MATH40005 Probability and Statistics

Spring 2023 Version 0.0.4

Dean Bodenham

2 January 2023

Contents

Preface

These notes provide a broad introduction to statistics. **Chapter 1** uses the concepts of the central tendency and dispersion of a distribution to introduce the notions of statistics, parameters and parameter estimation. **Chapter 2** introduces different ways to visualise different types of data, and briefly describes random number generation. **Chapter 3** provides theory for the special case when we can assume our sample of random variables are i.i.d. normal, and different ways of constructing confidence intervals in this case. **Chapter 4** is an introduction to statistical hypothesis testing. **Chapter 5** describes four pitfalls or common mistakes that are made in statistics. **Chapter 6** defines covariance and correlation for data. **Chapter 7** introduces the idea of a statistical model, while **Chapter 8** introduces the notion of likelihood and maximum likelihood estimation. **Chapter 9** introduces one of the most popular methods in statistics: regression. **Chapter 10** provides a short introduction to Bayesian inference, and finally, **Chapter 11** introduces the bootstrap. The appendix provides additional examples, results and code; the material in this chapter will not be examinable.

It soon becomes apparent that R code has been included in the notes. All the figures in the notes have been created using R (in fact, these notes were created using R, via R Markdown, which is something you will learn to use during this term), and sometimes the code has been included where it may aid in understanding the plot. **If the code is in the notes, the reader to is encouraged to try out the code and recreate the plot or analysis for themselves.**

Another feature of these notes is that there are two versions: one version which has gaps, usually in exercises or proofs, while another version of the notes is complete. We will fill in the 'gappy' version of the notes during lectures, and then after each chapter is completed in the lectures, the completed version of the notes will be released.

Also note while these notes contain all the material in the course, there are accompanying video lectures. Sometimes these video lectures will explain something in the notes in more detail, e.g. Section 4.4.2. The lecture videos will proceed through the notes mostly in order, although Chapters 2 and 5 will be spread between several lectures.

These notes are still a work in progress, and occasionally there may be a typo or error. If you spot one, please feel free to post or ask about it on the noticeboard.

Dean Bodenham, January 2023

Chapter 1

Central Tendency and Dispersion

The concept of a measure of **central tendency** refers to a typical value for a probability distribution. Different definitions for central tendency give rise to different statistics; the three most common measures of central tendency are the **mean**, the **median** and the **mode**. An alternative name for central tendency is location.

Dispersion is a measure of the extent to which the values of a distribution are spread out. Consequently, it is also referred to as the spread, variability or scale. The most common measures of dispersion are the **variance**, **standard deviation**, **range** and **interquartile range**.

In this chapter we shall look at both statistics for central tendency and dispersion both in terms of random variables, and the corresponding sample versions when a sample has been observed. The material in this chapter is mainly taken from [1, 2, 3, 7].

1.1 Mean, variance and higher order moments

In this section we review the notations of the expectation of a random variable X, as well as the variance and higher-order moments of a random variable X.

1.1.1 Review of expectation

For a random variable X, it was defined in Term 1 (Chapter 10 of the Prof. Veraart's lecture notes) that the **expected value** or **expectation** or **mean** of a discrete random variable X is

$$\mathrm{E}(X) = \sum_{x \in \mathrm{Im}(X)} x \mathrm{P}(X = x), \tag{1.1}$$

where $\mathrm{Im}(X)$ is the set of values in $\mathbb{R}$ that X can take, i.e. $\mathrm{Im}(X) = \{X(\omega) \mid \omega \in \Omega\}$, the image of the sample space Ω under X. Recall that $\mathrm{E}(X)$ exists only if the sum in the right-hand side of Equation (1.1) converges absolutely. See Section A.1 in the appendix for further discussion on the issue of when the expectation exists.

Recall also that for a continuous random variable X with density f_X, the expectation of X is defined as

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x, \tag{1.2}$$

provided that the integral exists and $\mathrm{E}(|X|) < \infty$. Note that, regarding notation, both parentheses and brackets will be used for expectation, i.e. $\mathrm{E}(X) = \mathrm{E}[X]$.

1.1.2 Notation for random variables and observed values

Recall the notational convention that random variables are denoted by **UPPERCASE** letters, while observed values, or realisations of the random variables, are denoted by **lowercase** letters. For example, the random variable X is observed during an experiment to have the value x.

In the case that we are observing or measuring several values of a single random variable, for example the toss of a coin, we might write that $x_1, x_2, \ldots, x_n$ are realisations of the random variable $X \sim \mathrm{Bern}(p)$.

In another situation, each of the observations might correspond to a different random variable, possibly from the same distribution. For example, consider the heights of a group of n individuals $x_1, x_2, \ldots, x_n$ as realisations of the random variables $X_1, X_2, \ldots, X_n$, where each X_i follows the distribution $\mathrm{N}(\mu, \sigma^2)$.

1.1.3 The minimum expected squared deviation

Suppose one decides to measure how much a random variable X deviates from a constant a by using the squared deviation, $(X-a)^2$. The closer a is to X, the smaller this quantity will be. To measure the deviation over all possible values for X, one considers the expected value of this quantity, $\mathrm{E}[(X-a)^2]$; the value of a that minimises it will provide a good predictor for X. One may expect this special value of a to be $\mathrm{E}(X)$, and we can manipulate the expression of the quantity

$$\begin{aligned}
\mathrm{E}[(X-a)^2] &= \mathrm{E}[(X - \mathrm{E}[X] + \mathrm{E}[X] - a)^2] \\
&= \mathrm{E}[([X - \mathrm{E}[X]] + [\mathrm{E}[X] - a])^2] \\
&= \mathrm{E}[(X - \mathrm{E}[X])^2] + 2\mathrm{E}[(X - \mathrm{E}[X])(\mathrm{E}[X] - a)] + \mathrm{E}[(\mathrm{E}[X] - a)^2] \\
&= \mathrm{E}[(X - \mathrm{E}[X])^2] + 0 + \mathrm{E}[(\mathrm{E}[X] - a)^2] \\
&= \mathrm{E}[(X - \mathrm{E}[X])^2] + (\mathrm{E}[X] - a)^2,
\end{aligned}$$

where we have used the fact that the term $(\mathrm{E}[X] - a)$ is a constant, and so since

$$\mathrm{E}[X - \mathrm{E}[X]] = \mathrm{E}[X] - \mathrm{E}[\mathrm{E}[X]] = \mathrm{E}[X] - \mathrm{E}[X] = 0,$$

the middle term is computed, using the linearity of expectation, as

$$\begin{aligned}
2\mathrm{E}[(X - \mathrm{E}[X])(\mathrm{E}[X] - a)] &= 2(\mathrm{E}[X] - a)\,\mathrm{E}[X - \mathrm{E}[X]] \\
&= 2(\mathrm{E}[X] - a) \cdot 0 \\
&= 0.
\end{aligned}$$

Now, since $(\mathrm{E}[X] - a)$ is a real number, and $(\mathrm{E}[X] - a)^2 \geq 0$,

$$\begin{aligned} \mathrm{E}[(X-a)^2] &= \mathrm{E}[(X - \mathrm{E}[X])^2] + (\mathrm{E}[X] - a)^2 \\ \Rightarrow \mathrm{E}[(X-a)^2] &\geq \mathrm{E}[(X - \mathrm{E}[X])^2]. \end{aligned}$$

Therefore, one has proved

Theorem 1.1.1. Given a random variable X, over all values $a \in \mathbb{R}$,

$$\min_a \mathrm{E}[(X-a)^2] = \mathrm{E}[(X - \mathrm{E}[X])^2]. \tag{1.3}$$

♦

In other words, $\mathrm{E}[X]$ is the value that minimises the expected squared deviation of X. This is a fundamental result that will be useful later in the course.

1.1.4 Review of variance

We recall the definition for the variance of a random variable X.

Definition 1.1.2. The **variance** of a random variable X is defined as

$$\mathrm{Var}\,(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]. \tag{1.4}$$

The positive square root of $\mathrm{Var}\,(X)$ is the **standard deviation** of X. ■

Remark 1.1.3. We have seen in Theorem 1.1.1 in Section 1.1.3 that the minimum of the quantity $\mathrm{E}[(X-a)^2]$ is obtained when this quantity is the variance of X. In some sense, this makes the variance a natural measure of dispersion, if we are taking our metric to be the squared deviation of X. □

Remark 1.1.4. Although the variance is often calculated and quoted, its square root, the standard deviation, is more easily interpretable. The reason for this is one of **units**: the standard deviation of X has the same units as X, but $\mathrm{Var}\,(X)$ has units which are the square of the unit of X. Let's consider a concrete example: suppose it is determined that the height of people from a certain country follow a normal distribution $\mathrm{N}\,(\mu, \sigma^2)$, where the mean is $\mu = 172\text{cm}$, and $\sigma = 3\text{cm}$. Therefore, the variance is 9cm^2, and so it does not even make sense to add the mean and variance in any way, since the units are different. However, it does make sense to add/subtract the mean and standard deviation. Recall that for a random variable $X \sim \mathrm{N}\,(\mu, \sigma^2)$, one has $\mathrm{P}\,(\mu - 2\sigma \leq X \leq \mu + 2\sigma) > 0.95$. For our example, if the assumption about the heights following that normal distribution is correct, that means that over 95% of the population has a height in the interval $(166\text{cm}, 178\text{cm})$. □

Exercise 1.1.5. For any random variable X, show that $\mathrm{Var}\,(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2 - (2\mathrm{E}[X])\,X + (\mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2] - (2\mathrm{E}[X])\,\mathrm{E}[X] + \mathrm{E}[(\mathrm{E}[X])^2] \\ &= \mathrm{E}[X^2] - 2\,(\mathrm{E}[X])^2 + (\mathrm{E}[X])^2 \\ &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2\end{aligned}$$

where we have used the linearity of the expectation and put constant terms in parentheses. △

Summary: Three things to remember about the variance

For any random variable X, the variance of X is denoted $\mathrm{Var}\,(X)$ and

- is **defined** as $\mathrm{Var}\,(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$,
- can be **rewritten** as $\mathrm{Var}\,(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$,
- can be **interpreted** as $\mathrm{Var}\,(X) = \min_{a \in \mathbb{R}} \mathrm{E}[(X - a)^2]$.

1.1.5 The variance of bounded random variables

The following result is useful for bounded random variables:

Proposition 1.1.6. Suppose that the random variable X is known to only take values in the bounded range $[a, b]$. Then

$$\mathrm{Var}\,(X) \leq \frac{(b-a)^2}{4}.$$

♦

Proof. See the solution to Question 2, Problem Sheet 8. □

This proposition is particularly useful for variables that follow a Bernoulli distribution:

Corollary 1.1.7. Suppose $X \sim \mathrm{Bern}(p)$, for some $p \in [0, 1]$. Then

$$\mathrm{Var}\,(X) = p(1-p) \leq \tfrac{1}{4}.$$

♦

Proof.

We know from Term 1 (Example 13.1.10 of Prof. Veraart's notes) that if $X \sim \mathrm{Bern}(p)$ then $\mathrm{Var}\,(X) = p(1-p)$. Then, there are at least three proofs:

- If $X \sim \mathrm{Bern}(p)$, then $X \in \{0, 1\} \subset [0, 1]$, and the result follows from Prop. 1.1.6.
- Without using Prop. 1.1.6, one can show the function $g(p) = p(1-p) \leq \frac{1}{4}$ on $[0, 1]$ using differentiation.
- For an elementary proof, one can write $p(1-p) = \frac{1}{4} - (p - \frac{1}{2})^2$ (completing the square), showing that $\frac{1}{4}$ is a global maximum of $p(1-p)$.

□

1.1.6 Review of moments

We also review the concept of **moments**:

Definition 1.1.8. For each positive integer k, the kth **moment** of the random variable X (or its distribution F_X) is denoted μ'_k and is defined by

$$\mu'_k = \mathrm{E}[X^k]. \tag{1.5}$$

Furthermore, the kth **central moment** of X, denoted μ_k, is defined by

$$\mu_k = \mathrm{E}[(X - \mu)^k], \tag{1.6}$$

where the μ is defined to be the first moment $\mu = \mu'_1 = \mathrm{E}[X]$. ■

There are two important moments we have already seen: the first moment μ'_1 is the mean, and the second central μ_2 moment is the variance. Some of the higher-order moments have special names and interpretations, but we shall not consider these here.

In the case that X is a continuous random variable with p.d.f. f_X,

$$\mu'_k = \mathrm{E}[X^k] = \int_{-\infty}^{\infty} x^k f_X(x)\,\mathrm{d}x,$$

as long as the integral exists, or alternatively, as long as $\mathrm{E}[|X^k|] < \infty$. See Section A.1 in the appendix for further discussion on the issue of when the expectation exists.

Remark 1.1.9. The moments μ'_k are sometimes referred to as the **raw** moments or non-central moments, in order to clearly distinguish them from the central moments. □

Exercise 1.1.10. Show that for a random variable X with mean μ and finite variance $\sigma^2 < \infty$ that the second raw moment is $\mu_2' = \mathrm{E}[X^2] = \mu^2 + \sigma^2$.

We start with the definition for the variance being the second central moment:

$$\begin{aligned}
\sigma^2 &= \mu_2 \\
&= \mathrm{E}[(X-\mu)^2] \\
&= \mathrm{E}[X^2 - 2X\mu + \mu^2] \\
&= \mathrm{E}[X^2] - 2\mu\mathrm{E}[X] + \mathrm{E}[\mu^2] \\
&= \mathrm{E}[X^2] - 2\mu^2 + \mu^2 \\
\Rightarrow \sigma^2 &= \mathrm{E}[X^2] - \mu^2 \\
\Rightarrow \mathrm{E}[X^2] &= \mu^2 + \sigma^2
\end{aligned}$$

where the fourth line used the linearity of the expectation.

Alternatively, one could have used Exercise 1.1.5:

$$\begin{aligned}
\sigma^2 = \mathrm{Var}\,(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\
\Rightarrow \mathrm{E}[X^2] &= \sigma^2 + (\mathrm{E}[X])^2 = \sigma^2 + (\mu)^2 = \mu^2 + \sigma^2
\end{aligned}$$

which agrees with the first calculation. △

1.2 Sample mean and variance

Definition 1.2.1. Given the random variables $X_1, X_2, \ldots, X_n$, the **sample mean** $\overline{X}$ is the statistic defined as the arithmetic mean of these variables,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.7}$$

■

Definition 1.2.2. Given the random variables $X_1, X_2, \ldots, X_n$, the **sample variance** S^2 is defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \tag{1.8}$$

The **sample standard deviation** is the statistic defined by $S = \sqrt{S^2}$. ■

Remark 1.2.3. Both $\overline{X}$ and S^2 are functions of the sample $X_1, X_2, \ldots, X_n$, and so should be written as $\overline{X}(X_1, \ldots, X_n)$ and $S^2(X_1, \ldots, X_n)$. However, when it is clear that these statistics relate to a particular sample, as is almost always the case, we shall simply write them as $\overline{X}$ and S^2. □

Remark 1.2.4. One may wonder why the sample variance is defined with a factor of $\frac{1}{n-1}$, instead of with a factor of $\frac{1}{n}$. This will be discussed in Section 1.2.1. □

Exercise 1.2.5. Given the definition of S^2 in Definition 1.2.2, show that

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2. \tag{1.9}$$

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) \\
&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\left(n\overline{X}\right) + n\overline{X}^2 \\
\Rightarrow (n-1)S^2 &= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2
\end{aligned}$$

as required.

△

1.2.1 Expected value of sample mean and variance

For the expectations of $\overline{X}$ and S^2, we have the following result.

Proposition 1.2.6. Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independently sampled from a distribution F_X that has mean μ and finite variance σ^2. Then

1. $\mathrm{E}\left(\overline{X}\right) = \mu$,
2. $\mathrm{Var}\left(\overline{X}\right) = \dfrac{\sigma^2}{n}$,
3. $\mathrm{E}\left(S^2\right) = \sigma^2$.

♦

Remark 1.2.7. Proposition 1.2.6 above provides one reason for choosing to define the sample variance as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, rather than as $S_b^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. The proposition shows, under the assumption that the random variables $X_1, X_2, \ldots, X_n$ are sampled i.i.d with mean μ and variance σ^2, that

$$\mathrm{E}\left(S_b^2\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) = \mathrm{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\mathrm{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2.$$

This shows that that S^2 is an **unbiased estimator** of σ^2. Similarly, $\overline{X}$ is an unbiased estimator of μ. The concept of biased and unbiased estimators will be discussed in Section 1.5.4. □

The proof of Proposition 1.2.6 is left as an exercise:

Proof of Proposition 1.2.6 (Part 1).

Using the linearity of the expectation,

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

□

Proof of Proposition 1.2.6 (Parts 2 and 3).

Part 2: Using the fact that the X_i are independent, and properties of the variance,

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

Part 3: Using the alternative expression for S^2 in Equation (1.9), from Exercise 1.2.5,

$$\mathrm{E}\left(S^2\right) = \mathrm{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]\right) = \frac{1}{n-1}\left[\mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\right]$$

We now use Exercise 1.1.10 twice. First, since each X_i is a random variable with mean μ and variance σ^2, $\mathrm{E}\left(X_i^2\right) = \mu^2 + \sigma^2$. Second, Parts 1 and 2 have just shown that $\overline{X}$ is a random variable with mean μ and variance $\frac{\sigma^2}{n}$, and so $\mathrm{E}(\overline{X}^2) = \mu^2 + \frac{\sigma^2}{n}$. Now, again using the linearity of the expectation,

$$\begin{aligned}
\mathrm{E}\left(S^2\right) &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathrm{E}\left(X_i^2\right) - n\mathrm{E}\left(\overline{X}^2\right)\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\
&= \frac{1}{n-1}\left[n\left(\mu^2+\sigma^2\right) - \left(n\mu^2+\sigma^2\right)\right] = \frac{1}{n-1}\left[(n-1)\,\sigma^2\right] \\
\Rightarrow \mathrm{E}\left(S^2\right) &= \sigma^2,
\end{aligned}$$

which completes the proof of all three parts.

□

Remark 1.2.8. Note that while Proposition 1.2.6 specifies that the random variables have mean μ and variance σ^2, there is no mention of the actual distribution that these random variables follow. Specifically, the random variables are **not necessarily** normally distributed; they could be Bernouilli, Gamma, etc. as long as the mean and (finite) variance of the distribution is known. □

1.2.2 The sample mean and variance for a set of observations

For a set (or collection) of observations $x_1, x_2, \ldots, x_n$, the sample variance s^2 is defined by

Definition 1.2.9. For real values $x_1, x_2, \ldots, x_n$, the sample variance s^2 is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \qquad \text{where} \qquad \overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j. \tag{1.10}$$

■

There is a result for the sample variance s^2 that is very similar to Theorem 1.1.1 for the (population) variance for the a random variable X which is described in the following exercise:

Exercise 1.2.10. Given a sample of observations $x_1, x_2, \ldots, x_n$, with the sample mean $\overline{x}$ defined in Equation (1.10), prove that

$$\min_a \left[\sum_{i=1}^{n}(x_i - a)^2\right] = \sum_{i=1}^{n}(x_i - \overline{x})^2 = (n-1)\,s^2. \tag{1.11}$$

We use a similar trick to that in the proof of Theorem 1.1.1. For any given a,

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}\left[(x_i - \overline{x}) + (\overline{x} - a)\right]^2 = \sum_{i=1}^{n}\left[(x_i - \overline{x})^2 + 2(x_i - \overline{x})(\overline{x} - a) + (\overline{x} - a)^2\right] \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - a)\sum_{i=1}^{n}(x_i - \overline{x}) + \sum_{i=1}^{n}(\overline{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - a)\cdot 0 + n(\overline{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - a)^2,
\end{aligned}$$

where in the third line we used $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\overline{x} = n\overline{x} - n\overline{x} = 0$.
Since $n(\overline{x} - a)^2 \geq 0$,

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \overline{x})^2,$$

with equality only when $\overline{x} = a$, which proves the result.

△

1.3 The Markov and Chebyshev Inequalities

In this section we derive two inequalities for random variables, each of which has very few assumptions. These inequalities are applicable to a wide range of distributions, and could even be used in cases where the distribution is not known; such results are sometimes called **distribution free**. Both inequalities can be used to make statements about how far a random variable deviates from it mean. In Example 1.3.6 we shall see how Chebyshev's inequality can be used to forecast the results of an election. The statement and proofs of these inequalities are from [1].

Theorem 1.3.1 (Markov's Inequality)**.** If a random variable X can only take nonnegative values, then

$$\mathrm{P}\left(X \geq a\right) \leq \frac{\mathrm{E}\left(X\right)}{a}, \qquad \text{for all } a > 0. \tag{1.12}$$

♦

Proof. Fix a positive number $a > 0$, and define the random variable

$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

This definition of Y_a ensures that $Y_a \leq X$ for all values of a and X, and therefore:

$$\mathrm{E}\left(Y_a\right) \leq \mathrm{E}\left(X\right). \tag{1.13}$$

On the other hand, since Y_a is a discrete random variable, one can computes its expectation as

$$\begin{aligned} \mathrm{E}\left(Y_a\right) &= 0 \cdot \mathrm{P}\left(Y_a = 0\right) + a \cdot \mathrm{P}\left(Y_a = a\right) \\ &= 0 \cdot \mathrm{P}\left(X < a\right) + a \cdot \mathrm{P}\left(X \geq a\right). \\ \Rightarrow \mathrm{E}\left(Y_a\right) &= a \cdot \mathrm{P}\left(X \geq a\right). \end{aligned} \tag{1.14}$$

Combining Equations (1.13) and (1.14), one obtains

$$a \cdot \mathrm{P}\left(X \geq a\right) \leq \mathrm{E}\left(X\right), \tag{1.15}$$

from which the inequality in Equation (1.12) follows. □

Remark 1.3.2. Note that this inequality holds for any **nonnegative** random variable X. Find the place in the proof where this nonnegative assumption is used. □

Remark 1.3.3. The proof of Markov's Inequality is fundamental. Note that there are other proofs of this result that appear to be simpler, but these proofs often make assumptions (e.g. that the random variable X is continuous). An advantage of this proof is that it works for any nonnegative random variable X, whether it be discrete, continuous, or otherwise. □

Theorem 1.3.4 (Chebyshev's Inequality)**.** If X is a random variable with mean μ and variance σ^2, then for all $c > 0$,

$$\mathrm{P}\left(|X - \mu| \geq c\right) \leq \frac{\sigma^2}{c^2}. \tag{1.16}$$

♦

Proof.

Although there is no restriction on the values the random variable X can take, the random variable $(X - \mu)^2$ is nonnegative, and applying the Markov Inequality to $(X - \mu)^2$ with $a = c^2$ yields

$$\mathrm{P}\left((X - \mu)^2 \geq c^2\right) \leq \frac{\mathrm{E}[(X - \mu)^2]}{c^2}.$$

One observes that the event $(X - \mu)^2 \geq c^2$ is identical to the event $|X - \mu| \geq c$, and that $\mathrm{E}[(X - \mu)^2] = \mathrm{Var}\left(X\right) = \sigma^2$. Therefore,

$$\mathrm{P}\left(|X - \mu| \geq c\right) = \mathrm{P}\left((X - \mu)^2 \geq c^2\right) \leq \frac{\mathrm{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}.$$

which proves the result.

□

Chebyshev's inequality is often stated in the following equivalent form:

Corollary 1.3.5. If X is a random variable with mean μ and variance σ^2, then for all $k > 0$,

$$\mathrm{P}\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}. \tag{1.17}$$

♦

Proof. In Equation (1.16) let $c = k\sigma$. □

We now have all the tools to see an example of how Chebyshev's inequality can be used to determine the accuracy of an estimate of a parameter. The following example is taken from [1, p. 270].

Example 1.3.6. Suppose that a population is taking part in a vote and an unknown proportion p of the voters supports a particular option, labelled A. Suppose it is possible to interview a sample of n randomly selected voters and record $\widehat{p}$, the proportion of that sample that supports option A. How close can we say $\widehat{p}$ is to p?

Let us label our sample of n voters from 1 to n, and let X_i be the random variable with value $x_i = 1$ if voter i supports option A, and $x_i = 0$ otherwise. By this construction, each $X_i \sim \text{Bern}(p)$, where p is the unknown parameter we wish to estimate, and $\widehat{p} = \overline{x}$. Since each X_i has mean $\text{E}(X_i) = p$ and variance $\text{Var}(X_i) = p(1-p)$, using Proposition 1.2.6, $\text{E}\left(\overline{X}\right) = p$ and $\text{Var}\left(\overline{X}\right) = p(1-p)/n$. Therefore, for any $\epsilon > 0$, Chebyshev's Inequality in Theorem 1.3.4 gives

$$\text{P}\left(\left|\overline{X} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Furthermore, using Corollary 1.1.7, one can remove the unknown p on the right-hand side to obtain

$$\text{P}\left(\left|\overline{X} - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}.$$

As a specific example, taking $\epsilon = 0.1$ and $n = 100$,

$$\text{P}\left(\left|\overline{X} - p\right| \geq 0.1\right) \leq \frac{1}{4 \cdot 100 \cdot (0.1)^2} = 0.25.$$

We can interpret this to mean that when the the sample size of our voters is $n = 100$ then the probability that our estimate of p is incorrect by more than 0.1 is not larger than 0.25. △

Chebyshev's inequality requires knowledge of (or bounds on) the mean and variance of the random variable under consideration. However, what if all that is available are estimates of the mean and variance from a sample? In such cases, there is a sample version of Chebyshev's inequality which is very useful and only requires a slight modification to Equation (1.17):

Theorem 1.3.7 (**Reading material**)**.** Suppose $X_1, X_2, \ldots, X_n$ and X_{n+1} are i.i.d random variables, and define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \qquad Q^2 = \left(\frac{n+1}{n}\right)S^2. \tag{1.18}$$

Then for all $\lambda \geq 1$,

$$\text{P}\left(\left|X_{n+1} - \overline{X}\right| > \lambda Q\right) \leq \frac{1}{\lambda^2} + \frac{1}{n}. \tag{1.19}$$

♦

Remark 1.3.8. The proof of Theorem 1.3.7 is beyond the scope of this course, but can be found in [13], which proves a more general version of Equation (1.19). □

1.4 Estimating the mean of a sample

Suppose one records the observed values $x_1, x_2, \ldots, x_n$ for the random variables $X_1, X_2, \ldots, X_n$, where the random variables are i.i.d. according to some distribution F_X with unknown mean θ. Suppose one wishes to estimate the value of θ. One could reason that the sample mean is a good estimate of θ because Proposition 1.2.6 shows that $\mathrm{E}(\overline{X}) = \theta$. One could then estimate the value of θ by computing the sample mean of the observations, i.e. an **estimate** $\widehat{\theta}$ of the parameter θ is

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1} x_i.$$

Example 1.4.1. To give a specific example of the above, suppose that $n = 10$ and the random variables $X_1, X_2, \ldots, X_n$ following distribution F_X are observed as $x_1, x_2, \ldots, x_{10}$ with values

$$\{169.6, 166.8, 157.1, 181.1, 158.4, 165.6, 166.7, 156.5, 168.1, 165.3\}.$$

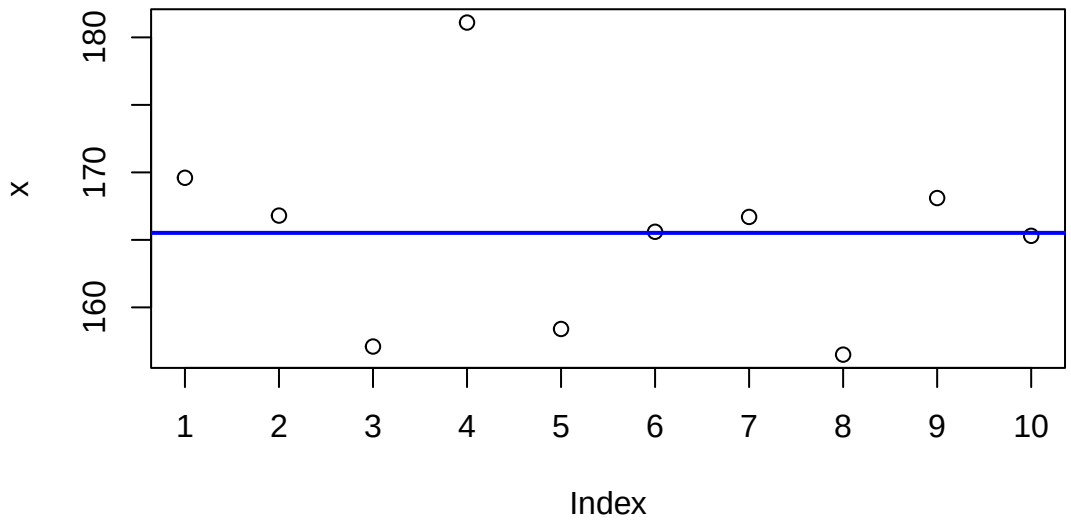
Then

$$\widehat{\theta} = \frac{1}{10}(169.6 + \cdots + 165.3) = 165.52.$$

We plot the data and $\widehat{\theta}$ below. Note that while this estimate $\widehat{\theta}$ may give us an idea for the true value of θ, there is no measure of how close the estimate $\widehat{\theta}$ is to the true value of θ.

△

```
x <- c(169.6, 166.8, 157.1, 181.1, 158.4, 165.6, 166.7, 156.5, 168.1, 165.3)
theta_hat <- mean(x)
plot(x, xaxp=c(1,10,9))   # plot data and force tick marks 1:10
abline(h=theta_hat, col="blue", lwd=2)
```



1.5 Parameter estimation

This section defines two ways of providing estimates of parameters: point estimates and interval estimates. First, we need to define what we mean by a **parameter**.

Definition 1.5.1. In a problem of statisical inference, a characteristic or combination of characteristics that determine the (joint) distribution for the random variable(s) of interest is called a **parameter** of the distribution. ■

Example 1.5.2. Consider the random variable X. Then the mean, $\mu = \mathrm{E}(X)$, is a parameter of the distribution of X. Another parameter is the variance, $\sigma^2 = \mathrm{Var}(X)$, and another is the standard deviation, $\sigma = \sqrt{\mathrm{Var}(X)}$. △

We contrast the definition of a parameter with the definition of a **statistic**.

Definition 1.5.3. Suppose that the observable random variables of interest are $X_1, X_2, \ldots, X_n$. Let r be an arbitrary real-valued function of n random variables. Then the random variable $T = r(X_1, X_2, \ldots, X_n) = r(\mathbf{X})$ is called a **statistic**. ■

Example 1.5.4. Consider the collection of random variables $X_1, X_2, \ldots, X_n$. One example of a statistic is the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. A second example is the maximum of these random variables, $Y = \max\{X_1, \ldots, X_n\}$. A third example is the function r on n variables which has the constant value $r(X_1, \ldots, X_n) = 7$ for all values of $X_1, X_2, \ldots, X_n$. △

1.5.1 Point estimation

The **frequentist** view of statistics is that a parameter has a true value, which is a certain fixed number, but this value is unknown to the experimenter or statistician.

The goal of the statistician is to estimate the true value of the parameter as closely as possible. Suppose there is a collection of observable random variables $X_1, X_2, \ldots, X_n$, all assumed to follow the same distribution F_X, and a parameter of the distribution F_X that one wishes to estimate is denoted θ. One frequentist approach is to determine a statistic $r(X_1, \ldots, X_n)$ that has an expected value close to the value of θ. Then, once the random variables $X_1, X_2, \ldots, X_n$ are observed to be $x_1, x_2, \ldots, x_n$, respectively, the value $\widehat{\theta} = r(x_1, \ldots, x_n)$ is taken to be an estimate of θ. This is an example of a point estimation:

Definition 1.5.5. Given a sample of random variables $X_1, X_2, \ldots, X_n$, a **point estimator** is any function $\widehat{\Theta}(X_1, X_2, \ldots, X_n)$. ■

Remark 1.5.6. Since an estimator is a function of random variables, it is itself a random variable. Note also that any statistic is a point estimator. □

Remark 1.5.7. If it is understood which sample is being used, we shall simply write the estimator as $\widehat{\Theta} = \widehat{\Theta}(X_1, X_2, \ldots, X_n) = \widehat{\Theta}(\mathbf{X})$. □

Remark 1.5.8. To indicate the number of variables being used, sometimes it will be necessary to write $\widehat{\Theta}_n = \widehat{\Theta}(X_1, X_2, \ldots, X_n)$. □

Remark 1.5.9. Throughout these notes, the parameter of interest will often be denoted by θ, while its estimate will be denoted by $\widehat{\theta}$. In other words, $\widehat{\theta}$ is the realisation of the point estimator $\widehat{\Theta}$, once the random variables $X_1, X_2, \ldots, X_n$ have been observed as $x_1, x_2, \ldots, x_n$. □

Finally, the standard error of an estimator is an important concept that will be used later:

Definition 1.5.10. The **standard error** of the estimator $\widehat{\Theta}$ is defined as the square root of its variance, i.e. $\mathrm{SE}_{\widehat{\Theta}} = \sqrt{\mathrm{Var}\left(\widehat{\Theta}\right)}$. ■

1.5.2 Interval estimation

While a point estimate provides a single number $\widehat{\theta}$ that estimates the true value of the parameter θ, it does not provide any information on how close $\widehat{\theta}$ is to θ. Another method for estimating the value of θ is to provide a range of values, usually in the form of an interval, which contains the true value of θ with a high level of probability.

Again, suppose that the random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is observed as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and that random variables in $\mathbf{X}$ each follow the distribution F_X which has a parameter θ.

Definition 1.5.11. An **interval estimate** of a real-valued parameter θ is any pair of functions $L(\mathbf{x})$ and $U(\mathbf{x})$ of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$, for all possible $\mathbf{x} = (x_1, \ldots, x_n)$. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called an **interval estimator**, and if $\mathbf{X} = \mathbf{x}$ is observed then the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made. ■

Remark 1.5.12. Although Definition 1.5.11 specifies a closed interval $[L(\mathbf{x}), U(\mathbf{x})]$, there are situations when an open interval or half-open interval can be used. Similarly, there are situations where $L(\mathbf{x}) = -\infty$ or $U(\mathbf{x}) = \infty$ and then the interval is one-sided; for example, with $L(\mathbf{x}) = -\infty$ one has the inference $\theta \leq U(\mathbf{x})$. □

Definition 1.5.13. For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of a parameter θ, the **coverage probability** of $[L(\mathbf{X}), U(\mathbf{X})]$ is the probability that the random interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the true parameter, θ. In symbols, it is denoted by either $\mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})] | \theta)$ or $\mathrm{P}_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$. ■

Definition 1.5.14. If the interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ is designed so that $L(\mathbf{X}) \leq U(\mathbf{X})$ and

$$\mathrm{P}_\theta(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) \geq 1 - \alpha,$$

for every possible value of θ, and some $\alpha \in (0, 1)$, then we call $[L(\mathbf{X}), U(\mathbf{X})]$ a $1 - \alpha$ **confidence interval**. ■

Remark 1.5.15. Setting the value of α is completely up to the statistician and the analysis under consideration. It is acceptable in some cases to set $\alpha = 0.1$, making $1 - \alpha = 0.9$, while in other cases one may set $\alpha = 0.01$, making $1 - \alpha = 0.99$. However, although there is no special reason for using this value, it is common to set $\alpha = 0.05$, making a $1 - \alpha = 0.95$ confidence interval. □

Remark 1.5.16. Often, it is more common to call a $1 - \alpha$ confidence interval a $100(1 - \alpha)\%$ confidence interval, i.e. the most common type of confidence interval is a 95% confidence interval (when $\alpha = 0.05$). □

Example 1.5.17. Let us return to the data of Example 1.4.1, where

$$\mathbf{x} = \{169.6, 166.8, 157.1, 181.1, 158.4, 165.6, 166.7, 156.5, 168.1, 165.3\}$$

are observations of random variables $\mathbf{X}$ following distribution F_X, which has unknown mean θ. Suppose, however, that the variance of F_X is known to be $\sigma^2 = 27.04$. Now, note that the version of Chebyshev's inequality given in Equation (1.17) can be rewritten (with Y) as

$$\mathrm{P}\left(|Y - \mu| < k\sigma\right) \geq 1 - \frac{1}{k^2} \qquad \Rightarrow \qquad \mathrm{P}\left(Y - k\sigma < \mu < Y + k\sigma\right) \geq 1 - \frac{1}{k^2}.$$

If we set $Y = \overline{X}$, which has $\mathrm{E}(\overline{X}) = \theta$ and $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$ by Proposition 1.2.6, then

$$\mathrm{P}\left(\overline{X} - k\tfrac{\sigma}{\sqrt{n}} < \theta < \overline{X} + k\tfrac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}.$$

Therefore, taking $k = 5$, the interval $(157.30, 173.74)$ contains θ with **confidence** 0.96, since $1 - \frac{1}{25} = 0.96$ and one can compute $157.30 \approx 165.52 - 5 \cdot \sqrt{27.04}/\sqrt{10}$ and similarly $173.74 \approx 165.52 + 5 \cdot \sqrt{27.04}/\sqrt{10}$. △

1.5.3 Real-world example: the Chesapeake and Ohio freight study

In the early 1950s, the Chesapeake and Ohio Railroad Company (C&O) undertook a study to determine the amount of revenue due them on interline, less-than-carload freight shipments. When a freight shipment travels over several railroads, the revenue from the freight charge is appropriately divided among those railroads. A **waybill**, which accompanies each shipment, provides the details on the goods, route and charges of that shipment, and allows the division of revenue to be calculated. However, these calculations in the 1950s were performed by hand, and were thus time consuming and costly. C&O were interested in discovering if this calculation could be accurately determined on the basis of a small sample and thereby saving clerical expense.

One experiment studied the division of revenue of less-than-carload shipments over the Pere Marquette district between C&O and another company, labelled A, over a six-month period. The total number of waybills for that period (22,984) was known, as was the total amount of revenue. The problem was to determine how to divide the revenue between C&O and A.

Using stratified sampling, 2,072 of the 22,984 waybills (roughly 9%) were sampled, and from this sample it was estimated that the total revenue due C&O was $\$64,568$. A second study examined all the waybills and calculated that the revenue due C&O was $\$64,651$ ($\$83$, or approximately 1.3%, more).

However, the first (small) study only cost $\$1,000$, while the second (complete) study cost $\$5,000$!

Given the two analysis options, clearly it does not make financial sense to spend $\$4,000$ to catch an error of less than $\$100$.

1.5.4 Estimators, bias and variance

This section defines a few important concepts regarding estimators.

Definition 1.5.18. The **estimation error** of the estimator $\widehat{\Theta}$ of a parameter θ is defined to be $\widehat{\Theta} - \theta$. ■

Definition 1.5.19. The **bias** of the estimator $\widehat{\Theta}$ of a parameter θ, denoted by $b_\theta(\widehat{\Theta})$, is the expected value of the estimation error:

$$b_\theta(\widehat{\Theta}) = \mathrm{E}[\widehat{\Theta}] - \theta \tag{1.20}$$

■

Remark 1.5.20. Since the parameter θ is assumed to be a constant (but unknown) value, Equation (1.20) follows from $\mathrm{E}[\widehat{\Theta} - \theta] = \mathrm{E}[\widehat{\Theta}] - \theta$. □

Definition 1.5.21. The estimator $\widehat{\Theta}$ of a parameter θ is called **unbiased** if $\mathrm{E}[\widehat{\Theta}] = \theta$, for every possible value of θ. ■

Example 1.5.22. The sample mean $\overline{X}$ of an i.i.d. sample $X_1, X_2, \ldots, X_n$ from a distribution with mean μ is an unbiased estimator of μ since $\mathrm{E}\left(\overline{X}\right) = \mu$ by Proposition 1.2.6. △

Definition 1.5.23. The **mean squared error** of the estimator $\widehat{\Theta}$ of a parameter θ is defined as the quantity $\mathrm{E}[(\widehat{\Theta} - \theta)^2]$. ■

Given the definitions above, we have the following important result:

Theorem 1.5.24. The mean squared error of an estimator $\widehat{\Theta}$ of a parameter θ can be expressed in terms of its bias and variance:

$$\mathrm{E}[(\widehat{\Theta}-\theta)^2] = \left[b_\theta(\widehat{\Theta})\right]^2 + \mathrm{Var}\left(\widehat{\Theta}\right). \tag{1.21}$$

♦

Proof.

For any random variable X, Exercise 1.1.5 gives us $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$. Rearranging, this identity is:

$$\mathrm{E}[X^2] = (\mathrm{E}[X])^2 + \mathrm{Var}(X)$$

Applying this identity to the estimation error $\widehat{\Theta}-\theta$, which is itself a random variable since the $\widehat{\Theta}$ is a random variable and θ is an unknown constant, and using the properties of the expectation and variance, one obtains

$$\begin{aligned}
\mathrm{E}[(\widehat{\Theta}-\theta)^2] &= \left(\mathrm{E}[\widehat{\Theta}-\theta]\right)^2 + \mathrm{Var}\left(\widehat{\Theta}-\theta\right) \\
&= \left(\mathrm{E}[\widehat{\Theta}]-\theta\right)^2 + \mathrm{Var}\left(\widehat{\Theta}\right) \\
\Rightarrow \mathrm{E}[(\widehat{\Theta}-\theta)^2] &= \left[b_\theta(\widehat{\Theta})\right]^2 + \mathrm{Var}\left(\widehat{\Theta}\right)
\end{aligned}$$

as required.

□

The concept of **support** will be used in the next section.

Definition 1.5.25. The support of a function is the set of values in the domain that are not mapped to zero, i.e. the support of the function $f : \mathbb{R} \to \mathbb{R}$ is the set $\{x \in \mathbb{R} \,|\, f(x) \neq 0\}$. ■

1.6 Other measures of central tendency and dispersion

While the mean and the variance are the two most popular statistics for central tendency and dispersion, respectively, there are other statistics that are also important: the mode, the median, the range and the interquartile range.

1.6.1 The mode

Definition 1.6.1. For a random variable X with probability density (or mass) function f_X, the **mode** of the distribution of X is defined as

$$\text{mode}\,(X) = \arg\max_x f_X\,(x) \tag{1.22}$$

where the 'arg max' function is described in Remark 1.6.2. ■

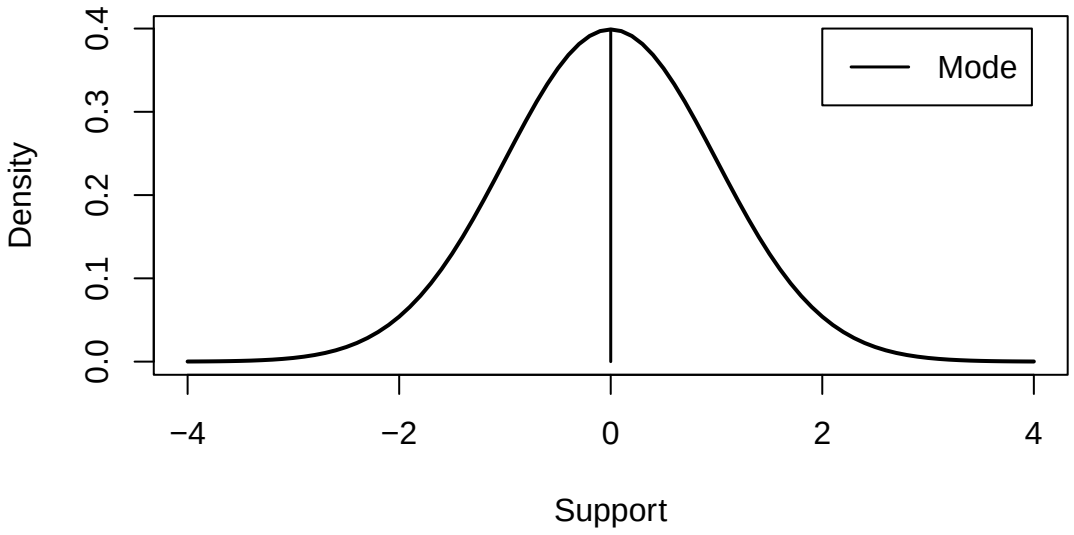
Remark 1.6.2. It is possible that the notation for 'arg max' is not so familiar. Rather than return the maximum value of an expression, it is the function that returns the **argument** that gives the maximum value of that expression. The following example should clarify this definition. □

Example 1.6.3. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = 1 - x^2$. Then

$$\max_x f(x) = 1$$
$$\arg\max_x f(x) = 0,$$

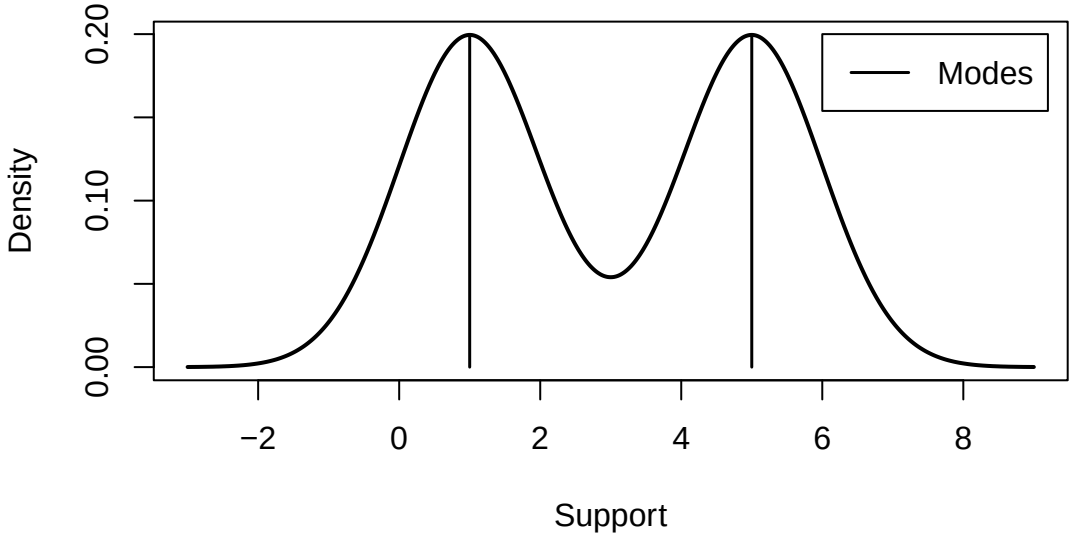
since the maximum of f is $f(0) = 1$. △

Example 1.6.4. A simple example of a mode for a continuous distribution is that the $\text{N}\,(0, 1)$ distribution has a mode at 0. This is shown in the figure below. △



Remark 1.6.5. Most distributions we will encounter have only one mode are are called **unimodal**. However, it is easy to create distributions with more than one mode by considering mixtures of distributions. □

Example 1.6.6. Consider the random variable X, which follows a $\mathrm{N}(1,1)$ distribution with probability 0.5 and a $\mathrm{N}(5,1)$ distribution with probability 0.5. This random variable is a mixture of the $\mathrm{N}(1,1)$ and $\mathrm{N}(5,1)$ normal distributions. The density of this mixture is shown below. Because it has two modes, it is called **bimodal**. △



Remark 1.6.7. The mode can be computed for sample of observations by selecting the value (or values) that occurs most often. It is therefore **more suitable for discrete data**, rather than continuous data, since we expect continuous data values to be distinct. □

Example 1.6.8. If we reconsider the data from Example 1.4.1,

$$\{169.6, 166.8, 157.1, 181.1, 158.4, 165.6, 166.7, 156.5, 168.1, 165.3\},$$

since all the values are different, they are all the sample mode (even though 166.7 and 166.8 are close, they are not equal). △

Example 1.6.9. Suppose that a small survey is conducted among a group of students to determine which is the most popular choice of smartphone, and the response options for the survey are Apple (A), Samsung (S), Huawei (H), Xiaomi (X) and other (O), for any other phone that is not one of these four. The data is collected and the responses are:

$$\{\mathrm{A}, \mathrm{S}, \mathrm{X}, \mathrm{A}, \mathrm{S}, \mathrm{H}, \mathrm{S}, \mathrm{O}, \mathrm{H}, \mathrm{S}, \mathrm{A}\}.$$

Since the value S, for Samsung, occurs the most times (4 times), S is the mode of the sample. △

1.6.2 The median

Definition 1.6.10. For a random variable X, a **median** of the distribution of X is defined as a value m such that

$$\mathrm{P}(X \geq m) \geq \tfrac{1}{2} \qquad \text{and} \qquad \mathrm{P}(X \leq m) \geq \tfrac{1}{2}. \tag{1.23}$$

While there is no standard notation for this, it is possible to write $m = \text{median}(X)$. ■

Example 1.6.11. Define the discrete random variable X to have support $\{1, 2, 3, 4\}$ and distribution specified by

$$\mathrm{P}(X = 1) = 0.1, \qquad \mathrm{P}(X = 2) = 0.2 \qquad \mathrm{P}(X = 3) = 0.3 \qquad \mathrm{P}(X = 4) = 0.4.$$

Then $\mathrm{P}(X \leq 3) = 0.6 \geq \frac{1}{2}$ and $\mathrm{P}(X \geq 3) = 0.7 \geq \frac{1}{2}$, which shows that 3 is a median of this distribution. △

Example 1.6.12. Note that the median is not unique. One can modify Example 1.6.11 by defining the discrete random variable X to have support $\{1, 2, 3, 4\}$ and distribution specified by

$$\mathrm{P}(X = 1) = 0.1, \qquad \mathrm{P}(X = 2) = 0.4 \qquad \mathrm{P}(X = 3) = 0.3 \qquad \mathrm{P}(X = 4) = 0.2.$$

Then $\mathrm{P}(X \leq 2) = 0.5 \geq \frac{1}{2}$ and $\mathrm{P}(X \geq 2) = 0.9 \geq \frac{1}{2}$. However, we also have $\mathrm{P}(X \leq 3) = 0.8 \geq \frac{1}{2}$, and $\mathrm{P}(X \geq 3) = 0.5 \geq \frac{1}{2}$. In fact, every value m in the interval $2 \leq m \leq 3$ is a median of this distribution. In such cases, it is common to choose the median to be midpoint of the interval, i.e. $m = 2.5$. △

The following result characterises the median as value that minimises the expected absolute deviation of a random variable; compare this result to Theorem 1.1.1.

Theorem 1.6.13. Suppose that m is a median of the distribution for the random variable X. Then, for any real value a,

$$\min_a \mathrm{E}(|X - a|) = \mathrm{E}(|X - m|).$$

♦

Exercise 1.6.14. Prove Theorem 1.6.13. △

See the solution to Question 2 on Problem Sheet 9.

1.6.3 The sample median

Just as there is a sample version of the mean, there is a sample version of the median.

Definition 1.6.15. Given a sample of observations $x_1, x_2, \ldots, x_n$, the **sample median** m is defined as

$$m = \begin{cases} x_{([n+1]/2)}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & \text{if } n \text{ is even,} \end{cases} \tag{1.24}$$

where $\left\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right\} = \{x_1, x_2, \ldots, x_n\}$ and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Note that $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are known as the **order statistics**. Note that when n is even, any value in the interval $\left[x_{(n/2)}, x_{(n/2+1)}\right]$ is a sample median. ■

Example 1.6.16. Suppose a group of students write a test which is marked out of 20 and their grades are recorded as

$$\{14, 16, 12, 19, 11, 15, 8, 18, 12, 16\}.$$

To compute the median, we first sort the grades $\{8, 11, 12, 12, 14, 15, 16, 16, 18, 19\}$. Then, since there are 10 grades, and since 14 is the 5th smallest grade and 15 is the 6th smallest grade, following Definition 1.6.15, any value in the interval $[14, 15]$ is considered a median. We could follow Equation (1.24) and declare 14.5 to be **the** sample median, although 14, 14.5, and 15 are all sample medians. The sample mean can be computed as 14.2 (and in fact 14.2 is also a sample median, since 14.2 is in the interval $[14, 15]$). The sample modes are 12 and 16. △

Remark 1.6.17. One interesting feature of the median is that it only requires the data to be **ordinal**, i.e. data that can be ordered. □

Example 1.6.18. Suppose we survey the same ten students from Example 1.6.9 and ask them how satisfied they are with their phones. The possible responses are, in increasing order of satisfaction: 'very unsatisfied' (VU), 'unsatisfied' (U), 'neither satisfied nor unsatisfied' (N), 'satisfied' (S) and 'very satisfied' (VS). Suppose the data is recorded as

$$\{\text{VS, U, S, S, N, VU, S, N, VS, U}\}.$$

Using the ordering VU < U < N < S < VS, we can reorder the data to

$$\{\text{VU, U, U, N, N, S, S, S, VS, VS}\},$$

and so the median is between N and S, i.e. between 'neither satisfied nor unsatisfied' and 'satisfied', inclusive. (Note how there is no way to average these two ordinal quantities!) △

When we have interval-valued data, the following is the analogue of Exercise 1.2.10:

Proposition 1.6.19. Given a sample of observations $x_1, x_2, \ldots, x_n$, with sample median m. Then, for any real value a,

$$\min_a \left(\sum_{i=1}^{n} |x_i - a| \right) = \sum_{i=1}^{n} |x_i - m| \,. \qquad \blacklozenge$$

Exercise 1.6.20. Prove Proposition 1.6.19. △

This is essentially an algebraic result, and the proof is in Section A.2 in the appendix.

1.6.4 The median vs the mean

A natural question to ask is: are there any advantages to using the median over the mean as a statistic? The median is often used as a measure of central tendency when the data may contain very extreme values which may skew the sample mean. One area where the median is often used is to summarise the reported earnings of a population (of a country, of a company, or even of a profession).

Example 1.6.21. Suppose nine statisticians report their salaries or earnings for the past year, and the data is recorded as

$$\{£52,000, £85,000, £45,000, £24,000, £120,000, £69,000, £71,000, £10,000,000, £37,000\}.$$

Since there are nine values, the median can be computed to be the 5th smallest value, which is $£69,000$. However, when one computes the mean, one obtains a value of $£1,167,000$, which is over one million pounds more! The reason for this is that the eighth respondent reported earnings of $£10,000,000$ last year, which skewed the mean upwards. From this example, if one asks for the representative annual salary of a statistician, it may be more sensible to report the median earnings rather than the mean earnings.

Code for computing the median and mean in R are given below. △

```
x <- c(52, 85, 45, 24, 120, 69, 71, 10000, 37) * 1000
cat("mean is: ", mean(x), "\n", sep="")
#> mean is: 1167000
cat("median is: ", median(x), "\n", sep="")
#> median is: 69000
```

1.6.5 The range

Definition 1.6.22. Given a sample of interval-valued data $x_1, x_2, \ldots, x_n$, the **range** is a measure of dispersion defined as the length of the smallest interval which contains all the

data, and can be computed as $\mathrm{R} = x_{(n)} - x_{(1)}$, where $x_{(1)}$ is the smallest value and $x_{(n)}$ is the largest value of the sample. ■

Example 1.6.23. For the data $\{3.2, 1.7, 5.4, 2.8, 4.1\}$, the range is $\mathrm{R} = 5.4 - 1.7 = 3.7$. △

1.6.6 Interquartile range

Definition 1.6.24. Let the cumulative distribution function for the random variable X be denoted by F_X. Then the function $F_X^{-1} : (0,1) \to \mathbb{R}$ is called the **quantile function** for the distribution X, for all $p \in (0,1)$ $F_X^{-1}(p)$ is defined to be the smallest value x such that $F_X(x) \geq p$. ■

Remark 1.6.25. Note that the symbol F_X can be interpreted in two ways: it either refers to the distribution of X, or to the function that is the cumulative distribution function of X. In this latter case, for some value x, $F_X(x) = \mathrm{P}(X \leq x)$. While this might seem to be an abuse of notation, the cumulative distribution function completely specifies a probability distribution. □

Remark 1.6.26. Given a quantile function F_X^{-1} for a random variable X, one can define the median to be $m = F_X^{-1}(0.5)$. □

Definition 1.6.27. Given a quantile function F_X^{-1} for a random variable X, the **lower quartile** is defined as $q_{0.25} = F_X^{-1}(0.25)$ while the **upper quartile** is defined as $q_{0.75} = F_X^{-1}(0.75)$ ■

Definition 1.6.28. Given a quantile function F_X^{-1} for a random variable X, the **interquartile range** is defined as $\mathrm{IQR} = F_X^{-1}(0.75) - F_X^{-1}(0.25)$. ■

Remark 1.6.29. When dealing with samples, one computes the upper and lower quartiles of the sample following similar logic to that for computing the sample median; let's briefly revisit the computation for the sample median. For a sample of size n, denoted $x_1, x_2, \ldots, x_n$, one computes the index $i_m = \frac{n+1}{2}$. If this value is an integer, then the sample median is simply $x_{(i_m)}$, where $x_{(i_m)}$ is the i_mth order statistic, or equivalently, the i_mth smallest element in the sample. If i_m is not an integer, but a half-integer, then the sample median is simply the average of $x_{(i_m-0.5)}$ and $x_{(i_m+0.5)}$. For example, if $i_m = 3.5$, then the sample median is the average of $x_{(3)}$ and $x_{(4)}$.

Now, for the lower quartile $q_{0.25}$, one computes the index $i_{0.25} = \frac{1}{2}(\lfloor i_m \rfloor + 1) = \frac{1}{2}\left(\left\lfloor \frac{n+1}{2} \right\rfloor + 1\right)$. If this is an integer, the lower quatile is $x_{(i_{0.25})}$, otherwise it is the average of $x_{(i_{0.25}-0.5)}$ and $x_{(i_{0.25}+0.5)}$. The upper quartile is defined by using the index $i_{0.75} = n - i_{0.25} + 1$. □

Example 1.6.30. Suppose the sample is $\{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, i.e. 9 elements. In this case, $i_m = 5$, $i_{0.25} = 3$ and $i_{0.75} = 7$. So, $m = 10$, $q_{0.25} = 6$ and $q_{0.75} = 14$. △

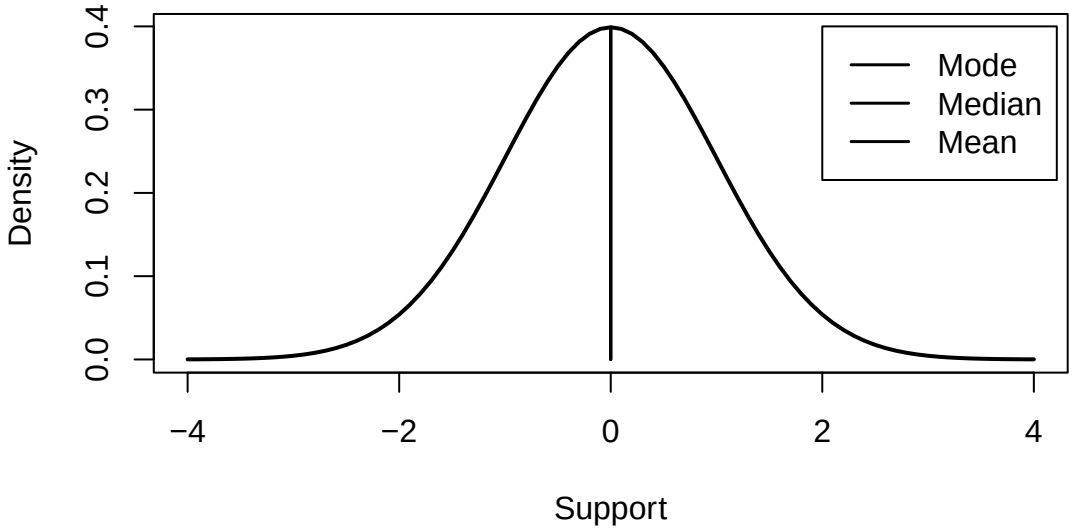
Example 1.6.31. Suppose the sample is $\{2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24\}$, i.e. 12 elements. In this case, $i_m = 6.5$, so the sample median is $\frac{1}{2}(12 + 14) = 13$. Since $\lfloor i_m \rfloor = 6$, $i_{0.25} = 3.5$ and $i_{0.75} = 9.5$. Therefore, $q_{0.25} = \frac{1}{2}(6 + 8) = 7$ and $q_{0.75} = \frac{1}{2}(18 + 20) = 19$. △

Remark 1.6.32. Note that there exist alternative definitions for the lower and upper quartiles that will give slightly different results. However, the definitions we give here are widely accepted and are the definitions used in the R programming language. □

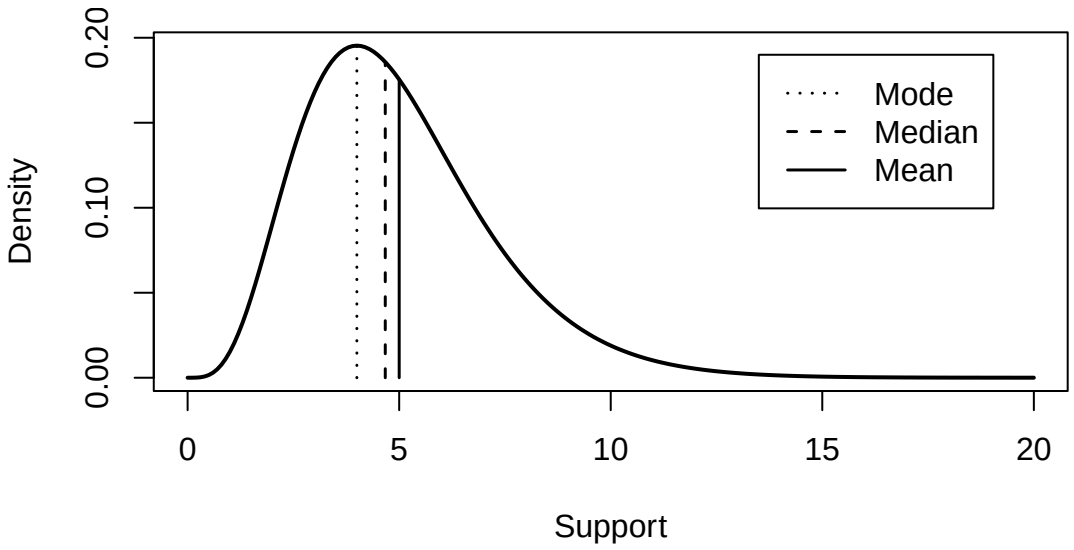
Remark 1.6.33. Note also that there is a difference between the the lower quartile of a random variable and the lower quartile of a sample. For a random variable, the lower quartile is precisely defined as $q_{0.25} = F_X^{-1}(0.25)$. However, as remarked above, for finite samples there is more than one way to compute the lower quartile. □

1.6.7 Comparing the mean, mode and median

For a normal distribution, $\mathrm{N}(\mu, \sigma^2)$, the mean, mode and median are all the same value μ. The probability density function of the standard normal distribution is shown in the figure below.



The probability density function of the $\Gamma(5, 1)$ distribution (where $k = 5$ is the shape and $\theta = 1$ is the scale) is shown in the figure below with the mode, median and mean illustrated by separate lines.



Chapter 2

Exploratory Data Analysis

This chapter described the different types of data we may encounter as well as several ways for visualising these data.

2.1 Types of data

There are two main types of data: **discrete** data and **continuous** data. Discrete data are measurements that only take on certain values that cannot be arbitrarily subdivided. Continuous data are measurements, on the other hand, can be measured to an arbitrary precision. The table below presents a few examples of each.

Beyond the simple division of data into discrete or continuous, there is a taxonomy (or classification) of data into four main types:

- **Nominal** data (also called categorical data)
- **Ordinal** data
- **Interval** data
- **Ratio** data

Nominal data consists of data which are recorded in two or more categories. For example, a fruit stall may sell apples, oranges and lemons, and a record of a day's sales may be {Apple, Orange, Apple, Lemon}. Note how these categories cannot be subdivided, or compared to each other in terms of relative degree (size); an apple is not necessarily greater than (or less than) an orange. The only possible comparison between two data points is equality, and two data points are either 'equal' or 'not equal'.

Ordinal data is discrete data on which there is a natural ordering. For example, in a satisfaction survey, the three categories {Dissatisfied, Neutral, Satisfied} have the natural ordering Dissatisfied $\leq$ Neutral $\leq$ Satisfied. However, there is no way to quantify the relative degree of difference between them; for example, the difference in satisfication between 'Dissatisfied' and 'Satisfied' is not necessarily twice the difference between 'Dissatisfied' and 'Neutral'.

Interval data is continuous data which has an arbitrary zero point, and so the ratio of two values cannot be directly interpreted. While this description may seem complicated, it is easily illustrated by the classic example of the measurement of temperature in degrees Celsius: the point 0°C is the temperature at which water freezes, while 100°C is the temparature at which water boils, and these two reference points provide a way to measure a unit of 1°C. Now, while the data is continuous, and between 10°C and 11°C we may measure a temperature to an arbitrary precision, e.g. 10.27°C, it does not really make sense to compare ratios of temperatures in degrees Celsius, and say '10°C is twice as hot as 5°C'.

Ratio data is continuous data for which has a meaningful zero value and therefore it is meaningful to compare ratios of these values. Measurements such as mass or length are on the ratio scale, and it makes sense to say, for example, that a laptop with mass 2.4 kg is twice as heavy as a laptop with mass 1.2 kg.

Remark 2.1.1. Note that if we measured temperature in degrees Kelvin, since 0 Kelvin is an absolute zero, then it would be possible to say 200 Kelvin is twice 100 Kelvin. □

Table 2.1: Examples of data for the four data types

Data type	Examples
Nominal	{UK, France, Germany, China}; {Apple, Microsoft, Dell}
Ordinal	{Disagree, Neutral, Agree}; Shoe sizes {3, 3.5, 4, . . . , 11, 11.5}
Interval	Temperature in Celsius; Time since start of calendar year
Ratio	Mass; length; duration; sales price

Remark 2.1.2. In this course we shall mainly focus on continuous data that is of ratio type, and although there may be occasions where we look at nominal data, we shall not often consider ordinal or interval data. □

It is also worth noting which appropriate measure of central tendency to use for each data type. For example, the mode is the appropriate measure for nominal data; since the categories have no natural ordering, one cannot even compute the median or mean of nominal data.

Table 2.2: The measurements of central tendency for the four data types with the appropriate statistic highlighted in bold.

Nominal	Ordinal	Interval	Ratio
Mode	Mode	Mode	Mode
	Median	Median	Median
		Mean	**Mean**

Remark 2.1.3. Note that 'data' is plural, and refers to multiple measurements. The singular is 'datum'. Therefore, grammatically we should always refer to 'these data', and avoid saying 'this data'. Often, to avoid this distinction, it is convenient to simply write 'the data'. □

2.2 The bar chart

Let's generate some data in R representing the mobile phones belonging to a group of people in a particular venue.

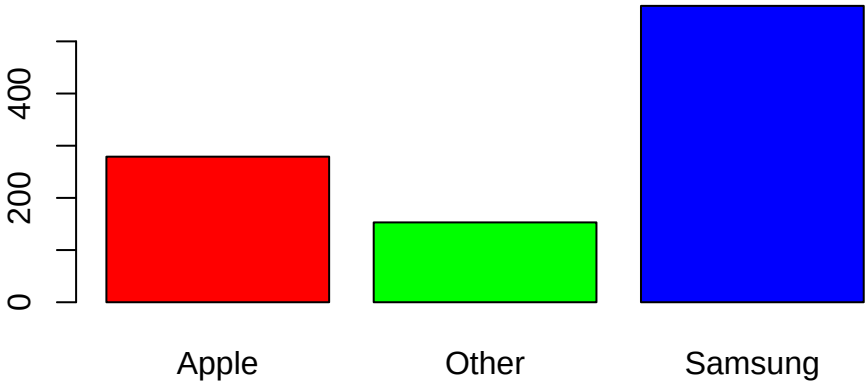
```
set.seed(1)
brands <- c("Apple", "Samsung", "Other")
prob <- c(0.30, 0.55, 0.15)
phone_counts <- sample(x=brands, size=1000, replace=TRUE, prob=prob)

# print the first 20 entries of the 1000 data points
head(phone_counts, n=20)
#>  [1] "Samsung" "Samsung" "Apple"   "Other"   "Samsung"
#>  [6] "Other"   "Other"   "Apple"   "Apple"   "Samsung"
#> [11] "Samsung" "Samsung" "Apple"   "Samsung" "Apple"
#> [16] "Samsung" "Apple"   "Other"   "Samsung" "Apple"
```

This data is of the nominal type, and can be quickly counted using the `table` function in R:

```
t <- table(phone_counts)
print(t)
#> phone_counts
#>   Apple   Other Samsung
#>     279     153     568
```

One useful visualisation for discrete data is a **bar chart**:

```
# R uses the command `barplot` rather than `barchart` with a table:
barplot(t, col=c("red", "green", "blue"))
```



This provides a nice visual representation of the counts of the nominal data.

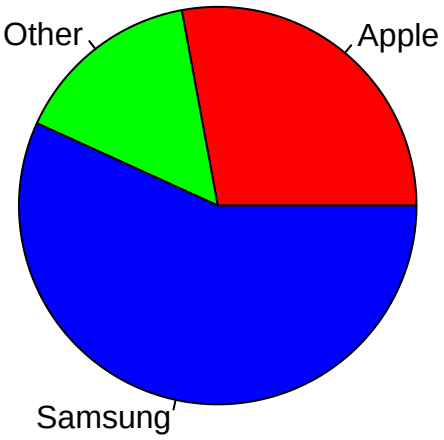
2.3 The pie chart

A variant of the bar chart for visualising discrete data is the **pie chart** which represents the data as proportions of a disc. We generate the same data again (the mobile phones belonging to a particular group of people):

```
# generate the data
set.seed(1)
brands <- c("Apple", "Samsung", "Other")
prob <- c(0.30, 0.55, 0.15)
phone_counts <- sample(x=brands, size=1000, replace=TRUE, prob=prob)

# create the table
t <- table(phone_counts)
print(t)
#> phone_counts
#>   Apple   Other Samsung
#>     279     153     568
```

And now plot the pie chart using the `pie` function in R:

```
# the `rainbow` function creates a vector of colours
pie( t, col=rainbow(length(t)) )
```



The pie chart makes it easy to see that the more than half of the phones are Samsung phones, while about a third are Apple phones.

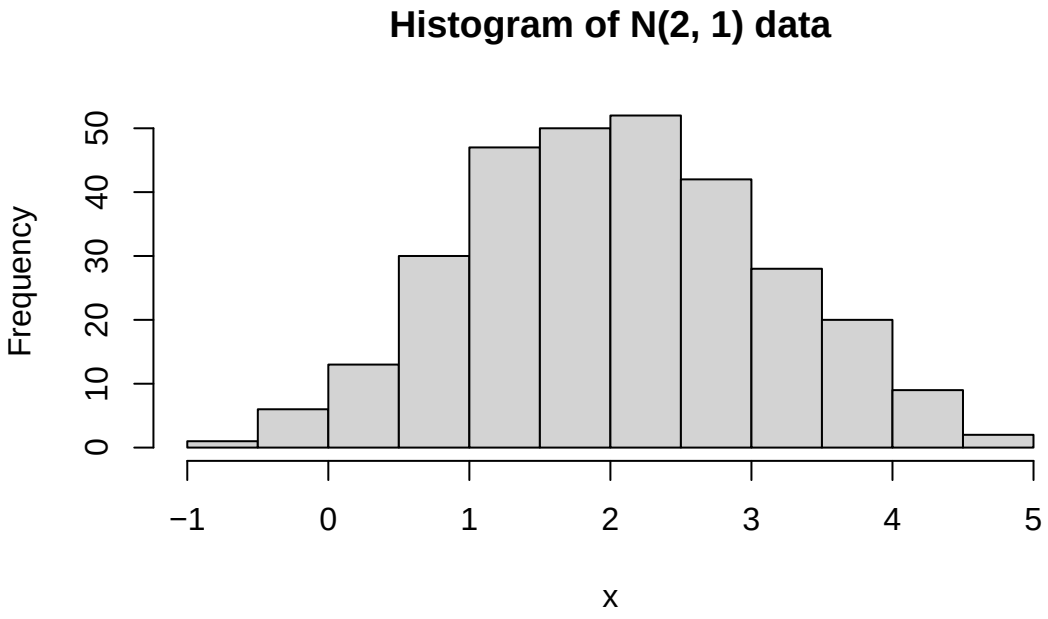
2.4 The histogram

For continuous data, where data is measured to arbitrary precision and therefore almost every value is different, a bar chart is not suitable (the bar plot of n values would be n bars of height 1, one for each different value).

However, a natural idea is to count values that are 'close together' as belonging to one category, and then let the x-axis indicate the value of the categories/bars. This is essentially the definition of a **histogram**.

As an example, we generate $n = 300$ values following a $\mathrm{N}(2, 1)$ distribution.

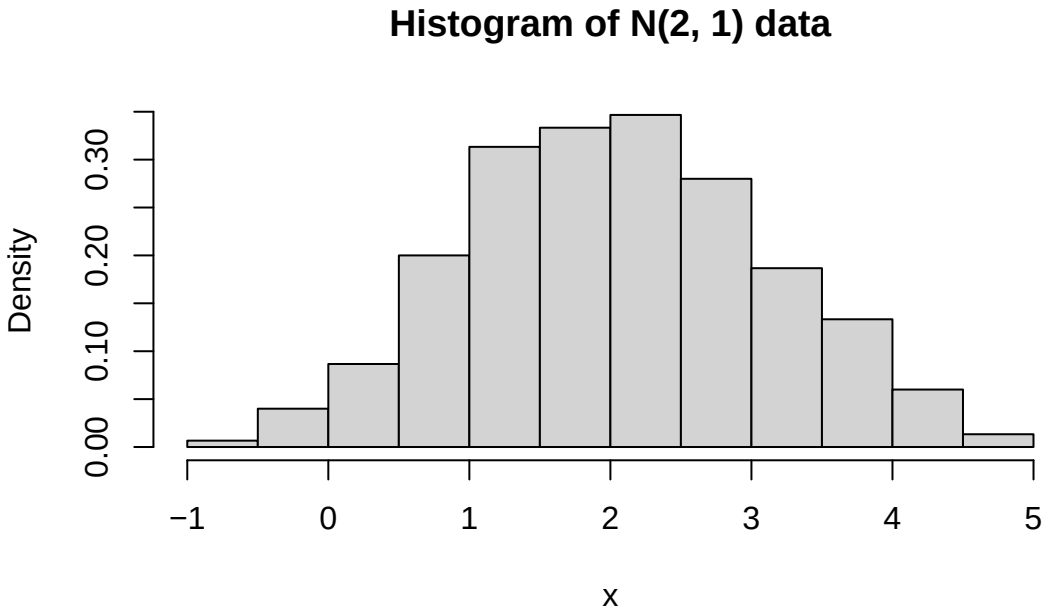
```
# generate the data, 300 values following N(2, 1) distribution
set.seed(2)
x <- rnorm(300, mean=2, sd=1)

# show the first 10 values
print(head(x, n=10))
#>  [1] 1.10309 2.18485 3.58785 0.86962 1.91975 2.13242 2.70795
#>  [8] 1.76030 3.98447 1.86121

# plot the histogram using the `hist` function
# note that setting the number of breaks (bars) is optional
hist(x, main="Histogram of N(2, 1) data")
```



Notice how most of the data is centred around the value 2, which is the mean, and how almost all of the data is within 3 units (or three standard deviations) of the mean.

Note also that this histogram displays the frequency, or counts of the data on the y-axis. One can use the option `freq=FALSE` to plot the histogram with the y-axis scaled so that the area enclosed by the histogram is 1; in other words, the histogram represents the empirical probability density function of the data. Indeed, a histogram is very useful for visualising the distribution of a (univariate) data set with continuous values.

```
# now using the option `freq=FALSE` which scales the y-axis to a density
hist(x, main="Histogram of N(2, 1) data", freq=FALSE)
```



One may wonder how the number of bars in the histogram is chosen. In fact, there are several algorithms for deciding how many bars (or breaks) to plot. For more information look at the R help by typing the command `?hist` in the R console. However, it is possible to force a certain number of breaks to be used in the histogram; however if too many or too few breaks are specified, the histogram will look quite different[1].

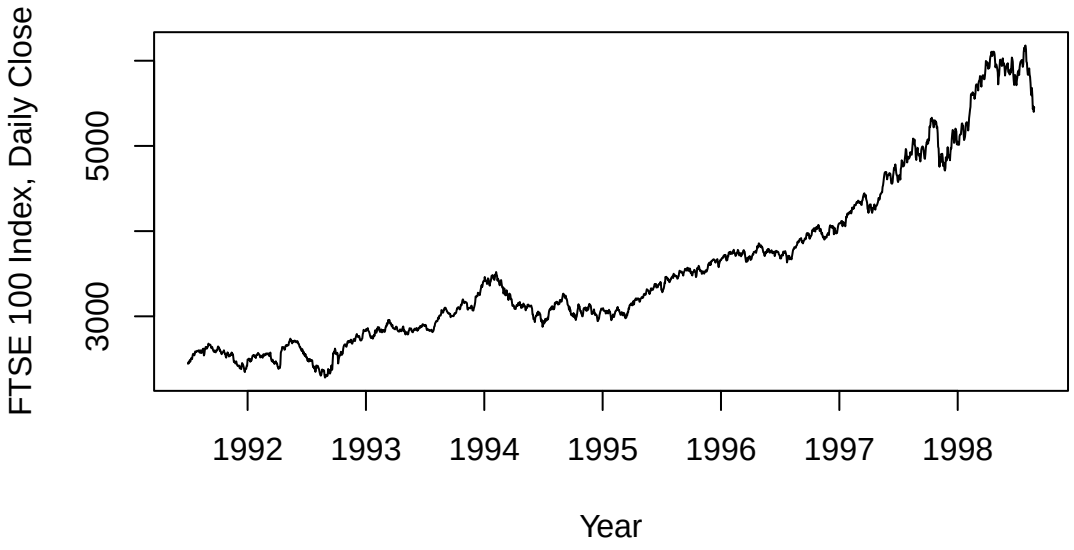
[1]Note that sometimes R will refuse to plot the histogram with a specified number of bars, if the number of bars specified is too high or too low.

2.5 The line plot

There are situations when a set of observations is recorded in a particular order; such sequences of data are called **time series**, because the data are often a set of measurements which are recorded at certain times. In these situations, it is important to visualise these measurements in the relevant order and we plot the data so that the x-axis is time, and the y-axis is the value of the measurement. Although the measurements are recorded at discrete times, in order to indicate that the underlying process does still exist between the discrete measurements, we often choose to join up the individual points with line segments, creating a **line plot**.

Perhaps the most familiar example of a line plot (and a time series) is the share price of a company. As an example, let us consider a financial times series data set that gives the value of the daily close of the Finanical Times Stock Exchange 100 Index (FTSE) for a period from mid-1991 to mid-1999; in total there are 1860 values. The FTSE index is a statistic that is a weighted average of the value of the 100 most valueable companies on the London Stock Exchange. The data is contained in the `EuStockMarkets` dataset, which is included as part of an R installation.

```
# For more info about the data, type `?EuStockMarkets` in the R console
# FTSE is the 4th column
FTSE <- EuStockMarkets[,4]
# the data set of class 'mts', which is means 'multivariate time series',
# there is a special function called 'time' to get the timestamps
year <- time(EuStockMarkets)

# The command `plot` with the option `type='l'` creates a line plot
ylab <- "FTSE 100 Index, Daily Close"
plot(x=year, y=FTSE, type='l', xlab="Year", ylab=ylab)
```



2.6 Plotting distribution functions

There are times when we wish to plot the probability density functions or cumulative distribution functions for random variables. This is not strictly exploratory data analysis, but there may be times when we wish to overlay a plot of a histogram of the data with a probability density function. For example, see Section A.6.

Every distribution in R has four functions associated with the distribution, starting with `d`, `p`, `q` and `r`. We shall look at the example of the normal distribution; further information can be found by typing `?rnorm` in the R terminal, which includes information on how to set parameter values (e.g. the mean and variance). The four functions are:

1. `dnorm`: given a value x, this computes the $f(x)$, where f is the probability density function of the normal distribution.

2. `pnorm`: given a value x, this computes the $F(x)$, where F is the cumulative distribution function of the normal distribution.

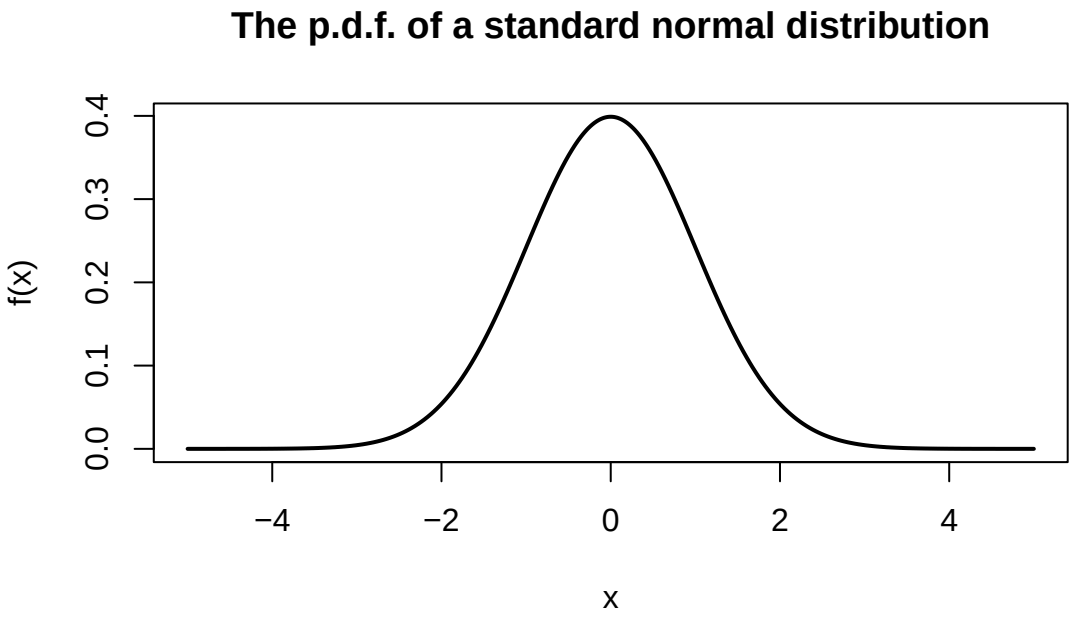
3. `qnorm`: given a value p, this computes $x = F^{-1}(p)$, i.e. the value x such that $F(x) = p$, where F is the cumulative distribution function of the normal distribution. So, `qnorm` is the inverse of `pnorm`.

4. `rnorm`: computes observations of a random variable following a normal distribution.

Remark 2.6.1. For other distributions, such as the gamma distribution, there are corresponding functions, i.e. `dgamma`, `pgamma`, `qgamma` and `rgamma`. □

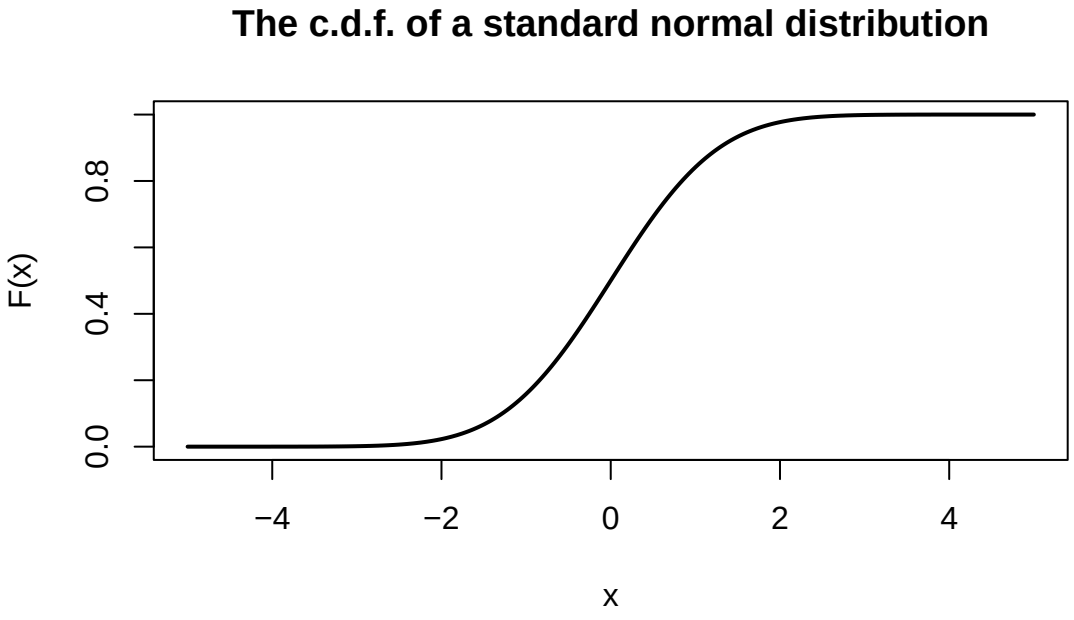
To create a plot of the distribution function, we need to specify the range of values to be evaluated by `dnorm`. This can be done with the `seq` function, which creates linearly-spaced sequences.

```
# We plot the pdf of a normal distribution with mean 0 and std. dev. 1
# on the interval [-5, 5];
# we want points on this interval that are 0.01 units apart.
x <- seq(from=-5, to=5, by=0.01)

# now we compute the pdf values using `dnorm`
y <- dnorm(x, mean=0, sd=1)

# and we plot the pdf, also specifying a title for the plot:
title <- 'The p.d.f. of a standard normal distribution'
plot(x, y, type='l', ylab='f(x)', main=title, lwd=2)
```



We can also plot the cumulative distribution function using the function `pnorm`.

```
# now we compute the cdf values using `pnorm`
# we still use the previously-created vector x
y <- pnorm(x, mean=0, sd=1)

# and we create the plot
title <- 'The c.d.f. of a standard normal distribution'
plot(x, y, type='l', ylab='F(x)', main=title, lwd=2)
```



2.7 The scatterplot

There are times when each data point consists of a collection of measurements. For example, the classic `mtcars` data set in R lists 11 aspects or attributes of 32 cars from 1973-1974.

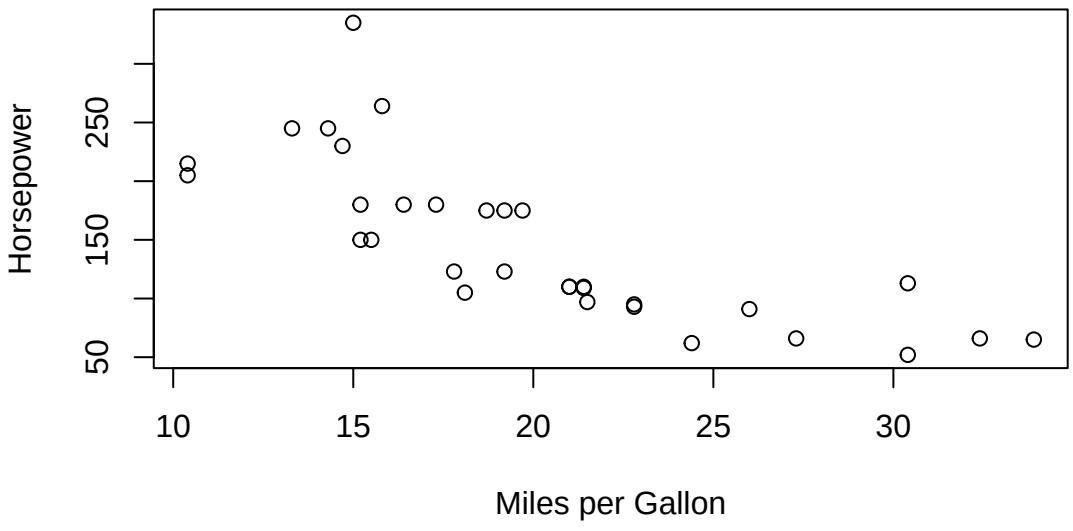
```
# extract all rows, and columns 1, 2, 3, 4, 6 and print first 4 rows
carData <- mtcars[, c(1, 2, 3, 4, 6)]
print(head(carData, n=4))
#>                   mpg cyl disp  hp    wt
#> Mazda RX4        21.0   6  160 110 2.620
#> Mazda RX4 Wag    21.0   6  160 110 2.875
#> Datsun 710       22.8   4  108  93 2.320
#> Hornet 4 Drive   21.4   6  258 110 3.215
```

Here we extracted the following five columns:

- **mpg**: Miles per US gallon, which represents fuel consumption
- **cyl**: Number of cylinders of the car
- **disp**: Displacement in cubic inches (related to size of cylinders)
- **hp**: Gross horsepower (related to power of engine)
- **wt**: Weight of vehicle (in units of 1000 lbs)

Suppose we wish to plot the cars' miles per gallon versus their horsepower. One can create a scatterplot in R using the `plot` function, but this time without using the `type` option. This will plot the data as separate points. Note that the `carData` object is in a format called a **data frame**, and it is very easy to extract individual columns using the '$' operator:

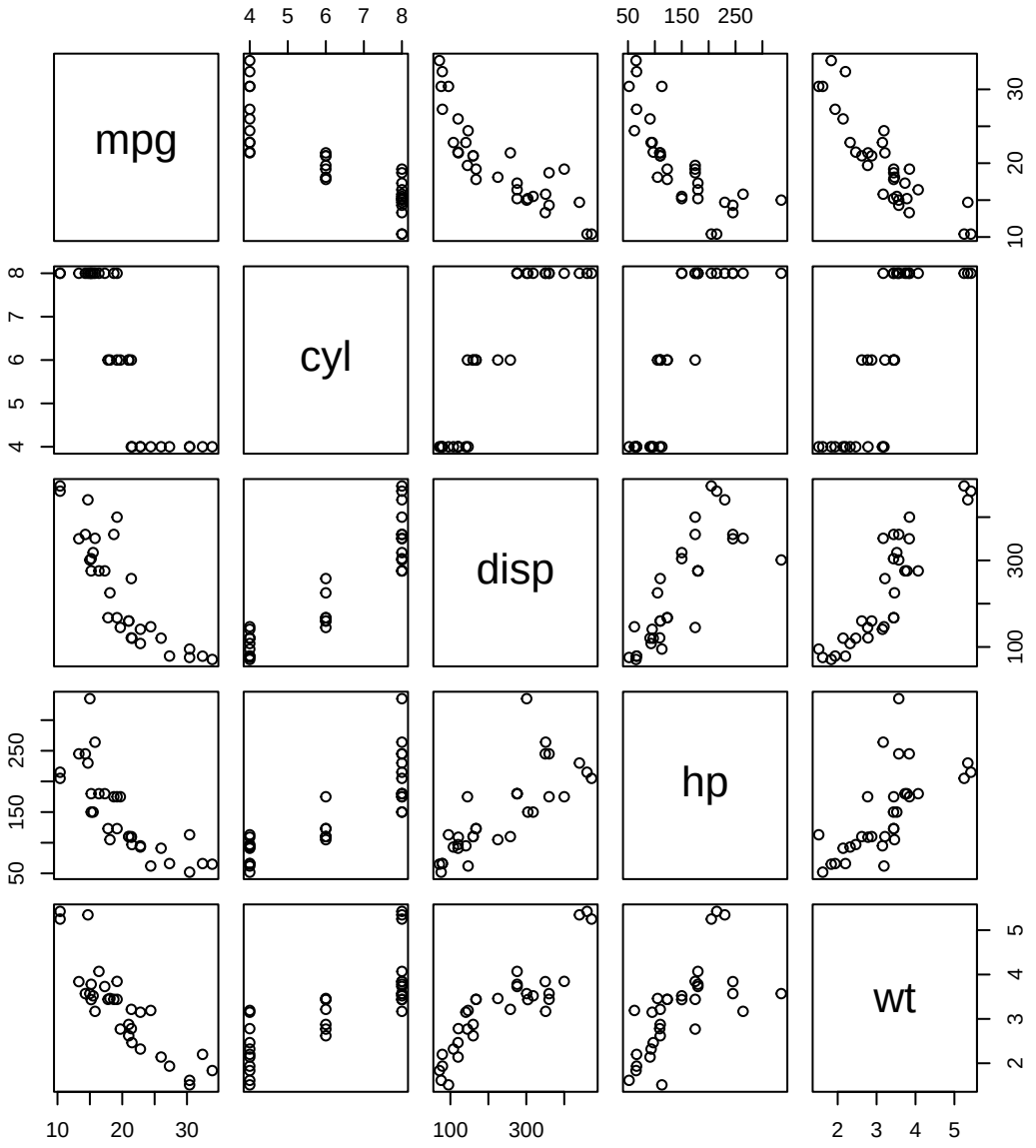
```
# plot hp versus mpg, specifying mpg on the x-axis and hp on the y-axis
plot(x=carData$mpg, y=carData$hp, xlab="Miles per Gallon", ylab="Horsepower")
```



First, note how a scatterplot plots the data as separate points; this is because there is not necessarily any relation between values of the individual attributes (compare to the line plot), and rather the pairs of values need to be compared to each other. From the above scatterplot, we can see that the higher the horsepower, the lower the fuel efficiency.

The `plot` function in R is very powerful, and if one attempts to plot the whole `carsData` data set, it will plot all pairs, as shown below. For example, the figure in the second row, first column plots `mpg` against `cyl`.

```
# plot all five variables of the reduced data set against each other
plot(carData)
```



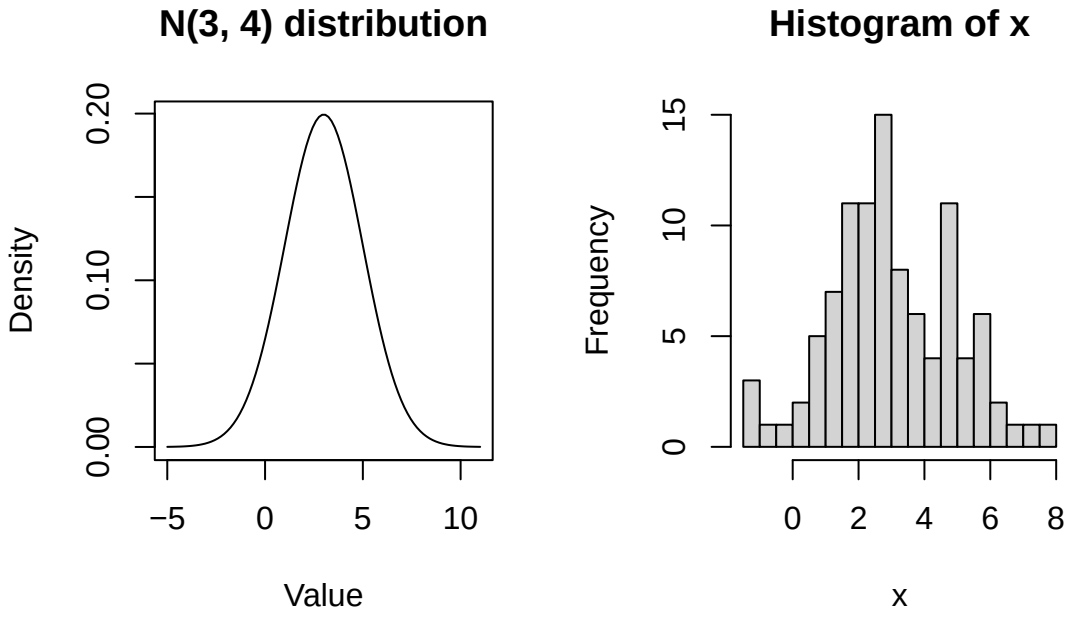
Note how the `cyl` data is discrete, while the data for the other variables are continuous.

2.8 The Q-Q plot

The **quantile-quantile plot** or **Q-Q plot** is a very useful plot that can be used to compare two probability distributions by plotting their quantiles (see Section 1.6.6) against each other.

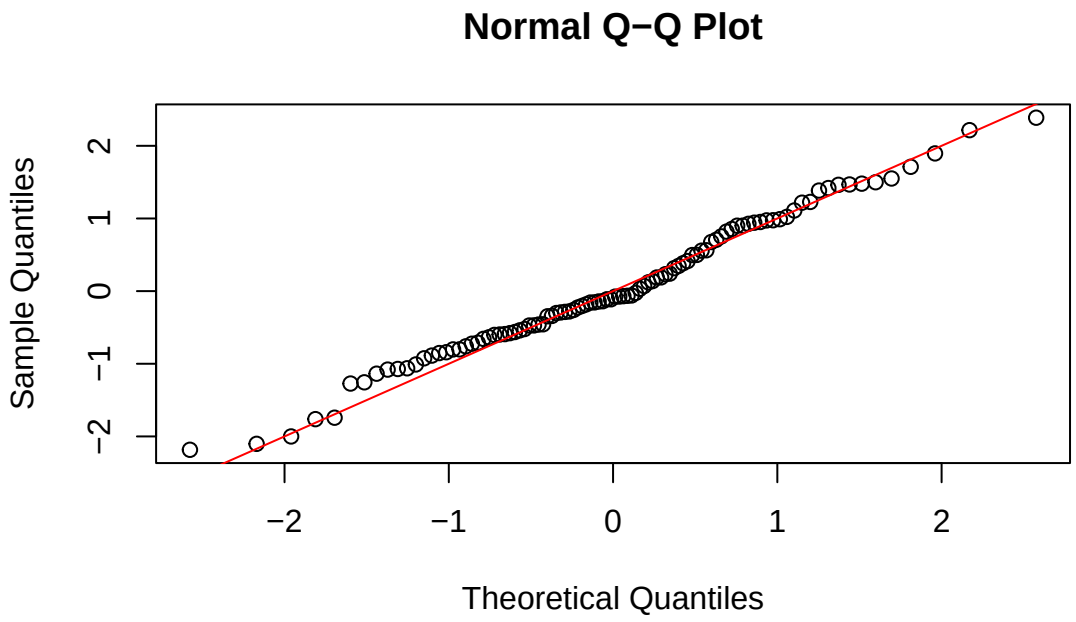
It is very useful, for example, for checking if data is approximately normally distributed. As an experiment, let us generate some data following $\mathrm{N}(3, 4)$.

```
# generate 100 observations following N(3, 4) distribution
set.seed(5)
mu <- 3
sigma <- 2
x <- rnorm(100, mean=mu, sd=sigma)

# density of normal
xx <- seq(from=-5, to=11, length=100)
yy <- dnorm(xx, mean=mu, sd=sigma)

# Create plot with 2 subplots; 1 row, 2 columns
layout( matrix(c(1,2), nrow=1, ncol=2, byrow = FALSE) )

#---- plot 1: Density of N(3, 4)----#
main <- paste0("N(", mu, ", ", sigma^2,") distribution")
plot(xx, yy, type='l', main=main, xlab="Value", ylab="Density")

#---- plot 2: histogram of x----#
hist(x, breaks=20)
```



The histogram above closely resembles the density plot. Now, we create the Q-Q plot after standardising the data, i.e. subtracting the mean and dividing by the standard deviation so that the data has mean 0 and variance 1. The code below shows the Q-Q plot:

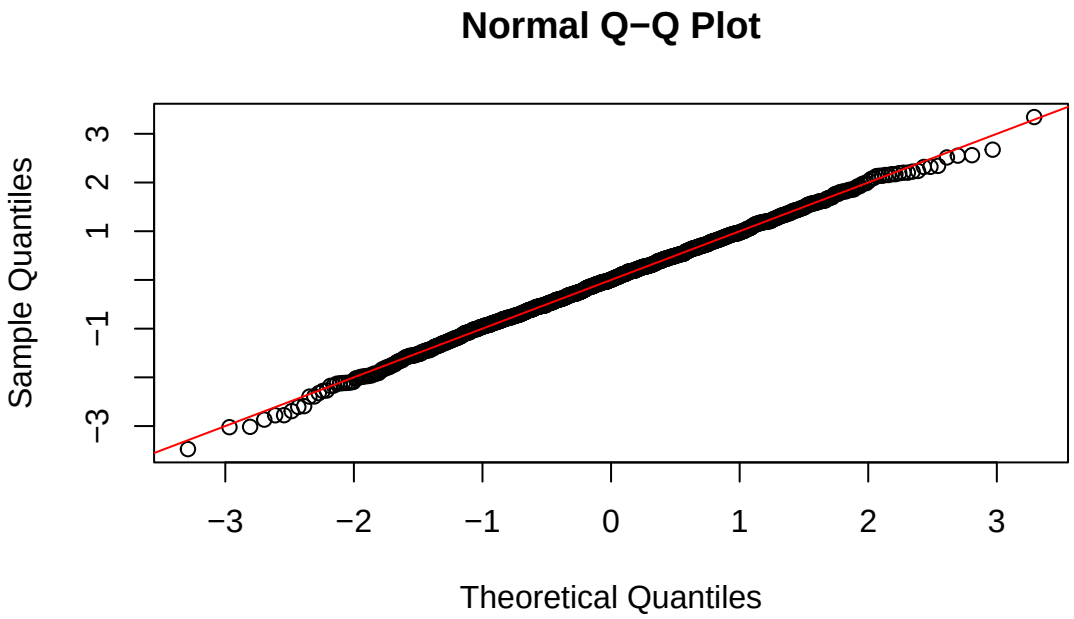
```
# qqplot of standardised data, subtracting mu and dividing by sigma
z <- (x-mu)/sigma
qqnorm(z)

# this plots the line y=x through the data in red
abline(0, 1, col="red")
```



The `qqplot` function first generates the **sample quantiles** of the data `x` by sorting the data. In other words, since there are 100 values, the smallest value represents the 0.01 (first) quantile, and the second smallest representsthe 0.02 (second) quantile, etc. Next, the **theoretical quantiles** of the standard normal distribution are created, i.e. the first quantile will be `qnorm(1/100)`, the second will be `qnorm(2/100)` etc. Then, the sample and theoretical quantiles are plotted against each other in a scatter plot. If the scatterplot lies roughly along the line $y = x$, then the distributions are very similar to each other. The Q-Q plot above shows that the two distributions seem to agree, as the data mostly follows the line $y = x$, given that the sample size of 100 is not very large. This is good, since the standardised data indeed follows a $\mathrm{N}(0, 1)$ distribution.

If we repeat the experiment using more observations, say 1000 instead of 100, then the agreement with the line $y = x$ becomes more pronounced:

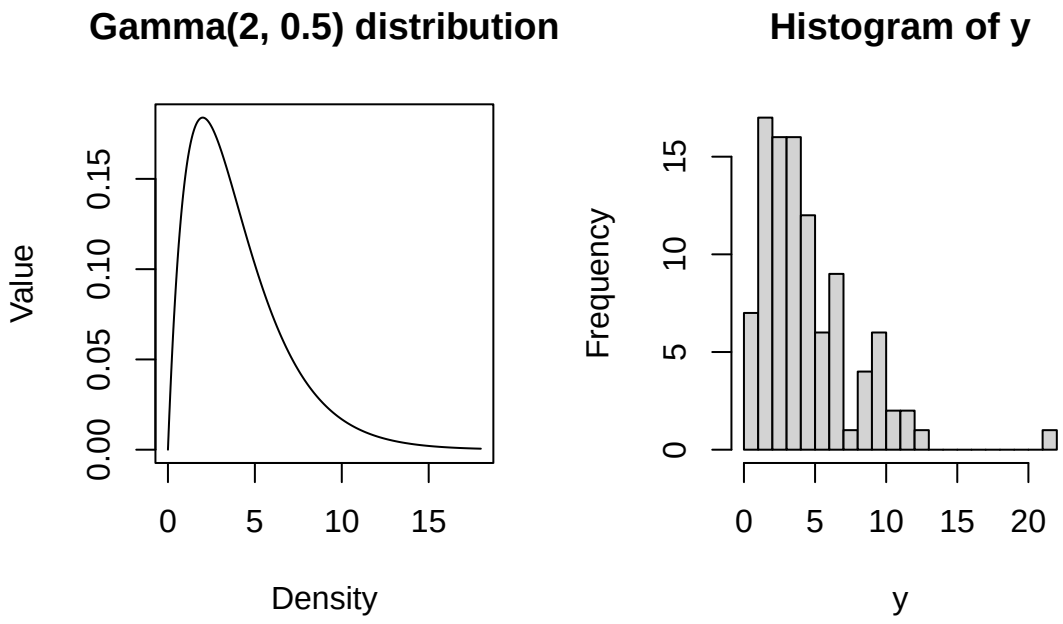
```r
# generate the data
set.seed(5)
mu <- 3
sigma <- 2
x <- rnorm(1000, mean=mu, sd=sigma)

# standardise the data using the sample mean and sample standard deviation
z <- (x-mean(x))/sd(x)

#create qqplot and plot line y=x in red
qqnorm(z)
abline(0, 1, col="red")
```



Now that we have seen an example of a Q-Q plot when the sample data is normally distributed (like the theoretical distribution), let us have a look at a Q-Q plot when the data is different from the theoretical distribution.

For this experiment, we shall sample 100 observations from a $\Gamma(2, 0.5)$ distribution using the shape-rate parametrisation. As before, the probability density function of this distribution is plotted along with the histogram of the data.

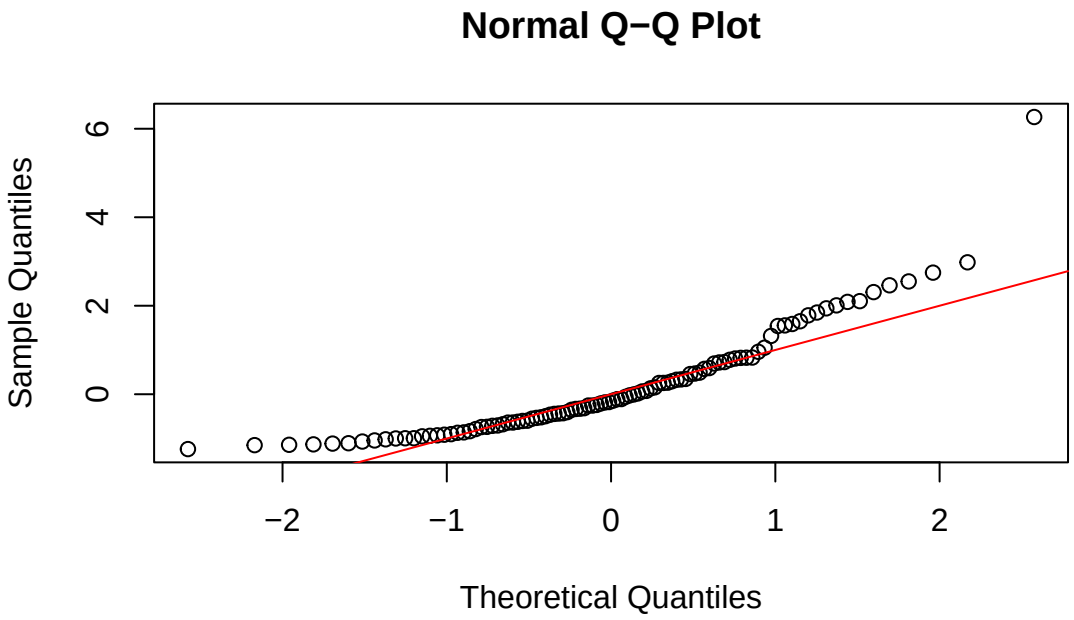
```
# Gamma distribution Gamma(2, 2), with shape k=2 and scale theta=2
alpha <- 2
beta <- 0.5
y <- rgamma(100, shape=alpha, rate=beta)

# Create plot with 2 subplots; 1 row, 2 columns
layout( matrix(c(1,2), nrow=1, ncol=2, byrow = FALSE) )

#---- plot 4 ----#
xx <- seq(from=0, to=18, by=0.01)
yy <- dgamma(xx, shape=alpha, rate=beta)
main <- paste0("Gamma(", alpha, ", ", beta,") distribution")
plot(xx, yy, type='l', main=main, xlab="Density", ylab="Value")

#---- plot 5 ----#
hist(y, breaks=20)
```



Next, we create the Q-Q plot of the standardised Gamma data, assuming the theoretical distribution is a standard normal.

Note that if $X \sim \Gamma(\alpha, \beta)$, then $\mathrm{E}(X) = \frac{\alpha}{\beta}$ and $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$.

```
# standardise the data
mu <- alpha/beta
sigma <- sqrt(alpha) / beta
z <- (y-mu)/sigma

#create the Q-Q plot
qqnorm(z)
abline(0, 1, col="red")
```



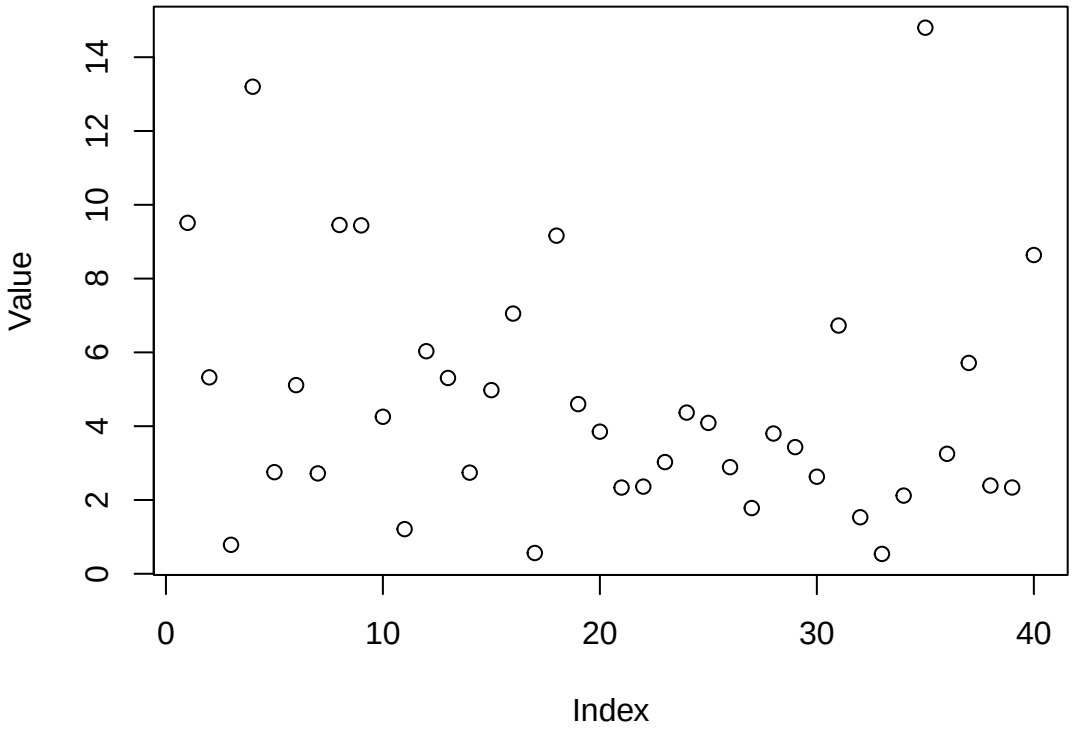
The data does not follow the line $y = x$ very well; the points representing the lower and upper quantiles are above the line. This leads us to conclude that the sample data does not follow a normal distribution. Of course, we know this is the correct conclusion because the sample data is Gamma-distributed.

Note that while the Q-Q plot provides graphical evidence that two distributions are the same or not, this is still a qualitative measure rather than a quantitative measure. While it is possible to extract a statistic from a Q-Q plot, this is beyond the scope of the course.

2.9 The Box plot

A box plot is graphical technique for visualising the distribution of a data set. It is particularly useful for visualising the spread of data and identifying values which can be considered **outliers**, where outliers are extreme values that are significantly different to the other data.

As an example, consider the following data set of $x_1, x_2, \ldots, x_{40}$ values which we plot using a scatterplot, where the x-axis is simply the index of each observation:



A boxplot is constructed as follows: a rectangular box is plotted between the lower and upper quartiles (see Section 1.6.6) with the median shown as a thick black line through the box, and two 'whiskers' are attached to the box showing the minimum and maximum values of the data. However, there is a slight caveat; these whiskers may not show the actual minimum or maximum value if the data set contains outliers.

Outliers in a box plot are determined according the following criterion: if we let $q_{0.25}$ denote the lower quartile, $q_{0.75}$ denote the upper quartile and $\text{IQR} = q_{0.75} - q_{0.25}$, as usual, then

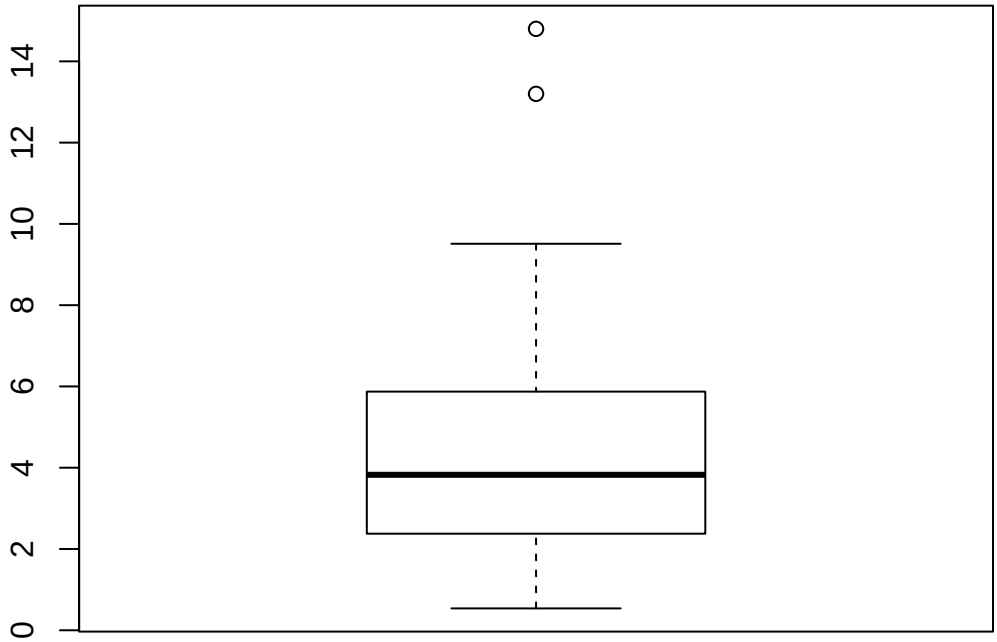
$$x \notin [q_{0.25} - 1.5 \cdot \text{IQR},\, q_{0.75} + 1.5 \cdot \text{IQR}] \qquad \Rightarrow \qquad x \text{ is an outlier}$$

In other words, if a value x is more than 1.5 times the IQR below the lower quartile (or above the upper quartile), then x is an outlier.

If any outliers are identified, then the whiskers show the smallest/largest values of the data set that are not outliers, and the outliers are indicated as circles at the outlier values above/below the whiskers.

Here we plot a box plot using R.

```
# while the default in R is grey, the box of the boxplot is usually white
boxplot(x, col="white")
```



Two outliers are identified around the value 14, while the other values are between 0 and 10. The lower quartile is approximately 2, the upper quartile is approximately 6, and the median is approximately 4.

It is possible to obtain these values directly using the R function `boxplot.stats`:

```
# print the five values used in the box plot:
# lower whisker, lower quartile, median, upper quartile, upper whisker
print( boxplot.stats(x)$stats )
#> [1] 0.53803 2.37641 3.82623 5.87291 9.51134

#print the outliers
print( boxplot.stats(x)$out )
#> [1] 13.2 14.8
```

Two points to note:

- It is possible to print multiple boxplots for a data set simultaneously; for example try `boxplot(mtcars[, c(3, 4)])`.
- It is possible to plot box plots horizontally using the option `horizontal=TRUE` inside the call of the box plot, i.e. `boxplot(x, horizontal=TRUE)`.

2.10 Pseudorandom Number Generators

This section provides a short overview into how 'random' numbers are generated in statistical programming languages such as R. Such numbers are called **pseudorandom**, because they are generated in a deterministic way and are therefore not truly random; however, sequences of pseudorandom numbers are carefully generated in such a way that they exhibit desirable properties of random numbers.

Most pseudorandom number generators (PRNGs) are concerned with generating sequences of integers that are uniformly-distributed in a bounded range. These values are then later transformed into sequences following a particular distribution, such as the normal distribution.

2.10.1 A simple approach: middle squares method

A simple approach developed by John von Neumann starts with a positive integer x with with $2n$ digits, and obtains the next number in the sequence by taking the middle $2n$ digits of x^2, perhaps after padding x^2 with zeros. One example is

$$6238 \to \cancel{38}\mathbf{9126}\cancel{44} \to \cancel{83}\mathbf{2838}\cancel{76} \to \cancel{08}\mathbf{0542}\cancel{44} \to \cancel{00}\mathbf{2937}\cancel{64} \to \cancel{08}\mathbf{6259}\cancel{69}.$$

However, this method suffers from all sequences having short **periods**, and either converging to zero,

$$8277 \to \cancel{68}\mathbf{5087}\cancel{29} \to \cancel{25}\mathbf{8775}\cancel{69} \to \cancel{77}\mathbf{0006}\cancel{25} \to \cancel{00}\mathbf{0000}\cancel{36} \to 0,$$

or reaching a state where the numbers repeat,

$$7953 \to \cancel{63}\mathbf{2502}\cancel{09} \to \cancel{06}\mathbf{2600}\cancel{04} \to \cancel{06}\mathbf{7600}\cancel{00} \to \cancel{57}\mathbf{7600}\cancel{00}.$$

2.10.2 A modern approach: Mersenne Twister

The Mersenne Twister algorithm [12] is one of the most widely-implemented algorithms for generating pseudorandom integers, and is the default choice in R and Python, as well as other languages such as MATLAB and Julia.

The most popular implementation of the Mersenne Twister algorithm has a period of $2^{19937}-1$ for carefully chosen parameters. For a w-bit implementation, the algorithm generates uniform pseudorandom integers in the range $[0, 2^w - 1]$ by representing an integer as a binary w-dimensional row vector and applying a linear transformation to obtain successive integers. In other words, if the starting integer is $\mathbf{x}$, and the linear transformation is represented by the matrix A, then the successive integers are $\mathbf{x}, \mathbf{x}A, \mathbf{x}A^2, \ldots, \mathbf{x}A^i, \ldots$. The linear transformation is chosen in such a way that it is both **fast** to compute and the resulting sequence is 'sufficiently' **uniformly-distributed**. Standard implementations of the algorithm are either for $w = 32$ or $w = 64$. Note that there are many alternative algorithms for generating pseudorandom integers.

```
# Let's see how big 2^19937 - 1 is in Python 3
from decimal import Decimal
print("{:.3E}".format(Decimal(str(2**19937 - 1))))
#> 4.315E+6001
```

2.10.3 Tests for randomness

There are several statistical sets of tests for pseudorandom number generators (PRNGs), such as the *diehard* set of tests, which test the uniformity of the sequences generated by a PRNG. Some tests explicitly look at the distribution of the integers generated, while others look at the binary sequences generated (which would be interpreted as integers). Most seem to involve computing a statistic that, in large samples, will follow a particular distribution if the values are uniformly distributed, and then using the chi-squared test to check if the statistic does not follow the particular distribution.

2.10.4 Setting the seed

For pseudorandom number generators (PRNGs) such as the Mersenne Twister algorithm which are deterministic, a pseudorandom sequence is essentially determined by the first value in the sequence. Many PRNGs allow a **random seed** to be specified by the user, which will determine the first value in the sequence, and consequently will determine the rest of the sequence. This is a useful feature for when a statistician wishes to perform a **reproducible analysis**, or to check whether two analyses produce the same result, or for debugging purposes. If a seed is not set, then R may start generating random values using a random seed based on the computer's clock time. Random seeds are usually, but not necessarily, integers. Below is an example of `rnorm` starting with the seed `1234`.

```
# generating a normally-distributed sequence with a particular seed
set.seed(1234)
print(rnorm(4))
#> [1] -1.20707  0.27743  1.08444 -2.34570

# generating more values
print(rnorm(4))
#> [1]  0.42912  0.50606 -0.57474 -0.54663

# generating more values, after resetting the seed
set.seed(1234)
print(rnorm(6))
#> [1] -1.20707  0.27743  1.08444 -2.34570  0.42912  0.50606
```

2.10.5 Pseudorandom numbers following a particular distribution

While pseudorandom number generators produce uniformly-distributed integers within a bounded range, statisticians often require random numbers which follow a distribution such as a normal or gamma distribution. Supposing the PRNG generates w-bit integers, the first step is to convert an integer $\mathbf{x} \in [0, 2^w - 1]$ into $\mathbf{x} \cdot 2^{-w} \in [0, 1)$ which will be $\mathrm{U}(0, 1)$-distributed. In order to obtain a sequence following a normal distribution, one option would then be to use the normal distribution's **inverse** cumulative distribution function ϕ^{-1} to compute $\phi^{-1}(\mathbf{x} \cdot 2^{-w})$. While the default setting for the `rnorm` function in R produces normally-distributed numbers by the procedure described, there are alternatives algorithms both for generating integers, and for converting $\mathrm{U}(0, 1)$ numbers to $\mathrm{N}(\mu, \sigma^2)$ numbers.

```
# Let's see how U(0,1) values come from integers
x <- runif(5)
print(x)
#> [1] 0.5858003 0.0089458 0.2937396 0.2773750 0.8135742
print(x * 2^32)
#> [1] 2515993152   38421900 1261602027 1191316373 3494274646

# Now, see how R generates observations following a normal distribution
set.seed(1)
print(rnorm(1, mean=2, sd=1))
#> [1] 1.3735

## Next, we set the seed and generate a U(0, 1) observation
set.seed(1)
u <- runif(1)
print(u)
#> [1] 0.26551

# Now we see how the inverse cdf function, qnorm, converts the U(0, 1)
# observation into a N(2, 1) observation; notice that it is the same value
print(qnorm(u, mean=2, sd=1))
#> [1] 1.3735
```

2.10.6 Other algorithms for pseudorandom number generation

There is third-year module called Stochastic Simulation which covers random number generation and other topics in much greater detail. To learn more about different methods for pseudorandom number generation in R and see a list of references, look at the R's documentation for `.Random.seed`:

```
# See the other methods available in R
?.Random.seed
```

Chapter 3

Samples of Normal Random Variables

In this chapter we extend the results of the Chapter 1 to the special case where the random variables $X_1, X_2, \ldots, X_n$ follow a normal distribution. We are able to derive the distributions for both the sample mean and the sample variance in this special case, and then we use these results to define Student's t-distribution which is useful for constructing confidence intervals.

3.1 The sample mean of normal random variables

We first prove a result about the sample mean of a collection of normal random variables:

Proposition 3.1.1. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, and that for $i \in \{1, 2, \ldots, n\}$, $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, where each μ_i and each σ_i is finite. Then, defining $Y = \sum_{i=1}^{n} X_i$,

$$Y \sim \mathrm{N}\left(\mu, \sigma^2\right),$$

where

$$\mu = \sum_{i=1}^{n} \mu_i,$$

and

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2.$$

♦

Remark 3.1.2. Note how general Proposition 3.1.1 is; each μ_i and each σ_i can be different for all $i = 1, 2, \ldots, n$, yet the sum still follows a normal distribution. □

Proof of Proposition 3.1.1.

Theorem 12.2.7 of Prof. Veraart's notes states that if X_i has moment generating function M_{X_i}, then

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

Each $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, therefore $M_{X_i}(t) = \exp(\mu_i t + \sigma_i^2 t^2/2)$, and since the X_i are independent,

$$\begin{aligned}
M_Y(t) &= M_{\sum_{i=1}^n X_i}(t) \\
&= \prod_{i=1}^n M_{X_i}(t) \\
&= \prod_{i=1}^n \exp\left(\mu_i t + \sigma_i^2 t^2/2\right) \\
&= \exp\left(\sum_{i=1}^n \left[\mu_i t + \sigma_i^2 t^2/2\right]\right) \\
&= \exp\left(t\sum_{i=1}^n \mu_i + (t^2/2)\sum_{i=1}^n \sigma_i^2\right) \\
&= \exp\left(\mu t + \sigma^2 t^2/2\right)
\end{aligned}$$

where we set $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$, and this shows that $Y \sim \mathrm{N}(\mu, \sigma^2)$, as required.

□

An important corollary of Proposition 3.1.1 looks at the special case where the μ_i are equal and the σ_i are equal.

Corollary 3.1.3. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables distributed according to $\mathrm{N}(\mu, \sigma^2)$. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. ♦

Proof.

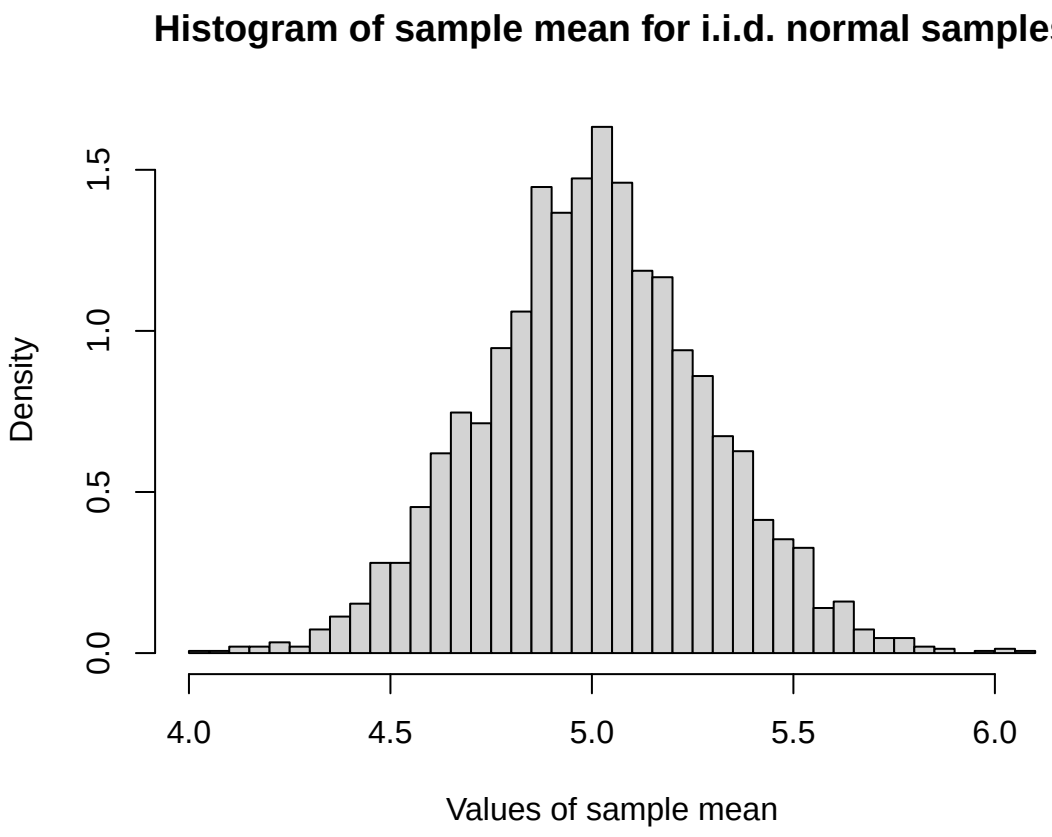
For each $i \in \{1, 2, \ldots, n\}$, $X_i \sim \mathrm{N}(\mu, \sigma^2)$, and so $\frac{1}{n}X_i \sim \mathrm{N}\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right)$. Then, using Proposition 3.1.1,

$$\overline{X} = \sum_{i=1}^{n} \frac{1}{n} X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} \frac{\mu}{n}, \sum_{i=1}^{n} \frac{\sigma^2}{n^2}\right) \qquad \Rightarrow \qquad \overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Note how $\mathrm{E}\left(\overline{X}\right) = \mu$ and $\mathrm{Var}\left(\overline{X}\right) = \sigma^2/n$, as shown in Proposition 1.2.6, when normality of the random variables was not assumed.

□

We can run an experiment in R to 'verify' this result; generate observations $x_1, x_2, \ldots, x_{50}$ for random variables $X_1, X_2, \ldots, X_{50} \sim \mathrm{N}(5, 2^2)$ and compute the sample mean $\overline{x}$, repeat this 3000 times and plot the results in a histogram. The results of such an experiment are shown below and seems to indicate the $\overline{X}$s are normally distributed. See Appendix A.3 for the code (but try it yourself first!).



3.2 The sample variance of normal random variables

This section investigates the distribution of the sample variance in the special case that the random variables $X_1, X_2, \ldots, X_n$ are i.i.d. according to a normal distribution.

First, we learn a useful result concerning orthogonal transformations of normal random variables. Although it will not immediately clear how this result is used, we shall soon use it in an important theorem.

Proposition 3.2.1. Suppose that $Z_1, Z_2, \ldots, Z_n$ are i.i.d. random variables each with a $\mathrm{N}(0,1)$ distribution, and write $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$. Suppose that $\mathbf{A}$ is an orthogonal $n \times n$ matrix, and define $\mathbf{Y} = \mathbf{AZ}$, with $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Then the $Y_1, Y_2, \ldots, Y_n$ are also i.i.d. random variables each with a $\mathrm{N}(0,1)$ distribution, and furthermore $\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2$. ♦

Proof. First, since $\mathbf{A}$ is an orthogonal matrix, $\mathbf{AA}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}_n$, the $n \times n$ identity matrix. Then,

$$\sum_{i=1}^n Y_i^2 = \mathbf{Y}^T\mathbf{Y} = \mathbf{Z}^T\mathbf{A}^T\mathbf{AZ} = \mathbf{Z}^T\mathbf{I}_n\mathbf{Z} = \mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^n Z_i^2. \tag{3.1}$$

Also, note that $|\det(\mathbf{A})| = 1$, since

$$1 = \det(\mathbf{I}_n) = \det\left(\mathbf{AA}^T\right) = \det(\mathbf{A})\det\left(\mathbf{A}^T\right) = (\det(\mathbf{A}))^2.$$

Now, the joint p.d.f. of the random variables $Z_1, Z_2, \ldots, Z_n$ is, for $-\infty < z_i < \infty$ ($i \in \{1, 2, \ldots, n\}$),

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^2\right).$$

Since $\mathbf{Y} = \mathbf{AZ}$, this is a linear change of variables. Since $\mathbf{A}$ is orthogonal it is also invertible, and we can write $\mathbf{z} = \mathbf{A}^{-1}\mathbf{y}$. For this transformation, the Jacobian is $[\det(\mathbf{A})]^{-1}$, and so the p.d.f. of $\mathbf{Y}$, denoted $g(\mathbf{y})$, is given by (from a theorem in multivariable calculus; see Theorem 9.2.2 in Prof. Veraart's notes for the same result for a single variable)

$$g(\mathbf{y}) = \frac{1}{|\det(\mathbf{A})|} f\left(\mathbf{A}^{-1}\mathbf{y}\right) = \left(\frac{1}{1}\right) f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^2\right) \tag{3.2}$$

$$\Rightarrow g(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n y_i^2\right), \tag{3.3}$$

where Equation (3.3) follows from Equation (3.2) by using Equation (3.1), which also holds for the realisations $\mathbf{z}$ and $\mathbf{y}$. Equation (3.3) shows that the joint p.d.f. of the $Y_1, Y_2, \ldots, Y_n$ random variables is the same as that for the $Z_1, Z_2, \ldots, Z_n$ random variables, and so the $Y_1, Y_2, \ldots, Y_n$ are independent and each Y_i has a $\mathrm{N}(0,1)$ distribution. □

Now we turn our attention to the sample variance S^2, and its relation to the sample mean $\overline{X}$, when the sample consists of i.i.d. normal random variables.

Theorem 3.2.2. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables distributed according to $\mathrm{N}(\mu, \sigma^2)$, with $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Then $\overline{X}$ and S^2 are independent random variables and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{3.4}$$

♦

Proof. We define the random variables $Z_1, Z_2, \ldots, Z_n$ by $Z_i = (X_i - \mu)/\sigma$. Then,

$$X_i \sim \mathrm{N}\left(\mu, \sigma^2\right) \Rightarrow Z_i = \frac{X_i - \mu}{\sigma} \sim \mathrm{N}(0,1).$$

Choose an orthogonal linear transformation $\mathbf{A}$ that has the first row equal to

$$\mathbf{u} = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

Note that the length of the vector $\mathbf{u}$ is 1. One way to construct such an $\mathbf{A}$ is to start with the $n \times n$ identity matrix $\mathbf{I}_n$, replace its first row with $\mathbf{u}$, and then use the Gram-Schmidt orthogonalisation procedure described in the linear algebra module. Now define the vector of random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ by $\mathbf{Y} = \mathbf{AZ}$. Then, using Proposition 3.2.1, the $Y_1, Y_2, \ldots, Y_n$ are also i.i.d. random variables each with distribution $\mathrm{N}(0,1)$, and $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2$. Furthermore,

$$Y_1 = \mathbf{uZ} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} Z_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = \sqrt{n}\left(\overline{Z}\right),$$

and therefore, using Equation (1.9) in Exercise 1.2.5,

$$\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2 = \sum_{i=1}^{n} Z_i^2 - n\left(\overline{Z}\right)^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2.$$

Since the Y_i are independent,

$$\begin{aligned} &\sum_{i=2}^{n} Y_i^2 \text{ is independent of } Y_1, \\ \Rightarrow &\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2 \text{ is independent of } \overline{Z}, \\ \Rightarrow &\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ is independent of } \overline{X}. \end{aligned}$$

The last implication follows since transformations of independent random variables are still independent (see Appendix A.4) and

$$\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \qquad \overline{Z} = \frac{1}{\sigma}\left(\overline{X} - \mu\right).$$

Now the distribution of S^2 follows from:

$$\begin{aligned}
&\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \frac{(n-1)\,S^2}{\sigma^2} \\
&\sum_{i=2}^{n} Y_i^2 \sim \chi^2_{n-1} \qquad \text{(the } Y_i \text{ are i.i.d. standard normal, see Remark 3.2.3)} \qquad (3.5) \\
\Rightarrow &\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}
\end{aligned}$$

which proves the result. This proof is taken from [3] and [11]. □

Remark 3.2.3. Equation (3.5) follows from the following two facts:

- If $Z \sim \mathrm{N}(0,1)$, then $Z^2 \sim \chi^2_1$ (Term 1 Problem Sheet 5, Exercise 8)
- If Y_1, Y_2 are independent and $Y_1^2, Y_2^2 \sim \chi^2_1$, then $Y_1^2 + Y_2^2 \sim \chi^2_2$ (Term 1 Problem Sheet 7, Exercise 3). As an exercise, show that if $Y_1, Y_2, \ldots, Y_n$ are independent and each $Y_i^2 \sim \chi^2_1$, then $\sum_{i=1}^{n} Y_i^2 \sim \chi^2_n$.

□

Remark 3.2.4. It is also worth reviewing why this proof went through the effort of using an orthogonal transformation, etc. We could define $W_i = X_i - \overline{X}$ for $i = 1, 2, \ldots, n$, and then write

$$(n-1)S^2 = \sum_{i=1}^{n} W_i^2,$$

however, the W_i are clearly not independent since each W_i involves an $\overline{X}$ term which contains all the variables $X_1, X_2, \ldots, X_n$. So, the reason for the orthogonal transformation was to rewrite S^2 as a sum of **independent** random variables, i.e. the random variables Y_i^2, for $i = 2, 3, \ldots, n$. This then allowed us to determine that the distribution of S^2 (after a rescaling) is χ^2_{n-1}. □

3.3 Confidence intervals for normal random variables

We have already seen in Section 1.4 how to construct a confidence intervals for the unknown mean of a sample using Chebyshev's inequality. In that case, the random variables $X_1, X_2, \ldots, X_n$ were observed as $x_1, x_2, \ldots, x_n$, and the random variables were assumed to follow the same distribution with unknown mean θ and known variance σ^2, but the distribution itself was unknown.

We now consider the situation where the random variables are known to be a normal distribution.

3.3.1 Case 1: normal distribution with variance known

Suppose that the random variables $X_1, X_2, \ldots, X_n$ follow a normal distribution with unknown mean θ and known variance σ^2. Our goal is to obtain confidence interval for the unknown mean θ. Suppose we wish to obtain a $(1-\alpha)$ confidence interval for some value α. If we consider the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

we know from Corollary 3.1.3 that

$$\overline{X} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right).$$

If one knew the value of θ, then one could define Z as

$$Z = \frac{\theta - \overline{X}}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(0, 1\right). \tag{3.6}$$

Using the cumulative distribution function F, we can find the value $z_{\alpha/2}$ where

$$\mathrm{P}\left(Z < z_{\alpha/2}\right) = \mathrm{P}\left(Z \leq z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

(Recall that $\mathrm{P}\left(Z < z_{\alpha/2}\right) = \mathrm{P}\left(Z \leq z_{\alpha/2}\right)$ because the normal distribution is continuous and therefore $\mathrm{P}\left(Z = z_{\alpha/2}\right) = 0$.)

We can similarly find $z_{1-\alpha/2}$, where

$$\mathrm{P}\left(Z < z_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}.$$

(In fact, using the symmetry of the normal distribution, $z_{1-\alpha/2} = -z_{\alpha/2}$.) Then,

$$\mathrm{P}\left(z_{\alpha/2} < Z < z_{1-\alpha/2}\right) = 1 - \alpha.$$

Therefore, we can obtain a confidence interval for θ by manipulating this equation as follows:

$$\begin{aligned}
&\mathrm{P}\left(z_{\alpha/2} < Z < z_{1-\alpha/2}\right) = 1-\alpha. \\
\Rightarrow &\mathrm{P}\left(z_{\alpha/2} < \frac{\theta - \overline{X}}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1-\alpha \\
\Rightarrow &\mathrm{P}\left(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta - \overline{X} < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \\
\Rightarrow &\mathrm{P}\left(\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \\
\Rightarrow &\mathrm{P}\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.
\end{aligned}$$

Remark 3.3.1. Usually one would define $Z = \frac{\overline{X}-\theta}{\sigma/\sqrt{n}}$, however defining Z as in Equation (3.6) above allows the confidence interval bounds to be derived more smoothly. □

Example 3.3.2. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables following a normal distribution with unknown mean θ and variance $\sigma^2 = 9$. Suppose we observe $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Given that $\overline{x} = 4$ and $n = 25$, let us construct a 95% confidence interval for θ.

First, for a 95% confidence interval, we have $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$. Therefore, $1 - \frac{\alpha}{2} = 0.975$, and we look at either Table 3.1 or Table 3.2 and find that $z_{0.975} = 1.96$, since $\mathrm{P}(Z < 1.96) = 0.975$. By symmetry, we have $z_{0.025} = -1.96$, i.e. $\mathrm{P}(Z < -1.96) = 0.025$.

Then, since $\overline{x} = 4$, $\sigma = 3$ and $\sqrt{n} = 5$, we can compute the 95% confidence interval for the unknown mean θ to be

$$\left(4 - 1.96 \cdot \frac{3}{5}, 4 + 1.96 \cdot \frac{3}{5}\right).$$

△

Remark 3.3.3. Notice how the confidence interval for the mean above is of the form

$$\left(\widehat{\Theta} - k \cdot \mathrm{SE}_{\widehat{\Theta}}, \widehat{\Theta} + k \cdot \mathrm{SE}_{\widehat{\Theta}}\right),$$

since $\widehat{\Theta} = \overline{X}$ and $\mathrm{SE}_{\widehat{\Theta}} = \sqrt{\mathrm{Var}\left(\widehat{\Theta}\right)} = \frac{\sigma}{\sqrt{n}}$. Here, where we can assume normality, we have $k = z_{1-\alpha/2}$. Also compare this to Example 1.5.17 which used Chebyshev's inequality to create the confidence interval for the mean and uses a value of k determined by Chebyshev's inequality. □

3.3.1.1 How to read Table 3.1

In Table 3.1, the values of $\mathrm{P}(Z < z)$ are given in the table, while one has to read the value of z off the row/column heading. One can read the table in two ways: First, suppose one wishes to know $\mathrm{P}(Z < 0.31)$. Then, one looks at the entry in the third row (marked 0.3) and the second column (marked 0.01), and one read the value 0.6217. Then, this means $\mathrm{P}(Z < 0.31) = 0.6217$.

For a second approach, suppose one wishes to find the z value for which $\mathrm{P}(Z < z) = 0.975$. One must then find the value in the table that is closest to 0.975. One searches the table, noticing that the values increase from top to bottom, and left to right. One actually finds the value 0.975 in the table in the row marked 1.9 and the column marked 0.06. This means that $\mathrm{P}(Z < 1.96) = 0.975$.

If the exact value you are looking for is not in the table, you can use the closest value or interpolate two values. For example, when trying to find the z value such that $\mathrm{P}(Z < z) = 0.95$, one ends up finding $\mathrm{P}(Z < 1.64) = 0.9495$ and $\mathrm{P}(Z < 1.65) = 0.9505$. One could choose either of these values, or interpolate and use 1.645 (which happens to be very close to the correct answer), but care must be used when interpolating.

These statistical tables were the traditional way one looked up values for cumulative distribution functions for different distributions. Now, if one has access to a computer, it is easy to compute the cumulative distribution function:

```
## Examples of computing critical values for standard normal distribution

# Find the value of z such that P(Z < z) = 0.975
qnorm(0.975)
#> [1] 1.96

# Find the value of P(Z < 1.96)
pnorm(1.96)
#> [1] 0.975

# Find the value of z such that P(Z < z) = 0.95
qnorm(0.95)
#> [1] 1.6449

# Find the value of P(Z < 1.645)
pnorm(1.645)
#> [1] 0.95002
```

Table 3.1: $\mathrm{P}(Z < z)$ where $Z \sim \mathrm{N}(0, 1)$ for values of z between 0.00 and 3.99

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3.2: Selected values of z for $\mathrm{P}(Z < z)$, where Z has a standard normal distribution

z	$\mathrm{P}(Z < z)$
1.281	0.900
1.645	0.950
1.960	0.975
2.326	0.990
2.576	0.995

3.3.1.2 How to read Table 3.2

Table 3.2 is straightforward to read, and provides several pairs of values (p, z) such that $\mathrm{P}(Z < z) = p$.

3.3.2 Confidence intervals experiment

Suppose $n = 25$ observations are sampled from a $\mathrm{N}(\theta, 1)$ distribution where θ is unknown. Suppose we construct a 95% confidence interval; will the **true** value of θ be contained in the resulting confidence interval? We know that for a given sample $x_1, x_2, \ldots, x_{25}$ that the 95% confidence interval will be constructed as

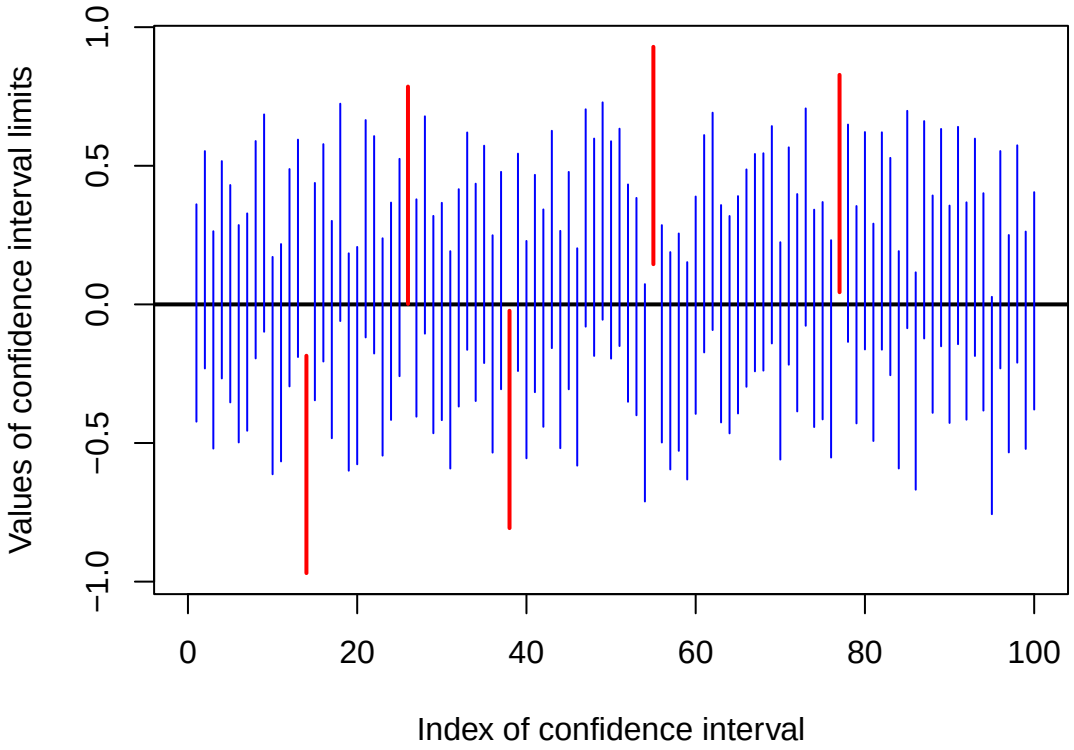
$$\left(\overline{x} - 1.96 \cdot \frac{1}{5}, \overline{x} + 1.96 \cdot \frac{1}{5}\right),$$

so the exact values of the endpoints of the confidence interval will depend on the values of the observed sample.

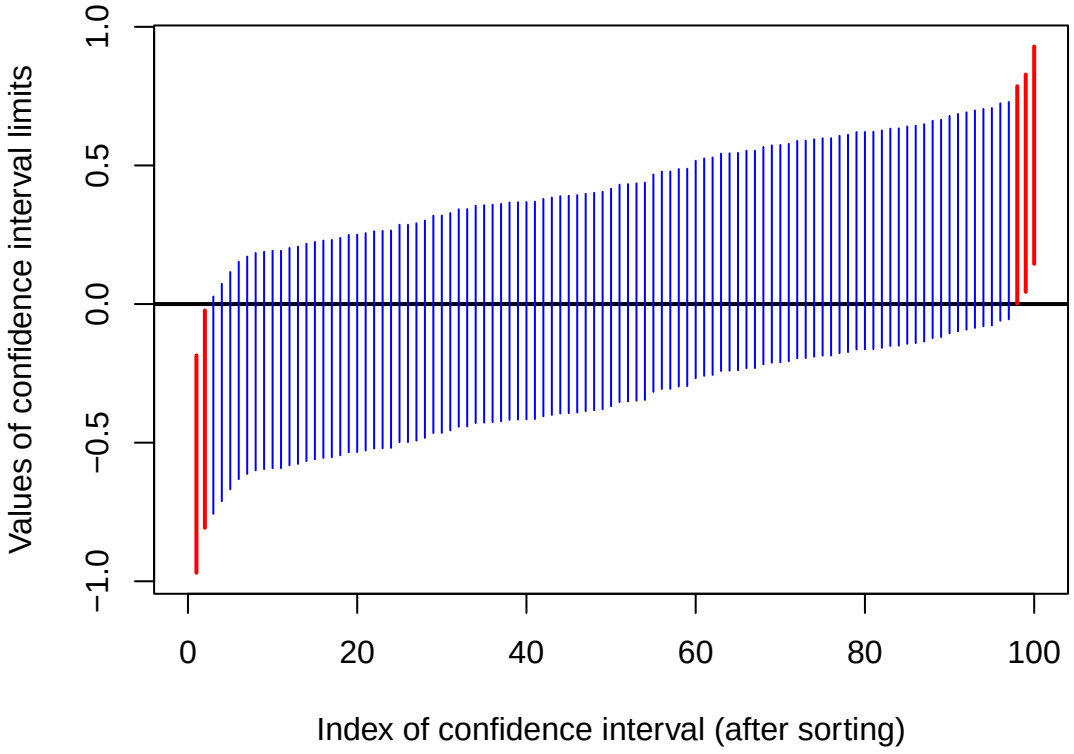
Let us run an experiment where we create 100 samples $x_1, x_2, \ldots, x_{25}$ from a $\mathrm{N}(0, 1)$ distribution. We then construct the confidence intervals as above, and plot these confidence intervals as vertical lines, where the lines are coloured blue if the true mean (in this case, 0) is contained in the confidence interval, and coloured red if the true mean is not contained in the constructed confidence interval.

The figure below shows that out of the 100 intervals, 95 contain the true value of the mean[1].

[1]Note however, that the number of trials (100) is relatively small, so re-running the experiment we may end up with, for example 94 or 97 constructed intervals which contain the true mean value. However, in the limit, the proportion would tend to 0.95 of the intervals containing the true mean value; so re-running the experiment with $10,000$ intervals would result in ~ 9500 intervals containing the true mean.



We can redraw this figure with confidence intervals sorted by the left endpoint, as shown below.



3.4 Student's t-distribution

Now suppose that the sample of random variables $X_1, X_2, \ldots, X_n$ can be assumed to follow a normal distribution with unknown mean θ, but suppose further that **the variance σ^2 is also unknown**. While we would again be able to deduce that $\overline{X} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right)$, even if one knew the value of θ, defining Z as

$$Z = \frac{\theta - \overline{X}}{\sigma/\sqrt{n}}$$

would not be possible as before because σ^2 is unknown. One may consider replacing σ^2 with S^2, where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

and then defining

$$T = \frac{\theta - \overline{X}}{S/\sqrt{n}}. \tag{3.7}$$

One may think that this 'minor' change of σ^2 for S^2 would still allow T to follow a $\mathrm{N}(0,1)$ distribution, but this is not the case. In fact, we will show that the statistic T follows a distribution called Student's t-distribution. We first show that T in Equation (3.7) can be written in a standard form.

Proposition 3.4.1. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables that follow a $\mathrm{N}(\mu, \sigma^2)$ distribution, and $\overline{X} = \frac{1}{n}\sum_{i=1} X_i$ and $S^2 = \frac{1}{n-1}\left(\overline{X} - X_i\right)^2$, as usual. Show that the random variable T, where

$$T = \frac{\mu - \overline{X}}{S/\sqrt{n}},$$

can be written in the form

$$T = \frac{U}{\sqrt{V/p}},$$

where

- $U \sim \mathrm{N}(0,1)$,
- $p = n - 1$,
- $V \sim \chi^2_p$, the chi-squared distribution with p degrees of freedom,
- U and V are independent random variables.

♦

Exercise 3.4.2. Prove Proposition 3.4.1. △

See the solution to Question 2, Problem Sheet 10, Week 18.

Remark 3.4.3. It is common to use the Greek letter ν instead of the letter p as the parameter of Student's t-distribution. This parameter is known as the **degrees of freedom** of Student's t-distribution. □

Theorem 3.4.4. If $U \sim \mathrm{N}(0,1)$ and $V \sim \chi^2_\nu$, and U and V are independent random variables, then the random variable T defined by

$$T = \frac{U}{\sqrt{V/\nu}},$$

follows Student's t-distribution with degrees of freedom ν and has probability density function

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{(\pi\nu)^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

♦

Proof. The proof follows a change of variables argument and that will be straighforward once the Calculus module has been completed; however, the proof **will not be examinable in this module**. It is included in the Section A.5 of the appendix for completeness. □

Remark 3.4.5. The distribution is named 'Student's' t-distribution because the British statistician who first published it, William Sealy Gosset, decided to publish the result under the pseudonym 'Student'. The reason for this is his employer, the Guinness Brewery in Dublin, Ireland, preferred that their staff publish scientific papers using pseudonyms. □

Remark 3.4.6. Note that the probability density function of the t-distribution is symmetric since $f_T(-t) = f_T(t)$. □

Table 3.3 provides critical values for the t-distribution.

Table 3.3: Values of t for $\mathrm{P}(T < t)$, where T has Student's t-distribution with ν degrees of freedom

ν	0.60	0.667	0.75	0.80	0.875	0.90	0.95	0.975	0.99	0.995	0.999
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

3.4.1 How to read Table 3.3

Each row of Table 3.3 shows values of t for a particular degrees of freedom ν, while each column header gives $p = \mathrm{P}(T < t)$.

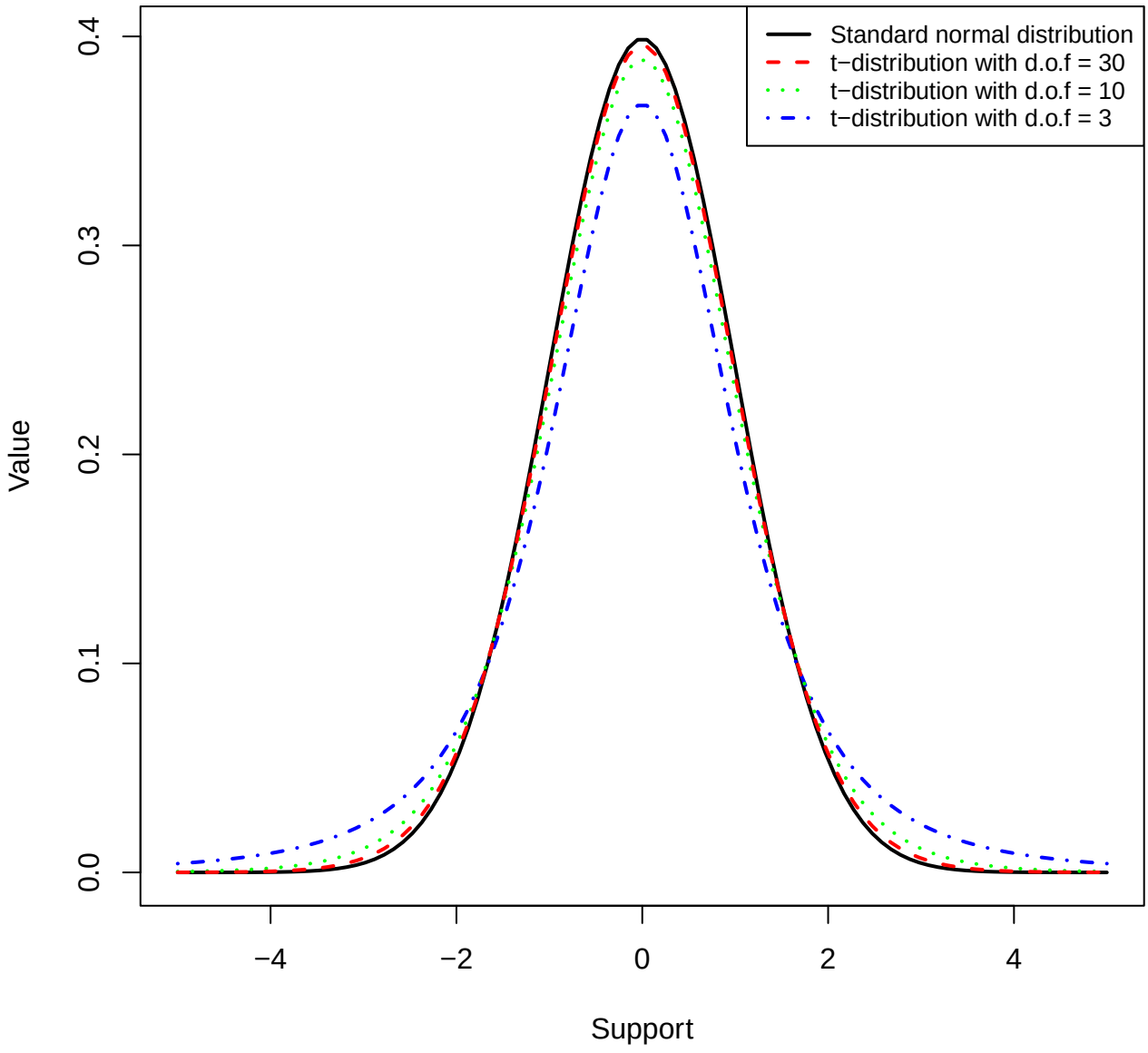
Suppose that one wishes to find the value t for a particular ν and probability value p. Then the entry in the corresponding ν row and column p will give the value t. For example, for $\nu = 9$ and $p = 0.975$, the value of t is 2.262. In other words, if $T \sim t_\nu$, where $\nu = 9$, then $\mathrm{P}(T < 2.262) = 0.975$.

One can also use R to compute the critical values of the t-distribution.

```
## Examples of computing critical values for Student's t-distribution

# Find the value of t such that P(T < t) = 0.975
# where the degrees of freedom (df) is 9
qt(0.975, df=9)
#> [1] 2.2622

# Find the value of P(T < 2.262) where the degrees of freedom (df) is 9
pt(2.262, df=9)
#> [1] 0.97499
```

3.4.2 The shape of the t-distribution

The plot below shows the probability density function of Student's t-distribution (dashed lines) for the degrees of freedom $\nu \in \{3, 10, 30\}$ superimposed on a plot of the probabiliy density function of the standard normal distribution, with mean 0 and variance 1. One immediately, notices that the of the t-distribution is very similar to that of the normal distribution, however One sees that when $\nu = 3$, the tails of the t-distribution are 'heavier'. As ν increases, the t-distribution becomes closer and closer to that of the standard normal distribution. Indeed, if we were to plot the probability density function of the t-distribution for $\nu = 50$, it would be almost indistinguishable from that of the standard normal.



3.4.3 Case 2: normal distribution with variance unknown

We can use Student's t-distribution to construct confidence intervals for the case where the data can be assumed normal, but the variance is unknown.

Example 3.4.7. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables following a normal distribution with unknown mean θ and unknown variance σ^2. Suppose we observe $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Given that $\overline{x} = 5$, and the sample variance $s^2 = 7$ and $n = 12$, let us construct a 95% confidence interval for θ.

First, for a 95% confidence interval, we have $1-\alpha = 0.95 \Rightarrow \alpha = 0.05$. Therefore, $1-\frac{\alpha}{2} = 0.975$. Next, since $n = 12$, the degrees of freedom $\nu = 11$ (see Proposition 3.4.1). Now we look at Table 3.3, in the row $\nu = 11$, and find that when $t_{11,0.975} = 2.201$ then $\mathrm{P}(T < 2.201) = 0.975$. By symmetry, (see Remark 3.4.6) we have $\mathrm{P}(T < -2.201) = 0.025$.

Since

$$T = \frac{\theta - \overline{X}}{S/\sqrt{n}},$$

then

$$\begin{aligned} &\mathrm{P}\left(-t_{\nu,1-\alpha/2} < T < t_{\nu,1-\alpha/2}\right) = 1-\alpha \\ \Rightarrow &\mathrm{P}\left(\overline{X} - t_{\nu,1-\alpha/2}\frac{S}{\sqrt{n}} < \theta < \overline{X} + t_{\nu,1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha. \end{aligned}$$

Then, since $\overline{x} = 5, s^2 = 7$ and $n = 12$, we can compute the 95% confidence interval for the unknown mean θ to be

$$\left(5 - 2.201 \cdot \frac{\sqrt{7}}{\sqrt{12}},\ 5 + 2.201 \cdot \frac{\sqrt{7}}{\sqrt{12}}\right).$$

△

Chapter 4

Hypothesis testing

4.1 Introduction

In this chapter we start by introducing the key concepts of hypotheses and p-values, before moving on to hypothesis testing for one sample. We next investigate the distribution of a p-value and define Type I and Type II errors. We finally look at a classic experiment in hypothesis testing before considering two-sample testing.

Definition 4.1. A **hypothesis** is a statement about a parameter (or parameters) of interest.

In a given experiment, there will always be at least two competing hypotheses. The **null hypothesis**, denoted H_0, is the so-called default position, which specifies the conditions under which the experiment is assumed to have taken place. The **alternative hypothesis**, denoted H_1, is the hypothesis that is complementary to the null hypothesis. A hypothesis test uses data from the experiment to decide which of these two competing hypotheses is supported given the data.

Example 4.1.1. Suppose the data $x_1, x_2, \ldots, x_n$ are observations of the random variable X, where X is assumed to follow a normal distribution with unknown mean θ and variance $\sigma^2 = 1$. We are interested in investigating the value of the unknown mean θ. A null hypothesis could be

$$H_0 : \theta = 0$$

and the alternative hypothesis is then

$$H_1 : \theta \neq 0.$$

△

While this example may appear a bit abstract, the next example illustrates the importance of hypothesis testing.

Example 4.1.2. Suppose a vaccine is developed to prevent infection of a particular disease. The 'vaccine efficacy' is defined as the proportionate reduction in infection rate between vaccinated and unvaccinated individuals. Denoting the vaccine efficacy as VE, one possible null hypothesis (and alternative hypothesis) is

$$H_0 : VE \leq 0.3,$$
$$H_1 : VE > 0.3.$$

This choice reflects the conservative view that a vaccine will only be considered effective if the efficacy is greater than 30%. △

4.1.1 p-values

The null hypothesis usually provides assumptions for the random variables which are observed. Then, once the data is observed, the probability of observing such data (or data at least as extreme as the observed data) can be computed. This probability is called the ***p*-value**.

Definition 4.2. Suppose the data $\mathbf{x}$ is observed, and the test statistic $t(\mathbf{x})$ is computed. Then the ***p*-value** is the probability of obtaining a test statistic at least as extreme as $t(\mathbf{x})$ under the assumption that the null hypothesis H_0 is true.

Remark 4.1.3. Under the null hypothesis, the random variables $\mathbf{X}$ are usually assumed to follow a particular distribution F_X. Then the test statistic is the random variable $T = t(\mathbf{X})$, which is assumed (or derived) to follow the distribution F_T, which has cumulative distribution function also denoted by F_T. Then the p-value is a transformation of T, usually $p = 1 - F_T(T)$, which is between 0 and 1 and is a probability. Once the random variables $\mathbf{X}$ are observed as the data $\mathbf{x}$, then the realized value of the p-value is $1 - F_T(t(\mathbf{x}))$. □

Remark 4.1.4. Since a p-value is a transformation of a random variable, and therefore is itself a random variable, given some data $\mathbf{x}$ and a null hypothesis H_0 one can compute the realization of this p-value, and this realized value is usually also denoted by p. To distinguish between this realization of the p-value and its random variable counterpart, sometimes this value p is called the **significance level**, or **the value of *p***, or the **observed *p*-value**. However, it is often the case that the realized value p will also simply be called **the *p*-value**, so it is worth keeping in mind that the term 'p-value' can have two meanings: it can either refer to the random variable or it can refer to the realized value. □

4.2 Decision making with p-values

Once the null and alternative hypotheses have been defined, one collects data and, under the assumption that the null hypothesis is true, the p-value is computed. We then need to make a decision; which hypothesis 'is true'? If the p-value, denoted by p, is very small (close to 0), this means that it was unlikely that the data are observations of random variables following the assumptions laid out by the null hypothesis; in other words, if our p-value is very small, then it is unlikely our null hypothesis is true.

However, how does one decide if the p-value is 'small enough'? There are two possible approaches:

1. Whatever the value of p-value p, we declare after the experiment that the data provides evidence to reject the null hypothesis at significance level p.
2. Before conducting the experiment, one decides in advance which significance level one would require in order to reject the null hypothesis. Suppose it is decided that the null hypothesis would only be rejected if the p-value were less than a value α. This value $\alpha \in (0, 1)$ is sometimes called the **significance threshold**. Then in order to decide whether or not to reject the null hypothesis, we compare the p-value to α. If the value of p is computed and $p < \alpha$, then the null hypothesis is **rejected** at significance threshold α. One may then be inclined to believe that the alternative hypothesis 'is true'. On the other hand, if $p \not< \alpha$, then the p-value is not extreme enough for the null hypothesis to be rejected. In this case, we **fail to reject** the null hypothesis.

Remark 4.2.1. In essence, the second approach changes the decision on whether or not to reject the null hypothesis given on observed value of p to determining a significance threshold in advance. It is common for the value of $\alpha = 0.05$ to be used, although **there is no particular reason a significance threshold of 0.05 needs to be used**. It is up to the statistician, and the application, to determine the value of α. □

Remark 4.2.2. Notice how in the second approach we either reject the null hypothesis or 'fail to reject' the null hypothesis. Some statisticians see the 'failing to reject' the null hypothesis as equivalent to 'accepting' the null hypothesis, while others are strongly against this idea. Or some may consider rejecting the null hypothesis to be equivalent to 'accepting' the alternative hypothesis. This issue will be discussed in the next section. □

Remark 4.2.3. There is an another interpretation of a p-value based on repeated experiments: suppose data $\mathbf{x}$ is observed which results in the p-value p being computed from the test statistic $t(\mathbf{x})$ assuming the null hypothesis is true. If the same experiment were repeated a large number of times, where in each trial of the experiment new data $\mathbf{x}'$ is observed, then p is the proportion of times that data leading to a test statistic at least as extreme as $t(\mathbf{x})$ is observed. To be more concrete, suppose an experiment records data $\mathbf{x}$ which results in a test statistic $t(\mathbf{x})$ and a p-value of 0.04. If the null hypothesis were true and the experiment were repeated $1,000,000$ times, then approximately 4% of the time data would be obtained that resulted in a test statistic at least as extreme as $t(\mathbf{x})$. □

4.2.1 Rejecting or accepting hypotheses?

There is some subtlety regarding the language used in the decision to reject the null hypothesis or not. Suppose that α is the significance threshold, and p is the computed p-value. If $p < \alpha$, then the null hypothesis **is rejected**. This does not mean that alternative hypothesis is 'accepted'; the most one can say is that the data **supports** the alternative hypothesis.

On the other hand, suppose that $p \not< \alpha$. Then the null hypothesis **fails to be rejected**. This also does not mean that the null hypothesis is 'accepted' as being true; we **never accept the null hypothesis** H_0 to be true just because there is insufficient evidence to reject H_0.

Example 4.2.4. In a criminal lawsuit, the null hypothesis is usually

$$H_0 : \text{the defendant is } \textbf{innocent},$$

while the alternative is

$$H_1 : \text{the defendant is } \textbf{guilty}.$$

In other words, the defendant is 'innocent until proven guilty'. The prosecution and defence then provide evidence and a judge or jury then decide whether or not to reject the null hypothesis H_0. If H_0 is rejected, then the defendant is found to be guilty, i.e. the evidence supports the alternative hypothesis H_1. On the other hand, if there is insufficient evidence to reject H_0, then the defendant is found **not guilty**; one does not declare the defendant to be innocent, i.e. one does not accept H_0. △

The table below shows, given a significance threshold α and a p-value p computed from data recorded during an experiment, which decisions we can make regarding the null hypothesis H_0 and which statements (if any) we can make regarding the alternative hypothesis H_1.

Result	Decisions and statements we can make regarding the	
	Null hypothesis H_0	**Alternative hypothesis** H_1
$p < \alpha$	Reject H_0	Data supports H_1
$p \not< \alpha$	Fail to reject H_0	–

Summary: making decisions and statements about hypotheses

- We **specify** our null and alternative hypotheses **in advance** of the experiment.
- Our decision about the null hypothesis is based on **data from an experiment**.
- We **never 'accept'** a hypothesis as being true after analysing the data.
- We either **reject** or **fail to reject** the null hypothesis H_0.
- Failing to reject H_0 does not mean it is 'accepted'; the result of the experiment is **inconclusive**.
- If H_0 is rejected, we may say that **the data supports** the alternative hypothesis H_1.

4.3 Setting up a hypothesis test

The following quote from Ronald Fisher [8], who invented hypothesis testing, shows that he regarded the only conclusive result of a hypothesis test to be one where the H_0 is rejected.

> 'Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.'

So, the goal of an experiment is to collect data to disprove the null hypothesis H_0. Although it is counterintuitive, we could use the following scheme when setting up our hypothesis tests:

1. Make the conclusion you wish the data to support the **alternative hypothesis** H_1.
2. Then make the opposite conclusion the **null hypothesis** H_0.
3. The null hypothesis then provides our assumptions for our random variables and the distribution of the test statistic.

The reason this is confusing is that this is 'thinking backwards'. But let us see an example of how this is used in the real world.

Example 4.3.1. A pharmaceutical company develops a drug for curing a particular disease. They need to run clinical trials to show the drug works. They then plan their hypothesis test as follows:

- The goal is to show that the drug is an effective treatment for the disease, so this will be the alternative hypothesis.
- The null hypothesis must then be the opposite, and neutral, so the null hypothesis is that the drug has no effect.
- The statistical analysis of the data will use the assumption that the drug has no effect.

After doing this planning, before collecting the data they specify the null hypothesis H_0 and alternative hypothesis H_1 as follows:

H_0 : the drug has **no effect** on the disease,
H_1 : the drug is **an effective treatment** the disease.

△

In practice, our hypothesis tests involve testing the values of certain parameters. Let us look at a few abstract examples.

Example 4.3.2. Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed according to a normal distribution with unknown mean θ and known variance σ^2. The goal is to show that θ does not equal some value θ_0 (we do not care if $\theta < \theta_0$ or $\theta > \theta_0$, we just want to show θ is not equal to θ_0). We would then specify the null hyothesis H_0 and alternative hypothesis H_1 is as follows:

$$H_0 : \theta = \theta_0,$$
$$H_1 : \theta \neq \theta_0.$$

△

Example 4.3.3. Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed according to a normal distribution with unknown mean θ and known variance σ^2. The goal is to show that θ is **greater than** some value θ_0. We would then specify the null hyothesis H_0 and alternative hypothesis H_1 is as follows:

$$H_0 : \theta \leq \theta_0,$$
$$H_1 : \theta > \theta_0.$$

△

Example 4.3.4. Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed according to a normal distribution with unknown mean θ and known variance σ^2. The goal is to show that θ is **less than** some value θ_0. We would then specify the null hyothesis H_0 and alternative hypothesis H_1 is as follows:

$$H_0 : \theta \geq \theta_0,$$
$$H_1 : \theta < \theta_0.$$

△

Remark 4.3.5. The alternative hypothesis H_1 in Example 4.3.3 is an example of a **one-sided** alternative, because in this case $\theta \in (\theta_0, \infty)$, which is a 'one-sided' interval. Similarly, the alternative hypothesis in Example 4.3.4 is a one-sided alternative. On the other hand, the hypothesis in Example 4.3.2 is a an example of a **two-sided** alternative, because if $\theta \neq \theta_0$, then $\theta < \theta_0$ or $\theta > \theta_0$. These different types of hypotheses give rise to different hypothesis tests which are computed a little differently; we shall see this in the next section. □

Remark 4.3.6. Note that in the case of one-sided hypotheses, the null hypothesis usually includes the value of the boundary point, e.g. $\theta \leq \theta_0$, rather than $\theta < \theta_0$, as in Examples 4.3.3 and 4.3.4 above. □

4.4 Hypothesis testing based on one sample

In this section all the hypothesis tests investigate the value of a parameter of independent and identically distributed random variables from a single sample $X_1, X_2, \ldots, X_n$. For simplicity, we assume that these independent random variables each follow a $\mathrm{N}(\theta, \sigma^2)$ distribution where the mean θ is unknown and the variance σ^2 is known, and our hypotheses concern the value of θ.

We start by looking at the case where the alternative is two-sided, since this will be more familiar to us from computing two-sided confidence intervals; in fact we will see that there is a relationship between confidence intervals and hypothesis tests. We then look at the case of testing one-sided hypotheses.

4.4.1 Testing two-sided hypotheses

Suppose that, as described above, $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. Suppose we specify our null hypothesis H_0 and alternative hypothesis H_1 as

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta \neq \theta_0. \end{aligned}$$

If we assume the null hypothesis H_0 is true, then

$$\begin{aligned} &\mathrm{E}[X_i] = \theta = \theta_0, \qquad \text{for } i = 1, 2, \ldots, n. \\ \Rightarrow &\mathrm{E}[\overline{X}] = \theta_0 \\ \Rightarrow &\overline{X} \sim \mathrm{N}\left(\theta_0, \frac{\sigma^2}{n}\right) \\ \Rightarrow &Z = \frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1). \end{aligned}$$

Suppose the signifance level is specified as α. Then, just as we derived the confidence interval in Section 3.3.1,

$$\begin{aligned} &\mathrm{P}\left(z_{\alpha/2} < Z < z_{1-\alpha/2}\right) = 1 - \alpha. \\ \Rightarrow &\mathrm{P}\left(z_{\alpha/2} < \frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha \\ \Rightarrow &\mathrm{P}\left(\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta_0 < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \end{aligned}$$

Then,

$$\begin{aligned} &\mathrm{P}\left(\theta_0 \in \left(\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha \\ \Rightarrow &\mathrm{P}\left(\theta_0 \notin \left(\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) = \alpha \end{aligned}$$

Then, suppose that the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are observed as the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we can compute the sample mean $\overline{x}$ and then the realization of the interval as

$$\left(\overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

If $\theta_0 \notin \left(\overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\right)$, then the null hypothesis H_0 is rejected, because the probability of this occurring was considered, in some sense, to be less than α.

Therefore, computing the $(1-\alpha)$-confidence interval allows us to decide whether or not to reject the null hypothesis.

Example 4.4.1. Suppose there are 250 students in a class and we wish to make a statement about the average height of a student in this class, where we denote the average height by θ. One way to determine the value of θ would be measure the height of all the students in the class and compute the mean of these heights. However, we guess that the average height is $\theta_0 = 180$cm, and we want to confirm whether or not our guess is correct within 10 minutes. So, we plan to measure the heights of 9 students and check our hypothesis is correct based on this sample (one minute to measure each student's height, and one minute to do the calculation!). We specify our null hypothesis H_0 and alternative hypothesis H_1 as follows:

$$H_0 : \theta = 180,$$
$$H_1 : \theta \neq 180.$$

Suppose it is assumed that the heights of these students follow a normal distribution with unknown mean θ and variance σ^2, and assume it is known that $\sigma^2 = 4$. We label the heights of the nine students we plan to measure as $X_1, X_2, \ldots, X_9$. Then we define

$$Z = \frac{\theta - \overline{X}}{\sigma/\sqrt{n}}, \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

where in this case $n = 9$. Before we collect the data, we set our significance threshold as $\alpha = 0.01$. Then, according to our work above, we would reject the null hypothesis if

$$\theta_0 \notin \left(\overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

where here we have $\theta_0 = 180$, $\alpha = 0.01$, and from Table 3.1 (or Table 3.2) $z_{0.995} = 2.576$ and $z_{0.005} = -2.576$.

Suppose we measure the heights of nine randomly selected students as $x_1, x_2, \ldots, x_9$, and from these nine values compute $\overline{x} = 177$. Then the interval becomes

$$\left(177 - 2.576 \cdot \frac{2}{3}, 177 + 2.576 \cdot \frac{2}{3}\right) = (175.28, 178.72).$$

Since $180 \notin (175.28, 178.72)$, we reject the null hypothesis. △

4.4.1.1 Computing the p-value for a two-sided hypothesis test

Computing the confidence interval above was enough to decide whether or not to reject H_0. However, we did not explicitly obtain the p-value above. Our test statistic $Z = \frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}}$ is assumed to follow a $\mathrm{N}(0,1)$ distribution, and we would compute a 'p-value' as $\widetilde{p} = 1 - F_Z(z)$, where $z = \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}}$ is the realization of Z after observing the data $\mathbf{x}$. Above we showed that $\theta_0 \notin \left(\overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \right)$ will result in H_0 being rejected. We see that

$$\theta_0 < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}} < z_{\alpha/2} \;\Rightarrow\; z < z_{\alpha/2}, \;\Rightarrow\; F_Z(z) < \frac{\alpha}{2} \;\Rightarrow\; \widetilde{p} = 1 - F_Z(z) > 1 - \frac{\alpha}{2}$$

or

$$\theta_0 > \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \;\Rightarrow\; \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \;\Rightarrow\; z > z_{1-\alpha/2} \;\Rightarrow F_Z(z) > 1 - \frac{\alpha}{2}$$
$$\Rightarrow \widetilde{p} = 1 - F_Z(z) < \frac{\alpha}{2}.$$

So the values of the quantity $\widetilde{p}$ that will result in the null hypothesis being rejected are $\widetilde{p} < \frac{\alpha}{2}$ or $\widetilde{p} > 1 - \frac{\alpha}{2}$. This is not really a p-value, since a p-value would be a quantity p such that $p < \alpha$ would result in the null hypothesis being rejected. So, what can we do? As the diagram shows below, we need to transform the quantity $\widetilde{p}$ to the quantity p, by mapping the two end regions below to a larger lower region:

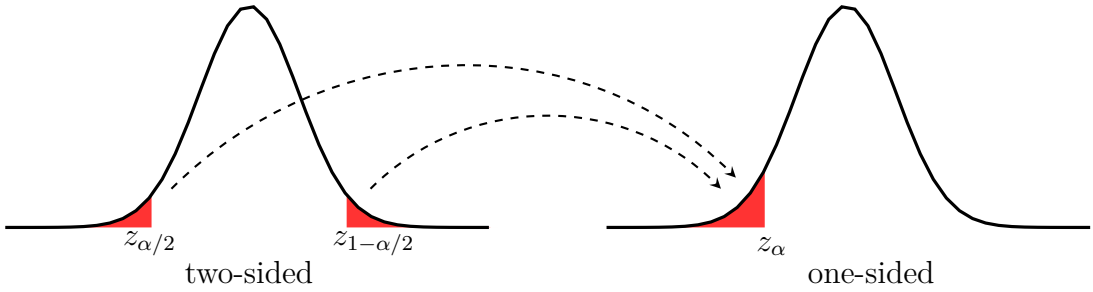


Figure 4.1: Mapping the two-sided critical regions to a one-sided critical region.

What would a suitable transformation be? The following proposition answers this question.

Proposition 4.4.2. Suppose that $\widetilde{p}$ is a quantity such that, for some value $\alpha \in (0,1)$ either $\widetilde{p} < \frac{\alpha}{2}$ or $\widetilde{p} > 1 - \frac{\alpha}{2}$. Then defining

$$p = 1 - 2\left|\widetilde{p} - \tfrac{1}{2}\right|,$$

results in $p < \alpha$. Moreover, if $\widetilde{p} = \frac{\alpha}{2}$ or $\widetilde{p} = 1 - \frac{\alpha}{2}$, then $p = \alpha$. ♦

Proof. The proof involves manipulating a few inequalities. Since $\alpha < 1$, this implies $\frac{\alpha}{2} < \frac{1}{2}$. Then

$$\begin{aligned}
&\widetilde{p} < \tfrac{\alpha}{2} \\
\Rightarrow &\widetilde{p} - \tfrac{1}{2} < \tfrac{\alpha}{2} - \tfrac{1}{2} < 0 \\
\Rightarrow &\left|\widetilde{p} - \tfrac{1}{2}\right| = \tfrac{1}{2} - \widetilde{p} > \tfrac{1}{2} - \tfrac{\alpha}{2} \\
\Rightarrow &-2\left|\widetilde{p} - \tfrac{1}{2}\right| < -2\left(\tfrac{1}{2} - \tfrac{\alpha}{2}\right) = -1 + \alpha \\
\Rightarrow &1 - 2\left|\widetilde{p} - \tfrac{1}{2}\right| < \alpha.
\end{aligned}$$

The other inequality is similar:

$$\begin{aligned}
&\widetilde{p} > 1 - \tfrac{\alpha}{2} \\
\Rightarrow &\widetilde{p} - \tfrac{1}{2} > 1 - \tfrac{\alpha}{2} - \tfrac{1}{2} = \tfrac{1}{2} - \tfrac{\alpha}{2} > 0 \\
\Rightarrow &\left|\widetilde{p} - \tfrac{1}{2}\right| = \widetilde{p} - \tfrac{1}{2} > \tfrac{1}{2} - \tfrac{\alpha}{2} \\
\Rightarrow &-2\left|\widetilde{p} - \tfrac{1}{2}\right| < -2\left(\tfrac{1}{2} - \tfrac{\alpha}{2}\right) = -1 + \alpha \\
\Rightarrow &1 - 2\left|\widetilde{p} - \tfrac{1}{2}\right| < \alpha.
\end{aligned}$$

So in either case, if $\widetilde{p} < \frac{\alpha}{2}$ or $\widetilde{p} > 1 - \frac{\alpha}{2}$, setting $p = 1 - 2\left|\widetilde{p} - \frac{1}{2}\right|$ will ensure $p < \alpha$.

One can check that if $\widetilde{p} = \frac{\alpha}{2}$ or $\widetilde{p} = 1 - \frac{\alpha}{2}$, then $p = \alpha$ which finishes proving the result. □

Remark 4.4.3. Checking that $\widetilde{p} = \frac{\alpha}{2}$ and $\widetilde{p} = 1 - \frac{\alpha}{2}$ are mapped to $p = \alpha$ ensures that the boundary of the 'two-sided' regions are mapped to the boundary of the 'one-sided' region. □

Example 4.4.4. Let us return to Example 4.4.1, where we assume that the random variables $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2)$, where the mean θ is unknown and it is assumed $\sigma^2 = 4$. We specify our null and alternative hypotheses as

$$\begin{aligned}
H_0 &: \theta = 180, \\
H_1 &: \theta \neq 180.
\end{aligned}$$

The significance threshold is specified as $\alpha = 0.01$, and then $n = 9$ observations $x_1, x_2, \ldots, x_9$ are measured and $\overline{x} = 177$ is computed. Defining

$$z = \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}} = \frac{180 - 177}{2/3} = 4.5,$$

and then one computes

$$\begin{aligned}
\widetilde{p} &= 1 - F_Z(4.5) = 3.4 \times 10^{-6}, \\
\Rightarrow p &= 1 - 2\left|\widetilde{p} - \tfrac{1}{2}\right| = 6.8 \times 10^{-6} < 0.01 = \alpha,
\end{aligned}$$

which means that the null hypothesis H_0 will be rejected. △

4.4.2 Testing one-sided hypotheses

Suppose as before that $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. However, now we wish to show that θ is greater than a certain value θ_0 (see Example 4.1.2 above for a real-world example). We therefore specify our null hypothesis H_0 and alternative hypothesis H_1 as

$$\begin{aligned} H_0 &: \theta \leq \theta_0, \\ H_1 &: \theta > \theta_0. \end{aligned}$$

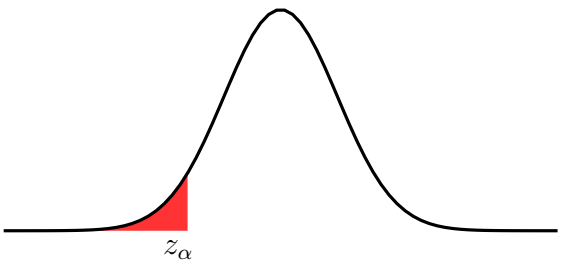
As before, assuming the null hypothesis is true, we have

$$\begin{aligned} &\overline{X} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right) \\ \Rightarrow &Z = \frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(\tfrac{\sqrt{n}}{\sigma}(\theta_0 - \theta), 1\right). \end{aligned}$$

Again, suppose we set the significance threshold as α. We now need to think a little carefully. Assuming H_0 is true,

$$\begin{aligned} &\mathrm{E}[\theta_0 - \overline{X}] = \theta_0 - \mathrm{E}[\overline{X}] = \theta_0 - \theta \geq 0 \\ \Rightarrow &\mathrm{E}[Z] = \tfrac{\sqrt{n}}{\sigma}(\theta_0 - \theta) \geq 0 \end{aligned}$$

Then it makes sense to consider the critical region (where it is unlikely for Z to occur) to be the lower critical region, where Z would take 'very' negative values, as shown in the figure below (recall that $z_\alpha < 0$).



Then

$$\begin{aligned} &\mathrm{P}(Z \geq z_\alpha) = 1 - \alpha. \\ \Rightarrow &\mathrm{P}\left(\frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} \geq z_\alpha\right) = 1 - \alpha \\ \Rightarrow &\mathrm{P}\left(\theta_0 \geq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \\ \Rightarrow &\mathrm{P}\left(\theta_0 < \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = \alpha. \end{aligned}$$

So if, after observing the data $x_1, x_2, \ldots, x_n$ and computing $\overline{x}$ we have $\theta_0 < \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$, then the null hypothesis would be rejected. In this case, note that it makes sense to compute

$$p = F_Z(z), \qquad \text{where } z = \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}}$$

as the p-value, since $p < \alpha$ will mean the null hypothesis is rejected.

One point to note: we are using z_α as the critical value rather than $z_{\alpha/2}$.

Example 4.4.5. As in Example 4.4.1, suppose again that we have 250 students in a class with heights denoted by the random variables $X_1, X_2, \ldots, X_n$ which we assume are independent and identically distributed with unknown mean θ and known variance $\sigma^2 = 4$. We wish to show that $\theta > \theta_0 = 174$cm, so we specify the null hypothesis H_0 and the alternative hypothesis H_1 by

$$H_0 : \theta \leq 174,$$
$$H_1 : \theta > 174.$$

We specify our significance threshold as $\alpha = 0.01$ and measure 9 student's heights as $x_1, x_2, \ldots, x_9$, and compute the sample mean $\overline{x} = 177$. We compute

$$\overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}} = 177 - 2.576 \cdot \frac{2}{3} \approx 175.28,$$

and since $174 < 175.28$, we reject the null hypothesis. Alternatively, we could compute the p-value as

$$z = \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}} = \frac{174 - 177}{2/3} = -4.5$$
$$\Rightarrow p = F_Z(z) = F_Z(-4.5) \approx 3.4 \times 10^{-6} < 0.01 = \alpha.$$

△

Example 4.4.6. Suppose everything is the same as in Example 4.4.5 except that when the data $x_1, x_2, \ldots, x_9$ was recorded, the sample mean was computed as $\overline{x} = 175$. Since the null hypothesis and alternative hypotheses are

$$H_0 : \theta \leq 174,$$
$$H_1 : \theta > 174,$$

at first glance once might consider that since the sample mean is an estimate of the mean and $\overline{x} = 175 > 174$, we should immediately reject the null hypothesis; otherwise how can $\theta \leq 174$? However, thinking in this way does not take stochastic variation, i.e. the noise or variance of the measurements, into account. Recall that need to compute

$$\overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}} = 175 - 2.576 \cdot \frac{2}{3} \approx 173.28,$$

(where again we specified $\alpha = 0.01$), and here $174 > 173.28$, so the null hypothesis in this case fails to be rejected. △

Remark 4.4.7. While all the examples so far have assumed normality and assumed that the variance σ^2 is known, this was merely for the convenience of referring to the critical values as z_α and $z_{1-\alpha}$, etc. In the next example, we assume σ^2 is unknown and use the t-distribution. □

Example 4.4.8. Suppose, similarly to Example 4.4.5, the random variables $X_1, X_2, \ldots, X_n$ are assumed to follow a normal distribution with unknown mean θ and variance σ^2, but in this case σ^2 is also unknown. We specify null hypothesis and alternative hypotheses

$$H_0 : \theta \leq 174,$$
$$H_1 : \theta > 174,$$

and $\alpha = 0.01$. Suppose the data $x_1, x_2, \ldots, x_9$ was recorded, and the sample mean was computed as $\overline{x} = 175$ and the sample variance was computed as $s^2 = 6$. We use Student's t-distribution with $9-1 = 8$ degrees of freedom, and from Table 3.3 we obtain $t_{8,0.99} = 2.896$, so by the symmetry of the t-distribution probability density function, $t_{8,0.01} = -2.896$. Therefore, similarly to above, we could derive

$$\begin{aligned}
&T = \frac{\theta_0 - \overline{X}}{S/\sqrt{n}} \sim t_{n-1} \\
\Rightarrow &\mathrm{P}\left(T \geq t_{n-1,\alpha}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\frac{\theta_0 - \overline{X}}{S/\sqrt{n}} \geq t_{n-1,\alpha}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\theta_0 \geq \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\theta_0 < \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right) = \alpha.
\end{aligned}$$

Here we have $\theta_0 = 174$, $n = 9$, $\alpha = 0.01$, $t_{8,0.01} = -2.896$, $\overline{x} = 175$ and $s^2 = 6$, so

$$\begin{aligned}
&\overline{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}} = 175 - 2.986 \cdot \frac{\sqrt{6}}{3} = 172.56 \\
\Rightarrow &174 \geq 172.56
\end{aligned}$$

so the null hypothesis fails to be rejected.

Alternatively,

$$\mathrm{P}\left(T < t_{n-1,\alpha}\right) = \alpha$$

so if $t = \frac{\theta_0 - \overline{x}}{s/\sqrt{n}} < t_{n-1,\alpha}$, then the null hypothesis would be rejected. However,

$$t = \frac{\theta_0 - \overline{x}}{s/\sqrt{n}} = \frac{174 - 175}{\sqrt{6}/3} = -1.2247 \geq -2.896 = t_{8,0.01},$$

and so the null hypothesis is not rejected. △

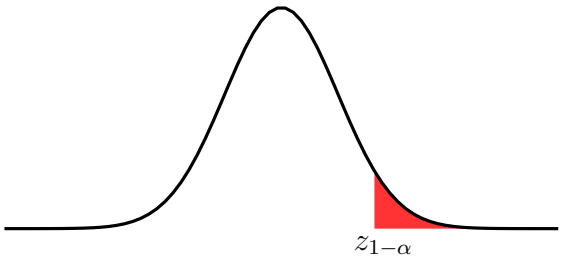
Remark 4.4.9. Another case to consider is if the null and alternative hypotheses are

$$H_0 : \theta \geq \theta_0,$$
$$H_1 : \theta < \theta_0.$$

In this case, the argument is symmetric, and we consider the critical region as in the figure below, since if $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\theta, \sigma^2)$, where θ is unknown and σ^2 is known,

$$\begin{aligned}
&\overline{X} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right) \\
\Rightarrow Z = \frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} &\sim \mathrm{N}(\theta_0 - \theta, 1) \\
\Rightarrow \mathrm{E}[Z] = \frac{\sqrt{n}}{\sigma}\mathrm{E}[\theta_0 - \overline{X}] = \frac{\sqrt{n}}{\sigma}\left[\theta_0 - \mathrm{E}[\overline{X}]\right] &= \frac{\sqrt{n}}{\sigma}\left[\theta_0 - \theta\right] \leq 0
\end{aligned}$$

under the null hypothesis.



Then as before,

$$\begin{aligned}
Z = \frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} &\sim \mathrm{N}(0, 1) \\
\Rightarrow &\mathrm{P}\left(Z \leq z_{1-\alpha}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\frac{\theta_0 - \overline{X}}{\sigma/\sqrt{n}} \leq z_{1-\alpha}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\theta_0 \leq \overline{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\theta_0 > \overline{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right) = \alpha.
\end{aligned}$$

□

Example 4.4.10. We return to Example 4.4.1, and again suppose there are 250 students in a class, and we wish to make a statement about the average height of a student in this class, which we denote by θ. We assume the heights of the students are independent and follow a normal distribution with mean θ, which is unknown, and variance σ^2, where we assume $\sigma^2 = 4$. Instead of just showing that $\theta \neq 180$ (where the units are cm), we decide that we want to show $\theta < 180$. So, we specify our null hyothesis H_0 and alternative hypothesis H_1 as follows:

$$H_0 : \theta \geq 180,$$
$$H_1 : \theta < 180.$$

We specify $\alpha = 0.01$, so $z_{1-\alpha} = z_{0.99} = 2.326$ We measure the heights of 9 students, denoting these heights as $x_1, x_2, \ldots, x_9$, and compute $\overline{x} = 177$. Then

$$\overline{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 177 + 2.326 \cdot \frac{2}{3} = 178.55.$$

Since

$$z = \frac{\theta_0 - \overline{x}}{\sigma/\sqrt{n}} = \frac{180 - 177}{2/3} = 4.5 > 2.326 = z_{1-\alpha},$$

so the null hypothesis is rejected. △

4.5 The distribution of a p-value

In this section we show that, under the null hypothesis, the p-value when considered as a random variable follows a uniform distribution on the interval $[0, 1]$; this is a fundamental result. Before showing this, we first prove the following general theorem:

Theorem 4.5.1. Let X be a random variable with continous cumulative distribution function F_X, and let $Y = F_X(X)$. Then Y is uniformly distributed on the interval $[0, 1]$. ♦

Proof. For simplicity, we assume that the inverse function of F_X, denoted F_X^{-1}, exists. The theorem can be proved without this assumption, but the proof then requires a few results from Analysis.

Now, since F_X is a c.d.f., for any $x \in \mathbb{R}$, $0 \leq F_X(x) \leq 1$. Therefore, $\mathrm{P}(Y \leq 0) = 0$ and $\mathrm{P}(Y > 1) = 0 \Rightarrow \mathrm{P}(Y \leq 1) = 1$. Now for any $y \in (0, 1)$, assuming the inverse function F_X^{-1} exists, one can compute

$$\begin{aligned} \mathrm{P}(Y \leq y) &= \mathrm{P}(F_X(X) \leq y) \\ &= \mathrm{P}\left(X \leq F_X^{-1}(y)\right) \\ &= F_X\left(F_X^{-1}(y)\right) \\ &= y. \end{aligned}$$

Therefore, the cumulative distribution function of the random variable Y is

$$F_Y(y) = \mathrm{P}(Y \leq y) = \begin{cases} 0, & \text{if } y \leq 0, \\ y, & \text{if } 0 < y < 1, \\ 1, & \text{if } y \geq 1. \end{cases} \tag{4.1}$$

Comparing this to the cumulative distribution function of a $\mathrm{Unif}(0, 1)$ random variable (see Definition 8.3.1 of Prof. Veraart's notes), this shows that $Y \sim \mathrm{Unif}(0, 1)$. □

Remark 4.5.2. Note that the same notation is used for the inverse function F_X^{-1} as for the quantile function of X (see Definition 1.6.24); however, although the quantile function always exists, the inverse of F_X^{-1} need not always exist. □

Corollary 4.5.3. Assuming the null hypothesis is true, a p-value based on a continuous test statistic follows a $\mathrm{Unif}(0, 1)$ distribution. ♦

Proof. Given data $\mathbf{x}$, under the null hypothesis the test statistic $t(\mathbf{x})$ is a realization of a random variable T which follows some distribution. The p-value $p = 1 - F_T(t(\mathbf{x}))$ is then a realization of $1 - F_T(T)$, which Theorem 4.5.1 proves is a $\mathrm{Unif}(0, 1)$ random variable. □

Remark 4.5.4. If the test statistic is discrete, rather than continuous, then the p-value also follows a discrete distribution, and so cannot be $\text{Unif}(0, 1)$; however, under the null hypothesis the p-value will follow a uniform discrete distribution that is 'as close as possible' to $\text{Unif}(0, 1)$. □

4.6 Type I and Type II errors

When deciding whether or not to reject the null hypothesis, we compare a p-value to a significance threshold α. We reject the null hypothesis if $p < \alpha$. Under the null hypothesis, however, the p-value follows a $\text{Unif}(0, 1)$ distribution so it is possible for $p < \alpha$ by chance, even when the null hypothesis is true. We would then have made an error in deciding to reject the null hypothesis. This error has a special name:

Definition 4.3. If the null hypothesis has been rejected, when in fact the null hypothesis is true, then we say a **Type I** error has occurred.

Remark 4.6.1. We read Type I error as 'type one error'. □

Remark 4.6.2. In practice, we **never know** if the null hypothesis is true or not, but it must be either true or false. We then calculate certain probabilities assuming we knew the truth about the null hypothesis. □

Remark 4.6.3. If a significance threshold of α is used, then the probability of this error occurring is $\text{P}(\text{Reject } H_0 \,|\, H_0 \text{ is true}) = \alpha$. So, whenever we reject a null hypothesis, we can be certain that the probability of a Type I error occurring is α. □

On the other hand, another type of error can occur. The null hypothesis may fail to be rejected, when it is actually false:

Definition 4.4. If the null hypothesis fails to be rejected, when in fact the null hypothesis is false, then we say a **Type II** error has occurred.

Remark 4.6.4. We read Type II error as 'type two error'. □

Remark 4.6.5. A Type II error could occur if we somehow got 'unlucky' with many of our observations coming from a tail of the distribution. For example, if in Example 4.4.10 where the null hypothesis is $H_0 : \theta \geq 180$ and our random selection of students that were measured were mostly tall students (when in fact there were very few tall students in the class), then the data provided would result in the null hypothesis failing to be rejected, when it is in fact false. □

Remark 4.6.6. We define the probability of a Type II error occurring as having value β (to go along with the Type I error probablity α), i.e. $\text{P}(\text{Fail to reject } H_0 \,|\, H_0 \text{ is false}) = \beta$. □

We would of course like our hypothesis tests to (correctly) reject the null hypothesis, i.e. when it is in fact false, with high probability. This probability has a special name:

Definition 4.5. The probability of correctly rejecting the null hypothesis, when in fact the null hypothesis is false, is defined as the **power** of the test, and is computed as $\mathrm{P}\left(\text{Reject } H_0 \,|\, H_0 \text{ is false}\right) = 1 - \beta$, where β is the probability of a Type II error occurring.

All of these quantities are summarised in the following table.

		Given that the null hypothesis H_0 is	
		True	False
Decision	Reject H_0	**Type I Error** $\mathrm{P}\left(\text{Reject } H_0 \,\vert\, H_0 \text{ is true}\right) = \alpha$	Correct decision: **Power** $\mathrm{P}\left(\text{Reject } H_0 \,\vert\, H_0 \text{ is false}\right) = 1 - \beta$
	Fail to reject H_0	Correct decision $\mathrm{P}\left(\text{Fail to reject } H_0 \,\vert\, H_0 \text{ is true}\right) = 1 - \alpha$	**Type II Error** $\mathrm{P}\left(\text{Fail to reject } H_0 \,\vert\, H_0 \text{ is false}\right) = \beta$

Table 4.1: A table summarising the definitions of Type I and Type II errors, as well as defining their respective probabilities of occurring, α and β. Note also the definition of statistical power, which is equal to $1 - \beta$.

Remark 4.6.7. In practice, we would like any test we perform to have low Type I error **and** low Type II error (i.e. high power). However, all we really have control over is the value of α; we cannot set β. In fact, the value of β turns out to rely on α, but in a non-obvious way: suppose we make α smaller and smaller, thereby reducing the Type I error of a test. This means that the data needs to be more and more extreme in order to obtain a p-value that will result in the null hypothesis being rejected. However, in this case only very extreme data will result in the null hypothesis being rejected. To go back to the height example, as we reduce α, we would need a smaller and smaller sample mean $\overline{x}$ in order to reject $H_0 : \theta \geq 180$. Even if the true (unknown) value of θ is $\theta = 178$, because of the low threshold α, it would be difficult to obtain data extreme enough to reject the null hypothesis, and so the p-value obtained after observing 'non-extreme' data would not be small enough to reject the null hypothesis; in other words, lowering the Type I error will result in increasing the Type II error. □

Remark 4.6.8. Remark 4.6.7 described why we have defined the quantities Type I error, Type II error and statistical power, and the problem with choosing a very small value for the significance threshold α. □

In the next section, we look at a classic real-world experiment that motivated some of the development of hypothesis testing.

4.7 An experiment: the lady tasting tea

This experiment, although not of a very serious nature, provides a good blueprint for all hypothesis testing problems.

Example 4.7.1. Two colleagues, a lady and a gentleman, take a break from work to enjoy a cup of tea. The gentleman makes two cups of tea for them, by putting the hot water in the cup followed by milk, and offers a cup to the lady. However, before making the second cup, the lady stops him and asks to rather put milk in the cup first, followed by the hot water. The gentleman is surprised, and asks the lady if she can taste the difference between the two processes of making tea. The lady asserts she can - and that she prefers a cup of tea to be made with milk poured into the cup before the hot water (we shall call this 'milk-first'). The gentleman is amazed, and asks if they can conduct an experiment to prove her claim. The lady agrees. △

In this experiment - which really took place between Ronald Fisher and his colleague Muriel Bristol - the lady and gentleman need to devise a procedure by which the lady can provide evidence that her claim is true.

The null hypothesis in this case is:

> H_0: The lady has no ability to discriminate between the different processes of making tea.

In other words, the assumption is that the lady cannot correctly identify if a cup of tea was made milk-first or tea-first.

The alternative hypothesis is the complementary hypothesis

> H_1: The lady has the ability to discriminate between the different processes of making tea and can identify how a cup of tea is made, either milk-first or tea-first.

4.7.1 The experimental setup

The lady and gentleman agree on the following experiment: the gentleman will make eight cups of tea, four of which will be milk-first and four of which will be tea first. The cups of tea will be presented to the lady in a **random** order, and her task will be to declare which four cups are made milk-first (the other four therefore being identified as tea-first).

Exercise 4.7.2. Under the null hypothesis, what is the probablity of the lady identifying all four of the milk-first cups correctly?

One can compute the probability using elementary principles. Choosing a group of 4 objects out of 8:

If one sequentially chooses the objects, one has 8 choices for the first object, then 7 choices for the second object, 6 choices for the third and 5 choices for the fourth, i.e. $8 \times 7 \times 6 \times 5 = 1680$ ways

However, one need to account for the order; there are $4 \times 3 \times 2 \times 1$ possible orders of 4 objects. Therefore, the number of ways of choosing 4 objects out of 8 is

$$\frac{8 \times 7 \times 6 \times 5}{4 \times 3 \times 2 \times 1} = \frac{8 \times 7 \times 6 \times 5 \times (4 \times 3 \times 2 \times 1)}{4 \times 3 \times 2 \times 1 \times (4 \times 3 \times 2 \times 1)} = \frac{8!}{4!4!} = \binom{8}{4} = 70$$

There is only one way of choosing 4 objects (tea-first cups) out of 4, so probability is $p = \frac{1}{70}$.

△

In other words, the p-value for the lady correctly identifying all four milk-first cups is $p = \frac{1}{70} \approx 0.014$. Therefore, if the lady correctly identified all four milk-first cups, the gentleman would be able to reject the null hypothesis at a threshold $\alpha = 0.05$, or even at the threshold $\alpha = 0.02$. In this case, there is then evidence to suggest that the lady can discriminate between the two processes of making tea.

Exercise 4.7.3. Under the null hypothesis, what is the probablity of the lady identifying exactly three out of the four milk-first cups correctly?

One already has that there are 70 ways of choosing 4 objects out of 8.

Now, in order to choose exactly 3 out of the 4 milk-first cups, the lady needs to select 3 milk-first cups and 1 tea-first cup.

The lady can choose exactly 3 out of the 4 milk-first cups in $\binom{4}{3} = 4$ ways.

Furthermore, the lady can choose exactly 1 out of the 4 tea-first cups in $\binom{4}{1} = 4$ ways.

Therefore, the probability of choosing exactly 3 out of the 4 milk-first cups is

$$p = \frac{4 \times 4}{70} = \frac{16}{70}.$$

△

So, if the lady manages to correctly identify **at least three** out of the four milk-first cups, the probability of this occurring is

$$p = \frac{1 + 16}{70} = \frac{17}{70} \approx 0.24.$$

This p-value is not significant at the $\alpha = 0.05$ level, so if the lady only managed to identify three out of the four milk-first cups, the gentleman would not reject the null hypothesis; in other words, the evidence is not strong enough to suggest that the lady can discriminate between the two processes of making tea.

Remark 4.7.4. Incidentally, when Fisher conducted this experiment with his colleague Muriel Bristol, she correctly identified all eight cups. □

In the next section we look at a popular statistical test.

4.8 Student's two-sample test

Consider the situation where we have two groups of random variables. The independent random variables $X_1, X_2, \ldots, X_n$ follow a $\mathrm{N}(\theta_1, \sigma_1^2)$ distribution where θ_1 is unknown and σ_1^2 is unknown, and the independent random variables $Y_1, Y_2, \ldots, Y_m$ follow a $\mathrm{N}(\theta_2, \sigma_2^2)$ distribution where θ_2 is unknown and σ_2^2 is unknown. (We also assume that each X_i is independent of each Y_j.) Suppose further that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are observed as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ are observed as $\mathbf{y} = (y_1, y_2, \ldots, y_m)$.

The question now is: is $\theta_1 = \theta_2$? Can we use the data $\mathbf{x}$ and $\mathbf{y}$ to answer this question? In general this question is difficult to answer, but in the special case where we can assume that $\sigma_1 = \sigma_2$ (even though their specific value is unknown), we can obtain an exact answer.

If we write $\sigma_1 = \sigma_2 = \sigma$ we know from Corollary 3.1.3 that

$$\overline{X} \sim \mathrm{N}\left(\theta_1, \frac{\sigma^2}{n}\right),$$

$$\overline{Y} \sim \mathrm{N}\left(\theta_2, \frac{\sigma^2}{m}\right) \qquad \Rightarrow \qquad -\overline{Y} \sim \mathrm{N}\left(-\theta_2, \frac{\sigma^2}{m}\right)$$

Then, using Proposition 3.1.1,

$$\overline{X} - \overline{Y} \sim \mathrm{N}\left(\theta_1 - \theta_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right).$$

If we defined

$$Z = \frac{\overline{X} - \overline{Y} - (\theta_1 - \theta_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

we could conclude that $Z \sim \mathrm{N}(0, 1)$. However, the problem is that we do not know the value of σ^2. We would like to mimic the construction of Student's t distribution in Section 3.4 by replacing σ^2 with some sort of sample variance S^2, but this would need to be a sample variance for the whole sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Defining

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2,$$

we know from Theorem 3.2.2 that

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1},$$

$$\Rightarrow \frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2},$$

since the X_i and Y_j random variables are all independent, and so S_X^2 and S_Y^2 are independent. Then using Theorem 3.4.4 it can be shown (exercise) that by defining

$$T = \frac{(\overline{X} - \overline{Y}) - (\theta_1 - \theta_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{4.2}$$

where

$$S_p^2 = \frac{1}{n+m-2}\left((n-1)S_X^2 + (m-1)S_Y^2\right), \tag{4.3}$$

is called the **pooled sample variance**, then $T \sim t_{n+m-2}$, i.e. Student's t-distribution with degrees of freedom equal to $n + m - 2$.

Let us summarise the assumptions:

- The two samples of random variables $\mathbf{X}$ and $\mathbf{Y}$ are independent and each follow a normal distribution.
- The means θ_1 and θ_2 of the two samples are unknown.
- The variances of the two samples are unknown but are assumed to be equal, i.e. $\sigma_1^2 = \sigma_2^2$.

Now, to answer the question of whether the two means θ_1 and θ_2 are equal or not, we define the null hypothesis H_0 and alternative hypothesis H_1 to be

$$H_0 : \theta_1 = \theta_2,$$
$$H_1 : \theta_1 \neq \theta_2.$$

Then, since under the null hypothesis $\theta_1 - \theta_2 = 0$, the t-statistic becomes

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Given the data, we can compute the statistic

$$t = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

and from this statistic, which under the null hypothesis follows a ***t*-distribution with $n+m-2$ degrees of freedom**, we can compute a p-value. If this p-value is below a desired significance threshold α, we would reject the null hypothesis at level α. If $T \sim t_{n+m-2}$, then

$$\mathrm{P}\left(|T| \leq t_{n+m-2,1-\alpha/2}\right) = 1 - \alpha,$$

so we would reject the null hypotheses if the realized statistic $|t| > t_{n+m-2,1-\alpha/2}$.

Note that rejecting the null hypothesis allows us to conclude that it is likely that the two means are not equal, i.e. $\theta_1 \neq \theta_2$; however, if the p-value is not significant and we fail to reject the null hypothesis it does not mean that $\theta_1 = \theta_2$! In such a case, where we fail to reject the null hypothesis, we cannot make any conclusion.

Example 4.8.1. A classic example would be to compare if the average heights of two groups of people are the same or not.

Suppose we visit country A and randomly select n people for a study. Their heights are the random variables $X_1, X_2, \ldots, X_n$, which are assumed to be independent and identically distributed following a normal distribution with unknown mean θ_1 and unknown variance σ_1^2. We measure these heights as $x_1, x_2, \ldots, x_n$, in cm.

Now suppose we visit a second county, B, and randomly select m people for a similar study. Their heights are the random variables $Y_1, Y_2, \ldots, Y_m$ which are assumed to follow a normal distribution with unknown mean θ_2 and unknown variance σ_2^2. Again, we measure these heights as $y_1, y_2, \ldots, y_m$, in cm.

We decide to assume $\sigma_1^2 = \sigma_2^2$, and specify the null hypothesis and alternative hypothesis as follows:

$$H_0 : \theta_1 = \theta_2,$$
$$H_1 : \theta_1 \neq \theta_2.$$

We specify the significance threshold as $\alpha = 0.05$. We have samples $n = 20$, $m = 25$, so $t_{n+m-2,1-\alpha/2} = t_{43,0.995}$. Now, Table 3.3 does not contain a row for 43 degrees of freedom, and it is worth discussing how to deal with such a situation. One option is to be more conservative and use the row for 40 degrees of freedom, so $t_{40,0.995} = 2.704$, we discuss other approaches in the remark below.

From the samples we compute that $\overline{x} = 178.5$, $\overline{y} = 176.2$, and the sample variance of the $\mathbf{x}$ sample is $s_x^2 = 4.5$ and the sample variance for the $\mathbf{y}$ sample is $s_y^2 = 3.8$. The pooled variance is computed as

$$\begin{aligned} s_p^2 &= \frac{1}{n+m-2}\left((n-1)s_x^2 + (m-1)s_y^2\right) \\ &= \frac{1}{20+25-2}\left((20-1)(4.5) + (25-1)(3.8)\right) \approx 4.109... \end{aligned}$$

Then,

$$t = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{178.5 - 176.2}{\sqrt{(4.109...)}\sqrt{\frac{1}{20} + \frac{1}{25}}} = 3.782$$

and since $|3.782| > 2.704$ the null hypothesis is rejected, and the data supports the case that the average height of people in countries A and B is different. $\triangle$

Remark 4.8.2. Above we used $t_{40,0.995}$ because $t_{43,0.995}$ was not available in our table. Another approach would be to attempt to get an approximate value for $t_{43,0.995}$ by taking the average of $t_{40,0.995}$ and $t_{45,0.995}$, i.e.

$$t_{43,0.995} \approx \tfrac{1}{2}\left(t_{45,0.995} + t_{40,0.995}\right) = \tfrac{1}{2}\left(2.690 + 2.704\right) = 2.697,$$

since we notice that the columns are all monotonically decreasing, which makes sense as the higher the degrees of freedom, the 'less fat' the tails. This is what a statistician might do in the past, before computers were so widely accessible. Now we could simply use R to compute the quantile value, and we see in the code below that $t_{43,0.995} = 2.695$ to three decimal places. Note that this is very close to our approximation 2.697. □

```
# compute the 0.995 quantile of the t-dist. with 43 degrees of freedom
qt(0.995, df=43)
#> [1] 2.6951
```

Remark 4.8.3. One may wonder if it is possible to relax the assumption that $\sigma_1 = \sigma_2$. Although it is beyond the scope of this course, there is a related test, known as Welch's test, which does not need to the assumption $\sigma_1 = \sigma_2$. However, this test is not exact, because the test statistic only **approximately** follows a t-distribution. □

Chapter 5

Pitfalls in Statistics

This section looks at four situations where one must be careful to interpret and make decisions based on statistics computed from data.

5.1 Correction for multiple hypothesis testing

In hypothesis testing, after assuming a null hypothesis to be true, one computes a p-value from the data given. If this p-value is close to 0, it indicates that data may not be generated according to the assumptions of the null hypothesis. One usually sets a significance threshold α, for example, $\alpha = 0.05$ or $\alpha = 0.01$ are common choices, and if our computed p-value is less than α, i.e. $p < \alpha$, we declare the p-value to be significant.

However, there are situations when we may compute multiple p-values simulataneously, and then care must be exercised before declaring a p-value to be significant.

Example 5.1.1. Suppose a pharmaceutical company is testing the effectiveness of 10 different medications for treating a particular disease, with each medication tested in its own clinical trial. The null hypothesis for each clinical trial is that the medication does not cure the disease. Data are collected, and the following p-values are collected:

Trial	p-value
1	0.020
2	0.300
3	0.003
4	0.006
5	0.400
6	0.010
7	0.100
8	0.700
9	0.250
10	0.090

The scientist in charge of all the clinical trials specified a desired significance threshold of $\alpha = 0.05$ in advance, and then declares the p-values from trials 1, 3, 4 and 6 are all significant (since 0.02, 0.003, 0.006 and 0.01 are all below the threshold $\alpha = 0.05$). However, is this conclusion correct? △

In fact, the scientist has made an error. Recall that for any two events A_1 and A_2,

$$\mathrm{P}(A_1 \cup A_2) = \mathrm{P}(A_1) + \mathrm{P}(A_2) - \mathrm{P}(A_1 \cap A_2) \leq \mathrm{P}(A_1) + \mathrm{P}(A_2),$$

since $\mathrm{P}(A_1 \cap A_2) \geq 0$. This can be generalised to n events $A_1, A_2, \ldots, A_n$, i.e.

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathrm{P}(A_i),$$

and in fact this is known as **Boole's Inequality**. Now, consider the example above. Let A_i be the event that $p_i < \alpha$, where p_i is the p-value computed from the data for the ith clinical trial. Then, since under the null hypothesis $p_i \sim U(0,1)$, $\mathrm{P}(A_i) = \alpha$.

Let A be the event that for at least one index $i \in \{1, 2, \ldots, n\}$, there is a p-value p_i such that $p_i < \alpha$. Then,

$$\mathrm{P}(A) = \mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathrm{P}(A_i) = \sum_{i=1}^{n} \alpha = n\alpha.$$

Therefore, in the example above, the probability of at least one p-value being significant is $\leq 10 \cdot 0.05 = 0.5$, in other words, up to a 50% chance!

There are several approaches to remedy this situation, but one of the simplest is to use the **Bonferroni correction**: if there are n tests, and the nominal signficance threshold is α, then one should use the adjusted significance threshold $\alpha' = \alpha/n$. Then, if we let $\widetilde{A}_i$ be the event that $p_i < \alpha/n$, and $\widetilde{A}$ be the event that at least one p_i is less than α/n, then

$$\mathrm{P}\left(\widetilde{A}\right) = \mathrm{P}\left(\bigcup_{i=1}^{n} \widetilde{A}_i\right) \leq \sum_{i=1}^{n} \mathrm{P}\left(\widetilde{A}_i\right) = \sum_{i=1}^{n} \frac{\alpha}{n} = n \cdot \frac{\alpha}{n} = \alpha.$$

Example 5.1.2. Returning to Example 5.1.1, if the desired signifance threshold is $\alpha = 0.05$, and there are $n = 10$ tests, then the adjusted significance threshold using the Bonferroni correction is $\frac{\alpha}{n} = \frac{0.05}{10} = 0.005$. Then, the only significant p-value is that of the third trial, $p_3 = 0.003$. △

Theorem 5.1.3 (Boole's Inequality)**.** For a set of events A_i, $i = 1, 2, \ldots n$,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathrm{P}(A_i).$$

♦

The proof of Boole's Inequality is left as an exercise (hint: use induction).

5.2 Spurious correlations

It is tempting to think that when two variables are highly correlated, e.g. $\rho_{XY} \geq 0.7$, that the two variables are related and perhaps one of the variables influences the other; in other words, that there is some form of **causation** between the two variables.

For example, it may be the case that the random (Bernoulli) variable of a customer in a supermarket buying bread or not is correlated with the random (Bernoulli) variable of a customer buying milk or not.

Now, while there are cases where there is causation between two random variables, please always remember this:

> **Correlation does not imply causation.**

The following are examples taken from the website [15]. Each examples shows two statistics that are highly correlated, but it should be clear that they are not related by any form of causation.

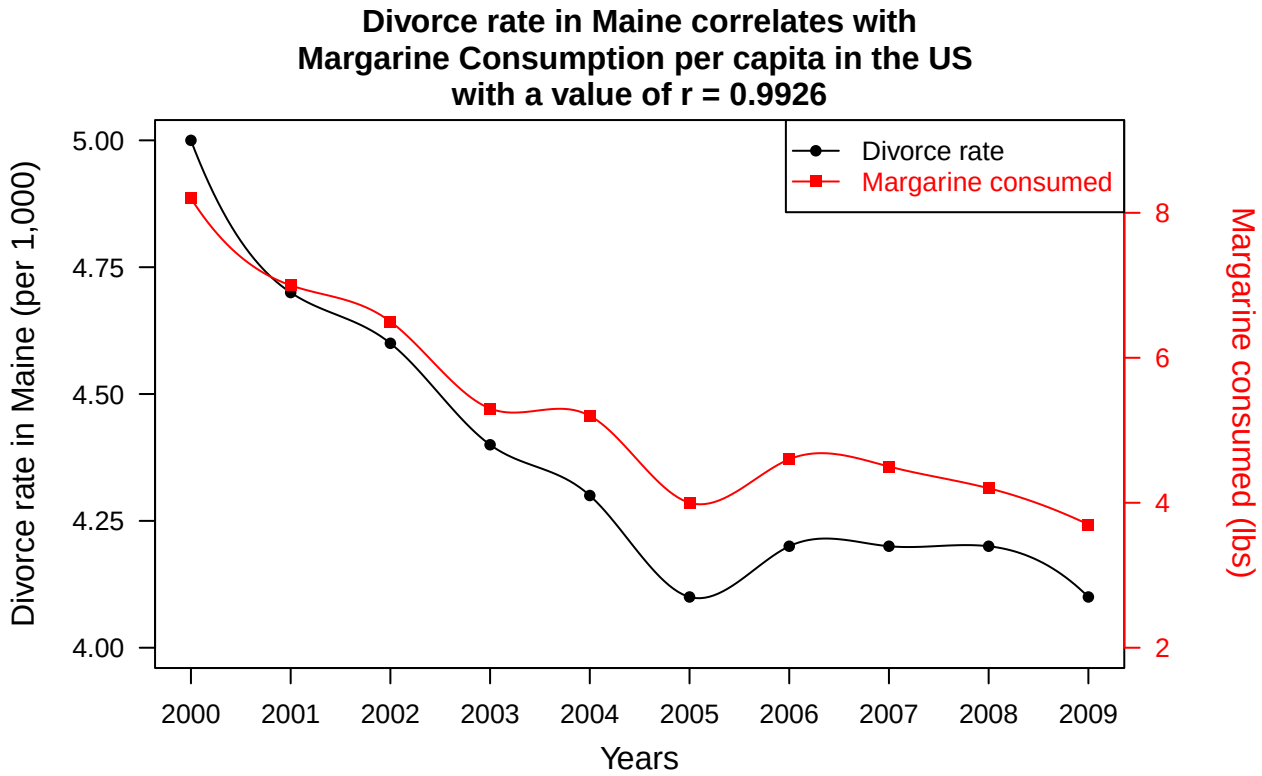
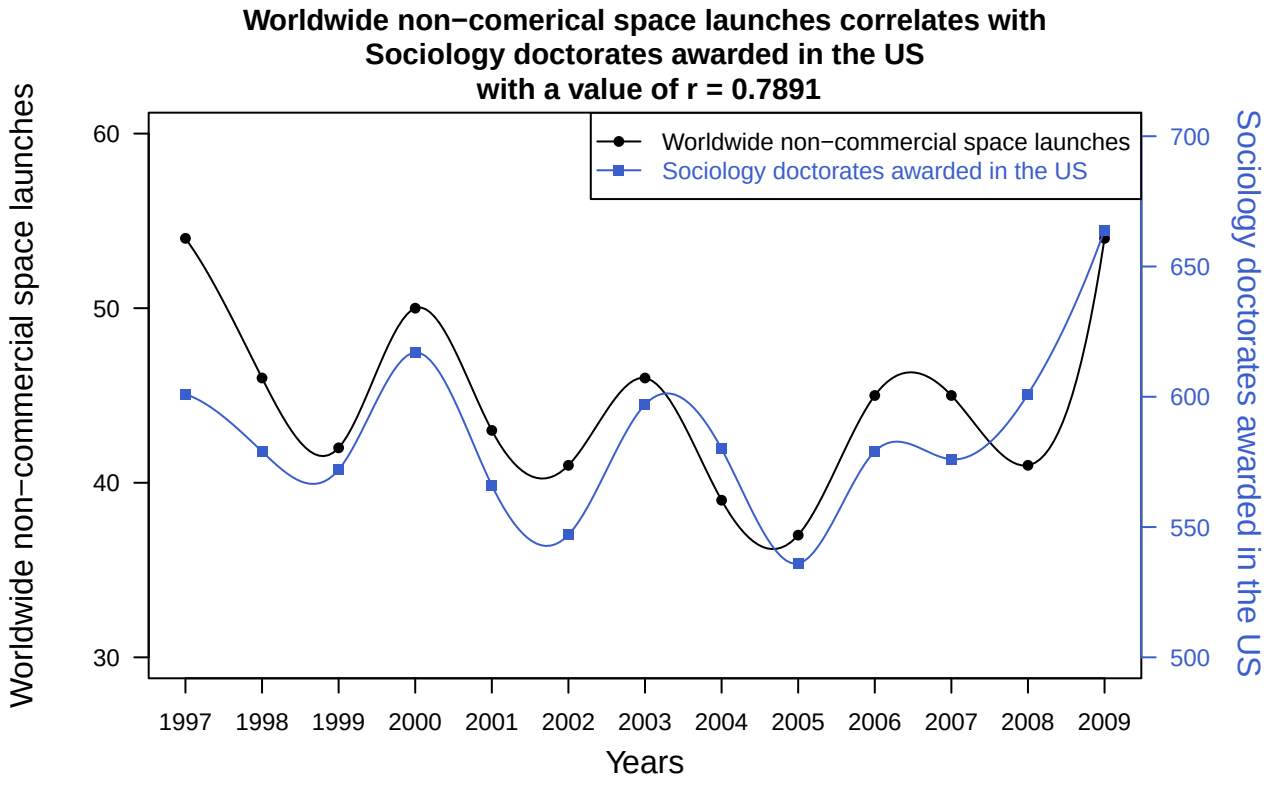
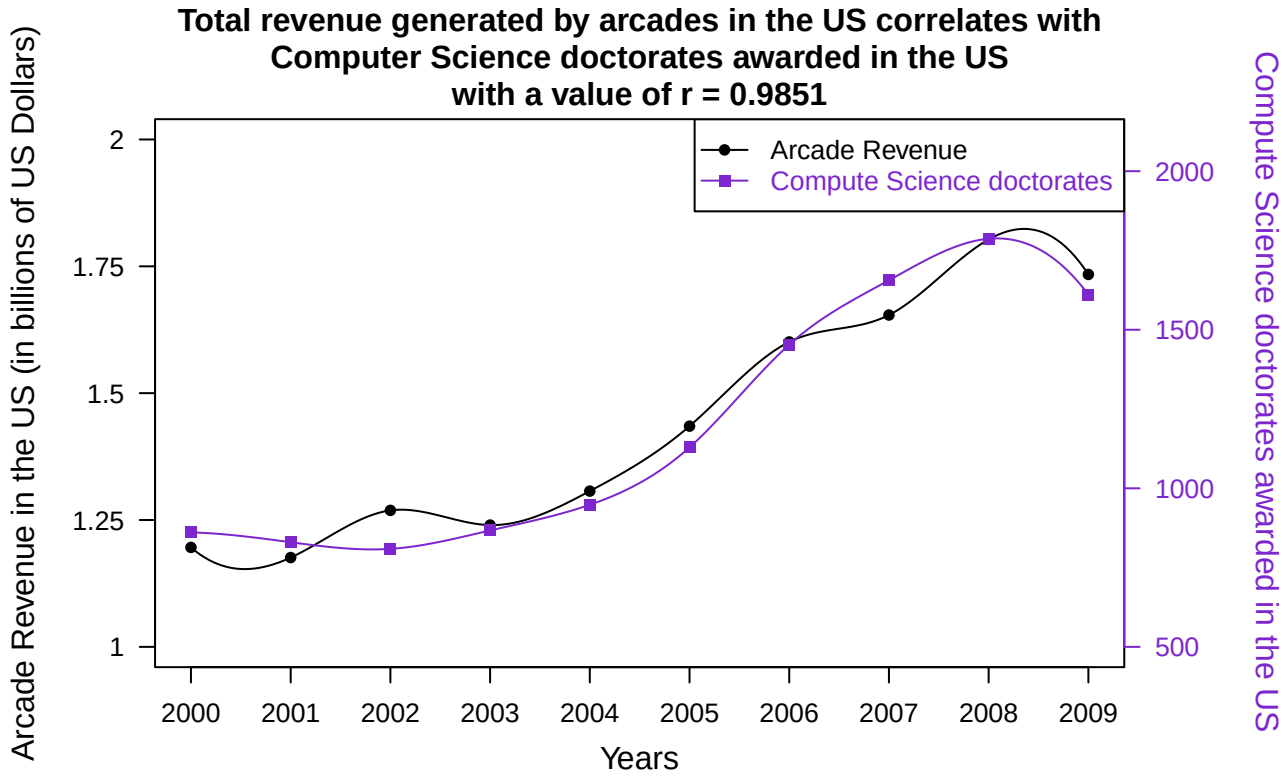


Figure 5.1: Margarine consumption and the divorce rate in Maine are highly correlated, but it is unlikely there is a causal relationship between these two quantities.


Worldwide non−comerical space launches correlates with
Sociology doctorates awarded in the US
with a value of r = 0.7891
Worldwide non−commercial space launches
Sociology doctorates awarded in the US
Worldwide non−commercial space launches
Sociology doctorates awarded in the US
60
50
40
30
700
650
600
550
500
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
Years


Total revenue generated by arcades in the US correlates with
Computer Science doctorates awarded in the US
with a value of r = 0.9851
Arcade Revenue
Compute Science doctorates
Arcade Revenue in the US (in billions of US Dollars)
Compute Science doctorates awarded in the US
2
1.75
1.5
1.25
1
2000
1500
1000
500
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
Years

5.3 Simpson's paradox

Simpson's paradox is a phenomenon that allows a statistical trend in groups of data to be reversed if these data are aggregated together. A classic example that illustrates this phenomenon comes from the survival records of the sinking of the ship the RMS *Titanic* in 1912. Table 5.1 shows the survival rates of third-class adult passengers (there were three classes of passenger, First, Second and Third) and crew members.

The survival rate in each group is computed by dividing the number in the group that survived by the total number in the group and expressing the result as a percentage, i.e. $151/627 = 24.08\%$. Table 5.1 shows that the survival rate among crew members was higher (and so is highlighted in bold) than among crew members.

Adult passengers and crew members				
Class	Survived	Did not survive	Total	Survival rate
Third	151	476	627	**24.08 %**
Crew	212	673	885	23.95 %

Table 5.1: The survival rate of adult 3rd-class passengers and crew members on the Titanic.

However, one may be interested in the survival rates of men and women in these two groups. If we disaggregate (split) Table 5.1 by gender, we obtain Table 5.2, and this is where something unusual is observed. In this table, looking at the subgroup of men, the numbers show that the survival rate of male crew members was higher than the survival rate of male third-class passengers. Not only that, the survival rate of female crew members was higher than the survival rate of female third-class passengers. But this seems impossible—how can the survival rate now be higher among crew in both the male and female subgroups? Check the numbers if you like—this is no mathematical error.

Men					Women				
Class	Survived	Did not survive	Total	Survival rate	Class	Survived	Did not survive	Total	Survival rate
Third	75	387	462	16.23%	Third	76	89	165	46.06%
Crew	192	670	862	**22.27%**	**Crew**	20	3	23	**86.96%**

Table 5.2: The survival rates of crew members and third-class passengers on the Titanic when the data is split into male and female subgroups.

This statistical phenomenon is actually caused by combining ratios when the two subgroups are very different in size. Notice how only $23/(862 + 23) < 3\%$ of the crew are women, but $165/(165 + 462) > 25\%$ of the third-class passengers are women. In other words, the proportions of men and women in the crew and third-class passengers are very different. Furthermore, the survival rate among men is far lower than the survival rate among women in the two categories. This is the main reason why Simpson's paradox occurs, because the ratios between the two groups are very different.

5.3.1 Proof of Simpson's paradox

An elegant proof for why Simpson's paradox can occur is from [10], and is shown in Figure 5.2 below.

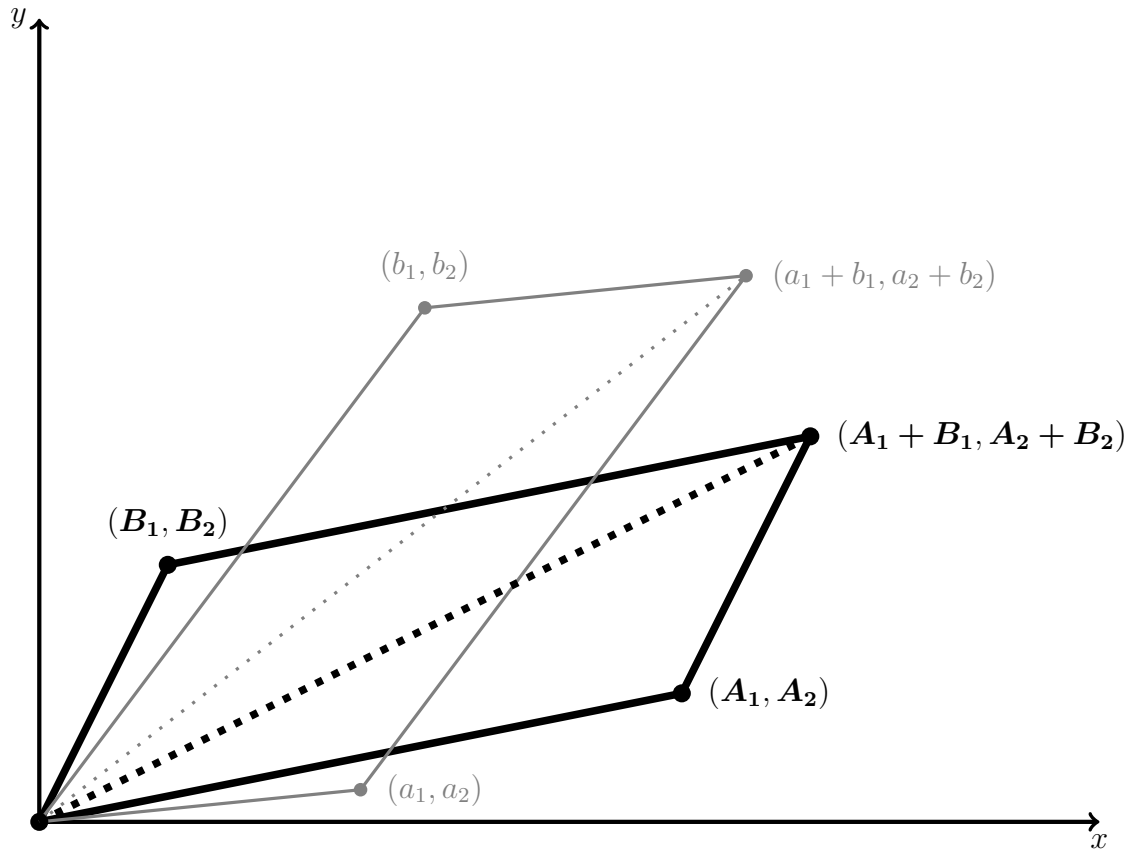


Figure 5.2: A pictorial proof of Simpson's paradox.

The two points (a_1, a_2) and (b_1, b_2) are two vertices of a parallelogram, which has its other two vertices given by $(0, 0)$ and $(a_1 + b_1, a_2 + b_2)$. The ratio $\frac{a_2}{a_1}$ is the gradient of the line from the origin $(0, 0)$ to the vertex (a_1, a_2); the other gradients are defined similarly.

There is also a second parallelogram defined with vertices $(0, 0)$, $(\boldsymbol{A_1}, \boldsymbol{A_2})$, $(\boldsymbol{B_1}, \boldsymbol{B_2})$ and $(\boldsymbol{A_1} + \boldsymbol{B_1}, \boldsymbol{A_2} + \boldsymbol{B_2})$.

Then, in this figure,

$$\frac{a_2}{a_1} < \frac{\boldsymbol{A_2}}{\boldsymbol{A_1}} \qquad \text{and} \qquad \frac{b_2}{b_1} < \frac{\boldsymbol{B_2}}{\boldsymbol{B_1}},$$

$$\text{but} \qquad \frac{a_2 + b_2}{a_1 + b_1} > \frac{\boldsymbol{A_2} + \boldsymbol{B_2}}{\boldsymbol{A_1} + \boldsymbol{B_2}}.$$

There are many real-world examples of Simpson's paradox from areas as diverse as sport, finance and university admissions. It is also relatively easy to come up with your own example now that you know how the phenomenon works.

Below is a artificial example from [3]. Two treatments for a disease, where one treatment is the standard treatment while the other is a new treatment, are administered to 80 patients, where 40 patients are given the standard treatment and 40 are given the new treatment. It is then recorded if a patient's condition improves or not. From Table 5.3 below, it seems that more patients improve when given the standard treatment compared to patients who are given the new treatment.

All patients				
Treatment	Improved	Not Improved	Total	Percent Improved
Standard	24	16	40	**60%**
New	20	20	40	50%

Table 5.3: Improvement percentage across patients for the standard and new treatments.

However, the treatments were not assigned randomly, and looking at the same data when disaggregated into two subgroups, Group A and Group B (which could be gender, age, height, eye colour, etc.), the data then reflects that a higher proporition of patients that are given the new treatment improve when compared to the standard treatment.

Group A				Group B			
Treatment	Improved	Not improved	Percent improved	Treatment	Improved	Not Improved	Percent improved
Standard	3	7	30%	Standard	21	9	70%
New	12	18	**40%**	**New**	8	2	**80%**

Table 5.4: Improvement percentage across patients for the standard and new treatments with the data disaggregated in to subgroups Group A and Group B.

Remark 5.3.1. Simpson's paradox shows that care must be taken when aggregating subgroups of data into a single group and then drawing statistical conclusions based on this aggregated sample. To avoid any issues, one can ensure that the subgroups of data follow similar proportions across the categories. □

5.4 Anscombe's quartet

Consider the following four datasets created by the statistician Frank Anscombe, where each dataset consists of 11 pairs of x- and y-values. Collectively, these datasets are known as Anscombe's quartet:

	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
1	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
2	8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
3	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
4	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
5	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
6	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
7	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
8	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
9	12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
10	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
11	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

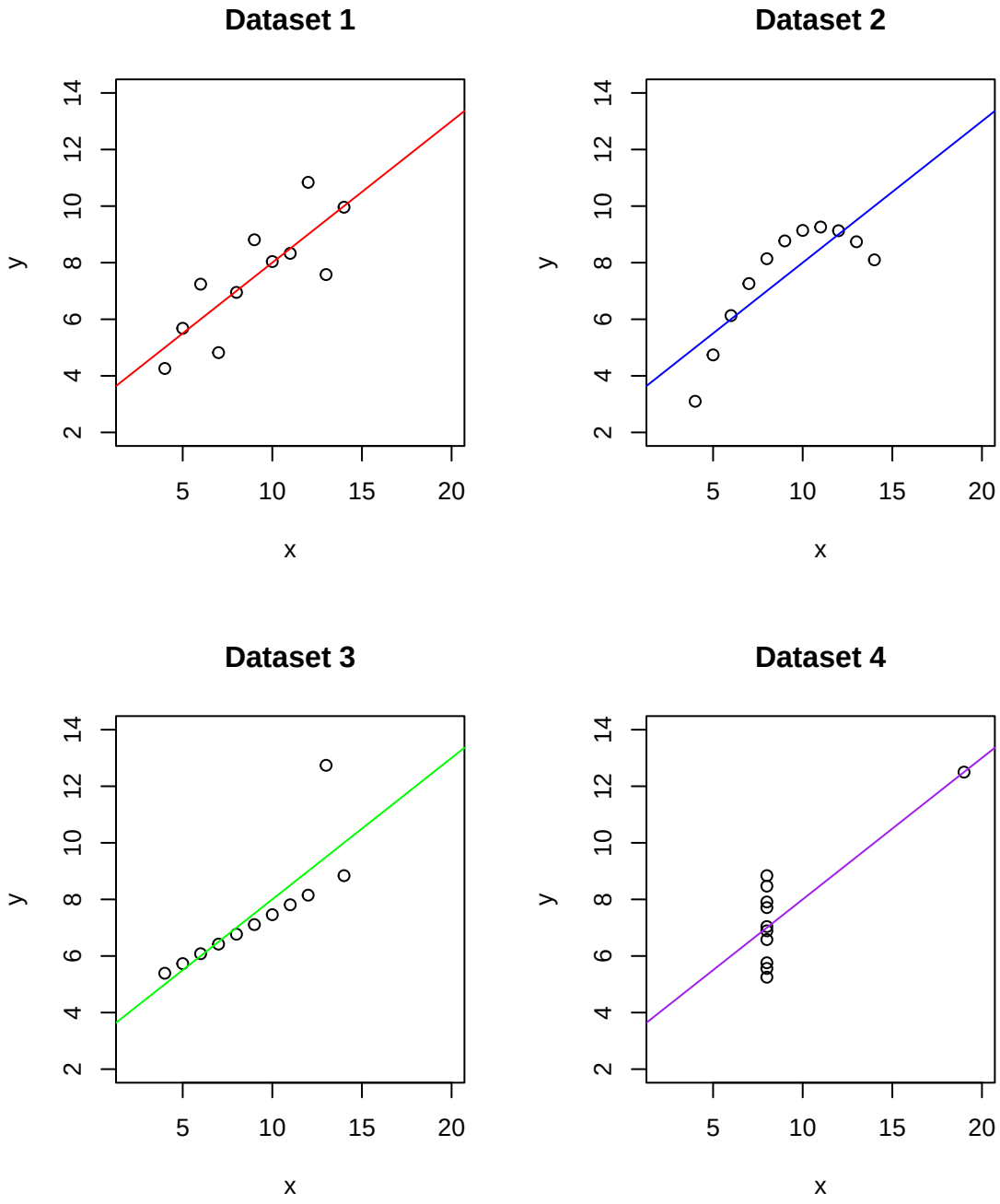
Table 5.5: The four datasets which make up Anscombe's quartet.

Interestingly, several summary statistics of these datasets are equal up to two decimal places:

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Mean of x	9.00	9.00	9.00	9.00
Mean of y	7.50	7.50	7.50	7.50
Variance of x	11.00	11.00	11.00	11.00
Variance of y	4.13	4.13	4.12	4.12
Correlation of x and y	0.82	0.82	0.82	0.82
Regression intercept	3.00	3.00	3.00	3.00
Regression slope	0.50	0.50	0.50	0.50
R^2	0.67	0.67	0.67	0.67

Table 5.6: Summary statistics for Anscombe's quartet.

However, if we plot the data, we see that these datasets are very different in character. **These data serves as a warning** to be careful of interpreting summary statistics in isolation, and reminds us of the need to create visualisations of the data whenever possible.



Chapter 6

Covariance and correlation

The covariance of two random variables X and Y, denoted $\mathrm{Cov}\,(X,Y)$, was already introduced by Prof. Veraart in Section 11.6 of the her notes last term. We shall briefly review the definition and some properties of $\mathrm{Cov}\,(X,Y)$ before defining a related quantity, the **correlation**.

Throughout this chapter we shall be referring to the mean and variance of both X and Y. In order to distinguish these values, we shall use the notation

$$\begin{aligned} \mathrm{E}\,(X) &= \mu_X, \qquad & \mathrm{Var}\,(X) &= \sigma_X^2 \\ \mathrm{E}\,(Y) &= \mu_Y, \qquad & \mathrm{Var}\,(Y) &= \sigma_Y^2 \end{aligned}$$

We shall also assume that the variances are non-zero and finite, i.e. $0 < \sigma_X^2, \sigma_Y^2 < \infty$.

6.1 Covariance

Definition 6.1. The **covariance** of two random variables X and Y, with means μ_X and μ_Y, respectively, is the number defined by

$$\mathrm{Cov}\,(X,Y) = \mathrm{E}[(X-\mu_X)\,(Y-\mu_Y)]. \tag{6.1}$$

Remark 6.1.1. From the definition, the covariance of a variable with itself is its variance:

$$\mathrm{Cov}\,(X,X) = \mathrm{E}[(X-\mu_X)^2] = \mathrm{Var}\,(X)\,.$$

□

Remark 6.1.2. From the definition, it is also immediate that the covariance is a symmetric function,

$$\mathrm{Cov}\,(Y,X) = \mathrm{Cov}\,(X,Y)$$

in other words, the covariance of X and Y is the same as the covariance of Y and X. □

Last term we saw the identity

$$\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y) + 2\mathrm{Cov}\,(X,Y)\,. \tag{6.2}$$

Since this is such a fundamental result, it is worth proving the following generalisation as an exercise:

Exercise 6.1.3. For any two random variables X and Y, and constants $a, b \in \mathbb{R}$,

$$\mathrm{Var}\,(aX+bY) = a^2\mathrm{Var}\,(X) + b^2\mathrm{Var}\,(Y) + 2ab\mathrm{Cov}\,(X,Y)\,. \tag{6.3}$$

The mean of $aX+bY$, which we will denote by μ_{aX+bY}, is

$$\mathrm{E}\,(aX+bY) = a\mathrm{E}\,(X) + b\mathrm{E}\,(Y) = a\mu_X + b\mu_Y = \mu_{aX+bY}.$$

Then, using the definitions of variance and covariance,

$$\begin{aligned}
\mathrm{Var}\,(aX+bY) &= \mathrm{E}[(aX+bY-\mu_{aX+bY})^2] \\
&= \mathrm{E}[(aX+bY-(a\mu_X+b\mu_Y))^2] \\
&= \mathrm{E}[(a\,(X-\mu_X)+b\,(Y-\mu_Y))^2] \\
&= \mathrm{E}[a^2\,(X-\mu_X)^2 + b^2\,(Y-\mu_Y)^2 + 2ab\,(X-\mu_X)\,(Y-\mu_Y)] \\
&= a^2\mathrm{E}[(X-\mu_X)^2] + b^2\mathrm{E}[(Y-\mu_Y)^2] + 2ab\mathrm{E}[(X-\mu_X)\,(Y-\mu_Y)] \\
&= a^2\mathrm{Var}\,(X) + b^2\mathrm{Var}\,(Y) + 2ab\mathrm{Cov}\,(X,Y)
\end{aligned}$$

△

We might wonder, for any two random variables X and Y, what possible values $\mathrm{Cov}\,(X,Y)$ can take. The next example provides an answer.

Example 6.1.4. Suppose that X and Y are random variables with $Y = aX+b$, for some constants $a, b \in \mathbb{R}$. Then, directly computing the variance,

$$\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X+aX+b) = \mathrm{Var}\,((a+1)X+b) = (a+1)^2\mathrm{Var}\,(X) \tag{6.4}$$

On the other hand, using Equation (6.2)

$$\begin{aligned}
2\mathrm{Cov}\,(X,Y) &= \mathrm{Var}\,(X+Y) - \mathrm{Var}\,(X) - \mathrm{Var}\,(Y) \\
&= (a+1)^2\mathrm{Var}\,(X) - \mathrm{Var}\,(X) - a^2\mathrm{Var}\,(X) \\
&= (a^2+2a+1)\mathrm{Var}\,(X) - \mathrm{Var}\,(X) - a^2\mathrm{Var}\,(X) \\
&= 2a\mathrm{Var}\,(X) \\
\Rightarrow \mathrm{Cov}\,(X,Y) &= a\mathrm{Var}\,(X)
\end{aligned}$$

This is true for any value of $a \in \mathbb{R}$. If $\mathrm{Var}(X)$ is finite, then for any $c \in \mathbb{R}$, set $a = c(\mathrm{Var}(X))^{-1}$; then $\mathrm{Cov}(X, Y) = c$. This shows that, for this example, we can choose a such that $\mathrm{Cov}(X, Y)$ can take any value in $\mathbb{R}$. △

Although this example shows that the covariance of two random variables can be any value in $\mathbb{R}$, the value is related to the variance. However, the covariance of X and Y is clearly related to the $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$. The next theorem describes this relationship.

Theorem 6.1.5. For any two random variables X and Y, with variances σ_X^2 and σ_Y^2, respectively,

$$|\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y.$$

♦

Proof. The proof of this theorem follows a very interesting approach using only elementary ideas. For any value $t \in \mathbb{R}$, define the function

$$f(t) = \mathrm{E}[((X - \mu_X)\, t + (Y - \mu_Y))^2],$$

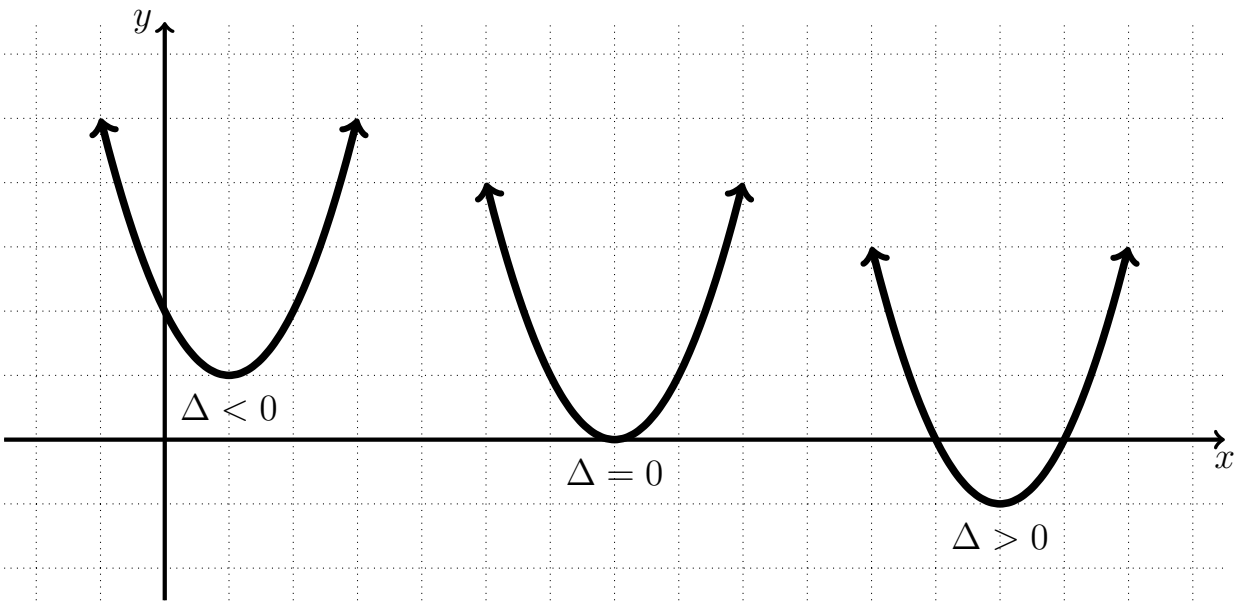
where μ_X and μ_Y are the means of X and Y, respectively. Expanding the square inside the expectation, and using the linearity of expectation, one obtains

$$\begin{aligned} f(t) &= \mathrm{E}[t^2 (X - \mu_X)^2 + 2t (X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2] \\ &= t^2 \mathrm{E}[(X - \mu_X)^2] + 2t\mathrm{E}[(X - \mu_X)(Y - \mu_Y)] + \mathrm{E}[(Y - \mu_Y)^2] \\ \Rightarrow f(t) &= t^2 \sigma_X^2 + 2t\mathrm{Cov}(X, Y) + \sigma_Y^2 \end{aligned} \tag{6.5}$$

Now, we notice two things:

1. $f(t)$ is quadratic in t, i.e. $f(t) = at^2 + bt + c$, where $a = \mathrm{Var}(X) = \sigma_X^2$, $b = 2\mathrm{Cov}(X, Y)$ and $c = \mathrm{Var}(Y) = \sigma_Y^2$.
2. $f(t)$ is defined as the expectation of a non-negative random variable, and is therefore itself non-negative, i.e. writing $Z = (X - \mu_X)\, t + (Y - \mu_Y)$, then $Z^2 \geq 0$, and therefore $f(t) = \mathrm{E}[Z^2] \geq 0$.

Since $f(t) = 0$ is a quadratic equation for real t, it has either no roots, one root or two roots. However, for any $t \in \mathbb{R}$, $f(t) \geq 0$ and therefore there is no value $t' \in \mathbb{R}$ such that $f(t') < 0$, and therefore there cannot be two roots. Therefore, the discriminant $\Delta = b^2 - 4ac \leq 0$. The figure on the next page illustrates this deduction:



Computing the discriminant $\Delta = b^2 - 4ac$ from Equation (6.5),

$$\begin{aligned} &\left(2\text{Cov}\left(X,Y\right)\right)^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 \\ \Rightarrow &\left(\text{Cov}\left(X,Y\right)\right)^2 \leq \sigma_X^2\sigma_Y^2 \\ \Rightarrow &\left|\text{Cov}\left(X,Y\right)\right| \leq \sigma_X\sigma_Y \end{aligned}$$

which proves the result. □

Remark 6.1.6. There is another proof of this result using the Cauchy-Schwartz inequality. □

6.1.1 The sample covariance

Suppose the random variables $X_1, X_2, \ldots, X_n$ follow distribution F_X, and the random variables $Y_1, Y_2, \ldots, Y_n$ follow distribution F_Y. Then the sample covariance is defined as

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right).$$

Suppose the random variables $X_1, X_2, \ldots, X_n$ are observed as $x_1, x_2, \ldots, x_n$, respectively, and the random variables $Y_1, Y_2, \ldots, Y_n$ are observed as $y_1, y_2, \ldots, y_n$, respectively. Then the observed sample covariance is defined as

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right).$$

6.2 Correlation

Given Theorem 6.1.5 the concept of correlation now arises naturally.

Definition 6.2. The **correlation** of the two random variables X and Y, with variances σ_X^2 and σ_Y^2, respectively, is the number defined as

$$\rho_{XY} = \frac{\mathrm{Cov}\,(X,Y)}{\sigma_X \sigma_Y} \tag{6.6}$$

Remark 6.2.1. There are a few slightly different notations for ρ_{XY}, such as $\rho_{X,Y}$ and $\rho\,(X,Y)$, but most involve the Greek letter 'rho' in some way. □

Remark 6.2.2. This definition of correlation is known as Pearson correlation, after the statistician Karl Pearson. There is another, similar definition of correlation called Spearman correlation, but we will not consider that here. □

There is now an immediate corollary to Theorem 6.1.5:

Corollary 6.2.3. For any random variables X and Y,

$$-1 \leq \rho_{XY} \leq 1.$$

♦

Proof. From the definition of correlation and Theorem 6.1.5,

$$\begin{aligned} |\rho_{XY}| &= \left| \frac{\mathrm{Cov}\,(X,Y)}{\sigma_X \sigma_Y} \right| \leq \left| \frac{\sigma_X \sigma_Y}{\sigma_X \sigma_Y} \right| = 1 \\ \Rightarrow -1 &\leq \rho_{XY} \leq 1 \end{aligned}$$

□

One may wonder under which circumstances $\rho_{XY} = \pm 1$. The next two results provide the answer.

Lemma 6.2.4. Suppose Z is a non-negative random variable. Then $\mathrm{E}[Z] = 0$ implies that $\mathrm{P}(Z = 0) = 1$. ♦

Proof. Assume that Z is a nonegative random variable, and further assume that $\mathrm{E}[Z] = 0$.

Recall Markov's Inequality, Theorem 1.3.1. For any $a > 0$,

$$\mathrm{P}(Z \geq a) \leq \frac{\mathrm{E}(Z)}{a}.$$

Rearranging this,

$$\begin{aligned} 0 = \mathrm{E}[Z] &\geq a \cdot \mathrm{P}(Z \geq a), \\ \Rightarrow \mathrm{P}(Z \geq a) &= 0. \end{aligned}$$

Now, if we consider the family of sets $A_i = \left\{Z \geq \frac{1}{i}\right\}$, then

$$\begin{aligned} \bigcup_{i=1}^{\infty} A_i &= \{Z > 0\}, \\ \Rightarrow \mathrm{P}(Z > 0) = \mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) &\leq \sum_{i=1}^{\infty} \mathrm{P}(A_i) = 0, \end{aligned}$$

since we have shown

$$\mathrm{P}(A_i) = \mathrm{P}\left(Z \geq \tfrac{1}{i}\right) = 0.$$

Then, since the sets $\{Z = 0\}$ and $\{Z > 0\}$ are disjoint, and $\{Z \geq 0\} = \{Z = 0\} \cup \{Z > 0\}$,

$$\begin{aligned} 1 &= \mathrm{P}(Z \geq 0) \\ &= \mathrm{P}(Z > 0) + \mathrm{P}(Z = 0) \\ &= 0 + \mathrm{P}(Z = 0) \\ &= \mathrm{P}(Z = 0), \end{aligned}$$

which proves the result. □

Remark 6.2.5. It is possible to formally show the reverse implication, that if Z is a non-negative random variable and $\mathrm{P}(Z = 0) = 1$ then $\mathrm{E}[Z] = 0$, but this is beyond the scope of this course. However, we will assume that this is true in the next proof. □

Corollary 6.2.6. For any two random variables X and Y, $|\rho_{XY}| = 1$ if and only if there exist numbers $a \neq 0$ and b such that $\mathrm{P}(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$. ♦

Proof.

Following the proof of Theorem 6.1.5, define

$$f(t) = \mathrm{E}[((X - \mu_X)\, t + (Y - \mu_Y))^2] = t^2\sigma_X^2 + 2t\mathrm{Cov}(X, Y) + \sigma_Y^2.$$

Then

$$\begin{aligned}
|\rho_{XY}| = 1 &\iff \left|\frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}\right| = 1 \\
&\iff \mathrm{Cov}(X, Y) = \pm\sigma_X\sigma_Y \\
&\iff (\mathrm{Cov}(X, Y))^2 = \sigma_X^2\sigma_Y^2 \\
&\iff (2\mathrm{Cov}(X, Y))^2 - 4\sigma_X^2\sigma_Y^2 = 0 \\
&\iff \text{ the discriminant of } f(t) \text{ is } \Delta = 0 \\
&\iff \ f(t) = 0 \text{ has a single root.}
\end{aligned}$$

Since the discriminant is 0, then t is the single root when

$$t = \frac{-\mathrm{Cov}(X, Y)}{\sigma_X^2} = -\left(\frac{\sigma_Y}{\sigma_X}\right)\left(\frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}\right) = -\left(\frac{\sigma_Y}{\sigma_X}\right)\rho_{XY}.$$

Furthermore, since $((X - \mu_X)\, t + (Y - \mu_Y))^2 \geq 0$, Lemma 6.2.4 gives us that

$$\begin{aligned}
&\mathrm{E}[((X - \mu_X)\, t + (Y - \mu_Y))^2] = 0 \\
\iff &\mathrm{P}\left([(X - \mu_X)\, t + (Y - \mu_Y)]^2 = 0\right) = 1 \\
\iff &\mathrm{P}((X - \mu_X)\, t + (Y - \mu_Y) = 0) = 1 \qquad (\text{since } Z^2 = 0 \text{ if and only if } Z = 0) \\
\iff &\mathrm{P}(Y = -tX + (\mu_X t + \mu_Y)) = 1 \\
\iff &\mathrm{P}(Y = aX + b) = 1
\end{aligned}$$

where $a = -t$ and $b = \mu_X t + \mu_Y$, with $t = -\left(\frac{\sigma_Y}{\sigma_X}\right)\rho_{XY}$, so $a = \left(\frac{\sigma_Y}{\sigma_X}\right)\rho_{XY}$, so a has the same sign as ρ_{XY}, which proves the final part of the result.

□

6.2.1 The sample correlation

Given a collection of pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, one can define the sample correlation.

Definition 6.3. Suppose the random variables $X_1, X_2, \ldots, X_n$ are observed as $x_1, x_2, \ldots, x_n$, respectively, and the random variables $Y_1, Y_2, \ldots, Y_n$ are observed as $y_1, y_2, \ldots, y_n$, respectively. Then the observed sample correlation is defined as

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}.$$

Remark 6.2.7. Every measurement from an experiment has a particular unit. For example, height in the UK is measured in feet and inches (or just inches), while height in France is measured in centimetres. Suppose we were computing the sample correlation from a collection of measurements of height with another quantity: would it matter which units were used? Would we get a different correlation value if we had recorded the heights in inches or in centimetres? The following proposition shows that as long as the units are linear functions of each other, it does not matter which units are used when computing correlation. □

Proposition 6.2.8. Suppose that the pairs of measurements $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are observed. Define the pairs $(u_1, v_1), (u_2, v_2), \ldots (u_n, v_n)$ by

$$u_i = ax_i + b, \qquad v_i = cy_i + d, \qquad i \in \{1, 2, \ldots, n\},$$

for some $a, b, c, d \in \mathbb{R}$ with $a > 0$ and $c > 0$. Then

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{\sum_{i=1}^{n} (u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{n} (u_i - \overline{u})^2} \sqrt{\sum_{i=1}^{n} (v_i - \overline{v})^2}} = r_{UV}.$$

♦

Proof.

$$\begin{aligned} \overline{u} &= \frac{1}{n} \sum_{i=1}^{n} u_i = \frac{1}{n} \sum_{i=1}^{n} (ax_i + b) = a \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + b = a\overline{x} + b \\ \Rightarrow u_i - \overline{u} &= a(x_i - \overline{x}) \end{aligned}$$

Similarly, $v_i - \overline{v} = c\,(y_i - \overline{y})$. Then

$$
\begin{aligned}
r_{UV} &= \frac{\sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \overline{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \overline{v})^2}} \\
&= \frac{\sum_{i=1}^{n} a\,(x_i - \overline{x}) \cdot c\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} a^2\,(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n} c^2\,(y_i - \overline{y})^2}} \\
&= \left(\frac{a}{|a|}\right)\left(\frac{c}{|c|}\right)\frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \left(\frac{a}{|a|}\right)\left(\frac{c}{|c|}\right) r_{XY} = r_{XY}
\end{aligned}
$$

since $a > 0$ and $c > 0$. This proves the result. □

Remark 6.2.9. If we allow either $a < 0$ or $c < 0$ (but not both) in Proposition 6.2.8, then $r_{XY} = -r_{UV}$. If we allow either $a = 0$ or $c = 0$ then r_{UV} is undefined, since the sample variance of the u_i or the v_i is 0. □

Remark 6.2.10. This result only applies to linear transformations of the variables. There will be times when we wish to use a non-linear transformation, such as $x \to \log(x)$; in such cases, the correlation value will change. □

6.2.2 A real data example: height and shoe size

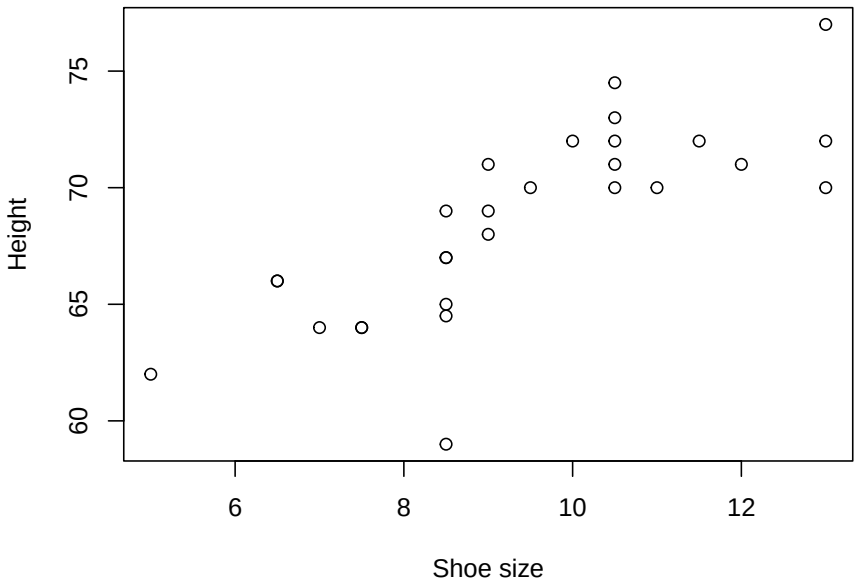
Let us look at some data from collected at Arizona State University, which records the heights and shoe sizes of 28 male and female students [17]. This data is saved in `shoesize.txt`; see Appendix A, Section A.7. The shoe sizes are given on the US scale, and height is given in inches.

Below we use R to read in and plot the data using a scatterplot.

```
# read in the data to a data frame and plot the data
df <- read.table(file="./chapters/data/shoesizes.txt", sep=",", header=TRUE)
plot(x=df$shoe.size, y=df$height, xlab="Shoe size", ylab="Height")

# compute the correlation
print(cor(df$shoe.size, df$height))
#> [1] 0.77661
```

Shoe size	Height	Gender	Shoe Size	Height	Gender
6.5	66.0	F	13.0	77.0	M
9.0	68.0	F	11.5	72.0	M
8.5	64.5	F	8.5	59.0	F
8.5	65.0	F	5.0	62.0	F
10.5	70.0	M	10.0	72.0	M
7.0	64.0	F	6.5	66.0	F
9.5	70.0	F	7.5	64.0	F
9.0	70.0	F	8.5	67.0	M
13.0	72.0	M	10.5	73.0	M
7.5	64.0	F	8.5	69.0	F
10.5	74.0	M	10.5	72.0	M
8.5	67.0	F	11.0	70.0	M
12.0	71.0	M	9.0	69.0	M
10.5	71.0	M	13.0	70.0	M

Table 6.1: Shoes sizes (US scale) and heights (inches) of 28 students from Arizona State University.



It seems as if taller students tend to have larger feet. Computing the sample correlation using the built-in `cor` function, we have $\rho_{XY} = 0.78$. In a later chapter we shall look at a method for computing the distribution of the correlation function, which will allow us to decide if this value is 'significant'.

Chapter 7

Statistical models

7.1 Review of probability models

Let us review the definitions of a **sample space**, **probability measure** and **probability model** [3, 7]. Although you have seen all of this before in Term 1 and in Prof. Veraart's notes, it is worth reviewing these definitions again to re-establish the notation.

Definition 7.1. A **sample space** Ω is a list of all possible outcomes or **responses** of some experiment. Collections of responses, which are subsets of Ω, are called **events**.

Definition 7.2. A **probability measure** P on a sample space Ω is a specification of numbers $\mathrm{P}(A)$ for all events $A \subset \Omega$ such that

1. for every event A, $\mathrm{P}(A) \geq 0$,
2. $\mathrm{P}(\Omega) = 1$,
3. for every finite or countable sequence of disjoint events $\{A_j \,|\, j \in J\}$, $\mathrm{P}\left(\bigcup_{j\in J} A_j\right) = \sum_{j\in J} \mathrm{P}(A_j)$.

Definition 7.3. A **probability model** consists of

1. a non-empty set called the sample space Ω;
2. a collection of events which are subsets of Ω;
3. a probability measure P assigning a probability to each event in Ω.

7.1.1 Notation for random variables and realisations

This is a good point at which to re-emphasise the distinction in notation between **random variables**, and **realisations** of random variables. Recall the definition of a random variable

Definition 7.4. A random variable is a function from a sample space Ω into the real numbers.

It is usual to denote random variables by uppercase letters, e.g. X. So, X is a function $X : \Omega \to \mathbb{R}$. Therefore, for some elementary event $\omega \in \Omega$, $X(\omega) \in \mathbb{R}$. Recall that elementary events are singleton subsets of the sample.

Definition 7.5. Let Ω be the sample space for an experiment and let X be a random variable defined on that sample space. Then for a particular elementary event $\omega \in \Omega$, the value $x = X(\omega) \in \mathbb{R}$ is called a **realisation** of the random variable X.

A realisation of x of X is also called an **observation** of X, and we shall use these terms interchangeably. It is usual to denote realisations with the lowercase letter corresponding to the random variable's uppercase letter. For example, $x_1, x_2, \ldots, x_n$ are realisations of the random variable X.

Finally, we shall use lowercase bold letters to denote collections of (or a vector of) observations, e.g. $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Similarly, uppercase bold letters will be used to denote collections of (or a vector of) random variables, e.g. $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

As an example, suppose may be interested in a random variable $X \sim \mathrm{N}(3, 2)$, and five realisations of this random variable can be generated in R:

```
print(rnorm(n=5, mean=3, sd=sqrt(2)))
#> [1] 1.6396 2.5863 3.3660 1.3706 3.2769
```

7.2 Inference using a probability model

Suppose we are in a situation where we know the probability model for a random variable of interest, but we are uncertain about a future response x. In this situation we may wish to make an **inference** about the value of this response x. There are several options for such an inference:

(a) Compute an estimate of a plausible value for x, e.g. using the expected value of x following our probability model.

(b) Construct a subset that has a high probability of containing the true value of x.

(c) Assess whether or not an observed value of x is an implausible value, given the known probability model.

Example 7.2.1. Suppose it is known that the lifespan X in years for a particular smartphone follows the distribution $X \sim \text{Exp}(\lambda)$ with $\lambda = 1$; see Figure 7.1 for a plot of this distribution.

(a) One option for estimating the lifespan of a new smartphone would be to compute $\text{E}(X) = 1$ year.

(b) Now suppose that one rather wishes to construct an interval $(0, c)$, such that this interval contains 95% of the probability for X and this interval is the smallest possible (i.e. the smallest such value of c needed). Then, one can compute c via the equation

$$\begin{aligned} 0.95 &= \int_0^c e^{-x} \mathrm{d}x = 1 - e^{-c} \\ \Rightarrow c &= -\log(0.05) = 2.996. \end{aligned}$$

One could interpret this to mean that the lifetime of the smartphone will be up to three years, with probability 0.95, assuming our model is correct.

(c) Suppose one was considering purchasing such a smartphone, but wondered whether it would last for (have a lifetime of at least) 5 years. This probability can be computed to be

$$\text{P}(X > 5) = \int_5^\infty e^{-x} \mathrm{d}x = e^{-5} = 0.0067,$$

which would lead one consider such a 5-year lifetime for this smartphone to be unlikely.

△

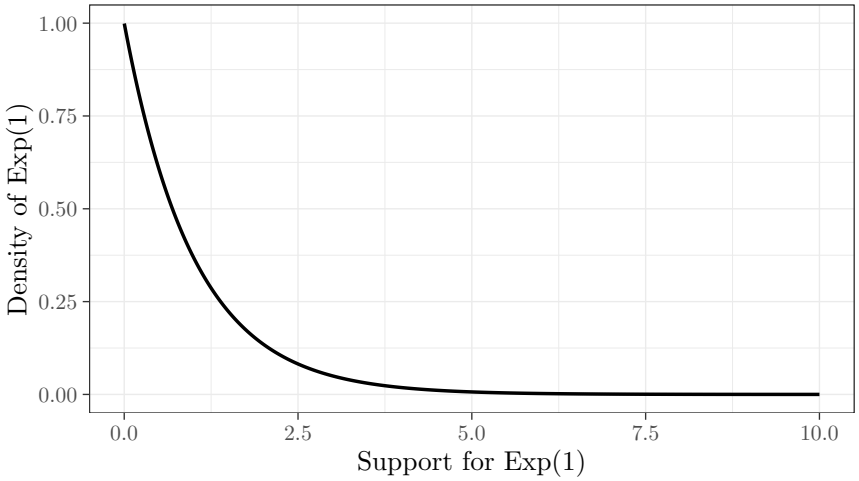


Figure 7.1: Plot showing density of $f(x) = e^{-x}$ for the Exp(1) distribution.

Let us review the situation for Example 7.2.1: we were certain about the probability model, but were uncertain about the data that would be observed.

While this is an interesting example, in practice one would often not have knowledge of the true distribution of the random variable under consideration. However, one may have access to **data** in the form of past observations of the random variable. For example, one may have knowledge of the lifetimes of a sample of smartphones of the same model. This leads us to consider statistical models in the next section. In other words, we may be certain about the data, but uncertain about the probability model generating the data.

7.3 Statistical models

We now consider a different situation to the one above: suppose that we observe data $\mathbf{x}$, but we are uncertain about the the mechanism generating the data $\mathbf{x}$. More explicitly, if we assume that the data $\mathbf{x}$ are observations of the random variables $\mathbf{X}$, then we are uncertain about the probability model for $\mathbf{X}$.

We could consider a **statistical model** for the data $\mathbf{x}$ to be a set $\{\mathrm{P}_\theta \,|\, \theta \in \Theta\}$ of probability measures, one of which is the true probability measure that resulted in data $\mathbf{x}$; however, this true probability measure and corresponding true parameter θ is unknown.

Definition 7.6. The space containing all possible values of the parameter θ is called the **parameter space** and is denoted by Θ, i.e. $\theta \in \Theta$.

Remark 7.3.1. Note the difference in notation between the parameter space Θ, and the notation for an estimator $\widehat{\Theta}$; the similarity in notation is unfortunate, but in practice it should be clear from the context which quantity one is dealing with. □

Example 7.3.2. Suppose five friends all purchased the same smartphone when it was released. The manufacturer of the smartphones claims that the lifespan of the phones (in years) follows an Exp(0.5) distribution, while another source claims the lifespan of the phones follows an Exp(1) distribution. Therefore, in this example the statistical model for the lifespan of the smartphones is $\{\mathrm{P}_{0.5}, \mathrm{P}_1\}$, where $\mathrm{P}_{0.5}$ is the Exp(0.5) probability measure and P_1 is the Exp(1) probability measure, i.e. our indexing parameter is $\theta \in \Theta = \{0.5, 1\}$. Suppose that the friends record the lifespans of their phones, i.e. they use the phones until they break, and obtain the sample $(0.76, 1.18, 0.15, 0.14, 0.44)$ number of years. Comparing the p.d.f.'s of the Exp(0.5) and Exp(1) distributions in Figure 7.2, which model would you be inclined to say is the correct one? What if the observed data had been $(1.91, 2.46, 1.08, 5.79, 0.29)$? △

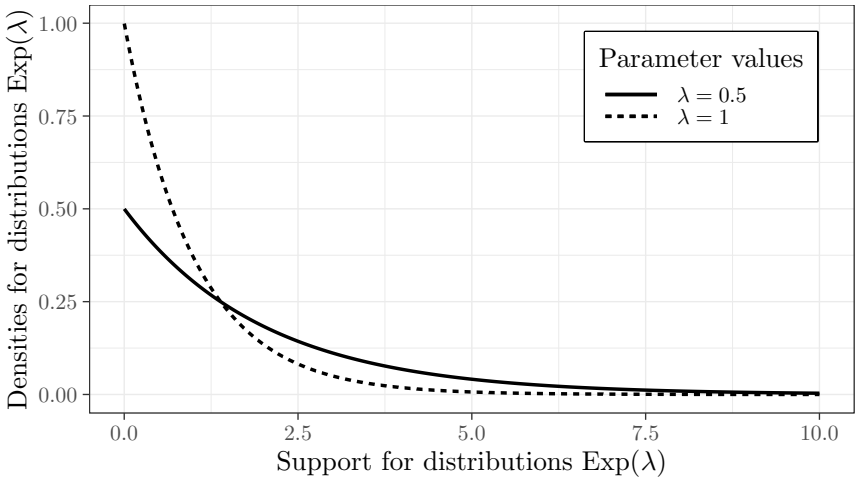


Figure 7.2: The density $f(x|\lambda) = \lambda e^{-\lambda}$ of the Exp(λ) distribution, for $\lambda \in \{0.5, 1\}$

This example raises several questions:

- Are five observations enough in order to make a firm conclusion? If not, then how many are 'enough'?
- Is it even fair to treat all five observations as coming from the same distribution?
- How can we decide with any certainty which of the two models is correct?
- What if we are not told that $\theta \in \Theta = \{0.5, 1\}$, but rather that θ is in the interval $\Theta = [0.2, 4]$?

In fact, the data in Example 7.3.2 were sampled from the Exp(0.5) and Exp(1) distributions using R:

```
print(rexp(n=5, rate=1))
#> [1] 0.75518 1.18164 0.14571 0.13980 0.43607
print(rexp(n=5, rate=0.5))
#> [1] 5.78994 2.45912 1.07937 1.91313 0.29409
```

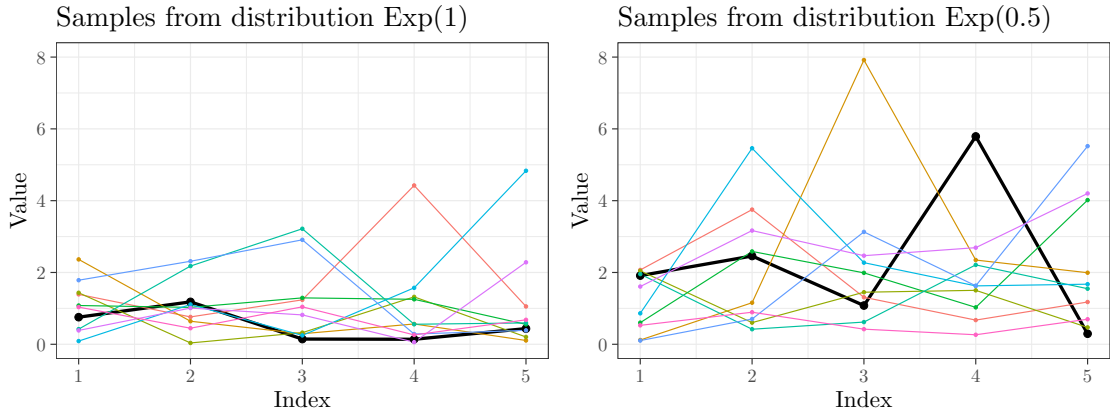


Figure 7.3: Ten samples from Exp(λ) distributions, for $\lambda \in \{1, 0.5\}$. The samples $(0.76, 1.18, 0.15, 0.14, 0.44)$ and $(1.91, 2.46, 1.08, 5.79, 0.29)$ from Example 7.3.2 are shown as **thick black lines**.

7.3.1 A detailed specification of a statistical model

A more concrete definition of a statistical model could be [3]:

Definition 7.7. A **statistical model** consists of

1. an identification of random variables of interest (both observable and hypothetically observable),
2. a specification of a joint distribution or family of possible distributions for the observable random variables,
3. the identification of any parameter(s) θ of those distributions that are assumed unknown and possibly hypothetically observable,
4. (Bayesian model) a specification of a (joint) distribution for the unknown parameters.

When one treats the unknown parameter(s) θ as random, the joint distribution of the observable random variables indexed by θ is understood as the conditional distribution of the observable random variables given the parameter(s) θ.

Remark 7.3.3. We can now contrast the different approaches of (pure) mathematics and statistics; mathematics starts with axioms and then develops the theory, while statistics starts with observable data and then tries to determine the underlying data-generating mechanism. In other words, mathematics starts with the truth and then discovers the world; statistics starts with observing the world and then discovers the truth [9]. □

Chapter 8

Likelihood

8.1 The likelihood function

Definition 8.1. Suppose we have a statistical model for the random variables $\mathbf{X}$ described by $\{\mathrm{P}_\theta \,|\, \theta \in \Theta\}$, and where each P_θ is specified by the probability density function (or probability mass function) f_θ. Having observed the data $\mathbf{x}$, the **likelihood function** $L(\cdot|\mathbf{x}) : \Theta \to \mathbb{R}$ is defined by $L(\theta|\mathbf{x}) = f_\theta(\mathbf{x})$ for any $\theta \in \Theta$.

Given the definition of a likelihood function, we also have

Definition 8.2. For any $\theta \in \Theta$, $L(\theta|\mathbf{x})$ is called the **likelihood** of θ given the observed data $\mathbf{x}$.

We can also redefine the probability density (mass) function f_θ:

Remark 8.1.1. The probability density (mass) function $f_\theta(\mathbf{x})$ in Definition 8.1 can also be denoted $f(\mathbf{x}|\theta) = f_\theta(\mathbf{x})$ and can be considered as the joint probability density (mass) function of $\mathbf{X}$ given the value of the model's parameter is θ. We then have the equation:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta). \tag{8.1}$$

□

Remark 8.1.2. In Definition 8.1 it seems that we are simply defining the likelihood to be the same as the probability density function or probability mass function. There is a subtle distinction, however, in terms of which of the pair $(\theta, \mathbf{x})$ is fixed and which is varying in each function. When we consider the p.d.f. or p.m.f. $f(\mathbf{x}|\theta)$, the parameter value θ is fixed and the data $\mathbf{x}$ (or random variable X) is varying. However, when we consider the likelihood function $L(\theta|\mathbf{x})$, we are considering the observed data $\mathbf{x}$ to be fixed and allowing $\theta \in \Theta$ to vary over all possible parameter values. □

Remark 8.1.3. Throughout this chapter we shall repeatedly use the fact that the joint probability density function (p.d.f.) for **independent** random variables is the product of the individual p.d.f.'s. Therefore, the likelihood function for an independent and identically distributed sample is the product of the individual likelihoods. □

Example 8.1.4. Suppose that the data $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ are independently sampled from a $\mathrm{N}(\theta, 1)$ distribution, i.e. a normal distribution with unknown mean θ and variance 1. Then the likelihood $L(\theta|\mathbf{x})$ is

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

We shall revisit this likelihood in Example 8.4.2. △

8.2 Interpreting the likelihood

Suppose we have a statisical model $\{\mathrm{P}_\theta \,|\, \theta \in \Theta\}$ where each P_θ is **discrete** and specified by the probability mass function f_θ. Then, given data $\mathbf{x}$, one can interpret $f_\theta(\mathbf{x})$ as the probability of obtaining the data $\mathbf{x}$ given that the true value of the parameter is θ.

Suppose furthermore that the likelihood function is defined as in Definition 8.1. Then if $\mathbf{x}$ is an observed sample of the random vector $\mathbf{X}$ with probability measure P_θ, then

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \mathrm{P}_\theta(\mathbf{X} = \mathbf{x}).$$

Therefore, in the case of **discrete** random variables, one interprets the value $L(\theta|\mathbf{x})$ as follows:

> $L(\theta|\mathbf{x})$ is the probability of observing the data $\mathbf{x}$ given that the true value of the parameter is θ.

It is **not** the probability that θ is the true value, given that we have observed the data $\mathbf{x}$.

Using this interpretation, for discrete random variables one can compare likelihoods for different values of the parameter θ and if, for example, one has

$$\mathrm{P}_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x}) = \mathrm{P}_{\theta_2}(\mathbf{X} = \mathbf{x}),$$

then the sample $\mathbf{x}$ that was observed is more likely to have occurred if $\theta = \theta_1$, rather than if $\theta = \theta_2$. One can then interpret this as saying that θ_1 is a more plausible value than θ_2 for the true value of the parameter θ.

Note the careful use of the word **plausible** rather than the word "probable". This is because here we consider θ to be a parameter with a fixed value which is unknown.

Example 8.2.1. Suppose that one has a (possibly unfair) coin and wishes to determine the probability θ of obtaining a head when the coin is tossed, with $\theta \in \Theta = [0, 1]$. The coin is tossed $n = 10$ times and exactly $x = 3$ heads are observed. An appropriate statistical model for the data is the $\text{Bin}(10, \theta)$ model, with likelihood function given by

$$L(\theta|3) = \binom{10}{3}\theta^3(1-\theta)^7.$$

This likelihood function is shown in Fig. 8.1. Notice the global maximum at $\theta = 0.3$, with value 0.27. △

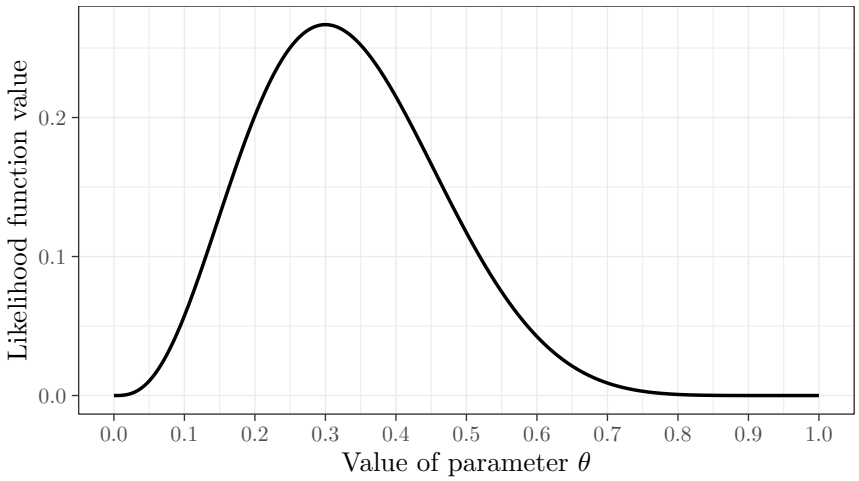


Figure 8.1: Plot showing likelihood of $L(\theta|3) = \binom{10}{3}\theta^3(1-\theta)^7$ for $\theta \in [0, 1]$.

8.3 Likelihood ratios

However, for a **continuous** random variable $\mathbf{X}$, evaluating the likelihood at a particular observed value $\mathbf{x}$ is $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \mathrm{P}_\theta(\mathbf{X} = \mathbf{x}) = 0$. Indeed, even for discrete distributions, likelihoods can be vanishingly small.

Example 8.3.1. Suppose that the sample space is the set of positive integers, $\Omega = \{1, 2, \ldots\}$ and that the statistical model is $\{\mathrm{P}_\theta \,|\, \theta \in \{1, 2\}\}$, where P_1 is the discrete uniform distribution on the set $\{1, 2, \ldots, 10^5\}$ and P_2 is the discrete uniform distribution on the set $\{1, 2, \ldots, 10^8\}$. Now, suppose the value $x = 100$ is observed. Then one can compute:

$$L(\theta = 1|100) = 10^{-5}, \qquad L(\theta = 2|100) = 10^{-8}, \qquad \frac{L(\theta = 1|100)}{L(\theta = 2|100)} = 1000.$$

Both of these likelihood values are very small, but one notices that the likelihood for $\theta = 1$ is one thousand times greater than the likelihood for $\theta = 2$. △

Therefore, rather than being interested in the likelihood $L(\theta|\mathbf{x})$ for a particular θ, we are more interested in **likehood ratios**, which for discrete random variables is simply the ratio of the probabilities

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \frac{\mathrm{P}_{\theta_1}(\mathbf{X} = \mathbf{x})}{\mathrm{P}_{\theta_2}(\mathbf{X} = \mathbf{x})}, \qquad \theta_1, \theta_2 \in \Theta, \qquad \mathbf{X} \text{ is } \textbf{discrete}.$$

One can also interpret likelihood ratios when the random variables are continuous. Suppose one has a statistical model $\{\mathrm{P}_\theta \,|\, \theta \in \Theta\}$ for a random variable X, and that for each θ, X has the p.d.f. $f(x|\theta)$ that is a continuous function of x. Then, for small enough $\delta > 0$, (see Exercise 8.3.2)

$$\mathrm{P}_\theta(x - \delta < X < x + \delta) \approx 2\delta f(x|\theta). \tag{8.2}$$

By Definition 8.1, $2\delta f(x|\theta) = 2\delta L(\theta|x)$, and so for two parameter values θ_1 and θ_2,

$$\frac{L(\theta_1|x)}{L(\theta_2|x)} \approx \frac{\mathrm{P}_{\theta_1}(x - \delta < X < x + \delta)}{\mathrm{P}_{\theta_2}(x - \delta < X < x + \delta)}, \qquad \theta_1, \theta_2 \in \Theta, \qquad X \text{ is } \textbf{continuous}.$$

Therefore, computing the ratio of likelihood functions for two parameter values gives an approximation of the ratio of probability values for observing the sample $\mathbf{x}$.

Exercise 8.3.2. Show that, for a continuous random variable X with continuous p.d.f. $f(x|\theta)$, and for small enough δ, the approximation in Equation (8.2) holds.

Fix a value x in the support of $f(x|\theta)$, and for ease of notation let us write $f(x|\theta) = f_\theta(x)$. Then, choose a small value $\epsilon > 0$. Since $f(x|\theta)$ is continuous, there exists a $\delta > 0$ such that for all $z \in \mathbb{R}$, if $|z - x| < \delta$ then $|f_\theta(z) - f_\theta(x)| < \epsilon$. This is just the definition of continuity from the Analysis course. Then

$$\begin{aligned}
\mathrm{P}_\theta\left(x-\delta<X<x+\delta\right) = \int_{x-\delta}^{x+\delta} f_\theta(z)\mathrm{d}z &= \int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x) + f_\theta(x)\right] \mathrm{d}z \\
&= \int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x)\right] \mathrm{d}z + \int_{x-\delta}^{x+\delta} f_\theta(x)\mathrm{d}z \\
&= \int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x)\right] \mathrm{d}z + 2\delta f_\theta(x),
\end{aligned}$$

because $f_\theta(x)$ is constant in z. Therefore,

$$\begin{aligned}
\mathrm{P}_\theta\left(x-\delta<X<x+\delta\right) - 2\delta f_\theta(x) &= \int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x)\right] \mathrm{d}z \\
\Rightarrow \left|\mathrm{P}_\theta\left(x-\delta<X<x+\delta\right) - 2\delta f_\theta(x)\right| &= \left|\int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x)\right] \mathrm{d}z\right| \\
&\leq \int_{x-\delta}^{x+\delta} \left|f_\theta(z) - f_\theta(x)\right| \mathrm{d}z \\
&< \int_{x-\delta}^{x+\delta} \epsilon \ \mathrm{d}z \\
\Rightarrow \left|\mathrm{P}_\theta\left(x-\delta<X<x+\delta\right) - 2\delta f_\theta(x)\right| &< 2\delta\epsilon.
\end{aligned}$$

So, for small enough δ, $\mathrm{P}_\theta\left(x-\delta<X<x+\delta\right) \approx 2\delta f_\theta(x) = 2\delta f(x|\theta)$. △

8.4 Equivalent likelihood functions

The preceding section motivates our interest in likelihood ratios, rather than the value of likelihoods themselves. One notices, however, that if instead of using the likelihood function $L(\theta|x)$, one instead used a function $L'(\theta|x) = cL(\theta|x)$, $c > 0$, one would obtain the same likelihood ratio:

$$\frac{L'(\theta_1|x)}{L'(\theta_2|x)} = \frac{cL(\theta_1|x)}{cL(\theta_2|x)} = \frac{L(\theta_1|x)}{L(\theta_2|x)}.$$

This leads to a natural definition of equivalence:

Definition 8.3. Given the likelihood function $L(\cdot|\mathbf{x})$ from Definition 8.1, any function $L'(\cdot|\mathbf{x}) = cL(\cdot|\mathbf{x})$ for $c > 0$ is an **equivalent likelihood function** for the parameter θ.

Exercise 8.4.1. Show that the relation $\sim$, where $L_1 \sim L_2$ if $L_1(\cdot|\mathbf{x})$ and $L_2(\cdot|\mathbf{x})$ are equivalent likelihood functions as in Definition 8.3, is an equivalence relation.

1. $L_1(\cdot|\mathbf{x}) = 1 \cdot L_1(\cdot|\mathbf{x}), \Rightarrow L_1 \sim L_1$.
2. $L_1 \sim L_2 \Rightarrow L_1(\cdot|\mathbf{x}) = c \cdot L_2(\cdot|\mathbf{x}) \Rightarrow L_2(\cdot|\mathbf{x}) = \frac{1}{c} L_1(\cdot|\mathbf{x}) \Rightarrow L_2 \sim L_1$.
3. $L_1 \sim L_2$ and $L_2 \sim L_3 \Rightarrow L_1(\cdot|\mathbf{x}) = c_1 \cdot L_2(\cdot|\mathbf{x})$ and $L_2(\cdot|\mathbf{x}) = c_2 \cdot L_3(\cdot|\mathbf{x})$
   $\Rightarrow L_1(\cdot|\mathbf{x}) = c_1 c_2 L_3(\cdot|\mathbf{x}) \Rightarrow L_1 \sim L_3$.

△

Example 8.4.2. Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a sample of observations i.i.d. according to a $\mathrm{N}(\theta, \sigma^2)$ distribution, where σ^2 is unknown but $\theta \in= \mathbb{R}$ is unknown. Then the likelihood is

$$\begin{aligned}
L(\theta|\mathbf{x}) = f_\theta(\mathbf{x}) &= \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left(x_i - \theta\right)^2\right) \\
\Rightarrow L(\theta|\mathbf{x}) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \theta\right)^2\right) \qquad (8.3) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n\left(\overline{x} - \theta\right)^2\right]\right) \qquad \text{(Exercise 8.4.3)} \\
&= \underbrace{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n-1}{2\sigma^2}s^2\right)}_{c>0} \exp\left(-\frac{n}{2\sigma^2}\left(\overline{x} - \theta\right)^2\right),
\end{aligned}$$

which shows that an equivalent likelihood is

$$L'(\theta|\mathbf{x}) = \exp\left(-\tfrac{n}{2\sigma^2}\left(\overline{x} - \theta\right)^2\right). \qquad (8.4)$$

△

Exercise 8.4.3. Show that

$$\sum_{i=1}^{n} (x_i - \theta)^2 = (n-1)s^2 + n(\overline{x} - \theta)^2,$$

where $\overline{x}$ and s^2 are defined in terms of $x_1, x_2, \ldots, x_n$ as usual.

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \theta)^2 &= \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x} - \theta)^2 = \sum_{i=1}^{n} \left[(x_i - \overline{x})^2 + 2(\overline{x} - \theta)(x_i - \overline{x}) + (\overline{x} - \theta)^2 \right] \\
&= \sum_{i=1}^{n} (x_i - \overline{x})^2 + 2(\overline{x} - \theta) \sum_{i=1}^{n} (x_i - \overline{x}) + \sum_{i=1}^{n} (\overline{x} - \theta)^2 \\
&= (n-1)s^2 + 2(\overline{x} - \theta) \cdot 0 + n(\overline{x} - \theta)^2 = (n-1)s^2 + n(\overline{x} - \theta)^2,
\end{aligned}$$

since $\sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = n\overline{x} - n\overline{x} = 0$.

△

8.5 Maximum likelihood estimation

Given a likelihood $L(\theta|\mathbf{x})$ of parameter θ, given the observed data $\mathbf{x}$, we might want to find the value of θ which maximises this likelihood.

Definition 8.4. Suppose $L(\theta|\mathbf{x})$ is the likelihood of the parameter θ given the observed data $\mathbf{x}$. Then the parameter value $\widehat{\theta}(\mathbf{x})$ at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed, is called the **maximum likelihood estimate** of θ.

Definition 8.5. If $L(\theta|\mathbf{x})$ is the likelihood of the parameter θ, given the observed data $\mathbf{x}$, with maximum likelihood estimate $\widehat{\theta}(\mathbf{x})$, then a **maximum likelihood estimator** of the parameter θ based on the random sample $\mathbf{X}$ is $\widehat{\theta}(\mathbf{X})$.

Remark 8.5.1. Note that we use MLE as an abbreviation for both the maximum likelihood **estimator** and the maximum likelihood **estimate**. □

8.5.1 Finding the maximum likelihood estimate

It is often the case that likelihoods have a nice analytical form. In such cases, we can use differential calculus to find the MLE $\widehat{\theta}(\mathbf{x})$. This will involve finding the first derivative of $L(\theta|\mathbf{x})$ with respect to θ. Suppose the values $(\theta_1, \theta_2, \ldots, \theta_m)$ satisfy equation $\frac{\mathrm{d}}{\mathrm{d}\theta}L(\theta|\mathbf{x}) = 0$; then these are **possible** candidates for the MLE. We would need to check that these values indeed maximize the likelihood, and we would also need to check if values on the boundary of the domain of the function maximize the likelihood.

Example 8.5.2. As in Example 8.1.4, suppose that the data $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ are independently sampled from a $\mathrm{N}(\theta, 1)$ distribution. The likelihood $L(\theta|\mathbf{x})$ is

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

Now, to find the maximum of this function, we use the chain rule to compute the first derivative:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right) \cdot \left(-\frac{1}{2}\sum_{i=1}^{n}2(x_i - \theta)(-1)\right)$$

If we set this to 0, since $\exp(z) > 0$ for all real z, this reduces to:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\theta}L(\theta|\mathbf{x}) = 0 \\ \Rightarrow &\sum_{i=1}^{n}(x_i - \theta) = 0 \\ \Rightarrow &\theta = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x} \end{aligned}$$

Now, $\theta = \overline{x}$ is the only value that is a solution to $\frac{\mathrm{d}}{\mathrm{d}\theta}L(\theta|\mathbf{x}) = 0$, but we need to check if it a maximum. This can be done by computing the second derivative, and evaluating it at $\theta = \overline{x}$. One can compute:

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}L(\theta|\mathbf{x}) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\cdot\left(\sum_{i=1}^{n}(x_i-\theta)\right)\right] \\ &= \frac{1}{(2\pi)^{n/2}}\left[\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\cdot\left(\sum_{i=1}^{n}(x_i-\theta)\right)^2\right. \\ &\qquad \left.+\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\cdot\left(\sum_{i=1}^{n}(-1)\right)\right]\end{aligned}$$

Evaluating at $\theta = \overline{x}$, and setting $A = \exp(-\frac{1}{2}\sum(x_i-\theta)^2) > 0$ we have

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}L(\theta|\mathbf{x})\right|_{\theta=\overline{x}} = (2\pi)^{-n/2}\left[A\cdot(0)^2 + A\cdot(-n)\right] = -nA\,(2\pi)^{-n/2} < 0$$

which shows that $\overline{x}$ is a maximum. Finally, we need to check that the boundary points for θ are not maxima. Since the mean θ is defined on the whole real line, the boundary points are $\pm\infty$. Evaluating $L(\theta|\mathbf{x})$ as $\theta \to \pm\infty$, we see that $\lim_{\theta\to\pm\infty} L(\theta|\mathbf{x}) = 0$. On the other hand, when $L(\theta|\mathbf{x})$ is evaluated at $\theta = \overline{x}$, we see that $L(\theta = \overline{x}|\mathbf{x}) = (2\pi)^{-n/2}\exp(\frac{-(n-1)s^2}{2}) > 0$.

Therefore $\widehat{\theta}(\mathbf{x}) = \overline{x}$ is the maximum likelihood estimate for θ, and the sample mean $\overline{X}$ is the maximum likelihood estimator. △

Remark 8.5.3. This example seemed to be a lot of work, computing derivatives, checking boundaries, etc. However, it is not always necessary to resort to calculus; our goal is simply to maximise the likelihood. It is generally harder to algebraically find a global upper bound on the likelihood function, but the following example shows that we should bear this alternative approach in mind. □

Example 8.5.4. (Continuation of Example 8.5.2) Let us try and find the MLE of the likelihood $L(\theta|\mathbf{x})$ in Example 8.5.2 in another way. Recall Exercise 1.2.10 which essentially stated that, given any number a,

$$\sum_{i=1}^{n}(x_i-a)^2 \geq \sum_{i=1}^{n}(x_i-\overline{x})^2,$$

with equality when $a = \overline{x}$. Using this result, for any value θ,

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \leq \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2\right),$$

with equality if and only if $\theta = \overline{x}$. Therefore, $\overline{X}$ is the MLE. △

8.5.2 The log-likelihood

There will be times when we do not see a direct method of maximisation (as in Example 8.5.4), but the likelihood $L(\theta|\mathbf{x})$ is in a form that is difficult to differentiate. In such cases, it may be worth trying to maximise a transform of the likelihood. A common transformation is to use the logarithm function, since it is monotonic, i.e.

$$\theta_1 \leq \theta_2 \Rightarrow \log(\theta_1) \leq \log(\theta_2),$$

and therefore finding the value $\widehat{\theta}$ that maximises $\log L(\theta|\mathbf{x})$ is equivalent to finding the value that maximises $L(\theta|\mathbf{x})$. We call the transformation $\log L(\theta|\mathbf{x})$ the **log-likelihood**.

Example 8.5.5. Suppose the random variables $X_1, X_2, \ldots, X_n$ follow a Bern(θ) distribution, and that we observe $\mathbf{X}$ as $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. Then the likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y(1-\theta)^{n-y}$$

where $y = \sum_{i=1}^{n} x_i = n\overline{x}$. While it is possible to differentiate $L(\theta|\mathbf{x})$ directly, the log-likelihood is simpler:

$$\log L(\theta|\mathbf{x}) = y \log \theta + (n-y)\log(1-\theta).$$

When $0 < y < n$ (i.e. the trials are not all 0 or not all 1), then differentiating $\log L(\theta|\mathbf{x})$ yields

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta|\mathbf{x}) = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

Setting this expression equal to 0 and solving, one obtains (exercise) $\widehat{\theta} = \frac{y}{n} = \overline{x}$.

However, we also need to check the cases $y = 0$ and $y = n$. When $y = 0 \Rightarrow x_i = 0$ for all i, $\log L(\theta|\mathbf{x}) = (n-y)\log(1-\theta)$. This is a decreasing function in θ, with no (local) maximum in the interval $(0, 1)$. Checking the boundary points $\theta \in \{0, 1\}$, we see the MLE occurs at $\theta = 0$; note this is still a special case of $\theta = \overline{x}$. Similarly, when $y = n$ (i.e. all $x_i = 1$), $\widehat{\theta} = 1 = \overline{x}$. Therefore in all cases, the MLE is $\widehat{\theta} = \overline{x}$. △

8.5.3 Finding the MLE for multiple unknown parameters

So far we have only discuss the case that θ is a single unknown parameter. In practice, the are many situations where we want to simultaneously maximise several parameters. This will require taking partial derivatives with respect to each variable, and setting these equations equal to zero, and then solving this simultaneous set of equations. This is beyond the scope of this module, but the following example is worth knowing.

Example 8.5.6. Suppose the observations $x_1, x_2, \ldots, x_n$ are independently sampled from a $\mathrm{N}(\mu, \sigma^2)$ distribution, with both $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ unknown. Similarly to Example 8.1.4, the likelihood can be written as

$$L(\mu, \sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Taking partial derivatives with respect to μ and σ^2 (not σ), one obtains the MLEs

$$\begin{aligned} \widehat{\theta}_1 &= \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \widehat{\theta}_2 &= \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \end{aligned} \tag{8.5}$$

Therefore the MLE of μ is $\overline{X}$ and the MLE of σ^2 is S_b^2. △

Chapter 9

Simple Linear Regression

In Chapter 6 we saw how, given two quantities X and Y for which we have pairs of recorded observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we can compute the correlation r_{XY} of two quantities to obtain a measure of association between X and Y.

In this section we introduce **linear regression** which takes this idea further and attempts to establish if there is a linear relationship between X and Y of the form $Y = \beta_0 + \beta_1 X$; if so, given a new sample point x_{n+1}, we may be able to predict the value of y_{n+1}.

This particular formulation is known as **simple** linear regression because there is only one predictor variable X. In later modules, you will learn how to extend this methodology to **multiple** linear regression, which fits a linear model of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. This chapter is based on [16, 2, 3].

9.1 Motivation for simple linear regression

Suppose we have two quantities of interest, X and Y, between which we believe there is some relationship which we would like to investigate. If we suppose that the relationship is specified by an unknown function f, where

$$Y = f(X),$$

then since it seems that the values of Y are obtained from the values of X through the function f, and we call X the **predictor** and Y the **response**.

Now, suppose we measure n pairs of values (x_i, y_i), for $i = 1, 2, \ldots, n$, where x_i is a measurement of X and y_i is a measurement of Y at time i. Then, if f were known, we could write for each pair (x_i, y_i),

$$y_i = f(x_i) + \gamma_i, \qquad i \in \{1, 2, \ldots, n\}, \tag{9.1}$$

where each γ_i is a random error in the observational process. For example, this could be a measurement error when recording the y_i values; for the moment, we assume that the x_i values are recorded without error.

Although the function f is unknown, suppose that $f(x)$ can be approximated by the straight line $\beta_0+\beta_1 x$ where β_0, β_1 are parameters to be determined. Since this is only an approximation, for the points $x_1, x_2, \ldots, x_n$ we write

$$f(x_i) = \beta_0 + \beta_1 x_i + \delta_i, \tag{9.2}$$

where each δ_i is a fixed error due to the lack of fit that occurs by approximating f by a straight line. In order for the simple linear regression model to be useful, it should be the case that $\delta_i \ll \gamma_i$, for $i = 1, 2, \ldots, n$. By combining Equations (9.1) and (9.2), and combining the errors into $e_i = \gamma_i + \delta_i$, we finally obtain the expression

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i \in \{1, 2, \ldots, n\}. \tag{9.3}$$

Remark 9.1.1. Note that the e_i terms in Equation (9.3) are the realisations of the random errors. □

9.2 Least squares estimation: an analytical approach

Suppose that we have n pairs of measurements of the quantities X and Y denoted by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We assume that there is a linear relationship of the form $Y = \beta_0 + \beta_1 X$, but the parameters β_0 and β_1 are unknown. For our data, we have the model given in Equation (9.3),

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i \in \{1, 2, \ldots, n\},$$

where the e_i are the (measurement and model) errors discussed in Section 9.1. Note that, since we do not know the true values of β_0 and β_1, the values of the errors e_i are unknown.

Least squares estimation is a purely analytical approach that finds the best estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the parameters β_0 and β_1, respectively. Of course, this method for finding the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ will depend on what is meant by 'best'.

Suppose we decided on the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Then we would define the **residuals** $\widehat{e}_i$ as

$$\widehat{e}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i), \qquad i \in \{1, 2, \ldots, n\}. \tag{9.4}$$

Now, since we have all pairs (x_i, y_i) and the estimates $\widehat{\beta}_0, \widehat{\beta}_1$, the residuals are observable. The goal of the least squares approach is to find the pair $(\widehat{\beta}_0, \widehat{\beta}_1)$ such that the **residual sum of squares (RSS)**

$$\sum_{i=1}^{n} (\widehat{e}_i)^2, \tag{9.5}$$

is as small as possible.

9.2.1 Solving the least squares problem

It will be useful to introduce notation for a few quantities before deriving expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Given pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we define the sample means $\overline{x}$ and $\overline{y}$ as usual

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The **sums of squares** are then defined as

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2,$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

The **sum of cross-products** is defined as

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

We now define the **residual sum of squares** function of β_0 and β_1 by

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 \tag{9.6}$$

We now derive expressions for the pair $(\widehat{\beta}_0, \widehat{\beta}_1)$ that minimises this function. One approach would be to compute partial derivatives $\frac{\partial}{\partial\beta_0}\mathrm{RSS}(\beta_0, \beta_1)$ and $\frac{\partial}{\partial\beta_1}\mathrm{RSS}(\beta_0, \beta_1)$, however there is a simpler approach.

9.2.1.1 Finding $\widehat{\beta}_0$

First, we find $\widehat{\beta}_0$. We note that Equation (9.6) can be rewritten as

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n}\left[(y_i - \beta_1 x_i) - \beta_0\right]^2.$$

It doesn't look as if much has changed, but if we write $z_i = y_i - \beta_1 x_i$, this becomes

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n}\left[z_i - \beta_0\right]^2.$$

And remember from Exercise 1.2.10 that for any value β_0, $\sum_{i=1}^{n}\left[z_i - \beta_0\right]^2 \geq \sum_{i=1}^{n}\left[z_i - \overline{z}\right]^2$. Therefore, setting

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i) = \overline{y} - \beta_1\overline{x},$$

we can conclude $\mathrm{RSS}(\widehat{\beta}_0, \beta_1) \leq \mathrm{RSS}(\beta_0, \beta_1)$, for all values of β_1.

9.2.1.2 Finding $\widehat{\beta}_1$

Having found $\widehat{\beta}_0$, we now need find the value of β_1 that minimises $\mathrm{RSS}(\widehat{\beta}_0, \beta_1)$.

$$\begin{aligned}
\mathrm{RSS}(\widehat{\beta}_0, \beta_1) &= \sum_{i=1}^{n} \left[(y_i - \beta_1 x_i) - \widehat{\beta}_0 \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \beta_1 x_i) - (\overline{y} - \beta_1 \overline{x}) \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \overline{y}) - \beta_1 (x_i - \overline{x}) \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \overline{y})^2 - 2\beta_1 (x_i - \overline{x})(y_i - \overline{y}) + \beta_1^2 (x_i - \overline{x})^2 \right] \\
&= S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx}
\end{aligned}$$

Completing the square (exercise) one can show

$$\begin{aligned}
\mathrm{RSS}(\widehat{\beta}_0, \beta_1) &= S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx} \\
&= S_{xx} \left(\beta_1 - \frac{S_{xy}}{S_{xx}} \right)^2 + \left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \right).
\end{aligned}$$

Therefore, $\mathrm{RSS}(\widehat{\beta}_0, \widehat{\beta}_1) \leq \mathrm{RSS}(\widehat{\beta}_0, \beta_1)$ where

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

9.2.1.3 Finding $\widehat{\beta}_0$ (continued)

Since we have now found the value for $\widehat{\beta}_1$, we have can substitute this value back into the expression for $\widehat{\beta}_0$:

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \beta_1 \overline{x} \\
\Rightarrow \widehat{\beta}_0 &= \overline{y} - \left(\frac{S_{xy}}{S_{xx}} \right) \overline{x}.
\end{aligned}$$

Summary

The values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of β_0 and β_1, respectively, that minimise the function

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i) \right]^2,$$

are given by

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \left(\frac{S_{xy}}{S_{xx}} \right) \overline{x}, \\
\widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}.
\end{aligned}$$

Exercise 9.2.1. Show that

$$S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx} = S_{xx}\left(\beta_1 - \frac{S_{xy}}{S_{xx}}\right)^2 + \left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right).$$

Rearranging and completing the square,

$$\begin{aligned}
S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx} &= S_{xx}\left[\beta_1^2 - 2\beta_1\frac{S_{xy}}{S_{xx}} + \frac{S_{yy}}{S_{xx}}\right] \\
&= S_{xx}\left[\beta_1^2 - 2\beta_1\frac{S_{xy}}{S_{xx}} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 - \left(\frac{S_{xy}}{S_{xx}}\right)^2 + \frac{S_{yy}}{S_{xx}}\right] \\
&= S_{xx}\left[\left(\beta_1 - \frac{S_{xy}}{S_{xx}}\right)^2 + \frac{S_{yy}}{S_{xx}} - \left(\frac{S_{xy}}{S_{xx}}\right)^2\right] \\
&= S_{xx}\left[\left(\beta_1 - \frac{S_{xy}}{S_{xx}}\right)^2 + \frac{1}{S_{xx}}\left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right)\right] \\
&= S_{xx}\left(\beta_1 - \frac{S_{xy}}{S_{xx}}\right)^2 + \left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right).
\end{aligned}$$

as desired.

△

9.2.2 Forbes' data with least squares

The Scottish physicist James D. Forbes sought to investigate the boiling point of water at different altitudes. His reason for doing this is that would provide a straightforward means for travellers, such as mountaineers, to determine their altitude above sea level.

It was already known that altitude could be determined from atmospheric air pressure measured with a barometer, but barometers in the 1840s were fragile instruments.

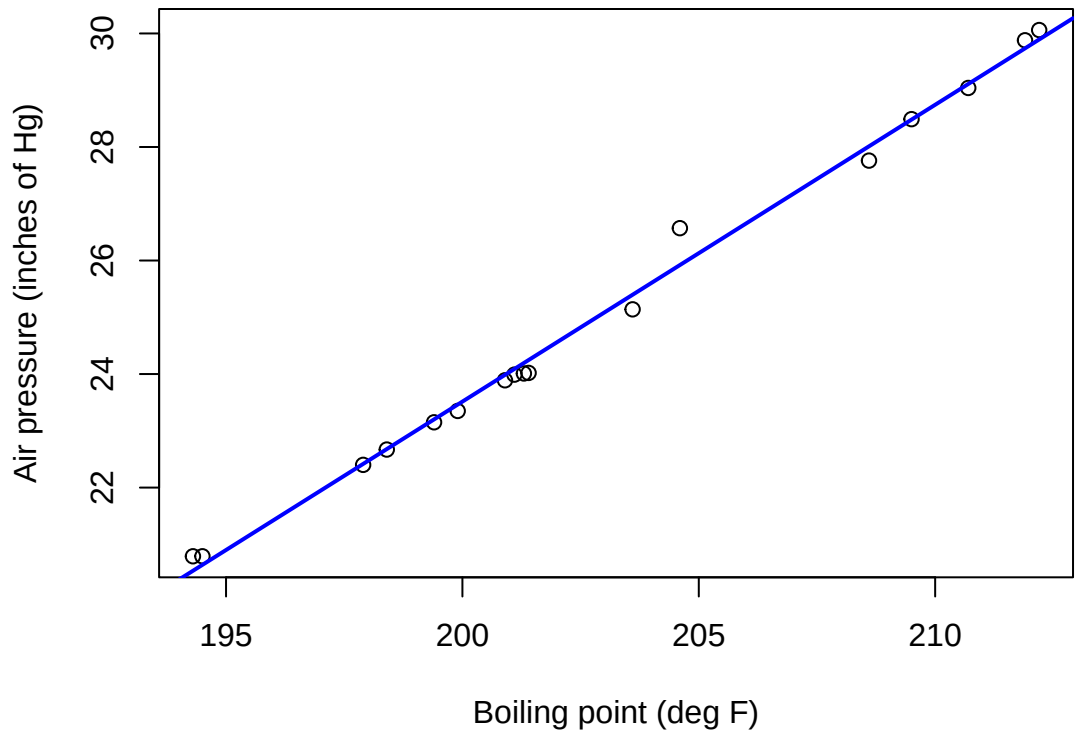
In fact, his approach was indirect: it was already known that altitude could be determined from barometric air pressure, so he decided to determine the relationship between the boiling point of water and air pressure (rather than trying to determine the relationship between the boiling point of water and altitude).

The units of air pressure that he used were 'inches of mercury', while the temperature for the boiling point of water was measured in Fahrenheit. The data he recorded are given in Table 9.1 below.

Boiling point (°F)	Air pressure (inches Hg)
194.50	20.79
194.30	20.79
197.90	22.40
198.40	22.67
199.40	23.15
199.90	23.35
200.90	23.89
201.10	23.99
201.40	24.02
201.30	24.01
203.60	25.14
204.60	26.57
209.50	28.49
208.60	27.76
210.70	29.04
211.90	29.88
212.20	30.06

Table 9.1: The data collected by James D. Forbes in the Alps and Scotland in the 1840s and 1850s. The boiling point of water is recorded in degrees Fahrenheit and the barometric air pressure is measured in inches of mercury (Hg).

```
library(MASS)
# print first 5 lines of the Forbes data; full set agrees with table above
print(head(forbes, n=5))
#>      bp  pres
#> 1 194.5 20.79
#> 2 194.3 20.79
#> 3 197.9 22.40
#> 4 198.4 22.67
#> 5 199.4 23.15

x <- forbes$bp
y <- forbes$pres
xbar <- mean(x)
ybar <- mean(y)
Sxx <- sum(  (x - xbar)^2  )
Syy <- sum(  (y - ybar)^2  )
Sxy <- sum(  (x - xbar) * (y - ybar)  )

beta1hat <- Sxy/Sxx
beta0hat <- ybar - beta1hat * xbar

plot(x, y, xlab="Boiling point (deg F)", ylab="Air pressure (inches of Hg)")
abline(a = beta0hat, b=beta1hat, col="blue", lwd=2)
```



9.3 The simple linear regression model

Suppose, as above, that there are two quantities, or variables, of interest, X and Y, between which we believe there is a linear relationship. Again, measure suppose we measure n pairs of values (x_i, y_i), for $i \in \{1, 2, \ldots, n\}$.

We now define a new model, known as the **simple linear regression** model. For $i \in \{1, 2, \ldots, n\}$,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{9.7}$$

where

- the x_i are assumed to be fixed, known values (we have measured them),
- the parameters β_0, β_1 are fixed, unknown parameters,
- the ϵ_i are independent random variables,
- $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, for some unknown σ^2.

The Y_i are random variables, depending on the observed x_i and the unobserved random errors ϵ_i. In the experiment, the Y_i are observed as y_i for $i \in \{1, 2, \ldots, n\}$.

Remark 9.3.1. The reason this model is called **simple** is not because it is 'easy', but rather because the model assumes only one independent quantity, the $\mathbf{x}$ values. Another type of regression is **multiple** linear regression, where there are multiple independent quantities, i.e. $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon$. □

Remark 9.3.2. The reason this model is called **linear** is that each random variable Y_i is a linear function of the parameters β_0, β_1. In fact, it is possible to transform the Y_i and the x_i using functions $f, g : \mathbb{R} \to \mathbb{R}$, i.e. $f(Y_i) = \beta_0 + \beta_1 g(x_i) + \epsilon_i$, and this would still be considered linear regression. Possibilities for f (and g) include $f(z) = z^2$, $f(z) = \sqrt{z}$, $f(z) = \log(z)$, $f(z) = \exp(z)$, etc. However, an example of a nonlinear regression model would be $Y_i = f(x_i, \beta_0, \beta_1) + \epsilon_i$, where the function $f(x_i, \beta_0, \beta_1)$ combines the quantities x_i, β_0 and β_1 in a nonlinear manner, e.g. $Y_i = \beta_0 + (1 - \beta_0)e^{-\beta_1(x_i - 2)} + \epsilon_i$. □

Remark 9.3.3. The choice of $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ is an **assumption**. It may or may not be a reasonable assumption (the errors may follow a different distribution), but this choice implies that the errors can be positive or negative, and that large positive values and small negative values (relative to σ^2) are unlikely. This is also known as the **conditional normal** model. □

Remark 9.3.4. The reason that the random errors ϵ_i are assumed to have mean $\mathrm{E}(\epsilon_i) = 0$ is that if one assumed that $\mathrm{E}(\epsilon_i) = \mu$ for some unknown value $\mu \neq 0$, then one could simply reparametrize $\epsilon_i' = \epsilon_i - \mu$ and $\beta_0' = \beta_0 + \mu$ and then use the alternate model $Y_i = \beta_0' + \beta_1 x_i + \epsilon_i'$, where $\mathrm{E}(\epsilon_i') = 0$. In other words, the intercept term β_0 'absorbs' any non-zero mean of the random errors. □

Using the linearity of the expectation and properties of the variance operator, We can compute

$$\begin{aligned} \mathrm{E}\left(Y_i\right) &= \mathrm{E}\left(\beta_0 + \beta_1 x_i + \epsilon_i\right) = \beta_0 + \beta_1 x_i + \mathrm{E}\left(\epsilon_i\right) = \beta_0 + \beta_1 x_i + (0) \\ \Rightarrow \mathrm{E}\left(Y_i\right) &= \beta_0 + \beta_1 x_i \end{aligned}$$

$$\begin{aligned} \mathrm{Var}\left(Y_i\right) &= \mathrm{Var}\left(\beta_0 + \beta_1 x_i + \epsilon_i\right) = \mathrm{Var}\left(\epsilon_i\right) = \sigma^2 \\ \Rightarrow \mathrm{Var}\left(Y_i\right) &= \sigma^2 \end{aligned}$$

Furthermore, since the ϵ_i are independent and normally-distributed, the Y_i are also independent and normally-distributed, and

$$Y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right),$$

with the parameters β_0, β_1 and σ^2 unknown.

9.3.1 Estimating the parameters

Since the model assumptions imply that the Y_i are independent and

$$Y_i \sim \mathrm{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right),$$

one can construct a likelihood and use use maximum likelihood estimation to obtain estimates for the parameters β_0, β_1 and σ^2. First, one needs to construct the likelihood. The probability density function for Y_i is

$$f(y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right],$$

for $i \in \{1, 2, \ldots, n\}$. Writing $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, using the independence of the $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$,

$$\begin{aligned} f(\mathbf{y}|\beta_0, \beta_1, \sigma^2) &= f(y_1, y_2, \ldots, y_n|\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i|\beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right], \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right]. \end{aligned}$$

Therefore, the likelihood is

$$L(\beta_0, \beta_1, \sigma^2|\mathbf{x}, \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right],$$

and one can also compute the log-likelihood to be

$$\log L(\beta_0, \beta_1, \sigma^2|\mathbf{x}, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2.$$

For a fixed value of σ^2, maximising the log-likelihood is equivalent to minimising the quantity

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2.$$

But, this expression is the same as the expression for the residual sum of squares in Equation (9.6),

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2.$$

We have already found in Section 9.2.1 that the maximum likelihood estimates for β_0 and β_1 (the estimates which minimies $\mathrm{RSS}(\beta_0, \beta_1)$) are

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y} - \left(\frac{S_{xy}}{S_{xx}}\right)\overline{x} \\ \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}}.\end{aligned}$$

All that is left is to find the maximum likelihood estimate of σ^2, i.e. the value of σ^2 that maximises the (log)-likelihood

$$\log L(\sigma^2 | \mathbf{x}, \mathbf{y}, \widehat{\beta}_0, \widehat{\beta}_1) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.$$

In this likelihood, the only unknown quantity is σ^2. In an almost identical approach to finding the MLEs for Example 8.5.6 (see the solution to Question 2 on Problem Sheet 13), we can take the derivative with respect to σ^2:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\log L(\sigma^2 | \mathbf{x}, \mathbf{y}, \widehat{\beta}_0, \widehat{\beta}_1) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2,$$

and setting

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\log L(\sigma^2 | \mathbf{x}, \mathbf{y}, \widehat{\beta}_0, \widehat{\beta}_1) = 0 \\ \Rightarrow &-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 = 0 \\ \Rightarrow &\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.\end{aligned}$$

Therefore, the maximum likelihood estimate for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2 = \frac{1}{n}\left(\widehat{\mathrm{RSS}}_{xy}\right),$$

if we define

$$\widehat{\mathrm{RSS}}_{xy} = \mathrm{RSS}(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2,$$

Remark 9.3.5. We include the subscripts x and y in $\widehat{\mathrm{RSS}}_{xy}$ to indicate that the quantity depends on the data $\mathbf{x}$ and $\mathbf{y}$. Later, we shall define a similar (but random) quantity $\widehat{\mathrm{RSS}}_{xY}$ which depends on the data $\mathbf{x}$ and the random variables $\mathbf{Y}$. □

Summary

Given the simple linear regression (conditional normal) model,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where the $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and are independent, the maximum likelihood estimators for the parameters β_0, β_1 and σ^2 are

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \left(\frac{S_{xy}}{S_{xx}}\right)\overline{x} \\
\widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
\widehat{\sigma}^2 &= \frac{1}{n}\left(\widehat{\mathrm{RSS}}_{xy}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.
\end{aligned}$$

9.3.2 Residuals

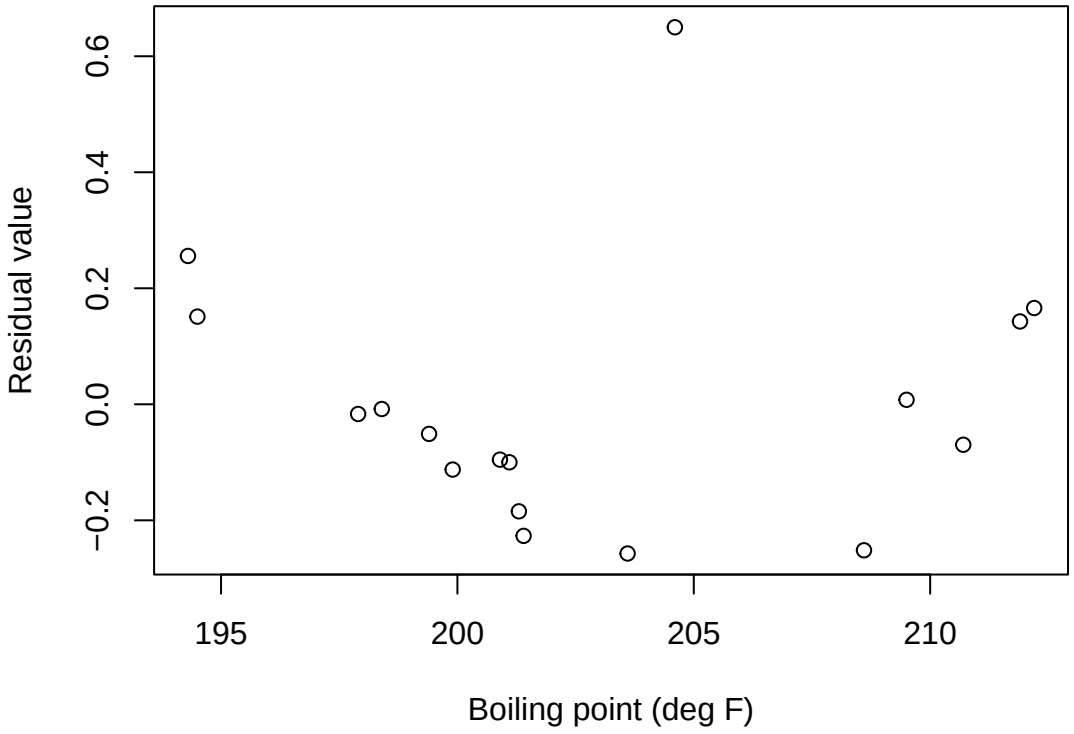
Given the maximum likelihood estimates $\widehat{\beta}_0, \widehat{\beta}_1$, we fit the straight line specified by the points $(x_i, f(x_i))$, where $f(x_i) = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. We can now observe the **residuals**

$$\widehat{\epsilon}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i).$$

If our model is correct, then these residuals $\widehat{\epsilon}_i$ are realisations of the (unobservable) random errors ϵ_i. According to our model, these $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and are independent. Therefore, **if our model is correct** the residuals $\widehat{\epsilon}_i$, when plotted, should appear to be independently distributed according to some $\mathrm{N}(0, \sigma^2)$.

If we return to the Forbes' data set, we decided to fit a line using a least squares approach relating the air pressure and boiling point. Since the estimates $\widehat{\beta}_0, \widehat{\beta}_1$ for the least squares approach are the same as that for the simple linear regression model, the fitted lines would be the same, and therefore the residuals would be the same. If we plot these residuals (below), we notice that they appear to follow more of a 'U' shape, rather than be randomly distributed around 0. This suggests that our model may be incorrect.

```
library(MASS)
x <- forbes$bp
y <- forbes$pres
xbar <- mean(x)
ybar <- mean(y)
Sxx <- sum(  (x - xbar)^2  )
Syy <- sum(  (y - ybar)^2  )
Sxy <- sum(  (x - xbar) * (y - ybar)  )

beta1hat <- Sxy/Sxx
beta0hat <- ybar - beta1hat * xbar

residuals <- y - (beta0hat + beta1hat * x)
plot(x, residuals, xlab="Boiling point (deg F)", ylab="Residual value")
```



9.3.3 The `lm` function in R

Before trying another model for the Forbes' data, we introduce the `lm` function (for **l**inear **m**odel) in R which is very useful for computing the parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The only unusual feature of this function is that it uses the tilde ~ in its function call. We again use the Forbes' data as an example.

```
library(MASS)
x <- forbes$bp
y <- forbes$pres
xbar <- mean(x)
ybar <- mean(y)
Sxx <- sum(  (x - xbar)^2  )
Syy <- sum(  (y - ybar)^2  )
Sxy <- sum(  (x - xbar) * (y - ybar)  )

# computing the parameters and the residuals
beta1hat <- Sxy/Sxx
beta0hat <- ybar - beta1hat * xbar
beta_hat <- c(beta0hat, beta1hat)
residuals <- y - (beta0hat + beta1hat * x)

# here we use the lm function; note the use of the tilde "~"
model <- lm(y ~ x)

# Comparing the parameters computed above and using lm, they agree
print(cbind(beta_hat, model$coefficients))
#>               beta_hat
#> (Intercept) -81.06373 -81.06373
#> x             0.52289   0.52289

# Compare the residuals computed above and using lm, they agree
# Only the first 5 are shown here using `head`, but all 17 agree
print(head(cbind(residuals, model$residuals), n=5))
#>    residuals
#> 1  0.1511552  0.1511552
#> 2  0.2557337  0.2557337
#> 3 -0.0166790 -0.0166790
#> 4 -0.0081252 -0.0081252
#> 5 -0.0510176 -0.0510176
```

This is just an example, but it shows how useful the `lm` function is. Simply calling `model <- lm(y ~ x)`, we can obtain the parameters $\widehat{\beta}_0, \widehat{\beta}_1$ from `model$coefficients` and the residuals ϵ_i from `model$residuals`.

9.3.4 Return to Forbes' data

Let us try fitting the linear model again after first transforming the y_i and/or the x_i. Specifically, if we define the x_i to be the observations for boiling point and the y_i to be the

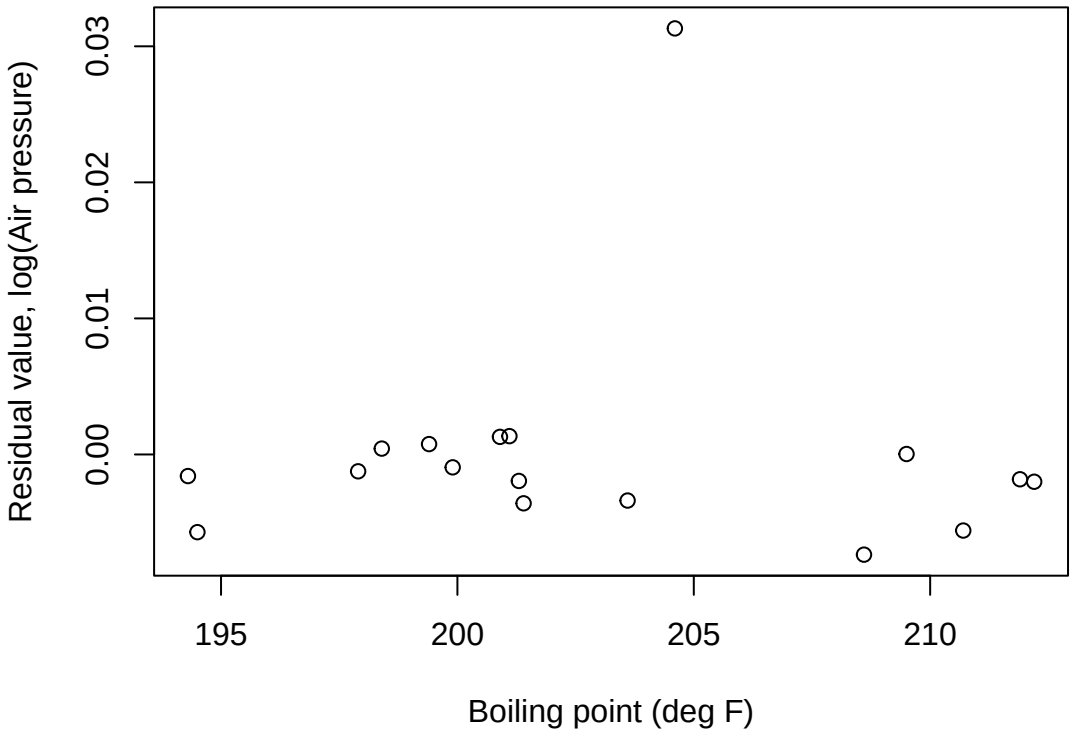
observations for the air pressure for $i \in \{1, 2, \ldots, n\}$, let's consider the transformation

$$Z_i = \log(Y_i),$$

and then fit the simple linear regression model

$$Z_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{9.8}$$

```
library(MASS)
x <- forbes$bp
y <- forbes$pres
z <- log(y)
model_2 <- lm(z ~ x)
residuals <- model_2$residuals
ylab <- "Residual value, log(Air pressure)"
plot(x, residuals, xlab="Boiling point (deg F)", ylab=ylab)
```



These residuals appear to be centred around 0 which suggests that this model is a better fit to the data and

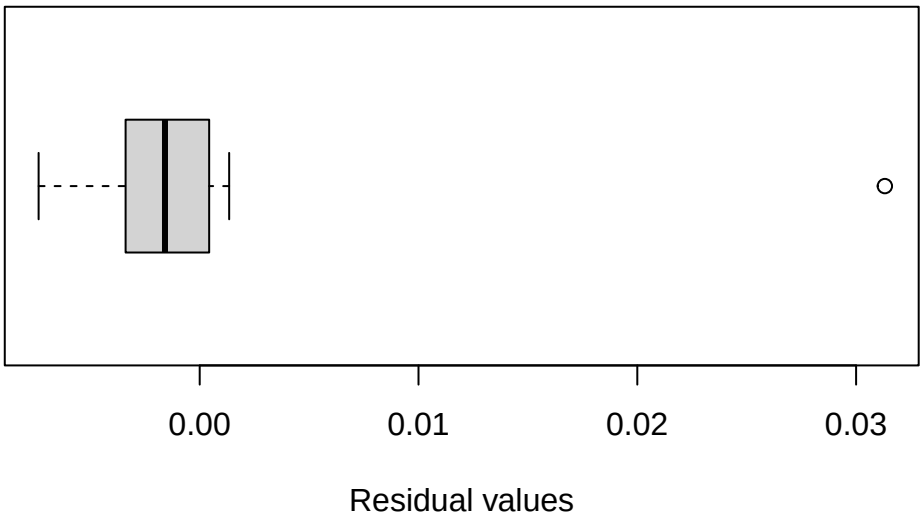
$$\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{9.9}$$

may be a better model than simply $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

9.3.4.1 Identifying any outliers

However, the 12th residual at boiling point 204.6 appears to be much larger than the others. It could be that this data point was recorded incorrectly, with a larger error than the others, or it could be a true data point and our model that assumes normally-distributed errors is incorrect. If we decide that the observation is the result of a measurement error, we could see if its residual value is an outlier using a boxplot and Tukey's criterion.

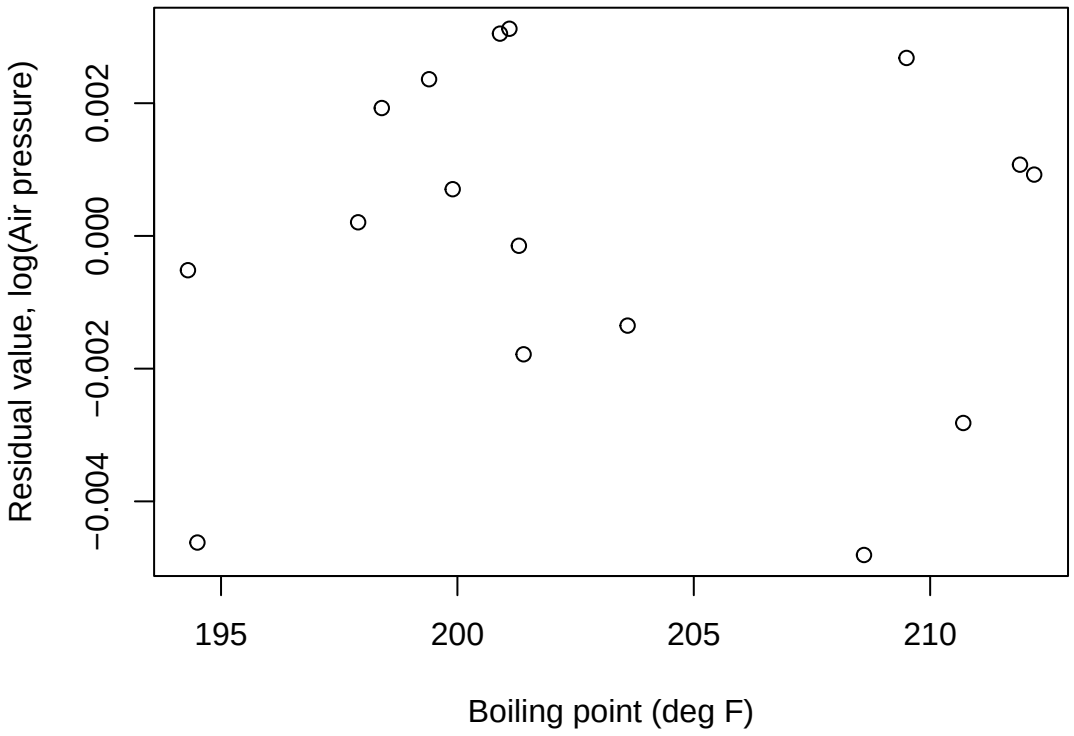
```
boxplot(model_2$residuals, horizontal=TRUE, xlab="Residual values")
```



This point appears to be an outlier, and if we remove it and replot the residuals, the residual plot appears to show values centred around 0, which suggests that this model is a better fit to the data.

9.3.4.2 The residuals with outliers removed

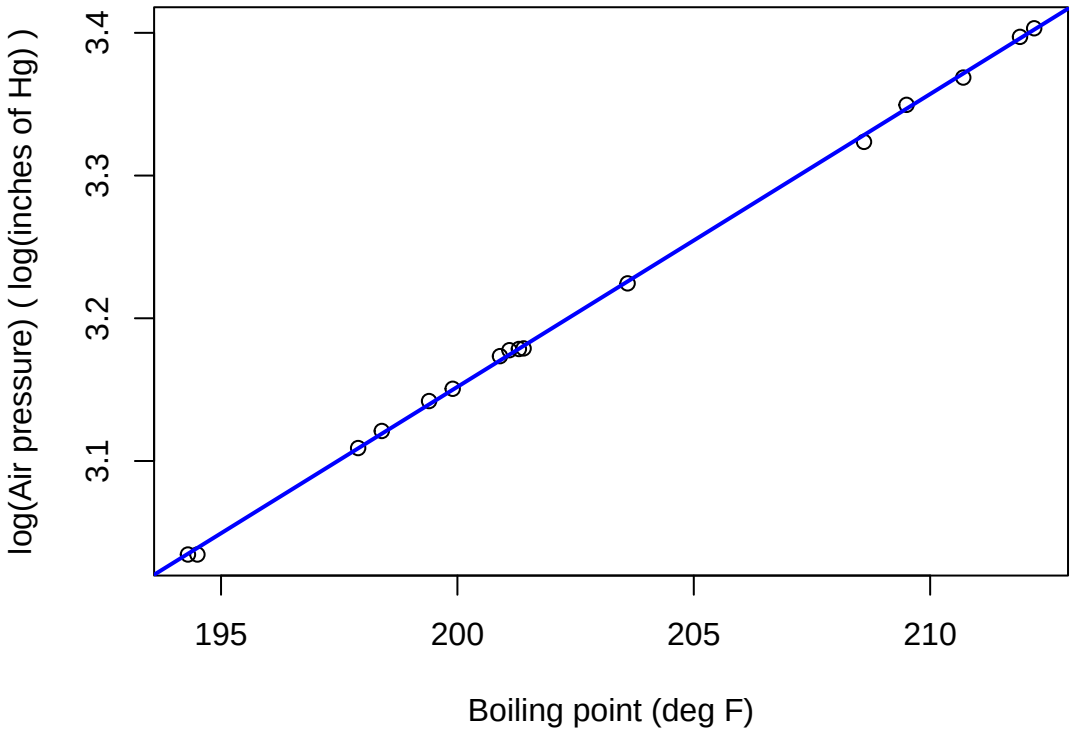
```
library(MASS)
# the outlier is the 12th value, so remove this value to define the inliers
x <- forbes$bp[-12]
y <- forbes$pres[-12]
z <- log(y)
model_3 <- lm(z ~ x)

residuals <- model_3$residuals
xlab <- "Boiling point (deg F)"
ylab <- "Residual value, log(Air pressure)"
plot(x, residuals, xlab=xlab, ylab=ylab)
```



9.3.4.3 Refitting the model

If we finally plot the fitted regression line for this model, we see that there is a good fit, as suggested by the plot of the residuals.

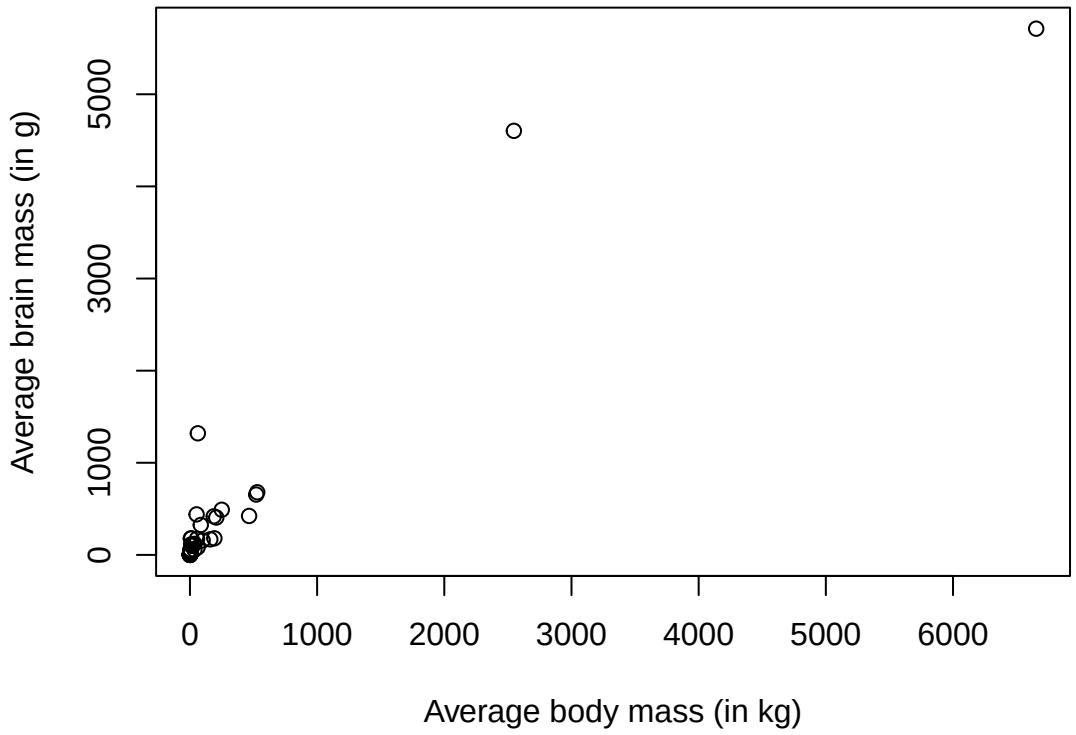
```
library(MASS)
# the outlier is the 12th value, so remove this value to define the inliers
x <- forbes$bp[-12]
y <- forbes$pres[-12]
z <- log(y)

# compute the parameters of the linear model
model_3 <- lm(z ~ x)

# extract parameter coefficients from model_3 object
beta0hat <- model_3$coefficients[1]
beta1hat <- model_3$coefficients[2]

# plot the (transformed) values and the regression line
ylab <- "log(Air pressure) ( log(inches of Hg) )"
plot(x, z, xlab="Boiling point (deg F)", ylab=ylab)
abline(a = beta0hat, b=beta1hat, col="blue", lwd=2)
```



9.3.5 Example: mammals data

Let us look at another dataset from the `MASS` package. The `mammals` dataset gives the average body mass (in kg) and the average brain mass (in g) for 62 land mammals. Plotting the raw data we see:

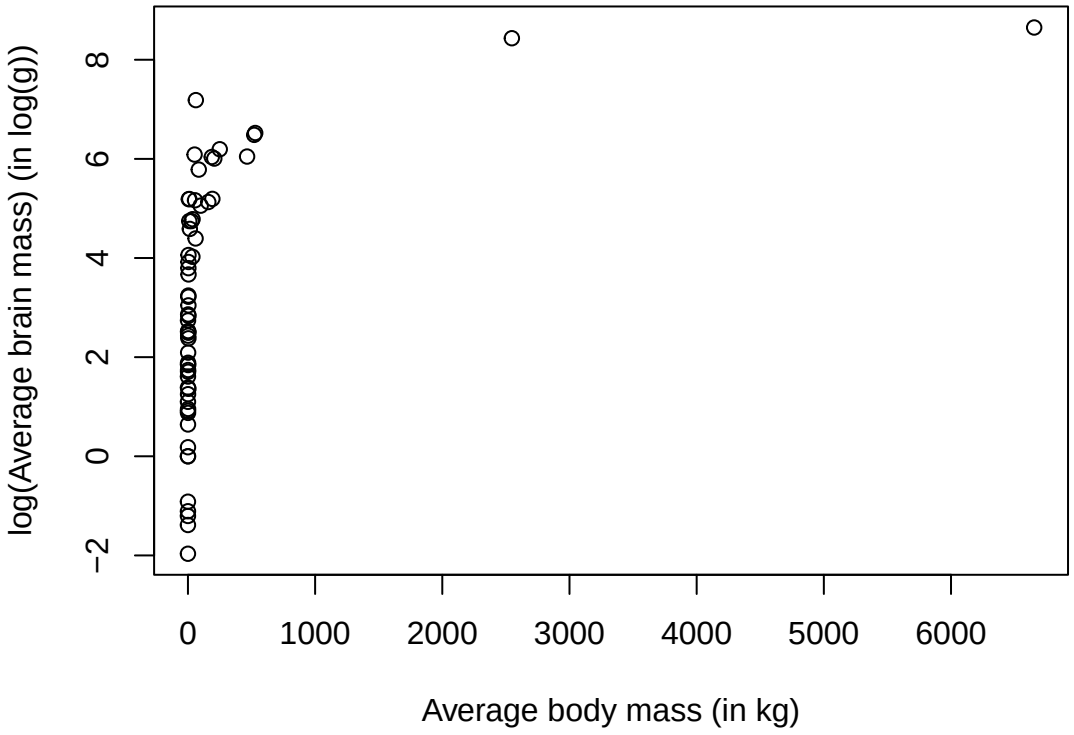
```
library(MASS)
xlab="Average body mass (in kg)"
ylab="Average brain mass (in g)"
plot(x=mammals$body, y=mammals$brain, xlab=xlab, ylab=ylab)
```



However, the plot seems to be distorted by two large values.

If we take the logarithm of the brain masses:

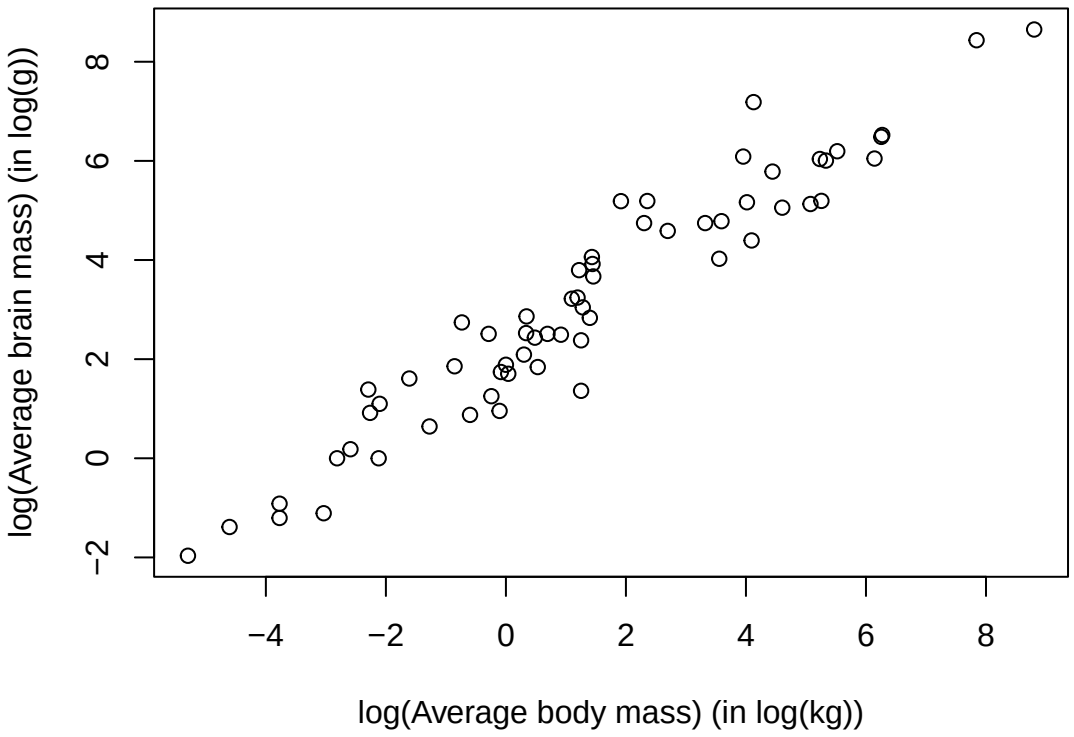
```
library(MASS)
xlab="Average body mass (in kg)"
ylab="log(Average brain mass) (in log(g))"
plot(x=mammals$body, y=log(mammals$brain), xlab=xlab, ylab=ylab)
```



If spreads out the values along the brain mass axis, but this is not much better.

If we take the logarithm of both the brain and body masses:

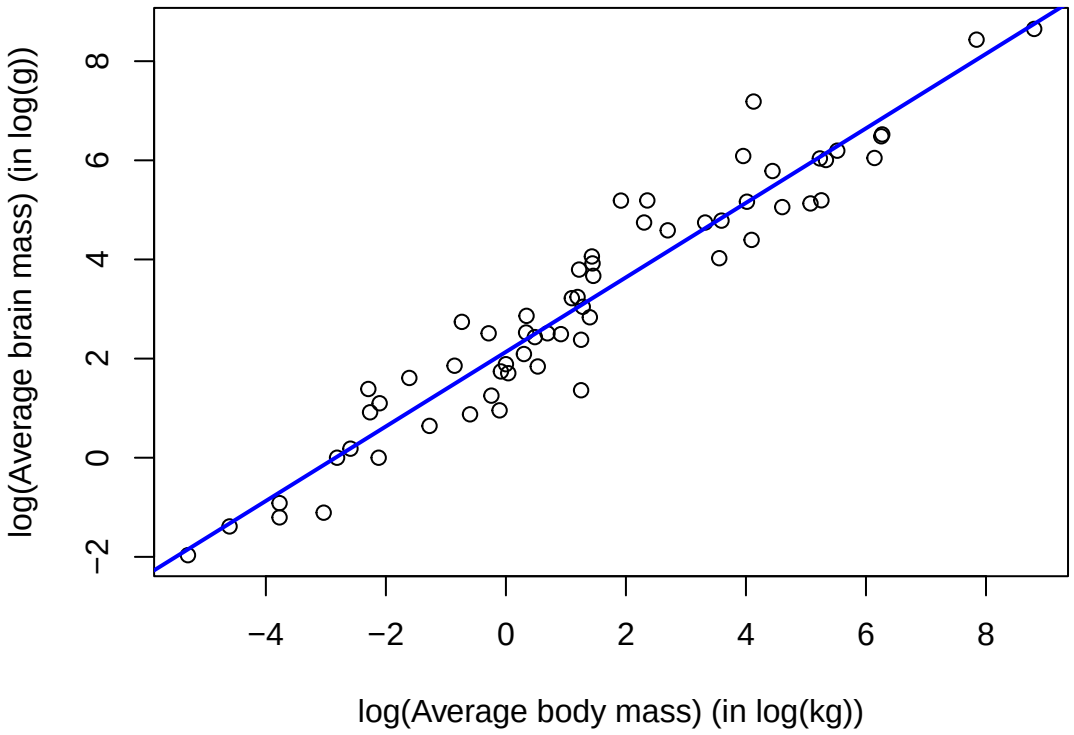
```
library(MASS)
xlab="log(Average body mass) (in log(kg))"
ylab="log(Average brain mass) (in log(g))"
plot(x=log(mammals$body), y=log(mammals$brain), xlab=xlab, ylab=ylab)
```



A clearer picture emerges.

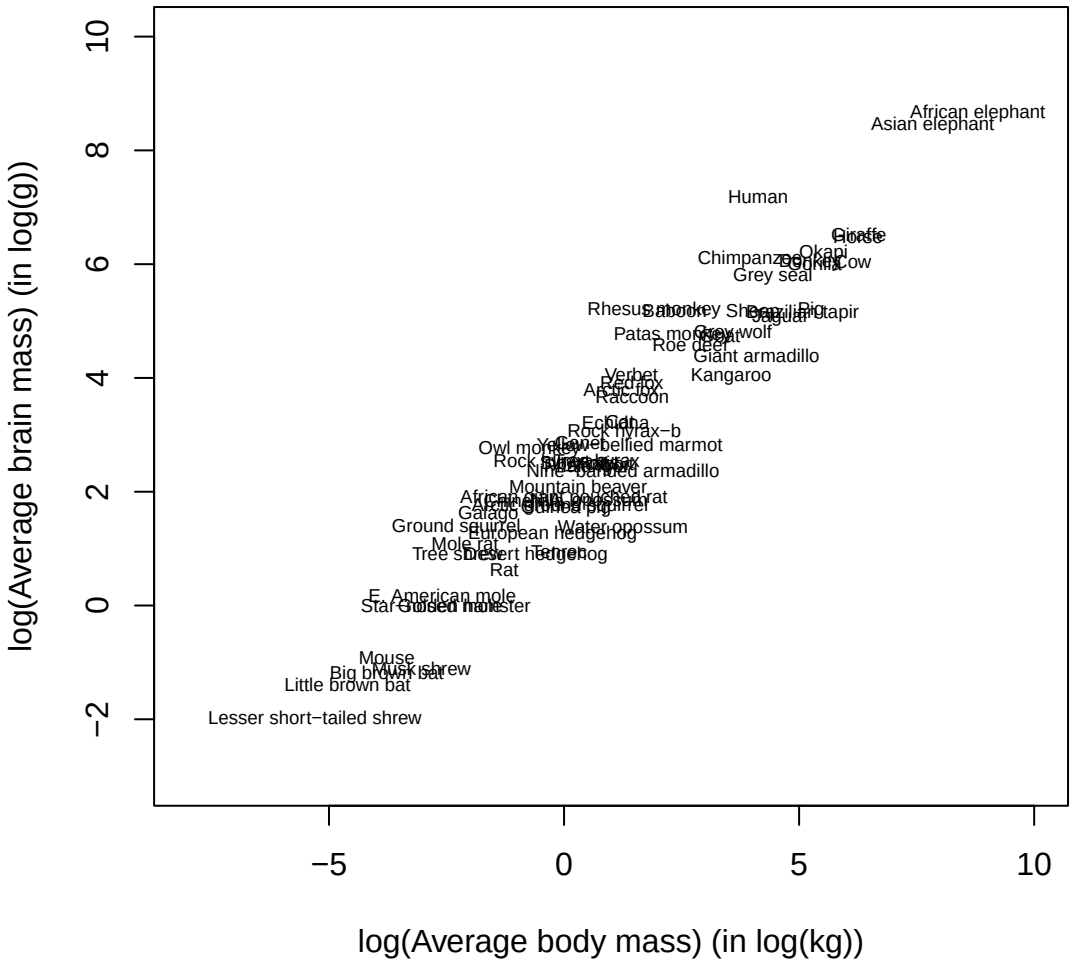
This plot suggests a linear relationship:

```
library(MASS)
x <- log(mammals$body)
y <- log(mammals$brain)
model <- lm(y ~ x)
xlab="log(Average body mass) (in log(kg))"
ylab="log(Average brain mass) (in log(g))"
plot(x=x, y=y, xlab=xlab, ylab=ylab)
abline(a=model$coefficients[1], b=model$coefficients[2], lwd=2, col="blue")
```



So, in this case, it appears that there is a linear relationship between the logarithm of a mammal's brain mass and the logarithm of a mammal's body mass. We can also plot the data with the names of the mammals (instead of circles) in order to see which data point corresponds to which mammal.

```
library(MASS)
x <- log(mammals$body)
y <- log(mammals$brain)
xlab="log(Average body mass) (in log(kg))"
ylab="log(Average brain mass) (in log(g))"

# plotting the data, but without points, and slightly extending the axes
# the purpose of this command is to get the correct axes
plot(x=x, y=y, xlab=xlab, ylab=ylab, pch=NA, xlim=c(-8,10), ylim=c(-3, 10))

#now adding the names, and making the text slightly smaller (cex=0.6)
names <- rownames(mammals)
text(x=x, y=y, label=names, cex=0.6)
```



We will not investigate the goodness of fit for this data set (try this yourself), it is meant to be an example to show how transforming the data can lead to the discovery of a linear relationship where one was unapparent from the raw (untransformed) data.

9.4 The R^2 statistic

Recall from Section 9.3.1 that the computed values for the fitted coefficients are

$$\widehat{\beta}_0 = \overline{y} - \left(\frac{S_{xy}}{S_{xx}}\right)\overline{x},$$
$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

and the estimated residual sum of squares is defined as

$$\widehat{\mathrm{RSS}}_{xy} = \sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.$$

Also recall the sum of squares S_{yy} defined in Section 9.2.1:;

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Definition 9.1. The statistic R^2 (pronounced '**R-squared**') is defined as

$$R^2 = \frac{S_{yy} - \widehat{\mathrm{RSS}}_{xy}}{S_{yy}}. \tag{9.10}$$

Remark 9.4.1. Interpreting the definition of R^2, it measures the difference between the sum of squares and the estimated residual sum of squares, normalised by the sum of squares. □

We may wonder what range of values R^2 can take. First, Exercise 9.4.3 shows that we can rewrite $\widehat{\mathrm{RSS}}_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$. Using this result,

$$R^2 = \frac{S_{yy} - \widehat{\mathrm{RSS}}_{xy}}{S_{yy}} = \frac{S_{yy} - \left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right)}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx}S_{yy}}.$$

Now, we recognise that this is the square of the sample correlation from Definition 6.3,

$$R^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = \frac{\left(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} = r_{XY}^2.$$

Therefore, $R^2 \in [0, 1]$, since $r_{XY} \in [-1, 1]$ (although this was not explicitly shown, it was shown in Corollary 6.2.3 that $\rho_{XY} \in [-1, 1]$. To prove $r_{XY} \in [-1, 1]$ directly, one could use the Cauchy-Schwartz inequality).

Remark 9.4.2. Sometimes the R^2 statistic is quoted as evidence that a model fits the data well. This usually happens when the R^2 statistic is close to 1. While an R^2 value close to 1 does indicate that the fitted line is 'close' to the data, we must remember that it R^2 is only the square of the correlation. The next section explores this further. □

Exercise 9.4.3. Show that

$$\widehat{\mathrm{RSS}}_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}.$$

$$\begin{aligned}
\widehat{\mathrm{RSS}}_{xy} &= \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right]^2 \\
&= \sum_{i=1}^{n} \left[y_i - \left(\overline{y} - \frac{S_{xy}}{S_{xx}} \overline{x} + \frac{S_{xy}}{S_{xx}} x_i \right) \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \overline{y}) - \frac{S_{xy}}{S_{xx}} (x_i - \overline{x}) \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \overline{y}) - 2\frac{S_{xy}}{S_{xx}} (x_i - \overline{x}) (y_i - \overline{y}) + \left(\frac{S_{xy}}{S_{xx}} \right)^2 (x_i - \overline{x})^2 \right] \\
&= \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2\frac{S_{xy}}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y}) + \left(\frac{S_{xy}}{S_{xx}} \right)^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= S_{yy} - 2\frac{S_{xy}}{S_{xx}} S_{xy} + \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx} \\
&= S_{yy} - 2\frac{(S_{xy})^2}{S_{xx}} + \left(\frac{(S_{xy})^2}{S_{xx}} \right) \\
&= S_{yy} - \frac{(S_{xy})^2}{S_{xx}}
\end{aligned}$$

as required.

△

9.5 Evaluating the fit of a model

Let us look at the Forbes' data again, with the 12th data point removed (remember, this point was an outlier). We reconsider the two models:

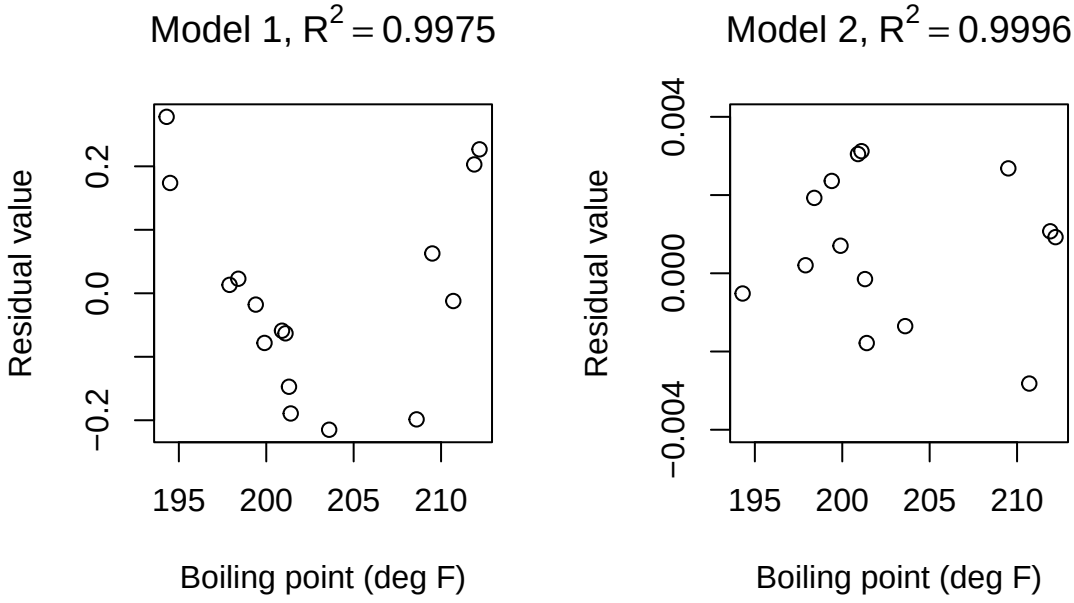
$$\begin{aligned} \text{Model 1:} \qquad & Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \\ \text{Model 2:} \qquad & \log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i, \end{aligned}$$

where $i \in \{1, 2, \ldots, 11, 13, 1415, 16, 17\}$.

Below we plot the residual plots of each model side-by-side. We note that for Model 1, which is incorrect, $R^2 = 0.9975$, while for Model 2, $R^2 = 0.9996$. Although the R^2 value for Model 2 is higher than that for Model 1, what is striking is that the R^2 value for Model 1 is still very close to 1, even though Model 1 is the incorrect model, as can be seen from its residual plot.

If one had only tried Model 1, and only considered the R^2 value without looking at the residual plot, one might have been tempted to consider Model 1 as fitting the data well because its R^2 value is close to 1. However, the residual plot for Model 1 has a 'U'-shape, and clearly shows that Model 1 does not fit the data well.

This shows that one must exercise caution when trying to interpret R^2 values, and one should always look at the residual plots in order to determine whether or not a model fits the data well.



Chapter 10

Bayesian Inference

This chapter provides a brief introduction to Bayesian inference, introducing the concepts of prior and posterior distributions. The definitions and examples are from [3, 2, 7].

10.1 Frequentist vs Bayesian inference

All of the statistics we have done now has been from the **frequentist** perspective, in that a (unknown) parameter θ is assumed to be a fixed but unknown value, and the goal is to estimate this value as closely as possible. A probability in this perspective is considered to be the frequency of an event occurring when an experiment is repeated a large number of times, e.g. the probability of a coin landing heads up is considered as the proportion of times the coin will land heads up over many coin tosses.

The **Bayesian** perspective is different: it treats a parameter θ as a random variable and attempts to estimate its **distribution**. At the beginning an initial (prior) distribution is assumed for the parameter and then after data is observed, this distribution is updated (to the posterior distribution). A probability in this perspective is considered as a **degree of belief**. If we considered the event 'the coin will land heads up on the next toss', then assigning a probability of 1 means we believe it will definitely lands heads up, and assigning a probability of 0 means the coin will definitely not land heads up. However, we may decide that θ, the probability of the coin landing heads up, is some value between 0 and 1.

If we had no particular belief for the value of θ, we may decide to initially assume that $\theta \sim \text{Unif}(0, 1)$. On the other hand, if we believed that θ was more likely to be close to 0.5, we may decide that to initially assume that $\theta \sim \text{Beta}(2, 2)$. After acquiring data by tossing the coin, and we would then update our belief about the distribution of θ using this data. These initial and updated distributions are called the **prior** and **posterior** distributions.

10.2 Prior and posterior distributions

We first recall the definitions for the likelihood function and the marginal distribution.

Definition 10.1. When the conditional joint p.d.f. (p.m.f.), denoted $f(\mathbf{x}|\theta)$, of the observations in a random sample is considered as a function of θ for given values $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, it is called the **likelihood function** and is denoted by $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$.

Definition 10.2. Suppose that the continuous random variables $\mathbf{X}$ and θ have a joint distribution denoted by $f(\mathbf{x}, \theta)$ and that the support of θ is the set Θ. Then the **marginal distribution** of $\mathbf{X}$ is the distribution of $\mathbf{X}$ derived from this joint distribution by

$$m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}, \theta) \mathrm{d}\theta.$$

In the case that θ is discrete, the marginal distribution is simply the summation $m(\mathbf{x}) = \sum_{\theta \in \Theta} f(\mathbf{x}, \theta)$.

likelihood /ˈlʌɪklɪhʊd/	**marginal** /ˈmɑːdʒɪn(ə)l/
The state or fact of something's being **likely**; probability	1. relating to or at the **edge** or margin
Origins: *likely* + *-hood* (state of being)	2. minor or not important
Source: Oxford English Dictionary	Origins: Latin, *margo, margin-*, meaning *edge*

Remark 10.2.1. While $f(\mathbf{x}|\theta)$ is the conditional p.d.f. (p.m.f.) of $\mathbf{X}$ given θ, it is often referred to as the **likelihood function**, as defined in Definitions 8.1 and 10.1 and Equation (8.1). Note that $f(\mathbf{x}|\theta)$ is sometimes referred to as the **sampling distribution** [2]. □

Definition 10.3. Suppose that one has a statistical model with parameter θ, and one treats θ as random. Then the distribution one assigns to θ before observing any other random variables of interest is called the **prior distribution** and its p.d.f. (p.m.f.) is denoted by $\pi(\theta)$.

Definition 10.4. Suppose that one has a statistical inference problem with unknown parameter θ, and there are random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ which are observed as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The conditional distribution of θ given $\mathbf{X}$ is called the **posterior distribution** of θ, and its p.d.f. (p.m.f.) of θ given $\mathbf{X} = \mathbf{x}$ is usually denoted by $\pi(\theta|\mathbf{x})$.

prior /ˈprʌɪə/	**posterior** /pɑˈstɪərɪə/
Existing or coming **before** in time, order, or importance.	Coming **after** in time or order; later.
Origins: Latin, *prior*, meaning *previous*, *earlier*, *preceding*, *former*	Origins: Latin, *post* means *after*
Source: Oxford English Dictionary	

These definitions lead to the following important result, which is also referred to as Bayes' Theorem; you have already seen Bayes' Theorem in Term 1 using probabilities of events, this is a different version using probability distributions.

Theorem 10.2.2. Suppose the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ have a joint distribution with p.d.f. (p.m.f.) $f(\mathbf{x}|\theta)$. Suppose also that the value of the parameter θ is unknown and the prior p.d.f. (p.m.f.) for θ is $\pi(\theta)$. Then the posterior p.d.f. (p.m.f.) is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}, \tag{10.1}$$

where $m(\mathbf{x})$ is the marginal joint p.d.f. (p.m.f.) of $\mathbf{X}$. ♦

Proof. For simplicity, assume that the parameter space Θ is either an interval of the real line (or the entire real line) and that $\pi(\theta)$ is a prior p.d.f. on Θ rather than a prior p.m.f.. However, the proof can be adapted to the case that $\pi(\theta)$ is a prior p.m.f..

Multiplying the conditional joint p.d.f. $f(\mathbf{x}|\theta)$ by the prior $\pi(\theta)$ results in the $(n+1)$-dimensional joint p.d.f. of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and θ,

$$f(\mathbf{x}, \theta) = f(\mathbf{x}|\theta)\pi(\theta). \tag{10.2}$$

(See Definition 14.2.1 in the Term 1 lecture notes.) The marginal joint p.d.f. $m(\mathbf{x})$ of $\mathbf{X}$ can then be obtained by integrating the right-hand side of Equation (10.2) over all values of $\theta \in \Theta$,

$$m(\mathbf{x}) = \int_\Theta f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta. \tag{10.3}$$

Then, the conditional p.d.f. of θ given that $\mathbf{X} = \mathbf{x}$, denoted $\pi(\theta|\mathbf{x})$, must then be

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

(The first equality uses Definition 14.2.1 in the Term 1 lecture notes.) This is Bayes's theorem restated for parameters and random samples. Although we assumed for Equation (10.3) that $\pi(\theta)$ was a p.d.f. on an interval of the real line, in the case it is discrete, one can replace the integral by the appropriate sum over all possible values of θ. □

Table 10.3: List of symbols in Equation (10.1) of Theorem 10.2.2.

Symbol	Meaning
$\pi(\theta)$	The **prior** distribution of θ.
$\pi(\theta\|\mathbf{x})$	The **posterior** distribution of θ given the data $\mathbf{x}$.
$f(\mathbf{x}\|\theta)$	The conditional joint p.d.f. (p.m.f.) of the random variables $\mathbf{X}$ given the parameter θ; often referred to as the **likelihood function**. Sometimes called the sampling distribution.
$m(\mathbf{x})$	The **marginal** joint distribution of $\mathbf{x}$.

Example 10.2.3. Suppose the proportion θ of defective lightbulbs produced in a particular large shipment is unknown, and one wishes to estimate θ. With no knowledge of θ, suppose one chooses the prior to be the uniform distribution on the interval $[0, 1]$, i.e.

$$\pi(\theta) = \begin{cases} 1, & \text{if } 0 < \theta < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Suppose a random sample of n lightbulbs is taken from the shipment, and for $i = 1, 2, \ldots, n$ let the random variable $X_i = 1$ if the ith lightbulb is defective and let $X_i = 0$ otherwise. Then the independent random variables $X_1, X_2, \ldots, X_n$ form n Bernoulli trials with parameter θ, and so the p.m.f. for each X_i is

$$f(x_i|\theta) = \begin{cases} \theta^{x_i}(1-\theta)^{1-x_i}, & \text{if } x_i \in \{0, 1\}, \\ 0, & \text{otherwise.} \end{cases}$$

Recall that $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, which implies $n\overline{x} = \sum_{i=1}^{n} x_i$. Then the joint pmf of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ can be written, for $x_1, x_2, \ldots, x_n \in \{0, 1\}$, as

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^{n\overline{x}}(1-\theta)^{n-n\overline{x}}, \qquad \text{if } 0 < \theta < 1,$$

where the independence of the X_i is used for the first equality. Therefore, for $0 < \theta < 1$, if $y = n\overline{x}$,

$$f(\mathbf{x}|\theta)\pi(\theta) = \theta^{y}(1-\theta)^{n-y}. \tag{10.4}$$

One could compute the marginal distribution $m(\mathbf{x})$ as in Equation (10.3) in order to obtain the posterior p.d.f. $\pi(\theta|\mathbf{x})$ directly. However, one notices from Equation (10.1) that

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta), \tag{10.5}$$

and so another way to arrive at the solution would be to compare Equation (10.4) to the p.d.f.'s (p.m.f.'s) of known distributions, and see if the p.d.f. (p.m.f.) matches one of the known ones up to a normalising constant. Returning to the example, recall that for a random variable with values $\theta \in (0, 1)$, the beta distribution with parameters $\alpha > 0$, $\beta > 0$ has the p.d.f.

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{10.6}$$

One notices that $f(\mathbf{x}|\theta)\pi(\theta)$ in Equation (10.4) is then proportional to the p.d.f. of a beta distribution with $\alpha = y + 1$ and $\beta = n - y + 1$, and therefore

$$\pi(\theta|\mathbf{x}) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y}(1-\theta)^{n-y}.$$

Note that since the statistic $Y = \sum_{i=1}^{n} X_i$ is used to construct the posterior distribution, it will be used in any inference that is based on the posterior distribution. △

Remark 10.2.4. Note how we used Equation (10.5) to determine the posterior distribution without explicitly computing the marginal distribution $m(\mathbf{x})$. This is a common approach that is useful in a variety of cases, but when one does not recognise the right-hand side of Equation (10.5) as being a well-known p.d.f. (p.m.f.) up to a normalising constant, then the marginal distribution $m(\mathbf{x})$ needs to be computed. □

Remark 10.2.5. Observe in Example 10.2.3 that the likelihood (sampling distribution) was a binomial distribution, the prior was a uniform distribution, which led to the posterior being a beta distribution. The next example looks at the normal distribution. □

Example 10.2.6. Suppose that X follows a $\mathrm{N}(\theta, \tau^2)$ distribution where τ^2 is known and the mean θ is unknown. Then a likelihood function is given by

$$f(x|\theta) = \exp\left(-\frac{1}{2\tau^2}(x-\theta)^2\right).$$

Suppose that the prior distribution for θ is chosen to be a $\mathrm{N}(\mu, \sigma^2)$ distribution for some known μ and σ^2:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)$$

Note that the variance of the prior, σ^2, is not necessarily related to the variance of the sampling distribution, τ^2. One then computes the posterior distribution as being proportional to:

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \pi(\theta) f(\mathbf{x}|\theta) = \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)\exp\left(-\frac{1}{2\tau^2}(x-\theta)^2\right)\\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta - \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2\right) \qquad \text{(Exercise 10.2.7)}.\end{aligned}$$

Defining

$$\begin{aligned}M &= \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right) = \frac{\tau^2}{\sigma^2+\tau^2}\mu + \frac{\sigma^2}{\sigma^2+\tau^2}x\\ V^2 &= \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2},\end{aligned}$$

simplifies the equation to

$$\pi(\theta|\mathbf{x}) \propto \exp\left(-\frac{1}{2V^2}[\theta - M]^2\right),$$

which shows that the posterior distribution is $\mathrm{N}(M, V^2)$, without computing the marginal distribution. △

Exercise 10.2.7. Show that the posterior distribution in Example 10.2.6 is

$$\pi(\theta|\mathbf{x}) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta-\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2\right).$$

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \pi(\theta) f(\mathbf{x}|\theta) \\
&= \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)\exp\left(-\frac{1}{2\tau^2}(x-\theta)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\theta^2-2\theta\mu+\mu^2\right)-\frac{1}{2\tau^2}\left(x^2-2x\theta+\theta^2\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\theta^2-2\theta\mu\right)-\frac{1}{2\tau^2}\left(\theta^2-2x\theta\right)\right)\exp\left(-\frac{\mu^2}{2\sigma^2}-\frac{x^2}{2\tau^2}\right) \\
&= \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\theta^2-2\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\theta\right]\right)\exp\left(-\frac{\mu^2}{2\sigma^2}-\frac{x^2}{2\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta^2-2\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\theta\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta^2-\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2\right).
\end{aligned}$$

△

Remark 10.2.8. Example 10.2.6 illustrates an interesting result: when the prior distribution for the unknown mean and likelihood are both normal, then the posterior is also normal. □

In Example 10.2.6 the prior distribution for parameter θ was chosen to be a normal distribution with parameters μ and σ^2. Such parameters have a special name:

Definition 10.5. If Ψ is the family of distributions from which the prior distribution is chosen, and if the distributions in Ψ are parametrised by further parameters, then these associated parameters of the prior distribution are called prior **hyperparameters**.

Remark 10.2.9. Similarly, if the posterior distribution belongs to a family Φ that is parametrised by certain parameters, then these are called posterior **hyperparameters**. □

hyper- /ˈhʌɪpə/

Over; beyond; above
Origins: Greek, *huper*, meaning *over* or *beyond*
Source: Oxford English Dictionary

10.3 Conjugate prior distributions

Rather than starting with the definition of what a conjugate prior is, let us start with an example.

Example 10.3.1. Let us reconsider Example 10.2.3, where we observe a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a Bernoulli distribution with unknown parameter $\theta \in [0, 1]$ which we wish to estimate. Again defining $y = \sum_{i=1}^{n} x_i$, one has that the conditional p.m.f. of y is that of a $\text{Bin}(y, \theta)$ distribution

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Suppose for the prior one chooses a general beta distribution $\text{Beta}(\alpha, \beta)$, for known **hyper-parameters** $\alpha, \beta > 0$. This distribution has p.d.f. given by Equation (10.6). Then the joint distribution of $\mathbf{x}$ and θ is

$$\begin{aligned} f(\mathbf{x}, \theta) = f(\mathbf{x}|\theta)\pi(\theta) &= \left[\binom{n}{y} \theta^y (1-\theta)^{n-y}\right] \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}\right] \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1}. \end{aligned}$$

To compute the marginal distribution, recall that the p.d.f. of a $\text{Beta}(\gamma, \delta)$ distribution must integrate to 1:

$$\begin{aligned} &\int_0^1 \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \theta^{\gamma-1} (1-\theta)^{\delta-1} \, \mathrm{d}\theta = 1 \\ \Rightarrow &\int_0^1 \theta^{\gamma-1} (1-\theta)^{\delta-1} \, \mathrm{d}\theta = \frac{\Gamma(\gamma)\Gamma(\delta)}{\Gamma(\gamma+\delta)}. \end{aligned}$$

Therefore the marginal distribution $m(\mathbf{x})$ is

$$m(\mathbf{x}) = \int_0^1 f(\mathbf{x}, \theta) \, \mathrm{d}\theta = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

Finally, this gives the posterior distribution of θ given $\mathbf{x}$ (or y) as

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1}.$$

One recognises this as a $\text{Beta}(y+\alpha, n-y+\beta)$ distribution. △

Remark 10.3.2. This example illustrates an interesting phenomenon: for a Bernoulli or Binomial likelihood function, if one starts with a Beta prior, one will obtain a Beta posterior. This is an example of a **conjugate** family of distributions, which can be defined as follows: □

Definition 10.6. Let $\mathbf{X}$ be conditionally distributed given θ with p.m.f. or p.d.f. $f(\mathbf{x}|\theta)$ in the family of distributions $\mathcal{F}$. Let Ψ be the family of distributions from which the prior distribution chosen. If, for any prior distribution $\pi(\theta)$ chosen from Ψ and any set of observations $\mathbf{x} \subset \Omega$, the posterior distribution $\pi(\theta|\mathbf{x})$ is also in Ψ, then Ψ is a **conjugate family** of prior distributions for samples with distributions in $\mathcal{F}$.

Example 10.3.3. From Example 10.3.1, one sees that the beta distribution is conjugate to the Bernoulli and binomial distributions. In fact, the beta distribution is also conjugate to the negative binomial and geometric distributions, and is a suitable choice of prior when the unknown parameter θ is a percentage or proportion. △

Remark 10.3.4. In Example 10.2.3, the likelihood was Bernoulli (or Binomial), the prior was Unif$[0, 1]$ and the posterior was a beta distribution. This is in fact a special case of the conjugacy illustrated in Example 10.3.1, because the Unif$[0, 1]$ distribution is simply the Beta $(1, 1)$ distribution. □

Example 10.3.5. From Example 10.2.6, one sees that the normal distribution is conjugate to itself. △

In fact, most of the commonly used distributions have a conjugate family of prior distributions. It is also natural to wonder if it is always the case that any choice of prior distribution will lead to an easily identifiable posterior distribution.

Exercise 10.3.6. Show that the conjugate prior for the exponential distribution is the gamma distribution.

Suppose that a sample of random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is observed as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Suppose further that each X_i follows an exponential distribution with the same unknown parameter θ, i.e. each X_i has the p.d.f. for $x_i > 0$,

$$f(x_i|\theta) = \theta \exp(-\theta x_i).$$

Therefore the likelihood is, for $x_i > 0$ for all $i = (1, 2, \ldots, n)$, and writing $y = \sum_{i=1}^{n} x_i$,

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} (\theta \exp(-\theta x_i)) = \theta^n \exp(-\theta y).$$

Suppose the prior for θ is a $\Gamma(\alpha, \beta)$. Then (using the shape-rate parametrisation):

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta)$$

Then the posterior p.d.f. is proportional to:

$$\begin{aligned}\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta) &= \theta^n \exp(-\theta y) \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \\ &\propto \theta^{n+\alpha-1} \exp(-\theta(y+\beta))\end{aligned}$$

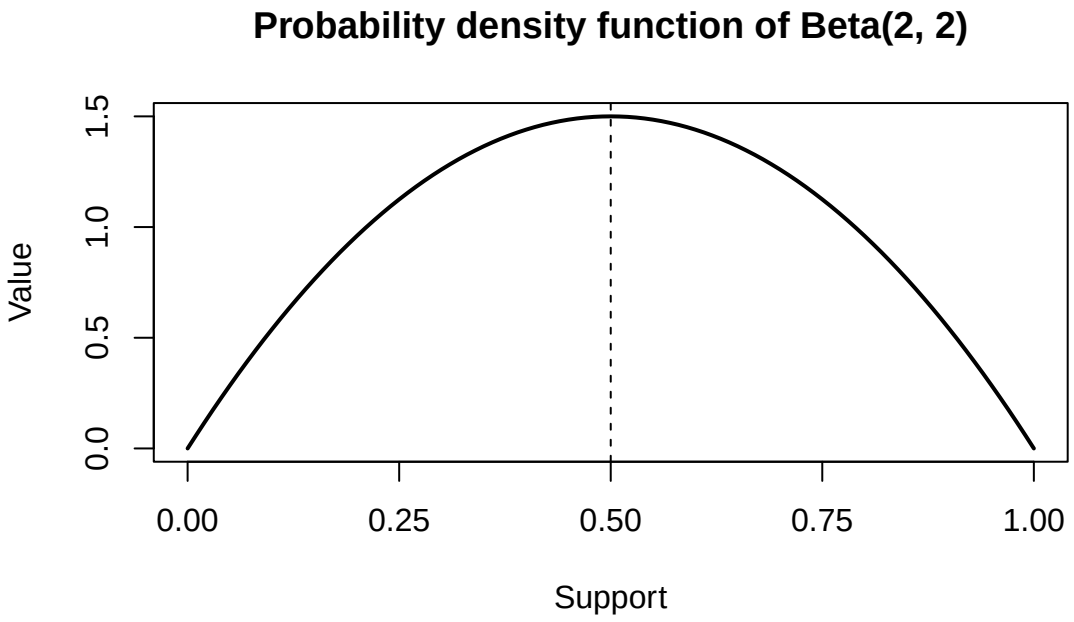
which shows that the posterior is a $\Gamma(n+\alpha, y+\beta)$ distribution, i.e.

$$\pi(\theta|\mathbf{x}) = \frac{(y+\beta)^{n+\alpha}}{\Gamma(n+\alpha)} \theta^{n+\alpha-1} \exp(-\theta(y+\beta)).$$

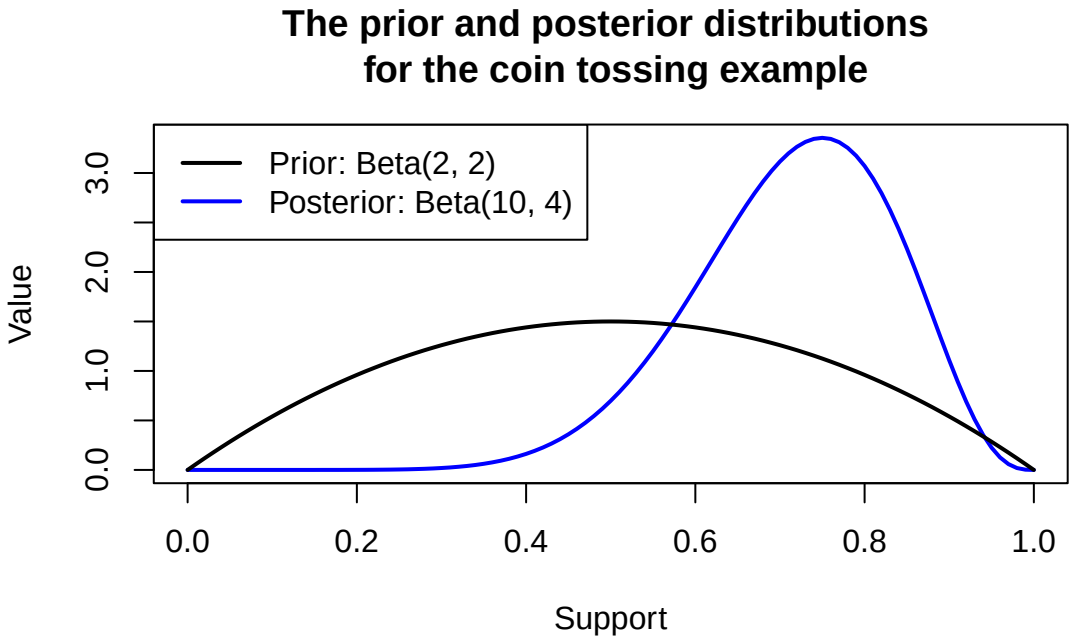
△

10.4 Return to the coin tossing example

Let us return to the coin tossing example described in the introduction. Suppose that we assume $\theta \sim \text{Beta}(2, 2)$, as shown in the figure below; i.e. this is our prior distribution $\pi(\theta)$.



Now suppose that we toss the coin ten times and the coin comes up heads exactly 8 times. Now, we can compute our posterior using Example 10.3.1; here we have $\alpha = \beta = 2$, $n = 10$ and $y = 8$, and so the posterior is a $\text{Beta}\,(y + \alpha, n - y + \beta) = \text{Beta}\,(10, 4)$ distribution:



10.5 Intractable posterior distributions

The examples we have seen in the previous section are all mathematically convenient: with an appropriate choice of prior for a certain likelihood, one obtains a posterior distribution belonging to a well-known family. However, there are many (even standard) examples where things are not so convenient.

If the posterior p.d.f. (p.m.f.) is difficult to integrate, i.e. has no known closed form solution, then one would need to resort to numerical methods of integration in order to obtain the cumulative distribution function (c.d.f.). Furthermore, it is possible that even numerical integration may be difficult. In either case, we call such posteriors **intractable**, and show two examples below.

Example 10.5.1. Let us return to Example 10.2.3 one last time. As before, suppose we observe a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a Bernoulli distribution with unknown parameter $\theta \in [0,1]$, and we wish to estimate θ. Defining $y = \sum_{i=1}^n x_i$, the likelihood is $f(\mathbf{x}|\theta) = \theta^y(1-\theta)^{n-y}$. Suppose that, rather than choosing the prior to be a beta distribution, one prefers the prior to be a truncated $\mathrm{N}(\mu, \sigma^2)$ distribution, i.e. a normal distribution restricted to the interval $[0,1]$. Then, the prior is

$$\pi(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right), \qquad \theta \in [0,1],$$

and the posterior is proportial to

$$\pi(\mathbf{x}|\theta) \propto f(\mathbf{x}|\theta)\pi(\theta) = \theta^y(1-\theta)^{n-y}\exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right).$$

No matter what the value of the normalising constant (from the marginal distribution) is, it does not seem as if the posterior $\pi(\mathbf{x}|\theta)$ belongs to any of the well-known families of distributions. △

Example 10.5.2. Suppose we have a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from a $\Gamma(\alpha, \beta)$ distribution with $\alpha = \theta$ unknown and $\beta = 1$ known. For x_i, the conditional distribution is:

$$f(x_i|\theta) = \frac{1}{\Gamma(\theta)} x_i^{\theta-1}\exp(-x_i), \qquad x_i > 0,$$

and therefore the likelihood function for the data $\mathbf{x}$ is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n \left(\frac{1}{\Gamma(\theta)} x_i^{\theta-1}\exp(-x_i)\right) = \frac{1}{(\Gamma(\theta))^n}\left(\prod_{i=1}^n x_i\right)^{\theta-1}\exp\left(-\sum_{i=1}^n x_i\right).$$

Since for any gamma distribution we must have $\theta > 0$, suppose we choose a $\Gamma(\gamma, \delta)$ prior for θ. Then the posterior $\pi(\theta|\mathbf{x})$ is proportional to

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta) = \frac{1}{(\Gamma(\theta))^n}\left(\prod_{i=1}^n x_i\right)^{\theta-1}\exp\left(-\sum_{i=1}^n x_i\right)\frac{\delta^\gamma}{\Gamma(\gamma)}\theta^{\gamma-1}\exp(-\delta\theta).$$

This certainly does not seem to be the p.d.f. (p.m.f.) for a standard distribution, nor does it seem to be easily integerable (analytically) in order to obtain a c.d.f.; of particular concern is the $(\Gamma(\theta))^{-n}$ term. Even numerical integration for this function seems challenging. $\triangle$

In fact, it is not unusual in Bayesian inference to end up with a situation Example 10.5.1 where the posterior is an unknown distribution and seemingly intractable. In later courses, one will learn numerical methods for sampling from such intractable distributions in order to be able to perform inference in the case where the posterior is an unknown distribution.

10.6 The effect of the prior on the posterior

It is natural to wonder how critical the choice of prior parameters is for the posterior. The next example investigates this issue.

Example 10.6.1. Suppose that the lifetime of a certain type of smartphone follows an exponential distribution with parameter θ. Suppose that a gamma distribution $\Gamma(\alpha, \beta)$ is selected as the prior for θ. Exercise 10.3.6 shows that posterior is a $\Gamma(n+\alpha, y+\beta)$ distribution with p.d.f. (using the shape-rate parametrisation):

$$\pi(\theta|\mathbf{x}) = \frac{(y+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}\theta^{n+\alpha-1}\exp\left(-\theta\left(y+\beta\right)\right), \tag{10.7}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the data and $y = \sum_{i=1}^{n} x_i$. Let's make the example more concrete. Suppose that we observe the following lifetimes for five smartphones, where the unit of measurement is hours:

$$\mathbf{x} = (2894, 3228, 3415, 3187, 3501) \qquad \Rightarrow \qquad n = 5 \text{ and } y = 16225.$$

Suppose that we choose as a prior a gamma distribution with parameters $\alpha_1 = 4, \beta_1 = 20000$; we shall call this Prior 1. Recall that for a random variable $X \sim \Gamma(\alpha, \beta)$, using the shape-rate parametrisation,

$$\mathrm{E}(X) = \frac{\alpha}{\beta}, \qquad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}.$$

Therefore, Prior 1 has mean $\alpha_1/\beta_1 = 0.0002$ and standard deviation $\sqrt{\alpha_1}/\beta_1 = 0.0001$. Suppose we also consider a second prior with parameters $\alpha_2 = 1, \beta_2 = 1000$. Then Prior 2 has mean $\alpha_2/\beta_2 = 0.001$ and standard deviation $\sqrt{\alpha_2}/\beta_2 = 0.001$; therefore, Prior 2 has mean five times as large as the mean of Prior 1, and Prior 2 has standard deviation ten times as large as that of Prior 1. So, although the two priors are from the same family of distributions, their means and standard deviations are very different. Using Equation (10.7), Posterior 1 (using Prior 1) is a $\Gamma(9, 36225)$ distribution, while Posterior 2 (using Prior 2) is a $\Gamma(6, 17225)$ distribution. The p.d.f.'s of the both priors and the resulting posteriors are plotted in Figure 10.1.

Note how the p.d.f.'s of the priors, which belong to the same family of distributions, are very different in shape, yet while the resulting posteriors are clearly not the same, their shapes are remarkably similar. Recall that this was only after five data points; as an exercise modify the `data` vector in the code (below) to contain more observations and see what happens. △

```
# create the data vector, and compute the priors and the posteriors
data <- c(2894, 3228, 3415, 3187, 3501)
prior_alpha <- c(4, 1)
prior_beta <- c(20000, 1000)
posterior_alpha <- prior_alpha + length(data)
posterior_beta <- prior_beta + sum(data)
```

```
# create the plots
x <- seq(from=0, to=0.0015, by=1e-6)
prior_1 <- dgamma(x, shape=prior_alpha[1], rate=prior_beta[1])
prior_2 <- dgamma(x, shape=prior_alpha[2], rate=prior_beta[2])
posterior_1 <- dgamma(x, shape=posterior_alpha[1], rate=posterior_beta[1])
posterior_2 <- dgamma(x, shape=posterior_alpha[2], rate=posterior_beta[2])

# plot the priors and posteriors in two subplots
lwd <- 1.5 # line widths
lty <- c("solid", "dashed") # line types
ylim <- c(0,5000) # y-axis limits
labs <- c("Support", "Value") # x- and y-axis labels

# make 1 row of 2 plots, cex controls size
par(mfrow=c(1,2), cex=0.75)
plot(x, prior_1, type='l', lty=lty[1], lwd=lwd,
     xlab=labs[1], ylab=labs[2], ylim=ylim)
lines(x, prior_2, type='l', lty=lty[2], lwd=lwd)
legend(0.0008, ylim[2], legend=paste0("Prior ", 1:2), lty=lty,
    lwd=rep(lwd, 2), cex=1)
plot(x, posterior_1, type='l', lty=lty[1], lwd=lwd,
    xlab=labs[1], ylab=labs[2], ylim=ylim)
lines(x, posterior_2, type='l', lty=lty[2], lwd=lwd)
legend(0.0006, ylim[2], legend=paste0("Posterior ", 1:2), lty=lty,
    lwd=rep(lwd, 2), cex=1)
```

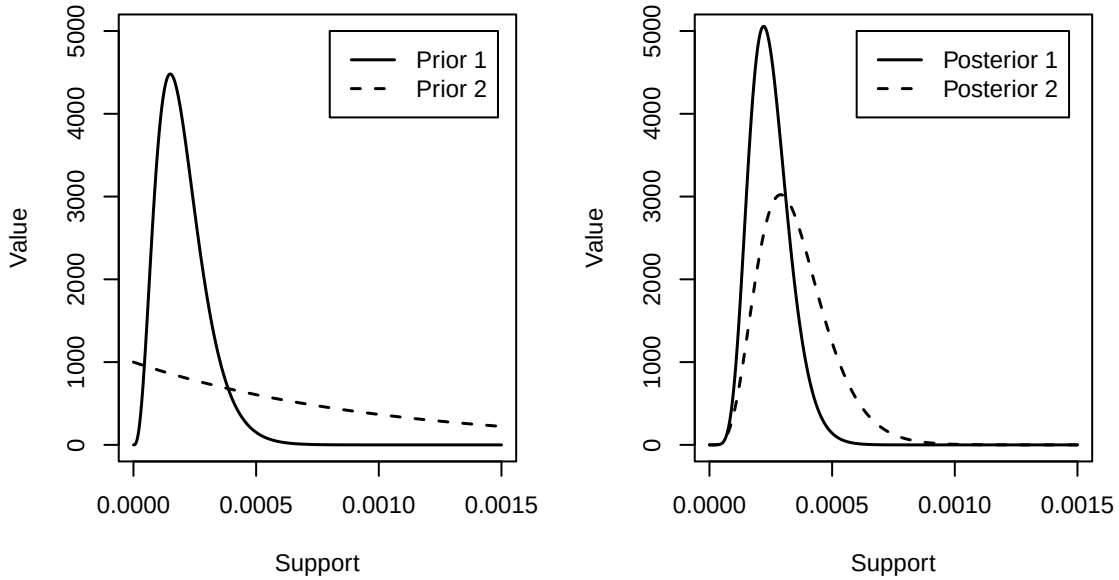


Figure 10.1: The two priors and the resulting posteriors from Example 10.6.1.

10.7 Choosing a prior distribution

In a Bayesian approach to a statistical analysis, one starts with a model for the data, which leads rise to a likelihood with unknown parameter(s). What is different from the frequentist approach is that these parameters are considered to be random variables. The statistician then chooses a prior distribution for each unknown parameter, and once data is observed, uses the likelihood and marginal distributions to obtain a posterior distribution, as Equation (10.1) in Theorem 10.2.2 shows. Leaving aside the choice of model for the likelihood, it is clear that the choice of prior—both the distributional family and any parameters for the prior p.d.f. (p.m.f.)—will have an effect on the posterior.

Which family of distributions should one select for a particular prior? Clearly, the answer depends on the actual problem, but it is important to at least choose a prior that has the same **support** as the unknown parameter being estimated (recall that the support of a function is the subset of the domain on which the function is not zero).

In Example 10.2.3, the parameter θ was a proportion, $\theta \in [0, 1]$. Therefore, one appropriate choice would be from the family of beta distributions, of which the uniform distribution is a special case ($\alpha = \beta = 1$), since all distributions in this family have support $[0, 1]$. For this example, it would not be a good idea to choose the normal or gamma distributions, since these have support on the whole of $\mathbb{R}$ or $\mathbb{R}^+$. However, Example 10.5.1 shows that it is possible to select a normal distribution as a prior, as long as one truncates the normal distribution to have support $[0, 1]$.

In Example 10.2.6, the unknown parameter θ was the mean of a normal distribution, which has support $\theta \in \mathbb{R}$. In this case, a normal distribution would be a suitable choice of prior for θ, since it also has support $\mathbb{R}$.

In Example 10.5.2, the unknown parameter θ for the exponential distribution has support $\theta > 0$. Therefore, a distribution from the family of gamma distributions is a good choice.

So, the first point to ensure when choosing a prior is that it has the same support as the unknown parameter being estimated. Furthermore, Section 10.3 shows that for certain likelihoods, choosing a prior from a conjugate family can lead to an easy computation for the posterior distribution: the posterior will be from the same family as the prior with parameters of the distribution updated from the data and the prior's parameters. Table 10.5 provides a list of the conjugate likelihoods that we have seen.

Table 10.5: A list of conjugate priors for a few well-known distributions.

Likelihood	Conjugate prior	Derivation
Bernoulli distribution	Beta distribution	Example 10.3.1
Binomial distribution	Beta distribution	Similar to Example 10.3.1
Normal distribution	Normal distribution	Example 10.2.6
Exponential distribution	Gamma distribution	Exercise 10.3.6

Suppose that one is starting a statistical analysis using the Bayesian approach for estimating a parameter θ. Suppose further that one has decided on a particular family of distributions for the prior, such as a Beta (α, β). How does one choose values for the hyperparameters α and β? On the one hand, if one has no strong belief or reliable information on the distribution of θ, then one could choose hyperparameter values that make the prior "flat", for example using $\alpha = \beta = 1$, which is then the uniform distribution, or $\alpha = 1, \beta = 1000$, which are the values for Prior 2 in Example 10.6.1 and is shown in Figure 10.1. On the other hand, if one has information from surveys or similar past experiments, these could inform the choice of hyperparameters; for example see Prior 1 in Example 10.6.1 and Figure 10.1.

Remark 10.7.1. It is beyond the scope of this course, but the Bernstein-von Mises theorem states that, under mild conditions and given enough data, for different choices of prior distributions $\pi_1(\theta)$ and $\pi_2(\theta)$ the posterior distributions $\pi_1(\theta|\mathbf{x})$ and $\pi_2(\theta|\mathbf{x})$ will be 'asymptotically the same'. Here $\pi_1(\theta)$ and $\pi_2(\theta)$ could be distributions from the same family (e.g. both gamma distributions) with different values for their hyperparameters, or they could be different distributions (e.g. a beta distribution and truncated normal distribution). □

Chapter 11

The Bootstrap

So far we have seen how to construct confidence intervals for estimators and compute p-values for hypothesis tests when the underlying distributions are known to be normal. However, there are times when we will be presented with a data set and cannot make assume the data follow a normal distribution.

This chapter discusses the nonparametric statistical method known as **the bootstrap** [5, 6] which allows, among other things, the computation of standard errors of estimators, the construction of confidence intervals and the computation of p-values for hypothesis tests when the underlying probability distribution of the data is unknown.

11.1 The empirical distribution

In earlier chapters, we have assumed that our data $x_1, x_2, \ldots, x_n$ are observations of the random variables $X_1, X_2, \ldots, X_n$, where these random variables are assumed to follow some known distribution F (e.g. normal), perhaps with unknown parameters which are to be estimated. However, there will be cases where the distribution is completely unknown.

The idea behind the bootstrap is to assume that we can use the **empirical distribution** $\widehat{F}$, which can be constructed from our data, in place of the unknown true distribution, F, in order to perform inference.

Definition 11.1. Given data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the empirical distribution $\widehat{F}$ on the (multi)set $\mathbf{x}$ is defined as the discrete distribution with probability mass function

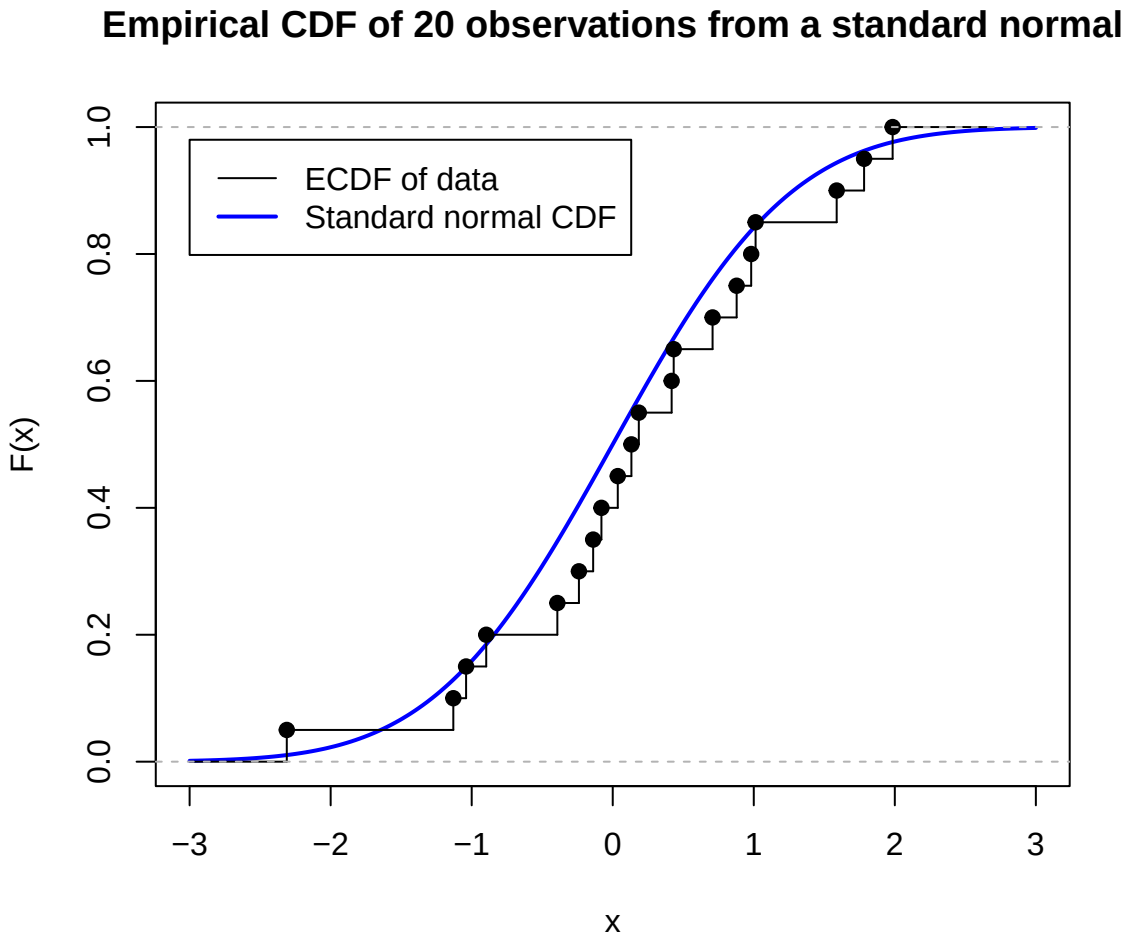
$$p_{\mathbf{X}}(x) = \begin{cases} \frac{1}{n}, & \text{if } x \in \mathbf{x}, \\ 0, & \text{otherwise.} \end{cases}$$

if all of the data in $\mathbf{x}$ are distinct. If the data are not distinct, then $p_{\mathbf{X}}(x) = \frac{1}{n} c_{\mathbf{x}}(x)$, where $c_{\mathbf{x}}(x)$ is the number of times that x occurs in the multiset $\mathbf{x}$, i.e. $c_{\mathbf{x}}(x) = \sum_{i=1}^{n} \mathbb{I}(x = x_i)$.

Remark 11.1.1. Since it is possible for $x_i = x_j$ for $i \neq j$, we refer to $\mathbf{x}$ as a 'multiset', rather than as a set, because multisets allow for values to occur more than once. □

In the figure below, we show an example of an empirical cumulative distribution function $\widehat{F}$, for data $\mathbf{x} = \{x_1, x_2, \ldots, x_{20}\}$ sampled from a standard normal distribution, superimposed on the cumulative distribution function of the standard normal.

```
# generate data following a standard normal
set.seed(2)
n <- 20
x <- rnorm(n)
#plot cdf of std normal between -3 and 3
z <- seq(from=-3, to=3, by=0.01)
main <- paste0("Empirical CDF of ",n," observations from a standard normal")
plot(x=z, y=pnorm(z), col="blue", type='l', lwd=2,
     xlab="x", ylab="F(x)", main=main)
# plot ecdf of data and add a legend
lines(ecdf(x), col="black", verticals=T)
legend(x=-3, y=0.98, legend=c("ECDF of data", "Standard normal CDF"),
    col=c("black", "blue"), lty=c(1, 1), lwd=c(1, 2))
```



11.2 Estimating the error: Aspirin data

The following dataset from [6] describes the result of a study that investigated if a small dose of aspirin could reduce the risk of hearts attacks in healthy middle-aged men. This study followed the 'gold standard' of clinical trials: a controlled, randomized, double-blind study, where half of the subjects received a dose of aspirin and the other half received a placebo (a pill with no therapeutic benefit). Not only were the subjects randomly assigned to receive either the aspirin or the placebo, but moreover, neither the subjects nor the physicians supervising the patients knew which treatment they were receiving, either the aspirin or placebo. The goal of the study is to estimate the **odds ratio** θ, where

$$\theta = \frac{\mathrm{P}\left(\text{Heart attack} \,|\, \text{taking aspirin}\right)}{\mathrm{P}\left(\text{Heart attack} \,|\, \text{not taking aspirin}\right)}.$$

If $\theta < 1$, then this means that taking aspirin reduces the risk of a heart attack occurring; but if $\theta > 1$, this means that taking aspirin increases the risk of a heart attack occurring. The results of the experiment are summarised in Table 11.1 below.

Treatment	Heart attacks fatal and non-fatal	Number of subjects
Aspirin	104	11037
Placebo	189	11034

Table 11.1: Results of aspirin study, showing the number of subjects receiving either aspirin or the placebo, and the number of heart attacks among subjects in each treatment group.

Now, an estimate $\widehat{\theta}$ of θ to two decimal places is

$$\widehat{\theta} = \frac{104/11037}{189/11034} = 0.55.$$

This seems to be great news (taking aspirin reduces the risk of heart attacks), except for one thing: we have no idea how accurate this point estimate is. In other words, we have no measure of the error of the measurement. To put it another way, we may have computed a point estimate based on a mean, but we have no measure of its variance.

We may think about using Chebyshev's inequality to obtain an interval estimate, but this will not work; we essentially have one observation, summarised by the table, while we would need several observations in order to compute a mean and standard deviation. In fact, we are really dealing with two unknown distributions, since there is one distribution F_a for subjects taking the aspirin, and another distribution F_p for subjects taking the placebo. So, it appears we are stuck.

11.2.1 Enter the Bootstrap (World)

Although we may only have one sample of data, we use it to create the empirical distributions $\widehat{F}_a$ and $\widehat{F}_p$. Let us first focus on $\widehat{F}_a$.

For $\widehat{F}_a$, we have 11037 Bernoulli trials (subjects), of which 104 were successful (heart attacks occurred). So, we create a sample of 104 1s and 10933 0s ($104 + 10933 = 11037$), and then sample 11037 values from this sample **with replacement**. This is now our first boostrap sample $\mathbf{x}_a^{*1}$.

For $\widehat{F}_p$, we have 11034 Bernoulli trials (subjects), of which 189 were successful (heart attacks occurred). So, we create a sample of 189 1s and 10845 0s ($189 + 10845 = 11034$), and then sample 11034 values from this sample **with replacement**. This is now our first boostrap sample $\mathbf{x}_p^{*1}$.

Then, we compute $\widehat{\theta}^{*1}$, our first bootstrap replication of $\widehat{\theta}$, as

$$\widehat{\theta}^{*1} = \frac{\text{Proportion of 1s in } \mathbf{x}_a^{*1}}{\text{Proportion of 1s in } \mathbf{x}_p^{*1}} = \frac{\text{Sample mean of } \mathbf{x}_a^{*1}}{\text{Sample mean of } \mathbf{x}_p^{*1}}.$$

We then repeat this process $B-1$ times to obtain bootstrap **replications** $\widehat{\theta}^{*1}, \widehat{\theta}^{*2}, \ldots, \widehat{\theta}^{*B}$. With these replications, we could compute the sample variance

$$\frac{1}{B-1}\sum_{i=1}^{B}\left(\widehat{\theta}^{*i} - \tfrac{1}{B}\sum_{j=1}^{B}\widehat{\theta}^{*j}\right)^2.$$

However, another approach to obtaining an interval estimate for $\widehat{\theta}$ would be to choose a value α, sort the replications, and then use

$$\left(\widehat{\theta}^*_{(\lfloor B[\alpha/2]\rfloor)}, \widehat{\theta}^*_{(\lfloor B[1-\alpha/2]\rfloor)}\right)$$

where $\widehat{\theta}^*_{(i)}$ is the ith order statistic, or equivalently, ith smallest replication in our set. If we run a simulation computing $B = 1000$ bootstrap replications and choose $\alpha = 0.05$, then we obtain

$$\left(\widehat{\theta}^*_{(25)}, \widehat{\theta}^*_{(975)}\right) = (0.43, 0.69).$$

This shows that the value of $\widehat{\theta}$ is (statistically significantly) lower than 1, and therefore the data shows that aspirin lowers the risk of having a heart attack.

See Section A.8 in the appendix for code that computed this interval.

Remark 11.2.1. There are other approaches to compute a confidence interval using the bootstrap [6], the one presented here is just the simplest method. □

11.2.2 Return to the aspirin data

The same study described above which measured how many patients given aspirin/the placebo had heart attacks also measured the number of strokes among the subjects. This data is summarised in Table 11.2.

Treatment	Strokes	Number of subjects
Aspirin	119	11037
Placebo	98	11034

Table 11.2: The number of strokes that occurred in the aspirin study.

For this data we compute

$$\widehat{\theta} = \frac{119/11037}{98/11034} = 1.21,$$

which appears to show that taking aspirin increases the risk of having a stroke. However, if we compute the 95% bootstrap confidence interval as before, we find that this interval is

$$(0.94, 1.57)\,,$$

which contains the value $\theta = 1$, which would indicate that there is no effect on the probability of having a stroke. Therefore, while aspirin significantly lowers the risk of heart attack, it does not significantly lower or increase the risk of a stroke.

See Section A.8 in the appendix for code that computed this interval.

11.3 The bootstrap procedure

Figure 11.1 below shows the general procedure for using the bootstrap to obtain a measure of the error of an estimate $\widehat{\theta} = s(\mathbf{x})$. Here, $s : \mathbb{R}^n \to \mathbb{R}$ is a statistic and such as the sample mean, sample variance or sample median. If we had paired data, this could be the correlation.

Given a dataset $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the empirical distribution $\widehat{F}$ is contructed. From this empirical distribution, B bootstrap samples of the data $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ are obtained by sampling **with replacement**. For example, if $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, then a bootstrap sample could be $\mathbf{x}^{*1} = (x_4, x_3, x_3, x_1, x_2)$. Then, we use these samples to compute the bootstrap replications $\widehat{\theta}^{*i} = s(\mathbf{x}^{*i})$, for $i = 1, 2, \ldots, B$. Now we essentially have B samples of our estimate, (at least in the 'bootstrap world') and can compute the standard error of $\widehat{\theta}$ or even an interval estimate, as we did for the aspirin data.

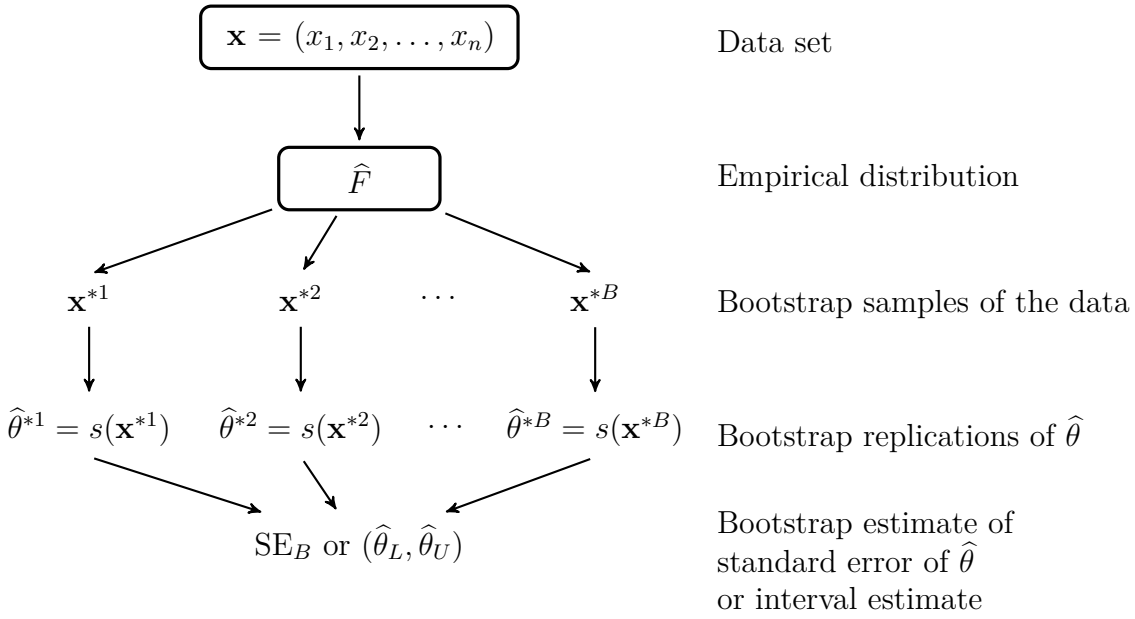


Figure 11.1: The process for obtaining bootstrap estimates, adapted from [6]. Here, s is some function on the data, e.g. the sample mean, sample variance or sample median. In the case of the aspirin data, this is simply the sample mean of the 1s and 0s.

11.4 Bootstrapping the median: the mouse data

The table below shows the results of a small experiment on a group of sixteen mice to test the efficacy of a new medical treatment. Seven of the mice were randomly selected to receive the treatment while the remaining nine mice were not given the treatment (control group). The goal of the treatment is to prolong survival time after a particular surgery, which all sixteen mice received. The survival time (in days) after the surgery was recorded for all of the mice.

Group	Data			Mean	Estimated standard error
Treatment	94	197	16		
(7 samples)	38	99	141	86.86	25.24
	23				
Control	52	104	146		
(9 samples)	10	50	31	56.22	14.14
	40	27	46		
Differences				30.63	28.93

Table 11.3: The mouse data set from [6], with the sample means and standard errors of the two groups.

To test if there is a significant difference in the survival times of these two groups, we may consider using Student's two-sample t-test[1] from Section 4.8.

If we label the sample mean and sample variance of the treatment group as $\overline{x}_t$ and s_t^2, respectively, and the sample mean and sample variance of the control group as $\overline{x}_c$ and s_c^2, then this table provides $\overline{x}_t$, $\overline{x}_c$, $s_t/\sqrt{n}$ and $s_c/\sqrt{m}$, where $n = 7$ and $m = 9$. The pooled sample variance can be computed as

$$\begin{aligned} s_p^2 &= \frac{1}{n+m-2}\left((n-1)s_t^2 + (m-1)s_c^2\right) \\ &= \frac{1}{n+m-2}\left((n-1)(n)\left(\frac{s_t}{\sqrt{n}}\right)^2 + (m-1)(m)\left(\frac{s_c}{\sqrt{m}}\right)^2\right) = 54.22 \\ \Rightarrow t &= \frac{\overline{x}_t - \overline{x}_c}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = 1.12. \end{aligned}$$

This t statistic value of 1.12 is not significant (see Table 3.3), so it would appear there is no statistically significant difference in the survival times between the two groups.

However, were we really justified in our use of Student's t-test? The data are positive integers, and are not necessarily normal. There are also perhaps too few data points to check this assumption with Q-Q plots.

[1] However, does this data really appear to be normally-distributed? Let's ignore this for the moment.

What if, instead of comparing means, we decided to compare medians? The median of the treatment group is 94 and the median of the control group is 46, so a difference in medians is 48. But, we would need some method of obtaining an estimate of error for this statistic (difference of medians): again, we can use the bootstrap.

Figure 11.2 shows us the empirical cumulative distribution functions for the treatment and control groups' survival times.

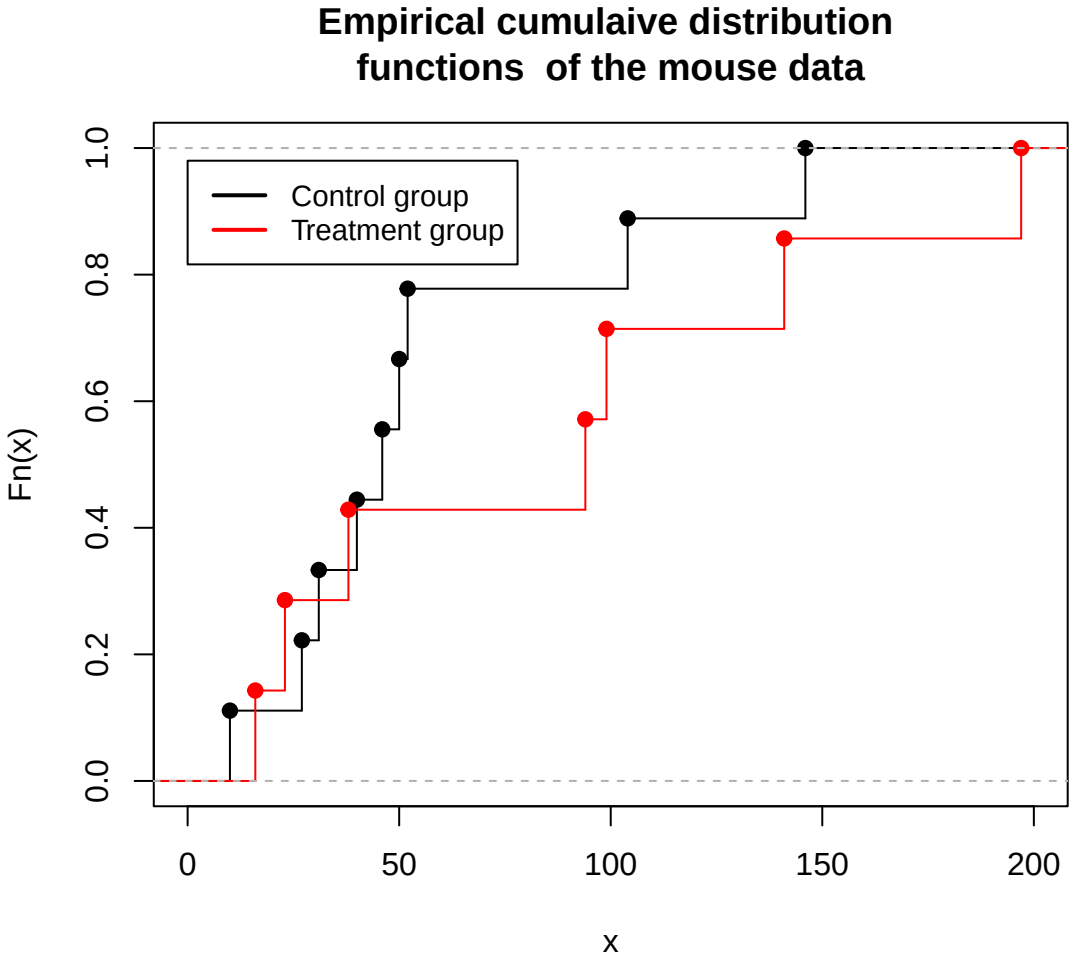


Figure 11.2: Plot of empirical cumulative distribution functions of the two groups of data in the mouse dataset.

We could create bootstrap samples from the treatment group $\mathbf{x}_t^{*1}, \mathbf{x}_t^{*2}, \ldots, \mathbf{x}_t^{*B}$ and then compute medians for each bootstrap sample, $m_t^{*1}, m_t^{*2}, \ldots, m_t^{*B}$. Similarly, we could create bootstrap samples for the control group $\mathbf{x}_c^{*1}, \mathbf{x}_c^{*2}, \ldots, \mathbf{x}_c^{*B}$, then compute medians for each bootstrap sample, $m_c^{*1}, m_c^{*2}, \ldots, m_c^{*B}$ Finally, we could compute the differences of these medians, $m_\Delta^{*i} = m_t^{*i} - m_c^{*i}$, for $i = 1, 2, \ldots, B$.

If we followed this procedure, we could compute a 95% confidence interval for this difference in medians using the bootstrap replications $m_\Delta^{*1}, m_\Delta^{*2}, \ldots, m_\Delta^{*B}$, as in the aspirin example above. If we did this (using the code in Appendix A.9), we would obtain the interval

$$(-29, 110).$$

This interval contains 0, which would mean that there is no increase in survival time, so we conclude that there is not sufficient evidence to state that the treatment is significantly beneficial. If the interval only consisted of positive numbers, then we could say that the treatment is beneficial.

Remark 11.4.1. The reason the difference in standard errors $28.93 = \sqrt{25.24^2 + 14.14^2}$ was included in the original table in [6] is that this value can be used instead of the (scaled) pooled standard deviation for a different t-test known as Welch's t-test, which we have not discussed. □

11.5 The bootstrap: outlook

This chapter is only a brief introduction to the bootstrap. There are many more applications, such as two-sample testing, where the bootstrap can be used. We also have not presented any theory for this technique: indeed, how do we know that any of these bootstrap intervals we have calculated are meaningful?

The good news is that there are theoretical results showing that, as the sample size of the data increases, then bootstrap estimates gets closer and closer to the true values they are estimating. The bad news is that there are not many results that provide statistical guarantees for finite samples. To think about a case where the bootstrap may not perform well, what if your sample of data from distribution F was (unfortunately) not representative and was sampled mostly from one of the tails of the distribution? We would not then expect the bootstrap to provide good estimates.

However, despite the lack of finite-sample guarantees, the bootstrap is a widely used and accepted method that has the major advantage of not requiring one to assume that the underlying distribution is known.

Remark 11.5.1. As stated in a book co-authored by the developer of the bootstrap [6], the origin of the name 'bootstrap' for this procedure is from the phrase 'to pull oneself up by one's bootstrap'. This phrase is thought to be from an 18th-century story by R. E. Raspe about his fictional character Baron Munchausen, who gets stuck in a lake/swamp but then pulls himself free by pulling upwards on his own bootstraps; of course, this clearly defies the laws of physics. (In previous eras, people often wore boots which has straps at the top of them in order to allow the wearer to pull to boots on.) □

Appendix A

Additional details

A.1 Expectation extended (Reading Material)

From the definition expectation it seems that the expected avlue of a random variable is either finite or it does not exist. We have seen an example of a random variable with a expectation that does not exist: when X follows a Cauchy distribution with probability density function

Remark A.1.1. Recall from your Analysis course that when dealing with improper integrals of the form

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = \int_{-\infty}^{a} g(x)\,\mathrm{d}x + \int_{a}^{\infty} g(x)\,\mathrm{d}x$$

that in order for the integral on the left hand-side of the equation to exist, then for any value a, either

- at least one of the two integrals on the right-hand side evaluates to a finite value, or
- both integrals on the right-hand side evaluate to $+\infty$, or both evaluate to $-\infty$ (i.e. same sign).

□

An alternative definition from [4] , which in essence is equivalent, is to first define for any nonnegative continuous random variable $X \geq 0$ that $\mathrm{E}[X] = \int_0^\infty xf(x)\,\mathrm{d}x$, where f is the probability density function. Note in this case that the expectation may have value ∞. Then, the definition is extended from the nonnegative case to the general case, where X can be negative, by defining $X^+ = \max\{X, 0\}$ to be the positive part of X and $X^- = \max\{-X, 0\}$ to be the negative part of \$X\$, noticing that $X = X^+ - X^-$, and then declaring that $\mathrm{E}[X]$ exists and has value $\mathrm{E}[X] = \mathrm{E}[X^+] - \mathrm{E}[X^-]$ whenever the subtraction makes sense, i.e. whenever either $\mathrm{E}[X^+] < \infty$ or $\mathrm{E}[X^-] < \infty$. This definition then allows $\mathrm{E}[X]$ to be finite, $+\infty$ or $-\infty$, and in the case we have both $\mathrm{E}[X^+] = \mathrm{E}[X^-] = \infty$, then the expectation is undefined (does not exist).

Although you have seen many examples where the expectation is finite, here are a few examples where the expectation is not finite.

Example A.1.2. Recall from Definition 8.3.8 in Prof. Veraart's notes that the Cauchy distribution has probability density function

$$f_X(x) = \frac{1}{\pi\,(1+x^2)}, \qquad \text{with support } x \in \mathbb{R}.$$

One can compute that the $\mathrm{E}[X]$ does not exist; see Problem Sheet 8, Question 1. △

Example A.1.3. Suppose $X \sim \mathrm{Exp}(1)$, the exponential distribution with parameter 1, which has probability density function (Definition 8.3.2 in Prof. Veraart's notes)

$$f_X(x) = \begin{cases} e^{-x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Now, define $Y = \frac{1}{X}$. One can show that $\mathrm{E}\,(X) = \infty$. Hint: first show that the probability density function of Y is

$$f_Y(y) = \begin{cases} \frac{1}{y^2} e^{-1/y} & \text{if } y \geq 0, \\ 0 & \text{if } y < 0. \end{cases}$$

△

Example A.1.4. Let X be the discrete random variable with probability mass function

$$p_X(x) = \mathrm{P}\,(X = x) = \begin{cases} 2^{-n}, & \text{for } x = 2^n, \text{where } n \in \{1, 2, \dots\}, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, X is equal to 2^n with probability 2^{-n}, for $n \in \mathbb{N}$, and is 0 otherwise. Since (a) $p_X(x) \geq 0$ and (b) $\sum_x p_X(x) = \sum_{n=1}^{\infty} p_X(2^n) = \sum_{n=1}^{\infty} 2^{-n} = 1$, this is a valid probability mass function. Now,

$$\mathrm{E}[X] = \sum_{x \in \mathrm{Im}(X)} x p_X(x) = \sum_{n=1}^{\infty} 2^n \cdot 2^{-n} = \sum_{n=1}^{\infty} 1 = \infty,$$

showing that X has infinite expected value. △

Remark A.1.5. Example A.1.4 forms the basis of the St Petersburg Paradox, where a gambler is invited by a casino to play the following game: toss a (fair) coin until it lands on tails. If it lands on tails on the nth toss (and is heads for tosses 1 to $n-1$), then the payout is 2^n units of currency. However, to play this game, the gambler needs to pay the casino an entry fee. What would be a fair value for the entry fee? Or ask yourself how much would you be willing to pay for the entry fee in order to play this game. □

A.2 Proof of Proposition 1.6.19 (Reading material)

We provide an elementary algebraic proof from [14]. Before proving the general case, we first prove the special case which we name

Lemma A.2.1. If $(x_1, x_2) \in \mathbb{R}^2$ and $x_1 \leq x_2$, then the function $g_{(x_1,x_2)} : \mathbb{R} \to \mathbb{R}$ defined by

$$g_{(x_1,x_2)}(z) = |x_1 - z| + |x_2 - z|$$

is minimised when $z \in [x_1, x_2]$. ♦

Proof. Start by defining $g_{(x_1,x_2)}(z) = |x_1 - z| + |x_2 - z|$.

In this special case, the value $a \in \mathbb{R}$ can be in one of three intervals:

$$\begin{aligned}
(1)\ a < x_1 (\leq x_2) \Rightarrow |x_1 - a| + |x_2 - a| &= (x_1 - a) + (x_2 - a) \\
&= x_1 + x_2 - 2a \\
&> x_1 + x_2 - 2x_1 \\
&= x_2 - x_1 \\
(2)\ a > x_2 (\geq x_1) \Rightarrow |x_1 - a| + |x_2 - a| &= -(x_1 - a) - [-(x_2 - a)] \\
&= 2a - x_1 - x_2 \\
&> 2x_2 - x_1 - x_2 \\
&= x_2 - x_1
\end{aligned}$$

$$\begin{aligned}
(3)\ a \in [x_1, x_2] \Rightarrow |x_1 - a| + |x_2 - a| &= -(x_1 - a) + (x_2 - a) \\
&= x_2 - x_1
\end{aligned}$$

This can be summarised as:

$$\begin{aligned}
z \notin [x_1, x_2] &\Rightarrow g_{(x_1,x_2)}(z) > x_1 - x_2, \\
z \in [x_1, x_2] &\Rightarrow g_{(x_1,x_2)}(z) = x_1 - x_2.
\end{aligned}$$

Therefore, $g_{(x_1,x_2)}(z)$ is minimised when $z \in [x_1, x_2]$.

□

We can now prove the general result:

Proposition 1.6.19. Given a sample of observations $x_1, x_2, \ldots, x_n$, with sample median m. Then, for any real value a,

$$\min_a \left(\sum_{i=1}^{n} |x_i - a| \right) = \sum_{i=1}^{n} |x_i - m| .$$

Proof.

For a fixed $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we define the function $g_{\mathbf{x}} : \mathbb{R} \to \mathbb{R}$ for any $z \in \mathbb{R}$ by

$$g_{\mathbf{x}}(z) = \sum_{i=1}^{n} |x_i - z| .$$

Our goal is to find the value $a \in \mathbb{R}$ such that

$$g_{\mathbf{x}}(a) = \min_z g_{\mathbf{x}}(z) , \tag{A.1}$$

and show that this point a is the median of the sample $\mathbf{x}$. There are now eight steps to completing the proof.

First, notice that we can rewrite $g_{\mathbf{x}}$ as

$$g_{\mathbf{x}}(z) = \sum_{i=1}^{n} |x_i - z| = \sum_{i=1}^{n} \left| x_{(i)} - z \right| ,$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the reordered $x_1, x_2, \ldots, x_n$ values such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ (and $\left\{ x_{(1)}, x_{(2)}, \ldots, x_{(n)} \right\} = \{x_1, x_2, \ldots, x_n\}$). We will rewrite this sum in a different way, but will need to consider whether n is even or odd.

Second, define c to be

$$c = \begin{cases} \frac{n}{2}, & \text{if } n \text{ is even,} \\ \frac{n+1}{2}, & \text{if } n \text{ is odd.} \end{cases} \tag{A.2}$$

Third, for $j \in \{1, 2, \ldots, c-1\}$, define the term

$$g_{\mathbf{x},j}(z) = \left| x_{(j)} - z \right| + \left| x_{(n+1-j)} - z \right| .$$

Using Lemma A.2.1, $g_{\mathbf{x},j}(z)$ is minimised when $z \in \left[x_{(j)}, x_{(n+1-j)} \right]$.

Fourth, we define $g_{\mathbf{x},c}(z)$:

$$\begin{aligned} n \text{ is even} : g_{\mathbf{x},c}(z) &= g_{\mathbf{x},\frac{n}{2}}(z) = \left| x_{(\frac{n}{2})} - z \right| + \left| x_{(\frac{n}{2}+1)} - z \right| \\ n \text{ is odd} : g_{\mathbf{x},c}(z) &= g_{\mathbf{x},\frac{n+1}{2}}(z) = \left| x_{(\frac{n+1}{2})} - z \right| \end{aligned}$$

Notice again that $g_{\mathbf{x},c}(z)$ is minimised when $z \in \left[x_{(c)}, x_{(n+1-c)}\right]$, and that when n is odd this interval $\left[x_{(c)}, x_{(n+1-c)}\right]$ is the point $x_{(\frac{n+1}{2})} = \left[x_{(\frac{n+1}{2})}, x_{(\frac{n+1}{2})}\right]$, which is the sample median. Notice how for even n the sample median also minimises $g_{\mathbf{x},c}(z)$ (use Lemma A.2.1!). The effort spent dealing with whether n is odd or even now allows us to write $g_{\mathbf{x}}(z)$ succinctly as

$$g_{\mathbf{x}}(z) = \sum_{j=1}^{c} g_{\mathbf{x},j}(z),$$

where c depends on n and is defined in Equation (A.2).

Fifth, note that the intervals $\left[x_{(j)}, x_{(n+1-j)}\right]$ are nested, i.e.

$$\left[x_{(1)}, x_{(n)}\right] \supset \left[x_{(2)}, x_{(n-1)}\right] \supset \cdots \supset \left[x_{(j)}, x_{(n+1-j)}\right] \supset \cdots \supset \left[x_{(c)}, x_{(n+1-c)}\right].$$

<!– and therefore,

$$\bigcap_{j=1}^{c} \left[x_{(j)}, x_{(n+1-j)}\right] = \left[x_{(c)}, x_{(n+1-c)}\right].$$

–>

Sixth, let $a \in \left[x_{(c)}, x_{(n+1-c)}\right]$ (i.e. a is in the innermost interval). Then $g_{\mathbf{x},j}(a)$ is the minimum of $g_{\mathbf{x},j}(z)$, for all $j \in \{1, 2, \ldots, c\}$.

Seventh, for any value $a' \in \mathbb{R}$, $g_{\mathbf{x}}(a') \geq \min_z g_{\mathbf{x}}(z)$. In particular, this is true for $a \in \left[x_{(c)}, x_{(n+1-c)}\right]$, so $g_{\mathbf{x}}(a) \geq \min_z g_{\mathbf{x}}(z)$.

Eighth, we also have

$$g_{\mathbf{x}}(a) = \sum_{j=1}^{c} g_{\mathbf{x},j}(a) = \sum_{j=1}^{c} \min_z g_{\mathbf{x},j}(z) \leq \min_z \sum_{j=1}^{c} g_{\mathbf{x},j}(z) = \min_z g_{\mathbf{x}}(z),$$

showing $g_{\mathbf{x}}(a) \leq \min_z g_{\mathbf{x}}(z)$, and therefore $g_{\mathbf{x}}(a) = \min_z g_{\mathbf{x}}(z)$, which proves the result.

The reason for the inequality

$$\sum_{j=1}^{c} \min_z g_{\mathbf{x},j}(z) \leq \min_z \sum_{j=1}^{c} g_{\mathbf{x},j}(z)$$

is that, in general, on the left-hand side there may be c different z values for minimizing the $g_{\mathbf{x},j}(z)$ functions, but on the right-hand side the same (minimising) z value is used for all $g_{\mathbf{x},j}(z)$ functions.

Note that when n is odd, we must have $a = x_{(\frac{n+1}{2})}$, which is the sample median. When n is even, any $a \in [x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$ minimises $g_{\mathbf{x}}(z)$; in particular the sample median $\frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]$ is in this interval and will minimise $g_{\mathbf{x}}(z)$.

□

This proof is essentially an algebraic exercise. Contrast this proof with the proof of Theorem 1.6.13 in the notes (the proof will be given in the Solution to Question 2 in Problem Sheet 9), where that proof was dealing with expectations of random variables.

A.3 Experiment for Corollary 3.1.3

Below is the R code for creating the histogram on page 65.

```
## Generate normal i.i.d. samples, compute sample menas
## and plot histogram of sample means

# set the seed for replicability; randomly chosen seed 1234
set.seed(1234)

# number of i.i.d. samples in one sample mean
n <- 50

# vector to store samples (not strictly necessary to create this)
x <- rep(0, n)

# number of sample means that will populate our histogram
num_samples <- 3000

# vector to store samples for histogram (necessary to initialise this)
xbar <- rep(0, num_samples)

# values for mu and sigma
mu <- 5
sigma <- 2

#generate samples and store in xbar
for (i in seq_len(num_samples)){
    # Note: seq_len is convenient way for creating vector c(1:n)

    # generate random draws
    x <- rnorm(n, mean=mu, sd=sigma)

    # save the mean in xbar
    xbar[i] <- mean(x)
}

# plot a histogram of the sampel means, xbar
num_breaks <- 30
main <- "Histogram of sample mean for i.i.d. normal samples"
xlab <- "Values of sample mean"
hist(xbar, breaks=num_breaks, main=main, xlab=xlab, freq=F)
```

A.4 A discussion on independence (Reading material)

The goal of this section is to prove the result:

Proposition A.4.1. Let X and Y be independent random variables, and let $g(x)$ be a function only of x, and let $h(y)$ be a function only of y. Then $U = g(X)$ and $V = h(Y)$ are independent random variables. ♦

This result is used in a step in the proof of Theorem 3.2.2, and it is a result one would intuitively expect to be true. However, it is still worth proving formally, and this will be done via a sequence of short results.

First, we recall the definition of independence for random variables from Prof. Veraart's notes (Definition 12.2.1). Suppose that X is a random variable with cumulative distribution function F_X and Y is a random variable with cumulative distribution function F_X. Let $F_{X,Y}$ denote the joint cumulative distribution function of X and Y. Then X and Y are independent random variables if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Although the proof of the Proposition A.4.1 applies to both discrete and continuous random variables, for simplicity we will provide a proof for the continuous case (the discrete case is similar). It will then be useful to have the alternative characterisation of independence provided in Section 12.4.1 of Prof. Veraart's notes: suppose that f_X and f_Y are the probability density functions of the random variables X and Y, respectively, and that $f_{X,Y}$ is the joint probability density function. Then X and Y are independent random variables if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y). \tag{A.3}$$

This result can actually be taken a bit further: [2] (Lemma 4.2.7):

Lemma A.4.2. Let (X, Y) be a bivariate random vector with joint probability density (or mass) function $f_{X,Y}(x, y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{X,Y}(x, y) = g(x)h(y).$$

♦

Proof. Try proving it yourself or see Lemma 4.2.7 in [2]. □

We now start with the following short result:

Proposition A.4.3. Let X and Y be two independent random variables, and let $g(x)$ be a function only of x, and let $h(y)$ be a function only of y. Then

$$\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\mathrm{E}[h(Y)]. \tag{A.4}$$

♦

Proof. This result, proved in [2] (Theorem 4.2.10), follows directly from the two-dimensional law of the unconscious statistician (Theorem 12.6.2 in Prof. Veraart's notes). Assuming X and Y are continuous as defined above,

$$\begin{aligned}
\mathrm{E}[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y && \text{(2D LOTUS)}\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y && \text{(independence and Eqn. (A.3))}\\
&= \int_{-\infty}^{\infty} h(y)f_Y(y)\left(\int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x\right)\mathrm{d}y\\
&= \left(\int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)\,\mathrm{d}y\right)\\
&= \mathrm{E}[g(X)]\mathrm{E}[h(Y)].
\end{aligned}$$

The same result can be proved in the discrete case by replacing integrals with summations. □

Proposition A.4.4. Let X and Y be independent random variables. For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$,

$$\mathrm{P}\left(X \in A, Y \in B\right) = \mathrm{P}\left(X \in A\right)\mathrm{P}\left(Y \in B\right),$$

in other words, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events. ♦

Proof. This result is also proved in [2] (Theorem 4.2.10). Let $\mathbb{I}_A$ be the indicator function for the set A, i.e.

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A,\\ 0, & \text{if } \omega \notin A. \end{cases}$$

and let $\mathbb{I}_B$ be the indicator function for the set B, defined similarly. Define the set

$$C = \left\{(x,y) \mid x \in A, y \in B\right\},$$

and let $\mathbb{I}_C$ be the indicator function for the set C. Note that $\mathbb{I}_C$ can be written in terms of $\mathbb{I}_A$ and $\mathbb{I}_B$,

$$\mathbb{I}_C(x,y) = \mathbb{I}_A(x)\mathbb{I}_B(y).$$

Recall also that the expected value of an indicator function is simply the probability of an event occurring, i.e. $\mathrm{E}[\mathbb{I}_A] = \mathrm{P}\left(A\right)$. Then, using Proposition A.4.3,

$$\begin{aligned}
\mathrm{P}\left(X \in A, Y \in B\right) &= \mathrm{P}\left((X,Y) \in C\right)\\
&= \mathrm{E}[\mathbb{I}_C]\\
&= \mathrm{E}[\mathbb{I}_A\mathbb{I}_B]\\
&= \mathrm{E}[\mathbb{I}_A]\mathrm{E}[\mathbb{I}_B]\\
&= \mathrm{P}\left(X \in A\right)\mathrm{P}\left(Y \in B\right).
\end{aligned}$$

□

We are now ready to prove Proposition A.4.1:

Proof of Proposition A.4.1. This proof is given in [2] (Theorem 4.3.5); for convenience we assume that $U = g(X)$ and $V = h(Y)$ are continuous random variables. Recall that we assume that X and Y are independent random variables. For any $u \in \mathbb{R}$ and $v \in \mathbb{R}$, we can define the sets

$$A_u = \{x \,|\, g(x) \leq u\}, \qquad B_v = \{y \,|\, h(y) \leq v\}.$$

Denoting the joint cumulative distribution function of (U, V) by $F_{U,V}$, by the definition of being a cumulative distribution function,

$$\begin{aligned} F_{U,V}(u, v) &= \mathrm{P}\left(U \leq u, V \leq v\right) \\ &= \mathrm{P}\left(X \in A_u, Y \in B_v\right) \qquad \text{(definition of } U \text{ and } V\text{)} \\ &= \mathrm{P}\left(X \in A_u\right)\mathrm{P}\left(Y \in B_v\right) \qquad \text{(by Proposition A.4.4).} \end{aligned}$$

Then the joint probability density function of (U, V) is

$$\begin{aligned} f_{U,V}(u, v) &= \frac{\partial^2}{\partial u \partial v} F_{U,V}(u, v) \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{P}\left(X \in A_u\right)\right)\left(\frac{\mathrm{d}}{\mathrm{d}v}\mathrm{P}\left(Y \in B_v\right)\right). \end{aligned} \tag{A.5}$$

Note that, in the last line, the first factor on the right-hand side is a function only of u and the second factor is a function only of v. Therefore, by Lemma A.4.2, the random variables U and V are independent.

□

A.5 Student's t-distribution

If $U \sim \mathrm{N}(0,1)$, and $V \sim \chi^2_p$, then

$$\frac{U}{\sqrt{V/p}} \sim t_p,$$

where t_p is called Student's t-distribution with p degrees of freedom and has probability density function

$$f(x) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi p}\Gamma\left(\frac{p}{2}\right)}\left(1+\frac{x^2}{p}\right)^{-\frac{p+1}{2}}$$

Remark A.5.1. While we have denoted the degrees of freedom above by p for clarity, it is convention to denote the degrees of freedom by the Greek letter ν. □

Remark A.5.2. The name of this distribution arises from the fact that the William Sealy Gosset published a paper on this distribution in 1908 in the journal Biometrika under the pseduonym 'Student'. He worked for the Guinness Brewery in Dublin at the time, and it has been suggested that Guinness asked him to publish under a pseduonym so that competitors would not know they were using the t-distribution to test the quality of their hops. □

A.5.1 Using the t-distribution

If the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed according to a $\mathrm{N}(\mu, \sigma^2)$ distribution with μ and σ^2 known, then defining

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

one can show

$$\mathrm{E}(Z) = 0, \qquad \mathrm{Var}(Z) = 1,$$

and since $\overline{X}$ is normally distributed and Z is a linear transformation of a $\overline{X}$, it follows that

$$Z \sim \mathrm{N}(0,1).$$

However, what is we know μ, but the variance σ^2 is unknown? Could we simply replace σ^2 with the sample variance s^2, and still consider the transformed random variable to be normally distributed?

While this would be a good approximation for large values of n, in fact the exact distribution of

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is not normal, but rather it is t-distributed with $n-1$ degrees of freedom.

Question 1 of Problem Sheet 10 shows that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{U}{\sqrt{V/(n-1)}},$$

where $U \sim \mathrm{N}(0,1)$, and $V \sim \chi^2_{n-1}$, which shows that $T \sim t_{n-1}$.

A.5.2 Review: Gamma distribution

We review that the Gamma distribution parametrised by shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, and denoted by $\mathrm{Gamma}(\alpha, \beta)$ or $\Gamma(\alpha, \beta)$; this was introduced in Prof. Veraart's notes, Definition 8.3.3. If $X \sim \mathrm{Gamma}(\alpha, \beta)$, then its probability density function f_X is defined by

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \qquad \text{with support } x > 0. \tag{A.6}$$

Note that since the support is $x > 0$, for $x \leq 0$, $f_X(x) = 0$. Also note that $\Gamma(\alpha)$ is the gamma function evaluated at α. Recall that the gamma function is defined for $z \in \mathbb{R}$ with $z > 0$ by

$$\Gamma(z) = \int_0^\infty x^z \exp(-z) \mathrm{d}x$$

Since f_X is a probability density function, we have

$$\begin{aligned} &\int_0^\infty f_X(x) \mathrm{d}x = 1 \\ \Rightarrow &\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathrm{d}x = 1 \\ \Rightarrow &\int_0^\infty x^{\alpha-1} \exp(-\beta x) \mathrm{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha} \end{aligned} \tag{A.7}$$

Equation (A.7) will be needed in the derivation of the t-distribution.

Remark A.5.3. Note that this parametrisation uses **shape** α and **scale** β. There is another parametrisation using shape $k = \alpha$ and **scale** $\theta = 1/\beta$, where the scale is the inverse of the rate. □

Remark A.5.4. The gamma function is also defined for complex values, but this will be covered in your course on omplex analysis. □

A.5.3 Derivation of Student's t-distribution (Reading Material)

Suppose that $U \sim \mathrm{N}(0,1)$ and $V \sim \chi^2_p$ for some $p > 0$, and suppose that U and V are independent (this is very important!). Define

$$T = \frac{U}{\sqrt{V/p}}.$$

The goal is to find the probability density function of T.

We first recall that the probability density functions f_U of U and f_V of V are

$$f_U(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$
$$f_V(v) = \frac{1}{2^{p/2}\Gamma(p/2)} v^{p/2-1} \exp\left(-\frac{v}{2}\right)$$

First, since U and V are independent, their joint probability density function factorises, i.e.

$$f_{U,V}(u,v) = f_U(u) f_V(v).$$

Second, since we wish to find the probability density function of T, let's define the transformations

$$T = \frac{U}{\sqrt{V/p}}, \qquad W = V.$$

We can rearrange these equations to

$$U = T\left(\frac{V}{p}\right)^{1/2}$$
$$\Rightarrow U = T\left(\frac{W}{p}\right)^{1/2}$$

and

$$V = W.$$

Then the joint distribution of T and W is

$$f_{T,W}(t,w) = f_{U,V}(u,v)\,|J| = f_U(u) f_V(v)\,|J|\,,$$

where J is the absolute value of the Jacobian. We can compute the Jacobian as

$$J = \det\begin{pmatrix} \frac{\partial u}{\partial t} & \frac{\partial v}{\partial t} \\ \frac{\partial u}{\partial w} & \frac{\partial v}{\partial w} \end{pmatrix} = \det\begin{pmatrix} \left(\frac{w}{p}\right)^{1/2} & 0 \\ -\frac{t}{2p}\left(\frac{w}{p}\right)^{-1/2} & 1 \end{pmatrix} = \left(\frac{w}{p}\right)^{1/2}$$

Then,

$$f_{T,W}(t,w) = f_U(u)f_V(v)\,|J| = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\left(\frac{w}{p}\right)\right)\frac{1}{2^{p/2}\Gamma(p/2)}v^{p/2-1}\exp\left(-\frac{w}{2}\right)\left(\frac{w}{p}\right)^{1/2}$$

$$= \frac{1}{(\pi p)^{1/2}2^{(p+1)/2}\Gamma(p/2)}w^{(\frac{p+1}{2})-1}\exp\left(-\frac{1}{2}\left(1+\frac{t^2}{p}\right)w\right)$$

If we now compute the marginal probability density functiopn of T by integrating over W, then (since W is χ^2_p distributed and so has support the positive real line)

$$f_T(t) = \int_0^\infty f_{T,W}(t,w)\mathrm{d}w$$

$$= \frac{1}{(\pi p)^{1/2}2^{(p+1)/2}\Gamma(p/2)}\int_0^\infty w^{(\frac{p+1}{2})-1}\exp\left(-\frac{1}{2}\left(1+\frac{t^2}{p}\right)w\right)\mathrm{d}w$$

and we notice that the integral on the right is in the same form as Equation (A.7), but with

$$x = w, \qquad \alpha = \frac{p+1}{2}, \qquad \beta = \frac{1}{2}\left(1+\frac{t^2}{p}\right).$$

Therefore,

$$f_T(t) = \frac{1}{(\pi p)^{1/2}2^{(p+1)/2}\Gamma(p/2)}\left(\Gamma\left(\tfrac{p+1}{2}\right)\left(\frac{1}{2}\left(1+\frac{t^2}{p}\right)\right)^{-\frac{p+1}{2}}\right),$$

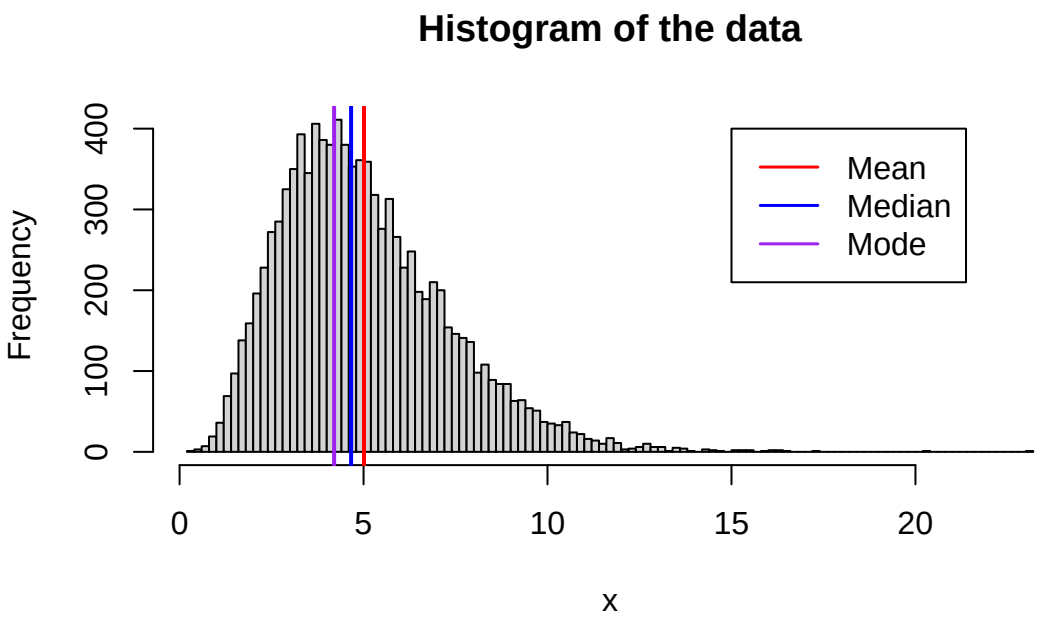
and the powers of 2 cancel out to give

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{(\pi p)^{1/2}\Gamma\left(\frac{p}{2}\right)}\left(1+\frac{t^2}{p}\right)^{-\frac{p+1}{2}}.$$

A.6 Visualising the mean, median and mode with R

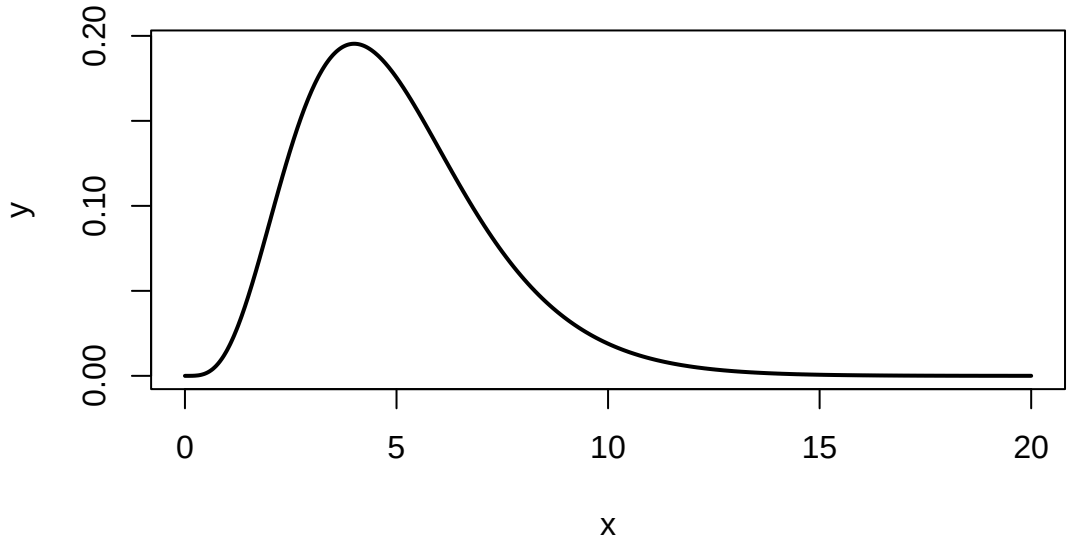
The following code shows a simulation to find the sample mean, median and mode for a $\Gamma(5, 1)$ distribution.

```
# set shape parameter (k) and scale parameter (theta)
# for Gamma distribution and setting the number of trials (n)
k <- 5
theta <- 1
n <- 10000

# generate the values, after setting the seed to ensure same sequence
# generated every time
set.seed(1)
x <- rgamma(n, shape=k, scale=theta)

# create and save the histogram to object h
h <- hist(x, breaks=100, main="Histogram of the data")

# add a verticals line showing the mean and median making them red/blue,
# and line width=2
abline(v=mean(x), col='red', lwd=2)
abline(v=median(x), col='blue', lwd=2)

# find the bin which has the largest count and find the value of the break
# with the largest count, i.e. the (empricial) mode
empirical_mode_index <- which.max(h$counts)
empirical_mode <- h$breaks[empirical_mode_index]

# add a vertical line showing the empirical mode, making it purple
abline(v=empirical_mode, col='purple', lwd=2)

# add legend
legend_x <- 15
legend_y <- 400
legend(legend_x, legend_y, legend=c("Mean", "Median", "Mode"),
   col=c("red", "blue", "purple"), lty=c(1, 1, 1), lwd=1.5, cex=1)

# true mode can be obtained in terms of parameters k and theta (exercise)
true_mode <- (k-1)*theta
cat("(empirical mode, true mode): ")
#> (empirical mode, true mode):
cat("(", empirical_mode, ", ", true_mode, ")\n", sep="")
#> (4.2, 4)
```



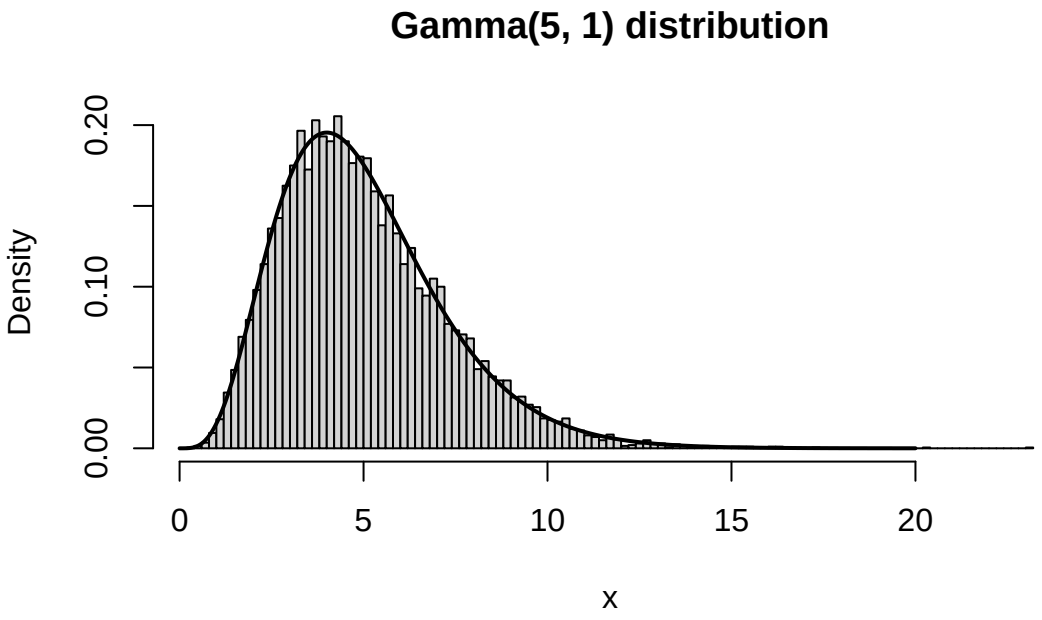
It is also possible to plot the probability density function using `dgamma`:

```
# set shape parameter (k) and scale parameter (theta) for Gamma dist.
k <- 5
theta <- 1

# set the range and generate evenly spaced points in this range
x <- seq(from=0, to=20, length.out=1000)

# Compute the values f(x), where f is the pdf of Gamma(k, theta)
y <- dgamma(x, shape=k, scale=theta)
plot(x, y, type='l', lwd=2)
```



We can redo the experiment, plotting both the histogram and overlaying the probability density function. However, in this case, we need to make the histogram display densities, rather than counts (densities are the normalised counts).

```
# set shape parameter (k) and scale parameter (theta) for Gamma dist.
k <- 5
theta <- 1

# set the range and generate evenly spaced points in this range
x <- seq(from=0, to=20, length.out=1000)

# Compute the values f(x), where f is the pdf of Gamma(k, theta)
y <- dgamma(x, shape=k, scale=theta)

# generate points again for histogram, and save in x_sample
set.seed(1)
x_sample <- rgamma(10000, shape=k, scale=theta)

# this time create histogram using frequencies, rather than counts
mainstring <- paste0("Gamma(", k, ", ", theta, ") distribution")
hist(x_sample, breaks=100, freq=F, main=mainstring, xlab="x")

# now use 'lines' to overlay the density
# (this adds a line to a plot, while 'plot' will start a new plot)
lines(x, y, type='l', lwd=2)
```



The observed sample fits the probability density function quite closely.

A.7 The shoe size data

In case you want the shoe size/height data, you can copy paste the following into a text file named `shoesize.txt`:

```
shoe.size,height,gender
6.5,66.0,F
9.0,68.0,F
8.5,64.5,F
8.5,65.0,F
10.5,70.0,M
7.0,64.0,F
9.5,70.0,F
9.0,71.0,F
13.0,72.0,M
7.5,64.0,F
10.5,74.5,M
8.5,67.0,F
12.0,71.0,M
10.5,71.0,M
13.0,77.0,M
11.5,72.0,M
8.5,59.0,F
5.0,62.0,F
10.0,72.0,M
6.5,66.0,F
7.5,64.0,F
8.5,67.0,M
10.5,73.0,M
8.5,69.0,F
10.5,72.0,M
11.0,70.0,M
9.0,69.0,M
13.0,70.0,M
```

and then read it into a data frame `df` in R, and compute the correlation, using:

```
df <- read.table("shoesize.txt", header=TRUE, sep=",")
shoesize <- df$shoe.size
height <- df$height
print( cor(shoesize, height) )
```

A.8 Bootstrap estimation with the aspirin data

This section contains the code for computing the bootstrap confidence interval for the aspirin data in Section 11.2.

It is possible to write a bootstrap script using a for loop, but here we will use a slightly different approach that is slightly more suitable for R. We start by writing a function that will compute a single bootstrap estimate $\widehat{\theta}^{*i}$.

```
# create a function for computing the bootstrap sample of theta
computeBootTheta <- function(){
    # the number of heart attacks in each group
    # a = aspirin group, p = placebo group
    x_a <- 104
    x_p <- 189
    # the total number of samples in each group
    n_a <- 11037
    n_p <- 11034

    # create the empirical distributions as vectors filled with 1s and 0s
    fhat_a <- c( rep(1, x_a), rep(0, n_a - x_a) )
    fhat_p <- c( rep(1, x_p), rep(0, n_p - x_p) )

    # compute a bootstrap sample from the aspirin and placebo groups;
    # sample is the function that randomly samples from a vector
    b_a <- sum( sample(x=fhat_a, size=n_a, replace=T) ) / n_a
    b_p <- sum( sample(x=fhat_p, size=n_p, replace=T) ) / n_p

    #compute theta_star, the ratio of the proportions, and return it
    theta_star <- b_a / b_p
    return(theta_star)
}
```

We then call this function in our 'main' programme below:

```
# the estimated value of theta
x_a <- 104
x_p <- 189
n_a <- 11037
n_p <- 11034
theta <- (x_a/n_a) / (x_p/n_p)

# compute bootstrap samples
set.seed(1)
numtrials <- 1000

# compute bootstrap samples using the `replicate` function and store the
# result in a vector.
# This runs a function a specified number of times. We could use a
# for loop here, but `replicate` (from the `apply` family of functions)
# is supposed to be faster.
theta_star_vec <- replicate(n=numtrials, computeBootTheta())

# set bootstrap limits
alpha <- 0.05
limits <- c(alpha/2, 1-alpha/2)

# obtains quantiles after sorting; the `type=1` will ensure the 25th and
# 975th values in the vector of 1000 are selected with no interpolation
# To see what I mean - try out the function yourself on `c(1:1000)`.
q <- quantile(theta_star_vec, probs=limits, type=1)

# print results
cat("theta: ",theta,", bootstrap interval: (", q[1],", ", q[2],")\n",sep="")
#> theta: 0.55011, bootstrap interval: (0.43446, 0.68844)
```

So, as the output shows, an estimatea to two decimal places is $\widehat{\theta} = 0.55$ and a 95% interval to two decimal places is

$$(0.43, 0.69)\,.$$

If we compute a confidence interval in the same way for the apsirin-stroke data in Table 11.2

```
# create a function for computing the bootstrap sample of theta
computeBootThetaStroke <- function(){
    # the number of heart attacks in each group
    # a = aspirin group, p = placebo group
    x_a <- 119
    x_p <- 98
    # the total number of samples in each group
    n_a <- 11037
    n_p <- 11034

    # create the empirical distributions as vectors filled with 1s and 0s
    fhat_a <- c( rep(1, x_a), rep(0, n_a - x_a) )
    fhat_p <- c( rep(1, x_p), rep(0, n_p - x_p) )

    # compute a bootstrap sample from the aspirin and placebo groups;
    # sample is the function that randomly samples from a vector
    b_a <- sum( sample(x=fhat_a, size=n_a, replace=T) ) / n_a
    b_p <- sum( sample(x=fhat_p, size=n_p, replace=T) ) / n_p

    #compute theta_star, the ratio of the proportions, and return it
    theta_star <- b_a / b_p
    return(theta_star)
}
```

We then call this function in our 'main' programme below:

```
# the estimated value of theta
x_a <- 119
x_p <- 98
n_a <- 11037
n_p <- 11034
theta <- (x_a/n_a) / (x_p/n_p)

# compute bootstrap samples
set.seed(1)
numtrials <- 1000

# compute bootstrap samples using the `replicate` function and store the
# result in a vector.
# This runs a function a specified number of times. We could use a
# for loop here, but `replicate` (from the `apply` family of functions)
# is supposed to be faster.
theta_star_vec <- replicate(n=numtrials, computeBootThetaStroke())

# set bootstrap limits
alpha <- 0.05
limits <- c(alpha/2, 1-alpha/2)

# obtains quantiles after sorting; the `type=1` will ensure the 25th and
# 975th values in the vector of 1000 are selected with no interpolation
# To see what I mean - try out the function yourself on `c(1:1000)`.
q <- quantile(theta_star_vec, probs=limits, type=1)

# print results
cat("theta: ",theta,", bootstrap interval: (", q[1],", ", q[2],")\n",sep="")
#> theta: 1.214, bootstrap interval: (0.94092, 1.5662)
```

So, as the output shows, an estimatea to two decimal places is $\widehat{\theta} = 1.21$ and a 95% interval to two decimal places is

$$(0.94, 1.57)\,.$$

A.9 Bootstrap estimation with the mouse data

Here is code for bootstrap estimation with the mouse data.

```
computeBootMouse <- function(){
x <- c(94, 197, 16, 38, 99, 141, 23)
y <- c(52, 104, 146, 10, 50, 31, 40, 27, 46)
nx <- length(x)
ny <- length(y)

# compute a bootstrap sample;
# sample is the function that randomly samples from a vector
xb <- sample(x=x, size=nx, replace=T)
yb <- sample(x=y, size=ny, replace=T)

#compute theta and return it
theta_star <- median(xb) - median(yb)
return(theta_star)
}
```

```
x <- c(94, 197, 16, 38, 99, 141, 23)
y <- c(52, 104, 146, 10, 50, 31, 40, 27, 46)
theta <- median(x) - median(y)

# compute bootstrap samples
set.seed(1)
numtrials <- 1000

theta_star_vec <- replicate(n=numtrials, computeBootMouse())
alpha <- 0.05
limits <- c(alpha/2, 1-alpha/2)
# obtains quantiles after sorting
q <- quantile(theta_star_vec, probs=limits, type=1)
cat("theta: ", theta, ", bootstrap interval: (",q[1],", ",q[2],")\n", sep="")
#> theta: 48, bootstrap interval: (-29, 110)
```

A.10 Computing the maximum likelihood

When computing the maximum of a likelihood function f, an argument that is commonly used is to check the points x where $f'(x) = 0$, or the boundary points of the domain of f. This section explains the reasoning behind this approach. In your Analysis module, you may have seen the following result, which is sometimes called **Fermat's Theorem**:

Proposition A.10.1. Let $f : [a, b] \to \mathbb{R}$ be a function. If f has a local minimum or a local maximum at some point $x \in (a, b)$, and if f is differentiable at x, then f'(x) = 0. ⧫

We now follow the argument provided in the book *Calculus* by Michael Spivak (see the first few pages of Chapter 11). A corollary to Proposition A.10.1 is

Corollary A.10.2. For a continuous function $f : [a, b] \to \mathbb{R}$, the maximum of f occurs at either:

1. one of the boundary points, a or b,
2. an interior point, i.e. $x \in (a, b)$.

⧫

This result can be refined a bit further

Corollary A.10.3. For a continuous function $f : [a, b] \to \mathbb{R}$, the maximum of f occurs at either:

1. one of the boundary points, a or b,
2. an interior point, i.e. $x \in (a, b)$, where the derivative $f'(x)$ does NOT exist,
3. an interior point, i.e. $x \in (a, b)$. where the derivative $f'(x)$ does exist.

⧫

It does not seem as if Corollary A.10.3 provides anything new, and it may seem as if we have made it result more complicated, splitting (2) into cases (2) and (3). However, we can now use Proposition A.10.1 to replace case (3), to finally obtain

Corollary A.10.4. For a continuous function $f : [a, b] \to \mathbb{R}$, the maximum of f occurs at either:

1. one of the boundary points, a or b,
2. an interior point, i.e. $x \in (a, b)$, where the derivative $f'(x)$ does NOT exist,
3. an interior point, i.e. $x \in (a, b)$. where $f'(x) = 0$.

⧫

When a function is differentiable everywhere, we are left with cases (1) and (3), and so just need to check points in these two cases. This concludes the reasoning behind the argument.

Bibliography

[1] Bertsekas, D. P. and Tsitsiklis, J. N. (2002). *Introduction to probability*, volume 1. Athena Scientific Belmont, MA.

[2] Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury, 2nd edition.

[3] DeGroot, M. H. and Schervish, M. J. (2012). *Probability and statistics*. Pearson Education.

[4] Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge University Press.

[5] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

[6] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[7] Evans, M. J. and Rosenthal, J. S. (2004). *Probability and statistics: The science of uncertainty*. W. H. Freeman and Company, New York.

[8] Fisher, R, A. (1935). *The Design of Experiments*. Oliver and Boyd, London and Edinburgh.

[9] Hand, D. J. (2020). Personal conversation.

[10] Kocik, J. (2001). Proof without words: Simpson's paradox. *Mathematics Magazine*, 74(5):399.

[11] Lunn, D. (2006). Lecture notes for A5, University of Oxford.

[12] Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30.

[13] Saw, J. G., Yang, M. C., and Mo, T. C. (1984). Chebyshev inequality with estimated mean and variance. *The American Statistician*, 38(2):130–132.

[14] Schwertman, N. C., Gilks, A. J., and Cameron, J. (1990). A simple noncalculus proof that the median minimizes the sum of the absolute deviations. *The American Statistician*, 44(1):38–39.

[15] Vigen, T. (2015). Spurious correlations, https://www.tylervigen.com/spurious-correlations.

[16] Weisberg, S. (1985). *Applied linear regression*. John Wiley & Sons, 2nd edition.

[17] Weiss, N. A. (2008). *Introductory Statistics*. Pearson-Addison-Wesley, 8th edition.