THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

School of Mathematics and Physics

STAT1301
Advanced Analysis of Scientific Data

Dirk P. Kroese   Benoit Liquet

September 17, 2014

Contents

Preface

These notes are intended for first-year science students who would like to more fully understand the logical reasoning and computational techniques behind statistics. Our intention was to make something that would be useful as a reference for advanced first-year students and statistics tutors, providing both a sound theoretical foundation of statistics as well as a comprehensive introduction to the statistical language R. The material requires a mathematics background comparable to that of Queensland's Mathematics C curriculum (which includes basic knowledge of functions, exponential, differentiation, integration, and matrices). No preknowledge of statistics is required.

It is important to realise that these notes do not replace the regular STAT1201 book, which is aimed at a much broader student audience. The STAT1301 notes will, however, explain several concepts on a much deeper level than is feasible in STAT1201. For this reason it is not possible to cover everything that is treated in STAT1201. For example, goodness of fit tests, contingency tables, logistic regression, and non-parametric methods are not discussed in the STAT1301 notes. Instead, more attention is paid to understanding randomness, probability distributions, statistical modeling, estimation, hypothesis testing, analysis of variance, regression, and linear models. Our guiding principle was that it is more important to know the "why" than the "how". Our thanks go out to many people who provided feedback on earlier versions of the notes, including Michael Bulmer, Alan Huang, Miranda Mortlock, Melanie Robertson–Dean, and Robert Salomone.

To get the most use out of these notes it is important that you carefully read the whole story from beginning to end, annotate the notes, check the results, make connections, do the exercises, try the R programs, visit the lectures, and *most importantly*, ask questions about things you do not understand. If you are frightened by the maths, it is good to remember that the mathematics is there to make life *easier*, not harder. Mathematics is the language of science, and many things can be said more precisely in one single formula, definition, or with a simple artificial example, than is possible in many pages of verbose text. Moreover, by using mathematics it becomes possible to build up statistical knowledge from very basic facts to a high level of sophistication. Of course in a first-year statistics course, however advanced, it is not possible to cover all the knowledge that has been built up over hundreds of years. We will sometimes only give a glimpse of new things to discover, but we have to leave something for your future studies! Knowing the mathematical reasoning behind statistics avoids using statistics only as a black box, with many "magic" buttons. Especially when you wish to do further research it is important to be able to develop your own statistical reasoning, separate from any statistical package.

The material in these notes were partly based on our text books:

- Dirk P. Kroese and Joshua C.C. Chan (2014). *Statistical Modeling and Computation*, Springer, New York.

- Pierre Lafaye de Micheaux, Rémy Drouilhet, and Benoit Liquet (2014). *The R Software: Fundamentals of Programming and Statistical Analysis*, Springer, New York.

Brisbane, *Dirk Kroese and Benoit Liquet*
September 17, 2014

Chapter 1

Introduction

Statistics is an essential part of science, providing the language and techniques necessary for understanding and dealing with chance and uncertainty in nature. It involves the design, collection, analysis, and interpretation of numerical data, with the aim of extracting patterns and other useful information. The purpose of this introductory chapter is to give a brief overview of the structure of the notes and the range of topics that will be discussed.

1.1 Statistical Studies

In science the typical steps that are taken to answer a real-life research question are:

Steps for a Statistical Study

1. Design an experiment to give information about the research question.
2. Conduct this experiment and collect the data.
3. Summarise and visualise the observed data.
4. Make a statistical model for the data.
5. Analyse this model and make decisions about the model based on the observed data.
6. Translate decisions about the model to decisions and predictions about the research question.

To fully understand statistics it is important that you follow the reasoning behind the steps above. Let's look at a concrete example.

Example 1.1 Suppose we have a coin and wish to know if it is fair — that is, if the probability of Heads is 1/2. Thus the research questions is here: is the coin fair? What we could do to investigate this question is to conduct an experiment where we toss the coin a number of times, say 100 times, and observe when Heads or Tails appears. The data is thus a sequence of Heads and Tails— or we could simply write a 1 for Heads and 0 for Tails. We thus have a sequence of 100 observations, such as 1 1 0 1 0 1 0 0 1 . . . 0 1 1. These are our data. We can visualise the data by drawing a bar graph such as in Figure 1.1.

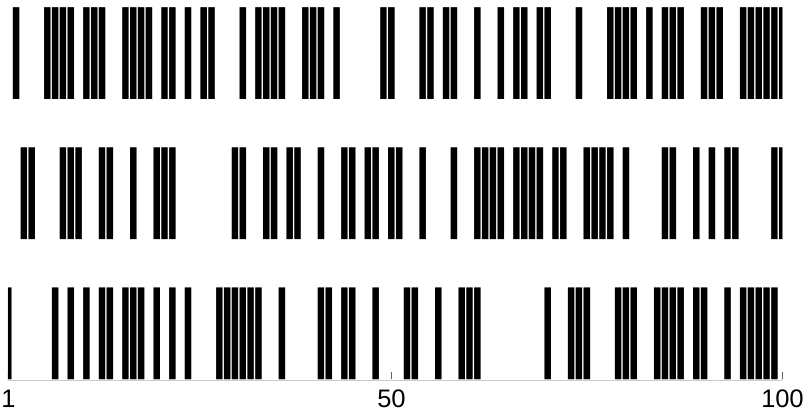


Figure 1.1: Three experiments where a fair coin is tossed 100 times. The dark bars indicate when Heads (=1) appears.

We can *summarise* the data by giving the total number of Heads, x say. Suppose we observe $x = 60$. Thus, we find 60 Heads in 100 tosses. Does this mean that the coin is not fair, or is this outcome simply due to chance?

Note that if we would repeat the experiment with the same coin, we would likely get a different series of Heads and Tails (see Figure 1.1) and therefore a different outcome for x. We can now reason as follows (and this is crucial for the understanding of statistics): if we denote by X (capital letter) the total number of Heads (out of 100) that we will observe *tomorrow*, then we can view $x = 60$ as just one possible outcome of the *random variable* X. To answer the question whether the coin is fair, we need to say something about how likely it is that X takes a value of 60 or more for a fair coin. To calculate probabilities and other quantities of interest involving X we need an appropriate statistical *model* for X, which tells us how X behaves probabilistically. Using such a model we can calculate, in particular, the probability that X takes a value of 60 or more, which is about 0.028 for a fair coin — so quite small. However, we *did* observe this quite unlikely event, providing reasonable evidence that the coin may not be fair.

1.2 Outline of the Notes

In these notes you will learn how to apply and understand the 6 steps of a statistical study mentioned above. We will introduce the topics in a linear fashion, starting in Chapter 2 with the summarisation and visualisation of data. This is in a way the easiest part, as it requires little preknowledge. We will use the statistical package `R` to read and structure the data and make figures and tables and other summaries. Chapter 3 is about *probability*, which deals with the modeling and understanding of randomness.

We will learn about concepts such as random variables, probability distributions, and expectations. Various important probability distributions in statistics, including the *binomial* and *normal* distributions, receive special attention in Chapter 4. We then continue with a few more probability topics in Chapter 5, including multiple random variables, independence, and the central limit theorem. At the end of that chapter we introduce some simple statistical models. After this chapter, we will have built up enough background to properly understand the statistical analysis of data (Steps 4 and 5). In particular, we discuss *estimation* in Chapter 6 and and *hypothesis testing* in Chapter 7, for basic models. The remaining chapters consider the statistical analysis of more advanced models, including *analysis of variance* models and *regression* models, both of which are special examples of a *linear model*. The `R` program will be of great help here. Appendix A gives a short introduction to `R`.

Chapter 2

Describing Data

This chapter describes how to structure data, calculate simple numerical summaries and draw standard summary plots.

2.1 Introduction

Data is often stored in a table or spreadsheet. A statistical convention is to denote variables as columns and the individual items (or units) as rows. It is useful to think of three types of columns in your spreadsheet:

1. The first column is usually an identifier column, where each unit/row is given a unique name or ID.

2. Certain columns can correspond to the design of the experiment, specifying for example to which experimental group the unit belongs. Often the entries in these columns are *deterministic*; that is, they stay the same if the experiment were to be repeated.

3. Other columns represent the observed measurements of the experiment. Usually, these measurements exhibit *variability*; that is, they would change if the experiment were to be repeated.

It will be convenient to illustrate various data concepts by using the Excel data file `nutrition_elderly.xls`, which contains nutritional measurements of thirteen variables (columns) for 226 elderly individuals (rows) living in Bordeaux (Gironde, South-West France) who were interviewed in the year 2000 for a nutritional study (see Table 2.1 for a description of the variables).

Table 2.1: Description of the variables in the nutritional study

Description	**Unit** or **Coding**	**Variable**
Gender	1=Male; 2=Female	`gender`
Family status	1=Single 2=Living with spouse 3=Living with family 4=Living with someone else	`situation`
Daily consumption of tea	Number of cups	`tea`
Daily consumption of coffee	Number of cups	`coffee`
Height	Cm	`height`
Weight (actually: mass)	Kg	`weight`
Age at date of interview	Years	`age`
Consumption of meat	0=Never 1=Less than once a week 2=Once a week 3=2/3 times a week 4=4/6 times a week 5=Every day	`meat`
Consumption of fish	Idem	`fish`
Consumption of raw fruits	Idem	`raw_fruit`
Consumption of cooked fruits and vegetables	Idem	`cooked_fruit_veg`
Consumption of chocolate	Idem	`chocol`
Type of fat used for cooking	1=Butter 2=Margarine 3=Peanut oil 4=Sunflower oil 5=Olive oil 6=Mix of vegetable oils (*e.g.* Isio4) 7=Colza oil 8=Duck or goose fat	`fat`

You can import this table into R using one of the methods described in Appendix A. ☞ 169 For example, you could first save the data file `nutrition_elderly.xls` in a csv file named `nutrition_elderly.csv` and then use the following command:

```
> nutri <- read.csv("nutrition_elderly.csv",header=TRUE)
```

This causes `nutri` to be stored as a so-called `data.frame` object in R — basically a list of columns. To check the type of your object you can used the R function `class()`.

```
> class(nutri)
 [1] "data.frame"
```

The R function `head()` gives the first few rows of the data table, including the variable names.

```
> head(nutri)
  gender situation tea coffee height weight age meat fish
1      2         1   0      0    151     58  72    4    3
2      2         1   1      1    162     60  68    5    2
3      2         1   0      4    162     75  78    3    1
4      2         1   0      0    154     45  91    0    4
5      2         1   2      1    154     50  65    5    3
6      2         1   2      0    159     66  82    4    2
  raw_fruit cooked_fruit_veg chocol fat
```

```
1         1                4      5   6
2         5                5      1   4
3         5                2      5   4
4         4                0      3   2
5         5                5      3   2
6         5                5      1   3
```

The names of the variables can also be obtained directly via the function `names()`, as in `names(nutri)`. This returns a list of all the names of the data frame. The data for each individual column (corresponding to a specific name) can be accessed by using `R`'s *list*$*name* construction. For example, `nutri$age` gives the vector of ages of the individuals in the nutrition data set.

☞ 180

Note that all the entries in `nutri` are *numerical* (that is, they are numbers). However, the *meaning* of each number depends on the respective columns. For example, a 1 in the "gender" column means here that the person is male (and 2 for female), while a 1 in the "fish" column indicates that this person eats fish less than once a week. Note also that it does not make sense to take the average of the values in the "gender" column, but it makes perfect sense for the "weights" column. To better manipulate the data it is important to specify exactly what the structure is of each variable. We discuss this next.

2.2 Structuring Variables According to Type

We can generally classify the measurement variables into two types: *quantitative* and *qualitative* (also called categorical). For quantitative variables we can make a distinction between continuous quantitative and discrete quantitative variables:

Continuous quantitative variables represent measurements that take values in a continuous range, such as the height of a person or the temperature of an environment. Continuous variables capture the idea that measurements can always be made more precisely.

Discrete quantitative variables have only a small number of possibilities, such as a count of some outcomes. For example in the data `nutri` the variable `tea` representing the daily number of tea cups is a discrete quantitative variable.

For qualitative variables (often called **factors**), we can distinguish between nominal and ordinal variables:

Nominal factors represent groups of measurements without order. For example, recording the sex of subjects is essentially the same as making a group of males and a group of females.

Ordinal factors represent groups of measurement that do have an order. A common example of this is the age group someone falls into. We can put these groups in order because we can put ages in order.

Example 2.1 The variable types for the data set nutri are given in Table 2.2.

Table 2.2: The variable types for the dataset nutri

Nominal	gender, situation, fat
Ordinal	meat, fish, raw_fruit, cooked_fruit_veg, chocol
Discrete quantitative	tea, coffee
Continuous quantitative	height, weight, age

Initially, all variables in nutri are identified as quantitative, because they happened to be entered as numbers[1]. You can check the type (or structure) of the variables with the command str(nutri).

```
> str(nutri)
 'data.frame':       226 obs. of  13 variables:
$ gender           : int  2 2 2 2 2 2 2 2 2 2 ...
$ situation        : int  1 1 1 1 1 1 1 1 1 1 ...
$ tea              : int  0 1 0 0 2 2 2 0 0 0 ...
$ coffee           : int  0 1 4 0 1 0 0 2 3 2 ...
$ height           : int  151 162 162 154 154 159 160 163 154
                          160 ...
$ weight           : int  58 60 75 45 50 66 66 66 60 77 ...
$ age              : int  72 68 78 91 65 82 74 73 89 87 ...
$ meat             : int  4 5 3 0 5 4 3 4 4 2 ...
$ fish             : int  3 2 1 4 3 2 3 2 3 3 ...
$ raw_fruit        : int  1 5 5 4 5 5 5 5 5 5 ...
$ cooked_fruit_veg: int  4 5 2 0 5 5 5 5 5 4 ...
$ chocol           : int  5 1 5 3 3 1 5 1 5 0 ...
$ fat              : int  6 4 4 2 2 3 6 6 6 3 ...
```

We shall now set up an adapted R structure for each variable.

2.2.1 Structuring Nominal Factors

For qualitative variables without order (that is, factor variables), set up the structure with the function as.factor(). It might be useful to also use the function levels() to recode the different levels of a qualitative variable. Let us perform these operations on the factor variables from our dataset:

```
> nutri$gender <- as.factor(nutri$gender)
> levels(nutri$gender) <- c("Male","Female")
> nutri$situation <- as.factor(nutri$situation)
> levels(nutri$situation) <- c("single","couple",
+                                   "family","other")
> nutri$fat <- as.factor(nutri$fat)
> levels(nutri$fat) <- c("butter","margarine","peanut",
+              "sunflower","olive","Isio4","rapeseed","duck")
```

[1] If gender had been entered as M and F, the variable would have automatically been structured as a factor. In the same way, the entries for the other two factor variables, situation and fat, could have been entered as letters or words.

2.2.2 Structuring Ordinal Factors

For ordinal factors, the structure can be set up with the function as.ordered(). As before it is possible to recode the different levels via the function levels(). Let us perform these operations on the ordinal factors from our dataset:

```
> nutri$meat <- as.ordered(nutri$meat)
> nutri$fish <- as.ordered(nutri$fish)
> nutri$raw_fruit <- as.ordered(nutri$raw_fruit)
> nutri$cooked_fruit_veg <- as.ordered(nutri$cooked_fruit_veg)
> nutri$chocol <- as.ordered(nutri$chocol)
> mylevels <- c("never","< 1/week.","1/week.","2-3/week.",
+              "4-6/week.","1/day")
> levels(nutri$chocol) <- levels(nutri$cooked_fruit_veg) <-
+                  levels(nutri$raw_fruit)  <- mylevels
> levels(nutri$fish) <- levels(nutri$meat) <- mylevels
```

2.2.3 Structuring Discrete Quantitative Data

For a quantitative variable that takes integer values, the structure is set up with the function as.integer().

```
> nutri$tea <- as.integer(nutri$tea)
> nutri$coffee <- as.integer(nutri$coffee)
```

Note that nutri$tea and nutri$coffee were initially classified as integer types anyway, so that the above assignments are superfluous.

2.2.4 Structuring continuous quantitative variables

For a continuous variable, the structure is set up with the function as.double().

```
> nutri$height <- as.double(nutri$height)
> nutri$weight <- as.double(nutri$weight)
> nutri$age <- as.double(nutri$age)
```

2.2.5 Good practice

We can now check using the R function str() the structure of our data.frame nutri.

```
> str(nutri)
 'data.frame':        226 obs. of  13 variables:
$ gender          : Factor w/ 2 levels "Male","Female": 2 2 2
                     2 2 2 2 2 2 2 ...
$ situation       : Factor w/ 4 levels "single","couple",..: 1
                     1 1 1 1 1 1 1 1 1 ...
$ tea             : int  0 1 0 0 2 2 2 0 0 0 ...
$ coffee          : int  0 1 4 0 1 0 0 2 3 2 ...
$ height          : num  151 162 162 154 154 159 160 163
                          154 160 ...
$ weight          : num  58 60 75 45 50 66 66 66 60 77 ...
$ age             : num  72 68 78 91 65 82 74 73 89 87 ...
$ meat            : Ord.factor w/ 6 levels "never"<"< 1/week."
                     <..: 5 6 4 1 6 5 4 5 5 3 ...
$ fish            : Ord.factor w/ 6 levels "never"<"< 1/week."
                     <..: 4 3 2 5 4 3 4 3 4 4 ...
$ raw_fruit       : Ord.factor w/ 6 levels "never"<"< 1/week."
                     <..: 2 6 6 5 6 6 6 6 6 6 ...
$ cooked_fruit_veg: Ord.factor w/ 6 levels "never"<"< 1/week."
                     <..: 5 6 3 1 6 6 6 6 6 5 ...
$ chocol          : Ord.factor w/ 6 levels "never"<"< 1/week."
                     <..: 6 2 6 4 4 2 6 2 6 1 ...
$ fat             : Factor w/ 8 levels "butter","margarine",..
                     : 6 4 4 2 2 3 6 6 6 3 ...
```

As the nature of each variables is well defined, we introduce the possibility to use the function `attach()` which gives direct access to the variables (columns) of a data.frame, by typing the name of a variable.

```
> attach(nutri)
> gender[1:3]
[1] Female Female Female
Levels: Male Female
> class(gender)
[1] "factor"
```

You can save your data in an R file (name extension must be .RData or .rda) by using the R function `save()`. This file could be loaded later by using the R function `load()`.

```
> save(nutri,file="nutrielderly.RData")
> detach(nutri)  # detach the nutri data.frame
> rm(list=ls())  # deleting all objects in the current workspace
> ls()           # list all objects

 character(0) #indicates that no objects are left.

> load("nutrielderly.RData")
> ls()
  [1] "nutri"
> attach(nutri) #access to the variables
```

In the remaining sections of this chapter we discuss various ways to extract summary information from our “raw” data in `nutri`. Which type of plots and numerical summaries can be performed depends strongly on the structure of the data table, and on the type of the variable(s) in play.

2.3 Summary Tables

It is often interesting to represent a large table of data in a more condensed form. A table of counts or a table of frequencies makes it easier to understand the underlying distribution of a variable, especially if the data are qualitative or ordinal. Such tables are obtained with the function `table()`.

```
> (tc <- table(fat))        # Table of counts.
fat
   butter margarine    peanut  sunflower     olive     Isio4
       15        27        48        68        40        23
 rapeseed      duck
        1         4
> (tf <- tc/length(fat))  # Frequency table.
fat
     butter   margarine      peanut   sunflower       olive
0.066371681 0.119469027 0.212389381 0.300884956 0.176991150
      Isio4    rapeseed        duck
0.101769912 0.004424779 0.017699115
```

Tip

The brackets around R expression as

```
> (x <- 2)
[1] 2
```

cause the value 2 that is stored in the object x to be printed. The same is achieved by

```
> x <- 2
> x
[1] 2
```

It is also possible to use table() to **cross tabulate** between two or more variables:

```
> (mytable <- table(gender,situation))
        situation
gender   single couple family other
  Male       20     63      2     0
  Female     78     56      7     0
```

To add summed margins to this table, use the function addmargins().

```
> (table.complete <- addmargins(mytable))
        situation
gender   single couple family other sum
  Male       20     63      2     0  85
  Female     78     56      7     0 141
  sum        98    119      9     0 226
```

2.4 Summary Statistics

In the following, $\boldsymbol{x} = (x_1, \dots, x_n)^\mathsf{T}$ is a column vector of numbers. For our nutri data set $\boldsymbol{x}$ could for example correspond to the heights of the $n = 226$ individuals.

Warning

Numerical summaries cannot be computed when some data are missing (`NA`). If necessary, missing data can be omitted with the function `na.omit()`.

```
> x <- na.omit(height) # Useless in this case since height
                       # has no NA.
```

The **mean** of the data of $x_1, \ldots, x_n$ is denoted by $\bar{x}$ and is simply the average of the data values:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

We will often refer to $\bar{x}$ as the **sample mean**, rather than "the mean of the data". Using the `mean` function in `R` for our `nutri` data, we have, for example:

```
> mean(height)
[1] 163.9602
```

The **median** of the data $x_1, \ldots, x_n$ is the value $\widetilde{x}$ "in the middle" of the data. More precisely, if we first *order* the data so that $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$, then

- if n is odd, then the median is the value $x_{\frac{n+1}{2}}$ — that is, the value at position $\frac{n+1}{2}$,
- if n is even, then any value between the values at positions $\frac{n}{2}$ and $\frac{n}{2} + 1$ can be used as a median of the series. In practice, the median is usually the average between these two values.

The `R` function to calculate the median is `median()`. For example,

```
> median(height)
[1] 163
```

The p-**quantile** $(0 < p < 1)$ of the data $x_1, \ldots, x_n$ is a value y such that a fraction p of the data is less than or equal to y and a fraction $1 - p$ of the data is greater than or equal to y. For example, the sample 0.5-quantile corresponds to the sample median. The p-quantile is also called the $100 \times p$ **percentile**. The 25, 50, and 75 sample percentiles are sometimes called the first, second, and third **quartiles**. Using `R` we have, for example,

```
> quantile(height,probs=c(0.1,0.9))
10% 90%
153 176
```

While the sample mean and median say something about the *location* of the data, it does not provide information about the *dispersion* (spread) of the data. The following summary statistics are useful for this purpose.

The **range** of the data $x_1, \ldots, x_n$ is given by

$$\text{range} = \max_{1 \leqslant i \leqslant n} x_i - \min_{1 \leqslant i \leqslant n} x_i \ .$$

In Rthe function `range()` returns the minimum and maximum of the data, so to get the actual range we have to take the difference of the two. For example,

```
> r <- range(height)
140 188
> diff(r)   # same as r[2] - r[1] or max(height) - min(height)
48
```

The **sample variance** of $x_1, \ldots, x_n$ is defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \,, \tag{2.1}$$

where $\bar{x}$ is the sample mean. We will see in later chapters that it plays an essential role in the analysis of statistical data. The square root of the sample variance $s = \sqrt{s^2}$ is called the **sample standard deviation**. In R we have, as an example,

```
> var(height)
81.06063
> sd(height)
9.003368
```

Tip

The function summary() applied to a vector of quantitative data calculates the minimum, maximum, mean and the three quartiles.

2.5 Graphical Representations

The way a variable is represented graphically should always be adapted to the type of variable.

2.5.1 Plotting Qualitative Variables

Bar charts

Suppose we wish to display graphically how many elderly are living by themselves, as a couple, with family, or other. Recall that, the data are given in the situation column of our nutri table. Assuming we already restructured the data as in Section 2.2, we can make a barplot of the number of people in each category using the R function barplot() as follows:

```
> col1 <- c("gray","orangered","lightgoldenrodyellow","red")
> barplot(table(situation),col=col1)
```

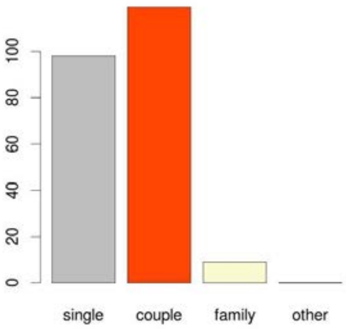


Figure 2.1: Bar chart for a qualitative variable.

Pie chart

A pie chart can be obtained with the function `pie()`. The following command makes a pie chart for the types of fat used for the `nutri` data set. Remember to first assign the type and levels for the `fat` column.

```
> pie(table(fat))
```

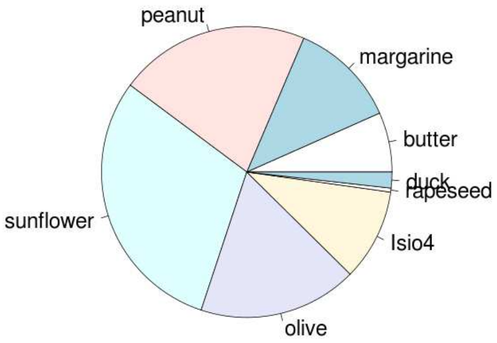


Figure 2.2: Pie chart for a qualitative variable.

2.5.2 Plotting Quantitative Variables

We now present a few useful graphs for exploring quantitative data. We will first focus on continuous variables (e.g., age) and then add some specific graphs related to discrete variables (e.g., tea). The aim is to describe the variability present in a single variable. The pattern of variability we see is called the *distribution* of the variable and this pattern typically involves a central tendency, where observations tend to gather around a central value, with fewer observations further away. The main aspects of the distribution are the *location* (or centre) of the variability, the *spread* of the variability (how far the values extend from the centre) and the *shape* of the variability in whether or not values are spread symmetrically on either side of the centre.

Boxplot

Such a chart can be obtained with function `boxplot()`, as shown below. The chart gives explanations.

```
> boxplot(age)
```

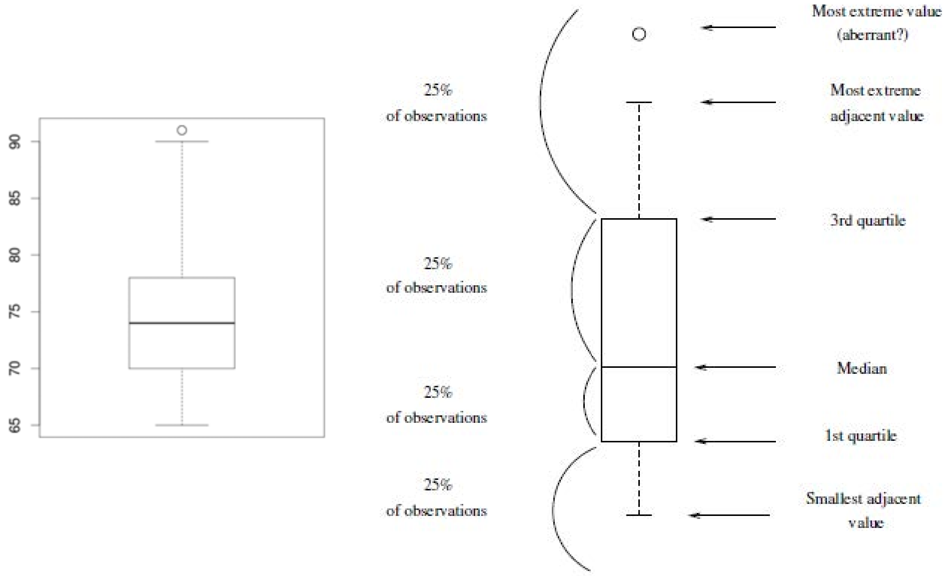


Figure 2.3: Boxplot produced by `R` (left side) and explanations (right side).

The box is drawn using the values of the three quartiles. Values represented as small circles are outliers, which might be suspect or deviant. These extreme values are those that are outside the box, further than 1.5 times the interquartile range (generally noted IQR). The interquartile range is defined as the distance between the third quartile (Q_3) and the first quartile (Q_1): IQR $= Q_3 - Q_1$. Note that values which are outside the box, but within 1.5 times the interquartile range, are called adjacent values. The whiskers are drawn at the largest and smallest adjacent value. Note that boxplot representation is also possible for discrete variables.

Histogram

A histogram is a main graphical representation of the distribution of a quantitative variable. We start by breaking the range of the values into a number of *bins* or *classes*. We tally the counts of the values falling in each bin and then make the plot by drawing rectangles whose bases are the bin intervals and whose heights are the counts. In `R` we can use the function `hist()`.

```
>  hist(age,prob=T,ylab="Density",col="orangered")
```

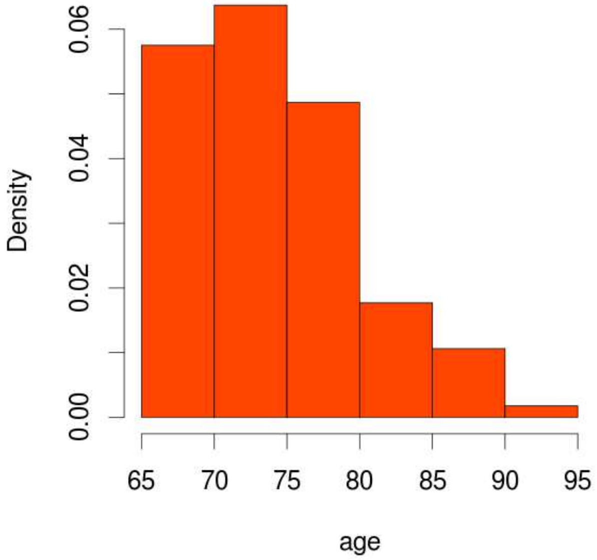


Figure 2.4: Histogram of variable age.

For example, Figure 2.4 shows a histogram of the 226 ages in data `nutri`. Here 6 bins were used. Rather than using raw counts, the vertical axis here gives the proportion in each class, the relative frequency, defined by $\frac{count}{total}$. An equivalent of a histogram for a discrete variable is obtained using the `plot()` function on a frequency table:

```
> plot(table(tea)/length(tea),col="darkolivegreen"
+                ,ylab="Probability",lwd=5)
```

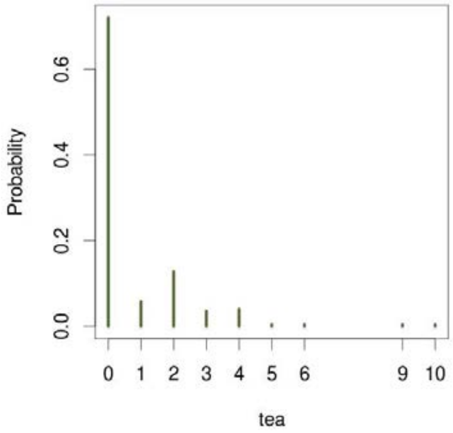


Figure 2.5: Bar chart for a discrete quantitative variable.

Empirical cumulative distribution function

The empirical cumulative distribution function, denoted by $F_n(\cdot)$, is a step function which jumps an amount k/n at observation values, where k is the number of tied observations at that value. For observations $(x_1, \ldots, x_n)$, $F_n(x)$ is the fraction of observations less than or equal to x, i.e.,

$$F_n(x) = \frac{\#\{x_i \leqslant x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leqslant x\}} \ ,$$

where $1_{\{x_i \leqslant x\}}$ is equal to 1 when $x_i \leqslant x$ and 0 otherwise. To produce the plot (see Figure 2.6) of the empirical cumulative distribution function using R, we need to combine the functions `plot()` and `ecdf()`.

```
> plot(ecdf(age),xlab="age")
```

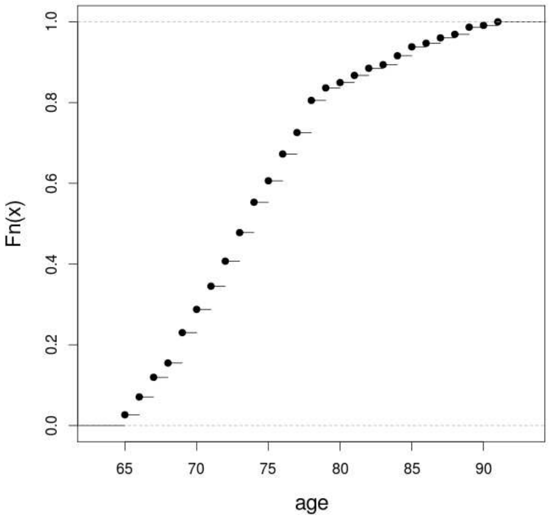


Figure 2.6: Plot of the empirical distribution function for a continuous quantitative variable.

The empirical distribution function for a discrete quantitative variable is obtained in the same way.

2.5.3 Graphical Representations in a Bivariate Setting

In this section, we present a few useful representations to explore relationships between two variables. The graphical representation will depend on the nature of the two variables.

Two-way plots for two qualitative variables

You can overlay two barplots, as shown on the next two figures.

```
> tss <- prop.table(table(gender,situation),1)
> barplot(tss,bes=T,leg=T)
```

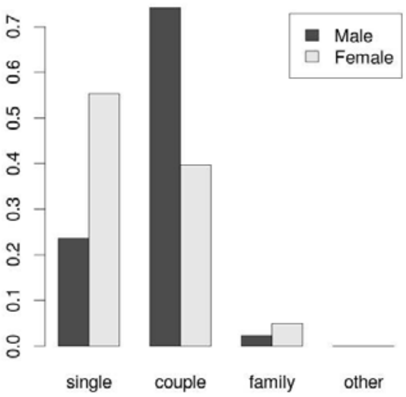


Figure 2.7: Bar plot for two qualitative variables.

Two-way plots for two quantitative variables

We can visualise patterns between two quantitative variables using a *scatter plot*. These are easily drawn by making two axes, one for each variable, and then using the values of the variables as the coordinates of a point to plot each case. We use the R function `plot()` to produce Figure 2.8.

```
> plot(weight~height)
```

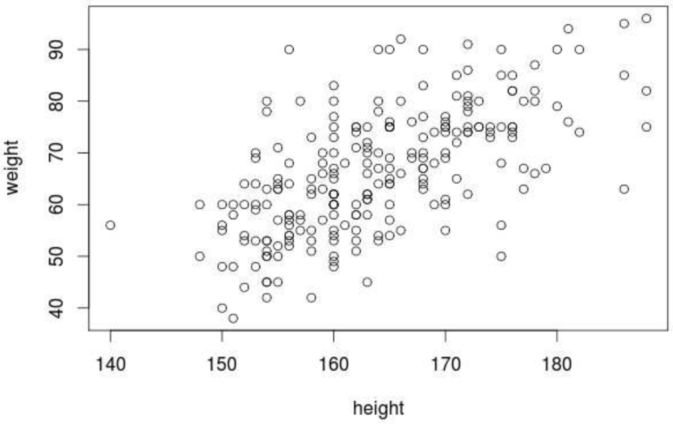


Figure 2.8: Plot of two quantitative variables.

Here we see the first use of an R **formula**: `weight ~ height`. It is a shorthand notation for describing the relation between two or more variables. Many statistics and plotting functions in R (such as the function `plot()`) allow formulas as input arguments. The output of the plotting or statistical function depends on the *types* (discrete qualitative, continuous quantitive, etc.) of the input variables. This is very important to realise when using R.

Two-way plots for one qualitative and one quantitative variable

In this setting, it is interesting to draw box plots of the quantitative variable for each level of the qualitative variable. If the variables are structured correctly in R, you simply need to call the function `plot()` which enables to produce Figure 2.9 .

```
> par(cex=1.5,bty="n")
> plot(coffee~gender,col="gray")
```

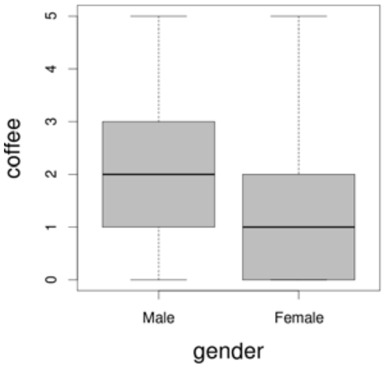


Figure 2.9: Box plots of a quantitative variable as a function of the levels of a qualitative variable.

R summary

`str()`: display the structure of each column of a data.frame

`attach()`: gives access to the variables of a data.frame

`as.factor()`: transform a variable into factors

`levels()`: display or change the levels of a factor

`as.ordered()`: transform a variable into ordered factors

`as.integer()`: structure a discrete variable

`as.double()`: structure a continuous variable

`table()`: table of counts for a variable, or contingency table between two variables

`addmargins()`: add margins to a contingency table

`margin.table()`: marginal distributions of a contingency table

`median()`: sample median

`mean()`: sample mean

`quantile()`: sample quantile

`summary()`: when applied to a numerical data, returns minimum, maximum, quartiles and mean of a sample

`range()`: minimum and maximum of a sample

`var()`: sample variance

`sd()`: sample standard deviation

`barplot()`: draw a bar chart

`pie()`: draw a pie chart

`plot(ecdf())`: plot the empirical cumulative distribution function

`boxplot()`: draw a box plot

`hist()`: draw a histogram

`plot()`: draw a scatter plot

Conclusion

- It is important to be able to identify the types of variables recorded in a study. Data from quantitative and qualitative variables are described and will be analysed in different ways.
- The sample mean and the sample median are used to measure the location of a distribution.
- The sample quantiles give a description of the distribution of a variable.
- The sample standard deviation gives a measure of the spread of the distribution.
- A box plot give a compact description of a distribution.
- Scatter plots are used to visualise the relationship between two quantitative variables.

2.6 Problems

1. Let us consider these observed values: $8, 6, 6, 4, 5, 6, 11, 8, 9$.

   (a) Calculate the mean and the standard deviation of these observations.

   (b) Calculate the median and the interquartile range.

(c) Check your results using R software

2. An experiment compared the reaction times of two groups of subjects, males and females. Each group was composed of twenty subjects selected on the criterion that they were between 18 and 21 years of age and participated in some form of sports/physical activity for at least three times a week. Each subject had a ruler placed between their thumb and forefinger at the 0 cm mark. The ruler was dropped and the distance it had travelled before they caught it was recorded. The results are shown in table 2.3.

Table 2.3: Reaction times (cm) between sexes

Male	14.2	16.0	19.8	21.9	15.3
	18.8	18.5	15.2	15.0	18.5
	16.1	15.2	17.4	12.8	17.3
	20.0	14.3	16.1	17.0	16.3
Female	18.9	14.1	15.5	13.4	17.3
	19.7	15.5	14.0	18.4	19.4
	16.5	17.8	14.7	15.2	16.6
	15.9	21.0	16.4	15.6	19.2

(a) Propose a summary of the reaction time for the male and for the female.

(b) Make side-by-side box plots of the reaction times.

(c) Describe the differences, if any, between male and female reaction times.

(d) Produce the same results using the R software.

3. A study asked 20 students to hold their breath for as long as was physically comfortable. The time breath was held was recorded with a stop-watch, along with the sex and height of the students. The results are given in Table 2.4. All participants were non-smokers and were seated during the experiment.

Table 2.4: Height (cm) and time breath held (s)

Female		Male	
Height	Breath Held	Height	Breath Held
175	22.22	184	60.75
158	30.57	182	67.41
166	17.47	180	42.19
175	22.39	191	59.74
160	26.90	189	52.64
165	36.85	181	43.37
166	27.33	180	73.27
170	29.55	170	59.09
170	13.87	176	51.15
172	34.66	185	58.32

(a) Make a histogram of the breath holding times, ignoring sex.

(b) Use the previous plot to describe the distribution of variability.

Chapter 3

Understanding Randomness

The purpose of this chapter is to introduce you to the language of *probability*, which is an indispensable tool for the understanding of randomness. You will learn how to think about random experiments in terms of probability models and how to calculate probabilities via counting. We will discuss how to describe random measurements via random variables and their distributions — specified by the cdf, pmf, and pdf. The expectation and variance of random variables provide important summary information about the distribution of a random variable.

3.1 Introduction

Statistical data is inherently random: if we would repeat the gathering of data we most likely would obtain (slightly) different measurements. There are various reasons why there is variability in the data.

To better understand the role that randomness plays in statistical analyses, we need to know a few things about the theory of *probability* first.

3.2 Random Experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but which is nevertheless subject to analysis. Examples of random experiments are:

1. tossing a die and observing its face value,
2. measuring the amount of monthly rainfall in a certain location,
3. choosing at random ten people and measuring their heights.
4. selecting at random fifty people and observing the number of left-handers,
5. conducting a survey on the nutrition of the elderly, resulting in a table of data such as `nutri` discussed in Chapter 2. ☞ 13

The goal of *probability* is to understand the behaviour of random experiments by analysing the corresponding *mathematical models*. Given a mathematical model for a random experiment one can calculate quantities of interest such as probabilities and expectations (defined later). Mathematical models for random experiments are also the basis of *statistics*, where the objective is to infer which of several competing models best fits the observed data. This often involves the estimation of model parameters from the data.

Example 3.1 (Coin Tossing) One of the most fundamental random experiments is the one where a coin is tossed a number of times. Indeed, much of probability theory can be based on this simple experiment. To better understand how this coin toss experiment behaves, we can carry it out on a computer, using programs such as R. The following R program simulates a sequence of 100 tosses with a fair coin (that is, Heads and Tails are equally likely), and plots the results in a bar chart.

```
> x <- runif(100)<0.5    # generate the coin tosses
> barplot(x)             # plot the results in a bar chart
```

The function `runif()` draws a vector of 100 uniform random numbers from the interval $[0, 1]$. By testing whether the uniform numbers are less than 0.5, we obtain a vector `x` of logical (TRUE or FALSE) variables, indicating Heads and Tails, say. Typical outcomes for three such experiments were given in Figure 1.1. ☞ 10

We can also plot the average number of Heads against the number of tosses. This is accomplished by adding two lines of code:

```
> y <- cumsum(x)/1:100 # calculate the cumulative  sum and divide
                       # elementwise by the vector 1:100
> plot(y,type="l")     # plot the result in a line graph
```

The result of three such experiments is depicted in Figure 3.1. Notice that the average number of Heads seems to converge to 0.5, but there is a lot of random fluctuation.

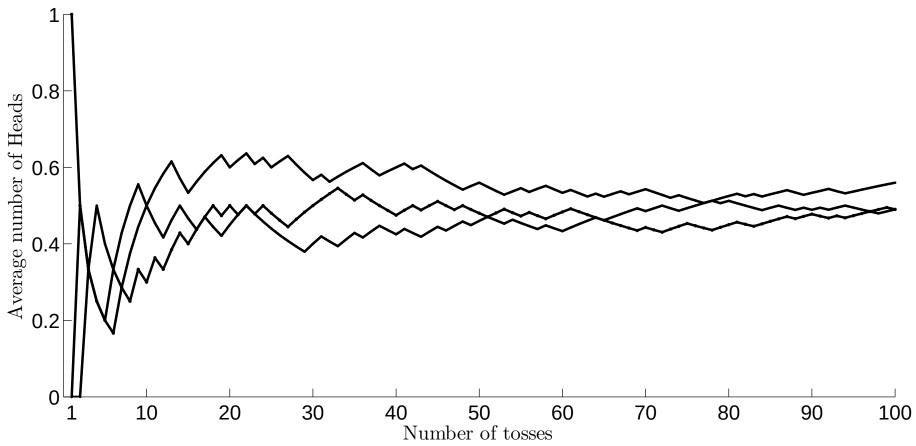


Figure 3.1: The average number of Heads in n tosses, where $n = 1, \ldots, 100$.

Similar results can be obtained for the case where the coin is *biased*, with a probability of Heads of p, say. Here are some typical *probability* questions.

- What is the probability of x Heads in 100 tosses?
- How many Heads would you expect to come up?
- What is the probability of waiting more than 4 tosses before the first Head comes up?

A statistical analysis would start from observed data of the experiment — for example, all the outcomes of 100 tosses are known. Suppose the probability of Heads p is not known. Typical *statistics* questions are:

- Is the coin fair?
- How can p be best estimated from the data?
- How accurate/reliable would such an estimate be?

To answer these questions we need to have a closer look at the models that are used to describe random experiments.

3.3 Probability Models

Although we cannot predict the outcome of a random experiment with certainty, we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

> **Definition 3.1** The **sample space** Ω of a random experiment is the set (collection) of all possible outcomes of the experiment.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively and observe their face values. A typical outcome could be written as a tuple (first die, second die). It follows that Ω is the set containing the outcomes $(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (6,6)$. There are thus $6 \times 6 = 36$ possible outcomes.
2. Measure the lifespan of a person in years. A possible outcome is for example 87.231 or 39.795. Any real number between 0 and, say, 140 would be possible. So, we could take Ω equal to the interval $[0, 140]$.
3. Measure the heights in metres of 10 people. We could write an outcome as a vector $(x_1, \ldots, x_{10})$, where the height of the first selected person is x_1, the height of the second person is x_2, and so on. We could take Ω to be the set of all positive vectors of length 10.

For modeling purposes it is often easier to take the sample space larger (but not smaller) than is strictly necessary. For example, in the second example we could have taken the set of real numbers as our sample space.

Often we are not interested in a single outcome but in whether or not one of a *group* of outcomes occurs.

Definition 3.2 An **event** is a subset of the sample space Ω to which a probability can be assigned.

Events will be denoted by capital letters $A, B, C, \ldots$. We say that event A **occurs** if the outcome of the experiment is one of the elements in A.

Examples of events for the three random experiments mentioned above are:

1. The event that the sum of two dice is 10 or more:

$$A = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\} \ .$$

2. The event that a person lives to become an octogenarian:

$$A = [80, 140) \ .$$

3. The event that the third selected person in the group of 10 is taller than 2 metres:

$$A = \{(x_1, \ldots, x_{10}) \text{ such that } x_3 > 2\} \, .$$

Note

Note that a list of numbers can be *ordered* or *unordered*. It is customary to write unordered lists (that is, sets) with curly brackets, and ordered lists (that is vectors) with round brackets. Hence, $\{1,2,3\}$ is the same as $\{3,2,1\}$, but the vector $(1,2,3)$ is not equal to $(3,2,1)$.

Since events are sets, we can apply the usual set operations to them, as illustrated in the *Venn diagrams* in Figure 3.2.

1. The set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur.

2. The set $A \cup B$ (A **union** B) is the event that A *or* B *or* both occur.

3. The event A^c (A **complement**) is the event that A does *not* occur.

4. If $B \subset A$ (B is a **subset** of A) then event B is said to *imply* event A.

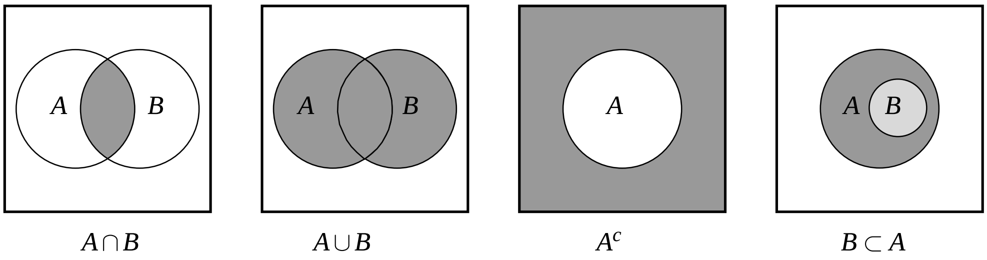


Figure 3.2: Venn diagrams of set operations. Each square represents the sample space Ω.

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$ (empty set), are called **disjoint** events.

Example 3.2 (Casting Two Dice) Suppose we cast two dice consecutively. The sample space is $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$. Let $A = \{(6,1),\ldots,(6,6)\}$ be the event that the first die is 6, and let $B = \{(1,6),\ldots,(6,6)\}$ be the event that the second die is 6. Then $A \cap B = \{(6,1),\ldots,(6,6)\} \cap \{(1,6),\ldots,(6,6)\} = \{(6,6)\}$ is the event that both dice are 6.

The third ingredient in the model for a random experiment is the specification of the probability of the events. It tells us how *likely* it is that a particular event will occur. We denote the probability of an event A by $\mathbb{P}(A)$ — note the special "black board bold" font. No matter how we define $\mathbb{P}(A)$ for different events A, the probability must always satisfy three conditions, given in the following definition.

Definition 3.3 A **probability measure** $\mathbb{P}$ is a function which assigns a number between 0 and 1 to each event, and which satisfies the following rules:

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$.
2. $\mathbb{P}(\Omega) = 1$.
3. For any sequence $A_1, A_2, \ldots$ of *disjoint* events we have

Sum Rule:
$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots \ . \tag{3.1}$$

The crucial property (3.1) is called the **sum rule** of probability. It simply states that if an event can happen in several distinct ways (expressed as a union of events, none of which are overlapping), then the probability that at least one of these events happens (that is, the probability of the union) is equal to the sum of the probabilities of the individual events. We see a similar property in an *area* measure: the total area of the union of nonoverlapping regions is simply the sum of the areas of the individual regions.

The following theorem lists some important consequences of the definition above. Make sure you understand the meaning of each of them, and try to prove them yourself, using *only* the three rules above; see Problem 1. ☞ 50

Theorem 3.1 (Properties of a Probability Measure). Let A and B be events and $\mathbb{P}$ a probability. Then,

1. $\mathbb{P}(\emptyset) = 0$,
2. if $A \subset B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$,
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

We have now completed our general model for a random experiment. Of course for any *specific* model we must carefully specify the sample space Ω and probability $\mathbb{P}$ that best describe the random experiment.

An important case where $\mathbb{P}$ is easily specified is where the sample space has a finite number of outcomes that are all *equally likely*. In this case the probability of an event $A \subset \Omega$ is simply

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of elements in } A}{\text{Number of elements in } \Omega}\,, \tag{3.2}$$

provided that the total number of elements in Ω is finite. The calculation of such probabilities thus reduces to *counting*.

3.4 Counting

Counting is not always easy. Let us first look at some examples:

1. A multiple choice form has 20 questions; each question has 3 choices. In how many possible ways can the exam be completed?

2. Consider a horse race with 8 horses. How many ways are there to gamble on the placings (1st, 2nd, 3rd).

3. Jessica has a collection of 20 CDs, she wants to take 3 of them to work. How many possibilities does she have?

To be able to comfortably solve a multitude of counting problems requires a lot of experience and *practice*, and even then, some counting problems remain exceedingly hard. Fortunately, many counting problems can be cast into the simple framework of drawing balls from an urn, see Figure 3.3.

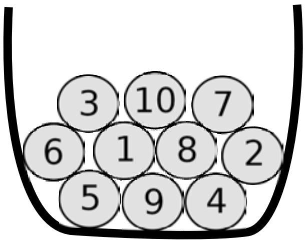


Figure 3.3: An urn with n balls

Consider an urn with n different balls, numbered $1, \ldots, n$ from which k balls are drawn. This can be done in a number of different ways. First, the balls can be drawn one-by-one, or one could draw all the k balls at the same time. In the first case the **order** in which the balls are drawn can be noted, in the second case that is not possible. In the latter case we can (and will) still assume the balls are drawn one-by-one, but that the order is not noted. Second, once a ball is drawn, it can either be put back into the urn (after the number is recorded), or left out. This is called, respectively, drawing with and without **replacement**. All in all there are 4 possible experiments: (ordered, with replacement), (ordered, without replacement), (unordered, without replacement) and (ordered, with replacement). The art is to recognise a seemingly unrelated counting problem as one of these four urn problems. For the three examples above we have the following

1. Example 1 above can be viewed as drawing 20 balls from an urn containing 3 balls, noting the order, and with replacement.

2. Example 2 is equivalent to drawing 3 balls from an urn containing 8 balls, noting the order, and without replacement.

3. In Example 3 we take 3 balls from an urn containing 20 balls, not noting the order, and without replacement.

We have left out the less important (and more complicated) unordered with replacement case. An example is counting how many different throws there are with 3 dice.

We now consider for each of the three cases how to count the number of arrangements. For simplicity we consider for each case how the counting works for $n = 4$ and $k = 3$, and then state the general situation. Recall the notation that we introduced in Remark 3.3, distinguishing ordered arrangements with round brackets and unordered ones with curly brackets. ☞ 34

Drawing with Replacement, Ordered

Here, after we draw each ball, note the number on the ball, and put the ball back. For our specific case $n = 4$ and $k = 3$ some possible outcomes are: $(1, 1, 1), (4, 1, 2)$, $(2, 3, 2)$, $(4, 2, 1), \ldots$ To count how many such arrangements there are, we can reason as follows: we have three positions $(\cdot, \cdot, \cdot)$ to fill. Each position can have the numbers 1, 2, 3, or 4, so the total number of possibilities is $4 \times 4 \times 4 = 4^3 = 64$. This is illustrated via the tree diagram in Figure 3.4.

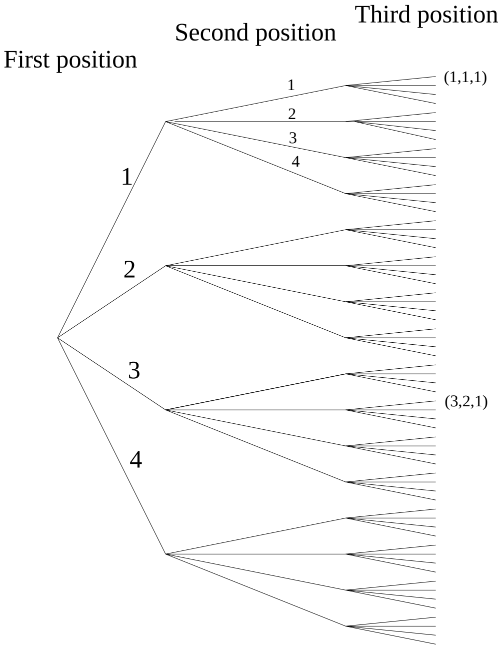


Figure 3.4: Enumerating the number of ways in which three ordered positions can be filled with 4 possible numbers, where repetition is allowed.

For general n and k we can reason analogously to find:

Theorem 3.2 The number of ordered arrangements of k numbers chosen from $\{1,\dots,n\}$, with replacement (repetition) is n^k.

Drawing Without Replacement, Ordered

Here we draw again k numbers (balls) from the set $\{1,2,\dots,n\}$, and note the order, but now do not replace them. Let $n = 4$ and $k = 3$. Again there are 3 positions to fill $(\cdot,\cdot,\cdot)$, but now the numbers cannot be the same, e.g., (1,4,2),(3,2,1), etc. Such an ordered arrangements called a **permutation** of size k from set $\{1,\dots,n\}$. (A permutation of $\{1,\dots,n\}$ of size n is simply called a permutation of $\{1,\dots,n\}$ (leaving out "of size n"). For the 1st position we have 4 possibilities. Once the first position has been chosen, we have only 3 possibilities left for the second position. And after the first two positions have been chosen there are 2 positions left. So the number of arrangements is $4 \times 3 \times 2 = 24$ as illustrated in Figure 3.5, which is the same tree as in Figure 3.4, but with all "duplicate" branches removed.

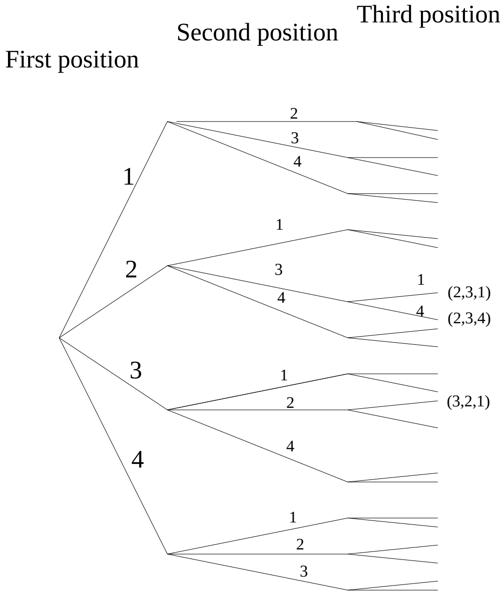


Figure 3.5: Enumerating the number of ways in which three ordered positions can be filled with 4 possible numbers, where repetition is NOT allowed.

For general n and k we have:

Theorem 3.3 The number of permutations of size k from $\{1,\ldots,n\}$ is ${}^nP_k = n(n-1)\cdots(n-k+1)$.

In particular, when $k = n$, we have that the number of ordered arrangements of n items is $n! = n(n-1)(n-2)\cdots 1$, where $n!$ is called ***n*-factorial**. Note that

$$^nP_k = \frac{n!}{(n-k)!}.$$

Drawing Without Replacement, Unordered

This time we draw k numbers from $\{1,\ldots,n\}$ but do not replace them (no replication), and do not note the order (so we could draw them in one grab). Taking again $n = 4$ and $k = 3$, a possible outcome is $\{1,2,4\}$, $\{1,2,3\}$, etc. If we noted the order, there would be ${}^nP_k$ outcomes, among which would be (1,2,4), (1,4,2), (2,1,4), (2,4,1), (4,1,2), and (4,2,1). Notice that these 6 permutations correspond to the single unordered arrangement $\{1,2,4\}$. Such unordered arrangements without replications are called **combinations** of size k from the set $\{1,\ldots,n\}$.

To determine the number of combinations of size k we simply need to divide ${}^nP_k$ by the number of permutations of k items, which is $k!$. Thus, in our example ($n = 4, k = 3$) there are $24/6 = 4$ possible combinations of size 3. In general we have:

Theorem 3.4 The number of combinations of size k from the set $\{1, \dots n\}$ is

$$^nC_k = \binom{n}{k} = \frac{^nP_k}{k!} = \frac{n!}{(n-k)!\,k!}\ .$$

Note the two different notations for this number.

Summarising, we have the following table:

Table 3.1: Number of ways k balls can be drawn from an urn containing n different balls.

	Replacement	
Order	Yes	No
Yes	n^k	^nP_k
No	—	^nC_k

Returning to our original three problems, we can now solve them easily:

1. The total number of ways the exam can be completed is $3^{20} = 3,486,784,401$.
2. The number of placings is $^8P_3 = 336$.
3. The number of possible combinations of CDs is $\binom{20}{3} = 1140$.

Once we know how to count, we can apply the equilikely principle to calculate probabilities:

1. What is the probability that out of a group of 40 people all have different birthdays?

   Answer: Choosing the birthdays is like choosing 40 balls with replacement from an urn containing the balls 1,. . . ,365. Thus, our sample space Ω consists of vectors of length 40, whose components are chosen from $\{1, \dots, 365\}$. There are $|\Omega| = 365^{40}$ such vectors possible, and all are *equally likely*. Let A be the event that all 40 people have different birthdays. Then, $|A| = {}^{365}P_{40} = 365!/325!$ It follows that $\mathbb{P}(A) = |A|/|\Omega| \approx 0.109$, so not very big!

2. What is the probability that in 10 tosses with a fair coin we get exactly 5 Heads and 5 Tails?

   Answer: Here Ω consists of vectors of length 10 consisting of 1s (Heads) and 0s (Tails), so there are 2^{10} of them, and all are *equally likely*. Let A be the event of exactly 5 heads. We must count how many binary vectors there are with exactly 5 1s. This is equivalent to determining in how many ways the positions of the 5 1s can be chosen out of 10 positions, that is, $\binom{10}{5}$. Consequently, $\mathbb{P}(A) = \binom{10}{5}/2^{10} = 252/1024 \approx 0.25$.

3. We draw at random 13 cards from a full deck of cards. What is the probability that we draw 4 Hearts and 3 Diamonds?

   Answer: Give the cards a number from 1 to 52. Suppose 1–13 is Hearts, 14–26 is Diamonds, etc. Ω consists of unordered sets of size 13, without repetition, e.g., $\{1, 2, \ldots, 13\}$. There are $|\Omega| = \binom{52}{13}$ of these sets, and they are all equally likely. Let A be the event of 4 Hearts and 3 Diamonds. To form A we have to choose 4 Hearts out of 13, and 3 Diamonds out of 13, followed by 6 cards out of 26 Spade and Clubs. Thus, $|A| = \binom{13}{4}\times\binom{13}{3}\times\binom{26}{6}$. So that $\mathbb{P}(A) = |A|/|\Omega| \approx 0.074$.

3.5 Conditional Probabilities

How do probabilities change when we know that some event B has occurred? Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$, which is called the *conditional probability* of A given B. The situation is illustrated in Figure 3.6.

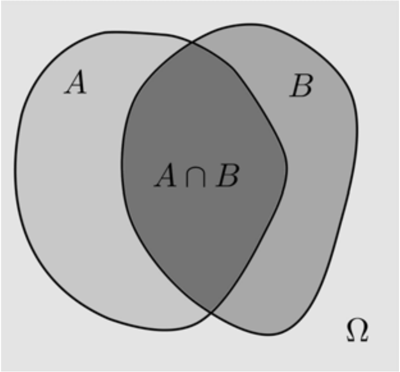


Figure 3.6: What is the probability that A occurs (that is, the outcome lies in A) given that the outcome is known to lie in B?

Definition 3.4 The **conditional probability** of A given B (with $\mathbb{P}(B) \neq 0$) is defined as:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} . \tag{3.3}$$

Example 3.3 (Casting Two Dice) We cast two fair dice consecutively. Given that the sum of the dice is 10, what is the probability that one 6 is cast? Let B be the event that the sum is 10:

$$B = \{(4, 6), (5, 5), (6, 4)\} .$$

Let A be the event that one 6 is cast:

$$A = \{(1,6),\dots,(5,6),(6,1),\dots,(6,5)\} .$$

Then, $A \cap B = \{(4,6),(6,4)\}$. And, since for this experiment all elementary events are equally likely, we have

$$\mathbb{P}(A \,|\, B) = \frac{2/36}{3/36} = \frac{2}{3} .$$

Independent Events

When the occurrence of B does not give extra information about A, that is $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$, the events A and B are said to be **independent**. A slightly more general definition (which includes the case $\mathbb{P}(B) = 0$) is that A and B are said to be independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) . \tag{3.4}$$

Example 3.4 (Casting Two Dice (Continued)) We cast two fair dice consecutively. Suppose A is the event that the first toss is 6 and B is the event that the second one is a 6, then naturally A and B are independent events, knowing that the first die is a 6 does not give any information about what the result of the second die will be. Let's check this formally. We have $A = \{(6,1),(6,2)\dots,(6,6)\}$ and $B = \{(1,6),(2,6),\dots,(6,6)\}$, so that $A \cap B = \{(6,6)\}$, and

$$\mathbb{P}(A \,|\, B) = \frac{1/36}{6/36} = \frac{1}{6} = \mathbb{P}(A) .$$

Product Rule

By the definition of conditional probability (3.3) we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \,|\, A) .$$

It is not difficult to generalise this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$, which we abbreviate as $A_1 A_2 \cdots A_n$. This gives the second major rule in probability: the **product rule**. We leave the proof as an exercise; see Problem 9.

Theorem 3.5 (Product Rule). Let $A_1, \dots, A_n$ be a sequence of events with $\mathbb{P}(A_1 \cdots A_{n-1}) > 0$. Then,

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}) . \tag{3.5}$$

Example 3.5 (Urn Problem) We draw consecutively 3 balls from an urn with 5 white and 5 black balls, without putting them back. What is the probability that all drawn balls will be black?

Let A_i be the event that the i-th ball is black. We wish to find the probability of $A_1 A_2 A_3$, which by the product rule (3.5) is

$$\mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) = \frac{5}{10}\,\frac{4}{9}\,\frac{3}{8} \approx 0.083 .$$

3.6 Random Variables and their Distributions

Specifying a model for a random experiment via a complete description of the sample space Ω and probability measure $\mathbb{P}$ may not always be necessary or convenient. In practice we are only interested in certain *numerical measurements* pertaining to the experiment. Such random measurements can be included into the model via the notion of a **random variable**. A random variable can be viewed as an observation of a random experiment that has not yet taken place. In other words, a random variable can be considered as a measurement that becomes available *tomorrow*, while all the thinking about the measurement can be carried out *today*. For example, we can specify today exactly the probabilities pertaining to the random variables.

We often denote random variables with *capital* letters from the last part of the alphabet, e.g., $X, X_1, X_2, \ldots, Y, Z$. Random variables allow us to use natural and intuitive notations for certain events, such as $\{X = 10\}$, $\{X > 1000\}$, $\{\max(X, Y) \leqslant Z\}$, etc.

Advanced

Mathematically, a random variable is a *function* which assigns a numerical value (measurement) to each outcome. An event such as $\{X > 1000\}$ is to be interpreted as the set of outcomes for which the corresponding measurement is greater than 1000.

We give some more examples of random variables without specifying the sample space.

1. The number of defective transistors out of 100 inspected ones.
2. The number of bugs in a computer program.
3. The amount of rain in a certain location in June.
4. The amount of time needed for an operation.

Similar to our discussion of the data types in Chapter 2, we distinguish between discrete and continuous random variables: ☞ 15

- **Discrete** random variables can only take *countably many* values.
- **Continuous** random variables can take a continuous range of values; for example, any value on the positive real line $\mathbb{R}_+$.

Let X be a random variable. We would like to designate the probabilities of events such as $\{X = x\}$ and $\{a \leqslant X \leqslant b\}$. If we can specify all probabilities involving X, we say that we have determined the **probability distribution** of X. One way to specify the probability distribution is to give the probabilities of all events of the form $\{X \leqslant x\}$. This leads to the following definition.

Definition 3.5 The **cumulative distribution function** (cdf) of a random variable X is the function F defined by

$$F(x) = \mathbb{P}(X \leqslant x), \quad x \in \mathbb{R} .$$

We have used $\mathbb{P}(X \leqslant x)$ as a shorthand notation for $\mathbb{P}(\{X \leqslant x\})$. From now on we will use this type of abbreviation throughout the notes. In Figure 3.7 the graph of a general cdf is depicted. Note that any cdf is increasing (if $x \leqslant y$ then $F(x) \leqslant F(y)$) and lies between 0 and 1. We can use any function F with these properties to specify the distribution of a random variable X.

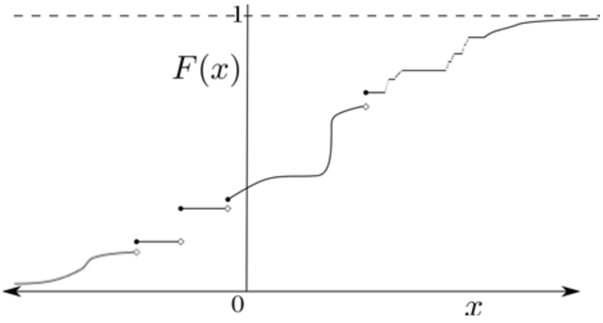


Figure 3.7: A cumulative distribution function (cdf).

If X has cdf F, then the probability that X takes a value in the interval $(a, b]$ (excluding a, including b) is given by

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) .$$

To see this, note that $\mathbb{P}(X \leqslant b) = \mathbb{P}(\{X \leqslant a\} \cup \{a < X \leqslant b\})$, where the events $\{X \leqslant a\}$ and $\{a < X \leqslant b\}$ are disjoint. Thus, by the sum rule: $F(b) = F(a) + \mathbb{P}(a < X \leqslant b)$, which leads to the result above.

Definition 3.6 A random variable X is said to have a **discrete distribution** if $\mathbb{P}(X = x_i) > 0$, $i = 1, 2, \ldots$ for some finite or countable set of values $x_1, x_2, \ldots$, such that $\sum_i \mathbb{P}(X = x_i) = 1$. The **probability mass function (pmf)** of X is the function f defined by $f(x) = \mathbb{P}(X = x)$.

We sometimes write f_X instead of f to stress that the pmf refers to the discrete random variable X. The easiest way to specify the distribution of a discrete random variable is to specify its pmf. Indeed, by the sum rule, if we know $f(x)$ for all x, then ☞ 35 we can calculate all possible probabilities involving X. In particular, the probability that X lies in some set B (say an interval (a, b)) is

$$\mathbb{P}(X \in B) = \sum_{x \in B} f(x) , \tag{3.6}$$

as illustrated in Figure 3.8. Note that $\{X \in B\}$ should be read as "X is an element of B".

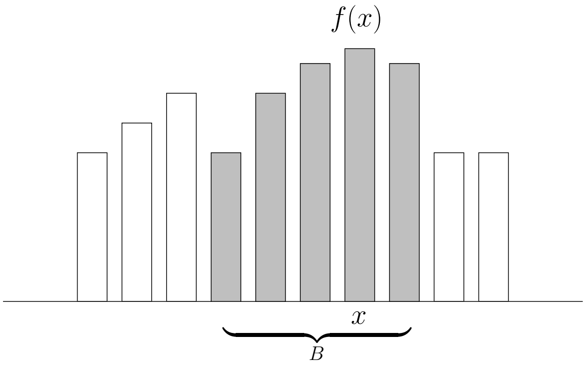


Figure 3.8: Probability mass function (pmf)

Example 3.6 (Sum of Two Dice) Toss two fair dice and let X be the sum of their face values. The pmf is given in Table 3.2; see Problem 16.

Table 3.2: Pmf of the sum of two fair dice.

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Definition 3.7 A random variable X with cdf F is said to have a **continuous distribution** if there exists a positive function f with *total integral 1* such that for all $a < b$,

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) = \int_a^b f(u)\,\mathrm{d}u\ . \tag{3.7}$$

Function f is called the **probability density function (pdf)** of X.

Warning

Note that we use the *same* notation f for both the pmf and pdf, to stress the similarities between the discrete and continuous case. Henceforth we will use the notation $X \sim f$ and $X \sim F$ to indicate that X is distributed according to pdf f or cdf F.

In analogy to the discrete case (3.6), once we know the pdf, we can calculate any probability that X lies in some set B by means of integration:

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x\ , \tag{3.8}$$

as illustrated in Figure 3.9.

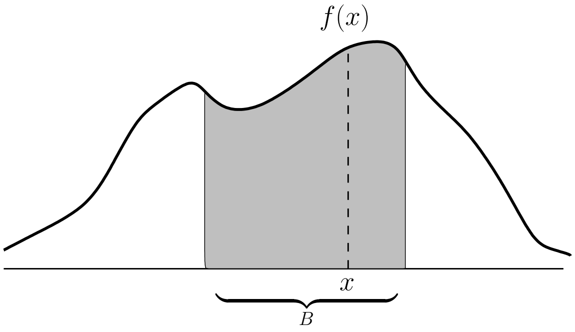


Figure 3.9: Probability density function (pdf)

Suppose that f and F are the pdf and cdf of a continuous random variable X, as in Definition 3.7. Then F is simply a *primitive* (also called anti-derivative) of f:

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u\ .$$

Conversely, f is the *derivative* of the cdf F:

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x) = F'(x)\ .$$

It is important to understand that in the continuous case $f(x)$ is not equal to the probability $\mathbb{P}(X = x)$, because the latter is 0 for all x. Instead, we interpret $f(x)$ as the *density* of the probability distribution at x, in the sense that for any small h,

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_{x}^{x+h} f(u)\,\mathrm{d}u \approx h\,f(x)\ . \tag{3.9}$$

Note that $\mathbb{P}(x \leqslant X \leqslant x + h)$ is equal to $\mathbb{P}(x < X \leqslant x + h)$ in this case.

Example 3.7 (Random Point in an Interval) Draw a random number X from the interval of real numbers $[0, 2]$, where each number is equally likely to be drawn. What are the pdf f and cdf F of X? We have

$$\mathbb{P}(X \leqslant x) = F(x) = \begin{cases} 0 & \text{if } x < 0,\\ x/2 & \text{if } 0 \leqslant x \leqslant 2,\\ 1 & \text{if } x > 2. \end{cases}$$

By differentiating F we find

$$f(x) = \begin{cases} 1/2 & \text{if } 0 \leqslant x \leqslant 2,\\ 0 & \text{otherwise.} \end{cases}$$

Note that this density is *constant* on the interval $[0, 2]$ (and zero elsewhere), reflecting the fact that each point in $[0, 2]$ is equally likely to be drawn.

3.7 Expectation

Although all probability information about a random variable is contained in its cdf or pmf/pdf, it is often useful to consider various numerical characteristics of a random variable. One such number is the *expectation* of a random variable, which is a "weighted average" of the values that X can take. Here is a more precise definition.

Definition 3.8 (Expectation of a Discrete Random Variable). Let X be a *discrete* random variable with pmf f. The **expectation** (or expected value) of X, denoted as $\mathbb{E}(X)$, is defined as

$$\mathbb{E}(X) = \sum_x x\,\mathbb{P}(X = x) = \sum_x x\,f(x)\,. \tag{3.10}$$

The expectation of X is sometimes written as μ_X. It is assumed that the sum in (3.10) is well-defined — possibly infinity (∞) or minus infinity ($-\infty$). One way to interpret the expectation is as a *long-run average payout*. Suppose in a game of dice the payout X (dollars) is the largest of the face values of two dice. To play the game a fee of d dollars must be paid. What would be a fair amount for d? If the game is played many times, the long-run fraction of tosses in which the maximum face value takes the value 1, 2,..., 6, is $\mathbb{P}(X = 1), \mathbb{P}(X = 2), \ldots, \mathbb{P}(X = 6)$, respectively. Hence, the long-run average payout of the game is the weighted sum of $1, 2, \ldots, 6$, where the weights are the long-run fractions (probabilities). So, the long-run payout is

$$\begin{aligned}\mathbb{E}(X) =& 1 \times \mathbb{P}(X = 1) + 2 \times \mathbb{P}(X = 2) + \cdots + 6 \times \mathbb{P}(X = 6)\\ &= 1 \times \frac{1}{36} + 2 \times \frac{3}{36} + 3 \times \frac{5}{36} + 4 \times \frac{7}{36} + 5 \times \frac{9}{36} + 6 \times \frac{11}{36} = \frac{161}{36} \approx 4.47\,.\end{aligned}$$

The game is "fair" if the long-run average profit $\mathbb{E}(X) - d$ is zero, so you should maximally wish to pay $d = \mathbb{E}(X)$ dollars.

Tip

For a *symmetric* pmf/pdf the expectation (if finite) is equal to the symmetry point.

For continuous random variables we can define the expectation in a similar way, replacing the sum with an integral.

Definition 3.9 (Expectation of a Continuous Random Variable). Let X be a *continuous* random variable with pdf f. The **expectation** (or expected value) of X, denoted as $\mathbb{E}(X)$, is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x\,. \tag{3.11}$$

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$, is also a random variable. The following theorem simply states that the expected value of a function of X is the weighted average of the values that this function can take.

Theorem 3.6 (Expectation of a Function of a Random Variable). If X is discrete with pmf f, then for any real-valued function g

$$\mathbb{E}(g(X)) = \sum_x g(x)\, f(x)\ .$$

Replace the sum with an integral for the continuous case.

Example 3.8 (Die Experiment and Expectation) Find $\mathbb{E}(X^2)$ if X is the outcome of the toss of a fair die. We have

$$\mathbb{E}(X^2) = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = \frac{91}{6}\ .$$

An important consequence of Theorem 3.6 is that the expectation is “linear”.

Theorem 3.7 (Properties of the Expectation). For any real numbers a and b, and functions g and h,

1. $\mathbb{E}(a\,X + b) = a\,\mathbb{E}(X) + b\ ,$
2. $\mathbb{E}(g(X) + h(X)) = \mathbb{E}(g(X)) + \mathbb{E}(h(X))\ .$

We show it for the discrete case. The continuous case is proven analogously, simply by replacing sums with integrals. Suppose X has pmf f. The first statement follows from

$$\mathbb{E}(aX + b) = \sum_x (ax + b) f(x) = a \sum_x x\, f(x) + b \sum_x f(x) = a\,\mathbb{E}(X) + b\ .$$

Similarly, the second statement follows from

$$\begin{aligned}\mathbb{E}(g(X) + h(X)) &= \sum_x (g(x) + h(x)) f(x) = \sum_x g(x) f(x) + \sum_x h(x) f(x)\\ &= \mathbb{E}(g(X)) + \mathbb{E}(h(X))\ .\end{aligned}$$

Another useful numerical characteristic of the distribution of X is the *variance* of X. This number, sometimes written as σ_X^2, measures the *spread* or dispersion of the distribution of X.

Definition 3.10 The **variance** of a random variable X, denoted as $\mathrm{Var}(X)$, is defined as

$$\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 \; , \tag{3.12}$$

where $\mu = \mathbb{E}(X)$. The square root of the variance is called the **standard deviation**. The number $\mathbb{E}X^r$ is called the r-th **moment** of X.

Theorem 3.8 (Properties of the Variance). For any random variable X the following properties hold for the variance.

1. $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
2. $\mathrm{Var}(a+bX) = b^2\,\mathrm{Var}(X)$.

To see this, write $\mathbb{E}(X) = \mu$, so that $\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. By the linearity of the expectation, the last expectation is equal to the sum $\mathbb{E}(X^2) - 2\,\mu\,\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - \mu^2$, which proves the first statement. To prove the second statement, note that the expectation of $a + bX$ is equal to $a + b\mu$. Consequently,

$$\mathrm{Var}(a+bX) = \mathbb{E}\left((a+bX-(a+b\mu))^2\right) = \mathbb{E}(b^2(X-\mu)^2) = b^2\mathrm{Var}(X) \; .$$

Conclusion

- Probability is used to describe random experiments: experiments whose outcome cannot be determined in advance.
- Important concepts in a probability model are: the sample space, events, the probability of an event, and random variables.
- Two of the most useful rules for calculating probabilities are the *sum rule* and the *product rule*. The first is about the union of (disjoint) events, the second about the intersection of events.
- The probability distribution of a random variable is completely specified by its cdf (cumulative distribution function). For discrete random variables it is more useful to specify the distribution via the pmf (probability mass function); for continuous random variables use instead the pdf (probability density function).
- The expectation (expected value) of a random variable is the weighted average of the values that a random variable can take. It is a measure for the locality of the distribution of the random variable.
- The variance is the expected squared distance from the random variable to its expected value, and is therefore a measure of the spread of the distribution of the random variable.

3.8 Problems

☞ 35 1. Prove Theorem 3.1 using only the three rules of probability in Definition 3.3.

2. Consider a random experiment with a finite number of equally likely outcomes ☞ 36 — so that the probabilities can be calculated via (3.2). Make a picture containing the sample space, the possible outcomes, and two disjoint sets A and B, that illustrates that the sum rule (3.1) clearly holds in this case.

3. We toss a fair coin three times.

   (a) Find the sample space, if we observe the exact sequences of Heads (= 1) and Tails (= 0).

   (b) Find the sample space, if we observe only the total number of Heads.

4. We randomly select 3 balls from an urn with 365 balls, numbered 1, . . . ,365, noting the order.

   (a) How many possible outcomes of the experiment are there, if we put each ball back into the urn before we draw the next?

   (b) Answer the same question as above, but now if we *don't* put the balls back.

   (c) Calculate the probability that in case (a) we draw 3 times the same ball.

5. Let $\mathbb{P}(A) = 0.9$ and $\mathbb{P}(B) = 0.8$. Show that $\mathbb{P}(A \cap B) \geqslant 0.7$.

6. What is the probability that none of 54 people in a room share the same birthday?

7. Consider the experiment of throwing 2 fair dice.

   (a) Find the probability that both dice show the same face.

   (b) Find the same probability, using the extra information that the sum of the dice is not greater than 4.

8. Solve the following counting problems by formulating them as drawing k balls from n.

   (a) In how many ways can the numbers 1,. . . ,5 be arranged, such as 13524, 25134, etc?

   (b) In a group of 20 people each person has a different birthday. How many different arrangements of these birthdays are there (assuming each year has 365 days)?

   (c) We draw three cards (at the same time) from a full deck of cards (52 cards). In how many different ways can we do this?

   (d) 10 balls, numbered 1,. . . ,10 are put in an urn. 5 are selected *without putting them back*, and are put in a row. How many possible arrangements are there of the ordered balls.

   (e) Same as above, but now the balls remain unordered.

(f) We again take 5 balls from the above urn, but now *put them back at each draw*. If we note the order, how many arrangements are there of the ordered balls.

9. Prove the product rule.

10. In Example 3.5 determine the probability $\mathbb{P}(A_1A_2A_3)$ via a counting argument.

11. Ten people stand in a line to "draw straws" (there are thus 9 long straws and 1 short straw).

    (a) What is the probability that the first person draws the short straw?

    (b) What is the probability that the 2nd person draws the short straw?

    (c) What is the probability that the 10th person draws the short straw?

12. We draw at random a number in the interval [0,1] such that each number is "equally likely". Think of the *random generator* on you calculator.

    (a) Determine the probability that we draw a number less than 1/2.

    (b) What is the probability that we draw a number between 1/3 and 3/4?

    (c) Suppose we do the experiment two times (independently), giving us two numbers in [0,1]. What is the probability that the sum of these numbers is greater than 1/2? Explain your reasoning.

13. Select at random 3 people from a large population. What is the probability that they all have the same birthday?

14. How many binary vectors are there of length 20 with exactly 5 ones?

15. Two fair dice are thrown and the smallest of the face values, Z say, is noted.

    (a) Give the pmf of Z in table form:

z	$*$	$*$	$*$	$\dots$
$\mathbb{P}(Z = z)$	$*$	$*$	$*$	$\dots$

    (b) Calculate the expectation of Z.

16. Let X be the sum of two dice, as in Example 3.6. Z be the sum of the face values.

☞ 45

    (a) Calculate $\mathbb{P}(Z \geqslant 9)$.

    (b) Calculate $\mathbb{P}(4 \leqslant Z \leqslant 8)$.

    (c) Determine the expectation of Z.

17. We select "uniformly" a point in the unit square: $\{(x, y) : 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1\}$. Let Z be the largest of the coordinates. Give the cdf and pdf of Z and draw their graphs.

18. A continuous random variable X has cdf F given by,

$$F(x) = \begin{cases} 0, & x < 0 \\ x^3, & x \in [0, 1] \\ 1 & x > 1\ . \end{cases}$$

(a) Determine the pdf of X.

(b) Calculate $\mathbb{P}(1/2 < X < 3/4)$.

(c) Calculate $\mathbb{E}[X]$.

Chapter 4

Common Distributions

This chapter presents four probability distributions that are the most frequently used in the study of statistics: the Bernoulli, Binomial, Uniform, and Normal distributions. We give various properties of these distributions and show how to compute probabilities of interest for them. You will also learn how to simulate random data from these distributions.

4.1 Introduction

In the previous chapter, we have seen that a random variable that takes values in a continuous set (such as an interval) is said to be *continuous* and a random variable that can have only a finite or countable number of different values is said to be discrete; see Section 3.6. Recall that the distribution of a continuous variable is specified by its ☞ 43 *probability density function* (pdf), and the distribution of a discrete random variable by its *probability mass function* (pmf).

In the following, we first present two distributions for discrete variables: the Bernoulli (or binary) and Binomial distributions. Then, we describe two popular distributions for continuous variables: the Uniform and Normal distributions. All of these distributions are actually *families* of distributions, which depend on a few (one or two in this case) **parameters**: fixed values that determine the shape of the distribution. Although in statistics we only employ a relatively small collection of distribution types (binomial, normal, etc), we can make an infinite amount of distributions through parameter selection.

4.2 Bernoulli Distribution

A **Bernoulli trial** is a random experiment that has only two possible outcomes, usually labeled “success” (or 1) and “failure” (or 0). The corresponding random variable X is called a Bernoulli variable. For example, a Bernoulli variable could model a single coin toss experiment by attributing the value 1 for Heads and 0 for Tails. Another example is selecting at random a person from some population and asking him if he/she approves of the prime minister or not.

Definition 4.1 A random variable X is said to have a **Bernoulli** distribution with success probability p if X can only assume the values 0 and 1, with probabilities

$$\mathbb{P}(X = 1) = p \quad \text{and} \quad \mathbb{P}(X = 0) = 1 - p\,.$$

We write $X \sim \mathsf{Ber}(p)$.

Figure 4.1 gives the pmf of a Bernoulli random variable.

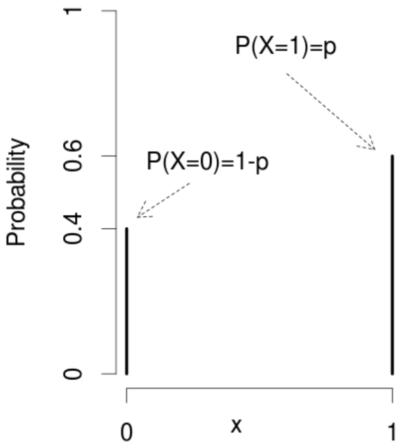


Figure 4.1: Probability mass function for the Bernoulli distribution, with parameter p (the case $p = 0.6$ is shown)

The expectation and variance of $X \sim \mathsf{Ber}(p)$ are easy to determine. We leave the proof as an exercise, as it is instructive do do it yourself, using the definitions of the expectation and variance; see (3.10) and (3.12). ☞ 47

Theorem 4.1 (Expectation and Variance of the Bernoulli Distribution). Let $X \sim \mathsf{Ber}(p)$. Then,

1. $\mathbb{E}(X) = p$
2. $\mathrm{Var}(X) = p(1 - p)$

4.3 Binomial Distribution

☞ 9 Let us go back the coin flip experiment of Example 1.1. In particular, we flip a coin 100 times and count the number of success (Heads), say X. Suppose that the coin is fair. What is the distribution of the total number of successes X? Obviously X can take any of the values 0,1,...,100. So let us calculate the probability of x successes: $\mathbb{P}(X = x)$ for $x = 0, 1, \ldots, 100$. In other words we wish to derive the pmf of X. In this case we can use a counting argument, as in Section 3.4. Namely, if we note the ☞ 36 sequence of 100 tosses, there are 2^{100} possible outcomes of the experiment, and they are all equally likely (with a fair coin). To calculate the probability of having exactly x successes (1s) we need to see how many of the possible outcomes have exactly x 1s and $100 - x$ 0s. There are $\binom{100}{x}$ of these, because we have to choose exactly x positions for the 1s out of 100 possible positions. In summary, we have derived

$$\mathbb{P}(X = x) = \frac{\binom{100}{x}}{2^{100}}, \quad x = 0, 1, 2, \ldots, 100 .$$

This is an example of a Binomial distribution. We can now calculate probabilities of interest such as $\mathbb{P}(X \geqslant 60)$, which we said in Example 1.1 was approximately equal to 0.028. Let us check this, using R as a calculator. We need to evaluate

$$\mathbb{P}(X \geqslant 60) = \sum_{x=60}^{100} \frac{\binom{100}{x}}{2^{100}} = \frac{\sum_{x=60}^{100} \binom{100}{x}}{2^{100}} .$$

We can do this in R in one line:

```
> sum(choose(100,60:100))/2^(100)


[1] 0.02844397
```

More generally, when we toss a coin n times and the probability of Heads is p (not necessarily 1/2), the outcomes are no longer equally likely (for example when p is close to 1 the sequence coin flips $1, 1, \ldots, 1$ is more likely to occur than $0, 0, \ldots, 0$). We can use the product rule (3.5) to find that the probability of having a particular sequence with x heads and $n - x$ tails is $p^x(1 - p)^{n-x}$; see Problem 8. Since there are $\binom{n}{x}$ of these sequences, we see that X has a $\mathsf{Bin}(n, p)$ distribution, as given in the following definition.

☞ 42

Definition 4.2 (Binomial Distribution). A random variable X is said to have a **Binomial** distribution with parameters n and p if X can only assume the integer values $x = 0, 1, \ldots, n$, with probabilities

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n . \tag{4.1}$$

We write $X \sim \mathsf{Bin}(n, p)$.

Figure 4.2 shows the pmf of the $\mathsf{Bin}(10, 0.7)$ distribution.

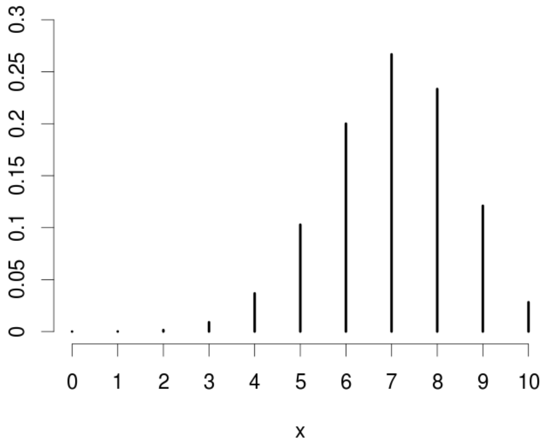


Figure 4.2: Probability mass function of the $\mathsf{Bin}(10, 0.7)$ distribution.

The following theorem lists the expectation and variance of the $\mathsf{Bin}(n, p)$. A simple proof will be given in the next chapter; see Example 5.4. In any case, expression for the expectation should come as no surprise, as we would expect np successes in a sequence of Bernoulli experiments (coin flips) with success probability p. Note that both the expectation and variance are n times the expectation and variance for a $\mathsf{Ber}(p)$. This is no coincidence, as a Binomial random variable can be seen as the sum of n independent Bernoulli random variables.

☞ 74

Theorem 4.2 (Expectation and Variance of the Binomial Distribution). Let $X \sim \mathsf{Bin}(n, p)$. Then,

1. $\mathbb{E}(X) = np$
2. $\mathrm{Var}(X) = np(1 - p)$

Tip

The number of successes in a series of n independent Bernoulli trials with success probability p has a $\mathsf{Bin}(n, p)$ distribution.

Counting the number of successes in a series of coin flip experiments might seem a bit artificial, but it is important to realise that many practical statistical situations can be treated exactly as a sequence of coin flips. For example, suppose we wish to conduct a survey of a large population to see what the proportion p is of males, where p is unknown. We can only know p if we survey *everyone* in the population, but suppose we do not have the resources or time to do this. Instead we select at random n people from the population and note their gender. We assume that each person is chosen with equal probability. This is very much like a coin flipping experiment. In fact if we allow the same person to be selected more than once, then the two situations are *exactly* the same. Consequently, if X is the total number of males in the group of n selected persons, then $X \sim \mathsf{Bin}(n, p)$. You might, rightly, argue that in practice we would not select the same person twice. But for a large population this would rarely happen, so the Binomial model is still a good model. For a small population, however, we should use a (more complicated) urn model to describe the experiment, where we draw balls (select people) without replacement and without noting the order. Counting for such experiments was discussed in Section 3.4. ☞ 36

4.4 Uniform Distribution

The simplest continuous distribution is the uniform distribution.

Definition 4.3 A random variable X is said to have a **uniform** distribution on the interval $[a, b]$ if its pdf is given by

$$f(x) = \frac{1}{b-a}, \quad a \leqslant x \leqslant b \qquad \text{(and } f(x) = 0 \text{ otherwise).} \tag{4.2}$$

We write $X \sim \mathsf{U}[a, b]$.

The random variable $X \sim \mathsf{U}[a, b]$ can model a randomly chosen point from the interval $[a, b]$, where each choice is equally likely. A graph of the density function is given in Figure 4.3. Note that the total area under the pdf is 1.

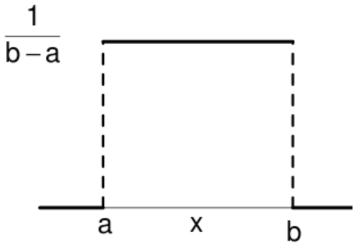


Figure 4.3: Probability density function for a uniform distribution on $[a, b]$

Theorem 4.3 (Properties of the Uniform Distribution). Let $X \sim \mathrm{U}[a, b]$. Then,

1. $\mathbb{E}(X) = (a + b)/2$
2. $\mathrm{Var}(X) = (b - a)^2/12$

Proof: The expectation is finite (since it must lie between a and b) and the pdf is symmetric. It follows that the expectation is equal to the symmetry point $(a + b)/2$. To find the variance, it is useful to write $X = a + (b - a)U$ where $U \sim \mathsf{U}[0, 1]$. In words: randomly choosing a point between a and b is equivalent to first randomly

☞ 49

choosing a point in [0, 1], multiplying this by $(b - a)$, and adding a. We can now write $\mathrm{Var}(X) = \mathrm{Var}(a + (b - a)U)$, which is the same as $(b - a)^2\mathrm{Var}(U)$, using the second property for the variance in Theorem 3.8. So, it suffices to show that $\mathrm{Var}(U) = 1/12$. Writing $\mathrm{Var}(U) = \mathbb{E}(U^2) - (\mathbb{E}(U))^2 = \mathbb{E}(U^2) - 1/4$, it remains to show that $\mathbb{E}(U^2) = 1/3$. This follows by direct integration:

$$\mathbb{E}(U^2) = \int_0^1 u^2 1\mathrm{d}u = \frac{1}{3}u^3\Big|_0^1 = \frac{1}{3}\,.$$

4.5 Normal Distribution

We now introduce the most important distribution in the study of statistics: the normal (or Gaussian) distribution.

Definition 4.4 A random variable X is said to have a **normal** distribution with parameters μ (expectation) and σ^2 (variance) if its pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ x \in \mathbb{R} \tag{4.3}$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$.

The parameters μ and σ^2 turn out to be the expectation and variance of the distribution, respectively. If $\mu = 0$ and $\sigma = 1$ then the distribution is known as the **standard normal** distribution. Its pdf is often denoted by φ (phi), so

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}.$$

The corresponding cdf is denoted by Φ (capital phi).

In Figure 4.4 the density function of the $\mathsf{N}(\mu, \sigma^2)$ distribution for various μ and σ^2 is plotted.

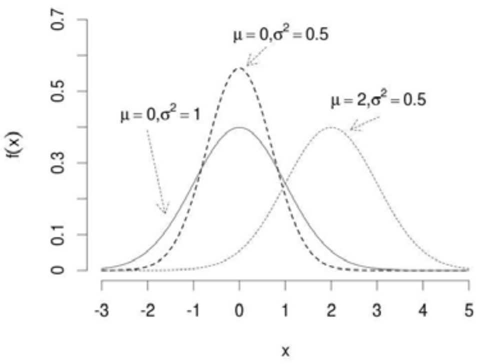


Figure 4.4: Probability density functions for various Normal distributions

Theorem 4.4 (Properties of the Normal Distribution). Let $X \sim \mathsf{N}(\mu, \sigma^2)$. Then,

1. $\mathbb{E}(X) = \mu$
2. $\mathrm{Var}(X) = \sigma^2$

We leave the proof as an advanced exercise; see Problem 14.

The normal distribution is symmetric about the expectation μ and the dispersion is controlled by the variance parameter σ^2, or the standard deviation σ (see Figure 4.4). An important property of the normal distribution is that any normal random variable can be thought of as a simple transformation of a standard normal random variable.

Theorem 4.5 If Z has standard normal distribution, then $X = \mu + \sigma Z$ has a $\mathsf{N}(\mu, \sigma^2)$ distribution. Consequently, if $X \sim \mathsf{N}(\mu, \sigma^2)$ then the **standardised** random variable

$$Z = \frac{X - \mu}{\sigma} \tag{4.4}$$

has a standard normal distribution.

Proof: Suppose Z is standard normal. So, $\mathbb{P}(Z \leqslant z) = \Phi(z)$ for all z. Let $X = \mu + \sigma Z$. We wish to derive the pdf f of X and show that it is of the form (4.3). We first derive the cdf F:

$$F(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(\mu + \sigma Z \leqslant x) = \mathbb{P}(Z \leqslant (x - \mu)/\sigma) = \Phi((x - \mu)/\sigma) .$$

By taking the derivative $f(x) = F'(x)$ we find (apply the chain rule of differentiation):

$$f(x) = F'(x) = \Phi'((x - \mu)/\sigma)\frac{1}{\sigma} = \varphi((x - \mu)/\sigma)/\sigma ,$$

which is the pdf of a $\mathsf{N}(\mu, \sigma^2)$-distributed random variable (replace x with $(x - \mu)/\sigma$ in the formula for φ and divide by σ. This gives precisely (4.3)).

By using the standardisation (4.4) we can simplify calculations involving arbitrary normal random variables to calculations involving only standard normal random variables.

Example 4.1 Standardisation can be viewed as a way to make comparisons between normal populations on the same scale. Suppose female heights are Normally distributed with mean 168 cm and variance 36 cm^2 and male heights are Normally distributed with mean 179 cm and variance 49 cm^2. Who is the more unusually tall for her/his gender, a female who is taller than 180 cm or a male who is taller than 200cm? Let us denote by X and Y the heights of a randomly selected woman and man, respectively. The probability that the female is taller than 180 cm is equal to

$$\begin{aligned} \mathbb{P}(X \geqslant 180) &= \mathbb{P}(X - 168 > 180 - 168) \\ &= \mathbb{P}\left(\frac{X - 168}{6} > \frac{180 - 168}{6}\right) \\ &= \mathbb{P}(Z \geqslant 2) = 1 - \mathbb{P}(Z \leqslant 2) = 1 - \Phi(2) . \end{aligned}$$

For the male we have, similarly,

$$\begin{aligned}\mathbb{P}(Y \geqslant 200) &= \mathbb{P}\left(\frac{Y-179}{7} > \frac{200-179}{7}\right)\\ &= \mathbb{P}(Z > 3) = 1 - \Phi(3)\ .\end{aligned}$$

Since $\Phi(3)$ is larger than $\Phi(2)$, finding a male to be taller than 2m is more unusual than finding a female taller than 180cm.

In the days before the computer it was customary to provide tables of $\Phi(x)$ for $0 \leqslant x \leqslant 4$, say. Nowadays we can simply use statistical software. For example, the cdf Φ is encoded in `R` as the function `pnorm()`. So to find $1 - \Phi(2)$ and $1 - \Phi(3)$ we can type:

```
> 1 - pnorm(2)
[1] 0.02275013


> 1 - pnorm(3)
[1] [1] 0.001349898
```

Unfortunately there is no simple formula for working out areas under the Normal density curve. However, as a rough rule for $X \sim \mathsf{N}(\mu, \sigma^2)$:

- the area within $c = 1$ standard deviation of the mean is 68%
- the area within $c = 2$ standard deviations of the mean is 95%
- the area within $c = 3$ standard deviations of the mean is 99.7%

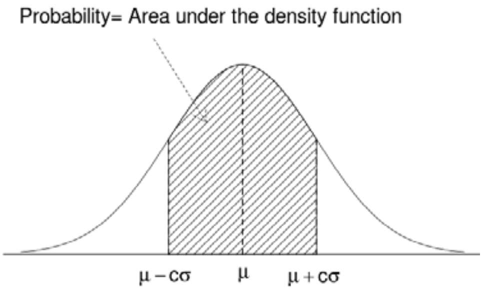


Figure 4.5: The area of the shaded region under the pdf is the probability $\mathbb{P}(|X - \mu| \leqslant c)$ that X lies less than c standard deviations (σ) away from its expectation (μ)

The function `pnorm()` can also be used to evaluate the cdf of general normal distribution. For example, let $X \sim \mathsf{N}(3, 4)$. Suppose we wish to find $\mathbb{P}(X \leqslant 3)$. In `R` we can enter:

```
> pnorm(3,mean=1,sd=2)
[1] 0.8413447
```

Note that `R` uses the standard deviation as an argument, not the variance!

We can also go the other way around: let $X \sim \mathsf{N}(3, 4)$. For what value z does it hold that $\mathbb{P}(X \leqslant z) = 0.9$. Such a value z is called a **quantile** of the distribution — in this case the 0.9-quantile. The concept is closely related to the *sample quantile* discussed in Section 2.4, but the two are not the same. Figure 4.6 gives an illustration. For the normal distribution the quantiles can be obtained via the R function qnorm(). ☞ 19

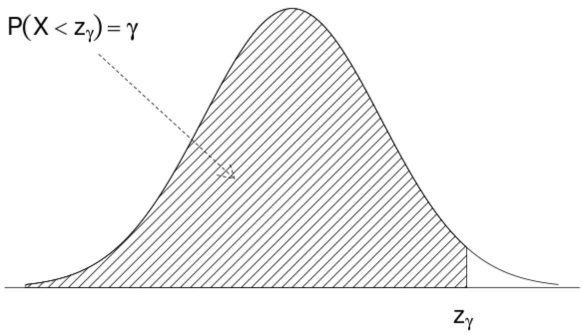


Figure 4.6: z_γ is the γ quantile of a normal distribution.

Here are some examples.

```
> qnorm(0.975)
[1] 1.959964

> qnorm(0.90,mean=1,sd=2)
[1] 3.563103

> qnorm(0.5,mean=2,sd=1)
[1] 2
```

4.6 Generating Random Variables

This section shows how to generate random variables on a computer. We first introduce R functions to generate observations from main distributions and then present some graphical tools to investigate the distribution of the simulated data.

Many computer programs have an inbuilt **random number generator**. This is a program that produces a stream of numbers between 0 and 1 that for all intent and purposes behave like independent draws from a uniform distribution on the interval [0,1]. Such numbers can be produced by the function runif(). For example

```
> runif(1)
[1] 0.6453129
```

Repeating gives a different number

```
> runif(1)
[1] 0.8124339
```

Or we could produce 5 such numbers in one go.

```
> runif(5)
[1] 0.1813849 0.9126095 0.2082720 0.1540227 0.9572725
```

We can use a uniform random number to simulate a toss with a fair coin by returning TRUE if $x < 0.5$ and FALSE if $x \geqslant 0.5$.

```
> runif(1) < 0.5
[1]  TRUE
```

We can turn the logical numbers into 0s and 1s by by using the function `as.integer()`

```
>  as.integer(runif(20)<0.5)
 [1] 1 1 0 1 0 0 1 0 1 0 0 1 1 1 1 1 0 0 0 0
```

We can, in principle, draw from *any* probability distribution including the normal distribution, using *only* uniform random numbers. However, to draw from a normal distribution we will use R's inbuilt `rnorm()` function. For example, the following generates 5 outcomes from the standard normal distribution:

```
> rnorm(5)
-1.1871560 -0.9576287 -1.2217339 -0.0412956  0.4981450
```

Tip

In R every function for generating **r**andom variables starts with an "`r`" (e.g., `runif()`, `rnorm()`). This is also holds for discrete random variables:

```
>  rbinom(1,size=10,p=0.5)
[1] 5
```

corresponds to the realisation of a random variable $X \sim \mathsf{Bin}(10, 0.5)$ and the instruction

```
>  rbinom(1,size=1,p=0.5)
[1] 1
```

corresponds to the realisation of a random variable $X \sim \mathsf{Ber}(0.5)$.

Generating artificial data can be a very useful way to understand probability distributions. For example, if we generate many realisations from a certain distribution, then
☞ 21 the histogram and empirical cdf of the data (see Section 2.5) will resemble closely the
true pdf/pmf and cdf of the distribution. Moreover the summary statistics (see Sec-
☞ 19 tion 2.4) of the simulated data such as the sample mean and sample quantiles will resemble the true distributional properties such as the expected value and the quantiles. Let us illustrate this by drawing one 10,000 samples from the $\mathsf{N}(2, 1)$ distribution.

```
> x <- rnorm(10e4,mean=2,sd=1)
> summary(x)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.573   1.328   1.997   1.997   2.670   5.865
```

The true first and third quartiles are 1.32551 and 2.67449, respectively, which are quite close to the sample quartiles. Similarly the true expectation and median are 2, which is again close to the sample mean and sample median.

The following R script (program) was used to produce Figure 4.7. We see a very close correspondence between the true pdf (on the left, in red) and a histogram of the 10,000 data points. The true cdf (on the right, in red) is virtually indistinguishable from the empirical cdf.

```
# simnorm.R
par(mfrow=c(1,2),cex=1.5)     # two plot windows, use larger font
x <- rnorm(10e4,mean=2,sd=1)  # generate data
hist(x,prob=TRUE,breaks=100)  # make histogram
curve(dnorm(x,mean=2,sd=1),col="red",ylab="",lwd=2,add=T)  #true pdf
plot(ecdf(x)) # draw the empirical cdf
curve(pnorm(x,mean=2,sd=1),col="red",lwd=1,add=TRUE)       #true cdf
```

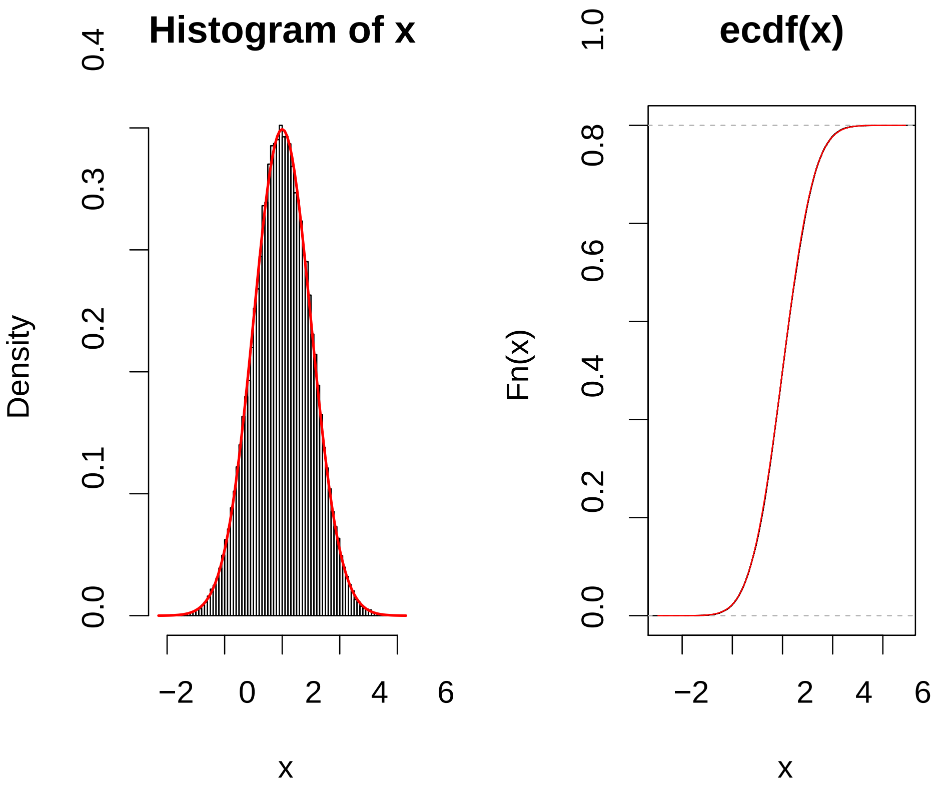


Figure 4.7: Left: pdf of the N(2, 1) distribution (red) and histogram of the generated data. Right: cdf of the N(2, 1) distribution (red) empirical cdf of the generated data.

Tip

Density functions (pmf or pdf) always start in R with “d” (e.g., dnorm(), dunif()). The cummulative distribution functions (cdf), which give a **p**robability, always start in R with “p” (e.g., pnorm(), punif()). **Q**uantiles start with “q” (e.g., qnorm(),qunif()).

To summarize, we present in table 4.1 the main R functions for the evaluation of densities, cumulative distribution functions, quantiles, and the generation of random variables for the distributions described in this chapter. Later on we will encounter more distributions such as the Student’s *t* distribution, the *F* distribution, and the chi-squared distribution. You can use the “d”, “p”, “q” and “r” construction to evaluate pmfs, cdfs, quantiles, and random numbers in exactly the same way!

Table 4.1: Standard discrete and continuous distributions. R functions for the mass or density function (d--), cumulative distribution function (p--) and quantile function (q--). Instruction to generate (r--) pseudo-random numbers from these distributions.

Distr.	**R functions**	**Distr.**	**R functions**
Ber(p)	dbinom(x,size=1,prob=p)	N(μ, σ^2)	dnorm(x,mean=μ,sd=σ)
	pbinom(x,size=1,prob=p)		pnorm(x,mean=μ,sd=σ)
	qbinom(γ,size=1,prob=p)		qnorm(γ,mean=μ,sd=σ)
	rbinom(n,size=1,prob=p)		rnorm(n,mean=μ,sd=σ)
Bin(n, p)	dbinom(x,size=1,prob=p)	U[a, b]	dunif(x,min=a,max=b)
	pbinom(x,size=n,prob=p)		punif(x,min=a,max=b)
	qbinom(γ,size=n,prob=p)		qunif(γ,min=a,max=b)
	rbinom(n,size=n,prob=p)		runif(n,min=a,max=b)

Conclusion

- A Bernoulli trial is a random experiment with only two possible outcomes (success/failure).
- The total number of successes in n independent Bernoulli trials with success probability p has a Bin(n, p) (binomial) distribution. The expectation and variance are np and $np(1 - p)$, respectively.
- The Uniform distribution has a pdf that is constant over an interval. It models drawing at random a point from this interval where all possible values are equally likely.
- The Normal distribution is often used to model continuous random measurements.
- $X \sim \mathsf{N}(\mu, \sigma^2)$ indicates that the continuous random variable X has a Normal distribution with expectation (mean) μ and variance σ^2.

- If $X \sim \mathsf{N}(\mu, \sigma^2)$ then the standardised random variable $Z = (X - \mu)/\sigma$ has a standard Normal distribution.
- In R software:
  - d–: density function, pmf/pdf (e.g.: `dnorm()`, `dunif()`)
  - p–: cumulative distribution function (e.g.: `pnorm()`, `punif()`)
  - q–: quantile function (e.g.: `qnorm()`, `qunif()`, `qbinom()`)
  - r–: random variable generation (e.g.: `rnorm()`, `runif()`, `rbinom()`)

4.7 Problems

1. Let $X \sim \mathsf{Ber}(p)$. Using the definition of expectation and variance,
   (a) show that $\mathbb{E}(X) = p$,
   (b) show that $\mathrm{Var}(X) = p(1 - p)$.

2. Let $X \sim \mathsf{Bin}(4, 1/2)$. Give the pmf of X in table form.

x	*	*	*	...
$\mathbb{P}(X = x)$	*	*	*	...

   What is the pmf of $Y = X^2$ in table form?

3. Let $X \sim \mathsf{U}[1, 4]$. Sketch the pdf and cdf. Calculate $\mathbb{P}(X \leqslant 2)$ and $\mathbb{P}(2 < X < 3)$.

4. Let $X \sim \mathsf{N}(0, 1)$, and $Y = 1 + 2X$. What is the pdf of Y?

5. Let $X \sim \mathsf{N}(0, 1)$. Find $\mathbb{P}(X \leqslant 1.3)$ using R. From this, and using the symmetry of the standard normal distribution, find $\mathbb{P}(X > -1.3)$.

6. Let $Y \sim \mathsf{N}(1, 4)$. Express $\mathbb{P}(Y \leqslant 3)$ in terms of the cdf of the standard normal distribution (Φ). Do the same for $\mathbb{P}(-1 \leqslant Y \leqslant 2)$. Check your answers with R.

7. We draw at random 5 numbers from 1,. . . , 100, *with replacement* (for example, drawing number 9 twice is possible). What is the probability that exactly 3 numbers are even?

8. Consider a coin toss experiment where a biased coin with success probability p is tossed n times. Let A_i be the event that the ith toss is Heads. The event that the first x tosses are Heads and the next $n - x$ are Tails can thus be written as

$$A = A_1 \cap A_2 \cap \cdots \cap A_x \cap A_{x+1}^c \cap \cdots \cap A_n^c,$$

   where A_k^c means that the event that the kth toss is tails (which is the complement of A_k). Using the product rule (3.5) and the independence of $A_1, A_2, \ldots$ show that $\mathbb{P}(A) = p^x(1 - p)^{n-x}$. ☞ 42

9. If $X \sim \mathsf{U}[0, 1]$, what is the expectation of $Y = 10 + 2X$?

10. Suppose the lymphocyte count from a blood test has a Normal distribution with mean $2.5{\times}10^9/\mathrm{L}$ and standard deviation $0.765 \times 10^9/L$. What is the probability that a randomly chosen blood test will have a lymphocyte count between $2.3 \times 10^9/\mathrm{L}$ and $2.9 \times 10^9/\mathrm{L}$?

11. Suppose the copper level from a blood test has a Normal distribution with mean 18.5μmol/L and standard deviation 3.827 μmol/L. What is the lowest copper level that would put a blood test result in the highest 1%?

12. Generate 10,000 draws from the Uniform distribution on the interval [1,2]. Compare the summary statistics with the true ones, exactly as was done for the $\mathsf{N}(2, 1)$ distribution in Section 4.6. Also compare the histogram and empirical cdf with the pdf and the cdf of the distribution, as in Figure 4.7.

13. The random numbers produced by the computer are not truly random, as they are generated by a deterministic algorithm. For this reason they are often referred to as *pseudo random* numbers. Sometimes it is useful to repeat a random sequence of numbers. We can do this by setting the so-called **seed** of the random number generator. Try the following in R:

```
> runif(5)
> runif(5)
> set.seed(1234)
> runif(5)
> set.seed(1234)
> runif(5)
```

What do you observe?

14. Let $X \sim \mathsf{N}(\mu, \sigma^2)$. We can write (see (4.4)) $X = \mu + \sigma Z$, where $Z \sim \mathsf{N}(0, 1)$. Using the rules for expectation and variance we have $\mathbb{E}(X) = \mu + \sigma\mathbb{E}(Z)$ and $\mathrm{Var}(X) = \sigma^2\mathrm{Var}(Z)$. To prove Theorem 4.4 it therefore suffices to show that $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

    (a) Show that $\mathbb{E}(Z) = 0$.

    (b) Show that if $\mathbb{E}(Z) = 0$, then $\mathrm{Var}(Z) = \mathbb{E}(Z^2)$.

    (c) Show that

$$\mathbb{E}(Z^2) = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z = \int_{-\infty}^{\infty} \underbrace{z}_{\text{take derivative}}\ \underbrace{z\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}z^2}}_{\text{take primitive}}\,\mathrm{d}z = 1$$

    using partial integration, as indicated.

Chapter 5

Multiple Random Variables

In this chapter you will learn how random experiments that involve more than one random variable can be described via their joint cdf and joint pmf/pdf. When the random variables are *independent* of each other, the joint density has a simple product form. We will discuss the most basic statistical model for data — independent and identically distributed (iid) draws from a common distribution. We will show that the expectation and variance of sums of random variables obey simple rules. We will also illustrate the *central limit theorem*, explaining the central role that the normal distribution has in statistics. The chapter concludes with the conceptual framework for statistical modeling and gives various examples of simple models.

5.1 Introduction

In the previous chapters we considered random experiments that involved only a single random variable, such as the number of heads in 100 tosses, the number of left-handers in 50 people, or the amount of rain on the 2nd of January 2014 in Brisbane. This is obviously a simplification: in practice most random experiments involve multiple random variables. Here are some examples of experiments that we could do "tomorrow".

1. We randomly select $n = 10$ people and observe their heights. Let $X_1, \ldots, X_n$ be the individual heights.

2. We toss a coin repeatedly. Let $X_i = 1$ if the ith toss is Heads and $X_i = 0$ otherwise. The experiment is thus described by the sequence $X_1, X_2, \ldots$ of Bernoulli random variables.

3. We randomly select a person from a large population and measure his/her mass X and height Y.

4. We simulate 10,000 realisations from the standard normal distribution using the `rnorm()` function. Let $X_1, \ldots, X_{10{,}000}$ be the corresponding random variables.

How can we specify the behavior of the random variables above? We should not just specify the pdf of the individual random variables, but also say something about the interaction (or lack thereof) between the random variables. For example, in the third experiment above if the height Y is large, then most likely the mass X is large as well. In contrast, in the first two experiments it is reasonable to assume that the random variables are "independent" in some way; that is, information about one of the random variables does not give extra information about the others. What we need to specify is the **joint distribution** of the random variables. The theory below for multiple random variables follows a similar path to that of a single random variable described in Section 3.6. ☞ 43

Let $X_1, \ldots, X_n$ be random variables describing some random experiment. Recall that the distribution of a *single* random variable X is completely specified by its cumulative distribution function. For *multiple* random variables we have the following generalisation.

Definition 5.1 The **joint cdf** of $X_1, \ldots, X_n$ is the function F defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n) \; .$$

Notice that we have used the abbreviation $\mathbb{P}(\{X_1 \leqslant x_1\} \cap \cdots \cap \{X_n \leqslant x_n\}) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n)$ to denote the probability of the intersection of events. We will use this abbreviation from now on.

As in the univariate (that is, single-variable) case we distinguish between *discrete* and *continuous* distributions.

5.2 Joint Distributions

Example 5.1 (Dice Experiment) In a box there are three dice. Die 1 is an ordinary die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces. The experiment consists of selecting a die at random followed by a toss with that die. Let X be the die number that is selected and let Y be the face value of that die. The probabilities $\mathbb{P}(X = x, Y = y)$ in Table 5.1 specify the joint distribution of X and Y. Note that it is more convenient to specify the joint probabilities $\mathbb{P}(X = x, Y = y)$ than the joint cumulative probabilities $\mathbb{P}(X \leqslant x, Y \leqslant y)$. The latter can be found, however, from the former by applying the sum rule. For example, $\mathbb{P}(X \leqslant 2, Y \leqslant 3) = \mathbb{P}(X = 1, Y = 1) + \cdots + \mathbb{P}(X = 2, Y = 3) = 6/18 = 1/3$. Moreover, by that same sum rule, the distribution of X is found by summing the $\mathbb{P}(X = x, Y = y)$ over all values of y — giving the last column of Table 5.1. Similarly, the distribution of Y is given by the column totals in the last row of the table.

Table 5.1: The joint distribution of X (die number) and Y (face value).

		y						
		1	2	3	4	5	6	Σ
	1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
x	2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	0	$\frac{1}{3}$
	3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{1}{9}$	$\frac{1}{3}$
	Σ	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

In general, for discrete random variables $X_1, \ldots, X_n$ the joint distribution is easiest to specify via the joint pmf.

Definition 5.2 The **joint pmf** f of discrete random variables $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) .$$

We sometimes write $f_{X_1,\ldots,X_n}$ instead of f to show that this is the pmf of the random variables $X_1, \ldots, X_n$. To save on notation, we can refer to the sequence $X_1, \ldots, X_n$ simply as a random "vector" $\mathbf{X} = (X_1, \ldots, X_n)$. If the joint pmf f is known, we can calculate the probability of any event via summation as

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{\mathbf{x} \in B} f(\mathbf{x}) . \tag{5.1}$$

That is, to find the probability that the random vector lies in some set B (of dimension n), all we have to do is sum up all the probabilities $f(\mathbf{x})$ over all $\mathbf{x}$ in the set B. This is simply a consequence of the sum rule and a generalisation of (3.6). In particular, as illustrated in Example 5.1, we can find the pmf of X_i — often referred to as a **marginal** pmf, to distinguish it from the joint pmf — by summing the joint pdf over all possible values of the other variables. For example,

☞ 44

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y) . \tag{5.2}$$

The converse is not true: from the marginal distributions one cannot in general reconstruct the joint distribution. For example, in Example 5.1 we cannot reconstruct the inside of the two-dimensional table if only given the column and row totals.

For the continuous case we need to replace the joint pmf with the joint pdf.

Definition 5.3 The **joint pdf** f of continuous random variables $X_1, \ldots, X_n$ (summarised as $\mathbf{X}$) is the positive function with total integral 1 such that

$$\mathbb{P}(\mathbf{X} \in B) = \int_{\mathbf{x} \in B} f(\mathbf{x})\, \mathrm{d}\mathbf{x} \quad \text{for all sets } B . \tag{5.3}$$

The integral in (5.3) is now a multiple integral — instead of evaluating the area under f, we now need to evaluate the (n-dimensional) volume. Figure 5.1 illustrates the concept for the 2-dimensional case.

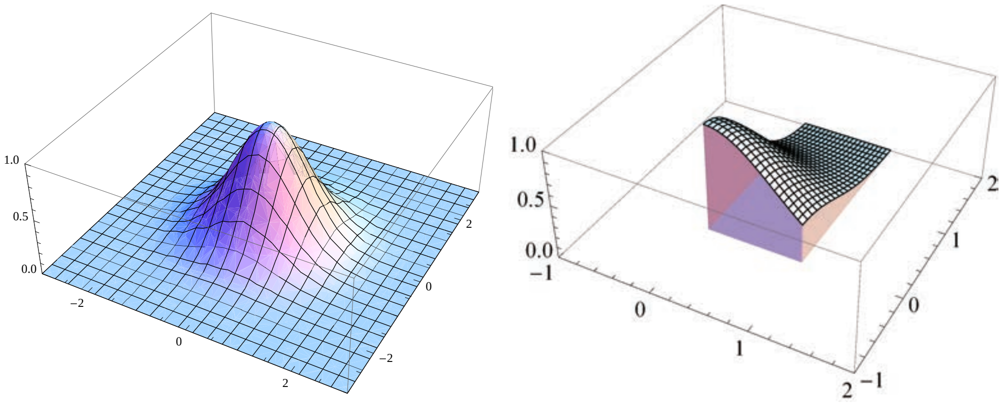


Figure 5.1: Left: a two-dimensional joint pdf of random variables X and Y. Right: the area under the pdf corresponds to $\mathbb{P}(0 \leqslant X \leqslant 1, Y \geqslant 0)$.

5.3 Independence of Random Variables

We have seen that in order to describe the behaviour of multiple random variables it is necessary to specify the joint distribution, not just the individual (that is, marginal) ones. However, there is one important exception, namely when the random variables are *independent*. We have so far only defined what independence is for *events* — see ☞ 42 (3.4). In the discrete case we define two random variables X and Y to be independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent for every choice of x and y; that is,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)\ .$$

This means that any information about what the outcome of X is does not provide any extra information about Y. For the pmfs this means that the joint pmf $f(x, y)$ is equal to the product of the marginal ones $f_X(x) f_Y(y)$. We can take this as the definition for independence, also for the continuous case, and when more than two random variables are involved.

Definition 5.4 Random variables $X_1, \ldots, X_n$ with joint pmf or pdf f are said to be **independent** if

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{5.4}$$

for all $x_1, \ldots, x_n$, where $\{f_{X_i}\}$ are the marginal pdfs.

Example 5.2 (Dice Experiment Continued) We repeat the experiment in Example 5.1 with three ordinary fair dice. Since the events $\{X = x\}$ and $\{Y = y\}$ are now independent, each entry in the pdf table is $\frac{1}{3} \times \frac{1}{6}$. Clearly in the first experiment not *all* events $\{X = x\}$ and $\{Y = y\}$ are independent.

Note

Many statistical models involve random variables $X_1, X_2, \ldots$ that are **independent and identically distributed**, abbreviated as **iid**. We will use this abbreviation throughout this book and write the corresponding model as

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathsf{Dist} \text{ (or } f \text{ or } F\text{)} ,$$

where Dist is the common distribution with pdf f and cdf F.

Example 5.3 Suppose X and Y are independent and both have a standard normal distribution. We say that (X, Y) has a bivariate standard normal distribution. What is the joint pdf? We have

$$f(x, y) = f_X(x) f_Y(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} = \frac{1}{2\pi} \mathrm{e}^{-\frac{1}{2}(x^2+y^2)} .$$

The graph of this joint pdf is the hat-shaped surface given in the left pane of Figure 5.1. We can also simulate independent copies $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(0, 1)$ and $Y_1, \ldots, Y_n \sim_{\text{iid}} \mathsf{N}(0, 1)$ and plot the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ to gain insight into the joint distribution. The following lines of R code produce the scatter plot of simulated data in Figure 5.2.

```
> x <- rnorm(2000)
> y <- rnorm(2000)
> plot(y~x,xlim = c(-3,3), ylim= c(-3,3))
```

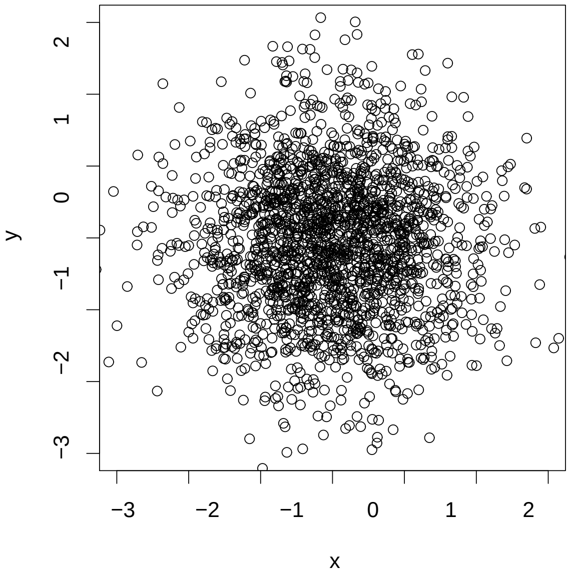


Figure 5.2: Scatter plot of 2000 points from the bivariate standard normal distribution.

5.4 Expectations for Joint Distributions

Similar to the univariate case in Theorem 3.6, the expected value of a real-valued ☞ 48 function h of $(X_1, \dots, X_n) \sim f$ is a weighted average of all values that $h(X_1, \dots, X_n)$ can take. Specifically, in the discrete case,

$$\mathbb{E}[h(X_1, \dots, X_n)] = \sum_{x_1,\dots,x_n} h(x_1, \dots, x_n)\, f(x_1, \dots, x_n), \tag{5.5}$$

where the sum is taken over all possible values of $(x_1, \dots, x_n)$. In the continuous case replace the sum above with a (multiple) integral.

Two important special cases are the expectation of the *sum* (or more generally any linear transformation plus a constant) of random variables and the *product* of random variables.

Theorem 5.1 (Properties of the Expectation). Let $X_1, \dots, X_n$ be random variables with expectations $\mu_1, \dots, \mu_n$. Then,

$$\mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] = a + b_1\mu_1 + \cdots + b_n\mu_n \tag{5.6}$$

for all constants a, $b_1, \dots, b_n$. Also, for *independent* random variables,

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mu_1\, \mu_2 \cdots \mu_n\ . \tag{5.7}$$

We show it for the discrete case with two variables only. The general case follows by analogy and, for the continuous case, by replacing sums with integrals. Let X_1 and X_2 be discrete random variables with joint pmf f. Then, by (5.5),

$$\begin{aligned}
\mathbb{E}[a + b_1 X_1 + b_2 X_2] &= \sum_{x_1,x_2} (a + b_1 x_1 + b_2 x_2)\, f(x_1, x_2) \\
&= a + b_1 \sum_{x_1}\sum_{x_2} x_1 f(x_1, x_2) + b_2 \sum_{x_1}\sum_{x_2} x_2 f(x_1, x_2) \\
&= a + b_1 \sum_{x_1} x_1 \left(\sum_{x_2} f(x_1, x_2) \right) + b_2 \sum_{x_2} x_2 \left(\sum_{x_1} f(x_1, x_2) \right) \\
&= a + b_1 \sum_{x_1} x_1 f_{X_1}(x_1) + b_2 \sum_{x_2} x_2 f_{X_2}(x_2) = a + b_1\mu_1 + b_2\mu_2\ .
\end{aligned}$$

Next, assume that X_1 and X_2 are independent, so that $f(x_1, x_2) = f_{X_1}(x_1)\, f_{X_2}(x_2)$. Then,

$$\begin{aligned}
\mathbb{E}[X_1\, X_2] &= \sum_{x_1,x_2} x_1\, x_2\, f_{X_1}(x_1) f_{X_2}(x_2) \\
&= \sum_{x_1} x_1 f_{X_1}(x_1) \times \sum_{x_2} x_2 f_{X_2}(x_2) = \mu_1\, \mu_2\ .
\end{aligned}$$

Definition 5.5 (Covariance). The **covariance** of two random variables X and Y with expectations $\mathbb{E}X = \mu_X$ and $\mathbb{E}Y = \mu_Y$ is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]\ .$$

The covariance is a measure of the amount of linear dependency between two random variables. A scaled version of the covariance is given by the **correlation coefficient**:

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y} \, , \tag{5.8}$$

where $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$.

For easy reference Theorem 5.2 lists some important properties of the variance and covariance.

Theorem 5.2 (Properties of the Variance and Covariance). For random variables X, Y and Z, and constants a and b, we have

1. $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.
2. $\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X)$.
3. $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$.
4. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
5. $\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$.
6. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
7. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.
8. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Proof. For simplicity of notation we write $\mathbb{E}Z = \mu_Z$ for a generic random variable Z. Properties 1 and 2 were already shown in Theorem 3.8. ☞ 49

3. $\mathrm{Cov}(X, Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[X\,Y - X\,\mu_Y - Y\,\mu_X + \mu_X\,\mu_Y] = \mathbb{E}[X\,Y] - \mu_X\,\mu_Y$.
4. $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[(Y - \mu_Y)(X - \mu_X)] = \mathrm{Cov}(Y, X)$.
5. $\mathrm{Cov}(aX + bY, Z) = \mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\,\mathbb{E}(Z) = a\,\mathbb{E}[XZ] - a\,\mathbb{E}(X)\mathbb{E}(Z) + b\,\mathbb{E}[YZ] - b\,\mathbb{E}(Y)\mathbb{E}(Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$.
6. $\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)] = \mathbb{E}[(X - \mu_X)^2] = \mathrm{Var}(X)$.
7. By Property 6, $\mathrm{Var}(X+Y) = \mathrm{Cov}(X+Y, X+Y)$. By Property 5, $\mathrm{Cov}(X+Y, X+Y) = \mathrm{Cov}(X, X) + \mathrm{Cov}(Y, Y) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, where in the last equation Properties 4 and 6 are used.
8. If X and Y are independent, then $\mathbb{E}[X\,Y] = \mu_X\,\mu_Y$. Therefore, $\mathrm{Cov}(X, Y) = 0$ follows immediately from Property 3.

In particular, combining Properties (7) and (8) we see that if X and Y are independent, then the variance of their sum is equal to the sum of their variances. It is not difficult to deduce from this (Problem 10) the following more general result.

Theorem 5.3 Let $X_1, \ldots, X_n$ be independent random variables with expectations $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$. Then,

$$\mathrm{Var}(a + b_1X_1 + b_2X_2 + \cdots + b_nX_n) = b_1^2\sigma_1^2 + \cdots + b_n^2\sigma_n^2 \tag{5.9}$$

for all constants $a, b_1, \ldots, b_n$.

☞ 56 **Example 5.4** We now show a simple way to prove Theorem 4.2; that is, to prove that the expectation and variance for the $\mathsf{Bin}(n, p)$ distribution are np and $np(1 - p)$, respectively. Let $X \sim \mathsf{Bin}(n, p)$. Hence, we can view X as the total number of successes in n Bernoulli trials (coin flips) with success probability p. Let us introduce Bernoulli random variables $X_1, \ldots, X_n$, where $X_i = 1$ is the ith trial is a success (and $X_i = 0$ otherwise). We thus have that $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{Ber}(p)$. The key to the proof is to observe that X is simply the sum of the $X_i' s$; that is

$$X = X_1 + \cdots + X_n \ .$$

Since we have seen that each Bernoulli variable has expectation p and variance $p(1-p)$, we have by Theorem 5.1 that

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np$$

and by Theorem 5.3 that

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = np(1 - p) \ ,$$

as had to be shown.

5.5 Limit Theorems

Two main results in probability are the *law of large numbers* and *the central limit theorem*. Both are limit theorems involving sums of independent random variables. In particular, consider a sequence $X_1, X_2, \ldots$ of iid random variables with finite expectation μ and finite variance σ^2. For each n define the sum $S_n = X_1 + \cdots + X_n$. What can we say about the (random) sequence of sums $S_1, S_2, \ldots$ or averages $S_1, S_2/2, S_3/3, \ldots$? By (5.6) and (5.9) we have $\mathbb{E}(S_n/n) = \mu$ and $\mathrm{Var}(S_n/n) = \sigma^2/n$. Hence, as n increases the variance of the (random) average S_n/n goes to 0. Informally, it means the following.

The average of a large number of iid random variables tends to their expectation as the sample size goes to infinity.

This is a nice property: if we wish to say something about the expectation of a random variable, we can simulate many independent copies and then take the average of these, to get a good approximation to the (perhaps unknown) expectation. The approximation will get better and better when the sample size gets larger.

Example 5.5 Let $U \sim \mathsf{N}(0, 1)$. What is the expectation of $\sqrt{U}$? We know that the expectation of U is $1/2$. Would the expectation of $\sqrt{U}$ be $\sqrt{1/2}$? We can determine in this case the expectation exactly (see Problem 5), but let us use simulation and the law of large numbers instead. All we have to do is simulate a large number of uniform numbers, take their square roots, and average over all values:

```
> u <- runif(10e6)
> x <- sqrt(u)
> mean(x)
```

```
[1] 0.6665185
```

Repeating the simulation gives consistently 0.666 in the first three digits behind the decimal point. You can check that the true expectation is 2/3, which is smaller than $\sqrt{1/2} \approx 0.7071$.

The central limit theorem describes the approximate distribution of S_n (or S_n/n), and it applies to both continuous and discrete random variables. Informally, it states the following.

> *The sum of a large number of iid random variables approximately has a normal distribution.*

Specifically, the random variable S_n has a distribution that is approximately normal, with expectation $n\mu$ and variance $n\sigma^2$. This is a truly remarkable result and is one of the great milestones in mathematics. We will not have enough background to prove it, but we can demonstrate it very nicely using simulation.

☞ 57 Let X_1 be a $\mathsf{U}[0, 1]$ random variable. Its pdf (see Section 4.4) is constant on the interval [0,1] and 0 elsewhere. If we simulate many independent copies of X_1 and take a histogram, the result will resemble the shape of the pdf (this, by the way, is a consequence of the law of large numbers). What about the pdf of $S_2 = X_1 + X_2$? We can generate many copies of both X_1 and X_2, add them up, and then make a histogram. Here is how you could do it in R and the result is given in Figure 5.3.

```
> x1 <- runif(10e6)
> x2 <- runif(10e6)
> hist(x1 +x2,breaks=100,prob=T)
```

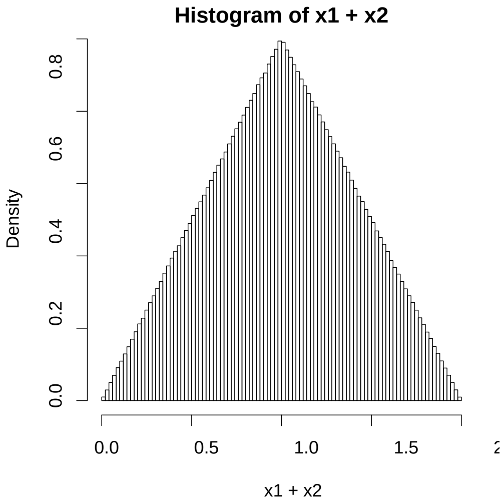


Figure 5.3: Histogram for the sum of 2 independent uniform random variables.

The pdf seems to be triangle shaped and, indeed, this is not so difficult to show; see Problem 3. Now let us do the same thing for sums of 3 and 4 uniform numbers. Figure 5.4 shows that the pdfs have assumed a bellshaped form reminiscent of the normal distribution. Indeed, if we superimpose the normal distribution with the same mean and variance as the sums, the agreement is excellent.

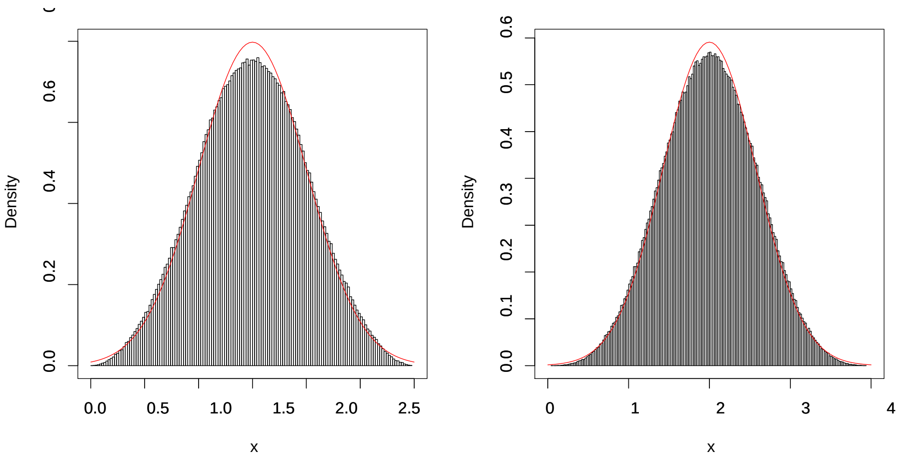


Figure 5.4: The histograms for the sums of 3 (left) and 4 (right) uniforms are in close agreement with normal pdfs.

The central limit theorem does not only hold if we add up continuous random variables, such as uniform ones, but it also holds for the discrete case. In particular, recall that a binomial random variable $X \sim \mathsf{Bin}(n, p)$ can be viewed as the sum of n iid $\mathsf{Ber}(p)$ random variables: $X = X_1 + \cdots + X_n$. As a direct consequence of the central limit theorem it follows that for large n, $\mathbb{P}(X \leqslant k) \approx \mathbb{P}(Y \leqslant k)$, where $Y \sim \mathsf{N}(np, np(1-p))$. As a rule of thumb, this normal approximation to the binomial distribution is accurate if both np and $n(1-p)$ are larger than 5.

Finally, when we add up independent *normal* random variables, then the resulting random variable has again a normal distribution. In fact any linear combination of independent normal random variables, such as $b_1X_1 + b_2X_2 + \cdots + b_nX_n$ can be shown to have again a normal distribution. This is quite an exceptional property, which makes the standard normal distribution stand out from most other distributions. The proof is outside the scope of a first-year course, but the central limit result should give you some confidence that it is true. And you can verify particular cases yourself via simulation; see Problem 9. Thus, the following theorem is one of the main reasons why the normal distribution is used so often in statistics.

Theorem 5.4 Let $X_1, X_2, \ldots, X_n$ be independent normal random variables with expectations $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$. Then, for any numbers $a, b_1, \ldots, b_n$ the random variable

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

has a normal distribution with expectation $a + \sum_{i=1}^n b_i\,\mu_i$ and variance $\sum_{i=1}^n b_i^2\,\sigma_i^2$.

Note that the expectation and variance of Y are a direct consequence of Theorems 5.1 and 5.3.

5.6 Statistical Modeling

Let us now return right to the beginning of these notes, to the steps for a statistical study in Section 1.1. Figure 5.5 gives a sketch of the conceptual framework for statistical ☞ 9 modeling and analysis. *Statistical modeling* refers to finding a plausible probabilistic model for the data. This model contains what we know about the reality and how the data were obtained. Once we have formulated the model, we can carry out our calculations and analysis and make conclusions.

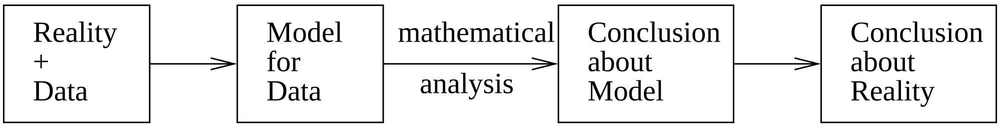


Figure 5.5: Statistical modeling and analysis.

The simplest class of statistical models is the one where the data $X_1, \ldots, X_n$ are assumed to be independent and identically distributed (iid), as we already mentioned. In many cases it is assumed that the sampling distribution is normal. Here is an ☞ 71 example.

Example 5.6 (One-sample normal model) From a large population we select 300 men between 40 and 50 years of age and measure their heights. Let X_i be the height of the i-th selected person, $i = 1, \ldots, 300$. As a model take,

$$X_1, \ldots, X_{300} \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$$

for some unknown parameters μ and σ^2. We could interpret these as the population mean and variance.

A simple generalisation of a single sample of iid data is the model where there are two independent samples of iid data, as in the examples below.

Example 5.7 (Two-sample Binomial Model) To assess whether there is a difference between boys and girls in their preference for two brands of cola, say *Sweet* and *Ultra* cola, we select at random 100 boys and 100 girls and ask whether they prefer *Sweet* or *Ultra*. We could model this via two independent Bernoulli samples. That is, for each $i = 1, \ldots, 100$ let $X_i = 1$ if the i-th boy prefers *Sweet* and let $X_i = 0$ otherwise. Similarly, let $Y_i = 1$ if the i-th girl prefers *Sweet* over *Ultra*. We thus have the model

$$\begin{aligned} X_1, \ldots, X_{100} &\overset{\text{iid}}{\sim} \mathsf{Ber}(p_1)\ , \\ Y_1, \ldots, Y_{100} &\overset{\text{iid}}{\sim} \mathsf{Ber}(p_2)\ , \\ X_1, \ldots, X_{100}, Y_1, \ldots, Y_{100} &\ \ \text{independent, with } p_1 \text{ and } p_2 \text{ unknown.} \end{aligned}$$

The objective is to assess the difference $p_1 - p_2$ on the basis of the observed values for $X_1, \ldots, X_{100}, Y_1, \ldots, Y_{100}$. Note that it suffices to only record the total number of boys or girls who prefer *Sweet* cola in each group; that is, $X = \sum_{i=1}^{100} X_i$ and $Y = \sum_{i=1}^{100} Y_i$.

This gives the **two-sample binomial model**:

$$\begin{aligned} X &\sim \mathsf{Bin}(100, p_1)\ , \\ Y &\sim \mathsf{Bin}(100, p_2)\ , \\ X, Y &\ \ \text{independent, with } p_1 \text{ and } p_2 \text{ unknown.} \end{aligned}$$

Example 5.8 (Two-sample Normal Model) From a large population we select 200 men between 25 and 30 years of age and measure their heights. For each person we also record whether the mother smoked during pregnancy or not. Suppose that 60 mothers smoked during pregnancy.

Let $X_1, \ldots, X_{60}$ be the heights of the men whose mothers smoked, and let $Y_1, \ldots, Y_{140}$ be the heights of the men whose mothers did not smoke. Then, a possible model is the **two-sample normal model**:

$$\begin{aligned} X_1, \ldots, X_{60} &\overset{\text{iid}}{\sim} \mathsf{N}(\mu_1, \sigma_1^2)\ , \\ Y_1, \ldots, Y_{140} &\overset{\text{iid}}{\sim} \mathsf{N}(\mu_2, \sigma_2^2)\ , \\ X_1, \ldots, X_{60}, Y_1, \ldots, Y_{140} &\quad \text{independent,} \end{aligned}$$

where the model parameters μ_1, μ_2, σ_1^2, and σ_2^2 are unknown. One would typically like to assess the difference $\mu_1 - \mu_2$. That is, does smoking during pregnancy affect the (expected) height of the sons? A typical simulation outcome of the model is given in Figure 5.6, using parameters $\mu_1 = 175, \mu_2 = 170, \sigma_1^2 = 200$, and $\sigma_2^2 = 100$.

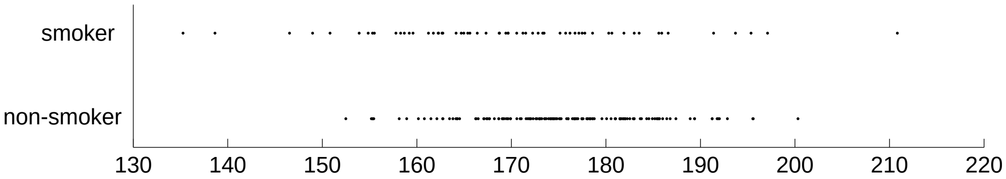


Figure 5.6: Simulated height data from a two-sample normal model.

Remark 5.1 (About Statistical Modeling) At this point it is good to emphasise a few points about statistical modeling.

- *Any* model for data is likely to be *wrong*. For example, in Example 5.8 the height would normally be recorded on a discrete scale, say 1000 – 2200 (mm). However, samples from a $\mathsf{N}(\mu, \sigma^2)$ can take any real value, including negative values! Nevertheless, the normal distribution could be a reasonable approximation to the real sampling distribution. An important advantage of using a normal distribution is that it has many nice mathematical properties as we have seen.

- Most statistical models depend on a number of *unknown* parameters. One of the main objectives of *statistical inference* — to be discussed in subsequent chapters — is to gain knowledge of the unknown parameters on the basis of the observed data.

- Any model for data needs to be checked for suitability. An important criterion is that data simulated from the model should resemble the observed data — at least for a certain choice of model parameters.

Conclusion

- Multiple random variables are specified by their joint cdf or pmf/pdf.

- The marginal pmf (pdf) is obtained from the joint pmf by summing (integrating) over all possible values of the other variables.

- For independent random variables the joint pmf/pdf is the product of the marginal ones.

- The expectation of the sum of random variables is equal to the sum of their expectations.

- The variance of the sum of *independent* random variables is equal to the sum of their variances.

- The expectation of the product of *independent* random variables is equal to the product of their expectations.

- The covariance and correlation coefficient are measures for the amount of linear dependence between two random variables.

- The law of large numbers implies that the average of a large number of iid random variables is approximately equal to their expectation.
- The central limit theorem says that the sum of a large number of iid random variables is approximately normal.
- Any linear combination of independent normal random variables is again a normal random variable.
- Many statistical models are formulated as a single or multiple iid samples.

5.7 Problems

1. The joint pmf of X and Y is given by the table

	y			
x	1	3	6	8
2	0	0.1	0.1	0
5	0.2	0	0	0
6	0	0.2	0.1	0.3

   (a) Determine the (marginal) pmf of X and of Y.

   (b) Are X and Y independent?

   (c) Calculate $\mathbb{E}[X^2 Y]$.

2. Explain how *in principle* we could calculate

$$\mathbb{P}(X_1 + X_2 > 1) \ ,$$

   if we knew the joint pdf of X_1 and X_2.

3. Suppose that X and X are independent and uniformly distributed on the interval [0,1]. Prove that their sum $X + Y$ has a triangular pdf, of the form in Figure 5.2.

4. Let $X_1, \dots, X_6$ be the masses of 6 people, selected from a large population. Suppose the masses have a normal distribution with a mean of 75 kg and a standard deviation of 10 kg. What do $Y_1 = 6X_1$ and $Y_2 = X_1 + \cdots + X_6$ represent, physically? Explain why Y_1 and Y_2 have different distributions. Which one has the smallest variance?

5. Let $U \sim \mathsf{U}[0, 1]$. Calculate $\mathbb{E}[\sqrt{U}] = \int_0^1 u^{\frac{1}{2}}\,\mathrm{d}u$.

6. Let $X \sim \mathsf{Bin}(100, 1/4)$. Approximate, using the central limit theorem, the probability $\mathbb{P}(20 \leqslant X \leqslant 30)$. Compare this with the exact probability. Use the `R` functions `pnorm()` and `pbinom()`.

7. A lift can carry a maximum of 650 kg. Suppose that the weight of a person is normally distributed with expectation 75 kg and standard deviation 10 kg. Let Z_n be the total weight of n randomly selected persons.

   (a) Determine the probability that $Z_8 \geqslant 650$.

   (b) Determine n such that $\mathbb{P}(Z_n \geqslant 650) \leqslant 0.01$.

8. Consider the following game: You flip 10 fair coins, all at once, and count how many Heads you have. I'll pay you out the squared number of Heads, in dollars. However, you will need to pay me some money in advance. How much would you be prepared to give me if you could play this game as many times as you'd like?

9. Let $X \sim \mathsf{N}(1, 2)$ and $Y \sim \mathsf{N}(3, 1)$ be independent. What is the distribution of $Z = X + Y$? Simulate many copies of Z and verify visually that the corresponding histogram is in close agreement with the true pdf.

10. Derive Theorem 5.3 from the rules in Theorem 5.2 .

11. A quick and dirty way to simulate a standard normal random variable is to take the sum of 6 independent $\mathsf{U}[0, 1]$ random variables, subtract 3, and multiply by $\sqrt{2}$. Explain why this is a sensible approach.

12. Formulate a statistical model for each of the situations below, in terms of one or more iid samples. If a model has more than one parameter, specify which parameter is of primary interest.

    (a) A ship builder buys each week hundreds of tins of paint, labeled as containing 20 liters. The builder suspects that the tins contain, on average, less than 20 liters, and decides to determine the volume of paint in nine randomly chosen tins.

    (b) An electronics company wishes to examine if the rate of productivity differs significantly between male and female employees involved in assembly work. The time of completion of a certain component is observed for 12 men and 12 women.

    (c) The head of a mathematics department suspects that lecturers A and B differ significantly in the way they assess student work. To test this, 12 exams are both assessed by lecturer A and B.

    (d) We wish to investigate if a certain coin is fair. We toss the coin 500 times and examine the results.

    (e) We investigate the effectiveness of a new teaching method, by dividing 20 students into two groups of 10, where the first group is taught by the old method, and the second group is taught by the new method. Each student is asked to complete an exam before and after the teaching period.

    (f) We wish to assess which of two scales is the more sensitive. We measure, for each scale, ten times a standard weight of 1kg.

(g) To investigate if the support for the *Honest* party is the same in two different cities, one hundred voters in each city are asked if they would vote for the *Honest* party or not.

(h) In a study on the effectiveness of an advertising campaign, a survey was conducted among 15 retail outlets. For each outlet the sales on a typical Saturday was recorded one month before and one month after the advertising campaign.

(i) To focus their marketing of remote controlled cars an electronics company wishes to investigate who in the end decides to buy: the child or the father. It records who decides in 400 transactions involving a father and a son.

Chapter 6

Estimation

In this chapter you will learn how to estimate parameters of simple statistical models from the observed data. The difference between estimate and estimator will be explained. Confidence intervals will be introduced to assess the accuracy of an estimate. We will derive confidence interval for a variety one- and two-sample models. Various probability distributions, such as the Student's t distributions, the F distribution, and the χ^2 distribution will make their first appearance.

6.1 Introduction

Recall the framework of statistical modeling in Figure 5.5. We are given some data (measurements) for which we construct a *model* that depends on one or more parameters. Based on the observed data we try to say something about the model parameters. For example, we wish to *estimate* the parameters. Here are some concrete examples. ☞ 77

Example 6.1 (Biased Coin) We throw a coin 1000 times and observe 570 Heads. Using this information, what can we say about the "fairness" of the coin? The data (or better, *datum*, as there is only one observation) here is the number $x = 570$. Suppose we view x as the outcome of a random variable X which describes the number of Heads in 1000 tosses. Our statistical model is then:

$$X \sim \mathsf{Bin}(1000, p) ,$$

where $p \in [0, 1]$ is unknown. Any statement about the fairness of the coin is expressed in terms of p and is assessed via this model. It is important to understand that p will *never be known*. The best we can do is to provide an *estimate* of p. A common sense estimate of p is simply the proportion of Heads $x/1000 = 0.570$. But how accurate is this estimate? Is it possible that the unknown p could in fact be 0.5? One can make sense of these questions through detailed analysis of the statistical model.

Example 6.2 (Iid Sample from a Normal Distribution) Consider the standard model for data

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2) ,$$

where μ and σ^2 are unknown. The random measurements $\{X_i\}$ could represent the masses of randomly selected teenagers, the heights of the dorsal fin of sharks, the dioxin concentrations in hamburgers, and so on. Suppose, for example that, with $n = 10$, the observed measurements $x_1, \ldots, x_n$ are:

77.01, 71.37, 77.15, 79.89, 76.46, 78.10, 77.18, 74.08, 75.88, 72.63.

A common-sense *estimate* (a number) for μ is the **sample mean**

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = 75.975 \ , \tag{6.1}$$

and σ^2 can be estimated via the **sample variance**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ . \tag{6.2}$$

Note that the estimates $\bar{x}$ and s^2 are functions of the data $\mathbf{x} = (x_1, \ldots, x_n)$ only. We ☞ 19 encountered these summary statistics already in Section 2.4.

Why are these numbers good estimates (guesses) for our unknown parameters μ and σ^2. How accurate are these numbers? That is, how far away are they from the true parameters? To answer these questions we need to investigate the statistical properties of the sample mean and sample variance.

Note

It is customary in statistics to denote the estimate of a parameter $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$; for example, $\widehat{\mu} = \bar{x}$ in the example above.

6.2 Estimates and estimators

If we have some data coming from some statistical model, how do we estimate the parameters? There are various systematic ways to construct sensible estimates for parameters of various models. Suppose we have n independent copies $X_1, \ldots, X_n$ of a random variable X whose distribution depends on p parameters (for example, $X \sim \mathsf{N}(\mu, \sigma^2)$, with $p = 2$ parameters). A useful general approach to estimate the parameters is the **method of moments**. Recall that the ***k*-th moment** of a random variable X is defined ☞ 48 as $\mathbb{E}(X^k)$; see Definition 3.10. For example, the expectation is the first moment. In the method of moments the estimated parameters are chosen such that the first p true moments $\mathbb{E}(X^k)$ are matched to their sample averages $\sum_{i=1}^{n} x_i^k / n$.

Example 6.3 Let $X_1, \ldots, X_n$ be an iid copies of $X \sim \mathsf{N}(\mu, \sigma^2)$. The first moment of each X is $\mathbb{E}(X) = \mu$, and the second moment of X is $\mathbb{E}(X^2) = \mathrm{Var}(X) + [\mathbb{E}(X)]^2 = \sigma^2 + \mu^2$.

To find the method of moments estimates for μ and σ^2, let us call them $\widehat{\mu}$ and $\widehat{\sigma^2}$, we need to match the first two moments to their sample averages. That is, we need to solve

$$\begin{aligned}\widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \widehat{\mu}^2 + \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \,.\end{aligned}$$

The first equation gives the sample mean $\widehat{\mu} = \bar{x}$ as our estimate for μ. Substituting $\widehat{\mu} = \bar{x}$ in the second equation, we find that the second equation gives

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \tag{6.3}$$

as an estimate for σ^2. This estimate seems quite different from the sample variance s^2 in (6.2). But the two estimates are actually very similar. To see this, expand the quadratic term in (6.2), to get

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \,.$$

Now break up the sum:

$$\begin{aligned}s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\bar{x}x_i + \sum_{i=1}^{n} \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1\right)\end{aligned}$$

and simplify

$$\begin{aligned}s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).\end{aligned}$$

Comparing this with (6.3), we see that $s^2 = n/(n-1)\,\widehat{\sigma^2}$, so they differ only in a factor $n/(n-1)$. For large n they are practically the same.

To find out how *good* an estimate is, we need to investigate the properties of the corresponding **estimator**. The estimator is obtained by replacing the fixed observations x_i with the random variables X_i in the expression for the estimate. For example, the estimator corresponding to the sample mean $\bar{x}$ is

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \,.$$

The interpretation is that $X_1, \ldots, X_n$ are the data that we will obtain if we carry out the experiment *tomorrow*, and $\bar{X}$ is the (random) sample mean of these data, which again will be obtained tomorrow.

Let us go back to the basic model were $X_1, \dots, X_n$ are independent and identically distributed with some unknown expectation μ and variance σ^2. We do not require that the $\{X_i\}$ are normally distributed — we are only interested in estimating the expectation and variance.

To justify why $\bar{x}$ is a good estimate of μ, think about what we can say (today) about the properties of the estimator $\bar{X}$. The expectation and variance of $\bar{X}$ follow easily from ☞ 72 the rules for expectation and variance in Chapter 5. In particular, by (5.6) we have

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n}\mathbb{E}\,(X_1 + \cdots + X_n) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) \\ &= \frac{1}{n}(\mu + \cdots + \mu) = \mu\end{aligned}$$

☞ 74 and from (5.9) we have

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2}\,(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) \\ &= \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}\ .\end{aligned}$$

The first result says that the estimator $\bar{X}$ is "on average" equal to the unknown quantity that we wish to estimate (μ). We call an estimator whose expectation is equal to the quantity that we wish to estimate **unbiased**. The second result shows that the larger we take n, the closer the variance of $\bar{X}$ is to zero, indicating that $\bar{X}$ goes to the constant ☞ 74 μ for large n. This is basically the law of large numbers; see Section 5.5.

To assess exactly how close $\bar{X}$ is to μ one needs to look at a confidence interval for μ.

6.3 Confidence Intervals

An essential part in any estimation procedure is to provide an assessment of the *accuracy* of the estimate. Indeed, without information on its accuracy the estimate itself would be meaningless. Confidence intervals (sometimes called **interval estimates**) provide a precise way of describing the uncertainty in the estimate.

Definition 6.1 Let $X_1, \dots, X_n$ be random variables with a joint distribution depending on a parameter θ. Let $T_1 < T_2$ be functions of the data $X_1, \dots, X_n$ but not of θ. A random interval (T_1, T_2) is called a **stochastic confidence interval** for θ with confidence $1 - \alpha$ if

$$\mathbb{P}(T_1 < \theta < T_2) \geqslant 1 - \alpha \quad \text{for all } \theta\ . \tag{6.4}$$

If t_1 and t_2 are the observed values of T_1 and T_2, then the interval (t_1, t_2) is called the **numerical confidence interval** for θ with confidence $1-\alpha$. If (6.4) only holds approximately, the interval is called an **approximate confidence interval**.

The actual *meaning* of a confidence interval is quite tricky. Suppose we find a 90% numerical confidence interval (9.5,10.5) for θ. Does this mean that $\mathbb{P}(9.5 < \theta < 10.5)$? No! Since θ is a fixed number the probability $\mathbb{P}(9.5 < \theta < 10.5)$ is either 0 or 1, and we don't know which one, because we don't know θ. To find the meaning we have to go back to the definition of a confidence interval. There we see that the interval (9.5,10.5) is an *outcome* of a *stochastic* (i.e., random) confidence interval (T_1, T_2), such that $\mathbb{P}(T_1 < \theta < T_2) = 0.9$. Note that θ is constant, but the interval bounds T_1 and T_2 are random. If we would repeat this experiment many times, then we would get many numerical confidence intervals, as illustrated in Figure 6.1

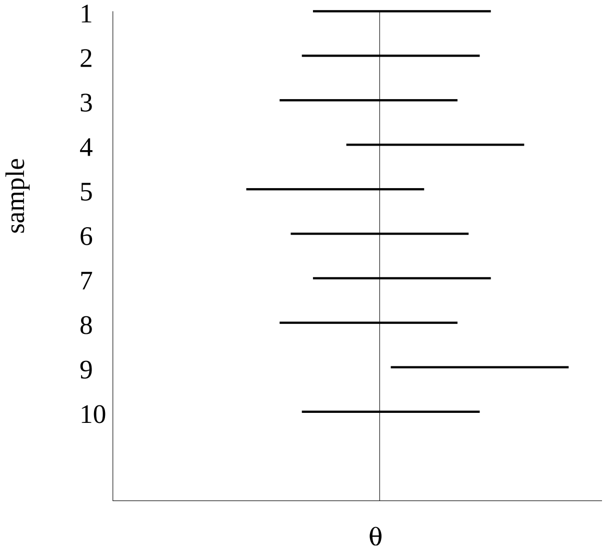


Figure 6.1: Possible outcomes of a stochastic confidence intervals.

Only in (on average) 9 out of 10 cases would these intervals contain our unknown θ. To put it in another way: Consider an urn with 90 white and 10 black balls. We pick at random a ball from the urn *but we do not open our hand to see what colour ball we have*. Then we are pretty confident that the ball we have in our hand is white. This is how confident you should be that the unknown θ lies in the interval (9.5, 10.5).

Note

Reducing α widens the confidence interval. A very large confidence interval is not very useful. Common choices for α are $0.01, 0.05$, and 0.1.

6.3.1 Approximate Confidence Interval for the Mean

Let $X_1, X_2, \ldots, X_n$ be an iid sample from a distribution with mean μ and variance $\sigma^2 < \infty$ (both assumed to be unknown). We assume that the sample size n is large. By the central limit theorem we know then that $X_1 + \cdots + X_n$ has approximately a normal distribution, so $\bar{X}$ also has approximately a normal distribution. We found the corresponding expectation and variance above, so ☞ 75

$$\bar{X} \overset{\text{approx.}}{\sim} \mathsf{N}(\mu, \sigma^2/n)\ .$$

Standardising $\bar{X}$ gives

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \overset{\text{approx.}}{\sim} \mathsf{N}(0,1)\ .$$

In order to construct a confidence interval for μ, we would like to create a so-called **pivot** variable that depends on all the data and on the parameter to be estimated, but nothing else. The above standardised form of $\bar{X}$ is not a pivot yet because it depends on σ^2. However, we can fix this by replacing σ^2 with its unbiased estimator S^2. By the law of large numbers S^2 looks more an more like the constant σ^2 as n grows larger. So, we have for large n

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \overset{\text{approx.}}{\sim} \mathsf{N}(0,1)\ , \tag{6.5}$$

where $S = \sqrt{S^2}$ is the sample standard deviation. Because T is approximately standard normal, we have, for example,

$$\mathbb{P}(T \leqslant 1.645) \approx 0.95 \quad \text{and} \quad \mathbb{P}(T \leqslant 1.96) \approx 0.975$$

because 1.645 is the 0.95 quantile of the normal distribution and 1.96 the 0.975 quantile, both of which are good to remember. See also Section 4.5. Because the standard normal distribution is symmetrical around 0, we also have, for example,

☞ 58

$$\mathbb{P}(-1.96 < T < 1.96) \approx 0.95\ .$$

Now, let us have a closer look at this, and plug back in the expression for the pivot T, so

$$\mathbb{P}\left(-1.96 < \frac{\bar{X}-\mu}{S/\sqrt{n}} < 1.96\right) \approx 0.95\ .$$

We can rearrange the event

$$A = \left\{-1.96 < \frac{\bar{X}-\mu}{S/\sqrt{n}} < 1.96)\right\}$$

as follows. Multiplying the left, middle, and right parts of the inequalities by $S/\sqrt{n}$ still gives the same event, so

$$A = \left\{-1.96\frac{S}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{S}{\sqrt{n}}\right\}.$$

Subtracting $\bar{X}$ from left, middle, and right parts still does not change anything about the event, so

$$A = \left\{-\bar{X} - 1.96\frac{S}{\sqrt{n}} < -\mu < -\bar{X} + 1.96\frac{S}{\sqrt{n}}\right\}.$$

Finally we multiply the left, middle, and right parts with -1. This will flip the $<$ signs to $>$. For example, $-3 < -2$ is the same as $3 > 2$. So we get,

$$A = \left\{\bar{X} + 1.96\frac{S}{\sqrt{n}} > \mu > \bar{X} - 1.96\frac{S}{\sqrt{n}}\right\},$$

which is the same as

$$A = \left\{ \bar{X} - 1.96\frac{S}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{S}{\sqrt{n}} \right\} .$$

If we write this as $A = \{T_1 < \mu < T_2\}$, with $\mathbb{P}(A) \approx 0.95$, then we see that (T_1, T_2) is an approximate 95% confidence interval for μ. We can repeat this procedure with any quantile of the normal distribution. This leads to the following result.

Theorem 6.1 Let $X_1, X_2, \ldots, X_n$ be an iid sample from a distribution with mean μ and variance $\sigma^2 < \infty$. Let q be the $1 - \alpha/2$ quantile of the standard normal distribution. An approximate stochastic confidence interval for μ is

$$\left(\bar{X} - q\frac{S}{\sqrt{n}},\ \bar{X} + q\frac{S}{\sqrt{n}} \right), \text{ abbreviated as } \bar{X} \pm q\frac{S}{\sqrt{n}} . \tag{6.6}$$

Since (6.6) is an asymptotic result only, care should be taken when applying it to cases where the sample size is small or moderate and the sampling distribution is heavily skewed.

Example 6.4 An oil company wishes to investigate how much on average each household in Melbourne spends on petrol and heating oil per year. The company randomly selects 51 households from Melbourne, and finds that these spent on average \$1136 on petrol and heating oil, with a sample standard deviation of \$178. We wish to construct a 95% confidence interval for the expected amount of money per year that the households in Melbourne spend on petrol and heating oil. Call this parameter μ.

We assume that the outcomes of the survey, $x_1, \ldots, x_{51}$, are realisations of an iid sample with expectation μ. Although we do not know the outcomes themselves, we know their sample mean $\bar{x} = 1136$ and standard deviation $s = 178$. An approximate numerical 95% confidence interval is thus

$$1136 \pm 1.96\frac{178}{\sqrt{51}} = (1087, 1185) .$$

6.3.2 Normal data, one sample

For an iid sample from the normal distribution, $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$ it is possible to construct *exact* confidence intervals for μ and σ^2, rather than only approximate.

Confidence interval for μ

For iid $\mathsf{N}(\mu, \sigma^2)$ data, the pivot variable T in (6.5) can be shown to have a **Student's t-distribution**. This distribution is named after its discoverer W.S. Gosset, who published under the pseudonym "Student". The t distribution is actually a family of distributions, depending on a single parameter called the (number of) **degrees of freedom**. We write $Z \sim \mathsf{t}_n$ to indicate that a random variable Z has a student distribution with n degrees of freedom. Here is the exact version of the approximate result (6.5).

Theorem 6.2 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathsf{t}_{n-1} \,. \tag{6.7}$$

Figure 6.2 gives graphs of the probability densities functions for the t_1, t_2, t_5, and t_{50}. Notice a similar bell-shaped curve as the normal distribution, but the tails of the distribution are a bit fatter than for the normal distribution. As n grows larger the pdf of the t_n gets closer and closer to the pdf of the $\mathsf{N}(0, 1)$ distribution.

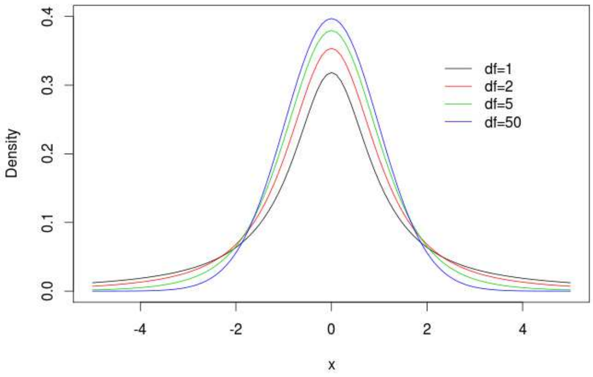


Figure 6.2: The pdfs of Student t distributions with various degrees of freedom (df).

We can use `R` to calculate the pdf, cdf, and quantiles for this distribution. For example, the following `R` script generates Figure 6.2.

```
curve(dt(x,df=1),ylim=c(0,0.4),xlim=c(-5,5),col=1,ylab="Density")
curve(dt(x,df=2),col=2,add=TRUE)
curve(dt(x,df=5),col=3,add=TRUE)
curve(dt(x,df=50),col=4,add=TRUE)
legend(2.1,0.35,lty=1,bty="n",
                  legend=c("df=1","df=2","df=5","df=50"),col=1:4)
```

To obtain the 0.975 quantile of the t_n distribution for $n = 1, 2, 5, 50$, and 100, enter the following commands.

```
> qt(0.975,df=c(1,2,5,50,100))
[1] 12.706205  4.302653  2.570582  2.008559  1.983972
```

As a comparison, the 0.975 quantile for the standard normal distribution is `qnorm(0.975)` = 1.959964 (say, 1.96).

Returning to the pivot T in (6.5), it has a t_{n-1} distribution. Simply repeating the arguments from Section 6.3.1 we find the following exact confidence interval for μ in terms of the quantiles of the t_{n-1} distribution.

Theorem 6.3 Let $X_1, X_2, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$ and let q be the $1 - \alpha/2$ quantile of the Student's t_{n-1} distribution. An exact stochastic confidence interval for μ is

$$\bar{X} \pm q \frac{S}{\sqrt{n}} . \tag{6.8}$$

Example 6.5 A buret is a glass tube with scales that can be used to add a specified volume of a fluid to a receiving vessel. We wish to determine a 95% confidence interval for the average volume of *one* drop of water that leaves the buret, based on the data in Table 6.1.

Table 6.1: An experiment with a buret

Volume in buret (ml)	
initial	25.36
after 50 drops	22.84
after 100 drops	20.36

Our model for the data is as follows: let X_1 be the volume of the first 50 drops, and X_2 the volume of the second 50 drops. We assume that X_1, X_2 are iid and $\mathsf{N}(\mu, \sigma^2)$ distributed, with unknown μ and σ^2. Note that μ is the expected volume of 50 drops, and therefore $\mu/50$ is the expected volume of one drop.

With $n = 2$ and $\alpha = 0.05$, we have that the 0.975 quantile of the t_1 distribution is $q = 12.71$. The outcomes of X_1 and X_2 are respectively $x_1 = 2.52$ and $x_2 = 2.48$. Hence,

$$s = \sqrt{(2.52 - 2.50)^2 + (2.48 - 2.50)^2} = 0.02\sqrt{2} .$$

Hence, a numerical 95% CI for μ is

$$2.50 \pm 12.71 \times 0.02 = (2.25, 2.75) .$$

However, we want a 95% CI for $\mu/50$! We leave it as an exercise (see Problem 2) to show that we can simply divide the 95% CI for μ by 50 to obtain a 95% CI for $\mu/50$. Thus, a 95% (numerical) confidence interval for the average volume of one drop of water is

$$(0.045, 0.055) \quad (\text{ml}) .$$

Confidence interval for σ^2

Next, we construct a confidence interval for σ^2. As before let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$. Consider the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

It turns out that $(n-1)S^2/\sigma^2$ has a known distribution, called the χ^2 **distribution**, where χ is the Greek letter *chi*. Hence, the distribution is also written (and pronounced) as the chi-squared distribution. Like the t distribution, the χ^2 distribution is actually a family of distributions, depending on a parameter that is again called the *degrees of freedom*. We write $Z \sim \chi^2_n$ to denote that Z has a chi-square distribution with n degrees of freedom. Figure 6.3 shows the pdf of the $\chi^2_1, \chi^2_2, \chi^2_5$, and χ^2_{10} distributions. Note that the pdf is not symmetric and starts at $x = 0$. The χ^2_1 has a density that is infinite at 0, but that is no problem — as long as the total integral under the curve is 1.

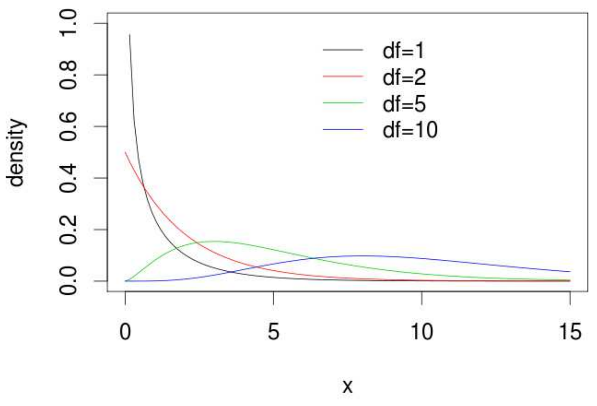


Figure 6.3: The pdfs of chi-square distributions with various degrees of freedom (df).

Figure 6.3 was made in a very similar way to Figure 6.2, mostly by replacing `dt` with `dchisq` in the `R` code. Here is the beginning of the script — you can work out the rest.

```
> curve(dchisq(x,df=1),xlim=c(0,15),ylim=c(0,1),ylab="density")
```

To obtain the 0.025 and 0.975 quantiles of the χ^2_{24} distribution, for example, we can issue the command.

```
> qchisq(p=c(0.025,0.975),24)
```

```
[1] 12.40115 39.36408
```

Returning to the pivot variable $(n-1)S^2/\sigma^2$, this can be shown to have a χ^2_{n-1} distribution. For future reference we list this as a theorem.

Theorem 6.4 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \,. \tag{6.9}$$

Hence, if we denote the $\alpha/2$ and $1-\alpha/2$ quantiles of this distribution by q_1 and q_2, then

$$\mathbb{P}\left(q_1 < \frac{(n-1)}{\sigma^2} S^2 < q_2\right) = 1 - \alpha \; .$$

Rearranging this shows

$$\mathbb{P}\left(\frac{(n-1)S^2}{q_2} < \sigma^2 < \frac{(n-1)S^2}{q_1}\right) = 1 - \alpha \; .$$

This gives the following exact confidence interval for σ^2 in terms of the quantiles of the χ^2_{n-1} distribution.

Theorem 6.5 Let $X_1, X_2, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$ and let q_1 and q_2 be the $\alpha/2$ and $1-\alpha/2$ quantiles of the χ^2_{n-1} distribution. An exact stochastic confidence interval for μ is

$$\left(\frac{(n-1)S^2}{q_2}, \frac{(n-1)S^2}{q_1}\right) \; . \tag{6.10}$$

Example 6.6 On the label of a certain packet of aspirin it is written that the standard deviation of the tablet weight (actually mass) is 1.0 mg. To investigate if this is true we take a sample of 25 tablets and discover that the sample standard deviation is 1.3mg. A 95% numerical confidence interval for σ^2 is

$$\left(\frac{24 \times 1.3^2}{39.4}, \frac{24 \times 1.3^2}{12.4}\right) = (1.04, 3.27) \; ,$$

where we have used (in rounded numbers) $q_1 = 12.4$ and $q_2 = 39.4$ calculated before with the `qchisq()` function. A 95% numerical confidence interval for σ is found by taking square roots (why?):

$$(1.02, 1.81) \; .$$

Note that this CI does not contain the asserted weight of 1.0 mg. We therefore have some doubt whether the "true" standard deviation is indeed equal to 1.0 mg.

6.3.3 Normal data, two samples

Consider now *two* independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, X_n$ from respectively a $\mathsf{N}(\mu_X, \sigma_X^2)$ and $\mathsf{N}(\mu_Y, \sigma_Y^2)$ distribution. We wish to make confidence intervals for $\mu_X - \mu_Y$ and σ_X^2/σ_Y^2. The difference $\mu_X - \mu_Y$ tells us how the two *means* relate to each other, and σ_X^2/σ_Y^2 gives an indication how the *variances* relate to each other.

Confidence interval for $\mu_X - \mu_Y$

Constructing an exact confidence interval for $\mu_X - \mu_Y$ is very similar to the 1-sample case *provided* that we assume the **extra model assumption** that *the variances of the two samples are the same.* That is, we assume that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, for some unknown

σ^2. The analysis now proceeds as follows. The obvious estimator for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$. Next, observe that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\,\sqrt{1/m + 1/n}} \sim \mathsf{N}(0, 1)\ .$$

However, if σ^2 is unknown, we should replace it with an appropriate estimator. For this we will use the **pooled sample variance**, S_p^2, which is defined as

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}\ , \tag{6.11}$$

where S_X^2 and S_Y^2 are the sample variances for the X_i's and Y_i's, respectively. It is not difficult to show that S_p^2 is an unbiased estimator of σ^2. Similar to Theorem 6.7 we have that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\,\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathsf{t}_{m+n-2}\ . \tag{6.12}$$

Using this pivot variable, we find (completely analogous to the 1-sample case) that

$$\bar{X} - \bar{Y} \pm q\,S_p\,\sqrt{\frac{1}{m} + \frac{1}{n}}$$

is a 100%$(1-\alpha)$ stochastic confidence interval for $\mu_X - \mu_Y$, where q is the $1 - \alpha/2$ quantile of the t_{m+n-2} distribution.

Example 6.7 A human movement student has a theory that the expected mass of 3rd year students differs from that of 1st years. To investigate this theory, random samples are taken from each of the two groups. A sample of 15 1st years has a mean of 62.0kg and a standard deviation of 15kg, while a sample of 10 3rd years has a mean of 71.5kg and a standard deviation of 12kg. Does this show that the expected masses are indeed different?

Here we have $m = 15$ and $n = 10$. The outcomes for $\bar{X} - \bar{Y}$ and S_p are respectively $62 - 71.5 = -9.5$ and

$$s_p = \sqrt{\frac{14 \times 15^2 + 9 \times 12^2}{23}} = 13.90339\ .$$

To construct a 95% numerical confidence interval for $\mu_X - \mu_Y$ we need to also evaluate the factor $\sqrt{1/m + 1/n}$ (here 0.4082483) and the 0.975 quantile of the t_{23} distribution (here 2.068658), using the `R` command `qt(0.975,23)`. So that the 95% numerical confidence interval for $\mu_X - \mu_Y$ is given by

$$-9.5 \pm 2.068658 \times 13.90339 \times 0.4082483 = (-21.24,\ 2.24)\ .$$

This contains the value 0, so there is not enough evidence to conclude that the two expectations are different.

Confidence interval for σ_X^2/σ_Y^2

It is important that you realise that the construction of the exact confidence interval for $\mu_X - \mu_Y$ only "works" when the variances σ_X^2 and σ_Y^2 are *equal*. To check this, we could compare the outcomes of S_X^2 and S_Y^2. More precisely, we could construct a confidence interval for σ_X^2/σ_Y^2, on the basis of the outcomes of the statistic S_X^2/S_Y^2 and see if it contains the number 1. The distribution of this statistic (after a scaling) is called the **F-distribution**, after R.A. Fisher, one of the founders of modern statistic. So, in addition to the Student's t distribution and the χ^2 distribution this is the third important distribution that appears in the study of statistics. Again this is a family of distributions, this time depending on two parameters (called, as usual, *degrees of freedom*). We write $\mathsf{F}(m,n)$ for an F distribution with degrees of freedom m and n. Figure 6.4 gives a plot of various pdfs of this family. Here is the beginning of the script — you can work out the rest.

```
> curve(df(x,df1=1,df2=3),xlim=c(0,8),ylim=c(0,1.5),ylab="density")
```

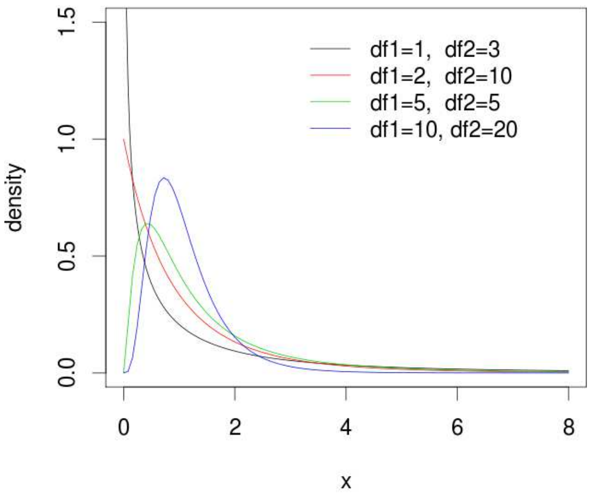


Figure 6.4: The pdfs of F distributions with various degrees of freedom (df).

It is out of the scope of this 1-st year course to discuss all the properties of the F distribution (or indeed the t and the χ^2), but the thing to remember is that it is just a probability distribution, like the normal and uniform one, and we can calculate pdfs, cdfs, and quantiles exactly as for the normal distribution, using the "d, p, q, r" construction, as in Table 4.1. ☞ 64

The precise result for the distribution of S_X^2/S_Y^2 is as follows.

Theorem 6.6 Let $X_1,\ldots,X_m \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X,\sigma_X^2)$, $Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu_Y,\sigma_Y^2)$, and $X_1,\ldots,X_m, Y_1,\ldots,Y_n$ independent. Then

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathsf{F}(m-1,n-1)\ . \tag{6.13}$$

If we want to give a confidence interval for σ_X^2/σ_Y^2, then we can use F in (6.13) as a pivot. In particular,

$$\mathbb{P}\left(q_1 < \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} < q_2\right) = 1 - \alpha\ ,$$

where q_1 and q_2 are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the $\mathsf{F}(m-1, n-1)$ distribution. Rearranging gives

$$\mathbb{P}\left(\frac{1}{q_2}\frac{S_X^2}{S_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{1}{q_1}\frac{S_X^2}{S_Y^2}\right) = 1 - \alpha\ ,$$

so that

$$\left(\frac{1}{q_2}\frac{S_X^2}{S_Y^2}, \frac{1}{q_1}\frac{S_X^2}{S_Y^2}\right)$$

is a $100(1-\alpha)\%$ stochastic confidence interval for σ_X^2/σ_Y^2 .

Example 6.8 In Example 6.7 we assumed the variances σ_X^2 and σ_Y^2 of the two samples were the same. To *check* this, we could construct a confidence interval for σ_X^2/σ_Y^2 and see if it contains the value 1. We have $s_X^2/s_Y^2 = 15^2/12^2$. For an 95% confidence interval we need to evaluate the 0.025 and 0.975 quantiles of the $\mathsf{F}(14, 9)$ distribution. Using the `R` command `qf(0.025,14,9)`, we find $q_1 = 0.3115944$, and `qf(0.975,14,9)` gives $q_2 = 3.797952$. The confidence interval is thus

$$\left(\frac{1}{3.797952}\frac{15^2}{12^2}, \frac{1}{0.3115944}\frac{15^2}{12^2}\right) \approx (0.41, 5.01),$$

which clearly contains 1, so that there is no ground to suspect that the true variances are different.

6.3.4 Binomial data, one sample

Example 6.9 In an opinion poll of 1000 registered voters, 227 voters say they will vote for the Greens. Give a 95% confidence interval for the proportion p of Green voters of the total population.

A systematic way to proceed is to view the data, 227, as the outcome of a random variable X (the number of green voters under 1000 registered voters) with a $\mathsf{Bin}(1000, p)$ distribution. In other words, we view X as the total number of "Heads" (= votes green) in a coin flip experiment with some unknown probability p of getting Heads. Note that this is only a *model* for the data. In practice it is not always possible to truly select 1000 people at random from the population and find their true party preference. For example a randomly selected person may not wish to participate or could deliberately give the "wrong answer".

Now, let us proceed to make a confidence interval for p, in the general situation that we have an outcome of some random variable X with a $\mathsf{Bin}(n, p)$ distribution. It is not so easy to find an exact confidence interval for p that satisfies (6.4) in Definition 6.1. Instead, when n is large we rely on the central limit theorem (see Section 5.5) to construct an *approximate* confidence interval. The reasoning is as follows: ☞ 74

For large n, X has approximately a $\mathsf{N}(np, np(1-p))$ distribution. Let $\widehat{P} = X/n$ denote the estimator of p. We use capital letter P to stress that an estimator is a random variable and we add a hat $\widehat{P}$ for estimator. The outcome of $\widehat{P}$ is denoted $\widehat{p}$ which is an estimate of the parameter p. Then $\widehat{P}$ has approximately a $\mathsf{N}(p, p(1-p)/n)$ distribution. For some small α (e.g., $\alpha = 0.05$) let q be the $1 - \alpha/2$ quantile of the standard normal distribution. Thus, with Φ the cdf of the standard normal distribution, we have

$$\Phi(q) = 1 - \alpha/2 \,.$$

Then, using the pivot variable

$$\frac{\widehat{P} - p}{\sqrt{p(1-p)/n}},$$

which is approximately standard normal, we have

$$\mathbb{P}\left(-q < \frac{\widehat{P} - p}{\sqrt{p(1-p)/n}} < q\right) \approx 1 - \alpha \,.$$

Rearranging gives:

$$\mathbb{P}\left(\widehat{P} - q\sqrt{\frac{p(1-p)}{n}} < p < \widehat{P} + q\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha \,.$$

This would suggest that we take $\widehat{p} \pm q\sqrt{\frac{p(1-p)}{n}}$ as an numerical (approximate) $(1 - \alpha)$ confidence interval for p, were it not for the fact that the bounds still contain the unknown p! However, for large n the estimator $\widehat{P}$ is close to the real p, so that we have

$$\mathbb{P}\left(\widehat{P} - q\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}} < p < \widehat{P} + q\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}\right) \approx 1 - \alpha \,.$$

Hence, an numerical *approximate* $(1 - \alpha)$-confidence interval for p is

$$\widehat{p} \pm q\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \,. \qquad (6.14)$$

Example 6.10 For Example 6.9 we have $\widehat{p} = 227/1000 = 0.227$, and $q = 1.960$, so that an approximate 95% numerical CI for p is given by

$$(0.227 - 1.960 \times 0.0132, 0.227 + 1.960 \times 0.0132) = (0.20, 0.25) \,.$$

6.3.5 Binomial data, two samples

Example 6.11 Two groups of men and women are asked whether they experience nightmares "often" (at least once a month) or "seldom" (less than once a month). The results are given in Table 6.2.

Table 6.2: Counts of people experiencing nightmares.

	Men	Women	Total
Often	55	60	115
Seldom	105	132	237
Total	160	192	

The observed proportions of frequent nightmares by men and women are 34.4% and 31.3%. Is this difference statistically significant, or due to chance? To assess this we could make a confidence interval for the difference of the true proportions p_X and p_Y.

The general model is as follows. Let X be the number of "successes" in group 1; $X \sim \mathsf{Bin}(m, p_X)$. ($p_X$ unknown.) Let Y be the number of "successes" in group 2; $Y \sim \mathsf{Bin}(n, p_Y)$. ($p_Y$ unknown.) Assume X and Y are independent.

We wish to compare the two proportions via a $(1-\alpha)$-confidence interval for $p_X - p_Y$.

The easiest way is to again rely on the central limit theorem. We assume from now on that m and n are sufficiently large (mp_X and $m(1-p_X) > 5$, np_Y and $n(1-p_Y) > 5$), so that the normal approximation the binomial distribution can be applied.

Let $\widehat{p}_X = X/m$ and $\widehat{p}_Y = Y/n$. By the central limit theorem,

$$\frac{\widehat{P}_X - \widehat{P}_Y - (p_X - p_y)}{\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}}$$

has approximately a $\mathsf{N}(0,1)$ distribution. Hence, with q the $(1-\alpha/2)$-quantile of the $\mathsf{N}(0,1)$ distribution (as in Section 6.3.4), we have

$$\mathbb{P}\left(-q \leqslant \frac{\widehat{P}_X - \widehat{P}_Y - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}} \leqslant q\right) \approx 1 - \alpha\ .$$

Rewriting, this gives

$$\begin{aligned}\mathbb{P}\Bigg(\widehat{P}_X - \widehat{P}_Y - q\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}} \leqslant p_X - p_Y \\ \leqslant \widehat{P}_X - \widehat{P}_Y + q\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}\Bigg) \\ \approx 1 - \alpha.\end{aligned}$$

As in the 1-sample case of Section 6.3.4, the same is *approximately* true, if we replace p_X and p_Y above by $\widehat{P}_X$ and $\widehat{P}_Y$ (law of large numbers). We now have stochastic bounds which only depend on the data.

Hence, an numerical *approximate* $100(1-\alpha)\%$ stochastic confidence interval for $p_X - p_Y$ is

$$\widehat{p}_X - \widehat{p}_Y \pm q\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{m} + \frac{\widehat{p}_Y(1-\widehat{p}_Y)}{n}}\ , \tag{6.15}$$

where q is the $1-\alpha/2$ quantile of the standard normal distribution.

Example 6.12 We continue Example 6.11. We have $\widehat{p}_X = 55/160$, $\widehat{p}_Y = 60/192$ and $q = 1.96$, so that a 95% numerical CI for $p_X - p_Y$ is given by

$$(0.031 - 0.099, 0.031 + 0.099) = (-0.07, 0.13)\ .$$

This interval contains 0, so there is no evidence that men and women are different in their experience of nightmares.

6.3.6 Summary

For one- and two-sample normal (Gaussian) data Table 6.3 provides *exact* confidence intervals for various parameters. The model for the two-sample data is $X_1, \ldots, X_m \sim_{\text{iid}} \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n \sim_{\text{iid}} \mathsf{N}(\mu_Y, \sigma_Y^2)$, where $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are independent. All parameters are assumed to be unknown. S_p is the pooled sample variance.

Table 6.3: Exact confidence intervals for normal data with unknown mean and variance. For the $\mu_X - \mu_Y$ confidence interval it must hold that $\sigma_X^2 = \sigma_Y^2$.

Parameter	$1-\alpha$ confidence interval	quantile(s)
μ_X	$\bar{X} \pm q\dfrac{S_X}{\sqrt{m}}$	$q = (1-\alpha/2)$-quantile of t_{m-1}
σ_X^2	$\left(\dfrac{(m-1)S_X^2}{q_2}, \dfrac{(m-1)S_X^2}{q_1}\right)$	$q_1 = \alpha/2$-quantile of χ_{m-1}^2 $q_2 = (1-\alpha/2)$-quantile of χ_{m-1}^2
$\mu_X - \mu_Y$	$\bar{X} - \bar{Y} \pm q\, S_p \sqrt{\dfrac{1}{m} + \dfrac{1}{n}}$	$q = (1-\alpha/2)$-quantile of t_{m+n-2}
σ_X^2/σ_Y^2	$\left(\dfrac{1}{q_2}\dfrac{S_X^2}{S_Y^2}, \dfrac{1}{q_1}\dfrac{S_X^2}{S_Y^2}\right)$	$q_1 = \alpha/2$-quantile of $\mathsf{F}(m-1, n-1)$ $q_2 = (1-\alpha/2)$-quantile of $\mathsf{F}(m-1, n-1)$

For one- and two-sample data from the binomial distribution, described by the model $X \sim \mathsf{Bin}(m, p_X)$ and $Y \sim \mathsf{Bin}(n, p_Y)$ independently, approximate $(1-\alpha)$ confidence intervals for p_X and $p_X - p_Y$ are given in Table 6.4. We use the notation $\widehat{p}_X = X/m$ and $\widehat{p}_Y = Y/n$. q is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Table 6.4: Approximate confidence intervals for binomial data.

Parameter	Approximate $1-\alpha$ confidence interval
p_X	$\widehat{p}_X \pm q\sqrt{\dfrac{\widehat{p}_X(1-\widehat{p}_X)}{m}}$
$p_X - p_Y$	$\widehat{p}_X - \widehat{p}_Y \pm q\sqrt{\dfrac{\widehat{p}_X(1-\widehat{p}_X)}{m} + \dfrac{\widehat{p}_Y(1-\widehat{p}_Y)}{n}}$

Conclusion

- Statistical models are often depend on parameters that need to be estimated from the data.
- An estimate (number) is an outcome of an estimator (random).

- To assess how good an estimate is, we can investigate the properties of the corresponding estimator.
- A $1-\alpha$ stochastic confidence interval contains the true parameter with $1-\alpha$ probability.
- A $1-\alpha$ numerical confidence interval is an outcome of the corresponding stochastic interval.
- A pivot variable (pivotal quantity) is a function of the data and of the unknown parameter.
- Pivot variables can be used to construct confidence interval.
- Using the central limit theorem approximate confidence intervals can be constructed for the mean (expectation) of a distribution.
- Exact confidence intervals can be constructed for the parameters of various 1- and 2-sample models involving Normal and Binomial distributions.
- These exact confidence intervals require the quantiles of the Student's t, the χ^2 distribution or the F distributions.

6.4 Problems

1. Suppose in Example 6.4 the data is known to be normal. Give an exact confidence interval for μ.

2. In Example 6.5 show that if (a, b) is a numerical confidence for μ, then $(a/50, b/50)$ is a numerical confidence interval for $\mu/50$. [Hint: construct a *stochastic* 95% CI for $\mu/50$.]

3. A study of iron deficiency among infants compared breast-fed with formula-fed babies. A sample of 25 breast-fed infants gave a mean blood haemoglobin level of 13.3 and a standard deviation of 1.4, while a sample of 21 formula-fed infants gave a mean and standard deviation of 12.4 and 2.0 respectively. From previous experience, haemoglobin levels can be assumed to follow a normal distribution. Assuming equal variances, calculate a 95% confidence interval for the difference between the mean haemoglobin levels of the two groups.

Chapter 7

Hypothesis Testing

Hypothesis testing involves making *decisions* about certain hypotheses on the basis of the observed data. In many cases we have to decide whether the observations are due to "chance" or due to an "effect". We will guide you through the steps that need to be taken to carry out a statistical test. Standard tests for various 1- and 2-sample problems involving Normal and Binomial random variables are provided.

7.1 Introduction

Suppose the model for the data $\mathbf{X}$ is described by a family of probability distributions that depend on a parameter $\boldsymbol{\theta}$. For example, for the 1-sample normal model $\mathbf{X} = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$. So in this case $\boldsymbol{\theta}$ is the vector (μ, σ^2).

The aim of *hypothesis testing* is to decide, on the basis of the observed data $\mathbf{x}$, which of two competing hypotheses on the parameters is true. For example, one hypothesis could be that $\mu = 0$ and the other that $\mu \neq 0$. Traditionally, the two hypotheses do not play equivalent roles. One of the hypothesis contains the "status quo" statement. This is the **null hypothesis**, often denoted by H_0. The **alternative hypothesis**, denoted H_1, contains the statement that we wish to show. A good analogy is found in a court of law. Here, H_0 (present state of affairs) could be the statement that a suspect is innocent, while H_1 is the statement that the suspect is guilty (what needs to be demonstrated). The legal terms such as "innocent until proven guilty", and "without reasonable doubt" show clearly the asymmetry between the hypotheses. We should only be prepared to reject H_0 if the observed data, that is the evidence, is very unlikely to have happened under H_0.

The decision whether to reject H_0 or not is dependent on the outcome of a **test statistic** T, which is a function of the data $\mathbf{X}$ only. The ***p*-value** is the probability that under H_0 the (random) test statistic takes a value as extreme as or more extreme than the one observed. Let t be the observed outcome of the test statistic T. We consider three types of tests:

- **Left one-sided test**. Here H_0 is rejected for small values of t, and the p-value is defined as $p = \mathbb{P}_{H_0}(T \leqslant t)$.

- **Right one-sided test**: Here H_0 is rejected for large values of t, and the p-value is defined as $p = \mathbb{P}_{H_0}(T \geqslant t)$,
- **Two-sided test**: In this test H_0 is rejected for small or large values of t, and the p-value is defined as $p = \min\{2\mathbb{P}_{H_0}(T \leqslant t),\ 2\mathbb{P}_{H_0}(T \geqslant t)\}$.

The smaller the p-value, the greater the strength of the evidence against H_0 provided by the data. As a rule of thumb:

$p < 0.10$	suggestive evidence,
$p < 0.05$	reasonable evidence,
$p < 0.01$	strong evidence.

The following decision rule is generally used to decide between H_0 and H_1:

Decision rule : *Reject H_0 if the p-*value *is smaller than some* **significance level** α.

In general, a statistical test involves the following steps.

Steps for a Statistical Test

1. Formulate a statistical model for the data.
2. Give the null and alternative hypotheses (H_0 and H_1).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under H_0.
5. Evaluate the outcome of the test statistic.
6. Calculate the p-value.
7. Accept or reject H_0 based on the p-value.

Choosing an appropriate test statistic is akin to selecting a good estimator for the unknown parameter $\boldsymbol{\theta}$. The test statistic should summarise the information about $\boldsymbol{\theta}$ and make it possible to distinguish between the two hypotheses.

Example 7.1 (Blood Pressure) Suppose the systolic blood pressure for white males aged 35–44 is known to be normally distributed with expectation 127 and standard deviation 7. A paper in a public health journal considers a sample of 101 diabetic males and reports a sample mean of 130. Is this good evidence that diabetics have on average a higher blood pressure than the general population?

To assess this, we could ask the question how likely it would be, *if diabetics were similar to the general population*, that a sample of 101 diabetics would have a mean blood pressure this far from 127.

Let us perform the seven steps of a statistical test. A reasonable model for the data is $X_1, \ldots, X_{101} \sim_{\text{iid}} \mathsf{N}(\mu, 49)$. Alternatively, the model could simply be $\bar{X} \sim \mathsf{N}(\mu, 49/101)$, since we only have an outcome of the sample mean of the blood pressures. The null hypothesis (the status quo) is $H_0 : \mu = 127$; the alternative hypothesis is $H_1 : \mu > 127$.

We take $\bar{X}$ as the test statistic. Note that we have a right one-sided test here, because we would reject H_0 for high values of $\bar{X}$. Under H_0 we have $\bar{X} \sim \mathsf{N}(127, 49/101)$. The outcome of $\bar{X}$ is 130, so that the p-value is given by

$$\mathbb{P}(\bar{X} \geqslant 130) = \mathbb{P}\left(\frac{\bar{X} - 127}{\sqrt{49/101}} \geqslant \frac{130 - 127}{\sqrt{49/101}}\right) = \underbrace{\mathbb{P}(Z \geqslant 4.31)}_{\texttt{1-pnorm(4.31)}} \approx 8.16 \cdot 10^{-6} ,$$

where $Z \sim \mathsf{N}(0, 1)$. So it is extremely unlikely that the event $\{\bar{X} \geqslant 130\}$ occurs if the two groups are the same with regard to blood pressure. However, the event *has* occurred. Therefore, there is is *strong* evidence that the blood pressure of diabetics differs from the general public.

Example 7.2 (Loaded Die) We suspect a certain die to be loaded. Throwing 100 times we observe 25 sixes. Is there enough evidence to justify our suspicion?

We ask ourselves a similar type of question as in the previous example: What if the die would indeed be fair. What would then be the probability that out of 100 tosses 25 or more sixes would appear? To calculate this, let X be the number of sixes out of 100. Our model is $X \sim \mathsf{Bin}(100, p)$, with p unknown. We would like to show the hypothesis $H_1 : p > 1/6$; otherwise, we do not reject (accept) the null hypothesis $H_0 : p = 1/6$. Our test statistic is simply X. Under H_0, $X \sim \mathsf{Bin}(100, 1/6)$, so that the p-value for this right one-sided test is

$$\mathbb{P}(X \geqslant 25) = \underbrace{\sum_{k=25}^{100} \binom{100}{k} (1/6)^k \, (5/6)^{100-k}}_{\texttt{1-pbinom(24,100,1/6)}} \approx 0.0217 .$$

This is quite small. Hence, we have *reasonable* evidence that the die is loaded.

In the rest of this chapter we are going to look at a selection of basic tests, involving one or two iid samples from either a normal or Bernoulli distribution.

7.2 Type-I error, Type-II error, Power

In any hypothesis test we can make two types of mistakes, illustrated in Table 7.1.

Table 7.1: Type-I and type-II errors

	True state of nature	
Decision	**H_0 is true**	**H_1 is true**
Accept H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

Ideally we would like construct tests which make both types of errors (let's call them type-I and type-II error, denoted E_I and E_{II}, respectively) as small as possible.

Unfortunately, this is not possible, because the two errors "compete" with each other. For example, if we make the critical region larger, we decrease E_{II} but at the same time increase E_I. On the other hand, if we make the critical region smaller, we decrease E_I but at the same time increase E_{II}.

Now, in classical statistics the null hypothesis and alternative hypothesis do not play equivalent roles. We are only prepared to reject the zero hypothesis if the observed value of the test statistic is very "unlikely" under the zero hypothesis. Only if this evidence is strong do we wish to reject H_0. In other words, we certainly do not wish to make a large type I error.

7.3 One-sample *t*-test

Example 7.3 One of the statements in a research article is that the amount of caffeine in regular cola is "19 mg per 6-oz serving". In a second study the caffeine content was determined for a sample of size $n = 40$ of a different brand of cola. The sample mean and standard deviation were 19.57 (mg) and 1.40 (mg) respectively. Should we conclude that the expected amount of caffeine in this brand is more than "19 mg per 6-oz serving"?

To answer this question, we again consider an appropriate model for this situation. We represent the observations by $X_1, \ldots, X_n$, and assume that they form an iid sample from a $\mathsf{N}(\mu, \sigma^2)$ distribution, where μ and σ^2 are *unknown*. The hypotheses can now be formulated as: $H_0 : \mu = 19$ against $H_1 : \mu > 19$.

Which test statistic should we choose? Since we wish to make a statement about μ, the test statistic should reflect this. We could take $\bar{X}$ as our test statistic and reject H_0 for large values of $\bar{X}$. However, this leads to a complication. It looks like our null hypothesis only contains one parameter value, but in fact it contains *many*, because we should have written

$$H_0 : \mu = 19, \quad 0 < \sigma^2 < \infty \,.$$

It is the unknown σ^2 that leads to the complication in choosing $\bar{X}$ as our test statistic. To see this, consider the following two cases. First consider the case where σ^2 is very small. In that case, $\bar{X}$ is under H_0 very much concentrated around 19, and therefore any deviation from 19, such as 19.57 would be most unlikely under H_0. We would therefore reject H_0. On the other hand, if σ^2 is very large, then a value of 19.57 could very well be possible under H_0, so we would not reject it.

This shows that $\bar{X}$ is not a good test statistic, but that we should "weigh" it with the standard deviation. That is, we should measure our deviation from 19 in units of σ rather than in units of 1. However, we do not know σ. But this is easily fixed by replacing σ with an appropriate estimator. This leads to the test statistic

$$T = \frac{\bar{X} - 19}{S/\sqrt{40}} \,.$$

☞ 90 The factor $\sqrt{40}$ is a "normalising" constant which enables us to utilise Theorem 6.7. Namely, under H_0 the random variable T has a $\mathsf{t}_{n-1} = \mathsf{t}_{39}$ distribution. Note that this is true for *any* value of σ^2. The observed outcome of T is

$$\frac{19.57 - 19}{1.166}\sqrt{40} = 3.09 \,.$$

Using R,

```
> 1 - pt(3.09,df=39)
[1] 0.001841394
```

we find the p-value

$$\mathbb{P}_{H_0}(T \geqslant 2.57) = 0.0018 \; .$$

Since this is very small, we reject H_0. Therefore, there is significant evidence that the average caffeine content is more for the non-regular brand. However, it is important to note that the "statistical significance" of this statement – the difference is not due to chance – is something completely different from its "practical significance" – the difference is important practically. First of all, the true value of μ could very well still be close to 19. To see this, verify that a 99% CI for μ is $(19.07, 20.07)$. Secondly, even if μ would be 20 or even 20.5, it is not at all clear that such a deviation is "practically" important. For example, is it unhealthy to have 1 mg more caffeine per ounce in your cola? Thus although there is strong evidence of the difference, the difference is so small that it is likely to be irrelevant in practice.

The above test is often called a **1-sample *t*-test**. In general, let $X_1, \ldots, X_n \sim_{\text{iid}} \mathsf{N}(\mu, \sigma^2)$. Let μ_0 be a given number. We wish to test the hypothesis $H_0 : \mu = \mu_0$ against left-, right-, and two-sided alternatives by using the test statistic

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \; , \tag{7.1}$$

with $\bar{X} = \frac{1}{n} \sum_{i-1}^{n} X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Under H_0 we have $T \sim \mathsf{t}_{n-1}$. Reject H_0 according to the p-value. Note that the p-value which depends on whether the test is left one-sided, right one-sided or two-sided. In the example above, the test is right-one sided (we reject H_0 for large value of T).

Using R

In the above example, we did not have access to the individual measurements $x_1, \ldots, x_n$ — only to the summaries $\bar{x}$ and s, which provided all the information required to carry out the hypothesis testing. In practice, however, the individual measurements will be available. In that case it is convenient to carry out the t-test using the R function `t.test()`. Let us generate some iid data (of size $n = 40$) from a $\mathsf{N}(19.65, 1.40)$ distribution.

```
> set.seed(3)
> x <- rnorm(40,19.65,1.40)
> x[1:10] # the first 10 observations
[1] 18.30329 19.24046 20.01230 18.03702 19.92410 19.69217
[7] 19.76958 21.21325 17.94360 21.42432
```

We could obtain the sample mean and the sample standard deviation using the following code:

```
> mean(x)
[1] 19.57043
```

```
> sd(x)

[1] 1.166238
```

The sample mean and the sample standard deviation are the same as we used previously (example 7.3) as summary statistics to compute a 1-sample t-test and a confidence interval. Note that if you repeat this yourself, you will get the same realisations, ☞ 66 as we have specified the random number generator using `set.seed()` function. Applying the `t.test()` function we get for the above data:

```
> t.test(x,mu=19,conf.level=0.99)

          One Sample t-test

data:  x
t = 3.0935, df = 39, p-value = 0.003648
alternative hypothesis: true mean is not equal to 19
99 percent confidence interval:
 19.07110 20.06977
sample estimates:
mean of x
 19.57043
```

The main output of the function `t.test()` are: the outcome of the T statistic (t = 3.0935), the p-value = 0.0036, the alternative hypothesis (*true mean is not equal to 19*), a 99% confidence interval ([19.07, 20.07]) and the sample mean $\bar{x}$ = 19.57. Note that we find the same results for the outcome of T, the 99% confidence interval and the sample mean. However, the p-value is associated to a two-sided hypothesis indicating by the alternative hypothesis: *true mean is not equal to 19*. To get a one-sided test, we need to specify the alternative by using the argument `alternative="greater"` in our case. By default, R performs a two-sided test.

```
> t.test(x,mu=19,alt="greater")$p.value

[1] 0.00182425 # same p-value found in the previous example
```

Note that the value of μ_0 for the null hypothesis $H_0 : \mu = \mu_0$ is specified by the argument `mu=19`. Finally, a 95% confidence interval of the mean μ could be obtained using the argument `conf.level=0.95` (level by default in `t.test()` function):

```
> t.test(x,mu=19,conf.int=0.95)$conf.int

[1] 19.19745 19.94341
attr(,"conf.level")
[1] 0.95
```

7.4 One-sample test for proportions

Example 7.4 In a certain market research study we wish to investigate whether people would prefer a new type of sweetener in a certain brand of yoghurt. Ten people were given two packets of yoghurt, one with the old sweetener and one with the new sweetener. Eight of the ten people preferred the yoghurt with the new sweetener and two preferred the old yoghurt. Is there enough evidence that the new style of yoghurt is preferred?

First we formulate the model. Let $X_1, \dots, X_{10}$ be such that

$$X_i = \begin{cases} 1 & \text{if person } i \text{ prefers the new yoghurt,} \\ 0 & \text{if person } i \text{ prefers the old yoghurt,} \end{cases}$$

$i = 1, \dots, 10$. We assume that $X_1, \dots, X_{10}$ are independent and that for all i, $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$, for some unknown p (between 0 and 1). We wish to test

$$H_0 : p = 0.5 \quad \text{against} \quad H_1 : p > 0.5 \; .$$

As test statistic we could use the total number of people preferring the new yoghurt, $X = \sum_{i=1}^{10} X_i$, and we would reject H_0 for large values of X. Under H_0 the test statistic has a $\mathsf{Bin}(10, 1/2)$ distribution. The p-value is thus, similar to Example 7.2,

$$\mathbb{P}_{H_0}(X \geqslant 8) = \sum_{k=8}^{10} \binom{8}{k} (1/2)^{10} \approx 0.0546875 \; .$$

Note that we can evaluate the probability above in `R` using `1 - pbinom(7,10,0.5)`. Since the p-value is reasonably small (0.055), there is some doubt about H_0.

Remark 7.1 Our model above is in a sense over-specific. We assume that we observe the preference X_i for each individual. But in fact, we only observe the total number of preferences $X = X_1 + \cdots + X_n$ for the new yoghurt. An alternative and simpler model would suffice here, namely: let X be the total number of preferences for the new type of yoghurt, we assume $X \sim \mathsf{Bin}(n, p)$, for some unknown p. The test now proceeds in exactly the same way as before.

We now describe the general situation for the **one-sample binomial test**. Suppose that $X_1, \dots, X_n$ are the results of n independent Bernoulli trials with success parameter p. That is the X_i's are independent and

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0) \; .$$

Then, $X := X_1 + \cdots + X_n \sim \mathsf{Bin}(n, p)$. We wish to test $H_0 : p = p_0$ against left-, right-, and two-sided alternatives.

As test statistic we can use X, which under H_0 has a $\mathsf{Bin}(n, p_0)$ distribution. We accept/reject H_0 based on the p-value of the test.

Using R

For one-sample binomial test, we use the R function `binom.test()` which arguments and output are very similar to those used with `t.test()` function for 1-sample t-test.

```
> binom.test(x=8,n=10,p=0.5,alternative="greater")

        Exact binomial test

data:  8 and 10
number of successes = 8, number of trials = 10, p-value = 0.05469
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4930987 1.0000000
sample estimates:
probability of success
                   0.8
```

Using the Normal Approximation

For large n, analogously to Sections 6.3.4, X has approximately a $\mathsf{N}(np, np(1-p))$ distribution and then the estimator $\widehat{P} = X/n$ has approximately a $\mathsf{N}(p, p(1-p)/n)$ distribution. It follows that

$$\frac{\widehat{P}-p}{\sqrt{p(1-p)/n}},$$

has approximately a $\mathsf{N}(0,1)$ distribution. Now, under $H_0: \ p = p_0$, our test statistic

$$Z = \frac{\widehat{P}-p_0}{\sqrt{p_0(1-p_0)/n}},$$

has approximately a $\mathsf{N}(0,1)$ distribution.

Example 7.5 Returning to Example 7.4, from our data we have the estimate $\widehat{p} = \frac{8}{10}$.Thus, the outcome of the test statistic is

$$z = \frac{0.8-0.5}{\sqrt{0.5(1-0.5)/10}} = 1.897367.$$

This gives a p-value of $\mathbb{P}_{H_0}(Z \geqslant 1.897367) \approx 0.02889$ (in R type `1 - pnorm(1.897367)`). This approximate p-value is far from the exact test as our sample size is not enough large to use the central limit theorem. It is possible to obtain this result by using the R function `prop.test`:

```
> prop.test(x=8,n=10,p=0.5,alternative="greater",correct=FALSE)$p.value

[1] 0.02888979
```

7.5 Two-sample *t*-test

Example 7.6 Let us return to the data of Example 6.7. That is, we have a sample of 15 1st-year students with a mean mass of 62.0kg and a standard deviation of 15kg, and a sample of 10 3rd-year students with a mean mass of 71.5kg and a standard deviation of 12kg. Suppose we wish to test if in expectation the 3rd-year students are *heavier* than the 1st-year students. Note that in Example 6.7 we did not express any idea which of the two groups were heavier. We were only interested in a confidence interval for the difference of the means. ☞ 94

The standard model choice is again the two-sample normal model: Let X_i be the mass of the i-th 1st-year student, $i = 1, \dots, m$. Let Y_i be the mass of i-th 3rd year students, $i = 1, \dots, n$. (Here $m = 15$ and $n = 10$). We assume that

- $X_1, \dots, X_m \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X, \sigma_X^2)$.
- $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu_Y, \sigma_Y^2)$.
- $X_1, \dots, X_m, Y_1, \dots, Y_n$ are *independent*,

where μ_X, μ_Y, σ_X^2, and σ_X^2 are unknown parameters. We wish to test $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X < \mu_Y$.

Suppose that the variances of the two samples are the *same*: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, for some unknown σ^2. In analogy with the 1-sample t-test and in view of (6.12) a sensible test statistic is then ☞ 94

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{1/m + 1/n}} , \tag{7.2}$$

where S_p is the pooled sample variance in (6.11). We reject the null hypothesis for small values of T. By (6.12) if $\mu_X = \mu_Y$, then T has a t_{23} distribution (the number of degrees of freedom is $m + n - 2 = 15 + 10 - 2 = 23$). This enables us to find the exact p-value. In particular, from our sample we have the following outcome of S_p^2:

$$s_p^2 = \frac{14 \times (15^2) + 9 \times (12^2)}{23} = 193.304$$

So we have an observed value of T of

$$t = \frac{62.0 - 71.5}{\sqrt{193.304 \left(\frac{1}{15} + \frac{1}{10}\right)}} = -1.674 .$$

Using the computer (in R: type `pt(-1.674,23)`) we find

$$\mathbb{P}_{H_0}(T \leqslant -1.674) \approx 0.0538 .$$

It follows that there is reasonable evidence to support the student's theory that the 3rd year students are in expectation heavier.

In general we have the following situation. Consider the model given by the dot-points in Example 7.6. But also assume $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. We wish to test the hypothesis

$H_0 : \mu_X = \mu_Y$ against various alternatives. The **two-sample *t*-test** is described as follows. As test statistic we take

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \, ,$$

where S_p^2 is the pooled sample variance:

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{m+n-2},$$

$\bar{X} = m^{-1} \sum_{i=1}^m X_i$ and $\bar{Y} = n^{-1} \sum_{j=1}^n Y_j$.

Under H_0 we have $T \sim \mathsf{t}_{n+m-2}$. We accept/reject H_0 depending on the p-value associated with the alternative (left-, right-, or two-sided).

Unequal Variances

For unequal variances $\sigma_X^2 \neq \sigma_Y^2$ the usual statistic test is then

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

This statistic follows approximately a Student distribution; the number of degrees of freedom (df) can be computed using Satterthwaites's approximation:

$$\text{df} = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_X^1}{m}\right)^2 + \frac{1}{m-1}\left(\frac{s_Y^1}{n}\right)^2}$$

Using R

The dataset `birthwt` from the MASS package in R contains information on the birth weights[1] of babies as well as various characteristics of the mother, such as whether she smokes, her age, etc.

First we load the data:

```
> library(MASS)        # load the package MASS
> help(birthwt)        # find information on the data set birthwt
> attach(birthwt)      # make variables available
```

The two variables of interest to us here are the birth weight in grams, `bwt`, and the smoking status of the mother, `smoke` (0 = no, 1= yes). Note that `smoke` divides the birth weight data into two groups. We can thus perform a 2-sample t-test to test whether the mean birthweight differs between the two groups. Let us first do a graphical inspection.

[1] We realise that we should have written birth *mass* instead of birth *weight*, but we decided to stick with the more informal label "weight" used in the actual data files. As long as you understand the difference between mass (measured in kg) and weight (a force, measured in Newton = kg m/s^2) there is no risk of confusion when "weight" is sometimes used in the colloquial sense.

```
> par(mfrow=c(1,2))
> plot(bwt~smoke)
> plot(bwt~factor(smoke))
```

The above R code makes the two plots in Figure 7.1. Note that in the left plot smoke is interpreted as a numerical variable. In the right plot smoke is make into a factor variable via the function factor() so that the plot function displays the relation between the two variables via a box plot.

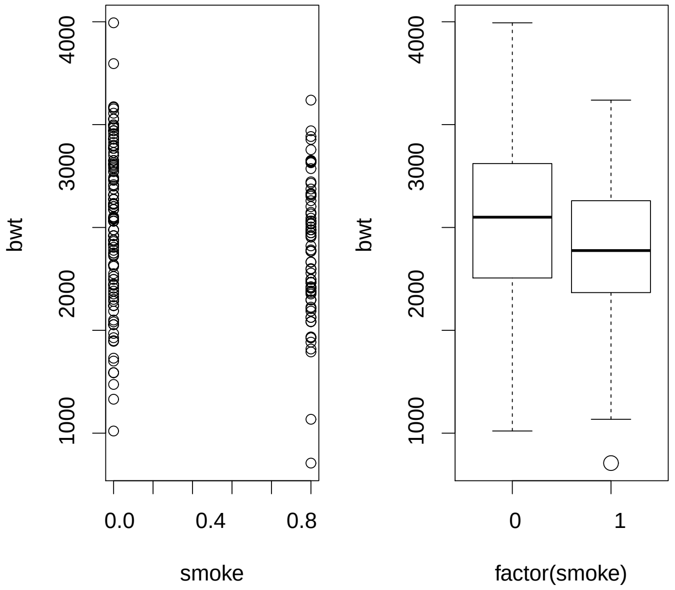


Figure 7.1: Two different plots for birthweights versus smoking status.

It is not entirely clear if there is a difference between the two groups, although the smoking group seems to have lower birth weights overall. Let us carry out the 7 steps of the 2-sample *t*-test manually first. Later on we will just give one R command to carry out the calculations.

1. The model that we use is the standard 2-sample normal model with equal variances (graphically the assumption of equal variances does not look too implausible, but we can test for this is we wish). Let the X_i's be the birthweights of the babies with non-smoking mothers and the Y_i's the birthweights of the babies with smoking mothers. To make this explicit, let's define vectors x and y accordingly using the subset() function.

```
> x <- subset(bwt,smoke==0)
> y <- subset(bwt,smoke==1)
> m <- length(x)
> n <- length(y)
```

2. The null hypothesis is $H_0 : \mu_X = \mu_Y$; that is, the expected birthweights for the two groups is the same. What should we take for H_1? Do we want to show

$\mu_X > \mu_Y$ or $\mu_X \neq \mu_Y$? If we are more interested in showing that smoking is detrimental for the baby (in terms of birth weight), then what we wish to demonstrate is $H_1 : \mu_X > \mu_Y$. Let's take this as the alternative hypothesis.

3. As a test statistics we take the statistic T in (7.2).

4. Under H_0 T has a t_{m+n-2} distribution. In this case $m = 115$ and $n = 74$, so we have a t_{187} distribution.

5. We next compute the outcome t of T.

```
> sp <- sqrt( ((m-1)*var(x) + (n-1)*var(y))/(m+n-2) )
> t = (mean(x) - mean(y))/sp/sqrt(1/m +1/n)
> t
[1] 2.652893
```

6. The p-value for this right one-sided test is $\mathbb{P}(T \geqslant 2.652893)$, where $T \sim \mathrm{t}_{187}$.

```
pval <- 1 - pt(t,m+n-2)
pval
[1] 0.004333363
```

7. The p-value is very small, so there is strong statistical evidence that smoking has a detrimental effect on birth weight.

As mentioned, we can let R do the above analysis with one line of code. This is done using the function `t.test()`.

```
> t.test(bwt~smoke,var.equal=TRUE,alternative="greater")

        Two Sample t-test

data:  bwt by smoke
t = 2.6529, df = 187, p-value = 0.004333
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 106.9528      Inf
sample estimates:
mean in group 0 mean in group 1
       3055.696        2771.919
```

It is important in this case to include the two options `var.equal = TRUE` and `alternative="greater"`, as by default R assumes unequal variances and a two-sided alternative.

Paired Data

Example 7.7 We wish to compare the results from two labs for a specific examination. Both labs made the necessary measurement on fifteen patients.

```
>  dose.lab1 <- c(22,18,28,26,13,8,21,26,27,29,25,24,22,28,15)
>  dose.lab2 <- c(25,21,31,27,11,10,25,26,29,28,26,23,22,25,17)
```

We wish to compare the theoretical means of two random variables X and Y based on two paired samples (for all subject X_i and Y_i are not *independent*. To this end, we use the difference random variable $D = X - Y$, and we compare the theoretical mean $\delta = \mu_X - \mu_Y$ of D with the reference value 0. We are thus back to the case of **one-sample *t*-test** by assuming a normal model for the difference $D_i = X_i - Y_i \sim \mathsf{N}(\mu_X - \mu_Y, \sigma^2)$. The hypotheses of the test are $H_0 : \mu_X - \mu_Y = 0$ and $H_1 : \mu_X - \mu_Y \neq 0$. Under H_0, the test statistic is:

$$T = \frac{\bar{D}}{S/\sqrt{n}} \sim \mathsf{t}_{n-1},$$

with $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$. We need to used the argument `paired=TRUE` to perform this test in R:

```
> t.test(dose.lab1,dose.lab2,paired=TRUE)

Paired t-test

data:  dose.lab1 and dose.lab2
t = -1.7618, df = 14, p-value = 0.09991
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0695338  0.2028671
sample estimates:
mean of the differences
             -0.9333333
```

At the specified risk level $\alpha = 5$ % (p-value=0.0999), we cannot decide that the two labs give different results on average.

7.6 Two-sample *F*-test

For the standard two-sample t test it is necessary that the variances are equal. We can test for this. Suppose we have two independent normal samples as in the previous section. We wish to test $H_0 : \sigma_X^2 = \sigma_Y^2$ against various alternatives; the most relevant one is $H_1 : \sigma_X^2 \neq \sigma_Y^2$. We choose the test statistic

$$F = \frac{S_X^2}{S_Y^2},$$

which, under H_0, has an F distribution with parameters (degrees of freedom) $m-1$ and $n-1$, by Theorem 6.6, We write $F \sim \mathsf{F}(m-1, n-1)$. For the alternative hypothesis ☞ 95 $H_1 : \sigma_X^2 \neq \sigma_Y^2$ we reject H_0 for large and small values of F (two-sided test).

Example 7.8 Consider again Example 7.6. The outcome of F is $\frac{15^2}{12^2} = 1.56$. Under H_0 F has an $\mathsf{F}(14, 19)$ distribution. The corresponding p-value for this two-sided test is thus $2\mathbb{P}_{H_0}(F \geqslant 1.56) = 0.508$ — which we have evaluated using the R command `2*(1 - pf(1.56,14,9))`. Since, the p-value is large we accept the null hypothesis. So our assumption of equal variances in Example 6.7 seems justified.

Using R

Consider the birthweight data of the previous section. We assumed there that the variances of the two groups of babies were the same. Is this justified? To carry out the F test, we use continue to use the variables that we used in the previous section; in particular, the birthweight data was collected in the vectors `x` and `y` of size `m` and `n`, respectively.

```
> f <- var(x)/var(y)
> pvalf <- 2*(1 - pf(f,m-1,n-1))  #p-value for 2-sided F test
> pvalf

[1] 0.2254372
```

Since the p-value is large we accept the hypothesis that the two variances are equal. We can also use the R function `var.test()` directly to test for equality of variances.

```
>  var.test(bwt~smoke)

        F test to compare two variances

data:  bwt by smoke
F = 1.3019, num df = 114, denom df = 73, p-value = 0.2254
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8486407 1.9589574
sample estimates:
ratio of variances
          1.301927
```

7.7 Two-sample test for proportions

Example 7.9 A politician believes that audience members of the ABC news are in general more left wing than audience members of a commercial news broadcast. A poll of two party preferences is taken. Of seventy ABC viewers, 40 claim left wing allegiance, while of 100 commercial station viewers, 50 claim left wing allegiance. Is there any evidence to support the politician's claim?

Our model is as follows. Let X be the number of left-wing ABC viewers out of $m = 70$, and let Y be the number of left-wing "commercial" viewers out of $n = 100$. We assume that X and Y are independent, with $X \sim \mathsf{Bin}(m, p_X)$ and $Y \sim \mathsf{Bin}(n, p_Y)$, for some unknown p_X and p_Y. We wish to test $H_0 : p_X = p_Y$ against $H_1 : p_X > p_Y$.

Since m and n are fairly large here, we proceed by using the central limit theorem ☞ 97 (CLT), analogously to Sections 6.3.4 and 6.3.5. Let $\widehat{P}_X := X/m$ and $\widehat{P}_Y := Y/n$ be the empirical proportions. By the CLT $\widehat{p}_X$ has approximately a $\mathsf{N}(p_X, p_X(1 - p_X)/m)$ distribution, and $\widehat{p}_Y$ has approximately a $\mathsf{N}(p_Y, p_Y(1 - p_Y)/n)$ distribution. It follows that

$$\frac{\widehat{P}_X - \widehat{P}_Y}{\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}}$$

has approximately a $\mathsf{N}(0,1)$ distribution. Now, under H_0, $p_X = p_Y = p$, say, and hence under H_0

$$\frac{\widehat{P}_X - \widehat{P}_Y}{\sqrt{\frac{p(1-p)}{m} + \frac{p(1-p)}{n}}}$$

has approximately a $\mathsf{N}(0,1)$ distribution. As we don't know what p is, we need to estimate it. If H_0 is true, then $X + Y \sim \mathsf{Bin}(m+n, p)$, and thus p can be estimated by the *pooled* success proportion

$$\widehat{P} := \frac{X+Y}{n+m} \,. \tag{7.3}$$

Concluding, we take as our test statistic:

$$Z = \frac{\widehat{P}_X - \widehat{P}_Y}{\sqrt{\widehat{P}(1-\widehat{P})\left(\frac{1}{m} + \frac{1}{n}\right)}}, \tag{7.4}$$

which under H_0 has approximately a $\mathsf{N}(0,1)$ distribution.

Our general formulation for the **2-sample binomial test** (also called the **test for proportions**) is as follows. Let X be the number of "successes" in group 1; $X \sim \mathsf{Bin}(m, p_X)$. ($p_X$ unknown.) Let Y be the number of "successes" in group 2; $Y \sim \mathsf{Bin}(n, p_Y)$. ($p_Y$ unknown.) Assume X and Y are independent. We wish to test $H_0 : p_X = p_Y$ against various alternatives (left one-sided, right one-sided, and two-sided). As test statistic we use Z given in (7.4). We accept/reject H_0 on the basis of the p-value.

Example 7.10 Returning to Example 7.9, from our data we have the estimates $\widehat{p_X} = \frac{40}{100}$, $\widehat{p_Y} = \frac{50}{100}$, and

$$\widehat{p} = \frac{40+50}{70+100} = \frac{90}{170} \,.$$

Thus, the outcome of the test statistic is

$$\frac{\frac{40}{70} - \frac{50}{100}}{\sqrt{\frac{90}{170} \times \frac{80}{170}\left(\frac{1}{70} + \frac{1}{100}\right)}} = 0.9183 \,.$$

This gives a p-value of $\mathbb{P}_{H_0}(Z \geqslant 0.9183) \approx 0.1792$ (in `R` type `1 - pnorm(0.9183)`), so there is no evidence to support the politician's claim.

Using R

As for one-sample test for proportions, the `R` function `prop.test()` enables us to compared two proportions:

```
> prop.test(x=c(40,50),n=c(70,100),alternative="greater",correct=F)

        2-sample test for equality of proportions without continuity
        correction

data:  c(40, 50) out of c(70, 100)
X-squared = 0.8433, df = 1, p-value = 0.1792
alternative hypothesis: greater
```

```
95 percent confidence interval:
 -0.05596576  1.00000000
sample estimates:
   prop 1    prop 2
0.5714286 0.5000000
```

Note that as expected we obtain the same p-value. However, the outcome of the statistic test corresponds to the square of the outcome of our statistic Z ($0.9183^2 = 0.8433$). The function provides also the sample proportion $\widehat{p}_X$ and $\widehat{p}_Y$.

7.8 Summary

This table lists all the tests we have introduced (Table 7.2).

Table 7.2: Standard tests.

Nature	Data	Conditions for validity	R function
mean	1 sample	$n > 30$ or normality	`t.test(x,...)`
	2 samples	normality and equal variances	`t.test(x,y,...)`
	2 samples	normality	`t.test(x,y,var.equal=F)`
	2 paired samples	$n > 30$ or normality	`t.test(x,y,paired=T)`
variance	2 samples	normality	`var.test(x,y,...)`
proportion	1 sample	$np \geqslant 5$ and $n(1-p) \geqslant 5$	`prop.test(x,...)`
	1 sample		`binom.test(x,...)`
	2 samples	large sample size	`prop.test(x,y,...)`

Conclusion

- Hypotheses are statements concerning the *parameters* of the model.
- For hypothesis testing always follow the “7 Steps”.
- A hypothesis test can be one-sided or two-sided.
- The p-value is the probability that under the null hypothesis the test statistic takes on a value as extreme as (or more extreme than) the one observed.
- Accept the null hypothesis for large p-values.

7.9 Problems

1. In 2000, a milk vendor found that 30% of his milk sales were of a low fat variety. Last week, of his 1500 milk sales, 500 were low fat. Is there any indication of a move towards low fat milk? Give the p-value associated with the test.

2. A study of iron deficiency among infants compared breast-fed with formula-fed babies. A sample of 25 breast-fed infants gave a mean blood haemoglobin level

of 13.3 and a standard deviation of 1.4, while a sample of 21 formula-fed infants gave a mean and standard deviation of 12.4 and 2.0 respectively.

(a) From previous experience, haemoglobin levels can be assumed to follow a normal distribution. Test, at the 5% level, whether the population variances for the breast-fed and formula-fed babies are equal.

(b) Assuming equal variances, calculate a 95% confidence interval for the difference between the mean haemoglobin levels of the two groups.

3. Brand A batteries cost more than brand B batteries. Their life lengths are normally distributed. A sample of 16 batteries of brand A gave a mean life length of 4.6 (hours) and a standard deviation of 1.1 (hours), while a sample of 26 batteries of brand B gave a mean and standard deviation of 4.2 and 0.9 respectively.

   (a) Give a statistical model for the experiment above.

   (b) Test the hypothesis that the expected lifetime of brand A batteries is greater than that of brand B batteries. State precisely what assumptions you make.

4. A foundry produces steel castings used in the automotive industry. An acceptable fraction of nonconforming castings from this process is 10%. If this fraction is found to be greater than 10%, the production process is examined for errors. In a random sample of 250 castings, 41 were found to be nonconforming. Test whether the production process needs to be examined.

5. Water witching, the practice of using the movement of a forked twig to locate underground water, dates back over 400 year. The following data shows the outcome of all the wells dug in a certain area. Recorded for each well was whether it proved to be successful (S) or unsuccessful (U) in providing water.

Witched Wells					Nonwitched Wells				
S	S	S	S	U	U	S	S	S	S
U	S	S	S	S	S	S	S	S	S
S	U	S	U	S	S	U	S	S	S
S	S	S	S	S	S	U	S	S	S
S	S	S	S	S	S	S	S	U	S
U	S	S	S		U	S	S	S	S
					S	S			

   Formulate and evaluate an appropriate test for these data. Give the usual "8 steps" of a statistical test explicitly.

6. Two catalysts are being analysed to determine how they affect the mean yield of a chemical process. Catalyst 1 is currently in use, but catalyst 2 is cheaper. A test is run to assess whether Catalyst 2 gives the same process yield as Catalyst 1. The result of the test is given in Table 7.3. The sample standard deviation of the yields are 2.385 and 2.983, respectively.

   (a) Give a statistical model for the experiment above.

Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75

Table 7.3: Process yields.

(b) Can we assume at an $\alpha = 0.05$ level of significance that the variances in the yield for the two catalysts are the same?

(c) Test the hypothesis that the expected yield for the two catalysts are the same. Specify the 8 steps of the statistical test. For step 6 specify the P-value, or give appropriate lower or upper bounds. Should catalyst 2 be adopted or not?

7. A researcher claims that 10% of all bicycle helmets have manufacturing flaws that could potentially cause injury to the wearer. A sample of 200 helmets revealed that 16 helmets contained such defects.

   (a) Give a 95% approximate confidence interval for the true percentage of faulty helmets.

   (b) Do the data support the researchers claim?

☞ 113

8. Let us return to Example 7.8. Consider now that your statistic test is $F = \frac{S_Y^2}{S_X^2}$. Determine the p-value associated to F for testing $H_0 : \sigma_X^2 = \sigma_Y^2$ against $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

9. Vinny from Vegas is very suspicious of biased coins. Before playing he examines each opponent's coin using the following self-devised statistical test: Throw the coin 10 times, and reject it as "biased" if the total number of Heads is more than 8 or less than 2.

   (a) Formulate the usual steps of a statistical test for Vinny's test (only steps 1–4 and 6 are required).

   (b) Compute the probability of a Type I error (=error of the first kind) of Vinny's test.

   (c) Calculate the Type II error of Vinny's test when the coin's probability of Heads is 0.6.

Chapter 8

Analysis of Variance

We present an introduction to the analysis of grouped data via an analysis of variance (ANOVA). We discuss ANOVA models with 1 factor and 2 factors with or without interaction. You will learn how to estimate parameters of the models and how to carry out hypothesis tests using R.

8.1 Introduction

Analysis of variance (ANOVA) is used to study the relationship between a *quantitative* variable of interest and one or several *categorical* variables. The variable of interest is called the **response** variable (or dependent variable) and the other variables are called **explanatory** variables (or independent variables). Recall (see Section 2.2) ☞ 15 that categorical variables take values in a *discrete* number of categories, such as yes/no, green/blue/brown, and male/female. In R, such variables are called **factors**. They often arise in factorial experiments: controlled statistical experiments in which the aim is to assess how a response variable is affected by one or more factors tested at several **levels**. A typical example is an agricultural experiment where one wishes to investigate how the yield of a food crop depends on two factors (1) pesticide, at two levels (yes and no), and (2) fertilizer, at three levels (low, medium, and high). Table 8.1 gives an example of data that is produced in such an experiment. Here three responses (crop yield) are collected from each of the six different combinations of levels.

Table 8.1: Crop yield data

Crop Yield	Pesticide	Fertilizer
3.23	No	Low
3.20	No	Low
3.16	No	Low
2.99	No	Medium
2.85	No	Medium
2.77	No	Medium
5.72	No	High
5.77	No	High
5.62	No	High
6.78	Yes	Low
6.73	Yes	Low
6.79	Yes	Low
9.07	Yes	Medium
9.09	Yes	Medium
8.86	Yes	Medium
8.12	Yes	High
8.04	Yes	High
8.31	Yes	High

Note that the pesticide factor only has two levels. To investigate whether using pesticide is effective (produces increased crop yield) we could simple carry out a 2-sample *t*-test; see Section 7.5. Let us carry out the usual steps for a statistical test here: ☞ 109

1. The model is a two-sample normal model. Let $Y_1, \ldots, Y_9 \sim_{\text{iid}} \mathsf{N}(\mu_1, \sigma^2)$ be the crop yields without pesticide and $Y_{10}, \ldots, Y_{18} \sim_{\text{iid}} \mathsf{N}(\mu_2, \sigma^2)$ be the crop yields with pesticide. Note that we assume equal variances for both groups and all variables are independent of each other.

2. H_0 is the hypothesis that there is no difference between the groups; that is, $\mu_1 = \mu_2$. The alternative hypothesis is that there is a difference: $\mu_1 \neq \mu_2$.

3. ☞ 105 As a test statistic we use the T statistic given in (7.1) (with $\mu_0 = 0$).

4. We find the outcome $t = -7.2993$ (e.g., using `t.test()`)

5. The p-value is $1.783 \cdot 10^{-6}$, which is very small.

6. We therefore fail to accept the null-hypothesis. There is very strong evidence that using pesticide makes a difference.

Note that the above *t*-test does not tell us whether the pesticide was *successful* (that is, gives a higher average yield). Think how you would assess this.

What if we consider instead whether fertilizer "explains" crop yield. For this factor we have three levels: low, medium, and high. So a 2-sample *t*-test does no longer work. Nevertheless, we would like to make a similar analysis as above. Steps 1 and 2 are easily adapted:

1. The model is a 3-sample normal model. Let $Y_1, Y_2, Y_3, Y_{10}, Y_{11}, Y_{12} \sim_{\text{iid}} \mathsf{N}(\mu_1, \sigma^2)$ be the crop yields with low fertilizer, $Y_4, Y_5, Y_6, Y_{13}, Y_{14}, Y_{15} \sim_{\text{iid}} \mathsf{N}(\mu_2, \sigma^2)$ be the crop yields with medium fertilizer, and $Y_7, Y_8, Y_9, Y_{16}, Y_{17}, Y_{18} \sim_{\text{iid}} \mathsf{N}(\mu_3, \sigma^2)$ be the crop yield with high fertilizer. We assume equal variances for all three groups, and that all variables are independent of each other.

2. H_0 is the hypothesis that there is no difference between the groups; that is, $\mu_1 = \mu_2 = \mu_3$. The alternative hypothesis is that there is a difference.

The question is now how to formulate a test statistic (a function of the data) that makes it easy to distinguish between the null and alternative hypothesis. This is where ANOVA comes. It will allow us to compare the means of any number of levels within a factor. Moreover, we will be able to explain the response variable using multiple factors at the same time. For example, how does the crop yield depend on both pesticide and fertilizer.

8.2 Single-Factor ANOVA

8.2.1 Model

Consider a response variable which depends on a single factor with d levels. Within each level i there are n_i independent measurements of the response variable. The data thus consist of d independent samples with sizes $n_1, \ldots, n_d$:

$$\underbrace{Y_1, \ldots, Y_{n_1}}_{\text{level 1}}, \underbrace{Y_{n_1+1}, \ldots, Y_{n_1+n_2}}_{\text{level 2}}, \ldots, \underbrace{Y_{n-n_d+1}, \ldots, Y_n}_{\text{level } d}, \tag{8.1}$$

where $n = n_1 + \cdots + n_d$. An obvious model for the data is that the $\{Y_i\}$ are assumed to be independent and normally distributed with a mean and variance which depend only on the level. Such a model is simply a d-sample generalization of the two-sample normal model in Example 5.8. To be able to analyse the model via ANOVA one needs however the additional model assumption that the *variances are all equal*; that is, they are the same for each level. Writing Y_{ik} as the response for the k-th replication at level i we can define the model as follows. ☞ 78

Definition 8.1 (Single-Factor ANOVA Model). Let Y_{ik} be the response for the k-th replication at level i. Then,

$$Y_{ik} = \mu_i + \varepsilon_{ik}, \quad k = 1, \ldots, n_i, \; i = 1, \ldots, d, \tag{8.2}$$

where $\{\varepsilon_{ik}\} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

Note that (8.2) is simply another way of writing that Y_{ik} has a normal distribution with mean μ_i and variance σ^2.

Instead of (8.2) one often sees the "factor effects" formulation

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik} \ , \quad k = 1, \ldots, n_i \ , \ i = 1, \ldots, d \ , \tag{8.3}$$

where μ is interpreted as the *overall* effect, common to all levels, and $\alpha_i = \mu_i - \mu$ is the *incremental effect* of level i. This model is not identifiable (which means that some parameters cannot be estimated). We therefore have to impose a (linear) constraint to make it identifiable, for example $\sum_{i=1}^d n_i \alpha_i = 0$, which corresponds to taking the overall effect μ as the reference.

8.2.2 Estimation

The model (8.2) has $d + 1$ unknown parameters: $\mu_1, \ldots, \mu_d$, and σ^2. Each μ_i can be estimated exactly as for the 1-sample normal model, by only taking into account the data in level i. In particular, the estimator of μ_i is the sample mean within the i-th level:

$$\widehat{\mu_i} = \overline{Y}_i = \sum_{n_i} \sum_{k=1}^{n_i} Y_{ik}, \quad i = 1, \ldots, d \ .$$

To estimate σ^2, we should utilise the fact that all $\{Y_{ik}\}$ have the same variance σ^2. So, as in the 2-sample normal model case, we should *pool* our data and not just calculate, say the sample variance of the first level only. The model (8.2) assumes that the errors $\varepsilon_{ik} = Y_{ik} - \mu_i$ are independent and normally distributed, with a constant variance σ^2. If we knew the $\{\varepsilon_{ik}\}$ we could just take the sample variance to estimate σ^2. Unfortunately, we do not know the $\{\mu_i\}$. However, we can estimate μ_i with $\overline{Y}_i$. This suggests that we replace the unknown true errors ε_{ik} with the estimated errors — the so-called **residual errors** (or simply residuals) $e_{ik} = Y_{ik} - \overline{Y}_i$. An estimator for σ^2 is therefore

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n-d} = \frac{\sum_{i=1}^d \sum_{k=1}^{n_i} (Y_{ik} - \overline{Y}_i)^2}{n-d} \ . \tag{8.4}$$

The quantities SSE and MSE defined implicitly above are called the **sum of squares of the errors** and **mean square error**. The first is thus the sum of the squared residual errors. The MSE is thus our estimator for σ^2. The reason why we divide by $n - d$ rather than n is that in this way MSE is an unbiased estimator for σ^2.

8.2.3 Hypothesis testing

The typical aim is to test whether the d levels have the same means; that is, to test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_d$$

versus the alternative that this is not the case (at least two different means).

Note that we, in effect, compare two models: Under H_0 we simply have the standard standard 1-sample model for data, and under H_1 we have the single-factor model (8.2) or (8.3). To assess which model is more appropriate, we could compare the variability of the data in the simpler model to the variability of the data in the second, more complex, model. More precisely, we would like to compare the variances σ^2 for both models. Let's call them σ_1^2 and σ_2^2 to distinguish between them. Because the first model is a special case of the second, $\sigma_1^2 > \sigma_2^2$ if H_1 is true, and $\sigma_1^2 = \sigma_2^2$. It therefore makes sense to base our test statistic on estimators of σ_1^2 and σ_2^2. We already saw that σ_2^2 is estimated via the mean square error (MSE). And we can estimate σ_1^2 simply via the sample variance

$$\frac{\text{SST}}{n-1} = \frac{\sum_{i=1}^{d}\sum_{k=1}^{n_i}(Y_{ik} - \overline{Y})^2}{n-1}, \tag{8.5}$$

where $\overline{Y}$ denotes the sample mean of all $\{Y_{ik}\}$ (the overall mean). Like SSE, the quantity SST is a sum of squared terms — the **total sum of squares**.

So, a sensible test statistic could be the fraction SST/SSE, where we would reject H_0 if this fraction becomes too large. Alternatively, one could use any simple function of SST and SSE whose distribution under H_0 can be computed. The actual test statistic that is used in this situation is

$$F = \frac{(\text{SST} - \text{SSE})/(d-1)}{\text{SSE}/(n-d)}, \tag{8.6}$$

where the difference SST − SSE is again a "sum of squares" (see Problem 3):

$$\text{SST} - \text{SSE} = \sum_{i=1}^{d}\sum_{k=1}^{n_i}(\overline{Y}_i - \overline{Y})^2 \,. \tag{8.7}$$

Let us denote this by SSF (Sum of Squares due to the Factor). It measures the variability *between* the different levels of the factor. If we further abbreviate SSF/$(d-1)$ to MSF (mean square factor) and SSE/$(n-d)$ to MSE (mean square error), then we can write our test statistic as

$$F = \frac{\text{MSF}}{\text{MSE}} \,.$$

The test statistic F thus compares the variability *between* levels with the variability *within* the levels. We reject H_0 for large values of F (right one-sided test). To actually carry out the test we need to know the distribution of F under H_0, which is given in the following theorem, the proof of which is beyond a 1-st year course.

Theorem 8.1 Under H_0, $F = \text{MST}/\text{MSE}$ has an $\mathsf{F}(d-1, n-d)$ distribution.

Recall that the parameters of the F distribution are referred to as the degrees of freedom (DF). ☞ 95

8.2.4 ANOVA table

The previous results are summarised in Table 8.2.

Table 8.2: One-factor ANOVA table. f is the outcome of the F statistic.

Source of Variation	DF	SS	Mean Squares	F	$\mathbb{P}[F > f]$
Treatment	$d-1$	SSF	MSF	$\frac{\text{MSF}}{\text{MSE}}$	p-value
Error	$n-d$	SSE	MSE		
Total	$n-1$	SST			

8.2.5 Example and graphical inspection

Five treatments ($T_1, \ldots, T_5$) against cold sore, including one placebo, were randomly assigned to thirty patients (six patients per treatment group). For each patient, the time (in days) between the apparition of the cold sore and complete scarring was measured. The results are given in Table 8.3.

Table 8.3: Cold sore healing times for 5 different treatments. T_1 is a placebo treatment.

T_1	T_2	T_3	T_4	T_5
5	4	6	7	9
8	6	4	4	3
7	6	4	6	5
7	3	5	6	7
10	5	4	3	7
8	6	3	5	6

The aim here is to compare the mean scarring times. The times in the placebo column seem a little higher. But is this due to chance or is there a real difference. To answer this question, let us first load the data into R.

```
> x <- data.frame(Placebo=c(5,8,7,7,10,8),T2=c(4,6,6,3,5,6),
+ T3=c(6,4,4,5,4,3),T4=c(7,4,6,6,3,5),T5=c(9,3,5,7,7,6))
```

The first important point to note is that while Table 8.3 (and the data frame `x`) is a perfectly normal table (and data frame) it is *in the wrong format* for an ANOVA study. Remember (see Chapter 2) that the measurements (the healing times) must be in a single column. In this case we should have a table with only two columns (apart from the index column): one for the response variable (healing time) and one for the factor (treatment). The factor has here 5 levels ($T_1, \ldots, T_5$). An example of a correctly formated table is Table 8.1.

We need to first "stack" the data using the `stack()` function. This creates a new data frame with only two columns: one for the healing times and the other for the factor (at levels $T_1, \ldots, T_5$). The default names for these columns are `values` and `ind`. We rename them to `times` and `treatment`.

```
> sx <- stack(x)
> names(sx)[1]  <- "times"
> names(sx)[2] <- "treatment"
> attach(sx)   # make the names available
```

The second important point is both columns in the reformated data frame `sx` have the correct type (check with `str(sx)`): the response is a qualitative variable (numerical) and the treatment is a categorical variable (factor) at five levels.

We can do a brief descriptive analysis, giving a data summary for the healing times withing each of the factor levels. In R this can be done conveniently via the function `tapply()`

```
> tapply(times,treatment,summary)
```

This applies the `summary()` function to the vector `times`, grouped into `treatment` levels. The output is as follows.

```
$Placebo
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0     7.0     7.5     7.5     8.0    10.0
$T2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00    4.25    5.50    5.00    6.00    6.00
$T3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   4.000   4.333   4.750   6.000
$T4
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.250   5.500   5.167   6.000   7.000
$T5
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   5.250   6.500   6.167   7.000   9.000
```

In particular, the level means (the $\bar{y}_i$) are given in the fourth column.

A plot of `times` versus `treatment` gives more information. Note that the command `plot(times ~ treatment)` automatically gives a box plot as the explanatory variable is a factor.

```
> plot(times ~ treatment)
```

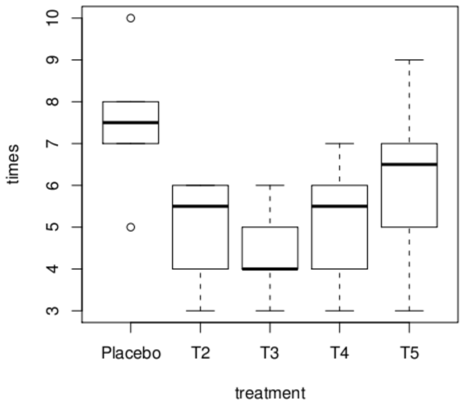


Figure 8.1: Box plot of scarring times for each treatment.

The construction `times ~ treatment` is an example of an `R` **formula**. It is a short-hand notation for the model of the data. In this case, where `times` is numeric and `treatment` a factor, it specifies a 1-factor ANOVA model. The tilde ~ in the formula is simply a symbol that separates the response variable (on the left) from the explanatory variable(s) (on the right). It has nothing to do with our mathematical notation for "is distributed as", as in $X \sim \mathrm{N}(0, 1)$.

8.2.6 ANOVA table

Using a 1-factor ANOVA model, we wish to test the hypothesis H_0 that all treatment levels have the same means versus the alternative that this is not the case. Our test statistic is $F = \text{MSL}/\text{MSE}$, which, if H_0 is true, we know has an F distribution; see Theorem 8.1. In this case $d = 5$ and $n = 30$, so F has an $\mathsf{F}(4, 25)$ distribution under H_0. The next step is to evaluate the outcome f of F based on the observed data, and then to calculate the p-value. Since we have a right one- sided test (we reject H_0 for large values of F), the p-value is $\mathbb{P}(F > f)$, where $F \sim \mathsf{F}(4, 25)$. Fortunately, `R` can do all these calculations for us, using for instance the function `aov()`. All we need to do is specify the `R` formula.

```
> my.aov <- aov(times~treatment)
> summary(my.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    4 36.467  9.1167   3.896 0.01359 *
Residuals   25 58.500  2.3400
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The values listed are the parameters (degrees of freedom, DF) for the F distribution (4 and 25), the sum of squares of the treatment SSF = 36.467 and the residuals SSE = 58.500, the corresponding mean squares MSL = 9.1167 and MSE = 2.3400 and, finally, the outcome of the test statistic $f = 3.896$, with corresponding p-value 0.01359, which is quite small. There is thus fairly strong evidence to believe that the treatments have an effect.

8.2.7 Validation of assumptions

The ANOVA model (8.2) assumes that the errors $\{\varepsilon_{ik}\}$ are independent and normally distributed with a constant variance. We can verify these assumption by investigating the residuals. If the model is correct, the residuals should be approximately be independent and normally distributed.

The assumptions of the model can be inspected graphically using the following commands.

```
> plot(my.aov)
```

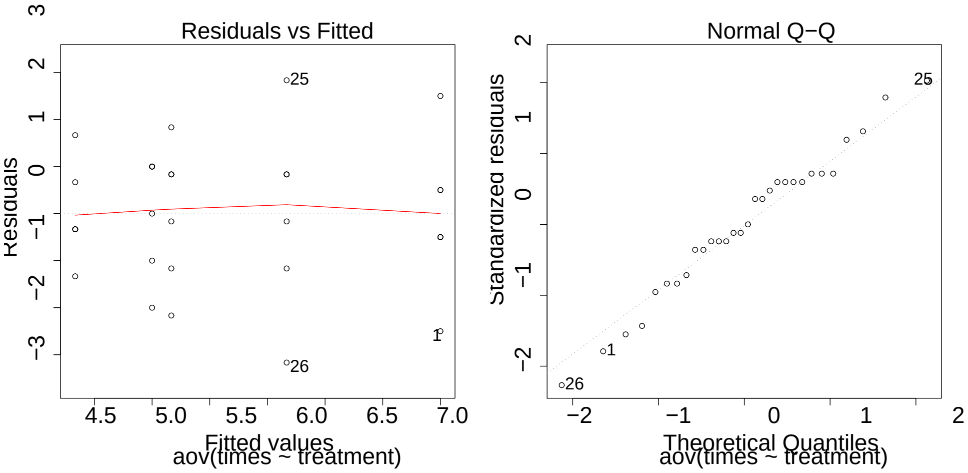


Figure 8.2: Analysing the residuals in single-factor ANOVA

R actually returns four diagnostic plots, but we have listed only two in Figure 8.2. Examining the residuals as a function of predicted values, the residuals are correctly spread, symmetrical about the x axis: the conditions of the model (i.e. independance and constant variance) seem valid. The normally of the residuals is indicated by the observed straight line in the QQ-plot representation.

8.2.8 Summary

The next table presents the main functions for single factor anaysis of variance.

Table 8.4: Basic functions for single-factor ANOVA

R instruction	Description
`plot(y~x)`	graphical inspection
`aov(y~x)`	analysis of variance
`summary(aov(y~x))`	analysis of variance table
`plot(aov(y~x))`	graphical analysis of residuals

8.3 Two-factor ANOVA

Many designed experiments deal with responses that depend on more than one factor. Think of the crop-yield data in Table 8.1. Here we have two factors (fertilizer and pesticide). We wish to investigate if either (or both) of them have any effect on the crop yield.

8.3.1 Model

Consider a response variable with depends on two factors. Suppose Factor 1 has d_1 levels and Factor 2 has d_2 levels. Within each pair of levels (i, j) we assume that there are n_{ij} replications. Let Y_{ijk} be the k-th observation at level (i, j). A direct generalization of (8.2) gives the following model.

Definition 8.2 (Two-factor ANOVA Model). Let Y_{ijk} be the response for the k-th replication at level (i, j). Then,

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}\ , \quad k = 1, \ldots, n_{ij}\ ,\ i = 1, \ldots, d_1, j = 1, \ldots, d_2\ , \tag{8.8}$$

where $\{\varepsilon_{ijk}\} \sim_{\text{iid}} \mathsf{N}(0, \sigma^2)$.

Note that the variances of the responses are assumed to be equal σ^2. To obtain a "factor effects" representation, we can reparameterise model (8.8) as follows:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}\ , \\ &\qquad k = 1, \ldots, n_{ij}\ ,\ i = 1, \ldots, d_1\ , j = 1, \ldots, d_2\ . \end{aligned} \tag{8.9}$$

The parameter μ can be interpreted as the overall mean response, α_i as the incremental effect due to Factor 1 at level i, and β_j as the incremental effect of Factor 2 at level j. The $\{\gamma_{ij}\}$ represent the interaction effects of the two factors.

8.3.2 Estimation

For the model (8.8), a natural estimator of μ_{ij} is the sample mean of all the responses at level i of Factor 1 and level j of Factor 2; that is,

$$\widehat{\mu}_{ij} = \overline{Y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}\ .$$

For the factor effects representation (8.9) the parameters can be defined in several ways. For the most important *balanced* case (all the n_{ij} are the same), the default choice for the parameters is as follows.

$$\mu = \frac{\sum_i \sum_j \mu_{ij}}{d_1 d_2}\ . \tag{8.10}$$

$$\alpha_i = \frac{\sum_j \mu_{ij}}{d_2} - \mu\ . \tag{8.11}$$

$$\beta_j = \frac{\sum_i \mu_{ij}}{d_1} - \mu\ . \tag{8.12}$$

$$\gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j\ . \tag{8.13}$$

For this case it is easy to see that $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ for all i and j. Note that under these restrictions model (8.9) has the same number of "free" parameters as model (8.8).

To estimate the overall effect μ we can simply take the overall mean $\overline{Y}$ of all the $\{Y_{ijk}\}$. The incremental effect α_i can be estimated via $\overline{Y}_{i\bullet} - \overline{Y}$, where $\overline{Y}_{i\bullet}$ is the average of all the $\{Y_{ijk}\}$ within level i of Factor 1. Similarly, β_j can be estimated via $\overline{Y}_{\bullet j} - \overline{Y}$, where $\overline{Y}_{\bullet j}$ is the average of all the $\{Y_{ijk}\}$ within level j of Factor 2. The estimation of the γ_{ij} is left as an exercise.

To estimate σ^2 we can reason similarly to the 1-factor case and consider the residuals $e_{ijk} = Y_{ijk} - \overline{Y}_{ij}$ as our best guess of the true model errors $\varepsilon_{ijk} = Y_{ijk} - \mu_{ij}$ for all
☞ 123 i,j, and k. Similar to (8.5) we have the unbiased estimator

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n - d_1\, d_2} = \frac{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y})^2}{n - d_1\, d_2} \; .$$

8.3.3 Hypothesis testing

The aim here is to detect

- whether Factor 1 has an effect on the response variable;
- whether Factor 2 has an effect on the response variable;
- and whether there is an interaction effect between Factors 1 and 2 on the response variable.

Following the usual steps for hypothesis testing, we need to formulate the questions above in terms of hypotheses on the model parameters. Let us take the model formulation (8.3). Remember that the null hypothesis should contain the "conservative" statement and the alternative hypothesis contains the statement that we wish to demonstrate. So, whether Factor 1 has an effect can be assessed by testing

$$H_0 : \alpha_i = 0 \quad \text{for all } i,$$

versus H_1: at least one α_i is not zero.

Similarly, we can assess the effectiveness of Factor 2 by testing

$$H_0 : \beta_j = 0 \quad \text{for all } j,$$

versus H_1: at least one β_j is not zero.

Finally, we can test for interaction by considering the hypothesis

$$H_0 : \gamma_{ij} = 0 \quad \text{for all } i, j,$$

versus H_1: at least one of the γ_{ij} is not zero.

Similar to the 1-factor ANOVA case we can again decompose the total sum of squares SST $= \sum_{i,j,k}(Y_{ijk} - \overline{Y})^2$ into the sum

$$\text{SST} = \text{SSF1} + \text{SSF2} + \text{SSF12} + \text{SSE},$$

where SSF1 measures the variability between the levels of Factor 1, SSF2 measures the variability between the levels of Factor 2, SSF12 measures the variability due to interaction between the factors, and SSE measures the variability within the levels.

As in the 1-factor ANOVA case, the test statistics for the above hypotheses are quotients of the corresponding mean square errors, and have an F distribution with a certain number of degrees of freedom.

8.3.4 ANOVA table

The various quantities of interest in an ANOVA table are summarised in Table 8.5.

Table 8.5: Two-factor ANOVA table. f is the outcome of the F statistic.

Source of Variation	DF	SS	Mean Squares	F	$\mathbb{P}[F > f]$
Factor 1	$d_1 - 1$	SSF1	MSF1	$\frac{\text{MSF1}}{\text{MSE}}$	p-value
Factor 2	$d_2 - d$	SSF2	MSF2	$\frac{\text{MSF1}}{\text{MSE}}$	p-value
Interaction	$(d_1 - 1)(d_2 - 1)$	SSF12	MSF12	$\frac{\text{MSF12}}{\text{MSE}}$	p-value
Error	$n - d_1 d_2$	SSE	MSE		
Total	$n - 1$	SST			

8.3.5 Example and graphical inspection

Consider the data in Table 8.6, representing the crop yield using four different crop treatments (e.g., strengths of fertilizer) on four different regions.

Table 8.6: Crop yield.

	Treatment		
Region	1	2	3
1	9.18, 8.26, 8.57	9.69, 8.25, 9.83	7.87, 8.91, 7.78
2	10.05, 8.92, 9.39	9.80, 10.90, 10.75	8.33, 8.18, 9.78
3	11.23, 11.11, 9.72	12.13, 12.01, 9.67	9.38, 10.10, 10.90
4	11.60, 9.83, 11.07	12.09, 10.15, 12.04	11.73, 8.86, 11.23

These data can be entered into R using the following script:

```
yield <- c(9.18, 8.26, 8.57, 10.05, 8.92, 9.39, 11.23, 11.11,
    9.72, 11.60, 9.83, 11.07, 9.69, 8.25, 9.83, 9.80, 10.90,
    10.75, 12.13, 12.01, 9.67, 12.09, 10.15, 12.04, 7.87,
    8.91, 7.78, 8.33, 8.18, 9.78, 9.38, 10.10, 10.90, 11.73,
    8.86, 11.23)
fertilizer <- gl(3,12,36,labels=paste("Fertilizer",1:3))
region <- gl(4,3,36,labels=paste("Region",1:4))
wheat <- data.frame(yield,fertilizer,region)
attach(wheat)
```

Here the function gl() generates factors by specifying the pattern of their levels.

We wish to study the effect of the type of fertilizer on the yield of the crop and whether there is a significantly different yield between the four regions. There could also be an interaction effect; for example, if a certain treatment works better in a specific region.

```
> interaction.plot(region,fertilizer,yield)     # use this
> interaction.plot(fertilizer,region,yield)     # or this
```

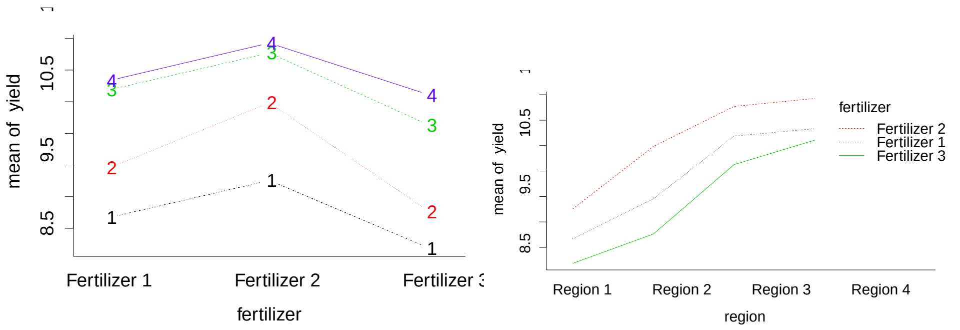


Figure 8.3: Exploration of interaction in two-way ANOVA.

These plots contain a lot of information. For example, the left figure makes it easier to investigate the Fertilizer effect. We can observe that the mean yield is always better with Fertilizer 2, whatever the region. A graph with horizontal lines would indicate no effect of the Fertilizer factor. The figure on the right may indicate an effect of the Region factor, as we can observe an increase of the mean yield from Region 1 to Region 4, whatever the Fertilizer used.

If there is no interaction between the two factors, the effect of one factor on the response variable is the same irrespective of the level of the second factor. This corresponds to observing parallel curves on both plots in figure (8.3). Indeed, the differences of the black dotted curve (Region 1) and the red dotted curve (Region 2) in the left plot represent the differential effects of the Region 2 versus Region 1 for each Fertilizer. If there is no interaction, these differences should be the same (i.e., parallel curves). Both plots in Figure 8.3 might indicate an absence of interaction as we can observe parallel curves. We will confirm it by testing the interaction effect in the next sub-section.

Tip

We plotted the two interaction plots in two different ways. Type `?interaction.plot` and `?par` to find out about the possible plotting parameters.

8.3.6 ANOVA table

Similarly of the 1-factor ANOVA case, the R function `aov()` provides the ANOVA table:

```
> summr <- summary(aov(yield~region*fertilizer))
> summr
         fertilizer region yield

                   Df  Sum Sq Mean Sq F value    Pr(>F)
region              3 29.3402  9.7801 11.2388 8.451e-05 ***
fertilizer          2  8.5596  4.2798  4.9182   0.01622 *
region:fertilizer   6  0.6954  0.1159  0.1332   0.99067
Residuals          24 20.8849  0.8702
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note

The formula `region*fertilizer`, used in `aov()`, corresponds in fact to the formula `region+fertilizer+region:fertilizer`, *i.e.* the factor region, the factor fertilizer, and the interaction of these two factors.

The p-value associated with the test of interaction is not significant (p-value= 0.99). This implies that the effect of fertilizer of yield is the same whatever the region. In this case, we perform an ANOVA without an interaction term which makes it easier to interpret the principal effect. The corresponding additive model is defined by:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \ . \tag{8.14}$$

```
> summr2 <- summary(aov(yield~region+fertilizer))
> summr2

            Df  Sum Sq Mean Sq F value    Pr(>F)
region       3 29.3402  9.7801 13.5959 8.883e-06 ***
fertilizer   2  8.5596  4.2798  5.9496  0.006664 **
Residuals   30 21.5802  0.7193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both p-values are significant which indicate a significant effect of region and fertilizer on crop yield.

Warning

When you have only one observation per combination of levels of the factors A and B (i.e., $n_{ij} = 1$ for all i, j), you can only estimate two-way ANOVA without interaction: `aov(yield~region+fertilizer)`.

Note that when there is interaction, we do not interpret the principal effects in the ANOVA table output. Suppose we found in our example an significant effect of the interaction term. This implies that the effect of fertilizer of yield can be different depending on the region. For example, we wish to know whether there is a fertilizer effect in region 1. To this end, we use the function `subset()`, which only uses data from a given region.

```
> fertilizer.region1 <- summary(aov(yield~fertilizer
+                             ,subset=region=="Region 1"))
> fertilizer.region1

            Df Sum Sq Mean Sq F value Pr(>F)
fertilizer   2  1.723  0.8613   1.875  0.233
Residuals    6  2.757  0.4595
```

Warning

The test in this ANOVA table corresponds to ANOVA with one factor (fertilizer) of the yield of wheat in region 1. It does not take into account any information from data in the other regions, which would allow for a better estimation of the residual variance. To test the fertilizer effect in region 1, divide the mean square of the fertilizer factor found in the ANOVA restricted to region 1 by the mean residual square of the ANOVA with interaction:

```
> F.fertilizer.region1 <- fertilizer.region1[[1]]$Mean[1]/
+       summr[[1]]$Mean[4]
> pvalue <- 1-pf(F.fertilizer.region1,df1=2,df2=24)
> pvalue
```

```
0.3863111
```

8.3.7 Validation of assumptions

As in one-way ANOVA, we validate the model with a study of the residuals of the underlying linear model.

```
> plot(my.aov)
```

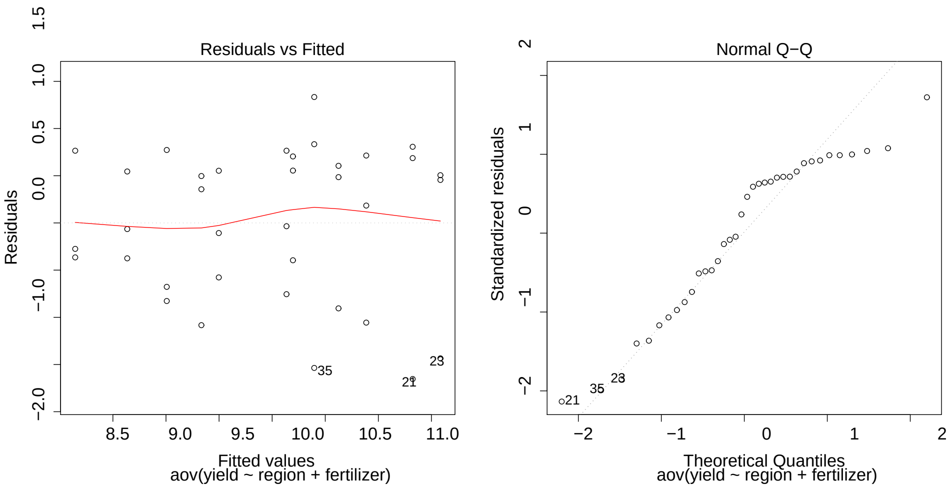


Figure 8.4: Residual analysis in two-way ANOVA.

However, if the data size is large enough for each pair of factor levels, it is better to check for normality in each subpopulation and for homoscedasticity.

8.3.8 Summary

This table lists the main functions for two-way ANOVA.

Table 8.7: Basic functions for two-way ANOVA.

`R` **instruction**	**Description**
`interaction.plot(x,z,y))`	graphical inspection
`aov(y~x*z))`	two-way ANOVA with interaction
`aov(y~x+z))`	two-way ANOVA without interaction
`summary(aov(y~x*z)))`	ANOVA table
`plot(aov(y~x*z)))`	residuals plots

Conclusion

1. ANOVA breaks down the total variability in a response into the residual error variability and the variability that can be explained by the factors.
2. One-factor ANOVA breaks down the total variability into the error variability within groups and the variability between groups.
3. Each variability component is summarised by a sum of squared deviations and a degrees of freedom
4. Two-factor ANOVA partitions the total variability in a response due to two categorical factors.
5. Two-factor ANOVA can help detect an interaction effect between two factors on the response.
6. The assumptions for ANOVA are homoscedasticity, normality and independence of the errors.

8.4 Problems

1. For each of the following situations, formulate a regression or ANOVA model.

   a. We wish to test if three different brands of compact cars have the same average fuel consumption. The fuel consumption for a traveled distance of 100km is measured for twenty cars of each brand.

   b. Heart rates were monitored for 20 laboratory rats during three different stages of sleep.

   c. We investigate the effectiveness of a new fertilizer, by dividing a large patch of land into 20 test plots, each of which is divided into 3 small sub-plots. In each of the 3 subplots a different concentration of fertilizer is tested: *weak, moderate*, and *strong*. The product yield for each subplot is recorded.

2. Analyse Table 8.1.

3. Show the sum of squares decomposition (8.7). Namely, write $\mathrm{SS}_{\mathrm{total}}$ as

$$\sum_{i=1}^{d}\sum_{k=1}^{n_i}\left[(\bar{Y}_i - \bar{Y}) + (Y_{ik} - \bar{Y}_i)\right]^2 , \tag{8.15}$$

and note that sum of the "cross product" vanishes, i.e.,

$$\sum_{i=1}^{d}\sum_{k=1}^{n_i}(\bar{Y}_i - \bar{Y})(Y_{ik} - \bar{Y}_i) = \sum_{i=1}^{d}(\bar{Y}_i - \bar{Y})\sum_{k=1}^{n_i}(Y_{ik} - \bar{Y}_i) = 0 .$$

4. In order to investigate the effectiveness of "walking exercises" for babies, 24 babies (of the same age and sex) were were randomly divided into 4 groups. Each group followed a different training program. The table shows the age (in months) when the infants first walked alone.

	Group		
A	B	C	D
9	11	11.5	13.25
9.5	10	12	11.5
9.75	10	9	12
10	11.75	11.5	13.5
13	10.5	13.25	11.5
9.5	15	13	11.5

A one-way analysis gives:

```
One-way ANOVA: A, B, C, D

Analysis of Variance
Source     DF       SS         MS        F       P
Factor     a      14.20        c         e       f
Error      20       b          d
Total      23     58.49
```

A and B underwent the same training program, but the progress of the children in group A was checked every week, whereas the children in group B were checked only once, at the end of the study.

(a) Evaluate the missing numbers a, b, c, d, e and f from the table. For f you may give an appropriate upper bound.

(b) Test whether there is a psychological effect of periodic testing by considering the contrast $C = \mu_A - \mu_B$. It is given that $SS_C = 1.12$.

5. An experiment was performed to determine the effects of four different chemicals on the strength of fabric. Five fabric samples were selected and each chemical type was tested once on each fabric sample. The data are shown below.

	Fabric Sample				
Chemical type	1	2	3	4	5
1	1.3	1.6	0.5	1.2	1.1
2	2.2	2.4	0.4	2.0	1.8
3	1.8	1.7	0.6	1.5	1.3
4	3.9	4.4	2.0	4.1	3.4

A two-way Analysis of Variance gave the following output:

```
Two-way ANOVA: strength versus fabric, chemical


Analysis of Variance for strength
Source        DF        SS         MS         F        P
fabric         a    6.6930          d         e        g
chemical       b   18.0440     6.0147         f        h
Error          c    0.9510     0.0792
Total         19   25.6880
```

(a) Formulate an appropriate statistical model for the data.

(b) Fill in the missing numbers a, . . . ,h. For the p-values you may use upper bounds (p-value < . . .).

(c) Is there an effect of the chemical on the fabric strength? State the appropriate hypotheses, give the test statistic, its degrees of freedom and p-value and state your conclusion.

(d) Examine the residuals and check for violations of basic assumptions that could invalidate the results.

6. Six different types of paint are tried on five different surfaces. For each combination of paint and surface we measure the quality of coverage. The (partial) results of a statistical analysis are depicted in the table below. There is no replication, thus we have 30 measurements in total. A two-way Analysis of Variance gave the following output:

```
Two-way ANOVA: coverage versus painttype, surface

Analysis of Variance for coverage
Source        DF        SS         MS         F        P
painttype             519.07
surface               166.87
Error                 383.93
Total
```

(a) Complete the ANOVA table

(b) Test whether there is any difference between the paint types with regard to the coverage quality.

(c) Is “surface type” an relevant factor in explaining the coverage quality?

Chapter 9

Regression

This chapter is a brief introduction to simple and multiple linear regression. We present how to get some confidence and prediction intervals for a new observations. We discuss model validation with a study of residuals.

9.1 Introduction

Francis Galton observed in an article in 1889 that the heights of adult offspring are, on the whole, more "average" than the heights of their parents. Galton interpreted this as a degenerative phenomenon, using the term *regression* to indicate this "return to mediocrity". Karl Pearson continued Galton's original work and conducted comprehensive studies comparing various relationships between members of the same family. Figure 9.1 depicts the measurements of the heights of 1078 fathers and their adult sons (one son per father).

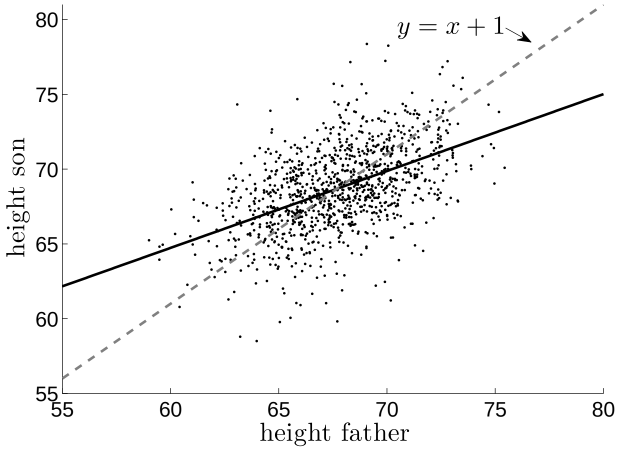


Figure 9.1: A scatter plot of heights from Pearson's data.

The average height of the fathers was 67 inches, and of the sons 68 inches. Because sons are on average 1 inch taller than the fathers we could try to "explain" the height of the son by taking the height of his father and adding 1 inch. However, the line $y = x + 1$ (dashed) does not seems to predict the height of the sons as accurately as the solid line in Figure 9.1. This line has a slope less than 1, and demonstrates Galton's "regression" effect. For example, if a father is 5% taller than average, then his son will be on the whole *less* than 5% taller than average.

Similar to ANOVA in Chapter 8, regression analysis is about finding relationships between a *response* variable which we would like to "explain" via one or more *explanatory* variables. However, whereas the response variable in ANOVA were qualitative (categorical, factors), in regression they are *quantitative* (numerical).

9.2 Simple Linear Regression

The most basic regression model involves a linear relationship between the response and a single explanatory variable. As in Pearson's height data, we have measurements $(x_1, y_1), \ldots, (x_n, y_n)$ that lie approximately on a straight line. It is assumed that these measurements are outcomes of vectors $(x_1, Y_1), \ldots, (x_n, Y_n)$, where, for each *deterministic* explanatory variable x_i, the response variable Y_i is a *random* variable with

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1\, x_i, \quad i = 1, \ldots, n \tag{9.1}$$

for certain *unknown* parameters β_0 and β_1. The (unknown) line

$$y = \beta_0 + \beta_1\, x \tag{9.2}$$

is called the **regression line**. To completely specify the model, we need to designate the joint distribution of $Y_1, \ldots, Y_n$. The most common linear regression model is given next. The adjective "simple" refers to the fact that a *single* explanatory variable is used to explain the response.

Definition 9.1 (Simple Linear Regression Model). The response data $Y_1, \ldots, Y_n$ depend on explanatory variables $x_1, \ldots, x_n$ via the linear relationship

$$Y_i = \beta_0 + \beta_1\, x_i + \varepsilon_i\,, \quad i = 1, \ldots, n\,, \tag{9.3}$$

where $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

This formulation makes it even more obvious that we view the responses as random variables which would lie exactly on the regression line, were it not for some "disturbance" or "error" term (represented by the $\{\varepsilon_i\}$).

To make things more concrete let us consider the student survey dataset `ssurv` which we can be found on the website. Suppose we wish to investigate the relation between the shoe size (explanatory variable) and the height (response variable) of a person.

First we load the data:

```
> rm(list=ls())   # good practice to clear the workspace
> ssurv <- read.csv("survey-summer.csv")
> attach(ssurv)
```

In the notation of Definition 9.1, x_i denotes the i-th shoe size in cm (stored in `shoe`) and y_i denotes the corresponding height in cm (stored in `height`). For the pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, we assume model (9.3). Note that the model has three unknown parameters: β_0, β_1, and σ^2. What can we say about the model parameters on the basis of the observed data $(x_1, y_1), \ldots, (x_n, y_n)$?

A first step in the analysis is to draw a scatterplot of the points (height versus shoe size) using the instruction `plot(height~shoe)`. As in the ANOVA chapter, `height~shoesize` is an `R` formula which indicates that `height` is the response variable and `shoe` an explanatory variable. Notice, however, that the `plot()` function interprets this relation differently in the regression case, because now `shoe` is numerical. As a result a scatterplot is returned — not a boxplot.

```
> plot(height~shoe,xlab="Height",ylab="Shoe size")
```

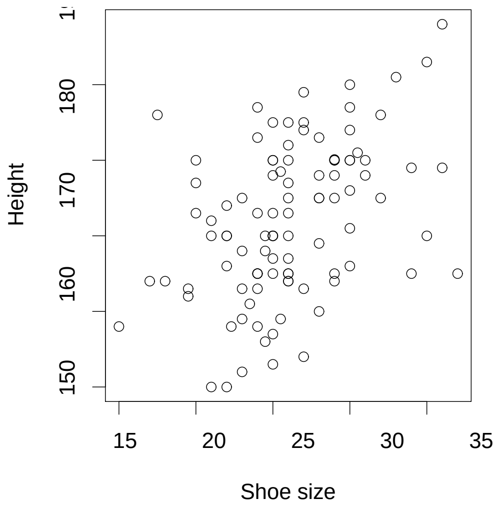


Figure 9.2: Scatter plot of height (in cm) against shoe size (in cm).

We observe a slight increase in the height as the shoe size increases, although this relationship is not very clear.

9.2.1 Estimation

Obviously we do not know the true regression line $y = \beta_0 + \beta_1 x$, but we can try to fit a line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ that best "fits" the data. Here $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimates for the unknown intercept β_0 and slope β_1. Note that by substituting x_i for x, we find that the corresponding y-value is $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. For each i, the difference $e_i = y_i - \widehat{y}_i$ is called a **residual error**, or simply **residual**. There are various measures for "best fit", but a very convenient one is minimise the sum of the squared residual errors, $\text{SSE} = \sum_{i=1}^{n} e_i^2$. This gives the following *least-squares* criterion:

$$\text{minimise SSE} . \tag{9.4}$$

The solution is given in the next theorem.

Theorem 9.1 The values for $\widehat{\beta}_1$ and $\widehat{\beta}_0$ that minimise the least-squares criterion (9.4) are:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{9.5}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}\,. \tag{9.6}$$

Proof: We seek to minimise the function $g(a, b) = \text{SSE} = \sum_{i=1}^n (y_i - a - bx_i)^2$ with respect to a and b. To find the optimal a and b, we take the derivative of SSE with respect to a, b and set it equal to 0. This leads to two linear equations:

$$\frac{\partial \sum_{i=1}^n (y_i - a - bx_i)^2}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

and

$$\frac{\partial \sum_{i=1}^n (y_i - a - bx_i)^2}{\partial b} = -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = 0\,.$$

From the first equation, we find $\overline{y} - a - b\overline{x} = 0$ and then $a = \overline{y} - b\overline{x}$. We put this expression for a in the second equation and get (omitting the factor -2):

$$\begin{aligned} \sum_{i=1}^n x_i (y_i - a - bx_i) &= \sum_{i=1}^n x_i \,(y_i - \overline{y} + b\overline{x} - bx_i) \\ &= \sum_{i=1}^n x_i y_i - \overline{y} \sum_{i=1}^n x_i + b \left(n\overline{x}^2 - \sum_{i=1}^n x_i^2 \right). \end{aligned}$$

Since this expression has to be 0, we can solve for b to obtain

$$b = \frac{\sum_{i=1}^n x_i y_i - n\overline{xy}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}\,.$$

Replacing a with $\widehat{\beta}_0$ and b with $\widehat{\beta}_1$, we have completed the proof.

If we replace in (9.5) and (9.6) the values y_i and $\overline{y}$ with the *random variables* Y_i and $\overline{Y}$, then we obtain the *estimators* of β_1 and β_0. Think of these as the parameters for the line of best fit that we would obtain if we would carry out the experiment *tomorrow*.

Warning

When dealing with parameters from the Greek alphabet, such as β, it is customary in the statistics literature —and you might better get used to it— to use the *same* notation (Greek letter) for the estimate and the corresponding estimator, both indicated by the “hat” notation: $\widehat{\beta}$. Whether $\widehat{\beta}$ is to be interpreted as random (estimator) or fixed (estimate) should be clear from the context.

Since both estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear combinations of normal random variables, their distributions are again normal. Moreover, it is not too difficult to calculate the corresponding expectations and variances. These are summarised in the next theorem.

Theorem 9.2 Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have a normal distribution. Their expected values are

$$\mathbb{E}(\widehat{\beta}_0) = \beta_0 \quad \text{and} \quad \mathbb{E}(\widehat{\beta}_1) = \beta_1 \ , \tag{9.7}$$

so both are *unbiased* estimators. Their variances are

$$\mathrm{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \tag{9.8}$$

and

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \ . \tag{9.9}$$

To estimate the unknown σ^2, we can reason in a similar way as we did for ANOVA models. For each x_i, σ^2 is the variance of the true error

$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$, where the $\{\varepsilon_i\}$ are iid with a $\mathsf{N}(0, \sigma^2)$ distribution. So, if we knew the true errors $\{\varepsilon_i\}$, we could estimate σ^2 via their sample variance, which is $\sum_{i=1}^n \varepsilon_i^2/(n-1)$. Unfortunately, we do not know the true errors, because the parameters β_0 and β_1 are unknown. However, we could replace the true error ε_i with the residual error $e_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$. Our estimator for σ^2 is then

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{1}{n-2} \sum e_i^2 \ . \tag{9.10}$$

Compare this with the single-factor ANOVA model in (8.4).) It can be shown that by dividing by $n-2$ rather than $n-1$ we obtain an unbiased estimator of σ^2. ☞ 122

9.2.2 Hypothesis testing

It is of interest to test whether the slope is 0. If this is the case then there is no association between the response and the explanatory variable. A standard case is to consider $H_0 : \beta_1 = 0$ and see if there is evidence against it. There are two approaches that we could use to construct a good test statistic.

The first approach is to utilise the fact that, by Theorem 9.2 the estimator $\widehat{\beta}_1$ has a normal distribution with expectation 0 and variance $\sigma^2 / \sum_{i=1}^n (x_i - \overline{x})^2$, under H_0. Hence, similar to the construction of the test statistic for the 1-sample normal model, we could use the test statistic

$$T = \frac{\widehat{\beta}_1 \sqrt{\sum (x_i - \overline{x})^2}}{\sqrt{\mathrm{MSE}}} \ . \tag{9.11}$$

It can be shown that under H_0, T has a Student's t distribution with $n-2$ degrees of freedom. A similar test statistic can be used to test whether β_0 is 0, but this is less relevant.

☞ 122

An alternative approach is to adopt an analysis of variance, similar to the 1-factor ANOVA in Section 8.2.3. Here we wish to compare two models: one with $\beta_1 = 0$ and the other with $\beta_1 \neq 0$. The first model is simply the "standard" 1-sample normal model. We can estimate the variance via the usual sample variance of the $\{Y_i\}$:

$$S^2 = \frac{\text{SST}}{n-1} = \frac{\sum(Y_i - \overline{Y})^2}{n-1},$$

where SST denotes the "total sum of squares", just as in the ANOVA case. For the second model we already estimated the variance via the MSE in (9.10). This suggest, by analogy with (8.6), the test statistic

$$F = \frac{\text{SSL}}{\text{SSE}/(n-2)}, \tag{9.12}$$

where SSL = SST − SSE. We reject H_0 for large values of F. It can be shown that under H_0, F has an $\mathsf{F}(1, n-2)$ distribution.

A useful quantity is the **coefficient of determination**:

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}\ .$$

It measures the fraction of the variance in the simpler model that is explained by using the more complicated model. The closer R^2 is to 1, the better the advanced model explains the data.

9.2.3 Using the computer

The relevant `R` function to do linear regression is `lm()` (abbreviation of *linear model*). The main parameter of this function is our usual `R` formula — in this case `height~shoe`.

```
> model1 <- lm(height ~ shoe) # We get an object of class "lm".
> model1

Call:
lm(formula = height ~ shoe)

Coefficients:
(Intercept)         shoe
    145.778        1.005
```

The above `R` output gives the least squares estimates of β_0 and β_1. For the above example, we get $\widehat{\beta_0} = 145.778$ and $\widehat{\beta_1} = 1.005$.

We can now draw the regression line on the scatter plot, using the function `abline()`:

```
> plot(height ~ shoe,ylab="Height",xlab="Shoe size")
> abline(model1,col='blue')
```

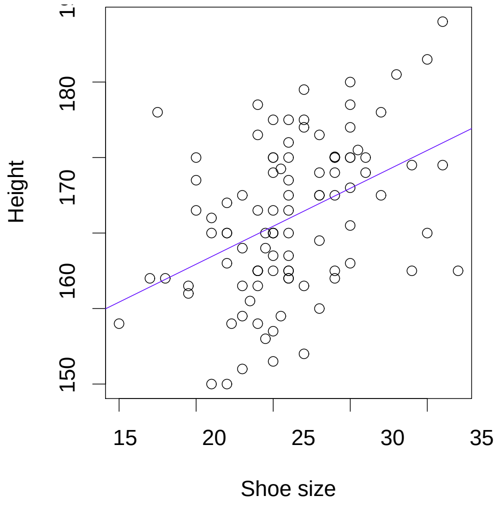


Figure 9.3: Scatter plot of height (in cm) against shoe size (in cm), with the fitted line.

The function `lm()` performs a complete analysis of the linear model and that you can get a summary of the calculations related to the data set with the function `summary()`.

```
> sumr1 <- summary(model1)
> sumr1

Call:
lm(formula = height ~ shoe)

Residuals:
     Min       1Q   Median       3Q      Max
-18.9073  -6.1465   0.1096   6.3626  22.6384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.7776     5.7629  25.296  < 2e-16 ***
shoe          1.0048     0.2178   4.613  1.2e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.299 on 98 degrees of freedom
Multiple R-squared: 0.1784,        Adjusted R-squared:  0.17
F-statistic: 21.28 on 1 and 98 DF,  p-value: 1.199e-05
```

Here is a description of the information in this output.

- `Call:` formula used in the model.
- `Residuals:` summary information for the residuals $e_i = y_i - \widehat{y_i}$.
- `Coefficients:` this table has four columns:
  - `Estimate` gives the estimates of the parameters of the regression line;

  - `Std. Error` gives the estimate of the standard deviation of the estimators of the regression line. These are the square roots of the variances in (9.8) and (9.9);
  - `t value` gives the realization of Student's test statistic associated with the hypotheses $H_0 : \beta_i = 0$ and $H_1 : \beta_i \neq 0$, $i = 0, 1$. In particular, the t-value for the slope corresponds to the outcome of T in (9.11);
  - `Pr(>|t|)` gives the p-value of Student's test (two-sided test).

- `Signif. codes:` codes for significance levels.
- `Residual standard error:` the estimate $\sqrt{\text{MSE}}$ of σ, and the associated degree of freedom $n - 2$.
- `Multiple R-Squared:` coefficient of determination R^2 (percentage of variation explained by the regression).
- `Adjusted R-Squared:` adjusted R^2 (of limited interest for simple linear regression).
- `F-Statistic:` realization of the F test statistic (9.12) associated with the hypotheses $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$. The associated degrees of freedom (1 and $n - 2$) are given, as is the p-value.

- The test associated with the intercept β_0 of the model is significant (p-value <0.05), it is therefore advised to keep the intercept (β_0) in the model. However, the intercept in this regression has no meaning. It might be better to use a regression on the variable shoe size, centred beforehand. In that case, β_0 would represent the mean height for students who have shoe size equal to the mean shoe size of observed students.

Tip

The instruction to perform linear regression without an intercept is `lm(y~x-1)` or equivalently `lm(y~0+x)`.

- The linear relationship between `height` and `shoe` is proven by the result of Student's test on coefficient β_1. The p-value<0.001 indicates a significant linear relationship between shoe size and height.
- The percentage of variance explained by the regression (r^2) is 0.1784. Only 17.8 % of the variability of child height is explained by shoe size. We therefore need to add to the model other explanatory variables (multiple linear regression), to increase the model's predictive power.
- The estimate of the slope indicates that the difference between the average height of students whose shoe size is different by one cm is 1.0048 cm.

You can access all the numerical values from the summary object directly. First check which names are available

```
> names(sumr1)
```

```
[1] "call"          "terms"       "residuals"    "coefficients"
[5] "aliased"       "sigma"       "df"           "r.squared"
[9] "adj.r.squared" "fstatistic" "cov.unscaled"
```

Then access the values via via the dollar ($) construction. For example, this extracts the p-value for the slope.

```
> sumr1$coefficients[2,4]
```

```
[1] 1.1994e-05
```

You can also do an analysis of variance either on the model or on the R formula. Use the function summary() to display the main results.

```
> summary(aov(model1))  # or summary(aov(height~shoe))
```

```
summary(aov(height~shoe))
            Df Sum Sq Mean Sq F value    Pr(>F)
shoe         1 1840.5 1840.47  21.284 1.199e-05 ***
Residuals   98 8474.4   86.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For example, SSE = 8474.4. Confidence intervals for β_0 and β_1 can be obtained via the R function confint().

```
> confint(model1)
```

```
                  2.5 %    97.5 %
(Intercept) 134.3412266 157.21391
shoe          0.5725866   1.43702
```

9.2.4 Confidence and prediction intervals for a new value

Linear regression is most useful when we wish to *predict* how a new response variable will behave, on the basis of a new explanatory variable x. For example, it may be difficult to measure the response variable, but by knowing the estimated regression line and the value for x, we will have a reasonably good idea what Y or the expected value of Y is going to be.

Thus, consider a new x and assume $Y \sim \mathsf{N}(\beta_0 + \beta_1 x, \sigma^2)$. First we're going to look at the *expected* value of Y, that is $y = \mathbb{E}(Y) = \beta_0 + \beta_1 x$. Since we do not know β_0 and β_1, we do not know (and will never know) the expected response y. However, we can *estimate* y via

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}\, x.$$

It is also possible to give a confidence interval for y:

$$\widehat{y} \pm c\,\sqrt{\text{MSE}}\,\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\,,$$

where c is a constant that depends on n and the confidence level α (it is the $1-\alpha/2$ quantile of the t_{n-2} distribution). Recall that MSE estimates the variance σ^2 of the model error.

If we wish to *predict* the value of Y for a given value of x, then we have *two* sources of variation:

1. Y itself is a random variable, which is normally distributed with variance σ^2,

2. we don't know the expectation $\beta_0 + \beta_1 x$ of Y. Estimating this number on the basis of previous observations $Y_1, \ldots, Y_n$ brings another source of variation.

Thus, instead of a confidence interval for $\beta_0 + \beta_1 x$ we need a *prediction interval* for a new response Y. We can construct a **prediction interval** at level $1-\alpha$ for Y, by finding two random bounds such that the random variable Y falls in the interval with probability $1-\alpha$:

$$\widehat{y} \pm c\sqrt{\text{MSE}}\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\,,$$

where c is the same constant as for the confidence interval above.

The `R` function to find the prediction interval and confidence interval for a new value x is `predict()`.

Continuing our height versus shoe size example, suppose we wish to find a confidence interval for the expected height for a shoe size $x = 30$. This is found as follows.

```
> predict(model1,data.frame(shoe=30),interval="confidence")

      fit      lwr      upr
1 175.9217 173.4261 178.4172
```

We calculate also predict the weight of a person whose shoe size is 30 to lie in the following interval, with probability 0.95.

```
> predict(model1,data.frame(shoe=30),interval="prediction")

       fit      lwr      upr
1 175.9217 157.2999 194.5434
```

Note that the prediction interval is much wider.

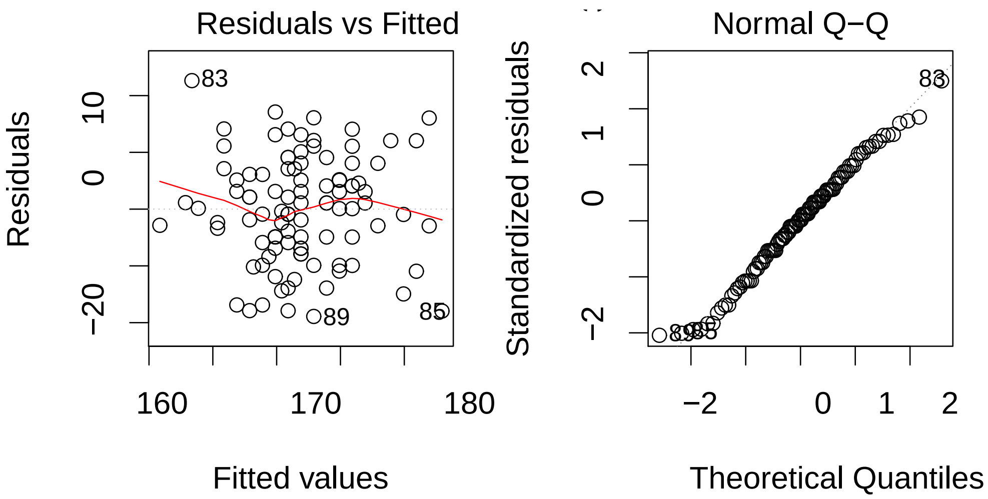
9.2.5 Validation of assumptions

As for the ANOVA model we can do an analysis of residuals to examine whether the underlying assumptions of the linear regression model are verified. Various plots of the residuals can be used to detect rather easily whether the assumptions on the errors $\{\varepsilon_i\}$ are respected:

- Plotting the histogram of residuals to check for normality. The QQ-plot is another approach.

- Plotting the residuals e_i as a function of the predicted values $\widehat{y}_i$. When all the model assumptions are verified, residuals and predicted values are uncorrelated. This plot should have no particular structure. This plot also gives indications on the validity on the linearity assumption, and on the homogeneity of error variance. The plot of e_i against $\widehat{y}_i$ should show a uniform spread of the residuals following a horizontal line on either side of the x axis.

```
> par(mfrow=c(1,2))
> plot(model1,1:2)
```



Examining the residuals as a function of predicted values, we that the residuals are correctly spread, symmetrical about the x axis: the conditions of the model seem valid.

Note that the instruction `plot(model1)` can draw four plots; some of these are for detection of outliers.

9.2.6 Summary

The table below presents the main functions to use for simple linear regression between the response variable `y` and the explanatory variable `x`.

Table 9.1: Main R functions for simple linear regression.

R **instruction**	**Description**
`plot(y~x)`	scatter plot
`lm(y~x)`	estimation of the linear model
`summary(lm(y~x))`	description of results of the model
`abline(lm(y~x))`	draw the estimated line
`confint(lm(y~x))`	confidence interval for regression parameters
`predict()`	function for predictions
`plot(lm(y~x))`	graphical analysis on residuals

9.3 Multiple linear regression

A linear regression model that contains more than one explanatory variable is called a *multiple linear regression model*.

Definition 9.2 (Multiple Linear Regression Model). In a **multiple linear regression model** the response data $Y_1, \ldots, Y_n$ depend on d-dimensional explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top$, via the linear relationship

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i , \quad i = 1, \ldots, n , \tag{9.13}$$

where $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$.

To explain things, let us go back to the student survey data set `ssurv`. Instead of "explaining" the student height via their shoe size, we could include other quantitative explanatory variables, such as the weight (stored in `weight`). The corresponding `R` formula for this model would be

$$\texttt{height} \sim \texttt{shoe} + \texttt{weight}$$

meaning that each random height `Height` satisfies

$$\texttt{Height} = \beta_0 + \beta_1 \texttt{shoe} + \beta_2 \texttt{weight} + \varepsilon,$$

where ε is a normally distributed error term with mean 0 and variance σ^2. The model has thus 4 parameters.

Before analysing the model we present a scatter plot of all pairs of variables, using the `R` function `pairs()`.

```
> pairs(height ~ shoe + weight)
```

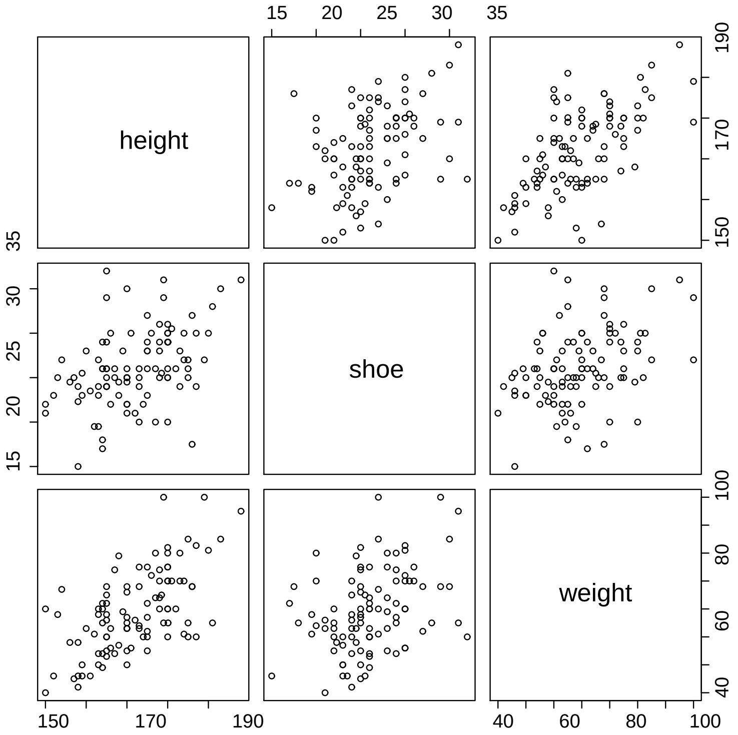


Figure 9.4: Scatter plot of all pairs of variables.

9.3.1 Analysis of the model

As for simple linear regression, the model can be analysed using the function `lm()`:

```
> model2 <- lm(height˜ shoe + weight)
> summary(model2)
Call:
lm(formula = height ˜ shoe + weight)

Residuals:
     Min        1Q    Median        3Q       Max
-21.4193   -4.0596   0.1891    4.8364   19.5371

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 132.2677     5.2473  25.207  < 2e-16 ***
shoe          0.5304     0.1962   2.703   0.0081 **
weight        0.3744     0.0572   6.546 2.82e-09 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 7.785 on 97 degrees of freedom
Multiple R-squared: 0.4301,        Adjusted R-squared: 0.4184
F-statistic: 36.61 on 2 and 97 DF,  p-value: 1.429e-12
```

The results output by summary() are presented in the same fashion as for simple linear regression. Note that the Fisher's global F test is used to test the global joint contribution of all explanatory variables in the model to "explaining" the variations of height. The null hypothesis is $H_0 : \beta_1 = \beta_2 = 0$. The assertion of interest is H_1 : at least one of the coefficients β_j ($j = 1, 2$) is significantly different from zero (at least one of the explanatory variables is associated with height after adjusting for the other explanatory variables).

Given the result of Fisher's global test (p-value = 1.429×10^{-12}), we can conclude that at least one of the explanatory variables is associated with height, after adjusting for the other variables. The individual Student tests indicate that:

- shoe size is linearly associated with student height, after adjusting for weight, with risk of error less than 5 % (p-value = 0.0081). At same weight, an increase of one cm in shoe size corresponds to an increase of 0.53 cm of average student height;

- weight is linearly associated with student height, after adjusting for shoe size (p-value = 2.82×10^{-09}). At same shoe size, an increase of one kg of the weight corresponds to an increase of 0.3744 cm of average student height.

Confidence intervals can again be found with confint()

```
confint(model2)
```

```
                  2.5 %      97.5 %
(Intercept) 121.8533072 142.6821199
shoe          0.1410087   0.9198251
weight        0.2608887   0.4879514
```

Confidence and prediction intervals can be obtained via the predict function. Suppose we wish to predict the height of a person with shoe size 30 and weight 75 kg. A confidence interval for the expected height is obtained as follows (notice that we can abbreviate "confidence" to "conf").

```
> predict(model2,data.frame(shoe=30,weight=75),interval="conf")
```

```
       fit      lwr      upr
1 176.2617 174.1698 178.3536
```

Similarly, the corresponding prediction interval is found as follows.

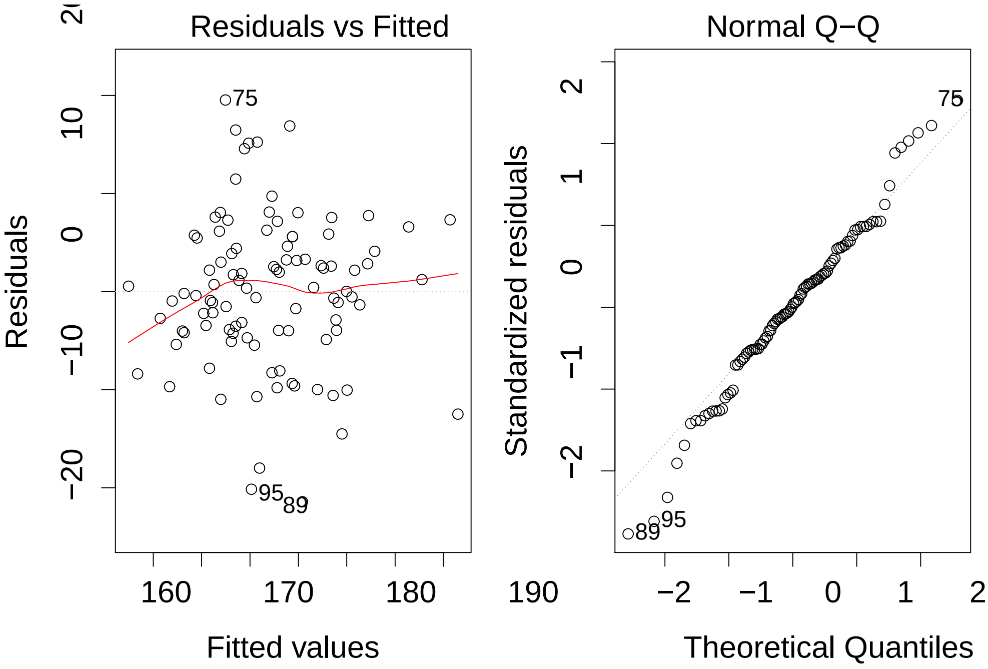
```
> predict(model2,data.frame(shoe=30,weight=75),interval="pred")
```

```
       fit      lwr      upr
1 176.2617 160.6706 191.8528
```

9.3.2 Validation of assumptions

We check the assumptions of this multivariate model by investigating the residuals plots.

```
> par(mfrow=c(1,2))
> plot(model2,1:2)
```



The residuals are correctly spread, symmetrical about the x axis: the conditions of the model seem valid. Moreover, the QQ-plot indicates no extreme departure from the normality.

9.3.3 Summary

Table 9.2 lists the main functions useful for multiple linear regression.

Table 9.2: Main R functions for multiple linear regression.

R instruction	Description
`pairs()`	graphical inspection
`lm(y~x1+x2+...+x3)`	estimation of the multiple linear model
`summary(lm())`	description of the results of the model
`confint(lm())`	confidence interval for regression parameters
`predict()`	function for predictions
`plot(lm())`	graphical analysis of residuals
`x1:x2`	interaction between x_1 and x_2

Conclusion

- Regression analysis involves modelling the response variable in terms of a linear relationship between the mean response and the predictor variables with residual variability about the mean.
- The standard inference for regression is to see whether the slope is zero or not. A slope of zero indicates no linear association between the variables while a significant nonzero slope indicates an association.
- The assumptions of linear regression are that the relationship is linear and that the residual variability is Normally distributed with constant standard deviation. These assumptions should be checked using plots of the residuals.
- Confidence intervals and prediction intervals can be calculated for estimated means and predicted outcomes, respectively, based on the least-squares fit.

9.4 Problems

1. Edwin Hubble discovered that the universe is expanding. If v is a galaxy's recession velocity (relative to any other galaxy) and d is its distance (from that same galaxy), Hubble's law states that

$$v = Hd, \tag{9.14}$$

where H is known as Hubble's constant. The following are distance (in millions of light-years) and velocity (thousands of miles per second) measurements made on 5 galactic clusters.

distance	68	137	315	405	700
velocity	2.4	4.7	12.0	14.4	26.0

2. For the following situations, formulate a regression or ANOVA model.

   (a) In a study of shipping costs, a company controller has randomly selected 9 air freight invoices from current shippers in order to assess the relationship between shipping costs and distance, for a given volume of goods.

   (b) Heart rates were monitored for 20 laboratory rats during three different stages of sleep.

Chapter 10

Linear model

Much of modeling in applied statistics is done via the versatile class of linear models. We will give a brief introduction to such models, which requires some knowledge of linear algebra (mostly vector/matrix notation). We will learn that both the ANOVA and linear regression models are special cases of linear models, so that these can be analysed in a similar way using `R`'s powerful `lm()` function. In addition to estimation and hypothesis testing, we consider model selection to determine which of many competing linear models is the most descriptive of the data.

10.1 Introduction

The ANOVA and linear regression models in Chapters 8 and 9 are both special cases of a **linear model**. Let $\mathbf{Y}$ be the column vector of response data $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$.

Definition 10.1 In a **linear model** the response data vector $\mathbf{Y}$ depends on a matrix $\mathcal{X}$ of explanatory variables (called the **design matrix**) via the linear relationship

$$\mathbf{Y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}$ is a vector of parameters and $\boldsymbol{\varepsilon}$ a vector of independent error terms, each $\mathsf{N}(0, \sigma^2)$ distributed.

Example 10.1 For the simple linear regression model (see Definition 9.1) we have ☞ 140

$$\mathcal{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} .$$

The situation for linear models in which the explanatory variables are *factors* is a little more complicated, requiring the introduction of dummy variables. We explain it with an example.

☞ 121 **Example 10.2** Consider a 1-factor ANOVA model (see Section 8.2) with 3 levels and 2 replications per levels. Denoting the responses by

$$\underbrace{Y_1, Y_2}_{\text{level 1}}, \underbrace{Y_3, Y_4}_{\text{level 2}}, \underbrace{Y_5, Y_6}_{\text{level 3}} ,$$

and the expectations within the levels by μ_1, μ_2, and μ_3, we can write the vector $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_3 \\ \mu_3 \end{pmatrix} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}}_{\boldsymbol{\varepsilon}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} .$$

If we denote for each response Y the level by x, then we can write

$$Y = \mu_1 \, \mathrm{I}(x = 1) + \mu_2 \, \mathrm{I}(x = 2) + \mu_3 \, \mathrm{I}(x = 3) + \varepsilon, \tag{10.1}$$

where $\mathrm{I}(x = k)$ is an **indicator** or *dummy* variable that is 1 if $x = k$ and 0 otherwise, $k = 1, 2, 3$. As an alternative to (10.1) we could use the "factor effects" representation

$$Y = \mu + \alpha_1 \mathrm{I}(x = 1) + \alpha_2 \, \mathrm{I}(x = 2) + \alpha_3 \, \mathrm{I}(x = 3) + \varepsilon, \tag{10.2}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 0$. Or we could use the representation

$$Y = \mu + \alpha_2 \, \mathrm{I}(x = 2) + \alpha_3 \, \mathrm{I}(x = 3) + \varepsilon, \tag{10.3}$$

where $\alpha_1 = 0$. In this case μ should be interpreted as the expected response in level 1.

In R, all data from a general linear model is assumed to be of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n , \tag{10.4}$$

where x_{ij} is the j-th explanatory variable for individual i and the errors ε_i are independent random variables such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. In matrix form, $\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Thus, the first column can always be interpreted as an "intercept" parameter. The corresponding R formula for this model would be

$$\texttt{y} \sim \texttt{x1} + \texttt{x2} + \cdots + \texttt{xp} .$$

It is important to note that R automatically treats quantitative and qualitative explanatory variables differently. For any linear model you can retrieve the design matrix via the function `model.matrix()`.

Let us look at some examples of linear models. In the first model variables x_1 and x_2 are both considered (by R) to be quantitative.

```
> my.dat <- data.frame(y = c(10,9,4,2,4,9),
+    x1=c(7.4,1.2,3.1,4.8,2.8,6.5),x2=c(1,1,2,2,3,3))
> mod1 <- lm(y~x1+x2,data = my.dat)
> print(model.matrix(mod1))
  (Intercept)  x1 x2
1           1 7.4  1
2           1 1.2  1
3           1 3.1  2
4           1 4.8  2
5           1 2.8  3
6           1 6.5  3
```

Suppose we want the second variable to be factorial instead. We can change the type as follows. Observe how this changes the design matrix.

```
> my.dat$x2 <- factor(my.dat$x2)
> mod2 <- lm(y~x1+x2,data=my.dat)
> print(model.matrix(mod2))
  (Intercept)  x1 x22 x23
1           1 7.4   0   0
2           1 1.2   0   0
3           1 3.1   1   0
4           1 4.8   1   0
5           1 2.8   0   1
6           1 6.5   0   1
```

Warning

By default, R sets the incremental effect α_i of the first-named level (in alphabetical order) to zero. To impose the model constraint $\sum_i \alpha_i = 0$ for a factor `x`, use `C(x,sum)` in the R formula, instead of `x`.

In this example, the variable $x2$ is a categorical variable:

```
> my.dat$x2
[1] 1 1 2 2 3 3
Levels: 1 2 3
```

The model `mod2` is an extension of the model presented in equation (10.3):

$$Y = \mu + \beta_1 x_1 + \alpha_2 \, \mathrm{I}(x_2 = 2) + \alpha_3 \, \mathrm{I}(x_2 = 3) + \varepsilon, \tag{10.5}$$

where $\alpha_1 = 0$, associated to the first-named level of `x2` (which is ”1” here). In this model, μ is interpreted as the expected response in level 1 in a model adjusted with the x_1 variable. The parameter α_2 should be interpreted as the expected difference between the response in level 2 and the response in level 1:

$$\begin{aligned} \mathbb{E}(Y \,|\, x_1 = a; x_2 = 2) - \mathbb{E}(Y \,|\, x_1 = a; x_2 = 1) &= \mu + a - (\mu + a - \alpha_2), \\ &= \alpha_2 \end{aligned}$$

where a is any value for the variable x_2. A similar interpretation holds for the parameter α_3.

The following code enables to use the constraint $\sum_i \alpha_i = 0$.

```
> mod3 <- lm(y~x1+C(x2,sum),data=my.dat)
> print(model.matrix(mod3))
```

```
  (Intercept)  x1 C(x2, sum)1 C(x2, sum)2
1           1 7.4           1           0
2           1 1.2           1           0
3           1 3.1           0           1
4           1 4.8           0           1
5           1 2.8          -1          -1
6           1 6.5          -1          -1
```

This design matrix $\mathcal{X}$ is associated with the parameter vector $\boldsymbol{\beta} = (\mu, \beta_1, \alpha_1, \alpha_2)^\top$, with $\alpha_1 + \alpha_2 + \alpha_3 = 0$, as the model `mod3` can be written as:

$$
\begin{aligned}
Y &= \mu + \beta_1 x_1 + \alpha_1 \,\mathrm{I}(x_2 = 1) + \alpha_2 \,\mathrm{I}(x_2 = 2) + \alpha_3 \,\mathrm{I}(x_2 = 3) + \varepsilon \\
&= \mu + \beta_1 x_1 + \alpha_1 \,\mathrm{I}(x_2 = 1) + \alpha_2 \,\mathrm{I}(x_2 = 2) - (\alpha_1 + \alpha_2)\,\mathrm{I}(x_2 = 3) + \varepsilon \\
&= \mu + \beta_1 x_1 + \alpha_1 \,(\mathrm{I}(x_2 = 1) - \mathrm{I}(x_2 = 3)) + \alpha_2 \,(\mathrm{I}(x_2 = 2) - \mathrm{I}(x_2 = 3)) + \varepsilon\ .
\end{aligned}
$$

In this parameterisation, μ is the overall effect, common to all levels on the factor x_2 in a model adjusted with the variable x_1, and α_i is the incremental effect of level i in a model adjusted with the variable x_1.

10.2 Estimation and hypothesis testing

Suppose we have a vector data $\mathbf{y}$ from a linear model $\mathbf{Y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathcal{X}$ is a known design matrix, and $\boldsymbol{\varepsilon}$ is a vector of iid $\mathsf{N}(0, \sigma^2)$ errors. We wish to estimate the parameter vector $\boldsymbol{\beta}$ and the model variance σ^2. We can again use a least-squares to estimate $\boldsymbol{\beta}$: Find $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^\top$ such that

$$
\sum_{i=1}^{n} (y_i - \{\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}\})^2 \quad \text{is minimal.}
$$

It can be shown that this gives the least squares estimate $\widehat{\boldsymbol{\beta}} = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top \mathbf{y}$, where $(\mathcal{X}^\top \mathcal{X})^{-1}$ is the inverse of the matrix $\mathcal{X}^\top \mathcal{X}$. The quantity

$$
e_i = y_i - \{\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}\}
$$

is the i-th residual error. Hence, the least squares criterion minimises the sum of the squares of the residual errors, denoted SSE. To estimate σ^2 we can, as in Chapters 8 and 9, take the mean square error

$$
\widehat{\sigma^2} = \mathrm{MSE} = \frac{\mathrm{SSE}}{n - (p + 1)},
$$

where $p + 1$ is the number of components in the vector $\boldsymbol{\beta}$.

For hypothesis testing, we can test whether certain parameters in $\boldsymbol{\beta}$ are zero or not. This can be investigated with an analysis of variance, where the residual variance of the full model is compared with the residual variance of the reduced model. The corresponding test statistics have an F distribution under the null hypothesis. The exact details are beyond a first introduction to statistics, but fortunately R provides all the information necessary to carry out a statistical analysis of quite complicated linear models.

If we are interested to a single parameter β_i, we also can use the same approach as the Student's test used to test if a single parameter is equal to zero or not; see (9.11).. ☞ 143 In a multivariate model, the individual test statistic used in R is following a Student's t distribution with $n-(p+1)$ degrees of freedom (p being the number of covariates in the model).

10.3 Using the computer

To make things more concrete we return to the dataset birthwt which we used at the end of Section 7.5. ☞ 109 We wish to explain the child's weight at birth using various characteristics of the mother, her family history, and her behaviour during pregnancy. The explained variable is weight at birth (quantitative variable btw, expressed in grammes); the explanatory variables are given below.

First we load the data:

```
> library(MASS)       # load the package MASS
> ls("package:MASS") # show all variables associated with this package
> help(birthwt)       # find information on the data set birthwt
```

Here is some information from help(birthwt) on the explanatory variables that we will investigate.

```
age:   mother‘s age in years
lwt:   mother‘s weight in lbs
race:  mother‘s race (1 = white, 2 = black, 3 = other)
smoke: smoking status during pregnancy (0 = no, 1 = yes)
ptl:   no. of previous premature labors
ht:    history of hypertension (0 = no, 1 = yes)
ui:    presence of uterine irritability (0 = no, 1 = yes)
ftv:   no. of physician visits during first trimester
bwt:   birth weight in grams
```

We can see the structure of the variables via str(birthwt). Check yourself that all variables are defined as *quantitative* (int). However, the variables race, smoke, ht, and ui should really be interpreted as *qualitative* (factors). To fix this, we could redefine redefine them with the function as.factor(), similar to what we did in Chapter 2. Alternatively, we could use the function factor() in the R formula to let the program know that certain variables are factors. We will use the latter approach.

Tip

For *binary* response variables (that is, variables taking the values 0 or 1) it does not matter whether the variables are interpreted as factorial or numerical, as R will return identical summary tables for both cases.

We can now investigate all kinds of models. For example, let us see if the mother's weight, her age, her race, and whether she smokes explain the baby's birthweight.

```
> attach(birthwt)
model1 <- lm(bwt~lwt+age+factor(race)+smoke)
sumr1 <- summary(model1)
sumr1

Call:
lm(formula = bwt ~ lwt + age + factor(race) + smoke)

Residuals:
    Min      1Q  Median      3Q     Max
-2281.9  -449.1    24.3   474.1  1746.2

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2839.433    321.435   8.834  8.2e-16 ***
lwt               4.000      1.738   2.301  0.02249 *
age              -1.948      9.820  -0.198  0.84299
factor(race)2  -510.501    157.077  -3.250  0.00137 **
factor(race)3  -398.644    119.579  -3.334  0.00104 **
smoke          -401.720    109.241  -3.677  0.00031 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 682.1 on 183 degrees of freedom
Multiple R-squared: 0.1483,       Adjusted R-squared: 0.125
F-statistic: 6.373 on 5 and 183 DF,  p-value: 1.758e-05
```

The results output by `summary()` are presented in the same fashion as for simple linear regression. Parameter estimates are given in the column `Estimate`.

The realisations of Student's test statistics associated with the hypotheses $H_0 : \beta_i = 0$ and $H_1 : \beta_i \neq 0$ are given in column `t value`; the associated p-values are in column `Pr(>|t|)`. `Residual standard error` gives the estimate of σ and the number of associated degrees of freedom $n-p-1$. The coefficient of determination R^2 (`Multiple R-squared`) and an adjusted version (`Adjusted R-squared`) are given, as are the realisation of Fisher's global test statistic (`F-statistic`) and the associated p-value.

Note

Fisher's global F test is used to test the global joint contribution of all explanatory variables in the model for "explaining" the variability in Y. The null hypothesis is $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ (under the linear model, the p explanatory variables give no useful information to predict Y). The assertion of interest is H_1 : at least one of the coefficients β_j ($j = 1, 2, \ldots, p$) is significantly different from zero (at least one of the explanatory variables is associated with Y after adjusting for the other explanatory variables).

Given the result of Fisher's global test (p-value = 1.758×10^{-5}), we can conclude that at least one of the explanatory variables is associated with child weight at birth, after adjusting for the other variables. The individual Student tests indicate that:

- mother weight is linearly associated with child weight, after adjusting for age, race and smoking status, with risk of error less than 5 % (p-value = 0.022). At the same age, race status and smoking status, an increase of one pound in the mother's weight corresponds to an increase of 4 g of average child weight at birth;

- the age of the mother is not significantly linearly associated with child weight at birth when mother weight, race and smoking status are already taken into account (p-value = 0.843);

- weight at birth is significantly lower for a child born to a mother who smokes, compared to children born to non-smoker mothers of same age, race and weight, with a risk of error less than 5 % (p-value=0.00031). At same age, race and mother weight, child weight at birth is 401.720 g less for a smoker mother than for a non-smoker mother.

- regarding the interpretation of the variable race we recall that the model performed used as reference the group race=1 (white). Then, the estimation of −510.501 g represents the difference of child birth weight between black mothers (`race=2`) and white mothers (reference group), and this result is significantly different from zero (p-value=0.001) in a model adjusted for mother weight, mother age and smoking status. Similarly, the difference in average weight at birth between group `race = 3` and the reference group is −398.644 g and is significantly different from zero (p-value=0.00104), adjusting for mother weight, mother age and smoking status.

Interaction

We can also include interaction terms in the model.Let us see whether there is any interaction effects between `smoke` and `age` via the model

$$\texttt{Bwt} = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{smoke} + \beta_3\texttt{age} \times \texttt{smoke} + \varepsilon\ .$$

In R this is done as follows:

```
model3 <- lm(bwt~age*smoke)
summary(model3)
```

```
Call:
lm(formula = bwt ~ age * smoke)

Residuals:
     Min       1Q   Median       3Q      Max
-2189.27  -458.46    51.46   527.26  1521.39

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2406.06     292.19   8.235 3.18e-14 ***
age            27.73      12.15   2.283   0.0236 *
smoke         798.17     484.34   1.648   0.1011
age:smoke     -46.57      20.45  -2.278   0.0239 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 709.3 on 185 degrees of freedom
Multiple R-squared: 0.06909,        Adjusted R-squared: 0.054
F-statistic: 4.577 on 3 and 185 DF,  p-value: 0.004068
```

We observe that the estimate for β_3 (-46.57) is significantly different from zero (p-value $= 0.024$). We therefore conclude that the effect of mother age on child weight is not the same depending on the smoking status of the mother. The results on association between mother age and child weight must therefore be presented separately for the smoker and the non-smoker group. For non-smoking mothers (`smoke = 0`), the mean child weight at birth increases on average by 27.73 grams for each year of the mother's age. A confidence interval can be found as follows.

```
confint(model3)[2,]
```

```
  2.5 %   97.5 %
 3.76278 51.69998
```

For smoking mothers, there seems to be a decrease in birthweight, $\widehat{\beta_1} + \widehat{\beta_3} = 27.73138 - 46.57191 = -18.84054$. To see if this is significant, we can again make a confidence interval and see if 0 is contained in it or not. A clever way of doing this is to create a new variable `nonsmoke` = 1-`smoke`, which reverses the encoding for the smokers and nonsmokers. Then, the parameter $\beta_1 + \beta_3$ in the original model is the same as the parameter β_1 in the following model

$$\texttt{Bwt} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{nonsmoke} + \beta_3 \texttt{age} \times \texttt{nonsmoke} + \varepsilon\ .$$

Hence the confidence interval can be found as follows.

```
nonsmoke <- 1 - smoke
confint(lm(bwt~age*nonsmoke))[2,]
```

```
    2.5 %     97.5 %
-51.28712  13.60605
```

Since 0 lies in this confidence interval, the effect of `age` on `bwt` is not significant for smoking mothers.

10.4 Variable selection

Among the large number of possible explanatory variables, we wish to select those which explain the observed responses the best. This way, we can decrease the number of predictors (giving a parsimonious model) and get good predictive power by eliminating redundant variables.

In this section, we briefly present two methods for variable selection available in R. They are illustrated on a few variables from data set `birthwt`. In particular, we consider the explanatory variables `lwt`, `age`, `ui`, `smoke`, `ht` and two recoded variables `ftv1` and `ptl1`. We note `ftv1` = 1 if there was at least one visit to a physician, and `ftv1`= 0 otherwise. Similarly, we note `ptl1` = 1 if there is at least one preterm birth in the family history, and `ptl1` = 0 otherwise.

```
> ftv1 <- as.integer(ftv>=1)
> ptl1 <- as.integer(ptl>=1)
```

10.4.1 Forward selection

The forward selection method is an iterative method. At each step, it selects the most significant explanatory variable (at level α) when we regress Y on all explanatory variables selected at previous steps and the newly chosen variable, as long as the marginal contribution of the new variable is significant.

Watch this method in action with function `add1()` for level $\alpha = 0.05$.

```
> form1 <- formula(bwt~lwt+age+ui+smoke+ht+ftv1+ptl1) #store formula
> add1(lm(bwt~1),form1,test="F")

Single term additions

Model:
bwt ~ 1
       Df Sum of Sq      RSS    AIC F value     Pr(F)
<none>               99969656 2492.8
lwt     1   3448639 96521017 2488.1  6.6814  0.010504 *
age     1    815483 99154173 2493.2  1.5380  0.216475
ui      1   8059031 91910625 2478.9 16.3968 7.518e-05 ***
smoke   1   3625946 96343710 2487.8  7.0378  0.008667 **
ht      1   2130425 97839231 2490.7  4.0719  0.045032 *
ftv1    1   1340387 98629269 2492.2  2.5414  0.112588
ptl1    1   4755731 95213925 2485.6  9.3402  0.002570 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`ui` is the most significant variable.

```
add1(lm(bwt~ui),form1,test="F")

Single term additions

Model:
bwt ~ ui
       Df Sum of Sq      RSS    AIC F value   Pr(F)
<none>               91910625 2478.9
lwt     1   2074421 89836203 2476.6  4.2950 0.03960 *
age     1    478369 91432256 2479.9  0.9731 0.32518
smoke   1   2996636 88913988 2474.6  6.2687 0.01315 *
ht      1   3162595 88748030 2474.3  6.6282 0.01082 *
```

```
ftv1     1    950090 90960534 2478.9  1.9428 0.16503
ptl1     1   2832244 89078381 2475.0  5.9139 0.01597 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ht is the most significant variable.

```
> add1(lm(bwt~ui+ht),form1,test="F")
Single term additions

Model:
bwt ~ ui + ht
       Df Sum of Sq      RSS    AIC F value    Pr(F)
<none>              88748030 2474.3
lwt     1   3556661 85191369 2468.5  7.7236 0.006013 **
age     1    420915 88327114 2475.4  0.8816 0.348988
smoke   1   2874044 85873986 2470.0  6.1916 0.013720 *
ftv1    1    698945 88049085 2474.8  1.4686 0.227120
ptl1    1   2678123 86069907 2470.5  5.7564 0.017422 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

lwt is the most significant variable.

```
add1(lm(bwt~ui+ht+lwt),form1,test="F")
Single term additions

Model:
bwt ~ ui + ht + lwt
       Df Sum of Sq      RSS    AIC F value   Pr(F)
<none>              85191369 2468.5
age     1     97556 85093813 2470.3  0.2109 0.64657
smoke   1   2623742 82567628 2464.6  5.8469 0.01658 *
ftv1    1    510128 84681241 2469.4  1.1084 0.29380
ptl1    1   2123998 83067371 2465.8  4.7048 0.03136 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

smoke is the most significant variable.

```
> add1(lm(bwt~ui+ht+lwt+smoke),form1,test="F")
Single term additions

Model:
bwt ~ ui + ht + lwt + smoke
       Df Sum of Sq      RSS    AIC F value   Pr(F)
<none>              82567628 2464.6
age     1     67449 82500178 2466.5  0.1496 0.69935
ftv1    1    274353 82293275 2466.0  0.6101 0.43576
ptl1    1   1425291 81142337 2463.3  3.2145 0.07464 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No further variable is significant. The method thus stops at the model with variables: ui, ht, lwt and smoke.

10.4.2 Backward elimination

This time, we start with the complete model and at each step, we delete the variable with lowest value of Student's test statistic (largest p-value) in absolute value, as long as it is not significant (at a specified level α).

Watch this method in action with function `drop1()` for level $\alpha = 0.05$.

```
> drop1(lm(form1),test="F")

Single term deletions

Model:
bwt ~ lwt + age + ui + smoke + ht + ftv1 + ptl1
       Df Sum of Sq      RSS    AIC F value     Pr(F)
<none>                80682074 2466.2
lwt     1   2469731 83151806 2469.9  5.5405 0.0196536 *
age     1     90142 80772217 2464.5  0.2022 0.6534705
ui      1   5454284 86136359 2476.6 12.2360 0.0005899 ***
smoke   1   1658409 82340484 2468.1  3.7204 0.0553149 .
ht      1   3883249 84565324 2473.1  8.7116 0.0035808 **
ftv1    1    270077 80952151 2464.9  0.6059 0.4373584
ptl1    1   1592757 82274831 2467.9  3.5731 0.0603190 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We delete variable `age`.

```
> drop1(lm(bwt~lwt+ui+smoke+ht+ftv1+ptl1),test="F")

Single term deletions

Model:
bwt ~ lwt + ui + smoke + ht + ftv1 + ptl1
       Df Sum of Sq      RSS    AIC F value     Pr(F)
<none>                80772217 2464.5
lwt     1   2737552 83509769 2468.8  6.1684 0.0139097 *
ui      1   5561240 86333456 2475.0 12.5309 0.0005082 ***
smoke   1   1680651 82452868 2466.3  3.7869 0.0531944 .
ht      1   3953082 84725299 2471.5  8.9073 0.0032306 **
ftv1    1    370120 81142337 2463.3  0.8340 0.3623343
ptl1    1   1521058 82293275 2466.0  3.4273 0.0657462 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We delete variable `ftv1`.

```
> drop1(lm(bwt~lwt+ui+smoke+ht+ptl1),test="F")
Single term deletions

Model:
bwt ~ lwt + ui + smoke + ht + ptl1
       Df Sum of Sq      RSS    AIC F value    Pr(F)
<none>                81142337 2463.3
```

```
lwt      1   2887694 84030031 2467.9  6.5126 0.011528 *
ui       1   5787979 86930316 2474.3 13.0536 0.000391 ***
smoke    1   1925034 83067371 2465.8  4.3415 0.038583 *
ht       1   4215957 85358294 2470.9  9.5082 0.002362 **
ptl1     1   1425291 82567628 2464.6  3.2145 0.074642 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We delete variable `plt1`. The method stops at the model with variables: `ui`, `ht`, `lwt` and `smoke`.

It should be noted that different methods of automatic selection may not lead to the same choice of variables in the final model. They have the advantage of being easy to use, and of treating the question of variable selection in a systematic manner. The main drawback is that variables are included or deleted based on purely statistical criteria, without taking into account the aim of the study. This usually leads to a model which may be satisfactory from a statistical point of view, but in which the variables are not the most relevant when it comes to understanding and interpreting the data in the study.

10.5 Analysis of residuals

We present here a few elements on analysis of residuals. Suppose for the `birthwt` data set we ended up with model represented by the following `R` formula.

$$\texttt{bwt} \sim \texttt{smoke + age + lwt + factor(race) + ui + ht + smoke:age}\,.$$

☞ 156

It is good to review what the actual model looks like, in terms of (10.4).

$$\texttt{Bwt} = \beta_0+\beta_1\texttt{smoke}+\beta_2\texttt{age}+\beta_3\texttt{lwt}+\beta_4\texttt{race2}+\beta_5\texttt{race3}+\beta_6\texttt{ui}+\beta_7\texttt{ht}+\beta_8\texttt{smoke}\times\texttt{age}+\varepsilon.$$

The following `R` code checks various model assumptions.

```
> finalmodel<-lm(bwt~smoke+age+lwt+factor(race)+ui+ht+smoke:age)
> par(mfrow=c(1:2))
> plot(finalmodel,1:2,col.smooth="red")
```

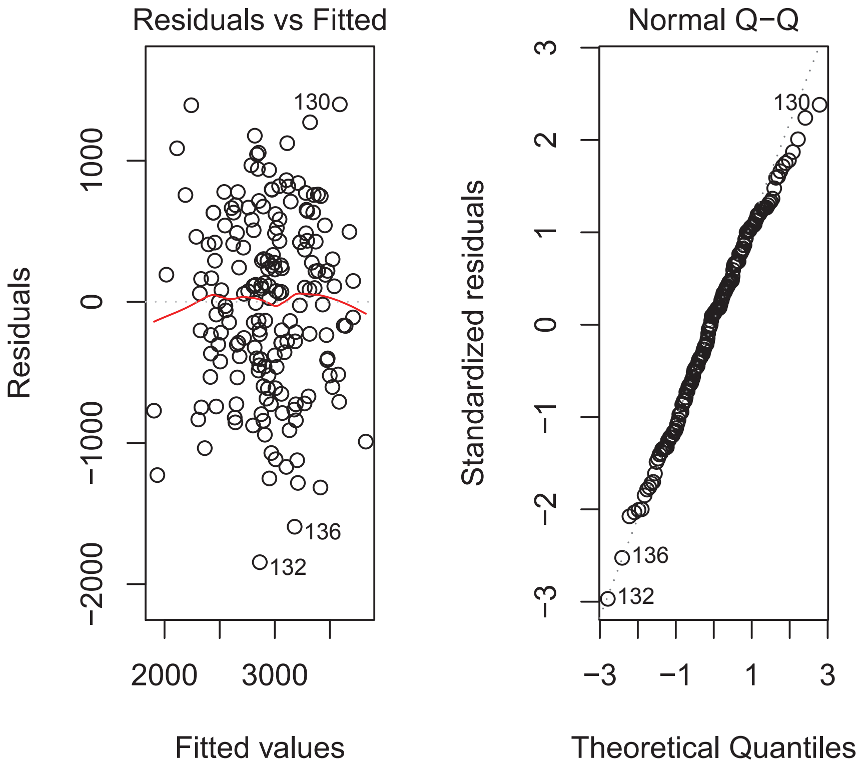


Figure 10.1: Checking the assumptions of homoscedasticity (left) and normality (right).

It can also be useful to plot the residuals as a function of each explanatory variable, as shown in Figure 10.2. This plot is useful to check whether there is a relationship between the error term and the explanatory variables, which would invalidate the assumption of independence between errors and explanatory variables. This plot is also useful to detect outliers.

```
> res <- residuals(finalmodel)
> par(mfrow=c(2,3))
> plot(res~smoke);plot(res~age);plot(res~lwt)
> plot(res~race);plot(res~ui);plot(res~ht)
```

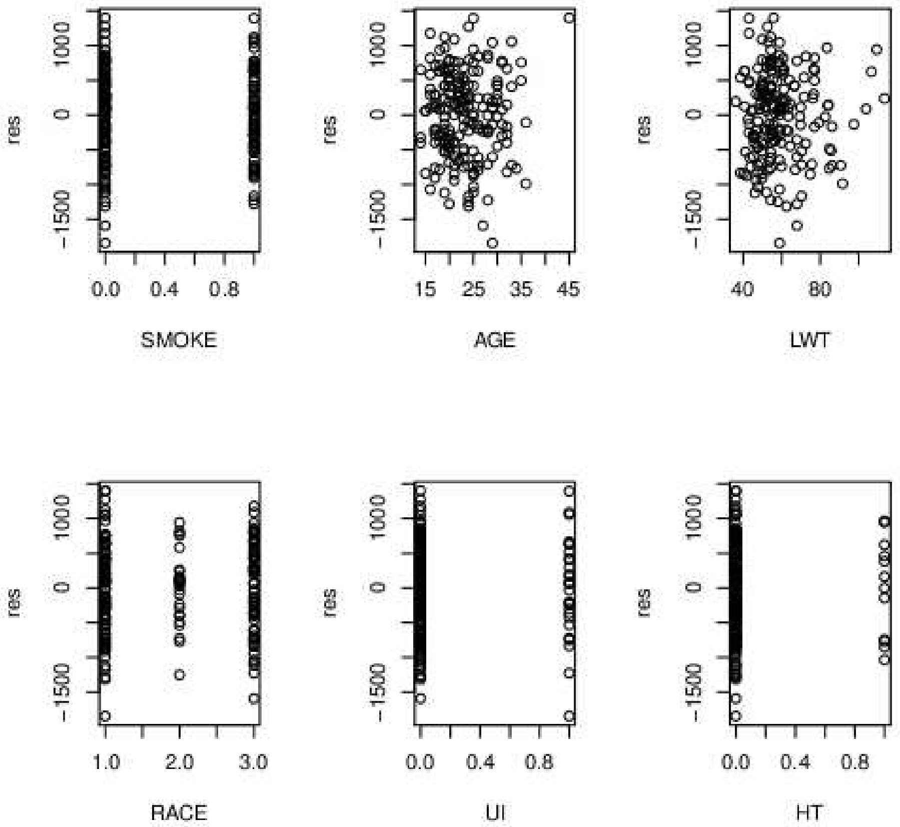


Figure 10.2: Residuals as a function of explanatory variables.

Conclusion

- The ANOVA and linear regression models are both special cases of a **linear model**.
- Categorical variables have to be included as dummy variables in the model.
- Interaction terms enable to get more flexible models.
- Forward and backward selection procedures select the variables which explain the best the observed responses.

Appendix A

Introduction to R

A.1 Presentation of the software

A.1.1 Origins

R is a piece of statistical software created by Ross Ihaka & Robert Gentleman. It is both a programming language and a work environment: commands are executed thanks to instructions coded in a relatively simple language; results are displayed as text; plots are visualized directly in their own window. It is a clone of the software S-plus, which is based on the object-oriented programming language S, developed by AT&T Bell Laboratories in 1988. This piece of software is used to manipulate data, draw plots and perform statistical analyses of data.

A.1.2 Why use R?

First of all, R is **free** and **open-source**. It works under UNIX (and Linux), Windows and Macintosh: it is **cross-platform**. It is being developed in the free software movement by a large and growing community of eager volunteers. Anyone can contribute to and improve R by integrating more functionalities or analysis methods. It is thus a quickly and constantly evolving piece of software.

It is also a very powerful and complete tool, especially adapted for **statistical methods**. The learning curve is steeper than for other software on the market (such as SPSS or Minitab): R is not the kind of software you can use out of the box with a few clicks of the mouse on menus. However, this approach has two advantages:

- it is **didactic**, since you have to understand a statistical method before you can put it at use;
- R is very **efficient** once you master it: you will then be able to create your own tools, enabling you to operate very sophisticated data analyses.

Warning

R is harder to comprehend than other software on the market. You need to spend time learning the syntax and commands.

R is especially powerful for data manipulation, calculations, and plots. Its features include:

- an integrated and very well conceived documentation system;
- efficient procedures for data treatment and storage;
- a suite of operators for calculations on tables, especially matrices;
- a vast and coherent collection of statistical procedures for data analysis;
- advanced graphical capabilities;
- a simple and efficient programming language, including conditioning, loops, recursion and input-output possibilities.

A.2 R and Statistics

Many classical and modern statistical techniques are implemented in R. The most common methods for statistical analysis, such as:

- descriptive statistics;
- hypothesis testing;
- analysis of variance;
- linear regression methods (simple and multiple).

are directly included at the core of the system. It should be noted than most advanced statistical methods are also available through external packages. These are easy to install, directly from a menu. They are all grouped and can be browsed on the website of the *Comprehensive R Archive Network* (CRAN) (`http://cran.r-project.org`). This website also includes, for some large domains of interest, a commented list of packages associated with a theme (called Task View). This facilitates the search for a package on a specific statistical method. Furthermore, detailed documentation for each package is available on the CRAN.

It should also be noted that recent statistical methods are added on a regular basis by the Statistics community itself.

A.3 R and plots

One of the main strengths of R is its capacity (much greater than that of other free software on the market) to combine a programming language with the ability to draw high–quality plots. Usual plots are easily drawn using predefined functions. These functions also include many parameters, for example to add titles, captions, colours,etc. But it is also possible to create more sophisticated plots to represent complex data, such as contour lines, volumes with a 3D effect, density curves, and many other things. It is also possible to add mathematical formulae. You can arrange or overlay several plots in the same window, and use many colour palettes.

You can get a demonstration of the graphical possibilities in R by typing in the following instructions:

```
>demo(image)
>example(contour)
>demo(graphics)
>demo(persp)
>demo(plotmath)
>demo(Hershey)
>require(lattice) # Load the package, which you must have
                 # previously installed by using the menu
                 # Packages/Install packages.
>demo(lattice)
>example(wireframe)
>require(rgl)     # Same remark as above.
>demo(rgl)        # You can interact by using your mouse.
>example(persp3d)
```

The next figures shows a few of these plots.

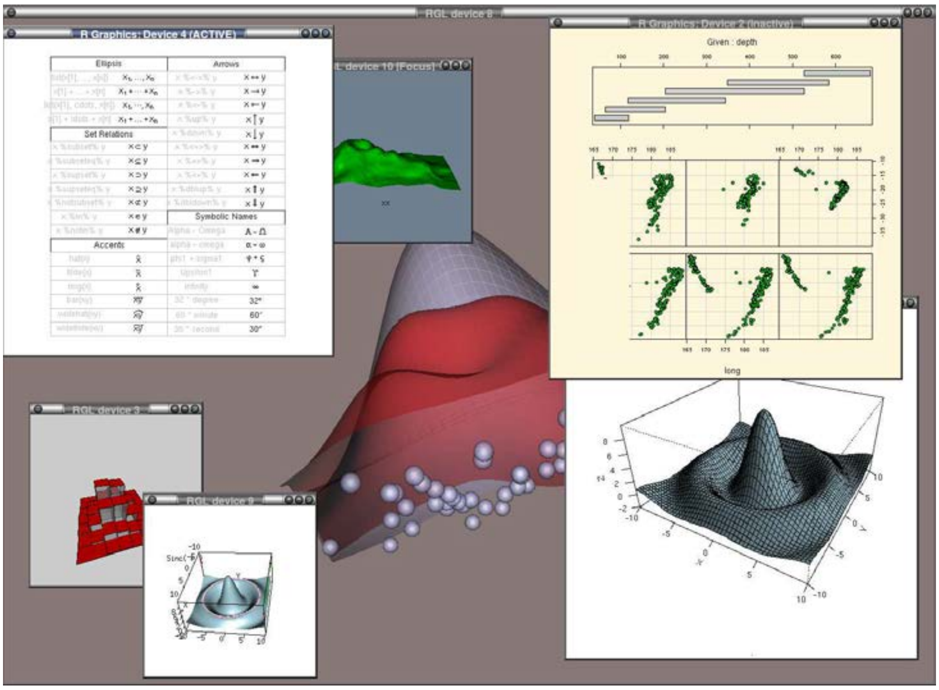


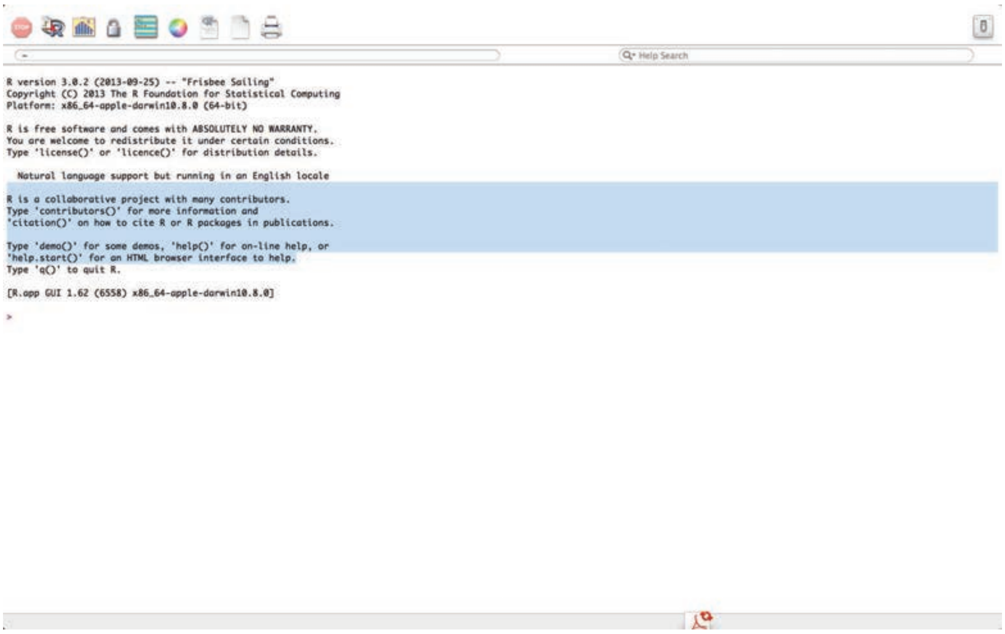
Figure A.1: A few of the graphical possibilities offered by R.

A.4 The R Graphical User Interface (GUI)

The R Graphical User Interface (*i.e.*, its set of menus) is very limited compared to other standard software, and completely non-existent on some operating systems. This minimalism can set back some new users. However, this drawback is limited since:

- it has the didactic advantage that it incites users to know well the statistical procedures they wish to use;
- there are additional tools which extend the GUI.

When you open the R GUI program by clicking on the R icon you should get a window that looks something like Figure A.2. This windows is the R **console**. Below the startup information (information about what version R you are using, license details, and so on) you should see a > (greater than sign). This prompt is where you enter R code. To run code that you have typed after the prompt hit the Enter or Return key. Now that we have a new R session open we can get started.



```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.62 (6558) x86_64-apple-darwin10.8.0]

>
```

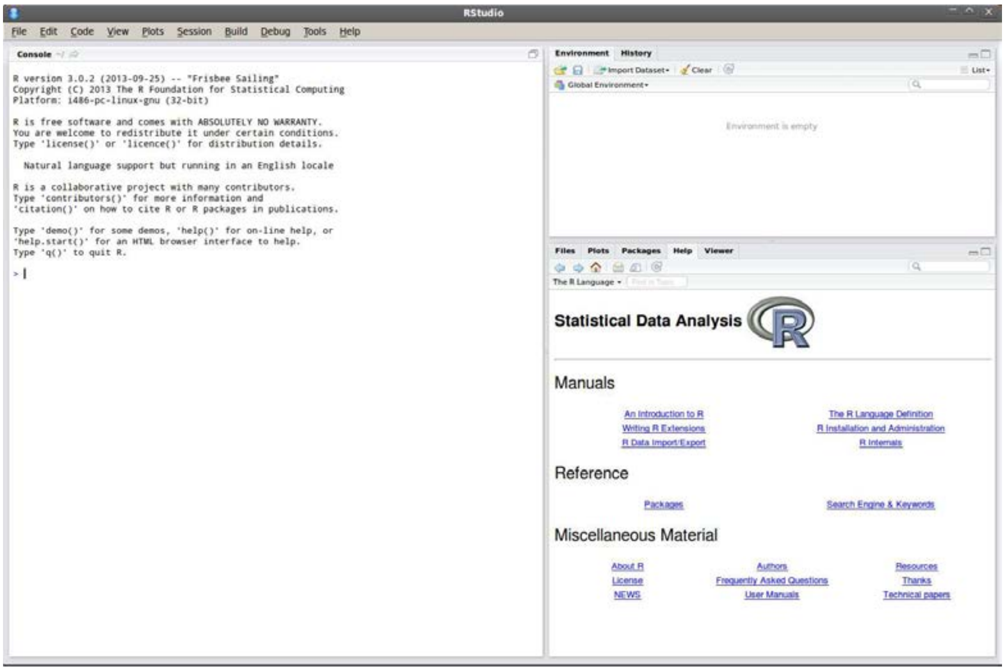
Figure A.2: R startup Console.

A.5 First steps in R

In the following, we give a brief overview of RStudio which is a relatively new editor specially targeted at R. It may be the best editor to start with for a beginner. RStudio is cross-platform, free and open-source software; it is available at http://www.rstudio.com. The Rstudio website (http://www.rstudio.com/ide/docs/) has a number of useful tutorials.

Presentation of Rstudio

Rstudio provides a centralized and well organized place to do almost anything you want to do with R. When you first open Rstudio you should see a default window that looks like Figure A.3. In this figure you see three window panes. The large one on the left is the *Console*. This pane functions exactly the same as the console in regular R. Other panes include the *Workspace/History* panes, in the upper right hand corner. The *Workspace* pane shows you all the objects in your current workspace and some of their characteristics, like how many observations a data frame has. You can click on an object in this pane to see its contents. This is especially useful for quickly looking at a data set. The *History* pane records all the commands you have run. It also allows you to return code and insert it into a source code file.



```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: i486-pc-linux-gnu (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Figure A.3: Rstudio Startup Panel.

In the lower right-hand corner you can see the *Files/Plots/Packages/Help* pane. Basically, it allows you to see and organize your files. The *Plots* pane is where figures you create in R appear. This panes allows you to see all of the figures you have created in a session using the right and left arrows icons. It also lets you save the figures in a variety of formats. The *Packages* pane shows the packages you have installed, allows you to load individual packages by clicking on the dialog box next to them, access their help files (just click on the package name), update the packages, and even install new packages. Finally, the *Help* pane shows you help files. You can search for help files and search within help files using this pane.

The important pane that does not show up when you open RStudio for the first time is the Source pane. The Source pane is where you create, edit, and run your source code files. R source code files have extension .R. When you create a new source code document RStudio will open a new *Source* pane. Do this by going to the menu bar and clicking on File then New File. In the New File drop down menu you can select the R Script option. You should now see a new pane with a bar across the top that looks like the first image in Figure A.3. To run the R code you have in your source code file simply highlight it (if you are only running one line code, you don’t need to highlight the code; you can simply put your cursor on that line) and click the Run icon on the top bar. This sends the code above where you have highlighted. The Source icon next to this runs all the code in the file using R’ source command.

A brief introduction of R syntax through some instructions to type

We advise the reader to play with these commands and try to understand how they work. Then we will give more explanations in the next section. In the *Source* pane write and run the following instructions.

• **Basic operations.**

```
> 1*2*3*4
[1] 24
> factorial(4)
[1] 24
> cos(pi)
[1] -1
> x <- 1:10
> x
 [1]  1  2  3  4  5  6  7  8  9 10
> exp(x)
 [1]     2.718282     7.389056    20.085537    54.598150
 [5]   148.413159   403.428793  1096.633158  2980.957987
 [9]  8103.083928 22026.465795
> x^2
 [1]   1   4   9  16  25  36  49  64  81 100
> chain <- "R is great!"
> chain
[1] "R is great!"
> nchar(chain)
[1] 11
> ?nchar
> M <- matrix(x,ncol=5,nrow=2)
> M
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9
[2,]    2    4    6    8   10
> M[2,3]
[1] 6
> L <- list(matrix=M,vector=x,chain=chain)
> L[[3]]
[1] "R is great!"
> mini.game <- function() {
+  while(TRUE) {
+  guess <- sample(0:1,1)
+  {cat("Choose a number between 0 and ");valeur <- scan(n=1)}
+   if (valeur==guess) {print("Well done!");break()}
+   else print("Try again.")
+   }
+ }
```

```
> # Type: mini.game()
> ls()
[1] "chain" "L"     "M"     "x"
> rm(chain)
```

The following commands perform matrix operations.

```
> A <- matrix(runif(9),nrow=3)
> 1/A
         [,1]     [,2]     [,3]
[1,] 2.270797 1.546875 1.422103
[2,] 1.268152 1.957924 1.057803
[3,] 1.642736 5.273120 2.174020
> A * (1/A)
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    1    1    1
[3,]    1    1    1
> B <- matrix(1:12,nrow=3)
> A * B
Error in A * B : non-conformable arrays
> A %*% B
         [,1]      [,2]      [,3]     [,4]
[1,] 3.842855  9.212923 14.582990 19.95306
[2,] 4.646105 11.380053 18.114001 24.84795
[3,] 2.367954  6.143031  9.918107 13.69318
> (invA <- solve(A))
          [,1]      [,2]      [,3]
[1,]  1.145642 -3.376148  5.187347
[2,]  4.379786 -4.641906  2.844607
[3,] -3.321872  6.381822 -5.863772
> A %*% invA
              [,1]         [,2] [,3]
[1,]  1.000000e+00 0.000000e+00    0
[2,]  0.000000e+00 1.000000e+00    0
[3,] -2.220446e-16 4.440892e-16    1
> det(A)
[1] 0.04857799
> eigen(A)
$values
[1]  1.6960690+0.000000i -0.1424863+0.091319i
[3] -0.1424863-0.091319i
$vectors
             [,1]                   [,2]                   [,3]
[1,] 0.5859852+0i  0.6140784-0.1816841i  0.6140784+0.1816841i
[2,] 0.7064296+0i  0.2234155+0.2505528i  0.2234155-0.2505528i
[3,] 0.3969616+0i -0.6908020+0.0000000i -0.6908020+0.0000000i
```

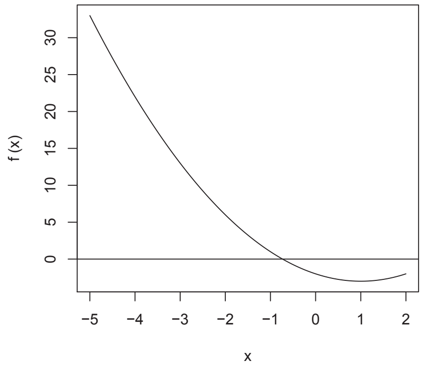
- **Statistics.**

Here are a few statistical calculations.

```
> weight <- c(70,75,74)
> mean(weight)
[1] 73
> height <- c(182,190,184)
> mat <- cbind(weight,height)
> mat
     weight height
[1,]     70    182
[2,]     75    190
[3,]     74    184
> apply(mat,MARGIN=2,FUN=mean)
  weight   height
 73.0000 185.3333
> ?apply
> colMeans(mat)
  weight   height
 73.0000 185.3333
> names <- c("Peter","Ben","John")
> data <- data.frame(Names=names,height,weight)
> summary(data)
   Names        height           weight
 Ben  :1   Min.   :182.0   Min.   :70.0
 John :1   1st Qu.:183.0   1st Qu.:72.0
 Peter:1   Median :184.0   Median :74.0
           Mean   :185.3   Mean   :73.0
           3rd Qu.:187.0   3rd Qu.:74.5
           Max.   :190.0   Max.   :75.0
```

- **Some plots.**

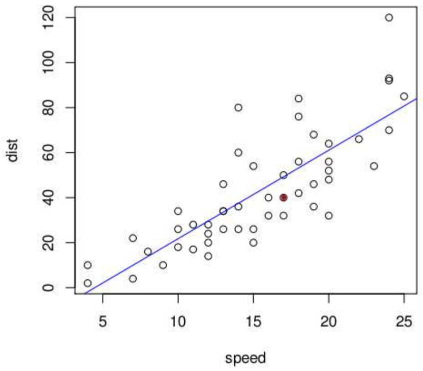
```
> f <- function(x) x^2-2*x-2
> curve(f,xlim=c(-5,2));abline(h=0)
> locator(1) # Click on the intersection of the two curves.
```



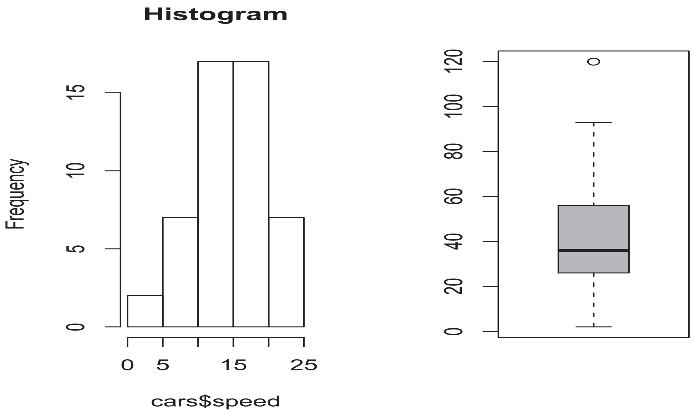
```
> uniroot(f,c(-5,2))
$root
```

```
[1] -0.7320503
$f.root
[1] -1.874450e-06
$iter
[1] 8
$estim.prec
[1] 6.103516e-05
```

```
> plot(cars)
> abline(lm(dist~speed,data=cars),col="blue")
> points(cars[30,],col="red",pch=20)
```



```
> par(mfrow=c(1,2))
> hist(cars$speed,main="Histogram")
> boxplot(cars$dist,col="orange")
```



A.6 Some R syntax used in this textbook

Before giving a brief overview of the main syntax used in this textbook let's note that R is referred to as an "object-oriented language". Objects are like the R language's nouns. They are things like a vector of numbers, a data set, a word, a table of results from some analysis, and so on. Saying that R is "object-oriented" means that R is focused on doing actions to objects. We will see some example of actions commands and functions later. First let's create a few objects.

Numeric and string objects

You can choose almost any name you want for our objects as long as it begins with an alphabetic character and does not contain spaces. To put something into the object we use the assignment operator <-:

```
> xx <- 10
```

To see the contents of our object xx, type its name.

```
> xx

[1] 10
```

10 is clearly the contents of xx, [1] is the row number of the object that 10 is on. You can create objects with words and other characters is the same way.

```
> my.text <- "I like R"
```

An object's type is important to keep in mind as it determines what we can do it. For example, you cannot take the mean of a character object like the my.text objects:

```
> mean(my.text)

[1] NA
Warning message:
In mean.default(my.text) : argument is not numeric or logical: returning NA
```

Trying to find the mean of your my.text object gives us a warning message and return NA: not applicable. To find out an object's type use the command:

```
> class(my.text)

[1] "character"
```

The function is enables to test the nature of an object:

```
> is(my.text,"character") # equivalent to is.character()

[1] TRUE
```

Vector and data frame objects

A vector is simply a group of numbers, character strings, and so on. Let's create a simple numeric vector containing the numbers 50, 38.5, 37.5. To do this we will use the c (concatenate) function:

```
> age <- c(50,38.5,37.5)
> age
[1] 50.0 38.5 37.5
```

Vectors of character strings are created in a similar way.

```
> Author <- c("Dirk","Benoit","Michael")
> Author
[1] "Dirk"   "Benoit"  "Michael"
```

Let's now combine the two vectors age and Author into a new object with the cbind function.

```
> AgeAuthorObject <- cbind(age,Author)
> AgeAuthorObject
 age    Author
[1,] "50"   "Dirk"
[2,] "38.5" "Benoit"
[3,] "37.5"   "Michael"
```

By binding these two objects together we've created a new matrix object. You can see that the numbers in the age column are between quotation marks. Matrices, like vectors, can only have one data type. You can also create a matrix object using the R function matrix. The functions colnames and rownames make it possible to retrieve or set column or row names of a matrix-like object. If you want to have an object with rows and columns and allow the columns to contain data with *different* types, we need to use data frames using the data.frame() function.

```
> AgeAuthorObject <- data.frame(age,Author)
> AgeAuthorObject
  age  Author
1 50.0   Dirk
2 38.5  Benoit
3 37.5 Michael
```

You can use the names command to see any data frames' names. The command names is not specific to data.frame object but could be applied to others R object such as list object which is latter defined.

```
> names(AgeAuthorObject)
 [1] "age"    "Author"
```

Notice that the first column of the data set has no name and it is a series of numbers. This is the row.names attribute. We can use the row.names command to set the row names from a vector.

```
> row.names(AgeAuthorObject) <- c("First","Second","Third")
 [1] "age"    "Author"
```

Component selection

The dollar sign ($) is called the component selector. It enables to extract any names from an R object.

```
> AgeAuthorObject$age

 [1] 50.0 38.5 37.5
```

In this example, it extracted the `age` column from the `AgeAuthorObject`. You can then compute for example the mean of the age by using

```
> mean(AgeAuthorObject$age)

 [1] 39.66667
```

Using the component selector can create long repetitive code if you want to select many components. You can streamline your code by using command such as `attach`. This command attaches a database to R's search path (you can see what is in your current search path with the `search` command; just type `search()` into your R console). R will then search the database for variables you specify. You don't need to use the component selector to tell R again to look in a particular data frame after you have attached it. For example, let's attach the `cars` data that comes with R. It has two variables, `speed` and `dist` (type ?cars for more information on this dataset)

```
> attach(cars)
> head(speed)  ## Display the first values of speed

[1] 4 4 7 7 8 9


> mean(speed)

[1] 15.4
```

It is a good idea to `detach` a data frame after you are done using it, to avoid confusing R.

```
> detach(cars)
```

Another way to select parts of an object is to use subscripts. They are denoted with squares brace `[]`. We can use subscripts to select not only columns from data frames but also rows and individuals values. Let's see it in action with the data frame `cars`

```
> head(cars)

 speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

```
> cars[3:7,]  ## select information from the third trough
                  ## seventh row
  speed dist
3     7    4
4     7   22
5     8   16
6     9   10
7    10   18

> cars[4,2]  ## select the fourth row of dist
[1] 22
```

An equivalent way is:

```
> cars[4,"dist"]
[1] 22
```

Also note the functions `which()`, `which.min()` and `which.max()`, which are often very useful to extract .

```
> mask <- c(TRUE,FALSE,TRUE,NA,FALSE,FALSE,TRUE)
> which(mask) # Outputs the indices corresponding to the values
                  # TRUE.
[1] 1 3 7

> x <- c(0:4,0:5,11)
> which.min(x)  # Outputs the index of the smallest value.
[1] 1

>  which.max(x)  # Outputs the index of the largest value.
[1] 12
```

We can also select the cars with a speed less than 9 mph by using

```
> cars[which(cars$speed<9),]
 speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
```

An another way is to use the function `subset`:

```
> subset(cars,speed<9)
 speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
```

List objects

The most flexible and richest structure in R is the list. Unlike the previous structures, lists can **group together in one structure data of different types** without altering them. Generally speaking, each element of a list can thus be a vector, a matrix or even a list. Here is a first example:

```
> A <- list(TRUE,-1:3,matrix(1:4,nrow=2),"A character string")
> A

[[1]]
[1] TRUE

[[2]]
[1] -1  0  1  2  3

[[3]]
     [,1] [,2]
[1,]    1    3
[2,]    2    4

[[4]]
[1] "A character string"
```

In such a structure, with heterogeneous data types, element ordering is often completely arbitrary. Elements can therefore be explicitly named, which makes the output more user-friendly. Here is an example:

```
> B <- list(my.matrix=matrix(1:4,nrow=2),my.numbers=-1:3)
> B

$my.matrix
     [,1] [,2]
[1,]    1    3
[2,]    2    4

$my.numbers
[1] -1  0  1  2  3
```

Naming elements will make it easier to extract elements from a list:

```
> B$my.matrix

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

Functions and commands

Functions and commands do things to objects. We will use “function” and “command” as synonyms, although there is a subtle difference between them. We have already used the function `mean()`. This function takes the sample mean of a numeric vector object. To find the mean of an numeric object `age` simply type:

```
> mean(x=age)
[[1] 39.66667
```

Notice that we typed the command's name then enclosed the object name in parentheses immediately afterward. This is the basic syntax that all commands use, i.e. COMMAND(ARGUMENTS). If you don't want to explicitly include an argument you still need to type the parentheses after the command. For example the function getwd() gives your current working directory:

```
> getwd()
"c:/Users/JohnSmith/STAT1301"
```

Arguments modify what commands do. In our example, we gave the mean command one argument (x=age) telling it that we wanted to find the mean of age. Arguments us the ARGUMENTLABEL=VALUE syntax. In this case x is the argument label. To find all of arguments that a command can accept look at Arguments section of the command's help file. To access to help file type: ?mean. Argument lables may be put in any order and also can be abbreviated provided there is no ambiguity. Let's see an another example with multiple arguments with the round command which rounds a vector numbers. We can use the digits (or abbreviated as dig or simply d) argument to specify how many decimal places we want the numbers rounded to.

```
> round(x=mean(age),digits=1)
[1]  39.7
```

Basic functions

Here are a few basic data manipulation functions. These are used very often; it is essential that you know them.

- length(): returns the length of a vector. We can use the digits argument to specify how many decimal places we want the numbers rounded to.

  ```
  > length(c(1,3,6,2,7,4,8,1,0))
  [1] 9
  ```

- sort(): sorts the elements of a vector, in increasing or decreasing order.

  ```
  > sort(c(1,3,6,2,7,4,8,1,0))
  [1] 0 1 1 2 3 4 6 7 8
  ```

  ```
  > sort(c(1,3,6,2,7,4,8,1,0),decreasing=TRUE)
  [1] 8 7 6 4 3 2 1 1 0
  ```

- rev(): rearranges the elements of a vector in reverse order.

  ```
  > rev(c(1,3,6,2,7,4,8,1,0))
  [1] 0 1 8 4 7 2 6 3 1
  ```

- order(), rank() : the first function returns the vector of (increasing or decreasing) ranking indices of the elements. The second function returns the vector of ranks of the elements. In case of a tie, the ordering is always from left to right.

  ```
  > vec <- c(1, 3, 6, 2, 7, 4, 8, 1, 0)
  > names(vec) <- 1:9
  > vec
  1 2 3 4 5 6 7 8 9
  1 3 6 2 7 4 8 1 0
  ```

  ```
  > sort(vec)
  9 1 8 4 2 6 3 5 7
  0 1 1 2 3 4 6 7 8
  ```

  ```
  >  order(vec)
  [1] 9 1 8 4 2 6 3 5 7
  ```

  ```
  >  rank(vec)
   1   2   3   4   5   6   7   8   9
  2.5 5.0 7.0 4.0 8.0 6.0 9.0 2.5 1.0
  ```

- unique(): as the name suggests, this function removes the duplicates of a vector.

  ```
  >  unique(c(1,3,6,2,7,4,8,1,0))
  [1] 1 3 6 2 7 4 8 0
  ```

Create your own functions

We have just seen some brief notions on executing functions in R. The R language can also be used to create your own functions. We give only a brief overview here. You should scrutinise the code below to ensure that you understand it well. To illustrate simply the function creation process, we shall focus on the computation of the Body Mass Index (BMI), from the weight (actually mass!) (in kg) and the height (in m), using the well-known formula

$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}.$$

This is easily programmed in R as follows:

```
> BMI <- function(weight,height){
 +          bmi <- weight/height^2
 +          names(bmi) <- "BMI"
 +          return(bmi)}
```

Warning

The function `return()` is optional in the code above, but you should take the habit of using it. Indeed, there are contexts where it is essential:

```
> f <- function(x){
+     res <- vector("numeric",length(x))
+     for (i in 1:10){
+     res[i] <- rnorm(1) + x[i]}
+      } # Forgot to include return(res)
>  f(1:10) # Does not output anything!
```

We can now execute the function `BMI()` we just created:

```
> BMI(70,1.82)

    BMI
21.13271


>  BMI(1.82,70) # Do not swap the arguments of a function

        BMI
0.0003714286


> BMI(height=1.82,weight=70) # unless they are preceded by their
                              # names.

 BMI
21.13271
```

This function only outputs a single value. The code below outputs a list of several variables.

```
>  BMI <- function(weight,height){
+      bmi <- weight/height^2
+      res <- list(Weight=weight,Height=height,BMI=bmi)
       return(res)}
```

The next instruction shows that the new function `BMI()` returns a list of three named elements (`Weight`, `Height` and `BMI`).

```
>  BMI(70,1.82)

 $Weight
[1] 70

$Height
[1] 1.82

$BMI
[1] 21.13271
```

Reading data

The following R instruction will read the data present in a file (to be chosen in a dialog window) and import them into R as a data.frame which we have chosen to call my.data.

```
> my.data <- read.table(file=file.choose(),header=T,sep="\t",
+                                dec=".",row.names=1)
```

The function read.table() accepts many arguments; the most common are described in the following table.

Table A.1: Main arguments to read.table().

Argument name	Description
file=path/to/file	Location and name of the file to be read.
header=TRUE	Logical value indicating whether the variable names are given on the first line of the file.
sep="\t"	The values on each line are separated by this character ("\t"=TABULATION; ""=whitespace; ","=,; etc.).
dec="."	Decimal mark for numbers ("." or ",").
row.names=1	The first column of the file gives the individuals' names. If this is not the case, simply omit this argument.

When using the function read.table(), you will need to specify the value of the argument file which must contain, in a character string, the name of the file and its complete path. You might have noticed that we used the function file.choose(), which opens up a dialog window to select a file and returns the required character string. This is an easy method to get the path to a file, but the path can also be specified explicitly:

```
> my.data <- read.table(file="C:/MyFolder/data.txt")
```

Warning

Note that file paths are specified using slashes (/). This notation comes from the UNIX environment. In R , you cannot use backslashes (\), as you would in Microsoft Windows, unless you double all the backslashes (\\).

Another option is using the function setwd() to change the work directory. The argument file will then accept the file name alone, without its path.

```
> setwd("C:/MyFolder")
> my.file <- "mydata.txt"
> data <- read.table(file=my.file)
```

Your data are now available in the R console: they are stored in the object which you have chosen to call data. You can visualize them by typing data; you can also type head(data) or tail(data) to display only the beginning or the end of the dataset. You can also use str(data) to see the nature of each column of your data.

Exporting data to an ASCII text file

The relevant function is `write.table()`. Suppose you have a data.frame called `mydata`, containing data that you wish to save in a text file. You would then use the instruction:

```
> write.table(mydata, file = "myfile.txt", sep = "\t")
```

Note

There also exists a function `write()`, which is used on vectors and matrices. This function has an interesting argument: `ncolumns` allows you to specify the number of columns in the resulting file. Note however that the file will contain the transpose of the matrix or vector you are writing.

The workspace and history

All of the objects you create become part of your workspace. Use the `ls()` command to list all of the objects in your current workspace.

```
> ls()
> [1] "age"              "AgeAuthorObject" "Author"           "my.text"
```

You can remove specific objects by using the `rm` command:

```
> rm(my.text)  #remove my.text object
```

If you want to remove all objects in the workspace use `rm(list=ls())`. To save the entire workspace into a `RData` or `rda` file recognise by `R` use the `save.image` command. The main argument of this function is the location and the name of the file you want to save the workspace into. If you don't specify the file path it will be saved into your current working directory (use `getwd()`). For example, to save the current workspace in a file called `first-try.RData` in the current working directory type:

```
> save.image(file="first-try.Rdata")
```

Use the `load` command to load a saved workspace back into `R` :

```
> rm(list=ls())  #remove all objects
> ls() #list the object in your workspace
> character(0)   #no object in your workspace


> load(first-try.Rdata)  #load objects in first-try.Rdata
> ls()
[1] "age"              "AgeAuthorObject" "Author"           "my.text"
```

If you want to save a specific object called, for example `mydata` (corresponding for example to a new data set you have created), you can save it to a file called for example `mydata.RData` by using `save` command:

```
> save(mydata,file="mydata.Rdata")
```

When you enter a command into R it becomes part of your history. To see the most recent commands in your history use the history command or use the History pane in Rstudio. You can also us the up and down arrows on your keyboard when your cursor is in the R console to scroll through your history.

To conclude this brief introduction for using R, we present the source() function which import a sequence of R instructions from a file to the console. This helps prevent overloading the console. Consider that you have the following code stored in a file called my-instruction.R which is located for example in your work directory.

```
x <- 13
y <- 45
my.list <- list(x=x,y=y)
```

Then you can execute this simple code by using the source() function. We first delete all objects in the work directory:

```
> rm(list=ls())  #remove all objects
> ls() #list the object in your workspace
```

And now execute the code from the file my-instruction.R.

```
> source("my-instruction.R")
> ls()
```

```
[1] "x"  "y" "my.list"
```

Appendix B

Installing R, Rstudio and R packages

This chapter explains how to install `R` version *x* (replace throughout *x* by the number of the latest version) and Rstudio under Microsoft Windows, Mac or Linux. We also present briefly how to install additional packages under Windows, Mac or Linux.

B.1 Installing `R` and Rstudio

For Microsoft Windows user:

- First download `R` (file `R-`*x*`-win.exe` where *x* is the number of the latest version) using your usual web browser from the URL `http://cran.r-project.org/bin/windows/base/`
  Save this executable file on the Windows Desktop and double-click the file `R-`*x*`-win.exe` (its icon is).

  The software then installs. All you have to do is follow the instructions displayed on your screen and keep the default options.

  When the icon is added to the Desktop, installation is complete.

- Download RStudio from `http://www.rstudio.com/ide/download/desktop` and install it. Leave all default settings in the installation options. You can now Open Rstudio

Tip

You should install `R` before installing `RSudio`. You can download the Mac or Linux version from the following websites:

- R : http://r-project.org
- RSudio: http://www.rstudio.com/ide/download

The download webpages for these programs have comprehensive information on how to install them, so please refer to those pages for more information.

After installing R and RStudio you will probably also want to install a number of user-written packages that are covered in this textbook.

B.2 Installing additional packages

Many additional modules (called packages or libraries) are available on the website http://cran.r-project.org/web/packages/

Packages extend the functionalities of R . We present several ways of installing a new package.

Installing from RStudio

Go to the *Packages* pane, click on *install Packages*. Then enter the name of the chosen packages (e.g., MASS) and hit the Enter key. The package is now installed and listed in the *Packages* pane. To load the package you need to tick the box related to the package.

Installing from the command line

You can use R without the menus of the graphical user interface. This is the case for example under Unix/Linux, where R does not have a graphical user interface. In that case, use the following command to install packages (say Rcmdr from the CRAN website:

```
> install.packages("Rcmdr")
```

B.3 Loading installed packages

Warning

To understand this section, you need to have a rough understanding of the difference between your computer's random access memory (RAM) and physical memory.

Installing a package means that its files are "written" physically on the hard disk: when you turn your computer off then on again, the files are still in the same place. You will not need to reinstall the package, unless you need an updated version.
On the contrary, loading a package (to the memory) means that it is temporarily at the user's disposal in R . But if you close and reopen R , the package is no longer available: you need to load it again.

To sum up, once you have installed a package on your computer's hard disk, you have to load it to R 's memory before you can use it.

For example, if you type in the R console:

```
> Commander()
```

you should see the following error message, which indicates that the package including this function cannot be accessed by R :

```
Error: cannot find function "Commander"
```

You must first load Rcmdr to the memory. To do this, you can either type

```
> require("Rcmdr")
```

or

```
> library("Rcmdr")
```

in the console. You can also go to the pane Packages in Rstudio and tick the box corresponding to the packages Rcmdr. Type once again in the R console:

```
> Commander()
```

Note that the package Rcmdr has been loaded and that this command no longer returns an error message.

General index