MATH40005 Probability and Statistics

Spring 2020 Version 0.1.0

Dean Bodenham

29 March 2020

Contents

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 and 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, 4].

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

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

where $\mathrm{Im}\left(X\right)$ is the set of values in $\mathbb{R}$ that X can take, i.e. $\mathrm{Im}\left(X\right) = \{X\left(\omega\right) \mid \omega \in \Omega\}$, the image of the sample space Ω under X. Recall that $\mathrm{E}\left(X\right)$ exists only if the right-hand side of Equation (1.1) converges absolutely, i.e. when $\sum_{x \in \mathrm{Im}(X)} |x| \, \mathrm{P}\left(X = x\right) < \infty$.

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

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

provided that $\int_{-\infty}^{\infty} |x| \, f_X\left(x\right) \mathrm{d}x < \infty$. Note that, regarding notation, both parentheses and brackets will be used for expectation, i.e. $\mathrm{E}\left(X\right) = \mathrm{E}[X]$.

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

1.1.3 The minimum expected squared deviation of a random variable

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 that the term $(\mathrm{E}[X]-a)$ is a constant, and so

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

due to the linearity of the expectation. 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. In particular, there is a useful generalisation of this result:

Theorem 1.1.2. Given two arbitrary random variables X and Y with a specified joint distribution, suppose that X and Y both have finite means. Then the function g of X that minimises $\mathrm{E}[(Y-g\,(X))^2]$ is $g(X) = \mathrm{E}[Y|X]$, i.e.

$$\min_g \mathrm{E}[(Y-g\,(X))^2] = \mathrm{E}[(Y-\mathrm{E}[Y|X])^2]$$

♦

Proof. The proof follows the argument above for Theorem 1.1.1, but using conditional expectations. **See the the solution to Question 2, Problem Sheet 9, Week 17.** □

Remark 1.1.3. We consider $\mathrm{E}[Y|X]$ to be a function of the random variable X (**not** Y), because when $X = x$ its value is $\mathrm{E}[Y|X = x]$, as defined in Definition 13.2.3 of Prof. Veraart's notes from Term 1. □

1.1.4 Review of conditional expectation

Conditional distribution and conditional expectation were defined in Chapter 13 of Prof. Veraart's notes in Term 1. We briefly review one of the main results, the Law of Total Expectation, and discuss notation.

Theorem 13.2.4 in Prof. Veraart's notes proved that for jointly continuous random variables X, Y,

$$\mathrm{E}\left(Y\right) = \int_{x: f_X(x)>0} \mathrm{E}\left(Y|X=x\right) f_X(x)\mathrm{d}x. \tag{1.4}$$

This is known as the Law of Total Expectation. Given Remark 1.1.3 above which describes how we can consider $\mathrm{E}[Y|X]$ to be a function of the random variable X, (and therefore is itself a random variable) we can rewrite this result as

Theorem 1.1.4. Given two jointly-distributed random variables X and Y,

$$\mathrm{E}[\mathrm{E}\left(Y|X\right)] = \mathrm{E}\left(Y\right). \tag{1.5}$$

⧫

Remark 1.1.5. In other words, the Law of Total Expectation computes the expected value of the random variable $\mathrm{E}\left(Y|X\right)$. Note that there is now no mention of whether X and Y are continuous or discrete; Equation (1.5) holds for both discrete and continous random variables. □

Exercise 1.1.6. Let X and Y be random variables and let $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be any two functions. Prove that $\mathrm{E}[g(X)h(Y)|X] = g(X)\mathrm{E}[h(Y)|X]$.

Assume that Y is continuous; the discrete case is similar. Suppose that X has the realised value x:

$$\begin{aligned}\mathrm{E}[g(X)h(Y)|X=x] &= \int_{-\infty}^{\infty} g(x)h(y)f_{Y|X}\left(y|x\right)\mathrm{d}y \\ &= g(x)\int_{-\infty}^{\infty} h(y)f_{Y|X}\left(y|x\right)\mathrm{d}y \\ &= g(x)\mathrm{E}[h(Y)|X=x].\end{aligned}$$

This shows that the realised values $\mathrm{E}[g(X)h(Y)|X=x]$ and $g(x)\mathrm{E}[h(Y)|X=x]$ of the random variables $\mathrm{E}[g(X)h(Y)|X] = g(X)\mathrm{E}[h(Y)|X]$ are always equal; therefore the two random variables are equal. △

Remark 1.1.7. Looking more closely at Equation (1.5), one notices that while there are three expectation operators, each is with respect to a different distribution. Suppose that X and Y are jointly continuous; then the inner expectation $\mathrm{E}\left(Y|X\right)$ on the left-hand side is using the p.d.f. $f_{Y|X}$ of Y given X, while the outer expectation is using the p.d.f. f_X of X, and the expectation on the right-hand side is using the p.d.f. f_Y of Y (although it could also be using the joint p.d.f. $f_{X,Y}$ of X and Y, the result would be the same). An alternative notation is to indicate the distribution being used as a subscript for the expectation operator, e.g.

$$\mathrm{E}_{Y|X}\left(Y\right) = \mathrm{E}\left(Y|X\right)$$

Note that it is then not necessary to indicate the conditioning inside the parentheses for the expectation. □

Using this alternative notation, the Law Of Total Expectation becomes

$$\mathrm{E}_X[\mathrm{E}_{Y|X}\left(Y\right)] = \mathrm{E}_Y\left(Y\right). \tag{1.6}$$

Note that, using this notation, $\mathrm{E}_{X,Y}\left(Y\right) = \mathrm{E}_Y\left(Y\right)$. This can easily be shown for the continuous case by

$$\mathrm{E}_{X,Y}\left(Y\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y = \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y = \mathrm{E}_Y\left(Y\right).$$

One can also prove a generalisation of the Law of Total Expectation:

Proposition 1.1.8. Given two random variables X and Y with some known joint distribution, consider the function h of X and Y that has the value $h(x, y)$ for the realised values $X = x$ and $Y = y$. Then

$$\mathrm{E}_X[\mathrm{E}_{Y|X}[h(X,Y)]] = \mathrm{E}_{X,Y}[h(X,Y)]. \tag{1.7}$$

♦

Proof. Suppose that the random variables are both continuous with joint distribution $f_{X,Y}$. The discrete case is similar. Further denote the p.d.f. of marginal distribution of X as f_X, and the p.d.f. of the conditional distribution of Y given X as $f_{Y|X}$. Then

$$\begin{aligned}\mathrm{E}_X[\mathrm{E}_{Y|X}[h(X,Y)]] &= \int_{-\infty}^{\infty} \Big(\mathrm{E}_{Y|X}[h(x,Y)]\Big) f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(x,y) f_{Y|X}(y|x)\,\mathrm{d}y\right) f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{Y|X}(y|x) f_X(x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \mathrm{E}_{X,Y}[h(X,Y)]\end{aligned}$$

where the last line used $f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x, y)$ (Definition 13.2.1 in Prof. Veraart's notes). □

Clearly, when $h(X, Y) = Y$, Proposition 1.1.8 reduces to the Law of Total Expectation, Equation (1.6).

Summary

For random variables X and Y,

- $\mathrm{E}(X)$ is a number.
- $\mathrm{E}(Y|X = x)$ is a number whose value depends on x.
- $\mathrm{E}(Y|X)$ is a function of the random variable X, and so is a random variable itself. Its value is $\mathrm{E}[Y|X = x]$ when the realised value of X is x. Alternative notation is $\mathrm{E}_{Y|X}(Y)$.
- The Law of Total Expectation states that $\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$. See also Equation (1.6).

1.1.5 Review of variance

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

Definition 1.1.9. The **variance** of a random variable X is defined as

$$\operatorname{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]. \tag{1.8}$$

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

Remark 1.1.10. 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 1.1.11. 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 $\operatorname{Var}(X)$ 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 $\mu = 172$cm, and $\sigma = 3$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 1.1.12. For any random variable X, show that $\operatorname{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.

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

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

The following result is useful for bounded random variables:

Proposition 1.1.13. Suppose that the random variable X is known to only take values in the bounded range $[a, b]$. Then $\operatorname{Var}(X) \leq \dfrac{(b-a)^2}{4}$. ♦

Proof. See the solution to Question 2, Problem Sheet 8, Week 16. □

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

Corollary 1.1.14. Suppose $X \sim \mathrm{Bern}(p)$, for some $p \in [0, 1]$. Then $\operatorname{Var}(X) = p(1-p) \leq \frac{1}{4}$. ♦

Proof.

If $X \sim \mathrm{Bern}(p)$, then $X \in \{0, 1\} \subset [0, 1]$, and the result follows from Propostion 1.1.13. Alternatively, one can show the function $g(p) = p(1-p) \leq \frac{1}{4}$ on $[0, 1]$ using differentiation. Or, notice that 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 1.1.15. 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

$$\mu'_k = \mathrm{E}[X^k]. \tag{1.9}$$

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

$$\mu_k = \mathrm{E}[(X-\mu)^k], \tag{1.10}$$

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,

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

Remark 1.1.16. 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 1.1.17. 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 equality in the first line used the linearity of the expectation.

Alternatively, one could have used Exercise 1.1.12:

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

as in the first calculation. △

Remark 1.1.18. 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).

□

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,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.11}$$

■

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

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \tag{1.12}$$

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 $\overline{X}\left(X_1, \ldots, X_n\right)$ and $S^2\left(X_1, \ldots, X_n\right)$. 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

$$(n-1)\,S^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2. \tag{1.13}$$

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

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,

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

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 (1.13), 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 1.1.17 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,

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

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 collection of observations

For a collection of observations $x_1, x_2, \ldots, x_n$, we define can the sample variance s^2 as follows
Definition 1.2.9. For real values $x_1, x_2, \ldots, x_n$, the sample variance s^2 is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2, \qquad \text{where} \qquad \overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j. \tag{1.14}$$

■

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 (1.14), prove that

$$\min_a \left[\sum_{i=1}^{n}\left(x_i - a\right)^2\right] = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = (n-1)\, s^2. \tag{1.15}$$

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

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

where in the third line we used

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\overline{x} = n\overline{x} - n\overline{x} = 0.$$

Since $n\left(\overline{x} - a\right)^2 \geq 0$,

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

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**. 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

$$\mathrm{P}\left(X \geq a\right) \leq \frac{\mathrm{E}\left(X\right)}{a}, \qquad \text{for all } a > 0. \tag{1.16}$$

♦

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:

$$\mathrm{E}\left(Y_a\right) \leq \mathrm{E}\left(X\right). \tag{1.17}$$

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

$$\mathrm{E}\left(Y_a\right) = 0 \cdot \mathrm{P}\left(X < a\right) + a \cdot \mathrm{P}\left(X \geq a\right). \tag{1.18}$$

Combining Equations (1.17) and (1.18), one obtains

$$a \cdot \mathrm{P}\left(X \geq a\right) \leq \mathrm{E}\left(X\right), \tag{1.19}$$

from which the inequality in Equation (1.16) follows. □

Theorem 1.3.2 (Chebyshev's Inequality)**.** If X is a random variable with mean μ and variance σ^2, then for all $c > 0$,

$$\mathrm{P}\left(|X - \mu| \geq c\right) \leq \frac{\sigma^2}{c^2}. \tag{1.20}$$

♦

Proof.

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

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

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

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

which proves the result.

□

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

Corollary 1.3.3. If X is a random variable with mean μ and variance σ^2, then for all $k > 0$,

$$\mathrm{P}\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}. \tag{1.21}$$

♦

Proof. In Equation (1.20) 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 1.3.4. 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 $X_i \sim \mathrm{Bern}(p)$, where p is the unknown parameter we wish to estimate, and $\widehat{p} = \overline{x}$. Since each X_i has mean $\mathrm{E}\left(X_i\right) = p$ and variance $\mathrm{Var}\left(X_i\right) = p(1-p)$, using Proposition 1.2.6, $\mathrm{E}\left(\overline{X}\right) = p$ and $\mathrm{Var}\left(\overline{X}\right) = p(1-p)/n$. Therefore, for any $\epsilon > 0$, Chebyshev's Inequality in Theorem 1.3.2 gives

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

Furthermore, using Corollary 1.1.14, one can remove the unknown p on the right-hand side to obtain

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

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

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

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 (1.21):

Theorem 1.3.5 (from [8])**.** Suppose $X_1, X_2, \ldots, X_n$ and X_{n+1} are i.i.d random variables, and define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \qquad Q^2 = \left(\frac{n+1}{n}\right)S^2. \tag{1.22}$$

Then for all $\lambda \geq 1$,

$$\mathrm{P}\left(\left|X_{n+1} - \overline{X}\right| > \lambda Q\right) \leq \frac{1}{\lambda^2} + \frac{1}{n}. \tag{1.23}$$

♦

Remark 1.3.6. The proof of Theorem 1.3.5 is beyond the scope of this course, but can be found in [8], which proves a more general version of Equation (1.23). □

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. distributed 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

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

To give a specific example, suppose that $n = 10$ and $x_1, x_2, \ldots, x_{10}$ are observed to be

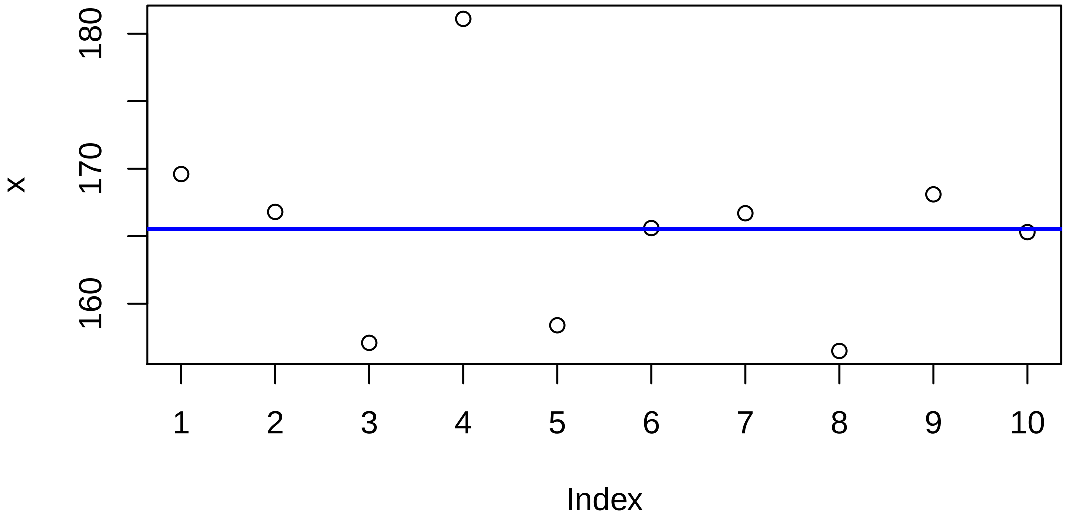
$$\{169.6, 166.8, 157.1, 181.1, 158.4, 165.6, 166.7, 156.5, 168.1, 165.3\}.$$

Then

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

Note that while this $\widehat{\theta}$ 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. We plot the data and $\widehat{\theta}$ below.

```r
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)
```



Now, although the mean θ of the distribution F_X is unknown, suppose that the variance of F_X is known to be $\sigma^2 = 27.04$. Note that the version of Chebyshev's inequality given in Equation (1.21) can be rewritten as

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

If we take our random variable to be $\overline{X}$, which has $\mathrm{E}(\overline{X}) = \theta$ and $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$ (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. (One can compute $157.30 \approx 165.52 - 5 \cdot \sqrt{27.04}/\sqrt{10}$ and $173.74 \approx 165.52 + 5 \cdot \sqrt{27.04}/\sqrt{10}$.)

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. Another parameter is the variance, $\sigma^2 = \mathrm{Var}\left(X\right)$. ■

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

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 $T = r\left(X_1, X_2, \ldots, X_n\right) = r(\mathbf{X})$ 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 the maximum of these random variables, $Y = \max\left\{X_1, \ldots, X_n\right\}$. A third example is the function r on n variables which has the constant value $r\left(X_1, \ldots, X_n\right) = 7$ 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 $r\left(X_1, \ldots, X_n\right)$ 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 $\widehat{\theta} = r\left(x_1, \ldots, x_n\right)$ 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 $\widehat{\Theta}\left(X_1, X_2, \ldots, X_n\right)$. ■

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 $\widehat{\Theta} = \widehat{\Theta}\left(X_1, X_2, \ldots, X_n\right) = \widehat{\Theta}\left(\mathbf{X}\right)$. □

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

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 $x_1, x_2, \ldots, x_n$. □

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 1.5.10. 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 1.5.11. Although Definition 1.5.10 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 1.5.12. 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 1.5.13. 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 1.5.14. 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 1.5.15. 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 $\alpha = 0.05$). □

Remark 1.5.16. In Section 1.4 we used Chebyshev's inequality to create a 96% confidence interval for the unknown mean θ. □

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 $\$5,000$!

1.5.4 Estimators, bias and variance

This section defines a few important concepts regarding estimators.

Definition 1.5.17. The **estimation error** of the estimator $\widehat{\Theta}$ of a parameter θ is defined to be $\widehat{\Theta} - \theta$. ■

Definition 1.5.18. 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:

$$b_\theta(\widehat{\Theta}) = \mathrm{E}[\widehat{\Theta}] - \theta \tag{1.24}$$

■

Remark 1.5.19. Since the parameter θ is assumed to be a constant (but unknown) value, Equation (1.24) follows from $\mathrm{E}[\widehat{\Theta} - \theta] = \mathrm{E}[\widehat{\Theta}] - \theta$. □

Definition 1.5.20. The estimator $\widehat{\Theta}$ of a parameter θ is called **unbiased** if $\mathrm{E}[\widehat{\Theta}] = \theta$, for every possible value of θ. ■

Example 1.5.21. 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 1.5.22. 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 these definitions, we have the following important result:

Theorem 1.5.23. The mean squared error of an estimator $\widehat{\Theta}$ of a parameter θ can be expressed in terms of its bias and variance:

$$\mathrm{E}[(\widehat{\Theta} - \theta)^2] = \left[b_\theta(\widehat{\Theta})\right]^2 + \mathrm{Var}\left(\widehat{\Theta}\right). \tag{1.25}$$

♦

Proof.

For any random variable X, Exercise 1.1.12 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.

□

1.6 The sample mean of normal random variables

In this section we look at the special case that the random variables $X_1, X_2, \ldots, X_n$ follow a normal distribution. We first prove the following result:

Proposition 1.6.1. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, and that for $i \in \{1, 2, \ldots, n\}$, $X_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, where each μ_i and each σ_i is finite. Then, defining $Y = \sum_{i=1}^n X_i$,

$$Y \sim \mathrm{N}\left(\mu, \sigma^2\right), \qquad \text{where } \mu = \sum_{i=1}^n \mu_i \text{ and } \sigma^2 = \sum_{i=1}^n \sigma_i^2.$$

♦

Proof.

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

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

Each $X_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, therefore $M_{X_i}(t) = \exp\left(\mu_i t + \sigma_i^2 t^2/2\right)$, 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 $Y \sim \mathrm{N}\left(\mu, \sigma^2\right)$, as required.

□

Corollary 1.6.2. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables distributed according to $\mathrm{N}\left(\mu, \sigma^2\right)$. 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 $i \in \{1, 2, \ldots, n\}$, $X_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$, and so $\frac{1}{n} X_i \sim \mathrm{N}\left(\frac{\mu}{n}, \frac{\sigma^2}{n^2}\right)$. Then, using Prop. 1.6.1,

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

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

□

1.7 The sample variance of normal random variables

Another useful result concerns 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 1.7.1. 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,

$$\sum_{i=1}^n Y_i^2 = \mathbf{Y}^T\mathbf{Y} = \mathbf{Z}^T\mathbf{A}^T\mathbf{A}\mathbf{Z} = \mathbf{Z}^T\mathbf{I}_n\mathbf{Z} = \mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^n Z_i^2. \tag{1.26}$$

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 $\mathbf{Y}$ is given by (using a theorem from multivariable calculus)

$$g(\mathbf{y}) = \frac{1}{|\det(\mathbf{A})|} f\left(\mathbf{A}^{-1}\mathbf{y}\right) = \left(\frac{1}{1}\right) f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n y_i^2\right), \tag{1.27}$$

where in the last equality we have used Equation (1.26), which also holds for the realisations $\mathbf{z}$ and $\mathbf{y}$. Equation (1.27) 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 1.7.2. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random variables distributed according to $\mathrm{N}\left(\mu, \sigma^2\right)$, 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

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{1.28}$$

♦

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 1.7.1, 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 (1.13) 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 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:

$$\begin{aligned} &\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n \left(Z_i - \overline{Z}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n \left(X_i - \overline{X}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \\ &\sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1} \qquad \text{(the } Y_i \text{ are i.i.d. standard normal, see Remark 1.7.3)} \qquad (1.29) \\ \Rightarrow &\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \end{aligned}$$

which proves the result. This proof is taken from [3] and [7]. □

Remark 1.7.3. Equation (1.29) 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 $\sum_{i=1}^n Y_i^2 \sim \chi^2_n$.

□

1.8 Other measures of central tendency and dispersion

1.8.1 The mode

Definition 1.8.1. For a random variable X with probability density function f_X, the **mode** of the distribution of X is defined as

$$\operatorname{mode}(X) = \arg\max_x f_X(x) \tag{1.30}$$

where 'argmax' is described in Remark 1.8.18. ■

1.8.2 The median

Definition 1.8.2. For a random variable X, a **median** of the distribution of X is defined as a value m such that

$$\mathrm{P}(X \geq m) \geq \tfrac{1}{2} \qquad \text{and} \qquad \mathrm{P}(X \leq m) \geq \tfrac{1}{2}. \tag{1.31}$$

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

Example 1.8.3. 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 1.8.4. Note that the median is not unique. One can modify Example 1.8.3 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 $\mathrm{P}(X \leq 2) \leq \frac{1}{2}$ and $\mathrm{P}(X \geq 3) \geq \frac{1}{2}$, and so 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 results will be referenced in later sections, but their proofs will be exercises.

Theorem 1.8.5. 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 1.8.6. Prove Theorem 1.8.5. △

See the solution to Question 4 on Problem Sheet 9, Week 17.

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

Definition 1.8.7. Given a sample of observations $x_1, x_2, \ldots, x_n$, the **sample median** m is defined as

$$m = \begin{cases} x_{([n+1]/2)}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & \text{if } n \text{ is even,} \end{cases} \tag{1.32}$$

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. ■

The following is the analogue of Exercise 1.2.10:
Proposition 1.8.8. Given a sample of observations $x_1, x_2, \dots, x_n$, with sample median m. Then, for any real value a,

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

♦

Exercise 1.8.9. Prove Proposition 1.8.8. △

See the solution to Question 3 on the Problem-based Learning Sheet 9, Week 18.

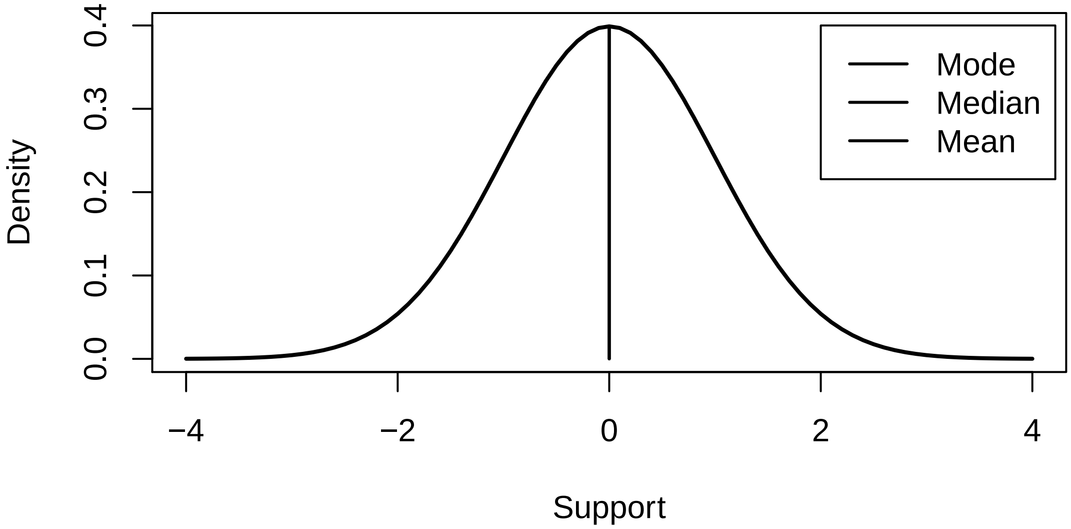
1.8.3 Interquartile range

Definition 1.8.10. 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 1.8.11. 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, $F_X(x) = \mathrm{P}\left(X \leq x\right)$. While this might seem to be an abuse of notation, the cumulative distribution function completely specifies a probability distribution. □
Remark 1.8.12. Given a quantile function F_X^{-1} for a random variable X, one can define the median to be $m = F_X^{-1}\left(0.5\right)$. □
Definition 1.8.13. Given a quantile function F_X^{-1} for a random variable X, the **lower quartile** is defined as $q_{0.25} = F_X^{-1}\left(0.25\right)$ while the **upper quartile** is defined as $q_{0.75} = F_X^{-1}\left(0.75\right)$ ■
Definition 1.8.14. Given a quantile function F_X^{-1} for a random variable X, the **interquartile range** is defined as $\mathrm{IQR} = F_X^{-1}\left(0.75\right) - F_X^{-1}\left(0.25\right)$. ■
Remark 1.8.15. 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 $x_1, x_2, \dots, x_n$, 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 $x_{(4)}$.

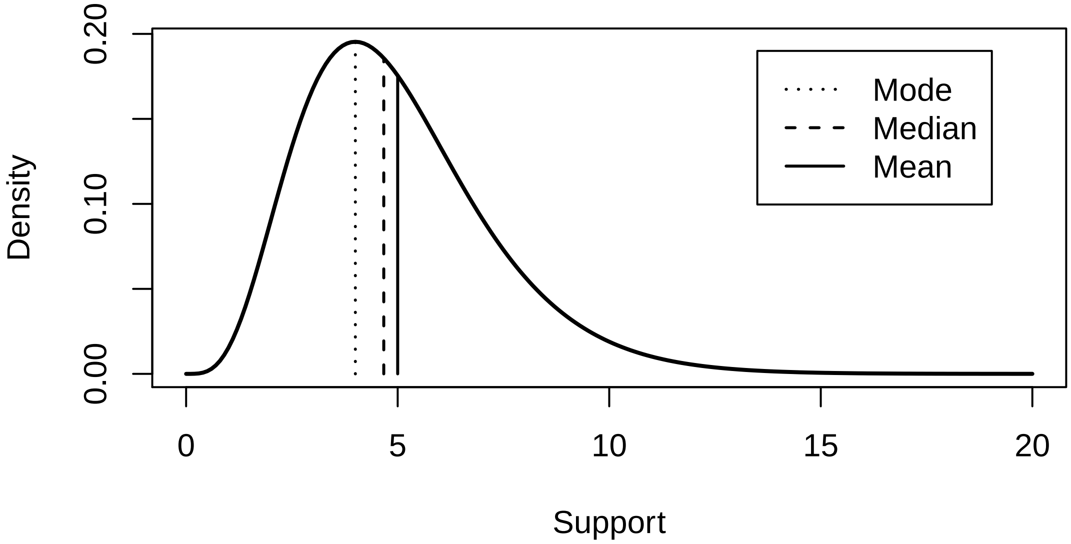
Now, for the lower quartile $q_{0.25}$, one computes the index $i_{0.25} = \frac{1}{2}\left(\lfloor i_m \rfloor + 1\right) = \frac{1}{2}\left(\left\lfloor \frac{n+1}{2} \right\rfloor + 1\right)$. If this is an integer, the lower quatile is $x_{(i_{0.25})}$, otherwise it is the average of $x_{(i_{0.25} - 0.5)}$ 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 1.8.16. 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 1.8.17. 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 $\frac{1}{2}\left(12 + 14\right) = 13$. Since $\lfloor i_m \rfloor = 6$, $i_{0.25} = 3.5$ and $i_{0.75} = 9.5$. Therefore, $q_{0.25} = \frac{1}{2}\left(6 + 8\right) = 7$ and $q_{0.75} = \frac{1}{2}\left(18 + 20\right) = 19$. △

1.8.4 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.



Remark 1.8.18. 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 1.8.19. 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 $f(0) = 1$. △

Chapter 2

Statistical models

2.1 Review of probability models

Let us review the definitions of a **sample space**, **probability measure** and **probability model** [3, 4].

Definition 2.1.1. 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 2.1.2. A **probability measure** P on a sample space Ω is a specification of numbers $\mathrm{P}\left(A\right)$ for all events $A \in \Omega$ such that

1. for every event A, $\mathrm{P}\left(A\right) \geq 0$,
2. $\mathrm{P}\left(\Omega\right) = 1$,
3. for every finite or countable sequence of disjoint events $\{A_j \mid j \in J\}$, $\mathrm{P}\left(\bigcup_{j\in J} A_j\right) = \sum_{j\in J} \mathrm{P}\left(A_j\right)$.

■

Definition 2.1.3. 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.

■

2.1.1 Notation for random variables and realisations

This is a good point at which to reemphasise the distinction in notation between **random variables**, and **realisations** of random variables. Recall the definition of a random variable

Definition 2.1.4. 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 sample.

Definition 2.1.5. 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
```

2.2 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 2.2.1. Suppose it is known that the lifespan X in years for a particular smartphone follows the distribution $X \sim \mathrm{Exp}(\lambda)$ with $\lambda = 1$; see Figure 2.1 for a plot of this distribution.

(a) One option for estimating the lifespan of a new smartphone would be to compute $\mathrm{E}\left(X\right) = 1$ 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

$$\mathrm{P}\left(X > 5\right) = \int_5^\infty e^{-x} \mathrm{d}x = e^{-5} = 0.0067,$$

which would lead one consider such a 5-year lifetime for this smartphone to be unlikely.

△

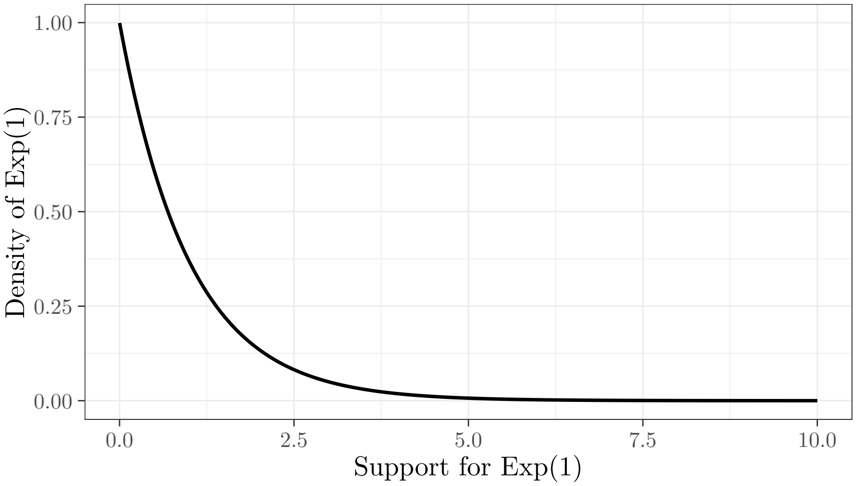


Figure 2.1: Plot showing density of $f(x) = e^{-x}$ for the Exp(1) distribution.

Let us review the situation for Example 2.2.1: 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.

2.3 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 2.3.1. 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 2.3.2. 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 2.3.3. 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 $\{\mathrm{P}_1, \mathrm{P}_2\}$, where P_1 is the Exp(0.5) probability measure and P_2 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 2.2, which model would you be inclined to say is the correct one? What if the observed data had been $(1.91, 2.46, 1.08, 5.79, 0.29)$? △

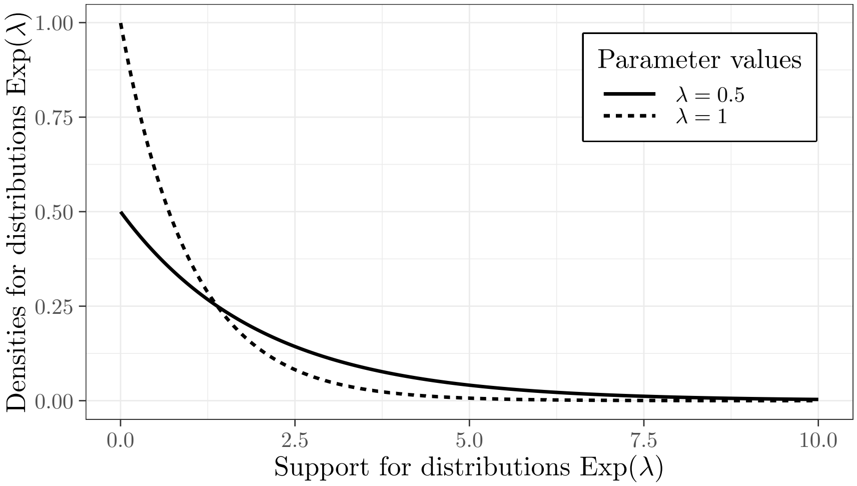


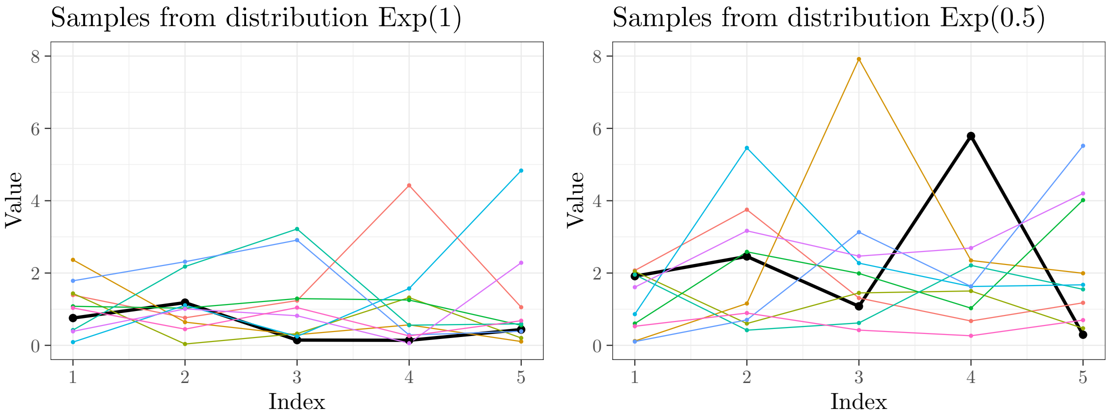
Figure 2.2: Plot showing density of $f(x|\lambda) = \lambda e^{-\lambda}$ for the Exp(λ) distribution, for $\lambda \in \{0.5, 1\}$

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 2.3.3 were sampled from the Exp(0.5) and Exp(1) distributions using R:

```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
```


Samples from distribution Exp(1)

Samples from distribution Exp(0.5)

Figure 2.3: 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 2.3.3 are shown as **thick black lines**.

A more concrete definition of a statistical model could be [3]:
Definition 2.3.4. 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. (if desired) 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 2.3.5. 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 [6]. □

Chapter 3

Likelihood

3.1 The likelihood function

Definition 3.1.1. Suppose we have a statistical model for the random variables $\mathbf{X}$ described by $\{\mathrm{P}_\theta \mid \theta \in \Theta\}$, 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 $L(\theta|\mathbf{x}) = f_\theta\left(\mathbf{x}\right)$ for any $\theta \in \Theta$. ■

Given the definition of a likelihood function, we also have
Definition 3.1.2. 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 3.1.3. The probability density (mass) function $f_\theta\left(\mathbf{x}\right)$ in Definition 3.1.1 can also be denoted $f(\mathbf{x}|\theta) = f_\theta\left(\mathbf{x}\right)$ 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:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta). \tag{3.1}$$

□

Remark 3.1.4. In Definition 3.1.1 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 3.1.5. Througtout 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 3.1.6. Suppose that the data $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ are independently sampled from a $\mathrm{N}\left(\theta, 1\right)$ 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 3.4.3. △

3.2 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 3.1.1. 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 3.2.1. 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 $\mathrm{Bin}(10, \theta)$ model, with likelihood function given by

$$L(\theta|3) = \binom{10}{3}\theta^3(1-\theta)^7.$$

This likelihood function is shown in Fig. 3.1. Notice the global maximum at $\theta = 0.3$, with value 0.27. △

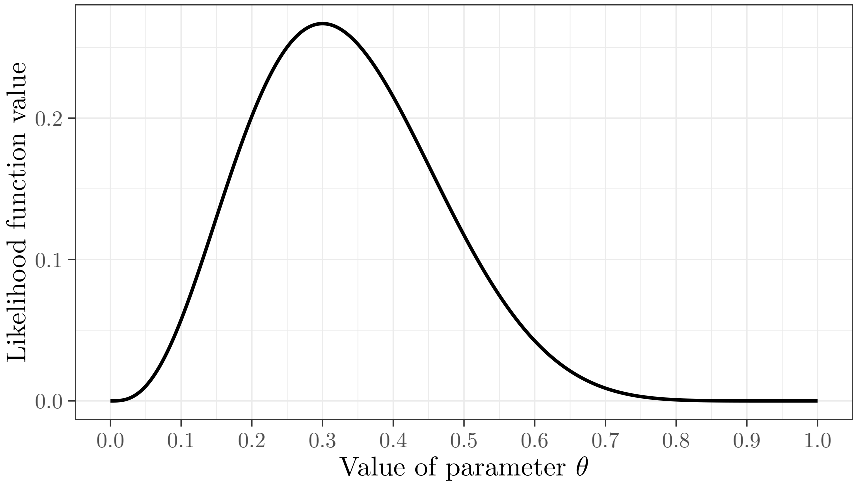


Figure 3.1: Plot showing likelihood of $L(\theta|3) = \binom{10}{3}\theta^3(1-\theta)^7$ for $\theta \in [0, 1]$.

3.3 Likelihood ratios

However, for a **continuous** random variable $\mathbf{X}$, evaluating the likelihood at a particular observed value $\mathbf{x}$ is $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \mathrm{P}_\theta\left(\mathbf{X} = \mathbf{x}\right) = 0$. Indeed, even for discrete distributions, likelihoods can be vanishingly small.

Example 3.3.1. 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 $\left\{1, 2, \ldots, 10^5\right\}$ and P_2 is the discrete uniform distribution on the set $\left\{1, 2, \ldots, 10^8\right\}$. 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 the 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

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})} = \frac{\mathrm{P}_{\theta_1}\left(\mathbf{X} = \mathbf{x}\right)}{\mathrm{P}_{\theta_2}\left(\mathbf{X} = \mathbf{x}\right)}, \qquad \theta_1, \theta_2 \in \Theta, \qquad \mathbf{X} \text{ is } \textbf{discrete}.$$

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 3.3.2)

$$\mathrm{P}_\theta\left(x - \delta < X < x + \delta\right) \approx 2\delta f(x|\theta). \tag{3.2}$$

By Definition 3.1.1, $2\delta f(x|\theta) = 2\delta L(\theta|x)$, and so for two parameter values θ_1 and θ_2,

$$\frac{L(\theta_1|x)}{L(\theta_2|x)} \approx \frac{\mathrm{P}_{\theta_1}\left(x - \delta < X < x + \delta\right)}{\mathrm{P}_{\theta_2}\left(x - \delta < X < x + \delta\right)}, \qquad \theta_1, \theta_2 \in \Theta, \qquad X \text{ is } \textbf{continuous}.$$

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 3.3.2. 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 (3.2) 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,

$$\begin{aligned}
\mathrm{P}_\theta\left(x - \delta < X < x + \delta\right) - 2\delta f_\theta(x) \quad &= \int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x)\right]\mathrm{d}z \\
\Rightarrow \left|\mathrm{P}_\theta\left(x - \delta < X < x + \delta\right) - 2\delta f_\theta(x)\right| &= \left|\int_{x-\delta}^{x+\delta} \left[f_\theta(z) - f_\theta(x)\right]\mathrm{d}z\right| \\
&\leq \int_{x-\delta}^{x+\delta} \left|f_\theta(z) - f_\theta(x)\right|\mathrm{d}z \\
&< \int_{x-\delta}^{x+\delta} \epsilon \ \mathrm{d}z \\
\Rightarrow \left|\mathrm{P}_\theta\left(x - \delta < X < x + \delta\right) - 2\delta f_\theta(x)\right| &< 2\delta\epsilon.
\end{aligned}$$

So, for small enough δ, $\mathrm{P}_\theta\left(x - \delta < X < x + \delta\right) \approx 2\delta f_\theta(x) = 2\delta f(x|\theta)$. △

Exercise 3.3.3. Show that

$$\sum_{i=1}^{n} \left(x_i - \theta\right)^2 = (n-1)s^2 + n\left(\overline{x} - \theta\right)^2,$$

where $\overline{x}$ and s^2 are defined in terms of $x_1, x_2, \ldots, x_n$ as usual.

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

since $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = n\overline{x} - n\overline{x} = 0$.

△

3.4 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 3.4.1. Given the likelihood function $L(\cdot|\mathbf{x})$ from Definition 3.1.1, any function $L'(\cdot|\mathbf{x}) = cL(\cdot|\mathbf{x})$ for $c > 0$ is an **equivalent likelihood function** for the parameter θ. ■

Exercise 3.4.2. 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 3.4.1, 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 3.4.3. Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a sample of observations i.i.d. according to a $\mathrm{N}\left(\theta, \sigma^2\right)$ distribution, where σ^2 is unknown but $\theta \in= \mathbb{R}$ is unknown. Then the likelihood is

$$\begin{aligned} L(\theta|\mathbf{x}) &= f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left(x_i - \theta\right)^2\right) \\ \Rightarrow L(\theta|\mathbf{x}) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \theta\right)^2\right) && (3.3) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n\left(\overline{x} - \theta\right)^2\right]\right) && \text{(Exercise 3.3.3)} \\ &= \underbrace{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n-1}{2\sigma^2}s^2\right)}_{c>0} \exp\left(-\frac{n}{2\sigma^2}\left(\overline{x} - \theta\right)^2\right), \end{aligned}$$

which shows that an equivalent likelihood is

$$L'(\theta|\mathbf{x}) = \exp\left(-\tfrac{n}{2\sigma^2}\left(\overline{x} - \theta\right)^2\right). \qquad (3.4)$$

△

3.5 The Likelihood Principle

The following principle offers an approach for inference based on the likelihood function.

Principle 3.5.1. If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $L(\theta|\mathbf{x})$ is proportional to $L(\theta|\mathbf{y})$, i.e. there exists a constant $C(\mathbf{x}, \mathbf{y})$ such that

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y})\, L(\theta|\mathbf{y}), \text{ for all } \theta, \tag{3.5}$$

then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical. ■

Remark 3.5.2. The Likelihood Principle can be restated: If two model, data combinations yield equivalent likelihood functions, then inferences about the unknown parameters must be the same. □

Yet another way to restate it is that if two sample points $\mathbf{x}$ and $\mathbf{y}$ have proportional likelihoods, then they contain equivalent information about θ.

Example 3.5.3. Suppose that for sample points $\mathbf{x}$, with likelihood function $L(\cdot|\mathbf{x})$, and parameter values θ_1, θ_2 one has $L(\theta_2|\mathbf{x}) = 2L(\theta_1|\mathbf{x})$. Then, one can interpret this as saying that the value θ_2 for θ is twice as plausible as the value θ_1 for θ. Furthermore, suppose that the Likelihood Principle holds and Equation (3.5) is true for this sample $\mathbf{x}$ and another sample $\mathbf{y}$. Then

$$L(\theta_2|\mathbf{y}) = \frac{1}{C(\mathbf{x}, \mathbf{y})} L(\theta_2|\mathbf{x}) = \frac{2}{C(\mathbf{x}, \mathbf{y})} L(\theta_1|\mathbf{x}) = 2L(\theta_1|\mathbf{y}).$$

Therefore, whether one observes $\mathbf{x}$ or $\mathbf{y}$, one can conclude that the value θ_2 is twice as plausible as θ_1. △

However, not all statisticians agree with the use of the likelihood principle. Let us consider the following example:

Example 3.5.4. 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 experiment will be: toss the coin until exactly 3 heads are observed. The experiment is performed, and exactly $x = 7$ tails are observed before the 3rd head is observed. An appropriate statistical model for x is the negative binomial distribution $\text{NegBin}(3, \theta)$ model, with likelihood function given by

$$L(\theta|3) = \binom{9}{2} \theta^3 (1-\theta)^7.$$

Notice that this likelihood function is a positive multiple of the likelihood function in Example 3.2.1. Therefore, the likelihood principle tells us that one can make the same inferences about the unknown value of the parameter θ, even though the data were obtained in different ways. Some statisticians believe that additional information, such as the sampling method, should be taken into account, which could lead to different inferences rather than if one solely relied on the likelihood function. △

3.6 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 3.6.1. 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 3.6.2. 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 3.6.3. Note that we use MLE as an abbreviation for both the maximum likelihood **estimator** and the maximum likelihood **estimate**. □

3.6.1 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 $(\theta_1, \theta_2, \dots, \theta_m)$ 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 3.6.4. As in Example 3.1.6, suppose that the data $x = \{x_1, x_2, \dots, x_n\}$ 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:

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}L(\theta|\mathbf{x}) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \cdot \left(\sum_{i=1}^{n}(x_i-\theta)\right)\right] \\ &= \frac{1}{(2\pi)^{n/2}}\left[\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \cdot \left(\sum_{i=1}^{n}(x_i-\theta)\right)^2 + \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \cdot \left(\sum_{i=1}^{n}(-1)\right)\right] \end{aligned}$$

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 $\lim_{\pm\infty} L(\theta|\mathbf{x}) = 0$. On the other hand, when $L(\theta|\mathbf{x})$ is evaluated at $\theta = \overline{x}$, we see that $L(\theta = \overline{x}|\mathbf{x}) = (2\pi)^{-n/2} > 0$.

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 3.6.5. 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 3.6.6. (Continuation of Example 3.6.4) Let us try and find the MLE of the likelihood $L(\theta|\mathbf{x})$ in Example 3.6.4 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. △

3.6.2 The log-likelihood

There will be times when we do not see a direct method of maximisation (as in Example 3.6.6), 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 3.6.7. 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 $\widehat{\theta} = \overline{x}$. △

3.6.3 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 3.6.8. Suppose the observations $x_1, x_2, \ldots, x_n$ are independently sampled from a $\mathrm{N}\left(\mu, \sigma^2\right)$ distribution, with both $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ unknown. Similarly to Example 3.1.6, 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

$$\begin{aligned} \widehat{\theta}_1 &= \widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \\ \widehat{\theta}_2 &= \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) \end{aligned} \tag{3.6}$$

Therefore the MLE of μ is $\overline{X}$ and the MLE of σ^2 is S_b^2. △

Chapter 4

Covariance and correlation

The covariance of two random variables X and Y, denoted $\operatorname{Cov}(X, Y)$, 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 $\operatorname{Cov}(X, Y)$ 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

$$\begin{aligned} \mathrm{E}(X) &= \mu_X, \qquad \operatorname{Var}(X) = \sigma_X^2 \\ \mathrm{E}(Y) &= \mu_Y, \qquad \operatorname{Var}(Y) = \sigma_Y^2 \end{aligned}$$

We shall also assume that the variances are non-zero and finite, i.e. $0 < \sigma_X^2, \sigma_Y^2 < \infty$.

4.1 Covariance

Definition 4.1.1. The **covariance** of two random variables X and Y, with means μ_X and μ_Y, respectively, is the number defined by

$$\operatorname{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]. \tag{4.1}$$

■

Remark 4.1.2. From the definition, the covariance of a variable with itself is its variance:

$$\operatorname{Cov}(X, X) = \mathrm{E}[(X - \mu_X)^2] = \operatorname{Var}(X).$$

□

Remark 4.1.3. From the definition, it is also immediate that the covariance is a symmetric function,

$$\operatorname{Cov}(Y, X) = \operatorname{Cov}(X, Y)$$

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

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y). \tag{4.2}$$

Since this is such a fundamental result, it is worth proving the following generalisation as an exercise:

Exercise 4.1.4. For any two random variables X and Y, and constants $a, b \in \mathbb{R}$,

$$\mathrm{Var}\left(aX + bY\right) = a^2\mathrm{Var}\left(X\right) + b^2\mathrm{Var}\left(Y\right) + 2ab\mathrm{Cov}\left(X, Y\right). \tag{4.3}$$

The mean of $aX + bY$, which we will denote by μ_{aX+bY}, is

$$\mathrm{E}\left(aX + bY\right) = a\mathrm{E}\left(X\right) + b\mathrm{E}\left(Y\right) = a\mu_X + b\mu_Y = \mu_{aX+bY}.$$

Then, using the definitions of variance and covariance,

$$\begin{aligned}
\mathrm{Var}\left(aX + bY\right) &= \mathrm{E}[(aX + bY - \mu_{aX+bY})^2] \\
&= \mathrm{E}[(aX + bY - (a\mu_X + b\mu_Y))^2] \\
&= \mathrm{E}[(a\left(X - \mu_X\right) + b\left(Y - \mu_Y\right))^2] \\
&= \mathrm{E}[a^2\left(X - \mu_X\right)^2 + b^2\left(Y - \mu_Y\right)^2 + 2ab\left(X - \mu_X\right)\left(Y - \mu_Y\right)] \\
&= a^2\mathrm{E}[(X - \mu_X)^2] + b^2\mathrm{E}[(Y - \mu_Y)^2] + 2ab\mathrm{E}[(X - \mu_X)\left(Y - \mu_Y\right)] \\
&= a^2\mathrm{Var}\left(X\right) + b^2\mathrm{Var}\left(Y\right) + 2ab\mathrm{Cov}\left(X, Y\right)
\end{aligned}$$

△

We might wonder, for any two random variables X and Y, what possible values $\mathrm{Cov}\left(X, Y\right)$ can take. The next example provides an answer.

Example 4.1.5. 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,

$$\mathrm{Var}\left(X + Y\right) = \mathrm{Var}\left(X + aX + b\right) = \mathrm{Var}\left((a+1)X + b\right) = (a+1)^2\mathrm{Var}\left(X\right) \tag{4.4}$$

On the other hand, using Equation (4.2)

$$\begin{aligned}
2\mathrm{Cov}\left(X, Y\right) &= \mathrm{Var}\left(X + Y\right) - \mathrm{Var}\left(X\right) - \mathrm{Var}\left(Y\right) \\
&= (a+1)^2\mathrm{Var}\left(X\right) - \mathrm{Var}\left(X\right) - a^2\mathrm{Var}\left(X\right) \\
&= (a^2 + 2a + 1)\mathrm{Var}\left(X\right) - \mathrm{Var}\left(X\right) - a^2\mathrm{Var}\left(X\right) \\
&= 2a\mathrm{Var}\left(X\right) \\
\Rightarrow \mathrm{Cov}\left(X, Y\right) &= a\mathrm{Var}\left(X\right)
\end{aligned}$$

This is true for any value of $a \in \mathbb{R}$. If $\mathrm{Var}\left(X\right)$ is finite, then for any $c \in \mathbb{R}$, set $a = c(\mathrm{Var}\left(X\right))^{-1}$; then $\mathrm{Cov}\left(X, Y\right) = c$. This shows that, for this example, we can choose a such that $\mathrm{Cov}\left(X, Y\right)$ 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 $\mathrm{Var}\left(X\right)$ and $\mathrm{Var}\left(Y\right)$. The next theorem describes this relationship.

Theorem 4.1.6. 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

$$f(t) = \mathrm{E}[((X-\mu_X)\,t + (Y-\mu_Y))^2],$$

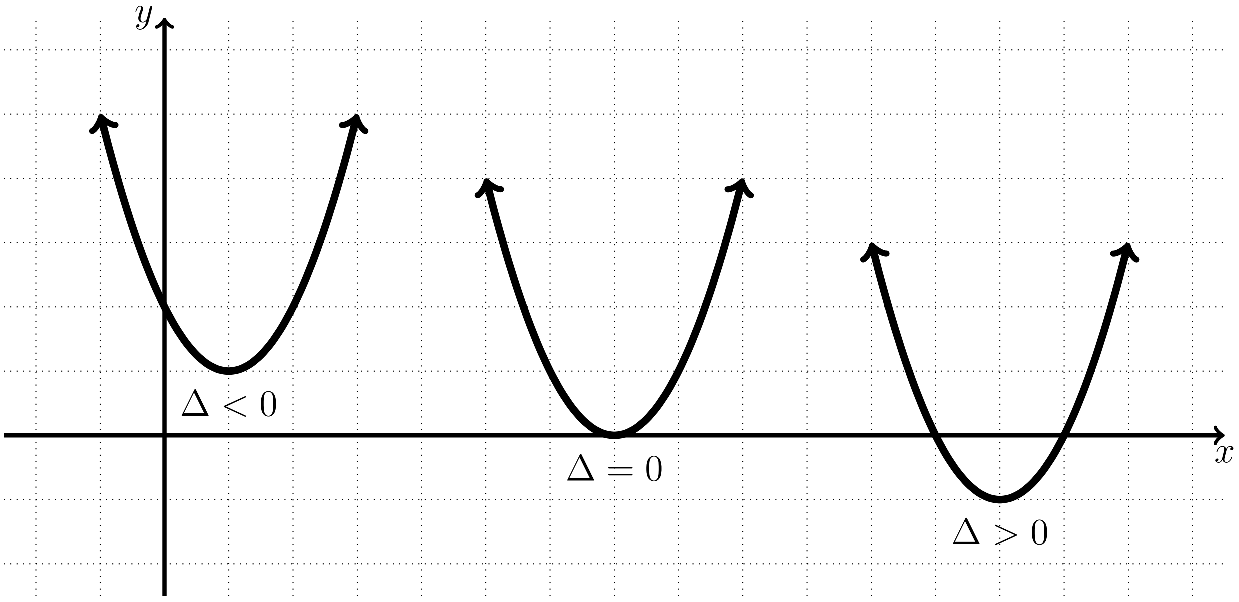
where μ_X and μ_Y are the means of X and Y, respectively. Expanding the square inside the expectation, and using the linearity of expectation, one obtains

$$\begin{aligned} f(t) &= \mathrm{E}[t^2\,(X-\mu_X)^2 + 2t\,(X-\mu_X)\,(Y-\mu_Y) + (Y-\mu_Y)^2] \\ &= t^2\mathrm{E}[(X-\mu_X)^2] + 2t\mathrm{E}[(X-\mu_X)\,(Y-\mu_Y)] + \mathrm{E}[(Y-\mu_Y)^2] \\ \Rightarrow f(t) &= t^2\sigma_X^2 + 2t\mathrm{Cov}\,(X,Y) + \sigma_Y^2 \end{aligned} \tag{4.5}$$

Now, we notice two things:

1. $f(t)$ is quadratic in t, i.e. $f(t) = at^2 + bt + c$, where $a = \mathrm{Var}\,(X)$, $b = 2\mathrm{Cov}\,(X,Y)$ and $c = \mathrm{Var}\,(Y)$.
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 below illustrates this deduction.



Computing the discriminant $\Delta = b^2 - 4ac$ from Equation (4.5),

$$\begin{aligned} &(2\mathrm{Cov}\,(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 \\ \Rightarrow\, &(\mathrm{Cov}\,(X,Y))^2 \leq \sigma_X^2\sigma_Y^2 \\ \Rightarrow\, &|\mathrm{Cov}\,(X,Y)| \leq \sigma_X\sigma_Y \end{aligned}$$

which proves the result. □

Remark 4.1.7. There is another proof of this result using the Cauchy-Schwartz inequality. □

4.1.1 The sample covariance

Suppose the random variables $X_1, X_2, \dots, X_n$ follow distribution F_X, and the random variables $Y_1, Y_2, \dots, Y_n$ 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, X_n are observed as $x_1, x_2, \dots, x_n$, respectively, and the random variables $Y_1, Y_2, \dots, Y_n$ are observed as $y_1, y_2, \dots, y_n$, respectively. Then the observed sample covariance is defined as

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right).$$

4.2 Correlation

Given Theorem 4.1.6 the concept of correlation now arises naturally.

Definition 4.2.1. 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

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{4.6}$$

■

Remark 4.2.2. 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 4.2.3. 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 4.1.6:

Corollary 4.2.4. For any random variables X and Y,

$$-1 \leq \rho_{XY} \leq 1.$$

♦

Proof. From the definition of correlation and Theorem 4.1.6,

$$\begin{aligned} |\rho_{XY}| &= \left|\frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}\right| \leq \left|\frac{\sigma_X \sigma_Y}{\sigma_X \sigma_Y}\right| = 1 \\ \Rightarrow -1 &\leq \rho_{XY} \leq 1 \end{aligned}$$

□

One may wonder under which circumstances $\rho_{XY} = \pm 1$. The next two results provide the answer.

Lemma 4.2.5. Suppose Z is a non-negative random variable. Then $\mathrm{E}[Z] = 0$ if and only if $\mathrm{P}(Z = 0) = 1$. ♦

Proof. We shall prove the result in the discrete case. Since $Z \geq 0$, for any realisation z in the image of Z, $\mathrm{Im}(Z)$, we have $z \geq 0$. Then

$$\mathrm{E}[Z] = \left(0 \cdot \mathrm{P}(Z = 0) + \sum_{z \in \mathrm{Im}(Z), z > 0} z\mathrm{P}(Z = z) \right) = \sum_{z \in \mathrm{Im}(Z), z > 0} z\mathrm{P}(Z = z)$$

Now, since $\mathrm{P}(Z = z) \geq 0$ for all $z \in \mathrm{Im}(Z)$, this implies $z\mathrm{P}(Z = z) \geq 0$ for all $z \in \mathrm{Im}(Z)$ with $z > 0$. Furthermore, we can write

$$\mathrm{P}(Z = 0) = 1 - \sum_{z \in \mathrm{Im}(Z), z > 0} \mathrm{P}(Z = z).$$

Therefore, we have

$$\begin{aligned}
&\mathrm{E}[Z] = 0 \\
\iff &(z\mathrm{P}(Z = z)) = 0 \text{ for all } z \in \mathrm{Im}(Z) \text{ with } z > 0 \\
\iff &\mathrm{P}(Z = z) = 0 \text{ for all } z \in \mathrm{Im}(Z) \text{ with } z > 0 \\
\iff &\left(\sum_{z \in \Omega, z > 0} \mathrm{P}(Z = z) \right) = 0 \\
\iff &\mathrm{P}(Z = 0) = 1
\end{aligned}$$

This proves the result in the discrete case; the continuous case is similar. □

Corollary 4.2.6. 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 4.1.6, 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 $|\rho_{XY}| = 1$ if and only if the discriminant of $f(t)$ is $\Delta = 0$ if and only if $f(t) = 0$ has a single root. 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 4.2.5 gives us that

$$\begin{aligned}
&\mathrm{E}[((X - \mu_X)t + (Y - \mu_Y))^2] = 0 \\
\iff &\mathrm{P}\left([(X - \mu_X)t + (Y - \mu_Y)]^2 = 0\right) = 1 \\
\iff &\mathrm{P}((X - \mu_X)t + (Y - \mu_Y) = 0) = 1 \\
\iff &\mathrm{P}(Y = -tX + (\mu_X t + \mu_Y)) = 1 \\
\iff &\mathrm{P}(Y = aX + b) = 1
\end{aligned}$$

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. □

4.2.1 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 4.2.7. 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 4.2.8. 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 4.2.9. 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.

$$\begin{aligned} \overline{u} &= \frac{1}{n}u_i = \frac{1}{n}(ax_i + b) = a\left(\frac{1}{n}x_i\right) + b = a\overline{x} + b \\ \Rightarrow & u_i - \overline{u} = a(x_i - \overline{x}) \end{aligned}$$

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 4.2.10. If we allow either $a < 0$ or $c < 0$ (but not both) in Proposition 4.2.9, 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 4.2.11. 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. □

4.2.2 A real data example: height and shoe size

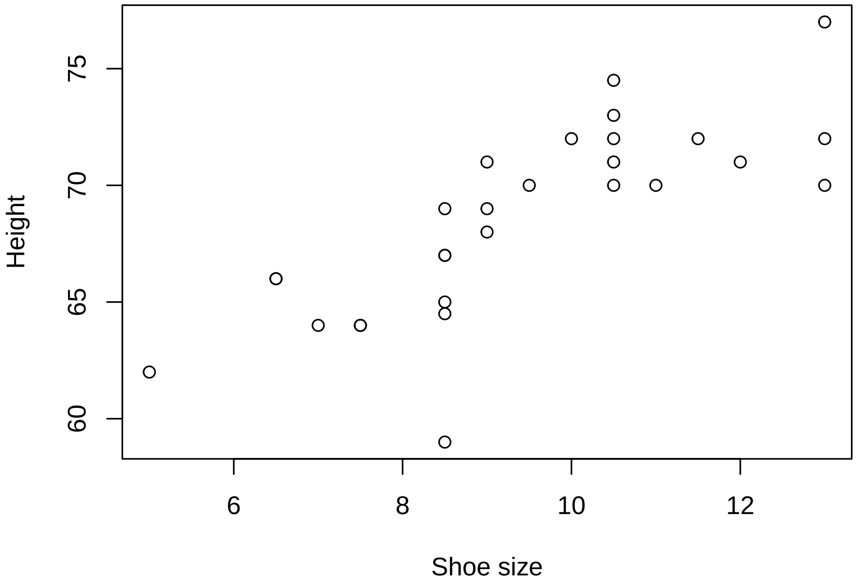
Let us look at some data from collected at Arizona State University, which records the heights and shoe sizes of 28 male and female students [10]. This data is saved in `shoesize.txt`; see Appendix A, Section A.3. The shoe sizes are given on the US scale, and height is given in inches.

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 4.1: Shoes sizes (US scale) and heights (inches) of 28 students from Arizona State University.

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="./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
```



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 5

Simple Linear Regression

In the previous chapter 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 [9, 2, 3].

5.1 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),

$$y_i = f(x_i) + \gamma_i, \qquad i \in \{1, 2, \ldots, n\}, \tag{5.1}$$

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

$$f(x_i) = \beta_0 + \beta_1 x_i + \delta_i, \tag{5.2}$$

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 (5.1) and (5.2), and combining the errors into $e_i = \gamma_i + \delta_i$, we finally obtain the expression

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i \in \{1, 2, \ldots, n\}. \tag{5.3}$$

Remark 5.1.1. Note that the e_i terms in Equation (5.3) are the realisations of the random errors. □

5.2 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 (5.3),

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i \in \{1, 2, \ldots, n\},$$

where the e_i are the (measurement and model) errors discussed in Section 5.1. 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 esimates $\widehat{\beta}_0$ 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

$$\widehat{e}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i), \qquad i \in \{1, 2, \ldots, n\}. \tag{5.4}$$

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)**

$$\sum_{i=1}^{n} (\widehat{e}_i)^2 \tag{5.5}$$

is as small as possible.

5.2.1 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

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 \tag{5.6}$$

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.

5.2.1.1 Finding $\widehat{\beta}_0$

First, we find $\widehat{\beta}_0$. We note that Equation (5.6) can be rewritten as

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[(y_i - \beta_1 x_i) - \beta_0\right]^2 .$$

It doesn't look as if much has changed, but if we write $z_i = y_i - \beta_1 x_i$, this becomes

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[z_i - \beta_0\right]^2 .$$

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

$$\widehat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_1 x_i) = \overline{y} - \beta_1 \overline{x},$$

we can conclude $\text{RSS}(\widehat{\beta}_0, \beta_1) \leq \text{RSS}(\beta_0, \beta_1)$, for all values of β_1.

5.2.1.2 Finding $\widehat{\beta}_1$

Having found $\widehat{\beta}_0$, we now need find the value of β_1 that minimises $\text{RSS}(\widehat{\beta}_0, \beta_1)$.

$$\begin{aligned}
\text{RSS}(\widehat{\beta}_0, \beta_1) &= \sum_{i=1}^{n} \left[(y_i - \beta_1 x_i) - (\overline{y} - \beta_1 \overline{x})\right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \overline{y}) - \beta_1 (x_i - \overline{x})\right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \overline{y})^2 - 2\beta_1 (x_i - \overline{x})(y_i - \overline{y}) + \beta_1^2 (x_i - \overline{x})^2\right] \\
&= S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx}
\end{aligned}$$

Completing the square (exercise) one can show

$$\begin{aligned}
\text{RSS}(\widehat{\beta}_0, \beta_1) &= S_{yy} - 2\beta_1 S_{xy} + \beta_1^2 S_{xx} \\
&= S_{xx} \left(\beta_1 - \frac{S_{xy}}{S_{xx}}\right)^2 + \left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right).
\end{aligned}$$

Therefore, $\text{RSS}(\widehat{\beta}_0, \widehat{\beta}_1) \leq \text{RSS}(\widehat{\beta}_0, \beta_1)$ where

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

5.2.1.3 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 $\widehat{\beta}_0$:

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \beta_1 \overline{x} \\
\Rightarrow \widehat{\beta}_0 &= \overline{y} - \left(\frac{S_{xy}}{S_{xx}}\right) \overline{x}.
\end{aligned}$$

Exercise 5.2.1. 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.

△

5.2.2 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. i

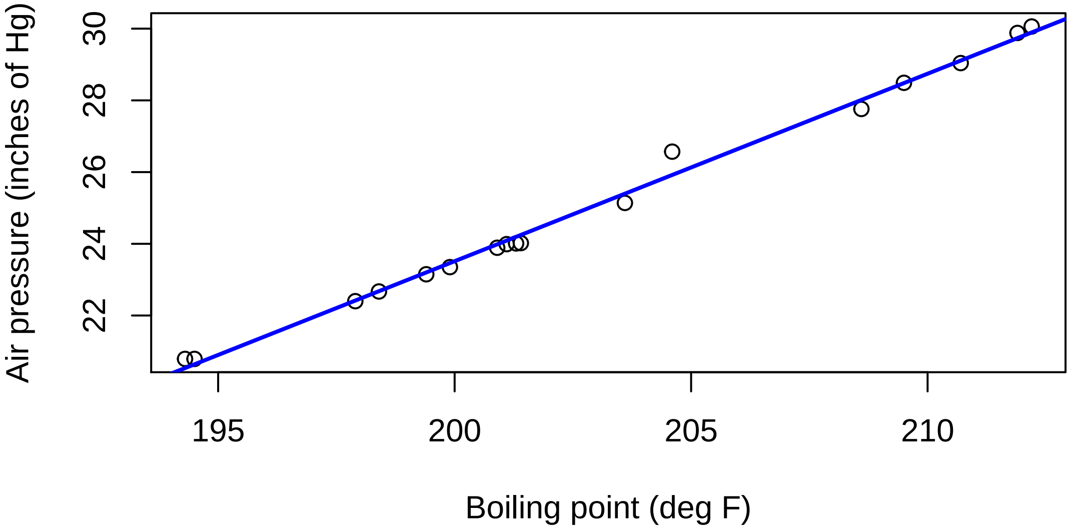
It was already known that altitude could be determined from atmospheric air pressure measured with a barometer, but barometers in the 1840s were fragile instruments.

In fact, his approach was indirect: it was already known that altitude could be determined from barometric air pressure, so embarked on exerp the relationship between the boiling point of water (m)

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 5.1: 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(forbes)
#>       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
#> 6  199.9 23.35
#> 7  200.9 23.89
#> 8  201.1 23.99
#> 9  201.4 24.02
#> 10 201.3 24.01
#> 11 203.6 25.14
#> 12 204.6 26.57
#> 13 209.5 28.49
#> 14 208.6 27.76
#> 15 210.7 29.04
#> 16 211.9 29.88
#> 17 212.2 30.06

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)
```



5.3 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 $i \in \{1, 2, \dots, n\}$.

We now define a new model, known as the **simple linear regression** model. For $i \in \{1, 2, \dots, n\}$,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{5.7}$$

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,
- $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$, 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 $i \in \{1, 2, \dots, n\}$.

Remark 5.3.1. The reason this model is called **simple** is not because it is 'easy', but rather because the model assumes only one independent quantity, the **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 5.3.2. 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 5.3.3. The choice of $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$ 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 5.3.4. The reason that the random errors ϵ_i are assumed to have mean $\mathrm{E}\left(\epsilon_i\right) = 0$ is that if one assumed that $\mathrm{E}\left(\epsilon_i\right) = \mu$ 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 $\mathrm{E}\left(\epsilon_i'\right) = 0$. 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.

5.3.1 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 $i \in \{1, 2, \dots, n\}$. Writing $\mathbf{y} = (y_1, y_2, \dots, y_n)$, using the independence of the $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$,

$$\begin{aligned} f(\mathbf{y}|\beta_0, \beta_1, \sigma^2) = f(y_1, y_2, \dots, y_n|\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(y_i|\beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right], \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right]. \end{aligned}$$

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 (5.6),

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2.$$

We have already found in Section 5.2.1 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 3.6.8 (see the solution to Question 4 on Problem-based Learning Sheet 11), we can take the derviative 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 5.3.5. 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 $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$ and are independent, the maximum likelihood estimators for the parameters β_0, β_1 and σ^2 are

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \left(\frac{S_{xy}}{S_{xx}} \right) \overline{x} \\
\widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
\widehat{\sigma}^2 &= \frac{1}{n} \left(\widehat{\mathrm{RSS}}_{xy} \right) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right]^2.
\end{aligned}$$

5.3.2 Residuals

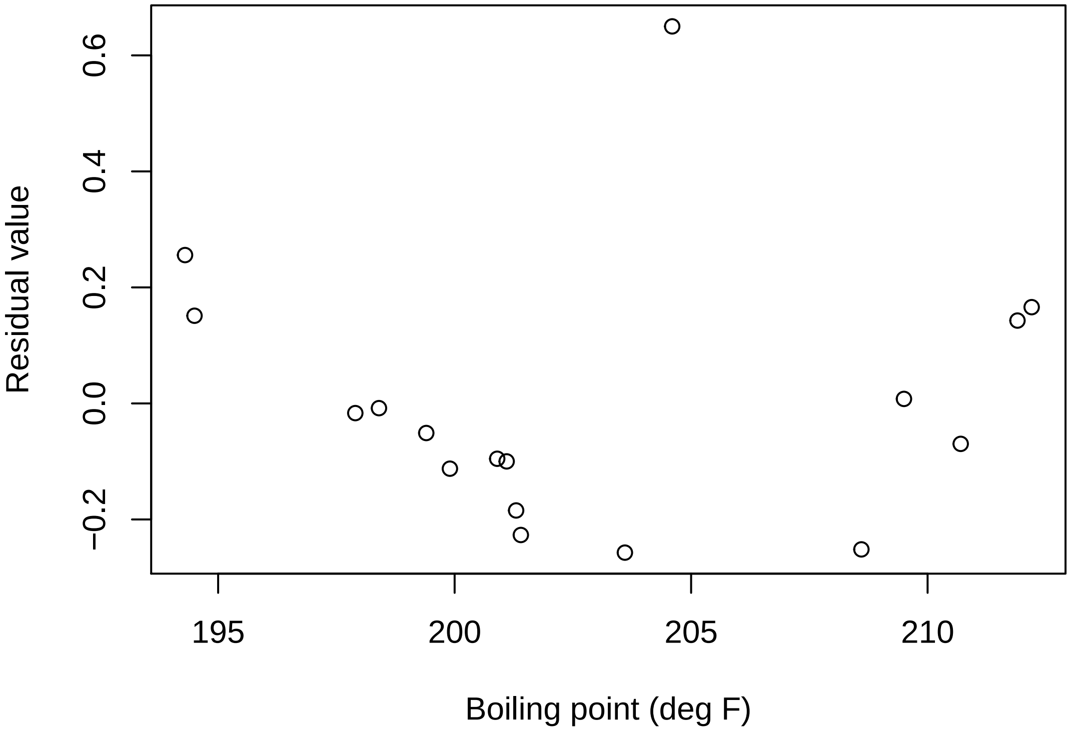
Given the maximum likelihood estimates $\widehat{\beta}_0, \widehat{\beta}_1$, we fit the straight line specified by the points $(x_i, f(x_i))$, where $f(x_i) = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. We can now observe the **residuals**

$$\widehat{e}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i).$$

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 $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$ and are independent. Therefore, **if our model is correct** the residuals $\widehat{\epsilon}_i$, when plotted, should appear to also be distributed according to some $\mathrm{N}\left(0, \sigma^2\right)$.

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.

```r
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")
```



5.3.3 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
beta1hat <- Sxy/Sxx
beta0hat <- ybar - beta1hat * xbar
beta_hat <- c(beta0hat, beta1hat)

# computing the residuals
residuals <- y - (beta0hat + beta1hat * x)

# here we use the lm function; not the use of the tilde "~"
model <- lm(y ~ x)

# now, to compare the parameters computed above and using lm
print(cbind(beta_hat, model$coefficients))
#>              beta_hat
#> (Intercept) -81.06373 -81.06373
#> x             0.52289   0.52289

# now, to compare the residuals computed above and using lm
print(cbind(residuals, model$residuals))
#>     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
#> 6  -0.1124638 -0.1124638
#> 7  -0.0953562 -0.0953562
#> 8  -0.0999347 -0.0999347
#> 9  -0.2268024 -0.2268024
#> 10 -0.1845131 -0.1845131
#> 11 -0.2571657 -0.2571657
#> 12  0.6499419  0.6499419
#> 13  0.0077692  0.0077692
#> 14 -0.2516277 -0.2516277
#> 15 -0.0697017 -0.0697017
#> 16  0.1428274  0.1428274
#> 17  0.1659597  0.1659597
```

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`.

5.3.4 Return to Forbes' data

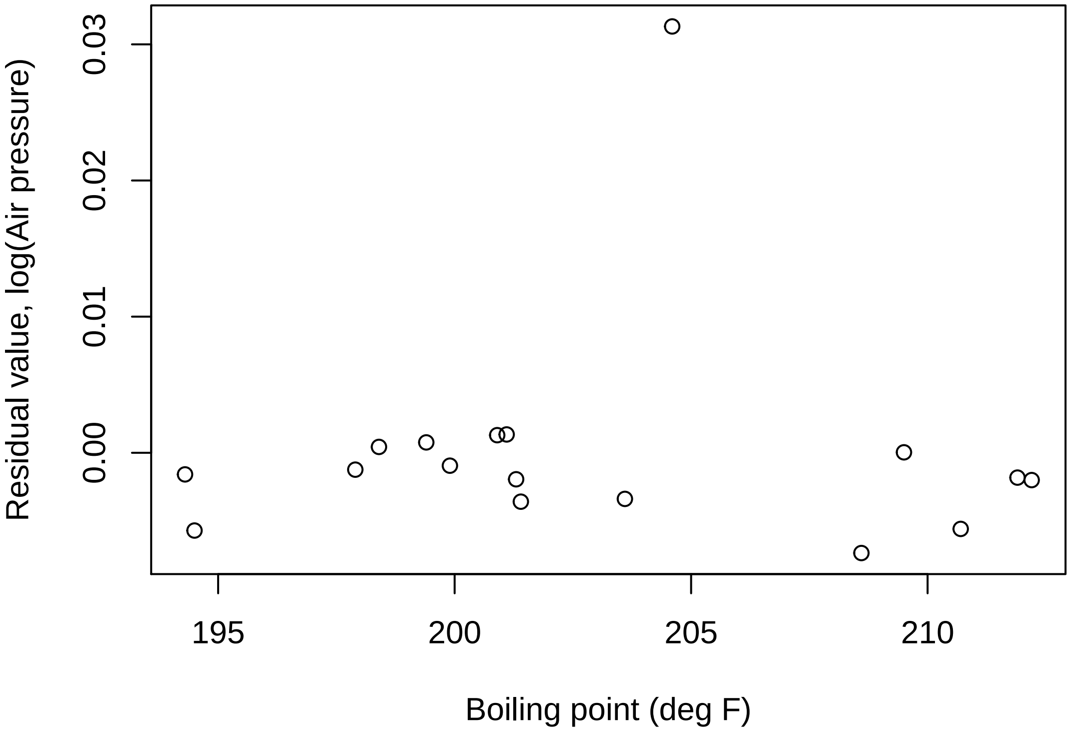
Let us try fitting the linear model again after first transforming the y_i and/or the x_i. Specifically, if we define the x_i to be the observations for boiling point and the y_i to be the observations for the air pressure for $i \in \{1, 2, \dots, n\}$, let's consider the transformation

$$Z_i = \log(Y_i),$$

and then fit the simple linear regression model

$$Z_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{5.8}$$

```r
library(MASS)
x <- forbes$bp
y <- forbes$pres
z <- log(y)
model_2 <- lm(z ~ x)

residuals <- model_2$residuals
plot(x, residuals, xlab="Boiling point (deg F)", ylab="Residual value, log(Air pressure)")
```



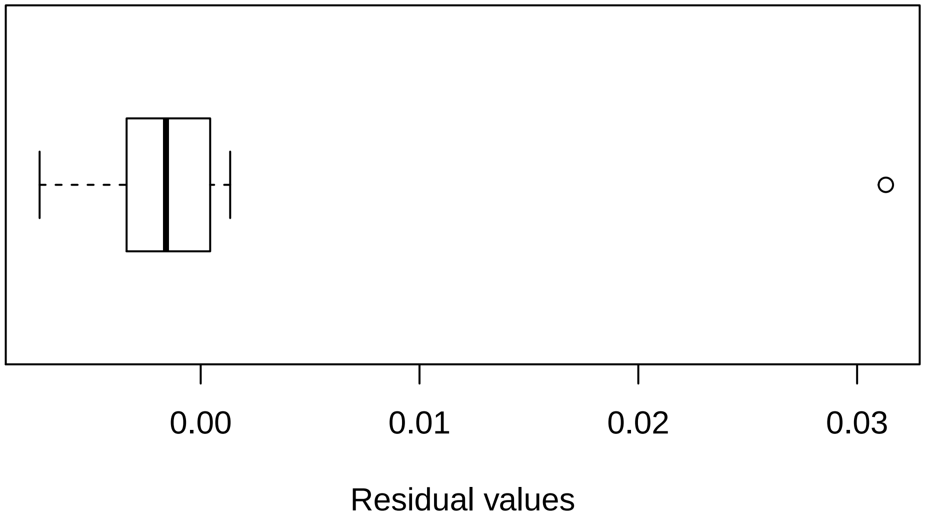
These residuals appear to be centred around 0 which suggests that this model is a better fit to the data and

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{5.9}$$

may be a better model than simply $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

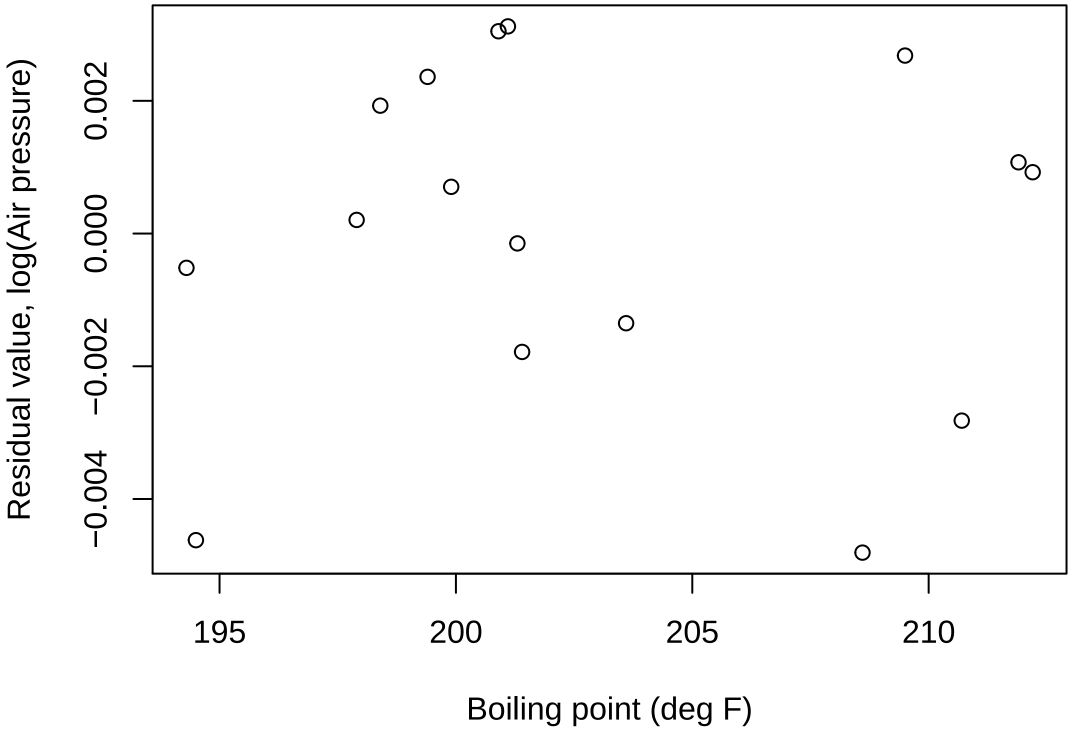
However, the 12th residual at boiling point 204.6 appears to be much larger than the others. It could be that this data point was recorded incorrectly, with a larger error than the others, or it could be a true data point and our model that assumes normally-distributed errors is incorrect. If we decide that the observation is the result of a measurement error, we coudl see if its residual value is an outlier using a boxplot and Tukey's criterion.

```
boxplot(model_2$residuals, horizontal=TRUE, xlab="Residual values")
```



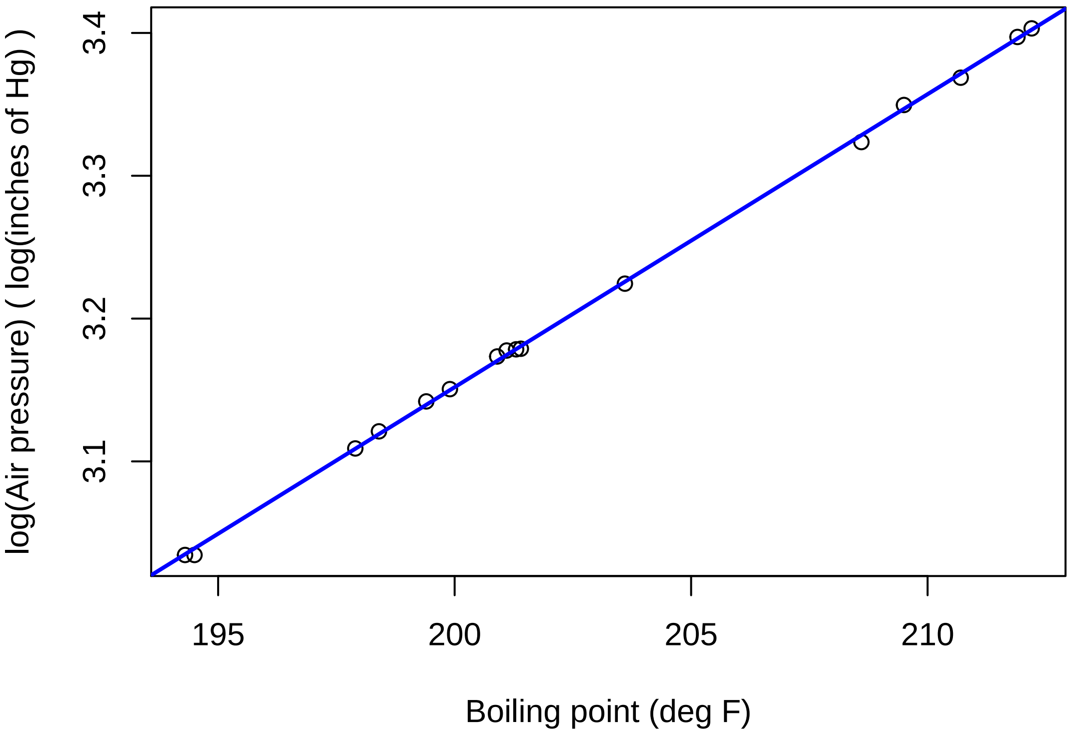
This point appears to be an outlier, and if we remove it and replot the residuals, the residual plot appears to show values centred around 0, which suggests that this model is a better fit to the data.

```
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
plot(x, residuals, xlab="Boiling point (deg F)", ylab="Residual value, log(Air pressure)")
```



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.

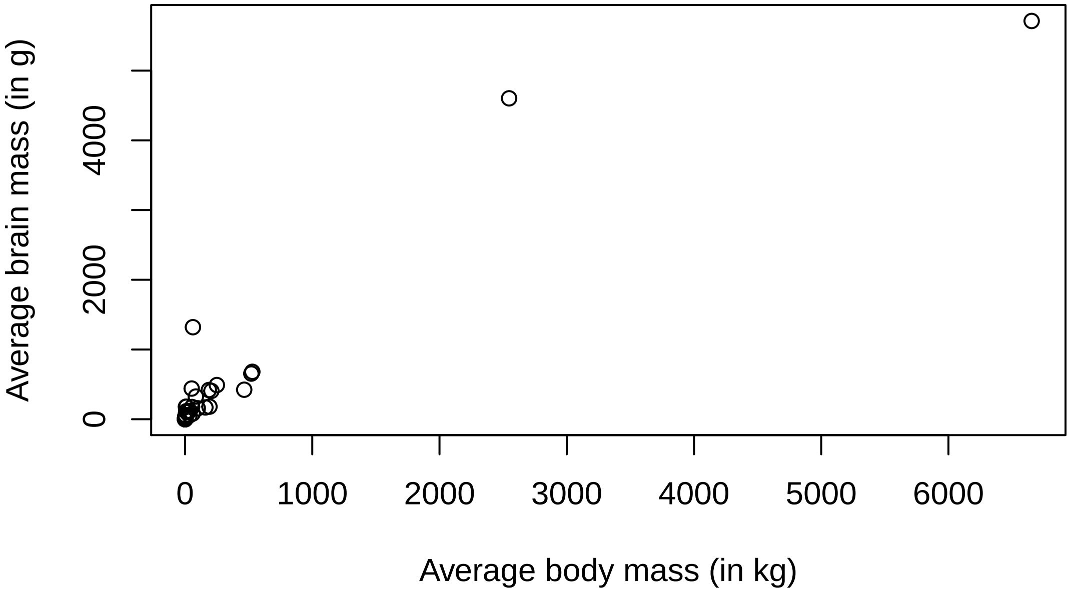
```r
library(MASS)
# the outlier is the 12th value, so remove this value to define the inliers
x <- forbes$bp[-12]
y <- forbes$pres[-12]
z <- log(y)

# 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)
```



5.3.5 Example: mammals data

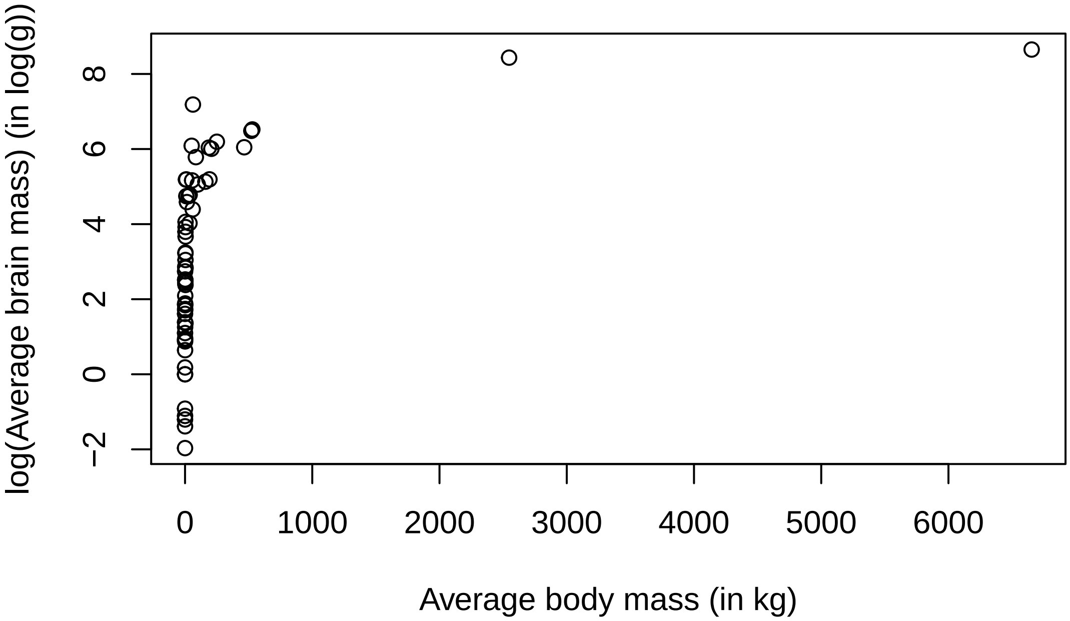
Let us look at another dataset from the `MASS` package. The `mammals` dataset gives the average body mass (in kg) and the average brain mass (in g) for 62 land mammals. Plotting the raw data we see:

```r
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)
```



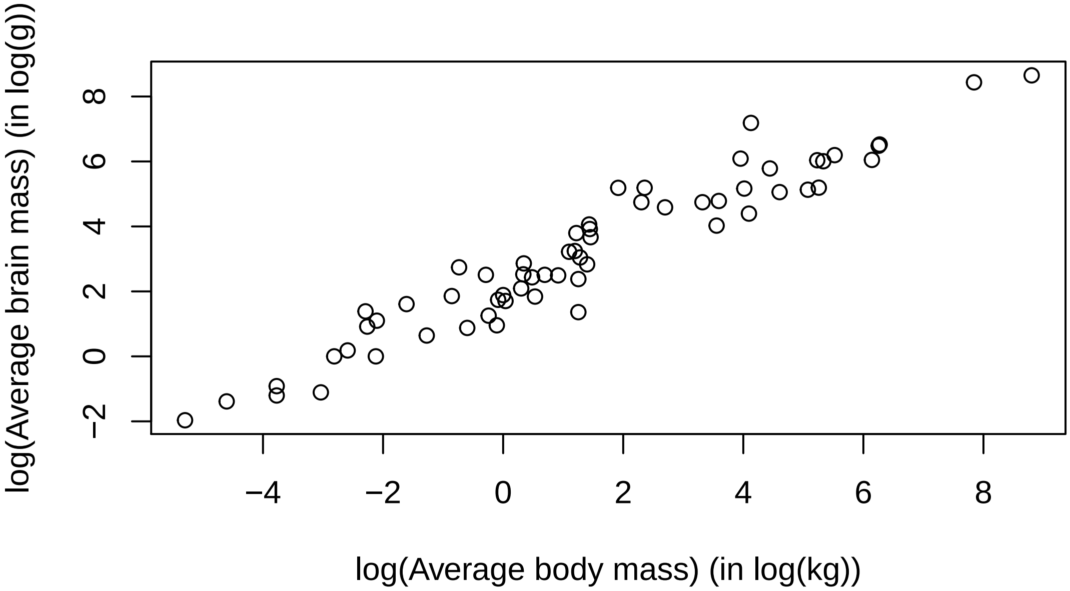
However, the plot seems to be distorted by two large values. If we take the logarithm of the brain masses:

```r
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)
```



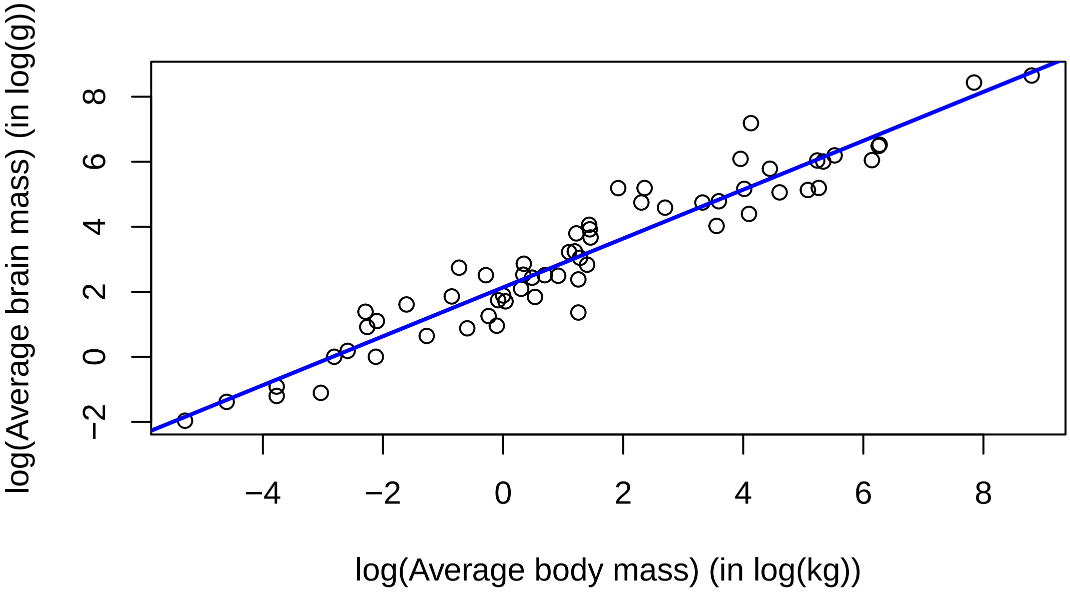
This is not much better. If take the logarithm of both the brain and body masses:

```r
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)
```



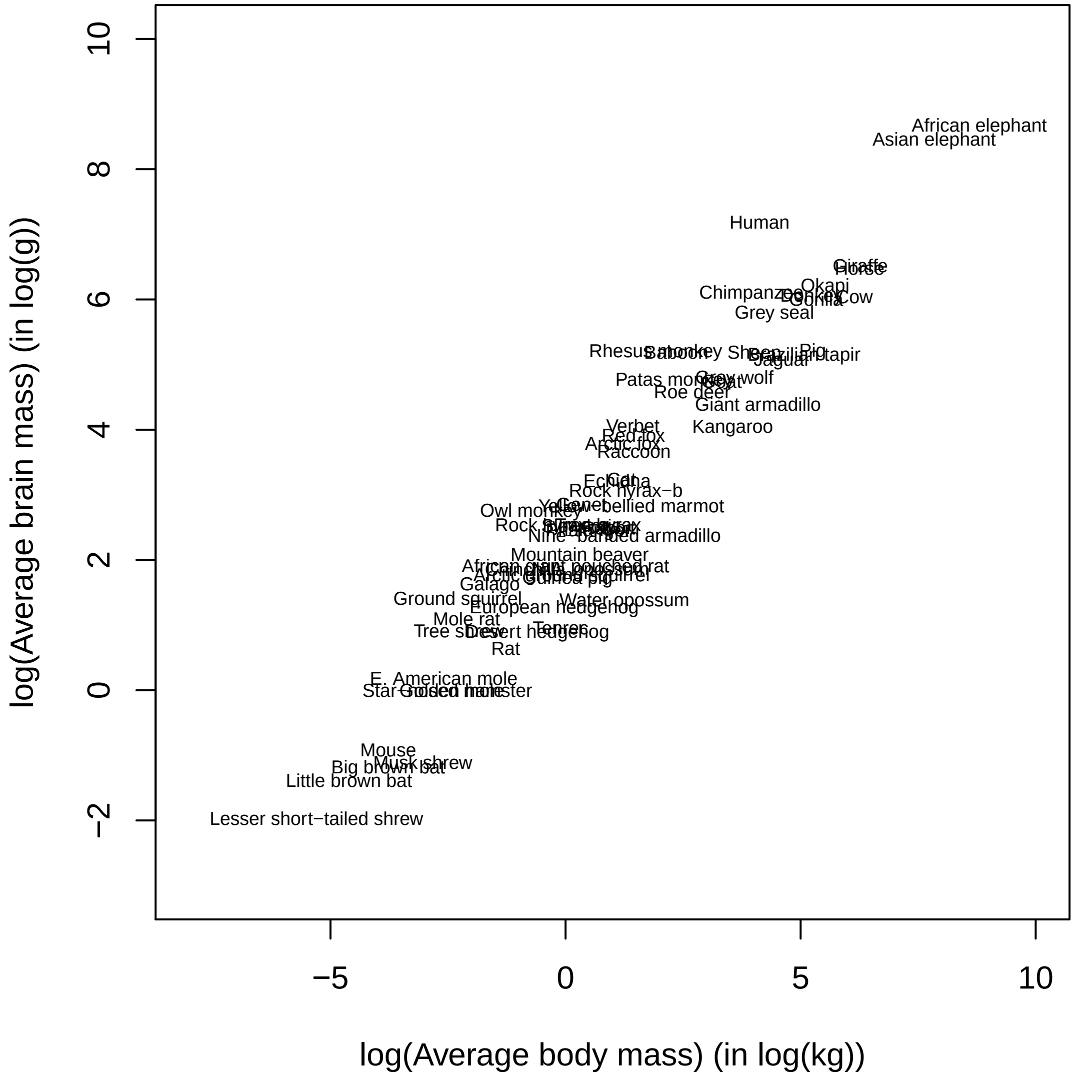
A clearer picture emerges, which suggests a linear relationship:

```r
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 mody 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.

```r
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
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
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.

5.3.6 The distribution of the parameters and the residuals

So far we have assumed that the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

has independent errors $\epsilon_i \mathrm{N}\left(0, \sigma^2\right)$, and have computed maximum likelihood estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2)$ for the parameters $(\beta_0, \beta_1, \sigma^2)$. However, these are point estimates and we would like to estimate the error in these estimates. To do so, we will need to determine the distribution that these estimates follow.

We first return to our computation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and rewrite these expressions. We first note that S_{xy} can be rewritten as

$$S_{xy} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i.$$

This then allows us to rewrite the maximum likelihood estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as

$$\begin{aligned}\widehat{\beta}_0 = \widehat{\beta}_0\left(\mathbf{x}, \mathbf{y}\right) &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\left(x_i - \overline{x}\right)\overline{x}}{S_{xx}}\right) y_i.\\ \widehat{\beta}_1 = \widehat{\beta}_1\left(\mathbf{x}, \mathbf{y}\right) &= \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i,\end{aligned}$$

This then allows us to write the maximum likelihood estimators for the pairs (x_i, Y_i) are

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_0 = \widehat{\beta}_0\left(\mathbf{x}, \mathbf{Y}\right) &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\left(x_i - \overline{x}\right)\overline{x}}{S_{xx}}\right) Y_i\\ \widehat{\boldsymbol{\beta}}_1 = \widehat{\beta}_1\left(\mathbf{x}, \mathbf{Y}\right) &= \frac{1}{S_{xx}} \sum_{i=1}^{n} \left(x_i - \overline{x}\right) Y_i\end{aligned}$$

Exercise 5.3.6. Show that one can rewrite S_{xy} as

$$S_{xy} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i.$$

$$\begin{aligned}S_{xy} &= \sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i - \overline{y} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)\\ &= \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i - \overline{y} \cdot 0\\ &= \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i\end{aligned}$$

△

Since the $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are linear combinations of the normal random variables $Y_1, Y_2, \ldots, Y_n$, they are themselves normal random variables. One can compute

$$\begin{aligned} \mathrm{E}(\widehat{\boldsymbol{\beta}}_0) = \beta_0, &\qquad \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}_0\right) = \left(\frac{1}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right) \sigma^2 \\ \mathrm{E}(\widehat{\boldsymbol{\beta}}_1) = \beta_1, &\qquad \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) \ = \left(\frac{1}{S_{xx}}\right) \sigma^2 \end{aligned}$$

to give the result

Proposition 5.3.7. The maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ are distributed according to the following normal distributions

$$\widehat{\boldsymbol{\beta}}_0 \sim \mathrm{N}\left(\beta_0, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right), \qquad \widehat{\boldsymbol{\beta}}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

♦

Since the Y_i are independent, they are uncorrelated, and so

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \mathrm{Cov}\left(Y_i, Y_i\right) = \mathrm{Var}(Y_i), & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

One can use this to prove

Proposition 5.3.8. The covariance between $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ is

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) = \left(-\frac{\overline{x}}{S_{xx}}\right) \sigma^2.$$

♦

Proof.

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) &= \mathrm{Cov}\left(\sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) Y_i, \frac{1}{S_{xx}} \sum_{j=1}^{n} (x_j - \overline{x})\, Y_j\right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) \frac{1}{S_{xx}} (x_j - \overline{x}) \,\mathrm{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) \frac{1}{S_{xx}} (x_i - \overline{x}) \,\mathrm{Var}(Y_i) \\ &= \left(\frac{1}{nS_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) - \frac{\overline{x}}{(S_{xx})^2} \sum_{i=1}^{n} (x_i - \overline{x})^2\right) \sigma^2 \\ &= \left(\frac{1}{nS_{xx}}(0) - \frac{\overline{x}}{(S_{xx})^2} S_{xx}\right) \sigma^2 = \left(-\frac{\overline{x}}{S_{xx}}\right) \sigma^2. \end{aligned}$$

□

Remark 5.3.9. In the proof we use the bilinearity of the covariance operator; for any random variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$, and constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots b_m$,

$$\operatorname{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{Cov}\left(X_i, Y_j\right).$$

□

In a similar manner to Proposition 5.3.8, one can the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ to prove

Proposition 5.3.10. The covariance between each Y_i and the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_0$ is

$$\operatorname{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) = \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)\sigma^2.$$

♦

Proposition 5.3.11. The covariance between each Y_i and the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_1$ is

$$\operatorname{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) = \left(\frac{x_i - \overline{x}}{S_{xx}}\right)\sigma^2.$$

♦

Having already defined the observed residuals for $i \in \{1, 2, \ldots, n\}$ as $\widehat{e}_i$, where

$$\widehat{e}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i),$$

one can now define

Definition 5.3.12. Given the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$, the residuals $\widehat{\epsilon}_i$ are defined as

$$\widehat{\epsilon}_i = Y_i - \left(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i\right).$$

The observed residuals $\widehat{e}_i$ are realisations of these random variables $\widehat{\epsilon}_i$. ■

Exercise 5.3.13. Show that one can rewrite $\widehat{\beta}_0$ as

$$\widehat{\beta}_0 = \widehat{\beta}_0\left(\mathbf{x}, \mathbf{y}\right) = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) Y_i.$$

$$\widehat{\beta}_0 = \widehat{\beta}_0\left(\mathbf{x}, \mathbf{y}\right) = \overline{y} - \widehat{\beta}_1 \overline{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \left(\frac{1}{S_{xx}}\sum_{i=1}^{n}\left(x_i - \overline{x}\right) y_i\right)\overline{x}$$

$$= \sum_{i=1}^{n}\frac{y_i}{n} - \left(\frac{1}{S_{xx}}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\overline{x}\right) y_i$$

$$= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) y_i$$

△

Remark 5.3.14. The residuals $\widehat{\epsilon}_i$ are random variables, since they are functions of the pairs (x_i, Y_i) and the maximum likeilhood estimators $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$. □

Recall that we obtained the maximum likelihood estimate for σ^2 as

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right]^2.$$

We can define the maximum likelihood estimator $\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}$ as

$$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \left(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i\right)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i^2.$$

We can also define

$$\widehat{\mathrm{RSS}}_{xY} = \mathrm{RSS}_{xY}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n}\left[Y_i - \left(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i\right)\right]^2 = \sum_{i=1}^{n}\widehat{\epsilon}_i^2.$$

One can easily show $\mathrm{E}\left(\widehat{\epsilon}_i\right) = 0$, however computing the variance requires more care. One can first show

$$\mathrm{Var}(\widehat{\epsilon}_i) = \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + x_i^2\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) + 2x_i\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1),$$

then subsituting in the expressions for each term one obtains

$$\mathrm{Var}(\widehat{\epsilon}_i) = \sigma^2\left[\frac{n-2}{n} + \frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i\overline{x} - 2\left(x_i - \overline{x}\right)^2\right)\right].$$

Noting that $\mathrm{E}\left(\widehat{\epsilon}_i\right) = 0$ implies $\mathrm{Var}(\widehat{\epsilon}_i) = \mathrm{E}\left(\widehat{\epsilon}_i^2\right) - \left(\mathrm{E}\left(\widehat{\epsilon}_i\right)\right)^2 = \mathrm{E}\left(\widehat{\epsilon}_i^2\right)$, one can compute

$$\mathrm{E}\left(\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\widehat{\epsilon}_i^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(\widehat{\epsilon}_i\right)$$

$$= \frac{1}{n}\sigma^2\left[\frac{n-2}{n} + \frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i\overline{x} - 2\left(x_i - \overline{x}\right)^2\right)\right]$$

$$= \left(\frac{n-2}{n}\right)\sigma^2.$$

This shows that the estimator $\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}$ is biased. However, we can show that the estimator $\frac{\widehat{\mathrm{RSS}}_{xY}}{n-2}$ is an unbiased estimator of σ^2, since

$$\mathrm{E}\left(\frac{\widehat{\mathrm{RSS}}_{xY}}{n-2}\right) = \mathrm{E}\left(\frac{n\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}}{n-2}\right) = \frac{n}{n-2}\mathrm{E}\left(\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}\right) = \sigma^2.$$

Exercise 5.3.15. Show that

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_0) = \beta_0.$$

$$\begin{aligned}
\mathrm{E}(\widehat{\boldsymbol{\beta}}_0) = \mathrm{E}\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)Y_i\right] &= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)\mathrm{E}[Y_i] \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)(\beta_0 + \beta_1 x_i) \\
&= \beta_0\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) + \beta_1\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)x_i \\
&= \beta_0
\end{aligned}$$

because

$$\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right) = \left(\sum_{i=1}^{n}\frac{1}{n} - \frac{\overline{x}}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})\right) = \left(n \cdot \frac{1}{n} - \frac{\overline{x}}{S_{xx}}(0)\right) = 1$$

and

$$\begin{aligned}
\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)x_i &= \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{\overline{x}}{S_{xx}}\sum_{i=1}^{n}\left(x_i^2 - \overline{x}x_i\right) \\
&= \overline{x} - \frac{\overline{x}}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right) \\
&= \overline{x} - \frac{\overline{x}}{S_{xx}}\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right) \\
&= \overline{x} - \frac{\overline{x}}{S_{xx}}(S_{xx}) \\
&= 0
\end{aligned}$$

△

Exercise 5.3.16. Show that

$$\operatorname{Var}\left(\widehat{\beta}_0\right) = \sigma^2 \left(\frac{1}{nS_{xx}} \sum_{i=1}^{n} x_i^2 \right)$$

$$\operatorname{Var}\left(\widehat{\beta}_0\right) = \operatorname{Var}\left(\sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} \right) Y_i \right) = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} \right)^2 \operatorname{Var}(Y_i)$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} \right)^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - \frac{2}{n} \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} + \frac{(x_i - \overline{x})^2\,\overline{x}^2}{(S_{xx})^2} \right)$$

$$= \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{n^2} - \frac{2\overline{x}}{nS_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) + \frac{\overline{x}^2}{(S_{xx})^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)$$

$$= \sigma^2 \left(n \cdot \frac{1}{n^2} - \frac{2\overline{x}}{nS_{xx}}(0) + \frac{\overline{x}^2}{(S_{xx})^2} S_{xx} \right)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)$$

This factor in front of σ^2 can be rewritten as

$$\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} = \frac{1}{nS_{xx}} \left(S_{xx} + n\overline{x}^2 \right)$$

$$= \frac{1}{nS_{xx}} \left(\sum_{i=1}^{n} (x_i - \overline{x})^2 + n\overline{x}^2 \right)$$

$$= \frac{1}{nS_{xx}} \left(\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right] + n\overline{x}^2 \right)$$

$$= \frac{1}{nS_{xx}} \sum_{i=1}^{n} x_i^2$$

This completes the exercise.

△

Exercise 5.3.17. Use the identity $\sum_{i=1}^{n} x_i\overline{x} = n\overline{x}^2$ to show that

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_1) = \beta_1$$

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_1) = \mathrm{E}(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i) = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,\mathrm{E}(Y_i)$$

$$= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,(\beta_0+\beta_1 x_i)$$

$$= \frac{1}{S_{xx}}\left(\beta_0\sum_{i=1}^{n}(x_i-\overline{x}) + \beta_1\sum_{i=1}^{n}\left(x_i^2-\overline{x}x_i\right)\right)$$

$$= \frac{1}{S_{xx}}\left(\beta_0(0) + \beta_1\left(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right)\right)$$

$$= \frac{1}{S_{xx}}\,(\beta_1 S_{xx}) = \beta_1$$

△

Exercise 5.3.18. Show that

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) = \mathrm{Var}\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i\right)$$

$$= \frac{1}{(S_{xx})^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\,\mathrm{Var}(Y_i) = \frac{1}{(S_{xx})^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\,\sigma^2$$

$$= \frac{\sigma^2}{(S_{xx})^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{\sigma^2}{(S_{xx})^2}S_{xx} = \frac{\sigma^2}{S_{xx}}$$

△

Exercise 5.3.19. Show that

$$\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) = \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)\sigma^2$$

$$\begin{aligned}
\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) &= \mathrm{Cov}\left(Y_i, \sum_{j=1}^{n}\left(\frac{1}{n} - \frac{(x_j - \overline{x})\,\overline{x}}{S_{xx}}\right)Y_j\right) \\
&= \sum_{j=1}^{n}\left(\frac{1}{n} - \frac{(x_j - \overline{x})\,\overline{x}}{S_{xx}}\right)\mathrm{Cov}\left(Y_i, Y_j\right) \\
&= \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)\mathrm{Var}\left(Y_i\right) \\
&= \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}}\right)\sigma^2
\end{aligned}$$

△

Exercise 5.3.20. Show that

$$\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) = \left(\frac{x_i - \overline{x}}{S_{xx}}\right)\sigma^2$$

$$\begin{aligned}
\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) &= \mathrm{Cov}\left(Y_i, \frac{1}{S_{xx}}\sum_{j=1}^{n}(x_j - \overline{x})\,Y_j\right) \\
&= \frac{1}{S_{xx}}\sum_{j=1}^{n}(x_j - \overline{x})\,\mathrm{Cov}\left(Y_i, Y_j\right) \\
&= \frac{1}{S_{xx}}\,(x_i - \overline{x})\,\mathrm{Var}\left(Y_i\right) \\
&= \left(\frac{x_i - \overline{x}}{S_{xx}}\right)\sigma^2
\end{aligned}$$

△

Exercise 5.3.21. Show that for $i \in \{1, 2, \ldots, n\}$,

$$\mathrm{E}\left(\widehat{\epsilon}_i\right) = 0.$$

$$\begin{aligned}
\mathrm{E}\left(\widehat{\epsilon}_i\right) &= \mathrm{E}\left(Y_i - \widehat{\boldsymbol{\beta}}_0 - \widehat{\boldsymbol{\beta}}_1 x_i\right) \\
&= \mathrm{E}\left(Y_i\right) - \mathrm{E}\left(\widehat{\boldsymbol{\beta}}_0\right) - x_i \mathrm{E}(\widehat{\boldsymbol{\beta}}_1) \\
&= (\alpha + \beta x_i) - \alpha - x_i \beta \\
&= 0.
\end{aligned}$$

$\triangle$

Exercise 5.3.22. Show that

$$\mathrm{Var}(\widehat{\epsilon}_i) = \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + x_i^2 \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) + 2x_i \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1)$$

$$\begin{aligned}
\mathrm{Var}(\widehat{\epsilon}_i) &= \mathrm{Var}(Y_i - (\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i)) \\
&= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i) \\
&= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_i) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) \\
&= \mathrm{Var}(Y_i) + \left[\mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1 x_i) + 2\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1 x_i)\right] - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) \\
&= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + x_i^2 \mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) + 2x_i \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i \mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1)
\end{aligned}$$

$\triangle$

Exercise 5.3.23. Show that

$$\mathrm{Var}(\widehat{\epsilon}_i) = \sigma^2 \left[\frac{n-2}{n} + \frac{1}{S_{xx}} \left(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i\overline{x} - 2\left(x_i - \overline{x}\right)^2 \right) \right]$$

$$\begin{aligned}
\mathrm{Var}(\widehat{\epsilon}_i) &= \mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{\boldsymbol{\beta}}_0) + x_i^2\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) + 2x_i\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\beta}}_1) - 2\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_0) - 2x_i\mathrm{Cov}(Y_i, \widehat{\boldsymbol{\beta}}_1) \\
&= \sigma^2 + \sigma^2 \left(\frac{1}{nS_{xx}} \sum_{j=1}^{n} x_j^2 \right) + x_i^2 \left(\frac{\sigma^2}{S_{xx}} \right) \\
&\qquad + 2x_i \left(-\frac{\overline{x}\sigma^2}{S_{xx}} \right) - 2\left(\frac{1}{n} - \frac{\left(x_i - \overline{x}\right)\overline{x}}{S_{xx}} \right)\sigma^2 - 2x_i\left(\frac{x_i - \overline{x}}{S_{xx}} \right)\sigma^2 \\
&= \sigma^2 \Bigg[1 - \frac{2}{n} + \frac{1}{nS_{xx}} \sum_{j=1}^{n} x_j^2 + \frac{x_i^2}{S_{xx}} \\
&\qquad + 2x_i \left(-\frac{\overline{x}}{S_{xx}} \right) + 2\frac{\left(x_i - \overline{x}\right)\overline{x}}{S_{xx}} - 2x_i\left(\frac{x_i - \overline{x}}{S_{xx}} \right) \Bigg] \\
&= \sigma^2 \left[\frac{n-2}{n} + \frac{1}{S_{xx}} \left(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i\overline{x} + 2\left(x_i - \overline{x}\right)\overline{x} - 2x_i\left(x_i - \overline{x}\right) \right) \right] \\
&= \sigma^2 \left[\frac{n-2}{n} + \frac{1}{S_{xx}} \left(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i\overline{x} + 2\left(x_i - \overline{x}\right)\left(\overline{x} - x_i\right) \right) \right] \\
&= \sigma^2 \left[\frac{n-2}{n} + \frac{1}{S_{xx}} \left(\frac{1}{n} \sum_{j=1}^{n} x_j^2 + x_i^2 - 2x_i\overline{x} - 2\left(x_i - \overline{x}\right)^2 \right) \right]
\end{aligned}$$

△

Exercise 5.3.24. Show that

$$\sum_{i=1}^{n}\left[\frac{n-2}{n}+\frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n}x_j^2+x_i^2-2x_i\overline{x}-2\left(x_i-\overline{x}\right)^2\right)\right]=n-2$$

$$\sum_{i=1}^{n}\left[\frac{n-2}{n}+\frac{1}{S_{xx}}\left(\frac{1}{n}\sum_{j=1}^{n}x_j^2+x_i^2-2x_i\overline{x}-2\left(x_i-\overline{x}\right)^2\right)\right]$$

$$=n\cdot\left\{\frac{n-2}{n}\right\}+\frac{1}{S_{xx}}\left(n\cdot\left\{\frac{1}{n}\sum_{j=1}^{n}x_j^2\right\}+\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\overline{x}-2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)$$

$$=(n-2)+\frac{1}{S_{xx}}\left(\sum_{j=1}^{n}x_j^2+\sum_{i=1}^{n}x_i^2-2n\overline{x}^2-2S_{xx}\right)$$

$$=(n-2)+\frac{2}{S_{xx}}\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2-S_{xx}\right)$$

$$=(n-2)+\frac{2}{S_{xx}}\left(S_{xx}-S_{xx}\right)$$

$$=n-2$$

△

5.4 The R^2 statistic (Reading material)

Recall from Section 5.3.1 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 5.2.1

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Definition 5.4.1. The statistic R^2 (pronounced '**R-squared**') is defined as

$$R^2 = \frac{S_{yy} - \widehat{\mathrm{RSS}}_{xy}}{S_{yy}}. \tag{5.10}$$

■

Remark 5.4.2. 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 5.4.4 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 4.2.7,

$$R^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = \frac{\left(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2\sum_{i=1}^{n}(y_i - \overline{y})^2} = r_{XY}^2.$$

Therefore, $R^2 \in [0, 1]$, since $r_{XY} \in [-1, 1]$ (although this was not explicitly shown, it was shown in Corollary 4.2.4 that $\rho_{XY} \in [-1, 1]$. To prove $r_{XY} \in [-1, 1]$ directly, one could use the Cauchy-Schwartz inequality).

Remark 5.4.3. Sometimes the R^2 statistic is quoted as evidence that a model fits the data well. This usually happends 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 5.4.4. 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.

△

5.5 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:

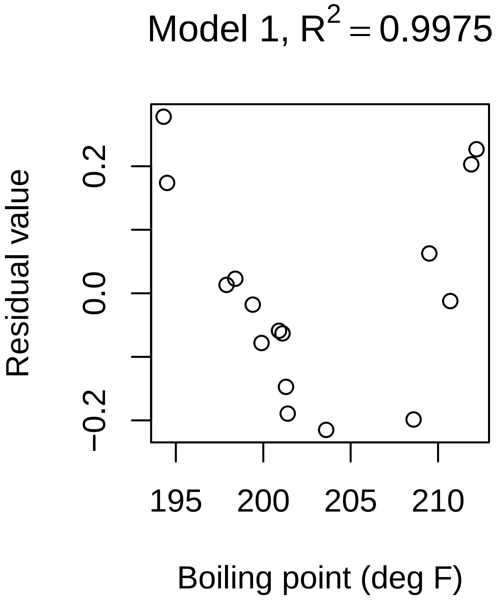
$$\begin{aligned} \text{Model 1:} \qquad Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \\ \text{Model 2:} \qquad \log(Y_i) &= \beta_0 + \beta_1 x_i + \epsilon_i, \end{aligned}$$

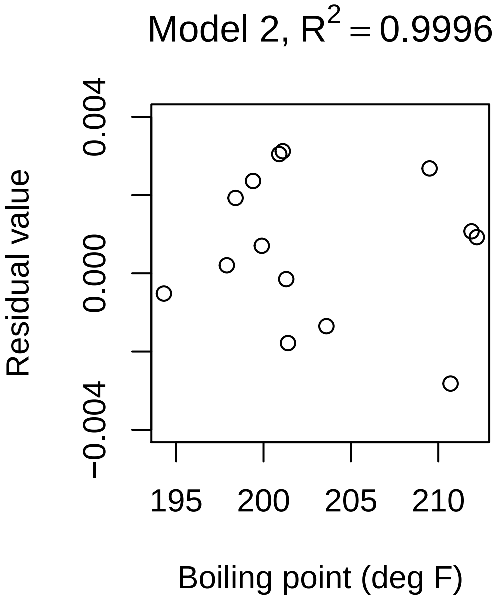
where $i \in \{1, 2, \dots, 12, 14, 15, 16, 17\}$.

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 6

Hypothesis testing

6.1 Introduction

We start with the definition of a statistical hypothesis:
Definition 6.1.1. 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 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 to accept as the truth. 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 as extreme as the observed data) can computed. This probability is called the **p-value**.

Once the p-value is computed, we need to make a decision on which hypothesis to accept as the truth, the null hypothesis or the alternative hypothesis. If the p-value, denoted by p, is very small (close to 0), this means that it was unlikely 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 small is too small? In order to make the decision, we compare the value of the p-value to a **significance threshold** (or significance level) α. If our p-value $p < α, we **reject** the null hypothesis at significance level α. We may then be inclined to believe that the alternative hypothesis is true.

On the other hand, if $p \nless \alpha$, our p-value is not extreme enough for us to decide to reject the null hypothesis. In this case, we **fail to reject** the null hypothesis.

Ronald Fisher was the statistician who invented hypothesis testing, and he described the null hypothesis as follows [5]:

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

Although this introduction has been somewhat abstract, the next section provides a concrete example of hypothesis testing.

6.2 The lady tasting tea

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

Example 6.2.1. 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 ‘mikl-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, i.e. can identify how a cup of tea is made, either milk-first or tea-first.

6.2.1 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 idetified as tea-first).

Exercise 6.2.2. 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 level $\alpha = 0.05$, or even at the level $\alpha = 0.02$. In this case, there is then evidence to suggest that the lady can discriminate between the two processes of making tea.

Exercise 6.2.3. 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 6.2.4. Incidentally, when Fisher conducted this experiment with his colleague Muriel Bristol, she correctly identified all eight cups. □

Before discussing Student's two-sample test in Section 6.4, we return to the topic of confidence intervals in the special case that the random variables are known to be normally distributed.

6.3 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.

6.3.1 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 a 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 1.6.2 that

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

Therefore,

$$Z = \frac{\theta - \overline{X}}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(0, 1\right). \tag{6.1}$$

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 6.3.1. Usually one would define $Z = \frac{\overline{X}-\theta}{\sigma/\sqrt{n}}$, however defining Z as in Equation (6.1) above allows the confidence interval bounds to be derived more smoothly. □

Example 6.3.2. 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 6.1 or Table 6.2 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

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

△

6.3.1.1 How to read Table 6.1

While Table 6.2 is straightforward to read. Table 6.1 requires a short explanation. In Table 6.1, 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:

```r
qnorm(0.975)
#> [1] 1.96
pnorm(1.96)
#> [1] 0.975
qnorm(0.95)
#> [1] 1.6449
pnorm(1.645)
#> [1] 0.95002
```

Table 6.1: $\mathrm{P}(Z < z)$ 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 6.2: 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

6.3.2 Case 2: normal distribution with variance unknown

Suppose again that the random variables $X_1, X_2, \ldots, X_n$ follow a normal distribution with unknown mean θ, but now suppose that the variance σ^2 **is also unknown**. Our goal is again to obtain a confidence interval for the unknown mean θ. One idea would be to use the sample variance

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

instead of the variance σ^2 and define

$$T = \frac{\overline{X} - \theta}{S/\sqrt{n}}. \tag{6.2}$$

In order to use the sample variance, we must assume $n \geq 2$. However, this quantity T does not follow a normal distribution. It follows a disribution known as Student's t-distribution with $n-1$ degrees of freedom. The probability density function of this distribution is derived in Section A.4 of the appendix to be

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

where m is the degrees of freedom parameter. Table 6.3 gives values for the cumulative distribution function for this distribution for a range of degrees of freedom. One notes that $f_T(-t) = f_T(t)$, i.e. the probability density function is symmetric. Therefore if we let t_α denote the value such that $\mathrm{P}\left(T < t_\alpha\right) = \alpha$, we have $-t_\alpha = t_{1-\alpha}$.

Example 6.3.3. Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables following a normal distribution with unknown mean θ and unknown variance σ^2. Suppose we observe $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ as $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Given that the sample mean is $\overline{y} = 7$, the sample variance is $s^2 = 16$ and $n = 9$, let us construct a 95% confidence interval for θ.

First, we calculate the degrees of freedom. Since $n = 9$, the degrees of freedom is $n - 1 = 8$. Next, for a 95% confidence interval, $\alpha = 0.05$ and so $1 - \alpha/2 = 0.975$. So, we wish to find $t_{0.975}$ such that $\mathrm{P}\left(T < t_{0.975}\right) = 0.975$. Looking at the table in the row for degrees of freedom equal to 8, we see that $\mathrm{P}\left(T < 2.306 = 0.975\right)$. We use symmetry to conclude $\mathrm{P}\left(T < -2.306 = 0.025\right)$. Then,

$$\begin{aligned}
&\mathrm{P}\left(t_{0.025} < \frac{\overline{X} - \theta}{S/\sqrt{n}} < t_{0.975}\right) = 0.95 \\
\Rightarrow &\mathrm{P}\left(-t_{0.025} > \frac{\theta - \overline{X}}{S/\sqrt{n}} > -t_{0.975}\right) = 0.95 \\
\Rightarrow &\mathrm{P}\left(-t_{0.975} < \frac{\theta - \overline{X}}{S/\sqrt{n}} < -t_{0.025}\right) = 0.95 \\
\Rightarrow &\mathrm{P}\left(t_{0.025} < \frac{\theta - \overline{X}}{S/\sqrt{n}} < t_{0.975}\right) = 0.95 \\
\Rightarrow &\mathrm{P}\left(t_{0.025}\frac{S}{\sqrt{n}} < \theta - \overline{X} < t_{0.975}\frac{S}{\sqrt{n}}\right) = 0.95 \\
\Rightarrow &\mathrm{P}\left(\overline{X} + t_{0.025} \cdot \frac{S}{\sqrt{n}} < \theta < \overline{X} + t_{0.975} \cdot \frac{S}{\sqrt{n}}\right) = 0.95
\end{aligned}$$

And so, using the values for the observed sample, a 95% confidence interval for the unknown mean θ is

$$\left(7 - 2.306\frac{4}{3}, 7 + 2.306\frac{4}{3}\right).$$

△

Table 6.3: 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	0.87	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

6.4 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 $\mathrm{N}\left(\theta_1, \sigma_1^2\right)$ distribution where θ_1 is unknown and σ_1^2 is known, and the independent random variables $Y_1, Y_2, \ldots, Y_m$ follow a $\mathrm{N}\left(\theta_2, \sigma_2^2\right)$ distribution where θ_2 is unknown and σ_2^2 is known. (We also assume that each X_i is independent of eahc Y_j.) Suppose further that $\mathbf{X} = (X_1, X_2, X_n)$ 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 that $\sigma_1 = \sigma_2$, we can obtain an exact answer.

If we write $\sigma_1 = \sigma_2 = \sigma$, and define

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

where

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

it can be shown that $T \sim t_{n+m-2}$, i.e. Student's t-distribution with degrees of freedom equal to $n + m - 2$.

Let us summarise the assumptions:

- The two groups of random variables are independent and each follow a normal distribution.
- The means of the two samples are unknown
- The variances of the two samples 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 to be

$$H_0 : \theta_1 = \theta_2.$$

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 we can compute a p-value. If this p-value is below a desired significance level α, we would reject the null hypothesis at level α.

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.

Remark 6.4.1. 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 7

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, 4].

7.1 Prior and posterior distributions

We first recall the definitions for the likelihood function and the marginal distribution.

Definition 7.1.1. 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 7.1.2. 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*

With these definitions in mind we are ready to define the concepts of prior and posterior distributions.

Definition 7.1.3. 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 7.1.4. 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. Origins: Latin, *prior*, meaning *previous*, *earlier*, *preceding*, *former* Source: Oxford English Dictionary	Coming **after** in time or order; later. Origins: Latin, *post* means *after*

These definitions lead to the following important result, which is also referred to as Bayes' Theorem.
Theorem 7.1.5. 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 pmf or pdf for θ is $\pi(\theta)$. Then the posterior p.d.f. (p.m.f.) is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}, \tag{7.1}$$

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 priorp.m.f..

Multiplying the conditional joint pdf or pmf $f(\mathbf{x}|\theta)$ by the prior pdf $\pi(\theta)$ results in the $(n+1)$-dimensional joint p.d.f. (p.m.f.) of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and θ,

$$f\left(\mathbf{x}, \theta\right) = f(\mathbf{x}|\theta)\pi(\theta). \tag{7.2}$$

The marginal joint p.d.f. (p.m.f.) $m(\mathbf{x})$ of $\mathbf{X}$ can then be obtained by integrating the right-hand side of Equation (7.2) over all values of $\theta \in \Theta$,

$$m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta. \tag{7.3}$$

Then, the conditional p.d.f. of θ given that $\mathbf{X} = \mathbf{x}$, denoted $\pi(\theta|\mathbf{x})$, must then be

$$\pi(\theta|\mathbf{x}) = \frac{f\left(\mathbf{x}, \theta\right)}{m(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

This is Bayes's theorem restated for parameters and random samples. Although we assumed for Equation (7.3) 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 7.3: List of symbols in Equation (7.1) of Theorem 7.1.5.

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 $\mathbf{x}$.

Remark 7.1.6. 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 3.1.1 and 7.1.1 and Equation (3.1). Note however, that $f(\mathbf{x}|\theta)$ is sometimes referred to as the **sampling distribution** [2]. The use of three separate names for the same quantity can be quite confusing; from now on we shall refer to it as the likelihood function. □

Example 7.1.7. 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 $i = 1, 2, \dots, n$ let the random variable $X_i = 1$ if the ith lightbulb is defective and let $X_i = 0$ otherwise. Then the independent random variables $X_1, X_2, \dots, X_n$ form n Bernoulli trials with parameter θ, and so the pmf 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 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ can be written, for $x_1, x_2, \dots, x_n \in \{0, 1\}$, 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 $0 < \theta < 1$,

$$f(\mathbf{x}|\theta)\pi(\theta) = \theta^{\overline{x}}(1-\theta)^{n-\overline{x}}. \tag{7.4}$$

One could compute the marginal distribution $m(\mathbf{x})$ as in Equation (7.3) in order to obtain the posterior p.d.f. $\pi(\theta|\mathbf{x})$ directly. However, one notices from Equation (7.1) that

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta), \tag{7.5}$$

and so another way to arrive at the solution would be to compare Equation (7.4) 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.

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{7.6}$$

One notices that $f(\mathbf{x}|\theta)\pi(\theta)$ in Equation (7.4) 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 7.1.8. Note how we used Equation (7.5) 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 (7.5) 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 7.1.9. Observe in Example 7.1.7 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 7.1.10. 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:

$$\begin{aligned}\pi(\theta|\mathbf{x}) \propto \pi(\theta) f(\mathbf{x}|\theta) &= \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)\exp\left(-\frac{1}{2\tau^2}(x-\theta)^2\right)\\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta - \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2\right) \qquad \text{(Exercise 7.1.11)}\,.\end{aligned}$$

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 7.1.11. Show that the posterior distribution in Example 7.1.10 is

$$\pi(\theta|\mathbf{x}) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta - \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2\right).$$

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto \pi(\theta) f(\mathbf{x}|\theta)\\ &= \exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right)\exp\left(-\frac{1}{2\tau^2}(x-\theta)^2\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\theta^2 - 2\theta\mu + \mu^2\right) - \frac{1}{2\tau^2}\left(x^2 - 2x\theta + \theta^2\right)\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\theta^2 - 2\theta\mu\right) - \frac{1}{2\tau^2}\left(x^2 - 2x\theta\right)\right)\exp\left(-\frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\tau^2}\right)\\ &= \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\theta^2 - 2\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\theta\right]\right)\exp\left(-\frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\tau^2}\right)\\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta^2 - 2\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\theta\right]\right)\\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\theta^2 - \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2}\right)\right]^2\right).\end{aligned}$$

△

Remark 7.1.12. Example 7.1.10 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 7.1.10 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 7.1.13. 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 7.1.14. 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

7.2 Conjugate prior distributions

Rather than starting with the definition of what a conjugate prior is, let us start with an example.

Example 7.2.1. Let us reconsider Example 7.1.7, 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 pmf 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 **hyperparameters** $\alpha, \beta > 0$. This distribution has p.d.f. given by Equation (7.6). Then the joint distribution of $\mathbf{x}$ and θ is

$$\begin{aligned} f\left(\mathbf{x}, \theta\right) = f(\mathbf{x}|\theta)\pi(\theta) &= \left[\binom{n}{y} \theta^y (1-\theta)^{n-y}\right] \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} \left(1-\theta\right)^{\beta-1}\right] \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1}. \end{aligned}$$

To compute the marginal distribution, recall that the p.d.f. of a $\text{Beta}\,(\gamma, \delta)$ distribution must integrate to 1:

$$\begin{aligned} \int_0^1 \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \theta^{\gamma-1} \left(1-\theta\right)^{\delta-1} \mathrm{d}\theta &= 1 \\ \Rightarrow \int_0^1 \theta^{\gamma-1} \left(1-\theta\right)^{\delta-1} \mathrm{d}\theta &= \frac{\Gamma(\gamma)\Gamma(\delta)}{\Gamma(\gamma+\delta)}. \end{aligned}$$

Therefore the marginal distribution $m(\mathbf{x})$ is

$$m(\mathbf{x}) = \int_0^1 f\left(\mathbf{x}, \theta\right) \mathrm{d}\theta = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

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 7.2.2. 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 7.2.3. 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 7.2.4. From Example 7.2.1, 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 7.2.5. In Example 7.1.7, the likelihood was Bernoulli (or Binomial), the prior was $\text{Unif}[0,1]$ and the posterior was a beta distribution. This is in fact a special case of the conjugacy illustrated in Example 7.2.1, because the $\text{Unif}[0,1]$ distribution is simply the $\text{Beta}\,(1,1)$ distribution. □

Example 7.2.6. From Example 7.1.10, 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 7.2.7. 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 $x_i > 0$,

$$f(x_i|\theta) = \theta \exp\left(-\theta x_i\right).$$

Therefore the likelihood is, for $x_i > 0$ for all $i = (1, 2, \ldots, n)$, and writing $y = \sum_{i=1}^n x_i$,

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \left(\theta \exp\left(-\theta x_i\right)\right) = \theta^n \exp\left(-\theta y\right).$$

Suppose the prior for θ is a $\Gamma\left(\alpha, \beta\right)$. Then (using the shape-rate parametrisation):

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp\left(-\beta\theta\right)$$

Then the posterior p.d.f. is proportional to:

$$\begin{aligned}\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta) &= \theta^n \exp\left(-\theta y\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp\left(-\beta\theta\right) \\ &\propto \theta^{n+\alpha-1} \exp\left(-\theta\left(y+\beta\right)\right)\end{aligned}$$

which shows that the posterior is a $\Gamma\left(n+\alpha, y+\beta\right)$ distribution, i.e.

$$\pi(\theta|\mathbf{x}) = \frac{(y+\beta)^{n+\alpha}}{\Gamma(n+\alpha)} \theta^{n+\alpha-1} \exp\left(-\theta\left(y+\beta\right)\right).$$

△

7.3 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 7.3.1. Let us return to Example 7.1.7 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 $\mathrm{N}\left(\mu, \sigma^2\right)$ 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. △

This is because the normalising constant will not have any θ terms.

Example 7.3.2. 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. △

In fact, it is not unusual in Bayesian inference to end up with a situation Example 7.3.1 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.

7.4 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 7.4.1. 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 7.2.7 shows that posterior is a $\Gamma(n+\alpha, y+\beta)$ distribution with p.d.f. (using the shape-rate parametrisation):

$$\pi(\theta|\mathbf{x}) = \frac{(y+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}\theta^{n+\alpha-1}\exp\left(-\theta\left(y+\beta\right)\right), \tag{7.7}$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 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,

$$\mathrm{E}\left(X\right) = \frac{\alpha}{\beta}, \qquad \mathrm{Var}\left(X\right) = \frac{\alpha}{\beta^2}.$$

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 (7.7), 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 7.1.

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. △

```r
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)
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
par(mfrow=c(1,2), cex=0.75) # make 1 row of 2 plots, cex controls size
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)
```

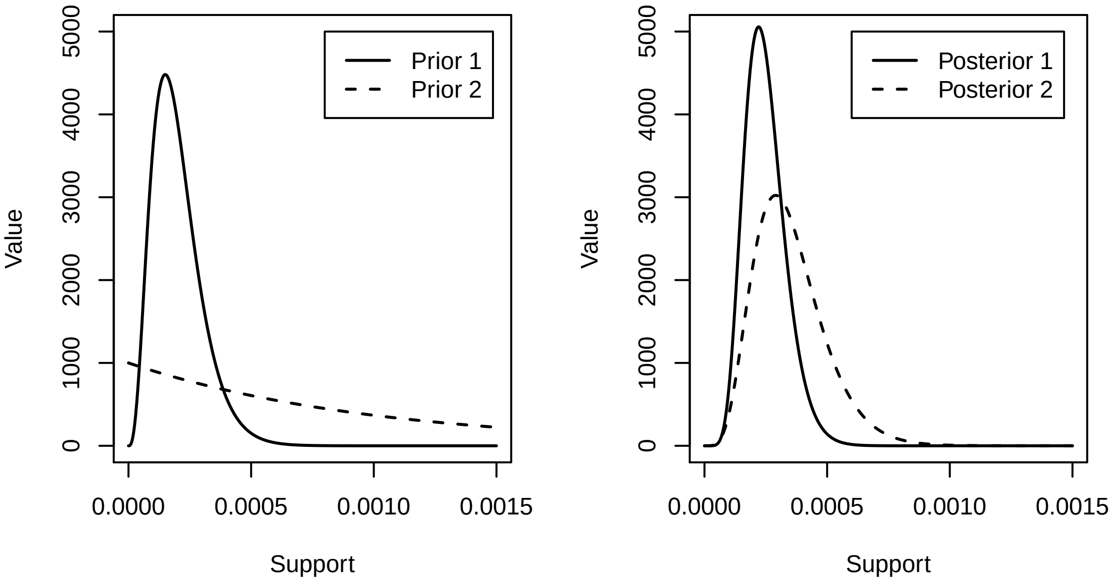


Figure 7.1: The two priors and the resulting posteriors from Example 7.4.1.

7.5 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 (7.1) in Theorem 7.1.5 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 7.1.7, 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 7.3.1 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 7.1.10, 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 7.3.2, 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 7.2 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 7.5 provides a list of the conjugate likelihoods that we have seen.

Table 7.5: A list of conjugate priors for a few well-known distributions.

Likelihood	Conjugate prior	Derivation
Bernoulli distribution	Beta distribution	Example 7.2.1
Binomial distribution	Beta distribution	Similar to Example 7.2.1
Normal distribution	Normal distribution	Example 7.1.10
Exponential distribution	Gamma distribution	Exercise 7.2.7

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 7.4.1 and is shown in Figure 7.1. 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 7.4.1 and Figure 7.1.

Remark 7.5.1. 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 8

Pitfalls in Statistics

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

8.1 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 8.1: The four datasets which make up Anscombe's quartet.

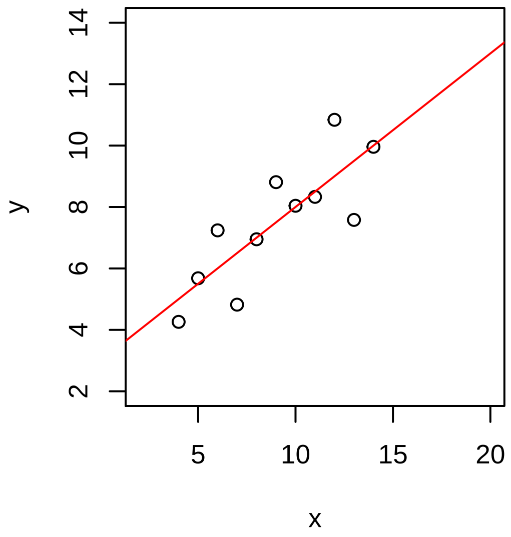
Interestingly, several summary statistics of these datasets are equal up to two decimal places:

	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

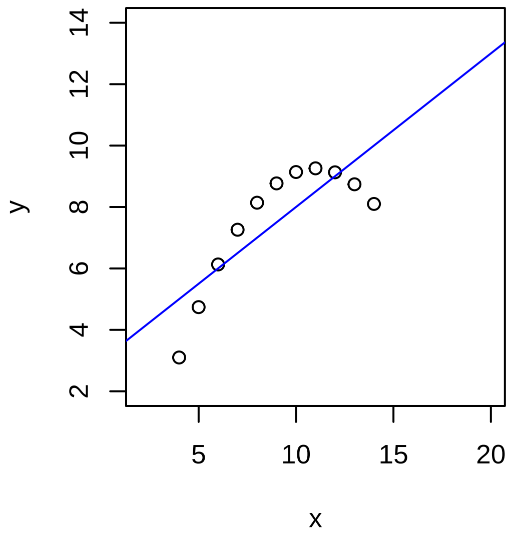
Table 8.2: Summary statistics for Anscombe's quartet.

However, if we plot the data, we see that these datasets are very different in character. This example serves as a warning to be careful of interpreting summary statistics in isolation, and reminds us of the need to create a visualisation of the data whenever possible.

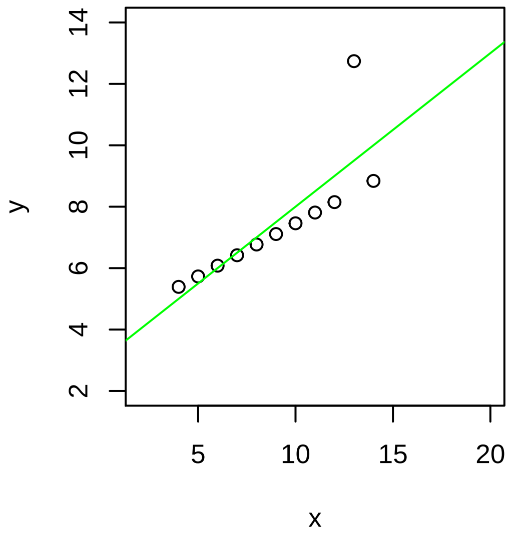
Dataset 1



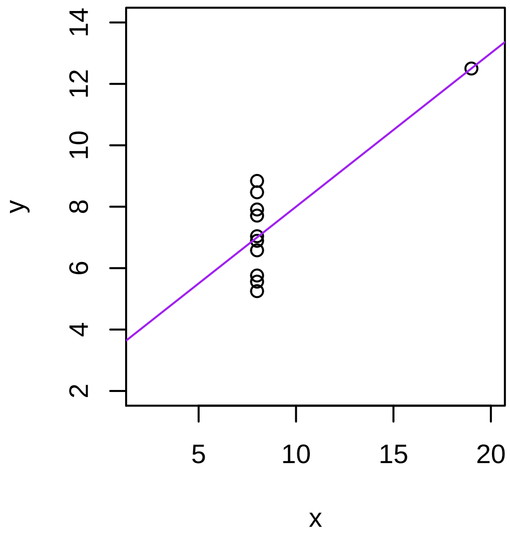
Dataset 2



Dataset 3



Dataset 4



8.2 Correction for multiple hypothesis testing

In hypothesis testing, it is often the case that 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 level α, for example, $\alpha = 0.05$ or $\alpha = 0.01$ are common choices, and if our compute 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 8.2.1. 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 level 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,

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

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

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

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 $p_i \sim U(0,1)$, $\mathrm{P}\left(A_i\right) = \alpha$.

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,

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

Therefore, in the example above, the probability of at least one p-value being significant is $10 \cdot 0.05 = 0.5$, in other words, 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 level is α, then one should use the adjusted significance level $\alpha' = \alpha/n$. 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 8.2.2. Returning to Example 8.2.1, if the desired signifance level is $\alpha = 0.05$, and there are $n = 10$ tests, then the adjusted significance level 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, $p_3 = 0.003$. △

Appendix A

Additional notes from lectures

A.1 Lecture 6, Monday 27 January 2020, Additional Examples

These additional examples were discussed in class in order to illustrate the importance of Theorem 1.5.23. If we defined new notation, $\mathrm{MSE}(\widehat{\Theta}, \theta)$, as the mean-squared error of the estimator $\widehat{\Theta}$ and the parameter θ, then

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

When choosing an estimator $\widehat{\Theta}$ for a parameter θ, ideally we would like the estimator to

- be unbiased,
- have small $\mathrm{MSE}(\widehat{\Theta}, \theta)$

Example A.1.1. Suppose the random variables $X_1, X_2, \dots X_n$ are independent with $\mathrm{E}\left(X_i\right) = \theta$, $\mathrm{Var}\left(X_i\right) = \sigma^2$, for $i \in \{1, 2, \dots, n\}$, where the mean θ is unknown.

One choice for an estimator for θ is the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, Proposition 1.2.6 gives

$$\begin{aligned} \mathrm{E}\left(\overline{X}\right) &= \theta \\ \Rightarrow b_\theta(\overline{X}) &= \mathrm{E}\left(\overline{X}\right) - \theta = 0 \end{aligned}$$

Therefore $\overline{X}$ is an unbiased estimator of the mean.

In order compute its mean-squared error, we again use Proposition 1.2.6 to obtain $\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$ and then Theorem 1.5.23 gives us

$$\mathrm{MSE}(\widehat{\Theta}, \theta) = \left[b_\theta(\widehat{\Theta})\right]^2 + \mathrm{Var}\left(\widehat{\Theta}\right) = (0)^2 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n}.$$

△

Example A.1.2. Again, consider the collection of independentrandom variables $X_1, X_2, \dots X_n$, with $\mathrm{E}\left(X_i\right) = \theta$, $\mathrm{Var}\left(X_i\right) = \sigma^2$, for $i \in \{1, 2, \dots, n\}$, where the mean θ is unknown.

Now, let's take X_1 as our estimator for the mean θ. We can compute $\mathrm{E}\left(X_1\right) = \theta \Rightarrow b_\theta(X_1) = 0$, so it is unbiased. Since $\mathrm{Var}\left(X_1\right) = \sigma^2$, together with the fact it is unbiased shows that $\mathrm{MSE}(X_1, \theta) = \sigma^2$. Therefore, as an estimator X_1 is unbiased, but it has a larger MSE than the estimator $\overline{X}$. △

Remark A.1.3. We might wonder if it is possible to improve on an estimator; for example, how do we know if $\overline{X}$ is the 'best' estimator for the mean? What do we mean by 'best'? We have stated that ideally our estimator will be unbiased and have small mean-squared error, but sometimes there is a tradeoff between the two, as the next example shows. □

Example A.1.4. Consider the collection of independent random variables $X_1, X_2, \dots X_n$, with $\mathrm{E}\left(X_i\right) = \mu$, $\mathrm{Var}\left(X_i\right) = \sigma^2$, for $i \in \{1, 2, \dots, n\}$, where the mean μ and variance σ^2 are both unknown. However, in addition, assume that the variables are normally distributed, i.e. $X_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$ (but, μ and σ^2 are unknown).

Suppose we wish to estimate σ^2. The sample variance

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

is a natural choice for an estimator, since from Proposition 1.2.6,

$$\mathrm{E}\left(S^2\right) = \sigma^2 \qquad \Rightarrow \qquad b_{\sigma^2}(S^2) = 0$$

It is an exercise in the PBL Sheet 10 to show that (with the normality assumption)

$$\mathrm{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}.$$

Therefore,

$$\mathrm{MSE}(S^2, \sigma^2) = b_{S^2}(\sigma^2) + \mathrm{Var}\left(S^2\right) = (0)^2 + \mathrm{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}.$$

△

Remark A.1.5. Considering the expression $\mathrm{MSE}(\widehat{\Theta}, \theta) = \left[b_\theta(\widehat{\Theta})\right]^2 + \mathrm{Var}\left(\widehat{\Theta}\right)$, if we insist on only using unbiased estimators, then the only way to improve the MSE is to find an (unbiased) estimator with a smaller variance. If on the other hand, we place more importance on minimising the MSE, and are willing to consider biased estimators, it may be possible to minimise the MSE further. □

Example A.1.6. As in Example A.1.4, assume that the random variables $X_1, X_2, \dots, X_n$ are independent and are normally distributed with mean μ and variance σ^2, which are both unknown. Again, suppose we wish to estimate the variance. Now, define the estimator S_b^2 as

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

Now, using properties of the expectation and variance, we can compute

$$\begin{aligned}
\mathrm{E}\left(S_b^2\right) &= \mathrm{E}\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n} \mathrm{E}\left(S^2\right) = \frac{n-1}{n}\sigma^2 \\
\mathrm{Var}\left(S_b^2\right) &= \mathrm{Var}\left(\frac{n-1}{n} S^2\right) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}\left(S^2\right) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}
\end{aligned}$$

Therefore, the bias is

$$b_{\sigma^2}(S_b^2) = \mathrm{E}\left(S_b^2\right) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n},$$

and we can compute the MSE as

$$\mathrm{MSE}(S_b^2, \sigma^2) = b_{S_b^2}(\sigma^2) + \mathrm{Var}\left(S_b^2\right) = \left(\frac{-\sigma^2}{n}\right)^2 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

Now, since $n > 0$, we get the sequence of inequalities

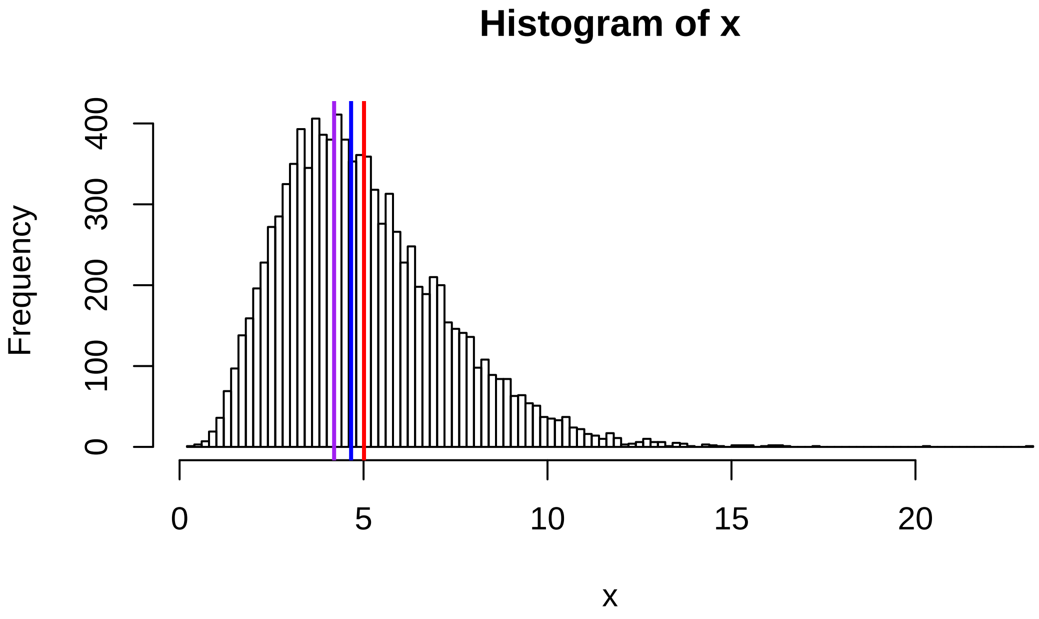
$$\begin{aligned}
&\mathrm{MSE}(S_b^2, \sigma^2) = \frac{(2n-1)\sigma^4}{n^2} < \frac{(2n-1+1)\sigma^4}{n^2} = \frac{2\sigma^4}{n} < \frac{2\sigma^4}{n-1} = \mathrm{MSE}(S^2, \sigma^2) \\
&\Rightarrow \mathrm{MSE}(S_b^2, \sigma^2) < \mathrm{MSE}(S^2, \sigma^2)
\end{aligned}$$

and so the estimator S_b^2 has a MSE that is smaller than S^2. △

A.2 R experiment: visualising the mean, median and mode

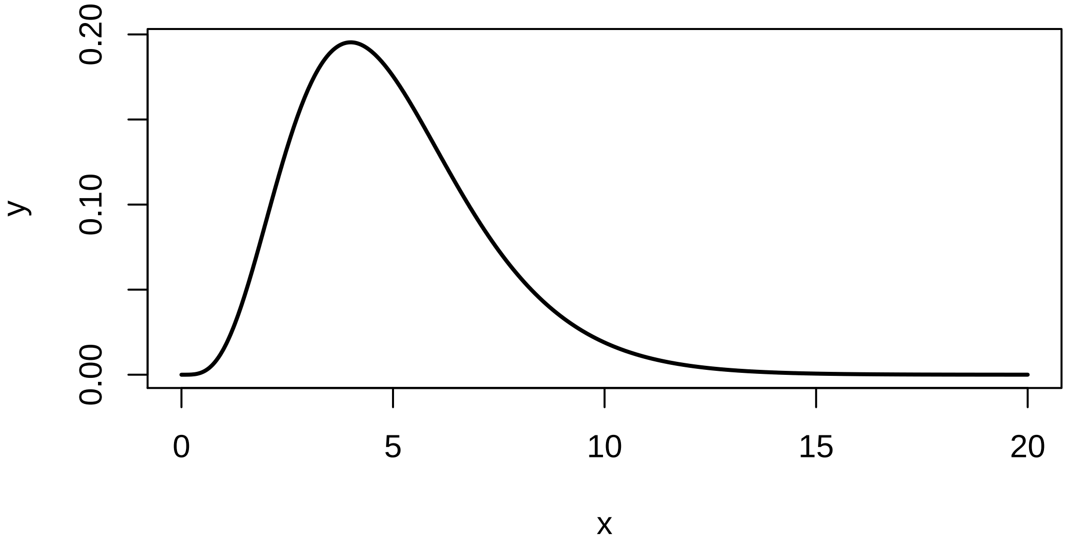
The following code shows a simulation to find the sample mean, median and mode for a $\Gamma(5, 1)$ distribution.

```
# set shape parameter (k) and scale parameter (theta) for Gamma distribution
k <- 5
theta <- 1
# setting the number of trials
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)

# 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
empirical_mode_index <- which.max(h$counts)

# now find the value of the break with the largest count, i.e. the (empricial) mode
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)

# the true mode can be obtained in terms of parameter k and parameter theta (exercise)
true_mode <- (k-1)*theta
cat("(empirical mode, true mode): (", empirical_mode, ", ", true_mode, ")\n", sep="")
#> (empirical mode, true mode): (4.2, 4)
```



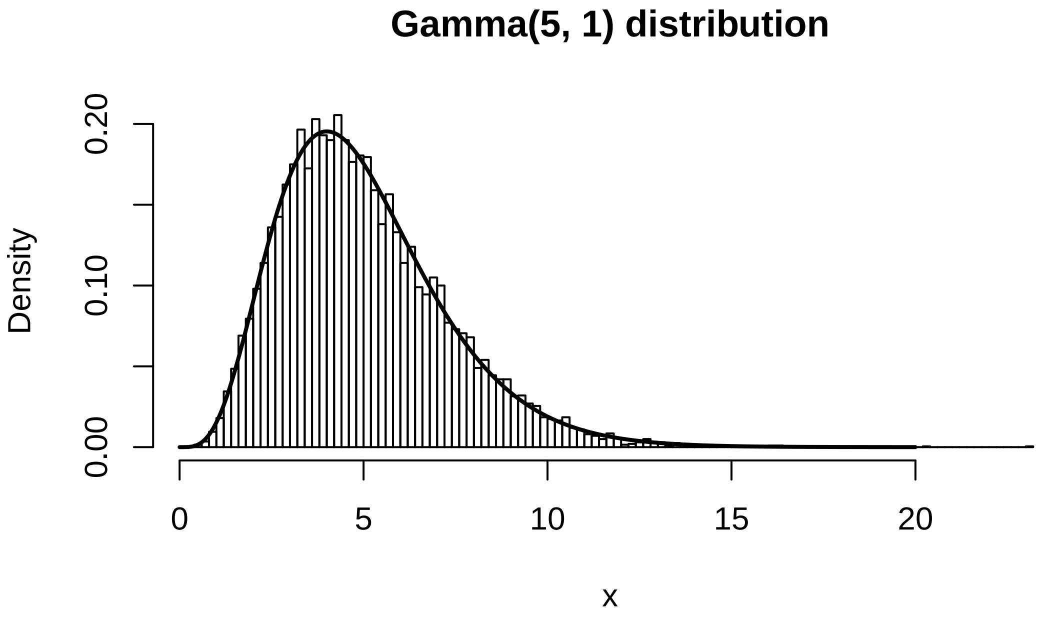
It is also possible to plot the probability density function using `dgamma`:

```
# set shape parameter (k) and scale parameter (theta) for Gamma distribution
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 desnity function. However, in this case, we need to make the histogram display densities, rather than counts (densities are the normalised counts).

```r
# set shape parameter (k) and scale parameter (theta) for Gamma distribution
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.3 Lecture 9, Friday 7 February 2020

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:

```r
df <- read.table("shoesize.txt", header=TRUE, sep=",")
shoesize <- df$shoe.size
height <- df$height
print( cor(shoesize, height) )
```

A.4 Student's t-distribution

If $U \sim \mathrm{N}(0,1)$, and $V \sim \chi^2_m$, then

$$\frac{U}{\sqrt{V/m}} \sim t_m,$$

where t_m is called Student's t-distribution with m degrees of freedom and has probability density function

$$f(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi m}\,\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}$$

Remark A.4.1. While we have denoted the degrees of freedom above by m for clarity, it is convention to denote the degrees of freedom by the Greek letter ν. □

Remark A.4.2. 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.4.1 Using the t-distribution

If the random variables $X_1, X_2, \dots, X_n$ are independent and identically distributed according to a $\mathrm{N}\left(\mu, \sigma^2\right)$ 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.

Recall that Question 2 of Problem Sheet 10 showed 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.4.2 Review: Gamma distribution

We review that the Gamma distribution parametrised by shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, and denoted by $\text{Gamma}(\alpha, \beta)$ or $\Gamma(\alpha, \beta)$; this was introduced in Prof. Veraart's notes, Definition 8.3.3. If $X \sim \text{Gamma}(\alpha, \beta)$, then its probability density function f_X is defined by

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \qquad \text{with support } x > 0. \tag{A.1}$$

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

$$\begin{aligned} &\int_0^\infty f_X(x) \mathrm{d}x = 1 \\ \Rightarrow &\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathrm{d}x = 1 \\ \Rightarrow &\int_0^\infty x^{\alpha-1} \exp(-\beta x) \mathrm{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha} \end{aligned} \tag{A.2}$$

Equation (A.2) will be needed in the derivation of the t-distribution.

Remark A.4.3. 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.4.4. The gamma function is also defined for complex values, but this will be covered in your course on omplex analysis. □

A.4.3 Derivation of Student's t-distribution (Reading Material)

Suppose that $U \sim \mathrm{N}(0, 1)$ and $V \sim \chi^2_m$ for some $m > 0$, and suppose that U and V are independent (this is very important!). Define

$$T = \frac{U}{\sqrt{V/m}}.$$

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

$$\begin{aligned} f_U(u) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \\ f_V(v) &= \frac{1}{2^{m/2}\Gamma(m/2)} v^{m/2-1} \exp\left(-\frac{v}{2}\right) \end{aligned}$$

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

$$T = \frac{U}{\sqrt{V/m}}, \qquad W = V.$$

We can rearrange these equations to

$$U = T\left(\frac{V}{m}\right)^{1/2}$$
$$\Rightarrow U = T\left(\frac{W}{m}\right)^{1/2}$$

and

$$V = W.$$

Then the joint distribution of T and W is

$$f_{T,W}(t,w) = f_{U,V}(u,v)\left|J\right| = f_U(u)f_V(v)\left|J\right|,$$

where J is the absolute value of the Jacobian. We can compute the Jacobian as

$$J = \det\begin{pmatrix} \frac{\partial u}{\partial t} & \frac{\partial v}{\partial t} \\ \frac{\partial u}{\partial w} & \frac{\partial v}{\partial w} \end{pmatrix} = \det\begin{pmatrix} \left(\frac{w}{m}\right)^{1/2} & 0 \\ -\frac{t}{2m}\left(\frac{w}{m}\right)^{-1/2} & 1 \end{pmatrix} = \left(\frac{w}{m}\right)^{1/2}$$

Then,

$$f_{T,W}(t,w) = f_U(u)f_V(v)\left|J\right| = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}t^2\left(\frac{w}{m}\right)\right)\frac{1}{2^{m/2}\Gamma(m/2)}v^{m/2-1}\exp\left(-\frac{w}{2}\right)\left(\frac{w}{m}\right)^{1/2}$$

$$= \frac{1}{(\pi m)^{1/2}2^{(m+1)/2}\Gamma(m/2)}w^{\left(\frac{m+1}{2}\right)-1}\exp\left(-\frac{1}{2}\left(1+\frac{t^2}{m}\right)w\right)$$

If we now compute the marginal probability density functiopn of T by integrating over W, then (since W is χ^2_m distributed and so has support the positive real line)

$$f_T(t) = \int_0^\infty f_{T,W}(t,w)\mathrm{d}w$$

$$= \frac{1}{(\pi m)^{1/2}2^{(m+1)/2}\Gamma(m/2)}\int_0^\infty w^{\left(\frac{m+1}{2}\right)-1}\exp\left(-\frac{1}{2}\left(1+\frac{t^2}{m}\right)w\right)\mathrm{d}w$$

and we notice that the integral on the right is in the same form as Equation (A.2), but with

$$x = w, \qquad \alpha = \frac{m+1}{2}, \qquad \beta = \frac{1}{2}\left(1+\frac{t^2}{m}\right).$$

Therefore,

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

and the powers of 2 cancel out to give

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

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] Evans, M. J. and Rosenthal, J. S. (2004). *Probability and statistics: The science of uncertainty*. W. H. Freeman and Company, New York.

[5] Fisher, R, A. (1935). *The Design of Experiments*. Oliver and Boyd, London and Edinburgh.

[6] Hand, D. J. (2020). Personal conversation.

[7] Lunn, D. (2006). Lecture notes for A5, University of Oxford.

[8] 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.

[9] Weisberg, S. (1985). *Applied linear regression*. John Wiley & Sons, 2nd edition.

[10] Weiss, N. A. (2008). *Introductory Statistics*. Pearson-Addison-Wesley, 8th edition.