ST102/ST110
Elementary Statistical Theory

Course pack

2022/23 (Lent term)

Dr James Abdey

lse.ac.uk/statistics

ST102/ST110

Elementary Statistical Theory

Course pack

Contents

Chapter 6
Sampling distributions of statistics

6.1 Synopsis of chapter

This chapter considers the idea of sampling and the concept of a sampling distribution for a statistic (such as a sample mean) which must be understood by all users of statistics.

6.2 Learning outcomes

After completing this chapter, you should be able to:

- demonstrate how sampling from a population results in a sampling distribution for a statistic
- prove and apply the results for the mean and variance of the sampling distribution of the sample mean when a random sample is drawn with replacement
- state the central limit theorem and recall when the limit is likely to provide a good approximation to the distribution of the sample mean.

6.3 Introduction

Suppose we have a *sample* of n observations of a random variable X:

$$\boldsymbol{\{X_1, X_2, \ldots, X_n\}}.$$

We have already stated that in statistical inference each individual observation X_i is regarded as a value of a random variable X, with some probability distribution (that is, the *population distribution*).

In this chapter we discuss how we define and work with:

- the joint distribution of the whole sample $\{X_1, X_2, \ldots, X_n\}$, treated as a multivariate random variable
- distributions of univariate functions of $\{X_1, X_2, \ldots, X_n\}$ (*statistics*).

6.4 Random samples

Many of the results discussed here hold for many (or even all) probability distributions, not just for some specific distributions.

It is then convenient to use generic notation.

- We use $f(x)$ to denote both the pdf of a continuous random variable, and the pf of a discrete random variable.
- The parameter(s) of a distribution are generally denoted as θ. For example, for the Poisson distribution θ stands for λ, and for the normal distribution θ stands for (μ, σ^2).
- Parameters are often included in the notation: $f(x;\theta)$ denotes the pf/pdf of a distribution with parameter(s) θ, and $F(x;\theta)$ is its cdf.

For simplicity, we may often use phrases like 'distribution $f(x;\theta)$' or 'distribution $F(x;\theta)$' when we mean 'distribution with the pf/pdf $f(x;\theta)$' and 'distribution with the cdf $F(x;\theta)$', respectively.

The simplest assumptions about the joint distribution of the sample are as follows.

1. $\{X_1, X_2, \ldots, X_n\}$ are **independent** random variables.
2. $\{X_1, X_2, \ldots, X_n\}$ are **identically distributed** random variables. Each X_i has the same distribution $f(x;\theta)$, with the same value of the parameter(s) θ.

The random variables $\{X_1, X_2, \ldots, X_n\}$ are then called:

- **independent and identically distributed (IID) random variables** from the distribution (population) $f(x;\theta)$
- **a random sample** of size n from the distribution (population) $f(x;\theta)$.

We will assume this most of the time from now. So you will see many examples and questions which begin something like:

> 'Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a normal distribution with mean μ and variance $\sigma^2 \ldots$'.

6.4.1 Joint distribution of a random sample

The **joint probability distribution** of the random variables in a random sample is an important quantity in statistical inference. It is known as the **likelihood function**. You will hear more about it in the chapter on point estimation.

For a random sample the joint distribution is easy to derive, because the X_is are independent.

The joint pf/pdf of a random sample is:

$$f(x_1, x_2, \ldots, x_n) = f(x_1; \theta)\, f(x_2; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Other assumptions about random samples

Not all problems can be seen as IID random samples of a single random variable. There are other possibilities, which you will see more of in the future.

- IID samples from *multivariate* population distributions. For example, a sample of (X_i, Y_i), with the joint distribution $\prod_{i=1}^{n} f(x_i, y_i)$.
- Independent but not identically distributed observations. For example, observations (X_i, Y_i) where Y_i (the 'response variable') is treated as random, but X_i (the 'explanatory variable') is not. Hence the joint distribution of the Y_is is $\prod_{i=1}^{n} f_{Y|X}(y_i \,|\, x_i; \theta)$ where $f_{Y|X}(y \,|\, x; \theta)$ is the conditional distribution of Y given X. This is the starting point of *regression modelling* (introduced later in the course).
- Non-independent observations. For example, a *time series* $\{Y_1, Y_2, \ldots, Y_T\}$ where $i = 1, 2, \ldots, T$ are successive time points. The joint distribution of the series is, in general:

$$f(y_1; \theta)\, f(y_2 \,|\, y_1; \theta)\, f(y_3 \,|\, y_1, y_2; \theta) \cdots f(y_T \,|\, y_1, y_2, \ldots, y_{T-1}; \theta).$$

Random samples and their observed values

Here we treat $\{X_1, X_2, \ldots, X_n\}$ as random variables. Therefore, we consider what values $\{X_1, X_2, \ldots, X_n\}$ *might* have in different samples.

Once a *real* sample is actually observed, the values of $\{X_1, X_2, \ldots, X_n\}$ in that specific sample are no longer random variables, but realised values of random variables, i.e. known numbers.

Sometimes this distinction is emphasised in the notation by using:

- $X_1, X_2, \ldots, X_n$ for the random variables
- $x_1, x_2, \ldots, x_n$ for the observed values.

6.5 Statistics and their sampling distributions

A **statistic** is a known function of the random variables $\{X_1, X_2, \ldots, X_n\}$ in a random sample.

Example 6.1 All of the following are statistics:

- the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$
- the sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ and standard deviation $S = \sqrt{S^2}$
- the sample median, quartiles, minimum, maximum etc.
- quantities such as:

$$\sum_{i=1}^{n} X_i^2 \quad \text{and} \quad \frac{\bar{X}}{S/\sqrt{n}}.$$

Here we focus on single (univariate) statistics. More generally, we could also consider vectors of statistics, i.e. multivariate statistics.

6.5.1 Sampling distribution of a statistic

A (simple) random sample is modelled as a sequence of IID random variables. A statistic is a function of these random variables, so it is also a random variable, with a distribution of its own.

In other words, if we collected several random samples from the same population, the values of a statistic would not be the same from one sample to the next, but would vary according to some probability distribution.

The **sampling distribution** is the probability distribution of the values which the statistic would have in a large number of samples collected (independently) from the same population.

Example 6.2 Suppose we collect a random sample of size $n = 20$ from a normal population (distribution) $X \sim N(5, 1)$.

Consider the following statistics:

- sample mean $\bar{X}$, sample variance S^2, and $\max_X = \max(X_1, X_2, \ldots, X_n)$.

Here is one such random sample (with values rounded to 2 decimal places):

```
6.28 5.22 4.19 3.56 4.15 4.11 4.03 5.81 5.43 6.09
4.98 4.11 5.55 3.95 4.97 5.68 5.66 3.37 4.98 6.58
```

For this random sample, the values of our statistics are:

- $\bar{x} = 4.94$
- $s^2 = 0.90$
- $\max_x = 6.58$.

Here is another such random sample (with values rounded to 2 decimal places):

```
5.44 6.14 4.91 5.63 3.89 4.17 5.79 5.33 5.09 3.90
5.47 6.62 6.43 5.84 6.19 5.63 3.61 5.49 4.55 4.27
```

For this sample, the values of our statistics are:

- $\bar{x} = 5.22$ (the first sample had $\bar{x} = 4.94$)
- $s^2 = 0.80$ (the first sample had $s^2 = 0.90$)
- $\max_x = 6.62$ (the first sample had $\max_x = 6.58$).

How to derive a sampling distribution?

The sampling distribution of a statistic is the distribution of the values of the statistic in (infinitely) *many* repeated samples. However, typically we only have *one* sample which was actually observed. Therefore, the sampling distribution seems like an essentially hypothetical concept.

Nevertheless, it is possible to derive the forms of sampling distributions of statistics under different assumptions about the sampling schemes and population distribution $f(x; \theta)$.

There are two main ways of doing this.

- Exactly or approximately through mathematical derivation. This is the most convenient way for subsequent use, but is not always easy.
- With *simulation*, i.e. by using a computer to generate (artificial) random samples from a population distribution of a known form.

Example 6.3 Consider again a random sample of size $n = 20$ from the population $X \sim N(5, 1)$, and the statistics $\bar{X}$, S^2 and $\max_X$.

We first consider deriving the sampling distributions of these by approximation through simulation.

- Here a computer was used to draw 10,000 independent random samples of $n = 20$ from $N(5, 1)$, and the values of $\bar{X}$, S^2 and $\max_X$ for each of these random samples were recorded.
- Figures 6.1, 6.2 and 6.3 show histograms of the statistics for these 10,000 random samples.

We now consider deriving the exact sampling distribution. Here this is possible. For a random sample of size n from $N(\mu, \sigma^2)$ we have:

(a) $\bar{X} \sim N(\mu, \sigma^2/n)$

(b) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$

(c) the sampling distribution of $Y = \max_X$ has the following pdf:

$$f_Y(y) = n(F_X(y))^{n-1} f_X(y)$$

where $F_X(x)$ and $f_X(x)$ are the cdf and pdf of $X \sim N(\mu, \sigma^2)$, respectively.

Curves of the densities of these distributions are also shown in Figures 6.1, 6.2 and 6.3.

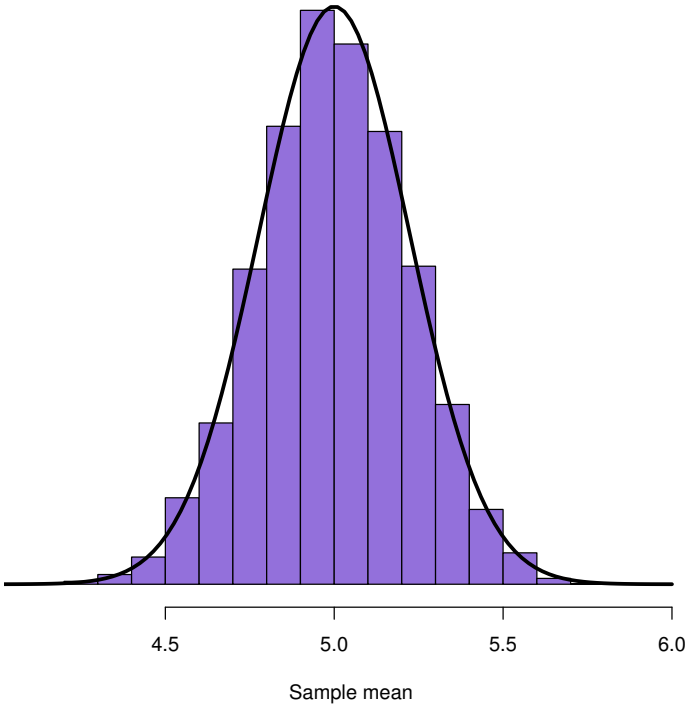


Figure 6.1: Simulation-generated sampling distribution of $\bar{X}$ to accompany Example 6.3.

6.6 Sample mean from a normal population

Consider one very common statistic, the sample mean:

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = \frac{1}{n}\boldsymbol{X}_1 + \frac{1}{n}\boldsymbol{X}_2 + \cdots + \frac{1}{n}\boldsymbol{X}_n.$$

What is the sampling distribution of $\bar{X}$?

We know from Section 5.10.2 that for independent $\{X_1, X_2, \ldots, X_n\}$ from any distribution:

$$\mathbf{E}\left(\sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{X}_i\right) = \sum_{i=1}^{n} \boldsymbol{a}_i \mathbf{E}(\boldsymbol{X}_i)$$

and:

$$\mathbf{Var}\left(\sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{X}_i\right) = \sum_{i=1}^{n} \boldsymbol{a}_i^2 \mathbf{Var}(\boldsymbol{X}_i).$$

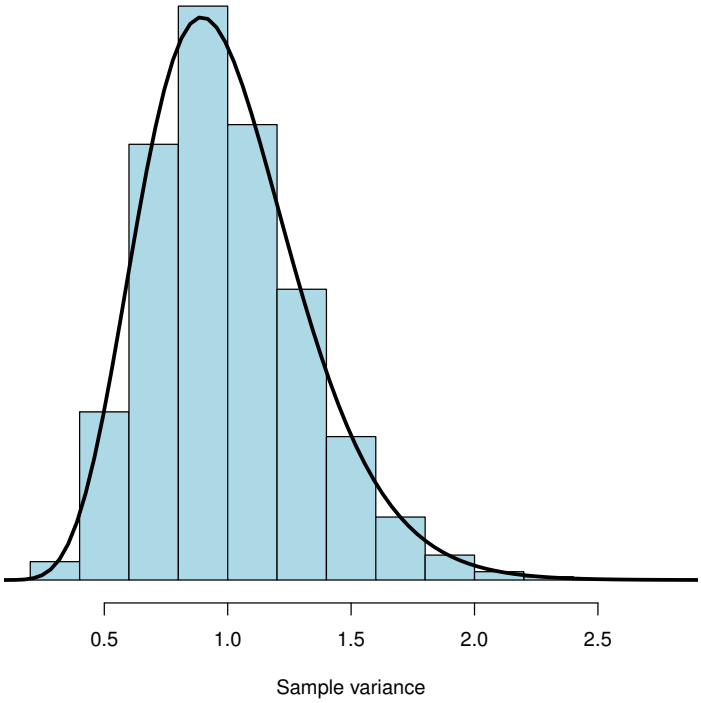


Figure 6.2: Simulation-generated sampling distribution of S^2 to accompany Example 6.3.

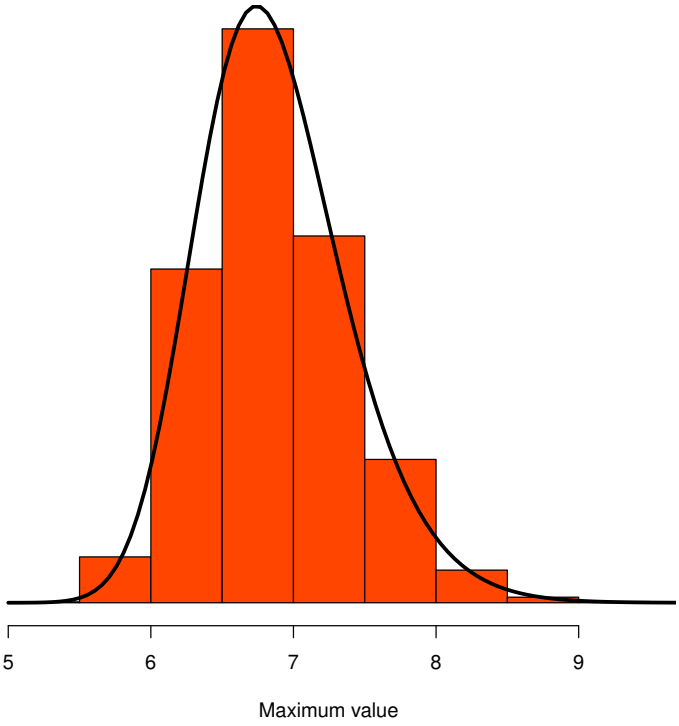


Figure 6.3: Simulation-generated sampling distribution of $\max_X$ to accompany Example 6.3.

For a random sample, all X_is are independent and $\mathrm{E}(X_i) = \mathrm{E}(X)$ is the same for all of them, since the X_is are identically distributed. $\bar{X} = \sum_i X_i/n$ is of the form $\sum_i a_i X_i$, with $a_i = 1/n$ for all $i = 1, 2, \ldots, n$.

Therefore:

$$\mathbf{E}(\bar{\boldsymbol{X}}) = \sum_{i=1}^{n} \frac{\mathbf{1}}{\boldsymbol{n}} \mathbf{E}(\boldsymbol{X}) = \boldsymbol{n} \times \frac{\mathbf{1}}{\boldsymbol{n}} \mathbf{E}(\boldsymbol{X}) = \mathbf{E}(\boldsymbol{X})$$

and:

$$\mathbf{Var}(\bar{\boldsymbol{X}}) = \sum_{i=1}^{n} \frac{\mathbf{1}}{\boldsymbol{n^2}} \mathbf{Var}(\boldsymbol{X}) = \boldsymbol{n} \times \frac{\mathbf{1}}{\boldsymbol{n^2}} \mathbf{Var}(\boldsymbol{X}) = \frac{\mathbf{Var}(\boldsymbol{X})}{\boldsymbol{n}}.$$

So the mean and variance of $\bar{X}$ are $\mathrm{E}(X)$ and $\mathrm{Var}(X)/n$, respectively, for a random sample from *any* population distribution of X. What about the form of the sampling distribution of $\bar{X}$?

This depends on the distribution of X, and is not generally known. However, when the distribution of X is normal, we do know that the sampling distribution of $\bar{X}$ is also normal.

Suppose that $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a normal distribution with mean μ and variance σ^2, then:

$$\bar{\boldsymbol{X}} \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}\right).$$

For example, the pdf drawn on the histogram in Figure 6.1 is that of $N(5, 1/20)$.

We have $\mathrm{E}(\bar{X}) = \mathrm{E}(X) = \mu$.

- In an individual sample, $\bar{x}$ is not usually equal to μ, the expected value of the population.
- However, *over repeated samples* the values of $\bar{X}$ are centred at μ.

We also have $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = \sigma^2/n$, and hence also $\mathrm{sd}(\bar{X}) = \sigma/\sqrt{n}$.

- The variation of the values of $\bar{X}$ in different samples (the **sampling variance**) is large when the population variance of X is large.
- More interestingly, the sampling variance gets smaller when the sample size n increases.
- In other words, when n is large the distribution of $\bar{X}$ is more tightly concentrated around μ than when n is small.

Figure 6.4 shows sampling distributions of $\bar{X}$ from $N(5, 1)$ for different n.

> **Example 6.4** Suppose that the heights (in cm) of men (aged over 16) in a population follow a normal distribution with some unknown mean μ and a known standard deviation of 7.39.

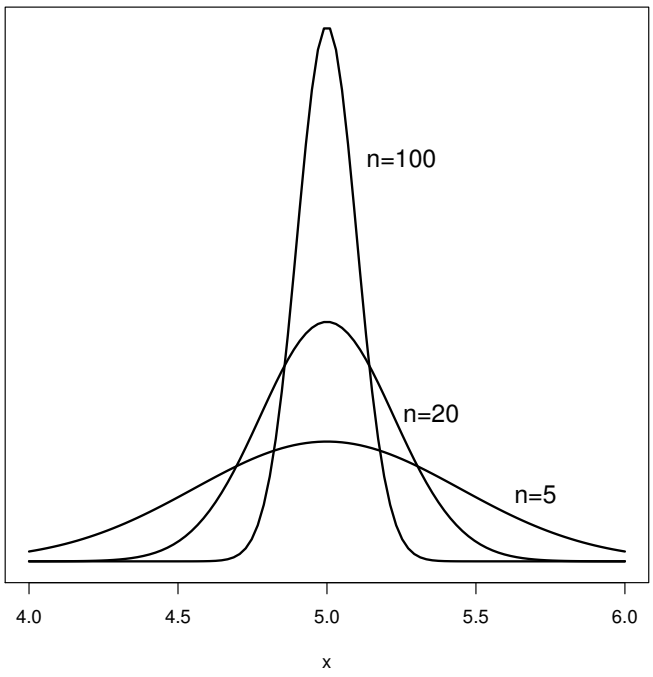


Figure 6.4: Sampling distributions of $\bar{X}$ from $N(5, 1)$ for different n.

We plan to select a random sample of n men from the population, and measure their heights. How large should n be so that there is a probability of at least 0.95 that the sample mean $\bar{X}$ will be within 1 cm of the population mean μ?

Here $X \sim N(\mu, (7.39)^2)$, so $\bar{X} \sim N(\mu, (7.39/\sqrt{n})^2)$. What we need is the smallest n such that:

$$P(|\bar{X} - \mu| \leq 1) \geq 0.95.$$

So:

$$\begin{aligned} P(|\bar{X} - \mu| \leq 1) &\geq 0.95 \\ P(-1 \leq \bar{X} - \mu \leq 1) &\geq 0.95 \\ P\left(\frac{-1}{7.39/\sqrt{n}} \leq \frac{\bar{X} - \mu}{7.39/\sqrt{n}} \leq \frac{1}{7.39/\sqrt{n}}\right) &\geq 0.95 \\ P\left(-\frac{\sqrt{n}}{7.39} \leq Z \leq \frac{\sqrt{n}}{7.39}\right) &\geq 0.95 \\ P\left(Z > \frac{\sqrt{n}}{7.39}\right) &< \frac{0.05}{2} = 0.025 \end{aligned}$$

where $Z \sim N(0, 1)$. From Table 3 of Murdoch and Barnes' *Statistical Tables*, we see that the smallest z which satisfies $P(Z > z) < 0.025$ is $z = 1.97$. Therefore:

$$\frac{\sqrt{n}}{7.39} \geq 1.97 \quad \Leftrightarrow \quad n \geq (7.39 \times 1.97)^2 = 211.9.$$

Therefore, n should be at least 212.

6.7 The central limit theorem

We have discussed the very convenient result that if a random sample comes from a normally-distributed population, the sampling distribution of $\bar{X}$ is also normal. How about sampling distributions of $\bar{X}$ from other populations?

For this, we can use a remarkable mathematical result, the **central limit theorem** (CLT). In essence, the CLT states that the normal sampling distribution of $\bar{X}$ which holds *exactly* for random samples from a normal distribution, also holds *approximately* for random samples from *nearly any* distribution.

The CLT applies to 'nearly any' distribution because it requires that the variance of the population distribution is finite. If it is not (such as for some Pareto distributions, introduced in Chapter 3), the CLT does not hold. However, such distributions are not common.

Suppose that $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a population distribution which has mean $\mathrm{E}(X_i) = \mu < \infty$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$, that is with a finite mean and finite variance. Let $\bar{X}_n$ denote the sample mean calculated from a random sample of size n, then:

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

for any z, where $\Phi(z)$ denotes the cdf of the standard normal distribution.

The '$\lim_{n\to\infty}$' indicates that this is an **asymptotic** result, i.e. one which holds increasingly well as n increases, and exactly when the sample size is infinite.

The full proof of the CLT is not straightforward. A partial (and non-examinable!) version is given in a note on the ST102 Moodle site.

In less formal language, the CLT says that for a random sample from *nearly any* distribution with mean μ and variance σ^2 then:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

approximately, when n is sufficiently large. We can then say that $\bar{X}$ is **asymptotically normally distributed** with mean μ and variance σ^2/n.

The wide reach of the CLT

It may appear that the CLT is still somewhat limited, in that it applies only to sample means calculated from random (IID) samples. However, this is not really true, for two main reasons.

- There are more general versions of the CLT which do not require the observations X_i to be IID.
- Even the basic version applies very widely, when we realise that the 'X' can also be a function of the original variables in the data. For example, if X and Y are

random variables in the sample, we can also apply the CLT to:

$$\sum_{i=1}^{n} \frac{\ln(X_i)}{n} \quad \text{or} \quad \sum_{i=1}^{n} \frac{X_i Y_i}{n}.$$

Therefore, the CLT can also be used to derive sampling distributions for many statistics which do not initially look at all like $\bar{X}$ for a single random variable in an IID sample. You may get to do this in future courses.

How large is 'large n'?

The larger the sample size n, the better the normal approximation provided by the CLT is. In practice, we have various rules-of-thumb for what is 'large enough' for the approximation to be 'accurate enough'. This also depends on the population distribution of X_i. For example:

- for symmetric distributions, even small n is enough
- for very skewed distributions, larger n is required.

For many distributions, $n > 30$ is sufficient for the approximation to be reasonably accurate.

Example 6.5 In the first case, we simulate random samples of sizes:

$$n = 1, 5, 10, 30, 100 \text{ and } 1{,}000$$

from the Exp(0.25) distribution (for which $\mu = 4$ and $\sigma^2 = 16$). This is clearly a skewed distribution, as shown by the histogram for $n = 1$ in Figure 6.5.

10,000 independent random samples of each size were generated. Histograms of the values of $\bar{X}$ in these random samples are shown in Figure 6.5. Each plot also shows the pdf of the approximating normal distribution, $N(4, 16/n)$. The normal approximation is reasonably good already for $n = 30$, very good for $n = 100$, and practically perfect for $n = 1{,}000$.

Example 6.6 In the second case, we simulate 10,000 independent random samples of sizes:

$$n = 1, 10, 30, 50, 100 \text{ and } 1{,}000$$

from the Bernoulli(0.2) distribution (for which $\mu = 0.2$ and $\sigma^2 = 0.16$).

Here the distribution of X_i itself is not even continuous, and has only two possible values, 0 and 1. Nevertheless, the sampling distribution of $\bar{X}$ can be very well-approximated by the normal distribution, when n is large enough.

Note that since here $X_i = 1$ or $X_i = 0$ for all i, $\bar{X} = \sum_{i=1}^{n} X_i/n = m/n$, where m is the number of observations for which $X_i = 1$. In other words, $\bar{X}$ is the **sample proportion** of the value $X = 1$.

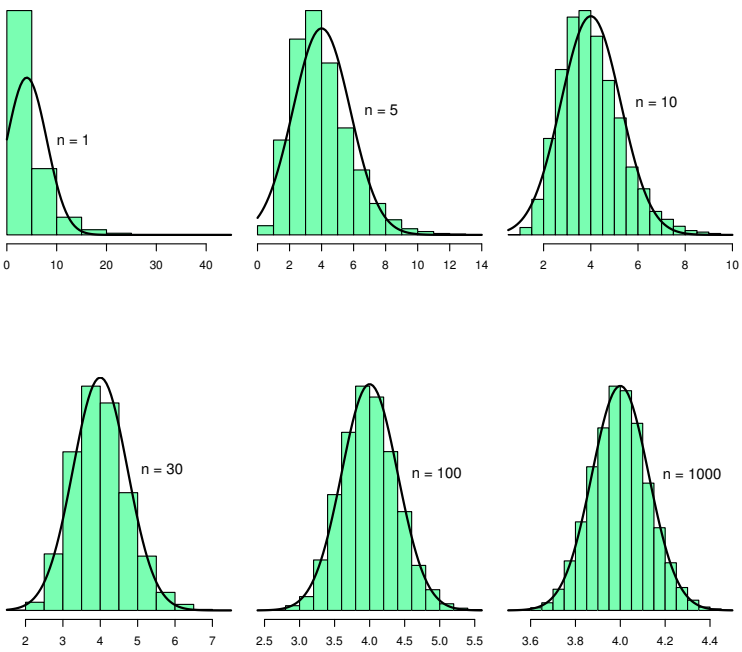


Figure 6.5: Sampling distributions of $\bar{X}$ for various n when sampling from the Exp(0.25) distribution.

> The normal approximation is clearly very bad for small n, but reasonably good already for $n = 50$, as shown by the histograms in Figure 6.6.

6.8 Some common sampling distributions

In the remaining chapters, we will make use of results like the following.

Suppose that $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ are two independent random samples from $N(\mu, \sigma^2)$, then:

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1}$$

$$\sqrt{\frac{n+m-2}{1/n+1/m}} \times \frac{\bar{X}-\bar{Y}}{\sqrt{(n-1)S_X^2+(m-1)S_Y^2}} \sim t_{n+m-2}$$

and:

$$\frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}.$$

Here 'χ^2', 't' and 'F' refer to three new families of probability distributions:

- the χ^2 ('chi-squared') distribution
- the t distribution
- the F distribution.

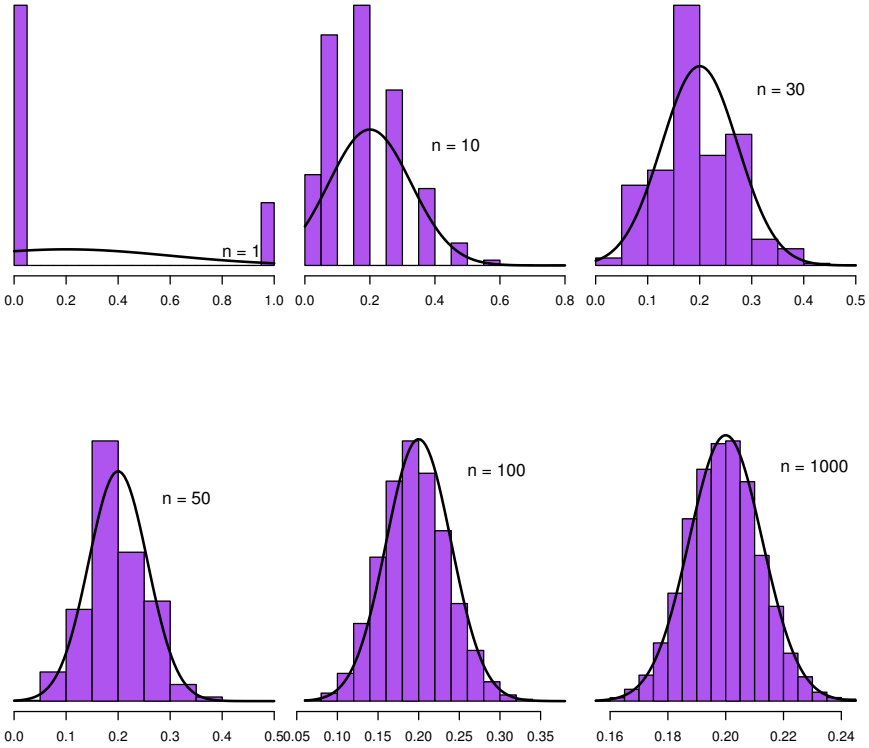


Figure 6.6: Sampling distributions of $\bar{X}$ for various n when sampling from the Bernoulli(0.2) distribution.

These are not often used as distributions of individual variables. Instead, they are used as sampling distributions for various statistics. Each of them arises from the normal distribution in a particular way. We will now briefly introduce their main properties. This is in preparation for statistical inference, where the uses of these distributions will be discussed at length.

6.8.1 The χ^2 distribution

Definition of the χ^2 distribution

Let $\boldsymbol{Z_1, Z_2, \ldots, Z_k}$ be *independent* $\boldsymbol{N(0,1)}$ random variables. If:

$$\boldsymbol{X = Z_1^2 + Z_2^2 + \cdots + Z_k^2 = \sum_{i=1}^{k} Z_i^2}$$

the distribution of X is the **χ^2 distribution with k degrees of freedom**. This is denoted by $X \sim \chi^2(k)$ or $X \sim \chi^2_k$.

The χ^2_k distribution is a continuous distribution, which can take values of $x \geq 0$. Its mean and variance are:

- $\boldsymbol{\mathrm{E}(X) = k}$
- $\boldsymbol{\mathrm{Var}(X) = 2k}$.

For reference, the probability density function of $X \sim \chi^2_k$ is:

$$f(x) = \begin{cases} (2^{k/2}\Gamma(k/2))^{-1} x^{k/2-1} \mathrm{e}^{-x/2} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \, \mathrm{d}x$$

is the *gamma function*, which is defined for all $\alpha > 0$. (Note the formula of the pdf of $X \sim \chi^2_k$ is *not* examinable.)

The shape of the pdf depends on the degrees of freedom k, as illustrated in Figure 6.7. In most applications of the χ^2 distribution the appropriate value of k is known, in which case it does not need to be estimated from data.

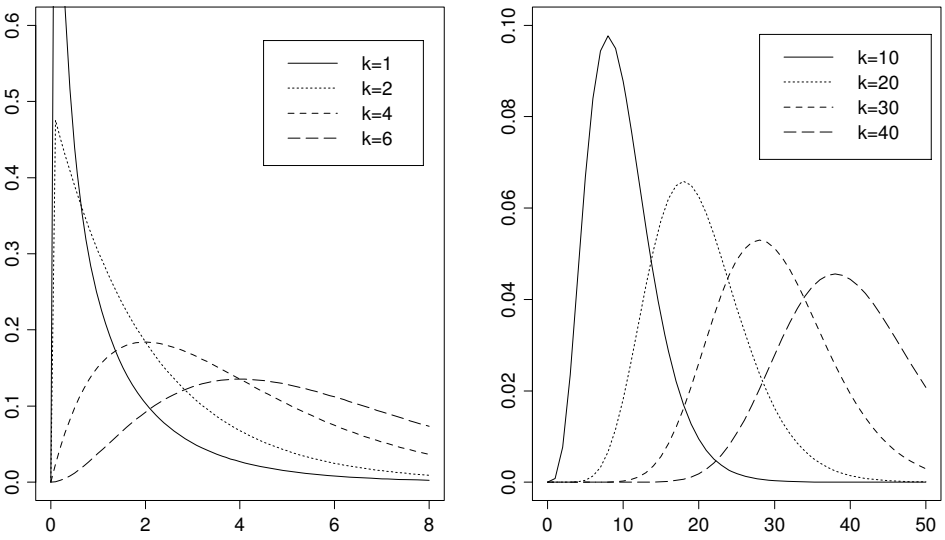


Figure 6.7: χ^2 pdfs for various degrees of freedom.

If $X_1, X_2, \ldots, X_m$ are independent random variables and $X_i \sim \chi^2_{k_i}$, then their sum is also χ^2-distributed where the individual degrees of freedom are added, such that:

$$\boldsymbol{X_1 + X_2 + \cdots + X_m \sim \chi^2_{k_1+k_2+\cdots+k_m}.}$$

The uses of the χ^2 distribution will be discussed later. One example though is if $\{X_1, X_2, \ldots, X_n\}$ is a random sample from the population $N(\mu, \sigma^2)$, and S^2 is the sample variance, then:

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

This result is used to derive basic tools of statistical inference for both μ and σ^2 for the normal distribution.

Tables of the χ^2 distribution

In exercises and the examination, you will need a table of some probabilities for the χ^2 distribution. Table 8 of Murdoch and Barnes' *Statistical Tables* shows the following information.

- The rows correspond to different degrees of freedom k (denoted in the table by ν). The table shows values of k up to 100.
- The columns correspond to the right-tail probability $P(X > x) = \alpha$, where $X \sim \chi^2_k$, for different values of α. The first page contains $\alpha = 0.995, 0.99, \ldots, 0.50$, and the second page contains $\alpha = 0.30, 0.25, \ldots, 0.001$.
- The numbers in the table are values of x such that $P(X > x) = \alpha$ for the k and α in that row and column.

Example 6.7 Consider two numbers in the '$\nu = 5$' row, the 2.675 in the '$\alpha = 0.75$' column and the 3.000 in the '$\alpha = 0.70$' column. These mean that for $X \sim \chi^2_5$ we have:

- $P(X > 2.675) = 0.75$ (and hence $P(X \leq 2.675) = 0.25$)
- $P(X > 3.000) = 0.70$ (and hence $P(X \leq 3.000) = 0.30$).

These also provide bounds for probabilities of other values. For example, since 2.8 is between 2.675 and 3.000, we can conclude that:

$$0.70 < P(X > 2.8) < 0.75.$$

The ways in which this table may be used in statistical inference will be explained in later chapters.

6.8.2 (Student's) t distribution

Definition of Student's t distribution

Suppose $\boldsymbol{Z \sim N(0,1)}$, $\boldsymbol{X \sim \chi^2_k}$, and Z and X are *independent*. The distribution of the random variable:

$$\boldsymbol{T = \frac{Z}{\sqrt{X/k}}}$$

is the **t distribution with k degrees of freedom**. This is denoted $T \sim t_k$ or $T \sim t(k)$. The distribution is also known as 'Student's t distribution'.

The t_k distribution is continuous with the pdf:

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

for all $-\infty < x < \infty$. Examples of $f(x)$ for different k are shown in Figure 6.8. (Note the formula of the pdf of t_k is *not* examinable.)

From Figure 6.8, we see the following.

- The distribution is symmetric around 0.

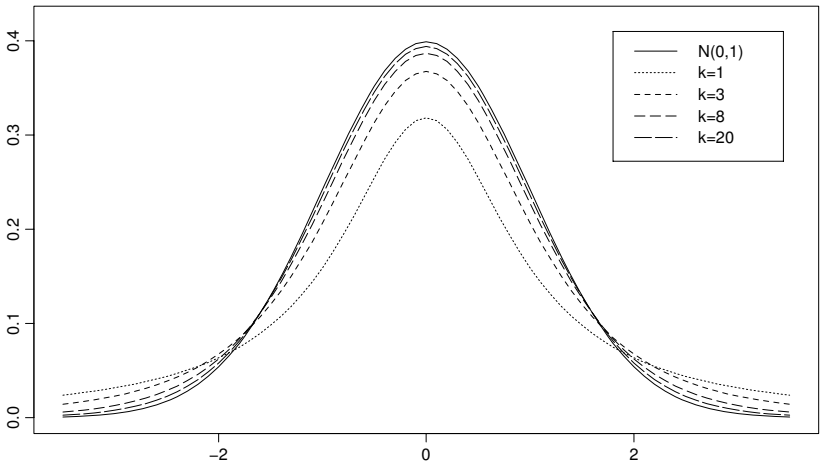


Figure 6.8: Student's t pdfs for various degrees of freedom.

- As $k \to \infty$, the t_k distribution tends to the standard normal distribution, so t_k with large k is very similar to $N(0, 1)$.
- For any finite value of k, the t_k distribution has heavier tails than the standard normal distribution, i.e. t_k places more probability on values far from 0 than $N(0, 1)$ does.

For $T \sim t_k$, the mean and variance of the distribution are:

$$\mathbf{E}(T) = 0 \quad \text{for } k > 1$$

and:

$$\mathbf{Var}(T) = \frac{k}{k-2} \quad \text{for } k > 2.$$

This means that for t_1 neither $\mathrm{E}(T)$ nor $\mathrm{Var}(T)$ exist, and for t_2, $\mathrm{Var}(T)$ does not exist.

Tables of the t distribution

In exercises and the examination, you will need a table of some probabilities for the t distribution. Table 7 of Murdoch and Barnes' *Statistical Tables* shows the following information.

- The rows correspond to different degrees of freedom k (denoted in the table by ν). The table shows values of k up to 120, and then '∞', which is $N(0, 1)$.
- If you need a t_k distribution for which k is not in the table, use the nearest value or use interpolation.
- The columns correspond to the right-tail probability $P(T > t) = \alpha$, where $T \sim t_k$, for $\alpha = 0.10, 0.05, \ldots, 0.0005$.
- The numbers in the table are values of t such that $P(T > t) = \alpha$ for the k and α in that row and column.

Example 6.8 Consider the number 2.132 in the '$\nu = 4$' row, and the '$\alpha = 0.05$' column. This means that for $T \sim t_4$ we have:

- $P(T > 2.132) = 0.05$ (and hence $P(T \leq 2.132) = 0.95$).

The table also provides bounds for other probabilities. For example, the number in the '$\alpha = 0.025$' column is 2.776, so $P(T > 2.776) = 0.025$. Since $2.132 < 2.5 < 2.776$, we know that $0.025 < P(T > 2.5) < 0.05$.

Results for left-tail probabilities $P(T < t) = \alpha$ can also be obtained, because the t distribution is symmetric around 0. This means that $P(T < t) = P(T > -t)$. For example:

$$P(T < -2.132) = P(T > 2.132) = 0.05$$

and $P(T < -2.5) < 0.05$ since $P(T > 2.5) < 0.05$.

This is the same trick we used for the standard normal distribution.

6.8.3 The F distribution

Definition of the F distribution

Let U and V be two independent random variables, where $\boldsymbol{U \sim \chi^2_p}$ and $\boldsymbol{V \sim \chi^2_k}$. The distribution of:

$$\boldsymbol{F = \frac{U/p}{V/k}}$$

is the ***F* distribution with degrees of freedom (p, k)**, denoted $F \sim F_{p,k}$ or $F \sim F(p, k)$.

The F distribution is a continuous distribution, with non-zero probabilities for $x > 0$. The general shape of its pdf is shown in Figure 6.9.

For $F \sim F_{p,k}$, $\mathrm{E}(F) = k/(k-2)$, for $k > 2$. If $F \sim F_{p,k}$, then $1/F \sim F_{k,p}$. If $T \sim t_k$, then $T^2 \sim F_{1,k}$.

Tables of F distributions will be needed for some purposes. They will be available in the examination. We will postpone practice with them until later in the course.

6.9 Prelude to statistical inference

We conclude Chapter 6 with a discussion of the preliminaries of statistical inference before moving on to point estimation. The discussion below will review some key concepts introduced previously.

So, just what *is* 'Statistics'? It is a scientific subject of collecting and 'making sense' of data.

- Collection: designing experiments/questionnaires, designing sampling schemes, and administration of data collection.

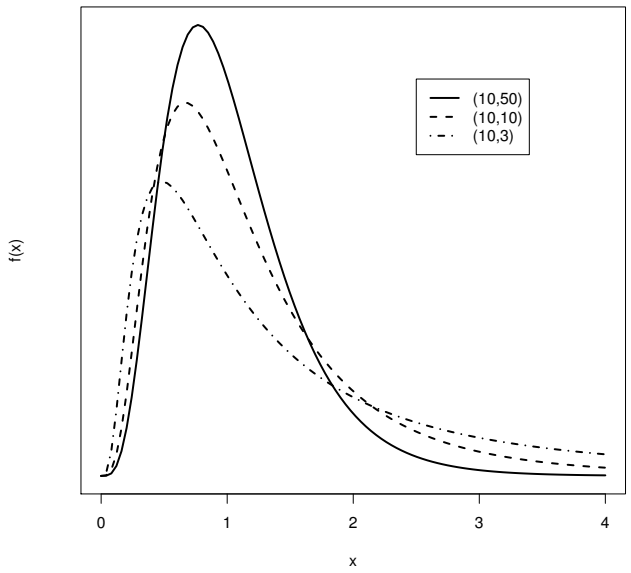


Figure 6.9: F pdfs for various degrees of freedom.

- Making sense: estimation, testing and forecasting.

So, 'Statistics' is an application-oriented subject, particularly useful or helpful in answering questions such as the following.

- Does a certain new drug prolong life for AIDS sufferers?
- Is global warming really happening?
- Are GCSE and A-level examination standards declining?
- Is the gap between rich and poor widening in Britain?
- Is there still a housing bubble in London?
- Is the Chinese yuan undervalued? If so, by how much?

These questions are difficult to study in a laboratory, and admit no self-evident axioms. Statistics provides a way of answering these types of questions using **data**.

What should we learn in 'Statistics'? The basic ideas, methods and theory. Some guidelines for learning/applying statistics are the following.

- Understand what data say in each specific context. All the methods are just tools to help us to understand data.
- Concentrate on what to do and why, rather than on concrete calculations and graphing.
- It may take a while to catch the basic idea of statistics – keep thinking!

6.9.1 Population versus random sample

Consider the following two practical examples.

> **Example 6.9** A new type of tyre was designed to increase its lifetime. The manufacturer tested 120 new tyres and obtained the average lifetime (over these 120 tyres) of 35,391 miles. So the manufacturer claims that the mean lifetime of new tyres is 35,391 miles.

> **Example 6.10** A newspaper sampled 1,000 potential voters, and 350 of them were Labour Party supporters. It claims that the proportion of Labour voters in the whole country is 350/1,000 = 0.35, i.e. 35%.

In both cases, the conclusion is drawn on a **population** (i.e. all the objects concerned) based on the information from a **sample** (i.e. a subset of the population).

In Example 6.9, it is impossible to measure the whole population. In Example 6.10, it is not economical to measure the whole population. Therefore, *errors are inevitable*!

The population is the entire set of objects concerned, and these objects are typically represented by some numbers. We *do not know* the entire population in practice.

In Example 6.9, the population consists of the lifetimes of all tyres, including those to be produced in the future. For the opinion poll in Example 6.10, the population consists of many '1's and '0's, where each '1' represents a voter for the Labour party, and each '0' represents a voter for other parties.

A sample is a (randomly) selected subset of a population, and is known in practice. The population is unknown. We represent a population by a **probability distribution**.

Why do we need a model for the entire population?

- Because the questions we ask concern the entire population, not just the data we have. Having a model for the population tells us that the remaining population is not much different from our data or, in other words, that the data are **representative** of the population.

Why do we need a *random* model?

- Because the process of drawing a sample from a population is a bit like the process of generating random variables. A different sample would produce different values. Therefore, the population from which we draw a random sample is represented as a probability distribution.

6.9.2 Parameter versus statistic

For a given problem, we typically assume a population to be a probability distribution $F(x;\theta)$, where the *form of distribution F is known* (such as normal or Poisson), and θ denotes some *unknown* characteristic (such as the mean or variance) and is called a **parameter**.

Example 6.11 Continuing with Example 6.9, the population may be assumed to be $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$, where μ is the 'true' lifetime.

Let:

$$X = \text{the lifetime of a tyre}$$

then we can write $X \sim N(\mu, \sigma^2)$.

Example 6.12 Continuing with Example 6.10, the population is a Bernoulli distribution such that:

$$P(X = 1) = P(\text{a Labour voter}) = \pi$$

and:

$$P(X = 0) = P(\text{a non-Labour voter}) = 1 - \pi$$

where:

$$\begin{aligned} \pi &= \text{the proportion of Labour supporters in the UK} \\ &= \text{the probability of a voter being a Labour supporter.} \end{aligned}$$

A sample: a set of data or random variables?

A sample of size n, $\{X_1, X_2, \ldots, X_n\}$, is also called a **random sample**. It consists of n real numbers in a practical problem. The word 'random' captures the fact that samples (of the same size) taken by different people or at different times may be different, as they are different subsets of a population.

Furthermore, a sample is also viewed as n **independent and identically distributed** (IID) random variables, when we assess the performance of a statistical method.

Example 6.13 For the tyre lifetime in Example 6.9, suppose the realised sample (of size $n = 120$) gives the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 35{,}391.$$

A different sample may give a different sample mean, such as 36,721.

Is the sample mean $\bar{X}$ a good **estimator** of the unknown 'true' lifetime μ? Obviously, we cannot use the real number 35,391 to assess how good this estimator is, as a different sample may give a different average value, such as 36,721.

By treating $\{X_1, X_2, \ldots, X_n\}$ as random variables, $\bar{X}$ is also a random variable. If the distribution of $\bar{X}$ concentrates closely around (unknown) μ, $\bar{X}$ is a good estimator of μ.

Definition of a statistic

Any known function of a random sample is called a **statistic**. Statistics are used for statistical inference such as estimation and testing.

Example 6.14 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the population $N(\mu, \sigma^2)$, then:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_1 + X_n^2 \quad \text{and} \quad \sin(X_3) + 6$$

are all statistics, but:

$$\frac{X_1 - \mu}{\sigma}$$

is *not* a statistic, as it depends on the unknown quantities μ and σ^2.

An observed random sample is often denoted as $\{x_1, x_2, \ldots, x_n\}$, indicating that they are n real numbers. They are seen as a **realisation** of n IID random variables $\{X_1, X_2, \ldots, X_n\}$.

The connection between a population and a sample is shown in Figure 6.10, where θ is a parameter. A known function of $\{X_1, X_2, \ldots, X_n\}$ is called a statistic.

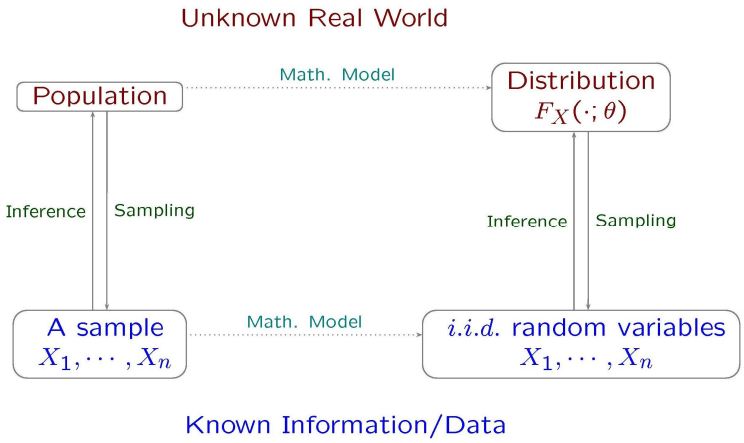


Figure 6.10: Representation of the connection between a population and a sample.

6.9.3 Difference between 'Probability' and 'Statistics'

'Probability' is a mathematical subject, while 'Statistics' is an application-oriented subject (which uses probability heavily).

Example 6.15 Let:

X = the number of lectures attended by a student in a term with 20 lectures

then $X \sim \text{Bin}(20, \pi)$, i.e. the pf is:

$$P(X = x) = \frac{20!}{x!\,(20-x)!}\pi^x(1-\pi)^{20-x} \quad \text{for } x = 0, 1, 2, \ldots, 20$$

and 0 otherwise.

Some probability questions are as follows. Treating π as known:

- what is $\text{E}(X)$ (the average number of lectures attended)?
- what is $P(X \geq 18)$ (the proportion of students attending at least 18 lectures)?
- what is $P(X < 10)$ (the proportion of students attending fewer than half of the lectures)?

Some statistics questions are as follows.

- What is π (the average attendance rate)?
- Is π larger than 0.9?
- Is π smaller than 0.5?

6.10 Overview of chapter

This chapter introduced sampling distributions of statistics which are the foundations to statistical inference. The sampling distribution of the sample mean was derived exactly when sampling from normal populations and also approximately for more general distributions using the central limit theorem. Three new families of distributions (χ^2, t and F) were defined.

6.11 Key terms and concepts

- **Central limit theorem**
- ***F* distribution**
- **Random sample**
- **Sampling variance**
- **(Student's) *t* distribution**
- **Chi-squared (χ^2) distribution**
- **IID random variables**
- **Sampling distribution**
- **Statistic**

Did you hear the one about the statistician? Probably.
(Anon)

Chapter 7
Point estimation

7.1 Synopsis of chapter

This chapter covers point estimation. Specifically, the properties of estimators are considered and the attributes of a desirable estimator are discussed. Techniques for deriving estimators are introduced.

7.2 Learning outcomes

After completing this chapter, you should be able to:

- summarise the performance of an estimator with reference to its sampling distribution
- use the concepts of bias and variance of an estimator
- define mean squared error and calculate it for simple estimators
- find estimators using the method of moments, least squares and maximum likelihood.

7.3 Introduction

The basic setting is that we assume a random sample $\{X_1, X_2, \ldots, X_n\}$ is observed from a population $F(x;\theta)$. The goal is to make inference (i.e. estimation or testing) for the unknown parameter(s) θ.

- Statistical inference is based on two things.
  1. A set of data/observations $\{X_1, X_2, \ldots, X_n\}$.
  2. An assumption of $F(x;\theta)$ for the joint distribution of $\{X_1, X_2, \ldots, X_n\}$.
- Inference is carried out using a statistic, i.e. a known function of $\{X_1, X_2, \ldots, X_n\}$.
- For *estimation*, we look for a statistic $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ such that the *value* of $\widehat{\theta}$ is taken as an *estimate* (i.e. an estimated value) of θ. Such a $\widehat{\theta}$ is called a **point estimator** of θ.
- For *testing*, we typically use a statistic to test if a hypothesis on θ (such as $\theta = 3$) is true or not.

Example 7.1 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a population with mean $\mu = \mathrm{E}(X_i)$. Find an estimator of μ.

Since μ is the mean of the population, a natural estimator would be the sample mean $\widehat{\mu} = \bar{X}$, where:

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

We call $\widehat{\mu} = \bar{X}$ a **point estimator** (or simply an estimator) of μ.

For example, if we have an observed sample of 9, 16, 15, 4 and 12, hence of size $n = 5$, the sample mean is:

$$\widehat{\mu} = \frac{9 + 16 + 15 + 4 + 12}{5} = 11.2.$$

The value 11.2 is a **point estimate** of μ. For an observed sample of 15, 16, 10, 8 and 9, we obtain $\widehat{\mu} = 11.6$.

7.4 Estimation criteria: bias, variance and mean squared error

Estimators are random variables and, therefore, have probability distributions, known as sampling distributions. As we know, two important properties of probability distributions are the mean and variance. Our objective is to create a formal criterion which combines both of these properties to assess the relative performance of different estimators.

Bias of an estimator

Let $\widehat{\theta}$ be an estimator of the population parameter θ.[1] We define the **bias** of an estimator as:

$$\mathbf{Bias}(\widehat{\boldsymbol{\theta}}) = \mathbf{E}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}. \tag{7.1}$$

An estimator is:

$$\textbf{positively biased if} \quad \mathbf{E}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta} > 0$$

$$\textbf{unbiased if} \quad \mathbf{E}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta} = 0$$

$$\textbf{negatively biased if} \quad \mathbf{E}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta} < 0.$$

A positively-biased estimator means the estimator would systematically overestimate the parameter by the size of the bias, *on average*. An unbiased estimator means the estimator would estimate the parameter correctly, *on average*. A negatively-biased

[1] The $\widehat{\ }$ (hat) notation is often used by statisticians to denote an estimator of the parameter beneath the $\widehat{\ }$. So, for example, $\widehat{\lambda}$ denotes an estimator of the Poisson rate parameter λ.

estimator means the estimator would systematically underestimate the parameter by the size of the bias, *on average*.

In words, the bias of an estimator is the difference between the expected (average) value of the estimator and the true parameter being estimated. Intuitively, it would be desirable, other things being equal, to have an estimator with zero bias, called an **unbiased** estimator. Given the definition of bias in (7.1), an unbiased estimator would satisfy:

$$\mathbf{E}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}.$$

In words, the expected value of the estimator is the true parameter being estimated, i.e. *on average*, under repeated sampling, an unbiased estimator correctly estimates θ.

We view bias as a 'bad' thing, so, other things being equal, the smaller an estimator's bias the better.

Example 7.2 Since $\mathrm{E}(\bar{X}) = \mu$, the sample mean $\bar{X}$ is an unbiased estimator of μ because:

$$\mathrm{E}(\bar{X}) - \mu = 0.$$

Variance of an estimator

The variance of an estimator, denoted $\mathbf{Var}(\widehat{\boldsymbol{\theta}})$, is obtained directly from the estimator's sampling distribution.

Example 7.3 For the sample mean, $\bar{X}$, we have:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}. \tag{7.2}$$

It is clear that in (7.2) increasing the sample size n decreases the estimator's variance (and hence the standard error, i.e. the square root of the estimator's variance), therefore increasing the *precision* of the estimator.[2] We conclude that variance is also a 'bad' thing so, other things being equal, the smaller an estimator's variance the better.

Estimator properties

Is $\widehat{\mu} = \bar{X}$ a 'good' estimator of μ?

Intuitively, X_1 or $(X_1 + X_2 + X_3)/3$ would not be good enough as estimators of μ. However, can we use other estimators such as the sample median:

$$\widehat{\mu}_1 = \begin{cases} X_{((n+1)/2)} & \text{for odd } n \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{for even } n \end{cases}$$

[2] Remember, however, that this increased precision comes at a cost – namely the increased expenditure on data collection.

or perhaps a trimmed sample mean:

$$\widehat{\mu}_2 = \frac{1}{n-k_1-k_2}(X_{(k_1+1)} + X_{(k_1+2)} + \cdots + X_{(n-k_2)})$$

or simply $\widehat{\mu}_3 = (X_{(1)} + X_{(n)})/2$, where $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics obtained by rearranging $X_1, X_2, \ldots, X_n$ into ascending order:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

and k_1 and k_2 are two small, positive integers?

To highlight the key idea, let θ be a scalar, and $\widehat{\theta}$ be a (point) estimator of θ. A good estimator would make $|\widehat{\theta} - \theta|$ as small as possible. However:

- θ is unknown
- the value of $\widehat{\theta}$ changes with the observed sample.

Mean squared error and mean absolute deviation

The **mean squared error (MSE)** of $\widehat{\theta}$ is defined as:

$$\mathbf{MSE}(\widehat{\boldsymbol{\theta}}) = \mathbf{E}\left((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right)$$

and the **mean absolute deviation (MAD)** of $\widehat{\theta}$ is defined as:

$$\mathbf{MAD}(\widehat{\boldsymbol{\theta}}) = \mathbf{E}\left(|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|\right).$$

Intuitively, MAD is a more appropriate measure for the error in estimation. However, it is technically less convenient since the function $h(x) = |x|$ is not differentiable at $x = 0$. Therefore, the MSE is used more often.

If $\mathrm{E}(\widehat{\theta}^2) < \infty$, it holds that:

$$\mathbf{MSE}(\widehat{\boldsymbol{\theta}}) = \mathbf{Var}(\widehat{\boldsymbol{\theta}}) + \left(\mathbf{Bias}(\widehat{\boldsymbol{\theta}})\right)^2$$

where $\mathrm{Bias}(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}) - \theta$.

Proof:

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathrm{E}\left((\widehat{\theta} - \theta)^2\right) \\
&= \mathrm{E}\left(\left((\widehat{\theta} - \mathrm{E}(\widehat{\theta})) + (\mathrm{E}(\widehat{\theta}) - \theta)\right)^2\right) \\
&= \mathrm{E}\left((\widehat{\theta} - \mathrm{E}(\widehat{\theta}))^2\right) + \mathrm{E}\left((\mathrm{E}(\widehat{\theta}) - \theta)^2\right) + 2\mathrm{E}\left((\widehat{\theta} - \mathrm{E}(\widehat{\theta}))(\mathrm{E}(\widehat{\theta}) - \theta)\right) \\
&= \mathrm{Var}(\widehat{\theta}) + \mathrm{E}\left((\mathrm{Bias}(\widehat{\theta}))^2\right) + 2\left((\mathrm{E}(\widehat{\theta}) - \mathrm{E}(\widehat{\theta}))(\mathrm{E}(\widehat{\theta}) - \theta)\right) \\
&= \mathrm{Var}(\widehat{\theta}) + \left(\mathrm{Bias}(\widehat{\theta})\right)^2 + 0.
\end{aligned}$$

■

We have already established that both bias and variance of an estimator are 'bad' things, so the MSE (being the sum of a bad thing and a bad thing squared) can also be viewed as a 'bad' thing.[3] Hence when faced with several competing estimators, we prefer the estimator with the smallest MSE.

So, although an unbiased estimator is intuitively appealing, it is perfectly possible that a biased estimator might be preferred if the 'cost' of the bias is offset by a substantial reduction in variance. Hence the MSE provides us with a formal criterion to assess the trade-off between the bias and variance of different estimators of the same parameter.

Example 7.4 A population is known to be normally distributed, i.e. $X \sim N(\mu, \sigma^2)$. Suppose we wish to estimate the population mean, μ. We draw a random sample $\{X_1, X_2, \ldots, X_n\}$ such that these random variables are IID. We have three candidate estimators of μ, T_1, T_2 and T_3, defined as:

$$T_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T_2 = \frac{X_1 + X_n}{2} \quad \text{and} \quad T_3 = \bar{X} + 3.$$

Which estimator should we choose?

We begin by computing the MSE for T_1, noting:

$$\mathrm{E}(T_1) = \mathrm{E}(\bar{X}) = \mu$$

and:

$$\mathrm{Var}(T_1) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Hence T_1 is an unbiased estimator of μ. So the MSE of T_1 is just the variance of T_1, since the bias is 0. Therefore, $\mathrm{MSE}(T_1) = \sigma^2/n$.

Moving to T_2, note:

$$\mathrm{E}(T_2) = \mathrm{E}\left(\frac{X_1 + X_n}{2}\right) = \frac{\mathrm{E}(X_1) + \mathrm{E}(X_n)}{2} = \frac{\mu + \mu}{2} = \mu$$

and:

$$\mathrm{Var}(T_2) = \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_n)}{2^2} = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2}.$$

So T_2 is also an unbiased estimator of μ, hence $\mathrm{MSE}(T_2) = \sigma^2/2$.

Finally, consider T_3, noting:

$$\mathrm{E}(T_3) = \mathrm{E}(\bar{X} + 3) = \mathrm{E}(\bar{X}) + 3 = \mu + 3$$

and:

$$\mathrm{Var}(T_3) = \mathrm{Var}(\bar{X} + 3) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

So T_3 is a positively-biased estimator of μ, with a bias of 3. Hence we have $\mathrm{MSE}(T_3) = \sigma^2/n + 3^2 = \sigma^2/n + 9$.

We seek the estimator with the smallest MSE. Clearly, $\mathrm{MSE}(T_1) < \mathrm{MSE}(T_3)$ so we can eliminate T_3. Now comparing T_1 with T_2, we note that:

[3] Or, for that matter, a 'very bad' thing!

- for $n = 2$, $\text{MSE}(T_1) = \text{MSE}(T_2)$, since the estimators are identical
- for $n > 2$, $\text{MSE}(T_1) < \text{MSE}(T_2)$, so T_1 is preferred.

So $T_1 = \bar{X}$ is our preferred estimator of μ. Intuitively this should make sense. Note for $n > 2$, T_1 uses all the **information** in the sample (i.e. all observations are used), unlike T_2 which uses the first and last observations only. Of course, for $n = 2$, these estimators are identical.

Some remarks are the following.

i. $\widehat{\mu} = \bar{X}$ is a better estimator of μ than X_1 as:

$$\textbf{MSE}(\widehat{\boldsymbol{\mu}}) = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}} < \textbf{MSE}(\boldsymbol{X_1}) = \boldsymbol{\sigma^2}.$$

ii. As $n \to \infty$, $\text{MSE}(\bar{X}) \to 0$, i.e. when the sample size tends to infinity, the error in estimation goes to 0. Such an estimator is called a (mean-square) **consistent estimator**.

Consistency is a reasonable requirement. It may be used to rule out some silly estimators.

For $\tilde{\mu} = (X_1 + X_4)/2$, $\text{MSE}(\tilde{\mu}) = \sigma^2/2$ which does not converge to 0 as $n \to \infty$. This is due to the fact that only a small portion of **information** (i.e. X_1 and X_4) is used in the estimation.

iii. For any random sample $\{X_1, X_2, \ldots, X_n\}$ from a population with mean μ and variance σ^2, it holds that $\text{E}(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$. The derivation of the expected value and variance of the sample mean was covered in Chapter 6.

iv. For any *independent* random variables $Y_1, Y_2, \ldots, Y_k$ and constants $a_1, a_2, \ldots, a_k$, then:

$$\textbf{E}\left(\sum_{i=1}^{k} \boldsymbol{a_i Y_i}\right) = \sum_{i=1}^{k} \boldsymbol{a_i \text{E}(Y_i)} \quad \textbf{and} \quad \textbf{Var}\left(\sum_{i=1}^{k} \boldsymbol{a_i Y_i}\right) = \sum_{i=1}^{k} \boldsymbol{a_i^2 \text{Var}(Y_i)}.$$

The proof uses the fact that:

$$\textbf{Var}\left(\sum_{i=1}^{k} \boldsymbol{a_i Y_i}\right) = \textbf{E}\left(\left(\sum_{i=1}^{k} \boldsymbol{a_i (Y_i - \text{E}(Y_i))}\right)^{\boldsymbol{2}}\right).$$

Example 7.5 Bias by itself cannot be used to measure the quality of an estimator. Consider two artificial estimators of θ, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, such that $\widehat{\theta}_1$ takes only the two values, $\theta - 100$ and $\theta + 100$, and $\widehat{\theta}_2$ takes only the two values θ and $\theta + 0.2$, with the following probabilities:

$$P(\widehat{\theta}_1 = \theta - 100) = P(\widehat{\theta}_1 = \theta + 100) = 0.5$$

and:

$$P(\widehat{\theta}_2 = \theta) = P(\widehat{\theta}_2 = \theta + 0.2) = 0.5.$$

Note that $\widehat{\theta}_1$ is an unbiased estimator of θ and $\widehat{\theta}_2$ is a positively-biased estimator of θ as:

$$\text{Bias}(\widehat{\theta}_2) = \text{E}(\widehat{\theta}_2) - \theta = ((\theta \times 0.5) + ((\theta + 0.2) \times 0.5)) - \theta = 0.1.$$

However:

$$\text{MSE}(\widehat{\theta}_1) = \text{E}((\widehat{\theta}_1 - \theta)^2) = (-100)^2 \times 0.5 + (100)^2 \times 0.5 = 10{,}000$$

and:

$$\text{MSE}(\widehat{\theta}_2) = \text{E}((\widehat{\theta}_2 - \theta)^2) = 0^2 \times 0.5 + (0.2)^2 \times 0.5 = 0.02.$$

Hence $\widehat{\theta}_2$ is a much better (i.e. more *accurate*) estimator of θ than $\widehat{\theta}_1$.

Example 7.6 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a population with mean $\mu = \text{E}(X_i)$ and variance $\sigma^2 = \text{Var}(X_i) < \infty$, for $i = 1, 2, \ldots, n$. Let $\widehat{\mu} = \bar{X}$. Find $\text{MSE}(\widehat{\mu})$.

We compute the bias and variance separately.

$$\text{E}(\widehat{\mu}) = \text{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\text{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Hence $\text{Bias}(\widehat{\mu}) = \text{E}(\widehat{\mu}) - \mu = 0$. For the variance, we note the useful formula:

$$\left(\sum_{i=1}^{k} a_i\right)\left(\sum_{j=1}^{k} b_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_i b_j = \sum_{i=1}^{k} a_i b_i + \sum_{1\le i\ne j\le k} a_i b_j.$$

Especially:

$$\left(\sum_{i=1}^{k} a_i\right)^2 = \sum_{i=1}^{k} a_i^2 + \sum_{1\le i\ne j\le k} a_i a_j.$$

Hence $\text{Var}(\widehat{\mu}) =$

$$\begin{aligned}
\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \text{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right) \\
&= \text{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)\right)^2\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{E}((X_i-\mu)^2) + \sum_{1\le i\ne j\le n}\text{E}\left((X_i-\mu)(X_j-\mu)\right)\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + \sum_{1\le i\ne j\le n}\text{E}(X_i-\mu)\,\text{E}(X_j-\mu)\right) = \frac{\sigma^2}{n}.
\end{aligned}$$

Hence $\text{MSE}(\widehat{\mu}) = \text{MSE}(\bar{X}) = \sigma^2/n$.

Finding estimators

In general, how should we find an estimator of θ in a practical situation?

There are three conventional methods:

- **method of moments estimation**
- **least squares estimation**
- **maximum likelihood estimation**.

7.5 Method of moments (MM) estimation

Method of moments estimation

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a population $F(x;\theta)$. Suppose θ has p components (for example, for a normal population $N(\mu, \sigma^2)$, $p = 2$; for a Poisson population with parameter λ, $p = 1$).

Let:

$$\boldsymbol{\mu_k = \mu_k(\theta) = \mathrm{E}(X^k)}$$

denote the kth **population moment**, for $k = 1, 2, \ldots$. Therefore, μ_k depends on the unknown parameter θ, as everything else about the distribution $F(x;\theta)$ is known.

Denote the kth **sample moment** by:

$$\boldsymbol{M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{X_1^k + X_2^k + \cdots + X_n^k}{n}.}$$

The **MM estimator (MME)** $\widehat{\theta}$ of θ is the solution of the p equations:

$$\boldsymbol{\mu_k(\widehat{\theta}) = M_k \quad \text{for } k = 1, 2, \ldots, p.}$$

Example 7.7 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Find the MM estimator of (μ, σ^2).

There are two unknown parameters. Let:

$$\widehat{\mu} = \widehat{\mu}_1 = M_1 \quad \text{and} \quad \widehat{\mu}_2 = M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

This gives us $\widehat{\mu} = M_1 = \bar{X}$.

Since $\sigma^2 = \mu_2 - \mu_1^2 = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, we have:

$$\widehat{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note we have:

$$\begin{aligned}\mathrm{E}(\widehat{\sigma}^2) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i^2) - \mathrm{E}(\bar{X}^2) \\ &= \mathrm{E}(X^2) - \mathrm{E}(\bar{X}^2) \\ &= \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \frac{(n-1)\sigma^2}{n}.\end{aligned}$$

Since:

$$\mathrm{E}(\widehat{\sigma}^2) - \sigma^2 = -\frac{\sigma^2}{n} < 0$$

$\widehat{\sigma}^2$ is a negatively-biased estimator of σ^2.

The sample variance, defined as:

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

is a more frequently-used estimator of σ^2 as it has zero bias, i.e. it is an unbiased estimator since $\mathrm{E}(S^2) = \sigma^2$. This is why we use the $n-1$ divisor when calculating the sample variance.

A useful formula for computation of the sample variance is:

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

Note the MME does not use any information on $F(x;\theta)$ beyond the moments.

The idea is that M_k should be pretty close to μ_k when n is sufficiently large. In fact:

$$\boldsymbol{M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k}$$

converges to:

$$\boldsymbol{\mu_k = \mathrm{E}(X^k)}$$

as $n \to \infty$. This is due to the **law of large numbers (LLN)**. We illustrate this phenomenon by simulation using R.

Example 7.8 For $N(2,4)$, we have $\mu_1 = 2$ and $\mu_2 = 8$. We use the sample moments M_1 and M_2 as estimators of μ_1 and μ_2, respectively. Note how the sample moments converge to the population moments as the sample size increases.

For a sample of size $n = 10$, we obtained $m_1 = 0.5145838$ and $m_2 = 2.171881$.

```
> x <- rnorm(10,2,2)
> x
 [1]  0.70709403 -1.38416864 -0.01692815  2.51837989 -0.28518898  1.96998829
 [7] -1.53308559 -0.42573724  1.76006933  1.83541490
> mean(x)
[1] 0.5145838
> x2 <- x^2
> mean(x2)
[1] 2.171881
```

For a sample of size $n = 100$, we obtained $m_1 = 2.261542$ and $m_2 = 8.973033$.

```
> x <- rnorm(100,2,2)
> mean(x)
[1] 2.261542
> x2 <- x^2
> mean(x2)
[1] 8.973033
```

For a sample of size $n = 500$, we obtained $m_1 = 1.912112$ and $m_2 = 7.456353$.

```
> x <- rnorm(500,2,2)
> mean(x)
[1] 1.912112
> x2 <- x^2
> mean(x2)
[1] 7.456353
```

Example 7.9 For a Poisson distribution with $\lambda = 1$, we have $\mu_1 = 1$ and $\mu_2 = 2$. With a sample of size $n = 500$, we obtained $m_1 = 1.09$ and $m_2 = 2.198$.

```
> x <- rpois(500,1)
> mean(x)
[1] 1.09
> x2 <- x^2
> mean(x2)
[1] 2.198
> x
 [1] 1 2 2 1 0 0 0 0 0 0 2 2 1 2 1 1 1 2 ...
```

7.6 Least squares (LS) estimation

Given a random sample $\{X_1, X_2, \ldots, X_n\}$ from a population with mean μ and variance σ^2, how can we estimate μ?

The MME of μ is the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$.

Least squares estimator of μ

The estimator $\bar{X}$ is also the **least squares estimator (LSE)** of μ, defined as:

$$\boldsymbol{\widehat{\mu} = \bar{X} = \min_{a} \sum_{i=1}^{n} (X_i - a)^2.}$$

Proof: Given that $S = \sum_{i=1}^{n}(X_i - a)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - a)^2$, where all terms are non-negative, then the value of a for which S is minimised is when $n(\bar{X} - a)^2 = 0$, i.e. $a = \bar{X}$.

■

Estimator accuracy

In order to assess the accuracy of $\widehat{\mu} = \bar{X}$ as an estimator of μ we calculate its MSE:

$$\boldsymbol{\mathrm{MSE}(\widehat{\mu}) = \mathrm{E}((\widehat{\mu} - \mu)^2) = \frac{\sigma^2}{n}.}$$

In order to determine the distribution of $\widehat{\mu}$ we require knowledge of the underlying distribution. Even if the relevant knowledge is available, one may only compute the *exact* distribution of $\widehat{\mu}$ explicitly for a limited number of cases.

By the central limit theorem, as $n \to \infty$, we have:

$$\boldsymbol{P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z)}$$

for any z, where $\Phi(z)$ is the cdf of $N(0, 1)$, i.e. when n is large, $\bar{X} \sim N(\mu, \sigma^2/n)$ approximately.

Hence when n is large:

$$\boldsymbol{P\left(|\bar{X} - \mu| \leq 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \approx 0.95.}$$

In practice, the standard deviation σ is unknown and so we replace it by the sample standard deviation S, where S^2 is the sample variance, given by:

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

This gives an approximation of:

$$\boldsymbol{P\left(|\bar{X} - \mu| \leq 1.96 \times \frac{S}{\sqrt{n}}\right) \approx 0.95.}$$

To be on the safe side, the coefficient 1.96 is often replaced by 2. The **estimated standard error** of $\bar{X}$ is:

$$\textbf{E.S.E.}(\bar{X}) = \frac{S}{\sqrt{n}} = \left(\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^{1/2}.$$

Some remarks are the following.

i. The LSE is a geometrical solution – it minimises the sum of squared distances between the estimated value and each observation. *It makes no use of any information about the underlying distribution.*

ii. Taking the derivative of $\sum_{i=1}^{n}(X_i - a)^2$ with respect to a, and equating it to 0, we obtain (after dividing through by -2):

$$\sum_{i=1}^{n}(X_i - a) = \sum_{i=1}^{n} X_i - na = 0.$$

Hence the solution is $\widehat{\mu} = \widehat{a} = \bar{X}$. This is another way to derive the LSE of μ.

7.7 Maximum likelihood (ML) estimation

We begin with an illustrative example. Maximum likelihood (ML) estimation generalises the reasoning in the following example to arbitrary settings.

Example 7.10 Suppose we toss a coin 10 times, and record the number of 'heads' as a random variable X. Therefore:

$$X \sim \text{Bin}(10, \pi)$$

where $\pi = P(\text{heads}) \in (0, 1)$ is the unknown parameter.

If $x = 8$, what is your best guess (i.e. estimate) of π? Obviously 0.8!

- Is $\pi = 0.1$ *possible*? Yes, but very unlikely.
- Is $\pi = 0.5$ *possible*? Yes, but not very likely.
- Is $\pi = 0.7$ or 0.9 *possible*? Yes, very likely.

Nevertheless, $\pi = 0.8$ is the *most likely*, or *'maximally' likely* value of the parameter. Why do we think '$\pi = 0.8$' is most likely?

Let:

$$L(\pi) = P(X = 8) = \frac{10!}{8!\,2!}\,\pi^8(1-\pi)^2.$$

Since $x = 8$ is the event which occurred in the experiment, this probability would be very large. Figure 7.1 shows a plot of $L(\pi)$ as a function of π.

The *most likely* value of π should make this probability as large as possible. This value is taken as the maximum likelihood estimate of π.

Maximising $L(\pi)$ is equivalent to maximising:

$$l(\pi) = \ln(L(\pi)) = 8\ln\pi + 2\ln(1-\pi) + c$$

where c is the constant $\ln(10!/(8!\,2!))$. Setting $\mathrm{d}l(\pi)/\mathrm{d}\pi = 0$, we obtain the maximum likelihood estimate $\widehat{\pi} = 0.8$.

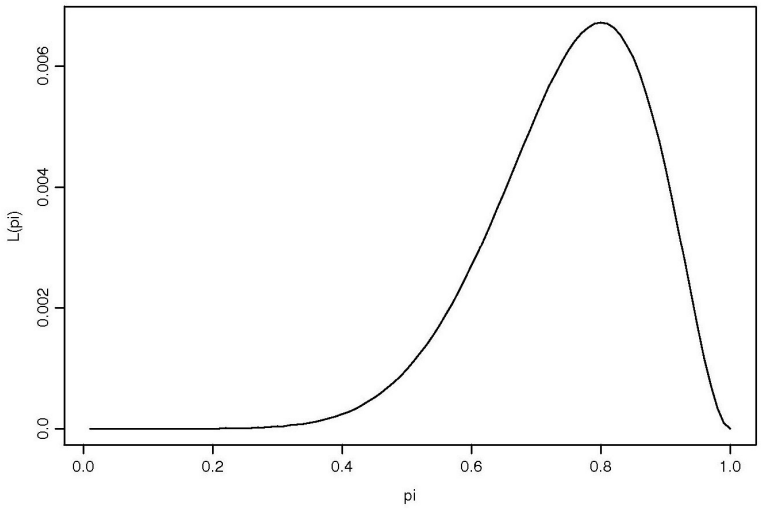


Figure 7.1: Plot of the likelihood function in Example 7.10.

Maximum likelihood definition

Let $f(x_1, x_2, \ldots, x_n; \theta)$ be the joint probability density function (or probability function) for random variables $(X_1, X_2, \ldots, X_n)$. The maximum likelihood estimator (MLE) of θ based on the observations $\{X_1, X_2, \ldots, X_n\}$ is defined as:

$$\boldsymbol{\widehat{\theta} = \max_{\theta} f(X_1, X_2, \ldots, X_n; \theta).}$$

Some remarks are the following.

i. The MLE depends only on the observations $\{X_1, X_2, \ldots, X_n\}$, such that:

$$\boldsymbol{\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n).}$$

Therefore, $\widehat{\theta}$ is a statistic (as it must be for an estimator of θ).

ii. If $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a population with probability density function $f(x; \theta)$, the joint probability density function for $(X_1, X_2, \ldots, X_n)$ is:

$$\boldsymbol{\prod_{i=1}^{n} f(x_i; \theta).}$$

- The joint pdf is a function of $(X_1, X_2, \ldots, X_n)$, while θ is a parameter.
- The joint pdf describes the probability distribution of $\{X_1, X_2, \ldots, X_n\}$.

The **likelihood function** is defined as:

$$\boldsymbol{L(\theta) = \prod_{i=1}^{n} f(X_i; \theta).} \tag{7.3}$$

- The likelihood function is a function of θ, while $\{X_1, X_2, \ldots, X_n\}$ are treated as constants (as given observations).
- The likelihood function reflects the information about the unknown parameter θ in the data $\{X_1, X_2, \ldots, X_n\}$.

Some remarks are the following.

i. The likelihood function is a function of the parameter. It is defined up to positive constant factors. A likelihood function is *not* a probability density function. It contains all the information about the unknown parameter from the observations.

ii. The MLE is $\boldsymbol{\widehat{\theta} = \max_{\theta} L(\theta)}$.

iii. It is often more convenient to use the **log-likelihood function**[4] denoted as:

$$\boldsymbol{l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i; \theta)}$$

as it transforms the product in (7.3) into a sum. Note that:

$$\boldsymbol{\widehat{\theta} = \max_{\theta} l(\theta).}$$

iv. For a *smooth* likelihood function, the MLE is often the solution of the equation:

$$\boldsymbol{\frac{\mathrm{d}}{\mathrm{d}\theta} l(\theta) = 0.}$$

v. If $\widehat{\theta}$ is the MLE and $\phi = g(\theta)$ is a function of θ, $\widehat{\phi} = g(\widehat{\theta})$ is the MLE of ϕ (which is known as the **invariance principle of the MLE**).

vi. Unlike the MME or LSE, the MLE uses all the information about the population distribution. It is often more *efficient* (i.e. more accurate) than the MME or LSE.

vii. In practice, ML estimation should be used whenever possible.

[4] Throughout where 'log' is used in log-likelihood functions, it will be assumed to be the logarithm to the base e, i.e. the natural logarithm.

Example 7.11 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a distribution with pdf:

$$f(x;\lambda) = \begin{cases} \lambda^2 x \mathrm{e}^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$ is unknown. Find the MLE of λ.

The joint pdf is $f(x_1, x_2, \ldots, x_n; \lambda) = \prod_{i=1}^{n} (\lambda^2 x_i \mathrm{e}^{-\lambda x_i})$ if all $x_i > 0$, and 0 otherwise.

The likelihood function is:

$$\begin{aligned} L(\lambda) &= \lambda^{2n} \exp\left(-\lambda \sum_{i=1}^{n} X_i\right) \prod_{i=1}^{n} X_i \\ &= \lambda^{2n} \exp(-n\lambda \bar{X}) \prod_{i=1}^{n} X_i. \end{aligned}$$

The log-likelihood function is $l(\lambda) = 2n \ln \lambda - n\lambda \bar{X} + c$, where $c = \ln \prod_{i=1}^{n} X_i$ is a constant.

Setting:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l(\lambda) = \frac{2n}{\widehat{\lambda}} - n\bar{X} = 0$$

we obtain $\widehat{\lambda} = 2/\bar{X}$.

Note the MLE $\widehat{\lambda}$ may be obtained from maximising $L(\lambda)$ directly. However, it is much easier to work with $l(\lambda)$ instead.

Example 7.12 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$.

The joint pdf is $(2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} (x_i - \mu)^2/(2\sigma^2)\right)$.

Case I: σ^2 is known.

The likelihood function is:

$$\begin{aligned} L(\mu) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{X} - \mu)^2\right). \end{aligned}$$

Hence the log-likelihood function is:

$$l(\mu) = \ln\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 - \frac{n}{2\sigma^2}(\bar{X} - \mu)^2.$$

Maximising $l(\mu)$ with respect to μ gives $\widehat{\mu} = \bar{X}$.

Case II: σ^2 is unknown.

The likelihood function is:

$$L(\mu, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right).$$

Hence the log-likelihood function is:

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 + c$$

where $c = -(n/2)\ln(2\pi)$. Regardless of the value of σ^2, $l(\bar{X}, \sigma^2) \geq l(\mu, \sigma^2)$. Hence $\widehat{\mu} = \bar{X}$.

The MLE of σ^2 should maximise:

$$l(\bar{X}, \sigma^2) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + c.$$

It follows from the lemma below that $\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$.

Lemma: Let $g(x) = -a\ln(x) - b/x$, where $a, b > 0$, then:

$$g\left(\frac{b}{a}\right) = \max_{x>0} g(x).$$

Proof: Letting $g'(x) = -a/x + b/x^2 = 0$ leads to the solution $x = b/a$.

■

Now suppose we wanted to find the MLE of $\gamma = \sigma/\mu$.

Since $\gamma = \gamma(\mu, \sigma)$, by the invariance principle the MLE of γ is:

$$\widehat{\gamma} = \gamma(\widehat{\mu}, \widehat{\sigma}) = \frac{\widehat{\sigma}}{\widehat{\mu}} = \frac{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}}{\sum_{i=1}^{n} X_i/n}.$$

Example 7.13 Consider a population with three types of individuals labelled 1, 2 and 3, and occurring according to the Hardy–Weinberg proportions:

$$p(1;\theta) = \theta^2, \quad p(2;\theta) = 2\theta(1-\theta) \quad \text{and} \quad p(3;\theta) = (1-\theta)^2$$

where $0 < \theta < 1$. Note that $p(1;\theta) + p(2;\theta) + p(3;\theta) = 1$.

A random sample of size n is drawn from this population with n_1 observed values equal to 1 and n_2 observed values equal to 2 (therefore, there are $n - n_1 - n_2$ values equal to 3). Find the MLE of θ.

Let us assume $\{X_1, X_2, \ldots, X_n\}$ is the sample (i.e. n observed values). Among them, there are n_1 '1's, n_2 '2's, and $n - n_1 - n_2$ '3's. The likelihood function is (where $\propto$

means 'proportional to'):

$$L(\theta) = \prod_{i=1}^{n} p(X_i;\theta) = p(1;\theta)^{n_1} p(2;\theta)^{n_2} p(3;\theta)^{n-n_1-n_2}$$
$$= \theta^{2n_1}(2\theta(1-\theta))^{n_2}(1-\theta)^{2(n-n_1-n_2)}$$
$$\propto \theta^{2n_1+n_2}(1-\theta)^{2n-2n_1-n_2}.$$

The log-likelihood is $l(\theta) \propto (2n_1 + n_2)\ln\theta + (2n - 2n_1 - n_2)\ln(1-\theta)$.

Setting:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} l(\theta) = \frac{2n_1 + n_2}{\widehat{\theta}} - \frac{2n - 2n_1 - n_2}{1-\widehat{\theta}} = 0$$

that is:

$$(1-\widehat{\theta})(2n_1 + n_2) = \widehat{\theta}(2n - 2n_1 - n_2)$$

leads to the MLE:

$$\widehat{\theta} = \frac{2n_1 + n_2}{2n}.$$

For example, for a sample with $n = 4$, $n_1 = 1$ and $n_2 = 2$, we obtain a point estimate of $\widehat{\theta} = 0.5$.

7.8 Asymptotic distribution of MLEs

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a population with a smooth pdf $f(x;\theta)$, and θ is a scalar. Denote as:

$$\boldsymbol{\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)}$$

the MLE of θ. Under some regularity conditions, the distribution of $\sqrt{n}(\widehat{\theta} - \theta)$ converges to $N(0, 1/I(\theta))$ as $n \to \infty$, where $I(\theta)$ is the **Fisher information** defined as:

$$\boldsymbol{I(\theta) = -\int_{-\infty}^{\infty} f(x;\theta) \frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} \mathrm{d}x.}$$

Some remarks are the following.

i. When n is large, $\boldsymbol{\widehat{\theta} \sim N(\theta, (nI(\theta))^{-1})}$ approximately.

ii. For a discrete distribution with probability function $p(x;\theta)$, then:

$$\boldsymbol{I(\theta) = -\sum_{x} p(x;\theta) \frac{\partial^2 \ln p(x;\theta)}{\partial\theta^2}.}$$

Example 7.14 For $N(\mu, \sigma^2)$ with σ^2 known, we have:

$$f(x;\mu) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Therefore:

$$\ln f(x;\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2.$$

Hence:

$$\frac{\mathrm{d}\ln f(x;\mu)}{\mathrm{d}\mu} = \frac{x-\mu}{\sigma^2} \quad \text{and} \quad \frac{\mathrm{d}^2\ln f(x;\mu)}{\mathrm{d}\mu^2} = -\frac{1}{\sigma^2}.$$

Therefore:

$$I(\mu) = -\int_{-\infty}^{\infty} -\frac{1}{\sigma^2} f(x;\mu)\,\mathrm{d}x = \frac{1}{\sigma^2}.$$

The MLE of μ is $\bar{X}$, and hence $\bar{X} \sim N(\mu, \sigma^2/n)$.

Example 7.15 For the Poisson distribution, $p(x;\lambda) = \lambda^x \mathrm{e}^{-\lambda}/x!$. Therefore:

$$\ln p(x;\lambda) = x\ln\lambda - \lambda - \ln(x!).$$

Hence:

$$\frac{\mathrm{d}\ln p(x;\lambda)}{\mathrm{d}\lambda} = \frac{x}{\lambda} - 1 \quad \text{and} \quad \frac{\mathrm{d}^2\ln p(x;\lambda)}{\mathrm{d}\lambda^2} = -\frac{x}{\lambda^2}.$$

Therefore:

$$I(\lambda) = \frac{1}{\lambda^2}\sum_{x=0}^{\infty} x\,p(x;\lambda) = \frac{1}{\lambda^2}\mathrm{E}(X) = \frac{1}{\lambda}.$$

The MLE of λ is $\bar{X}$. Hence $\bar{X} \sim N(\lambda, \lambda/n)$ approximately, when n is large.

7.9 Overview of chapter

This chapter introduced point estimation. Key properties of estimators were explored and the characteristics of a desirable estimator were studied through the calculation of the mean squared error. Methods for finding estimators of parameters were also described, including method of moments, least squares and maximum likelihood estimation.

7.10 Key terms and concepts

- Bias
- Consistent estimator
- Fisher information
- Information
- Invariance principle
- Law of large numbers (LLN)
- Least squares estimation
- Likelihood function
- Log-likelihood function
- Maximum likelihood estimation
- Mean absolute deviation (MAD)
- Mean squared error (MSE)
- Method of moments estimation
- Parameter
- Point estimate
- Point estimator
- Population moment
- Random sample
- Sample moment
- Standard error
- Statistic
- Unbiased

The group was alarmed to find that if you are a labourer, cleaner or dock worker, you are twice as likely to die than a member of the professional classes.
(The Sunday Times, 31 August 1980)

Chapter 8
Interval estimation

8.1 Synopsis of chapter

This chapter covers interval estimation – a natural extension of point estimation. Due to the almost inevitable sampling error, we wish to communicate the level of uncertainty in our point estimate by constructing confidence intervals.

8.2 Learning outcomes

After completing this chapter, you should be able to:

- explain the coverage probability of a confidence interval
- construct confidence intervals for means of normal and non-normal populations when the variance is known and unknown
- construct confidence intervals for the variance of a normal population
- explain the link between confidence intervals and distribution theory, and critique the assumptions made to justify the use of various confidence intervals.

8.3 Introduction

Point estimation is simple but not informative enough, since a point estimator is **always subject to errors**. A more scientific approach is to find an upper bound $U = U(X_1, X_2, \ldots, X_n)$ and a lower bound $L = L(X_1, X_2, \ldots, X_n)$, and hope that the unknown parameter θ lies between the two bounds L and U (life is not always as simple as that, but it is a good start).

An intuitive guess for estimating the population mean would be:

$$\boldsymbol{L = \bar{X} - k \times \text{S.E.}(\bar{X}) \quad \text{and} \quad U = \bar{X} + k \times \text{S.E.}(\bar{X})}$$

where $k > 0$ is a constant and $\text{S.E.}(\bar{X})$ is the standard error of the sample mean.

The (random) interval (L, U) forms an **interval estimator** of θ. For estimation to be as precise as possible, intuitively the width of the interval, $U - L$, should be small.

Typically, the **coverage probability**:

$$\boldsymbol{P(L(X_1, X_2, \ldots, X_n) < \theta < U(X_1, X_2, \ldots, X_n)) < 1.}$$

Ideally, we should choose L and U such that:

- the width of the interval is as small as possible
- the coverage probability is as large as possible.

8.4 Interval estimation for means of normal distributions

Let us consider a simple example. We have a random sample $\{X_1, X_2, \ldots, X_n\}$ from the distribution $N(\mu, \sigma^2)$, with σ^2 known.

From Chapter 7, we have reason to believe that $\bar{X}$ is a good estimator of μ. We also know $\bar{X} \sim N(\mu, \sigma^2/n)$, and hence:

$$\boldsymbol{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1).}$$

Therefore, supposing a 95% coverage probability:

$$\begin{aligned}
0.95 &= P\left(\frac{\sqrt{n}|\bar{X} - \mu|}{\sigma} \leq 1.96\right) \\
&= P\left(|\mu - \bar{X}| \leq 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(-1.96 \times \frac{\sigma}{\sqrt{n}} < \mu - \bar{X} < 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

Therefore, the interval covering μ with probability 0.95 is:

$$\boldsymbol{\left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)}$$

which is called a **95% confidence interval** for μ.

> **Example 8.1** Suppose $\sigma = 1$, $n = 4$, and $\bar{x} = 2.25$, then a 95% confidence interval for μ is:
>
> $$\left(2.25 - 1.96 \times \frac{1}{\sqrt{4}}, 2.25 + 1.96 \times \frac{1}{\sqrt{4}}\right) = (1.27, 3.23).$$
>
> Instead of a simple point estimate of $\widehat{\mu} = 2.25$, we say μ is between 1.27 and 3.23 at the 95% confidence level.

What is $P(1.27 < \mu < 3.23) = 0.95$ in Example 8.1? Well, this probability does not mean anything, since μ is an unknown constant!

We treat $(1.27, 3.23)$ as *one realisation of the random interval* $(\bar{X} - 0.98, \bar{X} + 0.98)$ which covers μ with probability 0.95.

What is the meaning of 'with probability 0.95'? If one repeats the interval estimation a large number of times, about 95% of the time the interval estimator covers the true μ.

Some remarks are the following.

i. The confidence level is often specified as 90%, 95% or 99%. Obviously the higher the confidence level, the wider the interval.

   For the normal distribution example:

$$\begin{aligned}
\mathbf{0.90} &= P\left(\frac{\sqrt{n}|\bar{X} - \mu|}{\sigma} \leq \mathbf{1.645}\right) \\
&= P\left(\bar{X} - 1.645 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right) \\
\mathbf{0.95} &= P\left(\frac{\sqrt{n}|\bar{X} - \mu|}{\sigma} \leq \mathbf{1.96}\right) \\
&= P\left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \\
\mathbf{0.99} &= P\left(\frac{\sqrt{n}|\bar{X} - \mu|}{\sigma} \leq \mathbf{2.576}\right) \\
&= P\left(\bar{X} - 2.576 \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2.576 \times \frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

   The *widths* of the three intervals are $2 \times 1.645 \times \sigma/\sqrt{n}$, $2 \times 1.96 \times \sigma/\sqrt{n}$ and $2 \times 2.576 \times \sigma/\sqrt{n}$, corresponding to the confidence levels of 90%, 95% and 99%, respectively.

   To achieve a 100% confidence level in the normal example, the width of the interval would have to be infinite!

ii. Among all the confidence intervals at the same confidence level, the one with the smallest width gives the *most accurate* estimation and is, therefore, optimal.

iii. For a distribution with a symmetric unimodal density function, optimal confidence intervals are *symmetric*, as depicted in Figure 8.1.

Dealing with unknown σ

In practice the standard deviation σ is typically *unknown*, and we replace it with the sample standard deviation:

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

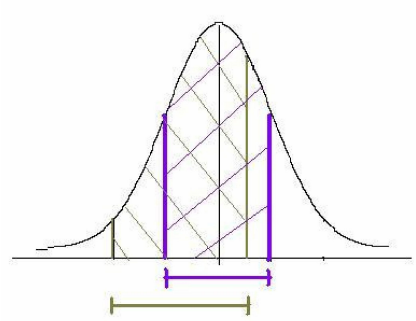


Figure 8.1: Symmetric unimodal density function showing that a given probability is represented by the narrowest interval when symmetric about the mean.

leading to a confidence interval for μ of the form:

$$\left(\bar{X} - k \times \frac{S}{\sqrt{n}}, \bar{X} + k \times \frac{S}{\sqrt{n}}\right)$$

where k is a constant determined by the confidence level and also by the distribution of the statistic:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}. \tag{8.1}$$

However, the distribution of (8.1) is no longer normal – it is the Student's t distribution.

8.4.1 An important property of normal samples

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$. Suppose:

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

where E.S.E.$(\bar{X})$ denotes the estimated standard error of the sample mean.

i. $\bar{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

ii. $\bar{X}$ and S^2 are independent, therefore:

$$\frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{(n-1)S^2/((n-1)\sigma^2)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\text{E.S.E.}(\bar{X})} \sim t_{n-1}.$$

An *accurate* $100(1-\alpha)\%$ confidence interval for μ, where $\alpha \in (0, 1)$, is:

$$\left(\bar{X} - c \times \frac{S}{\sqrt{n}}, \bar{X} + c \times \frac{S}{\sqrt{n}}\right) = (\bar{X} - c \times \text{E.S.E.}(\bar{X}), \bar{X} + c \times \text{E.S.E.}(\bar{X}))$$

where $c > 0$ is a constant such that $P(T > c) = \alpha/2$, where $T \sim t_{n-1}$.

8.5 Approximate confidence intervals

8.5.1 Means of non-normal distributions

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a non-normal distribution with mean μ and variance $\sigma^2 < \infty$.

When n is large, $\sqrt{n}(\bar{X} - \mu)/\sigma$ is $N(0, 1)$ approximately.

Therefore, we have an *approximate* 95% confidence interval for μ given by:

$$\left(\bar{\boldsymbol{X}} - \mathbf{1.96} \times \frac{\boldsymbol{S}}{\sqrt{\boldsymbol{n}}}, \bar{\boldsymbol{X}} + \mathbf{1.96} \times \frac{\boldsymbol{S}}{\sqrt{\boldsymbol{n}}}\right)$$

where S is the sample standard deviation. Note that it is a two-stage approximation.

1. Approximate the distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$ by $N(0, 1)$.
2. Approximate σ by S.

> **Example 8.2** The salary data of 253 graduates from a UK business school (in thousands of pounds) yield the following: $n = 253$, $\bar{x} = 47.126$, $s = 6.843$ and so $s/\sqrt{n} = 0.43$.
>
> A point estimate of the average salary μ is $\bar{x} = 47.126$.
>
> An approximate 95% confidence interval for μ is:
>
> $$47.126 \pm 1.96 \times 0.43 \quad \Rightarrow \quad (46.283, 47.969).$$

8.5.2 MLE-based confidence intervals

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a smooth distribution with unknown parameter θ. Let $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ be the MLE of θ.

Under some regularity conditions, it holds that $\widehat{\theta} \sim N(\theta, (nI(\theta))^{-1})$ approximately, when n is large, where $I(\theta)$ is the Fisher information.

This leads to the following *approximate* 95% confidence interval for θ:

$$\left(\widehat{\boldsymbol{\theta}} - \mathbf{1.96} \times (\boldsymbol{nI}(\widehat{\boldsymbol{\theta}}))^{-\mathbf{1/2}}, \widehat{\boldsymbol{\theta}} + \mathbf{1.96} \times (\boldsymbol{nI}(\widehat{\boldsymbol{\theta}}))^{-\mathbf{1/2}}\right).$$

8.6 Use of the chi-squared distribution

Let $X_1, X_2, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Therefore:

$$\frac{\boldsymbol{X_i} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \sim \boldsymbol{N(0, 1)}.$$

Hence:

$$\frac{\mathbf{1}}{\boldsymbol{\sigma^2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \boldsymbol{\mu})^{\mathbf{2}} \sim \boldsymbol{\chi^2_n}.$$

Note that:

$$\boldsymbol{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2+\frac{n(\bar{X}-\mu)^2}{\sigma^2}.} \tag{8.2}$$

Proof: We have:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2&=\sum_{i=1}^{n}((X_i-\bar{X})+(\bar{X}-\mu))^2\\&=\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu)\\&=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\&=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Hence:

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

■

Since $\bar{X}\sim N(\mu,\sigma^2/n)$, then $n(\bar{X}-\mu)^2/\sigma^2\sim\chi_1^2$. It can be proved that:

$$\boldsymbol{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\sim\chi_{n-1}^2.}$$

Therefore, decomposition (8.2) is an instance of the relationship:

$$\boldsymbol{\chi_n^2=\chi_{n-1}^2+\chi_1^2.}$$

8.7 Interval estimation for variances of normal distributions

Let $\{X_1,X_2,\ldots,X_n\}$ be a random sample from a population with mean μ and variance $\sigma^2<\infty$.

Let $M=\sum_{i=1}^{n}(X_i-\bar{X})^2=(n-1)S^2$, then $M/\sigma^2\sim\chi_{n-1}^2$.

For any given small $\alpha\in(0,1)$, we can find $0<k_1<k_2$ such that:

$$\boldsymbol{P(X<k_1)=P(X>k_2)=\frac{\alpha}{2}}$$

where $X\sim\chi_{n-1}^2$. Therefore:

$$\boldsymbol{1-\alpha=P\left(k_1<\frac{M}{\sigma^2}<k_2\right)=P\left(\frac{M}{k_2}<\sigma^2<\frac{M}{k_1}\right).}$$

Hence a $100(1-\alpha)\%$ confidence interval for σ^2 is:

$$\left(\frac{\boldsymbol{M}}{\boldsymbol{k_2}}, \frac{\boldsymbol{M}}{\boldsymbol{k_1}}\right).$$

Example 8.3 Suppose $n = 15$ and the sample variance is $s^2 = 24.5$. Let $\alpha = 0.05$.

From Table 8 of Murdoch and Barnes' *Statistical Tables*, we find:

$$P(X < 5.629) = P(X > 26.119) = 0.025$$

where $X \sim \chi^2_{14}$.

Hence a 95% confidence interval for σ^2 is:

$$\begin{aligned}\left(\frac{M}{26.119}, \frac{M}{5.629}\right) &= \left(\frac{14 \times S^2}{26.119}, \frac{14 \times S^2}{5.629}\right) \\ &= (0.536 \times S^2, 2.487 \times S^2) \\ &= (13.132, 60.934).\end{aligned}$$

In the above calculation, we have used the formula:

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

8.8 Overview of chapter

This chapter covered interval estimation. A confidence interval converts a point estimate of an unknown parameter into an interval estimate, reflecting the likely sampling error. The chapter demonstrated how to construct confidence intervals for means and variances of normal populations.

8.9 Key terms and concepts

- **Confidence interval**
- **Interval estimator**
- **Coverage probability**
- **Interval width**

A statistician took the Dale Carnegie Course, improving his confidence from 95% to 99%.
(Anon)

Chapter 9
Hypothesis testing

9.1 Synopsis of chapter

This chapter discusses hypothesis testing which is used to answer questions about an unknown parameter. We consider how to perform an appropriate hypothesis test for a given problem, determine error probabilities and test power, and draw appropriate conclusions from a hypothesis test.

9.2 Learning outcomes

After completing this chapter, you should be able to:

- define and apply the terminology of hypothesis testing
- conduct statistical tests of all the types covered in the chapter
- calculate the power of some of the simpler tests
- explain the construction of rejection regions as a consequence of prior distributional results, with reference to the significance level and power.

9.3 Introduction

Hypothesis testing, together with statistical estimation, are the two most frequently-used statistical inference methods. Hypothesis testing addresses a different type of practical question from statistical estimation.

Based on the data, a (statistical) test is to make a binary decision on a hypothesis, denoted by H_0:

reject H_0 or not reject H_0.

9.4 Introductory examples

> **Example 9.1** Consider a simple experiment – toss a coin 20 times.
>
> Let $\{X_1, X_2, \ldots, X_{20}\}$ be the outcomes where 'heads' $\rightarrow X_i = 1$, and 'tails' $\rightarrow X_i = 0$.
>
> Hence the probability distribution is $P(X_i = 1) = \pi = 1 - P(X_i = 0)$, for $\pi \in (0, 1)$.

Estimation would involve estimating π, using $\widehat{\pi} = \bar{X} = (X_1 + X_2 + \cdots + X_{20})/20$.

Testing involves assessing if a hypothesis such as 'the coin is fair' is true or not. For example, this particular hypothesis can be formally represented as:

$$\mathrm{H}_0 : \pi = 0.50.$$

We cannot be sure what the answer is just from the data.

- If $\widehat{\pi} = 0.90$, H_0 is *unlikely* to be true.
- If $\widehat{\pi} = 0.45$, H_0 *may* be true (and also may be untrue).
- If $\widehat{\pi} = 0.70$, what to do then?

Example 9.2 A customer complains that the amount of coffee powder in a coffee tin is less than the advertised weight of 3 pounds.

A random sample of 20 tins is selected, resulting in an average weight of $\bar{x} = 2.897$ pounds. Is this sufficient to substantiate the complaint?

Again statistical estimation cannot provide a firm answer, due to random fluctuations between different random samples. So we cast the problem into a hypothesis testing problem as follows.

Let the weight of coffee in a tin be a normal random variable $X \sim N(\mu, \sigma^2)$. We need to test the hypothesis $\mu < 3$. In fact, we use the data to test the hypothesis:

$$\mathrm{H}_0 : \mu = 3.$$

If we could reject H_0, the customer complaint would be vindicated.

Example 9.3 Suppose one is interested in evaluating the mean income (in £000s) of a community. Suppose income in the population is modelled as $N(\mu, 25)$ and a random sample of $n = 25$ observations is taken, yielding the sample mean $\bar{x} = 17$.

Independently of the data, three expert economists give their own opinions as follows.

- Dr A claims the mean income is $\mu = 16$.
- Ms B claims the mean income is $\mu = 15$.
- Mr C claims the mean income is $\mu = 14$.

How would you assess these experts' statements?

$\bar{X} \sim N(\mu, \sigma^2/n) = N(\mu, 1)$. We assess the statements based on this distribution.

If Dr A's claim is correct, $\bar{X} \sim N(16, 1)$. The observed value $\bar{x} = 17$ is one standard deviation away from μ, and may be regarded as a typical observation from the distribution. Hence there is *little inconsistency* between the claim and the data evidence. This is shown in Figure 9.1.

If Ms B's claim is correct, $\bar{X} \sim N(15, 1)$. The observed value $\bar{x} = 17$ begins to look a bit 'extreme', as it is two standard deviations away from μ. Hence there is *some inconsistency* between the claim and the data evidence. This is shown in Figure 9.2.

If Mr C's claim is correct, $\bar{X} \sim N(14, 1)$. The observed value $\bar{x} = 17$ is very extreme, as it is three standard deviations away from μ. Hence there is *strong inconsistency* between the claim and the data evidence. This is shown in Figure 9.3.

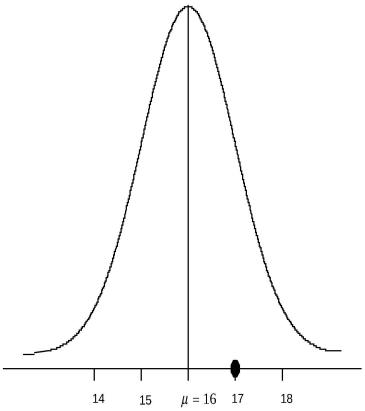


Figure 9.1: Comparison of claim and data evidence for Dr A in Example 9.3.

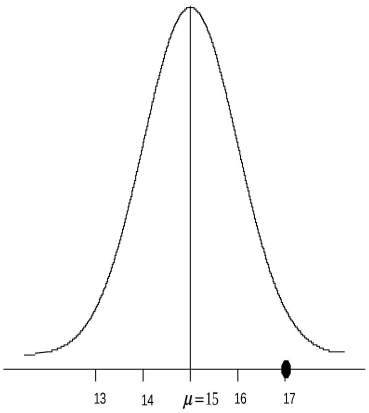


Figure 9.2: Comparison of claim and data evidence for Ms B in Example 9.3.

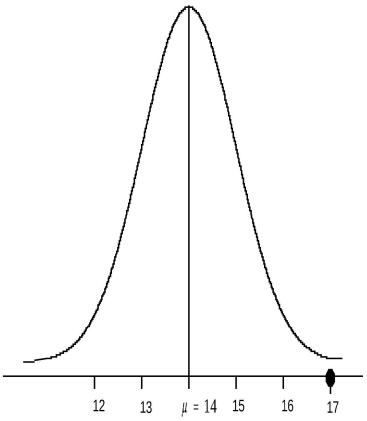


Figure 9.3: Comparison of claim and data evidence for Mr C in Example 9.3.

9.5 Setting p-value, significance level, test statistic

A measure of the discrepancy between the hypothesised (claimed) value of μ and the observed value $\bar{X} = \bar{x}$ is the probability of observing $\bar{X} = \bar{x}$ or more extreme values under the null hypothesis. This probability is called the **p-value**.

Example 9.4 Continuing Example 9.3:

- under $H_0 : \mu = 16$, $P(\bar{X} \geq 17) + P(\bar{X} \leq 15) = P(|\bar{X} - 16| \geq 1) = 0.317$
- under $H_0 : \mu = 15$, $P(\bar{X} \geq 17) + P(\bar{X} \leq 13) = P(|\bar{X} - 15| \geq 2) = 0.046$
- under $H_0 : \mu = 14$, $P(\bar{X} \geq 17) + P(\bar{X} \leq 11) = P(|\bar{X} - 14| \geq 3) = 0.003$.

In summary, we **reject** the hypothesis $\mu = 15$ or $\mu = 14$, as, for example, if the hypothesis $\mu = 14$ is true, the probability of observing $\bar{x} = 17$, or more extreme values, would be as small as 0.003. We are comfortable with this decision, *as a small probability event would be very unlikely to occur in a single experiment*.

On the other hand, we cannot reject the hypothesis $\mu = 16$. However, this does not imply that this hypothesis is necessarily true as, for example, $\mu = 17$ or 18 are at least as likely as $\mu = 16$. Remember:

not reject $\neq$ accept.

A statistical test is incapable of 'accepting' a hypothesis.

Definition of p-values

A p-value is the probability of the event that the test statistic takes the observed value or more extreme (i.e. more unlikely) values under H_0. It is a measure of the discrepancy between the hypothesis H_0 and the data.

- **A 'small' p-value indicates that H_0 is not supported by the data.**
- **A 'large' p-value indicates that H_0 is not inconsistent with the data.**

So p-values may be seen as a risk measure of rejecting H_0, as shown in Figure 9.4.

9.5.1 General setting of hypothesis tests

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a distribution with cdf $F(x; \theta)$. We are interested in testing the hypotheses:

$$\mathbf{H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1}$$

where θ_0 is a fixed value, Θ_1 is a set, and $\theta_0 \notin \Theta_1$.

- H_0 is called the **null hypothesis**.
- H_1 is called the **alternative hypothesis**.

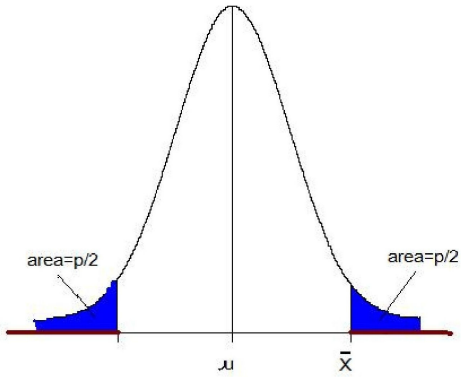


Figure 9.4: Interpretation of p-values as a risk measure.

The **significance level** is based on α, which is a small number between 0 and 1 selected subjectively. Often we choose $\alpha = 0.10$, 0.05 or 0.01, i.e. tests are often conducted at the significance levels of 10%, 5% or 1%, respectively. So we test at the $100\alpha\%$ significance level.

Our **decision** is to **reject H_0 if the p-value is $\leq \alpha$**.

9.5.2 Statistical testing procedure

1. Find a test statistic $T = T(X_1, X_2, \ldots, X_n)$. Denote by t the value of T for the given sample of observations under H_0.

2. Compute the p-value:

$$p = P_{\theta_0}(T = t \text{ or more 'extreme' values})$$

   where P_{θ_0} denotes the probability distribution such that $\theta = \theta_0$.

3. If $p \leq \alpha$ we reject H_0. Otherwise, H_0 is not rejected.

Our understanding of 'extremity' is defined by the alternative hypothesis H_1. This will become clear in subsequent examples. The significance level determines which p-values are considered 'small'.

Example 9.5 Let $\{X_1, X_2, \ldots, X_{20}\}$, taking values either 1 or 0, be the outcomes of an experiment of tossing a coin 20 times, where:

$$P(X_i = 1) = \pi = 1 - P(X_i = 0) \quad \text{for } \pi \in (0, 1).$$

We are interested in testing:

$$H_0 : \pi = 0.50 \quad \text{vs.} \quad H_1 : \pi \neq 0.50.$$

Suppose there are 17 X_is taking the value 1, and 3 X_is taking the value 0. Will you reject the null hypothesis at the 5% significance level?

Let $T = X_1 + X_2 + \cdots + X_{20}$. Therefore, $T \sim \text{Bin}(20, \pi)$. We use T as the test statistic. With the given sample, we observe $t = 17$. What are the more extreme values of T if H_0 is true?

Under H_0, $\text{E}(T) = n\pi_0 = 10$. Hence 3 is as extreme as 17, and the more extreme values are:

$$0, \quad 1, \quad 2, \quad 18, \quad 19 \quad \text{and} \quad 20.$$

Therefore, the p-value is:

$$\begin{aligned}\left(\sum_{i=0}^{3} + \sum_{i=17}^{20}\right) P_{\text{H}_0}(T = i) &= \left(\sum_{i=0}^{3} + \sum_{i=17}^{20}\right) \frac{20!}{i!\,(20-i)!}(0.50)^i(1-0.50)^{20-i} \\ &= 2 \times (0.50)^{20} \sum_{i=0}^{3} \frac{20!}{i!\,(20-i)!} \\ &= 2 \times (0.50)^{20} \times \left(1 + 20 + \frac{20 \times 19}{2!} + \frac{20 \times 19 \times 18}{3!}\right) \\ &= 0.0026.\end{aligned}$$

So we reject the null hypothesis of a fair coin at the 1% significance level.

9.5.3 Two-sided tests for normal means

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$. Assume $\sigma^2 > 0$ is known. We are interested in testing the hypotheses:

$$\mathbf{H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0}$$

where μ_0 is a given constant.

Intuitively if H_0 is true, $\bar{X} = \sum_i X_i/n$ should be close to μ_0. Therefore, large values of $|\bar{X} - \mu_0|$ suggest a departure from H_0.

Under H_0, $\bar{X} \sim N(\mu_0, \sigma^2/n)$, i.e. $\sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0, 1)$. Hence the **test statistic** may be defined as:

$$\boldsymbol{T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)}$$

and we reject H_0 for sufficiently 'large' values of $|T|$.

How large is 'large'? This is determined by the significance level.

Suppose $\mu_0 = 3$, $\sigma = 0.148$, $n = 20$ and $\bar{x} = 2.897$. Therefore, the observed value of T is $t = \sqrt{20} \times (2.897 - 3)/0.148 = -3.112$. Hence the p-value is:

$$P_{\mu_0}(|T| \geq 3.112) = P(|Z| > 3.112) = 0.0019$$

where $Z \sim N(0, 1)$. Therefore, the null hypothesis of $\mu = 3$ will be rejected even at the 1% significance level.

Alternatively, for a given $100\alpha\%$ significance level we may find the **critical value** c_α such that $P_{\mu_0}(|T| > c_\alpha) = \alpha$. Therefore, the p-value is $\leq \alpha$ if and only if the observed value of $|T| \geq c_\alpha$.

Using this alternative approach, we do not need to compute the p-value.

For this example, $c_\alpha = z_{\alpha/2}$, that is the top $100\alpha/2$th percentile of $N(0,1)$, i.e. the z-value which cuts off $\alpha/2$ probability in the upper tail of the standard normal distribution.

For $\alpha = 0.10$, 0.05 and 0.01, $z_{\alpha/2} = 1.645$, 1.96 and 2.576, respectively. Since we observe $|t| = 3.112$, the null hypothesis is rejected at all three significance levels.

9.5.4 One-sided tests for normal means

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$ with $\sigma^2 > 0$ known. We are interested in testing the hypotheses:

$$\mathbf{H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu < \mu_0}$$

where μ_0 is a known constant.

Under H_0, $T = \sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0,1)$. We continue to use T as the test statistic. For $H_1 : \mu < \mu_0$ we should reject H_0 when $t \leq c$, where $c < 0$ is a constant.

For a given $100\alpha\%$ significance level, the critical value c should be chosen such that:

$$\boldsymbol{\alpha = P_{\mu_0}(T \leq c) = P(Z \leq c).}$$

Therefore, c is the 100αth percentile of $N(0,1)$. Due to the symmetry of $N(0,1)$, $c = -z_\alpha$, where z_α is the top 100αth percentile of $N(0,1)$, i.e. $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0,1)$. For $\alpha = 0.05$, $z_\alpha = 1.645$. We reject H_0 if $t \leq -1.645$.

Example 9.6 Suppose $\mu_0 = 3$, $\sigma = 0.148$, $n = 20$ and $\bar{x} = 2.897$, then:

$$t = \frac{\sqrt{20} \times (2.897 - 3)}{0.148} = -3.112 < -1.645.$$

So the null hypothesis of $\mu = 3$ is rejected at the 5% significance level as there is significant evidence from the data that the true mean is likely to be smaller than 3.

Some remarks are the following.

i. We use a one-tailed test when we are only interested in the departure from H_0 in one direction.

ii. The distribution of a test statistic under H_0 *must be known* in order to calculate p-values or critical values.

iii. A test may be carried out by either computing the p-value *or* determining the critical value.

iv. The probability of incorrect decisions in hypothesis testing is typically positive. For example, the significance level is the probability of rejecting a true H_0.

9.6 *t* tests

***t* tests** are one of the most frequently-used statistical tests.

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$, where both μ and $\sigma^2 > 0$ are unknown. We are interested in testing the hypotheses:

$$\mathbf{H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu < \mu_0}$$

where μ_0 is known.

Now we cannot use $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ as a statistic, since σ is *unknown*. Naturally we replace it by S, where:

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

The test statistic is then the famous t statistic:

$$\boldsymbol{T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \sqrt{n}(\bar{X} - \mu_0) \Big/ \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}.}$$

We reject H_0 if $t < c$, where c is the critical value determined by the significance level:

$$\boldsymbol{P_{H_0}(T < c) = \alpha}$$

where P_{H_0} denotes the distribution under H_0 (with mean μ_0 and unknown σ^2).

Under H_0, $T \sim t_{n-1}$. Hence:

$$\boldsymbol{\alpha = P_{H_0}(T < c)}$$

i.e. c is the 100αth percentile of the t distribution with $n-1$ degrees of freedom. By symmetry, $c = -t_{\alpha, n-1}$, where $t_{\alpha,k}$ denotes the top 100αth percentile of the t_k distribution.

Example 9.7 To deal with the customer complaint that the average amount of coffee powder in a coffee tin is less than the advertised 3 pounds, 20 tins were weighed, yielding the following observations:

2.82, 3.01, 3.11, 2.71, 2.93, 2.68, 3.02, 3.01, 2.93, 2.56,
2.78, 3.01, 3.09, 2.94, 2.82, 2.81, 3.05, 3.01, 2.85, 2.79.

The sample mean and standard deviation are, respectively:

$$\bar{x} = 2.897 \quad \text{and} \quad s = 0.148.$$

To test $H_0 : \mu = 3$ vs. $H_1 : \mu < 3$ at the 1% significance level, the critical value is $c = -t_{0.01,19} = -2.539$.

Since $t = \sqrt{20} \times (2.897 - 3)/0.148 = -3.112 < -2.539$, we reject the null hypothesis that $\mu = 3$ at the 1% significance level.

We conclude that there is highly significant evidence which supports the claim that the mean amount of coffee is less than 3 pounds.

> Note the hypotheses tested are in fact:
>
> $$\mathrm{H}_0 : \mu = \mu_0,\ \sigma^2 > 0 \quad \text{vs.} \quad \mathrm{H}_1 : \mu \neq \mu_0,\ \sigma^2 > 0.$$
>
> Although H_0 does not specify the population distribution completely ($\sigma^2 > 0$), the distribution of the test statistic, T, under H_0 is completely known. This enables us to find the critical value or p-value.

9.7 General approach to statistical tests

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the distribution $F(x;\theta)$. We are interested in testing:

$$\mathbf{H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1}$$

where Θ_0 and Θ_1 are two non-overlapping sets. A general approach to test the above hypotheses at the $100\alpha\%$ significance level may be described as follows.

1. Find a test statistic $T = T(X_1, X_2, \ldots, X_n)$ such that the distribution of T under H_0 is known.
2. Identify a **critical region** $\mathcal{C}$ such that:
$$\boldsymbol{P_{\mathrm{H}_0}(T \in \mathcal{C}) = \alpha.}$$
3. If the observed value of T with the given sample is in the critical region $\mathcal{C}$, H_0 is rejected. Otherwise, H_0 is not rejected.

In order to make a test powerful in the sense that the chance of making an incorrect decision is small, the critical region should consist of those values of T which are *least supportive* of H_0 (i.e. which lie in the direction of H_1).

9.8 Two types of error

Statistical tests are often associated with two kinds of decision errors, which are displayed in the following table:

		Decision made	
		H_0 not rejected	H_0 rejected
True state of nature	H_0 true	Correct decision	Type I error
	H_1 true	Type II error	Correct decision

Some remarks are the following.

i. Ideally we would like to have a test which minimises the probabilities of making both types of error, which unfortunately is not feasible.

ii. The probability of making a Type I error is the significance level, which is under our control.

iii. We do not have explicit control over the probability of a Type II error. For a given significance level, we try to choose a test statistic such that the probability of a Type II error is small.

iv. The **power function** of the test is defined as:

$$\boldsymbol{\beta(\theta) = P_\theta\,(\mathrm{H_0 \text{ is rejected}}) \quad \text{for } \theta \in \Theta_1}$$

i.e. $\beta(\theta) = 1 - P(\text{Type II error})$.

v. The null hypothesis $\mathrm{H_0}$ and the alternative hypothesis $\mathrm{H_1}$ are not treated equally in a statistical test, i.e. there is an asymmetric treatment. The choice of $\mathrm{H_0}$ is based on the subject matter concerned and/or technical convenience.

vi. It is more conclusive to end a test with $\mathrm{H_0}$ rejected, as the decision of 'not reject $\mathrm{H_0}$' does not imply that $\mathrm{H_0}$ is accepted.

9.9 Tests for variances of normal distributions

Example 9.8 A container-filling machine is used to package milk cartons of 1 litre ($= 1{,}000$ cm^3). Ideally, the amount of milk should only vary slightly. The company which produced the filling machine claims that the variance of the milk content is *not greater* than 1 cm^3. To examine the veracity of the claim, a random sample of 25 cartons is taken, resulting in 25 measurements (in cm^3) as follows:

1,000.3,	1,001.3,	999.5,	999.7,	999.3,
999.8,	998.3,	1,000.6,	999.7,	999.8,
1,001.0,	999.4,	999.5,	998.5,	1,000.7,
999.6,	999.8,	1,000.0,	998.2,	1,000.1,
998.1,	1,000.7,	999.8,	1,001.3,	1,000.7.

Do these data support the claim of the company?

Turning Example 9.8 into a statistical problem, we assume that the data form a random sample from $N(\mu, \sigma^2)$. We are interested in testing the hypotheses:

$$\boldsymbol{\mathrm{H_0} : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad \mathrm{H_1} : \sigma^2 > \sigma_0^2.}$$

Let $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, then $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Under $\mathrm{H_0}$ we have:

$$\boldsymbol{T = \frac{(n-1)S^2}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} \sim \chi^2_{n-1}.}$$

Since we will reject $\mathrm{H_0}$ against an alternative hypothesis $\sigma^2 > \sigma_0^2$, we should reject $\mathrm{H_0}$ for *large* values of T.

H_0 is rejected if $t > \chi^2_{\alpha,n-1}$, where $\chi^2_{\alpha,n-1}$ denotes the top 100αth percentile of the χ^2_{n-1} distribution, i.e. we have:

$$\boldsymbol{P(T \geq \chi^2_{\alpha,n-1}) = \alpha.}$$

For any $\sigma^2 > \sigma_0^2$, the **power** of the test at σ is:

$$\begin{aligned}\boldsymbol{\beta(\sigma)} &\boldsymbol{= P_\sigma(\text{H}_0 \text{ is rejected})} \\ &\boldsymbol{= P_\sigma(T > \chi^2_{\alpha,n-1})} \\ &\boldsymbol{= P_\sigma\left(\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}\right)} \\ &\boldsymbol{= P_\sigma\left(\frac{(n-1)S^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2} \times \chi^2_{\alpha,n-1}\right)}\end{aligned}$$

which is greater than α, as $\sigma_0^2/\sigma^2 < 1$, where $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ when σ^2 is the *true* variance, instead of σ_0^2. Note that here $1 - \beta(\sigma)$ is the probability of a Type II error.

Suppose we choose $\alpha = 0.05$. For $n = 25$, $\chi^2_{\alpha,n-1} = \chi^2_{0.05,24} = 36.415$.

With the given sample, $s^2 = 0.8088$ and $\sigma_0^2 = 1$, $t = 24 \times 0.8088 = 19.41 < \chi^2_{0.05,24}$. Hence we do not reject H_0 at the 5% significance level. There is no significant evidence from the data against the company's claim that the variance is not beyond 1.

With $\sigma_0^2 = 1$, the power function is:

$$\boldsymbol{\beta(\sigma) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{\chi^2_{0.05,24}}{\sigma^2}\right) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{36.415}{\sigma^2}\right)}$$

where $(n-1)S^2/\sigma^2 \sim \chi^2_{24}$.

For any given values of σ^2, we may compute $\beta(\sigma)$. We list some specific values next.

σ^2	1	1.5	2	3	4
$\chi^2_{0.05,24}/\sigma^2$	36.415	24.277	18.208	12.138	9.104
$\beta(\sigma)$	0.05	0.446	0.793	0.978	0.997
Approximate $\beta(\sigma)$	0.05	0.40	0.80	0.975	0.995

Clearly, $\beta(\sigma) \nearrow$ as $\sigma^2 \nearrow$. Intuitively, it is easier to reject $H_0 : \sigma^2 = 1$ if the true population, which generates the data, has a larger variance σ^2.

Due to the sparsity of the available χ^2 tables, we may only obtain some approximate values for $\beta(\sigma)$ – see the entries in the last row in the above table. The more accurate values of $\beta(\sigma)$ were calculated using a computer.

Some remarks are the following.

i. The significance level is selected subjectively by the statistician. To make the conclusion more convincing in the above example, we may use $\alpha = 0.10$ instead. As $\chi^2_{0.10,24} = 33.196$, H_0 is not rejected at the 10% significance level. In fact the p-value is:

$$P_{H_0}(T \geq 19.41) = 0.73$$

where $T \sim \chi^2_{24}$.

ii. As σ^2 increases, the power function $\beta(\sigma)$ also increases.

iii. For $H_1 : \sigma^2 \neq \sigma_0^2$, we should reject H_0 if:

$$\boldsymbol{t \leq \chi^2_{1-\alpha/2,\,n-1} \quad \text{or} \quad t \geq \chi^2_{\alpha/2,\,n-1}}$$

where $\chi^2_{\alpha,\,k}$ denotes the top 100αth percentile of the χ^2_k distribution.

9.10 Summary: tests for μ and σ^2 in $N(\mu, \sigma^2)$

Null hypothesis, H_0	$\mu = \mu_0$ (σ^2 known)	$\mu = \mu_0$	$\sigma^2 = \sigma_0^2$
Test statistic, T	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	$\frac{(n-1)S^2}{\sigma_0^2}$
Distribution of T under H_0	$N(0,1)$	t_{n-1}	χ^2_{n-1}

In the above table, $\bar{X} = \sum_{i=1}^{n} X_i/n$, $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, and $\{X_1, X_2, \ldots, X_n\}$ is a random sample from $N(\mu, \sigma^2)$.

9.11 Comparing two normal means with paired observations

Suppose that the observations are paired:

$$\boldsymbol{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)}$$

where all X_is and Y_is are independent, $X_i \sim N(\mu_X, \sigma_X^2)$ and $Y_i \sim N(\mu_Y, \sigma_Y^2)$.

We are interested in testing the hypothesis:

$$\boldsymbol{H_0 : \mu_X = \mu_Y.} \tag{9.1}$$

Example 9.9 The following are some practical examples.

- Do husbands make more money than wives?
- Is the increased marketing budget improving sales?
- Are customers willing to pay more for the new product than the old one?
- Does TV advertisement A have higher average effectiveness than advertisement B?
- Will promotion method A generate higher sales than method B?

> Observations are paired together for good reasons: husband-wife, before-after, A-vs.-B (from the same subject).

Let $Z_i = X_i - Y_i$, for $i = 1, 2, \ldots, n$, then $\{Z_1, Z_2, \ldots, Z_n\}$ is a random sample from the population $N(\mu, \sigma^2)$, where:

$$\boldsymbol{\mu = \mu_X - \mu_Y \quad \text{and} \quad \sigma^2 = \sigma_X^2 + \sigma_Y^2.}$$

The hypothesis (9.1) can also be expressed as:

$$\boldsymbol{\mathrm{H}_0 : \mu = 0.}$$

Therefore, we should use the test statistic $T = \sqrt{n}\bar{Z}/S$, where $\bar{Z}$ and S^2 denote, respectively, the sample mean and the sample variance of $\{Z_1, Z_2, \ldots, Z_n\}$.

At the $100\alpha\%$ significance level, for $\alpha \in (0, 1)$, we reject the hypothesis $\mu_X = \mu_Y$ when:

- $|t| > t_{\alpha/2, n-1}$, if the alternative is $\mathrm{H}_1 : \mu_X \neq \mu_Y$
- $t > t_{\alpha, n-1}$, if the alternative is $\mathrm{H}_1 : \mu_X > \mu_Y$
- $t < -t_{\alpha, n-1}$, if the alternative is $\mathrm{H}_1 : \mu_X < \mu_Y$

where $P(T > t_{\alpha, n-1}) = \alpha$, for $T \sim t_{n-1}$.

9.11.1 Power functions of the test

Consider the case of testing $\mathrm{H}_0 : \mu_X = \mu_Y$ vs. $\mathrm{H}_1 : \mu_X > \mu_Y$ only. For $\mu = \mu_X - \mu_Y > 0$, we have:

$$\begin{aligned}
\boldsymbol{\beta(\mu)} &= \boldsymbol{P_\mu(\mathrm{H}_0 \text{ is rejected})} \\
&= \boldsymbol{P_\mu(T > t_{\alpha, n-1})} \\
&= \boldsymbol{P_\mu\left(\frac{\sqrt{n}\bar{Z}}{S} > t_{\alpha, n-1}\right)} \\
&= \boldsymbol{P_\mu\left(\frac{\sqrt{n}(\bar{Z} - \mu)}{S} > t_{\alpha, n-1} - \frac{\sqrt{n}\mu}{S}\right)}
\end{aligned}$$

where $\sqrt{n}(\bar{Z} - \mu)/S \sim t_{n-1}$ under the distribution represented by P_μ.

Note that for $\mu > 0$, $\beta(\mu) > \alpha$. Furthermore, $\beta(\mu)$ increases as μ increases.

9.12 Comparing two normal means

Let $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ be two *independent* random samples drawn from, respectively, $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. We seek to test hypotheses on $\mu_X - \mu_Y$.

We cannot pair the two samples together, because of the different sample sizes n and m.

Let the sample means be $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{Y} = \sum_{i=1}^{m} Y_i/m$, and the sample variances be:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \textbf{and} \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

Some remarks are the following.

- $\bar{X}$, $\bar{Y}$, S_X^2 and S_Y^2 are independent.
- $\bar{X} \sim N(\mu_X, \sigma_X^2/n)$ and $(n-1)S_X^2/\sigma_X^2 \sim \chi_{n-1}^2$.
- $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/m)$ and $(m-1)S_Y^2/\sigma_Y^2 \sim \chi_{m-1}^2$.

Hence $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$. If $\sigma_X^2 = \sigma_Y^2$, then:

$$\frac{(\bar{X} - \bar{Y} - (\mu_X - \mu_Y))/\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}{\sqrt{((n-1)S_X^2/\sigma_X^2 + (m-1)S_Y^2/\sigma_Y^2)/(n+m-2)}}$$

$$= \sqrt{\frac{n+m-2}{1/n + 1/m}} \times \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sim t_{n+m-2}.$$

9.12.1 Tests on $\mu_X - \mu_Y$ with known σ_X^2 and σ_Y^2

Suppose we are interested in testing:

$$\mathrm{H}_0 : \mu_X = \mu_Y \quad \text{vs.} \quad \mathrm{H}_1 : \mu_X \neq \mu_Y.$$

Note that:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0,1).$$

Under H_0, $\mu_X - \mu_Y = 0$, so we have:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0,1).$$

At the $100\alpha\%$ significance level, for $\alpha \in (0,1)$, we reject H_0 if $|t| > z_{\alpha/2}$, where $P(Z > z_{\alpha/2}) = \alpha/2$, for $Z \sim N(0,1)$.

A $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is:

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \times \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

9.12.2 Tests on $\mu_X - \mu_Y$ with $\sigma_X^2 = \sigma_Y^2$ but unknown

This time we consider the following hypotheses:

$$\mathrm{H}_0 : \mu_X - \mu_Y = \delta_0 \quad \text{vs.} \quad \mathrm{H}_1 : \mu_X - \mu_Y > \delta_0$$

where δ_0 is a given constant. Under H_0, we have:

$$\boldsymbol{T = \sqrt{\frac{n+m-2}{1/n+1/m}} \times \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sim t_{n+m-2}.}$$

At the $100\alpha\%$ significance level, for $\alpha \in (0,1)$, we reject H_0 if $t > t_{\alpha, n+m-2}$, where $P(T > t_{\alpha, n+m-2}) = \alpha$, for $T \sim t_{n+m-2}$.

A $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is:

$$\boldsymbol{\bar{X} - \bar{Y} \pm t_{\alpha/2, n+m-2} \times \sqrt{\frac{1/n+1/m}{n+m-2}((n-1)S_X^2 + (m-1)S_Y^2).}}$$

Example 9.10 Two types of razor, A and B, were compared using 100 men in an experiment. Each man shaved one side, chosen at random, of his face using one razor and the other side using the other razor. The times taken to shave, X_i and Y_i minutes, for $i = 1, 2, \ldots, 100$, corresponding to the razors A and B, respectively, were recorded, yielding:

$$\bar{x} = 2.84, \quad s_X^2 = 0.48, \quad \bar{y} = 3.02 \quad \text{and} \quad s_Y^2 = 0.42.$$

Also available is the sample variance of the differences, $Z_i = X_i - Y_i$, which is $s_Z^2 = 0.6$.

Test, at the 5% significance level, if the two razors lead to different mean shaving times. State clearly any assumptions used in the test.

Assumption: Suppose $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_n\}$ are two independent random samples from, respectively, $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$.

The problem requires us to test the following hypotheses:

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X \neq \mu_Y.$$

There are three approaches – a paired comparison method and two two-sample comparisons based on different assumptions. Since the data are recorded in pairs, the paired comparison is most relevant and effective to analyse these data.

Method I: paired comparison

We have $Z_i = X_i - Y_i \sim N(\mu_Z, \sigma_Z^2)$ with $\mu_Z = \mu_X - \mu_Y$ and $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$. We want to test:

$$H_0 : \mu_Z = 0 \quad \text{vs.} \quad H_1 : \mu_Z \neq 0.$$

This is the standard one-sample t test, where:

$$\frac{\sqrt{n}(\bar{Z} - \mu_Z)}{S_Z} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_Z/\sqrt{n}} \sim t_{n-1}.$$

H_0 is rejected if $|t| > t_{0.025,\,99} = 1.98$, where under H_0 we have:

$$T = \frac{\sqrt{n}\bar{Z}}{S_Z} = \frac{\sqrt{100} \times (\bar{X} - \bar{Y})}{S_Z}.$$

With the given data, we observe $t = 10 \times (2.84 - 3.02)/\sqrt{0.6} = -2.327$. Hence we reject the hypothesis that the two razors lead to the same mean shaving time at the 5% significance level.

A 95% confidence interval for $\mu_X - \mu_Y$ is:

$$\bar{x} - \bar{y} \pm t_{0.025,\,n-1} \times \frac{s_Z}{\sqrt{n}} = -0.18 \pm 0.154 \quad \Rightarrow \quad (-0.334, -0.026).$$

Some remarks are the following.

i. Zero is not in the confidence interval for $\mu_X - \mu_Y$.

ii. $t_{0.025,\,99} = 1.98$ is pretty close to $z_{0.025} = 1.96$.

Method II: two-sample comparison with known variances

A further assumption is that $\sigma_X^2 = 0.48$ and $\sigma_Y^2 = 0.42$.

Note $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma_X^2/100 + \sigma_Y^2/100)$, i.e. we have:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/100 + \sigma_Y^2/100}} \sim N(0, 1).$$

Hence we reject H_0 when $|t| > 1.96$ at the 5% significance level, where:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/100 + \sigma_Y^2/100}}.$$

For the given data, $t = -0.18/\sqrt{0.009} = -1.90$. Hence we *cannot* reject H_0.

A 95% confidence interval for $\mu_X - \mu_Y$ is:

$$\bar{x} - \bar{y} \pm 1.96 \times \sqrt{\frac{\sigma_X^2}{100} + \frac{\sigma_Y^2}{100}} = -0.18 \pm 0.186 \quad \Rightarrow \quad (-0.366, 0.006).$$

The value 0 is now contained in the confidence interval.

Method III: two-sample comparison with equal but unknown variance

A different additional assumption is that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

Now $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma^2/50)$ and $99(S_X^2 + S_Y^2)/\sigma^2 \sim \chi^2_{198}$. Hence:

$$\frac{\sqrt{50} \times (\bar{X} - \bar{Y} - (\mu_X - \mu_Y))}{\sqrt{99 \times (S_X^2 + S_Y^2)/198}} = 10 \times \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2 + S_Y^2}} \sim t_{198}.$$

Hence we reject H_0 if $|t| > t_{0.025,\,198} = 1.97$ where:

$$T = \frac{10 \times (\bar{X} - \bar{Y})}{\sqrt{S_X^2 + S_Y^2}}.$$

For the given data, $t = -1.897$. Hence we *cannot* reject H_0 at the 5% significance level.

A 95% confidence interval for $\mu_X - \mu_Y$ is:

$$\bar{x} - \bar{y} \pm t_{0.025,\,198} \times \sqrt{\frac{s_X^2 + s_Y^2}{100}} = -0.18 \pm 0.1870 \quad \Rightarrow \quad (-0.367, 0.007)$$

which contains 0.

Some remarks are the following.

i. Different methods lead to different but not contradictory conclusions, as remember:

$$\textbf{not reject} \neq \textbf{accept.}$$

ii. The paired comparison is intuitively the most relevant, requires the least assumptions, and leads to the most conclusive inference (i.e. rejection of H_0). It also produces the narrowest confidence interval.

iii. Methods II and III ignore the pairing of the data. Consequently, the inference is less conclusive and less accurate.

iv. A general observation is that H_0 is rejected at the $100\alpha\%$ significance level if and only if the value hypothesised by H_0 is not within the corresponding $100(1-\alpha)\%$ confidence interval.

v. It is much more challenging to compare two normal means with unknown and unequal variances. This will not be discussed in this course.

9.13 Tests for correlation coefficients

We now consider a test for the correlation coefficient of two random variables X and Y where:

$$\rho = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{(\mathrm{Var}(X)\,\mathrm{Var}(Y))^{1/2}} = \frac{\mathrm{E}((X-\mathrm{E}(X))(Y-\mathrm{E}(Y)))}{(\mathrm{E}((X-\mathrm{E}(X))^2)\,\mathrm{E}((Y-\mathrm{E}(Y))^2))^{1/2}}.$$

Some remarks are the following.

i. $\rho \in [-1,1]$, and $|\rho| = 1$ if and only if $Y = aX + b$ for some constants a and b. Furthermore, $a > 0$ if $\rho = 1$, and $a < 0$ if $\rho = -1$.

ii. ρ measures only the *linear relationship* between X and Y. When $\rho = 0$, X and Y are *linearly independent*, that is *uncorrelated*.

iii. If X and Y are independent (in the sense that the joint pdf is the product of the two marginal pdfs), $\rho = 0$. However, if $\rho = 0$, X and Y are not necessarily independent, as there may exist some *non-linear* relationship between X and Y.

iv. If $\rho > 0$, X and Y tend to increase (or decrease) together. If $\rho < 0$, X and Y tend to move in opposite directions.

Sample correlation coefficient

Given paired observations (X_i, Y_i), for $i = 1, 2, \ldots, n$, a natural estimator of ρ is defined as:

$$\widehat{\rho} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \right)^{1/2}}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$.

Example 9.11 The measurements of height, X, and weight, Y, are taken from 69 students in a class. ρ should be positive, intuitively!

In Figure 9.5, the vertical line at $\bar{x}$ and the horizontal line at $\bar{y}$ divide the 69 points into 4 quadrants: northeast (NE), southwest (SW), northwest (NW) and southeast (SE). Most points are in either NE or SW.

- In the NE quadrant, $x_i > \bar{x}$ and $y_i > \bar{y}$, hence:
$$\sum_{i \in \text{NE}} (x_i - \bar{x})(y_i - \bar{y}) > 0.$$
- In the SW quadrant, $x_i < \bar{x}$ and $y_i < \bar{y}$, hence:
$$\sum_{i \in \text{SW}} (x_i - \bar{x})(y_i - \bar{y}) > 0.$$
- In the NW quadrant, $x_i < \bar{x}$ and $y_i > \bar{y}$, hence:
$$\sum_{i \in \text{NW}} (x_i - \bar{x})(y_i - \bar{y}) < 0.$$
- In the SE quadrant, $x_i > \bar{x}$ and $y_i < \bar{y}$, hence:
$$\sum_{i \in \text{SE}} (x_i - \bar{x})(y_i - \bar{y}) < 0.$$
- Overall:
$$\sum_{i=1}^{69} (x_i - \bar{x})(y_i - \bar{y}) > 0$$
and hence $\widehat{\rho} > 0$.

Figure 9.6 shows examples of different sample correlation coefficients using scatterplots of bivariate observations.

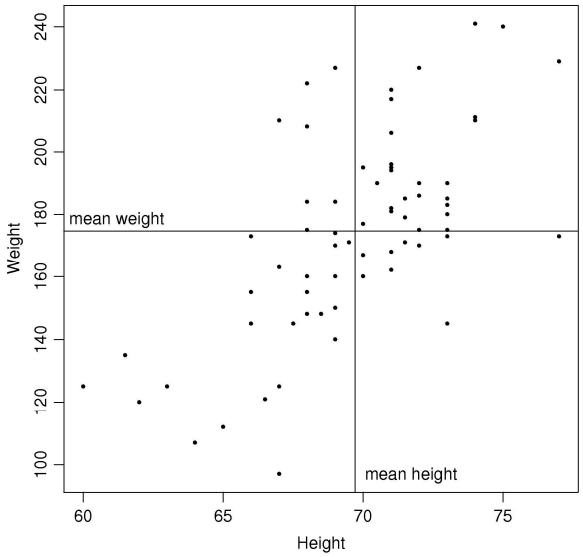


Figure 9.5: Scatterplot of height and weight in Example 9.11.

9.13.1 Tests for correlation coefficients

Let $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be a random sample from a two-dimensional normal distribution. Let $\rho = \text{Corr}(X_i, Y_i)$. We are interested in testing:

$$\mathbf{H_0 : \rho = 0 \quad vs. \quad H_1 : \rho \neq 0.}$$

It can be shown that under H_0 the test statistic is:

$$\boldsymbol{T = \widehat{\rho}\sqrt{\frac{n-2}{1-\widehat{\rho}^2}} \sim t_{n-2}.}$$

Hence we reject H_0 at the $100\alpha\%$ significance level, for $\alpha \in (0, 1)$, if $|t| > t_{\alpha/2, n-2}$, where:

$$\boldsymbol{P(T > t_{\alpha/2, n-2}) = \frac{\alpha}{2}.}$$

Some remarks are the following.

i. $|T| = |\widehat{\rho}|\sqrt{(n-2)/(1-\widehat{\rho}^2)}$ increases as $|\widehat{\rho}|$ increases.

ii. For $\text{H}_1 : \rho > 0$, we reject H_0 if $t > t_{\alpha, n-2}$.

iii. Two random variables X and Y are jointly normal if $aX + bY$ is normal for any constants a and b.

iv. For jointly normal random variables X and Y, if $\text{Corr}(X, Y) = 0$, X and Y are also *independent*.

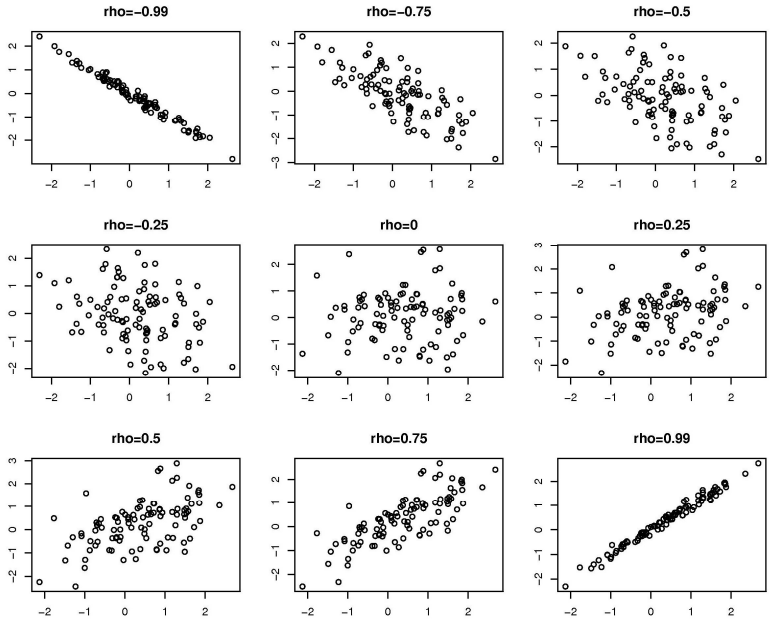


Figure 9.6: Scatterplots of bivariate observations with different sample correlation coefficients.

9.14 Tests for the ratio of two normal variances

Let $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ be two independent random samples from, respectively, $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. We are interested in testing:

$$\mathbf{H_0 : \frac{\sigma_Y^2}{\sigma_X^2} = k \quad vs. \quad H_1 : \frac{\sigma_Y^2}{\sigma_X^2} \neq k}$$

where $k > 0$ is a given constant. The case with $k = 1$ is of particular interest since this tests for equal variances.

Let the sample means be $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{Y} = \sum_{i=1}^{m} Y_i/m$, and the sample variances be:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

We have $(n-1)S_X^2/\sigma_X^2 \sim \chi_{n-1}^2$ and $(m-1)S_Y^2/\sigma_Y^2 \sim \chi_{m-1}^2$. Therefore:

$$\mathbf{\frac{\sigma_Y^2}{\sigma_X^2} \times \frac{S_X^2}{S_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.}$$

Under H_0, $T = kS_X^2/S_Y^2 \sim F_{n-1,m-1}$. Hence H_0 is rejected if:

$$\mathbf{t < F_{1-\alpha/2,n-1,m-1} \quad or \quad t > F_{\alpha/2,n-1,m-1}}$$

where $F_{\alpha,p,k}$ denotes the top 100αth percentile of the $F_{p,k}$ distribution, that is:

$$\mathbf{P(T > F_{\alpha,p,k}) = \alpha}$$

available from Table 9 of Murdoch and Barnes' *Statistical Tables*.

Since:

$$\boldsymbol{P\left(F_{1-\alpha/2,n-1,m-1} \leq \frac{\sigma_Y^2}{\sigma_X^2} \times \frac{S_X^2}{S_Y^2} \leq F_{\alpha/2,n-1,m-1}\right) = 1-\alpha}$$

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

$$\boldsymbol{\left(F_{1-\alpha/2,n-1,m-1} \times \frac{S_Y^2}{S_X^2}, F_{\alpha/2,n-1,m-1} \times \frac{S_Y^2}{S_X^2}\right).}$$

Example 9.12 Here we practise use of Table 9 of Murdoch and Barnes' *Statistical Tables* to obtain critical values for the F distribution.

Table 9 can be used to find the top 100αth percentile of the F_{ν_1,ν_2} distribution for $\alpha = 0.05$, 0.025, 0.01 and 0.001.

For example, for $\nu_1 = 3$ and $\nu_2 = 5$, then:

$$P(F_{3,5} > 5.41) = 0.05$$

$$P(F_{3,5} > 7.76) = 0.025$$

$$P(F_{3,5} > 12.06) = 0.01$$

and:

$$P(F_{3,5} > 33.20) = 0.001.$$

To find the bottom 100αth percentile, we note that $F_{1-\alpha,\nu_1,\nu_2} = 1/F_{\alpha,\nu_2,\nu_1}$. So, for $\nu_1 = 3$ and $\nu_2 = 5$, we have:

$$P\left(F_{3,5} < \frac{1}{F_{0.05,5,3}} = \frac{1}{9.01} = 0.111\right) = 0.05$$

$$P\left(F_{3,5} < \frac{1}{F_{0.025,5,3}} = \frac{1}{14.90} = 0.067\right) = 0.025$$

$$P\left(F_{3,5} < \frac{1}{F_{0.01,5,3}} = \frac{1}{28.20} = 0.035\right) = 0.01$$

and:

$$P\left(F_{3,5} < \frac{1}{F_{0.001,5,3}} = \frac{1}{134.60} = 0.007\right) = 0.001.$$

Example 9.13 The daily returns (in percentages) of two assets, X and Y, are recorded over a period of 100 trading days, yielding average daily returns of $\bar{x} = 3.21$ and $\bar{y} = 1.41$. Also available from the data are the following quantities:

$$\sum_{i=1}^{100} x_i^2 = 1{,}989.24, \quad \sum_{i=1}^{100} y_i^2 = 932.78 \quad \text{and} \quad \sum_{i=1}^{100} x_i y_i = 661.11.$$

Assume the data are normally distributed. Are the two assets positively correlated with each other, and is asset X riskier than asset Y?

With $n = 100$ we have:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 9.69$$

and:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = 7.41.$$

Therefore:

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_X s_Y} = 0.249.$$

First we test:

$$\mathrm{H}_0 : \rho = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \rho > 0.$$

Under H_0, the test statistic is:

$$T = \widehat{\rho}\sqrt{\frac{n-2}{1-\widehat{\rho}^2}} \sim t_{98}.$$

Setting $\alpha = 0.01$, we reject H_0 if $t > t_{0.01,98} = 2.37$. With the given data, $t = 2.545$ hence we reject the null hypothesis of $\rho = 0$ at the 1% significance level. We conclude that there is highly significant evidence indicating that the two assets are positively correlated.

We measure the risks in terms of variances, and test:

$$\mathrm{H}_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs.} \quad \mathrm{H}_1 : \sigma_X^2 > \sigma_Y^2.$$

Under H_0, we have that:

$$T = \frac{S_X^2}{S_Y^2} \sim F_{99,99}.$$

Hence we reject H_0 if $t > F_{0.05,99,99} = 1.39$ at the 5% significance level, using Table 9 of Murdoch and Barnes' *Statistical Tables*.

With the given data, $t = 9.69/7.41 = 1.308$. Therefore, we cannot reject H_0. As the test is not significant at the 5% significance level, we may not conclude that the variances of the two assets are significantly different. Therefore, there is no significant evidence indicating that asset X is riskier than asset Y.

Strictly speaking, the test is valid only if the two samples are independent of each other, which is not the case here.

9.15 Summary: tests for two normal distributions

Let $(X_1, X_2, \ldots, X_n) \sim_{IID} N(\mu_X, \sigma_X^2)$, $(Y_1, Y_2, \ldots, Y_m) \sim_{IID} N(\mu_Y, \sigma_Y^2)$, and $\rho = \text{Corr}(X, Y)$.

A summary table of tests for two normal distributions is:

Null hypothesis, H_0	$\mu_X - \mu_Y = \delta$ (σ_X^2, σ_Y^2 known)	$\mu_X - \mu_Y = \delta$ ($\sigma_X^2 = \sigma_Y^2$ unknown)	$\rho = 0$ ($n = m$)	$\frac{\sigma_Y^2}{\sigma_X^2} = k$
Test statistic, T	$\frac{\bar{X}-\bar{Y}-\delta}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}$	$\sqrt{\frac{n+m-2}{1/n+1/m}} \times \frac{\bar{X}-\bar{Y}-\delta}{\sqrt{(n-1)S_X^2+(m-1)S_Y^2}}$	$\widehat{\rho}\sqrt{\frac{n-2}{1-\widehat{\rho}^2}}$	$k\frac{S_X^2}{S_Y^2}$
Distribution of T under H_0	$N(0,1)$	t_{n+m-2}	t_{n-2}	$F_{n-1,m-1}$

9.16 Overview of chapter

This chapter has discussed hypothesis tests for parameters of normal distributions – specifically means and variances. In each case an appropriate test statistic was constructed whose distribution under the null hypothesis was known. Concepts of hypothesis testing errors and power were also discussed, as well as how to test correlation coefficients.

9.17 Key terms and concepts

- Alternative hypothesis
- Decision
- *p*-value
- Power function
- *t* test
- Type I error
- Critical value
- Null hypothesis
- Paired comparison
- Significance level
- Test statistic
- Type II error

To p, or not to p?
(James Abdey, Ph.D. Thesis 2009.[1])

[1] Available at `http://etheses.lse.ac.uk/31`

Chapter 10
Analysis of variance (ANOVA)

10.1 Synopsis of chapter

This chapter introduces analysis of variance (ANOVA) which is a widely-used technique for detecting differences between groups based on continuous dependent variables.

10.2 Learning outcomes

After completing this chapter, you should be able to:

- explain the purpose of analysis of variance
- restate and interpret the models for one-way and two-way analysis of variance
- conduct small examples of one-way and two-way analysis of variance with a calculator, reporting the results in an ANOVA table
- perform hypothesis tests and construct confidence intervals for one-way and two-way analysis of variance
- explain how to interpret residuals from an analysis of variance.

10.3 Introduction

Analysis of variance (ANOVA) is a popular tool which has an applicability and power which we can only start to appreciate in this course. The idea of analysis of variance is to investigate how variation in structured data can be split into pieces associated with components of that structure. We look only at one-way and two-way classifications, providing tests and confidence intervals which are widely used in practice.

10.4 Testing for equality of three population means

We begin with an illustrative example to test the hypothesis that three populations means are equal.

Example 10.1 To assess the teaching quality of class teachers, a random sample of 6 examination marks was selected from each of three classes. The examination marks for each class are listed in the table below.

Can we infer from these data that there is no significant difference in the examination marks among all three classes?

Class 1	Class 2	Class 3
85	71	59
75	75	64
82	73	62
76	74	69
71	69	75
85	82	67

Suppose examination marks from Class j follow the distribution $N(\mu_j, \sigma^2)$, for $j = 1, 2, 3$. So we assume examination marks are normally distributed with the same variance in each class, but possibly different means.

We need to test the hypothesis:

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3.$$

The data form a 6×3 array. Denote the data point at the (i, j)th position as X_{ij}. We compute the column means first where the jth column mean is:

$$\bar{X}_{\cdot j} = \frac{X_{1j} + X_{2j} + \cdots + X_{n_j j}}{n_j}$$

where n_j is the sample size of group j (here $n_j = 6$ for all j).

This leads to $\bar{x}_{\cdot 1} = 79$, $\bar{x}_{\cdot 2} = 74$ and $\bar{x}_{\cdot 3} = 66$. Transposing the table, we get:

	Observation						
	1	2	3	4	5	6	Mean
Class 1	85	75	82	76	71	85	79
Class 2	71	75	73	74	69	82	74
Class 3	59	64	62	69	75	67	66

Note that similar problems arise from other practical situations. For example:

- comparing the returns of three stocks
- comparing sales using three advertising strategies
- comparing the effectiveness of three medicines.

If H_0 is true, the three observed sample means $\bar{x}_{\cdot 1}$, $\bar{x}_{\cdot 2}$ and $\bar{x}_{\cdot 3}$ should be very close to each other, i.e. all of them should be close to the overall sample mean, $\bar{x}$, which is:

$$\bar{x} = \frac{\bar{x}_{\cdot 1} + \bar{x}_{\cdot 2} + \bar{x}_{\cdot 3}}{3} = \frac{79 + 74 + 66}{3} = 73$$

i.e. the mean value of all 18 observations.

So we wish to perform a hypothesis test based on the variation in the sample means such that the greater the variation, the more likely we are to reject H_0. One possible measure for the variation in the sample means $\bar{X}_{\cdot j}$ about the overall sample mean $\bar{X}$, for $j = 1, 2, 3$, is:

$$\sum_{j=1}^{3} (\bar{X}_{\cdot j} - \bar{X})^2. \tag{10.1}$$

However, (10.1) is *not scale-invariant*, so it would be difficult to judge whether the realised value is large enough to warrant rejection of H_0 due to the magnitude being dependent on the units of measurement of the data. So we seek a *scale-invariant test statistic*.

Just as we scaled the covariance between two random variables to give the scale-invariant correlation coefficient, we can similarly scale (10.1) to give the following possible test statistic:

$$T = \frac{\sum_{j=1}^{3} (\bar{X}_{\cdot j} - \bar{X})^2}{\text{sum of the three sample variances}}.$$

Hence we would reject H_0 for large values of T. (Note $t = 0$ if $\bar{x}_{\cdot 1} = \bar{x}_{\cdot 2} = \bar{x}_{\cdot 3}$ which would mean that there is no variation at all between the sample means. In this case all the sample means would equal $\bar{x}$.)

It remains to determine the distribution of T under H_0.

10.5 One-way analysis of variance

We now extend Example 10.1 to consider a general setting where there are k independent random samples available from k normal distributions $N(\mu_j, \sigma^2)$, for $j = 1, 2, \ldots, k$. (Example 10.1 corresponds to $k = 3$.)

Denote by $X_{1j}, X_{2j}, \ldots, X_{n_j j}$ the random sample with sample size n_j from $N(\mu_j, \sigma^2)$, for $j = 1, 2, \ldots, k$.

Our goal is to test:

$$\mathbf{H_0 : \mu_1 = \mu_2 = \cdots = \mu_k}$$

vs.

$$\mathbf{H_1 : \text{not all } \mu_j\text{s are the same.}}$$

One-way analysis of variance (one-way ANOVA) involves a continuous dependent variable and one categorical independent variable (sometimes called a factor, or treatment), where the k different levels of the categorical variable are the k different groups.

We now introduce statistics associated with one-way ANOVA.

Statistics associated with one-way ANOVA

The jth sample mean is:

$$\bar{X}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}.$$

The overall sample mean is:

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij} = \frac{1}{n} \sum_{j=1}^{k} n_j \bar{X}_{\cdot j}$$

where $n = \sum_{j=1}^{k} n_j$ is the total number of observations across all k groups.

The total variation is:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

with $n - 1$ degrees of freedom.

The between-groups variation is:

$$B = \sum_{j=1}^{k} n_j (\bar{X}_{\cdot j} - \bar{X})^2$$

with $k - 1$ degrees of freedom.

The within-groups variation is:

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2$$

with $n - k = \sum_{j=1}^{k} (n_j - 1)$ degrees of freedom.

The ANOVA decomposition is:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \sum_{j=1}^{k} n_j (\bar{X}_{\cdot j} - \bar{X})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\cdot j})^2.$$

We have already discussed the jth sample mean and overall sample mean. The total variation is a measure of the overall (total) variability in the data from all k groups about the overall sample mean. The ANOVA decomposition decomposes this into two components: between-groups variation (which is attributable to the factor level) and within-groups variation (which is attributable to the variation within each group and is assumed to be the same σ^2 for each group).

Some remarks are the following.

i. B and W are also called, respectively, **between-treatments variation** and

within-treatments variation. In fact W is effectively a *residual (error) sum of squares*, representing the variation which cannot be explained by the treatment or group factor.

ii. The ANOVA decomposition follows from the identity:

$$\sum_{i=1}^{m}(a_i - b)^2 = \sum_{i=1}^{m}(a_i - \bar{a})^2 + m(\bar{a} - b)^2.$$

However, the actual derivation is not required for this course.

iii. The following are some useful formulae for manual computations.

- $n = \sum_{j=1}^{k} n_j$.
- $\bar{X}_{\cdot j} = \sum_{i=1}^{n_j} X_{ij}/n_j$ and $\bar{X} = \sum_{j=1}^{k} n_j\bar{X}_{\cdot j}/n$.
- Total variation = Total SS $= B + W = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - n\bar{X}^2$.
- $B = \sum_{j=1}^{k} n_j\bar{X}_{\cdot j}^2 - n\bar{X}^2$.
- Residual (Error) SS $= W = \sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k} n_j\bar{X}_{\cdot j}^2 = \sum_{j=1}^{k}(n_j - 1)S_j^2$ where S_j^2 is the jth sample variance.

We now note, without proof, the following results.

i. $B = \sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X})^2$ and $W = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\cdot j})^2$ are independent of each other.

ii. $W/\sigma^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\cdot j})^2/\sigma^2 \sim \chi^2_{n-k}$.

iii. Under $\mathrm{H}_0 : \mu_1 = \cdots = \mu_k$, then $B/\sigma^2 = \sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X})^2/\sigma^2 \sim \chi^2_{k-1}$.

In order to test $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, we define the following test statistic:

$$\boldsymbol{F = \frac{\sum_{j=1}^{k} n_j(\bar{X}_{\cdot j} - \bar{X})^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\cdot j})^2/(n-k)} = \frac{B/(k-1)}{W/(n-k)}.}$$

Under H_0, $F \sim F_{k-1,n-k}$. We reject H_0 at the $100\alpha\%$ significance level if:

$$\boldsymbol{f > F_{\alpha,k-1,n-k}}$$

where $F_{\alpha,k-1,n-k}$ is the top 100αth percentile of the $F_{k-1,n-k}$ distribution, i.e. $P(F > F_{\alpha,k-1,n-k}) = \alpha$, and f is the observed test statistic value.

The p-value of the test is:

$$\boldsymbol{p\text{-value} = P(F > f).}$$

It is clear that $f > F_{\alpha,k-1,n-k}$ if and only if the p-value $< \alpha$, as we must reach the same conclusion regardless of whether we use the critical value approach or the p-value approach to hypothesis testing.

One-way ANOVA table

Typically, one-way ANOVA results are presented in a table as follows:

Source	DF	SS	MS	F	p-value
Factor	$k-1$	B	$B/(k-1)$	$\frac{B/(k-1)}{W/(n-k)}$	p
Error	$n-k$	W	$W/(n-k)$		
Total	$n-1$	$B+W$			

Example 10.2 Continuing with Example 10.1, for the given data, $k = 3$, $n_1 = n_2 = n_3 = 6$, $n = n_1 + n_2 + n_3 = 18$, $\bar{x}_{\cdot 1} = 79$, $\bar{x}_{\cdot 2} = 74$, $\bar{x}_{\cdot 3} = 66$ and $\bar{x} = 73$. The sample variances are calculated to be $s_1^2 = 34$, $s_2^2 = 20$ and $s_3^2 = 32$. Therefore:

$$b = \sum_{j=1}^{3} 6(\bar{x}_{\cdot j} - \bar{x})^2 = 6 \times ((79-73)^2 + (74-73)^2 + (66-73)^2) = 516$$

and:

$$\begin{aligned} w = \sum_{j=1}^{3}\sum_{i=1}^{6}(x_{ij} - \bar{x}_{\cdot j})^2 &= \sum_{j=1}^{3}\sum_{i=1}^{6} x_{ij}^2 - 6\sum_{j=1}^{3}\bar{x}_{\cdot j}^2 \\ &= \sum_{j=1}^{3} 5s_j^2 \\ &= 5 \times (34 + 20 + 32) \\ &= 430. \end{aligned}$$

Hence:

$$f = \frac{b/(k-1)}{w/(n-k)} = \frac{516/2}{430/15} = 9.$$

Under $H_0 : \mu_1 = \mu_2 = \mu_3$, $F \sim F_{k-1,n-k} = F_{2,15}$. Since $F_{0.01,2,15} = 6.359 < 9$, using Table 9 of Murdoch and Barnes' *Statistical Tables*, we reject H_0 at the 1% significance level. In fact the p-value (using a computer) is $P(F > 9) = 0.003$. Therefore, we conclude that there is a significant difference among the mean examination marks across the three classes.

The one-way ANOVA table is as follows:

Source	DF	SS	MS	F	p-value
Class	2	516	258	9	0.003
Error	15	430	28.67		
Total	17	946			

Example 10.3 A study performed by a Columbia University professor counted the number of times per minute professors from three different departments said 'uh' or 'ah' during lectures to fill gaps between words. The data listed in 'UhAh.csv' were derived from observing 100 minutes from each of the three departments. If we assume that the more frequent use of 'uh' or 'ah' results in more boring lectures, can we conclude that some departments' professors are more boring than others?

The counts for English, Mathematics and Political Science departments are stored. As always in statistical analysis, we first look at the summary (descriptive) statistics of these data.

```
> attach(UhAh)
> summary(UhAh)
   Frequency                 Department
 Min.   : 0.00   English           :100
 1st Qu.: 4.00   Mathematics       :100
 Median : 5.00   Political Science:100
 Mean   : 5.48
 3rd Qu.: 7.00
 Max.   :11.00
> xbar <- tapply(Frequency, Department, mean)
> s <- tapply(Frequency, Department, sd)
> n <- tapply(Frequency, Department, length)
> sem <- s/sqrt(n)
> list(xbar,s,n,sem)
[[1]]
          English       Mathematics Political Science
             5.81              5.30              5.33

[[2]]
          English       Mathematics Political Science
         2.493203          2.012587          1.974867

[[3]]
          English       Mathematics Political Science
              100               100               100

[[4]]
          English       Mathematics Political Science
        0.2493203         0.2012587         0.1974867
```

Surprisingly, professors in English say 'uh' or 'ah' more on average than those in Mathematics and Political Science (compare the sample means of 5.81, 5.30 and 5.33), but the difference *seems* small. However, we need to formally test whether the (seemingly small) differences are statistically significant.

Using the data, R produces the following one-way ANOVA table:

```
> anova(lm(Frequency ~ Department))
Analysis of Variance Table

Response: Frequency
            Df  Sum Sq Mean Sq F value Pr(>F)
Department   2   16.38  8.1900  1.7344 0.1783
Residuals  297 1402.50  4.7222
```

Since the p-value for the F test is 0.1783, we cannot reject the following hypothesis:

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3.$$

Therefore, there is no evidence of a difference in the mean number of 'uh's or 'ah's said by professors across the three departments.

In addition to a one-way ANOVA table, we can also obtain the following.

- An estimator of σ is:
$$\boldsymbol{\widehat{\sigma} = S = \sqrt{\frac{W}{n-k}}.}$$
- 95% confidence intervals for μ_j are given by:
$$\boldsymbol{\bar{X}_{\cdot j} \pm t_{0.025,\, n-k} \times \frac{S}{\sqrt{n_j}} \quad \text{for } j = 1, 2, \ldots, k}$$
where $t_{0.025,\, n-k}$ is the top 2.5th percentile of the Student's t_{n-k} distribution, which can be obtained from Table 7 of Murdoch and Barnes' *Statistical Tables*.

Example 10.4 Assuming a common variance for each group, from the preceding output in Example 10.3 we see that:

$$\widehat{\sigma} = s = \sqrt{\frac{1{,}402.50}{297}} = \sqrt{4.72} = 2.173.$$

Since $t_{0.025,\,297} \approx t_{0.025,\,\infty} = 1.96$, using Table 7 of Murdoch and Barnes' *Statistical Tables*, we obtain the following 95% confidence intervals for μ_1, μ_2 and μ_3, respectively:

$$j = 1: \quad 5.81 \pm 1.96 \times \frac{2.173}{\sqrt{100}} \quad \Rightarrow \quad (5.38, 6.24)$$

$$j = 2: \quad 5.30 \pm 1.96 \times \frac{2.173}{\sqrt{100}} \quad \Rightarrow \quad (4.87, 5.73)$$

$$j = 3: \quad 5.33 \pm 1.96 \times \frac{2.173}{\sqrt{100}} \quad \Rightarrow \quad (4.90, 5.76).$$

R can produce the following:

```
> stripchart(Frequency ~ Department,pch=16,vert=T)
> arrows(1:3,xbar+1.96*2.173/sqrt(n),1:3,xbar-1.96*2.173/sqrt(n),
angle=90,code=3,length=0.1)
> lines(1:3,xbar,pch=4,type="b",cex=2)
```

These 95% confidence intervals can be seen plotted in the R output below. Note that these confidence intervals all overlap, which is consistent with our failure to reject the null hypothesis that all population means are equal.

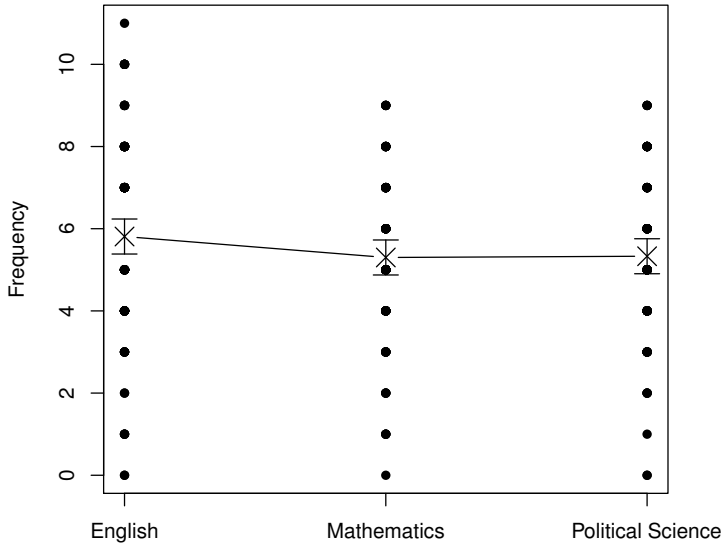


Figure 10.1: Overlapping confidence intervals.

Example 10.5 In early 2001, the American economy was slowing down and companies were laying off workers. A poll conducted during February 2001 asked a random sample of workers how long (in months) it would be before they faced significant financial hardship if they lost their jobs, with the data available in the file '`GallupPoll.csv`'. They are classified into four groups according to their incomes. Below is part of the R output of the descriptive statistics of the classified data. Can we infer that income group has a significant impact on the mean length of time before facing financial hardship?

```
   Hardship          Income.group
Min.   : 0.00   $20 to 30K: 81
1st Qu.: 8.00   $30 to 50K:114
Median :15.00   Over $50K : 39
Mean   :16.11   Under $20K: 67
3rd Qu.:22.00
Max.   :50.00
```

```
> xbar <- tapply(Hardship, Income.group, mean)
> s <- tapply(Hardship, Income.group, sd)
> n <- tapply(Hardship, Income.group, length)
> sem <- s/sqrt(n)
> list(xbar,s,n,sem)
[[1]]
$20 to 30K $30 to 50K  Over $50K Under $20K
 15.493827  18.456140  22.205128   9.313433

[[2]]
$20 to 30K $30 to 50K  Over $50K Under $20K
  9.233260   9.507464  11.029099   8.087043

[[3]]
$20 to 30K $30 to 50K  Over $50K Under $20K
        81        114         39         67

[[4]]
$20 to 30K $30 to 50K  Over $50K Under $20K
 1.0259178  0.8904556  1.7660693  0.9879896
```

Inspection of the sample means suggests that there is a difference between income groups, but we need to conduct a one-way ANOVA test to see whether the differences are statistically significant.

We apply one-way ANOVA to test whether the means in the $k = 4$ groups are equal, i.e. $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, from highest to lowest income groups.

We have $n_1 = 39$, $n_2 = 114$, $n_3 = 81$ and $n_4 = 67$, hence:

$$n = \sum_{j=1}^{k} n_j = 39 + 114 + 81 + 67 = 301.$$

Also $\bar{x}_{\cdot 1} = 22.21$, $\bar{x}_{\cdot 2} = 18.456$, $\bar{x}_{\cdot 3} = 15.49$, $\bar{x}_{\cdot 4} = 9.313$ and:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{X}_{\cdot j} = \frac{39 \times 22.21 + 114 \times 18.456 + 81 \times 15.49 + 67 \times 9.313}{301} = 16.109.$$

Now:

$$\begin{aligned} b &= \sum_{j=1}^{k} n_j (\bar{x}_{\cdot j} - \bar{x})^2 \\ &= 39 \times (22.21 - 16.109)^2 + 114 \times (18.456 - 16.109)^2 \\ &\quad + 81 \times (15.49 - 16.109)^2 + 67 \times (9.313 - 16.109)^2 \\ &= 5{,}205.097. \end{aligned}$$

We have $s_1^2 = (11.03)^2 = 121.661$, $s_2^2 = (9.507)^2 = 90.383$, $s_3^2 = (9.23)^2 = 85.193$ and

$s_4^2 = (8.087)^2 = 65.400$, hence:

$$\begin{aligned} w = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j})^2 &= \sum_{j=1}^{k}(n_j - 1)s_j^2 \\ &= 38 \times 121.661 + 113 \times 90.383 + 80 \times 85.193 + 66 \times 65.400 \\ &= 25{,}968.24. \end{aligned}$$

Consequently:

$$f = \frac{b/(k-1)}{w/(n-k)} = \frac{5{,}205.097/3}{25{,}968.24/(301-4)} = 19.84.$$

Under H_0, $F \sim F_{k-1,\,n-k} = F_{3,\,297}$. Since $F_{0.01,\,3,\,297} \approx 3.848 < 19.84$, we reject H_0 at the 1% significance level, i.e. there is strong evidence that income group has a significant impact on the mean length of time before facing financial hardship.

The pooled estimate of σ is:

$$s = \sqrt{w/(n-k)} = \sqrt{25{,}968.24/(301-4)} = 9.351.$$

A 95% confidence interval for μ_j is:

$$\bar{x}_{\cdot j} \pm t_{0.025,\,297} \times \frac{s}{\sqrt{n_j}} = \bar{x}_{\cdot j} \pm 1.96 \times \frac{9.351}{\sqrt{n_j}} = \bar{x}_{\cdot j} \pm \frac{18.328}{\sqrt{n_j}}.$$

Hence, for example, a 95% confidence interval for μ_1 is:

$$22.21 \pm \frac{18.328}{\sqrt{39}} \quad \Rightarrow \quad (19.28, 25.14)$$

and a 95% confidence interval for μ_4 is:

$$9.313 \pm \frac{18.328}{\sqrt{67}} \quad \Rightarrow \quad (7.07, 11.55).$$

Notice that these two confidence intervals do not overlap, which is consistent with our conclusion that there is a difference between the group means.

R output for the data is:

```
> anova(lm(Hardship ~ Income.group))
Analysis of Variance Table

Response: Hardship
              Df  Sum Sq Mean Sq F value    Pr(>F)
Income.group   3  5202.1 1734.03  19.828 9.636e-12 ***
Residuals    297 25973.3   87.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that minor differences are due to rounding errors in calculations.

10.6 From one-way to two-way ANOVA

One-way ANOVA: a review

We have independent observations $X_{ij} \sim N(\mu_j, \sigma^2)$ for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, k$. We are interested in testing:

$$\mathbf{H_0 : \mu_1 = \mu_2 = \cdots = \mu_k.}$$

The variation of the X_{ij}s is driven by a factor at different levels $\mu_1, \mu_2, \ldots, \mu_k$, in addition to random fluctuations (i.e. **random errors**). We test whether such a factor effect exists or not. We can model a one-way ANOVA problem as follows:

$$\boldsymbol{X_{ij} = \mu + \beta_j + \varepsilon_{ij} \quad \text{for } i = 1, 2, \ldots, n_j \text{ and } j = 1, 2, \ldots, k}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and the ε_{ij}s are independent. μ is the average effect and β_j is the factor (or treatment) effect at the jth level. Note that $\sum_{j=1}^{k} \beta_j = 0$. The null hypothesis (i.e. that the group means are all equal) can also be expressed as:

$$\mathbf{H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.}$$

10.7 Two-way analysis of variance

Two-way analysis of variance (two-way ANOVA) involves a continuous dependent variable and two categorical independent variables (factors). Two-way ANOVA models the observations as:

$$\boldsymbol{X_{ij} = \mu + \gamma_i + \beta_j + \varepsilon_{ij} \quad \text{for } i = 1, 2, \ldots, r \text{ and } j = 1, 2, \ldots, c}$$

where:

- μ represents the average effect
- $\beta_1, \beta_2, \ldots, \beta_c$ represent c different treatment (column) levels
- $\gamma_1, \gamma_2, \ldots, \gamma_r$ represent r different block (row) levels
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ and the ε_{ij}s are independent.

In total, there are $n = r \times c$ observations. We now consider the conditions to make the parameters μ, γ_i and β_j identifiable for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. The conditions are:

$$\boldsymbol{\gamma_1 + \gamma_2 + \cdots + \gamma_r = 0 \quad \text{and} \quad \beta_1 + \beta_2 + \cdots + \beta_c = 0.}$$

We will be interested in testing the following hypotheses.

- The 'no treatment (column) effect' hypothesis of $\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_c = 0$.
- The 'no block (row) effect' hypothesis of $\mathrm{H}_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_r = 0$.

We now introduce statistics associated with two-way ANOVA.

Statistics associated with two-way ANOVA

The sample mean at the ith block level is:

$$\bar{X}_{i\cdot} = \frac{1}{c}\sum_{j=1}^{c} X_{ij} \quad \text{for } i = 1, 2, \ldots, r.$$

The sample mean at the jth treatment level is:

$$\bar{X}_{\cdot j} = \sum_{i=1}^{r} \frac{1}{r} X_{ij} \quad \text{for } j = 1, 2, \ldots, c.$$

The overall sample mean is:

$$\bar{X} = \bar{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}.$$

The total variation (with $rc - 1$ degrees of freedom) is:

$$\text{Total SS} = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X})^2.$$

The between-blocks (rows) variation (with $r - 1$ degrees of freedom) is:

$$B_{\text{row}} = c\sum_{i=1}^{r}(\bar{X}_{i\cdot} - \bar{X})^2.$$

The between-treatments (columns) variation (with $c - 1$ degrees of freedom) is:

$$B_{\text{col}} = r\sum_{j=1}^{c}(\bar{X}_{\cdot j} - \bar{X})^2.$$

The residual (error) variation (with $(r - 1)(c - 1)$ degrees of freedom) is:

$$\text{Residual SS} = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2.$$

The (two-way) ANOVA decomposition is:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X})^2 = c\sum_{i=1}^{r}(\bar{X}_{i\cdot} - \bar{X})^2 + r\sum_{j=1}^{c}(\bar{X}_{\cdot j} - \bar{X})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2.$$

The total variation is a measure of the overall (total) variability in the data and the (two-way) ANOVA decomposition decomposes this into three components: between-blocks variation (which is attributable to the row factor level), between-treatments variation (which is attributable to the column factor level) and residual variation (which is attributable to the variation not explained by the row and column factors).

The following are some useful formulae for manual computations.

- Row sample means: $\bar{X}_{i\cdot} = \sum_{j=1}^{c} X_{ij}/c$, for $i = 1, 2, \ldots, r$.
- Column sample means: $\bar{X}_{\cdot j} = \sum_{i=1}^{r} X_{ij}/r$, for $j = 1, 2, \ldots, c$.
- Overall sample mean: $\bar{X} = \sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}/n = \sum_{i=1}^{r} \bar{X}_{i\cdot}/r = \sum_{j=1}^{c} \bar{X}_{\cdot j}/c$.
- Total SS $= \sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}^2 - rc\bar{X}^2$.
- Between-blocks (rows) variation: $B_{\text{row}} = c\sum_{i=1}^{r} \bar{X}_{i\cdot}^2 - rc\bar{X}^2$.
- Between-treatments (columns) variation: $B_{\text{col}} = r\sum_{j=1}^{c} \bar{X}_{\cdot j}^2 - rc\bar{X}^2$.
- Residual SS $=$ (Total SS) $- B_{\text{row}} - B_{\text{col}} = \sum_{i=1}^{r}\sum_{j=1}^{c} X_{ij}^2 - c\sum_{i=1}^{r} \bar{X}_{i\cdot}^2 - r\sum_{j=1}^{c} \bar{X}_{\cdot j}^2 + rc\bar{X}^2$.

In order to test the 'no block (row) effect' hypothesis of $\text{H}_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_r = 0$, the test statistic is defined as:

$$\boldsymbol{F = \frac{B_{\text{row}}/(r-1)}{(\text{Residual SS})/((r-1)(c-1))} = \frac{(c-1)B_{\text{row}}}{\text{Residual SS}}.}$$

Under H_0, $F \sim F_{r-1,\,(r-1)(c-1)}$. We reject H_0 at the $100\alpha\%$ significance level if:

$$\boldsymbol{f > F_{\alpha,\,r-1,\,(r-1)(c-1)}}$$

where $F_{\alpha,\,r-1,\,(r-1)(c-1)}$ is the top 100αth percentile of the $F_{r-1,\,(r-1)(c-1)}$ distribution, i.e. $P(F > F_{\alpha,\,r-1,\,(r-1)(c-1)}) = \alpha$, and f is the observed test statistic value.

The p-value of the test is:

$$\boldsymbol{p\text{-value} = P(F > f).}$$

In order to test the 'no treatment (column) effect' hypothesis of $\text{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_c = 0$, the test statistic is defined as:

$$\boldsymbol{F = \frac{B_{\text{col}}/(c-1)}{(\text{Residual SS})/((r-1)(c-1))} = \frac{(r-1)B_{\text{col}}}{\text{Residual SS}}.}$$

Under H_0, $F \sim F_{c-1,\,(r-1)(c-1)}$. We reject H_0 at the $100\alpha\%$ significance level if:

$$\boldsymbol{f > F_{\alpha,\,c-1,\,(r-1)(c-1)}.}$$

The p-value of the test is defined in the usual way.

Two-way ANOVA table

As with one-way ANOVA, two-way ANOVA results are presented in a table as follows:

Source	DF	SS	MS	F	p-value
Row factor	$r-1$	B_{row}	$B_{\text{row}}/(r-1)$	$\frac{(c-1)B_{\text{row}}}{\text{Residual SS}}$	p
Column factor	$c-1$	B_{col}	$B_{\text{col}}/(c-1)$	$\frac{(r-1)B_{\text{col}}}{\text{Residual SS}}$	p
Residual	$(r-1)(c-1)$	Residual SS	$\frac{\text{Residual SS}}{(r-1)(c-1)}$		
Total	$rc-1$	Total SS			

10.8 Residuals

Before considering an example of two-way ANOVA, we briefly consider **residuals**. Recall the original two-way ANOVA model:

$$\boldsymbol{X_{ij} = \mu + \gamma_i + \beta_j + \varepsilon_{ij}.}$$

We now decompose the observations as follows:

$$\boldsymbol{X_{ij} = \bar{X} + (\bar{X}_{i\cdot} - \bar{X}) + (\bar{X}_{\cdot j} - \bar{X}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})}$$

for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, where we have the following point estimators.

- $\widehat{\mu} = \bar{X}$ is the point estimator of μ.
- $\widehat{\gamma}_i = \bar{X}_{i\cdot} - \bar{X}$ is the point estimator of γ_i, for $i = 1, 2, \ldots, r$.
- $\widehat{\beta}_j = \bar{X}_{\cdot j} - \bar{X}$ is the point estimator of β_j, for $j = 1, 2, \ldots, c$.

It follows that the residual, i.e. the estimator of ε_{ij}, is:

$$\boldsymbol{\widehat{\varepsilon}_{ij} = X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}}$$

for $i = 1, 2, \ldots r$ and $j = 1, 2, \ldots, c$.

The two-way ANOVA model assumes $\varepsilon_{ij} \sim N(0, \sigma^2)$ and so, if the model structure is correct, then the $\widehat{\varepsilon}_{ij}$s should behave like independent $N(0, \sigma^2)$ random variables.

Example 10.6 The following table lists the percentage annual returns (calculated four times per annum) of the Common Stock Index at the New York Stock Exchange during 1981–85, available in the data file 'NYSE.csv'.

	1st quarter	2nd quarter	3rd quarter	4th quarter
1981	5.7	6.0	7.1	6.7
1982	7.2	7.0	6.1	5.2
1983	4.9	4.1	4.2	4.4
1984	4.5	4.9	4.5	4.5
1985	4.4	4.2	4.2	3.6

(a) Is the variability in returns from year to year statistically significant?

(b) Are returns affected by the quarter of the year?

Using two-way ANOVA, we test the no row effect hypothesis to answer (a), and test the no column effect hypothesis to answer (b). We have $r = 5$ and $c = 4$.

The row sample means are calculated using $\bar{X}_{i\cdot} = \sum_{j=1}^{c} X_{ij}/c$, which gives 6.375, 6.375, 4.4, 4.6 and 4.1, for $i = 1, 2, \ldots, 5$, respectively.

The column sample means are calculated using $\bar{X}_{\cdot j} = \sum_{i=1}^{r} X_{ij}/r$, which gives 5.34, 5.24, 5.22 and 4.88, for $j = 1, 2, 3, 4$, respectively.

The overall sample mean is $\bar{x} = \sum_{i=1}^{r} \bar{x}_{i\cdot}/r = 5.17$.

The sum of the squared observations is $\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 = 559.06$.

Hence we have the following.

$$\text{Total SS} = \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}^2 - rc\bar{x}^2 = 559.06 - 20 \times (5.17)^2 = 559.06 - 534.578 = 24.482.$$

$$b_{\text{row}} = c\sum_{i=1}^{r} \bar{x}_{i\cdot}^2 - rc\bar{x}^2 = 4 \times 138.6112 - 534.578 = 19.867.$$

$$b_{\text{col}} = r\sum_{j=1}^{c} \bar{x}_{\cdot j}^2 - rc\bar{x}^2 = 5 \times 107.036 - 534.578 = 0.602.$$

$$\text{Residual SS} = (\text{Total SS}) - b_{\text{row}} - b_{\text{col}} = 24.482 - 19.867 - 0.602 = 4.013.$$

To test the no row effect hypothesis $\text{H}_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_5 = 0$, the test statistic value is:

$$f = \frac{(c-1)b_{\text{row}}}{\text{Residual SS}} = \frac{3 \times 19.867}{4.013} = 14.852.$$

Under H_0, $F \sim F_{r-1,\,(r-1)(c-1)} = F_{4,\,12}$. Using Table 9 of Murdoch and Barnes' *Statistical Tables*, since $F_{0.01,\,4,\,12} = 5.412 < 14.852$, we reject H_0 at the 1%

significance level. We conclude that there is strong evidence that the return does depend on the year.

To test the no column effect hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, the test statistic value is:

$$f = \frac{(r-1)b_{\text{col}}}{\text{Residual SS}} = \frac{4 \times 0.602}{4.013} = 0.600.$$

Under H_0, $F \sim F_{c-1,\,(r-1)(c-1)} = F_{3,\,12}$. Since $F_{0.10,\,3,\,12} = 2.606 > 0.600$, we cannot reject H_0 even at the 10% significance level. Therefore, there is no significant evidence indicating that the return depends on the quarter.

The results may be summarised in a two-way ANOVA table as follows:

Source	DF	SS	MS	F	p-value
Year	4	19.867	4.967	14.852	< 0.01
Quarter	3	0.602	0.201	0.600	> 0.10
Residual	12	4.013	0.334		
Total	19	24.482			

We could also provide 95% confidence interval estimates for each block and treatment level by using the pooled estimator of σ^2, which is:

$$S^2 = \frac{\text{Residual SS}}{(r-1)(c-1)} = \text{Residual MS}.$$

For the given data, $s^2 = 0.334$.

R produces the following output:

```
> anova(lm(Return ~ Year + Quarter))
Analysis of Variance Table

Response: Return
          Df Sum Sq Mean Sq F value    Pr(>F)
Year       4 19.867  4.9667  14.852 0.0001349 ***
Quarter    3  0.602  0.2007   0.600 0.6271918
Residuals 12  4.013  0.3344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the confidence intervals for years 1 and 2 (corresponding to 1981 and 1982) are separated from those for years 3 to 5 (that is, 1983 to 1985), which is consistent with rejection of H_0 in the no row effect test. In contrast, the confidence intervals for each quarter all overlap, which is consistent with our failure to reject H_0 in the no column effect test.

Finally, we may also look at the residuals:

$$\widehat{\varepsilon}_{ij} = X_{ij} - \widehat{\mu} - \widehat{\gamma}_i - \widehat{\beta}_j \quad \text{for } i = 1, 2, \ldots r \text{ and } j = 1, 2, \ldots, c.$$

If the assumed normal model (structure) is correct, the $\widehat{\varepsilon}_{ij}$s should behave like independent $N(0, \sigma^2)$ random variables.

10.9 Overview of chapter

This chapter introduced analysis of variance as a statistical tool to detect differences between group means. One-way and two-way analysis of variance frameworks were presented depending on whether one or two independent variables were modelled, respectively. Statistical inference in the form of hypothesis tests and confidence intervals was conducted.

10.10 Key terms and concepts

- ANOVA decomposition
- Between-groups variation
- One-way ANOVA
- Residual
- Total variation
- Within-groups variation
- Between-blocks variation
- Between-treatments variation
- Random errors
- Sample mean
- Two-way ANOVA

A total of 4,000 cans are opened around the world every second. Ten babies are conceived around the world every second. Each time you open a can, you stand a 1-in-400 chance of falling pregnant.
(True or false?)

Chapter 11
Linear regression

11.1 Synopsis of chapter

This chapter covers linear regression whereby the variation in a continuous dependent variable is modelled as being explained by one or more continuous independent variables.

11.2 Learning outcomes

After completing this chapter, you should be able to:

- derive from first principles the least squares estimators of the intercept and slope in the simple linear regression model
- explain how to construct confidence intervals and perform hypothesis tests for the intercept and slope in the simple linear regression model
- demonstrate how to construct confidence intervals and prediction intervals and explain the difference between the two
- summarise the multiple linear regression model with several explanatory variables, and explain its interpretation
- provide the assumptions on which regression models are based
- interpret typical output from a computer package fitting of a regression model.

11.3 Introduction

Regression analysis is one of the most frequently-used statistical techniques. It aims to model an explicit relationship between one **dependent variable**, often denoted as y, and one or more **regressors** (also called covariates, or independent variables), often denoted as $x_1, x_2, \ldots, x_p$.

The goal of regression analysis is to understand how y depends on $x_1, x_2, \ldots, x_p$ and to predict or control the unobserved y based on the observed $x_1, x_2, \ldots, x_p$. We start with some simple examples with $p = 1$.

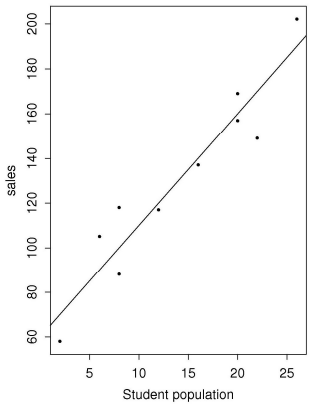
11.4 Introductory examples

Example 11.1 In a university town, the sales, y, of 10 Armand's Pizza Parlour restaurants are closely related to the student population, x, in their neighbourhoods. The data file '`Armand.csv`' contains the sales (in thousands of euros) in a period of three months together with the numbers of students (in thousands) in their neighbourhoods.

We plot y against x, and draw a straight line through the middle of the data points:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε stands for a random error term, β_0 is the intercept and β_1 is the slope of the straight line.



For a given student population, x, the predicted sales are $\widehat{y} = \beta_0 + \beta_1 x$.

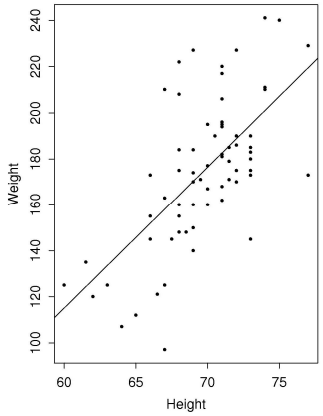
Example 11.2 The data file '`WeightHeight.csv`' contains the heights, x, and weights, y, of 69 students in a class.

We plot y against x, and draw a straight line through the middle of the data cloud:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε stands for a random error term, β_0 is the intercept and β_1 is the slope of the straight line.

For a given height, x, the predicted value $\widehat{y} = \beta_0 + \beta_1 x$ may be viewed as a kind of 'standard weight'.



Example 11.3 Some other possible examples of y and x are shown in the following table.

y	x
Sales	Price
Weight gain	Protein in diet
Present FTSE 100 index	Past FTSE 100 index
Consumption	Income
Salary	Tenure
Daughter's height	Mother's height

In most cases, there are several x variables involved. We will consider such situations later in this chapter.

Some questions to consider are the following.

- How to draw a line through data clouds, i.e. how to **estimate** β_0 and β_1?
- How accurate is the fitted line?
- What is the error in predicting a future y?

11.5 Simple linear regression

We now present the simple linear regression model. Let the paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be drawn from the model:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

where:

$$\boldsymbol{\mathrm{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = \sigma^2 > 0.}$$

Furthermore, suppose $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$. That is, the ε_is are assumed to be uncorrelated (remembering that a zero covariance between two random variables implies that they are uncorrelated).

So the model has three parameters: β_0, β_1 and σ^2.

For convenience, we will treat $x_1, x_2, \ldots, x_n$ as **constants**.[1] We have:

$$\boldsymbol{\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i \quad \text{and} \quad \mathrm{Var}(y_i) = \sigma^2.}$$

Since the ε_is are uncorrelated (by assumption), it follows that $y_1, y_2, \ldots, y_n$ are also uncorrelated with each other.

Sometimes we assume $\varepsilon_i \sim N(0, \sigma^2)$, in which case $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, and $y_1, y_2, \ldots, y_n$ are independent. (Remember that a linear transformation of a normal random variable is also normal, and that for jointly normal random variables if they are uncorrelated then they are also independent.)

[1] If you study an econometrics course, you will explore regression models in much more detail than is covered here. For example, $x_1, x_2, \ldots, x_n$ will be treated as random variables in an econometrics course.

Our tasks are two-fold.

- Statistical inference for β_0, β_1 and σ^2, i.e. (point) estimation, confidence intervals and hypothesis testing.
- Prediction intervals for future values of y.

We derive estimators of β_0 and β_1 using least squares estimation (introduced in Chapter 7). The least squares estimators (LSEs) of β_0 and β_1 are the values of (β_0, β_1) at which the function:

$$\boldsymbol{L(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}$$

obtains its minimum.

We proceed to partially differentiate $L(\beta_0, \beta_1)$ with respect to β_0 and β_1, respectively. Firstly:

$$\boldsymbol{\frac{\partial}{\partial \beta_0} L(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i).}$$

Upon setting this partial derivative to zero, this leads to:

$$\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \quad \text{or} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Secondly:

$$\boldsymbol{\frac{\partial}{\partial \beta_1} L(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i).}$$

Upon setting this partial derivative to zero, this leads to:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \\ &= \sum_{i=1}^{n} x_i (y_i - \bar{y} - (\widehat{\beta}_1 x_i - \widehat{\beta}_1 \bar{x})) \\ &= \sum_{i=1}^{n} x_i (y_i - \bar{y}) - \widehat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}). \end{aligned}$$

Hence:

$$\boldsymbol{\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.}$$

The estimator $\widehat{\beta}_1$ above is based on the fact that for any constant c, we have:

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - c)(y_i - \bar{y})$$

since:

$$\sum_{i=1}^{n} c(y_i - \bar{y}) = c\sum_{i=1}^{n}(y_i - \bar{y}) = 0.$$

Given that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, it follows that $\sum_{i=1}^{n} c(x_i - \bar{x}) = 0$ for any constant c.

In order to calculate $\widehat{\beta}_1$ numerically, often the following formula is convenient:

$$\boldsymbol{\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.}$$

An alternative derivation is as follows. Note $L(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. For any β_0 and β_1, we have:

$$\begin{aligned} L(\beta_0, \beta_1) &= \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i + \widehat{\beta}_0 - \beta_0 + (\widehat{\beta}_1 - \beta_1)x_i)^2 \\ &= L(\widehat{\beta}_0, \widehat{\beta}_1) + \sum_{i=1}^{n}(\widehat{\beta}_0 - \beta_0 + (\widehat{\beta}_1 - \beta_1)x_i)^2 + 2B \end{aligned} \tag{11.1}$$

where:

$$\begin{aligned} B &= \sum_{i=1}^{n}(\widehat{\beta}_0 - \beta_0 + (\widehat{\beta}_1 - \beta_1)x_i)(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \\ &= (\widehat{\beta}_0 - \beta_0)\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) + (\widehat{\beta}_1 - \beta_1)\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i). \end{aligned}$$

Now let $(\widehat{\beta}_0, \widehat{\beta}_1)$ be the solution to the equations:

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \tag{11.2}$$

such that $B = 0$. By (11.1), we have:

$$L(\beta_0, \beta_1) = L(\widehat{\beta}_0, \widehat{\beta}_1) + \sum_{i=1}^{n}(\widehat{\beta}_0 - \beta_0 + (\widehat{\beta}_1 - \beta_1)x_i)^2 \geq L(\widehat{\beta}_0, \widehat{\beta}_1).$$

Hence $(\widehat{\beta}_0, \widehat{\beta}_1)$ are the least squares estimators (LSEs) of β_0 and β_1, respectively.

To find the explicit expression from (11.2), note the first equation can be written as:

$$n\bar{y} - n\widehat{\beta}_0 - n\widehat{\beta}_1\bar{x} = 0.$$

Hence $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}$. Substituting this into the second equation, we have:

$$0 = \sum_{i=1}^{n} x_i(y_i - \bar{y} - \widehat{\beta}_1(x_i - \bar{x})) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \widehat{\beta}_1\sum_{i=1}^{n} x_i(x_i - \bar{x}).$$

Therefore:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i(\boldsymbol{y}_i - \bar{\boldsymbol{y}})}{\sum_{i=1}^{n} \boldsymbol{x}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})} = \frac{\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})}{\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2}.$$

This completes the derivation.

Remember $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Hence $\sum_{i=1}^{n} c(x_i - \bar{x}) = 0$ for any constant c.

We also note the estimator of σ^2, which is:

$$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}} = \frac{\sum_{i=1}^{n} (\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}}_{\mathbf{0}} - \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\boldsymbol{x}_i)^2}{\boldsymbol{n - 2}}.$$

We now explore the properties of the LSEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$. We now proceed to show that the means and variances of these LSEs are:

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}_{\mathbf{0}}) = \boldsymbol{\beta}_{\mathbf{0}} \quad \textbf{and} \quad \mathbf{Var}(\widehat{\boldsymbol{\beta}}_{\mathbf{0}}) = \frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\boldsymbol{n}} \frac{\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathbf{2}}}{\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2}$$

for $\widehat{\beta}_0$, and:

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}) = \boldsymbol{\beta}_{\mathbf{1}} \quad \textbf{and} \quad \mathbf{Var}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}) = \frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2}$$

for $\widehat{\beta}_1$.

Proof: Recall we treat the x_is as constants, and we have $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$ and also $\mathrm{Var}(y_i) = \sigma^2$. Hence:

$$\mathrm{E}(\bar{y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i) = \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x}.$$

Therefore:

$$\mathrm{E}(y_i - \bar{y}) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \bar{x}) = \beta_1(x_i - \bar{x}).$$

Consequently, we have:

$$\mathrm{E}(\widehat{\beta}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathrm{E}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\beta_1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1.$$

Now:

$$\mathrm{E}(\widehat{\beta}_0) = \mathrm{E}(\bar{y} - \widehat{\beta}_1\bar{x}) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0.$$

Therefore, the LSEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively.

To work out the variances, the key is to write $\widehat{\beta}_1$ and $\widehat{\beta}_0$ as **linear estimators** (i.e. linear combinations of the y_is):

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{k=1}^{n}(x_k - \bar{x})^2} = \sum_{i=1}^{n} a_i y_i$$

where $a_i = (x_i - \bar{x}) \Big/ \sum_{k=1}^{n}(x_k - \bar{x})^2$ and:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \bar{y} - \sum_{i=1}^{n} a_i \bar{x} y_i = \sum_{i=1}^{n} \left(\frac{1}{n} - a_i \bar{x} \right) y_i.$$

Note that:

$$\sum_{i=1}^{n} a_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} a_i^2 = \frac{1}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

Now we note the following lemma, without proof. Let $y_1, y_2, \ldots, y_n$ be uncorrelated random variables, and $b_1, b_2, \ldots, b_n$ be constants, then:

$$\mathrm{Var}\left(\sum_{i=1}^{n} b_i y_i \right) = \sum_{i=1}^{n} b_i^2 \mathrm{Var}(y_i).$$

By this lemma:

$$\mathrm{Var}(\widehat{\beta}_1) = \mathrm{Var}\left(\sum_{i=1}^{n} a_i y_i \right) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

and:

$$\mathrm{Var}(\widehat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - a_i \bar{x} \right)^2 = \sigma^2 \left(\frac{1}{n} + \sum_{i=1}^{n} a_i^2 \bar{x}^2 \right) = \frac{\sigma^2}{n} \left(1 + \frac{n\bar{x}^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} \right)$$

$$= \frac{\sigma^2}{n} \frac{\sum_{k=1}^{n} x_k^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

The last equality uses the fact that:

$$\sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n}(x_k - \bar{x})^2 + n\bar{x}^2.$$

■

11.6 Inference for parameters in normal regression models

The normal simple linear regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where:

$$\boldsymbol{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim_{IID} N(0, \sigma^2).}$$

$y_1, y_2, \ldots, y_n$ are independent (but not identically distributed) and:

$$\boldsymbol{y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).}$$

Since any linear combination of normal random variables is also normal, the LSEs of β_0 and β_1 (as linear estimators) are also normal random variables. In fact:

$$\boldsymbol{\widehat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)} \quad \textbf{and} \quad \boldsymbol{\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).}$$

Since σ^2 is unknown in practice, we replace σ^2 by its estimator:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2}$$

and use the *estimated* standard errors:

$$\boldsymbol{\text{E.S.E.}(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{\sqrt{n}} \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{1/2}}$$

and:

$$\boldsymbol{\text{E.S.E.}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{1/2}}.}$$

The following results all make use of distributional results introduced earlier in the course. Statistical inference (confidence intervals and hypothesis testing) for the normal simple linear regression model can then be performed.

i. We have:

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{\sigma^2} \sim \chi^2_{n-2}.$$

ii. $\widehat{\beta}_0$ and $\widehat{\sigma}^2$ are independent, hence:

$$\frac{\widehat{\beta}_0 - \beta_0}{\text{E.S.E.}(\widehat{\beta}_0)} \sim t_{n-2}.$$

iii. $\widehat{\beta}_1$ and $\widehat{\sigma}^2$ are independent, hence:

$$\frac{\widehat{\beta}_1 - \beta_1}{\text{E.S.E.}(\widehat{\beta}_1)} \sim t_{n-2}.$$

Confidence intervals for the simple linear regression model parameters

A $100(1-\alpha)\%$ confidence interval for β_0 is:

$$\left(\widehat{\boldsymbol{\beta}}_{\mathbf{0}} - \boldsymbol{t}_{\boldsymbol{\alpha/2, n-2}} \times \mathbf{E.S.E.}(\widehat{\boldsymbol{\beta}}_{\mathbf{0}}), \widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \boldsymbol{t}_{\boldsymbol{\alpha/2, n-2}} \times \mathbf{E.S.E.}(\widehat{\boldsymbol{\beta}}_{\mathbf{0}})\right)$$

and a $100(1-\alpha)\%$ confidence interval for β_1 is:

$$\left(\widehat{\boldsymbol{\beta}}_{\mathbf{1}} - \boldsymbol{t}_{\boldsymbol{\alpha/2, n-2}} \times \mathbf{E.S.E.}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}), \widehat{\boldsymbol{\beta}}_{\mathbf{1}} + \boldsymbol{t}_{\boldsymbol{\alpha/2, n-2}} \times \mathbf{E.S.E.}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}})\right)$$

where $t_{\alpha,k}$ denotes the top 100αth percentile of the Student's t_k distribution, obtained from Table 7 of Murdoch and Barnes' *Statistical Tables*.

Tests for the regression slope

The relationship between y and x in the regression model hinges on β_1. If $\beta_1 = 0$, then $y \sim N(\beta_0, \sigma^2)$.

To validate the use of the regression model, we need to make sure that $\beta_1 \neq 0$, or more practically that $\widehat{\beta}_1$ is significantly non-zero. This amounts to testing:

$$\mathbf{H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0.}$$

Under H_0, the test statistic is:

$$\boldsymbol{T = \frac{\widehat{\beta}_1}{\text{E.S.E.}(\widehat{\beta}_1)} \sim t_{n-2}.}$$

At the $100\alpha\%$ significance level, we reject H_0 if $|t| > t_{\alpha/2, n-2}$, where t is the observed test statistic value.

Alternatively, we could use $H_1 : \beta_1 < 0$ or $H_1 : \beta_1 > 0$ if there was a rationale for doing so. In such cases, we would reject H_0 if $t < -t_{\alpha, n-2}$ and $t > t_{\alpha, n-2}$ for the lower-tailed and upper-tailed t tests, respectively.

Some remarks are the following.

i. For testing $H_0 : \beta_1 = b$ for a given constant b, the above test still applies, but now with the following test statistic:

$$\boldsymbol{T = \frac{\widehat{\beta}_1 - b}{\text{E.S.E.}(\widehat{\beta}_1)}.}$$

ii. Tests for the regression intercept β_0 may be constructed in a similar manner, replacing β_1 and $\widehat{\beta}_1$ with β_0 and $\widehat{\beta}_0$, respectively.

In the normal regression model, the LSEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are also the MLEs of β_0 and β_1, respectively.

Since $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \sim_{IID} N(0, \sigma^2)$, the likelihood function is:

$$\boldsymbol{L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right)}$$

$$\boldsymbol{\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right).}$$

Hence the log-likelihood function is:

$$\boldsymbol{l(\beta_0, \beta_1, \sigma^2) = \frac{n}{2}\ln\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + c.}$$

Therefore, for any β_0, β_1 and $\sigma^2 > 0$, we have:

$$l(\beta_0, \beta_1, \sigma^2) \leq l(\widehat{\beta}_0, \widehat{\beta}_1, \sigma^2).$$

Hence $(\widehat{\beta}_0, \widehat{\beta}_1)$ are the MLEs of (β_0, β_1).

To find the MLE of σ^2, we need to maximise:

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

Setting $u = 1/\sigma^2$, it is equivalent to maximising:

$$g(u) = n\ln u - ub$$

where $b = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$.

Setting $\mathrm{d}g(u)/\mathrm{d}u = n/\widehat{u} - b = 0$, $\widehat{u} = n/b$, i.e. $g(u)$ attains its maximum at $u = \widehat{u}$. Hence the MLE of σ^2 is:

$$\widetilde{\sigma}^2 = \frac{1}{\widehat{u}} = \frac{b}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Note the MLE $\widetilde{\sigma}^2$ is a *biased* estimator of σ^2. In practice, we often use the unbiased estimator:

$$\boldsymbol{\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.}$$

We now consider an empirical example of the normal simple linear regression model.

Example 11.4 The dataset '`Cigarette.csv`' contains the annual cigarette consumption, x, and the corresponding mortality rate, y, due to coronary heart disease (CHD) of 21 countries. Some useful summary statistics calculated from the data are:

$$\sum_{i=1}^{21} x_i = 45{,}110, \quad \sum_{i=1}^{21} y_i = 3{,}042.2, \quad \sum_{i=1}^{21} x_i^2 = 109{,}957{,}100,$$

$$\sum_{i=1}^{21} y_i^2 = 529{,}321.58 \quad \text{and} \quad \sum_{i=1}^{21} x_i y_i = 7{,}319{,}602.$$

Do these data support the suspicion that smoking contributes to CHD mortality?

(Note the assertion 'smoking is harmful for health' is largely based on statistical, rather than laboratory, evidence.)

We fit the regression model $y = \beta_0 + \beta_1 x + \varepsilon$. Our least squares estimates of β_1 and β_0 are, respectively:

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i x_i y_i - \sum_i x_i \sum_j y_j / n}{\sum_i x_i^2 - (\sum_i x_i)^2 / n}$$

$$= \frac{7{,}319{,}602 - 45{,}110 \times 3{,}042.2/21}{109{,}957{,}100 - (45{,}110)^2/21}$$

$$= 0.06$$

and:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \frac{3{,}042.2 - 0.06 \times 45{,}110}{21} = 15.77.$$

Also:

$$\widehat{\sigma}^2 = \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 / (n-2)$$

$$= \left(\sum y_i^2 + n\widehat{\beta}_0^2 + \widehat{\beta}_1^2 \sum x_i^2 - 2\widehat{\beta}_0 \sum y_i - 2\widehat{\beta}_1 \sum x_i y_i + 2\widehat{\beta}_0 \widehat{\beta}_1 \sum x_i \right) / (n-2)$$

$$= 2{,}181.66.$$

We now proceed to test $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 > 0$. (If indeed smoking contributes to CHD mortality, then $\beta_1 > 0$.)

We have calculated $\widehat{\beta}_1 = 0.06$. However, is this deviation from zero due to sampling error, or is it significantly different from zero? (The magnitude of $\widehat{\beta}_1$ itself is not important in determining if $\beta_1 = 0$ or not – changing the scale of x may make $\widehat{\beta}_1$ arbitrarily small.)

Under H_0, the test statistic is:

$$T = \frac{\widehat{\beta}_1}{\text{E.S.E.}(\widehat{\beta}_1)} \sim t_{n-2} = t_{19}$$

where $\text{E.S.E.}(\widehat{\beta}_1) = \widehat{\sigma} / (\sum_i (x_i - \bar{x})^2)^{1/2} = 0.01293$.

Since $t = 0.06/0.01293 = 4.64 > 2.54 = t_{0.01,\,19}$, we reject the hypothesis $\beta_1 = 0$ at the 1% significance level and we conclude that there is strong evidence that smoking contributes to CHD mortality.

11.7 Regression ANOVA

In Chapter 10 we discussed ANOVA, whereby we decomposed the total variation of a continuous dependent variable. In a similar way we can decompose the total variation of y in the simple linear regression model. It can be shown that the regression ANOVA decomposition is:

$$\boldsymbol{\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\widehat{\beta}_1^2(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}$$

where, denoting sum of squares by 'SS', we have the following.

- **Total SS** is $\boldsymbol{\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2}$.
- **Regression (explained) SS** is $\boldsymbol{\sum_{i=1}^{n}\widehat{\beta}_1^2(x_i-\bar{x})^2=\widehat{\beta}_1^2\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)}$.
- **Residual (error) SS** is $\boldsymbol{\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2=}$ **Total SS − Regression SS**.

If $\varepsilon_i\sim N(0,\sigma^2)$ and $\beta_1=0$, then it can be shown that:

- $\sum_{i=1}^{n}(y_i-\bar{y})^2/\sigma^2\sim\chi^2_{n-1}$
- $\sum_{i=1}^{n}\widehat{\beta}_1^2(x_i-\bar{x})^2/\sigma^2\sim\chi^2_1$
- $\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2/\sigma^2\sim\chi^2_{n-2}$.

Therefore, under $\mathrm{H}_0:\beta_1=0$, we have:

$$\boldsymbol{F=\frac{\text{(Regression SS)}/1}{\text{(Residual SS)}/(n-2)}=\frac{(n-2)\widehat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2}=\left(\frac{\widehat{\beta}_1}{\text{E.S.E.}(\widehat{\beta}_1)}\right)^2\sim F_{1,n-2}.}$$

We reject H_0 at the $100\alpha\%$ significance level if $f>F_{\alpha,1,n-2}$, where f is the observed test statistic value and $F_{\alpha,1,n-2}$ is the top 100αth percentile of the $F_{1,n-2}$ distribution, obtained from Table 9 of Murdoch and Barnes' *Statistical Tables*.

A useful statistic is the **coefficient of determination**, denoted as R^2, defined as:

$$\boldsymbol{R^2=\frac{\text{Regression SS}}{\text{Total SS}}=1-\frac{\text{Residual SS}}{\text{Total SS}}.}$$

If we view Total SS as the total variation (or energy) of y, then R^2 is the proportion of the total variation of y explained by x. Note that $R^2\in[0,1]$. The closer R^2 is to 1, the better the explanatory power of the regression model.

11.8 Confidence intervals for E(y)

Based on the observations (x_i, y_i), for $i = 1, 2, \ldots, n$, we fit a regression model:

$$\boldsymbol{\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.}$$

Our goal is to predict the *unobserved* y corresponding to a *known* x. The point prediction is:

$$\boldsymbol{\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.}$$

For the analysis to be more informative, we would like to have some 'error bars' for our prediction. We introduce two methods as follows.

- A **confidence interval** for $\boldsymbol{\mu(x) = \mathrm{E}(y) = \beta_0 + \beta_1 x}$.
- A **prediction interval** for $\boldsymbol{y}$.

A confidence interval is an interval estimator of an unknown parameter (i.e. for a constant) while a prediction interval is for a random variable. They are different and serve different purposes.

We assume the model is normal, i.e. $\varepsilon = y - \beta_0 - \beta_1 x \sim N(0, \sigma^2)$ and let $\widehat{\mu}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$, such that $\widehat{\mu}(x)$ is an unbiased estimator of $\mu(x)$. We note without proof that:

$$\boldsymbol{\widehat{\mu}(x) \sim N\left(\mu(x), \frac{\sigma^2}{n} \frac{\sum_{i=1}^{n} (x_i - x)^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}\right).}$$

Standardising gives:

$$\frac{\widehat{\mu}(x) - \mu(x)}{\sqrt{(\sigma^2/n)\left(\sum_{i=1}^{n}(x_i - x)^2 / \sum_{j=1}^{n}(x_j - \bar{x})^2\right)}} \sim N(0, 1).$$

In practice σ^2 is unknown, but it can be shown that $(n-2)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$, where $\widehat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2/(n-2)$. Furthermore, $\widehat{\mu}(x)$ and $\widehat{\sigma}^2$ are *independent*. Hence:

$$\frac{\widehat{\mu}(x) - \mu(x)}{\sqrt{(\widehat{\sigma}^2/n)\left(\sum_{i=1}^{n}(x_i - x)^2 / \sum_{j=1}^{n}(x_j - \bar{x})^2\right)}} \sim t_{n-2}.$$

Confidence interval for $\mu(x)$

A $100(1-\alpha)\%$ confidence interval for $\mu(x)$ is:

$$\boldsymbol{\widehat{\mu}(x) \pm t_{\alpha/2,n-2} \times \widehat{\sigma} \times \left(\frac{\sum_{i=1}^{n}(x_i - x)^2}{n\sum_{j=1}^{n}(x_j - \bar{x})^2} \right)^{1/2}}.$$

Such a confidence interval contains the true expectation $\mathrm{E}(y) = \mu(x)$ with probability $1-\alpha$ over repeated samples. It does *not* cover y with probability $1-\alpha$.

11.9 Prediction intervals for y

A $100(1-\alpha)\%$ **prediction interval** is an interval which contains y with probability $1-\alpha$.

We may assume that the y to be predicted is independent of $y_1, y_2, \ldots, y_n$ used in the estimation of the regression model.

Hence $y - \widehat{\mu}(x)$ is normal with mean 0 and variance:

$$\boldsymbol{\mathrm{Var}(y) + \mathrm{Var}(\widehat{\mu}(x)) = \sigma^2 + \frac{\sigma^2}{n}\frac{\sum_{i=1}^{n}(x_i - x)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}}.$$

Therefore:

$$(y - \widehat{\mu}(x)) \Big/ \left(\widehat{\sigma}^2 \left(1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n\sum_{j=1}^{n}(x_j - \bar{x})^2} \right) \right)^{1/2} \sim t_{n-2}.$$

Prediction interval for y

A $100(1-\alpha)\%$ prediction interval covering y with probability $1-\alpha$ is:

$$\boldsymbol{\widehat{\mu}(x) \pm t_{\alpha/2,n-2} \times \widehat{\sigma} \times \left(1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n\sum_{j=1}^{n}(x_j - \bar{x})^2} \right)^{1/2}}.$$

Some remarks are the following.

i. It holds that:

$$P\left(y \in \widehat{\mu}(x) \pm t_{\alpha/2,n-2} \times \widehat{\sigma} \times \left(1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)^{1/2}\right) = 1 - \alpha.$$

ii. The prediction interval for y is *wider* than the confidence interval for $\mathrm{E}(y)$. The former contains the unobserved **random variable** y with probability $1 - \alpha$, the latter contains the unknown **constant** $\mathrm{E}(y)$ with probability $1 - \alpha$ over repeated samples.

Example 11.5 The dataset '`UsedFord.csv`' contains the prices (y, in \$000s) of 100 three-year-old Ford Tauruses together with their mileages (x, in thousands of miles) when they were sold at auction. Based on these data, a car dealer needs to make two decisions.

1. To prepare cash for bidding on *one* three-year-old Ford Taurus with a mileage of $x = 40$.
2. To prepare buying *several* three-year-old Ford Tauruses with mileages close to $x = 40$ from a rental company.

For the first task, a *prediction interval* would be more appropriate. For the second task, the car dealer needs to know the average price and, therefore, a *confidence interval* is appropriate. This can be easily done using R.

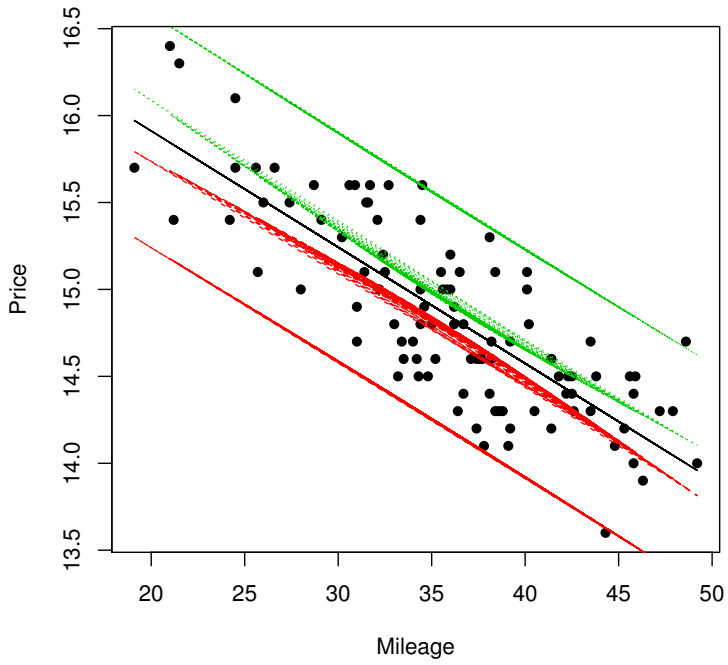
```
> reg <- lm(Price~ Mileage)
> summary(reg)

Call:
lm(formula = Price ~ Mileage)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68679 -0.27263  0.00521  0.23210  0.70071

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.248727   0.182093   94.72   <2e-16 ***
Mileage     -0.066861   0.004975  -13.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared:  0.6483,    Adjusted R-squared:  0.6447
F-statistic: 180.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
> new.Mileage <- data.frame(Mileage = c(40))
> predict(reg, newdata = new.Mileage, int = "c")
       fit      lwr      upr
1 14.57429 14.49847 14.65011
> predict(reg, newdata = new.Mileage, int = "p")
       fit      lwr      upr
1 14.57429 13.92196 15.22662
```

We predict that a Ford Taurus will sell for between \$13,922 and \$15,227. The average selling price of several three-year-old Ford Tauruses is estimated to be between \$14,498 and \$14,650. Because predicting the selling price for one car is more difficult, the corresponding prediction interval is *wider* than the confidence interval.

To produce the plots with confidence intervals for E(y) and prediction intervals for y, we proceed as follows:

```
> pc <- predict(reg,int="c")
> pp <- predict(reg,int="p")
> plot(Mileage,Price,pch=16)
> matlines(Mileage,pc)
> matlines(Mileage,pp)
```



11.10 Multiple linear regression models

For most practical problems, the variable of interest, y, typically depends on several explanatory variables, say $x_1, x_2, \ldots, x_p$, leading to the **multiple linear regression model**. In this course we only provide a brief overview of the multiple linear regression model. Subsequent econometrics courses would explore this model in much greater depth.

Let $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, for $i = 1, 2, \ldots, n$, be observations from the model:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i}$$

where:

$$\mathbf{E}(\boldsymbol{\varepsilon_i}) = \mathbf{0}, \quad \mathbf{Var}(\boldsymbol{\varepsilon_i}) = \boldsymbol{\sigma^2 > 0} \quad \textbf{and} \quad \mathbf{Cov}(\boldsymbol{\varepsilon_i, \varepsilon_j}) = \mathbf{0} \textbf{ for all } \boldsymbol{i \neq j}.$$

The multiple linear regression model is a natural extension of the simple linear regression model, just with more parameters: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and σ^2.

Treating all of the x_{ij}s as constants as before, we have:

$$\mathbf{E}(\boldsymbol{y_i}) = \boldsymbol{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}} \quad \textbf{and} \quad \mathbf{Var}(\boldsymbol{y_i}) = \boldsymbol{\sigma^2}.$$

$y_1, y_2, \ldots, y_n$ are uncorrelated with each other, again as before.

If in addition $\varepsilon_i \sim N(0, \sigma^2)$, then:

$$\boldsymbol{y_i \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right).}$$

Estimation of the intercept and slope parameters is still performed using least squares estimation. The LSEs $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$ are obtained by minimising:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2$$

leading to the fitted regression model:

$$\boldsymbol{\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p.}$$

The residuals are expressed as:

$$\boldsymbol{\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j x_{ij}.}$$

Just as with the simple linear regression model, we can decompose the total variation of y such that:

$$\boldsymbol{\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \widehat{\varepsilon}_i^2}$$

or, in words:

$$\textbf{Total SS} = \textbf{Regression SS} + \textbf{Residual SS}.$$

An unbiased estimator of σ^2 is:

$$\boldsymbol{\widehat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j x_{ij}\right)^2 = \frac{\textbf{Residual SS}}{n-p-1}.}$$

We can test a single slope coefficient by testing:

$$\mathbf{H_0} : \boldsymbol{\beta_i = 0} \quad \textbf{vs.} \quad \mathbf{H_1} : \boldsymbol{\beta_i \neq 0}.$$

Under H_0, the test statistic is:

$$\boldsymbol{T = \frac{\widehat{\beta}_i}{\text{E.S.E.}(\widehat{\beta}_i)} \sim t_{n-p-1}}$$

and we reject H_0 if $|t| > t_{\alpha/2, n-p-1}$. However, note the slight difference in the interpretation of the slope coefficient β_j. In the multiple regression setting, β_j is the effect of x_j on y, holding all other independent variables fixed – this is unfortunately not always practical.

It is also possible to test whether all the regression coefficients are equal to zero. This is known as a **joint test of significance** and can be used to test the overall significance of the regression model, i.e. whether there is at least one significant explanatory (independent) variable, by testing:

$$\mathbf{H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad vs. \quad H_1 : At\ least\ one\ \beta_i \neq 0.}$$

Indeed, it is preferable to perform this joint test of significance *before* conducting t tests of individual slope coefficients. Failure to reject H_0 would render the model useless and hence the model would not warrant any further statistical investigation.

Provided $\varepsilon_i \sim N(0, \sigma^2)$, under $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, the test statistic is:

$$\boldsymbol{F = \frac{(\text{Regression SS})/p}{(\text{Residual SS})/(n-p-1)} \sim F_{p, n-p-1}.}$$

We reject H_0 at the $100\alpha\%$ significance level if $f > F_{\alpha, p, n-p-1}$.

It may be shown that:

$$\text{Regression SS} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{\beta}_1(x_{i1} - \bar{x}_1) + \widehat{\beta}_2(x_{i2} - \bar{x}_2) + \cdots + \widehat{\beta}_p(x_{ip} - \bar{x}_p))^2.$$

Hence, under H_0, f should be very small.

We now conclude the chapter with worked examples of linear regression using R.

11.11 Regression using R

To solve practical regression problems, we need to use statistical computing packages. All of them include linear regression analysis. In fact all statistical packages, such as R, make regression analysis much easier to use.

> **Example 11.6** We illustrate the use of linear regression in R using the dataset '`Armand.csv`', introduced in Example 11.1.

```
> reg <- lm(Sales ~ Student.population)
> summary(reg)
```

```
Call:
lm(formula = Sales ~ Student.population)

Residuals:
   Min     1Q Median     3Q    Max
-21.00  -9.75  -3.00  11.25  18.00

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         60.0000     9.2260   6.503 0.000187 ***
Student.population   5.0000     0.5803   8.617 2.55e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.83 on 8 degrees of freedom
Multiple R-squared:  0.9027,    Adjusted R-squared:  0.8906
F-statistic: 74.25 on 1 and 8 DF,  p-value: 2.549e-05
```

The fitted line is $\widehat{y} = 60 + 5x$. We have $\widehat{\sigma}^2 = (13.83)^2$. Also, $\widehat{\beta}_0 = 60$ and E.S.E.$(\widehat{\beta}_0) = 9.2260$. $\widehat{\beta}_1 = 5$ and E.S.E.$(\widehat{\beta}_1) = 0.5803$.

For testing $\mathrm{H}_0 : \beta_0 = 0$ we have $t = \widehat{\beta}_0/\text{E.S.E.}(\widehat{\beta}_0) = 6.503$. The p-value is $P(|T| > 6.503) = 0.000187$, where $T \sim t_{n-2}$.

For testing $\mathrm{H}_0 : \beta_1 = 0$ we have $t = \widehat{\beta}_1/\text{E.S.E.}(\widehat{\beta}_1) = 8.617$. The p-value is $P(|T| > 8.617) = 0.0000255$, where $T \sim t_{n-2}$.

The F test statistic value is 74.25 with a corresponding p-value of:

$$P(F > 74.25) = 0.00002549$$

where $F_{1,\,8}$.

Example 11.7 We apply the simple linear regression model to study the relationship between two series of financial returns – a regression of Cisco Systems stock returns, y, on S&P500 Index returns, x. This regression model is an example of the **capital asset pricing model (CAPM)**.

Stock returns are defined as:

$$\textbf{return} = \frac{\textbf{current price} - \textbf{previous price}}{\textbf{previous price}} \approx \ln\left(\frac{\textbf{current price}}{\textbf{previous price}}\right)$$

when the difference between the two prices is small.

The data file '`Returns.csv`' contains daily returns over the period 3 January – 29 December 2000 (i.e. $n = 252$ observations). The dataset has 5 columns: Day, S&P500 return, Cisco return, Intel return and Sprint return.

Daily prices are definitely not independent. However, daily returns may be seen as a sequence of uncorrelated random variables.

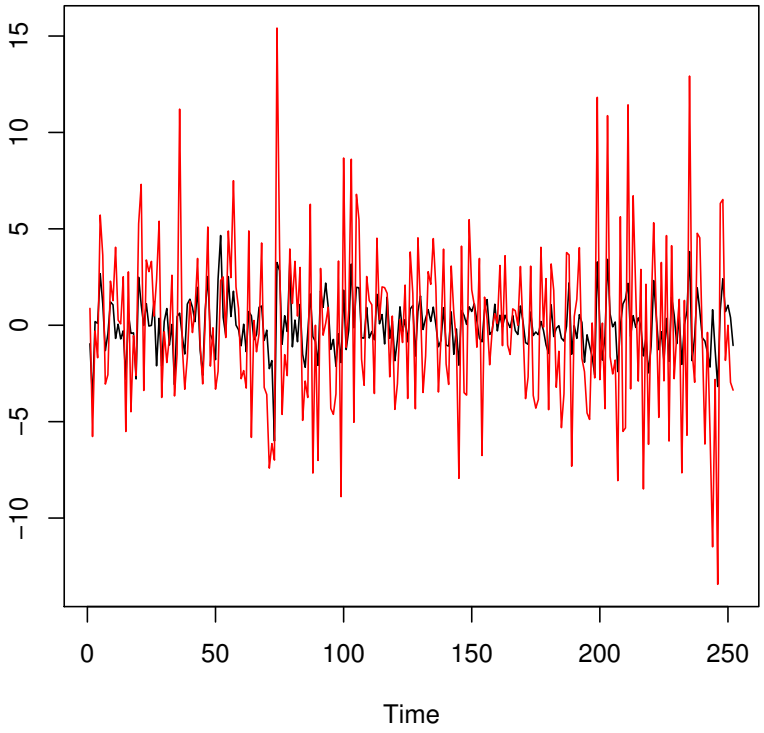
```
> summary(S.P500)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-6.00451 -0.85028 -0.03791 -0.04242  0.79869  4.65458

> summary(Cisco)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.4387  -3.0819  -0.1150  -0.1336   2.6363  15.4151
```

For the S&P500, the average daily return is -0.04%, the maximum daily return is 4.46%, the minimum daily return is -6.01% and the standard deviation is 1.40%.

For Cisco, the average daily return is -0.13%, the maximum daily return is 15.42%, the minimum daily return is -13.44% and the standard deviation is 4.23%.

We see that Cisco is much more volatile than the S&P500.

```
> sandpts <- ts(S.P500)
> ciscots <- ts(Cisco)
> ts.plot(sandpts,ciscots,col=c(1:2))
```



There is clear *synchronisation* between the movements of the two series of returns, as evident from examining the sample correlation coefficient.

```
> cor.test(S.P500,Cisco)

        Pearson's product-moment correlation

data:  S.P500 and Cisco
t = 14.943, df = 250, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 0.6155530 0.7470423
sample estimates:
     cor
0.686878
```

We fit the regression model: Cisco $= \beta_0 + \beta_1$S&P500 $+ \varepsilon$.

Our rationale is that part of the fluctuation in Cisco returns was driven by the fluctuation in the S&P500 returns.

R produces the following regression output.

```
> reg <- lm(Cisco ~ S.P500)
> summary(reg)

Call:
lm(formula = Cisco ~ S.P500)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1175  -2.0238   0.0091   2.0614   9.9491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04547    0.19433  -0.234    0.815
S.P500       2.07715    0.13900  14.943   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 250 degrees of freedom
Multiple R-squared:  0.4718,    Adjusted R-squared:  0.4697
F-statistic: 223.3 on 1 and 250 DF,  p-value: < 2.2e-16
```

The estimated slope is $\widehat{\beta}_1 = 2.07715$. The null hypothesis $\mathrm{H}_0 : \beta_1 = 0$ is rejected with a p-value of 0.000 (to three decimal places). Therefore, the test is extremely significant.

Our interpretation is that when the market index goes up by 1%, Cisco stock goes up by 2.07715%, on average. However, the error term ε in the model is large with an estimated $\widehat{\sigma} = 3.083\%$.

The p-value for testing $\mathrm{H}_0 : \beta_0 = 0$ is 0.815, so we cannot reject the hypothesis that $\beta_0 = 0$. Recall $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ and both $\bar{y}$ and $\bar{x}$ are very close to 0.

$R^2 = 47.18\%$, hence 47.18% of the variation of Cisco stock may be explained by the variation of the S&P500 index, or, in other words, 47.18% of the risk in Cisco stock is the *market-related risk*.

The capital asset pricing model (CAPM) is a simple asset pricing model in finance given by:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

where y_i is a stock return and x_i is a market return at time i.

The *total risk of the stock* is:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

The *market-related (or systematic) risk* is:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\frac{1}{n}\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

The *firm-specific risk* is:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Some remarks are the following.

i. β_1 measures the market-related (or systematic) risk of the stock.

ii. Market-related risk is unavoidable, while firm-specific risk may be 'diversified away' through *hedging*.

iii. Variance is a simple measure (and one of the most frequently-used) of risk in finance.

Example 11.8 The data in the file '`Foods.csv`' illustrate the effects of marketing instruments on the weekly sales volume of a certain food product over a three-year period. Data are real but transformed to protect the innocent!

There are observations on the following four variables:

$y = LVOL$: logarithms of weekly sales volume

$x_1 = PROMP$: promotion price

$x_2 = FEAT$: feature advertising

$x_3 = DISP$: display measure.

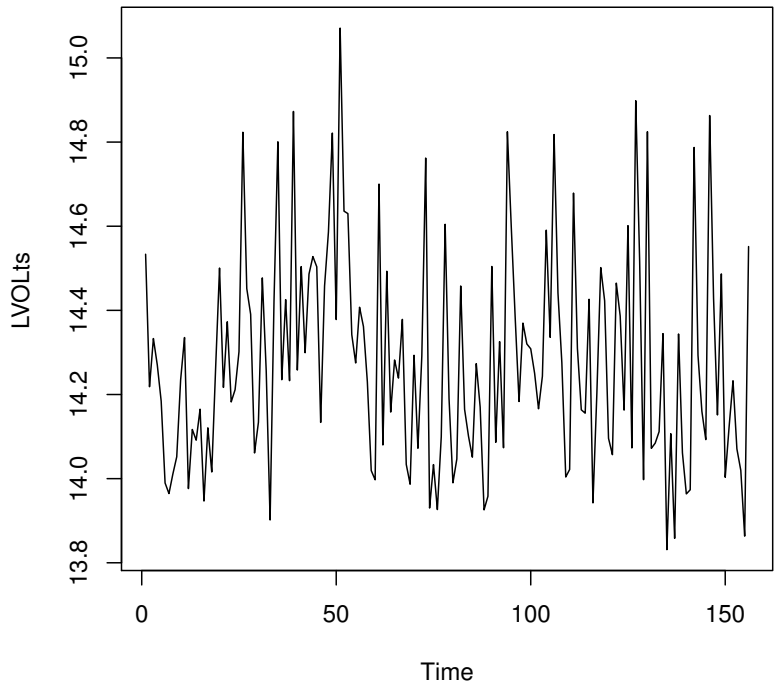
R produces the following descriptive statistics.

```
> summary(Foods)
      LVOL            PROMP            FEAT            DISP
 Min.   :13.83   Min.   :3.075   Min.   : 2.84   Min.   :12.42
 1st Qu.:14.08   1st Qu.:3.330   1st Qu.:15.95   1st Qu.:20.59
 Median :14.24   Median :3.460   Median :22.99   Median :25.11
 Mean   :14.28   Mean   :3.451   Mean   :24.84   Mean   :25.31
 3rd Qu.:14.43   3rd Qu.:3.560   3rd Qu.:33.49   3rd Qu.:29.34
 Max.   :15.07   Max.   :3.865   Max.   :57.10   Max.   :45.94
```

$n = 156$. The values of $FEAT$ and $DISP$ are much larger than $LVOL$.

As always, first we plot the data to ascertain basic characteristics.

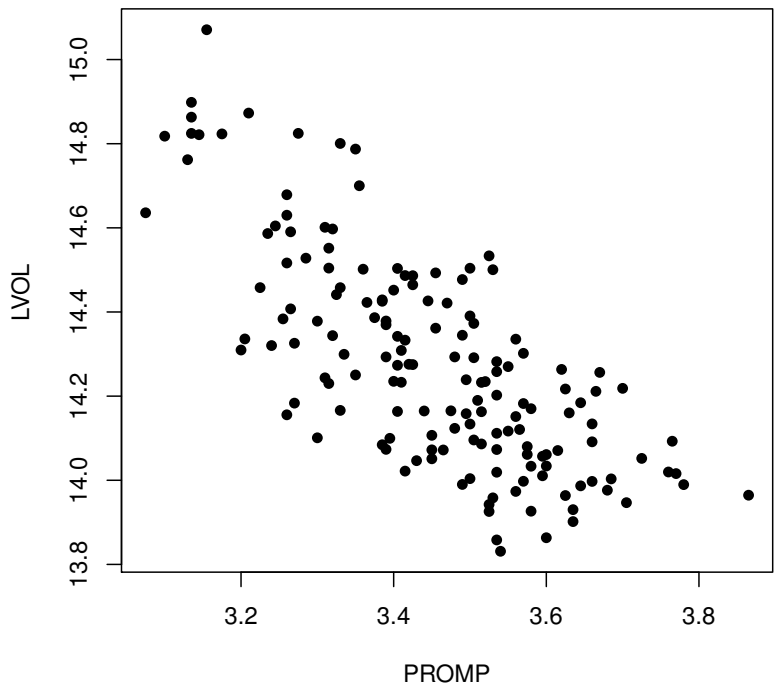
```
> LVOLts <- ts(LVOL)
> ts.plot(LVOLts)
```



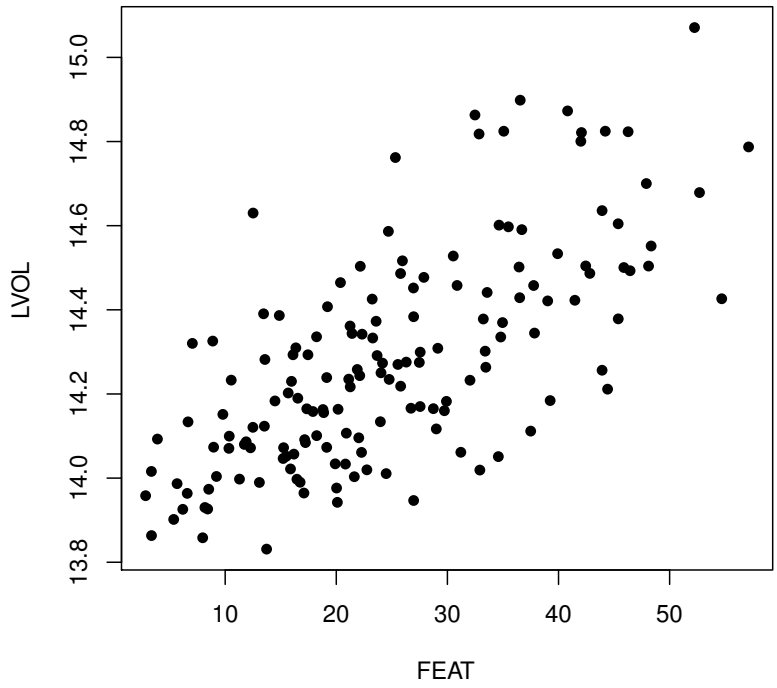
The time series plot indicates *momentum* in the data.

Next we show scatterplots between y and each x_i.

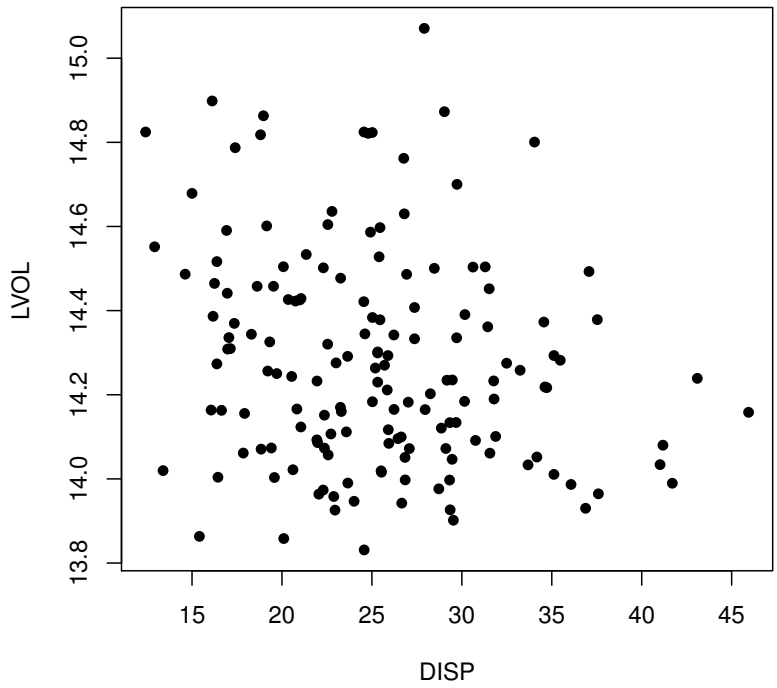
```
> plot(PROMP,LVOL,pch=16)
```



```
> plot(FEAT,LVOL,pch=16)
```




```
> plot(DISP,LVOL,pch=16)
```



What can we observe from these pairwise plots?

- There is a *negative correlation* between *LVOL* and *PROMP*.
- There is a *positive correlation* between *LVOL* and *FEAT*.
- There is *little or no correlation* between *LVOL* and *DISP*, but this might have been blurred by the other input variables.

Therefore, we should regress *LVOL* on *PROMP* and *FEAT* first.

We run a multiple linear regression model using x_1 and x_2 as explanatory variables:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.}$$

```
> reg <- lm(LVOL~PROMP + FEAT)
> summary(reg)

Call:
lm(formula = LVOL ~ PROMP + FEAT)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32734 -0.08519 -0.01011  0.08471  0.30804

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.1500102  0.2487489   68.94   <2e-16 ***
PROMP       -0.9042636  0.0694338  -13.02   <2e-16 ***
FEAT         0.0100666  0.0008827   11.40   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1268 on 153 degrees of freedom
Multiple R-squared:  0.756,     Adjusted R-squared:  0.7528
F-statistic:   237 on 2 and 153 DF,  p-value: < 2.2e-16
```

We begin by performing a joint test of significance by testing $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$. The test statistic value is given in the regression ANOVA table as $f = 237$, with a corresponding p-value of 0.000 (to three decimal places). Hence H_0 is rejected and we have strong evidence that at least one slope coefficient is not equal to zero.

Next we consider individual t tests of $\mathrm{H}_0 : \beta_1 = 0$ and $\mathrm{H}_0 : \beta_2 = 0$. The respective test statistic values are -13.02 and 11.40, both with p-values of 0.000 (to three decimal places) indicating that both slope coefficients are non-zero.

Turning to the estimated coefficients, $\widehat{\beta}_1 = -0.904$ (to three decimal places) which indicates that *LVOL* decreases as *PROMP* increases controlling for *FEAT*. Also, $\widehat{\beta}_2 = 0.010$ (to three decimal places) which indicates that *LVOL* increases as *FEAT* increases, controlling for *PROMP*.

We could also compute 95% confidence intervals, given by:

$$\boldsymbol{\widehat{\beta}_i \pm t_{0.025,\, n-3} \times \text{E.S.E.}(\widehat{\beta}_i).}$$

Since $n - 3 = 153$ is large, $t_{0.025,\, n-3} \approx z_{0.025} = 1.96$.

$R^2 = 0.756$. Therefore, 75.6% of the variation of *LVOL* can be explained (jointly) with *PROMP* and *FEAT*. However, a large R^2 does not necessarily mean that the fitted model is useful. For the estimation of coefficients and predicting y, the absolute measure 'Residual SS' (or $\widehat{\sigma}^2$) plays a critical role in determining the accuracy of the model.

Consider now introducing *DISP* into the regression model to give three explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

The reason for adding the third variable is that one would expect *DISP* to have an impact on sales and we may wish to estimate its magnitude.

```
> reg <- lm(LVOL~PROMP + FEAT + DISP)
> summary(reg)

Call:
lm(formula = LVOL ~ PROMP + FEAT + DISP)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33363 -0.08203 -0.00272  0.07927  0.33812

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.2372251  0.2490226  69.220   <2e-16 ***
PROMP       -0.9564415  0.0726777 -13.160   <2e-16 ***
FEAT         0.0101421  0.0008728  11.620   <2e-16 ***
DISP         0.0035945  0.0016529   2.175   0.0312 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1253 on 152 degrees of freedom
Multiple R-squared:  0.7633,    Adjusted R-squared:  0.7587
F-statistic: 163.4 on 3 and 152 DF,  p-value: < 2.2e-16
```

All the estimated coefficients have the right sign (according to commercial common sense!) and are statistically significant. In particular, the relationship with *DISP* seems real when the other inputs are taken into account. On the other hand, the addition of *DISP* to the model has resulted in a very small reduction in $\widehat{\sigma}$, from $\sqrt{0.0161} = 0.1268$ to $\sqrt{0.0157} = 0.1253$, and correspondingly a slightly higher R^2 (0.7633, i.e. 76.33% of the variation of *LVOL* is explained by the model). Therefore, *DISP* contributes very little to ‘explaining’ the variation of *LVOL* after the other two explanatory variables, *PROMP* and *FEAT*, are taken into account.

Intuitively, we would expect a higher R^2 if we add a further explanatory variable to the model. However, the model has become more complex as a result – there is an additional parameter to estimate. Therefore, strictly speaking, we should consider the ‘adjusted R^2’ statistic, although this will not be considered in this course.

Special care should be exercised when predicting with x out of the range of the observations used to fit the model, which is called **extrapolation.**

11.12 Overview of chapter

This chapter has covered the linear regression model with one or more explanatory variables. Least squares estimators were derived for the simple linear regression model, and statistical inference procedures were also covered. The multiple linear regression model and applications using R concluded the chapter.

11.13 Key terms and concepts

- ANOVA decomposition
- Confidence interval
- Independent variable
- Least squares estimation
- Multiple linear regression
- Regression analysis
- Residual
- Slope coefficient
- Coefficient of determination
- Dependent variable
- Intercept
- Linear estimators
- Prediction interval
- Regressor
- Simple linear regression

Facts are stubborn, but statistics are more pliable.
(Mark Twain)

Appendix A
Sampling distributions of statistics

A.1 Worked examples

1. Suppose A, B and C are independent chi-squared random variables with 5, 7 and 10 degrees of freedom, respectively. Calculate:

   (a) $P(B < 12)$

   (b) $P(A + B + C < 14)$

   (c) $P(A - B - C < 0)$

   (d) $P(A^3 + B^3 + C^3 < 0)$.

   In this question, you should use the closest value given in Murdoch and Barnes' *Statistical Tables*. Further approximation is not required.

   Solution:

   (a) $P(B < 12) \approx 0.9$, directly from Table 8, where $B \sim \chi^2_7$.

   (b) $A + B + C \sim \chi^2_{5+7+10} = \chi^2_{22}$, so $P(A + B + C < 14)$ is the probability that such a random variable is less than 14, which is approximately 0.1 from Table 8.

   (c) Transforming and rearranging the probability, we need:

$$P(A < B + C) = P\left(\frac{A}{5} < \frac{B + C}{17} \times \frac{17}{5}\right)$$
$$= P\left(\frac{A/5}{(B + C)/17} < 3.4\right) = P(F < 3.4) \approx 0.975$$

   where $F \sim F_{5,17}$, using Table 9 (practice of which will be covered later in the course[1]).

   (d) A chi-squared random variable only assumes non-negative values. Hence each of A, B and C is non-negative, so $A^3 + B^3 + C^3 \geq 0$, and:

$$P(A^3 + B^3 + C^3 < 0) = 0.$$

2. Suppose $\{Z_i\}$, for $i = 1, 2, \ldots, k$, are independent and identically distributed standard normal random variables, i.e. $Z_i \sim N(0, 1)$, for $i = 1, 2, \ldots, k$.

[1] Although we have yet to 'formally' introduce Table 9 of Murdoch and Barnes' *Statistical Tables*, you should be able to see how this works.

State the distribution of:

(a) Z_1^2

(b) Z_1^2/Z_2^2

(c) $Z_1/\sqrt{Z_2^2}$

(d) $\sum_{i=1}^{k} Z_i/k$

(e) $\sum_{i=1}^{k} Z_i^2$

(f) $(3/2) \times (Z_1^2 + Z_2^2)/(Z_3^2 + Z_4^2 + Z_5^2)$.

Solution:

(a) $Z_1^2 \sim \chi_1^2$

(b) $Z_1^2/Z_2^2 \sim F_{1,1}$

(c) $Z_1/\sqrt{Z_2^2} \sim t_1$

(d) $\sum_{i=1}^{k} Z_i/k \sim N(0, 1/k)$

(e) $\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$

(f) $(3/2) \times (Z_1^2 + Z_2^2)/(Z_3^2 + Z_4^2 + Z_5^2) \sim F_{2,3}$.

3. X_1, X_2, X_3 and X_4 are independent normally distributed random variables each with a mean of 0 and a standard deviation of 3. Find:

(a) $P(X_1 + 2X_2 > 9)$

(b) $P(X_1^2 + X_2^2 > 54)$

(c) $P((X_1^2 + X_2^2) > 99(X_3^2 + X_4^2))$.

Solution:

(a) We have $X_1 \sim N(0, 9)$ and $X_2 \sim N(0, 9)$. Hence $2X_2 \sim N(0, 36)$ and $X_1 + 2X_2 \sim N(0, 45)$. So:

$$P(X_1 + 2X_2 > 9) = P\left(Z > \frac{9}{\sqrt{45}}\right) = P(Z > 1.34) = 0.0901.$$

(b) We have $X_1/3 \sim N(0, 1)$ and $X_2/3 \sim N(0, 1)$. Hence $X_1^2/9 \sim \chi_1^2$ and $X_2^2/9 \sim \chi_1^2$. Therefore, $X_1^2/9 + X_2^2/9 \sim \chi_2^2$. So:

$$P(X_1^2 + X_2^2 > 54) = P(Y > 6) = 0.05$$

where $Y \sim \chi_2^2$.

(c) We have $X_1^2/9 + X_2^2/9 \sim \chi_2^2$ and also $X_3^2/9 + X_4^2/9 \sim \chi_2^2$. So:

$$\frac{X_1^2 + X_2^2}{X_3^2 + X_4^2} = \frac{(X_1^2 + X_2^2)/18}{(X_3^2 + X_4^2)/18} \sim F_{2,2}.$$

Hence:

$$P((X_1^2 + X_2^2) > 99(X_3^2 + X_4^2)) = P(Y > 99) = 0.01$$

where $Y \sim F_{2,2}$.

4. The independent random variables X_1, X_2 and X_3 are each normally distributed with a mean of 0 and a variance of 4. Find:

(a) $P(X_1 > X_2 + X_3)$

(b) $P(X_1^2 > 9.25(X_2^2 + X_3^2))$

(c) $P(X_1 > 5(X_2^2 + X_3^2)^{1/2})$.

Solution:

(a) We have $X_i \sim N(0, 4)$, for $i = 1, 2, 3$, hence:

$$X_1 - X_2 - X_3 \sim N(0, 12).$$

So:

$$P(X_1 > X_2 + X_3) = P(X_1 - X_2 - X_3 > 0) = P(Z > 0) = 0.5.$$

(b) We have $X_i/2 \sim N(0, 1)$, so $X_i^2/4 \sim \chi_1^2$ for $i = 1, 2, 3$. Hence:

$$\frac{2X_1^2}{X_2^2 + X_3^2} = \frac{(X_1^2/4)/1}{((X_2^2 + X_3^2)/4)/2} \sim F_{1,2}.$$

So:

$$P(X_1^2 > 9.25(X_2^2 + X_3^2)) = P\left(\frac{2X_1^2}{X_2^2 + X_3^2} > 9.25 \times 2\right) = P(Y > 18.5) = 0.05$$

where $Y \sim F_{1,2}$.

(c) We have:

$$P(X_1 > 5(X_2^2 + X_3^2)^{1/2}) = P\left(\frac{X_1}{2} > 5\left(\frac{X_2^2}{4} + \frac{X_3^2}{4}\right)^{1/2}\right)$$

$$= P\left(\frac{X_1}{2} > 5\sqrt{2}\left(\left(\frac{X_2^2}{4} + \frac{X_3^2}{4}\right)^{1/2}\right) \Big/ \sqrt{2}\right)$$

i.e. $P(Y_1 > 5\sqrt{2}Y_2)$, where $Y_1 \sim N(0, 1)$ and $Y_2 \sim \sqrt{\chi_2^2/2}$, or $P(Y_3 > 7.07)$, where $Y_3 \sim t_2$. From Table 7, this is approximately 0.01.

5. The independent random variables X_1, X_2, X_3 and X_4 are each normally distributed with a mean of 0 and a variance of 4. Using Murdoch and Barnes' *Statistical Tables*, derive values for k in each of the following cases:

(a) $P(3X_1 + 4X_2 > 5) = k$

(b) $P(X_1 > k\sqrt{X_3^2 + X_4^2}) = 0.025$

(c) $P(X_1^2 + X_2^2 + X_3^2 < k) = 0.9$

(d) $P(X_2^2 + X_3^2 + X_4^2 > 19X_1^2 + 20X_3^2) = k$.

Solution:

(a) We have $X_i \sim N(0,4)$, for $i = 1,2,3,4$, hence $3X_1 \sim N(0,36)$ and $4X_2 \sim N(0,64)$. Therefore:

$$\frac{3X_1 + 4X_2}{10} = Z \sim N(0,1).$$

So, $P(3X_1 + 4X_2 > 5) = k = P(Z > 0.5) = 0.3085$.

(b) We have $X_i/2 \sim N(0,1)$, for $i = 1,2,3,4$, hence $(X_3^2 + X_4^2)/4 \sim \chi_2^2$. So:

$$P\left(X_1 > k\sqrt{X_3^2 + X_4^2}\right) = 0.025 = P(T > k\sqrt{2})$$

where $T \sim t_2$ and hence $k\sqrt{2} = 4.303$, so $k = 3.04268$.

(c) We have $(X_1^2 + X_2^2 + X_3^2)/4 \sim \chi_3^2$, so:

$$P(X_1^2 + X_2^2 + X_3^2 < k) = 0.9 = P\left(X < \frac{k}{4}\right)$$

where $X \sim \chi_3^2$. Therefore, $k/4 = 6.251$. Hence $k = 25.004$.

(d) $P(X_2^2 + X_3^2 + X_4^2 > 19X_1^2 + 20X_3^2) = k$ simplifies to:

$$P(X_2^2 + X_4^2 > 19(X_1^2 + X_3^2)) = k$$

and:

$$\frac{X_2^2 + X_4^2}{X_1^2 + X_3^2} \sim F_{2,2}.$$

So, from Table 9, $k = 0.05$.

6. Suppose that the heights of students are normally distributed with a mean of 68.5 inches and a standard deviation of 2.7 inches. If 200 random samples of size 25 are drawn from this population with means recorded to the nearest 0.1 inch, find:

(a) the expected mean and standard deviation of the sampling distribution of the mean

(b) the expected number of recorded sample means which fall between 67.9 and 69.2 inclusive

(c) the expected number of recorded sample means falling below 67.0.

Solution:

(a) The sampling distribution of the mean of 25 observations has the same mean as the population, which is 68.5 inches. The standard deviation (standard error) of the sample mean is $2.7/\sqrt{25} = 0.54$.

(b) Notice that the samples are random, so we cannot be sure exactly how many will have means between 67.9 and 69.2 inches. We can work out the probability that the sample mean will lie in this interval using the sampling distribution:

$$\bar{X} \sim N(68.5, (0.54)^2).$$

We need to make a continuity correction, to account for the fact that the recorded means are rounded to the nearest 0.1 inch. For example, the probability that the recorded mean is ≥ 67.9 inches is the same as the probability that the sample mean is > 67.85. Therefore, the probability we want is:

$$\begin{aligned} P(67.85 < X < 69.25) &= P\left(\frac{67.85 - 68.5}{0.54} < Z < \frac{69.25 - 68.5}{0.54}\right) \\ &= P(-1.20 < Z < 1.39) \\ &= \Phi(1.39) - \Phi(-1.20) \\ &= 0.9177 - (1 - 0.1151) \\ &= 0.8026. \end{aligned}$$

As usual, the values of $\Phi(1.39)$ and $\Phi(-1.20)$ can be found from Table 3 of Murdoch and Barnes' *Statistical Tables*. Since there are 200 independent random samples drawn, we can now think of each as a single trial. The recorded mean lies between 67.9 and 69.2 with probability 0.8026 at each trial. We are dealing with a binomial distribution with $n = 200$ trials and probability of success $\pi = 0.8026$. The expected number of successes is:

$$n\pi = 200 \times 0.8026 = 160.52.$$

(c) The probability that the recorded mean is < 67.0 inches is:

$$P(X < 66.95) = P\left(Z < \frac{66.95 - 68.5}{0.54}\right) = P(Z < -2.87) = \Phi(-2.87) = 0.00205$$

so the expected number of recorded means below 67.0 out of a sample of 200 is:

$$200 \times 0.00205 = 0.41.$$

7. If Z is a random variable with a standard normal distribution, what is $P(Z^2 < 3.841)$?

Solution:

We can compute the probability in two different ways. Working with the standard normal distribution, we have:

$$\begin{aligned} P(Z^2 < 3.841) &= P\left(-\sqrt{3.841} < Z < \sqrt{3.841}\right) \\ &= P(-1.96 < Z < 1.96) \\ &= \Phi(1.96) - \Phi(-1.96) \\ &= 0.9750 - (1 - 0.9750) = 0.95. \end{aligned}$$

Alternatively, we can use the fact that Z^2 follows a χ^2_1 distribution. From Table 8 of Murdoch and Barnes' *Statistical Tables* we can see that 3.841 is the 5% right-tail value for this distribution, and so $P(Z^2 < 3.84) = 0.95$, as before.

8. Suppose that X_1 and X_2 are independent $N(0, 4)$ random variables. Compute $P(X_1^2 < 36.84 - X_2^2)$.

Solution:

Rearrange the inequality to obtain:

$$\begin{aligned} P(X_1^2 < 36.84 - X_2^2) &= P(X_1^2 + X_2^2 < 36.84) \\ &= P\left(\frac{X_1^2 + X_2^2}{4} < \frac{36.84}{4}\right) \\ &= P\left(\left(\frac{X_1}{2}\right)^2 + \left(\frac{X_2}{2}\right)^2 < 9.21\right). \end{aligned}$$

Since $X_1/2$ and $X_2/2$ are independent $N(0, 1)$ random variables, the sum of their squares will follow a χ^2_2 distribution. Using Table 8 of Murdoch and Barnes' *Statistical Tables*, we see that 9.210 is the 1% right-tail value, so the probability we are looking for is 0.99.

9. Suppose that X_1 X_2 and X_3 are independent $N(0, 1)$ random variables, while Y (independently) follows a χ^2_5 distribution. Compute $P(X_1^2 + X_2^2 < 7.236Y - X_3^2)$.

Solution:

Rearranging the inequality gives:

$$\begin{aligned} P(X_1^2 + X_2^2 < 7.236Y - X_3^2) &= P(X_1^2 + X_2^2 + X_3^2 < 7.236Y) \\ &= P\left(\frac{X_1^2 + X_2^2 + X_3^2}{Y} < 7.236\right) \\ &= P\left(\frac{(X_1^2 + X_2^2 + X_3^2)/3}{Y/5} < \frac{5}{3} \times 7.236\right) \\ &= P\left(\frac{(X_1^2 + X_2^2 + X_3^2)/3}{Y/5} < 12.060\right). \end{aligned}$$

Since $X_1^2 + X_2^2 + X_3^2 \sim \chi^2_3$, we have a ratio of independent χ^2_3 and χ^2_5 random variables, each divided by its degrees of freedom. By definition, this follows an $F_{3,5}$ distribution. From Table 9 of Murdoch and Barnes' *Statistical Tables*, we see that 12.060 is the 1% upper-tail value for this distribution, so the probability we want is equal to 0.99.

10. Compare the normal distribution approximation to the exact values for the upper-tail probabilities for the binomial distribution with 100 trials and probability of success 0.1.

Solution:

Let $R \sim \text{Bin}(100, 0.1)$ denote the exact number of successes. It has mean and variance:

$$\text{E}(R) = n\pi = 100 \times 0.1 = 10$$

and:

$$\text{Var}(R) = n\pi(1-\pi) = 100 \times 0.1 \times 0.9 = 9$$

so we use the approximation $R \mathrel{\dot\sim} N(10, 9)$ or, equivalently:

$$\frac{R-10}{\sqrt{9}} = \frac{R-10}{3} \mathrel{\dot\sim} N(0,1).$$

Applying a continuity correction of 0.5 (for example, 7.8 successes are rounded up to 8) gives:

$$P(R \geq r) \approx P\left(Z > \frac{r-0.5-10}{3}\right).$$

The results are summarised in the following table. The first column is the number of successes; the second gives the exact binomial probabilities; the third column lists the corresponding z-values (with the continuity correction); and the fourth gives the probabilities for the normal approximation.

Although the agreement between columns two and four is not too bad, you may think it is not as close as you would like for some applications.

r	$P(R \geq r)$	$z = (r-0.5-10)/3$	$P(Z > z)$
1	0.999973	−3.1667	0.999229
2	0.999678	−2.8333	0.997697
3	0.998055	−2.5000	0.993790
4	0.992164	−2.1667	0.984870
5	0.976289	−1.8333	0.966624
6	0.942423	−1.5000	0.933193
7	0.882844	−1.1667	0.878327
8	0.793949	−0.8333	0.797672
9	0.679126	−0.5000	0.691462
10	0.548710	−0.1667	0.566184
11	0.416844	0.1667	0.433816
12	0.296967	0.5000	0.308538
13	0.198179	0.8333	0.202328
14	0.123877	1.1667	0.121673
15	0.072573	1.5000	0.066807
16	0.039891	1.8333	0.033376
17	0.020599	2.1667	0.015130
18	0.010007	2.5000	0.006210
19	0.004581	2.8333	0.002303
20	0.001979	3.1667	0.000771
21	0.000808	3.5000	0.000233
22	0.000312	3.8333	0.000063
23	0.000114	4.1667	0.000015
24	0.000040	4.5000	0.000003
25	0.000013	4.8333	0.000001
26	0.000004	5.1667	0.000000

A.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix G.

1. (a) Suppose $\{X_1, X_2, X_3, X_4\}$ is a random sample of size $n = 4$ from the Bernoulli(0.2) distribution. What is the distribution of $\sum_{i=1}^{n} X_i$ in this case?

   (b) Write down the sampling distribution of $\bar{X} = \sum_{i=1}^{n} X_i/n$ for the sample considered in (a). In other words, write down the possible values of $\bar{X}$ and their probabilities.

   Hint: what are the possible values of $\sum_i X_i$, and their probabilities?

   (c) Suppose we have a random sample of size $n = 100$ from the Bernoulli(0.2) distribution. What is the approximate sampling distribution of $\bar{X}$ suggested by the central limit theorem in this case? Use this distribution to calculate an approximate value for the probability that $\bar{X} > 0.3$. (The true value of this probability is 0.0061.)

2. Suppose that we plan to take a random sample of size n from a normal distribution with mean μ and standard deviation $\sigma = 2$.

   (a) Suppose $\mu = 4$ and $n = 20$.

      i. What is the probability that the mean $\bar{X}$ of the sample is greater than 5?

      ii. What is the probability that $\bar{X}$ is smaller than 3?

      iii. What is $P(|\bar{X} - \mu| \leq 1)$ in this case?

   (b) How large should n be in order that $P(|\bar{X} - \mu| \leq 0.5) \geq 0.95$ for every possible value of μ?

   (c) It is claimed that the true value of μ is 5 in a population. A random sample of size $n = 100$ is collected from this population, and the mean for this sample is $\bar{x} = 5.8$. Based on the result in (b), what would you conclude from this value of $\bar{X}$?

3. A random sample of 25 audits is to be taken from a company's total audits, and the average value of these audits is to be calculated.

   (a) Explain what you understand by the sampling distribution of this average and discuss its relationship to the population mean.

   (b) Is it reasonable to assume that this sampling distribution is normal?

   (c) If the population of all audits has a mean of £54 and a standard deviation of £10, find the probability that:

      i. the sample mean will be greater than £60

      ii. the sample mean will be within 5% of the population mean.

Did you hear the one about the statistician? Probably.
(Anon)

Appendix B
Point estimation

B.1 Worked examples

1. Let X_1 and X_2 be two independent random variables with the same mean, μ, and the same variance, $\sigma^2 < \infty$. Let $\widehat{\mu} = aX_1 + bX_2$ be an estimator of μ, where a and b are two non-zero constants.

   (a) Identify the condition on a and b to ensure that $\widehat{\mu}$ is an unbiased estimator of μ.

   (b) Find the minimum mean squared error (MSE) among all unbiased estimators of μ.

   Solution:

   (a) Let $\mathrm{E}(\widehat{\mu}) = \mathrm{E}(aX_1 + bX_2) = a\,\mathrm{E}(X_1) + b\,\mathrm{E}(X_2) = (a+b)\mu$. Hence $a + b = 1$ is the condition for $\widehat{\mu}$ to be an unbiased estimator of μ.

   (b) Under this condition, noting that $b = 1 - a$, we have:

   $$\mathrm{MSE}(\widehat{\mu}) = \mathrm{Var}(\widehat{\mu}) = a^2\mathrm{Var}(X_1) + b^2\mathrm{Var}(X_2) = (a^2 + b^2)\sigma^2 = (2a^2 - 2a + 1)\sigma^2.$$

   Setting $\mathrm{d}\,\mathrm{MSE}(\widehat{\mu})/\mathrm{d}a = (4a - 2)\sigma^2 = 0$, we have $a = 0.5$, and hence $b = 0.5$. Therefore, among all unbiased *linear* estimators, the sample mean $(X_1 + X_2)/2$ has the minimum variance.

   Remark: Let $\{X_1, X_, \ldots, X_n\}$ be a random sample from a population with finite variance. The sample mean $\bar{X}$ has the minimum variance among *all* unbiased linear estimators of the form $\sum_{i=1}^{n} a_i X_i$, hence it is the *best linear unbiased estimator* (**BLUE**(!)).

2. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the (continuous) uniform distribution such that $X \sim \mathrm{Uniform}[0, \theta]$, where $\theta > 0$. Find the method of moments estimator (MME) of θ.

   Solution:

   The pdf of X_i is:

   $$f(x_i; \theta) = \begin{cases} \theta^{-1} & \text{for } 0 \le x_i \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

   Therefore:

   $$\mathrm{E}(X_i) = \frac{1}{\theta}\int_0^{\theta} x_i\,\mathrm{d}x_i = \frac{1}{\theta}\left[\frac{x_i^2}{2}\right]_0^{\theta} = \frac{\theta}{2}.$$

Therefore, setting $\widehat{\mu}_1 = M_1$, we have:

$$\frac{\widehat{\theta}}{2} = \bar{X} \quad \Rightarrow \quad \widehat{\theta} = 2\bar{X} = 2\sum_{i=1}^{n} \frac{X_i}{n}.$$

3. Let $X \sim \text{Bin}(n, \pi)$, where n is known. Find the methods of moments estimator (MME) of π.

Solution:

The pf of the binomial distribution is:

$$P(X = x) = \frac{n!}{x!\,(n-x)!}\,\pi^x(1-\pi)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$

and 0 otherwise. Therefore:

$$\text{E}(X) = \sum_{x=0}^{n} x\,P(X = x) = \sum_{x=1}^{n} x\,\frac{n!}{x!\,(n-x)!}\,\pi^x(1-\pi)^{n-x} = \sum_{x=1}^{n} \frac{n!}{(x-1)!\,(n-x)!}\,\pi^x(1-\pi)^{n-x}.$$

Let $m = n - 1$ and write $j = x - 1$, then $(n - x) = (m - j)$, and:

$$\text{E}(X) = \sum_{j=0}^{m} \frac{nm!}{j!\,(m-j)!}\,\pi\pi^j(1-\pi)^{m-j} = n\pi \sum_{j=0}^{m} \frac{m!}{j!\,(m-j)!}\,\pi^j(1-\pi)^{m-j}.$$

Therefore, $\text{E}(X) = n\pi$, and hence $\widehat{\pi} = X/n$.

4. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the distribution with pdf:

$$f(x) = \begin{cases} \lambda \exp(-\lambda(x-a)) & \text{for } x \geq a \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$. Find the method of moments estimators (MMEs) of λ and a.

Solution:

We have:

$$\text{E}(X) = \int_a^\infty x\,\lambda \exp(-\lambda(x-a))\,\text{d}x = \frac{1}{\lambda}\int_0^\infty (y + \lambda a)\text{e}^{-y}\,\text{d}y = \frac{1}{\lambda} + a$$

and:

$$\text{E}(X^2) = \int_a^\infty x^2\lambda \exp(-\lambda(x-a))\,\text{d}x = \int_0^\infty \left(\frac{y}{\lambda} + a\right)^2 \text{e}^{-y}\,\text{d}y = \frac{2}{\lambda^2} + \frac{2a}{\lambda} + a^2.$$

Therefore, the MMEs are the solutions to the equations:

$$\bar{X} = \frac{1}{\widehat{\lambda}} + \widehat{a} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{2}{\widehat{\lambda}^2} + \frac{2\widehat{a}}{\widehat{\lambda}} + \widehat{a}^2.$$

Actually, the explicit solutions may be obtained as follows:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{2}{\widehat{\lambda}^2} + \frac{2\widehat{a}}{\widehat{\lambda}} + \widehat{a}^2 - \left(\frac{1}{\widehat{\lambda}} + \widehat{a}\right)^2 = \frac{1}{\widehat{\lambda}^2}.$$

Hence:

$$\widehat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right)^{-1/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{-1/2}.$$

Consequently:

$$\widehat{a} = \bar{X} - \frac{1}{\widehat{\lambda}}.$$

5. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the distribution $N(\mu, 1)$. Find the maximum likelihood estimator (MLE) of μ.

Solution:

The joint pdf of the observations is:

$$f(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

We write the above as a function of μ only:

$$L(\mu) = C \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

where $C > 0$ is a constant. The MLE $\widehat{\mu}$ maximises this function, and also maximises the function:

$$l(\mu) = \ln L(\mu) = -\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2 + \log(C).$$

Therefore, the MLE effectively minimises $\sum_{i=1}^{n}(X_i - \mu)^2$, i.e. the MLE is also the least squares estimator (LSE), i.e. $\widehat{\mu} = \bar{X}$.

6. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a Poisson distribution with mean $\lambda > 0$. Find the maximum likelihood estimator (MLE) of λ.

Solution:

The probability function is:

$$P(X = x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}.$$

The likelihood and log-likelihood functions are, respectively:

$$L(\lambda) = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \frac{\mathrm{e}^{-n\lambda}\lambda^{n\bar{X}}}{\prod_{i=1}^{n} X_i!}$$

and:

$$l(\lambda) = \ln L(\lambda) = n\bar{X}\ln(\lambda) - n\lambda + C = n(\bar{X}\ln(\lambda) - \lambda) + C$$

where C is a constant (i.e. it may depend on X_i but cannot depend on the parameter). Setting:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l(\lambda) = n\left(\frac{\bar{X}}{\widehat{\lambda}} - 1\right) = 0$$

we obtain the MLE $\widehat{\lambda} = \bar{X}$, which is also the MME.

7. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from the (continuous) uniform distribution Uniform$[0, \theta]$, where $\theta > 0$ is unknown.
   (a) Find the maximum likelihood estimator (MLE) of θ.
   (b) If $n = 3$, $x_1 = 0.2$, $x_2 = 3.6$ and $x_3 = 1.1$, what is the maximum likelihood estimate of θ?

Solution:

(a) The pdf of Uniform$[0, \theta]$ is:

$$f(x;\theta) = \begin{cases} \theta^{-1} & \text{for } 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

The joint pdf is:

$$f(x_1, x_2, \ldots, x_n; \theta) = \begin{cases} \theta^{-n} & \text{for } 0 \le x_1, x_2, \ldots, x_n \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

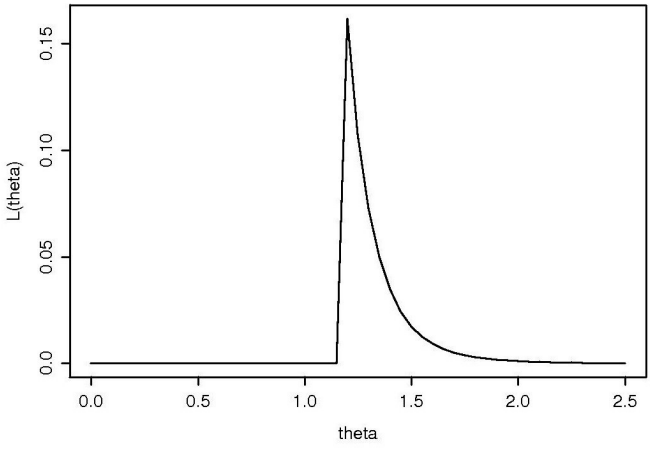
In fact $f(x_1, x_2, \ldots, x_n; \theta)$, as a function of θ, is the likelihood function, $L(\theta)$. The maximum likelihood estimator of θ is the value at which the likelihood function $L(\theta)$ achieves its maximum. Note:

$$L(\theta) = \begin{cases} \theta^{-n} & \text{for } X_{(n)} \le \theta \\ 0 & \text{otherwise} \end{cases}$$

where:

$$X_{(n)} = \max_i X_i.$$

Hence the MLE is $\widehat{\theta} = X_{(n)}$, which is different from the MME. For example, if $x_{(n)} = 1.16$, we have:



(b) For the given data, the maximum observation is $x_{(3)} = 3.6$. Therefore, the maximum likelihood estimate is $\widehat{\theta} = 3.6$.

8. Use the observed random sample $x_1 = 8.2$, $x_2 = 10.6$, $x_3 = 9.1$ and $x_4 = 4.9$ to calculate the maximum likelihood estimate of λ in the exponential pdf:

$$f(x;\lambda) = \begin{cases} \lambda \mathrm{e}^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Solution:

We derive a general formula with a random sample $\{X_1, X_2, \ldots, X_n\}$ first. The joint pdf is:

$$f(x_1, x_2, \ldots, x_n; \lambda) = \begin{cases} \lambda^n \mathrm{e}^{-\lambda n \bar{x}} & \text{for } x_1, x_2, \ldots, x_n \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

With all $x_i \geq 0$, $L(\lambda) = \lambda^n \mathrm{e}^{-\lambda n \bar{X}}$, hence the log-likelihood function is:

$$l(\lambda) = \ln L(\lambda) = n \ln(\lambda) - \lambda n \bar{X}.$$

Setting:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l(\lambda) = \frac{n}{\widehat{\lambda}} - n\bar{X} = 0 \quad \Rightarrow \quad \widehat{\lambda} = \frac{1}{\bar{X}}.$$

For the given sample, $\bar{x} = (8.2 + 10.6 + 9.1 + 4.9)/4 = 8.2$. Therefore, $\widehat{\lambda} = 0.1220$.

9. The following data show the number of occupants in passenger cars observed during one hour at a busy junction. It is assumed that these data follow a geometric distribution with pf:

$$p(x;\pi) = \begin{cases} (1-\pi)^{x-1}\pi & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Number of occupants	1	2	3	4	5	≥ 6	Total
Frequency	678	227	56	28	8	14	1,011

Find the maximum likelihood estimate of π.

Solution:

The sample size is $n = 1{,}011$. If we know all the 1,011 observations, the joint probability function for $x_1, x_2, \ldots, x_{1,011}$ is:

$$L(\pi) = \prod_{i=1}^{1,011} p(x_i;\pi).$$

However, we only know that there are 678 x_is equal to 1, 227 x_is equal to 2, ..., and 14 x_is equal to some integers not smaller than 6.

Note that:

$$P(X_i \geq 6) = \sum_{x=6}^{\infty} p(x;\pi) = \pi(1-\pi)^5(1+(1-\pi)+(1-\pi)^2+\cdots)$$

$$= \pi(1-\pi)^5 \times \frac{1}{\pi}$$

$$= (1-\pi)^5.$$

Hence we may only use:

$$\begin{aligned} L(\pi) &= p(1,\pi)^{678}p(2,\pi)^{227}p(3,\pi)^{56}p(4,\pi)^{28}p(5,\pi)^{8}((1-\pi)^5)^{14} \\ &= \pi^{1,011-14}(1-\pi)^{227+56\times 2+28\times 3+8\times 4+14\times 5} \\ &= \pi^{997}(1-\pi)^{525} \end{aligned}$$

hence:

$$l(\pi) = \ln L(\pi) = 997\ln(\pi) + 525\ln((1-\pi)).$$

Setting:

$$\frac{\mathrm{d}}{\mathrm{d}\pi} l(\pi) = \frac{997}{\widehat{\pi}} - \frac{525}{1-\widehat{\pi}} = 0 \quad \Rightarrow \quad \widehat{\pi} = \frac{997}{997+525} = 0.655.$$

Remark: Since $P(X_i = 1) = \pi$, $\widehat{\pi} = 0.655$ indicates that about 2/3 of cars have only one occupant. Note $\mathrm{E}(X_i) = 1/\pi$. In order to ensure that the average number of occupants is not smaller than k, we require $\pi < 1/k$.

10. Let $\{X_1, X_2, \ldots, X_n\}$, where $n > 2$, be a random sample from an unknown population with mean θ and variance σ^2. We want to choose between two estimators of θ, $\widehat{\theta}_1 = \bar{X}$ and $\widehat{\theta}_2 = (X_1 + X_2)/2$. Which is the better estimator of θ?

Solution:

Let us consider the bias first. The estimator $\widehat{\theta}_1$ is just the sample mean, so we know that it is unbiased. The estimator $\widehat{\theta}_2$ has expectation:

$$\mathrm{E}(\widehat{\theta}_2) = \mathrm{E}\left(\frac{X_1+X_2}{2}\right) = \frac{\mathrm{E}(X_1)+\mathrm{E}(X_2)}{2} = \frac{\theta+\theta}{2} = \theta$$

so it is also an unbiased estimator of θ.

Next, we consider the variances of the two estimators. We have:

$$\mathrm{Var}(\widehat{\theta}_1) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

and:

$$\mathrm{Var}(\widehat{\theta}_2) = \mathrm{Var}\left(\frac{X_1+X_2}{2}\right) = \frac{\mathrm{Var}(X_1)+\mathrm{Var}(X_2)}{4} = \frac{\sigma^2+\sigma^2}{4} = \frac{\sigma^2}{2}.$$

Since $n > 2$, we can see that $\widehat{\theta}_1$ has a lower variance than $\widehat{\theta}_2$, so it is a better estimator. Unsurprisingly, we obtain a better estimator of θ by considering the whole sample, rather than just the first two values.

11. Show that the MSE of an estimator $\widehat{\theta}$ can be written as:

$$\text{MSE}(\widehat{\theta}) = \text{Var}(\widehat{\theta}) + \Big(\text{Bias}(\widehat{\theta})\Big)^2.$$

Solution:

We need to introduce the term $\text{E}(\widehat{\theta})$ inside the expectation, so we add and subtract it to obtain:

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \text{E}((\widehat{\theta} - \theta)^2) \\
&= \text{E}\left(((\widehat{\theta} - \text{E}(\widehat{\theta})) - (\theta - \text{E}(\widehat{\theta})))^2\right) \\
&= \text{E}\left((\widehat{\theta} - \text{E}(\widehat{\theta}))^2 - 2(\widehat{\theta} - \text{E}(\widehat{\theta}))(\theta - \text{E}(\widehat{\theta})) + (\theta - \text{E}(\widehat{\theta}))^2\right) \\
&= \text{E}((\widehat{\theta} - \text{E}(\widehat{\theta}))^2) - 2\text{E}((\widehat{\theta} - \text{E}(\widehat{\theta}))(\theta - \text{E}(\widehat{\theta}))) + \text{E}((\theta - \text{E}(\widehat{\theta}))^2).
\end{aligned}$$

The first term in this expression is, by definition, the variance of $\widehat{\theta}$. The final term is:

$$\text{E}((\theta - \text{E}(\widehat{\theta}))^2) = (\theta - \text{E}(\widehat{\theta}))^2 = (\text{E}(\widehat{\theta}) - \theta)^2 = (\text{Bias}(\widehat{\theta}))^2$$

because θ and $\text{E}(\widehat{\theta})$ are both constants, and are not affected by the expectation operator. It remains to be shown that the middle term is equal to zero. We have:

$$\text{E}\left((\widehat{\theta} - \text{E}(\widehat{\theta}))(\theta - \text{E}(\widehat{\theta}))\right) = (\theta - \text{E}(\widehat{\theta}))\,\text{E}(\widehat{\theta} - \text{E}(\widehat{\theta})) = (\theta - \text{E}(\widehat{\theta}))(\text{E}(\widehat{\theta}) - \text{E}(\widehat{\theta})) = 0$$

which concludes our proof.

12. Find the MSEs of the estimators in Question 10.

Solution:

The MSEs are:

$$\text{MSE}(\widehat{\theta}_1) = \text{Var}(\widehat{\theta}_1) + (\text{Bias}(\widehat{\theta}_1))^2 = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

and:

$$\text{MSE}(\widehat{\theta}_2) = \text{Var}(\widehat{\theta}_2) + (\text{Bias}(\widehat{\theta}_2))^2 = \frac{\sigma^2}{2} + 0 = \frac{\sigma^2}{2}.$$

Note that the MSE of an unbiased estimator is equal to its variance.

13. Are the estimators in Question 10 (mean-square) consistent?

Solution:

The estimator $\widehat{\theta}_1$ has MSE equal to σ^2/n, which converges to 0 as $n \to \infty$. The estimator $\widehat{\theta}_2$ has MSE equal to $\sigma^2/2$, which stays constant as $n \to \infty$. Therefore, $\widehat{\theta}_1$ is a (mean-square) consistent estimator of θ, whereas $\widehat{\theta}_2$ is not.

14. Suppose that we have a random sample $\{X_1, X_2, \ldots, X_n\}$ from a Uniform$[-\theta, \theta]$ distribution. Find the method of moments estimator of θ.

Solution:

The mean of the Uniform$[a, b]$ distribution is $(a + b)/2$. In our case, this gives $\mathrm{E}(X) = (-\theta + \theta)/2 = 0$. The first population moment does not depend on θ, so we need to move to the next (i.e. second) population moment.

Recall that the variance of the Uniform$[a, b]$ distribution is $(b - a)^2/12$. Hence the second population moment is:

$$\mathrm{E}(X^2) = \mathrm{Var}(X) + \mathrm{E}(X)^2 = \frac{(\theta - (-\theta))^2}{12} + 0^2 = \frac{\theta^2}{3}.$$

We set this equal to the second sample moment to obtain:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{\widehat{\theta}^2}{3}.$$

Therefore, the method of moments estimator of θ is:

$$\widehat{\theta}_{MM} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}.$$

15. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a Bin(m, π) distribution, with both m and π unknown. Find the method of moments estimators of m, the number of trials, and π, the probability of success.

Solution:

There are two unknown parameters, so we need two equations. The expectation and variance of a Bin(m, π) distribution are $m\pi$ and $m\pi(1 - \pi)$, respectively, so we have:

$$\mu_1 = \mathrm{E}(X) = m\pi$$

and:

$$\mu_2 = \mathrm{Var}(X) + \mathrm{E}(X)^2 = m\pi(1 - \pi) + (m\pi)^2.$$

Setting the first two sample and population moments equal gives:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{m}\widehat{\pi} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \widehat{m}\widehat{\pi}(1 - \widehat{\pi}) + (\widehat{m}\widehat{\pi})^2.$$

The two equations need to be solved simultaneously. Solving the first equation for $\widehat{\pi}$ gives:

$$\widehat{\pi} = \frac{\sum_{i=1}^{n} X_i/n}{\widehat{m}} = \frac{\bar{X}}{\widehat{m}}.$$

Now we can substitute $\widehat{\pi}$ into the second moment equation to obtain:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \widehat{m}\,\frac{\bar{X}}{\widehat{m}}\left(1 - \frac{\bar{X}}{\widehat{m}}\right) + \left(\widehat{m}\,\frac{\bar{X}}{\widehat{m}}\right)^2$$

which we now solve for $\widehat{m}$ to find the method of moments estimator:

$$\widehat{m}_{MM} = \frac{\bar{X}^2}{\bar{X}^2 - \left(\sum_{i=1}^{n} X_i^2/n - \bar{X}\right)}.$$

16. Consider again the Uniform$[-\theta, \theta]$ distribution from Question 5. Suppose that we observe the following data:

$$1.8, \quad 0.7, \quad -0.2, \quad -1.8, \quad 2.8, \quad 0.6, \quad -1.3 \quad \text{and} \quad -0.1.$$

Estimate θ using the method of moments.

Solution:

The point estimate is:

$$\widehat{\theta}_{MM} = \sqrt{\frac{3}{8}\sum_{i=1}^{8} x_i^2} \approx 2.518$$

which implies that the data came from a Uniform$[-2.518, 2.518]$ distribution. However, this clearly cannot be true since the observation $x_5 = 2.8$ falls outside this range! The method of moments does not take into account that all of the observations need to lie in the interval $[-\theta, \theta]$, and so it fails to produce a useful estimate.

17. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from an Exp(λ) distribution. Find the MLE of λ.

Solution:

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda X_i} = \lambda^n \mathrm{e}^{-\lambda \sum_i X_i} = \lambda^n \mathrm{e}^{-\lambda n \bar{X}}$$

so the log-likelihood function is:

$$l(\lambda) = \ln(\lambda^n \mathrm{e}^{-\lambda n \bar{X}}) = n \ln(\lambda) - \lambda n \bar{X}.$$

Differentiating and setting equal to zero gives:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} l(\lambda) = \frac{n}{\widehat{\lambda}} - n\bar{X} = 0 \quad \Rightarrow \quad \widehat{\lambda} = \frac{1}{\bar{X}}.$$

The second derivative of the log-likelihood function is:

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} l(\lambda) = -\frac{n}{\lambda^2}$$

which is always negative, hence the MLE $\widehat{\lambda} = 1/\bar{X}$ is indeed a maximum. This happens to be the same as the method of moments estimator of λ.

18. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a $N(\mu, \sigma^2)$ distribution. Find the MLE of σ^2 if:

(a) μ is known

(b) μ is unknown.

In each case, work out if the MLE is an unbiased estimator of σ^2.

Solution:

The likelihood function is:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

so the log-likelihood function is:

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Differentiating with respect to σ^2 and setting the derivative equal to zero gives:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} l(\mu, \sigma^2) = -\frac{n}{2}\frac{1}{\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0.$$

If μ is known, we can solve this equation for $\widehat{\sigma}^2$:

$$\frac{n}{2}\frac{1}{\widehat{\sigma}^2} = \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{n}(X_i - \mu)^2 \quad \Rightarrow \quad \frac{n}{2}\widehat{\sigma}^2 = \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2 \quad \Rightarrow \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

The second derivative is always negative, so we conclude that the MLE:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

is indeed a maximum. We can work out the bias of this estimator directly:

$$\mathrm{E}(\widehat{\sigma}^2) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \sigma^2 \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^2}\right)$$

$$= \frac{\sigma^2}{n}\sum_{i=1}^{n}\mathrm{E}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

$$= \frac{\sigma^2}{n}\sum_{i=1}^{n}\mathrm{E}(Z_i^2)$$

$$= \frac{\sigma^2}{n}\, n = \sigma^2$$

where $Z_i = (X_i - \mu)/\sigma$, for $i = 1, 2, \ldots, n$. Therefore, the MLE of σ^2 is an unbiased estimator in this case.

If μ is unknown, we also need to maximise the likelihood function with respect to μ. Here, we consider an alternative method. The likelihood function is:

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

so, whatever the value of σ^2, we need to ensure that $\sum_{i=1}^{n}(X_i - \mu)^2$ is minimised. However, we have:

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

Only the second term on the right-hand side depends on μ and, because of the square, its minimum value is zero. It is minimised when μ is equal to the sample mean, so this is the MLE of μ:

$$\widehat{\mu} = \bar{X}.$$

The resulting MLE of σ^2 is:

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

This is not the same as the sample variance S^2, where we divide by $n-1$ instead of n. The expectation of the MLE of σ^2 is:

$$\begin{aligned}\mathrm{E}(\widehat{\sigma}^2) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) &= \frac{1}{n}\,\mathrm{E}\left((n-1)\,\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)\\ &= \frac{1}{n}\,\mathrm{E}((n-1)S^2)\\ &= \frac{\sigma^2}{n}\,\mathrm{E}\left(\frac{(n-1)S^2}{\sigma^2}\right).\end{aligned}$$

The term inside the expectation, $(n-1)S^2/\sigma^2$, follows a χ^2_{n-1} distribution, and so:

$$\mathrm{E}(\widehat{\sigma}^2) = \frac{\sigma^2}{n}(n-1).$$

This is not equal to σ^2, so the MLE of σ^2 is a biased estimator in this case. (Note that the estimator $\widehat{\sigma}^2 = S^2$ is an unbiased estimator of σ^2.) The bias of the MLE is:

$$\mathrm{Bias}(\widehat{\sigma}^2) = \mathrm{E}(\widehat{\sigma}^2) - \sigma^2 = \frac{\sigma^2}{n}\,(n-1) - \sigma^2 = -\frac{\sigma^2}{n}$$

which tends to zero as $n \to \infty$. In such cases, we say that the estimator is *asymptotically unbiased*.

B.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix G.

1. Based on a random sample of two independent observations from a population with mean μ and standard deviation σ, consider two estimators of μ, X and Y, defined as:
$$X = \frac{X_1}{2} + \frac{X_2}{2} \quad \text{and} \quad Y = \frac{X_1}{3} + \frac{2X_2}{3}.$$
Are X and Y unbiased estimators of μ?

2. Prove that, for normally distributed data, S^2 is an unbiased estimator of σ^2, but that S is a biased estimator of σ.
Hint: if $\bar{X}$ is the sample mean for a random sample of size n, the fact that the observations $\{X_1, X_2, \ldots, X_n\}$ are independent can be used to prove that (in the standard notation):
$$\mathrm{E}(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}.$$

3. A random sample of n independent Bernoulli trials with success probability π results in R successes. Derive an unbiased estimator of $\pi(1-\pi)$.

4. Given a random sample of n values from a normal distribution with unknown mean and variance, consider the following two estimators of σ^2 (the unknown population variance), where $S_{xx} = \sum(X_i - \bar{X})^2$:
$$T_1 = \frac{S_{xx}}{n-1} \quad \text{and} \quad T_2 = \frac{S_{xx}}{n}.$$
For each of these determine its bias, its variance and its mean squared error. Which has the smaller mean squared error?
Hint: use the fact that $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$ for a random sample of size n, or some equivalent formula.

5. Suppose that you are given observations y_1, y_2, y_3 and y_4 such that:
$$y_1 = \alpha + \beta + \varepsilon_1$$
$$y_2 = -\alpha + \beta + \varepsilon_2$$
$$y_3 = \alpha - \beta + \varepsilon_3$$
$$y_4 = -\alpha - \beta + \varepsilon_4.$$
The random variables ε_i, for $i = 1, 2, 3, 4$, are independent and normally distributed with mean 0 and variance σ^2.

(a) Find the least squares estimators of the parameters α and β.

(b) Verify that the least squares estimators in (a) are unbiased estimators of their respective parameters.

(c) Find the variance of the least squares estimator of α.

The group was alarmed to find that if you are a labourer, cleaner or dock worker, you are twice as likely to die than a member of the professional classes.
(The Sunday Times, 31 August 1980)

Appendix C
Interval estimation

C.1 Worked examples

1. (a) Find the length of a 95% confidence interval for the mean of a normal distribution with known variance σ^2.

   (b) Find the minimum sample size such that the width of a 95% confidence interval is not wider than d, where $d > 0$ is a prescribed constant.

Solution:

(a) With an available random sample $\{X_1, X_2, \ldots, X_n\}$ from the normal distribution $N(\mu, \sigma^2)$ with σ^2 known, a 95% confidence interval for μ is of the form:

$$\left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right).$$

Hence the width of the confidence interval is:

$$\left(\bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 2 \times 1.96 \times \frac{\sigma}{\sqrt{n}} = 3.92 \times \frac{\sigma}{\sqrt{n}}.$$

(b) Let $3.92 \times \sigma/\sqrt{n} \leq d$, and so we obtain the condition for the required sample size:

$$n \geq \left(\frac{3.92 \times \sigma}{d}\right)^2 = \frac{15.37 \times \sigma^2}{d^2}.$$

Therefore, in order to achieve the required accuracy, the sample size n should be *at least as large* as $15.37 \times \sigma^2/d^2$.

Note that as the variance $\sigma^2 \nearrow$, the confidence interval width $d \nearrow$, and as the sample size $n \nearrow$, the confidence interval width $d \searrow$. Also, note that when σ^2 is unknown, the width of a confidence interval for μ depends on S. Therefore, the *width is a random variable*.

2. The data below are from a random sample of size $n = 9$ taken from the distribution $N(\mu, \sigma^2)$:

   3.75, 5.67, 3.14, 7.89, 3.40, 9.32, 2.80, 10.34 and 14.31.

   (a) Assume $\sigma^2 = 16$. Find a 95% confidence interval for μ. If the width of such a confidence interval must not exceed 2.5, at least how many observations do we need?

   (b) Suppose σ^2 is now unknown. Find a 95% confidence interval for μ. Compare the result with that obtained in (a) and comment.

   (c) Obtain a 95% confidence interval for σ^2.

Solution:

(a) We have $\bar{x} = 6.74$. For a 95% confidence interval, $\alpha = 0.05$ so we need to find the top $100\alpha/2 = 2.5$th percentile of $N(0,1)$, which is 1.96. Since $\sigma = 4$ and $n = 9$, a 95% confidence interval for μ is:

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}} \Rightarrow \left(6.74 - 1.96 \times \frac{4}{3}, 6.74 + 1.96 \times \frac{4}{3}\right) = (4.13, 9.35).$$

In general, a $100(1-\alpha)\%$ confidence interval for μ is:

$$\left(\bar{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

where z_α denotes the top 100αth percentile of the standard normal distribution, i.e. such that:

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

where $Z \sim N(0,1)$. Hence the width of the confidence interval is:

$$2 \times z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}.$$

For this example, $\alpha = 0.05$, $z_{0.025} = 1.96$ and $\sigma = 4$. Setting the width of the confidence interval to be at most 2.5, we have:

$$2 \times 1.96 \times \frac{\sigma}{\sqrt{n}} = \frac{15.68}{\sqrt{n}} \leq 2.5.$$

Hence:

$$n \geq \left(\frac{15.68}{2.5}\right)^2 = 39.34.$$

So we need a sample of at least 40 observations in order to obtain a 95% confidence interval with a width not greater than 2.5.

(b) When σ^2 is unknown, a 95% confidence interval for μ is:

$$\left(\bar{X} - t_{\alpha/2,\,n-1} \times \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,\,n-1} \times \frac{S}{\sqrt{n}}\right)$$

where $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, and $t_{\alpha,\,k}$ denotes the top 100αth percentile of the Student's t_k distribution, i.e. such that:

$$P(T > t_{\alpha,\,k}) = \alpha$$

for $T \sim t_k$. For this example, $s^2 = 16$, $s = 4$, $n = 9$ and $t_{0.025,\,8} = 2.306$. Hence a 95% confidence interval for μ is:

$$6.74 \pm 2.306 \times \frac{4}{3} \quad \Rightarrow \quad (3.67, 9.81).$$

This confidence interval is much wider than the one obtained in (a). Since we do not know σ^2, we have *less information* available for our estimation. It is only natural that our estimation becomes less accurate.

Note that although the sample size is n, the Student's t distribution used has only $n-1$ degrees of freedom. The loss of 1 degree of freedom in the sample variance is due to not knowing μ. Hence we estimate μ using the data, for which we effectively pay a 'price' of one degree of freedom.

(c) Note $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1} = \chi^2_8$. From Table 8 of Murdoch and Barnes' *Statistical Tables*, for $X \sim \chi^2_8$, we find that:

$$P(X < 2.180) = P(X > 17.535) = 0.025.$$

Hence:

$$P\left(2.180 < \frac{8 \times S^2}{\sigma^2} < 17.535\right) = 0.95.$$

Therefore, the lower bound for σ^2 is $8 \times s^2/17.535 = 7.298$, and the upper bound is $8 \times s^2/2.180 = 58.701$. Therefore, a 95% confidence interval for σ^2, noting $s^2 = 16$, is:

$$(7.30, 58.72).$$

Note that the estimation in this example is rather inaccurate. This is due to two reasons.

i. The sample size is small.

ii. The population variance, σ^2, is large.

3. Assume that the random variable X is normally distributed and that σ^2 is known. What confidence level would be associated with each of the following intervals?

(a) $(\bar{x} - 1.645 \times \sigma/\sqrt{n}, \bar{x} + 2.326 \times \sigma/\sqrt{n})$.

(b) $(-\infty, \bar{x} + 2.576 \times \sigma/\sqrt{n})$.

(c) $(\bar{x} - 1.645 \times \sigma/\sqrt{n}, \bar{x})$.

Solution:

We have $\bar{X} \sim N(\mu, \sigma^2/\sqrt{n})$, hence $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$.

(a) $P(-1.645 < Z < 2.326) = 0.94$, hence a 94% confidence level.

(b) $P(-\infty < Z < 2.576) = 0.995$, hence a 99.5% confidence level.

(c) $P(-1.645 < Z < 0) = 0.45$, hence a 45% confidence level.

4. Five independent samples, each of size n, are to be drawn from a normal distribution where σ^2 is known. For each sample, the interval:

$$\left(\bar{x} - 0.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.06 \times \frac{\sigma}{\sqrt{n}}\right)$$

will be constructed. What is the probability that at least four of the intervals will contain the unknown μ?

Solution:

The probability that the given interval will contain μ is:

$$P(-0.96 < Z < 1.06) = 0.6869.$$

The probability of four or five such intervals is binomial with $n = 5$ and $\pi = 0.6869$, so let the number of such intervals be $Y \sim \text{Bin}(5, 0.6869)$. The required probability is:

$$P(Y \geq 4) = \binom{5}{4}(0.6869)^4(0.3131) + \binom{5}{5}(0.6869)^5 = 0.5014.$$

5. A personnel manager has found that historically the scores on aptitude tests given to applicants for entry-level positions are normally distributed with $\sigma = 32.4$ points. A random sample of nine test scores from the current group of applicants had a mean score of 187.9 points.

   (a) Find an 80% confidence interval for the population mean score of the current group of applicants.

   (b) Based on these sample results, a statistician found for the population mean a confidence interval extending from 165.8 to 210.0 points. Find the confidence level of this interval.

Solution:

(a) We have $n = 9$, $\bar{x} = 187.9$, $\sigma = 32.4$ and $1 - \alpha = 0.80$, hence $\alpha/2 = 0.10$ and, from Table 3 of Murdoch and Barnes' *Statistical Tables*, $P(Z > 1.282) = 1 - \Phi(1.282) = 0.10$. So an 80% confidence interval is:

$$187.9 \pm 1.282 \times \frac{32.4}{\sqrt{9}} \quad \Rightarrow \quad (174.05, 201.75).$$

(b) The half-width of the confidence interval is $210.0 - 187.9 = 22.1$, which is equal to the margin of error, i.e. we have:

$$22.1 = k \times \frac{\sigma}{\sqrt{n}} = k \times \frac{32.4}{\sqrt{9}} \quad \Rightarrow \quad k = 2.05.$$

$P(Z > 2.05) = 1 - \Phi(2.05) = 0.02018 = \alpha/2 \;\Rightarrow\; \alpha = 0.04036$. Hence we have a $100(1 - \alpha)\% = 100(1 - 0.04036)\% \approx 96\%$ confidence interval.

6. A manufacturer is concerned about the variability of the levels of impurity contained in consignments of raw materials from a supplier. A random sample of 10 consignments showed a standard deviation of 2.36 in the concentration of impurity levels. Assume normality.

   (a) Find a 95% confidence interval for the population variance.

   (b) Would a 99% confidence interval for this variance be wider or narrower than that found in (a)?

Solution:

(a) We have $n = 10$, $s^2 = (2.36)^2 = 5.5696$, $\chi^2_{0.975,\,9} = 2.700$ and $\chi^2_{0.025,\,9} = 19.023$. Hence a 95% confidence interval for σ^2 is:

$$\left(\frac{(n-1)s^2}{\chi^2_{0.025,\,n-1}}, \frac{(n-1)s^2}{\chi^2_{0.975,\,n-1}} \right) = \left(\frac{9 \times 5.5696}{19.023}, \frac{9 \times 5.5696}{2.700} \right) = (2.64, 18.57).$$

(b) A 99% confidence interval would be wider since:

$$\chi^2_{0.995,\,n-1} < \chi^2_{0.975,\,n-1} \quad \text{and} \quad \chi^2_{0.005,\,n-1} > \chi^2_{0.025,\,n-1}.$$

7. Why do we not always choose a very high confidence level for a confidence interval?

Solution:

We do not always want to use a very high confidence level because the confidence interval would be very wide. We have a trade-off between the width of the confidence interval and the coverage probability.

8. Suppose that 9 bags of sugar are selected from the supermarket shelf at random and weighed. The weights in grammes are 812.0, 786.7, 794.1, 791.6, 811.1, 797.4, 797.8, 800.8 and 793.2. Construct a 95% confidence interval for the mean weight of all the bags on the shelf. Assume the population is normal.

Solution:

Here we have a random sample of size $n = 9$. The mean is 798.30. The sample variance is $s^2 = 72.76$, which gives a sample standard deviation $s = 8.53$. From Table 7 of Murdoch and Barnes' *Statistical Tables*, the top 2.5th percentile of the t distribution with $n - 1 = 8$ degrees of freedom is 2.306. Therefore, a 95% confidence interval is:

$$\left(798.30 - 2.306 \times \frac{8.53}{\sqrt{9}}, 798.30 + 2.306 \times \frac{8.53}{\sqrt{9}}\right) = (798.30 - 6.56, 798.30 + 6.56)$$
$$= (791.74, 804.86).$$

It is sometimes more useful to write this as 798.30 ± 6.56.

9. Continuing Question 2, suppose we are now told that σ, the population standard deviation, is known to be 8.5 g. Construct a 95% confidence interval using this information.

Solution:

From Table 7 of Murdoch and Barnes' *Statistical Tables*, the top 2.5th percentile of the standard normal distribution $z_{0.025} = 1.96$ (recall $t_\infty = N(0, 1)$) so a 95% confidence interval for the population mean is:

$$\left(798.30 - 1.96 \times \frac{8.5}{\sqrt{9}}, 798.30 + 1.96 \times \frac{8.5}{\sqrt{9}}\right) = (798.30 - 6.53, 798.30 + 6.53)$$
$$= (792.75, 803.85).$$

Again, it may be more useful to write this as 798.30 ± 5.55. Note that this confidence interval is less wide than the one in Question 2, even though our initial estimate s turned out to be very close to the true value of σ.

10. Construct a 90% confidence interval for the variance of the bags of sugar in Question 2. Does the given value of 8.5 g for the population standard deviation seem plausible?

Solution:

We have $n = 9$ and $s^2 = 72.76$. For a 90% confidence interval, we need the bottom and top 5th percentiles of the chi-squared distribution on $n - 1 = 8$ degrees of freedom. These are:

$$\chi^2_{0.95,\,8} = 2.733 \quad \text{and} \quad \chi^2_{0.05,\,8} = 15.507.$$

A 90% confidence interval is:

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\,n-1}}\right) = \left(\frac{(9-1)\times 72.76}{15.507}, \frac{(9-1)\times 72.76}{2.733}\right)$$

$$= (37.536, 213.010).$$

The corresponding values for the standard deviation are:

$$(\sqrt{37.536}, \sqrt{213.010}) = (6.127, 14.595).$$

The given value falls well within this confidence interval, so we have no reason to doubt it.

C.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix G.

1. A business requires an inexpensive check on the value of stock in its warehouse. In order to do this, a random sample of 50 items is taken and valued. The average value of these is computed to be £320.41 with a (sample) standard deviation of £40.60. It is known that there are 9,875 items in the total stock.
   (a) Estimate the total value of the stock to the nearest £10,000.
   (b) Construct a 95% confidence interval for the mean value of all items and hence construct a 95% confidence interval for the total value of the stock.
   (c) You are told that the confidence interval in (b) is too wide for decision-making purposes and you are asked to assess how many more items would need to be sampled to obtain a confidence interval with the same level of confidence, but with half the width.

2. (a) A sample of 954 adults in early 1987 found that 23% of them held shares. Given a UK adult population of 41 million and assuming a proper random sample was taken, construct a 95% confidence interval estimate for the number of shareholders in the UK.
   (b) A 'similar' survey the previous year had found a total of 7 million shareholders. Assuming 'similar' means the same sample size, construct a 95% confidence interval estimate of the increase in shareholders between the two years.

A statistician took the Dale Carnegie Course, improving his confidence from 95% to 99%.
(Anon)

Appendix D
Hypothesis testing

D.1 Worked examples

1. A manufacturer has developed a new fishing line which is claimed to have an average breaking strength of 7 kg, with a standard deviation of 0.25 kg. Assume that the standard deviation figure is correct and that the breaking strength is normally distributed. Suppose that we carry out a test, at the 5% significance level, of $H_0 : \mu = 7$ vs. $H_1 : \mu < 7$. Find the sample size which is necessary for the test to have 90% power if the true breaking strength is 6.95 kg.

   Solution:

   The critical value for the test is $z_{0.95} = -1.645$ and the probability of rejecting H_0 with this test is:

   $$P\left(\frac{\bar{X} - 7}{0.25/\sqrt{n}} < -1.645\right)$$

   which we rewrite as:

   $$P\left(\frac{\bar{X} - 6.95}{0.25/\sqrt{n}} < \frac{7 - 6.95}{0.25/\sqrt{n}} - 1.645\right)$$

   because $\bar{X} \sim N(6.95, (0.25)^2/n)$.

   To ensure power of 90% we need $z_{0.10} = 1.282$ since:

   $$P(Z < 1.282) = 0.90.$$

   Therefore:

   $$\begin{aligned} \frac{7 - 6.95}{0.25/\sqrt{n}} - 1.645 &= 1.282 \\ 0.2 \times \sqrt{n} &= 2.927 \\ \sqrt{n} &= 14.635 \\ n &= 214.1832. \end{aligned}$$

   So to ensure that the test power is at least 90%, we should use a sample size of 215.

   Remark: We see a rather large sample size is required. Hence investigators are encouraged to use sample sizes large enough to come to rational decisions.

2. A doctor claims that the average European is more than 8.5 kg overweight. To test this claim, a random sample of 12 Europeans were weighed, and the difference between their actual weight and their ideal weight was calculated. The data are:

$$14, \quad 12, \quad 8, \quad 13, \quad -1, \quad 10, \quad 11, \quad 15, \quad 13, \quad 20, \quad 7 \quad \text{and} \quad 14.$$

Assuming the data follow a normal distribution, conduct a t test to infer at the 5% significance level whether or not the doctor's claim is true.

Solution:

We have a random sample of size $n = 12$ from $N(\mu, \sigma^2)$, and we test $\mathrm{H}_0 : \mu = 8.5$ vs. $\mathrm{H}_1 : \mu > 8.5$. The test statistic, under H_0, is:

$$T = \frac{\bar{X} - 8.5}{S/\sqrt{n}} = \frac{\bar{X} - 8.5}{S/\sqrt{12}} \sim t_{11}.$$

We reject H_0 if $t > t_{0.05,\,11} = 1.796$. For the given data:

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = 11.333 \quad \text{and} \quad s^2 = \frac{1}{11}\left(\sum_{i=1}^{12} x_i^2 - 12\bar{x}^2\right) = 26.606.$$

Hence:

$$t = \frac{11.333 - 8.5}{\sqrt{26.606/12}} = 1.903 > 1.796 = t_{0.05,\,11}$$

so we reject H_0 at the 5% significance level. There is significant evidence to support the doctor's claim.

3. $\{X_1, X_2, \ldots, X_{21}\}$ represents a random sample of size 21 from a normal population with mean μ and variance σ^2.

(a) Construct a test procedure with a 5% significance level to test the null hypothesis that $\sigma^2 = 8$ against the alternative that $\sigma^2 > 8$.

(b) Evaluate the power of the test for the values of σ^2 given below.

$\sigma^2 =$	8.84	10.04	10.55	11.03	12.99	15.45	17.24

Solution:

(a) We test:

$$\mathrm{H}_0 : \sigma^2 = 8 \quad \text{vs.} \quad \mathrm{H}_1 : \sigma^2 > 8.$$

The test statistic, under H_0, is:

$$T = \frac{(n-1)S^2}{\sigma_0^2} = \frac{20 \times S^2}{8} \sim \chi^2_{20}.$$

With a 5% significance level, we reject the null hypothesis if:

$$t \geq 31.410$$

since $\chi^2_{0.05,\,20} = 31.410$.

(b) To evaluate the power, we need the probability of rejecting H_0 (which happens if $t \geq 31.410$) conditional on the actual value of σ^2, that is:

$$P(T \geq 31.410 \,|\, \sigma^2 = k) = P\left(T \times \frac{8}{k} \geq 31.410 \times \frac{8}{k}\right)$$

where k is the true value of σ^2, noting that:

$$T \times \frac{8}{k} \sim \chi^2_{20}.$$

$\sigma^2 = k$	8.84	10.04	10.55	11.03	12.99	15.45	17.24
$31.410 \times 8/k$	28.4	25.0	23.8	22.8	19.3	16.3	14.6
$\beta(\sigma^2)$	0.10	0.20	0.25	0.30	0.50	0.70	0.80

4. The weights (in grammes) of a group of five-week-old chickens reared on a high-protein diet are 336, 421, 310, 446, 390 and 434. The weights of a second group of chickens similarly reared, except for their low-protein diet, are 224, 275, 393, 282 and 365. Is there evidence that the additional protein has increased the average weight of the chickens? Assume normality.

Solution:

Assuming normally-distributed populations with possibly different means, but the same variance, we test:

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X > \mu_Y.$$

The sample means and standard deviations are $\bar{x} = 389.5$, $\bar{y} = 307.8$, $s_X = 55.40$ and $s_Y = 69.45$. The test statistic and its distribution under H_0 are:

$$T = \sqrt{\frac{n+m-2}{1/n + 1/m}} \times \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sim t_{n+m-2}$$

and we obtain, for the given data, $t = 2.175 > 1.833 = t_{0.05,9}$ hence we reject H_0 that the mean weights are equal and conclude that the mean weight for the high-protein diet is greater at the 5% significance level.

5. Suppose that we have two independent samples from normal populations with known variances. We want to test the H_0 that the two population means are equal against the alternative that they are different. We could use each sample by itself to write down 95% confidence intervals and reject H_0 if these intervals did not overlap. What would be the significance level of this test?

Solution:

Let us assume $H_0 : \mu_X = \mu_Y$ is true, then the two 95% confidence intervals do not overlap if and only if:

$$\bar{X} - 1.96 \times \frac{\sigma_X}{\sqrt{n}} \geq \bar{Y} + 1.96 \times \frac{\sigma_Y}{\sqrt{m}} \quad \text{or} \quad \bar{Y} - 1.96 \times \frac{\sigma_Y}{\sqrt{m}} \geq \bar{X} + 1.96 \times \frac{\sigma_X}{\sqrt{n}}.$$

So we want the probability:

$$P\left(|\bar{X}-\bar{Y}| \geq 1.96 \times \left(\frac{\sigma_X}{\sqrt{n}}+\frac{\sigma_Y}{\sqrt{m}}\right)\right)$$

which is:

$$P\left(\left|\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}\right| \geq 1.96 \times \frac{\sigma_X/\sqrt{n}+\sigma_Y/\sqrt{m}}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}\right).$$

So we have:

$$P\left(|Z| \geq 1.96 \times \frac{\sigma_X/\sqrt{n}+\sigma_Y/\sqrt{m}}{\sqrt{\sigma_X^2/n+\sigma_Y^2/m}}\right)$$

where $Z \sim N(0,1)$. This does not reduce in general, but if we assume $n=m$ and $\sigma_X^2=\sigma_Y^2$, then it reduces to:

$$P(|Z| \geq 1.96 \times \sqrt{2}) = 0.0056.$$

The significance level is about 0.6%, which is much smaller than the usual conventions of 5% and 1%. Putting variability into two confidence intervals makes them more likely to overlap than you might think, and so your chance of incorrectly rejecting the null hypothesis is smaller than you might expect!

6. The following table shows the number of salespeople employed by a company and the corresponding value of sales (in £000s):

Number of salespeople (x)	210	209	219	225	232	221
Sales (y)	206	200	204	215	222	216
Number of salespeople (x)	220	233	200	215	205	227
Sales (y)	210	218	201	212	204	212

Compute the sample correlation coefficient for these data and carry out a formal test for a (linear) relationship between the number of salespeople and sales.

Note that:

$$\sum x_i = 2{,}616, \quad \sum y_i = 2{,}520, \quad \sum x_i^2 = 571{,}500,$$
$$\sum y_i^2 = 529{,}746 \quad \text{and} \quad \sum x_i y_i = 550{,}069.$$

Solution:

We test:

$$\mathrm{H}_0 : \rho = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \rho > 0.$$

The corresponding test statistic and its distribution under H_0 are:

$$T = \frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} \sim t_{n-2}.$$

We find $\widehat{\rho} = 0.8716$ and obtain $t = 5.62 > 2.764 = t_{0.01,\,10}$ and so we reject H_0 at the 1% significance level. Since the test is highly significant, there is overwhelming evidence of a (linear) relationship between the number of salespeople and the value of sales.

7. Two independent samples from normal populations yield the following results:

Sample 1	$n = 5$	$\sum(x_i - \bar{x})^2 = 4.8$
Sample 2	$m = 7$	$\sum(y_i - \bar{y})^2 = 37.2$

Test at the 5% signficance level whether the population variances are the same based on the above data.

Solution:

We test:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad \mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2.$$

Under H_0, the test statistic is:

$$T = \frac{S_1^2}{S_2^2} \sim F_{n-1,\,m-1} = F_{4,\,6}.$$

Critical values are $F_{0.975,\,4,\,6} = 1/F_{0.025,\,6,\,4} = 1/9.20 = 0.11$ and $F_{0.025,\,4,\,6} = 6.23$, using Table 9 of Murdoch and Barnes' *Statistical Tables*. The test statistic value is:

$$t = \frac{4.8/4}{37.2/6} = 0.1935$$

and since $0.11 < 0.1935 < 6.23$ we do not reject H_0, which means there is no evidence of a difference in the variances.

8. Why does it make no sense to use a hypothesis like $\bar{x} = 2$?

Solution:

We can see *immediately* if $\bar{x} = 2$ by calculating the sample mean. Inference is concerned with the population from which the sample was taken. We are not very interested in the sample mean in its own right.

9. (a) Of 100 clinical trials, 5 have shown that wonder-drug 'Zap2' is better than the standard treatment (aspirin). Should we be excited by these results?

   (b) Of the 1,000 clinical trials of 1,000 different drugs this year, 30 trials found drugs which seem better than the standard treatments with which they were compared. The television news reports only the results of those 30 'successful' trials. Should we believe these reports?

   (c) A child welfare officer says that she has a test which always reveals when a child has been abused, and she suggests it be put into general use. What is she saying about Type I and Type II errors for her test?

Solution:

(a) If 5 clinical trials out of 100 report that Zap2 is better, this is consistent with there being no difference whatsoever between Zap2 and aspirin if a 5% Type I error probability is being used for tests in these clinical trials. With a 5% significance level we expect 5 trials in 100 to show spurious significant results.

(b) If the television news reports the 30 successful trials out of 1,000, and those trials use tests with a significance level of 5%, we may well choose to be very cautious about believing the results. We would expect 50 spuriously significant results in the 1,000 trial results.

(c) The welfare officer is saying that the Type II error has probability zero. The test is always positive if the null hypothesis of no abuse is false. On the other hand, the welfare officer is saying *nothing* about the probability of a Type I error. It may well be that the probability of a Type I error is high, which would lead to many false accusations of abuse when no abuse had taken place. One should always think about both types of error when proposing a test.

10. A machine is designed to fill bags of sugar. The weight of the bags is normally distributed with standard deviation σ. If the machine is correctly calibrated, σ should be no greater than 20 g. We collect a random sample of 18 bags and weigh them. The sample standard deviation is found to be equal to 32.48 g. Is there any evidence that the machine is incorrectly calibrated?

Solution:

This is a hypothesis test for the variance of a normal population, so we will use the chi-squared distribution. Let:

$$X_1, X_2, \ldots, X_{18} \sim N(\mu, \sigma^2)$$

be the weights of the bags in the sample. An appropriate test has hypotheses:

$$\mathrm{H}_0 : \sigma^2 = 400 \quad \text{vs.} \quad \mathrm{H}_1 : \sigma^2 > 400.$$

This is a one-sided test, because we are interested in detecting an increase in variance. We compute the value of the test statistic:

$$t = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(18-1) \times (32.48)^2}{(20)^2} = 44.385.$$

At the 5% significance level, the upper-tail value of the chi-squared distribution on $\nu = 18 - 1$ degrees of freedom is $\chi^2_{0.05,\,17} = 27.587$. Our test statistic exceeds this value, so we reject the null hypothesis.

We now move to the 1% significance level. The upper-tail value is $\chi^2_{0.01,\,17} = 33.409$, so we reject H_0 again. We conclude that there is very strong evidence that the machine is incorrectly calibrated.

11. After the machine in Question 3 is calibrated, we collect a new sample of 21 bags. The sample standard deviation of their weights is 23.72 g. Based on this sample, can you conclude that the calibration has reduced the variance of the weights of the bags?

Solution:

Let:

$$Y_1, Y_2, \ldots, Y_{21} \sim N(\mu_Y, \sigma_Y^2)$$

be the weights of the bags in the new sample, and use σ_X^2 to denote the variance of the distribution of the previous sample, to avoid confusion. We want to test for a reduction in variance, so we set:

$$\mathrm{H}_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1 \quad \text{vs.} \quad \mathrm{H}_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1.$$

The value of the test statistic in this case is:

$$\frac{s_X^2}{s_Y^2} = \frac{(32.48)^2}{(23.72)^2} = 1.875.$$

If the null hypothesis is true, the test statistic will follow an $F_{18-1,\,21-1} = F_{17,\,20}$ distribution.

At the 5% significance level, the upper-tail critical value of the $F_{17,\,20}$ distribution is $F_{0.05,\,17,\,20} = 2.17$. Our test statistic does not exceed this value, so we cannot reject the null hypothesis.

We move to the 10% significance level. The upper-tail critical value is $F_{0.10,\,17,\,20} = 1.821$, so we can now reject the null hypothesis (if only barely). We conclude that there is some evidence that the variance is reduced, but it is not very strong evidence.

Notice the difference between the conclusions of these two tests. We have a much more powerful test when we compare our standard deviation of 32.48 g to a *fixed* standard deviation of 25 g, than when we compare it to an *estimated* standard deviation of 23.78 g, even though the values are similar.

D.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix G.

1. A random sample of fibres is known to come from one of two environments, A or B. It is known from past experience that the lengths of fibres from A have a log-normal distribution so that the log-length of an A-type fibre is normally distributed about a mean of 0.80 with a standard deviation of 1.00. (Original units are in microns.)

   The log-lengths of B-type fibres are normally distributed about a mean of 0.65 with a standard deviation of 1.00. In order to identify the environment from which the given sample was taken a subsample of n fibres are to be measured and the classification is to be made on the evidence of these measurements.

   Do not be put off by the log-normal distribution. This simply means that it is the logs of the data, rather than the original data, which have a normal distribution. If X represents the log of a fibre length for fibres from A, then $X \sim N(0.8, 1)$.

   (a) If $n = 50$ and the sample is attributed to type A if the sample mean of log-lengths exceeds 0.75, determine the error probabilities.

   (b) What sample size and decision procedures should be used if it is desired to have error probabilities such that the chance of misclassifying as A is to be 5% and the chance of misclassifying as B is to be 10%?

(c) If the sample is classified as A if the sample mean of log-lengths exceeds 0.75, and the misclassification as A is to have a probability of 2%, what sample size should be used and what is the probability of a B-type misclassification?

(d) If the sample comes from neither A nor B but from an environment with a mean log-length of 0.70, what is the probability of classifying it as type A if the decision procedure determined in (b) is applied?

2. In a wire-based nail manufacturing process the target length for cut wire is 22 cm. It is known that widths vary with a standard deviation equal to 0.08 cm. In order to monitor this process, a random sample of 50 separate wires is accurately measured and the process is regarded as operating satisfactorily (the null hypothesis) if the sample mean width lies between 21.97 cm and 22.03 cm so that this is the decision procedure used (i.e. if the sample mean falls within this range then the null hypothesis is not rejected, otherwise the null hypothesis is rejected).

(a) Determine the probability of a Type I error for this test.

(b) Determine the probability of making a Type II error when the process is actually cutting to a length of 22.05 cm.

(c) Find the probability of rejecting the null hypothesis when the true cutting length is 22.01 cm. (This is the power of the test when the true mean is 22.01 cm.)

3. A sample of seven is taken at random from a large batch of (nominally 12-volt) batteries. These are tested and their true voltages are shown below:

12.9, 11.6, 13.5, 13.9, 12.1, 11.9 and 13.0.

(a) Test if the mean voltage of the whole batch is 12 volts.

(b) Test if the mean batch voltage is less than 12 volts.

Which test do you think is the more appropriate?

4. To instil customer loyalty, airlines, hotels, rental car companies, and credit card companies (among others) have initiated frequency marketing programmes which reward their regular customers. In the United States alone, millions of people are members of the frequent-flier programmes of the airline industry. A large fast food restaurant chain wished to explore the profitability of such a programme. They randomly selected 12 of their 1,200 restaurants nationwide and instituted a frequency programme which rewarded customers with a $5.00 gift certificate after every 10 meals purchased at full price.

They ran the trial programme for three months. The restaurants not in the sample had an average increase in profits of $1,047.34 over the previous three months, whereas the restaurants in the sample had the following changes in profit:

$2,232.90	$545.47	$3,440.70	$1,809.10
$6,552.70	$4,798.70	$2,965.00	$2,610.70
$3,381.30	$1,591.40	$2,376.20	−$2,191.00

Note that the last number is negative, representing a decrease in profits. Specify the appropriate null and alternative hypotheses for determining whether the mean profit change for restaurants with frequency programmes is significantly greater (in a statistical sense which you should make clear) than \$1,047.34.

5. Two companies supplying a television repair service are compared by their repair times (in days). Random samples of recent repair times for these companies gave the following statistics:

	Sample size	Sample mean	Sample variance
Company A	44	11.9	7.3
Company B	52	10.8	6.2

   (a) Is there evidence that the companies differ in their true mean repair times? Give an appropriate hypothesis test to support your conclusions.

   (b) What is the p-value of your test?

   (c) What difference would it have made if the sample sizes had each been smaller by 35 (i.e. sizes 9 and 17, respectively)?

6. A museum conducts a survey of its visitors in order to assess the popularity of a device which is used to provide information on the museum exhibits. The device will be withdrawn if less than 30% of all of the museum's visitors make use of it. Of a random sample of 80 visitors, 20 chose to use the device.

   (a) Carry out a hypothesis test at the 5% significance level to see if the device should be withdrawn or not and state your conclusions.

   (b) Determine the p-value of the test.

   (c) What is the power of this test if the actual percentage of all visitors who would use this device is only 20%?

To p, or not to p?
(James Abdey, Ph.D. Thesis 2009.[1])

[1] Available at `http://etheses.lse.ac.uk/31`

Appendix E
Analysis of variance (ANOVA)

E.1 Worked examples

1. Three trainee salespeople were working on a trial basis. Salesperson A went in the field for 5 days and made a total of 440 sales. Salesperson B was tried for 7 days and made a total of 630 sales. Salesperson C was tried for 10 days and made a total of 690 sales. Note that these figures are total sales, not daily averages. The sum of the squares of all 22 daily sales ($\sum x_i^2$) is 146,840.

   (a) Construct a one-way analysis of variance table.

   (b) Would you say there is a difference between the mean daily sales of the three salespeople? Justify your answer.

   (c) Construct a 95% confidence interval for the mean difference between salesperson B and salesperson C. Would you say there is a difference?

   Solution:

   (a) The means are $440/5 = 88$, $630/7 = 90$ and $690/10 = 69$. We will perform a one-way ANOVA. First, we calculate the overall mean. This is:

   $$\frac{440 + 630 + 690}{22} = 80.$$

   We can now calculate the sum of squares between salespeople. This is:

   $$5 \times (88 - 80)^2 + 7 \times (90 - 80)^2 + 10 \times (69 - 80)^2 = 2{,}230.$$

   The total sum of squares is:

   $$146{,}840 - 22 \times (80)^2 = 6{,}040.$$

   Here is the one-way ANOVA table:

Source	DF	SS	MS	F	p-value
Salesperson	2	2,230	1,115	5.56	≈ 0.01
Error	19	3,810	200.53		
Total	21	6,040			

   (b) As $5.56 > 3.52 = F_{0.05,\,2,\,19}$, which is the top 5th percentile of the $F_{2,\,19}$ distribution (interpolated from Table 9 of Murdoch and Barnes' *Statistical Tables*), we reject $H_0 : \mu_1 = \mu_2 = \mu_3$ and conclude that there is evidence that the means are not equal.

(c) We have:

$$90 - 69 \pm 2.093 \times \sqrt{200.53 \times \left(\frac{1}{7} + \frac{1}{10}\right)} = 21 \pm 14.61.$$

Here 2.093 is the top 2.5th percentile point of the t distribution with 19 degrees of freedom. A 95% confidence interval is $(6.39, 35.61)$. As zero is not included, there is evidence of a difference.

2. The total times spent by three basketball players on court were recorded. Player A was recorded on three occasions and the times were 29, 25 and 33 minutes. Player B was recorded twice and the times were 16 and 30 minutes. Player C was recorded on three occasions and the times were 12, 14 and 16 minutes. Use analysis of variance to test whether there is any difference in the average times the three players spend on court.

Solution:

We have $\bar{x}_{.A} = 29$, $\bar{x}_{.B} = 23$, $\bar{x}_{.C} = 14$ and $\bar{x} = 21.875$. Hence:

$$3 \times (29 - 21.875)^2 + 2 \times (23 - 21.875)^2 + 3 \times (14 - 21.875)^2 = 340.875.$$

The total sum of squares is:

$$4{,}307 - 8 \times (21.875)^2 = 478.875.$$

Here is the one-way ANOVA table:

Source	DF	SS	MS	F	p-value
Players	2	340.875	170.4375	6.175	≈ 0.045
Error	5	138	27.6		
Total	7	478.875			

We test $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$ (i.e. the average times they play are the same) vs. H_1: The average times they play are not the same.

As $6.175 > 5.79 = F_{0.05,\,2,\,5}$, which is the top 5th percentile of the $F_{2,\,5}$ distribution, we reject H_0 and conclude that there is evidence of a difference between the means.

3. Three independent random samples were taken. Sample A consists of 4 observations taken from a normal distribution with mean μ_A and variance σ^2, sample B consists of 6 observations taken from a normal distribution with mean μ_B and variance σ^2, and sample C consists of 5 observations taken from a normal distribution with mean μ_C and variance σ^2.

The average value of the first sample was 24, the average value of the second sample was 20, and the average value of the third sample was 18. The sum of the squared observations (all of them) was 6,722.4. Test the hypothesis:

$$\mathrm{H}_0 : \mu_A = \mu_B = \mu_C$$

against the alternative that this is not so.

Solution:

We will perform a one-way ANOVA. First we calculate the overall mean:

$$\frac{4 \times 24 + 6 \times 20 + 5 \times 18}{15} = 20.4.$$

We can now calculate the sum of squares between groups:

$$4 \times (24 - 20.4)^2 + 6 \times (20 - 20.4)^2 + 5 \times (18 - 20.4)^2 = 81.6.$$

The total sum of squares is:

$$6{,}722.4 - 15 \times (20.4)^2 = 480.$$

Here is the one-way ANOVA table:

Source	DF	SS	MS	F	p-value
Sample	2	81.6	40.8	1.229	≈ 0.327
Error	12	398.4	33.2		
Total	14	480			

As $1.229 < 3.89 = F_{0.05,\,2,\,12}$, which is the top 5th percentile of the $F_{2,\,12}$ distribution, we see that there is no evidence that the means are not equal.

4. Four suppliers were asked to quote prices for seven different building materials. The average quote of supplier A was 1,315.8. The average quote of suppliers B, C and D were 1,238.4, 1,225.8 and 1,200.0, respectively. The following is the calculated two-way ANOVA table with some entries missing.

Source	DF	SS	MS	F	p-value
Materials			17,800		
Suppliers					
Error					
Total		358,700			

(a) Complete the table using the information provided above.

(b) Is there a significant difference between the quotes of different suppliers? Explain your answer.

(c) Construct a 90% confidence interval for the difference between suppliers A and D. Would you say there is a difference?

Solution:

(a) The average quote of all suppliers is:

$$\frac{1{,}315.8 + 1{,}238.4 + 1{,}225.8 + 1{,}200.0}{4} = 1{,}245.$$

Hence the sum of squares (SS) due to suppliers is:

$$7\times((1{,}315.8-1{,}245)^2+(1{,}238.4-1{,}245)^2+(1{,}225.8-1{,}245)^2+(1{,}200.0-1{,}245)^2] = 52{,}148.88$$

and the MS due to suppliers is $52{,}148.88/(4-1) = 17{,}382.96$.

The degrees of freedom are $7-1=6$, $4-1=3$, $(7-1)(4-1)=18$ and $7\times 4-1=27$ for materials, suppliers, error and total sum of squares, respectively.

The SS for materials is $6\times 17{,}800 = 106{,}800$. We have that the SS due to the error is given by $358{,}700 - 52{,}148.88 - 106{,}800 = 199{,}751.12$ and the MS is $199{,}751.12/18 = 11{,}097.28$. The F values are:

$$\frac{17{,}800}{11{,}097.28} = 1.604 \quad \text{and} \quad \frac{17{,}382.96}{11{,}097.28} = 1.567$$

for materials and suppliers, respectively. The two-way ANOVA table is:

Source	DF	SS	MS	F	p-value
Materials	6	106,800	17,800	1.604	≈ 0.203
Suppliers	3	52,148.88	17,382.96	1.567	≈ 0.232
Error	18	199,751.12	11,097.28		
Total	27	358,700			

(b) We test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (i.e. there is no difference between suppliers) vs. H_1: There is a difference between suppliers. The F value is 1.567 and at a 5% significance level the critical value from Table 9 (degrees of freedom 3 and 18) is 3.16, hence we do not reject H_0 and conclude that there is not enough evidence that there is a difference.

(c) The top 5th percentile of the t distribution with 18 degrees of freedom is 1.734 and the MS value is 11,097.28. So a 90% confidence interval is:

$$1{,}315.8 - 1{,}200 \pm 1.734 \times \sqrt{11{,}097.28\left(\frac{1}{7}+\frac{1}{7}\right)} = 115.8 \pm 97.64$$

giving $(18.16, 213.44)$. Since zero is not in the interval, there appears to be a difference between suppliers A and D.

5. Blood alcohol content (BAC) is measured in milligrams per decilitre of blood (mg/dL). A researcher is looking into the effects of alcoholic drinks. Four different individuals tried five different brands of strong beer (A, B, C, D and E) on different days, of course! Each individual consumed 1L of beer over a 30-minute period and their BAC was measured one hour later. The average BAC for beers A, C, D and E were 83.25, 95.75, 79.25 and 99.25, respectively. The value for beer B is not given. The following information is provided as well.

Source	DF	SS	MS	F	p-value
Drinker				1.56	
Beer			303.5		
Error		695.6			
Total					

(a) Complete the table using the information provided above.

(b) Is there a significant difference between the effects of different beers? What about different drinkers?

(c) Construct a 90% confidence interval for the difference between the effects of beers C and D. Would you say there is a difference?

Solution:

(a) We have:

Source	DF	SS	MS	F	p-value
Drinker	3	271.284	90.428	1.56	≈ 0.250
Beer	4	1214	303.5	5.236	≈ 0.011
Error	12	695.6	57.967		
Total	19	2,180.884			

(b) We test the hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_5$ (i.e. there is no difference between the effects of different beers) vs. the alternative H_1: There is a difference between the effects of different beers. The F value is 5.236 and at a 5% significance level the critical value from Table 9 is $F_{0.05,\,4,\,12} = 3.26$, so since $5.236 > 3.26$ we reject H_0 and conclude that there is evidence of a difference.

For drinkers, we test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (i.e. there is no difference between the effects on different drinkers) vs. the alternative H_1: There is a difference between the effects on different drinkers. The F value is 1.56 and at a 5% significance level the critical value from Table 9 is $F_{0.05,\,3,\,12} = 3.49$, so since $1.56 < 3.49$ we fail to reject H_0 and conclude that there is no evidence of a difference.

(c) The top 5th percentile of the t distribution with 12 degrees of freedom is 1.782. So a 90% confidence interval is:

$$95.75 - 79.25 \pm 1.782 \times \sqrt{57.967\left(\frac{1}{4}+\frac{1}{4}\right)} = 16.5 \pm 9.59$$

giving $(6.91, 26.09)$. As the interval does not contain zero, there is evidence of a difference between the effects of beers C and D.

6. A motor manufacturer operates five continuous-production plants: A, B, C, D and E. The average rate of production has been calculated for the three shifts of each plant and recorded in the table below. Does there appear to be a difference in production rates in different plants or by different shifts?

	A	B	C	D	E
Early shift	102	93	85	110	72
Late shift	85	87	71	92	73
Night shift	75	80	75	77	76

Solution:

Here $r = 3$ and $c = 5$. We may obtain the two-way ANOVA table as follows:

```
Source   DF         SS         MS        F
Shift     2     652.13     326.07     5.62
Plant     4     761.73     190.43     3.28
Error     8     463.87      57.98
Total    14   1,877.73
```

Under the null hypothesis of no shift effect, $F \sim F_{2,8}$. Since $F_{0.05,2,8} = 4.46 < 5.62$, we can reject the null hypothesis at the 5% significance level. (Note the p-value $= 0.030$.)

Under the null hypothesis of no plant effect, $F \sim F_{4,8}$. Since $F_{0.05,4,8} = 3.84 > 3.28$, we cannot reject the null hypothesis at the 5% significance level. (Note the p-value $= 0.072$.)

Overall, the data collected show some evidence of a shift effect but little evidence of a plant effect.

7. Complete the two-way ANOVA table below. In the places of p-values, indicate in the form such as '< 0.01' appropriately and use the closest value which you may find from Murdoch and Barnes' *Statistical Tables*.

Source	DF	SS	MS	F	p-value
Row factor	4	?	234.23	?	?
Column factor	6	270.84	45.14	1.53	?
Residual	?	708.00	?		
Total	34	1,915.76			

Solution:

First, C2 SS = (C2 MS)$\times$4 = 936.92.

The degrees of freedom for Error is $34 - 4 - 6 = 24$. Therefore, Error MS $= 708.00/24 = 29.5$.

Hence the F statistic for testing no C2 effect is $234.23/29.5 = 7.94$. From Table 9 of Murdoch and Barnes' *Statistical Tables*, $F_{0.001,4,24} = 6.59 < 7.94$. Therefore, the corresponding p-value is smaller than 0.001.

Since $F_{0.05,6,24} = 2.51 > 1.53$, the p-value for testing the C3 effect is greater than 0.05.

The complete ANOVA table is as follows:

```
Two-way ANOVA: C1 versus C2, C3

Source   DF         SS         MS        F        P
C2        4     936.92     234.23     7.94   <0.001
C3        6     270.84      45.14     1.53    >0.05
Error    24     708.00       29.5
Total    34   1,915.76
```

E.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix G.

1. An executive of a prepared frozen meals company is interested in the amounts of money spent on such products by families in different income ranges. The table below lists the monthly expenditures (in dollars) on prepared frozen meals from 15 randomly selected families divided into three groups according to their incomes.

Under \$15,000	\$15,000 – \$30,000	Over \$30,000
45.2	53.2	52.7
60.1	56.6	73.6
52.8	68.7	63.3
31.7	51.8	51.8
33.6	54.2	
39.4		

   (a) Based on these data, can we infer at the 5% significance level that the population mean expenditures on prepared frozen meals are the same for the three different income groups?

   (b) Produce a one-way ANOVA table.

   (c) Construct 95% confidence intervals for the mean expenditures of the first (under \$15,000) and the third (over \$30,000) income groups.

2. Does the level of success of publicly-traded companies affect the way their board members are paid? The annual payments (in \$000s) of randomly selected publicly-traded companies to their board members were recorded. The companies were divided into four quarters according to the returns in their stocks, and the payments from each quarter were grouped together. Some summary statistics are provided below.

```
Descriptive Statistics: 1st quarter, 2nd quarter, 3rd quarter, 4th quarter

Variable       N     Mean  SE Mean  StDev
1st quarter   30    74.10     2.89  15.81
2nd quarter   30    75.67     2.48  13.57
3rd quarter   30    78.50     2.79  15.28
4th quarter   30    81.30     2.85  15.59
```

   (a) Can we infer that the amount of payment differs significantly across the four groups of companies?

   (b) Construct 95% confidence intervals for the mean payment of the 1st quarter companies and the 4th quarter companies.

A total of 4,000 cans are opened around the world every second. Ten babies are conceived around the world every second. Each time you open a can, you stand a 1-in-400 chance of falling pregnant.
(True or false?)

Appendix F
Linear regression

F.1 Worked examples

1. Consider the simple linear regression model representing the linear relationship between two variables, y and x:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$, where ε_i are independent and identically distributed random variables with mean 0 and variance σ^2. Prove that the least squares straight line must necessarily pass through the point $(\bar{x}, \bar{y})$.

Solution:

The estimated regression line is:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

where $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$. When $\bar{x}$ is substituted for x_i, we obtain:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x} = \bar{y} - \widehat{\beta}_1 \bar{x} + \widehat{\beta}_1 \bar{x} = \bar{y}.$$

Therefore, the least squares straight line must necessarily pass through the point $(\bar{x}, \bar{y})$.

2. The following linear regression model is proposed to represent the linear relationship between two variables, y and x:

$$y_i = \beta x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$, where ε_i are independent and identically distributed random variables with mean 0 and variance σ^2, and β is an unknown parameter to be estimated.

(a) A proposed estimator of β is:

$$\widehat{\beta} = \min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

Explain why this estimator is sensible.

(b) Another proposed estimator of β is:

$$\tilde{\beta} = \min_{\beta} \sum_{i=1}^{n} |y_i - \beta x_i|.$$

Explain why $\widehat{\beta}$ would be preferred to $\tilde{\beta}$.

(c) Express $\widehat{\beta}$ explicitly as a function of y_i and x_i only.

(d) Using the estimator $\widehat{\beta}$:

i. what is the value of $\widehat{\beta}$ if $y_i = x_i$ for all i? What if they are the exact opposites of each other, i.e. $y_i = -x_i$ for all i?

ii. is it always the case that $-1 \leq \widehat{\beta} \leq 1$?

Solution:

(a) The estimator $\widehat{\beta}$ is sensible because it is the least squares estimator of β, which provides the 'best' fit to the data in terms of minimising the sum of squared residuals.

(b) The estimator $\widehat{\beta}$ is preferred to $\tilde{\beta}$ because the estimator $\tilde{\beta}$ is the least absolute deviations estimator of β, which is also an option, but unlike $\widehat{\beta}$ it cannot be computed explicitly via differentiation as the function $f(x) = |x|$ is not differentiable at zero. Therefore, $\tilde{\beta}$ is harder to compute than $\widehat{\beta}$.

(c) We need to minimise a convex quadratic, so we can do that by differentiating it and equating the derivative to zero. We obtain:

$$-2\sum_{i=1}^{n}(y_i - \widehat{\beta}x_i)x_i = 0$$

which yields:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(d) i. If $x_i = y_i$, then $\widehat{\beta} = 1$. If $x_i = -y_i$, then $\widehat{\beta} = -1$.

ii. Not true. A counterexample is to take $n = 1$, $x_1 = 1$ and $y_1 = 2$.

3. Let $\{(x_i, y_i)\}$, for $i = 1, 2, \ldots, n$, be observations from the linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

(a) Suppose that the slope, β_1, is *known*. Find the least squares estimator (LSE) of the intercept, β_0.

(b) Suppose that the intercept, β_0, is *known*. Find the LSE of the slope, β_1.

Solution:

(a) When β_1 is known, let $z_i = y_i - \beta_1 x_i$. The model then reduces to $z_i = \beta_0 + \varepsilon_i$. The LSE $\widehat{\beta}_0$ minimises $\sum_{i=1}^{n}(z_i - \beta_0)^2$, hence:

$$\widehat{\beta}_0 = \bar{z} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i).$$

(b) When β_0 is known, we may write $z_i = y_i - \beta_0$. The model is reduced to $z_i = \beta_1 x_i + \varepsilon_i$. Note that:

$$\begin{aligned}\sum_{i=1}^{n}(z_i - \beta_1 x_i)^2 &= \sum_{i=1}^{n}(z_i - \widehat{\beta}_1 x_i + (\widehat{\beta}_1 - \beta_1)x_i)^2 \\ &= \sum_{i=1}^{n}(z_i - \widehat{\beta}_1 x_i)^2 + (\widehat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} x_i^2 + 2D\end{aligned}$$

where $D = (\widehat{\beta}_1 - \beta_1)\sum_{i=1}^{n} x_i(z_i - \widehat{\beta}_1 x_i)$. Suppose we choose $\widehat{\beta}_1$ such that:

$$\sum_{i=1}^{n} x_i(z_i - \widehat{\beta}_1 x_i) = 0 \quad \text{i.e.} \quad \sum_{i=1}^{n} x_i z_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Hence:

$$\sum_{i=1}^{n}(z_i - \beta_1 x_i)^2 = \sum_{i=1}^{n}(z_i - \widehat{\beta}_1 x_i)^2 + (\widehat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n}(z_i - \widehat{\beta}_1 x_i)^2.$$

Therefore, $\widehat{\beta}_1$ is the LSE of β_1. Note now:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i(y_i - \beta_0)}{\sum_{i=1}^{n} x_i^2}.$$

4. Suppose an experimenter intends to perform a regression analysis by taking a total of $2n$ data points, where the x_is are restricted to the interval $[0, 5]$. If the xy-relationship is assumed to be linear and if the objective is to estimate the slope with the greatest possible precision, what values should be assigned to the x_is?

Solution:

Since:

$$\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

in order to minimise the variance of the sampling distribution of $\widehat{\beta}_1$, we must maximise:

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

To accomplish this, take half of the x_is to be 0, and the other half to be 5.

5. Suppose a total of $n = 9$ observations are to be taken on a simple linear regression model, where the x_is will be set equal to $1, 2, \ldots, 9$. If the variance associated with the xy-relationship is known to be 45, what is the probability that the estimated slope will be within 1.5 units of the true slope?

Solution:

Since $\bar{x} = (1 + 2 + \cdots + 9)/9 = 5$, then $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 60$ and so:

$$\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{45}{60} = 0.75.$$

Therefore:

$$\widehat{\beta}_1 \sim N(\beta_1, 0.75).$$

We require:

$$P(|\widehat{\beta}_1 - \beta_1| < 1.5) = P\left(|Z| < \frac{1.5}{\sqrt{0.75}}\right) = P(|Z| < 1.73) = 1 - 2 \times 0.0418 = 0.9164.$$

6. A researcher wants to investigate whether there is a significant link between GDP per capita and average life expectancy in major cities. Data have been collected in 30 major cities, yielding average GDPs per capita $x_1, x_2, \ldots, x_{30}$ (in \$000s) and average life expectancies $y_1, y_2, \ldots, y_{30}$ (in years). The following linear regression model has been proposed:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the ε_is are independent and $N(0, \sigma^2)$. Some summary statistics are:

$$\sum_{i=1}^{30} x_i = 620.35, \quad \sum_{i=1}^{30} y_i = 2{,}123.00, \quad \sum_{i=1}^{30} x_i y_i = 44{,}585.1$$

$$\sum_{i=1}^{30} x_i^2 = 13{,}495.62 \quad \text{and} \quad \sum_{i=1}^{30} y_i^2 = 151{,}577.3.$$

(a) Find the least-squares estimates of β_0 and β_1 and write down the fitted regression model.

(b) Compute a 95% confidence interval for the slope coefficient β_1. What can be concluded?

(c) Compute R^2. What can be said about how 'good' the model is?

(d) With $x = 30$, find a prediction interval which covers y with probability 0.95. With 97.5% confidence, what minimum average life expectancy can a city expect once its GDP per capita reaches \$30,000?

Solution:

(a) We have:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 1.026$$

and:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = 49.55.$$

Hence the fitted model is:

$$\widehat{y} = 49.55 + 1.026x.$$

(b) We first need E.S.E.$(\widehat{\beta}_1)$, for which we need $\widehat{\sigma}^2$. For $\widehat{\sigma}^2$, we need the Residual SS (from the Total SS and the Regression SS). We compute:

$$\text{Total SS} = \sum_i y_i^2 - n\bar{y}^2 = 1{,}339.67$$

$$\text{Regression SS} = \widehat{\beta}_1^2 \left(\sum_i x_i^2 - n\bar{x}^2 \right) = 702.99$$

$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = 636.68$$

$$\widehat{\sigma}^2 = \frac{636.68}{28} = 22.74$$

$$\text{E.S.E.}(\widehat{\beta}_1) = \left(\frac{\widehat{\sigma}^2}{\sum_i x_i^2 - n\bar{x}^2} \right)^{1/2} = 0.184.$$

Hence a 95% confidence interval for β_1 is:

$$(\widehat{\beta}_1 - t_{0.025,\,28} \times \text{E.S.E.}(\widehat{\beta}_1), \widehat{\beta}_1 + t_{0.025,\,28} \times \text{E.S.E.}(\widehat{\beta}_1))$$

which gives:

$$1.026 \pm 2.05 \times 0.184 \quad \Rightarrow \quad (0.65, 1.40).$$

The confidence interval does not contain zero. Therefore, we would reject the hypothesis of β_1 being zero at the 5% significance level. Hence there does appear to be a significant link.

(c) The model can explain 52% of the variation of y, since:

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{702.99}{1{,}339.67} = 0.52.$$

Whether or not the model is 'good' is subjective. It is not necessarily 'bad', although we may be able to determine a 'better' model with better explanatory power, possibly using multiple linear regression.

(d) The prediction interval has the form:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{0.025,\,n-2} \times \widehat{\sigma} \times \left(1 + \frac{\sum_i x_i^2 - 2x \sum_i x_i + nx^2}{n(\sum_i x_i^2 - n\bar{x}^2)} \right)^{1/2}$$

which gives:

$$(69.79, 90.87).$$

Therefore, we can be 97.5% confident that the average life expectancy lies above 69.79 years once GDP per capita reaches \$30,000.

7. The following is partial regression output:

```
The regression equation is
y = 2.1071 + 1.1263x

Predictor      Coef        SE Coef
Constant       2.1071      0.2321
x              1.1263      0.0911

Analysis of Variance

SOURCE           DF    SS
Regression       1     2011.12
Residual Error   40     539.17
```

In addition, $\bar{x} = 1.56$.

(a) Find an estimate of the error term variance, σ^2.

(b) Calculate and interpret R^2.

(c) Test at the 5% significance level whether or not the slope in the regression model is equal to 1.

(d) For $x = 0.8$, find a 95% confidence interval for the expectation of y.

Solution:

(a) Noting $n = 40 + 1 + 1 = 42$, we have:

$$\widehat{\sigma}^2 = \frac{\text{Residual SS}}{n-2} = \frac{539.17}{40} = 13.479.$$

(b) We have Total SS = Regression SS + Residual SS = 2,550.29. Hence:

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{2{,}011.12}{2{,}550.29} = 0.7886.$$

Therefore, 78.86% of the variation of y is explained by x.

(c) Under $\text{H}_0 : \beta_1 = 1$, the test statistic is:

$$T = \frac{\widehat{\beta}_1 - 1}{\text{E.S.E.}(\widehat{\beta}_1)} \sim t_{n-2} = t_{40}.$$

We reject H_0 if $|t| > 2.021 = t_{0.025,\,40}$. As $t = 0.1263/0.0911 = 1.386$, we cannot reject the null hypothesis that $\beta_1 = 1$ at the 5% significance level.

(d) Note $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\text{Regression SS})/(\widehat{\beta}_1)^2 = 2{,}011.12/(1.1263)^2 = 1{,}585.367$. Also:

$$\sum_{i=1}^{n}(x_i - x)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - x)^2 = 1{,}585.367 + 42 \times (1.56 - 0.8)^2$$
$$= 1{,}609.626.$$

Hence a 95% confidence interval for $\mathrm{E}(y)$ given $x = 0.8$ is:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{0.025,\,n-2} \times \widehat{\sigma} \times \left(\frac{\sum_{i=1}^{n} (x_i - x)^2}{n \sum_{j=1}^{n} (x_j - \bar{x})^2} \right)^{1/2}$$

$$= 2.1071 + 1.1263 \times 0.8 \pm 2.021 \times \sqrt{\frac{13.479 \times 1{,}609.626}{42 \times 1{,}585.367}}$$

$$= 3.0081 \pm 1.1536 \quad \Rightarrow \quad (1.854, 4.162).$$

8. Why is the squared sample correlation coefficient between the y_is and x_is the same as the squared sample correlation coefficient between the y_is and $\widehat{y}_i$s? No algebra is needed for this.

   Solution:

   The only difference between the x_is and $\widehat{y}_i$s is a rescaling by multiplying by $\widehat{\beta}_1$, followed by a relocation by adding $\widehat{\beta}_0$. Correlation coefficients are not affected by a change of scale or location, so it will be the same whether we use the x_is or the $\widehat{y}_i$s.

9. If the model fits, then the fitted values and the residuals from the model are independent of each other. What do you expect to see if the model fits when you plot residuals against fitted values?

   Solution:

   If the model fits, one would expect to see a random scatter with no particular pattern.

F.2 Practice questions

Try to solve the questions before looking at the solutions – promise?! Solutions are located in Appendix G.

1. The table below shows the cost of fire damage for ten fires together with the corresponding distances of the fires to the nearest fire station:

Distance in miles (x)	4.9	4.5	6.3	3.2	5.0
Cost in £000s (y)	31.1	31.1	43.1	22.1	36.2
Distance in miles (x)	5.7	4.0	4.3	2.5	5.2
Cost in £000s (y)	35.8	25.9	28.0	22.9	33.5

   (a) Fit a straight line to these data and construct a 95% confidence interval for the increase in cost of a fire for each mile from the nearest fire station.

   (b) Test the hypothesis that the 'true line' passes through the origin.

2. The yearly profits made by a company, over a period of eight consecutive years are shown below:

Year	1	2	3	4	5	6	7	8
Profit (in £000s)	18	21	34	31	44	46	60	75

   (a) Fit a straight line to these data and compute a 95% confidence interval for the 'true' yearly increase in profits.

   (b) The company accountant forecasts the profits for year 9 to be £90,000. Is this forecast reasonable if it is based on the above data?

3. The data table below shows the yearly expenditure (in £000s) by a cosmetics company in advertising a particular brand of perfume:

Year (x)	1	2	3	4	5	6	7	8
Expenditure (y)	170	170	275	340	435	510	740	832

   (a) Fit a regression line to these data and construct a 95% confidence interval for its slope.

   (b) Construct an analysis of variance table and compute the R^2 statistic for the fit.

   (c) Comment on the goodness of fit of the linear regression model.

   (d) Predict the expenditure for Year 9 and construct a 95% prediction interval for the actual expenditure.

4. Let X and ε be two independent random variables, and $\mathrm{E}(\varepsilon) = 0$. Let $Y = \beta_0 + \beta_1 X + \varepsilon$. Show that:

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \mathrm{Corr}(X, Y) \times \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}.$$

Facts are stubborn, but statistics are more pliable.
(Mark Twain)

Appendix G
Solutions to Practice questions

G.1 Chapter 6 – Sampling distributions of statistics

1. (a) The sum of n independent Bernoulli random variables, each with success probability π, is $\text{Bin}(n, \pi)$. Here $n = 4$ and $\pi = 0.2$, so $\sum_{i=1}^{4} X_i \sim \text{Bin}(4, 0.2)$.

   (b) The possible values of $\sum X_i$ are 0, 1, 2, 3 and 4, and their probabilities can be calculated from the binomial distribution. For example:

   $$P\left(\sum_{i=1}^{4} X_i = 1\right) = \binom{4}{1}(0.2)^1(0.8)^3 = 4 \times 0.2 \times 0.512 = 0.4096.$$

   The other probabilities are shown in the table below.

   Since $\bar{X} = \sum X_i/4$, the possible values of $\bar{X}$ are 0, 0.25, 0.5, 0.75 and 1. Their probabilities are the same as those of the corresponding values of $\sum X_i$. For example, $P(\bar{X} = 0.25) = P(\sum X_i = 1) = 0.4096$. The values and their probabilities are:

$\bar{X} = \bar{x}$	0.0	0.25	0.50	0.75	1.0
$P(\bar{X} = \bar{x})$	0.4096	0.4096	0.1536	0.0256	0.0016

   (c) For $X_i \sim \text{Bernoulli}(\pi)$, $\text{E}(X_i) = \pi$ and $\text{Var}(X_i) = \pi(1 - \pi)$. Therefore, the approximate normal sampling distribution of $\bar{X}$, derived from the central limit theorem, is $N(\pi, \pi(1 - \pi)/n)$. Here this is:

   $$N\left(0.2, \frac{0.2 \times 0.8}{100}\right) = N(0.2, 0.0016) = N(0.2, (0.04)^2).$$

   Therefore, the probability requested by the question is approximately:

   $$P(\bar{X} > 0.3) = P\left(\frac{\bar{X} - 0.2}{0.04} > \frac{0.3 - 0.2}{0.04}\right) = P(Z > 2.5) = 0.0062$$

   using Table 3 of Murdoch and Barnes' *Statistical Tables*. This is very close to the probability obtained from the exact sampling distribution, which is about 0.0061.

2. (a) Let $\{X_1, X_2, \ldots, X_n\}$ denote the random sample. We know that the sampling distribution of $\bar{X}$ is $N(\mu, \sigma^2/n)$, here $N(4, 2^2/20) = N(4, 0.2)$.

i. The probability we need is:

$$P(\bar{X} > 5) = P\left(\frac{\bar{X} - 4}{\sqrt{0.2}} > \frac{5 - 4}{\sqrt{0.2}}\right) = P(Z > 2.24) = 0.0126$$

where, as usual, $Z \sim N(0, 1)$.

ii. $P(\bar{X} < 3)$ is obtained similarly. Note that this leads to $P(Z < -2.24) = 0.0126$, which is equal to the $P(\bar{X} > 5) = P(Z > 2.24)$ result obtained above. This is because 5 is one unit above the mean $\mu = 4$, and 3 is one unit below the mean, and because the normal distribution is symmetric around its mean.

iii. One way of expressing this is:

$$P(\bar{X} - \mu > 1) = P(\bar{X} - \mu < -1) = 0.0126$$

for $\mu = 4$. This also shows that:

$$P(\bar{X} - \mu > 1) + P(\bar{X} - \mu < -1) = P(|\bar{X} - \mu| > 1) = 2 \times 0.0126 = 0.0252$$

and hence:

$$P(|\bar{X} - \mu| \leq 1) = 1 - 2 \times 0.0126 = 0.9748.$$

In other words, the probability is 0.9748 that the sample mean is within one unit of the true population mean, $\mu = 4$.

(b) We can use the same ideas as in (a). Since $\bar{X} \sim N(\mu, 4/n)$ we have:

$$\begin{aligned} P(|\bar{X} - \mu| \leq 0.5) &= 1 - 2 \times P(\bar{X} - \mu > 0.5) \\ &= 1 - 2 \times P\left(\frac{\bar{X} - \mu}{\sqrt{4/n}} > \frac{0.5}{\sqrt{4/n}}\right) \\ &= 1 - 2 \times P(Z > 0.25\sqrt{n}) \\ &\geq 0.95 \end{aligned}$$

which holds if:

$$P(Z > 0.25\sqrt{n}) \leq \frac{0.05}{2} = 0.025.$$

From Table 3 of Murdoch and Barnes' *Statistical Tables*, we see that this is true when $0.25\sqrt{n} \geq 1.96$, i.e. when $n \geq (1.96/0.25)^2 = 61.5$. Rounding up to the nearest integer, we get $n \geq 62$. The sample size should be at least 62 for us to be 95% confident that the sample mean will be within 0.5 units of the true mean, μ.

(c) Here $n > 62$, yet $\bar{x}$ is further than 0.5 units from the claimed mean of $\mu = 5$. Based on the result in (b), this would be quite unlikely *if* μ is really 5. One explanation of this apparent contradiction is that μ is *not* really equal to 5. This kind of reasoning will be the basis of statistical hypothesis testing, which will be discussed later in the course.

3. (a) The sample average is composed of 25 randomly sampled data which are subject to sampling variability, hence the average is also subject to this variability. Its sampling distribution describes its probability properties. If a large number of such averages were independently sampled, then their histogram would be the sampling distribution.

   (b) It is reasonable to assume that this sampling distribution is normal due to the CLT, although the sample size is rather small. If $n = 25$ and $\mu = 54$ and $\sigma = 10$, then the CLT says that:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(54, \frac{100}{25}\right).$$

   (c) i. We have:

$$P(\bar{X} > 60) = P\left(Z > \frac{60 - 54}{\sqrt{100/25}}\right) = P(Z > 3) = 0.0013$$

   using Table 3 of Murdoch and Barnes' *Statistical Tables*.

   ii. We are asked for:

$$\begin{aligned} P(0.95 \times 54 < \bar{X} < 1.05 \times 54) &= P\left(\frac{-0.05 \times 54}{2} < Z < \frac{0.05 \times 54}{2}\right) \\ &= P(-1.35 < Z < 1.35) \\ &= 0.8230 \end{aligned}$$

   using Table 3 of Murdoch and Barnes' *Statistical Tables*.

G.2 Chapter 7 – Point estimation

1. We have:

$$\mathrm{E}(X) = \mathrm{E}\left(\frac{X_1}{2} + \frac{X_2}{2}\right) = \frac{1}{2} \times \mathrm{E}(X_1) + \frac{1}{2} \times \mathrm{E}(X_2) = \frac{1}{2} \times \mu + \frac{1}{2} \times \mu = \mu$$

   and:

$$\mathrm{E}(Y) = \mathrm{E}\left(\frac{X_1}{3} + \frac{2X_2}{3}\right) = \frac{1}{3} \times \mathrm{E}(X_1) + \frac{2}{3} \times \mathrm{E}(X_2) = \frac{1}{3} \times \mu + \frac{2}{3} \times \mu = \mu.$$

   It follows that both estimators are unbiased estimators of μ.

2. The formula for S^2 is:

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

This means that $(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$, hence:

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

Because the sample is random, $\mathrm{E}(X_i^2) = \mathrm{E}(X^2)$ for all $i = 1, 2, \ldots, n$ as all the variables are identically distributed. From the standard formula $\mathrm{Var}(X) = \sigma^2 = \mathrm{E}(X^2) - \mu^2$, so (using the hint):

$$\mathrm{E}(X^2) = \sigma^2 + \mu^2 \quad \text{and} \quad \mathrm{E}(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}.$$

Hence:

$$(n-1)\,\mathrm{E}(S^2) = n(\sigma^2 + \mu^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right) = (n-1)\sigma^2$$

so $\mathrm{E}(S^2) = \sigma^2$, which means that S^2 is an unbiased estimator of σ^2, as stated.

The standard formula for $\mathrm{Var}(X)$, applied to S, states that:

$$\mathrm{E}(S^2) = \mathrm{Var}(S) + (\mathrm{E}(S))^2$$

which means that:

$$\mathrm{E}(S) = \sqrt{\mathrm{E}(S^2) - \mathrm{Var}(S)} = \sqrt{\sigma^2 - \mathrm{Var}(S)} < \sigma = \sqrt{\sigma^2}$$

since all variances are strictly positive. It follows that S is a biased estimator of σ (with its average value lower than the true value σ).

3. As defined, R is a random variable, and $R \sim \mathrm{Bin}(n, \pi)$, so that $\mathrm{E}(R) = n\pi$ and hence $\mathrm{E}(R/n) = \pi$. It also follows that:

$$1 - \mathrm{E}\left(\frac{R}{n}\right) = \mathrm{E}\left(1 - \frac{R}{n}\right) = \mathrm{E}\left(\frac{n-R}{n}\right) = 1 - \pi.$$

So the first obvious guess is that we should try $R/n \times (1 - R/n) = R/n - (R/n)^2$. Now:

$$n\pi(1-\pi) = \mathrm{Var}(R) = \mathrm{E}(R^2) - (\mathrm{E}(R))^2 = \mathrm{E}(R^2) - (n\pi)^2.$$

So:

$$\mathrm{E}\left(\left(\frac{R}{n}\right)^2\right) = \frac{1}{n^2}\,\mathrm{E}(R^2) = \frac{1}{n^2}(n\pi(1-\pi) + n^2\pi^2)$$

$$\Rightarrow \quad \mathrm{E}\left(\frac{R}{n} - \left(\frac{R}{n}\right)^2\right) = \frac{1}{n}\,\mathrm{E}(R) - \frac{1}{n^2}(n\pi(1-\pi) + n^2\pi^2)$$

$$= \frac{n\pi}{n} - \frac{n^2\pi^2}{n^2} - \frac{\pi(1-\pi)}{n}$$

$$= \pi - \pi^2 - \frac{\pi(1-\pi)}{n}.$$

However, $\pi(1-\pi) = \pi - \pi^2$, so:

$$\mathrm{E}\left(\frac{R}{n} - \left(\frac{R}{n}\right)^2\right) = \pi(1-\pi) - \frac{\pi(1-\pi)}{n} = \pi(1-\pi) \times \frac{n-1}{n}.$$

It follows that:

$$\pi(1-\pi) = \frac{n}{n-1} \times \mathrm{E}\left(\frac{R}{n} - \left(\frac{R}{n}\right)^2\right) = \mathrm{E}\left(\frac{R}{n-1} - \frac{R^2}{n(n-1)}\right).$$

So we have found the unbiased estimator of $\pi(1-\pi)$, but it could do with tidying up! When this is done, we see that:

$$\frac{R(n-R)}{n(n-1)}$$

is the required unbiased estimator of $\pi(1-\pi)$.

4. For T_1:

$$\mathrm{E}(T_1) = \mathrm{E}\left(\frac{S_{xx}}{n-1}\right) = \frac{1}{n-1}\mathrm{E}(S_{xx}) = \frac{1}{n-1} \times (n-1)\sigma^2 = \sigma^2.$$

Hence T_1 is an unbiased estimator of σ^2. Turning to the variance:

$$\mathrm{Var}(T_1) = \mathrm{Var}\left(\frac{S_{xx}}{n-1}\right) = \left(\frac{1}{n-1}\right)^2 \times \mathrm{Var}(S_{xx}) = \left(\frac{1}{n-1}\right)^2 \times (2\sigma^4(n-1)) = \frac{2\sigma^4}{n-1}.$$

By definition, $\mathrm{MSE}(T_1) = \mathrm{Var}(T_1) + (\mathrm{Bias}(T_1))^2 = 2\sigma^4/(n-1) + 0^2 = 2\sigma^4/(n-1)$.

For T_2:

$$\mathrm{E}(T_2) = \mathrm{E}\left(\frac{S_{xx}}{n}\right) = \frac{1}{n}\mathrm{E}(S_{xx}) = \frac{1}{n} \times (n-1)\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2.$$

It follows that $\mathrm{Bias}(T_2) = -\sigma^2/n$, hence T_2 is a biased estimator of σ^2.

$$\mathrm{Var}(T_2) = \mathrm{Var}\left(\frac{S_{xx}}{n}\right) = \left(\frac{1}{n}\right)^2 \times \mathrm{Var}(S_{xx}) = \left(\frac{1}{n}\right)^2 \times (2\sigma^4(n-1)) = \frac{2(n-1)\sigma^4}{n^2}.$$

By definition, $\mathrm{MSE}(T_2) = 2(n-1)\sigma^4/n^2 + (-\sigma^2/n)^2 = (2n-1)\sigma^4/n^2$.

It can be seen that $\mathrm{MSE}(T_1) > \mathrm{MSE}(T_2)$ since:

$$\frac{2}{n-1} - \frac{2n-1}{n^2} = \frac{2n^2 - (n-1)(2n-1)}{n^2(n-1)} = \frac{2n^2 - (2n^2 - 3n + 1)}{n^2(n-1)} = \frac{3n-1}{n^2(n-1)} > 0.$$

So, although T_2 is a biased estimator of σ^2, it is preferable to T_1 due to the dominating effect of its smaller variance.

5. (a) We start off with the sum of squares function:

$$S = \sum_{i=1}^{4} \varepsilon_i^2 = (y_1 - \alpha - \beta)^2 + (y_2 + \alpha - \beta)^2 + (y_3 - \alpha + \beta)^2 + (y_4 + \alpha + \beta)^2.$$

Now take the partial derivatives:

$$\begin{aligned}\frac{\partial S}{\partial \alpha} &= -2(y_1 - \alpha - \beta) + 2(y_2 + \alpha - \beta) - 2(y_3 - \alpha + \beta) + 2(y_4 + \alpha + \beta) \\ &= -2(y_1 - y_2 + y_3 - y_4) + 8\alpha\end{aligned}$$

and:

$$\begin{aligned}\frac{\partial S}{\partial \beta} &= -2(y_1 - \alpha - \beta) - 2(y_2 + \alpha - \beta) + 2(y_3 - \alpha + \beta) + 2(y_4 + \alpha + \beta) \\ &= -2(y_1 + y_2 - y_3 - y_4) + 8\beta.\end{aligned}$$

The least squares estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are the solutions to $\partial S/\partial\alpha = 0$ and $\partial S/\partial\beta = 0$. Hence:

$$\widehat{\alpha} = \frac{y_1 - y_2 + y_3 - y_4}{4} \quad \text{and} \quad \widehat{\beta} = \frac{y_1 + y_2 - y_3 - y_4}{4}.$$

(b) $\widehat{\alpha}$ is an unbiased estimator of α since:

$$\mathrm{E}(\widehat{\alpha}) = \mathrm{E}\left(\frac{y_1 - y_2 + y_3 - y_4}{4}\right) = \frac{\alpha + \beta + \alpha - \beta + \alpha - \beta + \alpha + \beta}{4} = \alpha.$$

$\widehat{\beta}$ is an unbiased estimator of β since:

$$\mathrm{E}(\widehat{\beta}) = \mathrm{E}\left(\frac{y_1 + y_2 - y_3 - y_4}{4}\right) = \frac{\alpha + \beta - \alpha + \beta - \alpha + \beta + \alpha + \beta}{4} = \beta.$$

(c) We have:

$$\mathrm{Var}(\widehat{\alpha}) = \mathrm{Var}\left(\frac{y_1 - y_2 + y_3 - y_4}{4}\right) = \frac{4\sigma^2}{16} = \frac{\sigma^2}{4}.$$

G.3 Chapter 8 – Interval estimation

1. (a) The *total* value of the stock is $9{,}875\mu$, where μ is the mean value of an item of stock. From Chapter 6, $\bar{X}$ is the obvious estimator of μ, so $9{,}875\bar{X}$ is the obvious estimator of $9{,}875\mu$. Therefore, an estimate for the total value of the stock is $9{,}875 \times 320.41 = £3{,}160{,}000$ (to the nearest £10,000).

(b) In this question $n = 50$ is large, and σ^2 is unknown so a 95% confidence interval for μ is:

$$\bar{x} \pm 1.96 \times \frac{s}{\sqrt{n}} = 320.41 \pm 1.96 \times \frac{40.6}{\sqrt{50}} = 320.41 \pm 11.25 \quad \Rightarrow \quad (£309.16, £331.66).$$

Note that because n is large we have used the standard normal distribution. It is more accurate to use a t distribution with 49 degrees of freedom. This gives an interval of (£308.87, £331.95) – not much of a difference.

To obtain a 95% confidence interval for the total value of the stock, $9{,}875\mu$, multiply the interval by 9,875. This gives (to the nearest £10,000):

$$(£3{,}050{,}000,\ £3{,}280{,}000).$$

(c) Increasing the sample size by a factor of k reduces the width of the confidence interval by a factor of $\sqrt{k}$. Therefore, increasing the sample size by a factor of 4 will reduce the width of the confidence interval by a factor of 2 ($= \sqrt{4}$). Hence we need to increase the sample size from 50 to $4 \times 50 = 200$. So we should collect another 150 observations.

2. (a) Let π be the proportion of shareholders in the population. Start by estimating π. We are estimating a proportion and n is large, so an approximate 95% confidence interval for π is, using the central limit theorem:

$$\widehat{\pi} \pm 1.96 \times \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}} \Rightarrow 0.23 \pm 1.96 \times \sqrt{\frac{0.23 \times 0.77}{954}} = 0.23 \pm 0.027 \ \Rightarrow \ (0.203, 0.257).$$

Therefore, a 95% confidence interval for the *number* (rather than the proportion) of shareholders in the UK is obtained by multiplying the above interval endpoints by 41 million and getting the answer 8.3 million to 10.5 million. An alternative way of expressing this is:

$$9{,}400{,}000 \pm 1{,}100{,}000 \quad \Rightarrow \quad (8{,}300{,}000, 10{,}500{,}000).$$

Therefore, we estimate there are about 9.4 million shareholders in the UK, with a margin of error of 1.1 million.

(b) Let us start by finding a 95% confidence interval for the difference in the two proportions. We use the formula:

$$\widehat{\pi}_1 - \widehat{\pi}_2 \pm 1.96 \times \sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}}.$$

The estimates of the proportions π_1 and π_2 are 0.23 and 0.171, respectively. We know $n_1 = 954$ and although n_2 is unknown we can assume it is approximately equal to 954 (noting the 'similar' in the question), so an approximate 95% confidence interval is:

$$0.23 - 0.171 \pm 1.96 \times \sqrt{\frac{0.23 \times 0.77}{954} + \frac{0.171 \times 0.829}{954}} = 0.059 \pm 0.036 \ \Rightarrow \ (0.023, 0.094).$$

By multiplying by 41 million, we get a confidence interval of:

$$2{,}400{,}000 \pm 1{,}500{,}000 \quad \Rightarrow \quad (900{,}000, 3{,}900{,}000).$$

We estimate that the number of shareholders has increased by about 2.4 million in the two years. There is quite a large margin of error, i.e. 1.5 million, especially when compared with a point estimate (i.e. interval midpoint) of 2.4 million.

G.4 Chapter 9 – Hypothesis testing

1. (a) We have $n = 50$ and $\sigma = 1$. We wish to test:

$$\mathrm{H}_0 : \mu = 0.65 \text{ (sample is from '}B\text{')} \quad \text{vs.} \quad \mathrm{H}_1 : \mu = 0.80 \text{ (sample is from '}A\text{')}.$$

The decision rule is that we reject H_0 if $\bar{x} > 0.75$.

The probability of a Type I error is:

$$P(\bar{X} > 0.75 \,|\, \mathrm{H}_0) = P\left(Z > \frac{0.75 - 0.65}{1/\sqrt{50}}\right) = P(Z > 0.71) = 0.2389.$$

The probability of a Type II error is:

$$P(\bar{X} < 0.75 \,|\, \mathrm{H}_1) = P\left(Z < \frac{0.75 - 0.80}{1/\sqrt{50}}\right) = P(Z < -0.35) = 0.3632.$$

(b) To find the sample size n and the value a, we need to solve two conditions:

- $\alpha = P(\bar{X} > a \,|\, \mathrm{H}_0) = P(Z > (a - 0.65)/(1/\sqrt{n})) = 0.05 \Rightarrow$ $(a - 0.65)/(1/\sqrt{n}) = 1.645$.
- $\beta = P(\bar{X} < a \,|\, \mathrm{H}_1) = P(Z < (a - 0.80)/(1/\sqrt{n})) = 0.10 \Rightarrow$ $(a - 0.80)/(1/\sqrt{n}) = -1.28$.

Solving these equations gives $a = 0.734$ and $n = 381$, remembering to round up!

(c) A sample is classified as being from A if H_1 if $\bar{x} > 0.75$. We have:

$$\alpha = P(\bar{X} > 0.75 \,|\, \mathrm{H}_0) = P\left(Z > \frac{0.75 - 0.65}{1/\sqrt{n}}\right) = 0.02 \;\Rightarrow\; \frac{0.75 - 0.65}{1/\sqrt{n}} = 2.05.$$

Solving this equation gives $n = 421$, remembering to round up! Therefore:

$$\beta = P(\bar{X} < 0.75 \,|\, \mathrm{H}_1) = P\left(Z < \frac{0.75 - 0.80}{1/\sqrt{421}}\right) = P(Z < -1.026) = 0.1515.$$

(d) The rule in (b) is 'take $n = 381$ and reject H_0 if $\bar{x} > 0.734$'. So:

$$P(\bar{X} > 0.734 \,|\, \mu = 0.7) = P\left(Z > \frac{0.734 - 0.7}{1/\sqrt{381}}\right) = P(Z > 0.66) = 0.2546.$$

2. (a) We have:

$$\begin{aligned}
\alpha &= 1 - P(21.97 < \bar{X} < 22.03 \,|\, \mu = 22) \\
&= 1 - P\left(\frac{21.97 - 22}{0.08/\sqrt{50}} < Z < \frac{22.03 - 22}{0.08/\sqrt{50}}\right) \\
&= 1 - P(-2.65 < Z < 2.65) \\
&= 1 - 0.992 \\
&= 0.008.
\end{aligned}$$

(b) We have:

$$\begin{aligned}\beta &= P(21.97 < \bar{X} < 22.03 \,|\, \mu = 22.05) \\ &= P\left(\frac{21.97 - 22.05}{0.08/\sqrt{50}} < Z < \frac{22.03 - 22.05}{0.08/\sqrt{50}}\right) \\ &= P(-7.07 < Z < -1.77) \\ &= P(Z < -1.77) - P(Z < -7.07) \\ &= 0.0384.\end{aligned}$$

(c) We have:

$$\begin{aligned}P(\text{rejecting } \mathrm{H}_0 \,|\, \mu = 22.01) &= 1 - P(21.97 < \bar{X} < 22.03 \,|\, \mu = 22.01) \\ &= 1 - P\left(\frac{21.97 - 22.01}{0.08/\sqrt{50}} < Z < \frac{22.03 - 22.01}{0.08/\sqrt{50}}\right) \\ &= 1 - P(-3.53 < X < 1.77) \\ &= 1 - (P(Z < 1.77) - P(Z < -3.53)) \\ &= 1 - (0.9616 - 0.00023) \\ &= 0.0386.\end{aligned}$$

3. (a) We are to test $\mathrm{H}_0 : \mu = 12$ vs. $\mathrm{H}_1 : \mu \neq 12$. The key points here are that n is small and that σ^2 is unknown. We can use the t test and this is valid provided the data are normally distributed. The test statistic value is:

$$t = \frac{\bar{x} - 12}{s/\sqrt{7}} = \frac{12.7 - 12}{0.858/\sqrt{7}} = 2.16.$$

This is compared to a Student's t distribution on 6 degrees of freedom. The critical value corresponding to a 5% significance level is 2.447. Hence we cannot reject the null hypothesis at the 5% significance level. (We can reject at the 10% significance level, but the convention on this course is to regard such evidence merely as casting doubt on H_0, rather than justifying rejection as such, i.e. such a result would be 'weakly significant'.)

(b) We are to test $\mathrm{H}_0 : \mu = 12$ vs. $\mathrm{H}_1 : \mu < 12$. There is no need to do a formal statistical test. As the sample mean is 12.7, which is greater than 12, there is no evidence whatsoever for the alternative hypothesis.

In (a) you are asked to do a two-sided test and in (b) it is a one-sided test. Which is more appropriate will depend on the purpose of the experiment, and your suspicions before you conduct it.

- If you suspected *before* collecting the data that the mean voltage was less than 12 volts, the one-sided test would be appropriate.
- If you had no prior reason to believe that the mean was less than 12 volts you would perform a two-sided test.

- General rule: decide on whether it is a one- or two-sided test *before* performing the statistical test!

4. It is useful to discuss the issues about this question before giving the solution.
   - We want to know whether a loyalty programme such as that at the 12 selected restaurants would result in an increase in mean profits greater than that observed (during the three-month test) at the other sites within the chain.
   - So we can model the profits across the chain as $\$1{,}047.34 + x$, where $\$x$ is the supposed effect of the promotion, and if the true mean value of x is μ, then we wish to test:
     $$\mathrm{H}_0 : \mu = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \mu > 0$$
     which is a one-tailed test since, clearly, there are (preliminary) grounds for thinking that there is an increase due to the loyalty programme.
   - We know nothing about the variability of profits across the rest of the chain, so we will have to use the sample data, i.e. to calculate the sample variance and to employ the t distribution with $\nu = 12 - 1 = 11$ degrees of freedom.
   - Although we shall want the variance of the data 'sample value − 1,047.34', this will be the same as the variance of the sample data, since for any random variable X and constant k we have:
     $$\mathrm{Var}(X + k) = \mathrm{Var}(X)$$
     because in calculating the variance every value $(x_i - \bar{x})$ is 'replaced' by $((x_i + k) - (\bar{x} + k))$, which is in fact the same value.
   - So we need to calculate $\bar{x}$, $\sum_{i=1}^{12} x_i^2$, $S_{xx} = \sum_{i=1}^{12} x_i^2 - n\bar{x}^2$ and s^2.

   The total change in profit for restaurants in the programme is $\sum_{i=1}^{12} x_i = 30{,}113.17$. Since $n = 12$, the mean change in profit for restaurants in the programme is:
   $$\frac{30{,}113.17}{12} = 2{,}509.431 = 1{,}047.34 + 1{,}462.091$$
   hence use $\bar{x} = 1{,}462.091$.

   The raw sum of squares is $\sum_{i=1}^{12} x_i^2 = 126{,}379{,}568.8$. So, the 'corrected' sum of squares is:
   $$S_{xx} = \sum_{i=1}^{12} x_i^2 - n\bar{x}^2 = 126{,}379{,}568.8 - 12 \times (2{,}509.431)^2 = 50{,}812{,}651.51.$$
   Therefore:
   $$s^2 = \frac{S_{xx}}{n-1} = \frac{50{,}812{,}651.51}{11} = 4{,}619{,}331.956.$$
   Hence the estimated standard error is:
   $$\frac{s}{\sqrt{n}} = \sqrt{\frac{4{,}619{,}331{,}956}{12}} = \sqrt{384{,}944.3296} = 620.439.$$

So, the test statistic value is:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1{,}462.091 - 0}{620.439} = 2.3565.$$

The relevant critical values for t_{11} in this one-tailed test are:

$$5\%: \quad 1.796 \quad \text{and} \quad 1\%: \quad 2.718.$$

So we see that the test is significant at the 5% significance level, but not at the 1% significance level, so reject H_0 and conclude that the loyalty programme does have an effect. (In fact, this means the result is moderately significant that the programme has had a beneficial effect for the company.)

5. (a) We test $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A \neq \mu_B$, where we use a two-tailed test since there is no prior reason to suggest the direction of the difference, if any. The test statistic value is:

$$\frac{11.9 - 10.8}{\sqrt{7.3/44 + 6.3/52}} = 2.06$$

where we assume the sample variances are equal to the population variances due to the large sample sizes (and hence we would expect accurate variance estimates). For a two-tailed test, this is significant at the 5% significance level ($1.96 < 2.06$), but not at the 1% significance level ($2.06 < 2.576$). We reject H_0 and conclude that company A is slower in repair times on average than company B, with a moderately significant result.

(b) The p-value for this two-tailed test is $2 \times P(Z > 2.06) = 0.0394$.

(c) For small samples, we should use a pooled estimate of the population standard deviation:

$$s = \sqrt{\frac{(9-1) \times 7.3 + (17-1) \times 6.2}{(9-1) + (17-1)}} = 2.5626 \quad \text{on 24 degrees of freedom.}$$

Hence the test statistic value in this case is:

$$\frac{11.9 - 10.8}{2.5626 \times \sqrt{1/9 + 1/17}} = 1.04.$$

This should be compared with the t_{24} distribution and is clearly not significant, even at the 10% significance level. With the smaller samples we fail to detect the difference.

Comparing the two test statistic calculations shows that the different results flow from differences in the estimated standard errors, hence ultimately (and unsurprisingly) from the differences in the sample sizes used in the two situations.

6. (a) Let π be the population proportion of visitors who would use the device. We test $H_0 : \pi = 0.3$ vs. $H_1 : \pi < 0.3$. The sample proportion is $p = 20/80 = 0.25$. Standard error of the sample proportion is $\sqrt{0.3 \times 0.7/80} = 0.0512$. The test statistic value is:

$$z = \frac{0.25 - 0.30}{0.0512} = -0.976.$$

For a one-sided (lower-tailed) test at the 5% significance level, the critical value is -1.645, so the test is not significant – and not even at the 10% significance level (the critical value is -1.282). On the basis of the data, there is no reason to withdraw the device.

The critical region for the above test is to reject H_0 if the sample proportion is less than $0.3 - 1.645 \times 0.0512$, i.e. if the sample proportion, p, is less than 0.2157.

(b) The p-value of the test is the probability of the test statistic value or a more extreme value conditional on H_0 being true. Hence the p-value is:

$$P(Z \leq -0.976) = 0.1645.$$

So for any $\alpha < 0.1645$ we would fail to reject H_0.

(c) The power of the test when $\pi = 0.2$ is the conditional probability:

$$P(P < 0.2157 \,|\, \pi = 0.2).$$

When $\pi = 0.2$, the standard error of the sample proportion is $\sqrt{0.2 \times 0.8/80} = 0.0447$. Therefore, the power when $\pi = 0.2$ is:

$$P\left(Z < \frac{0.2157 - 0.2}{0.0447}\right) = P(Z < 0.35) = 0.6368.$$

G.5 Chapter 10 – Analysis of variance

1. (a) For this example, $k = 3$, $n_1 = 6$, $n_2 = 5$, $n_3 = 4$ and $n = n_1 + n_2 + n_3 = 15$.

We have $\bar{x}_{\cdot 1} = 43.8$, $\bar{x}_{\cdot 2} = 56.9$, $\bar{x}_{\cdot 3} = 60.35$ and $\bar{x} = 52.58$.

Also, $\sum_{j=1}^{3} \sum_{i=1}^{n_j} x_{ij}^2 = 43{,}387.85$.

Total SS $= \sum_{j=1}^{3} \sum_{i=1}^{n_j} x_{ij}^2 - n\bar{x}^2 = 43{,}387.85 - 41{,}469.85 = 1{,}918$.

$w = \sum_{j=1}^{3} \sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{3} n_j \bar{x}_{\cdot j}^2 = 43{,}387.85 - 42{,}267.18 = 1{,}120.67$.

Therefore, $b =$ Total SS $- w = 1{,}918 - 1{,}120.67 = 797.33$.

To test $H_0 : \mu_1 = \mu_2 = \mu_3$, the test statistic value is:

$$f = \frac{b/(k-1)}{w/(n-k)} = \frac{797.33/2}{1{,}120.67/12} = 4.269.$$

Under H_0, $F \sim F_{2,12}$. Since $F_{0.05,2,12} = 3.89 < 4.269$, we reject H_0 at the 5% significance level, i.e. there exists evidence indicating that the population mean expenditures on frozen meals are not the same for the three different income groups.

(b) The ANOVA table is as follows:

Source	DF	SS	MS	F	P
Income	2	797.33	398.67	4.269	<0.05
Error	12	1,120.67	93.39		
Total	14	1,918.00			

(c) A 95% confidence interval for μ_j is of the form:

$$\bar{X}_{\cdot j} \pm t_{0.025,n-k} \times \frac{S}{\sqrt{n_j}} = \bar{X}_{\cdot j} \pm t_{0.025,12} \times \frac{\sqrt{93.39}}{\sqrt{n_j}} = \bar{X}_{\cdot j} \pm \frac{21.056}{\sqrt{n_j}}.$$

For $j = 1$, a 95% confidence interval is $43.8 \pm 21.056/\sqrt{6} \Rightarrow (35.20, 52.40)$.

For $j = 3$, a 95% confidence interval is $60.35 \pm 21.056/\sqrt{4} \Rightarrow (49.82, 70.88)$.

2. (a) Here $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 30$. We have $\bar{x}_{\cdot 1} = 74.10$, $\bar{x}_{\cdot 2} = 75.67$, $\bar{x}_{\cdot 3} = 78.50$, $\bar{x}_{\cdot 4} = 81.30$, $b = 909$, $w = 26{,}408$ and the pooled estimate of σ is $s = 15.09$.

Hence the test statistic value is:

$$f = \frac{b/(k-1)}{w/(n-k)} = 1.33.$$

Under $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, $F \sim F_{k-1,n-k} = F_{3,116}$. Since $F_{0.05,3,116} = 2.68 > 1.33$, we cannot reject H_0 at the 5% significance level. Hence there is no evidence to support the claim that payments among the four groups are significantly different.

(b) A 95% confidence interval for μ_j is of the form:

$$\bar{X}_{\cdot j} \pm t_{0.025,n-k} \times \frac{S}{\sqrt{n_j}} = \bar{X}_{\cdot j} \pm t_{0.025,116} \times \frac{15.09}{\sqrt{30}} = \bar{X}_{\cdot j} \pm 5.46.$$

For $j = 1$, a 95% confidence interval is $74.10 \pm 5.46 \Rightarrow (68.64, 79.56)$.

For $j = 4$, a 95% confidence interval is $81.30 \pm 5.46 \Rightarrow (75.84, 86.76)$.

G.6 Chapter 11 – Linear regression

1. (a) We first calculate $\bar{x} = 4.56$, $\sum x_i^2 = 219.46$, $\bar{y} = 30.97$, $\sum y_i^2 = 9{,}973.99$ and $\sum x_i y_i = 1{,}475.1$. The estimated regression coefficients are:

$$\widehat{\beta}_1 = \frac{1{,}475.1 - 10 \times 4.56 \times 30.97}{219.46 - 10 \times (4.56)^2} = 5.46 \quad \text{and} \quad \widehat{\beta}_0 = 30.97 - 5.46 \times 4.56 = 6.07.$$

The fitted line is:

$$\widehat{\text{Cost}} = 6.09 + 5.46 \times \text{Distance}.$$

In order to perform statistical inference, we need to find:

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2} \\ &= \left(\sum y_i^2 + n\widehat{\beta}_0^2 + \widehat{\beta}_1^2 \sum x_i^2 - 2\widehat{\beta}_0 \sum y_i - 2\widehat{\beta}_1 \sum x_i y_i + 2\widehat{\beta}_0 \widehat{\beta}_1 \sum x_i\right)/(n-2) \\ &= (9{,}973.99 + 10 \times (6.07)^2 + (5.46)^2 \times 219.46 - 2 \times 6.07 \times 309.7 \\ &\quad - 2 \times 5.46 \times 1475.1 + 2 \times 6.07 \times 5.46 \times 45.6)/(10-2) \\ &= 4.95.\end{aligned}$$

The estimated standard error of $\widehat{\beta}_1$ is:

$$\frac{\sqrt{4.95}}{\sqrt{219.46 - 10 \times (4.56)^2}} = 0.66.$$

Hence a 95% confidence interval for β_1 is $5.46 \pm 2.306 \times 0.66 \;\Rightarrow\; (3.94, 6.98)$.

(b) To test $\text{H}_0 : \beta_0 = 0$ vs. $\text{H}_1 : \beta_0 \neq 0$, we first determine the estimated standard error of $\widehat{\beta}_0$, which is:

$$\frac{\sqrt{4.95}}{\sqrt{10}} \left(\frac{219.46}{219.46 - 10 \times (4.56)^2}\right)^{1/2} = 3.07.$$

Therefore, test statistic value is:

$$\frac{6.07}{3.07} = 1.98.$$

Comparing with the t_8 distribution, this is not significant at the 5% significance level ($1.98 < 2.306$), but it is significant at the 10% significance level ($1.860 < 1.98$).

There is only weak evidence against the null hypothesis. Note though that in practice this hypothesis is not really of interest. A line through the origin implies that there is zero cost of a fire which takes place right next to a fire station. This hypothesis does not seem sensible!

2. The question implies that we want to explain changes in profitability by the passage of time. If we let x represent years and y represent profits (in £000s) then we need to perform a regression of y on x.

(a) We first calculate $\bar{x} = 4.5$, $\sum x_i^2 = 204$, $\bar{y} = 41.125$, $\sum y_i^2 = 16{,}159$ and $\sum x_i y_i = 1{,}802$. The estimated regression coefficients are:

$$\widehat{\beta}_1 = \frac{1{,}802 - 8 \times 4.5 \times 41.125}{204 - 8 \times (4.5)^2} = 7.65 \quad \text{and} \quad \widehat{\beta}_0 = 41.125 - 7.65 \times 4.5 = 6.70.$$

The fitted line is:

$$\widehat{\text{Profit}} = 6.70 + 7.65 \times \text{Year}.$$

In order to perform statistical inference, we need to find:

$$\begin{aligned}\widehat{\sigma}^2 &= \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2/(n-2) \\ &= \left(\sum y_i^2 + n\widehat{\beta}_0^2 + \widehat{\beta}_1^2 \sum x_i^2 - 2\widehat{\beta}_0 \sum y_i - 2\widehat{\beta}_1 \sum x_i y_i + 2\widehat{\beta}_0 \widehat{\beta}_1 \sum x_i\right)/(n-2) \\ &= (16{,}159 + 8 \times (6.70)^2 + (7.65)^2 \times 204 - 2 \times 6.70 \times 329 \\ &\quad - 2 \times 7.65 \times 1{,}802 + 2 \times 6.70 \times 7.65 \times 36)/(8-2) \\ &= 27.98.\end{aligned}$$

The estimated standard error of $\widehat{\beta}_1$ is:

$$\frac{\sqrt{27.98}}{\sqrt{204 - 8 \times (4.5)^2}} = 0.82.$$

Hence a 95% confidence interval for β_1 is $7.65 \pm 2.447 \times 0.82 \;\Rightarrow\; (5.64, 9.66)$.

(b) Substituting $x = 9$ we find the predicted year 9 profit (in £000s) is 75.55. The estimated standard error of this prediction is:

$$\sqrt{27.98} \times \left(1 + \frac{204 - 2 \times 9 \times 36 + 8 \times 9^2}{8 \times (204 - 8 \times (4.5)^2)}\right)^{1/2} = 6.71.$$

It follows that (using $t_{n-2} = t_6$) a 95% prediction interval for the predicted profit (in £000s) is:

$$75.55 \pm 2.447 \times 6.71 \quad \Rightarrow \quad (59.13, 91.97).$$

As 90 is in this prediction interval, we cannot reject the accountant's forecast out of hand. However, it is right at the top end of the prediction interval, and hence seems rather optimistic.

3. (a) We first calculate $\bar{x} = 4.5$, $\sum x_i^2 = 204$, $\bar{y} = 434$, $\sum y_i^2 = 1{,}938{,}174$ and $\sum x_i y_i = 19{,}766$. The estimated regression coefficients are:

$$\widehat{\beta}_1 = \frac{19{,}766 - 8 \times 4.5 \times 434}{204 - 8 \times (4.5)^2} = 98.62 \quad \text{and} \quad \widehat{\beta}_0 = 434 - 98.62 \times 4.5 = -9.79.$$

The fitted line is:

$$\widehat{\text{Expenditure}} = -9.79 + 98.62 \times \text{Year}.$$

In order to perform statistical inference, we need to find:

$$\begin{aligned}\widehat{\sigma}^2 &= \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2/(n-2) \\ &= \left(\sum y_i^2 + n\widehat{\beta}_0^2 + \widehat{\beta}_1^2 \sum x_i^2 - 2\widehat{\beta}_0 \sum y_i - 2\widehat{\beta}_1 \sum x_i y_i + 2\widehat{\beta}_0 \widehat{\beta}_1 \sum x_i\right)/(n-2) \\ &= (1{,}938{,}174 + 8 \times (-9.79)^2 + (98.62)^2 \times 204 - 2 \times (-9.79) \times 3{,}472 \\ &\quad - 2 \times 98.62 \times 19{,}766 + 2 \times (-9.79) \times 98.62 \times 36)/(8-2) \\ &= 3{,}807.65.\end{aligned}$$

The estimated standard error of $\widehat{\beta}_1$ is:

$$\frac{\sqrt{3{,}807.65}}{\sqrt{204 - 8 \times (4.5)^2}} = 9.52.$$

Hence a 95% confidence interval for β_1 is:

$$98.62 \pm 2.447 \times 9.52 \quad \Rightarrow \quad (75.32, 121.92).$$

(b) The ANOVA table is:

Source	DF	SS	MS	F
Regression	1	408,480	408,480	107.269
Residual Error	6	22,846	3,808	
Total	7	431,326		

Hence $R^2 = 408{,}480/431{,}326 = 0.947$.

(c) As R^2 is very close to 1, the linear regression model provides a very good fit.

(d) Substituting $x = 9$ we find the predicted year 9 profit (in £000s) is 877.79. The estimated standard error of this prediction is:

$$\sqrt{3{,}807.65} \times \left(1 + \frac{204 - 2 \times 9 \times 36 + 8 \times 9^2}{8 \times (204 - 8 \times (4.5)^2)}\right)^{1/2} = 78.23.$$

It follows that (using $t_{n-2} = t_6$) a 95% prediction interval for the predicted profit (in £000s) is:

$$877.79 \pm 2.447 \times 78.23 \quad \Rightarrow \quad (686.36, 1{,}069.22).$$

4. We first note $\mathrm{E}(Y) = \beta_0 + \beta_1 \mathrm{E}(X)$ and $Y - \mathrm{E}(Y) = (X - \mathrm{E}(X))\beta_1 + \varepsilon$. Hence:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \\
&= \mathrm{E}((X - \mathrm{E}(X))(X - \mathrm{E}(X))\beta_1) \\
&= \beta_1 \mathrm{Var}(X).
\end{aligned}$$

Therefore, $\beta_1 = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$. The second equality follows from the fact that $\mathrm{Corr}(X, Y) = \mathrm{Cov}(X, Y)/(\mathrm{Var}(X)\,\mathrm{Var}(Y))^{1/2}$.

Also, note that the first equality resembles the estimator:

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

although in the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, x is assumed to be fixed (to make the inference easier). Otherwise $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are no longer linear estimators, for example. The second equality reinforces the fact that $\beta_1 > 0$ if and only if x and y are positively correlated.

Appendix H
Formula sheet in the summer examination

Simple linear regression

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

LSEs: $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$, $\widehat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{j=1}^{n}(x_j - \bar{x})^2$ and:

$$\mathrm{Var}(\widehat{\beta}_0) = \frac{\sigma^2}{n} \frac{\sum_{j=1}^{n} x_j^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Estimator for the variance of ε_i: $\widehat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2/(n-2)$.

Regression ANOVA:

$$\text{Total SS} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \text{Regression SS} = \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and}$$

$$\text{Residual SS} = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Squared regression correlation coefficients:

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} \quad \text{and} \quad R^2_{adj} = 1 - \frac{(\text{Residual SS})/(n-2)}{(\text{Total SS})/(n-1)}.$$

For a given x, the expectation of y is $\mu(x) = \beta_0 + \beta_1 x$. A **$100(1-\alpha)\%$ confidence interval for $\mu(x)$** is:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{\alpha/2,\, n-2} \times \widehat{\sigma} \times \left\{ \frac{\sum_{i=1}^{n}(x_i - x)^2}{n \sum_{j=1}^{n}(x_j - \bar{x})^2} \right\}^{1/2}$$

and a **$100(1-\alpha)\%$ prediction interval covering y with probability $(1-\alpha)$** is:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{\alpha/2,\, n-2} \times \widehat{\sigma} \times \left\{ 1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n \sum_{j=1}^{n}(x_j - \bar{x})^2} \right\}^{1/2}.$$

One-way ANOVA:

Total variation: $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}^2-n\bar{X}^2$.

Between-treatments variation: $B=\sum_{j=1}^{k}n_j(\bar{X}_{\cdot j}-\bar{X})^2=\sum_{j=1}^{k}n_j\bar{X}_{\cdot j}^2-n\bar{X}^2$.

Within-treatments variation: $W=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\cdot j})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}^2-\sum_{j=1}^{k}n_j\bar{X}_{\cdot j}^2$.

Two-way ANOVA:

Total variation: $\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X})^2=\sum_{i=1}^{r}\sum_{j=1}^{c}X_{ij}^2-rc\bar{X}^2$.

Between-blocks (rows) variation: $B_{\text{row}}=c\sum_{i=1}^{r}(\bar{X}_{i\cdot}-\bar{X})^2=c\sum_{i=1}^{r}\bar{X}_{i\cdot}^2-rc\bar{X}^2$.

Between-treatments (columns) variation: $B_{\text{col}}=r\sum_{j=1}^{c}(\bar{X}_{\cdot j}-\bar{X})^2=r\sum_{j=1}^{c}\bar{X}_{\cdot j}^2-rc\bar{X}^2$.

Residual (error) variation:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X})^2=\sum_{i=1}^{r}\sum_{j=1}^{c}X_{ij}^2-c\sum_{i=1}^{r}\bar{X}_{i\cdot}^2-r\sum_{j=1}^{c}\bar{X}_{\cdot j}^2+rc\bar{X}^2.$$

lse.ac.uk/statistics

Department of Statistics
The London School of Economics
and Political Science
Houghton Street
London WC2A 2AE

Email: statistics@lse.ac.uk

Telephone: +44 (0)20 7852 3709