THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

STAT1301
Advanced Analysis of Scientific Data

by the
Statistics Group of
the School of Mathematics and Physics

July 27, 2025

CONTENTS

PREFACE

These notes are intended for first-year 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 self-learning guide for advanced first-year students, providing both a sound theoretical foundation of statistics as well as a comprehensive introduction to the statistical language R.

STAT1301 is the advanced version of STAT1201, and will explain several concepts on a deeper level than is feasible in STAT1201. Our guiding principle was that it is just as important to know the "why" as the "how". 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. 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 was partly based on:

- Joshua C.C. Chan and Dirk P. Kroese (2025). *Statistical Modeling and Computation*, Second Edition, 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.

We will introduce the topics in these notes in a linear fashion. We being with chapter 1 on *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 2. We then continue with a few more probability topics in Chapter 3, including multiple random variables, independence, and the central limit theorem. At the end of that chapter we introduce some simple statistical models. We then give a brief introduction to studies, data and evidence in Chapter 4. In Chapter 5 we describe how to summarize and visualize data. We will use the statistical package `R` to read and structure the data and make figures and tables and other summaries.

We then discuss the statistical analysis of data, with *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* (Chapter 8) and *regression* (Chapter 9), both of which are special examples of a *linear model* (Chapter 10). The final Chapter 11 touches on additional statistical techniques, such as goodness of fit tests and the chi-squared independence test. The `R` program will be of great help here. Appendix A gives a short introduction to `R`.

Statistics Group
School of Mathematics and Physics
The University of Queensland
July 27, 2025

CHAPTER 1

UNDERSTANDING RANDOMNESS

The purpose of this chapter is to give a short introduction to statistical studies and then 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.

1.1 Introduction

Statistics is an essential part of science, providing the language and techniques necessary for understanding and dealing with chance and uncertainty in the real and abstract worlds. It involves the design, collection, analysis, and interpretation of data, with aims including estimation, prediction, quantification of uncertainty and production of evidence-based inference. It has applications in every field, including science, engineering, business, medicine, economics and social sciences.

1.2 Statistical Studies

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. Summarize and visualize 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 (Biased Coin)** 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 here is: is the coin fair or biased? 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 0 0 1 0 1 0 0 1 . . . 0 1 1. These are our data. We can visualize the data by drawing a bar graph such as in Figure 1.1.

Figure 1.1: Outcome of an experiment where a fair coin is tossed 100 times. The dark bars indicate when Heads (=1) appears.

Think about the pros and cons of this plot. If we are only interested in the biasedness of the coin, then a simple chart that shows the total numbers of Heads and Tails would suffice, as knowing exactly where the Heads or Tails appeared is irrelevant. Thus, we can *summarize* the data by giving only 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 1.2) and therefore a different outcome for x.

Figure 1.2: Outcomes of three different experiments where a fair coin is tossed 100 times.

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.

Interestingly, we don't actually need any formulas to calculate this probability. Computers have become so fast and powerful that we can quickly approximate probabilities via *simulations*. Simulating this fair coin flip experiment in R is equivalent to sampling 100 times (with replacement) from a "population" $\{0, 1\}$ and counting how many 1s there are. In R:

```
> coin = c(0,1)

> sample(coin, 100, replace = T)

  [1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1 1 0 1 0 0 0 0 0
 [29] 1 0 0 1 0 0 1 1 1 0 1 0 1 1 1 1 1 1 0 0 1 1 0 1 0 0 0 0
 [57] 0 1 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 0
 [85] 1 0 1 0 0 0 0 0 1 1 1 1 0 0 1 1

> sum(sample(coin, 100, replace = T))

  [1] 48
```

In this case we had 48 Heads out of 100. If we repeat it two times:

```
> sum(sample(coin, 100, replace = T))

  [1] 54

> sum(sample(coin, 100, replace = T))

  [1] 38
```

Now, let's repeat this 1000 times and save the output in a variable:

```
> data.we.could.have.seen
      = replicate(1000, sum(sample(coin, 100, replace = T)))

> data.we.could.have.seen

   [1] 43 47 56 54 49 45 46 51 41 47 48 44 54 53 43 54 46 49
  [19] 48 44 47 52 53 39 44 52 53 45 52 57 49 54 48 56 42 47
  [37] 42 46 44 47 49 46 51 53 59 57 50 45 51 55 50 53 60 53
  ...
 [973] 45 49 42 53 54 51 56 46 49 48 53 46 55 37 47 49 51 54
 [991] 50 49 49 50 57 35 44 49 45 52

> data.we.could.have.seen >= 60

   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [10] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [19] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [28] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
  [46] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
  ...
 [991] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[1000] FALSE

> sum(data.we.could.have.seen >= 60)

  [1] 21

> mean(data.we.could.have.seen >= 60)

  [1] 0.021
```

So, *without any theoretical knowledge of probability*, we have found that the probability that X takes a value of 60 or more is approximately 0.021 for a fair coin. If the coin is indeed fair, then what we have witnessed was quite a rare event — entirely possible, but rather rare. We can either:

- accept that the coin is fair and that we just happened to see a rather rare occurrence; or
- do not accept that we've been so unlucky, and instead suspect that the coin is rigged.

■

You have already carried out your first scientific study and statistical hypothesis test! The rest of this course will build up your foundational knowledge in probability and statistics so that you can tackle a wider range of research questions and data types.

1.3 Random Experiments

Statistical data is inherently random: if we would repeat the process of collecting the data, we would most likely obtain different measurements. Various reasons why there is variability in the data will be discussed in Section 4.2. ☞ 69

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

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 an observational study on post-heart failure patients, resulting in a set of measurements for each patient, as discussed in Chapter 5. ☞ 81

The word experiment often means a planned scientific study, but here were are using it in a broader sense with focus on an outcome or set of outcomes which are treated as random.

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 one fits an appropriate model to the observed data.

■ **Example 1.2 (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
☞ 10 can be based on this simple experiment. In Section 1.2 we view this experiment from a statistical point of view (is the coin fair?). To better understand how this coin toss experiment behaves, we can carry it out on a computer. 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 <- sample(c(0,1),n=100,replace=T) # simulate coin tosses
> barplot(x)            # plot the results in a bar chart
```

The first line of code chooses 100 values with equal probability from the set {0, 1}, with replacement. Typical outcomes for three such experiments are given in Figure 1.3.

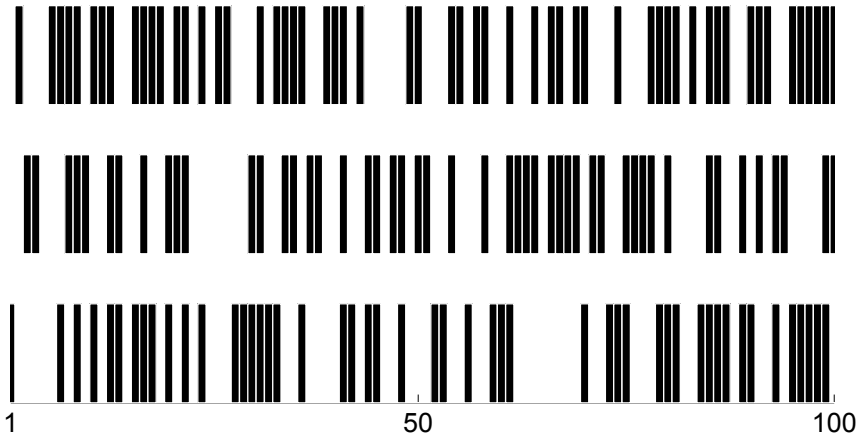


Figure 1.3: Outcomes of three different experiments where a fair coin is tossed 100 times.

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 1.4. Notice that the average number of Heads seems to converge to 0.5, but there is a lot of random fluctuation.

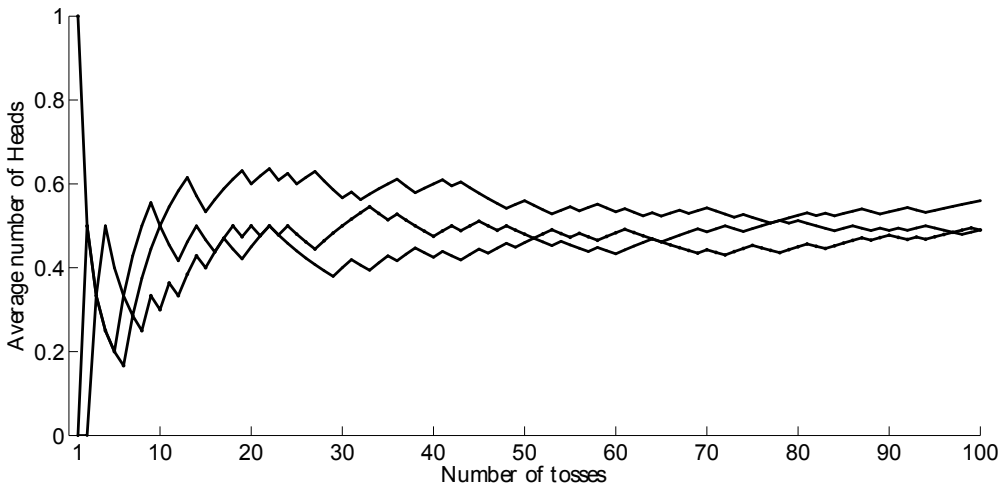


Figure 1.4: The average number of Heads in n tosses, where $n = 1, \ldots, 100$.

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 types of questions, we need to have a closer look at the models that are used to describe random experiments.

1.4 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 1.1: Sample Space

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.

We need to be able to describe situations where one of a *group* of outcomes occurs.

Definition 1.2: Event

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 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 1.5.

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 $B \subseteq A$ (B is a **subset** of A) then event B is said to *imply* event A.
5. $B \setminus A = B \cap A^C$ **(set difference of B and A)**

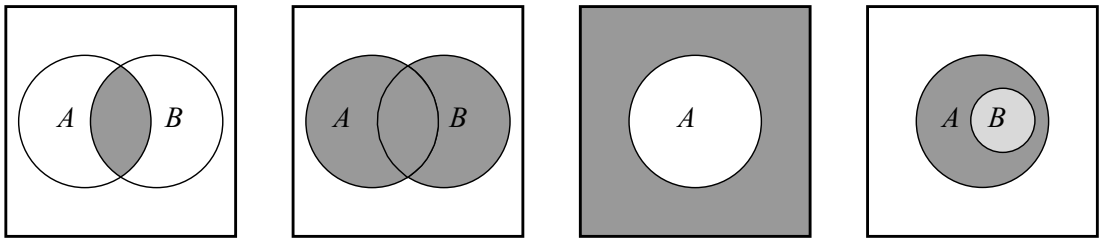


Figure 1.5: Venn diagrams of set operations. Each square shows the sample space Ω.

The set difference $B \setminus A$ can be seen on the $A \cap B$ diagram as the white part of B, excluding any part of A.

■ **Example 1.3 (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. ■

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$ (empty set), are called **disjoint** or **mutually exclusive** events. A collection of set properties and laws are given in the following theorem.

Theorem 1.1: Set Properties and Laws

Let A, B and C be sets. Then

$$A \cup B = B \cup A, \quad \text{Commutative union} \tag{1.1}$$
$$A \cap B = B \cap A, \quad \text{Commutative intersection} \tag{1.2}$$
$$A \cup (B \cup C) = (A \cup B) \cup C, \quad \text{Associative union} \tag{1.3}$$
$$A \cap (B \cap C) = (A \cap B) \cap C, \quad \text{Associative intersection} \tag{1.4}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \quad \text{Distributive union} \tag{1.5}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad \text{Distributive intersection} \tag{1.6}$$
$$(A \cap B)^c = A^c \cup B^c, \quad \text{De Morgan 1} \tag{1.7}$$
$$(A \cup B)^c = A^c \cap B^c, \quad \text{De Morgan 2.} \tag{1.8}$$

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, which are the axioms of probability, proposed by Andrey Kolmogorov in 1933.

Definition 1.3: Probability Measure

A **probability measure** $\mathbb{P}$ is a function which assigns a number to each event and satisfies the following axioms:

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

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \; . \tag{1.9}$$

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 axioms, any properties you have derived and any relevant set rules.

Theorem 1.2: Properties of a Probability Measure

Let A and B be events, $C_1, \cdots, C_N$ be disjoint events and $\mathbb{P}$ be a probability measure. Then,

1. $\mathbb{P}(\emptyset) = 0$,
2. $\mathbb{P}\left(\bigcup_{i=1}^{N} C_i\right) = \sum_{i=1}^{N} \mathbb{P}(C_i)$, Finite Disjoint Union Rule
3. if $A \subseteq B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$, Monotonicity
4. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, Complement Rule
5. $\mathbb{P}(A) \leqslant 1$,
6. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, Addition Rule.

The second of these can be termed the **finite disjoint union rule** of probability. If an event can happen in several distinct ways, 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.

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*. The probability of an event $A \subseteq \Omega$ is in this case simply

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of elements in } A}{\text{Number of elements in } \Omega} . \tag{1.10}$$

The calculation of such probabilities thus reduces to *counting*.

1.5 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 1.6.

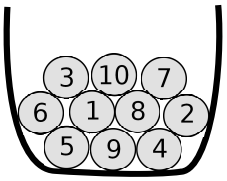


Figure 1.6: An urn with n balls.

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, with replacement) and (unordered, without replacement). The art is to recognize 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
☞ 17 in Remark 1.4: ordered arrangements are enclosed by round brackets and unordered ones by curly brackets.

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 1.7.

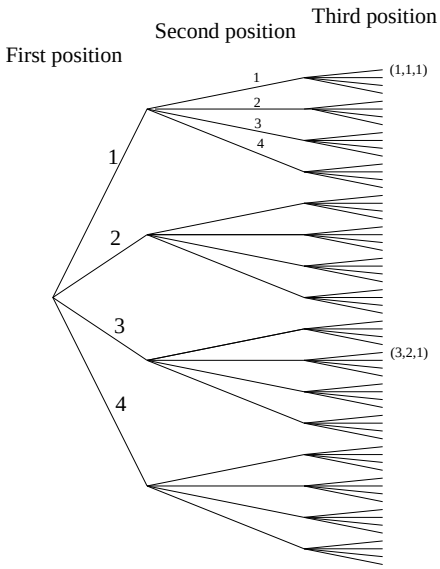


Figure 1.7: Enumerating the number of ways in which three ordered positions can be filled with 4 possible numbers, where repetition is allowed.

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

Theorem 1.3: Arrangements with Order and Replacment

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

Drawing Without Replacement, Ordered

Here we draw again k numbers (balls) from the set $\{1, 2, \ldots, n\}$, 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 $\{1, \ldots, n\}$. (A permutation of $\{1, \ldots, n\}$ of size n is simply called a permutation of $\{1, \ldots, n\}$ (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 possibilities left. So the number of arrangements is $4 \times 3 \times 2 = 24$ as illustrated in Figure 1.8, which is the same tree as in Figure 1.7, but with all "duplicate" branches removed.

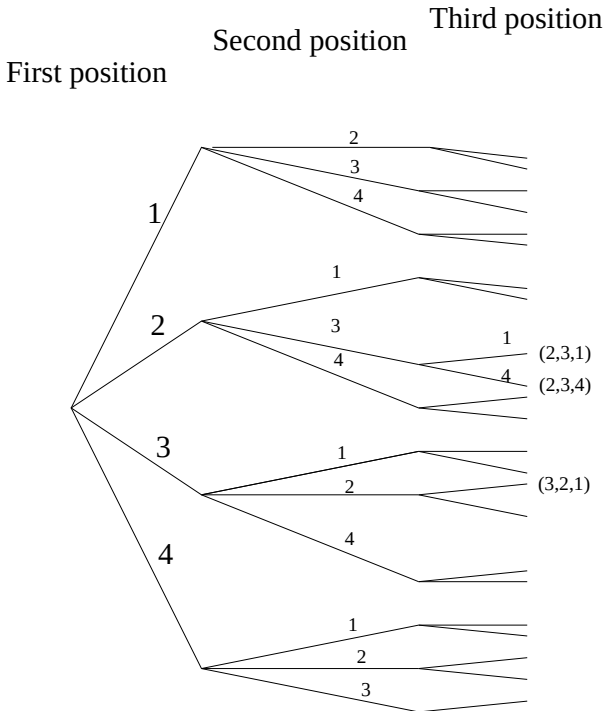


Figure 1.8: Enumerating the number of ways in which three ordered positions can be filled with 4 possible numbers, where repetition is NOT allowed.

For general n and k we have:

Theorem 1.4: Arrangements with Order and without Replacement

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 1.5: Arrangements without Order and without Replacement

The number of combinations of size k from the set $\{1,\ldots n\}$ 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 1.1: Number of ways k balls can be drawn from an urn containing n 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 $\{1, \ldots, 365\}$. 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 $\mathbb{P}(A) = |A|/|\Omega| \approx 0.074$.

1.6 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 1.9.

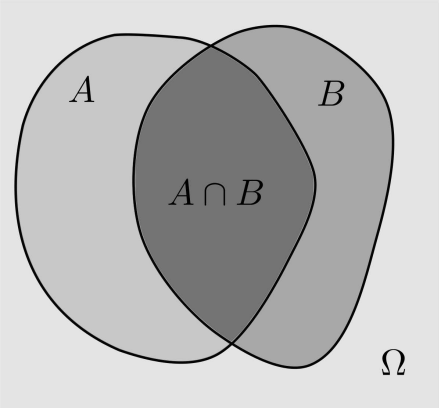


Figure 1.9: What is the probability that A occurs (that is, the outcome lies in A) given that the outcome is known to lie in B?

Definition 1.4: Conditional Probability

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

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \,. \tag{1.11}$$

■ **Example 1.4 (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:

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

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

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

■

Independent Events

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

Definition 1.5: Independent Events

Events A and B are said to be **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)\,. \tag{1.12}$$

■ **Example 1.5 (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 $A = \{(6,1),(6,2)\ldots,(6,6)\}$ and $B = \{(1,6),(2,6),\ldots,(6,6)\}$, so that $A \cap B = \{(6,6)\}$, and

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

■

For **multiple events** $A_1, A_2, \ldots, A_n$ are independent if and only if for any subset of these events $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})\,P(A_{i_2})\;\ldots\;P(A_{i_k}). \tag{1.13}$$

Product Rule

By the definition of conditional probability (1.11) we have

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

It is not difficult to generalize 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.

Theorem 1.6: Product Rule

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

$$\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1})\,. \tag{1.14}$$

■ **Example 1.6 (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 (1.14) is

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

■

1.7 Law of Total Probability and Bayes' Rule

Suppose that $B_1, B_2, \ldots, B_n$ is a **partition** of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω; see Figure 1.10.

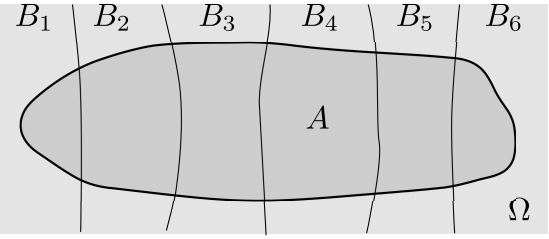


Figure 1.10: A partition $B_1, \ldots, B_6$ of the sample space Ω. Event A is partitioned into events $A \cap B_1, \ldots, A \cap B_6$.

A partitioning of the state space can sometimes make it easier to calculate probabilities via the following theorem.

Theorem 1.7

(Law of Total Probability). Let A be an event and let $B_1, B_2, \ldots, B_n$ be a partition of Ω. Then,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\, \mathbb{P}(B_i)\,. \tag{1.15}$$

Combining the law of total probability with the definition of conditional probability gives **Bayes' Rule**:

Theorem 1.8

(Bayes Rule). Let A be an event with $\mathbb{P}(A) > 0$ and let $B_1, B_2, \ldots, B_n$ be a partition of Ω. Then,

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\, \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \,|\, B_i)\, \mathbb{P}(B_i)}\,. \tag{1.16}$$

■ **Example 1.7 (Quality Control Problem)** A company has three factories (1, 2, and 3) that produce the same chip, each producing 15%, 35%, and 50% of the total production. The probability of a faulty chip at factory 1, 2, 3 is 0.01, 0.05, 0.02, respectively. Suppose we select randomly a chip from the total production and this chip turns out to be faulty. What is the conditional probability that this chip has been produced in factory 1?

Let B_i denote the event that the chip has been produced in factory i. The $\{B_i\}$ form a partition of Ω. Let A denote the event that the chip is faulty. We are given the information that $\mathbb{P}(B_1) = 0.15, \mathbb{P}(B_2) = 0.35, \mathbb{P}(B_3) = 0.5$ as well as $\mathbb{P}(A \mid B_1) = 0.01$, $\mathbb{P}(A \mid B_2) = 0.05$, $\mathbb{P}(A \mid B_3) = 0.02$.

We wish to find $\mathbb{P}(B_1 \mid A)$, which by Bayes' rule is given by

$$\mathbb{P}(B_1 \mid A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052 \,.$$

■

A simpler form of Bayes' rule is appropriate in some contexts. This can be derived directly from the Product Rule since

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

Hence

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B)\,\mathbb{P}(A \mid B)}{\mathbb{P}(A)} \,, \text{ provided } \mathbb{P}(A) \neq 0. \tag{1.17}$$

1.8 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.

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 5, we distinguish between discrete and continuous random variables: ☞ 84

- **Discrete** random variables can only take *countably many* values.
- **Continuous** random variables can take any value in a continuous interval or a union of disjoint intervals, such as the positive real line $\mathbb{R}_+$, *and* we have $P(X = c) = 0$ for every possible value c.

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 1.6: Cumulative Distribution Function

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 1.11 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.

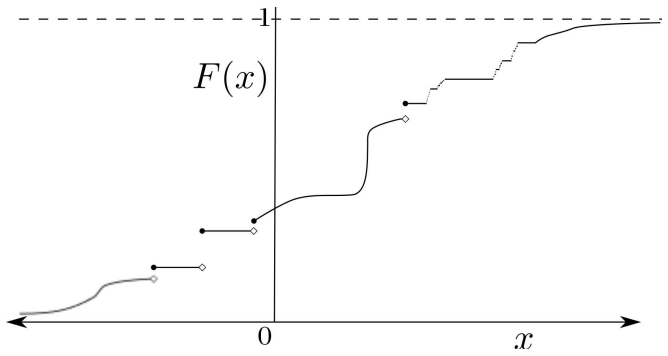


Figure 1.11: A cumulative distribution function (cdf).

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 1.7: Probability Mass Function

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
☞ 19 variable is to specify its pmf. Indeed, by the sum rule, if we know $f(x)$ for all x, then 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

$$\mathbb{P}(X \in B) = \sum_{x \in B} f(x) \,, \tag{1.18}$$

as illustrated in Figure 1.12. Note that $\{X \in B\}$ should be read as "X is an element of region B".

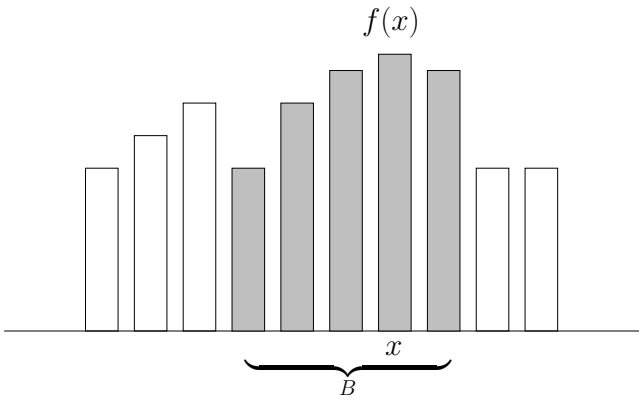


Figure 1.12: Probability mass function (pmf).

■ **Example 1.8 (Sum of Two Dice)** Toss two fair dice and let X be the sum of their face values. The pmf is given in Table 1.2.

Table 1.2: Pmf of the sum of two fair dice.

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

For a continuous random variable, it makes no sense to consider probabilities of the form $\mathbb{P}(X = x)$, as every such probability is zero! Instead of a probability mass function, we have to use a probability density function, which is defined as follows.

Definition 1.8: Probability Density Function

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 $a < b$,

$$\mathbb{P}(a < X \leqslant b) = F(b) - F(a) = \int_a^b f(u)\,\mathrm{d}u\,. \tag{1.19}$$

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

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 (1.18), once we know the pdf, we can calculate any probability that X lies in some set B by means of integration:

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x\,, \tag{1.20}$$

as illustrated in Figure 1.13.

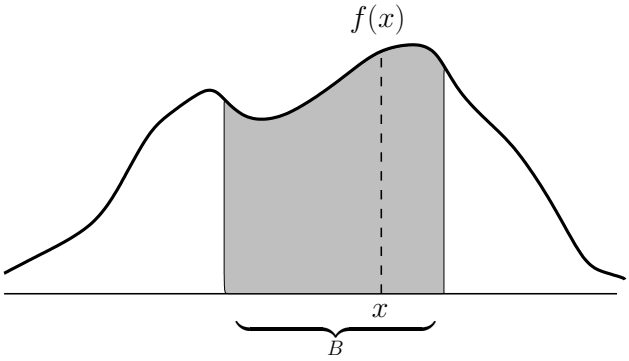


Figure 1.13: Probability density function (pdf).

Suppose that f and F are the pdf and cdf of a continuous random variable X, as in Definition 1.8. 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,

$$\mathbb{P}(x \leqslant X \leqslant x + h) = \int_{x}^{x+h} f(u)\,\mathrm{d}u \approx h\,f(x)\ . \tag{1.21}$$

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

■ **Example 1.9 (Random Point in an Interval)** Assume that we have a random variable X whose cdf is given below. What is the pdf for X?

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

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. Thus each point in $[0, 2]$ is equally likely to be drawn. Moreover, every subinterval of $[0, 2]$ of width a, with $0 < a < 2$, is equally likely. ■

Definition 1.9: Quantiles and Percentiles

Let X be a continuous random variable with cdf $F(x)$ and pdf $f(x)$, and p be a number between 0 and 1. Then the ***p*th quantile** or the **100*p*th percentile** of the distribution of X, denoted by $q(p)$, is defined by

$$p = F\left(q(p)\right) = \int_{-\infty}^{q(p)} f(u)du. \tag{1.22}$$

Assuming that $F()$ can be inverted, this can also be written as $q(p) = F^{-1}(p)$, which is called the **quantile function**.

1.9 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 1.10: Expectation (Discrete)

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

$$\mathbb{E}(X) = \sum_x x\,\mathbb{P}(X = x) = \sum_x x\,f(x)\,. \tag{1.23}$$

The expectation of X is sometimes written as μ_X. It is assumed that the sum in (1.23) is well-defined — possibly infinity (∞) or minus infinity ($-\infty$). One way to interpret the expectation is as a *long-run average payout*, as illustrated in the following example.

■ **Example 1.10 (Expected 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. ■

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 1.11: Expectation (Continuous)

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

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x\ . \tag{1.24}$$

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 1.9: 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 1.11 (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 1.9 is that the expectation is "linear".

Theorem 1.10: 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. $\mathbb{E}(g(X) + h(X)) = \mathbb{E}(g(X)) + \mathbb{E}(h(X))$.

Proof: 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 1.12: Variance

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

$$\mathrm{Var}(X) = \mathbb{E}(X - \mu)^2\ , \tag{1.25}$$

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 1.11

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

Proof: 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,

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

□

1.10 Exercises

1. Vinny from Vegas has received a "lucky" coin from "Tricky" Trish. Tricky has assured Vinny that the coin is biased towards heads. Vinny is skeptical, so decides to flip the coin eight times with the aim to investigate Tricky's claim.

   (a) How many possible outcomes are there for this random experiment?

   (b) Let X be the number of heads observed. Assuming the coin is fair, what is the probability distribution of X?

   (c) Calculate $\mathbb{P}(X = 7)$ for a fair coin.

   (d) Vinny conducts his experiment and finds 7 heads out of 8 tosses. What is the P-value for Vinny's test? What should he conclude?

2. Ten percent of people in King's Landing own both a dagger and a sword, whereas 30% owns neither. Of the people that own a dagger, 25% own also a sword. What is the probability that an arbitrarily selected person owns a sword?

3. We draw at random a point in a square. What is the probability that the point will lie in the right half of the square?

   Suppose we have extra information that the point lies below the line connecting the upper-left and bottom right corners. What is the conditional probability that the point will lie in the right half of the square, given this information?

4. We draw at random a point X in the interval [0,1], such that each point is equally likely. Give:

   (a) $\mathbb{P}(-1 < X < 0.5)$

   (b) $\mathbb{E}(X)$

   (c) If $Y = 2 + 3X$. What would the pdf of Y look like?

   (d) What is $\mathbb{E}(Y)$?

5. Suppose that the radius of a sphere, R, takes a value in the interval $[0.5, 1.5]$ with equal probability. Recalling the volume V of the sphere is given by $\frac{4}{3}\pi R^3$, determine the *expected* volume $\mathbb{E}V$.

6. We randomly take three balls *with* replacement from an urn containing 5 white and 5 black balls. What is the expected number of black balls?

7. We randomly take three balls *without* replacement from an urn containing 5 white and 5 black balls. What is the probability that two of the balls will be black?

8. A random variable X has probability density function f given by

$$f(x) = \begin{cases} c\,x, & \text{if} \quad x \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

   The constant c has to be equal to what?

9. One thousand possums are distributed randomly over 100 patches of forest (numbered from 1 to 100), so that each possum has a probability of 1/100 of ending up in any of the patches. What is the probability that patch number 23 will have 5 possums?

CHAPTER 2

COMMON PROBABILITY 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.

2.1 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 1.8. Recall that the distribution of a continuous variable is specified by its *probability density function* (pdf), and the distribution of a discrete random variable by its *probability mass function* (pmf).

☞ 28

In the following, we first present two distributions for discrete variables: they are the Bernoulli and Binomial distributions. Then, we describe two key 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 families (binomial, normal, etc.), we can make an infinite amount of distributions through parameter selection.

2.2 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 a population and asking them if they approve of the prime minister or not.

Definition 2.1: Bernoulli Distribution

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 2.1 gives the pmf of a Bernoulli random variable.

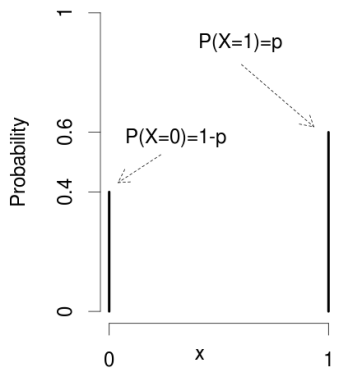


Figure 2.1: Probability mass function for the Bernoulli distribution, with parameter p (the case $p = 0.6$ is shown)

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
☞ 33 expectation and variance; see (1.23) and (1.25).

Theorem 2.1: Expectation and Variance of the Bernoulli Distribution

Let $X \sim \mathsf{Ber}(p)$. Then,

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

2.3 Binomial Distribution

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 1.5. Namely, if we note the 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

☞ 10
☞ 19

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

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 (1.14) to find that the probability of having a particular

☞ 26

sequence with x heads and $n - x$ tails is $p^x(1-p)^{n-x}$. 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.

Definition 2.2: 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

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n\,. \tag{2.1}$$

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

Figure 2.2 shows the pmf of the $\mathsf{Bin}(10, 0.7)$ distribution.

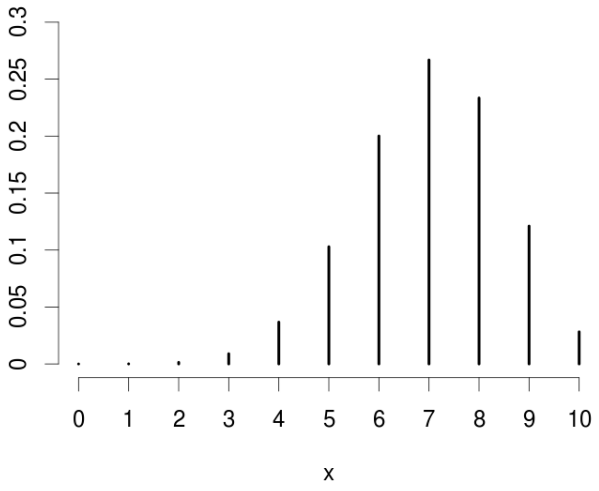


Figure 2.2: Probability mass function of the $\mathsf{Bin}(10, 0.7)$ distribution.

The following theorem lists the expectation and variance for the $\mathsf{Bin}(n, p)$ distribu-
☞ 61 tion. A simple proof will be given in the next chapter; see Example 3.4. In any case, the 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 of a $\mathsf{Ber}(p)$ random variable. This is no coincidence, as a Binomial random variable can be seen as the sum of n independent Bernoulli random variables.

Theorem 2.2: Expectation and Variance of the Binomial Distribution

Let $X \sim \mathsf{Bin}(n, p)$. Then,

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

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 realize 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 sex. 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 1.5.

☞ 19

2.4 Uniform Distribution

The simplest continuous distribution is the uniform distribution.

Definition 2.3: Uniform Distribution

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

$$f(x) = \frac{1}{b-a}, \quad a \leqslant x \leqslant b \qquad \text{(and } f(x) = 0 \text{ otherwise).} \tag{2.2}$$

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

A random variable $X \sim \mathcal{U}[a, b]$ 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 2.3. Note that the total area under the pdf is 1.

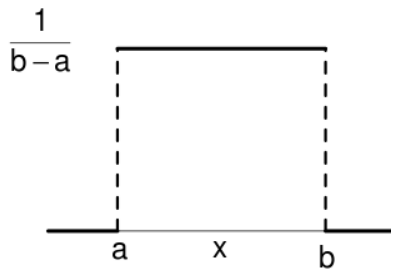


Figure 2.3: Probability density function for a uniform distribution on $[a, b]$

Theorem 2.3: Properties of the Uniform Distribution

Let $X \sim \mathcal{U}[a, b]$. 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 $U \sim \mathcal{U}[0, 1]$. In words: randomly choosing a point between a and b is equivalent to first randomly 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 1.9. 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:

☞ 35

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

□

2.5 Normal Distribution

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

Definition 2.4: Normal (or Gaussian) Distribution

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

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; x \in \mathbb{R} \tag{2.3}$$

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

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

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

The corresponding cdf is denoted by Φ (capital phi). In Figure 2.4 the density function of the $\mathcal{N}(\mu, \sigma^2)$ distribution for various μ and σ^2 is plotted.

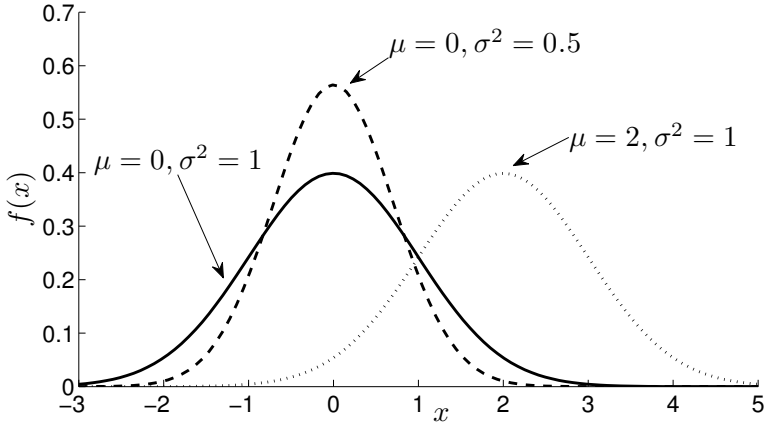


Figure 2.4: Probability density functions for various Normal distributions

You may verify yourself, by applying the definitions of expectation and variance, that indeed the following theorem holds:

Theorem 2.4: Properties of the Normal Distribution

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then,

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

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 2.4). 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 2.5: Standardization

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

$$Z = \frac{X - \mu}{\sigma} \tag{2.4}$$

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 (2.3). 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 $\mathcal{N}(\mu, \sigma^2)$-distributed random variable (replace x with $(x - \mu)/\sigma$ in the formula for φ and divide by σ. This gives precisely (2.3)). □

By using the standardization (2.4) we can simplify calculations involving arbitrary normal random variables to calculations involving only standard normal random variables.

■ **Example 2.1 (Standardization)** Standardization 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 sex, a female who is taller than 180 cm or a male who is taller than 200 cm? 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 $X \sim \mathcal{N}(\mu, \sigma^2)$:

- 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%

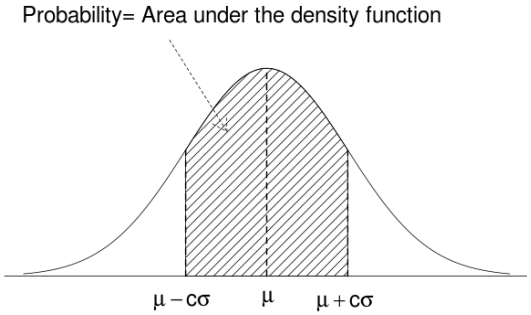


Figure 2.5: The area of the shaded region under the pdf is the probability $\mathbb{P}(|X - \mu| \leqslant c)$ that X lies less than c standard deviations (σ) away from its expectation (μ)

The function **pnorm** can also be used to evaluate the cdf of general normal distribution. For example, let $X \sim \mathcal{N}(1, 4)$. 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 $X \sim \mathcal{N}(1,4)$. For what value z does it hold that $\mathbb{P}(X \leqslant z) = 0.9$, i.e. $\Phi(z) = 0.9$? 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
☞ 87 *sample quantile* discussed in Section 5.4, but the two are not the same. For the normal distribution the quantiles can be obtained via the R function **qnorm**.

In statistical inference, we need the values which capture small tail areas under the standard normal curve. z_α denotes the value of z for which the area under the standard normal (z) curve to the *right* of z_α is α. Since the area under any pdf is 1, the area to the *left* of z_α under the standard normal curve is $1-\alpha$. So z_α is the $1-\alpha$ quantile of the standard normal distribution or $\Phi(z_\alpha) = 1-\alpha$. Figure 2.6 gives an illustration.

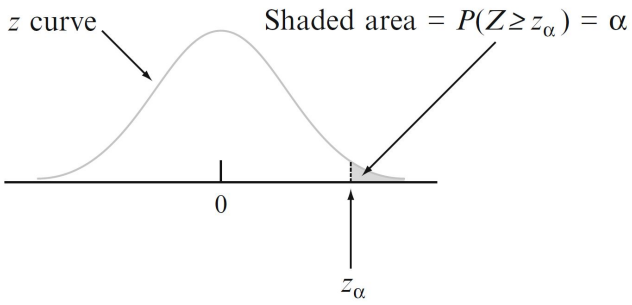


Figure 2.6: Illustration of z_α notation
(from Devore and Berk p184)

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
```

2.6 Simulating Random Variables

This section shows how to generate (simulate) 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
```

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 realization 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 realization 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 realizations from a certain distribution, then the histogram or density plot will resemble closely the true pdf/pmf of the distri-
☞ 87
bution. Moreover the summary statistics (see Section 5.4) 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 $\mathcal{N}(2, 1)$ 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 2.7. 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
```

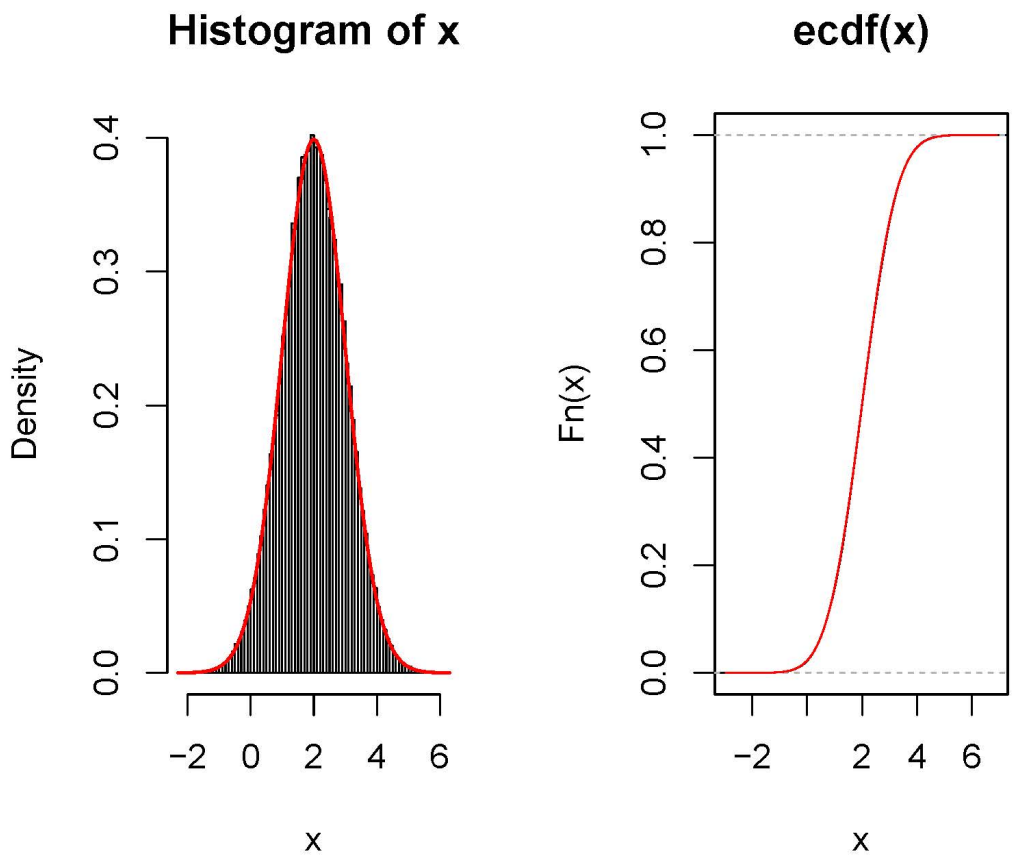


Figure 2.7: Left: pdf of the $\mathcal{N}(2,1)$ distribution (red) and histogram of the generated data. Right: cdf of the $\mathcal{N}(2,1)$ distribution (red) empirical cdf of the generated data.

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 2.1 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 2.1: 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

$\mathsf{Ber}(p)$	`dbinom(x,size=1,prob=`p`)`
	`pbinom(x,size=1,prob=`p`)`
	`qbinom(`γ`,size=1,prob=`p`)`
	`rbinom(n,size=1,prob=`p`)`

$\mathsf{Bin}(n, p)$	`dbinom(x,size=`n`,prob=`p`)`
	`pbinom(x,size=`n`,prob=`p`)`
	`qbinom(`γ`,size=`n`,prob=`p`)`
	`rbinom(n,size=`n`,prob=`p`)`

Distr.	R functions

$\mathcal{N}(\mu, \sigma^2)$	`dnorm(x,mean=`μ`,sd=`σ`)`
	`pnorm(x,mean=`μ`,sd=`σ`)`
	`qnorm(`γ`,mean=`μ`,sd=`σ`)`
	`rnorm(n,mean=`μ`,sd=`σ`)`

$\mathcal{U}[a, b]$	`dunif(x,min=`a`,max=`b`)`
	`punif(x,min=`a`,max=`b`)`
	`qunif(`γ`,min=`a`,max=`b`)`
	`runif(n,min=`a`,max=`b`)`

2.7 Exercises

1. In a city there are 100,000 couples. How many couples do you expect there to be in which both partners have the same birthday?

2. Let $X \sim \mathsf{Bin}(10, 1/4)$.

   (a) What is $\mathbb{E}(X)$ and $\mathrm{sd}(X)$?

   (b) What is $\mathbb{P}(X = 3)$?

   (c) Calculate $\mathbb{P}(X \leqslant 3)$ exactly.

3. Let X have a normal distribution with expectation 3 and variance 4. Using the standard normal distribution, what is the 95% quantile of X?

4. We randomly select 100 people from a large population in which 10% are left-handed. Using the standard normal distribution, what is the probability that 15 or more of these people are left-handed?

5. Suppose the maximum temperature on a random day in December in Brisbane follows a normal distribution with mean (i.e., expectation) 32 and standard deviation 2 (degrees Celsius).

   (a) What is the probability that the temperature exceeds 36 degrees on a given day?

   (b) A temperature of 36 degrees is considered dangerous for the elderly and infants. Suppose that due to global warming the mean temperature rises to μ degrees, while the standard deviation remains 2 degrees. What μ would increase the probability of exceeding 36 degrees to 10%?

CHAPTER 3

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.

3.1 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 in Brisbane over a given month. 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 realizations 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 behaviour 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 1.8.

☞ 28

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 generalization.

Definition 3.1: Joint Cumulative Distribution Function

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.

3.2 Joint Distributions

■ **Example 3.1 (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 3.1 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 3.1. Similarly, the distribution of Y is given by the column totals in the last row of the table.

Table 3.1: The joint distribution of X (die number) and Y (face value).

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

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

Definition 3.2: Joint Probability Mass Function

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

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{\mathbf{x} \in B} f(\mathbf{x}) \,. \tag{3.1}$$

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 generalization of (1.18). In particular, as illustrated in Example 3.1, 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 pmf over all possible values of the other variables. For example,

☞ 30

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y) \,. \tag{3.2}$$

The converse is not true: from the marginal distributions one cannot in general reconstruct the joint distribution. For example, in Example 3.1 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 3.3: Joint Probability Density Function

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

$$\mathbb{P}(\mathbf{X} \in B) = \int_{\mathbf{x} \in B} f(\mathbf{x})\, \mathrm{d}\mathbf{x} \quad \text{for all sets } B\,. \tag{3.3}$$

The integral in (3.3) is now a multiple integral — instead of evaluating the area under f, we now need to evaluate the (n-dimensional) volume. Figure 3.1 illustrates the concept for the 2-dimensional case.

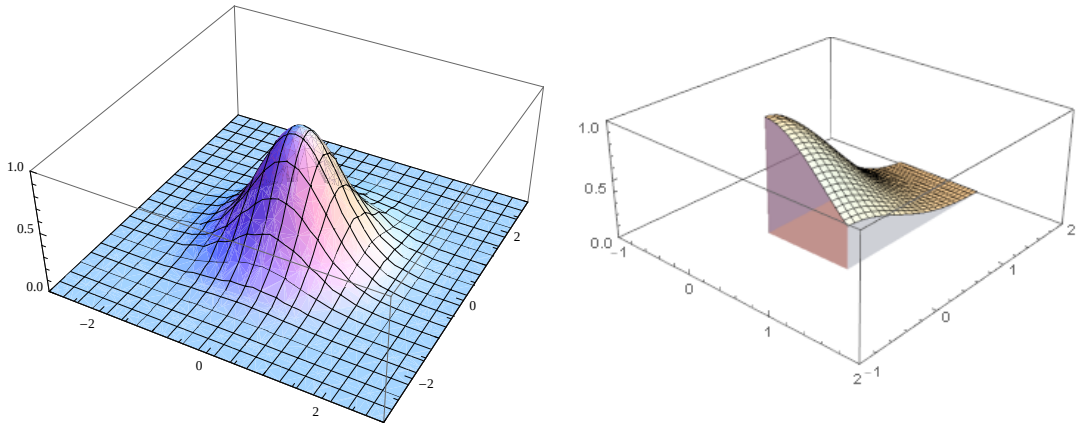


Figure 3.1: Left: a two-dimensional joint pdf of random variables X and Y. Right: the volume under the pdf corresponds to $\mathbb{P}(0 \leqslant X \leqslant 1, Y \geqslant 0)$.

3.3 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
☞ 26 (1.12). 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 3.4: Independent Random Variables

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

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{3.4}$$

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

■ **Example 3.2 (Dice Experiment Continued)** We repeat the experiment in Example 3.1 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. ■

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

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

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

■ **Example 3.3 (Bivariate Standard Normal Distribution)** 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

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

The graph of this joint pdf is the hat-shaped surface given in the left pane of Figure 3.1. We can also simulate independent copies $X_1, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(0, 1)$ and $Y_1, \ldots, Y_n \sim_{\text{iid}} \mathcal{N}(0, 1)$ 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 3.2.

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

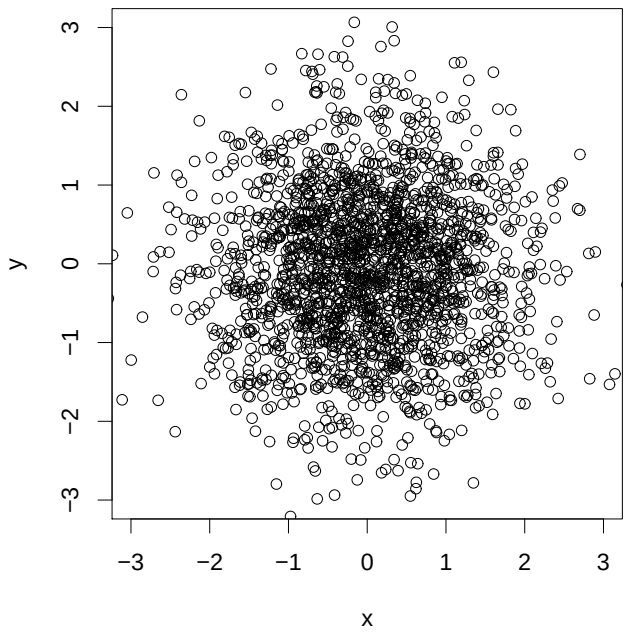


Figure 3.2: Scatter plot of 2000 points from the bivariate standard normal distribution.

We see a "spherical" pattern in the data. This is corroborated by the fact that the joint pdf has contour lines that are circles. ■

3.4 Expectations for Joint Distributions

☞ 34 Similar to the univariate case in Theorem 1.9, the expected value of a real-valued function h of $(X_1, \ldots, X_n) \sim f$ is a weighted average of all values that $h(X_1, \ldots, X_n)$ can take. Specifically, in the discrete case,

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \sum_{x_1, \ldots, x_n} h(x_1, \ldots, x_n)\, f(x_1, \ldots, x_n), \tag{3.5}$$

where the sum is taken over all possible values of $(x_1, \ldots, x_n)$. 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 3.1: Properties of the Expectation

Let $X_1, \ldots, X_n$ be random variables with expectations $\mu_1, \ldots, \mu_n$. Then,

$$\mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] = a + b_1\mu_1 + \cdots + b_n\mu_n \tag{3.6}$$

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

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mu_1\,\mu_2 \cdots \mu_n\ . \tag{3.7}$$

Proof: 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 (3.5),

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

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 3.5: 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**:

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\,\sigma_Y}\ , \tag{3.8}$$

where $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$. The correlation coefficient always lies between -1 and 1.

Theorem 3.2 lists some important properties of the variance and covariance.

Theorem 3.2: 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. ☞ 35 Properties 1 and 2 were already shown in Theorem 1.9.

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 the following more general result.

Theorem 3.3: Variance for Linear Combinations of Random Variables

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,

$$\mathrm{Var}(a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n) = b_1^2 \sigma_1^2 + \cdots + b_n^2 \sigma_n^2 \tag{3.9}$$

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

■ **Example 3.4 (Expectation and Variance for the Binomial Distribution)** We now show a simple way to prove Theorem 2.2; 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

☞ 43

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

Since we have seen that each Bernoulli variable has expectation p and variance $p(1-p)$, we have by Theorem 3.1 that

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

and by Theorem 3.3 that

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

as had to be shown. ■

3.5 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 (3.6) and (3.9) 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.

Theorem 3.4: Law of Large Numbers

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 3.5 (Square Root of a Uniform)** Let $U \sim \mathcal{U}(0,1)$. 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, 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.

Theorem 3.5: Central Limit Theorem

The sum of a large number of iid random variables approximately has a normal distribution. More precisely:
Central limit theorem (CLT) for sums: If $X_1, \ldots, X_n$ are iid with finite expectation μ and variance $\sigma^2 > 0$, then for large n:

$$X_1 + \cdots + X_n = S_n \overset{\text{approx.}}{\sim} \mathcal{N}(n\mu, n\sigma^2).$$

Central limit theorem (CLT) for means (conditions as above):

$$\frac{X_1 + \cdots + X_n}{n} = \bar{X} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

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.

☞ 43 Let X_1 be a $\mathcal{U}[0,1]$ random variable. Its pdf (see Section 2.4) 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 3.3.

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

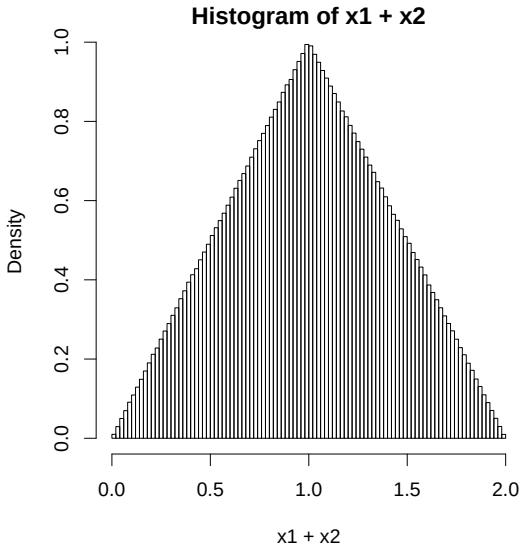


Figure 3.3: Histogram for the sum of 2 independent uniform random variables.

The pdf seems to be triangle-shaped and this is not so difficult to show. Now let us do the same thing for sums of 3 and 4 uniform numbers. Figure 3.4 shows that the pdfs have assumed a bell-shaped 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.

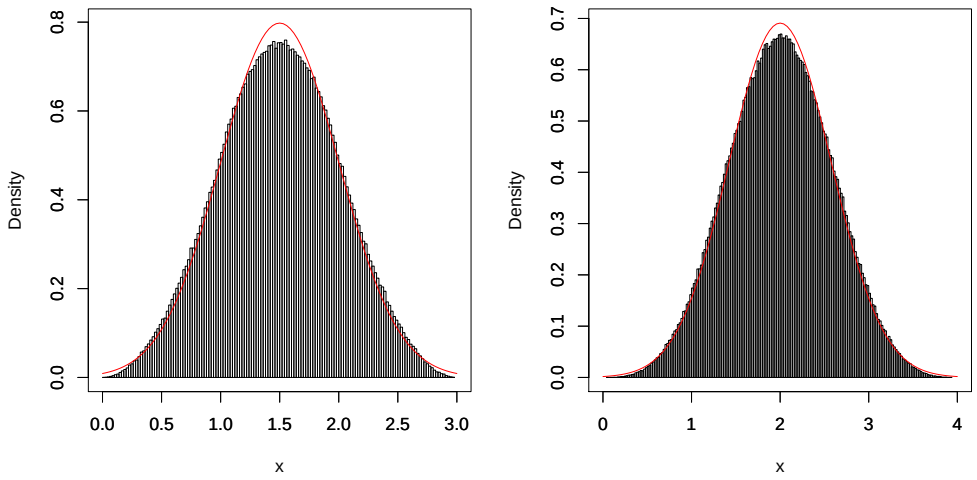


Figure 3.4: The histograms for the sums of 3 (left) and 4 (right) uniforms are in close agreement with normal pdfs.

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 $Y \sim \mathcal{N}(np, np(1-p))$. 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 $b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$ 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 this course, but the central limit result should give you some confidence that it is true. And you can verify particular cases yourself via simulation. Thus, the following theorem is one of the main reasons why the normal distribution is used so often in statistics.

Theorem 3.6: Linear Combinations of Normals are Again Normal

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

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

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 Theo-

rems 3.1 and 3.3.

3.6 Exercises

1. The weight (actually, mass!) of a randomly selected person has a normal distribution with expectation 75 kg and standard deviation 10 kg.

   (a) What is the probability that the weight is greater than 90 kg?

   (b) What is the probability that two persons together weigh more than 180 kg?

2. The distribution of the weight (in grams) of eggs given by the Easter Bunny is normal with mean 60 and standard deviation 6.

   (a) Hillary collects ten such eggs. What is the expectation and standard deviation of the total weight of the eggs in her basket? Provide working, using expectation and variance rules.

   (b) What is the probability that a single egg will weigh more than 63 grams?

   (c) What is the probability that the collection of ten eggs will weigh more than 630 grams?

3. A large survey of shoppers at a Brisbane beach clothing retailer were used to estimate the following probabilities of visiting a beach for 0, 1 or 2 days over the first weekend of December, and whether or not these shoppers liked to swim in the sea. Let X be the number of days a randomly chosen shopper at this retailer would report that they were going to visit a beach over this weekend. Let Y represent whether or not a randomly chosen shopper at this retailer would report that they liked to swim in the sea, with 1 representing that they do like it and 0 representing that they don't. The following table lists the probabilities for each possible combination of X and Y.

Y \ X	0	1	2
0	0.420	0.175	0.105
1	0.180	0.075	0.045

   (a) Determine the marginal probability mass function of X.

   (b) Determine the marginal probability mass function of Y.

   (c) Determine the expectation and standard deviation of X, and separately, of Y.

   (d) Determine the covariance and correlation between X and Y.

   (e) Are X and Y independent? Give reasons for your answer.

4. Harry writes down $f(x, y) = 2e^{-y-x}e^{-y}$ and claims that it is the joint pdf on $[0, \infty)^2$ of independent random variables.

   (a) Verify or debunk Harry's claim.

   (b) Determine $\mathbb{E}[XY]$.

5. Let $X \sim \mathsf{Ber}(p)$ and $Y \sim \mathsf{Ber}(q)$ be independent and let $Z = XY$ be their product. What is the variance of Z?

6. Suppose that you have two independent but unreliable temperature gauges, one which returns a measurement X in degrees Celsius that is normally distributed with mean equal to the true temperature and standard deviation of 3 degrees Celsius. The other is in degrees Fahrenheit and returns a measurement Y that is normally distributed with mean equal to the true temperature at standard deviation of 10 degrees Fahrenheit. Note that Y can be converted to Celsius via the formula
$$\widetilde{Y} = \frac{5}{9}(Y - 32) \ .$$
In the anticipation that it will make your temperature reading more accurate, you decide to combine the measurements (in degrees Celsius) as the average of the two: $Z = (X + \widetilde{Y})/2$.

   (a) What is the standard deviation of $\widetilde{Y}$?

   (b) What is the standard deviation of Z (in degrees Celsius)? Is Z indeed more accurate than either X or $\widetilde{Y}$?

7. A pair of random variables (X, Y) is drawn uniformly from the unit circle.

   (a) What is the joint pdf of X and Y?

   (b) What are the marginal pdfs of X and Y?

   (c) Calculate $\mathrm{Cov}(X, Y)$.

   (d) Are X and Y independent? Explain why or why not.

CHAPTER 4

STUDIES, DATA, AND EVIDENCE

The aim of this chapter is to give a short introduction to the statistical reasoning that we will be developing during the second part of this course. We will discuss the typical steps taken in a statistical study, give an overview of three simple statistical models, emphasize the distinction between observational and designed statistical experiments, introduce you to the language of hypothesis testing, and look at several data collection methods.

4.1 Statistical Modeling

Let us now return right to the beginning of these notes, to the steps for a statistical study in Section 1.2. Figure 4.1 gives a sketch of the conceptual framework for statistical 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.

☞ 9

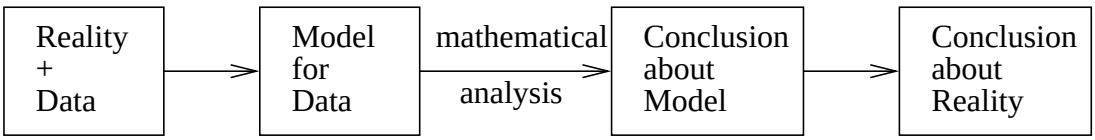


Figure 4.1: Statistical modeling and analysis.

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 example.

☞ 57

■ **Example 4.1 (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,

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

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

A simple generalization 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 4.2 (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 4.3 (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**:

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

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 4.2, using parameters $\mu_1 = 170, \mu_2 = 175, \sigma_1^2 = 200$, and $\sigma_2^2 = 100$.

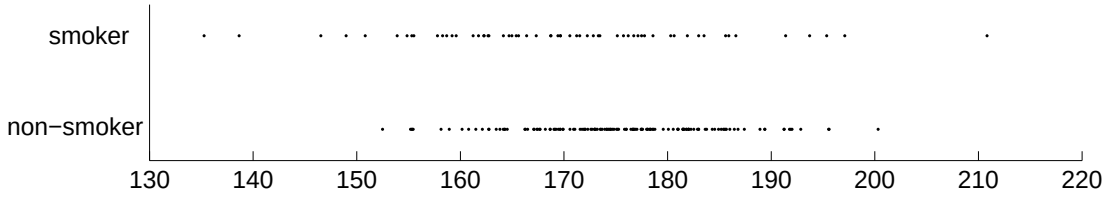


Figure 4.2: Simulated height data from a two-sample normal model.

Remark 4.1 (About Statistical Modeling) At this point it is good to emphasize a few points about statistical modeling.

- *Any* model for data is likely to be *wrong*. For example, in Example 4.3 the height would normally be recorded on a discrete scale, say 1000 – 2200 (mm). However, samples from a $\mathcal{N}(\mu, \sigma^2)$ 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.

4.2 Data

Data comes in many shapes and forms, but is often represented as a spreadsheet in a "standard format", where columns represent features or **variables** such as height, sex, and income, and rows represent individuals or units. Some of the most common and useful forms of datasets can be considered samples from a population of interest.

The data in a spreadsheet could be the result of an **observational study**, where we have no control over each variable (corresponding to a column in a spread sheet). A typical example is survey data. If we would repeat the whole study, the values in all columns would change.

Alternatively, The data in a spreadsheet could be the result of a **designed experiment**, where certain experimental conditions (variables) are controlled (fixed) to reduce unwanted variability in the measurements. If we would repeat the whole experiment, the experimental variable columns would stay the same.

Other forms of data might fit neither of the above categories, and this may limit our ability to draw meaningful conclusions about any underlying population.

If we have data from an observational study or a designed experiment, some measurements would *change* if the data were collected again. There may be various causes of variability/randomness in measurements. A great deal of variability could be due to the inclusion of different individuals or units in the data. Even if the same individuals or units were measured again, there could be changes due to time, other effects or measurement error.

For example, in height data the main source of variability is the natural diversity of heights in a population due to differences in genetics, diet and environmental factors. Another source of variability is the measurement variability (due to limitations in how accurately we can measure each height). Later on in this course we will consider statistical models that aim to explain the variability in the data. For example, we could try to explain the variability in heights not only via the natural variability in the population, but also taking into account variables such as sex, ethnicity, and shoe size. Any remaining variability that cannot be explained by the model is called the **residual variability**. Ideally, a model will produce small residual errors and predict new data well.

Among the set of variables measured, one or more will be of particular interest to us. We would like to explain the values of these variables in terms of the values of other variables and maybe predict their value based on the other variables. The variables we wish to explain are called **response variables** (or sometimes dependent, endpoint, outcome or output variables). The remaining variables are called **explanatory variables** (or sometimes independent variables, predictors, covariates, factors or inputs). We may believe that the value of the explanatory variable(s) can cause much of the value of the response variable(s). This is difficult to show, but evidence can be found via randomised experimental designs.

In a study, we will collect one or more measurements from each experimental unit. This might often be a person, but could instead be for example a plant, an area of land, a household, a star or a manufactured object.

A randomised experimental design involves each experimental unit being given one of a number of possible treatments (e.g. new treatment and old treatment), with the choice of treatment determined randomly, with equal probability of any experimental unit receiving any of the available treatments. The random assignment of treatments to subjects largely eliminates bias due to differences in the types of subjects in each treatment group. Each treatment group should then be similar in all respects before the treatments are applied. In particular, the distribution of any variable should be similar in all treatment groups. Importantly, this is true for all variables which have not been measured, as well as those that have.

For large enough studies, this ensures that any differences seen between the treatment groups must be due to the difference in treatments. Hence we can make causative conclusions from these studies about the effect of the treatment on the response variable(s). These types of studies are often called randomised controlled trials (RCTs) in medicine and form the basis of most medical evidence. You can read more about this if interested at the Cochrane Library website `https://www.cochranelibrary.com/` .

While a randomised controlled trial design is desirable, it is not always feasible or affordable. In medical research, business and government, there are often large amounts of existing data which can potentially be analysed for new purposes. This would form a retrospective study, with no opportunity to assign treatments to experimental units. Assuming no-one assigned anything originally, this would be an observational study. Observational studies can also be prospective, that is, planned, with measurements carried out later on. An example is the series of US Nurses' Heath studies in which over a hundred thousand US nurses responded to multiple questionnaires on a range of health topics beginning in 1976. There are many attributes which cannot be assigned to participants randomly due to physical or ethical barriers. Examples include sex, beliefs and age. Comparisons between groups defined by categories which have not been assigned randomly to experimental units can at most show association between that variable and the response variable(s).

One major difficulty with observational studies is the presence of confounding variables, that is, those which have not been measured or recorded, but which have an important effect on the relationship between the explanatory and the response variables. For example, we might study the relationship between height and hair length and conclude that taller people tend to have shorter hair. However, we would have ignored gender, with females on average having longer hair than males and also on average being shorter than males. Within each gender, we might find no relationship between height and hair length. So the original relationship found was confounded by gender. The proportions of people above a certain height (e.g. 180 cm) would be quite different in the male and female groups. These issues don't tend to arise in well executed randomised experiments since all treatment groups have similar distributions of all variables.

Analysis of data from randomised experimental studies is often fairly straightforward, with the treatment groups needing to be compared against each other. Techniques suitable for this will be covered in chapters 6, 7 and 8. Analysis of data from observational studies is often more difficult and needs methods such as those taught in chapters 9 and 10.

4.3 Designed Experiment: Alice's Caffeine Data

In this section we consider a simple designed experiment, to which we will come back several times during this course, and which we will refer to as Alice's Caffeine Data

experiment. Alice's research question is: does drinking caffeinated cola increase the heart rate compared to drinking decaf cola? To answer this question, she designed the following experiment:

- Measure the heart rates of 20 friends as subjects, using a pulse meter.
- Give 10 friends 250mL of *caffeinated* Diet Coke while the other 10 friends are given 250mL of *caffeine-free* (decaf) Diet Coke.
- Wait half an hour after the drink, and then measure the heart rates again.
- Record the *difference* in heart rates for each subject.

Even for such a seemingly simple study, there are a lot of design issues to think about. Let us discuss a few points.

1. Alice chose 20 subjects in her study. Is this a sufficient sample size? There are many considerations for choosing a "good" sample size. In general, the sample size is determined by (a) the size of the effect that we are trying to detect and (b) the variability in the data. If the data has little variability, it may suffice to only use a small sample size to detect a certain effect. In contrast, if the data exhibits a large amount of variability it may require a large sample size to detect any effects, especially if they are small. However, bigger sample sizes do not always make for better experiments. Running an overly large study often leads to poor quality of data, as it is difficult to enforce compliance to the study protocols at all levels of the study. A large sample size may also be impractical, potentially dangerous (e.g., for experimental medical treatments), costly, or time-consuming. Think of all the cola Alice would have to buy for a study with 1000 individuals!

2. Alice chose her friends as test subjects. Is that fair? Let us go back to the research question: to detect if caffeinated cola increases the heart rate. If the population of interest is not just Alice's friends, but the general population, then choosing the subjects within her circle of friends may introduce a sampling bias. For example, suppose that the effect of caffeine would depend on age and sex. If most of Alice's friends are between 19 and 22 years old and female, the sample group would no longer be representative of the general population. Any conclusions from this study would pertain to the smaller population of people that are similar to Alice's friends. But maybe the effect of caffeine does not depend on factors such as age, weight, sex, or the subject's cola drinking behaviour, and then the conclusions might be applicable to a larger population.

   There is another possible source of bias in Alice's experiment. Recall that she gives 10 friends caffeinated Diet Coke and 10 friends decaf Diet Coke. How are the two groups chosen? Perhaps she divides the groups into 10 males and 10 females. But this could lead to a bias in the results, if the difference in heart rate would depend on sex. To avoid any bias in the group selection, we can randomly

select the treatment and placebo group, by using the random number generator of R, for example. Such a **randomization** process is an important ingredient in many designed experiments.

Another issue is whether Alice's friends know if they are getting caffeinated or decaf cola. This may influence the measurements (increase in heart rate) due to the person's reaction to their knowledge of the treatment they have received and their subsequent personal reaction to that information. In a **blind experiment** the subjects do not know which treatment they have received. Even in this case, the experimenter may inadvertently influence the outcome if they know which subjects are assigned which treatments. To also remove this bias, the gold-standard procedure is to employ a **double-blind experiment**, where the experimenters also do not know how the treatments are distributed over the subjects.

In medical studies, it is common to compare a new treatment (e.g. a drug) against a **placebo** - a control treatment which resembles the new treatment in every manageable way, but is missing the crucial new ingredient. In a drug trial, the placebo might be a harmless sugar pill. In some studies, subjects who were unknowingly given the placebo improved with respect to the main measurement of interest, e.g. self-reported pain levels, and this is called the **placebo effect**.

3. Why did Alice wait half an hour after consuming drinks before measuring the pulse rates again? In this experiment, Alice used "subject-specific" knowledge. Namely, she did a thorough literature search on the average time it takes for humans to absorb and metabolize caffeine in drink form — most sources claimed this to be around 20–30 minutes.

4. Why did Alice choose to compare Diet Coke with decaf Diet Coke instead of the regular (non-Diet) cola? The reason is given in Figure 4.3: the only difference between caffeinated Diet Coke and decaf Diet Coke is the amount of caffeine. In contrast, caffeinated Regular Coke and decaf Regular Coke have, in addition to the caffeine content, different energy, protein, carbohydrate, and sodium contents. So a change in heart rate could be the result of the sugar content, for example, rather than caffeine content. We say that decaf Diet Coke serves as a suitable **control** for caffeinated Diet Coke, as their only difference is the caffeine content.

Let us examine Alice's data and introduce some experimental design terminology on the way. Alice's experiment involves actively applying treatments to subjects and observing their responses. An experimental **treatment** is a combination of factors set at certain levels. The variables describing the treatments are the explanatory variables in the study. The response from an experiment is the variable(s) of interest.

For Alice's experiment, the response variable is the change in pulse rate, and the explanatory variable is the caffeine content, which is considered at two levels (yes and no). The resulting changes in pulse rate are given in Table 4.1:

http://www.coca-cola.com.au/ourdrinks/nutrition-comparison-tool.jsp

Figure 4.3: Nutritional information for caffeinated/decaf Regular Coke (top) and caffeinated/decaf Diet Coke (bottom).

Table 4.1: Changes in pulse rate for Alice's caffeine experiment.

Caffeinated	17	22	21	16	6	−2	27	15	16	20
Decaf	4	10	7	−9	5	4	5	7	6	12

We mentioned in Section 4.2 that when using software such as R to analyse and display data it is important that the data is represented/stored in a standard format. Table 4.1 is not in a standard format. To convert it to standard format, we should store the pulse beat changes of all 20 subjects (Alice's friends) in a single column — called `pb`, for example. And a second column, `Caffeine`, indicates whether a subject

receives the caffeinated or decaf cola. Of course such a table is very tall and skinny and does not present as well on paper as Table 4.1.

Do the results in Table 4.1 provide any *evidence* that caffeine increases pulse rate? Let us now go through the same 6 steps of a statistical study as in Section 1.2. The first two steps (designing the study and collecting the data) have already been discussed above.

Step 3 is about visualizing and summarizing data. Figure 4.4 shows a possible visualization of the data in a so-called stripplot.

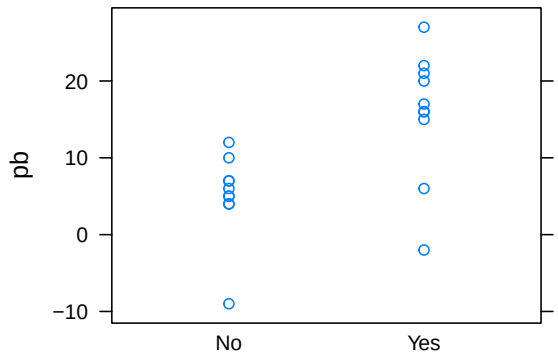


Figure 4.4: A visualization of Alice's caffeine data.

It was made using the following code:

```
> alice = read.csv("alice.csv")
> library(lattice)
> stripplot(pb ~ Caffeine, data=alice)
```

Two key summaries of the data are:

- The mean (i.e., average) increase in pulse rate for the decaf group is 5.1 bpm.
- The mean increase in pulse rate for caffeine group is 15.8 bpm.

The *group difference* is thus $15.8 - 5.1 = 10.7$ bpm. Is this evidence that caffeine increases pulse rate? In other words, if we summarize our data via the group difference, what evidence is there that the group difference 10.7 was due to the effect of the caffeine presence, rather than this happening by chance? We will consider this in chapters 6 onwards.

4.4 Sampling from a Population

We'll now consider how to obtain data. This is important for both observational and experimental studies. Most statistical theory assumes we are working with a simple

random sample. That is, if the population of interest is finite, say of size N, and we want to collect n observations from it, then each member of the population has the same probability of being included in the sample. Collecting data will cost a certain amount of time and/or money per experimental unit and you have a limited budget. Maybe you have enough resources to afford $n = 200$ observations from a population of size $100,000$. How do you get this done? If the measurements can all be done over the phone or internet, the physical location of each experimental unit may not matter. In this case, we could take a true random sample from the population. In R, we can set up a list with the entries 1 to $100,000$, then collect a random sample from this (without replacement) to find the chosen indexes of units (e.g. people) we will contact.

```
> indexes <- seq(1:100000)
> chosen <- sample(indexes, 200)
> sort(chosen)
```

Try running all of the above commands. You will obtain a completely different set of 200 randomly chosen index values each time. There's no need to sort the values, but it might make the task easier in some contexts. These numbers are not truly random like radioactive decay. However, they are produced by excellent pseudo-random number generator algorithms which, for our purposes, behave in the same way as true random number generation.

In many cases, location will matter greatly for the cost of data collection. Imagine you are measuring people and you had to visit each person in your sample all over Australia, with many being the only person on your list from a given area. This would be very expensive and slow, and we'd be keen to consider alternatives. What are these and what compromises do they entail?

One option is to try to visit a limited number of locations and obtain a number of observations at each location. This has the potential to be misleading, since most locations won't be visited at all. However, the price will be constrained and we can potentially use a sampling scheme to give all members of the population an equal chance of being included, and/or other desirable statistical properties. Two classic approaches to more affordable or flexible random sampling are:

1) Stratified sampling. We start by breaking the population up into subgroups called strata. Some examples of strata would be capital city and regional locations, males and females, and age groups. If a stratum has population size N_s, then we would usually take nN_s/N observations from it via simple random sampling. However, if we needed a higher level of accuracy from a particular stratum, we could oversample it, i.e. collect more observations from it. To implement stratified sampling, we need accurate population counts for each stratum, and an ability to take a simple random sample from each.

2) Cluster sampling. Here we divide the population into many fairly small clusters, such as the people living in one city block. We then select a set of clusters at random, and then either collect measurements from every individual in each selected cluster or take a simple random sample from each cluster.

A drawback of each method is that the assumption that all the individual measurements taken are independent of others is then violated to some extent because people within a cluster might have something in common which is not in common with the rest of the population. This tends to lead to a slightly higher variance in estimates than would have been seen in data from a simple random sample. However, being able to collect more data for the same amount of money and/or time usually results in more precise estimates overall. There are specific models and methods to analyse data from cluster sampling, but we will not cover them here. If you decide to use either of these methods in your project, you can analyse the data as if it came from a simple random sample.

■ **Example 4.4 (Villagers)** Assume that the population of interest all live in 10 small villages, with the Eastern 5 villages having 50 people each and the Western 5 villages having 100 people each. The total population size is $N = 5 * 50 + 5 * 100 = 750$. Assume we would like to collect a sample of size $n = 100$ from this population.

If using stratified sampling, we might choose East and West as two strata, noting that these are of size 250 and 500, respectively. From the Eastern villages, we would then need to take a simple random sample of size $100 * 250/750 \approx 33.3$, which we'd round to 33. From the Western villages, we'd need the remainder, which is $100 - 33 = 67$. It's also $100 * 500/750 \approx 66.7$, rounded to 67. A simple random sample from each stratum would likely result in having some participants chosen from every village. While this is quite fair and representative, it would be time consuming and expensive.

If using cluster sampling, we could make each Eastern village a cluster (size 50) and could split each of the larger Western villages into two clusters of 50 each. That would give 15 clusters. We might then choose a random two clusters from these 15 and fully sample each cluster. This would mean visiting at most two villages and ignoring the rest. We'd obtain our desired sample size and limit travel time, but not have seen the full variety of villages. We could take a less-extreme option and choose say four clusters, taking us to at most four villages. We would then take a random sample of half of each cluster, resulting in $4 * 25 = 100$ observations. We have to trade off efficiency against representation somehow.

■

It is possible to have multiple levels of stratification or clustering, or combinations of each. The cluster sampling example above runs the risk of containing no observations from either the Eastern or the Western villages. We could stratify first into East and West and then use cluster sampling in each region to try to ensure some representation and efficiency. Ultimately we reach a group where we either try to include everyone or we are taking a simple random sample. If it's the latter, R code like that shown earlier can help.

There are many other ways to collect data which are far from a random sample. These include:

- Convenience sample

Here, subjects are chosen because they're available. In these cases we hope that the results can be generalised to a wider population. However, there might easily be important differences with a wider population which stop this from being reasonable.

• Self-selected sample

In these cases, a broad group of people are potentially able to join the study, but only a limited set of people want to be part of it. A study of this kind welcomes them all. Many TV polls are of this type, where only an unusual sub-population will participate, requiring maybe high interest in the TV show and the highlighted issue, and possibly payment of a fee. Their results rarely even represent typical viewers of the show, let alone the broader public. In some cases, there is nothing stopping keen individuals from registering their views more than once.

From a statistical point of view, the main problem with these two sampling schemes is that the observations are not sampled independently, with equal probability, from the population. This tends to induce bias in any estimates derived from such samples.

4.5 Exercises

1. The Alice data (the difference in heart rate after and before drinking caffeinated diet or decaf diet cola) were provided in the following form:

Caffeinated	17	22	21	16	6	−2	27	15	16	20
Decaf	4	10	7	−9	5	4	5	7	6	12

   Write the data in *standard form* and explain how to read this into R.

2. A group of 20 people has to be randomised into a control and test group, with 10 people each. How many different test groups can we make?

3. A research group wishes to study the effect of a particular drug on blood pressure versus an existing drug. They plan to run a large *randomised controlled trial* ($n = 1000$ or more), recruiting subjects with an initial systolic blood pressure of at least 140 mm Hg, aged 30–60, both males and females, with any level of fitness, occupation or type of diet. To answer their research question, for each subject they only need record which drug was given and the blood pressure before and after treatment. Why do they not need to record values for the other variables (age, gender, etc.)?

4. A group of scientists were carrying out a study on the effectiveness of a new medicine in treating Tasmanian devil facial tumor disease. Thirty devils with facial tumors were split in to two groups of 15 each, and one group was given the new medicine whilst the other group was not treated. Observations on the size of the facial tumors of each devil were taken at the start of the study and after 12 weeks.

(a) Is this an observational or designed experiment? Describe some strengths and weaknesses of the experiment or aspects where you could do with more information to judge its quality.

(b) Which is the response variable and why? What type is it?

(c) Which is the explanatory variable and why? What type is it?

5. Middle-of-the-night (MOTN) insomnia is a common form of insomnia whose prevalence increases with age. This study was conducted to determine the efficacy of different dosages of triazolam when taken after a MOTN awakening with difficulty returning to sleep. In the study 24 patients (mean age 41.00±10.40, 10 female and 14 male) affected by MOTN insomnia were recruited and randomly allocated to one of three dosage groups: A (placebo: no triazolam), B (low dose triazolam), C (high dose triazolam). After 2 weeks of treatment, both low dose and high dose triazolam groups showed increased total time of sleep (in mins) relative to the placebo group (low dose: $p = 0.029$, high dose: $p = 0.004$).[1]

   (a) What type of study is this? Describe some strengths and weaknesses of the experiment or aspects where you could do with more information to judge its quality.

   (b) What values can the variable *total time of sleep* take? What type is it?

   (c) What values can the variable *dosage* take? What type is it?

   (d) which is the variable of interest; that is, the *response variable*? Why?

   (e) How might the response variable have been summarised across the dataset?

6. A study investigates the association between alcohol use and high school attendance. A large sample of adolescents ($n = 7, 874$) was taken and linked to an official school registry. The adolescents were asked "Have you ever consumed so much alcohol that you were clearly intoxicated (drunk)?" The original item had five categories ranging from "No, never" to "Yes, more than 10 times." Frequent intoxication was defined as drinking so much that one was clearly intoxicated more than 10 times (Skogen et al. , 2014), and on this basis a dichotomous variable was created. It was found that, after accounting for age, gender, socioeconomic status and mental health problems, adolescents who were frequently intoxicated missed a greater number of hours from school ($p < 0.001$).[2]

[1] Adapted from L. Ferini Strambi, S. Marelli, M. Zucconi, A. Galbiati, and G. Biggio (2017) *Effects of different doses of triazolam in the middle-of-the-night insomnia: a double-blind, randomized, parallel group study*, J Neurol. 264(7):1362-1369. doi: 10.1007/s00415-017-8530-z.

[2] Adapted from O. Heradstveit, J. C. Skogen, J. Hetland, and M. Hysing (2017) *Alcohol and Illicit Drug Use Are Important Factors for School-Related Problems among Adolescents*, Front. Psychol. 8:1023. doi: 10.3389/fpsyg.2017.01023

(a) What type of study is this? Describe some strengths and weaknesses of the experiment or aspects where you could do with more information to judge its quality.

(b) What values can the variable *frequent intoxication* take? What type is it?

(c) What values can the variable *number of hours missed* take? What type is it?

(d) Which is the variable of interest; that is, the response variable and what is it's type? Why do you think it's the response variable?

(e) How has the response variable been summarised across the dataset?

(f) What are the explanatory variables and their types?

7. In the next chapter, we'll be looking closely at data from the following study. The first half of the abstract is shown below.

   This study was focused on survival analysis of heart failure patients who were admitted to Institute of Cardiology and Allied hospital Faisalabad-Pakistan during April-December (2015). All the patients were aged 40 years or above, having left ventricular systolic dysfunction, belonging to NYHA class III and IV. Cox regression was used to model mortality considering age, ejection fraction, serum creatinine, serum sodium, anemia, platelets, creatinine phosphokinase, blood pressure, gender, diabetes and smoking status as potentially contributing for mortality.

   (a) What type of study is this? Describe some strengths and weaknesses of the experiment or aspects where you could do with more information to judge its quality.

   (b) What values can the variable *frequent intoxication* take? What type is it?

   (c) What is/are the response variable(s)? Are there any relevant issues to be aware of?

CHAPTER 5

DESCRIPTIVE STATISTICS

Exploratory data analysis is primarily about producing numerical and graphical summaries of the data. This will allow us to understand the ranges and distributions of all variables, to check for outliers and form an initial impression, particularly of any problems that might arise in subsequent analysis. This chapter will give examples of how to prepare and communicate the results of exploratory data analysis, making extensive use of R, which is both a free piece of software and a programming language.

5.1 Introduction

In this chapter we will perform exploratory data analysis of a dataset which describes a variety of health attributes for a group of post-heart failure patients from the Institute of Cardiology in Faisalabad, Pakistan. It was first analysed in the following paper:

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. PloS One, 12(7), e0181001. `https://search.library.uq.edu.au/permalink/f/tbms52/TN_cdi_plos_journals_1921149407`,

with further analysis in

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 1–16. `https://search.library.uq.edu.au/permalink/f/tbms52/TN_cdi_pubmed_primary_32013925`.

The study was approved by a university review board, with informed consent given by all patients included. The data was made available in 2020 at the University of California, Irvine, Machine Learning Repository at `https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records` and can be accessed on Blackboard under Learning Resources as `heart.csv`.

All subjects in this study were aged 40 years or more, and had been diagnosed with heart failure. More specifically, they had been diagnosed with left ventricular systolic

dysfunction, and given New York Heart Association functional classifications of III or IV, which are the most severe. This observational study measured 11 variables during the patients admission to the Institute for heart failure, and then whether or not the patient died sometime over the next year.

The paper by Ahmad *et al.* used survival analysis and Cox regression, which is slightly beyond the reasonable scope of this course. We will analyse the same dataset with a range of aims, particularly trying to determine which variables are associated with death and with ejection fraction, i.e. the fraction of blood ejected by the left ventricle of the heart with each beat. Each requires a different form of statistical analysis.

For each patient, there is information about age, sex and a number of health conditions, as well as some measurements from blood samples.

The previous chapter argued that the best source of evidence from a single study comes when it can be performed as a randomised controlled trial (RCT), since these can offer causative evidence. Such studies are often fairly straightforward to analyse, because we likely only have a small number of treatments included (often just two) and a single response variable. The other variables are largely ignorable since their distributions are initially the same for each treatment group due to the randomisation process.

However, many studies are observational, for reasons including ethics, practicality and cost. These can't usually lead to causative conclusions, but can still provide interesting findings worthy of follow-up. Many RCTs can partly trace their origins to observational studies. In an observational study, every variable potentially matters, possibly including some which haven't been measured. Suitable forms of analysis depend heavily on the type of the response variable, and to some extent on the explanatory variables also.

Methods suited to RCT analysis include t-tests and ANOVA, while observational studies will often need some form of regression. All of these are covered in later chapters. Exploratory data analysis is useful to check, summarise and communicate about the data before this, and to determine which analysis methods might be suitable.

5.2 Data as a Spreadsheet

Data is often stored in a table or spreadsheet. A statistical convention is to denote variables as columns and the individual experimental 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, after using a randomization procedure.

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.

In this course, we will store data in CSV (**Comma Separated Values**) format. That is, the data is given as a text file where, as the name suggests, values are separated by commas. You can open and create a CSV file/spreadsheet via Excel, a text editor or, better, via R. We will focus on the CSV file `heart.csv` which contains the data on 299 post-heart failure patients.

If you have not done so already, please install R and RStudio on your own computer. See `https://github.com/uqlibrary/technology-training/blob/master/R/Installation.md#r--rstudio-installation-instructions` for instructions.

Create a folder on your own computer to store your datasets. Download the `heart.csv` file from Blackboard and save it into this folder. Then open RStudio, and set your data folder to be the working directory via the menus with with Session > Set Working Directory > Choose Directory. If this is working, the corresponding text R commands will appear in the console window. For example:

```
> setwd("C:/STAT1301/data")
```

You can import the data into R using for example the function **`read.csv`**, as in:

```
> heart <- read.csv("heart.csv")
```

In R, `<-` means we assign the term on the right to the variable on the left. You can use `=` instead if you prefer. There are different read commands for different data formats. The most common format is .csv and **`read.csv`** is the main command to read in such files. Here we've read in all the heart patient data and stored it all in a data frame we've called heart. A `data.frame` object in R is basically a list of columns. To check the type of any object you can use the R function **`class`**.

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

The R function **`head`** gives the first few rows of the data frame, including the variable names.

```
> head(heart)
  age anaemia   cp diabetes ejection_fraction hbp platelets serum_creatinine serum_sodium sex smoking time death
1  75       0  582        0                20   1    265000             1.9          130   1       0    4     1
2  55       0 7861        0                38   0    263358             1.1          136   1       0    6     1
3  65       0  146        0                20   0    162000             1.3          129   1       1    7     1
4  50       1  111        0                20   0    210000             1.9          137   1       0    7     1
5  65       1  160        1                20   0    327000             2.7          116   0       0    8     1
6  90       1   47        0                40   1    204000             2.1          132   1       1    8     1
```

The names of the variables can also be obtained directly via the function **`names`**, as in `names(heart)`. 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.

☞ 242

We can see all the variable names at the top and make an educated guess on what they all mean, referring to Ahmad *et al.* for more information as needed. We abbreviated creatinine_phosphokinase level to cp and high blood pressure to hbp. For the binary (0 or 1) variables, which include anaemia, hbp, smoking and death, 1 indicates that the condition was present and 0 that it was absent. All patients were alive when initially measured, but those with a 1 recorded died sometime over the next year. When the sex variable is 1, the patient's sex was male, and 0 indicates that the patient's sex was female.

The time variable refers to the time of follow-up. For this dataset, the death variable was recorded at this time. Ideally, all patients would have been followed up at the same time, but all have been followed up within a year. It's unclear how the follow-up time was chosen. It seems likely that if a patient died, the Institute would be notified and record the death immediately. As a result, the time variable is partly determined by the event of death and highly correlated with it. Thus the time variable should not be used to try to predict death or any other variable, so we will leave it out of our analysis.

See Table 5.1 for a description of the variables.

Table 5.1: Variables in the heart study

Description	**Unit** or **Coding**	**Variable**
Age at date of admission	Years	`age`
Anaemia	0=absent; 1=present	`anaemia`
Diabetes	0=absent; 1=present	`diabetes`
High Blood Pressure	0=absent; 1=present	`hbp`
Sex	0=Female; 1=Male	`sex`
Smoking	0=absent; 1=present	`smoking`
Death	0=survived; 1=died	`death`
Creatinine Phosphokinase	$\mu g/L$	`cp`
Left Ventricular Ejection Fraction	%	`ejection_fraction`
Platelets	kiloplatelets/mL	`platelets`
Serum Creatinine	mg/dL	`serum_creatinine`
Serum Sodium	milliequivalents/L	`serum_sodium`
Time of follow-up	days since admitted	`time`

Note that all the entries in **`heart`** are *numerical* - they are numbers. However, the *meaning* of each number depends on the respective columns. To help decide how to summarise and analyse the data it is important to specify exactly what the type is of each variable. We discuss this next.

5.3 Variable Types

We can generally classify the measurement variables into two types: *quantitative* and *qualitative*, with the latter also called *categorical*. For quantitative variables we can

make a distinction between continuous and discrete variables. [1]

Continuous 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 variables have only a small number of possibilities, such as a count of some outcomes.

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

Nominal factors represent groups of measurements without an agreed or relevant order, such as names e.g. of prescriptions, people, diseases.

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

We usually call binary variables nominal, even if there is some natural order to the categories. This is mainly because even if we use the "wrong" order, reversing it is easy. Try to work out the type of each of the heart data variables. This will be discussed in lectures.

Initially, all variables in `heart` are identified by `R` as quantitative, because they happened to be entered as numbers. [2] You can check the structure of an object using the command `str` and the type of an `R` object using the command `typeof`. For example:

```
> str(heart)
 'data.frame':        299 obs. of  13 variables:
  $ age               : num  75 55 65 50 65 90 75 60 65 80 ...
  $ anaemia           : int  0 0 0 1 1 1 1 1 0 1 ...
  $ cp                : int  582 7861 146 111 160 47 246 315 157 123 ...
  $ diabetes          : int  0 0 0 0 1 0 0 1 0 0 ...
  $ ejection_fraction: int  20 38 20 20 20 40 15 60 65 35 ...
  $ hbp               : int  1 0 0 0 0 1 0 0 0 1 ...
  $ platelets         : num  265000 263358 162000 210000 327000 ...
  $ serum_creatinine : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
  $ serum_sodium      : int  130 136 129 137 116 132 137 131 138 133 ...
  $ sex               : int  1 1 1 1 0 1 1 1 0 1 ...
  $ smoking           : int  0 0 1 0 0 1 0 1 0 1 ...
  $ time              : int  4 6 7 7 8 8 10 10 10 10 ...
  $ death             : int  1 1 1 1 1 1 1 1 1 1 ...
> typeof(heart$age)
 [1] "double"
> typeof(heart$diabetes)
 [1] "integer"
```

[1] As all measurements are recorded to a finite level of accuracy, one could argue that all quantitative variables are discrete. The issue is how we *model* the data, via discrete and continuous random variables.

[2] If `sex` had been entered as words or letters such as M and F, the variable would have automatically been structured as a factor (categorical).

Some variables have been represented by integers and others by numeric variables, here with double precision, i.e. using twice the memory used for a standard floating point number. When a variable is categorical, we want R to treat is as such. In R, such variables are called "factors". We can change the type of a variable into a factor by e.g.

```
> heart$sex <- factor(heart$sex)
```

We need to do this for all the categorical variables. A quicker alternative is to tell R that some of the variables are factors as they're read in. We can specify the type of each variable to R as part of the **read.csv** command using the **colClasses** argument, as follows:

```
> heart <- read.csv('heart.csv', colClasses=c('numeric','factor',
 'numeric','factor','numeric','factor','numeric','numeric',
 'numeric','factor','factor','numeric','factor'))
```

In R, **c(a,b,..)** concatenates a list of items into a vector or list object – it's used as part of many commands. Having read the data in again and set the types of variables, we can now check the structure of the `data.frame` **heart**.

```
> str(heart)
 'data.frame':        299 obs. of  13 variables:
$ age              : num  75 55 65 50 65 90 75 60 65 80 ...
$ Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 1 2 ...
$ num  582 7861 146 111 160 ...
$ Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
$ num  20 38 20 20 20 40 15 60 65 35 ...
$ Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 1 1 2 ...
$ num  265000 263358 162000 210000 327000 ...
$ num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
$ num  130 136 129 137 116 132 137 131 138 133 ...
$ Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 1 2 ...
$ Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 2 ...
$ num  4 6 7 7 8 8 10 10 10 10 ...
$ Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Note that the factor variables retain levels of 0 and 1, but the data has been replaced by the factor values 1 and 2, meaning the first and second factor levels.

We can access the variables (columns) of a data frame via the $ construction.

```
> heart$diabetes[1:3]  #diabetes status of first three patients


[1] 0 0 0
Levels: 0 1
```

```
> class(heart$diabetes)
```

```
[1] "factor"
```

If you make any changes to your data and want to save it in another CSV file, you can do so via the **write.csv** function, as in:

```
> write.csv(heart,"heart2.csv")
```

5.4 Summary Statistics

In the following, $\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T}$ is a column vector of numbers. For our `heart` data set $\boldsymbol{x}$ could for example correspond to the ages of the $n = 299$ individuals when they were admitted to the Institute.

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 `heart` data, we have, for example:

```
> mean(heart$age)
[1] 60.83389
```

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. We will assume that the median is the average of these two values.

The R function to calculate the median is **median**. For example,

```
> median(heart$age)
[1] 60
```

The median is a more robust measure of distribution location than the mean, because it is unaffected by any rare extreme values.

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

```
> quantile(heart$age,probs=c(0.1,0.9))
10% 90%
45.0 75.4
```

While the sample mean and median say something about the *location* of the data, they do not provide information about the *dispersion* or *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 R, the 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.

```
> range(heart$age)
[1] 40 95
```

Typically, when the sample size increases, the range becomes wider, and so it is difficult to compare the spreads of two data sets via their ranges, when when the sample sizes are different. A more robust measure for the spread of the data is the **interquartile range** (IQR), which is the difference between the third and first quartile.

```
> IQR(heart$age)

[1] 19
```

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

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \tag{5.1}$$

where $\bar{x}$ is the sample mean. It is roughly the average squared distance from the mean. The use of $\frac{1}{n-1}$ rather than $\frac{1}{n}$ is to avoid bias in the estimation. Considering a random sample whose sample variance is S^2, it can be shown (see tutorials) that the expected value $\mathbb{E}(S^2) = \sigma^2$, where σ^2 is the variance of the distribution which generates the sample. We will see in later chapters that variance 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(heart$age)
[1] 141.4865
> sd(heart$age)
[1] 11.89481
```

Note that in this case, the units of sample variance are years^2 and the units of standard deviation are years. For interpretability, the standard deviation is more commonly reported.

5.5 Categorical Variables

There are a large number of possible numerical and graphical summaries we can make of the heart data. Some will be interesting, some will allow us to check conditions for particular statistical methods, but none will allow us to make formal conclusions. We will turn to confidence intervals and hypothesis testing for those. We'd particularly like to know something about the distribution of each variable. For categorical variables, the counts and proportions in each group are useful. For quantitative variables, plots and summaries of location and scale are useful, such as mean and standard deviation.

We are lucky with this dataset that it is complete, i.e. there are no missing entries. Missing data is quite common in practice. This might be due to a failure to record one or more variables for a patient while they were initially seen, or a failure to find them or have them cooperate on follow-up to record other variables. One of the simplest approaches to missing data is to just remove patients who have any missing variable values. There are other more complex imputation approaches which attempt to fill in suitable predictions. Whatever is done, missing data results in less data, i.e. effectively a smaller sample size.

Let's start by summarising the categorical variables numerically. We'll start with the anaemia variable.

```
> summary(heart$anaemia)

  0   1
170 129
> proportions(summary(heart$anaemia))
        0         1
0.5685619 0.4314381
```

We notice that there are a total of $170 + 129 = 299$ patients in the dataset, with 170 having no anaemia and 129 with anaemia. It's more informative to see the proportions, with a proportion of around 0.431 having anaemia. We rounded to 3 significant figures here. This is because there were only (at most) 3 significant figures in the counts in the first place and anything more than 3 significant figures is usually an uninformative distraction. The main exception would be if we are comparing two similar numbers, i.e. essentially focusing on their difference, and we might want 3 significant figures for that difference.

Use R to calculate the proportions (to 3 significant figures) of each of the other conditions of interest for this dataset: diabetes, high blood pressure, maleness, smoking and death within 1 year.

You should determine the following proportions:
diabetes: 0.418, high blood pressure: 0.351, sex male: 0.649, smoking: 0.321, death: 0.321.

The most important of these variables for our study is death, which is the main response variable. We can check to see if there appears to be a relationship between

any of these variables and the proportion of patients who died in the year following their initial examination after heart failure.

We'll do this first with anaemia, starting with a table of counts and moving on to a suitably calculated table of proportions. In the following, we use the **table** command to calculate and format counts for each combination of categories. We need to tell it the dimension (or variable) names via the **dnn** argument.

```
> table(heart$anaemia,heart$death,dnn=c("anaemia","death"))
```

```
       death
anaemia   0   1
      0 120  50
      1  83  46
```

```
> proportions(table(heart$anaemia,heart$death,
dnn=c("anaemia","death")),margin=1)
```

```
       death
anaemia         0         1
      0 0.7058824 0.2941176
      1 0.6434109 0.3565891
```

We processed the table into proportions via the **proportions** command. The **margins=1** argument ensures that the row proportions sum to 1. This is useful to see the proportion of patients who died within a year in both the anaemia and no anaemia groups. The patients with anaemia seem to have a death rate that is 0.06 or 6% higher than the patients without anaemia. However, we shouldn't overly believe such apparent differences at present. Formal analysis could be conducted using hypothesis tests and/or confidence intervals, as described in later chapters.

If you were going to include tables resembling these in a report, they'd need further formatting. You might find it useful to write the table to a .csv file as an intermediate step via something like:

```
> write.csv(table(heart$anaemia,heart$death),
  "counts_anaemia_death.csv")
> write.csv(proportions(table(heart$anaemia,heart$death),
margin=1), "proportions_anaemia_death.csv")
```

This gives you tables with category headings and the numeric content. The variable names are not included in the .csv file, whether **dnn** is used earlier or not. You can open either .csv file in e.g. Excel and then copy it over to Word or Latex, and then improve the content and appearance. For example:

	Survived	Died
No Anaemia	120	50
Anaemia	83	46

and

	Survived	Died
No Anaemia	0.706	0.294
Anaemia	0.643	0.357

These would be further improved with captions and numbering, and then would be referenced in the relevant document.

Note that when applying the R **proportions** command to a table, you could make the column proportions sum to 1 by using the argument **margin = 2**. For example:

```
> proportions(table(heart$anaemia,heart$death),margin=2)

       death
anaemia         0         1
      0 0.5911330 0.5208333
      1 0.4088670 0.4791667
```

From this we can see that in the group who died, a 0.479 proportion had anaemia when initially measured, while in the proportion who were alive at follow-up, a 0.409 proportion had anaemia when initially measured. This has some meaning, but since our analysis has death as the response variable, the previous table (with row proportions summing to 1) is more informative.

Without either margin, the whole table of proportions sum to 1 and it is hard to glean anything from the table. For example:

```
> proportions(table(heart$anaemia,heart$death))

       death
anaemia         0         1
      0 0.4013378 0.1672241
      1 0.2775920 0.1538462
```

This is the truth about the proportions of patients in each of the groups, but doesn't help us take into account that anaemia is less common than its absence in these patients. This sort of table would mainly be useful if we wanted to know the overall proportion of patients who fall into a particular category pair, e.g. had no anaemia and were alive at follow-up: 0.167.

For the other categorical variables, we might be tempted to only look at proportions from a table with the death variable. However, it is wise to look at the counts also, just in case some category pair has less than 10 observations, which would render some analyses unsuitable. This is done for all these other variables below.

```
> table(heart$diabetes,heart$death,
  dnn=c("diabetes","death"))
```

```
        death
diabetes   0   1
       0 118  56
       1  85  40
```

```
> proportions(table(heart$diabetes,heart$death,
  dnn=c("diabetes","death")),margin=1)
```

```
        death
diabetes         0         1
       0 0.6781609 0.3218391
       1 0.6800000 0.3200000
```

```
> table(heart$hbp,heart$death,dnn=c("hbp","death"))
```

```
  death
hbp   0   1
  0 137  57
  1  66  39
```

```
> proportions(table(heart$hbp,heart$death,
  dnn=c("hbp","death")),margin=1)
```

```
   death
hbp         0         1
  0 0.7061856 0.2938144
  1 0.6285714 0.3714286
```

```
> table(heart$sex,heart$death,dnn=c("sex","death"))
```

```
   death
sex   0   1
  0  71  34
  1 132  62
```

```
> proportions(table(heart$sex,heart$death,
  dnn=c("sex","death")),margin=1)


   death
sex         0         1
  0 0.6761905 0.3238095
  1 0.6804124 0.3195876


> table(heart$smoking,heart$death,dnn=c("smoking","death"))


       death
smoking   0   1
      0 137  66
      1  66  30


> proportions(table(heart$smoking,heart$death,
  dnn=c("smoking","death")),margin=1)


       death
smoking         0         1
      0 0.6748768 0.3251232
      1 0.6875000 0.3125000
```

We notice that all of the counts of category pairs with death have at least 30 observations, so we should have enough data to meet conditions of upcoming tests and confidence intervals. Relevant tests include tests of proportions and chi-squared, and these work best with 10 or more observations per category or combination of categories.

The proportions information seems to show that death rates for heart failure patients are very similar for those with diabetes vs those without, males vs females and smokers vs non-smokers. The sample proportion of patients with high blood pressure who died was 0.371, compared to 0.294 for those without high blood pressure - around an 8% difference. Of the categorical variables, high blood pressure and anaemia show the most potential in being able to distinguish between patients who will survive the next year and those who will not.

We can also present categorical variable data visually with bar charts. This can give a quick visual impression of the counts or proportions across various categories. The following R command creates a bar chart for counts of death and survival for subjects with and without anaemia, including some colours, a title, labelling and a legend. Note that since the categorical variables were coded in binary (0/1), we have to translate those into category labels as part of the plot command.

```
> barplot(table(heart$death,heart$anaemia),beside=T,main=
"Counts of heart failure patient survival, grouped by
anaemia status",col=c("cyan","brown"),
legend.text=c("Survived","Died"),
names.arg=c("No anaemia","Anaemia"))
```

It should result in Figure 5.1. All the comma-separated arguments to the **barplot** function have a useful role here. Please try this command yourself in R (having run previous commands which set things up). Then try modifying parts of it or removing parts to see what happens. You can look up all the options with **?barplot**. Note that title for a plot could be put either on the plot or in the caption. Captions are necessary for reports and publication. However, if you generate many plot files, it is useful to also have an identifying description in the plot title.

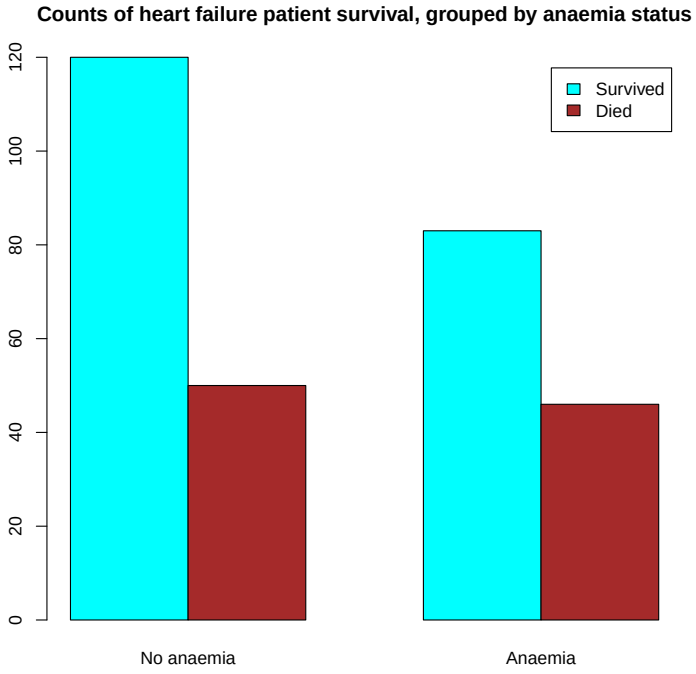


Figure 5.1: Survival and death counts by anaemia status

The following R command creates another bar chart based on a table of proportions for the relationship between anaemia and death, with the output given in Figure 5.2.

```
> barplot(proportions(table(heart$death,heart$anaemia),margin=2),
beside=T,main="Proportions of heart failure patient survival,
by anaemia status",col=c("cyan","brown"), legend.text=
c("Survived","Died"),names.arg=c("No anaemia","Anaemia"))
```

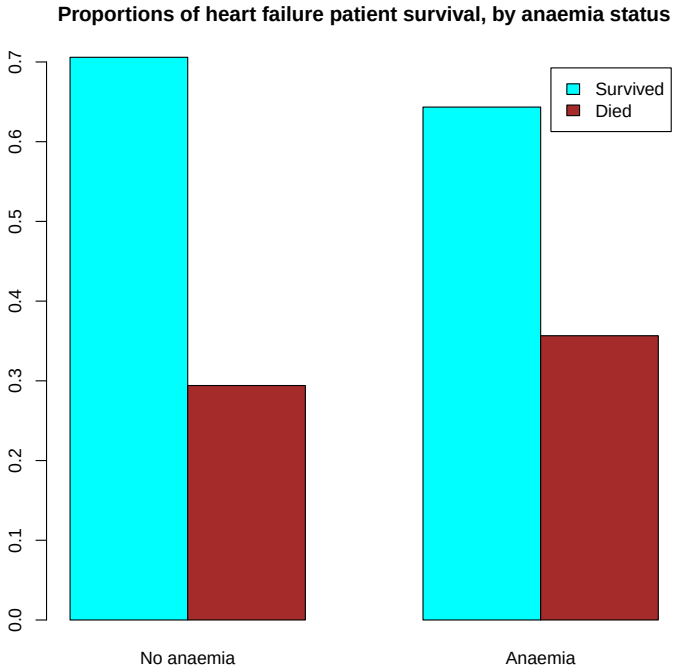


Figure 5.2: Survival proportions by anaemia status

As with the tables, the second plot showing proportions is easier to interpret. We can see that a higher proportion of patients with anaemia died compared to those without anaemia. However, the first plot is also useful to see the group sizes for each combination of categorical variables. All are above 40.

A variation on the above which uses area instead of bar height to represent counts is given by a mosaic plot. The following code and plot in Figure 5.3 provide an example.

```
> antab <- table(heart$anaemia,heart$death)
> rownames(antab) <- c("normal","anaemia")
> colnames(antab) <- c("survived","died")
> mosaicplot(antab,main="Heart failure patient survival,
grouped by anaemia status",col=c("cyan","brown"),
xlab="",ylab="",cex.axis=1)
```

For this plot, we first create a table object containing the data to be plotted, and alter the row and column names to give the category meanings instead of 0/1. The **cex** commands in R are expansion constants for text, which here are set to 1 to give normal sized text for category labels (the default with **mosaicplot** is 0.66).

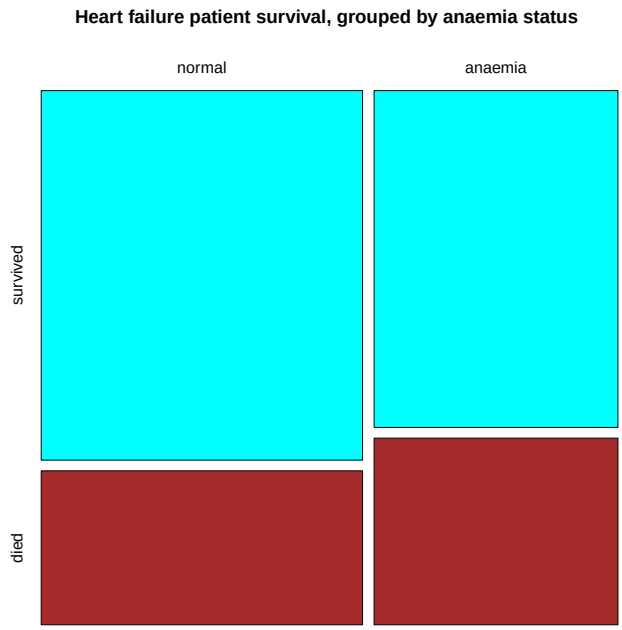


Figure 5.3: Survival by anaemia status

Mosaic plots may be most useful when there are two categorical variables and each has many categories. It uses area to illustrate relative counts, but also illustrates proportions both vertically and horizontally. We can immediately see that a higher proportion of the subjects with anaemia died compared to those without anaemia. We can also see that anaemia is less common than its absence in this dataset. When one categorical variable is explanatory (here anaemia) and the other is a response (here death), we think it's better to put the explanatory variable horizontally and the response variable vertically. We do that routinely with linear regression for continuous variables.

5.6 Quantitative variables

We next move to the quantitative variables – namely age (years), cp (creatinine phosphokinase mcg/L), ejection fraction (%), platelets (kiloplatelets/mL), serum creatinine (mg/dL) and serum sodium (mEq/L). The paper by Chicco et al. (2020) gives details of these variables. The paper by Ahmad et al. (2017) turns some of these into categorical variables by considering ranges, e.g. for ejection fraction (EF), with $EF \leqslant 30$, $30 < EF \leqslant 45$, $EF > 45$. We'll leave all variables as quantitative for now.

For each variable, the R command **summary** tells you a lot of what you'd initially like to know – enough to create a box plot.

For now, we'll focus on just one of the variables: serum creatinine, which Chicco *et al.* found to be an important predictor of survival. A full analysis would look at each of the quantitative variables. After reading and working through the following, you should choose another quantitative variable from this dataset and try to produce similar numerical and graphical summaries with R.

```
> summary(heart$serum_creatinine)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.500   0.900   1.100   1.394   1.400   9.400
```

In particular, this tells us that the median is around 1.1 mg/dL. We can also work out that the interquartile range (IQR), which is the third quartile minus the first quartile is $1.4 - 0.9 = 0.5$, and this is a measure of the spread of this variable.

It would be nice to see this information on a boxplot instead. In a boxplot, a central box covers the first quartile (Q1) of the data to the the third quartile (Q3). Note that these are also called the 25th and 75th percentiles, respectively. The distance between them (Q3 - Q1) is termed the interquartile range (IQR), a robust measure of spread. A line across the box shows the median (Q2). Observations more than 1.5 times the IQR *beyond* either edge of the box are called (box plot) outliers, and are shown as a *. Lines ('whiskers') extend from the box out to the smallest and largest observations that are not suspected outliers. Note that to obtain the qth percentile of the data, we average the $\lfloor (n+1)q/100 \rfloor$ and $\lceil (n+1)q/100 \rceil$ values, where $\lfloor\rfloor$ and $\lceil\rceil$ denote floor and ceiling (round down and round up), respectively. Some software, including R, uses interpolation instead of averaging.

To see a boxplot of the serum creatinine data, we can run

```
> boxplot(heart$serum_creatinine)
```

and produce the following plot in Figure 5.4.

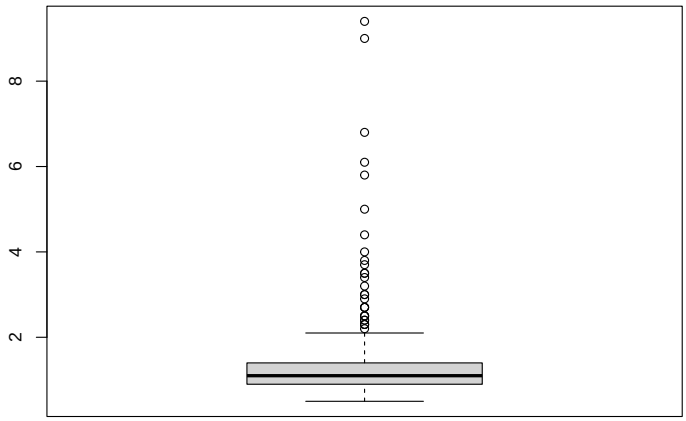


Figure 5.4: Basic boxplot of serum creatinine

That's fine for a quick check. We can see from both the numerical summary and the boxplot that the shape of this variable's distribution is highly right-skewed. I.e. the serum creatinine values are much more widely spread for higher values (>2) than they are below e.g. 1. That's not surprising, given that the values must be positive (have a lower limit of 0) and have no nearby upper limit. Other common shapes for distributions include symmetrical (left and right tails are almost the mirror of each other) and left-skewed, with a long tail of values towards the left (lower values).

If we were going to show a boxplot for serum creatinine to someone else, it would be better if it had some meaningful title or description, axis labels and some colour. It also might look better sideways. This can all be done via some tweaks below, resulting in Figure 5.5.

```
> boxplot(heart$serum_creatinine,main="Serum Creatinine in Heart
Failure Patients",xlab="Serum Creatinine (mg/dL)",
col="red",horizontal=T)
```

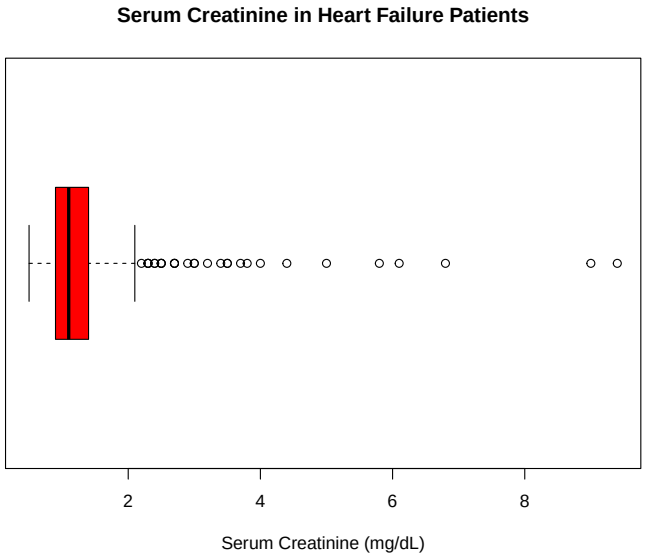


Figure 5.5: Boxplot of serum creatinine levels

You can get help about the options for any R command by typing e.g. `?boxplot`.

With the numerical summary and boxplot above, we are trying to find out about the distribution of a quantitative variable, as seen through some sample data. Other plots exist which can help with this: we will consider histograms and density estimate plots. A histogram is similar to a bar chart in that it plots counts of observations. However, bar charts only really apply to data with a small number of categories or maybe a small number of discrete values. Histograms count observations in a number of ranges called bins. Both the number of bins and their sizes are generally chosen automatically by software with the aim of visual clarity and reasonableness. We can create a histogram with R for the serum creatinine data as follows, with the result in Figure 5.6.

```
> hist(heart$serum_creatinine,col="red",main="Serum Creatinine Levels
 in Heart Failure Patients",xlab="Serum Creatinine (mg/dL)")
```

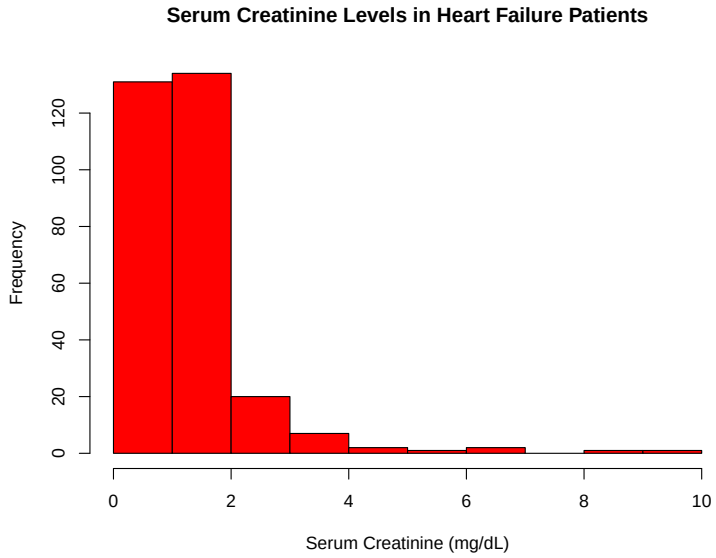


Figure 5.6: Histogram of serum creatinine levels

We can see many of the same aspects shown by the boxplot: the median serum creatinine level near 1, the right skewness, the short left tail. We also see a fast drop from the peak down towards the right after 2 mg/dL. We can see that there aren't many patients with serum creatinine levels above 4 mg/dL, but that values as high as 10 mg/dL are present in the data.

The next plot for quantitative variables we'd like you to consider is a kernel density estimate plot. This is an attempt to approximate the probability density function of the variable being considered. A histogram is a simpler way of doing this, but a kernel density plot smooths this out. The R code below produces a kernel density plot, leading to Figure 5.7.

```
> plot(density(heart$serum_creatinine),col="red",main="Kernel density
 estimate for serum creatinine in heart failure patients",
 xlab="Serum Creatinine (mg/dL)")
```

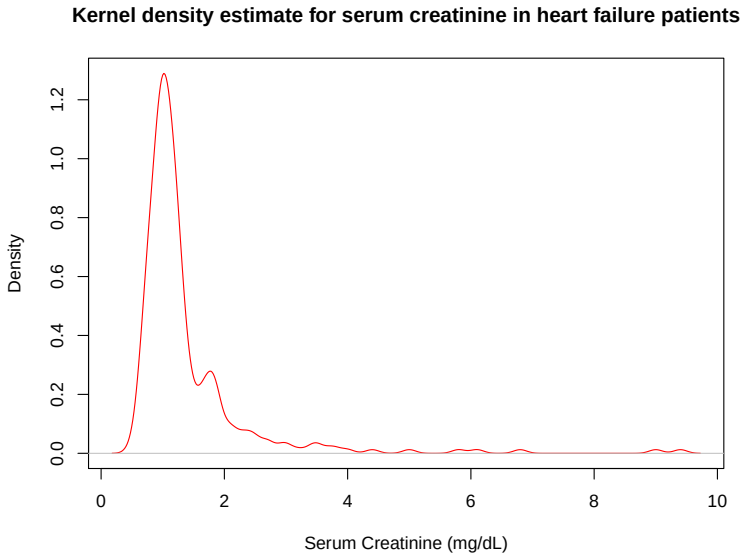


Figure 5.7: Kernel density estimate of serum creatinine levels

This looks quite similar to the histogram. You see a single peak (smoothed out from the histogram) and a highly right-skewed shape. The exact position of the peak in the histogram is to be ignored – the differences in counts are too small between the first two bars to be sure that such details are due to anything more than chance. As a result, smoothing into one peak in the kernel density plot is a good decision.

Some statistical techniques assume that the data they are applied to come from a normal distribution (symmetric, bell-shaped). In many cases, anything vaguely close to a normal distribution will do. This serum creatinine data is quite far from being normally distributed. If it was important for the data to be close to a normal distribution, we could consider a transformation of the data so that the result was closer to normal. One transformation which is commonly applied to right-skewed data is a logarithmic transformation, here replacing serum_creatinine with $\log_{10}$(serum_creatinine). The base of the log doesn't really matter, although base e (natural log) and base 10 are common, with the latter being easier to interpret. For values greater than 1, log transformations essentially squeeze the right side and have less effect on the left side. For values between 0 and 1, a log transformation stretches out the left side and has less effect on the right side.

The following R code applies a base 10 transformation to the serum creatinine data, before producing a kernel density plot, shown in Figure 5.8.

```
> plot(density(log10(heart$serum_creatinine)),col="red",main=
"Kernel density estimate for log base 10 serum creatinine
level",xlab="Log base 10 of Serum Creatinine (mg/dL)")
```

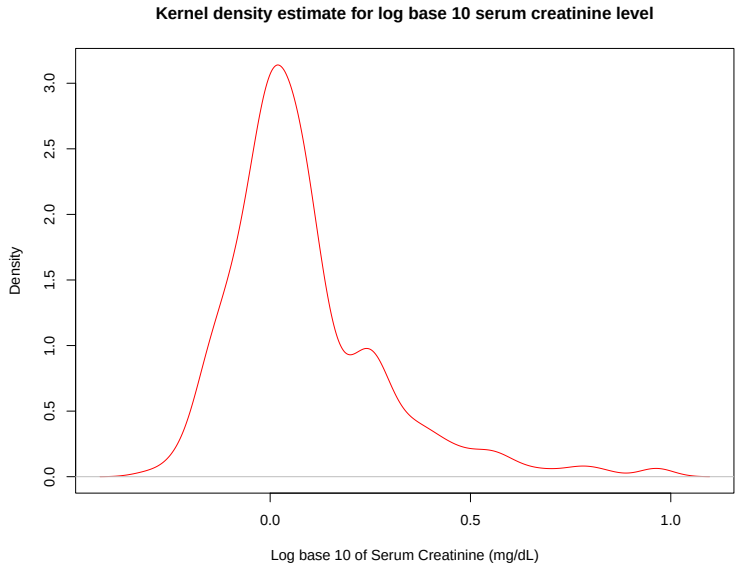


Figure 5.8: Kernel density estimate of $\log_{10}$ serum creatinine levels

The result is a more symmetric distribution, not so far from the shape of a normal distribution. For reference, we can produce a plot of the standard normal distribution probability density function via the following code, with the result shown in Figure 5.9.

```
> x <- seq(-4, 4, length=1000)
# create a sequence of 1000 equally spaced values from -4 to 4
> y <- dnorm(x) # find the normal density value for each x
> plot(x,y, type = "l", lwd = 2, xlab = "", ylab = "",
col="blue", main="Standard normal probability density function")
# plot with a title, using a blue line to connect the dots,
with line width 2
```

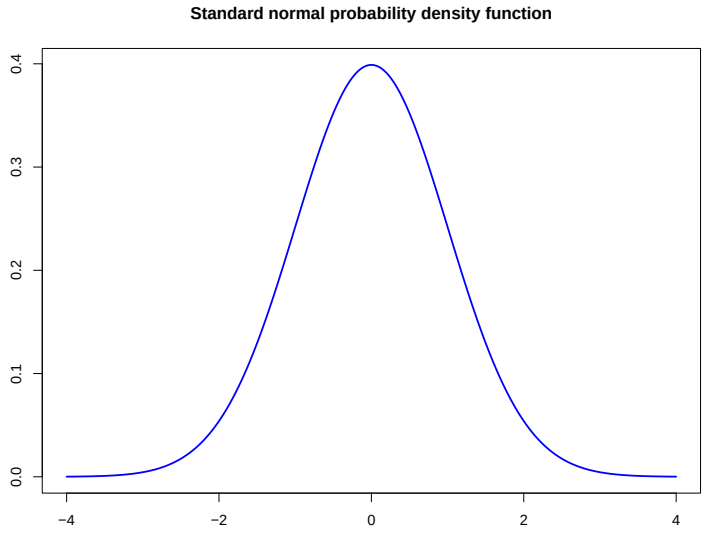


Figure 5.9: Standard normal probability density function

It's important to bear in mind that after a log transformation, the units are not the same and the log has to be reversed with an exponential function to get back to the original units. Similarly, if you had left-skewed data and wanted to make it more normal, you might use an exponential transformation, i.e. exp(x) in R. Other common transformations include square roots and reciprocals, i.e. $x^{0.5}$ and $1/x$.

We might decide that the log transformation has not brought the serum creatinine data distribution sufficiently close to normal. The reciprocal transformation $(1/x)$ is potentially better for this purpose, although it does reverse the order of observations on the x axis, i.e. the largest become the smallest and vice versa. The following R code and Figures 5.10 and 5.11 show the results, which seem to be an improvement.

```
> plot(density(1/heart$serum_creatinine),col="red",main="Kernel
density estimate for reciprocal of serum creatinine level",
xlab="Reciprocal of Serum Creatinine (mg/dL)")
```

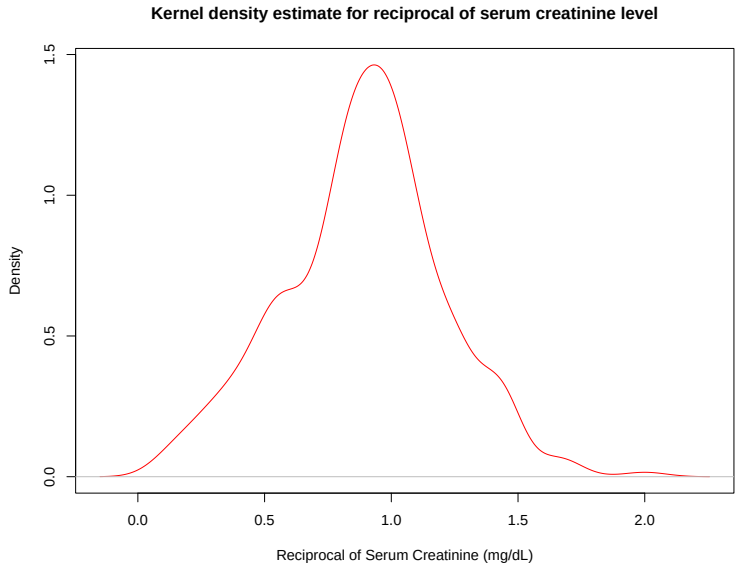


Figure 5.10: Kernel density estimate for reciprocal of serum creatinine level

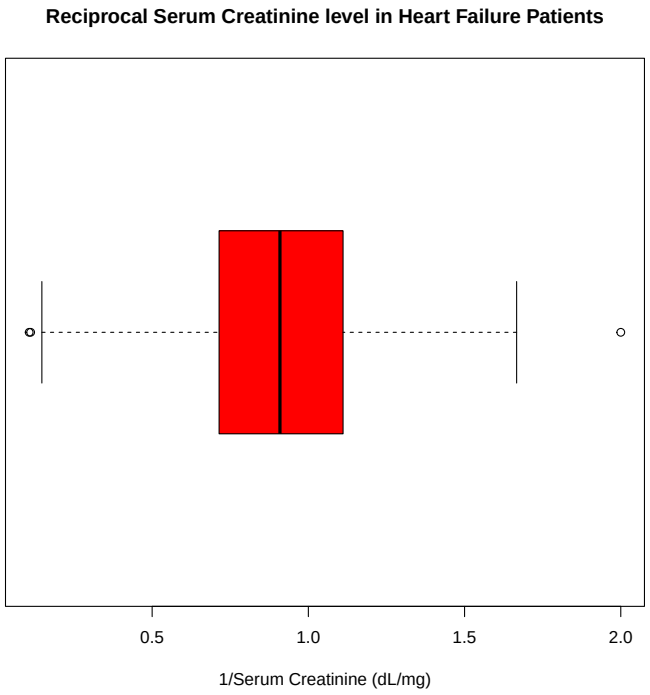


Figure 5.11: Boxplot for reciprocal of serum creatinine level

For right-skewed data, transformations such as log can also have the desirable effect of letting us see the data more evenly spaced along the x axis. Another way to do

this is to plot the data using a log scale. I.e. don't transform the data – just how the x axis is displayed. For boxplots with base R (i.e. without packages), the only such transformation available is log. The following alters the boxplot to put the x axis on a log scale, as seen in Figure 5.12.

```
> boxplot(heart$serum_creatinine,main="Serum Creatinine in
Heart Failure Patients",xlab="Serum Creatinine (mg/dL)",
col="red",horizontal=T,log="x")
```

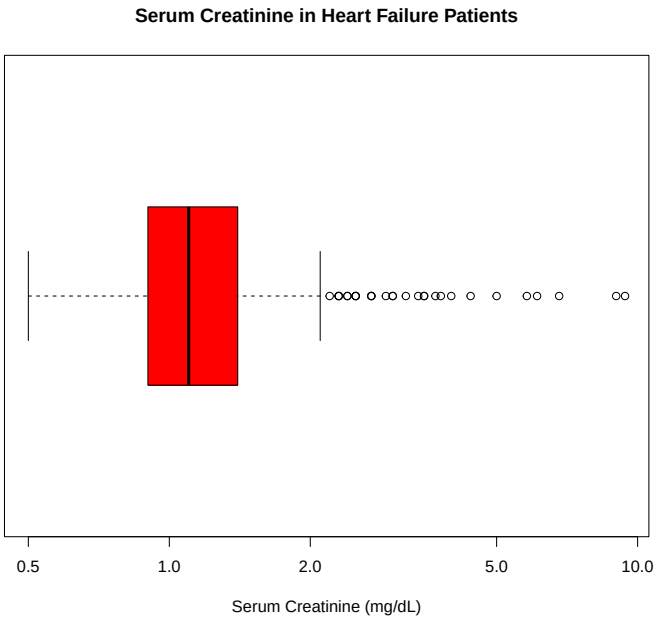


Figure 5.12: Boxplot of Serum Creatinine Levels with Log Scale

This has retained the original units while compressing the data into a readable plot. It's a good compromise, but we do have to remember that we're seeing a log scale. For example, the right whisker is further from the box than the left is.

The `log="x"` option is not available with `hist`, but it is with the `plot` command, as used above with kernel density estimate plots. Using it there with the following code results in Figure 5.13.

```
> plot(density(heart$serum_creatinine),col="red",main="Kernel
 density estimate for serum creatinine in heart failure
 patients",xlab="Serum Creatinine (mg/dL)", log="x")
```

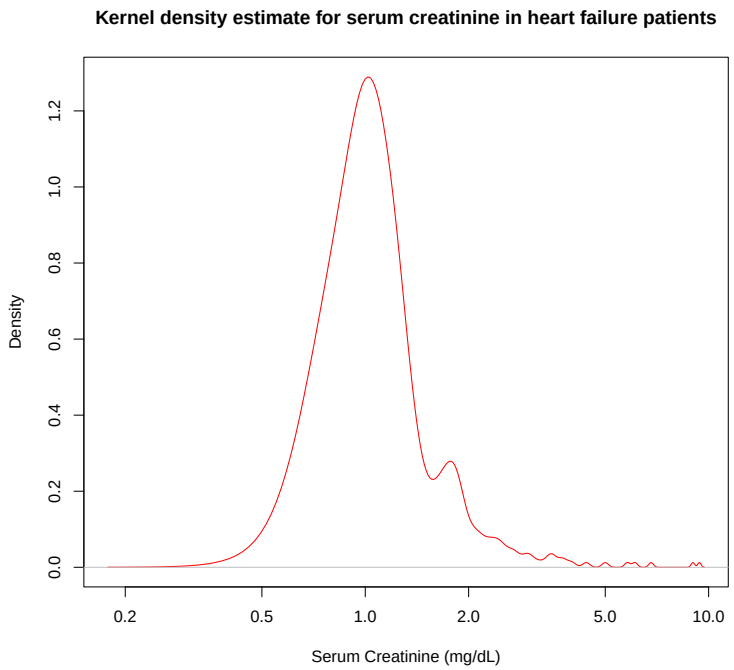


Figure 5.13: Kernel density estimate of serum creatinine levels with log scale

When we want to compare the data distribution visually to a reference distribution, we have one further option: the quantile-quantile plot. This compares the quantiles of the reference distribution (x-axis) to the quantiles of the data (y-axis). If the two distributions match, we should see a straight line. For distributions whose parameters change only location and/or scale, such as the normal and uniform, this straight line should occur if the reference distribution is from the same family as the data (e.g. both normally distributed), even if the parameters are quite different. For other distributions, the parameters of a reference distribution might have to be estimated from the data.

The following R code compares the serum creatinine levels and the reciprocal serum creatinine levels to a standard normal distribution, with the resulting plots shown in Figures 5.14 and 5.15

```
> qqnorm(heart$serum_creatinine)
> qqnorm(1/(heart$serum_creatinine))
```

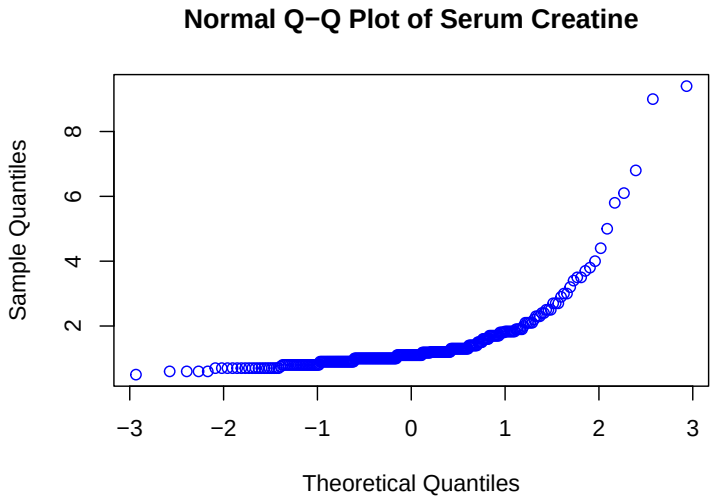


Figure 5.14: Normal Q-Q Plot of Serum Creatinine Levels

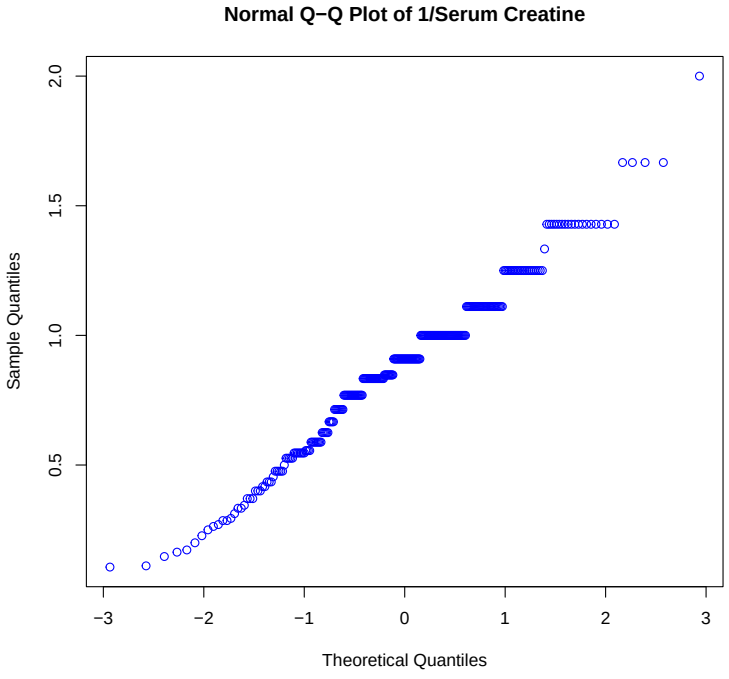


Figure 5.15: Normal Q-Q Plot of Reciprocal of Serum Creatinine Levels

The left part of Figure 5.14 (standard normal -3 to 1) is close to a straight line, which corroborates our impression that the peak of the kernel density plot is roughly normal in shape. However, the curve upward on the right hand side shows that the

sample quantiles are further out than would be expected under a normal distribution, which corresponds to the long right tail of the data distribution.

Figure 5.15 shows approximately a straight line, as would be expected after seeing the kernel density plot of the reciprocal serum creatinine levels. The stair-like appearance on the right is due to many patients with low serum creatinine levels having the same values recorded, which is likely due to limited accuracy of the measurements.

More advanced plotting options exist in the built-in `lattice` package and the downloadable `ggplot2` package, but we won't cover those in this course.

In the plots and summaries, we have occasionally seen observations which could be called outliers, with a formal definition of these used in boxplots. Before transformation, there were many boxplot outliers for the serum creatinine data. Transformations including log and reciprocal brought the data distribution closer to normal and reduced the number of boxplot outliers.

So what is an outlier? An outlier is a value that is not consistent with the majority of the data. It might be an error, or it might indicate the presence of diversity beyond what was expected. For example, there could be a subgroup of the population which is noticeably different, which had not been taken into account earlier. In the latter case, we have to accept that the population is more complicated than we first thought. It is worth checking possible outliers to see if they might be errors in either measurement or recording. If they are thought to be errors, re-taking or re-recording the measurement would be desirable. If this is not possible and the value is impossible, such as a person having a height of 10 meters, it can be deleted, resulting in missing data.

With the serum creatinine data, there were probably no meaningful outliers. With the most effective transformation tried (reciprocal), the boxplot showed very few outliers and these were very close to the other data points. Even data from a standard normal distribution will show outliers on a boxplot if the sample size is large enough.

Outliers can often be easily seen on plots. If extreme enough, they can have a substantial influence on summaries, models and conclusions. In general, we cannot discard outliers without justification. Finding outliers might even be the aim of a study. For example, a geologist searching for ore deposits may look for outliers in geological data.

5.7 One Categorical and One Quantitative Variable

We might be interested in the relationships between multiple variables. We've seen that already with pairs of categorical variables. We need to also consider pairs where one variable is categorical and the other is quantitative. In this case, a common choice is to choose a plot type which would suit the quantitative variable and produce one of these for each possible category, preferably all on one plot.

As an example with the heart dataset, let's consider the anaemia categorical variable and the serum creatinine quantitative variable. We can split the boxplot seen in Figure 5.12 into groups according to survival status by using an R formula as follows

to produce Figure 5.16.

```
> boxplot(heart$serum_creatinine~heart$death,main="Serum Creatinine in
  Heart Failure Patients by Survival Status",xlab="Serum
  Creatinine (mg/dL)",col="red",horizontal=T,log="x",
  names=c("Survived","Died"),ylab="")
```

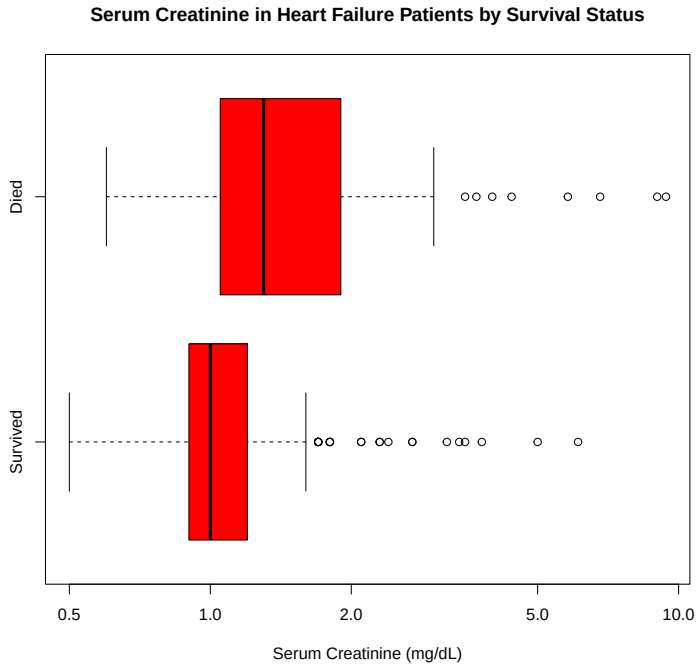


Figure 5.16: Boxplot of Serum Creatinine Levels versus Survival with Log Scale

The R formula is usually of the type: response variable ~ explanatory variable1 + explanatory variable2, etc. In this case, death is the response variable and serum creatinine is the only explanatory variable, but the distinction is unimportant for many plots.

Here serum_creatinine~death effectively means: produce boxplots of the serum creatinine data, split up by groups defined by values of the variable death. It's unfortunate that ~ means "has distribution" in written statistics and with R means "can be explained by".

The boxplots in Figure 5.16 show that the distribution of serum creatinine levels is quite different among the group of patients who died compared to those who survived. The group who died seem to have a higher median and a greater spread of values. These claims could be checked formally with hypothesis tests, but they stand out sufficiently with a reasonable amount of data for us to expect that the claims are true.

If we'd like to see kernel density estimates for each group on one plot, the following code will do this.

```
> plot(NULL, xlim=c(1e-2,1e1), ylim=c(0,1.6), ylab="Density",
  xlab="Serum Creatinine (mg/dL)",log="x",main="Kernel density
  estimate for serum creatinine in heart failure patients")
> lines(density(subset(heart$serum_creatinine,heart$death==1)),
  col="red",lwd=2)
> lines(density(subset(heart$serum_creatinine,heart$death==0)),
  col="blue",lwd=2)
> legend(x="topright",legend=c("died","survived"),
  col=c("red","blue"),lwd=2)
```

The first plot command sets up a blank plot other than the x and y axis labels and the title, but with log-spaced tick marks. This is done to set up the plot area to make sure that both the following lines plots will fit on it. That took trial and error after seeing how it worked with later lines. The first lines command adds the red density line which is constructed using only the subset of patients who died, while the next lines command adds a blue density line for the patients who survived. The last command adds a legend to the plot to explain the coloured lines. It results in Figure 5.17.

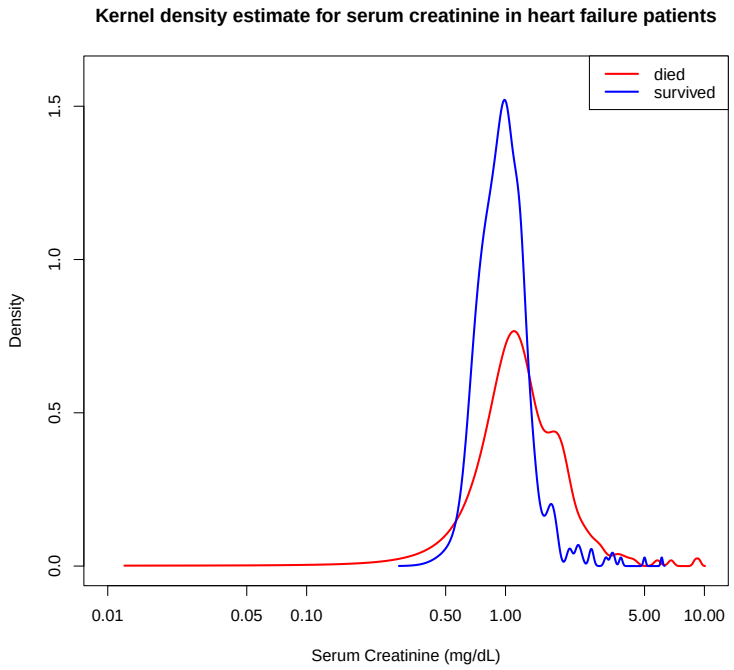


Figure 5.17: Kernel density estimate of serum creatinine levels by survival status with log scale

This plot illustrates the narrower range of serum creatinine levels for the surviving patients compared to the patients who died, as well as the relative absence of higher

values. The boxplot is probably the better option for clarity for an audience less familiar with statistics. The density plot contains more detailed information about the distributions in each group.

We can obtain numerical summaries for each group as follows using the **aggregate** function. This applies the listed function (here we've used summary) to each group, and then constructs a table showing the function output for each group.

```
> aggregate(heart$serum_creatinine, list(heart$death), FUN=summary)
```

```
  Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
1       0 0.500000  0.900000 1.000000 1.184877  1.200000 6.100000
2       1 0.600000  1.075000 1.300000 1.835833  1.900000 9.400000
```

5.8 Pairs of quantitative variables

One quantitative variable of some interest is the ejection fraction. In healthy people, this is generally over 50%. Numbers below this often indicate problems with the heart's ability to move blood around. This is a candidate to be a response variable, in addition to the death variable. The ejection fraction has only been recorded in a limited set of whole number percentage values. We can see this in R by running the following:

```
> sort(unique(heart$ejection_fraction))
> length(sort(unique(heart$ejection_fraction)))
```

```
 [1] 14 15 17 20 25 30 35 38 40 45 50 55 60 62 65 70 80
[1] 17
```

The unique command drops any duplicate values, sort puts the list of values in order from smallest to largest and length counts how many items are in the list.

Among 299 patients, only 17 different ejection fraction percentages have been recorded. This suggests that there is some rounding going on, either by the machine doing the estimate or people writing down the results. This doesn't stop us analysing ejection fraction like any other continuous variable, but the limited set of values will be visible in some plots and it's best to understand the origin.

We'll consider the pair: ejection fraction and serum sodium. Serum sodium also has only a small number of unique values in this dataset as can be seen below.

```
> sort(unique(heart$serum_sodium))
> length(unique(heart$serum_sodium))
```

```
 [1] 113 116 121 124 125 126 127 128 129 130 131 132 133 134
  135 136 137 138 139 140
[21] 141 142 143 144 145 146 148
[1] 27
```

This effectively means that the maximum number of combinations available for (serum creatinine, ejection fraction) pairs in this dataset is $17 * 27 = 459$. We have 299 observations, so some of these paired values might occur more than once. We can check this via the following command:

```
> nrow(unique(cbind(heart$serum_sodium,heart$ejection_fraction)))
```

```
[1] 125
```

So many of the pairs are unique, but $299 - 125 = 174$, so 174 of the patients have exactly the same pair of values recorded for these two variables as another patient. We'll consider the effect of this on plots shortly.

A scatterplot is a standard way of illustrating the relationship between pairs of quantitative variables. We'll view serum sodium as the explanatory variable and ejection fraction as the response variable. Convention puts the explanatory variable on the horizontal axis and the response variable on the vertical axis, with this being particularly common with linear regression analyses.

The R plot command below produces the scatterplot shown in Figure 5.18.

```
> plot(heart$serum_sodium, heart$ejection_fraction)
```

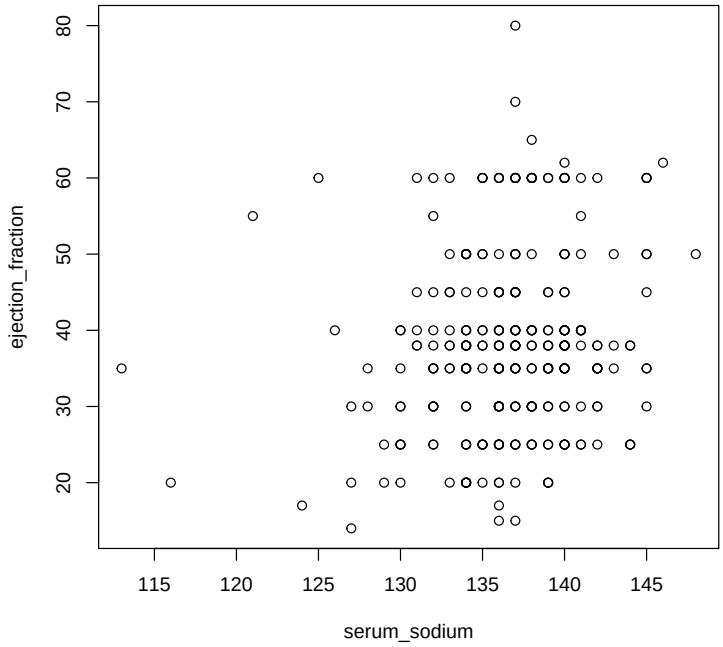


Figure 5.18: Scatterplot of serum sodium level against ejection fraction

Presumably there are 125 dots in Figure 5.18, trying to illustrate 299 patient records. We can improve this plot slightly by adding jitter to all of the points. I.e. a tiny random offset on one or both variables with the aim of showing where duplicate points are. With the code below, we add jitter to just the serum sodium variable, keeping the ejection fraction unchanged. One could instead apply jitter to both or just the ejection fraction values. We might as well also add a title, improve the axis labels and the appearance of the points. The R code below results in Figure 5.19.

```
> plot(jitter(heart$serum_sodium), heart$ejection_fraction,
xlab="Serum Sodium (mEq/L) - jittered",ylab="Left Ventricle
Ejection Fraction (%)", main="Ejection Fraction vs Serum
Sodium in Heart Failure Patients",col="green3")
```

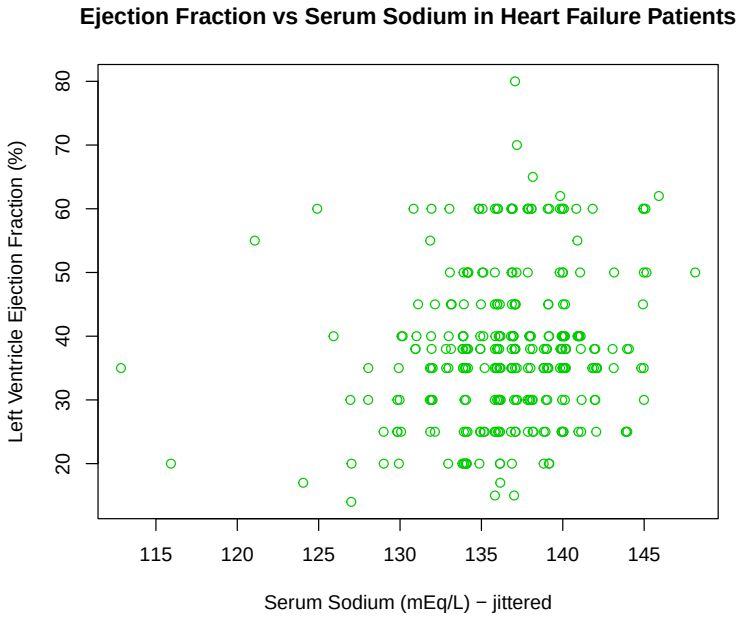


Figure 5.19: Jittered scatterplot of serum sodium level against ejection fraction

The jitter has separated observations with the same pair of values, highlighting areas of higher density. Here these occur mostly in the middle - from 130 to 140 mEq/L serum sodium and 25 to 60% ejection fraction. The main thing to look at on this plot is whether or not there is some sign of a trend, e.g. of increasing ejection fraction as serum sodium increases. There are small signs of such a trend. When the ejection fraction is lower, there seem to be more low values of serum sodium. Note that we should focus on the main mass of the points, not the occasional outlier.

We could produce similar scatterplots for every pair of quantitative variables. Since there are 12 variables overall, the data really exists in 12 dimensions. Any two variables can only offer a two-dimensional projection of this and can't tell the whole story.

For the 6 quantitative variables, we have two remaining powerful tools in R to summarise their relationships numerically and graphically. If you look at boxplots of each variable, you find that two benefit noticeably from a log transformation, in terms of having a more normal spread. These are serum creatinine (seen earlier) and cp (creatinine phosphokinase). We will start by putting all these quantitative variables together in one large dataframe of 6 columns and 299 rows.

```
> log10cp = log10(heart$cp)
> log10sc = log10(heart$serum_creatinine)
> qvars <-cbind(heart$age,log10cp,heart$ejection_fraction,
   heart$platelets,log10sc,heart$serum_sodium)
> dimnames(qvars)[[2]] <- c("age","log10cp","ejection
 fraction","platelets","log10sc","serum sodium")
```

We then produce a matrix of scatterplots showing every quantitative variable against every other quantitative variable with the following code, resulting in Figure 5.20.

```
> pairs(qvars)
```

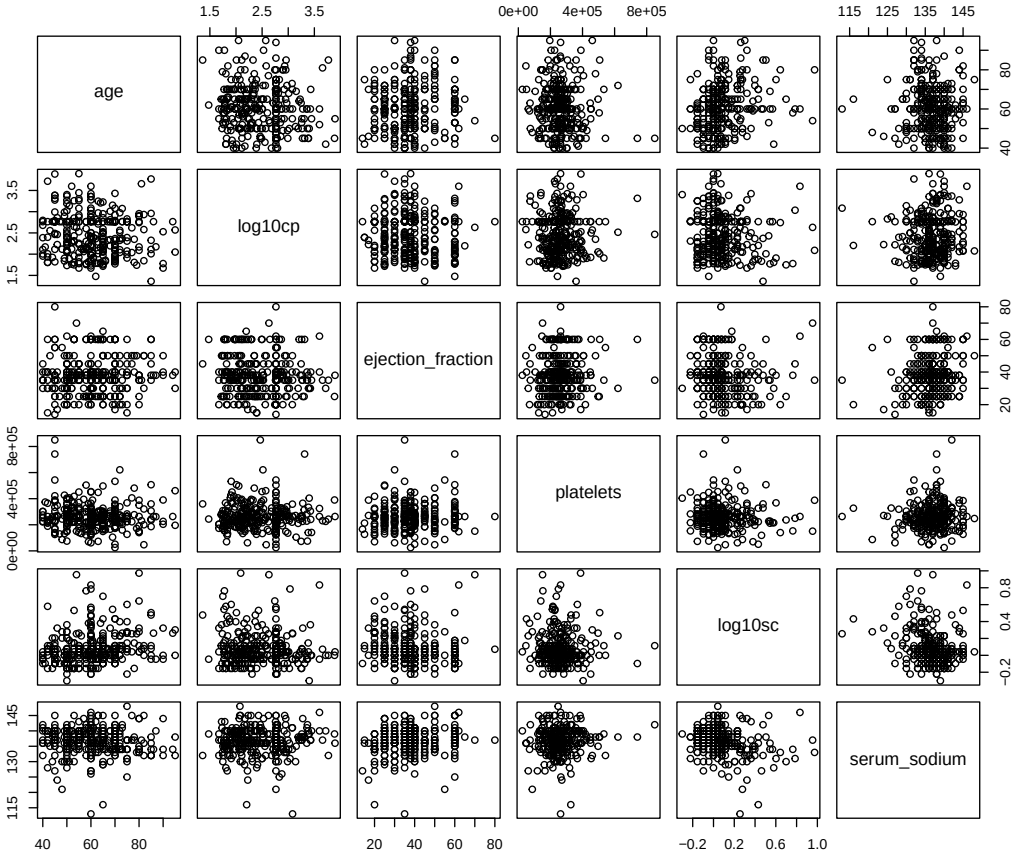


Figure 5.20: Scatterplots of quantitative variable pairs from heart dataset

The lower left triangle of this scatterplot matrix is the transpose of the upper right triangle, and so offers no new information. If desired, the lower left triangle can be suppressed using the following argument.

```
> pairs(qvars,lower.panel=NULL)
```

The pairs plot lets you look for variables with interesting relationships fairly quickly. None of the variable pairs have particularly strong relationships, but some look somewhat linked, such as log10sc ($\log_{10}$ of serum creatinine) vs serum sodium. It's not particularly easy to judge though, and it can be useful to complement this with a numerical counterpart: a sample correlation matrix, which is also symmetric.

This can be produced with the following command and for this data yields the result below.

```
> cor(qvars)
                          age     log10cp ejection fraction   platelets     log10sc serum_sodium
age               1.00000000 -0.09686136        0.06009836 -0.05235437  0.22909957  -0.04596584
log10cp          -0.09686136  1.00000000       -0.06912671  0.02151339 -0.06880571   0.01903598
ejection_fraction 0.06009836 -0.06912671        1.00000000  0.07217747 -0.09995198   0.17590228
platelets        -0.05235437  0.02151339        0.07217747  1.00000000 -0.04932872   0.06212462
log10sc           0.22909957 -0.06880571       -0.09995198 -0.04932872  1.00000000  -0.25670829
serum_sodium     -0.04596584  0.01903598        0.17590228  0.06212462 -0.25670829   1.00000000
```

A reminder that correlation is a value in the range $[-1, 1]$ which assesses the strength of a linear relationship between two variables. A value of +1 means that the two variables increase in lock step together. -1 means that as one increases, the other decreases. 0 means that there is no linear relationship – the best line through the data would be flat, meaning that the value of one variable tells you nothing about the value of the other. Otherwise, values with larger magnitude (absolute value) indicate stronger linear relationships. Here, the strongest correlation is estimated to be -0.257 between the $\log_{10}$ of serum creatinine and serum sodium. This suggests that they are anticorrelated – as one rises, the other tends to fall and vice versa.

The next strongest correlation is between $\log_{10}$ serum creatinine and age, estimated to be 0.229. Since the sign is positive, the variables are positively correlated, and so as one increases, so does the other, on average. Neither of these correlations is very strong, but all the others are weaker. Note that not every relationship need be linear and so correlation can miss substantial non-linear relationships. There's no sign of these from the pairs plots here though.

Both the pairs plot and correlation can be used with binary variables if desired. However, a pairs plot would need jitter to be clear.

Exploratory data analysis can give an initial impression of the data and consider issues such as outliers. It is not sufficient to make formal conclusions. For that, we need inferential statistics, particularly hypothesis testing and confidence intervals. These will be covered in later chapters, along with the construction of predictive models. Many of these methods make assumptions about the data, which can be checked via exploratory data analysis.

5.9 Exercises

1. Table 5.2 describes the 13 variables in a nutritional survey of elderly people.

Table 5.2: Description of the variables in a 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`

(a) Classify the variables.

(b) The data file **`nutri_restructured.csv`** contains the data of the nutritional study in Table 5.2, in which the variable are restructed to reflect their actual types and values. For comparision, the "raw" data file **`nutri_elderly.csv`** contains only the numerical data. Specify how to read the structured data into a data frame `nutri`.

(c) How many elderly are single, in a couple, in a family, or otherwise? Make a table of counts for the variable `situation`.

(d) Visualise the above table using a barchart or barplot.

(e) Make a box and whiskers plot of `age` against `situation`. Do the groups differ significantly in age?

(f) Plot `age` against `height`. What is the sample correlation between `age` and `height`?

(g) Give a 5-number summary of the variable `height`.

2. Measurements of the enzyme lactate dehydrogenase (LDH) in the blood were taken on seven subjects before fasting and after fasting. The results (mol/L) are shown in the following table:

Subject	1	2	3	4	5	6	7
Before	0.95	0.96	0.92	1.07	1.38	0.89	1.06
After	0.76	1.16	0.88	0.90	1.17	0.88	1.00
Decrease	0.19	−0.20	0.04	0.17	0.21	0.01	0.06

Calculate the five-number summary of the decreases in LDH and sketch a box plot for the distribution below.

3. Give an R command to calculate the sample variance for the set of data $\{1.23, 3.91, -2.15\}$.

4. Consider again the Alice data (the difference in heart rate after and before drinking caffeinated or decaf cola):

Caffeinated	17	22	21	16	6	−2	27	15	16	20
Decaf	4	10	7	−9	5	4	5	7	6	12

Identify the features and classify them as either quantitative or categorical.

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 intervals for a variety one- and two-sample models. Various probability distributions, such as the Student's t and the χ^2 distribution will make their first appearance.

6.1 Introduction

Recall the framework of statistical modeling in Figure 4.1. 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. ☞ 67

■ **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 here (or better, *datum*, as there is only one observation) 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

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

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
☞ 87 encountered these summary statistics already in Section 5.4.

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. ■

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, $X \sim \mathcal{N}(\mu, \sigma^2)$, 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
☞ 35 as $\mathbb{E}(X^k)$; see Definition 1.12. 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 (Method of Moments)** Let $X_1, \ldots, X_n$ be iid copies of $X \sim \mathcal{N}(\mu, \sigma^2)$. 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

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

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 $X_1, \ldots, X_n$ 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 the rules for expectation and variance in Chapter 3. In particular,

☞ 59 by (3.6) we have

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

☞ 61 and from (3.9) 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

☞ 61 μ for large n. This is in essence the law of large numbers; see Section 3.5.

To assess 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: Confidence Interval

Let $X_1, \ldots, X_n$ be random variables with a joint distribution depending on a parameter θ. Let $T_1 < T_2$ be functions of the data $X_1, \ldots, X_n$ 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 $\mathbb{P}(9.5 < \theta < 10.5) = 0.9$? 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

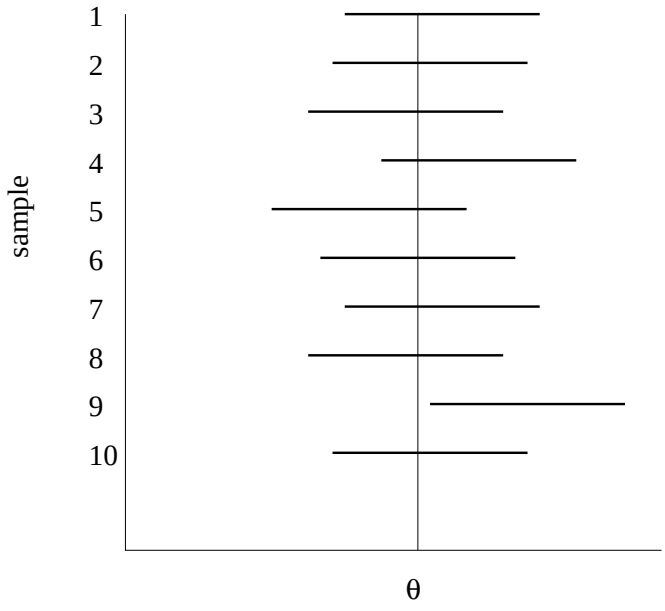


Figure 6.1: Possible outcomes of stochastic confidence intervals.

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 $(9.5, 10.5)$.

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.
☞ 62 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

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

Standardising $\bar{X}$ gives

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

In order to construct a confidence interval for μ, we would like to create a so-called **pivot** variable that (1) depends on all the data and on the parameter to be estimated and (2) has a distribution that does not depend on any unknown parameters. The above standardized 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 and more like the constant σ^2 as n grows larger. So, we have for large n

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

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 quan-
☞ 44 tile, both of which are good to remember; see also Section 2.5. Because the standard normal distribution is symmetrical around 0, we also have, for example,

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

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, $-3 < -2$ implies $3 > 2$. 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: Approximate Confidence Interval for μ

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

$$\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right), \text{ abbreviated as } \bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}. \tag{6.6}$$

The quantity $z_{\alpha/2}S/\sqrt{n}$ is called the **margin of error** for the confidence interval.

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 (Oil Company)** 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 realizations 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, $X_1, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(\mu, \sigma^2)$, it is possible to construct *exact* confidence intervals for μ and σ^2, rather than only approximate ones.

Confidence Interval for μ

For iid $\mathcal{N}(\mu, \sigma^2)$ 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 $Z \sim \mathsf{t}_{\text{df}}$ to indicate that a random variable Z has a student distribution with df degrees of freedom. Here is the exact version of the approximate result (6.5).

Theorem 6.2: Standardized Mean and Student *t*-distribution

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. 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 for the normal distribution, but the tails of the distribution are "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 $\mathcal{N}(0, 1)$ distribution.

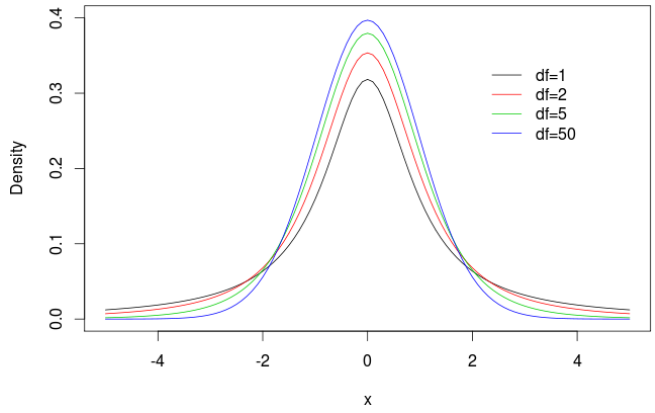


Figure 6.2: The pdfs of Student t distributions with various degrees of freedom (df).

We can use R to calculate the pdf, cdf, and quantiles for this distribution. For example, the following R script produces 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_{df} distribution for df $= 1, 2, 5, 50$, 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 given by `qnorm(0.975)` $= 1.959964$ (≈ 1.96).

Returning to the pivot T in (6.5), it has a t_{n-1} distribution. By repeating the rearrangement steps 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. ☞ 124

Theorem 6.3: Exact Confidence Interval for μ

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

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}. \tag{6.8}$$

■ **Example 6.5 (Volume of a Drop of Water)** 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 $\mathcal{N}(\mu, \sigma^2)$ 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 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 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Consider the sample variance

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

It turns out that $(n-1)S^2/\sigma^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/\sigma^2$ 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 $Z \sim \chi^2_{\text{df}}$ to denote that Z has a chi-square distribution with df 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.

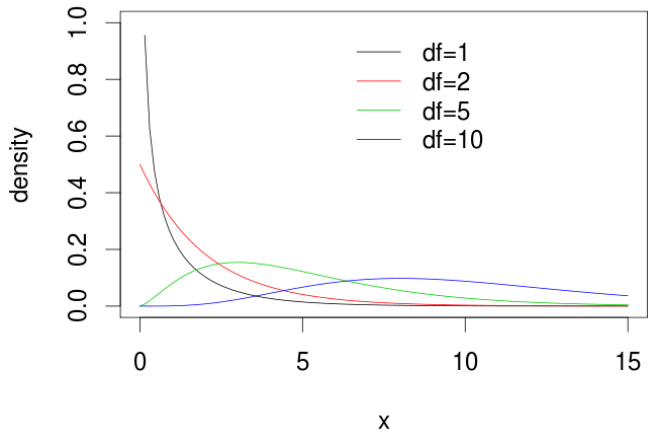


Figure 6.3: The pdfs of chi-square distributions with various degrees of freedom (df).

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
```

Because $(n-1)S^2/\sigma^2$ has a χ^2_{n-1} distribution, 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 6.4

Let $X_1, X_2, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(\mu, \sigma^2)$ 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 σ^2 is

$$\left(\frac{(n-1)S^2}{q_2}, \frac{(n-1)S^2}{q_1} \right) . \tag{6.9}$$

■ **Example 6.6 (Aspirin)** 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 $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ from respectively a $\mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathcal{N}(\mu_Y, \sigma_Y^2)$ distribution. We wish to make a confidence interval (approximate or exact) for $\mu_X - \mu_Y$.

Approximate Confidence Interval for $\mu_X - \mu_Y$

To make an approximate confidence interval for $\mu_X - \mu_Y$, we can reason in similar way as in Section 6.3.1. By the central limit theorem, we have

$$\bar{X} - \bar{Y} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} \right).$$

So, if we standardize and replace σ_X^2 and σ_Y^2 with their sample variances, we have

$$\frac{(\bar{X} - \bar{Y}) - (\mu_Y - \mu_Y)}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \overset{\text{approx.}}{\sim} \mathcal{N}(0, 1).$$

For small m and n the standard normal approximation may not be very accurate. Fortunately, it is possible to obtain a much better approximation using a Student distribution where df is given by the so-called **effective degrees of freedom**:

$$\mathrm{df} = \frac{\left(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_X^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_Y^2}{n_2}\right)^2}. \tag{6.10}$$

We thus have the following approximate confidence interval for $\mu_X - \mu_Y$, using the above *Welch-Satterthwaite approximation*.

Theorem 6.5: Confidence Interval for $\mu_X - \mu_Y$

Let $X_1, X_2, \ldots, X_{n_1} \sim_{\text{iid}} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_{n_2} \sim_{\text{iid}} \mathcal{N}(\mu_Y, \sigma_Y^2)$ be independent. Let $t_{\alpha/2,\text{df}}$ be the $1-\alpha/2$ quantile of the Student's t_{df} distribution, with df as in (6.10). An approximate $1-\alpha$ stochastic confidence interval for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm t_{\alpha/2,\text{df}} \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}. \tag{6.11}$$

Non-normal distributions

If the two samples come from non-normal distributions, which need not be the same, we can usually use 6.11 to produce approximate confidence intervals for the difference in the means $\mu_X - \mu_Y$. This is due to the effect of the central limit theorem, which is to give $\overline{X}$ and $\overline{Y}$ approximately normal distributions provided n_1 and n_2 are reasonably large (say 30 or more) or the data distributions for X_i and Y_j are approximately normal, $i \in \{1, \ldots, n_1\}$, $j \in \{1, \ldots, n_2\}$. For most applications, we will not know whether or not the data came from a normal distribution. However, we can check the sample size and visually determine whether the data distribution is close to normal using plots such as histograms and Q-Q plots. If the sample sizes are particularly small and/or the data distributions are highly skewed, there are nonparametric alternatives such as the Wilcoxon rank-sum test.

■ **Example 6.7 (Human Movement Study)** 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. The distribution of weights in each group is approximately normal. Are the average masses of the two groups different?

Here we have $n_1 = 15$ and $n_2 = 10$. The outcomes $\bar{X} - \bar{Y}$ is $\bar{x} - \bar{y} = 62 - 71.5 = -9.5$. Using (6.10), the effective degrees of freedom is df $= 22.09993$. You may verify also that

$$\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}} = 5.422177.$$

To construct a 95% numerical confidence interval for $\mu_X - \mu_Y$, we need to also evaluate the 0.975 quantile of the t_{df} distribution, using the `R` command `qt(0.975,22.09993)`. This gives 2.073329. So that the 95% numerical confidence interval for $\mu_X - \mu_Y$ is given by

$$-9.5 \pm 2.073329 \times 5.422177 = (-20.74,\ 1.74)\ .$$

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

6.3.4 Binomial Data, One Sample

How do we construct an approximate confidence interval for binomial data? Let us look at a concrete example first.

■ **Example 6.8 (Opinion Poll)** In an opinion poll of 1000 registered voters, 227 voters say they will vote for the Greens. How can we construct a 95% confidence interval for the proportion p of Green voters of the total population? A systematic way to proceed is to view the datum, 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.
☞ 61 Instead, for large n we rely on the central limit theorem (see Section 3.5) to construct an *approximate* confidence interval. The reasoning is as follows:

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

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2\ .$$

Then, using the pivot variable

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

which is approximately standard normal, we have

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\widehat{P} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha \ .$$

Rearranging gives:

$$\mathbb{P}\left(\widehat{P} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \widehat{P} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha \ .$$

This would suggest that we take $\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ 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

$$\mathbb{P}\left(\widehat{P} - z_{\alpha/2}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}} < p < \widehat{P} + z_{\alpha/2}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}\right) \approx 1 - \alpha \ .$$

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

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \ . \tag{6.12}$$

We can use this approximate confidence interval provided the normal approximation for the binomial is reasonable. We will use the conditions $np \geqslant 5$ and $n(1-p) \geqslant 5$. Since p is unknown, we can approximate these conditions by checking that $n\widehat{p} \geqslant 5$ and $n(1-\widehat{p}) \geqslant 5$, i.e. checking that we have at least 5 successes and 5 failures.

■ **Example 6.9 (Opinion Poll (Continued))** In Example 6.8, we have $n = 1000$, $\widehat{p} = 227/1000 = 0.227$, 227 successes and 773 failures, so easily meet the conditions. We would like an approximate 95% numerical CI for p, so use $z_{0.025} = 1.960$ and obtain

$$(0.227 - 1.960 \times 0.0132, 0.227 + 1.960 \times 0.0132) = (0.20, 0.25) \ .$$

■

6.3.5 Binomial Data, Two Samples

We next wish to construct an approximate confidence interval for the difference of two proportions. We again start with an example.

■ **Example 6.10 (Nightmares)** 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; $X \sim \mathsf{Bin}(n_1, p_X)$, where p_X is unknown.
- Let Y be the number of "successes" in Group 2; $Y \sim \mathsf{Bin}(n_2, p_Y)$, where p_Y is unknown.
- Assume X and Y are independent.

We wish to compare the two proportions via an approximate $(1-\alpha)$-confidence interval for $p_X - p_Y$. The easiest way is to again rely on the central limit theorem. The normal approximation to the binomial distribution can be applied if $n_1\widehat{p}_X$, $n_1(1-\widehat{p}_X)$, $n_2\widehat{p}_Y$ and $n_2(1-\widehat{p}_Y) \geqslant 5$.

Let $\widehat{P}_X = X/n_1$ and $\widehat{P}_Y = Y/n_2$. By the central limit theorem,

$$\frac{\widehat{P}_X - \widehat{P}_Y - (p_X - p_y)}{\sqrt{\frac{p_X(1-p_X)}{n_1} + \frac{p_Y(1-p_Y)}{n_2}}}$$

has approximately a $\mathcal{N}(0,1)$ distribution. Hence, with $z_{\alpha/2}$ the $(1-\alpha/2)$-quantile of the $\mathcal{N}(0,1)$ distribution (as in Section 6.3.4), we have

$$\mathbb{P}\left(-z_{\alpha/2} \leqslant \frac{\widehat{P}_X - \widehat{P}_Y - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{n_1} + \frac{p_Y(1-p_Y)}{n_2}}} \leqslant z_{\alpha/2}\right) \approx 1-\alpha\ .$$

Rewriting, this gives

$$\begin{aligned}\mathbb{P}\Bigg(\widehat{P}_X - \widehat{P}_Y - z_{\alpha/2}\sqrt{\frac{p_X(1-p_X)}{n_1} + \frac{p_Y(1-p_Y)}{n_2}} \leqslant p_X - p_Y \\ \leqslant \widehat{P}_X - \widehat{P}_Y + z_{\alpha/2}\sqrt{\frac{p_X(1-p_X)}{n_1} + \frac{p_Y(1-p_Y)}{n_2}}\Bigg) \\ \approx 1-\alpha.\end{aligned}$$

As in the 1-sample case of Section 6.3.4, the same is *approximately* true, if we replace p_X and p_Y in the square root terms 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, a numerical *approximate* $100(1-\alpha)\%$ confidence interval for $p_X - p_Y$ is

$$\widehat{p}_X - \widehat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n_1} + \frac{\widehat{p}_Y(1-\widehat{p}_Y)}{n_2}}\,. \tag{6.13}$$

■ **Example 6.11 (Nightmares (Continued))** We continue Example 6.10. We have $\widehat{p}_X = 55/160$, $\widehat{p}_Y = 60/192$, so plenty of successes and failures in each group. We have $z_{0.025} = 1.96$, so that an approximate 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. ■

6.4 Exercises

1. The following data were drawn from a normal distribution with unknown mean μ and standard deviation σ:

   2.86, 4.16, 6.35, 3.92, 5.96, 5.12, 6.43, 3.03, 4.80, 3.79,

   Give a 95% confidence interval for μ.

2. An oil company wishes to investigate how much on average each house-hold 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. Construct a 99% confidence interval for the expected amount of money per year that the households in Melbourne spend on petrol and heating oil.

3. In a pilot study of size $n = 100$ it was found that a 95% confidence interval for the mean weight of elderly people in nursing homes was (59.3, 82.5) kg. For the actual study, the minister for health would like to see a 95% confidence interval that is much narrower. What sample size should be taken to achieve a margin of error of 1 kg for the actual study?

4. Your printer has started to jam when you try to print. You suspect that the actual thickness of your printer paper is not as advertised, namely 100 μm. You select 10 sheets of the paper and measure their thicknesses to be:

Thickness (μm)	97	102	104	96	108	105	101	99	96	102

   Construct a 95% confidence interval for the paper thickness. Is there any evidence to suggest the paper thickness is not as advertised?

5. A sample of 4 students were asked how many hours they slept during the previous weekend. From their responses, a 95% confidence interval for the mean number of hours slept by all students in this population was 16.753 ± 7.39 hours. Based on this estimate, what is the smallest number of students who would need to be surveyed to estimate the mean hours of sleep with a margin of error of no more than 2 hours?

6. In an election poll 350 out of 980 women and 420 out of 1108 men said they would vote for "Shady" Shane. Give a 90% confidence interval for the difference in the proportion of "Shady" voters amongst women and men.

7. A juice company wants to find out the variation, as measured by variance σ^2, of the amount of juice in their 1.5L bottles. The company statistician took a random sample of 27 bottles from the production line and the sample standard deviation was 5.86ml. Find the 95% confidence interval for σ^2.

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 one- and two-sample problems involving Normal and Binomial random variables are provided.

7.1 Introduction

We had a first look at hypothesis testing in Chapter 4. Namely, in Section 1.2 we investigated a coin flip experiment (is the coin fair?) and in Section 4.3 we studied Alice's cola experiment (does drinking caffeinated Diet cola increase the heart rate?). In this chapter we will revisit both these experiments and describe their analysis in a framework that is more generally applicable.

☞ 9
☞ 71

In particular, suppose that we have a general model for data $\mathbf{X}$ that is described by a family of probability distributions that depend on a parameter $\boldsymbol{\theta}$. For example, in the one sample normal model, we have $\mathbf{X} = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(\mu, \sigma^2)$. 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 $p = \min\{2\mathbb{P}_{H_0}(T \leqslant t),\ 2\mathbb{P}_{H_0}(T \geqslant t)\}$ or $p = 2\mathbb{P}_{H_0}(T \geqslant |t|)$.

The smaller the P-value, the greater the strength of the evidence against H_0 provided
☞ 138 by the data. As a rule of thumb (see also Figure 7.1 below):

$p < 0.10$ weak evidence,
$p < 0.05$ moderate evidence,
$p < 0.01$ strong evidence.

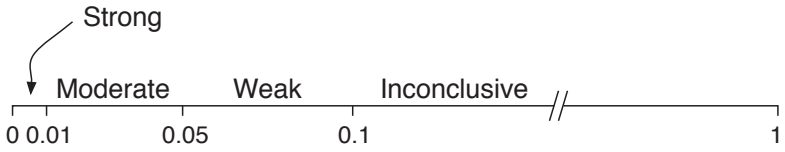


Figure 7.1: Strength of evidence for a P-value.

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 summarize 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 $X_1, \ldots, X_{101} \sim_{\text{iid}} \mathcal{N}(\mu, 49)$. Alternatively, the model could simply be $\bar{X} \sim \mathcal{N}(\mu, 49/101)$, 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 $\bar{X} \sim \mathcal{N}(127, 49/101)$. 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 $Z \sim \mathcal{N}(0, 1)$. 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 *strong* evidence that the blood pressure of diabetics differs from the general public. ■

■ **Example 7.2 (Biased Coin (Revisited))** We revisit Example 1.1, where we observed 60 out of 100 Heads for a coin that we suspect to be biased towards Heads. Is there enough evidence to justify our suspicion? ☞ 10

What are the 7 hypothesis steps in this case? A good model (step 1) for the data X (the total number of Heads in 100 tosses) is: $X \sim \mathsf{Bin}(100, p)$, with the probability of Heads, p, is unknown. We would like to show (step 2) the hypothesis $H_1 : p > 1/2$; otherwise, we do not reject (accept) the null hypothesis $H_0 : p = 1/2$. Our test statistic

(step 3) could simply be X. Under H_0, $X \sim \mathsf{Bin}(100, 1/2)$ (step 4). The outcome of X (step 5) is $x = 60$, so the P-value for this right one-sided test is

$$\mathbb{P}(X \geqslant 60) = \underbrace{\sum_{k=60}^{100} \binom{100}{k} \left(\frac{1}{2}\right)^{100}}_{\texttt{1-pbinom(59,100,1/2)}} \approx 0.02844397 \,.$$

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 One-sample t-test

☞ 71 Let us return to Alice's cola experiment in Section 4.3, and consider only the changes in pulse rate for the Decaf (control) group; see Table 7.1. Is there evidence that the expected change in pulse rate is greater than 0 for this group? If we found such evidence for the control group, this would put doubt on any conclusion that an increase in pulse rate for the treatment group is only due to caffeine — there could be other factors involved.

Table 7.1: Changes in pulse rate for the Decaf group in Alice's cola experiment.

4	10	7	–9	5	4	5	7	6	12

To answer this question, we again consider an appropriate model for this situation. We represent the observations by $X_1, \ldots, X_{10}$, and assume that they form an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, where both μ and σ^2 are *unknown*; note that is different to Example 7.1, where the variance is known. The hypotheses can now be formulated as: $H_0 : \mu = 0$ against $H_1 : \mu > 0$.

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

$$H_0 : \mu = 0, \quad 0 < \sigma^2 < \infty \,.$$

It is the unknown variance 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 the standard deviation σ is small, say 1. In that case, $\bar{X}$ is under H_0 very much concentrated around 0, and therefore any deviation from 0, such as 7 would be most

unlikely under H_0. We would therefore reject H_0. On the other hand, if σ is large, say 10, then a value of 7 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 "scale" it with the standard deviation. That is, we should measure our deviation from 0 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

$$T = \frac{\bar{X}}{S/\sqrt{10}} .$$

The factor $\sqrt{10}$ is a "standardising" constant which enables us to utilize Theorem 6.7. ☞ 126
Namely, under H_0 the random variable T has a $\mathsf{t}_{n-1} = \mathsf{t}_9$ distribution. Note that this is true for *any* value of σ^2. The observed outcome of T is

$$\frac{5.1}{5.59/\sqrt{10}} \approx 2.89 .$$

Using R,

```
> 1 - pt(2.89,df=9)
```

```
[1]  0.008942135
```

we find the P-value

$$\mathbb{P}_{H_0}(T \geqslant 2.89) \approx 0.0089 .$$

Since this is rather small, we reject H_0. Therefore, there is strong evidence that the expected difference in pulse rate is greater than 0.

The above test is often called a **one sample *t*-test**. In general, let $X_1, \ldots, X_n \sim_{\text{iid}} \mathcal{N}(\mu, \sigma^2)$. 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/accept H_0 based on the magnitude of the P-value. Note that the P-value 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).

Let us look more closely at the way the P-value is calculated. It depends on the distribution of the test statistic under the null hypothesis, the value of the test statistic and the choice of alternative hypothesis. The area under the pdf corresponding to the P-value is illustrated in Figure 7.2.

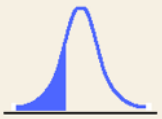
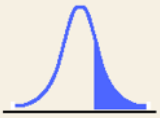
Statement of H_1		*p*-Value Area	*t*-Curve Region
$\mu < \mu_0$	(less than)	Area to the left of t (even if $t > 0$)	
$\mu > \mu_0$	(greater than)	Area to the right of t (even if $t < 0$)	
$\mu \neq \mu_0$	(not equal)	$2 \times$ area to the right of $\lvert t \rvert$	

Figure 7.2: Calculation of P-values.

Consider for example $H_1 : \mu < \mu_0$. This means that the extreme direction of interest is to the left. I.e. if we knew σ, the P-value would be the probability under H_0 that the sample mean $\overline{X}$ would be less than the observed $\overline{x}$, i.e. $p = \mathbb{P}_{H_0}(\overline{X} < \overline{x})$. When we do not know σ, we use the test statistic $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$. Both S and n are positive, so the direction of the extreme is the same for T. I.e. after the transformation, $p = \mathbb{P}_{H_0}(T < t)$, where $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$. The direction of the extreme is also maintained for the other two possible alternative hypotheses.

Using R

Note that in to order to carry out a one sample t-test, we only need the summary statistics $\bar{x}$ and s of the data. When the individual measurements are available, it is convenient to carry out the t-test using the R function **t.test**. As an example, we enter the data in Table 7.1 into R and print out the sample mean and standard deviation using the function **sprintf**, which can be used to format output neatly:

```
> x = c(4, 10, 7, -9, 5, 4, 5, 7, 6, 12)
> sprintf(fmt="mean=%s   sd=%3.3f", mean(x), sd(x))


"mean=5.1   sd=5.587"
```

Applying the **t.test** function, we get for the above data:

```
> t.test(x,alternative="greater")


  data:  x
t = 2.8868, df = 9, p-value = 0.008989
alternative hypothesis: true mean is greater than 0
```

```
95 percent confidence interval:
 1.8615    Inf
sample estimates:
mean of x
      5.1
```

The main output of the function `t.test` are: the outcome of the T statistic (t = 2.8868), the P-value = 0.008989, the alternative hypothesis (*true mean is greater than 0*) and the sample mean $\bar{x}$ = 5.1. To output just the P-value, we can use:

```
> t.test(x,alternative="greater")$p.value
```

```
[1] 0.008988979
```

By default, the `t.test` function takes a two-sided alternative. The option `alternative = "greater"` forces a right-one-sided alternative. Note that in this case `t.test` returns a one-sided confidence interval. To obtain a 99% two-sided confidence interval for μ we can use:

```
> t.test(x,conf.level=0.99)$conf.int
```

```
[1] -0.6413761 10.8413761
attr(,"conf.level")
[1] 0.99
```

To find the variable names that are returned by a function, use `names()`, as in

```
h = t.test(x)
names(h)
```

```
[1] "statistic" "parameter" "p.value" "conf.int" "estimate"
[6] "null.value" "alternative" "method" "data.name"
```

7.3 Type-I Error, Type-II Error, and Power

In any hypothesis test we can make two types of mistakes, illustrated in Table 7.2.

Table 7.2: Type-I and type-II errors

	True state of nature	
Decision	$\boldsymbol{H_0}$ **is true**	$\boldsymbol{H_1}$ **is true**
Accept $\boldsymbol{H_0}$	Correct	Type II Error
Reject $\boldsymbol{H_0}$	Type I Error	Correct

Whether we make a right or wrong decision is the result of a random process. Thus, for any statistical test where we make a decision in the end, there is a *probability* of a Type I error, Type II error, or correct decision. Ideally, we would like to construct tests which make the probabilities of Type-I and Type-II errors, (let's call them e_I and e_{II}) as small as possible. Unfortunately, this is not possible, because the two errors "compete" with each other: if we make e_I smaller, e_{II} will increase, and vice versa.

Because, as mentioned, the null and alternative hypothesis do not play equivalent roles, a standard approach is to keep the probability e_I of a Type I error at (or below) a certain threshold: the significance level, say 0.05. The decision rule: *reject* H_0 *if the* P-value *is smaller than some significance level* α ensures that $e_I \leqslant \alpha$.

Next, given that e_I remains at (or below) level α, we should try to make e_{II} as small as possible. The probability $1 - e_{II}$ is called the **power** of the test. It is the probability of making the right decision (reject H_0) under some alternative in H_1. So, minimizing the probability of a Type II error is the same as maximizing the power. Note that the power heavily depends on what alternative is used.

■ **Example 7.3 (Simulating the Power)** Suppose we have a one-sample t-test, where we want to test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. Our test statistic is $\bar{X}/(S/\sqrt{n})$ and under H_0, this test statistic has a t_{n-1} distribution. Suppose we have a significance level of $\alpha = 0.05$. What is the power of the test when the real parameters are $\mu = 1$ and $\sigma = 2$, for example?

Imagine what would happen if we conducted the test tomorrow, with the data $X_1, \ldots, X_n$ coming from $\mathcal{N}(1, 4)$. We would form the test statistic $T = \bar{X}/(S/\sqrt{n})$ and then calculate the corresponding P-value for this right-one-sided test. In `R` we would do it via: `pt(T,df=n-1)`. Finally, we would reject the null hypothesis if the P-value is less than 0.05. So let's do this many times on a computer and see how many times we correctly reject the null hypothesis. In the program below we use a sample size of $n = 5$.

```
R = 1e5   # number of repeats
n = 5     # sample size
mu = 1    # the actual mu
sigma = 2 # the actual standard deviation
pval = vector(mode="numeric", length=R) # initialize P-value vector
t = vector(mode="numeric", length=R)    # initialize test statistic vector

for (i in 1:R){
  x = rnorm(n,mean=mu,sd=sigma) # simulate the data
  t[i] = mean(x)/(sd(x)/sqrt(n)) # test statistic
  pval[i] = 1-pt(t[i],df=n-1)  # P-value
}

pow = print(length(which(pval < 0.05))/R) # estimate of the power

# Or we can use the power.t.test function:
power.t.test(n=n, delta=mu, sig.level=0.05, alternative="one.sided", type="
    one.sample", sd=sigma)
```

In this way, we calculate a power of 0.24 for the alternative $\mu = 1, \sigma = 2$. This is not so high! With this standard deviation and small sample size it will be very difficult to detect that $\mu = 1$. Let's repeat it with $n = 50$. We now get a power of 0.97, so close to 1. Hence a sample size of 50 is enough to detect a difference of 1 unit, if the standard deviation is 2.

The above example illustrates that the power depends on various factors: the sample size, the significance level, as well as μ and σ. In fact (you can verify it yourself) the power in the above code only depends on μ/σ, which is sometimes called the "signal to strength ratio". In R, we can make power calculations via the **power.t.test** function. Here is the output of the last lines in the code above:

```
One-sample t test power calculation

              n = 5
          delta = 1
             sd = 2
      sig.level = 0.05
          power = 0.2389952
    alternative = one.sided
```

A power analysis, as carried out above, allows us to choose a sample size large enough to determine some minimal effect, as long as we have an idea of the standard deviation. The latter can be estimated with a trial run, for example.

7.4 One-sample Test for Proportions

The statistical test in Example 7.2 is an example of a one-sample test for proportions. In this section we explore such tests in more detail, using the following example.

■ **Example 7.4 (Market Research)** 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 $X_1, \ldots, X_{10}$ be such that

$$X_i = \begin{cases} 1 & \text{if person } i \text{ prefers the new yoghurt,} \\ 0 & \text{if person } i \text{ prefers the old yoghurt,} \end{cases}$$

$i = 1, \ldots, 10$. We assume that $X_1, \ldots, X_{10}$ 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,

$$\mathbb{P}_{H_0}(X \geqslant 8) = \sum_{k=8}^{10} \binom{10}{k} (1/2)^{10} \approx 0.0546875 \, .$$

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 $X_1, \ldots, X_n$ 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.

■ **Example 7.5 (Market Research (Continued))** For one-sample binomial test, we can use the R function **binom.test**. The parameters and output are very similar to those used with **t.test** function for one-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 $\mathcal{N}(np, np(1-p))$ distribution and then the estimator $\widehat{P} = X/n$ has approximately a $\mathcal{N}(p, p(1-p)/n)$ distribution. It follows that

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

has approximately a $\mathcal{N}(0,1)$ 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 $\mathcal{N}(0,1)$ distribution.

■ **Example 7.6 (Market Research (Continued))** 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 quite different from the one for the exact test (0.05469), as our sample size is not enough large to use the central limit theorem. The R function **prop.test** uses this normal approximation, as in: `prop.test`:

```
> prop.test(x=8,n=10,p=0.5,alternative="greater",correct=FALSE)$p.value

[1] 0.02888979
```

Hence, for small sample sizes it is recommended to use **binom.test**. ■

7.5 Two-sample *t*-test

We next look at two-sample data, again using Alice's cola experiment as a guiding example. Below we repeat the table and stripplot from Section 4.3. Do the results in Table 7.3 provide any *evidence* that caffeine increases pulse rate?

Table 7.3: Changes in pulse rate for Alice's caffeine experiment.

Caffeinated	17	22	21	16	6	−2	27	15	16	20
Decaf	4	10	7	−9	5	4	5	7	6	12

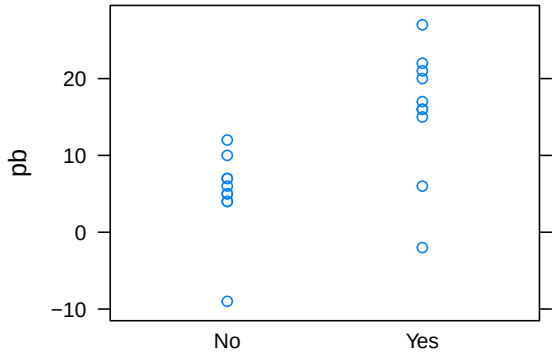


Figure 7.3: A visualization of Alice's caffeine data.

Let us go through the 7 steps of a hypothesis test. First, we could model the data as coming from different normal distributions. Let $X_1, \ldots, X_{n_1}$ (with $n_1 = 10$) be the change in heartbeat for the caffeinated Diet cola (treatment) group and let $Y_1, \ldots, Y_{n_2}$ (with $n_2 = 10$) be the change in heartbeat for the decaf Diet cola (control) group. We assume that

- $X_1, \ldots, X_{n_1} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$.
- $Y_1, \ldots, Y_{n_2} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$.
- $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ 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 $H_1 : \mu_X > \mu_Y$. It is useful to rewrite the alternative hypothesis as $H_1 : \mu_X - \mu_Y > 0$ so that we can keep track of the direction of the extreme in calculating the P-value.

☞ 130 Following the reasoning in Section 6.3.3, we use as our test statistic:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}},$$

which under H_0 has approximately a Student t_{df} distribution where df is given by

$$\text{df} = \frac{\left(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_X^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_Y^2}{n_2}\right)^2}, \tag{7.2}$$

which we already encountered in (6.10). Even for small n_1 and n_2 this approximation is very accurate. This two-sample *t*-test is attributed to *Bernard Welch*. This completes steps 1–4. Let us finish the remaining steps of the test by using R as a calculator. Note that the P-value is calculated using the area under the H_0 test statistic pdf curve to the right of the test statistic, matching the alternative hypothesis.

```
x = c(17,22,21,16,6,-2,27,15,16,20)
y = c(4,10,7,-9,5,4,5,7,6,12)
mx = mean(x)
my = mean(y)
sx = sd(x)
sy = sd(y)
a = sx^2/10
b = sy^2/10
t = (mx - my)/sqrt(a + b)
df = (a + b)^2/(a^2/9 + b^2/9)
pval = 1 - pt(t,df=df)
cat("t = ", t, ", df =", df,", pval=", pval)  #print the values
```

This gives the output (using the **cat** (for concatenate) function):

```
t = 3.37521 , df = 15.74042 , pval = 0.001965818
```

We conclude that there is strong evidence that the caffeine has an effect on the change in pulse beat.

Having defined `x` and `y` as in the above code, we can obtain the same results by using the **t.test** function:

```
> t.test(x,y, alternative="greater")

        Welch Two Sample t-test

data:  x and y
t = 3.3752, df = 15.74, p-value = 0.001966
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
```

```
 5.159642      Inf
sample estimates:
mean of x mean of y
     15.8       5.1
```

If using statistical tables instead, the calculated degrees of freedom should be rounded down to the nearest integer.

Equal Variance Assumption

In the above analysis, we did not assume that the variances for both groups were equal. If we *do* make such an assumption, it is possible to obtain a test statistic with an *exact* (not just approximate) Student distribution. The reasoning is as follows. To estimate the common variance of the groups (σ^2, say), we should "pool" the squared deviations from the means, giving the **pooled sample variance**

$$S_p^2 = \frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2 + \sum_{j=1}^{n_2}(Y_j-\bar{Y})^2}{n_1+n_2-2},$$

where $\bar{X} = \sum_{i=1}^{n_1} X_i/n_1$ and $\bar{Y} = \sum_{j=1}^{n_2} Y_j/n_2$. Since, under $H_0 : \mu_X = \mu_Y$ the random variable $\bar{X} - \bar{Y}$ has a $\mathcal{N}(0, \sigma^2\,(1/n_1 + 1/n_2))$ distribution, a natural test statistic in this case is

$$T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{1/n_1+1/n_2}}\,, \tag{7.3}$$

It turns out that, under H_0 this test statistic has exactly a $\mathrm{t}_{n_1+n_2-2}$ distribution. We accept/reject H_0 depending on the P-value associated with the alternative (left-, right-, or two-sided).

For the Alice example, we can perform this test using:

```
t.test(x,y,alternative = "greater", var.equal = T)
```

```

        Two Sample t-test

data:  x and y
t = 3.3752, df = 18, p-value = 0.001686
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.202718      Inf
sample estimates:
mean of x mean of y
     15.8       5.1
```

We see that the P-value is slightly smaller under the assumption of equal variances, but that in essence we come to the same conclusions.

Paired Data

When conducting a two-sample t-test, it is important to ascertain that the random variables are not *paired*. Such data often arises in "before–after" experiments or on replicated experiments involving the same subjects, as in the following example.

■ **Example 7.7 (Paired Lab Data)** We wish to compare the results from two labs for a specific examination. Both labs made the necessary measurement on *the same* fifteen patients.

```
> lab1 = c(22,18,28,26,13,8,21,26,27,29,25,24,22,28,15)
> lab2 = c(25,21,31,27,11,10,25,26,29,28,26,23,22,25,17)
```

In this case the measurements between the goups are not independent, as the measurements are conducted on the same patient. For example, both labs report high measurements (29 and 28) for patient 10, and both labs reported low measurements (8 and 10) for patient 6. ■

In general, suppose we wish to compare the difference in the expectations of two *dependent* random variables X and Y, based on paired samples $\{X_i\}$ and $\{Y_i\}$. To this end, we use the difference random variable $D = X - Y$, and we compare the expected difference $\delta = \mu_X - \mu_Y$ of D with the reference value 0. We are thus back to the case of a **one-sample *t*-test** if we assume a normal model for the difference $D_i = X_i - Y_i \sim \mathcal{N}(\mu_X - \mu_Y, \sigma^2)$. 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:

$$T = \frac{\bar{D}}{S/\sqrt{n}} \sim \mathrm{t}_{n-1},$$

with $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$.

■ **Example 7.8 (Paired Lab Data (Continued))** To use `t.test` on the pair lab data, we need to set the parameter `paired=TRUE`:

```
> t.test(lab1,lab2,paired=TRUE)

Paired t-test

data:  lab1 and 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
```

Since the P-value is rather high (0.1) there is not enough evidence to conclude that the two labs give different results on average. ■

7.6 Two-sample Test for Proportions

In this section we consider two-sample binomial data and construct a test to compare the two proportions.

■ **Example 7.9 (Are ABC Viewers More Left-wing?)** 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 $n_1 = 70$, and let Y be the number of left-wing "commercial" viewers out of $n_2 = 100$. We assume that X and Y are independent, with $X \sim \mathsf{Bin}(n_1, p_X)$ and $Y \sim \mathsf{Bin}(n_2, p_Y)$, 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 n_1 and n_2 are fairly large here, we proceed by using the central limit theorem
☞ 133 (CLT), analogously to Sections 6.3.4 and 6.3.5. Let $\widehat{P}_X := X/n_1$ and $\widehat{P}_Y := Y/n_2$ be the empirical proportions. By the CLT $\widehat{p}_X$ has approximately a $\mathcal{N}(p_X, p_X(1-p_X)/n_1)$ distribution, and $\widehat{p}_Y$ has approximately a $\mathcal{N}(p_Y, p_Y(1-p_Y)/n_2)$ distribution. It follows that

$$\frac{\widehat{P}_X - \widehat{P}_Y}{\sqrt{\frac{p_X(1-p_X)}{n_1} + \frac{p_Y(1-p_Y)}{n_2}}}$$

has approximately a $\mathcal{N}(0,1)$ distribution. Now, under H_0, $p_X = p_Y = p$, say, and hence under H_0

$$\frac{\widehat{P}_X - \widehat{P}_Y}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

has approximately a $\mathcal{N}(0,1)$ distribution. As we don't know what p is, we need to estimate it. If H_0 is true, then $X + Y \sim \mathsf{Bin}(n_1 + n_2, p)$, and thus p can be estimated by the *pooled* success proportion

$$\widehat{P} := \frac{X+Y}{n_1+n_2}\,. \tag{7.4}$$

Concluding, we take as our test statistic:

$$Z = \frac{\widehat{P}_X - \widehat{P}_Y}{\sqrt{\widehat{P}(1-\widehat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{7.5}$$

which under H_0 has approximately a $\mathcal{N}(0,1)$ distribution. ■

Our general formulation for the **two-sample binomial test** (also called the **test for proportions**) is as follows. First, the model is:

- $X \sim \mathsf{Bin}(n_1, p_X)$, where p_X is unknown.
- $Y \sim \mathsf{Bin}(n_2, p_Y)$, where p_Y is unknown.
- X and Y 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 (7.5). We accept/reject H_0 on the basis of the P-value.

■ **Example 7.10 (Are ABC Viewers More Left-wing? (Continued))** Returning to Example 7.9, from our data we have the estimates $\widehat{p}_X = \frac{40}{70}$, $\widehat{p}_Y = \frac{50}{100}$, 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.

As for one-sample test for proportions, we can also use the `R` function **`prop.test`** to compare 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, this function uses a test statistic which is the square of the one we used ($0.9183^2 = 0.8433$). This function also provides the sample proportions $\widehat{p}_X$ and $\widehat{p}_Y$. ■

7.7 Exercises

1. One of the statements in a research article is that the amount of caffeine in regular cola is "19 mg per 6-oz serving". Suppose we determine the caffeine content 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"?

   (a) What is the model for the data?

   (b) What are the null and alternative hypotheses?

   (c) What is the test statistic random variable?

   (d) What is the probability distribution of the test statistic under the null hypothesis?

   (e) What is the outcome of the test statistic?

   (f) This is a left-one-sided/right-one-sided/two-sided test. What is the P-value?

   (g) Do we accept or reject the null hypothesis? Why?

   (h) Provide a conclusion.

2. The British *Brexit* referendum resulted in an overall vote to leave the European Union (EU), as opposed to remaining an EU member, by 51.9% to 48.1%, respectively. Last month a poll was conducted with a sample size of 500 that showed 54% of the people in this poll would prefer to remain in the EU, while 46% would prefer to leave the EU. Does this poll suggest that the people of Britain have changed their mind?

   (a) Formulate the null and alternative hypotheses.

   (b) Conduct a statistical test for this problem, specifying the test statistic, its (approximate) distribution under the null hypothesis, the outcome of the test statistic, the P-value of the test, and your conclusion.

   (c) Give an (approximate) 95% confidence interval for the proportion of British people in favour of leaving the EU at the time of the poll.

3. The following data was drawn from a Bernoulli distribution with success probability p.
   1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1,

   (a) Estimate p via the sample proportion $\widehat{P} = X/n$, and give an approximate 95% confidence interval for p.

(b) Test $H_0 : p = 1/2$ versus $H_1 : p > 1/2$, using the total number of successes (ones), X, as the test statistic.

(c) Test $H_0 : p = 1/2$ versus $H_1 : p > 1/2$, using the test statistic $(\widehat{P} - 0.5)/\sqrt{0.5 \times 0.5/20}$.

4. A company has just set up a manufacturing line for steel rods, used as concrete reinforcing for a new construction. Each steel rod must be at least 6 cm in diameter to meet the client's specifications, but it is of no concern if they are slightly thicker. From a sample of 100 steel rods, the mean thickness was 6.1 cm and the standard deviation was 0.1 cm.

   (a) Estimate the probability that a given steel rod is thinner than 6 cm.

   (b) Assuming the standard deviation in thickness holds constant, what should the mean thickness be to ensure that 99% of rods are at least 6 cm in diameter?

   (c) Returning to the original question, carry out an appropriate statistical test, specifying the null and alternate hypotheses, the test statistic and its approximate distribution under the null hypothesis, the outcome of the test statistic, the P-value, and your conclusion.

5. A company is testing a new card shuffling machine. Taking 20 new packs of cards, exactly 17 had a red card on top after shuffling. Does this suggest that the card shuffling machine is not carrying out a perfect random shuffle? Carry out an appropriate statistical test, specifying the null and alternate hypotheses, the test statistic and its approximate distribution under the null hypothesis, the outcome of the test statistic, the P-value, and your conclusion.

6. A toy chain store claims that at least 80% of boys under 8 years old prefer Lego over other types of toys. But we feel that this claim is inflated. To investigate this, we observed the toy preference of 20 randomly selected boys under 8 years old.

   Let X be the number of boys who chose Lego. We wish to test the hypothesis $H_0 : p = 0.8$ against $H_1 : p < 0.8$. Suppose we decide to reject H_0 if $X \leqslant 12$.

   (a) Find the probability of a Type I error α.

   (b) Find the probability of a Type II error β for $p = 0.6$.

   (c) Find β for $p = 0.4$.

7. A certain medicine is supposed to have 5 mg of active ingredient, and a pharmacist decides to test this for both the brand name and the generic version. From 10 measurements on the brand name version, the mean amount of active ingredient was 4.7 mg with a sample standard deviation of 0.2 mg. From 10 measurements

on the generic version, the mean amount of active ingredient was 4.6 mg with a sample standard deviation of 0.4 mg.

(a) Construct 95% confidence intervals for each version of the medicine. Is there any reason to suspect that the mean active ingredient content differs from 5 mg for any version?

(b) The pharmacist only has these two versions available, and wants to compare the mean level of active ingredient. Carry out an appropriate statistical test, specifying the null and alternate hypotheses, the test statistic and its approximate distribution under the null hypothesis, the outcome of the test statistic, the P-value, and your conclusion.

8. Several species of bird show hardly any *sexual dimorphism*; that is, female and male specimens look very similar. One such species is *Fischer's lovebird* — a species of small parrot that is native to part of Africa. To see if there is any difference in the mean height between adult male and female lovebirds, researchers measured the heights of 100 male and 150 female lovebirds (randomly selected), giving a sample mean of 15.33 cm with a sample standard deviation of 2.1 cm for the male birds and a sample mean of 14.95 cm with a sample standard deviation of 1.9 cm for the female birds.

(a) Assuming that the group standard deviations may be different, give an approximate 90% confidence interval for the difference in mean height between male and female lovebirds.

(b) We assume that the standard deviations in the two groups are the same. Show that the pooled standard deviation is (to 2 decimal places) equal to 1.98 cm.

(c) Using the pooled standard deviation conduct a two-sample t-test to test if there is enough statistical evidence at an $\alpha = 0.05$ level of significance that male lovebirds are on average taller than female ones.

9. Bill wonders if are males more likely to be left-handed than females. He asks his STAT1301 class to fill out a survey, the results of which are summarised in the table below. He decides to do a proportion test to test his hypothesis.

	Left-Handed	Right-Handed
Male	20	133
Female	16	138

(a) State the null and alternative hypotheses in terms of the population parameters.

(b) Conduct the two-sample proportion test. What is the P-value? What can Bill conclude?

(c) Construct an (approximate) 99% confidence interval for the difference in proportions.

10. Joanna has been enthusiastically following a certain university's cricket team, and wants to test whether the "home ground advantage" is present in data. Of 100 games played, 25 were played on home ground, with 20 wins for Joanna's team. Of the other 75 games played away, 55 were wins for Joanna's team. Joanna models the outcome of each game as an independent Bernoulli trial, with probability that only depends on whether the game is a 'home' or 'away' game.

    Carry out an appropriate statistical test, specifying the null and alternate hypotheses, the test statistic and its approximate distribution under the null hypothesis, the outcome of the test statistic, the P-value, and your conclusion.

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 one factor and two 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** (in some fields confusingly called "dependent variable") and the other variables are called **explanatory variables** (or "independent variables"). Recall (see Section 5.3) that categorical variables take values in a *finite* number of categories, such as yes/no, green/blue/brown, and male/female. In R, such variables are called **factors**. They often arise in designed 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). Treatment pairs were assigned to plots via randomization. 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.

☞ 84

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 two-
☞ 148 sample t-test; see Section 7.5. Let us carry out the usual steps for a statistical test here:

1. The model is a two-sample normal model. Let $X_1, \ldots, X_9 \sim_{\text{iid}} \mathcal{N}(\mu_1, \sigma^2)$ be the crop yields without pesticide and $Y_1, \ldots, Y_9 \sim_{\text{iid}} \mathcal{N}(\mu_2, \sigma^2)$ be the crop yields with pesticide; all variables are assumed to be independent of each other. Note that we assumed here equal variances for both groups; you may verify graphically that this assumption is reasonable.

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: $\mu_1 \neq \mu_2$.

☞ 150 3. As a test statistic we use the T statistic given in (7.3).

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 two-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 three-sample normal model. Let $Y_1, Y_2, Y_3, Y_{10}, Y_{11}, Y_{12} \sim_{\text{iid}} \mathcal{N}(\mu_1, \sigma^2)$ be the crop yields with low fertilizer, $Y_4, Y_5, Y_6, Y_{13}, Y_{14}, Y_{15} \sim_{\text{iid}} \mathcal{N}(\mu_2, \sigma^2)$ be the crop yields with medium fertilizer, and $Y_7, Y_8, Y_9, Y_{16}, Y_{17}, Y_{18} \sim_{\text{iid}} \mathcal{N}(\mu_3, \sigma^2)$ 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.

The following code reads the data and produces Figure 8.1. We have used a few tricks in this code that you might find useful to know. Firstly, we plotted the levels in the order from "Low" to "High". This is done in Line 5. Without this line, the levels would be taken in alphabetical order, starting with "High". Secondly, we indicated what the Pesticide level was for the data in each Fertilizer group. This is done by specifying the plotting characters (numbers) in Line 6, for each data point, and in Line 7, we use these characters via the " pch = " option. Note that the **rep** function replicates numbers or strings; in this case nine 4s (producing crosses) and nine 1s (producing circles).

```
1 crop = read.csv("cropyield.csv")
2 library(lattice)
3 # reorder the levels from low to high
4 crop$Fertilizer = factor(crop$Fertilizer,
5                          levels = c("Low", "Medium", "High"))
6 chs = c(rep(4,9),rep(1,9))  # define two groups of plotting characters
7 stripplot(Yield~Fertilizer,pch=chs,cex=1.5,data=crop,xlab="Fertilizer")
8 #stripplot(Yield~Fertilizer,groups=Pesticide,cex=1.5,data=crop,
9 #          xlab="Fertilizer")
```

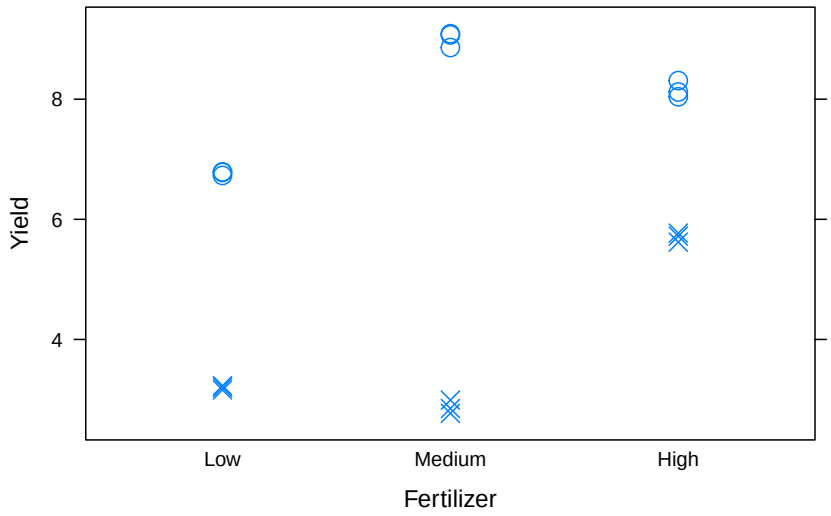


Figure 8.1: Strip plot of crop yield against fertilizer level. Whether pesticide was applied is also indicated (circle for Yes, cross for No).

8.2 Single-Factor (one-way) ANOVA

We start with single-factor experiments. Think of the crop yield example where we only consider the fertilizer factor, which is applied at 3 levels (low, medium, high).

8.2.1 Model

Consider a response variable which depends on a single factor with d levels, denoted $1, \ldots, d$. Let us use the letter i to indicate a level. So, $i \in \{1, \ldots, d\}$. Within each level i there are n_i independent measurements of the response variable. Let us use the letter k to indicate the independent replications in each level, so $k \in \{1, \ldots, n_i\}$. The total number of measurements is thus $n = n_1 + \cdots + n_d$. An obvious model for the data is that the $\{Y_{ik}\}$ 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
☞ 68 generalization of the two-sample normal model in Example 4.3. 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. Using this notation, we can write the one-way ANOVA model as follows.

Definition 8.1: Single-Factor ANOVA Model

Let Y_{ik} denote the kth measurement of the ith level in the response data. Then

$$Y_{ik} = \mu_i + \varepsilon_{ik}\ , \quad k = 1, \ldots, n_i,\ i = 1, \ldots, d, \tag{8.1}$$

where $\varepsilon_{11}, \ldots, \varepsilon_{1n_d} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, and μ_i is the mean effect of level i.

So, for a single response Y with explanatory factor x, we have

$$Y_{ik} = \mu_i + \varepsilon_{ik}\ . \tag{8.2}$$

Instead of (8.2), one often sees the "factor effects" formulation

$$Y_{ik} = \mu_1 + \alpha_i + \varepsilon_{ik}, \tag{8.3}$$

where μ_1 is the mean effect of the *reference* level (level 1 in this case) and $\alpha_i = \mu_i - \mu_1$ is the *incremental effect* of level i, relative to the reference level. The latter approach is used in R.

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:

$$\widehat{\mu_i} = \overline{Y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}, \quad i = 1, \ldots, d\ .$$

To estimate σ^2, we should utilize the fact that all $\{Y_{ik}\}$ are assumed to have the same variance σ^2. So, as in the two-sample normal model case, we should *pool* our data and not just calculate, say, the sample variance of the first level only. The model (8.1) assumes that the errors $\{\varepsilon_{ik}\}$ are independent and normally distributed, with a constant variance σ^2. If we knew each $\varepsilon_{ik} = Y_{ik} - \mu_i$, we could just calculate the variance of these observations to estimate σ^2. Unfortunately, we do not know the $\{\mu_i\}$. However, we can estimate each μ_i with $\overline{Y}_i$. This suggests that, we replace the unknown true errors ε_{ik} with the **residual errors** (or simply residuals), which is the difference between the observed value and the predicted value of the response, and are here given by $e_{ik} = Y_{ik} - \overline{Y}_i$. In the process of determining the e_{ik} values, we needed to estimate d group means. This leads to the loss of d degrees of freedom, leaving $n - d$ independent observations. We thus obtain the unbiased estimator

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{d} \sum_{k=1}^{n_i} e_{ik}^2}{n-d} = \frac{\sum_{i=1}^{d} \sum_{k=1}^{n_i} \left(Y_{ik} - \overline{Y}_i\right)^2}{n-d} = \frac{\text{SSE}}{n-d} = \text{MSE}. \tag{8.4}$$

This quantity is called the **mean squared residual error** (MSE).

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 compare here two models: Under H_0 we simply have the standard 1-sample normal model for data, and under H_1 we have the single-factor ANOVA model. 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 of the error terms 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 $\sigma_1^2 = \sigma_2^2$ if H_0 is true. It therefore makes sense to base the test statistic on estimators of σ_1^2 and σ_2^2. We already saw that σ_2^2 is estimated via $\mathrm{SSE}/(n-d)$. And we can estimate σ_1^2 simply via the sample variance

$$\frac{\sum_{i=1}^{d}\sum_{k=1}^{n_i}(Y_{ik}-\overline{Y})^2}{n-1} = \frac{\mathrm{SST}}{n-1}, \tag{8.5}$$

where $\overline{Y}$ denotes the sample mean of all $\{Y_{ik}\}$, and $\mathrm{SST} = \sum_{i=1}^{d}\sum_{k=1}^{n_i}(Y_{ik}-\overline{Y})^2$ is called the **total sum of squares**.

So, a sensible test statistic could be based on a 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

$$F = \frac{(\mathrm{SST}-\mathrm{SSE})/(d-1)}{\mathrm{SSE}/(n-d)}, \tag{8.6}$$

where the difference SST − SSE is again a "sum of squares":

$$\mathrm{SST}-\mathrm{SSE} = \sum_{i=1}^{d}\sum_{k=1}^{n_i}(\overline{Y}_i-\overline{Y})^2 \,. \tag{8.7}$$

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 $\mathrm{SSF}/(d-1)$ to MSF (mean square factor) and $\mathrm{SSE}/(n-d)$ to MSE (mean square error), then we can write our test statistic as

$$F = \frac{\mathrm{MSF}}{\mathrm{MSE}} \,.$$

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 = MSF/MSE has an $\mathsf{F}(d-1, n-d)$ distribution.

This **F-distribution** is named after R.A. Fisher — one of the founders of modern statistics. 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 $\mathsf{F}(\mathrm{df}_1, \mathrm{df}_2)$ for an F distribution with degrees of freedom df_1 and df_2. Figure 8.2 gives a plot of various pdfs of this family. We used a similar script as for the plotting of Figure 6.2. Here is the beginning of the script — you can work out the rest. ☞ 127

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

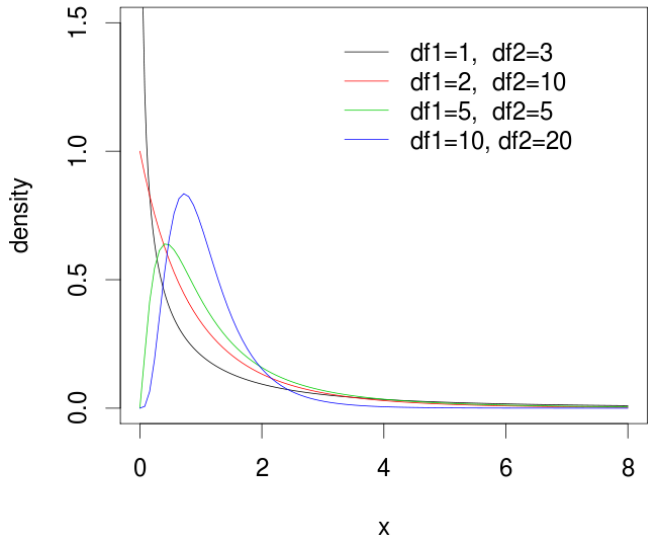


Figure 8.2: The pdfs of F distributions with various degrees of freedom (df).

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 2.1. ☞ 52

Fortunately, software can do all the calculations for us and summarize the results in an **ANOVA table**. For the one-factor case, it is of the form given 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			

8.2.4 Worked Example

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) for the cold sore to completely heal 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 healing 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 5) 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`.

```
> coldsore = stack(x)
> names(coldsore)  = c("times", "treatment")
```

The second important point is that both columns in the reformated data frame **coldsore** now have the correct type (check with `str(coldsore)`): the response is a quantitative 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 within each of the factor levels. In R this can be done conveniently via the function **tapply**, which applies a function to a table.

```
> tapply(coldsore$times,coldsore$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 $\bar{y}_i, i = 1, \ldots, 5$) are given in the fourth column.

A boxplot of `times` versus `treatment` gives more information:

```
> library(lattice)
> bwplot(times~treatment, data = coldsore, xlab="treatment")
```

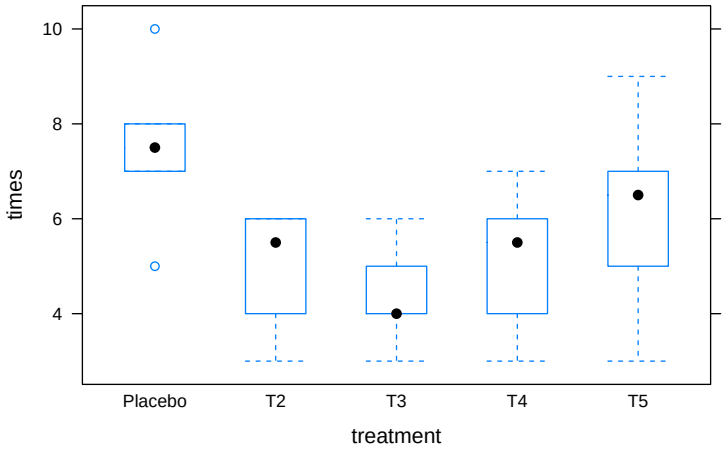


Figure 8.3: Box plot of healing times for each treatment.

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 $F = \text{MSF}/\text{MSE}$, which, if H_0 is true, we know has an F distribution; see Theorem 8.2.3. 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 functions **anova** and **lm**. All we need to do is specify the R formula.

```
>coldsore.lm = lm(times~treatment, data = coldsore)
>anova(coldsore.lm)
Analysis of Variance Table

Response: times
          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 MSF = 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.

Validation of the Assumptions

The ANOVA model (8.1) assumes that the errors $\{\varepsilon_{ik}\}$ are independent and normally distributed with a constant variance. Independence is difficult to check. In the context of an experiment, independence needs to be ensured by using an appropriate experimental design (e.g., via randomization). We can verify the other model assumptions by investigating the residuals. If the model is correct, the residuals should behave as independent random variables from a $\mathcal{N}(0, \sigma^2)$ distribution, for some σ^2.

The assumptions of the model can be inspected graphically using the following commands.

```
> par(mforw=c(1,2))
> plot(coldsore.lm, 1:2)
```

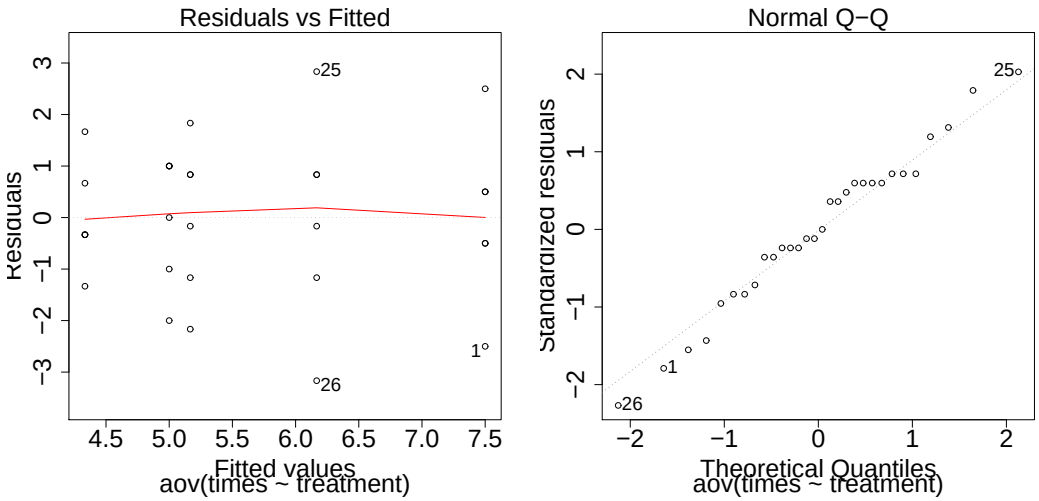


Figure 8.4: Analysing the residuals in single-factor ANOVA

R actually returns four diagnostic plots, but we have listed only two in Figure 8.4. 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., zero mean and constant variance) seem valid. The normality of the residuals is indicated by the observed straight line in the Q-Q plot. We will discuss more about assumptions checking when we study regression in the Chapter 9.

☞ 195

8.3 Two-factor (two-way) 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) there are n_{ij} replications, so that the total number of observations is $n = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} n_{ij}$. A direct generalization of (8.2) gives the following model.

Definition 8.2: Two-Factor ANOVA Model

Let Y_{ijk} denotes kth measurement of the (i, j)th level in the response data. 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 $\varepsilon_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

This is just saying that the response variables are independent of each other and that for each combination of Factor 1 level and Factor 2 level (i, j), the corresponding response Y has a $\mathcal{N}(\mu_{ij}, \sigma^2)$ distribution. The model thus has $d_1 d_2 + 1$ parameters. Note that the variances of the responses are all assumed to be the same (equal to σ^2).

To obtain a "factor effects" representation, we can reparameterize the model for a single response Y with explanatory pair (i, j) as follows:

$$\begin{aligned} Y_{ijk} \quad &= \quad \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \\ &\qquad i = 1, \ldots, d_1, \quad j = 1, \ldots, d_2, \quad k = 1, 2, \ldots, n_{ij}. \end{aligned} \tag{8.9}$$

The parameter μ is the overall (grand) mean response. For every explanatory pair (i, j), we *add* to this overall mean response:

- an incremental effect α_i due to Factor 1,
- an incremental effect β_j due to Factor 2,
- an interaction effect γ_{ij} due to both factors.

Notice that there are now $1 + d_1 + d_2 + d_1 d_2 + 1$ parameters (corresponding to μ, the α_i's, the β_j's, the γ_{ij}'s, and σ^2 respectively), more than that in the original parameterization in (8.8). This means that the model (8.9) is *overparameterized* and the model parameters cannot be estimated uniquely. One solution is to impose $d_1 + d_2 + 1$ constraints on the parameters. Similar to (8.3) in one-way ANOVA, we can set the first level of Factor 1 and the first level of Factor 2 as the "reference" level. This implies $\alpha_1 = 0$, $\beta_1 = 0$, $\gamma_{i1} = 0$ (for $i = 1, \ldots, d_1$) and $\gamma_{1j} = 0$ (for $j = 1, \ldots, d_2$). Notice that this also implies $\mu = \mu_{1,1}$. This is the approach used in R .

The advantage of the formulation (8.9) is that we can consider "nested" models by setting some parameters to zero. For example, if no interaction terms are included, we get the model

$$Y_{ijk} = \mu_{1,1} + \alpha_i + \beta_j + \varepsilon_{ijk}\,. \tag{8.10}$$

The assumption that there is no interaction and Factor 2 has no effect leads to the model:

$$Y_{ijk} = \mu_{1,1} + \alpha_i + \varepsilon_{ijk} \ , \tag{8.11}$$

which is a 1-factor ANOVA model. The simplest model is the default normal model, where neither of the factors has an effect:

$$Y_{ijk} = \mu_{1,1} + \varepsilon_{ijk} \ . \tag{8.12}$$

Which of these models is most appropriate can be investigated via statistical tests.

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 estimated in a similar way. The reference mean is estimated via $\overline{Y}_{1,1}$, as given above. The incremental effect α_i can be estimated via $\overline{Y}_{i\bullet} - \overline{Y}_{1,1}$, 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 $\overline{Y}_{\bullet j} - \overline{Y}_{1,1}$, where $\overline{Y}_{\bullet j}$ is the average of all the $\{Y_{ijk}\}$ within level j of Factor 2. Finally, γ_{ij} is estimated by taking the average of all responses at the level pair (i, j) and subtracting from this the estimates for $\mu_{1,1}$, α_i and β_j.

To estimate σ^2 we can reason similarly to the 1-factor case and consider the residuals $e_{ijk} = Y_{ijk} - \widehat{Y}_{ijk}$ as our best guess of the true model errors, where $\widehat{Y}_{ijk}$ is the fitted value to the ijk-th response. Similar to (8.5) we have the unbiased estimator ☞ 164

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n - d_1\, d_2} = \frac{\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{n_{i,j}} e_{ijk}^2}{n - d_1\, d_2} \ .$$

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.

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 $\text{SST} = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \overline{Y}_{ij})^2$ 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 residual variability (i.e., 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. The various quantities of interest in an ANOVA table are summarized in Table 8.4.

Table 8.4: 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	$d_2 - 1$	SSF2	MSF2	$\frac{\text{MSF2}}{\text{MSE}}$	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			

8.3.4 Worked Example

Consider the data in Table 8.5, representing the crop yield using four different crop treatments (e.g., strengths of fertilizer) on four different regions.

Table 8.5: 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. The code shows also a few "tricks of the trade". The **attach** function makes the variables `region, fertilizer, yield` available without having to use the $ construction, such as in `yield$region`. The function **paste** concatenates (joins) strings, after converting numbers into strings. So, we can get the string "Region 1", for example. The function **gl** generates factors by specifying the pattern of their levels.

Alternatively, you could of course enter the data in a CSV file with appropriate headers, and read the file into a data frame with **read.csv**.

```
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)
```

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
```

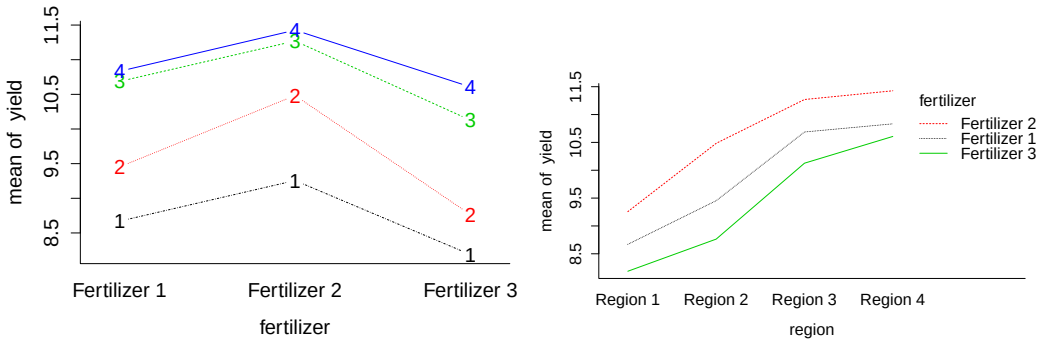


Figure 8.5: Exploration of interaction in two-way ANOVA.

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 (8.5). 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 8.5 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.

We plotted the two interaction plots in two different ways. To find out about the possible plotting parameters, type: `?interaction.plot` and `?par`.

ANOVA Table

Similar to the 1-factor ANOVA case, the `R` functions **`anova`** and **`lm`** provides the ANOVA table:

```
> yield.lm = lm(yield~region*fertilizer)
> anova(yield.lm)

Analysis of Variance Table

Response: 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
```

The formula `region*fertilizer`, used in `lm`, corresponds in fact to the formula `region+fertilizer+region:fertilizer`; i.e., the factor `region`, the factor `fertilizer`, and the interaction between 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 given in (8.10): ☞ 170

$$Y_{ijk} = \mu_{1,1} + \alpha_i + \beta_j + \varepsilon_{ijk} \, .$$

In R, we specify this by the model `yield ~ region+fertilizer`, as in:

```
> yield.lm.no_interaction = lm(yield~region+fertilizer)
> lm(yield.lm.no_interaction)


Analysis of Variance Table

Response: yield
           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.

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: `anova(lm(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.

```
anova(lm(yield~fertilizer,subset=region=="Region 1"))

Analysis of Variance Table

Response: yield
           Df Sum Sq Mean Sq F value Pr(>F)
fertilizer  2 1.7227 0.86134  1.8746 0.2331
Residuals   6 2.7569 0.45949
```

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.

Validation of Assumptions

As in one-way ANOVA, we validate the model with a study of the residuals of the underlying linear model.

```
> plot(yield.lm.no_interaction, 1:2)
```

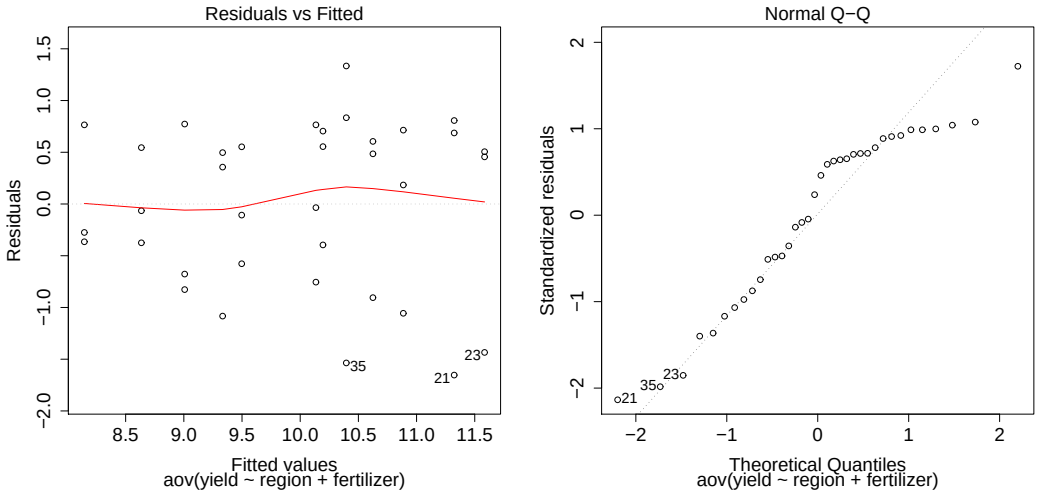


Figure 8.6: Residual analysis in two-way ANOVA.

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.

Unbalanced data

The crop yield data in this example is a **balanced data**, meaning that there are equal number of observations for each level of a factor. If the data is unbalanced, there are at least three ways to calculate the sum of squares (SS) for ANOVA, known as **type I, II, and III sums of squares**. In brief, type I (also called *"sequential"*) SS involves testing the factors sequentially in the order in which they are listed in the model. Type II (also called *"partial"* or *"hierarchical"*) SS tests for each of the factors in light of the other factor and without an interaction term. Type III (also called *"unique"*) SS is similar to Type II SS but the factors are also tested in light of the interaction term.

In-depth details of these are beyond the scope of this course, but you should be aware that they may lead to different results for unbalanced data. When the data is balanced, the three types of SS will give the same results. By default, the R function `anova` uses type I SS. ANOVA with types II and III SS can be accessed through the `Anova` function from the `car` package.

8.4 Randomization and Blocking

Recall that, in the Alice experiment (Section 4.3), Alice divided her 20 friends in two groups of 10 using *randomization*. One group was given the treatment of caffeinated diet cola, and the other was the control group who received decaf diet cola. Designed experiments such as Alice's often involve some form of randomization to alleviate the effect of **nuisance** factors. These are factors that are not of primary interest to the researcher, but may influence the response. An example of a nuisance factor for crop data could be the *plant location* of the crop. ☞ 71

As an example, consider crop yield data such as in Table 8.5. Suppose we wish to test the effectiveness of 3 treatments on the crop yield. We could plant the crop in a test plot with subplots arranged in rows and columns; for example, 4 rows and 12 columns in the figures below. The question is now how to assign treatments to the test plots. It seems reasonable to allocate each treatment 16 times. In a **completely randomized design**, we allocate the treatments in such a way that the each 3×16 allocation is equally likely. This can be done in $\mathbb{R}$ as follows. We first simulate a random permutation of the numbers 1,..., 48:

```
> x = sample(c(1:48),48)
> x

 [1] 14 38 19 40 42  2 23 37 44 18 41 17 25 21  4 30  8 43
[19] 10 28 36 46 47 29 16 26 12 13  6  3 39 24 22 45  1  7
[37] 33 27 34 11 31  9 35 32 48 15 20  5
```

Then we assign treatment 1 to the first 16 subplots, treatment 2 to the next 16, and treatment 3 to the last 16. If we colour the subplots red, green, and yellow, this gives the left panel in Figure 8.7

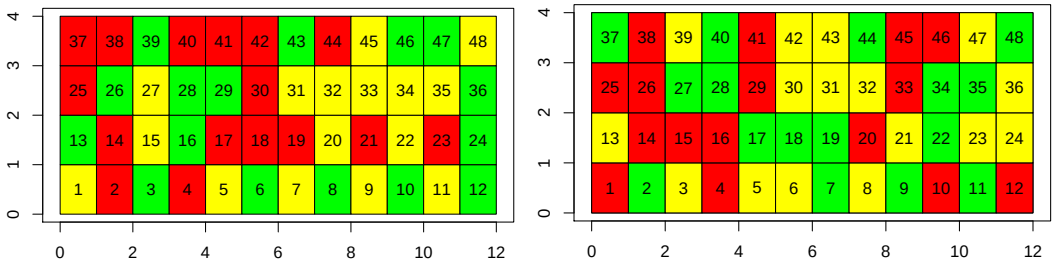


Figure 8.7: Left: completely randomized design. Right: randomized block design, with 4 treatments per block (row).

Now suppose that the soil conditions vary a lot within each column; for example the bottom row could lie at the bottom of a hill and the top row on the top of a hill. Then the soil condition of the row in which the crop was plotted could be an important factor (but a nuisance factor) in explaining the crop yield. Complete randomization as described above would alleviate the bias caused by the row soil conditions. However, note that in the left panel of Figure 8.7 rows 1 and 3 only have two treatments of type 1 (red). If the rows indeed are a factor, it would be better (less variability in the data) if we chose our design to *block* the treatments such that each block (i.e., row) has the same number of treatments. Of course we still should randomize within each block. The right panel of Figure 8.7 shows such a **randomized block design**. We made the design and figure with the following code.

```
colv = c(rep("white",48))  # a vector of colours
for (i in 1:4) {           # for each row
  x = sample(c(1:12), 12)  # random permutation 1,...,12
  p1 = x[1:4]              # column indices for treatment 1
  p2 = x[5:8]
  p3 = x[9:12]
  colv[(i-1)*12 + p1] = "red"   # assign colours to indices in row i
  colv[(i-1)*12 + p2]= "green"
  colv[(i-1)*12 + p3] = "yellow"
}
# plot the coloured rectangles
for (j in 0:3){
  for (i in 0:11 ){
    col = colv[1 + j*12 + i]   # colour of the rectangle
    rect(xleft = i, ybottom = j, xright = i+1, ytop = j+1, col = col)
    text(i+.5, j+.5, labels = 1+j*12 +i)  # numbers of the rectangles
  }
}
```

8.5 Multiple Comparisons

In the cold sore example in Section 8.2.4, we concluded that there is fairly strong evidence to believe that the treatments have an effect. But which treatment(s) has a significant effect?

When we reject the null hypothesis in an ANOVA, it only showed that the differences between the treatment means are significantly different. But it does not tell us which of the treatment means are significantly different from each other. We can follow up with post hoc tests to do pairwise comparisons between the treatments. This would involve performing multiple tests on different combinations of treatments. However, if we repeatedly do (for example) t-tests for every possible pairwise combination of treatments, the probability of making a type I error will increase considerably.

One approach is to use multiple comparison techniques. There are many such techniques, but we will focus on the Bonferroni procedure. This is a simple and commonly used method, but is very conservative for controlling type I error. If we want to carry out m comparisons, then we use α/m in place of α as the significance level for each test. The logic behind this is quite intuitive: if we have k tests, each with a type I error rate of at most α/m, then the *total* type I error rate cannot exceed α.

Alternatively, this is equivalent to multiplying the P-value of the tests by m, with a maximum of 1. This is the approach is used in R. The Bonferroni-adjusted P-values for ANOVA can be obtained using the `pairwise.t.test` function with the setting `p.adjust.method = "bonferroni"`. For the cold sore data, we see that only T3 has a significant effect.

```
pairwise.t.test(coldsore$times, coldsore$treatment, p.adj = "bonferroni")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  coldsore$times and coldsore$treatment

   Placebo T2    T3    T4
T2 0.090   -     -     -
T3 0.014   1.000 -     -
T4 0.140   1.000 1.000 -
T5 1.000   1.000 0.483 1.000

P value adjustment method: bonferroni
```

8.6 Exercises

1. The following table shows measurements of the haemoglobin (Hb) levels for patients with different types of sickle-cell disease: HbSS, HbST, and HbSC.

HbSS	HbST	HbSC
7.2	8.1	10.7
7.7	9.2	11.3
8.0	10.0	11.5
8.1	10.4	11.6
8.3	10.6	11.7
8.4	10.9	11.8
8.4	11.1	12.0
8.5	11.9	12.1
8.6	12.0	12.3
8.7	12.1	12.6
9.1		12.6
9.1		13.3
9.1		13.3
9.8		13.8
10.1		13.9
10.3		

(a) We wish to carry out 1-factor ANOVA test to asses whether the haemoglobin levels differ between the three types. How would you enter the data into a spreadsheet, in order to apply statistical software such as R to carry out the ANOVA?

(b) State the null and alternate hypotheses.

(c) Software produced the ANOVA table below. Complete the missing values.

	DF	SS	MS	F
HbType		99.89		
Error			1.00	
Total				

(d) What assumptions underlie the F-test?

(e) What is the P-value for this F-test? What can the researcher conclude?

2. Researchers were interested in whether risk-taking behaviours such as drinking, gambling and illicit drug use were related to rates of unintentional injuries. Using 180 subjects who were being treated for addiction, the study recorded the type of addiction (alcohol, cocaine, tobacco or gambling) along with the sex and age (years) of each subject. A questionnaire then asked whether the subject had an unintentional injury in the last year while also obtaining measures of stress using a standard scale.

(a) An initial interest for the researchers was whether subjects with different addiction type tended to have different levels of stress. Complete the seven

blank boxes in the following ANOVA table for testing for a difference in mean level of stress between the four addiction groups.

	DF	SS	MS	F
Addiction		55		
Error				
Total		6510		

(b) What is the pooled estimate for the standard deviation of stress levels based on these four groups?

(c) Give the P-value and F statistic in the ANOVA table. What do you conclude?

3. Researchers were interested in factors that affect the growth of *Gracilaria parvispora*, an edible seaweed of great economic value for which several commercial culture systems have been developed. Photosynthesis is an effective indicator of plant growth and maximum photosynthetic rate in particular is frequently used to identify preferable conditions for plant development. In one experiment, the researchers measured the maximum photosynthetic rate ($mg^{-1}\, O_2\, g\, dw^{-1}\, h^{-1}$) at temperatures of 20, 25, 30, and 35°C with six replicates at each temperature. The observations are shown in the following figure.

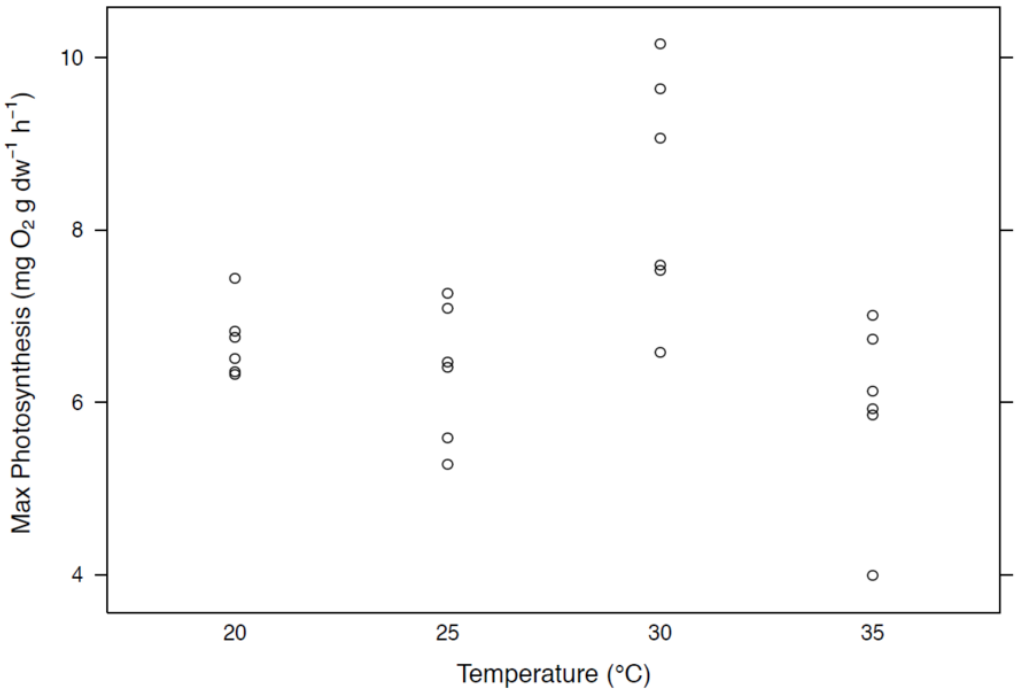


Figure 8.8: Maximum photosynthetic rate of *Gracilaria parvispora* at various temperatures.

A one-way analysis of variance in R produced the following partial output:

```
               Df    Sum Sq    Mean Sq    F value    Pr(>F)
Temperature          21.55
Residuals            19.42
```

(a) Briefly discuss the plot, describing the shape of the distribution for each temperature group and comparing their relative locations.

(b) Clearly state the null and alternative hypotheses for the analysis of variance.

(c) Calculate the F statistic and the P-value. What do you conclude?

(d) What is the R^2 value for this model?

(e) Which assumption for analysis of variance can be checked roughly from the figure above? Do you think that assumption is reasonable for this data?

(f) What is the pooled standard deviation?

(g) What statistical procedure would you use next to determine which means were significantly different?

(h) Temperature is a continuous predictor variable so this data could also have been modelled using linear regression.

i. Briefly explain why the researchers might have treated temperature as categorical rather than using linear regression.

ii. If you did use linear regression to model this data, would the R^2 value be higher or lower than the value in (b)?

4. To see if fertilizer level has an effect on the yield of wheat, three different plots (blocks A, B, and C) were divided into three subplots. Each of the subplots was assigned a Low, Medium, and High level of fertilizer. The crop yield (per 1000 kg) for each combination of Block and fertilizer level is given in Table 8.6.

Table 8.6: Crop yield.

	Fertilizer		
Block	Low	Medium	High
A	0.3702	0.9362	0.8483
B	0.4781	0.7309	1.0855
C	0.8883	1.2357	1.4606

Using R we obtained the following ANOVA table for the data, where we have omitted most of the top row values -.

```
Response: Yield
            Df Sum Sq  Mean Sq  F value  Pr(>F)
Fertilizer -- 0.48338        --       --     --
Block       2 0.41425 0.207125  15.150  0.01360
Residuals   4 0.05469 0.013672
```

(a) The R commands to create the ANOVA table above was

```
crop = read.csv("crop.csv")
anova(lm(Yield ~ Fertilizer + Block, data=crop))
```

Put the data in Table 8.6 in "standard form", as would be the case for the `"crop.csv"` file.

(b) Complete the ANOVA table.

(c) What does the value 0.01360 in the table indicate about the blocks?

(d) Is there statistical evidence that the fertilizer level has an effect on the crop yield?

5. An experiment on opiate withdrawal using rats investigated whether levorphanol reduces stress as reflected in measurements of plasma corticosterone levels. Twenty rats were randomly allocated to four treatment groups. Two of the groups were then given adrenaline to simulate stress. The resulting corticosterone levels (ng/ml) are shown in the following table.

Treatment	n	$\bar{x}$	s
Control	5	225	105
Levorphanol only	5	175	97
Adrenaline only	5	528	49
Levorphanol and Adrenaline	5	257	134

(a) Draw an interaction effects plot in the space below to show the relationship between corticosterone levels and the two factors.

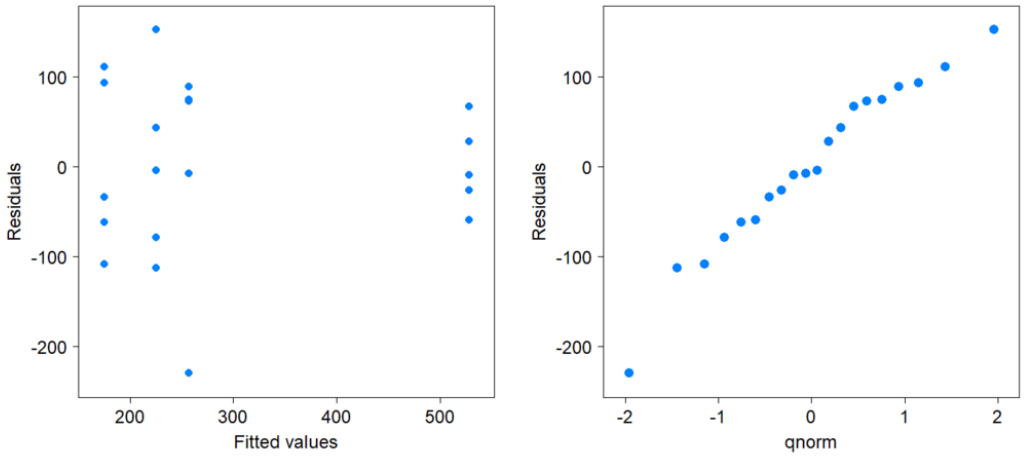
(b) The data, stored in the data frame `Rats`, was analysed in R. Partial output is given below. Complete all the missing items (A)–(H).

```
ModelRats = lm(Corticosterone~Levorphanol*Adrenaline,data=Rats)
summary(aov(ModelRats))
                      Df    Sum Sq Mean Sq F value    Pr(>F)
Levorphanol            1    128801 128801   12.630  0.002645
Adrenaline             1    185281 185281       (E)       (G)
Levorphanol:Adrenaline (A)   61051     (C)       (F)       (H)
Residuals              (B)   163164     (D)
---
```

(c) Based on this analysis, what are you able to conclude?

(d) What is the R^2 value for the model?

(e) Based on the following plots, comment on the validity of the assumptions underlying ANOVA.



6. Return to Exercise 1 of this chapter. Suppose that the researcher concludes that the mean Hb levels are not the same across the disease types. Perform further tests to investigate which pair of disease types are significantly different at the 5% significance level.

CHAPTER 9

REGRESSION

This chapter gives an introduction to simple and multiple linear regression. We present how to get confidence and prediction intervals for 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).

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 seem 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.

Regression analysis is about finding relationships between a *response* variable which we would like to "explain" via one or more *explanatory* variables. In regression, the response variable is usually a *quantitative* (numerical) variable.

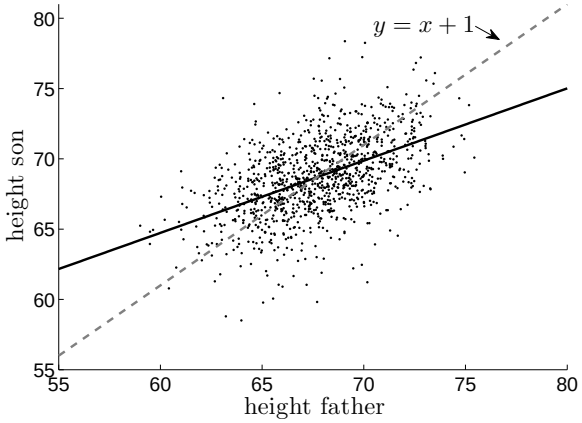


Figure 9.1: A scatter plot of heights from Pearson's data.

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

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 $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

This formulation makes it clear 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 $\{\varepsilon_i\}$). This is illustrated in Figure 9.2.

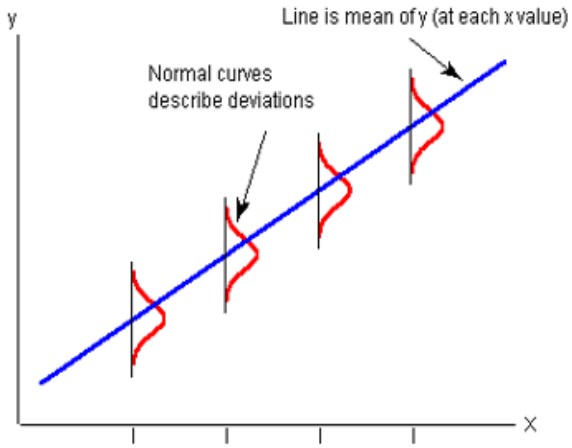


Figure 9.2: Linear regression model.

To make things more concrete let us consider the student survey dataset stored in the dataset `studentsurvey.csv`, which can be found on Blackboard. 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
> survey = read.csv("studentsurvey.csv")
> names(survey) # check the names
```

```
 [1] "sex"         "laptop"     "height"     "weight"
 [5] "pulserate"   "forearm"    "shoe"       "sleep"
 [9] "eyes"        "piercings"  "attractive" "country"
[13] "pizza"       "residence"  "work"       "grade"
[17] "life"        "superpower" "kiss"       "handed"
```

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). Here we use the **`xyplot`** function from the **`lattice`** library:

```
> library(lattice) # load the lattice library
> xyplot(height~shoe, data=survey)
```

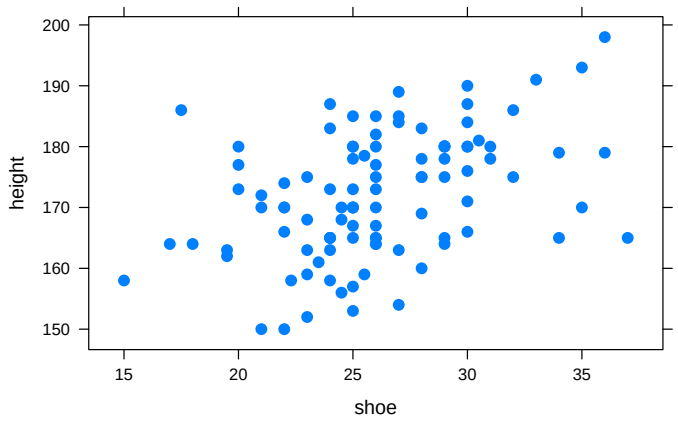


Figure 9.3: Scatter plot of height (in cm) against shoe size (in cm).

We observe a slight increase in the height as the shoe size increases, although this relationship is not very clear.

9.2.1 Estimation for Linear Regression

Obviously, we do not know the true regression line $y = \beta_0 + \beta_1 x$, but we can try to find 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 convenient and principled one is to minimize the Sum of the Squared residual Errors,

$$\text{SSE} = \sum_{i=1}^{n} e_i^2. \tag{9.4}$$

This gives the following *least-squares* criterion:

$$\text{minimize SSE}\,. \tag{9.5}$$

The solution is given in the next theorem.

Theorem 9.1: Least-squares Estimates

The values for $\widehat{\beta}_1$ and $\widehat{\beta}_0$ that minimize the least-squares criterion (9.5) are:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{9.6}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \, . \tag{9.7}$$

Proof: We seek to minimize 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):

$$\begin{aligned} \sum_{i=1}^n x_i (y_i - a - bx_i) &= \sum_{i=1}^n x_i \left(y_i - \overline{y} + b\overline{x} - bx_i\right) \\ &= \sum_{i=1}^n x_i y_i - \overline{y} \sum_{i=1}^n x_i + b \left(n\overline{x}^2 - \sum_{i=1}^n x_i^2 \right). \end{aligned}$$

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 (9.6) and (9.7) 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*.

When dealing with parameters from the Greek alphabet, such as β, it is customary in the statistics literature 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 (Theorem 3.6). Moreover, it is not too difficult to calculate the corresponding expectations and variances. These are summarized in the next theorem. We leave the proofs to later statistics courses.

☞ 64

Theorem 9.2: Properties of the Estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$

Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have a normal distribution. Their expected values are

$$\mathbb{E}(\widehat{\beta}_0) = \beta_0 \quad \text{and} \quad \mathbb{E}(\widehat{\beta}_1) = \beta_1 \,, \tag{9.8}$$

so both are *unbiased* estimators. Their variances are

$$\mathrm{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \tag{9.9}$$

and

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \,. \tag{9.10}$$

To estimate the unknown σ^2, we can reason as follows: 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 $\mathcal{N}(0, \sigma^2)$ 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 $\frac{1}{n-1} \sum e_i^2$. It turns out we need to scale this slightly to $\frac{1}{n-2} \sum e_i^2$ to obtain an *unbiased* estimator for σ^2. This is sometimes called the **mean squared error** (MSE) or **residual squared error** (RSE).

9.2.2 Hypothesis Testing for Linear Regression

It is of interest to test whether the slope β_1 is 0. If this is the case, then there is no association between the response and the explanatory variable. There are two approaches that we could use to construct a good test statistic.

One approach is to utilize 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 one-sample normal model, we could use the test statistic

$$T = \frac{\widehat{\beta}_1 \sqrt{\sum (x_i - \overline{x})^2}}{\widehat{\sigma}} \,. \tag{9.11}$$

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.

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 the usual R formula — in this case `height~shoe`.

```
> model1 = lm(height ~ shoe, data = survey)
> model1


Call:
lm(formula = height ~ shoe, data = survey)

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:

```
> xyplot(height~shoe,data=survey,type = c("p","r"),
        cex=1.2,pch=16, col.line="red", lwd =3)
```

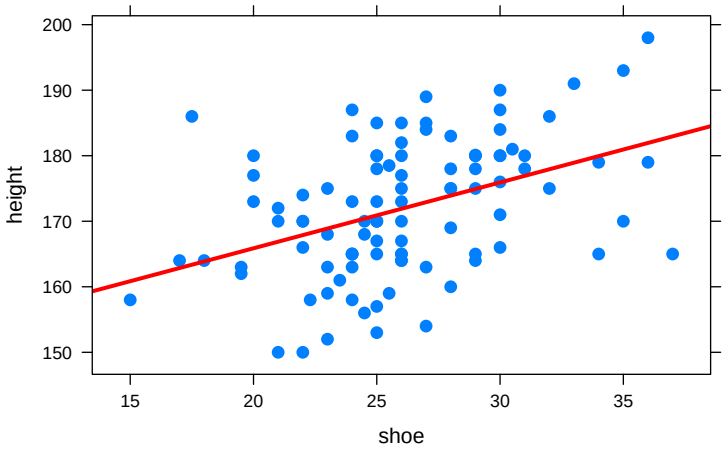


Figure 9.4: Scatter plot of height (in cm) against shoe size (in cm), with the fitted line.

The function `lm` performs a complete analysis of the linear model. The function `summary` provides a summary of the calculations:

```
> sumr1 = summary(model1)
> sumr1


Call:
lm(formula = height ~ shoe, data = survey)

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 (9.9) and (9.10);
  - `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 degrees of freedom $n - 2$.

The `R-squared` value indicates how well the linear model fits the data, in the sense that it gives the fraction of variance that is explained by the regression model, compared with the "default" model where the height data follow a normal distribution with some μ and σ^2. The closer this value is to 1, the better the fit.

The "default" model is essentially a straight horizontal line (no slope) with an error variance. The sum of squared residuals for that model is called the total sum of squares, defined as

$$\text{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{9.12}$$

A regression model can be compared to the default model in terms of the size of the sum of squared residuals using

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \tag{9.13}$$

which is known as the **coefficient of determination**. For simple linear regression, $R^2 = r^2$, i.e. it is equal to the square of the sample correlation between X and Y.

In this case, the fraction of variance explained by the regression (R^2) is 0.1784. Only 17.8 % of the variability of the height is explained by shoe size. If possible, it would be desirable to measure and include other explanatory variables (multiple linear regression) to increase the model's predictive power, i.e. increase R^2.

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, the following code extracts the P-value for the slope:

```
> sumr1$coefficients[2,4]
```

```
[1] 1.1994e-05
```

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 $Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$. 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 (numerical) confidence interval for y:

$$\widehat{y} \pm t_{\alpha/2,n-2}\, \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\,,$$

where $t_{\alpha/2,n-2}$ 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 (not just its expectation) 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. A random prediction interval is an interval (U, V) where U and V depend only on the (random) data and are chosen such that $\mathbb{P}(U \leqslant Y \leqslant V) = 1 - \alpha$. A corresponding outcome (u, v) is a numerical prediction interval. Using Theorem 9.2, we can find the following numerical prediction interval:

$$\widehat{y} \pm t_{\alpha/2,n-2}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \,.$$

The R function to find the prediction interval and confidence interval for a new value x is **predict**. For example, for our student survey data, 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 can 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

Linear regression assumes that, for the ith observation, $(Y_i | X_i = x_i) \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ with Y_i independent of Y_j, for all possible values $j \neq i$. This can also be stated as $(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. This is equivalent to saying that the assumptions of a linear regression model are that the error terms $\{\varepsilon_i\}$:

1. have *constant expectation* = 0 (linearity).
2. have *constant variance*.
3. are *normally distributed*.
4. are *independent* of each other.

For linear regression, the first of these is equivalent to the assumption of linearity. It can only be met if the relationship between x and $E(Y|X = x)$ is linear.

Although we do not know each error term $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$, the observed residual $e_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$ will be an estimate of it. Hence, we can do an analysis of residuals to examine whether the underlying assumptions of the linear regression model are verified. Suggested methods for checking each assumption are given below.

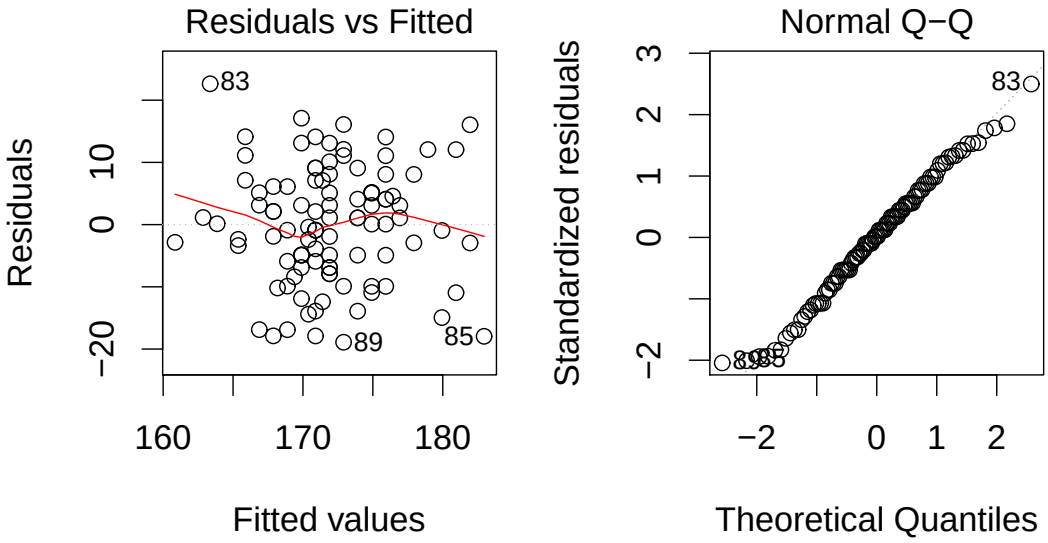
- Constant expectation 0 or linearity. This can be examined via a plot of y_i or e_i against x_i or the predicted (or fitted) values $\widehat{y}_i$. Note that the fitted value $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, so it is just a scaled, offset version of the x_i's and visually looks the same, with mirroring left to right if $\widehat{\beta}_1 < 0$. On a plot of y_i against $\widehat{y}_i$, one would expect a straight line to fit well through the middle of the data. On a plot of e_i against $\widehat{y}_i$, one would expect a flat line through $e = 0$ to pass through the middle of the data all the way along.

- Normality. This is best checked via a quantile–quantile plot (Q-Q plot) of the standardized residuals. Here, the sample quantiles of the standardized residuals are plotted against the theoretical quantiles of the standard normal distribution. Under the normality assumption the points should lie approximately on a straight line. One could also look for normality on a histogram or density plot of the residuals.

- Constant variance, also called homoscedasticity. This is easiest to examine via a plot of e_i against $\widehat{y}_i$ (residual vs fitted). One should expect to see the points having a similar standard deviation or variance all the way along from left to right, i.e. for all values of x or $\widehat{y}_i$.

- Independence. It is difficult to notice any dependence, but it can be looked for on a plot of e_i against either the predicted values $\widehat{y}_i$ or observation number (the order in which the observations were collected). When the residuals are independent, they should be uncorrelated with each other. The residuals should be found above and below the horizontal axis randomly, without any indication that nearby residuals have similar values.

```
> par(mfrow=c(1,2))
> plot(model1,1:2)
```



Examining the residuals as a function of predicted values, we see 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 outlier detection.

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

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i \,, \quad i = 1, \ldots, n \,, \tag{9.14}$$

where $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

To put across the idea, let us go back to the student survey data set `survey`. 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

```
height~shoe + 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, data = survey)
```

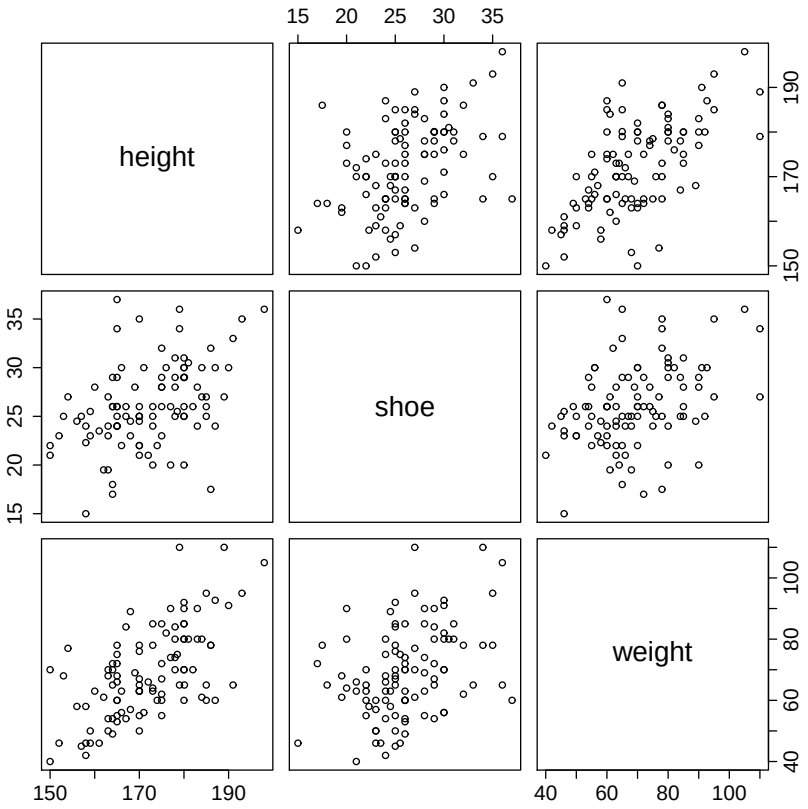


Figure 9.5: Scatter plots for all pairs of variables.

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 returned by **summary** are presented in the same fashion as for simple linear regression. The individual Student tests indicate that:

- shoe size is linearly associated with student height, after adjusting for weight (P-value = 0.0081). At the 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 the 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 for regression parameters can 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.

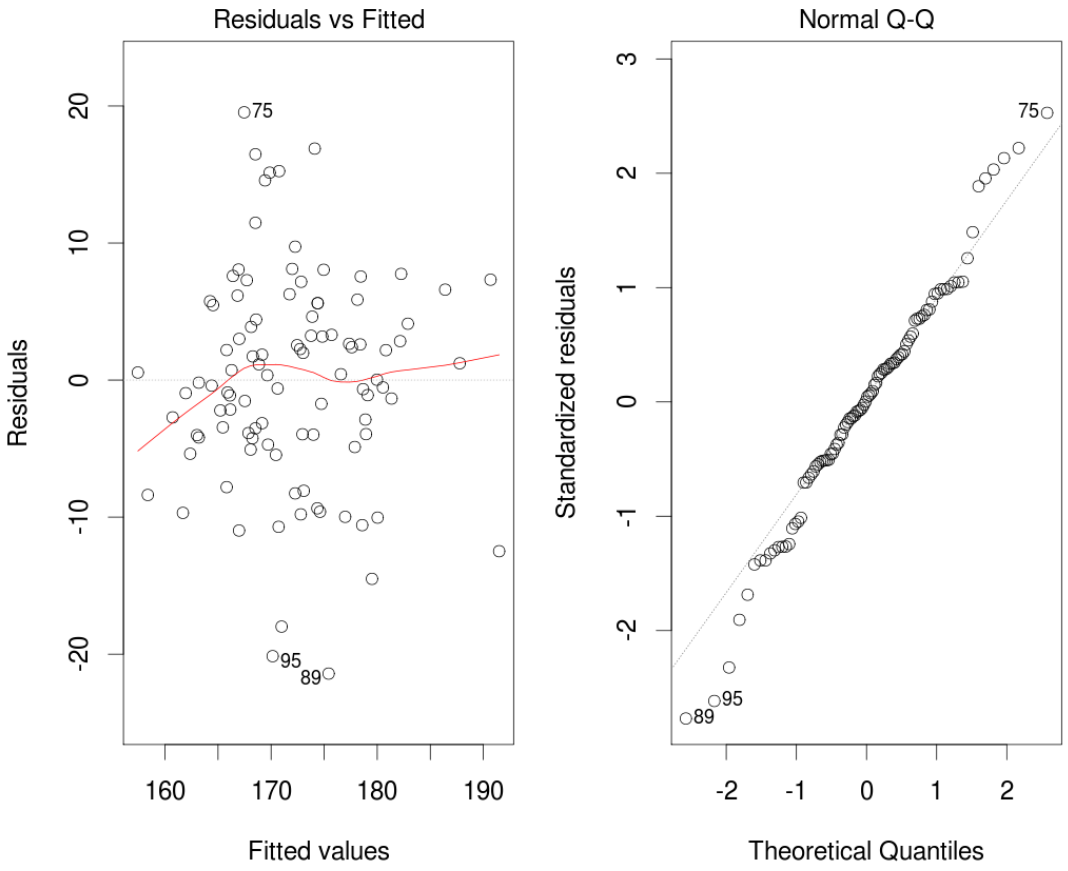
```
> predict(model2,data.frame(shoe=30,weight=75),interval="pred")
```

```
       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 Q-Q plot indicates no extreme departure from normality.

9.4 Exercises

1. For each of the plots in Figure 9.6, identify those that suggest (reasonably) that the assumptions of linear regression are satisfied. For those that don't, identify the single assumption that is **most clearly** violated by the data for each.

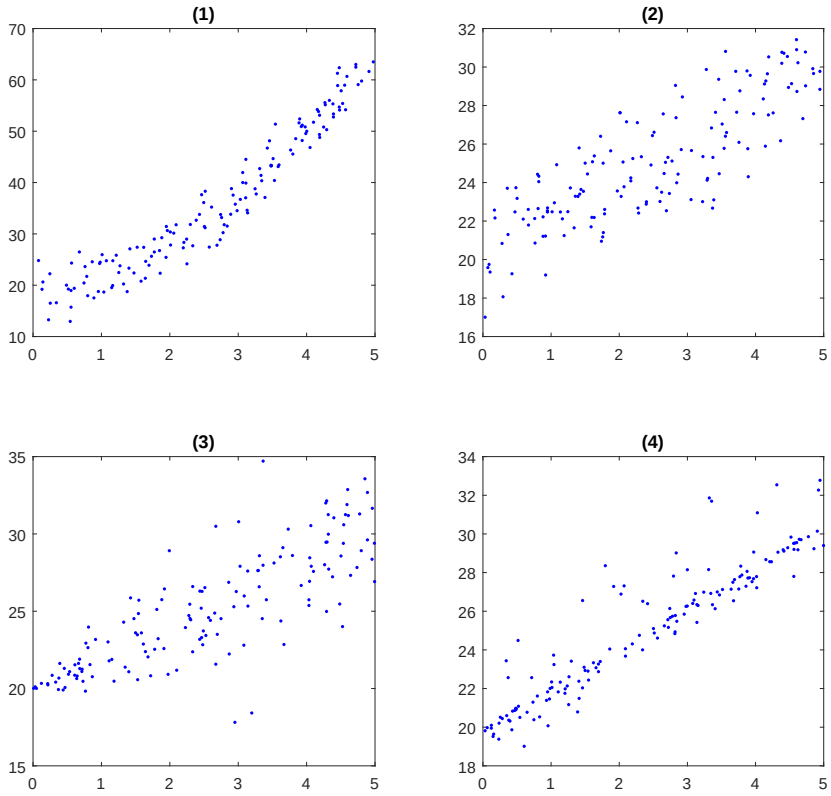


Figure 9.6: Data plots.

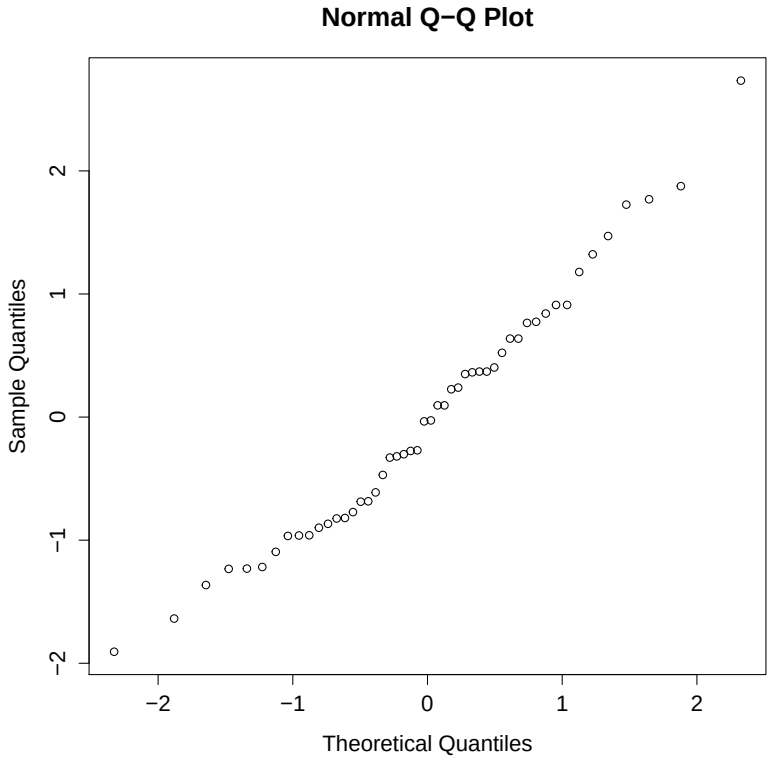
2. Researchers collected data from 50 pine trees, specifically their height (in metres) and age (in years), giving the columns Height and Age in their spread sheet. The data were entered into R, and the command lm (Height~ Age, data=Trees) was used to fit a linear model to the (Height, Age) pairs. The (partial) R output is as follows:

```
Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   5.5990       ----    6.082  1.88e-07
Age           3.3949     0.3899     ----       ----
```

   (a) Give the estimated slope and intercept in the linear relationship Height = $\beta_0 + \beta_1$Age.

(b) Using the fitted model, estimate how tall a 4-year old tree would be.

(c) The following is a quantile plot of the residuals of the fitted model. Does it suggest that the residual variability is normal? Explain why or why not.



(d) Is there a statistically significant relationship between tree age and height? Conduct the relevant statistical test and explain.

3. Figure 9.7 shows the birth weight (mass) of 74 babies in grams against the mother's age. All mothers smoked during pregnancy. Does this figure indicate that the birth weight decreases with age for smoking mothers?

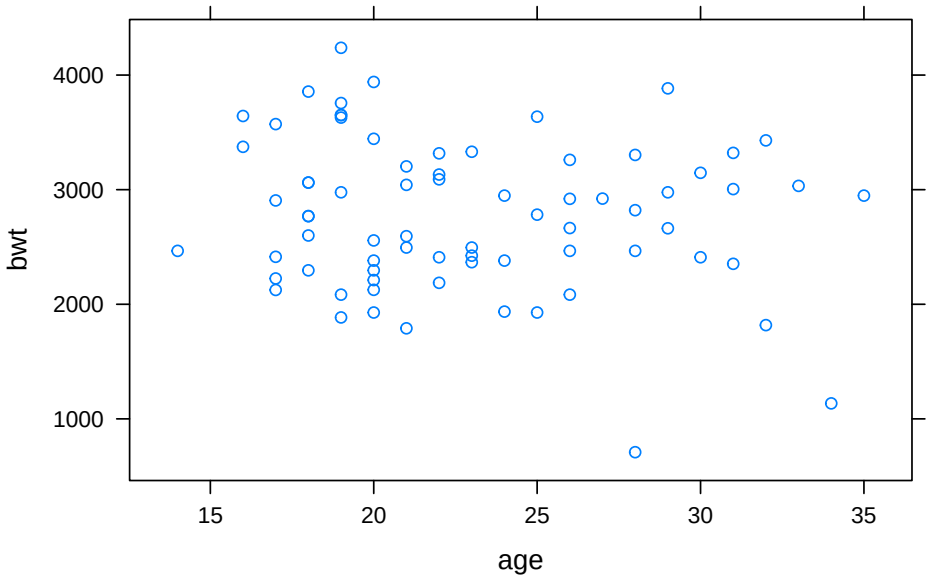


Figure 9.7: Birthweight vs age of smoking mothers.

A regression analysis with R gave the following output.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3204.23     357.96   8.951 2.57e-13
age           -18.84      15.24  -1.236     0.22


Residual standard error: 657.3 on 72 degrees of freedom
Multiple R-squared:  0.02078,Adjusted R-squared:  0.007183
F-statistic: 1.528 on 1 and 72 DF,  p-value: 0.2204
```

(a) Formulate the regression model, specifying the model assumptions.

(b) Write down the fitted regression model.

(c) Draw the estimated regression line in the figure, after computing the predicted birth weights for ages 15 and 35.

(d) Is there statistical evidence to conclude that the birth weight decreases with age?

4. Polychlorinated biphenyls (PCBs) were once used in industry but were banned in the 1970s because of concerns about their toxicity. Despite the ban, PCBs can still be detected in most people because they are persistent in the environment. A

team of researchers recorded the amount of PCBs detected in maternal milk from mothers who had eaten fish from a particular lake considered to be contaminated with PCBs. They subsequently administered an IQ test to the children when they were 11 years old. The scatter plot in Figure 9.8 shows the results along with the least-squares line fitting a linear relationship between the two variables.

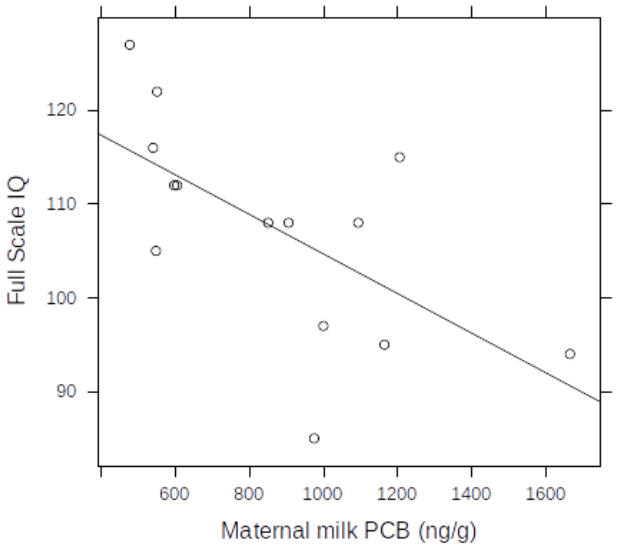


Figure 9.8: Full Scale IQ vs Maternal milk PCB (ng/g).

A regression analysis in R produced the following edited summary.

```
Coefficients:
             Estimate  Std. Error
(Intercept) 125.773972   7.008028
PCB          -0.021109   0.007538

Residual standard error: 9.314 on 12 degrees of freedom
Multiple R-squared: 0.3952,        Adjusted R-squared: 0.3448
F-statistic: 7.842 on 1 and 12 DF
```

(a) Based on the degrees of freedom, how many pairs of observations were used in the analysis?

(b) Based on the coefficients for the least-squares line provided by R, estimate the mean IQ of children if their mothers had a maternal milk PCB measurement of 1400 ng/g.

(c) State the assumptions underlying linear regression.

(d) Referring to the following Normal quantile plot of the residuals, comment on the validity of one of your assumptions.

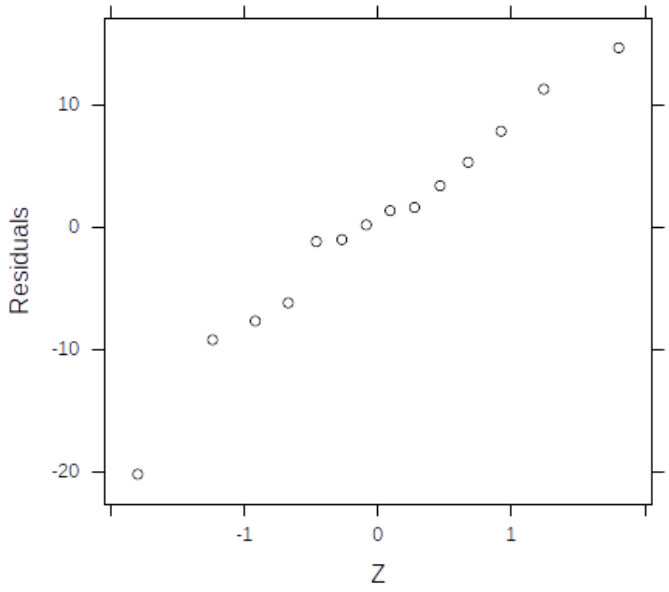


Figure 9.9: Residuals vs z-score.

(e) Is there evidence of a negative association between maternal milk PCB levels and IQ outcome?

5. Consider the following output from R.

```
Call:
lm(formula = Distance ~ Height, data = jump)

Residuals:
Min       1Q   Median       3Q      Max
-159.201  -25.580    2.489   31.213  130.076

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -218.495     27.351  -7.988 4.15e-15 ***
Height         2.207      0.158  13.966  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.52 on 903 degrees of freedom
Multiple R-squared:  0.1776,        Adjusted R-squared:  0.1767
F-statistic: 195.1 on 1 and 903 DF,  p-value: < 2.2e-16
```

What is a 95% confidence interval for the simple linear regression model of Distance for a Height of 188 cm? What is the corresponding 95% prediction interval?

6. The following multiple linear regression model for the typical yield (kg) of a grove of 100 10 year-old orange trees was proposed:

$$\texttt{Yield} = \beta_0 + \beta_1 \texttt{Rainfall} + \beta_2 \texttt{Food} + \varepsilon\,,$$

where `Rainfall` was the annual rainfall (mm) and `Food` was the annual amount of "citrus food" supplement (grams).

The output from R is given below.

```
Call:
lm(formula = Yield ~ Rainfall + Food)

Residuals:
Min       1Q   Median       3Q      Max
-12.8537  -2.7424   0.5358   2.9008  11.7751

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 99.504058   7.929248  12.549  < 2e-16 ***
Rainfall     0.017676   0.005193   3.404 0.000968 ***
Food         0.040197   0.005005   8.031 2.31e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.991 on 97 degrees of freedom
Multiple R-squared:  0.454,         Adjusted R-squared:  0.4427
F-statistic: 40.33 on 2 and 97 DF,  p-value: 1.793e-13
```

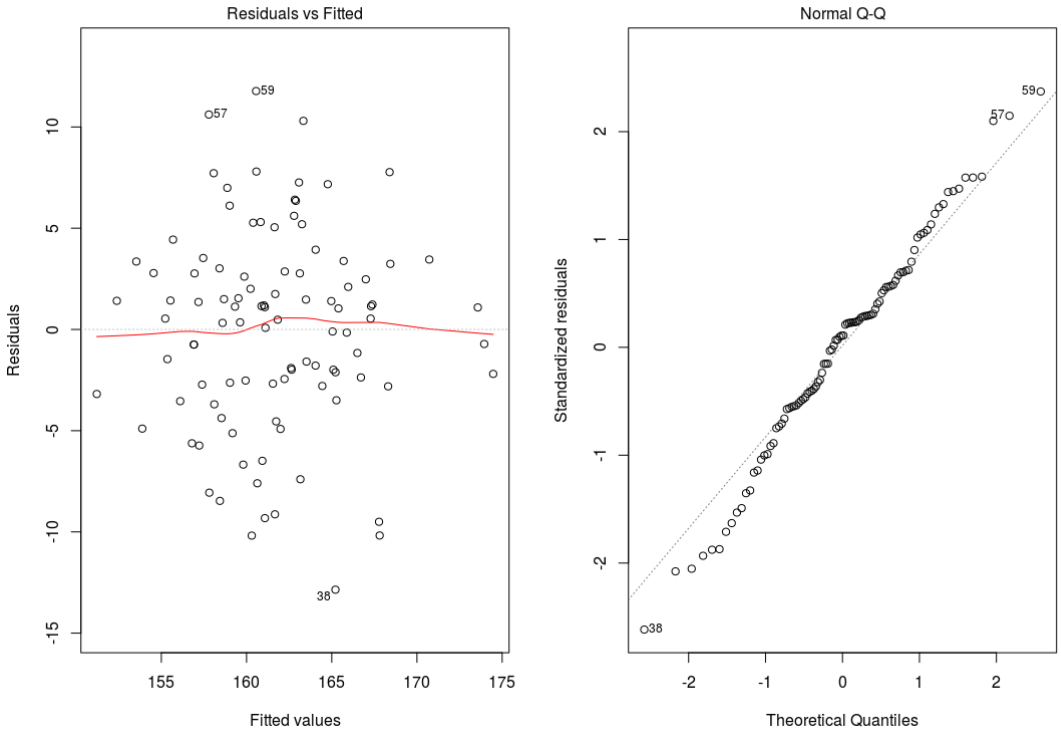
(a) Write down the fitted model.

(b) Use the fitted model to determine the estimated yield of a tree with 1000 mm of rainfall and 500 grams of citrus food.

(c) Continuing, construct a 95% prediction interval for the same values as in (a).

(d) According to the model and for the same values as in (a) and (b), what is the predicted probability that the yield would exceed 140 kg?

(e) Diagnostic plots for the residuals are displayed below.



Use these plots to comment on the validity of the model assumptions.

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 linear regression and ANOVA models are special cases of linear models, so that these can be analysed in a similar way (i.e., using the `lm` and `aov` functions). In addition to estimation and hypothesis testing, we will also consider transformation techniques that allow linear regression to more effectively deal with nonlinear data.

10.1 Introduction

The linear regression and ANOVA models in Chapters 8 and 9 are both special cases of a (normal) **linear model**. Let $\mathbf{Y}$ be the column vector of response data $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$.

Definition 10.1: Normal Linear Model

In a **normal linear model** the response data vector $\mathbf{Y}$ depends on a matrix $\boldsymbol{\mathcal{X}}$ of explanatory variables (called the **model matrix** or **design matrix**) via the linear relationship

$$\mathbf{Y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}$ is a vector of parameters and $\boldsymbol{\varepsilon}$ a vector of independent error terms, each $\mathcal{N}(0, \sigma^2)$ distributed.

■ **Example 10.1 (Simple Linear Regression)** For the simple linear regression model
☞ 186 (see Definition 9.1) we have

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} .$$

■

The situation for linear models in which the explanatory variables are *factors* is a little more complicated, requiring the introduction of indicator variables. We explain it with an example.

■ **Example 10.2 (One-factor ANOVA)** Consider a one-factor ANOVA model (see
☞ 162 Section 8.2) with 3 levels and 2 replications per level. Denote the responses by

$$\underbrace{Y_1, Y_2}_{\text{level 1}}, \underbrace{Y_3, Y_4}_{\text{level 2}}, \underbrace{Y_5, Y_6}_{\text{level 3}} .$$

Let μ_1 be the mean (i.e., expected) response at level — the reference level — and let α_2 and α_3 be the incremental effects of the other two levels. We can write the vector $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_1 + \alpha_2 \\ \mu_1 + \alpha_2 \\ \mu_1 + \alpha_3 \\ \mu_1 + \alpha_3 \end{pmatrix} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}}_{\boldsymbol{\varepsilon}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}}_{\boldsymbol{\mathcal{X}}} \underbrace{\begin{pmatrix} \mu_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} .$$

■

In R, a factor with d levels is represented by $d - 1$ **indicator** variables $x_2, \ldots, x_d$, such that if Y belongs to level k then $x_k = 1$ and all the other indicator variables have value 0. Similarly, if Y belongs to the reference level, then $x_1 = \ldots = x_d = 0$. Using this notation, we could rewrite the model in Example 10.2 as

$$Y = \mu_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon. \tag{10.1}$$

Hence all data from a general linear model is assumed to be of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n , \tag{10.2}$$

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,

$\mathbf{Y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\mathcal{X}} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Thus, the first column can always be interpreted as an "intercept" parameter. The corresponding R formula for this model would be

$$\texttt{y} \sim \texttt{x1} + \texttt{x2} + \cdots + \texttt{xp} \ .$$

Examples 10.1 and 10.2 show that it is important to treat quantitative (numbers) and qualitative (factors) explanatory variables differently. Fortunately, R automatically introduces indicator variables when the explanatory variable is a factor. We illustrate this with a few examples in which we print the model matrix, obtained via the function **model.matrix**.

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, using the function **factor**. Observe how this changes the model 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
```

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 (10.1):

$$Y = \mu + \beta_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon, \tag{10.3}$$

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. A similar interpretation holds for the parameter α_3.

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`.

10.2 Estimation and Hypothesis Testing

Suppose we observe a data vector $\mathbf{y}$ from a linear model $\mathbf{Y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mathcal{X}}$ is a known model matrix, and $\boldsymbol{\varepsilon}$ is a vector of iid $\mathcal{N}(0, \sigma^2)$ errors. We wish to estimate the parameter vector $\boldsymbol{\beta}$ and the model variance σ^2. We can again use a least-squares approach 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 $\widehat{\boldsymbol{\beta}} = (\boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}^\top \mathbf{y}$, where $(\boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{X}})^{-1}$ is the inverse of the matrix $\boldsymbol{\mathcal{X}}^\top \boldsymbol{\mathcal{X}}$. 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 minimizes 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

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n - (p+1)},$$

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 in 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). ☞ 190 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. We wish to explain the child's weight at birth using various characteristics of the mother, her family history, and her behaviour during pregnancy. ☞ 148 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 them with the function `as.factor`, similar to what we did in Chapter 5.

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.

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.

```
> model1 = lm(bwt~lwt+age+factor(race)+smoke, data = birthwt)
> sumr1 = summary(model1)
> sumr1


Call:
lm(formula = bwt ~ lwt + age + factor(race) + smoke, data = birthwt)

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 returned by **summary** are presented in the same fashion as for simple linear regression. Parameter estimates are given in the column `Estimate`.

The realizations 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 realization of Fisher's global test statistic (`F-statistic`) and the associated P-value.

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 the same age, race and mother weight, the child weight at birth is 401.720 g less for a smoking mother than for a non-smoking 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:

```
> model2 = lm(bwt~age*smoke, data=birthwt)
> summary(model2)


Call:
lm(formula = bwt ~ age * smoke, data = birthwt)

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(model2)[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 - birthwt$smoke
> confint(lm(bwt~age*nonsmoke, data=birthwt))[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 Analysis of Residuals

We present here a few elements on analysis of residuals. Suppose for the `birthwt` data set we ended up with the model represented by the following `R` formula.

```
bwt ~ smoke + age + lwt + factor(race) + ui + ht + smoke:age .
```

It is good to review what the actual model looks like, in terms of (10.2). ☞ 210

$$\texttt{Bwt} = \beta_0+\beta_1\texttt{smoke}+\beta_2\texttt{age}+\beta_3\texttt{lwt}+\beta_4\texttt{race2}+\beta_5\texttt{race3}+\beta_6\texttt{ui}+\beta_7\texttt{ht}+\beta_8\texttt{smoke}\times\texttt{age}+\varepsilon.$$

The following `R` code checks various model assumptions.

```
> finalmodel=lm(bwt~smoke+age+lwt+factor(race)+ui+ht+smoke:age,data=birthwt)
> par(mfrow=c(1:2))
> plot(finalmodel,1:2,col.smooth="red")
```

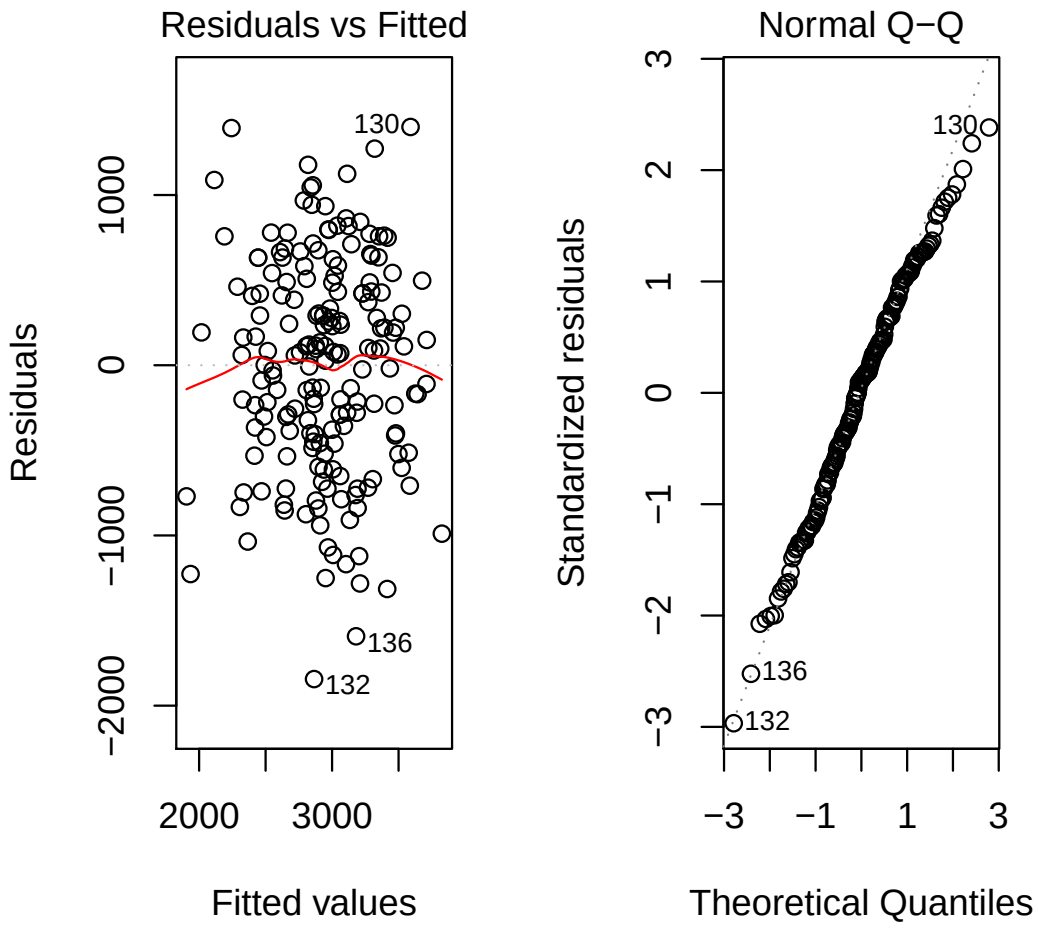


Figure 10.1: Checking the assumptions of homoscedasticity (left) and normality (right).

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. 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)
```

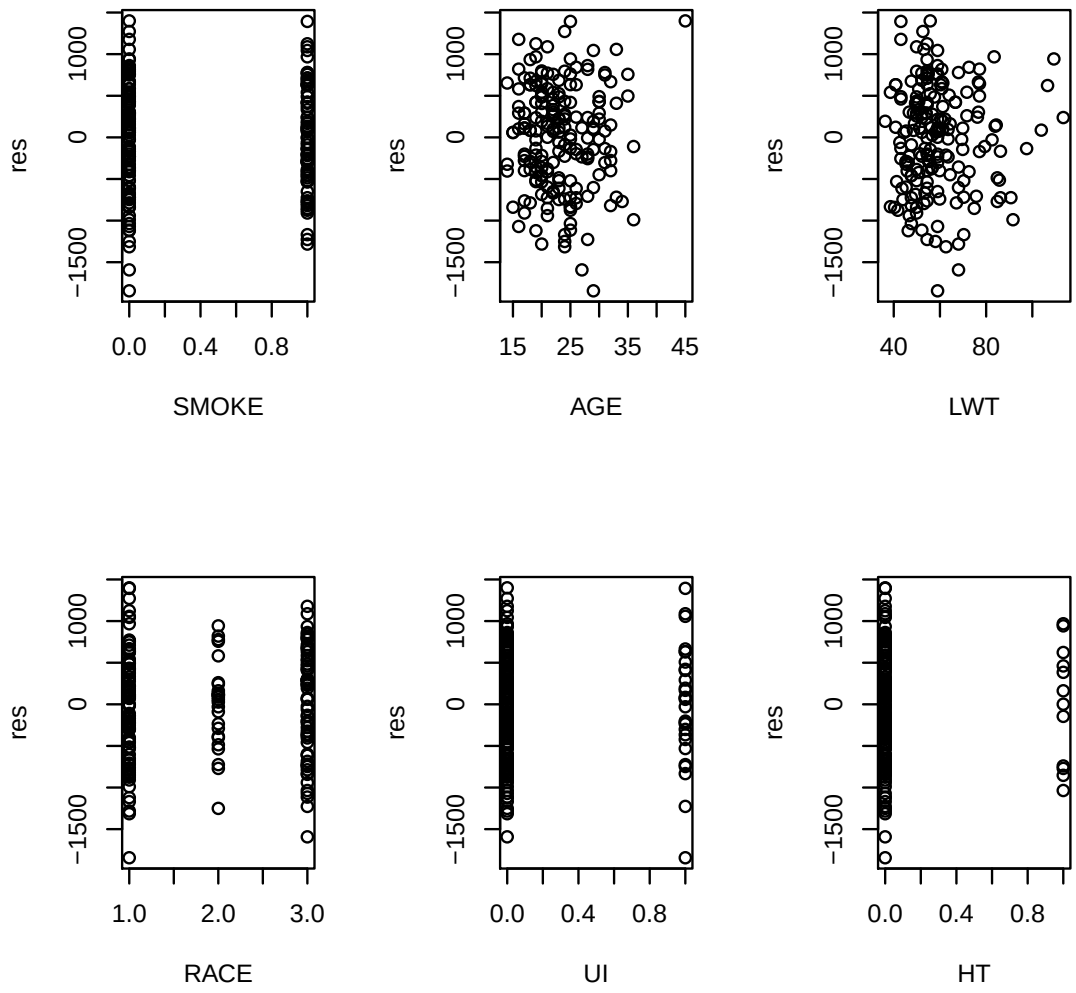


Figure 10.2: Residuals as a function of explanatory variables.

10.5 Transformations of data

When the relationship between the variables appear to be nonlinear, (for example, as revealed by scatter plots and/or residuals plots), it is sometimes possible to transform the data to make the relationship more linear. This strategy is often used to remedy the problem of not meeting the linearity and/or constant variance conditions in linear regression.

Some commonly used transformations include logarithmic, square, square root, and reciprocal. Depending on the situation, we may consider applying a transformation to the response data, to the explanatory variables, or to both the response data and the

response variables.

You will see that the data transformation approach is a trial-and-error process. After trying a transformation, we need to check the residuals plots and other model diagnostics to see if the transformed model has improved the situation. If not, we will try another transformation. We continue this process until we find an appropriate model for the data.

Take care when interpreting a transformed model, as we have changed the scale of the transformed variables. Sometimes it is useful to write down the "back-transformed" model.

Let's consider an example. The dataset `test.csv` comprises the test score (between 0 and 400) and the training time (in weeks) of 100 participants in a learning program. A scatterplot of the data reveals a nonlinear relationship between `score` and `week`. This is also evident in the residual plot.

```
> test <- read.csv("test.csv", header=TRUE)
> library(lattice)
> xyplot(score~week, data=test, pch=16)
> test.lm <- lm(score ~ week, data= test)
> plot(test.lm,1)
```

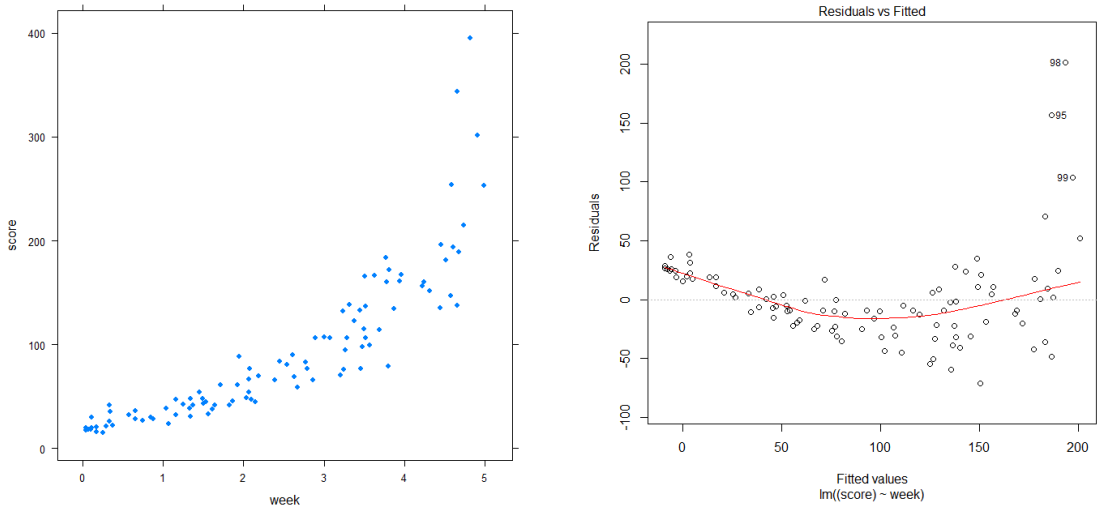


Figure 10.3: Fitting a linear regression to the `test` dataset.

Looking at the scatterplot, what type of transformation(s) could be suitable? Consider applying a logarithmic transform to `score`.

```
> xyplot(log(score)~week, data=test, pch=16)
> test.log.lm <- lm(log(score) ~ week, data= test)
> plot(test.log.lm,1)
```

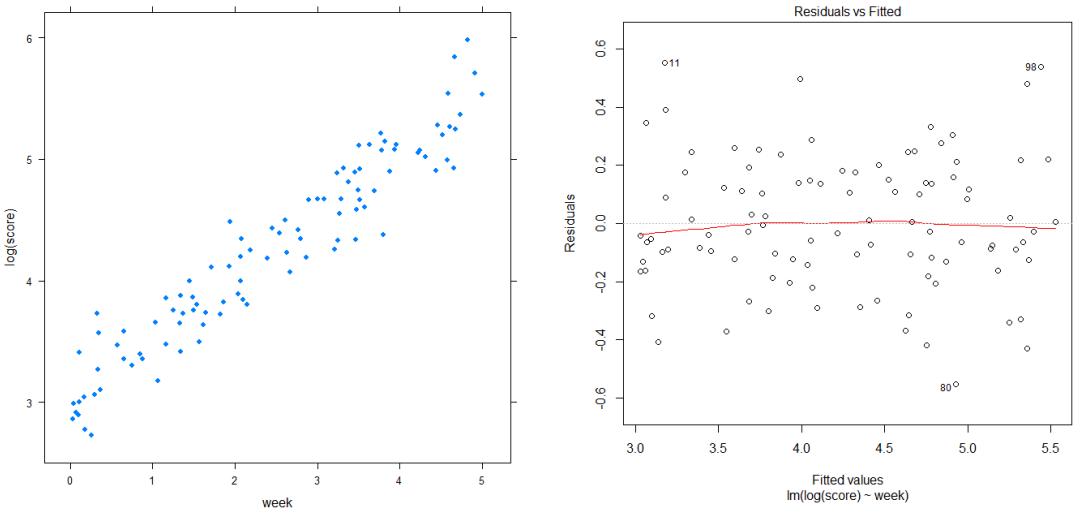


Figure 10.4: Applying a log transform to `score`.

This transformed model looks more reasonable.

```
summary(test.log.lm)

Call:
lm(formula = log(score) ~ week)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55106 -0.13114 -0.02827  0.15291  0.55200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.01087    0.04404   68.36   <2e-16 ***
week         0.50370    0.01544   32.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2276 on 98 degrees of freedom
Multiple R-squared:  0.9156,    Adjusted R-squared:  0.9148
F-statistic:  1064 on 1 and 98 DF,  p-value: < 2.2e-16
```

The model being fitted is in the form of

$$\ln(\texttt{score}) = \beta_0 + \beta_1 \texttt{week} + \varepsilon,$$

and its mean (or median) is $\mathbb{E}[\ln(\texttt{score})] = \beta_0 + \beta_1 \texttt{week}$. To obtain a model in the original scale of `score`, we can 'back-transform' this model, which takes the form of

$$\texttt{score} = e^{\beta_0} e^{\beta_1 \texttt{week}} e^{\varepsilon}.$$

In this case, the regression coefficients can be interpreted in a *multiplicative* manner rather than an *additive* manner. For example, we could say that a one-week increase of training time is associated with a $100 \times \left(e^{\widehat{\beta}_1} - 1\right)\% \approx 65\%$ increase in the `score` (in the original scale). The median of the back-transformed model is $e^{\beta_0 + \beta_1 \texttt{week}}$. Notice that the error term ε also becomes multiplicative (and not additive) in the back-transformed model.

10.6 Exercises

1. The simple linear regression model for data $\{(x_i, y_i)\}_{i=1}^n$ could be viewed as a normal linear model with design matrix $\boldsymbol{X}$ and parameter vector $\boldsymbol{\beta}$ given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

   Use the formula for $\widehat{\boldsymbol{\beta}}$ and simplify if possible to get an explicit expression for the intercept and slope of the simple linear regression model.

2. Consider the two-factor ANOVA model represented with "factor effects".

   Suppose there are two replications for each combination, the first factor has 2 levels and the second factor has three levels.

   View it as a normal linear model by writing down the design matrix $\boldsymbol{X}$ and corresponding parameter vector $\boldsymbol{\beta}$.

3. A team of scientists are investigating a huge population of large sharks near the *Island of Dr. Moreau*. The team measured the "weight" (actually mass in kg) and sex of 300 randomly selected sharks, as well as the distance (in km) these sharks swim in a week.

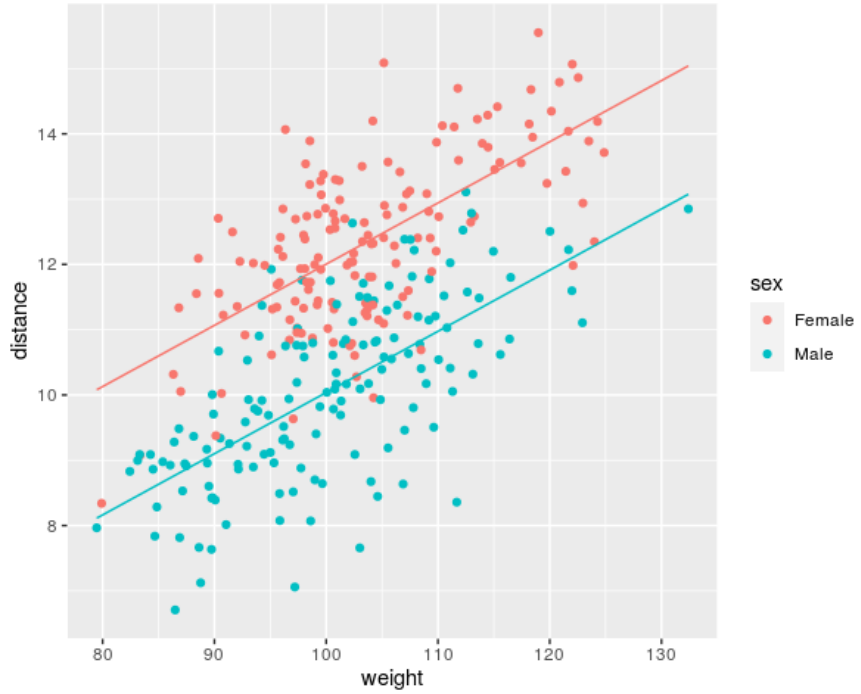


Figure 10.5: Swimming distance against weights of sharks.

Based on the sample of 151 female sharks, the scientists found a 95% confidence interval for the mean population weight of female sharks of (102.1, 105.1).

(a) What is the average weight of the 151 female sharks? What is the margin of error of this confidence interval?

(b) How many female sharks should the scientists catch to obtain a margin of error that is 4 times smaller?

(c) Do the male and female populations differ in their mean weights? To assess this, the team conducted a two-sample t-test on the weights of the 151 female and 149 male sharks. The value of the test statistic was $t = 3.178$, on 296 degrees of freedom. Formulate the null and alternative hypotheses, and calculate the P-value for this test. What is your conclusion?

A linear model analysis in R gives the following output:

```
Call:
lm(formula = dist ~ weight + sex, data = sharks)

Residuals:
     Min       1Q   Median       3Q      Max
-2.77432 -0.59452  0.04493  0.63498  2.59997
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.631348   0.632384   4.161 4.15e-05 ***
weight       0.093725   0.006055  15.478  < 2e-16 ***
sexMale     -1.966000   0.116240 -16.913  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.99 on 297 degrees of freedom
Multiple R-squared:  0.6835,        Adjusted R-squared:  0.6814
F-statistic: 320.7 on 2 and 297 DF,  p-value: < 2.2e-16
```

(d) Write down the fitted model for the female shark and the fitted model for the male shark.

(e) Based on the R output above, estimate how far a 100 kg female shark would swim in a week and how far a 100 kg male shark would swim in a week.

4. A survey in the country of Isotopia recorded the yearly wages of 856 individuals (in 1000 Isotopian dollars) as well as their gender and their perceived social class, with 5 choices: upper, upper-middle, middle, working, and lower class. For the R analysis shown below, the data was stored in a data frame `iso`, with columns `wage`, `class`, and `gender`. We have removed some output (indicated by letters `A, B,...,F`).

```
> mod = lm(wage~class + gender,data=iso)
> summary(mod)


Call:
lm(formula = wage ~ class + gender, data = iso)

Residuals:
    Min      1Q  Median      3Q     Max
-32.387  -6.928  -0.181   6.586  38.753

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         93.242      1.625  57.377  < 2e-16 ***
classUpper-middle   -5.503      1.740  -3.163  0.00162 **
classMiddle         -4.981      1.705  -2.922  0.00357 **
classWorking        -4.247      1.652  -2.572  0.01029 *
classLower         -13.448      1.993  -6.749 2.76e-11 ***
genderFemale         3.927      0.703    A       B
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 10.12 on 850 degrees of freedom
Multiple R-squared:  0.09971,Adjusted R-squared:  0.09441
F-statistic: 18.83 on 5 and 850 DF,  p-value: < 2.2e-16
```

```
> anova(mod)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
class         4   6441    C     15.73 1.98e-12 ***
gender        1   D      3194     E     F
Residuals   850  86988    102
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Estimate the expected income of working-class males.

(b) Is there statistical evidence that the expected income of upper class people differs from that of working class people?

(c) What is the meaning of the number 3.927 in the `genderFemale` row of the R output?

(d) Complete the missing numbers `A`, `C`, `D`, and `E`. Based on these, find the P-values in `B` and `F`. Do these indicate that on average women earn more than men in Isotopia?

5. Researchers were interested in modelling the relationship between the `Weight` (kg) and the `Length` (cm) of a certain fish species. A scatterplot of the data is shown in Figure 10.6(a). The researchers fitted a linear model to the data:

$$\texttt{Weight} = \beta_0 + \beta_1 \texttt{Length} + \varepsilon\,.$$

A diagnostic plot for the residuals is shown in Figure 10.6(b).

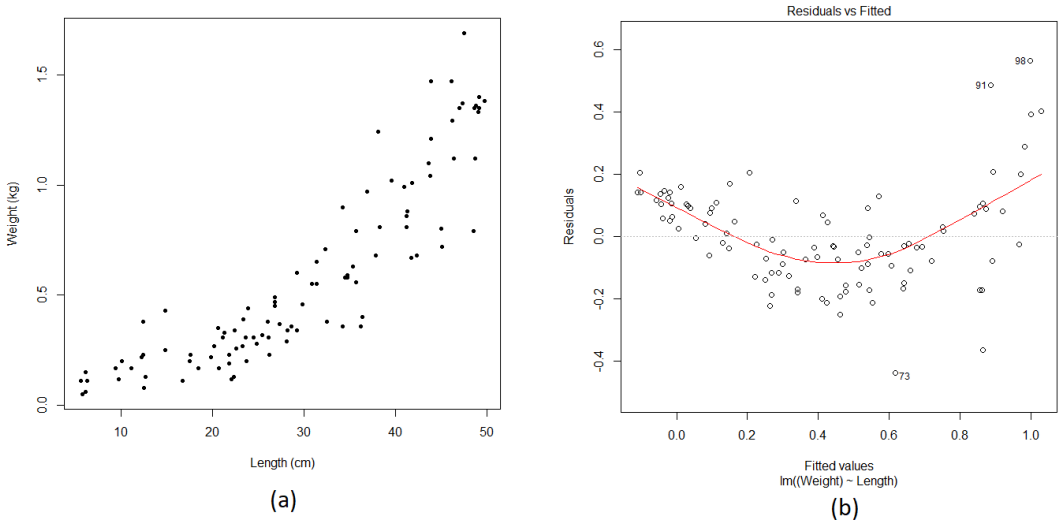


Figure 10.6: Scatterplot of the fish data and diagnostic plot for the fitted model.

(a) Explain why the above model does not provide a good fit to the data.

(b) Now consider a logarithmic transformation on the response variable. Read the data in to `R` from the file `fish.csv`. Fit the transformed linear model to the data using the `lm` command in `R`. Produce a plot similar to Figure 10.2 for the transformed model. Discuss whether this model is more appropriate for the data.

(c) Write down the fitted (transformed) model.

(d) What is the predicted weight of a fish that is 35cm long?

(e) Find the 95% prediction interval for a the weight of a fish that is 35cm long.

(f) A research caught a fish that is 35cm long. It weighted 0.893kg. Is this unusual?

(g) What would be the expected change in weight if a fish grew 1cm longer?

CHAPTER 11

CHI-SQUARED TESTS

The purpose of this chapter is to introduce you to some other useful techniques for data analysis. We will look at two tests that are related to the chi-squared distribution. The first is known under the name of *goodness of fit* (GoF) test and can be used for verifying that the data comes from a described distribution. The second test is known as an independence test and can be used to test whether two random variables are independent.

The concept behind the goodness of fit tests that we will consider in this chapter is quite straightforward. Essentially, GoF tests are comparing observed values with expected values. We first group the observed values into k classes, then the occupancy in each class is compared to its "expected" occupancy that is calculated based on the prescribed model. If the data do come from the described distribution, then we would expect the "observed" and expected counts to be close to each other. Hence large discrepancies between these counts would suggest otherwise.

There are various types of goodness of fit tests, but the ones we will consider here all rely, in some way or the other, on the properties of the *multinomial distribution*; and the corresponding test statistics have asymptotically a χ^2-distribution.

Another useful application of the χ^2-distribution is in contingency (or dependency) testing. Here, we are interested in testing the null hypothesis that two random variables are independent.

You will see the ideas behind goodness of fit tests and the independence test extend the concepts that you have learned in previous chapters. Emphasis will be more on the practical side (how can we apply the techniques, for example in R) rather than on full mathematical proofs, which would be out of scope for a first-year course.

11.1 Multinomial Distribution

We return to the very beginning of the course where, in Example 1.1, we discussed ☞ 10

how to simulate 100 coin tosses with a fair coin in order to estimate the probability of obtaining 60 or more Heads. Later on, we found that the number of Heads out of 100 tosses with a fair coin followed a $\mathsf{Bin}(100, 1/2)$ distribution. More generally, we found that the number X of Heads in n tosses, with a coin that has probability p of Heads, has a $\mathsf{Bin}(n, p)$ distribution. An equivalent way of simulating X is by throwing n balls in two boxes (0 and 1) with probability $1 - p$ and p, respectively, and counting how many balls are in box 1. The number in box 0 is then $n - X$.

We want to generalize this to throwing n balls into k boxes, numbered $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$. The resulting counts $X_1, \ldots, X_k$ turn out to have a *multinomial* distribution.

Definition 11.1: Multinomial Distribution

Random variables $X_1, \ldots, X_k$ are said to have a **multinomial** distribution, with parameters $n, p_1, \ldots, p_k$ if

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}\,,$$

for all $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$ for which $x_1 + x_2 + \cdots + x_k = n$. We write $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$.

■ **Example 11.1 (Army Recruits)** Suppose the IQ of army recruits is $\mathcal{N}(100, 16^2)$ distributed. Army recruits are classified as

$$\begin{array}{lcl} \text{Class 1} & : & \text{IQ} \leqslant 90 \\ \text{Class 2} & : & 90 < \text{IQ} \leqslant 110 \\ \text{Class 3} & : & \text{IQ} > 110 \end{array}$$

The *proportion* p_1, p_2 and p_3 of army recruits in the three categories are given by $\mathbb{P}(Y \leqslant 90) = p_1$, $\mathbb{P}(90 < Y \leqslant 110) = p_2$ and $\mathbb{P}(Y > 110) = p_3$, where $Y \sim \mathcal{N}(100, 16^2)$. It follows that we have the following proportions:

$$\begin{array}{lcl} \text{Class 1} & : & p_1 = 0.266 \\ \text{Class 2} & : & p_2 = 0.468 \\ \text{Class 3} & : & p_3 = 0.266 \end{array}$$

Now suppose we have 7 new recruits. What is the probability that of these 7 new recruits, two are Class 1; four are Class 2 and one is Class 3?

To answer this, let X_i be the number in class $i, i = 1, 2, 3$. Then, $(X_1, X_2, X_3) \sim \mathsf{Mnom}(7, p_1, p_2, p_3)$. Thus, it follows immediately that

$$\mathbb{P}(X_1 = 2, X_2 = 4, X_3 = 1) = \frac{7!}{2!\, 4!\, 1!}\, p_1^2\, p_2^4\, p_3^1 \approx 0.0957.$$

■

For the goodness-of-fit tests that we will discuss next, the following theorem is of utmost importance. The proof relies on the Central Limit Theorem and the fact that the square of a standard normal random variable has a χ^2_1 distribution.

Theorem 11.1: Multinomial Data, Known Parameters

Let $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$, then the random variable

$$\sum_{i=1}^{k} \frac{(X_i - n\,p_i)^2}{n\,p_i}$$

has approximately a χ^2_{k-1} distribution, for large n.

Remark 11.1 As a rule of thumb, we can use the approximation above provided that

$$np_i \geqslant 5, \qquad \text{for all } i.$$

■ **Example 11.2 (Simulation Experiment)** It is relatively easy to verify Theorem 11.1 for specific cases through simulation — in the same way that we verified the central limit theorem in Figure 3.4. In particular, suppose we throw $n = 100$ balls into $k = 10$ boxes, numbered 1,...,10, with equal probability. Let $X_1, \ldots, X_{10}$ be the counts. A typical count outcome is $14, 6, 13, 12, 11, 12, 5, 11, 7, 9$. The corresponding outcome of the random variable in Theorem 11.1 is 8.6 (check yourself). Figure 11.1 shows a histogram for $R = 1000$ such outcomes. We see that it matches well the density of the χ^2_9 distribution. The following code was used.

☞ 64

```
x = 1:10
n = 100
R = 1000
t = vector()   # initialize a vector t
for (i in 1:R){
  s = sample(x, size=n, replace=TRUE)
  h = hist(s, breaks=0:10) # create a histogram object
  ob  = h$counts   # contains the observed counts
  ex = 10
  t[i] = sum((ob - ex)^2/ex)
}

hist(t,breaks=30,freq = FALSE,main="")
curve(dchisq(x,df=9),xlim=c(0,30), col=2,lw=2,add = TRUE)
```

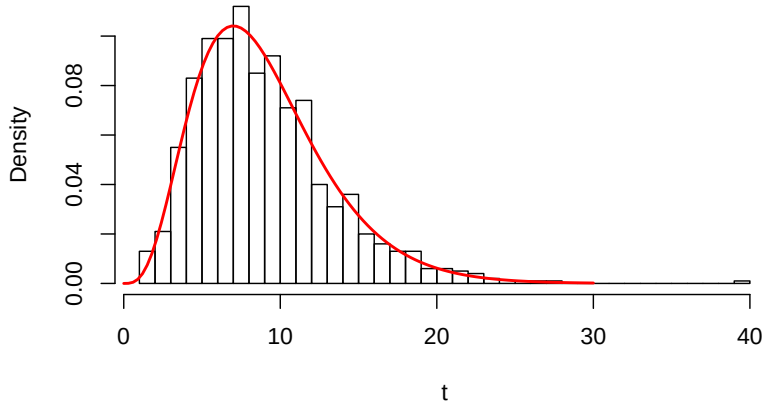


Figure 11.1: The histogram of the test statistic values closely matches the pdf of the χ^2_9 distribution (red curve).

11.2 Goodness of Fit with Known Parameters

We can use Theorem 11.1 to formulate a **goodness of fit test** for count data. Specifically, suppose we have a multinomial data

$$(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k).$$

We can test $H_0 : p_1 = \pi_1, \ldots, p_k = \pi_k$ against the alternative hypothesis that H_0 is not true by using the test statistic

$$T = \sum_{i=1}^{k} \frac{(X_i - n\,\pi_i)^2}{n\,\pi_i},$$

which, under H_0, has a χ^2_{k-1} distribution, by Theorem 11.1. We reject H_0 at the α level of significance if

$$T \geqslant q,$$

where q is the $(1-\alpha)$-quantile of the χ^2_{k-1} distribution.

Remark 11.2 We can symbolically write the test statistic as

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the *observed* number of observations in class i and E_i is the *expected* number of observations in class i. This form for the test statistic is found in any goodness of fit test.

■ **Example 11.3 (Frizzled Chickens)** The phenomenon of *complete dominance* predicts that progeny whose genetic component is (F, F) will be extremely frizzled, progeny with (F, f) slightly frizzled and (f, f) will be normal. According to the genetics theory, the proportions FF : Ff : ff should be 1 : 2 : 1. Out of 93 randomly selected chickens the observed frequencies phenotypes are 23 (extremely frizzled), 50 (slightly frizzled) and 20 (normal). Is this in accordance with the theory?

We can test this with a χ^2-goodness-of-fit test. Let us go through the usual steps of a statistical test (see Section 7.1). ☞ 138

1. Let X_1, X_2, X_3 be the total number of FF, Ff and ff chickens out of 93. Our model is: $(X_1, X_2, X_3) \sim \mathsf{Mnom}(93, p_1, p_2, p_3)$, for unknown p_1, p_2, p_3.
2. We want to test: $H_0 : p_1 = \frac{1}{4}, p_2 = \frac{1}{2}, p_3 = \frac{1}{4}$ against the alternative hypothesis that H_0 is not true.
3. As test statistic we use
$$T := \frac{(X_1 - 93/4)^2}{93/4} + \frac{(X_2 - 93/2)^2}{93/2} + \frac{(X_3 - 93/4)^2}{93/4} .$$
4. Under H_0 this has approximately a χ^2_2 distribution, see Theorem 11.1.
5. The outcome of T is
$$t = \frac{(23 - 93/4)^2}{93/4} + \frac{(50 - 93/2)^2}{93/2} + \frac{(20 - 93/4)^2}{93/4} = 0.72 .$$
6. The P-value for this right-one-sided test is $\mathbb{P}_{H_0}(T \geqslant 0.72) = 0.70$.
7. Because the value is high (0.70), we accept H_0; that is, we find no evidence to reject the theory.

■

11.3 Testing Independence

In Theorem 11.1, it is assumed that the probabilities $\{p_i\}$ are *known*. In many cases, however, these probabilities need to be *estimated* from the data, giving rise to the following modification of Theorem 11.1.

Theorem 11.2: Multinomial Data, Unknown Parameters

Let $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$, where the $p_i = p_i(\theta)$ depend on an unknown r-dimensional parameter vector θ. Denoting $\widehat{p_i} = p_i(\widehat{\theta})$ the (maximum likelihood) estimate of $p_i(\theta)$, the random variable

$$\sum_{i=1}^{k} \frac{(X_i - n\,\widehat{p_i})^2}{n\,\widehat{p_i}}$$

has approximately a χ^2_{k-1-r} distribution, for large n.

Comparing this with Theorem 11.1 we see that we apparently "lose" r degrees of freedom if we have to estimate r parameters.

An important application of Theorem 11.2 occurs in a *two-way table* of counts (also called a **contingency table**), where we wish to test for an association (i.e., dependence) between the two variables. We explain the idea via a specific example first.

■ **Example 11.4 (ESP Belief)** We wish to examine whether artists differ from non-artists in Extra-Sensory Perception (ESP) belief. Table 11.1 lists the amount of belief in ESP for a group of 114 Artists and a group of 344 Non-artists. We wish to investigate whether being an artist or not is "independent" of the ESP belief (strong, moderate or not).

Table 11.1: ESP belief

	ESP belief			
	Strong	Moderate	Not	total
Artists	67	41	6	114
Non-artists	129	183	32	344
	196	224	38	458

To see that this is a type of goodness of fit situation, we need to properly formulate a model for the data and express the null and alternative hypotheses in terms of the parameters in the model.

If we ignore the row and column totals, we have a table with $r = 2$ rows and $c = 3$ columns. We can imagine the table to be filled in the following way: We randomly select 458 people and ask whether they are an artist or not and what their ESP belief is. Let (U_k, V_k) denote the response for the kth selected person, where $U_k \in \{1, 2\}$, where (1 = artist, 2=non-artist), and $V_k \in \{1, 2, 3\}$, where, (1 = strong belief , 2 = medium belief, 3= no belief). We assume that $(U_1, V_1), \ldots, (U_n, V_n)$ are independent and distributed as a random vector (U, V) that can take values $(1, 1), (1, 2), (1, 3), (2, 1), (2, 2)$ and $(2, 3)$ with probabilities $p_{11}, p_{12}, \ldots, p_{23}$.

Now, instead of recording all (U_k, V_k), we could instead *count* how many people are artist with a strong ESP belief, artist with a Moderate ESP belief, etc. Let $X_{i,j}$ be the *count* in row i and column j. That is, the total number of observations out of $n = 458$ that fall in "cell" (i, j). For example, the outcome of $X_{2,2}$ is 183. From the model above we have

$$(X_{11}, \ldots, X_{23}) \sim \mathsf{Mnom}(n, p_{11}, \ldots, p_{23}) \,.$$

We wish to test whether null hypothesis that the random variables U and V are *independent*. In terms of the parameters of the model, the null hypothesis can be written as

$$H_0 : \quad p_{ij} = p_i \, q_j, \qquad \text{for all } i, j,$$

where p_1, p_2, q_1, q_2 and q_3 are unknown probabilities. Using Theorem 11.2, we can test the null hypothesis against the alternative hypothesis that $p_{ij} \neq p_i \, q_j$ for some i and j, by using the test statistic

$$T = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(X_{ij} - E_{ij})^2}{E_{ij}},$$

where E_{ij} is an estimator of np_{ij}, the expected number of observations in cell (i, j). Under H_0, this is $np_i \, q_j$. The natural estimators for p_i and q_j are

$$\widehat{p_i} = \frac{\sum_{j=1}^{3} X_{ij}}{458} \quad \text{and} \quad \widehat{q_j} = \frac{\sum_{i=1}^{2} X_{ij}}{458};$$

and hence we estimate the expected count as $E_{ij} = n\widehat{p_i}\widehat{q_j}$. In other words,

$$E_{ij} = \frac{\text{total count in row } i \times \text{total count in column } j}{\text{total count}}. \tag{11.1}$$

By Theorem 11.2 the test statistic T has under H_0 approximately a χ^2_2 distribution, because the total number of parameters to be estimated is $r = 1 + 2 = 3$. We have to subtract r from the total number of classes $r \times c$ minus 1, i.e., $6 - 1 = 5$ to get number of degrees of freedom: is $5 - 3 = 2$. We reject the null hypothesis for large values of T. The various estimated counts (in brackets) are given in the table below.

For example, $E_{11} = 114 \times 196/458 \approx 48.8$. It follows that the outcome of T is $t = 6.79 + 3.93 + 1.27 + 2.20 + 1.34 + 0.43 = 15.96$. The p-value is 0.00034. Hence, we strongly reject H_0. Artists indeed seem to differ from non-artists in ESP belief. ■

For the *general* contingency table, we have count data in r rows and c columns. Again, we wish to test for association between the variables. Let X_{ij} be the total number of observations (out of n) that fall in *cell* (i, j) (i.e., in the ith row and jth column). We have

$$(X_{11}, \ldots, X_{rc}) \sim \mathsf{Mnom}(n, p_{11}, \ldots, p_{rc}) \,.$$

Table 11.2: Observed and estimated counts for the ESP belief.

	ESP belief			
	Strong	Moderate	Not	total
Artists	67 (48.8)	41 (55.8)	6 (9.46)	114
Non-artists	129 (147)	183 (168)	32 (28.5)	344
	196	224	38	458

If there is no association between the two variables (null hypothesis), then

$$p_{ij} = p_i\, q_j, \quad \forall i, j,$$

for some (unknown) $p_1, \ldots, p_r$ and $q_1, \ldots, q_c$. We can test this by using the test statistic

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - E_{ij})^2}{E_{ij}},$$

where E_{ij} is as in (11.1). Under the null hypothesis of no association, the test statistics has approximately a χ^2_{df} distribution with the degrees of freedom parameter equal to

$$\text{df} = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)\ .$$

And we reject H_0 for large values of T.

11.4 Exercises

1. Dr Good counts the number of possums in 5 areas of similar size and habitat. She finds the following counts: 234, 301, 256, 315, 274. Does this indicate that certain areas are preferred by the possums? Carry out a χ^2 goodness of fit test.

   (a) What are the expected counts if all areas are equally attractive to the possums?

   (b) What is the outcome of the χ^2 test statistic?

   (c) What is the (approximate) distribution of the test statistic under the null hypothesis?

   (d) Using the table for the χ^2 distribution, what is the P-value for this test?

   (e) What is your conclusion about the area preferences of the possums?

2. Certain pheromones may influence male mating success in a species of fruit fly (*Drosophila melanogaster*). A group of behavioural geneticists performs a series of 50 mating trials to determine whether the *elongase* genotype is involved in determining mating success. Below is a two-way table of counts for genotype (either AA, AB, or BB), and Mating Success (either No or Yes).

	No	Yes
AA	6	6
AB	15	14
BB	2	7

We will conduct a χ^2 test to seek evidence for an association between genotype and the mating success.

(a) If there was no such association, what would be the expected counts for each cell?

(b) Compute the χ^2 statistic and give the p-value for the hypothesis test. What do you conclude?

APPENDIX A

R PRIMER

A.1 Installing R and RStudio

R is both a programming language and a work environment specifically developed for data analysis. The creators, Ross Ihaka & Robert Gentleman, wished to make available a free and open-source version of the statistical package 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. R works across multiple platforms (Windows, Mac, Linux) and is constantly evolving due to the contributions of a large and growing community of volunteers. Some advantages of using R:

- R is *free* and *easy to install and maintain*, especially when using integrated development environments (IDEs) such as RStudio.
- R has many external *packages*: collections of functions and data tailored to certain tasks. These packages can be easily installed, e.g., via RStudio.
- R has many efficient inbuilt procedures for *statistics*, data management and visualization.
- R has an integrated and accessible *documentation* system.

Install R from the *Comprehensive R Archive Network* (CRAN): `http://cran.r-project.org`.

R's base system comes with a rudimentary Graphical User Interface. We recommend instead the use of RStudio's integrated development environment (IDE), depicted in Figure A.1.

Install RStudio from `https://www.rstudio.com/`.

This IDE comprises (customizable) windows for R programs (top-left), the R console (bottom-left), environment variables and history (top-right), and plotting and packages information (bottom-right).

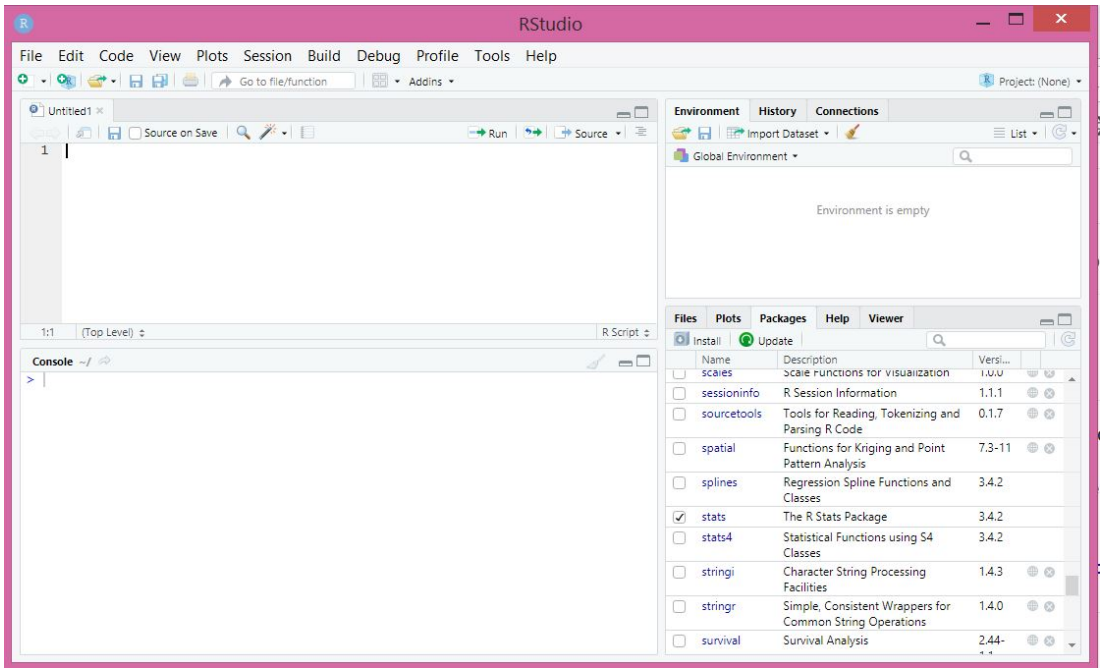


Figure A.1: A typical `RStudio` layout.

A.2 Learning R

There are many resources available to help you learn R. In RStudio, for example, the *Help>R Help* menu gives access to the comprehensive tutorial "An Introduction to R", as well as to the precise "R Language Definition". In this section we will merely give an overview of R. If at any point you need help, the first thing to do is consult R's **`help`** function. For example `help("sin")` will show information about the **`sin`** function and other trigonometric functions. An Internet search is nowadays also a good alternative, which often will bring you to a *Stack Exchange* question and answer website.

The URL `https://www.stat.berkeley.edu/~spector/Rcourse.pdf` by Phil Spector gives a comprehensive 103-slide introduction to the R language.

A.2.1 R as a Calculator

The simplest thing you can do with R is to use it as a basic calculator, as in

```
> 1*2*3*4
```

```
[1] 24
```

and

```
> sin(1)
```

```
[1] 0.841471
```

Here **sin** is the built-in trigonometric function. As on your calculator, numbers can be stored in memory. This is done via the **assignment operator** =, as in:

```
> xx = 10
```

To see the contents of 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 class function:

```
> class(my.text)
```

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

Names of objects are case-sensitive, and must begin with a letter and not contain spaces. Names may include fullstops, such as `my.name`.

A.2.2 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"
```

Vectors consisting of a sequence of numbers can be created via the "colon" operator, e.g., `1:5` is the same as `c(1,2,3,4,5)`, or via the `seq` function, as in:

```
> my.sequence = seq(from=1, to=20, by=2)
> my.sequence
```

```
 [1]  1  3  5  7  9 11 13 15 17 19
```

A vector can be rearranged into a matrix via the `matrix` function:

```
> matrix(my.sequence,ncol=5,nrow=2)
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    5    9   13   17
[2,]    3    7   11   15   19
```

If the number of elements in the vector is smaller than the number of elements in the matrix, the vector elements will be "cycled":

```
> matrix(1:5, ncol=5, nrow=2)
```

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

Let's now combine the two vectors `age` and `Author` into a new object with the **`cbind`** (column bind) function.

```
> AgeAuthorObject = cbind(age,Author)
> AgeAuthorObject
```

```
 age    Author
[1,] "50"   "Dirk"
[2,] "38.5" "Benoit"
[3,] "37.5"   "Michael"
```

We have created again a matrix object. Since, matrix objects must have the same type of objects, R has coerced (cast) the numerical age vector into a vector of strings. You can see that the numbers in the `age` column are between quotation marks. In R, a matrix object is seen as vector with extra attributes, in particular the dimension of the matrix, and possibly the row and column names. The attributes of an object can be obtained and set via the **`attributes`** function. 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, you need to use **data frame** objects, which can be constructed via 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 the data frame's name. The command **`names`** is not specific to the `data.frame` object but can be applied to other R objects as well, such as the `list` object, which is defined later.

```
> names(AgeAuthorObject)
```

```
[1] "age"    "Author"
```

Notice that the first column of the data set has no name and is a series of numbers. This is the `row.names` attribute of the data frame. We can use the `rownames` command to set the row names from a vector.

```
> rownames(AgeAuthorObject) = c("First","Second","Third")
> AgeAuthorObject
```

```
        age  Author
First    50    Dirk
Second 38.5  Benoit
Third  37.5 Michael
```

A.2.3 Component Selection

The dollar sign ($) is called the **component selector**. It enables to extract any column of a matrix-type object via its name.

```
> 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 the **attach** command. 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 the default packages of 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 brackets `[]`. 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 through
                  # 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 information.

```
> mask = c(TRUE,FALSE,TRUE,NA,FALSE,FALSE,TRUE)
> which(mask) # Outputs the indices corresponding 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
```

A.2.4 List Objects

The most flexible and richest data structure in R is the **list**. Lists can group together 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
```

A.2.5 Linear Algebra

We can do the usual linear algebra operations in R. Here are some examples.

```
> (A = matrix(1:6,nrow = 2)) # define matrix A and show output
```

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

Notice that we have used brackets around the assignments of A to force the output to be shown in the console. You can also achieve this by typing the name of the variable in the console or by using `print(A)` in the source code.

A vector in R is always interpreted as a column vector, even though it is printed as a row vector.

```
> (x = 3:1)  # define vector x and show output
```

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

Multiplying A with x results in a matrix object with 2 rows and 1 column.

```
> A %*% x  # multiply  matrix A with (column) vector x
```

```
     [,1]
[1,]   14
[2,]   20
```

We can coerce this matrix back into a vector (if needed) with the `as.vector` function. The transpose of a matrix is found via the `t` function.

```
> t(A)  # transpose of A
```

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

Functions of a matrix are foremost treated in an *elementwise* way, as in

```
> 1/A
```

```
     [,1]      [,2]      [,3]
[1,]  1.0 0.3333333 0.2000000
[2,]  0.5 0.2500000 0.1666667
```

and

```
  A * A
```

```
     [,1] [,2] [,3]
[1,]    1    9   25
[2,]    4   16   36
```

A common mistake is to use * instead of %*% for matrix multiplication. This need not give an error message, as the elementwise operation may be perfectly legitimate.

Let us introduce another matrix, B, which is a square matrix.

```
> B = matrix(1:4,nrow = 2)
> print(B)
```

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

The inverse of an invertible square matrix B can be found by solving the linear equation $BX = I$ for matrix X. In R we use the `solve` function:

```
> (solve(B))    # compute and print the inverse of B
```

```
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
```

Next, we solve the linear equation $B\mathbf{x} = (1, 2)^\top$:

```
> x  = solve(B,c(1,2))
> x
```

```
[1] -0.5  0.5
```

The function **apply** can be used to apply a function to the rows or columns of a matrix. For example, with matrix A as above, the column means are:

```
> apply(A,MARGIN=2,FUN=mean)  #column means
```

```
[1] 1.5 3.5 5.5
```

and the row means are

```
> apply(A,1,mean) #row means
```

```
[1] 3 4
```

A.2.6 Flow Control

R has the usual flow control statements for programming, including "if", "for" and "while" statements:

- if (*condition*) { *expression* } else { *expression* }
- for (*var* in *seq*) { *expression* }
- while (*condition*) { *expression* }

The R code below gives some examples. We can execute the code in RStudio via the "source" button. Typing `source("filename")` in the console, where `filename` is to be specified by you, will execute the code as well. The code illustrates also the use of the **cat** and **print** functions to output results. See the help files for their different uses. The function **scan** can be used to input data, from file, URL, or keyboard. To output a new line, use the special character "\n". Note that $x == y$ (that is, double equal sign) is used to compare x with y. In the code below, two strings are compared.

```
cat("Input name");
name = scan(,what="char",nmax=1)  # read the name from keyboard input
if (name == "Dirk"){ print("Welcome back Dirk")
  } else {cat("Hello",name,"\n")}  #important to have "} else"

for (i in 1:10) cat(i^2," ")   # output numbers in a row
cat("\n")                      # put newline in output

# this does the same but prints the results as a column
i = 1
while (i <= 10){
  print(i^2)
  i = i+1
}
```

A common mistake in "if else" statements is to start the "else" as the first word of a new line. This confuses R, as it deals with the previous statement as an "if" statement without the "else" part.

A.2.7 Functions

Functions are simply a set of statements that transform an "input" objects into an "output" object. We have already seen several examples of functions, such as the function **mean**. The input to this function is a vector of numbers, and the output is the mean (i.e., average) of these numbers.

The standard way to use a function is to assign the result of the function f to an object y, as in `y = f(x)`. However, some functions in `R` can change the attribute $f(x)$ of x to z via an assignment `f(x) = z`. A common example is the **`names`** function, which not only shows the names of an object, but can be used to change the names of that object as well. Another example is the **`levels`** function.

Some functions can be called with and "empty" argument. For example, **`getwd`** gives the current working directory:

```
> getwd()
```

```
"c:/Users/JohnSmith/DataProject"
```

Arguments are the input into a function, and use the `ARGUMENTLABEL=VALUE` syntax. To find all of arguments that a command can accept look at `Arguments` section of the command's help file. Argument labels may be put *in any order* and also can be abbreviated provided there is no ambiguity. It is advised, though, to keep the labels and the order exactly as in the specified argument list.

```
> ?rnorm  #open help file for rnorm
> x = rnorm(n=10,mean=3,sd=2)  #generate normal random variables
> (q = quantile(x, probs = c(0.25,0.75))) #output two quantiles
```

```
     25%      75%
1.021414 4.237529
```

Basic Functions

Here are some important data manipulation functions. See the help files for extra arguments.

- **`length`**: returns the length of a vector.

  ```
  > length(c(1,3,6,2,7,4,8,1,0))
  ```

  ```
  [1] 9
  ```

- **`sort`**: sorts the elements of a vector, in increasing order.

  ```
  > sort(c(1,3,6,2,7,4,8,1,0))
  ```

  ```
  [1] 0 1 1 2 3 4 6 7 8
  ```

- **order**, **rank**: the first function returns the vector of ranking indices of the elements. In case of a tie, the ordering is always from left to right. The second function returns the vector of ranks of the elements. In case of a tie, the ranks are shared and can be non-integer.

```
> 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**: 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
```

- **which**: gives the indices of a boolean vector that are TRUE.

```
> x = c(1,3,6,2,7,4,8,1,0)
> which(x > 2)

[1] 2 3 5 6 7

> x[ind]
```

```
[1] 3 6 7 4 8
```

- **rep**: replicates the values of a vector or object.

```
> rep(1:3,4)

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

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 scrutinize 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

$$\text{BMI} = \frac{\text{Weight}}{\text{Height}^2}.$$

The function BMI defined below returns a list of three named elements (`Weight`, `Height` and `BMI`).

```
BMI = function(weight,height){
            bmi = weight/height^2
            res = list(Weight=weight,Height=height,BMI=bmi)
            return(res)}
```

We can now execute the function `BMI` we just created:

```
>  BMI(weight=70, height=1.82)


 $Weight
[1] 70

$Height
[1] 1.82

$BMI
[1] 21.13271
```

A.2.8 Graphics

R comes with a "base" `graphics` package. To see the available functions and variables, type:

```
> ls(package:graphics)
```

Some of these are *high-level* functions, which produce complete plots with a single or just a few commands. Examples inlcude **plot**, **boxplot**, **contour**, **barplot**, and **hist**. Other plotting functions are *low-level*, plotting only parts of plots, such as **abline**, **points**, **curve**, **frame**, **axis**, **text**, and **legend**.

The function **plot** is a *generic* function for plotting R objects. Each object invokes its own plot function, called a *method*. Type `methods(plot)` to see all the methods that are associated with the **plot** function.

Various parameters can be used to change the appearance of a plot. Although in general R does a good job at selecting the right layout, when the plots need to be incorporated in a pdf for example, it may be important to change the size of the fonts. This is generally done via the function **par**. Type `?par` to find the many plotting parameters.

Plotting parameters of particular use are:

- `cex` : changes the size of the characters (especially useful when exporting graphs to be included in a LaTeX document).
- `mar`: a vector of form `c(bottom, left, top, right)` giving the margins of the plot.
- `pch` : the point type: either specified by a character or an integer. See `?points` for a list.
- `lty`: the line type, specified by an integer.
- `lwd`: the line width.
- `col`: the color of a line or character, specified as a character string or a number. Type **colors** for available colors.

The following code gives the graph in Figure A.2.

```
f = function(x) sin(x)
g = function(x) sin(x)*exp(-x)
windows(width=8,height=5)  # draws an external window in MS Windows
par(cex=1.5,mar = c(4,2,0.2,0.2))
curve(f,0,pi, lwd = 3,col="blue",xlab = "x",ylab="")
```

```
curve(g,0,pi, lwd = 3,lty="dashed",col="darkorange",xlab = "x",add=T)
legend(-0.05,1.02,c("f(x)","g(x)"),col=c("blue","darkorange"),lwd=2,
       lty=c("solid","dashed"),bty="n")
```

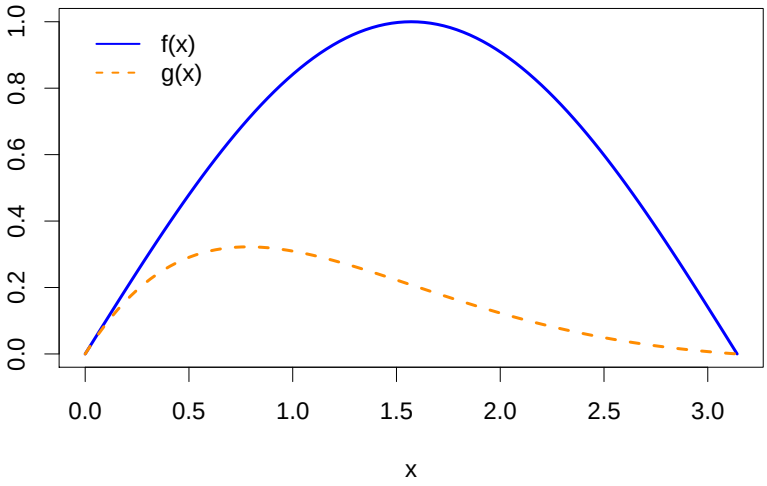


Figure A.2: A simple plot.

Using the commands below we can quickly plot a fitted line to the `cars` regression data. The result is depicted in Figure A.3.

```
> plot(cars)
> abline(lm(dist~speed,data=cars),col="blue")
> points(cars[30,],col="red",pch=20)
```

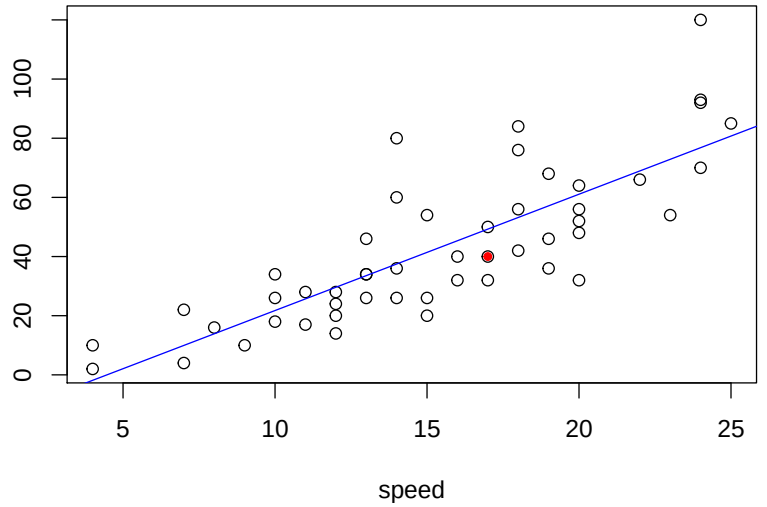


Figure A.3: Plotting a fitted line to regression data, and highlighting one point.

Instead of writing to a window, R can also write to other devices, such as a pdf or postscript file. For example:

```
> pdf("cars.pdf")
> plot(cars)
> dev.off()
```

Plots the cars data into a pdf file called cars.pdf. The file is not written until the command `dev.off()` is issued.

A.2.9 Reading and Writing 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; see the helpfile. For CSV (Comma Separated Values) file, you can use instead **`read.csv`**.

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")
```

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.

For writing data, 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")
```

A.2.10 Workspace, Batch Files, Package Installation

All of the objects you create become part of your workspace. Use the `ls` function to list all of the objects in your current workspace.

```
> ls()   # the brackets () are essential!
```

```
> [1] "age"  "AgeAuthorObject" "Author"  "my.text"
```

You can remove specific objects by using the `rm` function:

```
> rm(my.text)  #remove my.text object
```

If you want to remove all objects in the workspace use `rm(list=ls())`.

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 use the up and down arrows on your keyboard when your cursor is in the R console to scroll through your history.

Finally, we mention that R can be called in a shell. In RStudio, a shell can be created under the Tools menu. In the shell, you can type

```
> R CMD --help
```

to get a list of possible commands. In particular,

```
> R CMD BATCH infile.R outfile.txt
```

executes the statements from `infile.R` and writes the results to `outfile.txt`, or to standard output when the output file is not provided. You can also use `Rscript` instead of `R CMD BATCH`. To try things out, run the following batch file, e.g, named `batch.R`, in a command shell:

```
x = seq(0,2*pi,by=0.1)
print(x)      # write to standard output
windows()     # open a window
plot(sin(x))
Sys.sleep(5)  # wait 5 seconds before exiting
```

Before you can load a package with `library()`, you will need to *install* it first. In R studio the "Packages" tab in the lower-right IDE pane shows the packages that have already been installed. Clicking on the "Install" tab opens a window to search for new packages, which then will be automatically installed. Packages can also be manually installed via the `install.package` function.

APPENDIX B

ANSWERS TO EXERCISES

B.1 Understanding Randomness

1. (a) 256

   (b) $\mathbb{P}(X = x) = \dfrac{\binom{8}{x}}{256}$

   (c) $\frac{1}{32}$

   (d) $\frac{9}{256}$

2. 0.2

3. 0.25

4. (a) 0.5

   (b) 0.5

   (c) $f(y) = \begin{cases} \frac{1}{3}, & \text{if } y \in [2, 5], \\ 0, & \text{otherwise.} \end{cases}$

   (d) 3.5

5. $\frac{5\pi}{3}$

6. 1.5

7. $\frac{5}{12}$

8. 2

9. 0.037

B.2 Common Probability Distributions

1. 274
2. (a) 1.3693
   (b) 0.25
   (c) 0.776
3. 6.29
4. 0.0475 (or 0.0668 with continuity correction)
5. (a) 0.0228
   (b) $\mu \in [33.42, 33.44]$

B.3 Multiple Random Variables

1. (a) 0.0668
   (b) 0.017
2. (a) 600, 18.97
   (b) 0.3085
   (c) 0.057
3. (a) $f_X(x) = \begin{cases} 0.6 & \text{if } x = 0, \\ 0.25 & \text{if } x = 1, \\ 0.15 & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$
   (b) $f_Y(y) = \begin{cases} 0.7 & \text{if } y = 0, \\ 0.3 & \text{if } y = 1, \\ 0 & \text{otherwise.} \end{cases}$
   (c) $\mathbb{E}(X) = 0.55, \mathrm{SD}(X) = 0.74, \mathbb{E}(Y) = 0.3, \mathrm{SD}(Y) = 0.458$
   (d) $\mathrm{Cov}(X, Y) = 0, \varrho(X, Y) = 0$
   (e) Yes
4. (a) X and Y are independent.
   (b) $\mathbb{E}(XY) = 0.5$
5. $\mathrm{Var}(Z) = pq(1 - pq)$

6. (a) $\mathrm{SD}(\widetilde{Y}) = 5.556$
   (b) $\mathrm{SD}(Z) \approx 3.157$
7. (a) $f_{X,Y}(x, y) = \pi^{-1}$ for $x^2 + y^2 \leqslant 1$
   (b) $f_X(x) = 2\pi^{-1}\sqrt{1 - x^2}$ for $x \in [-1, 1]$, $f_Y(y) = 2\pi^{-1}\sqrt{1 - y^2}$ for $y \in [-1, 1]$
   (c) 0
   (d) Yes

B.4 Studies, Data, and Evidence

Chapter 4 exercises will be discussed during lectures.

B.5 Descriptive Statistics

Chapter 5 exercises will be discussed during lectures.

B.6 Estimation

1. $(3.71, 5.58)$
2. $(1069.2, 1202.8)$
3. at least 13188
4. $(98.1, 103.9)$
5. 24
6. $(-0.057, 0.013)$
7. $(21.3, 64.5)$

B.7 Hypothesis Testing

1. (a) one-sample normal data
   (b) $H_0 : \mu = 19$ vs $H_1 : \mu > 19$
   (c) $T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
   (d) $T \sim t_{39}$
   (e) 2.57

(f) 0.007

(g) reject

2. (a) $H_0 : p = 0.519$ vs $p \neq 0.519$

   (b) $z = -2.64$, P-value ≈ 0.0082

   (c) $(0.42, 0.5)$

3. (a) $\widehat{p} = 0.7$, CI $\approx (0.5, 0.9)$

   (b) P-value=0.058

   (c) P-value ≈ 0.037

4. (a) 0.159

   (b) 6.23

   (c) P-value=1

5. p-value = 0.00174

6. (a) 0.0321

   (b) 0.416

   (c) 0.021

7. (a) $(4.32, 4.89)$

   (b) P-value = 0.49

8. (a) $(-0.05, 0.81)$

   (b) to be discussed during lectures

   (c) P-value ≈ 0.07

9. (a) $H_0 : p_X = p_Y$ vs $H_1 : p_X > p_Y$

   (b) P-value = 0.23

   (c) $(-0.067, 0.121)$

10. P-value = 0.2526

B.8 Analysis of Variance

1. (a) see lecture annotations or solutions

   (b) $H_0 : \mu_{HbSS} = \mu_{HbST} = \mu_{HbSC}$

   (c) see lecture annotations or solutions

(d) independence, normality, constant variance

(e) P-value $\approx 2.3 \times 10^{-11}$

2. (a) see lecture annotations or solutions

(b) 6.06

(c) 0.6828

3. (a) to be discussed during lectures

(b) see lecture annotations or solutions

(c) 7.4, P-value = 0.0016

(d) 0.526

(e) constant variance

(f) 0.9854

(g) Bonferroni correction

(h) to be discussed during lectures

4. Question 4 will be discussed during lectures.

5. Question 5 will be discussed during lectures.

6. Perform pairwise comparisons with Bonferroni correction. See lecture annotations or solutions.

B.9 Regression

1. Question 1 will be discussed during lectures.

2. (a) 3.3949, 5.599

(b) 19.18

(c) see lecture annotations or solutions

(d) P-value $\approx 1.92 \times 10^{-11}$

3. (a) $\text{Weight}_i = \beta_0 + \beta_1 \text{Age}_i + \varepsilon_i$ where $\varepsilon_i \sim iid\ N(0, \sigma^2)$ for $i = 1, 2, \ldots, 74$

(b) Weight = $3204.23 - 18.84 \times$ Age

(c) 2921.18, 2543.78

(d) P-value = 0.1102

4. (a) 14

(b) 92.23

(c) linearity, independence, normality, equal variance

(d) see lecture annotations or solutions

(e) P-value = 0.008

5. (190.84, 201.97), (106.91, 285.91)

6. (a) $\widehat{\text{Yield}} = 99.504058 + 0 : 017626 \times \text{Rainfall} + 0 : 040197 \times \text{Food}$

(b) 137.28

(c) (127.38, 147.18)

(d) 0.2928

(e) see lecture annotations or solutions

B.10 Linear Model

1. see lecture annotations or solutions

2. see lecture annotations or solutions

3. (a) 103.6, 1.5

(b) 2416

(c) P-value = 0.00164

(d) 0.2928

(e) see lecture annotations or solutions

(f) 12,10

4. (a) 89

(b) P-value = 0.01029

(c) see lecture annotations or solutions

(d) $3.127 \times 10^{-8}, 2.96 \times 10^{-3}$

5. Question 5 will be discussed during lectures.

B.11 Chi-squared Tests

1. (a) 276

(b) 15.63

(c) χ^2_4

(d) 0.0036

	No	Yes
AA	5.52	6.48
AB	13.34	15.66
BB	4.14	4.86

(e) see lecture annotations or solutions

2. (a) see table above

(b) P-value = 0.2853

INDEX