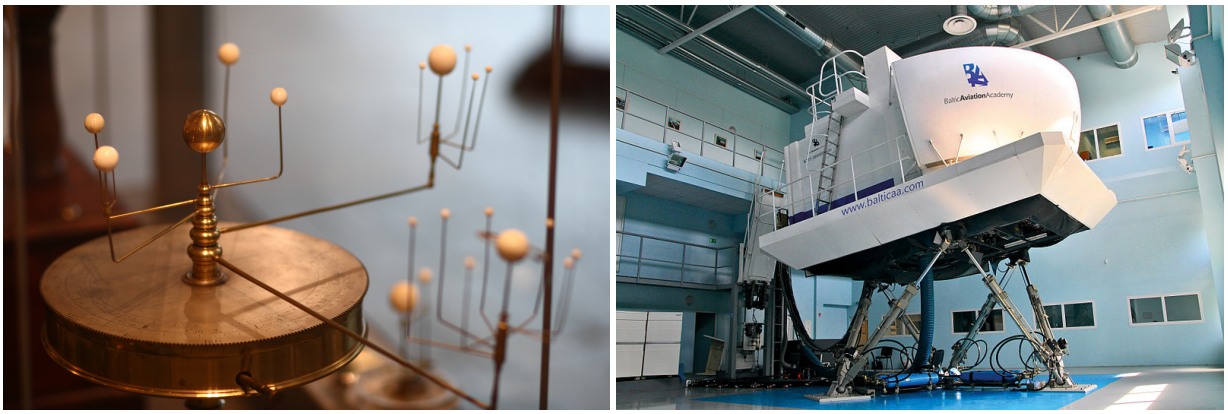



**Figure 12.5:** Harvard orrery [12], and Baltic Aviation Academy Airbus B737 Full Flight Simulator (FFS) in Vilnius (public domain).

### Mechanical Simulations

A mechanical simulation is one that is physically implemented but which is not necessarily **full-scale**, **to-scale**, or physically **realistic** in some respects. It simulates dynamic behaviours using electro-mechanical components. Mechanical simulations were popular prior to the advent of computers.

The 'orrery', a classic type of clockwork model of the solar system, is a typical example of a mechanical simulation (see Figure 12.5, left). Another example would be a CPR dummy that can be used to practice proper CPR technique, and which may have sensors to simulate certain heart behaviours and then provide feedback regarding the effectiveness of the applied CPR.

### Computer (Programmatic) Simulations

Programmatic simulations represent the target system or process using **data structures** and **algorithms**. The data structures are sets of variables that represent the properties of system objects, and the algorithms determine how these properties change over time. When quantitative analysts and consultants produce simulations, they are usually programmatic.

- **Event-Centric Computer Simulations:** this type of computer simulation models **activity** (and is dynamic in this sense), but the focus is not accurate modeling of time. The goal, rather, is to represent an event or sequence of events. For example, we might simulate the selection, and result, of sampling a population, or simulate possible outcomes of a series of events that themselves occur with particular probabilities.
- **Discrete Time Computer Simulations:** as suggested by the name, discrete time simulations treat time as a **discrete series of consecutive steps**, rather than continuously. A common example of this is the **agent-based** model (or multi-agent simulation); in this type of simulation, the time step may range from seconds to years, and the goal of the simulation is to explore how individual agents interact with each other over this time span.

- **Continuous Time Computer Simulations:** In contrast to discrete time simulations, continuous time simulations treat time as a **continuous property**. The challenge is that continuous time simulations are generally implemented on a computer, and computers are necessarily discrete. Thus, in practice, a continuous time simulation is one where the discrete time steps are simply **very small**. Note that this is not equivalent to implementing a continuous-time mathematical model on a computer and solving it using mathematical methods implemented as algorithms.

#### Hybrid Models

It is also possible to create a model of a system where one part of the model is of one type and another part is of another type. A realistic flight simulator, for instance, might consist of a few full-scale physical components such as the cockpit, seats, etc.,[15] while the experience of actually flying the plane is simulated via computer, and perhaps integrated with the physical part of the simulation by projecting a computer controlled image onto the cockpit window (see Figure 12.5, right). The computer simulation might also controls the physical behaviour of the motion of the cockpit – its pitch, yaw, and roll, for example.

15: Possibly using part of an actual plane.

## 12.2 Modeling Strategies

Among practitioners, it sometimes said that modeling is as much an art as it is a science. While there are no tested and true approaches that will work no matter the situation under consideration, the following steps, illustrated in Figures 12.6 to 12.11 with the simulation of a school of fish, often end up having practical importance in the process:

1. gather information about the target system;
2. create a conceptual model;
3. build the model;
4. verify and validate, and
5. run and analyze.

### 12.2.1 Information Gathering

As domain experts or modeling specialists, it can be tempting to believe that the understanding of the target system is so strong that that we can forgo collecting and validating information about that system and jump right into implementing a model of the system. However, modelers tend to be experts in specific techniques rather than in the behaviour of the target system, and *vice-versa* for the domain experts – **teamwork** is usually required to properly construct the model.

In such a case, the modeler and domain expert must work together closely to **gather the information** about the system that the domain expert believes will be required to understand or predict the relevant behaviours of the target system. The modeler must also keep in mind the types of information required to create a comprehensive and consistent model

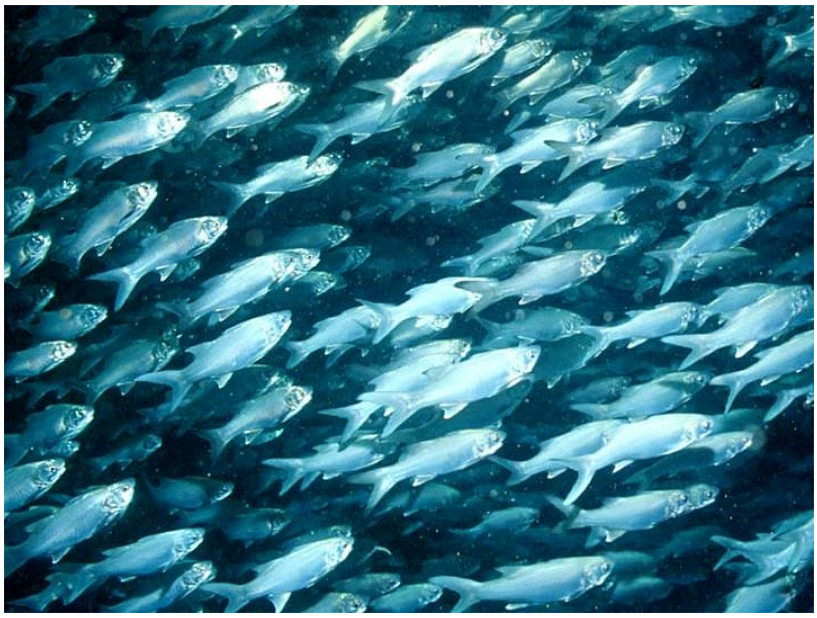


**Figure 12.6:** A school of fish – an example of a target system to simulate [15].

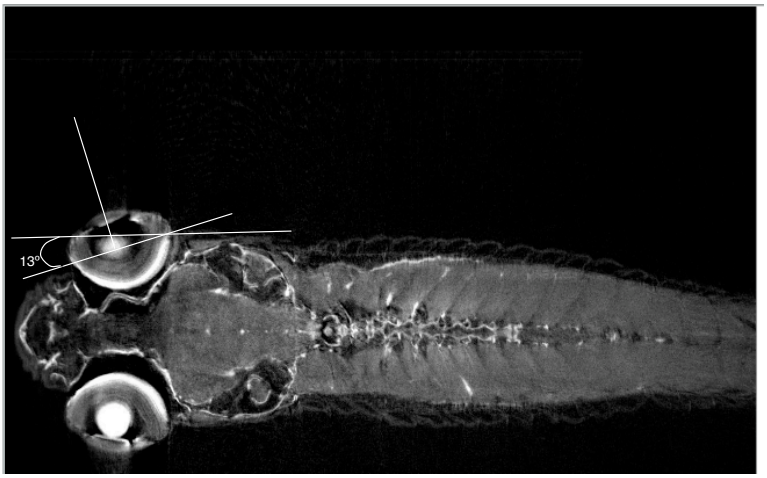



**Figure 12.7:** Gathering information: relevant perceptual mechanics information about a single fish, to be incorporated into the model [13].

of the system, given the proposed model type. Creating a **conceptual model** (see below) will greatly assist with the process of determining what information is necessary to properly **represent** the target system.

There is also an opportunity to **validate** the structure of the model at this stage. Even when a domain expert is involved, ensuring that the information being incorporated into the model comes from **rigorous and reliable sources**, and **documenting these sources** early on, will enhance the likelihood that the model will be valid, as well as **increasing the credibility** of the model in the eyes of those using the model.

### 12.2.2 Conceptual Model

A **conceptual model** is a clearly defined description of those **components**, **properties**, and **relationships** of the system that are believed to be important, relative to the system behaviours or properties of interest (i.e., the **modeling context**). A conceptual model may be a:

- **verbal description** of the system, structured in some way;

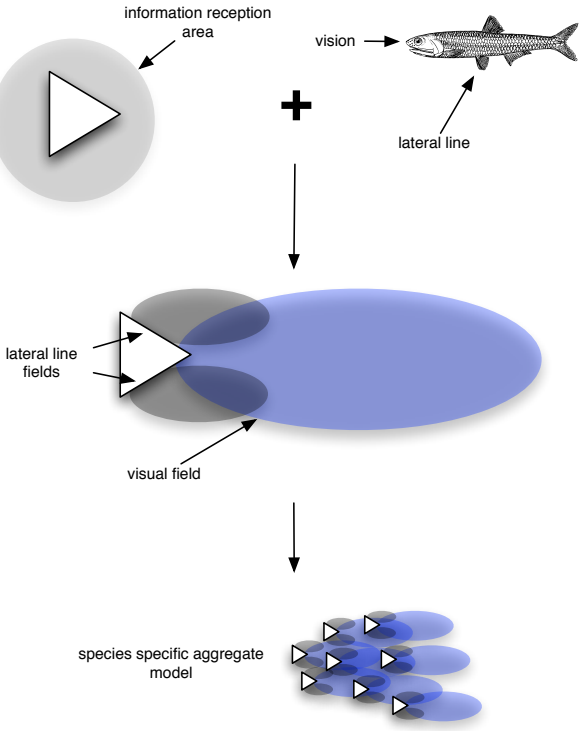



**Figure 12.8:** Creating a conceptual model: a conceptual model showing how elements of the target system – the fish in a fish school – will be represented in the model of the fish school [13].

- **collection of diagrams** depicting elements of the system and their relationships, or
- **combination** of both.

The conceptual model can be thought of as the **blueprint** that will be followed during construction of the model.

At this stage, the modeler will also often discover that it is necessary to **concretely define** the more abstract or less well-defined elements of the target system, in preparation for implementing the model. During the construction of the conceptual model, it may be determined that there are **gaps in the understanding** of the system itself, which prevent the construction of a complete model of the system.

If this occurs, it may be necessary to return to **gathering information** about the target system itself. If the required information is not readily available it is important at this step to indicate which parts of the model are based on reliable knowledge about the system and which parts are speculative.

This step can be challenging from an interdisciplinary perspective because, as we have already mentioned, it requires the modeler and the domain expert to **work together** to create the conceptual model. This requires, in a sense, the domain expert to **enter the modeler's world**, just as the modeler must **enter into the world of the domain expert**.

This can be difficult to achieve, for a variety of reasons, and as a result it can be tempting to skip this step outright – to leave the conceptual model in an **implicit stage** rather than in an **explicit stage** – and to jump straight into building the model.

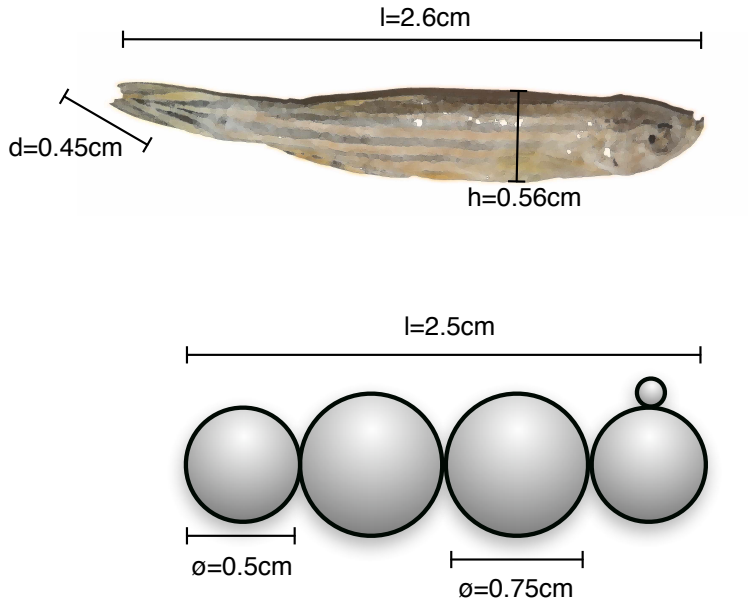



**Figure 12.9:** Creating a conceptual model: determining how specific relevant physical characteristics of individual fish will be represented and incorporated into the model [13].

However, unless the modeler is also a domain expert and the system itself is relatively simple, this can lead to models that do not perform satisfactorily in the final analysis.

### 12.2.3 Building the Model

Once the conceptual model is in place, a model type (e.g., mathematical, simulation) can be selected in order to **build the model** itself, using the conceptual model as a blueprint. Target system **objects**, **properties**, and **relationships** are translated into **model structures**.

### 12.2.4 Verification and Validation

Verifying the model means going over the model in order to confirm that it has been **constructed as intended**, given the conceptual blueprint that has been developed.

Validation refers to a process of confirming that the constructed model is in fact a **good match** for the target system. Thus, a model could be verified as having been constructed as intended, but the model might still be invalid if, for example, the modeler was misinformed about the actual workings of the target system.

A thoughtful discussion of model validation, in the context of building population-based disease simulation models, can be found in [9].

### 12.2.5 Analysis of Results

Once the model has been verified and validated, it may then be analysed in order to **draw conclusions** about the target system.

```
TIMESTEP(O, λ, Φ, N)
1  for each agent in λ
2       do L' ← ATTENTION(O)
3          I ← COGNITIVE-PROCESSING(L')
4          ACTION(I)
5  for each agent in Φ ▷ perception deprived agents
6       do L' ← ()
7          I ← COGNITIVE-PROCESSING(L')
8          ACTION(I)

ATTENTION(O)
1  L' ← MERGE-LISTS(O)
2  L' ← PICK-NEIGHBOURS(N, L')
3  ▷ the appropriate PICK-NEIGHBOURS procedure (below) is called for each scenario
4  return L'

PICK-NEIGHBOURS-RANDOM(N, L')
1  return RANDOM(N, L')

PICK-NEIGHBOURS-NEAREST(N, L')
1  return NEAREST(N, L')
```

**Figure 12.10:** Building the model: pseudo-code describing how the simulation of the fish school is created [13].

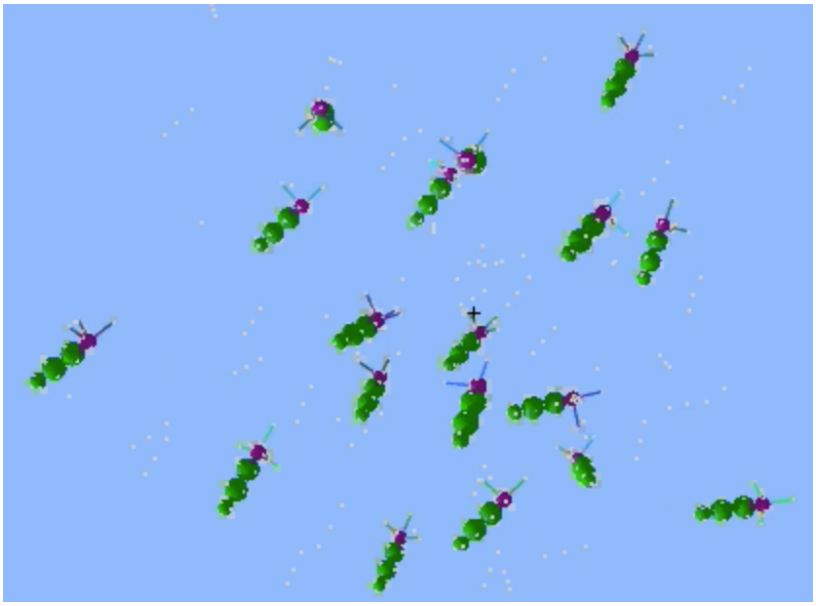


**Figure 12.11:** Building the model: the resulting simulation of the fish school. The schooling behaviour is an emergent property of the simulation, coming out of programmed individual simulated-fish behaviours [13].

In the case of simulations, model parameters have to be selected, and 'runs' of the model carried out for each set of parameters.[16]

If the model has **stochastic components**, it may be necessary to carry out multiple runs using the same parameter settings in order to produce

16: By 'run' we mean that the model is given certain **initial starting conditions** and then the behaviour of the simulation allowed to proceed and produce various outputs of interest.

posterior distributions for the outputs. Once the simulation has been run with all of the relevant parameter settings, the resulting output of the simulation can be analysed.

At this point, the analysis may follow a vast number of methods: **trend extraction and forecasting**, **classification**, **data visualization**, etc.

## 12.3 Practical Considerations

As with most applied disciplines, the implementation and applications are fraught with unsuspected challenges. While it remains important to have a good handle on the conceptual foundations of the field, the best way to become a competent practitioner is to continuously attempt to conduct simulations, and to learn from the inevitable mistakes made along the way.

### 12.3.1 Computational Complexity

Because simulations are computer programs, it remains crucial to be aware of the broader issue of **computational complexity** when constructing simulations. The computational complexity of an algorithm is based on the number of possible steps in the algorithm and how they interact with different types of data to lead to different run times.

Although a detailed discussion of computational complexity is beyond the scope of this section, understanding that the manner in which the simulation is programmed will influence its run time is very important, as this might limit the options for the exploration of parameter space.

As previously discussed, when a simulation is created, a set of parameters to vary has to be explicitly selected in order to explore the behaviour of the simulation. However, because specific parameter values have to be chosen for each run of the simulation, and because multiple simulations have to be run in order to get a general sense of the behaviour of the simulation (i.e. building a posterior distribution for the behaviour), and by extension the system, the problem of **combinatorial explosion** is encountered very quickly.

This problem cannot always be overcome, and it might be that the best that can be hoped for is to maximise the number of simulation runs that the computer can support in the available time.

### 12.3.2 Applications and Use Cases

**Science** The appropriate role of models and simulations within science is a topic for debate within scientific circles. Statistical models are well accepted and used extensively. Mathematical models are generally accepted if used in a theoretical context. In our experience, however, the use of simulations is currently not encouraged. In situations where carrying out actual experiments would be difficult,[17] simulations may be viewed as a type of **virtual experiment**. In such situations the results of the virtual experiment, although not viewed in the same light as actual experimental

17: For ethical reasons, notably.

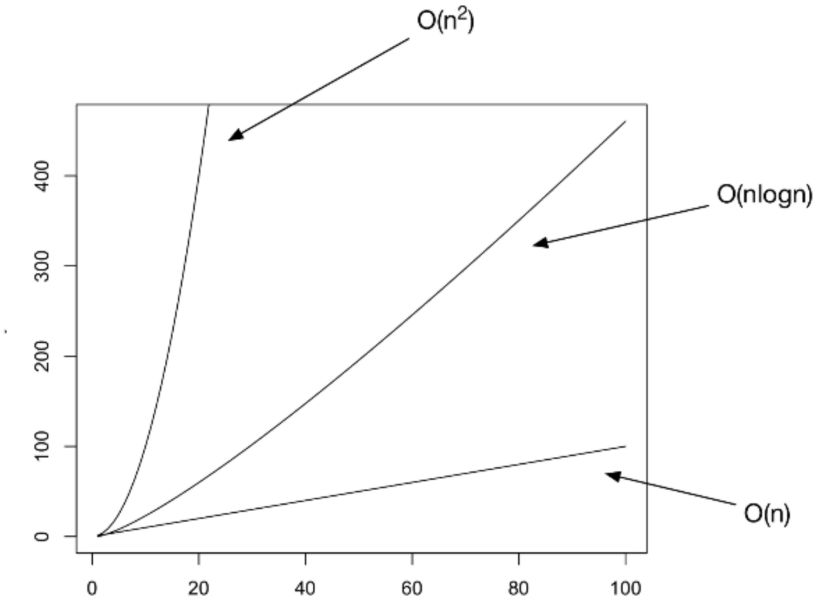



**Figure 12.12:** A sketch of some different possible computational complexities of a computer program, as represented in Big-O notation.

results, may, at the very least, fuel the discovery of hypotheses, which may then be tested using other methods.

**Business** Accurate prediction of events is highly valued in a business context. Consequently, the emphasis for models in this domain is on **predictive accuracy**, rather than on being able to use the model for explanatory purposes. Businesses use models to, for example, predict customer behaviour, how their business will be affected in certain market situations, and how they might reorganise their business structure to reduce overhead and increase profitability.

**Government** Setting policy is a major governmental activity. Within this context, it is often important to explore different possible policy scenarios, and gain a better understanding of which policies will be effective in a variety of circumstances. Models that provide **explanatory power** can be particularly helpful in this type of work, because it allows for an understanding of why one approach might work better than another. This can then be taken into account in order to ensure good policy.

In addition, as with businesses, governments are usually interested in making its own operations **more efficient and effective**. From an organisational perspective, models can help determine the best strategies for internal structures and processes, as well as the conditions under which such structures may function optimally.

**Education** Simulations play an important role in education, allowing students to explore and experience scenarios virtually, which may decrease the risks associated with "learning through doing", and increase the rewards of "learn from experience" in **controlled** and **monitored conditions**.[18]

18: For a very thorough discussion of the role of simulations in education, see [10].

**Entertainment** It might be argued that most forms of entertainment are simply reflections or **representations** of real world experiences, and are thus, in some sense, **models of life**. More specifically, simulations and models frequently play an important role in theatre, television, and film – allowing creators to convincingly mimic real life situations without needing to entirely re-create or enact them, using **physical small-scale models** (e.g., a small-scale model of a cityscape), **life-size models** of particular environments (e.g., a life-size model of a submarine), or **computer simulations** (e.g. simulated flocks of birds and artificially generated clouds, added to provide more realism and detail to the backdrop of a scene).

### 12.3.3 Modeling and Simulation Software

It is quite possible to create models **by hand**, without the use of computers, and it is also possible to create computer models or simulations without using a **particular programming environment**. But some programming environments have been specifically designed for creating simulations. Some of these currently available (as of 2018) include:

- Matlab Simulink (commercial simulation software)
- Simio (commercial simulation software)
- Netlogo (free software, mainly for teaching and prototyping)
- SymPy (a python library for discrete time simulations)

## 12.4 Case Study: NWMO

Canada has a long history with nuclear power: the first self-sustained Canadian nuclear reaction was achieved at Chalk River's ZEEP reactor in 1945. Over the years, numerous research reactors and power reactors have been built and decommissioned – as of 2014, electricity is currently being produced by 19 CANDU reactors in Ontario and New Brunswick. Given that the existence of high energy nuclear waste in Canada is a *fait accompli*,[19] it is paramount that we find ways to safely dispose of this waste.

19: We have already chosen, as a society, to use nuclear power and create nuclear waste

In 2002, the *Nuclear Fuel Waste Act* (NFWA) was enacted to study possible strategies for the management of Canada's used nuclear fuel. As a result, the *Nuclear Waste Management Organization* (NWMO) was formed by the Canadian nuclear power companies, with the mandate to provide recommendations to the Canadian Government for the long-term management of used nuclear fuel. One such recommendation, which was accepted in 2007, was the establishment of Adaptive Phased Management (APM) as both a social and technical approach to permanently manage Canada's used nuclear fuel. Canadian citizens determined that the optimal strategy, given the current state of technology in Canada, is the construction of a deep geological repository to contain and isolate the fuel.

This decision puts the NWMO in a unique and demanding position, as it is the first group in Canada to design and build a unique but extremely performance-critical engineering structure: a long term Canadian repository for high energy nuclear waste. By its very nature, this structure as

a whole cannot be tested in advance of use and essentially cannot be maintained once it is built. Furthermore, the environment and materials involved are themselves volatile and their long term behaviour is difficult to predict.

Under such challenging circumstances, engineers must do their best to use all of the expertise at their disposal to create as perfect a design as possible for the required structure. Despite the uniqueness of the structure, they need to produce a design that will meet the requirements that have been set out, and then, once built, function exactly as predicted on the first try. Such a design process is necessarily a lengthy one, involving many designers with high levels of expertise. Many designs would be proposed and rejected before a final design is selected, based on all the evidence and expertise the design team have at their disposal.

At the end of the process the engineering team will have high confidence in the final design that is put forward. The success of the structure in question is critical, and, as responsible, professional engineers, they would not put forward a design for such a structure without being entirely certain, to the best of their collective ability, that this structure will not fail.

Despite this confidence, due diligence requires more than the simple assurance (and belief) from the design team that the structure will not fail. It is not enough, from a societal perspective, for the team to simply provide a "vote of confidence:" it also requires the provision of more quantitative information about the failure aspects of the structure. Those responsible for the structure need to be able to determine (and to help the stakeholders understand) what are the structure's necessary and sufficient conditions for failure (and by extension, the conditions for non-failure). To produce these answers they need to be able to quantitatively examine what circumstances the structure might encounter, and under these circumstances, what the probability of failure is.

From an ideal testing point of view, the entire proposed structure would be built many times over to run trials relating to each of the foreseen circumstances. Data would then be gathered and analyzed to determine the failure tolerance of the structure. Failure probabilities would be calculated based on this data, along with an understanding of possible failure circumstances – the structure might even be redesigned to take into account the results of the testing.

However, as we have already noted, this idealistic testing scenario is simply not an option in this case. The structure as a whole cannot be directly tested even once, let alone multiple times; even were many replications of the structure itself available for testing, not all failure circumstances would be possible to re-create in a test environment.[20]

20: In particular those involving major geological forces and long time spans.

An alternative strategy is centered around a combination of physical testing and modeling of the behaviour of the structure and environment. More specifically, a larger structure is built up of many component parts, which themselves may be built up of many components. The failure parameters of these component parts may be tested, even if the structure as a whole cannot.

Similarly, while the structure itself, and perhaps even in some cases the components themselves, cannot be tested repeatedly, there remains the

option of creating models of the structure and components in question, and then using the behaviour of these models to predict the behaviour of the components and, in turn, of the structure at large.

In the absence of the ideal testing scenario, understanding and quantifying the failure of the system as a whole can be carried out by understanding and quantifying the failure circumstances of the components of the system, understanding the causal relationships between these components, creating models of the system as a whole based on these relationships, determining the failure circumstances and probabilities of the constructed structure level models and then transferring these findings over to the structure itself. This results in an estimate of the failure circumstances and probabilities of the actual engineered structure as a whole.

The end result of this exercise will thus be, rather than a simple yes/no statement (such as "No, the structure will not fail", for instance), a list of the possible failure circumstances and an estimate of the failure probabilities for both the structure components and the structure itself, along with a confidence measure indicating a level of confidence in the failure probabilities calculated for each failure circumstance.

Such a table of failure circumstances, probabilities, and confidence measures will allow those building the structure to open a legitimate dialogue with those responsible for, and those being affected by, the resulting structure. In essence, this deliverable will allow the designers of the structure to provide their stakeholders with a clearer and more detailed picture of the risks they are likely to encounter when undertaking the construction of such a structure.

**General Objectives**

The general objective of this Failure Analysis project as a whole is to estimate the failure probability of the Mark II canister and engineered barrier system immediately surrounding the canister. In order to achieve that larger objective, we anticipate that we will be using a combination of statistical analysis, mathematical modeling, and simulations, much as in this prototype. More specifically, we will take the approach that our model is meant to answer a specific question, as well as to provide outputs that can be fed into other models, as may be required by already-developed NWMO models.

In this prototype phase, however, the objective is to develop a methodology and implementation framework to confirm that interactions (both planned and emergent) can in principle be captured by the modeling process, both at the repository and the manufacturing level. For both the manufacturing process and the interactions models, a specific selection of a small number of sub-components of the entire system will be considered in this phase, in order to maintain focus on the development and testability of the methodology itself.

In [2], we report on a simulation approach for the *Failure Analysis Simulation Model for the APMRD-II*, we discuss some of the strategies that could be used to extract information and knowledge about the engineered barrier system, which could then be incorporated in any interaction model

of its components. A discussion of system complexity and the effect it had on our choice of modeling approach is also provided. We also provide a prototype UFC manufacturing process model: potential states, actions and variables are introduced, as well as the underlying modeling assumptions and families of parameters. The model is illustrated *via* a specific parameter set; a series of 8 scenarios showcase the effect of various parameter combinations.[21]

21: It should be noted that due to the uncertainty relating the manufacturing process parameters, the numbers presented are placeholders: reasonable estimates for a large number of these parameters will be required before the model can output meaningful failure estimates.

## 12.5 Exercise

Create a simulation of pre-board screening (PBS) wait-time at Borealian airfields (as described in Section 24.6).

## Chapter References


[1] D. Beloriszky. 'Application pratique des méthodes de M. Sundman à un cas particulier du problème des trois corps'. In: *Bulletin Astronomique* 6 (series 2) (1930), pp. 417–434.

[2] P. Boily and J. Schellinck. *Introduction to Quantitative Consulting*. Quadrangle/Data Action Lab, 2025.

[3] L. Cixin. *The Three Body Problem*. Chongqing Press, 2008.

[4] J.J. Clement. *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation*. Springer Netherlands, 2008.

[5] *Fourdee*. https://en.wikipedia.org/w/index.php?curid=8650757 .

[6] D.R. Hofstadter. 'Analogy as the core of cognition'. In: *The Analogical Mind: Perspectives from Cognitive Science*. Ed. by D. Gentner, K. Holyoak, and B. Kokinov. Cambridge MA: The MIT Press Bradford Book, 2001.

[7] K. Holyoak, D. Gentner, and B. Kokinov. 'Introduction: The place of analogy in cognition'. In: *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*. Sofia: NBU Press, 1998.

[8] *https://en.wikipedia.org/wiki/N-body_problem* .

[9] J.A. Kopec et al. 'Validation of population-based disease simulation models: A review of concepts and methods'. In: *BMC Public Health* 10 (2010). DOI: 10.1186/1471-2458-10-710.

[10] F. Landriscina. *Simulation and Learning: A Model-Centered Approach*. New York, NY: Springer, 2013.

[11] T. Momkov. *Planetary System Formation 2, N−body simulation*. *n*−Body Simulation . 2013.

[12] S. Ross. *A 1766 Benjamin Martin Orrery, used at Harvard (photo)*. Putnam Gallery Planetarium. 2009.

[13] J. Schellinck. 'A general perception based framework for modelling animal aggregation'. Ottawa: Carleton University, 2009.

[14] weusemath.org. *New Discoveries: n−body problem* .

[15] Wikipedia. *Shoaling and schooling* .

[16] L. Yilmaz, ed. *Concepts and Methodologies for Modeling and Simulation: A Tribute to Tuncer Ören*. Switzerland: Springer International Publishing, 2015.