

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 μ With σ Known	351
7.3.3	Confidence Level	356
7.3.4	Sample Size	358
7.3.5	C.I. for μ With σ 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	Miscellanea	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	731
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 μ	740
11.2.2	Inference on $\mu_1 - \mu_2$	745
11.2.3	Inference on σ^2	751
11.2.4	Inference on σ_1^2/σ_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