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

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.

[1] A fair number of the examples and exercises we provide in the chapter also come from those references.

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]

- **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**.

[2] The **skew** and the **kurtosis** are also sometimes used.

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:

$$\begin{aligned}\text{HM}(x_1, \ldots, x_n) &= \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}\\ \text{GM}(x_1, \ldots, x_n) &= \sqrt[n]{x_1 \cdots x_n}.\end{aligned}$$

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).

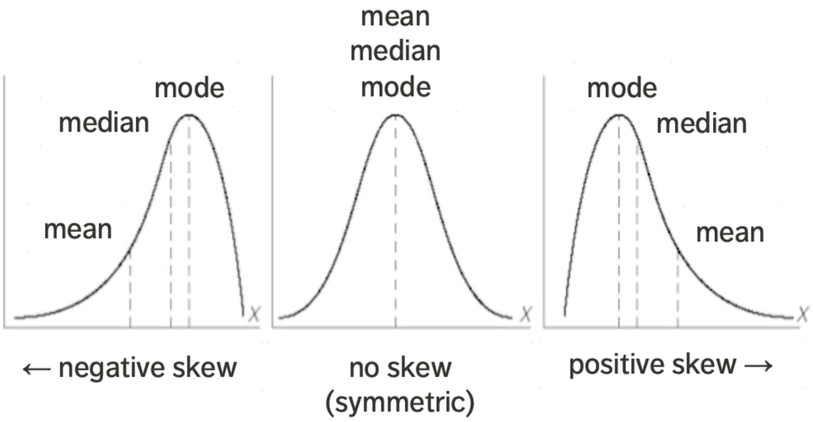


Figure 7.1: Mean, median, and mode in various skewness scenarios. [modified from unknown source]

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.

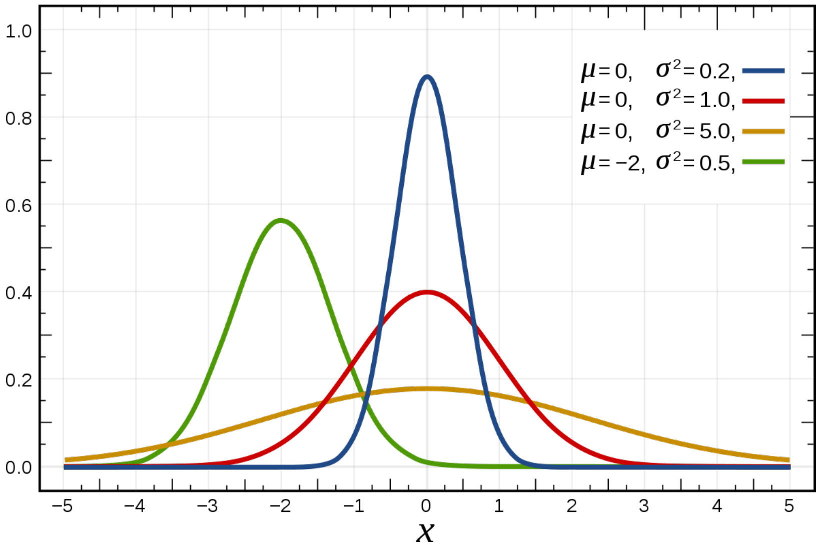


Figure 7.2: Normal distributions, with various means and standard deviations. [Wikipedia]

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 σ and σ^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]

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
```

[3] In statistical parlance, we say that 1 **degree of freedom** is lost when we use the sample to estimate the sample mean.

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

$$x < Q_1 - 1.5\,\mathrm{IQR} \quad \text{or} \quad x > Q_3 + 1.5\,\mathrm{IQR},$$

- **definite outlier** if

$$x < Q_1 - 3\,\mathrm{IQR} \quad \text{or} \quad x > Q_3 + 3\,\mathrm{IQR}.$$

Example In the set $\{1, 3, 4, 6, 7, 23\}$, $Q_1 = 2$, $Q_3 = 6.5$, and $\mathrm{IQR} = 4.5$. Thus

$$\begin{aligned} Q_1 - 1.5\mathrm{IQR} &= 2 - 1.5(4.5) = -4.75 \\ Q_3 + 1.5\mathrm{IQR} &= 6.5 + 1.5(4.5) = 13.25 \\ Q_1 - 3\mathrm{IQR} &= 2 - 3(4.5) = -11.5 \\ Q_3 + 3\mathrm{IQR} &= 6.5 + 3(4.5) = 20.0 \end{aligned}$$

Since $23 > Q_3 + 3\mathrm{IQR}$ (and $23 > Q_3 + 1.5\mathrm{IQR}$), 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):

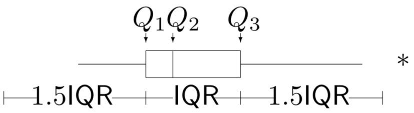


Figure 7.3: Boxplot with one (suspected) outlier.

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).

[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.

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.

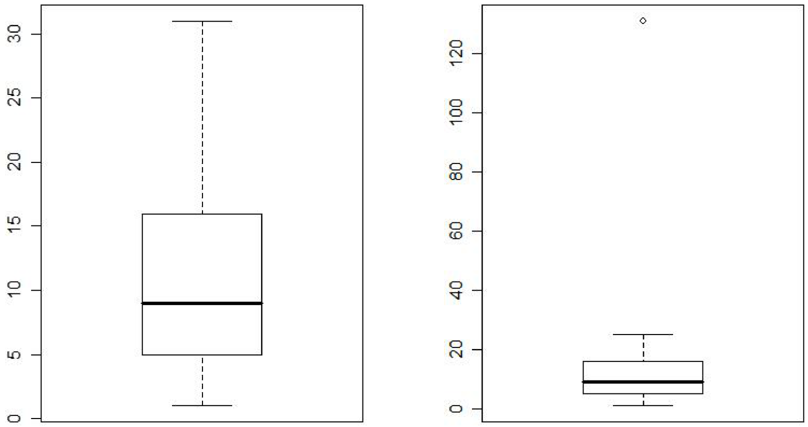


Figure 7.4: Boxplot of positively skewed datasets.

Histograms

Visual information about the distribution of the sample can also be provided *via* **histograms**.

A histogram for the sample $\{x_1, \dots, x_n\}$ 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

$$\min = y_1 = 1, \quad Q_1 = \frac{1}{2}(y_{10} + y_{11}) = 5, \quad \text{med} = \frac{1}{2}(y_{20} + y_{21}) = 9,$$
$$Q_3 = \frac{1}{2}(y_{30} + y_{31}) = 16, \quad \max = y_{40} = 31.$$

A corresponding histogram and boxplot are shown in Figure 7.5.

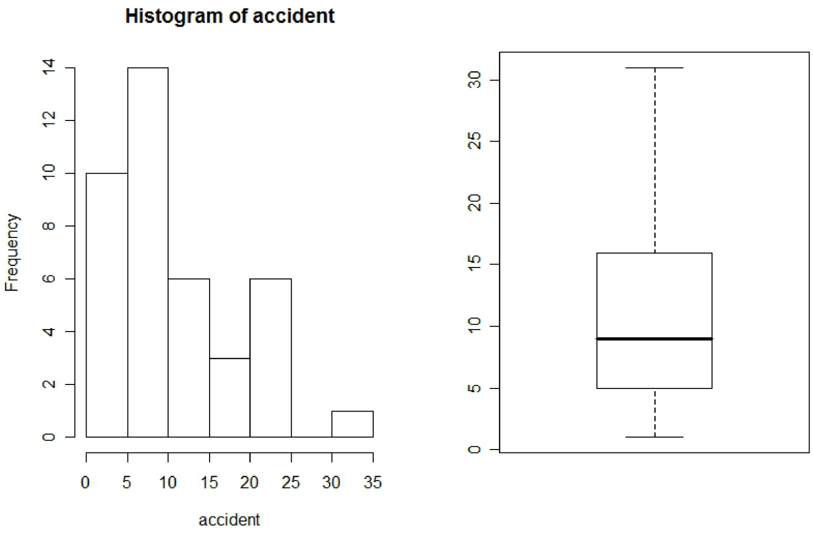
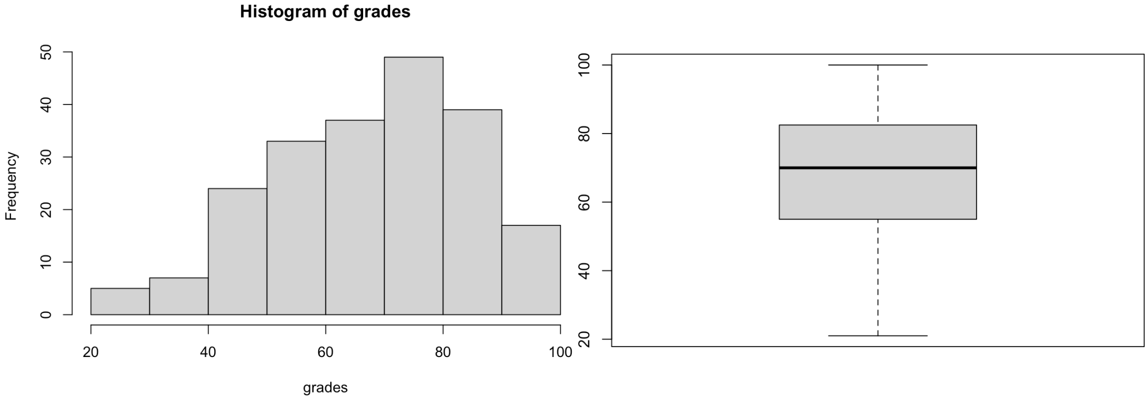


Figure 7.5: Histogram and boxplot of the Sydney accident dataset.

2. We can also visualize the `grades` dataset:

```
hist(grades, breaks = seq(20,100,10))
boxplot(grades)
```



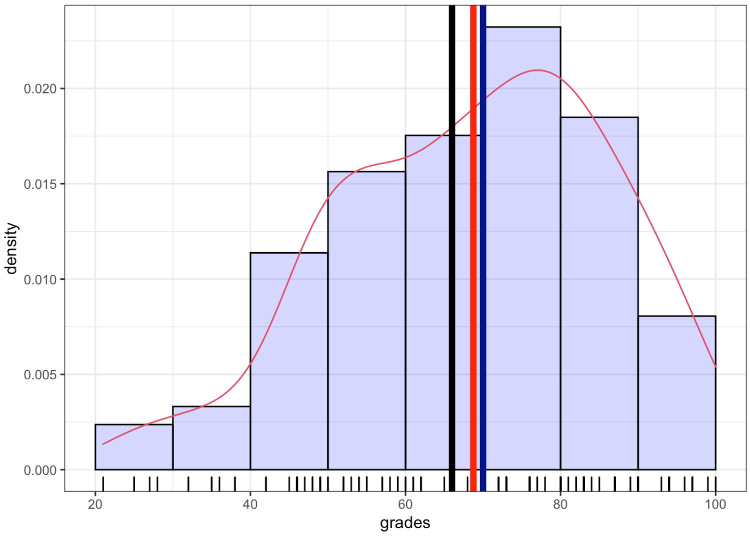
Here is a fancier version of the histogram, constructed with the `ggplot2` package.[6]

```
# 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
```

[6] See Section [1] for details on the use of this R package.

```
        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:

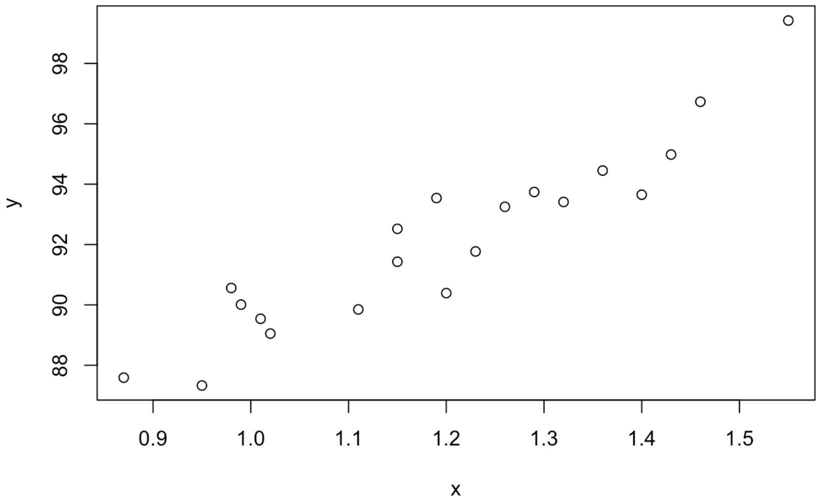
```
x = c(
  0.99,1.02,1.15,1.29,1.46,1.36,0.87,1.23,
  1.55,1.40,1.19,1.15,0.98,1.01,1.11,1.20,
  1.26,1.32,1.43,0.95
  )
y = c(
  90.01,89.05,91.43,93.74,96.73,94.45,87.59,91.77,
  99.42,93.65,93.54,92.52,90.56,89.54,89.85,90.39,
  93.25,93.41,94.98,87.33
)

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), $i = 1, \dots, n$, 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 ρ_{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

- ρ_{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;
- ρ_{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 ρ_{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 ρ_{XY} reflecting it (see Figure 7.6).

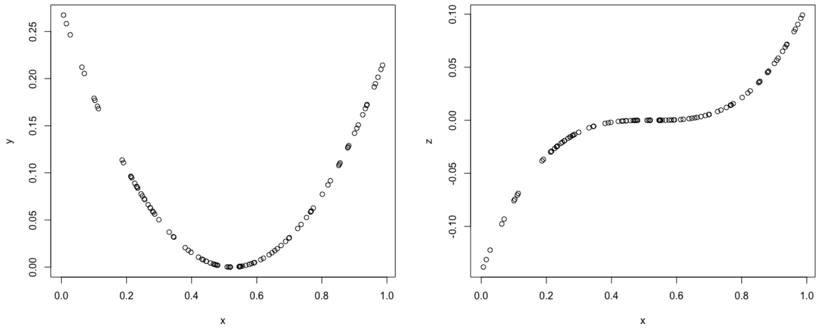


Figure 7.6: Examples of strong relationships that are not reflected by the coefficient of correlation.

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** θ, 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.

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** μ and **known variance** σ^2, then $\mathrm{Var}(\overline{X}) = \sigma^2/n$ and the **standard error of** $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

[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, χ^2−statistics, f−statistics, etc.);
- **order statistics** (sample maximum and minimum, sample range, etc.);
- **sample moments** and functions thereof (skewness, kurtosis, etc.);
- etc.

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]

```
(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 μ is not. What is the best estimate of μ? 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 μ and standard deviation 50, which is to say, $\overline{X} \sim \mathcal{N}(\mu, 5^2) = \mathcal{N}(\mu, 25)$, according to the CLT.

[8] Note that `var()` always treats the underlying dataset as a **sample**, not as a **population**.

7.3.2 Confidence Intervals for μ When σ is Known

Consider a sample $\{x_1, \ldots, x_n\}$ drawn from a **normal population** with **known** variance σ^2 and **unknown** mean μ. The sample mean

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

is a **point estimate** of μ.[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 μ.

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.

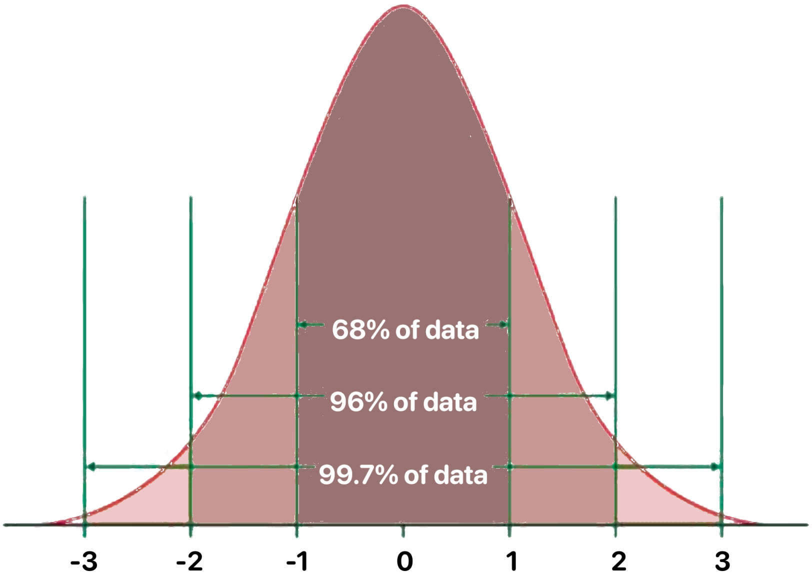


Figure 7.7: The 68-96-99.7 rule on the standard normal distribution. [source unknown]

In other words, whenever we observe a sample mean $\overline{X}$ (with sample size n) from a normal population with mean μ, 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** μ:

$$\overline{X} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + k\frac{\sigma}{\sqrt{n}} \quad \Longrightarrow \quad \text{C.I.}(\mu; g(k)) \equiv \overline{X} \pm k\frac{\sigma}{\sqrt{n}}.$$

Examples

1. Consider a sample $\{X_1, \dots, X_{64}\}$ 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 μ.

   According to the formula, the symmetric 68.3% confidence interval ($k = 1$) for μ 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 μ 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 μ is built, μ will fall between the endpoints of the interval about 68.3% of the time.[10]

2. Build a symmetric 95.5% confidence interval for μ.

   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 μ.

   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**.

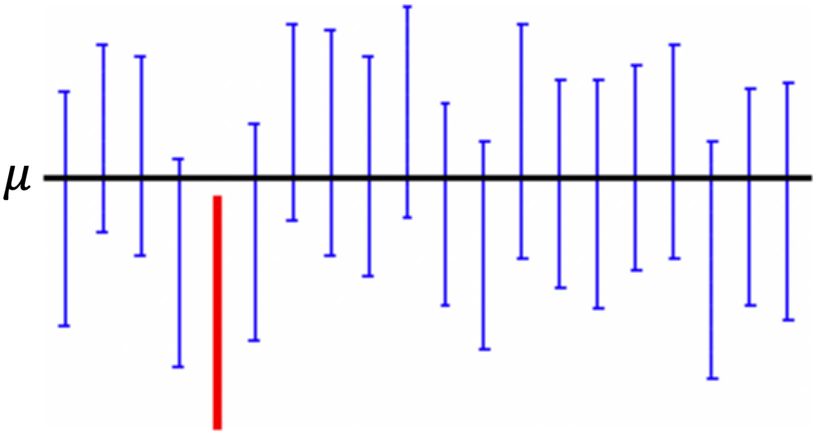


Figure 7.8: Frequentist interpretation of confidence intervals: out of 20 experiments, we would expect the true population mean to fall in the confidence interval about 19 times, on average. [source unknown]

Confidence Interval for μ when σ 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** μ, 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** μ **when** σ **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** μ **when** σ **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 μ.

   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 μ 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 μ 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
}
```

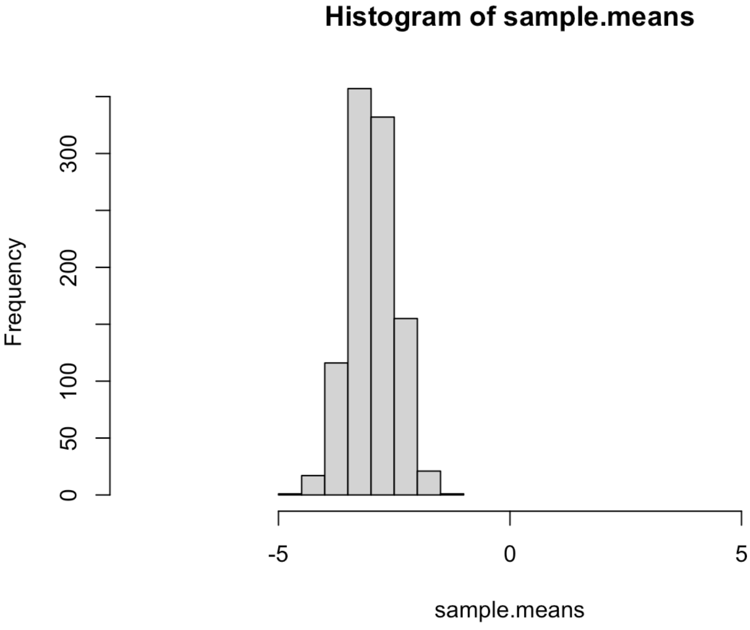
The proportion of the times when it does can thus be obtained *via*:

```
table(is.mu.in)/M
```

```
is.mu.in
FALSE  TRUE
0.055 0.945
```

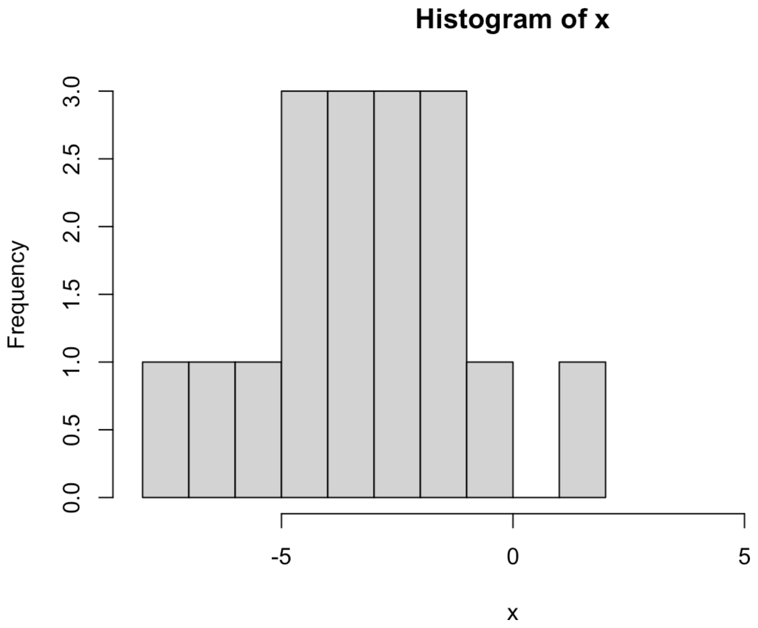
This is indeed very close to 95%. We can also verify the conclusion of the CLT: look at the histogram of the sample means!

```
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))
```

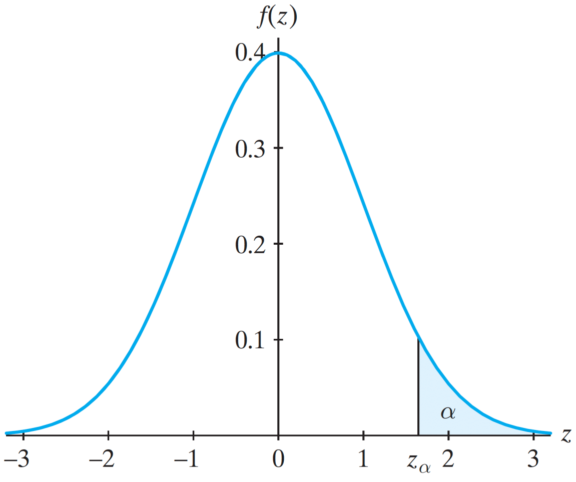


The sample variance is significantly larger than the standard error.

7.3.3 Confidence Level

The **confidence level** $1-\alpha$ is usually expressed in terms of a **small** α, so that $\alpha = 0.05$ corresponds to a confidence level of $1-\alpha = 0.95$.

For $\alpha \in (0,1)$, the value z_α 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.



$$P(Z > z_\alpha) = \alpha$$

$$P(Z > z) = 1 - \Phi(z) = \Phi(-z)$$

Figure 7.9: Quantiles of the standard normal distribution [5].

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.

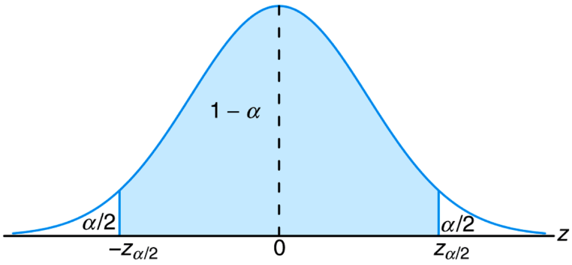


Figure 7.10: Two-sided quantiles of the standard normal distribution [5].

The most commonly-used cases are for $\alpha = 0.05$ and $\alpha = 0.01$:

$$P(|Z| > z_{0.025}) = 0.05 \implies z_{0.025} = 1.96$$
$$P(|Z| > z_{0.005}) = 0.01 \implies z_{0.005} = 2.575.$$

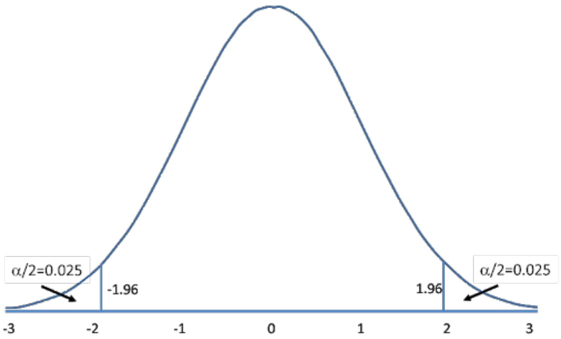


Figure 7.11: Two-sided quantiles of the standard normal distribution, for confidence level 0.05.

The symmetric $100(1-\alpha)\%$ C.I. for μ 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 α, **shorter confidence intervals are better** in relation to estimating the mean:

- estimates improve when the sample size n increases;
- estimates improve when σ 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.

[11] The only ones we will consider in these notes.

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 μ.

  The estimate of μ 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,
  $$\begin{aligned} \text{C.I.}(\mu; 0.95) &\equiv 19.93 \pm 1.96\frac{5}{\sqrt{9}} = (16.66, 23.20) \\ \text{C.I.}(\mu; 0.99) &\equiv 19.93 \pm 2.575\frac{5}{\sqrt{9}} = (15.64, 24.22). \end{aligned}$$

- 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 μ.

  The estimate of μ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)\%$ C.I. are:
  $$\begin{aligned} \text{C.I.}(\mu; 0.95) &\equiv 19.93 \pm 1.96\frac{5}{\sqrt{25}} = (17.97, 21.89) \\ \text{C.I.}(\mu; 0.99) &\equiv 19.93 \pm 2.575\frac{5}{\sqrt{25}} = (17.35, 22.51). \end{aligned}$$

- 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 μ.

  The estimate of μ is the sample mean $\overline{X} = 19.93$. The $100(1-\alpha)\%$ C.I. are:
  $$\begin{aligned} \text{C.I.}(\mu; 0.95) &\equiv 19.93 \pm 1.96\frac{10}{\sqrt{25}} = (16.01, 23.85) \\ \text{C.I.}(\mu; 0.99) &\equiv 19.93 \pm 2.575\frac{10}{\sqrt{25}} = (14.78, 25.08). \end{aligned}$$

Note how the confidence intervals are affected by α, n, and σ.

7.3.4 Sample Size

The **error** E we commit by estimating μ *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).

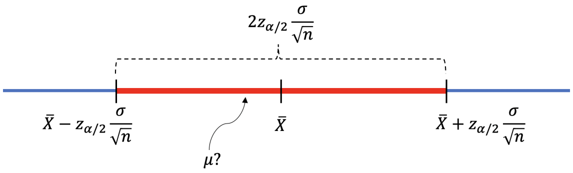

$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
$\bar{X}$
$\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
μ?

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]

$$E > z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

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 μ 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 α, σ, E, and n is not always intuitive, but it follows a simple rule.

7.3.5 Confidence Intervals for μ When σ is Unknown

So far, we have been in the fortunate situation of sampling from a population with **known** variance σ^2. What do we do when the population variance is **unknown** (a situation which occurs much more frequently in real world applications)?

[12] Sampling strategies can also help, but this is a topic for another day (see Chapter 10).

The solution is to estimate σ 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 μ 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]

If σ 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 α, 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 μ and unknown variance σ^2, then $t_{0.025}(5-1) = 2.776$ and

$$P\left(-2.776 < \frac{\overline{X}-\mu}{S/\sqrt{5}} < 2.776\right) = 0.95.$$

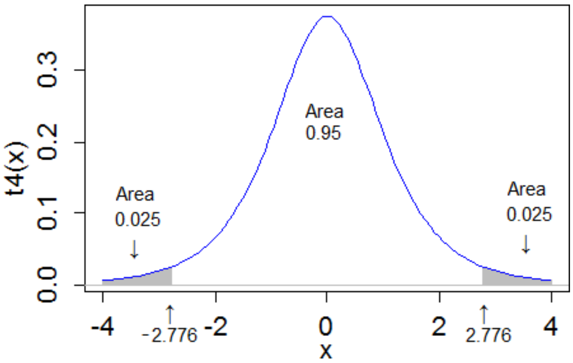


Figure 7.13: Critical value for Student distribution with 4 degrees of freedom, at confidence level 0.05. [source unknown]

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 μ. Note that $\overline{X} = 5.01$ and that $S = 0.97$.

[13] Remember, when σ is known (and n is large enough), we already know from the CLT that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is approximately $\mathcal{N}(0,1)$.

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 μ.

   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 ± 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.

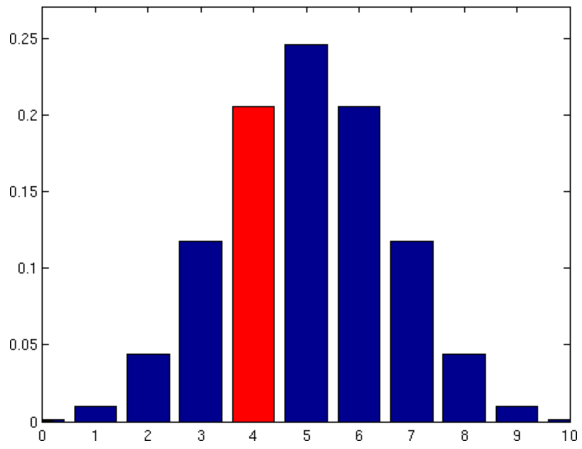


Figure 7.14: Binomial distribution for 10 trials, with probability of success 1/2. The probability of exactly 4 successes is highlighted in red.

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**.

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 $\text{E}[Y] = 50$ 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**.

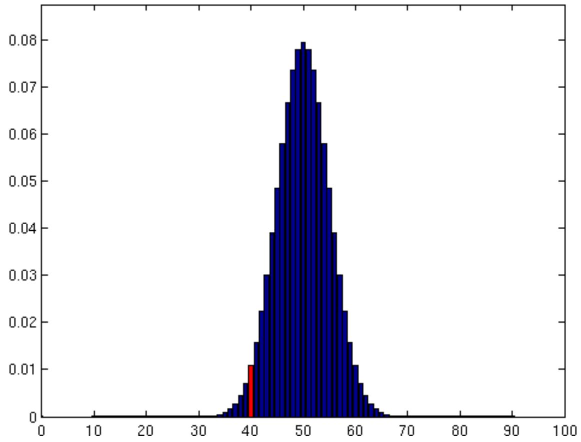


Figure 7.15: Binomial distribution for 100 trials, with probability of success 1/2. The probability of exactly 40 successes is highlighted in red.

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) ≈ 38% of the time (simply by chance).

[14] Which is to say, that the coin is symmetric.

[15] Compare this with the previous value $P(X = 4) = 0.205$.

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*:

> 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!

[17] The crisis concerns the prevalence of positive findings that are contradicted in subsequent studies [4].

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]

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 α and β are usually 0.05 and 0.2, respectively, although that is not a hard and fast rule.

Types of Null and Alternative Hypotheses

Let μ 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 μ_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.

[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.

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 μ 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 $\overline{X} < 4970$.

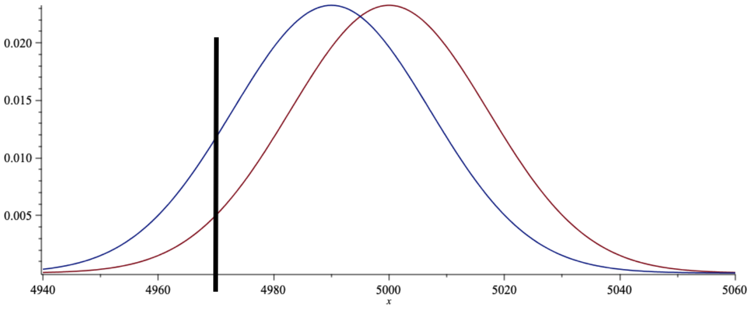
1. Find the probability of committing a type I error when H_0 is true.

   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

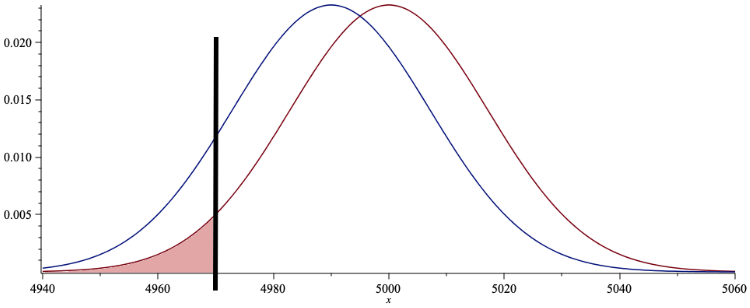
$$\alpha \approx P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{4970-5000}{120/7}\right) \approx P(Z < -1.75) \approx 0.0401\,.$$



   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 μ is actually 4990, say (and not 5000, as assumed in H_0).

   By definition, we have

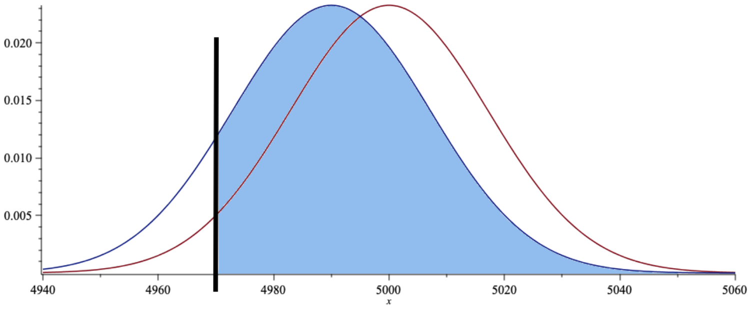
$$\begin{aligned}\beta &= P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})\\ &= P(\overline{X} > 4970 \mid \mu = 4990).\end{aligned}$$

Thus, according to the CLT, we have

$$\begin{aligned}\beta = P(\overline{X} > 4970) &= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{4970-4990}{120/7}\right)\\ &\approx P(Z > -1.17) = 1 - P(Z < -1.17) \approx 0.879\,.\end{aligned}$$

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:

$$\beta = P(\text{fail to reject } H_0 \mid H_0 \ \text{ is false}) = P(\overline{X} > 4970 \mid \mu = 4990).$$

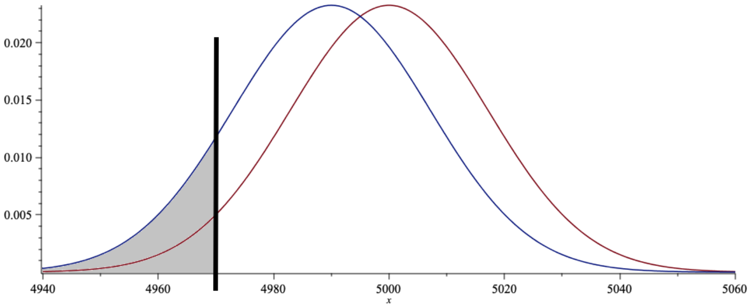


We would thus fail to reject H_0 if the observed value of $\overline{X}$ falls to the right of $\overline{X} = 4970$ (outside the critical region).

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 μ is actually 4950, say (and not 5000, as in H_0).

By definition, we have

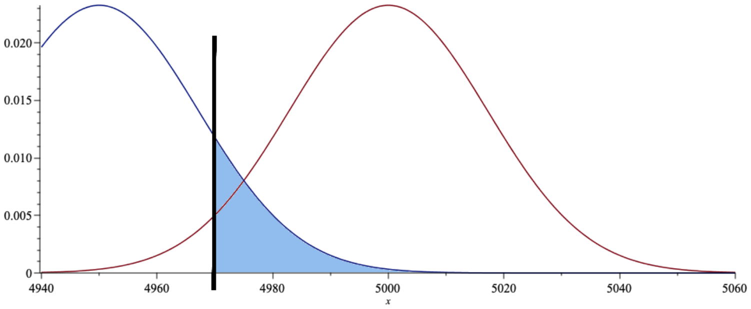
$$\beta = P(\text{fail to reject } H_0 \mid H_0 \ \text{ is false}) = P(\overline{X} > 4970|\mu = 4950).$$

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 μ and variance σ^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 σ^2 is **known**, and that the hypothesis concerns the **unknown** population mean μ.

Explanation: Left-Sided Alternative

Consider the unknown population mean μ. Suppose that we test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu < \mu_0,$$

where μ_0 is some candidate value for μ. To evaluate the evidence against H_0, we compare $\overline{X}$ to μ_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 α, 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.

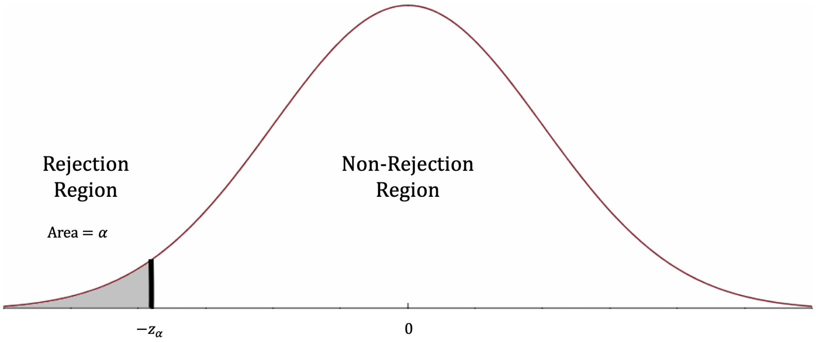


Figure 7.16: Critical test region, left-sided test.

An equivalent **right-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu > \mu_0;$$

at significance α, 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.

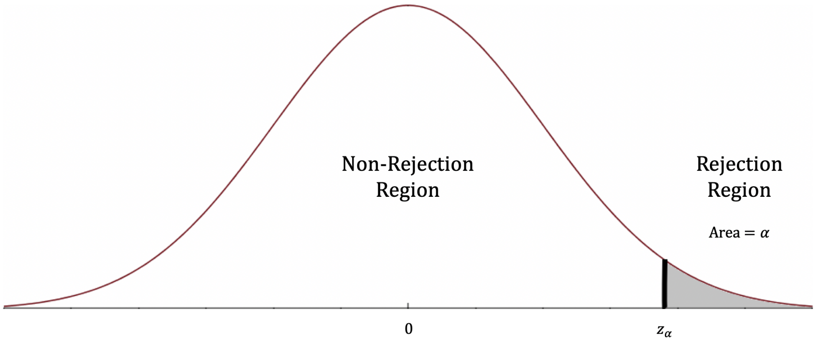


Figure 7.17: Critical test region, right-sided test.

The **two-sided test** pits

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0;$$

at significance α, 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.

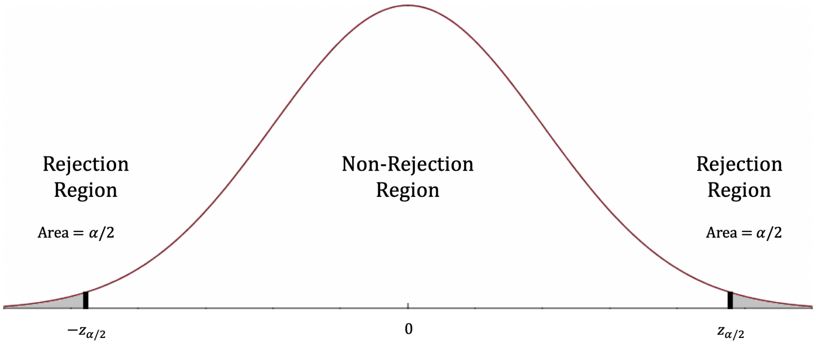


Figure 7.18: Critical test region, two-sided test.

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).

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 $z_0 = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$.

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_α 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.

α	z_α	$z_{\alpha/2}$
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).

[20] What we are trying to show using the data.

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 α, 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 μ 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 μ 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 μ be the true mean of the subgroup. We are testing for $H_0 : \mu = 60$ against $H_1 : \mu < 60$.

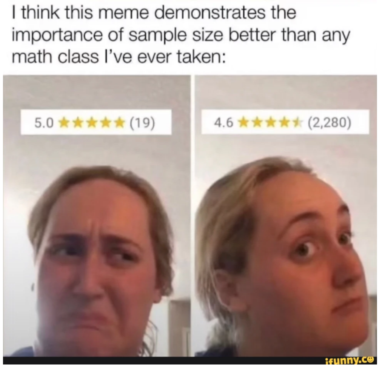
   The observed sample test statistic is
   $$z_0 = \frac{55-60}{10/\sqrt{9}} = -1.5.$$
   The corresponding $p-$value is
   $$P(\overline{X} \leq 55) = P(Z \leq -1.5) = 0.07.$$
   Thus there is not enough evidence to reject the claim that the subgroup is 'average', regardless of whether we use $\alpha = 0.05$ 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 μ 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. . .



I think this meme demonstrates the importance of sample size better than any math class I've ever taken:

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]

For a given α, we reject $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ if, and only if, μ_0 is **not** in the $100(1-\alpha)\%$ C.I. for μ.

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 μ.

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 μ 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 σ 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
$H_1 : \mu > \mu_0$	$t_0 > t_\alpha(n-1)$
$H_1 : \mu < \mu_0$	$t_0 < -t_\alpha(n-1)$
$H_1 : \mu \neq \mu_0$	$|t_0| > t_{\alpha/2}(n-1)$

[22] In order to avoid the controversy surrounding the crisis of replication?

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 μ 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 σ^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 ≤ 0.05, which is strong evidence against $H_0 \colon \mu = 16.6$.

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]

Paired Test

Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean μ_1 and unknown variance σ^2; let $X_{2,1}, \ldots, X_{2,n}$ be a random sample from a normal population with unknown mean μ_2 and unknown variance σ^2, with both populations **not necessarily independent** of one another.[24] We would like to test for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

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:

[23] These populations are potentially distinct.

[24] It is possible that the 2 samples arise from the same population, or represent two different measurements on the same units, say.

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 μ_1 and μ_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 $\alpha = 0.01$.

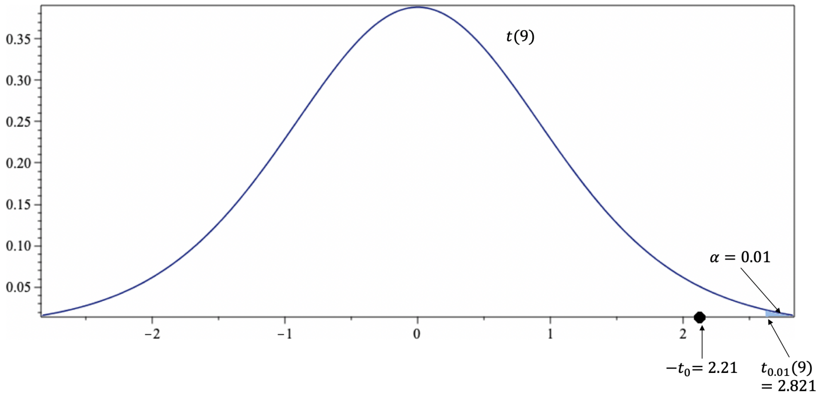


Figure 7.19: Critical test regions for the right-sided test, with $n = 10$ observations: confidence levels 0.05 (left) and 0.01 (right).

Unpaired Test

Let $X_{1,1}, \ldots, X_{1,n}$ be a random sample from a normal population with unknown mean μ_1 and variance σ_1^2; let $Y_{2,1}, \ldots, Y_{2,m}$ be a random sample from a normal population with unknown mean μ_2 and variance σ_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 σ_1^2 and σ_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_α 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 σ_1^2 and σ_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

$$t_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2/n + S_p^2/m}} \quad \text{and} \quad S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

$t_\alpha(n+m-2)$ 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 σ_1^2 and σ_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_α 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 σ 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 σ 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 σ 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** α if the p−value of the test is **below** α.[25]

If the p−value of the test is **greater** than the significance level α, then we **fail to reject the null hypothesis** H_0 **at significance level** α.[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.

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 μ_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)
   ```

[25] Which means that the probability of wrongly rejecting H_0 when H_0 is in fact true is below α, usually taken to be 0.05 or 0.01).

[26] Which, it is worth recalling, is not the same as accepting the null hypothesis.

```
    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 μ_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]

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 μ_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 $\mu_Y \approx 2.5$.

[27] Is it problematic that the sample size $n = 7$ is small?

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]

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)?

[28] Alas, we cannot. All that we can say is that we do not have enough evidence to reject the null hypothesis $H_0 : \mu_Y = 2.5$.

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 μ_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.

Let us return to the teaching method example of Section 7.4.7.

The mean of the grades, for all students, is:

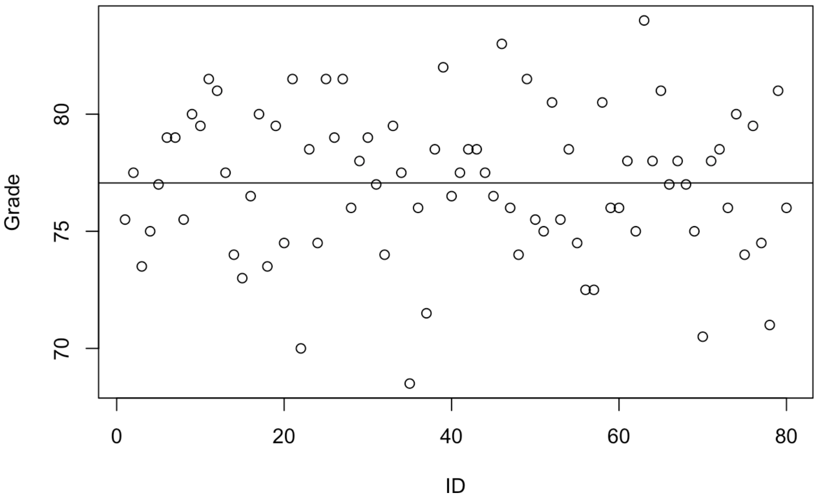
[29] We will have more to say on the topic in Chapter 11.

```
(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**

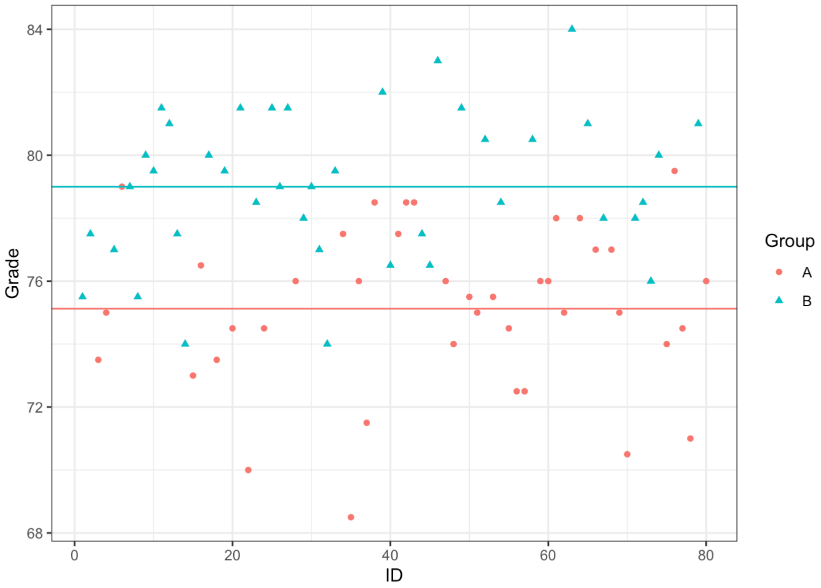
$$y_{i,j} = \mu + \varepsilon_{i,j},$$

where μ is the **overall mean**, $i = A, B$, and $j = 1, \dots, 40$, 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 μ_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

$$\begin{aligned}\mathrm{SST} &= \sum_{i,j}(y_{i,j}-\overline{y})^2 = \sum_{i,j}(y_{i,j}-\overline{y}_i+\overline{y}_i-\overline{y})^2 \\ &= \sum_i N_i(\overline{y}_i-\overline{y})^2 + \sum_{i,j}(y_{i,j}-\overline{y}_i)^2 = \mathrm{SSA}+\mathrm{SSE}\end{aligned}$$

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 × 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	$p-1$	$\mathrm{MSA} = \mathrm{SSA}/(p-1)$	MSA/MSE	$P(F_0 > F^*)$
Error	SSE	$N-p$	$\mathrm{MSE} = \mathrm{SSE}/(N-p)$		
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	7.2×10^{-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., $\varepsilon_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$).

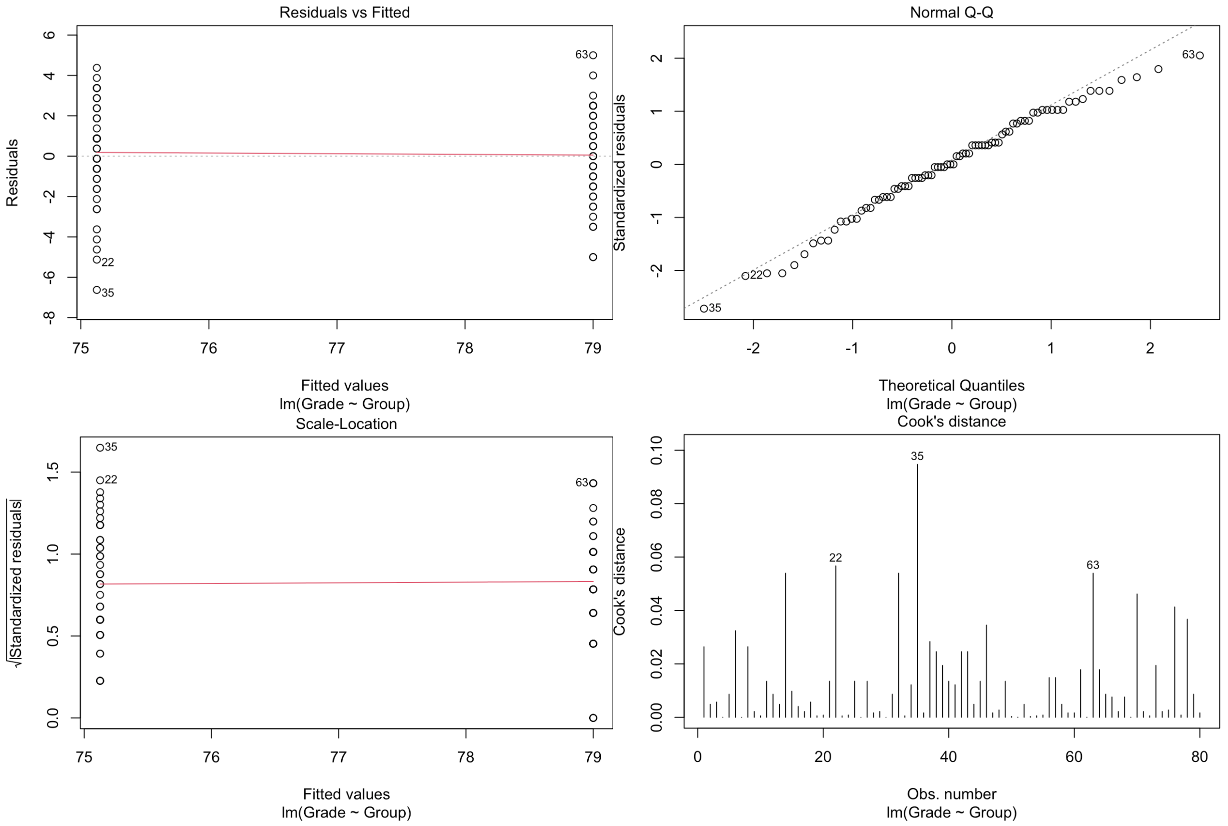
Assuming independence, we are required to verify three additional assumptions:

- 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]

In other words, if the errors are normally distributed with mean 0 and variance σ^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).

[32]: A straight line indicates normality.

[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 δ and variance σ_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, SS_{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]

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^{th} subject in the j^{th} treatment group;
- μ is the overall mean;
- τ_j is the j^{th} treatment effect, subject to a constraint

$$\sum_{j=1}^{p} \tau_j = 0;$$

- γ is the coefficient for the **covariate effect**;
- $(x_{i,j} - \overline{x})$ is the covariate value of the i^{th} subject in the j^{th} treatment group, adjusted by the mean, and
- $\varepsilon_{i,j}$ is the error of i^{th} subject in the j^{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 σ_ε^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.

[34] The medication may have strong side-effects which cannot be ignored.

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 χ^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 necessarly 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 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**, **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 α 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 α 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.

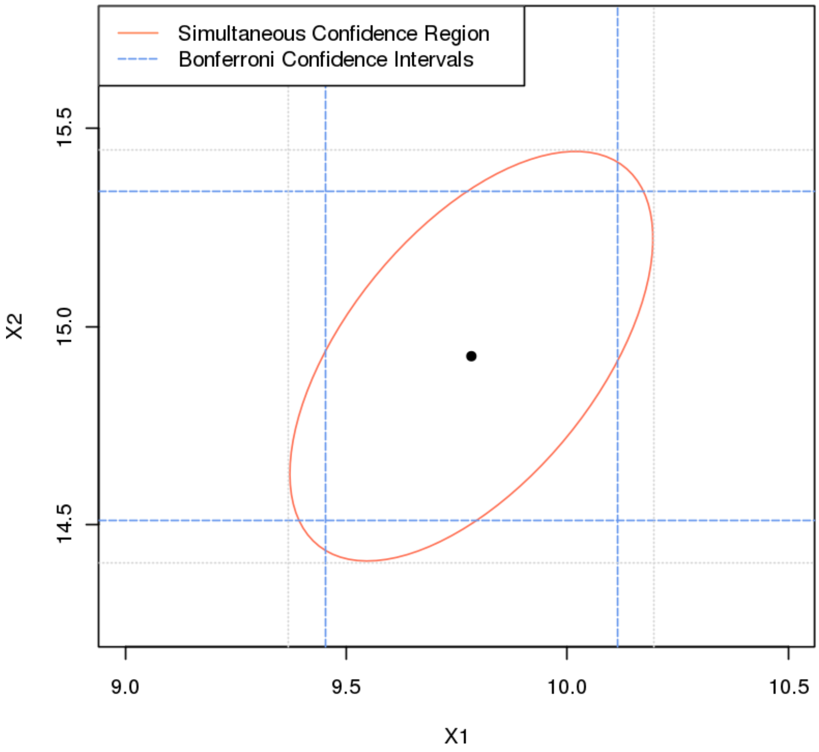


Figure 7.20: Confidence region for a bivariate normal random sample (sample not shown).

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 $i = 1, \dots, I$.

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^{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** SSP_{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 Λ^* is below some pre-agreed upon threshold, which depends on p, I, and N_i, $i = 1, \dots, I$.

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^{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 χ^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.

$$\begin{gathered}
80, 73, 83, 60, 49, 96, 87, 87, 60, 53, 66, 83, 32, 80, 66 \\
90, 72, 55, 76, 46, 48, 69, 45, 48, 77, 52, 59, 97, 76, 89 \\
73, 73, 48, 59, 55, 76, 87, 55, 80, 90, 83, 66, 80, 97, 80 \\
55, 94, 73, 49, 32, 76, 57, 42, 94, 80, 90, 90, 62, 85, 87 \\
97, 50, 73, 77, 66, 35, 66, 76, 90, 73, 80, 70, 73, 94, 59 \\
52, 81, 90, 55, 73, 76, 90, 46, 66, 76, 69, 76, 80, 42, 66 \\
83, 80, 46, 55, 80, 76, 94, 69, 57, 55, 66, 46, 87, 83, 49 \\
82, 93, 47, 59, 68, 65, 66, 69, 76, 38, 99, 61, 46, 73, 90, \\
66, 100, 83, 48, 97, 69, 62, 80, 66, 55, 28, 83, 59, 48, 61 \\
87, 72, 46, 94, 48, 59, 69, 97, 83, 80, 66, 76, 25, 55, 69 \\
76, 38, 21, 87, 52, 90, 62, 73, 73, 89, 25, 94, 27, 66, 66 \\
76, 90, 83, 52, 52, 83, 66, 48, 62, 80, 35, 59, 72, 97, 69 \\
62, 90, 48, 83, 55, 58, 66, 100, 82, 78, 62, 73, 55, 84, 83 \\
66, 49, 76, 73, 54, 55, 87, 50, 73, 54, 52, 62, 36, 87, 80, 80
\end{gathered}$$

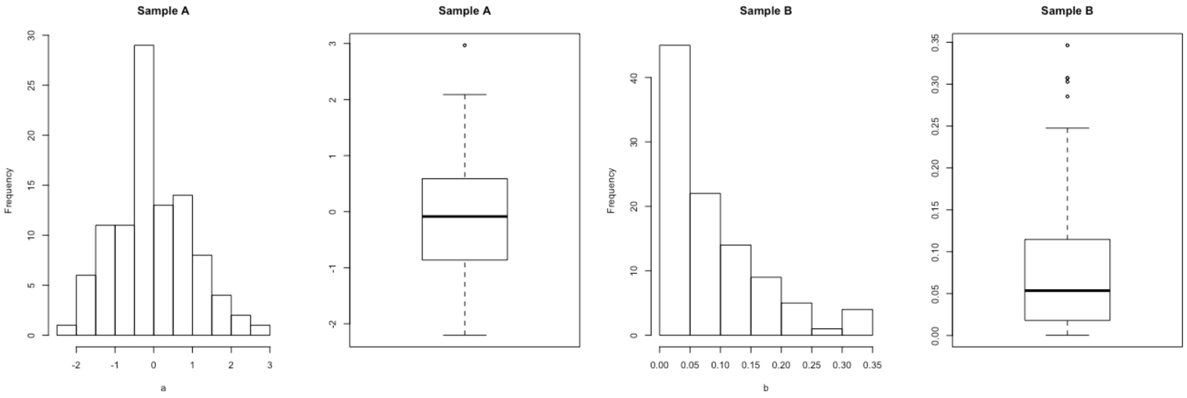
   a) Compute the sample mean and sample variance.
   b) Find the 5-point summary of the data. Is the distribution skewed?
   c) Are there any likely outliers in the sample? If so, indicate their values.
   d) Build and display the sample's boxplot chart.
   e) Build and display a sample histogram.
   f) 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 $\mathrm{kg/cm^2}$ 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 μ?

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 α.

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 μ_1 and μ_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 α 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 α 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 α 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 α 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 α 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 α;
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 α;
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.