# MATH10282: Introduction to Statistics
## Supplementary Lecture Notes

# 1 Introduction: What is Statistics?

Statistics is:

> 'the science of learning from **data**, and of measuring, controlling, and communicating **uncertainty**; and it thereby provides the navigation essential for controlling the course of scientific and societal advances.'
>
> Davidian, M. and Louis, T.A. (2012), *Science*.
> `http://dx.doi.org/10.1126/science.1218685`

There are two basic forms: descriptive statistics and inferential statistics. In this course we will discuss both, with inferential statistics being the major emphasis.

- *Descriptive Statistics* is primarily about summarizing a given data set through numerical summaries and graphs, and can be used for exploratory analysis to visualize the information contained in the data and suggest hypotheses etc.

  It is useful and important. It has become more exciting nowadays with people regularly using fancy interactive computer graphics to display numerical information (e.g. Hans Rosling's visualisation of the change in countries' health and wealth over time – see Youtube).

- *Inferential Statistics* is concerned with methods for making conclusions about a **population** using information from a **sample**, and assessing the reliability of, and uncertainty in, these conclusions.

  This allows us to make judgements in the presence of uncertainty and variability, which is extremely important in underpinning evidence-based decision making in science, government, business etc.

Many statistical analyses and calculations are easiest to perform using a computer. We will learn how to use the **statistical software R**, which is freely available to download from `http://r-project.org` for use on your own computer. A good introductory guide is 'Introduction to R' by Venables et al. (2006), which can be downloaded as a PDF from the R project website, or accessed from the R software itself via the menu (Help→Manuals).

To interact with R, we type commands into the console, or write script files which contain several commands for longer analyses. These commands are written in the R computer programming language, whose syntax is fairly easy to learn. In this way, we can perform mathematical and statistical calculations. R has many existing built-in functions, and users are also able to create their own functions. The R software also has very good graphical facilities, which can produce high quality statistical plots. Datasets for use in the R sessions are available from the course website `https://minerva.it.manchester.ac.uk/~saralees/intro.html` You can download these and store them for use in the lab sessions.

# 2 Populations and samples

A **population** is the collection of all individuals or items under consideration in the study. For a given population there will typically be one or more variables in which we are interested. For example, consider the following populations together with corresponding variables of interest:

(i) All adults in the UK who are eligible to vote; the variable of interest is the political party supported.

(ii) Car batteries of a particular type manufactured by a particular company; the variable of interest is the lifetime of the battery before failure.

(iii) All adult males working full-time in Manchester; the variable of interest is the person's gross income.

(iv) All potential possible outcomes of a planned laboratory experiment; the variable of interest is the value of a particular measurement.

In general, the variables of interest may be either **qualitative** or **quantitative**. Qualitative variables are either **nominal**, e.g. gender or political party supported, or **ordinal**, e.g. a measurement of size grouped into three categories: small, medium or large. Quantitative variables are either **discrete**, for example a count, or **continuous**, such as the variables income and lifetime above.

We wish to make conclusions, or **inferences**, about the population characteristics of variables of interest. One way to do so is to conduct a census, i.e. to collect data for each individual in the population. However often this is not feasible, due to one or more of the following:

- It may be too expensive or time consuming to do so, e.g. (i)
- Testing may be destructive, e.g. (ii), and we need to have some products left to sell!
- The population may be purely conceptual, e.g. (iv)

Instead, we collect data only for a **sample**, i.e. a subset of the population. We then use the characteristics of the sample to estimate the characteristics of the population. In order for this procedure to give a good estimate, the sample must be **representative** of the population. Otherwise, if an unrepresentative or 'biased' sample is used the conclusions will be systematically incorrect.

Some examples of samples from populations are given below:

(i) In an opinion poll in May 2015, a sample of 1000 adults was obtained and asked which political party they intended to vote for in the upcoming UK General Election on 7 May 2015. A summary of these responses is:

| Party | Number of supporters |
|---|---|
| Conservative | 369 |
| Labour | 314 |
| Lib Dem | 75 |
| UKIP | 118 |
| Other | 124 |

(ii) A random sample of 40 manufactured car batteries was taken from the production line, and their lifetimes (in years) determined. The data are as follows, arranged in ascending order for convenience:

1.6, 1.9, 2.2, 2.5, 2.6, 2.6, 2.9, 3.0, 3.0, 3.1,
3.1, 3.1, 3.1, 3.2, 3.2, 3.2, 3.3, 3.3, 3.3, 3.4,
3.4, 3.4, 3.5, 3.5, 3.6, 3.7, 3.7, 3.7, 3.8, 3.8,
3.9, 3.9, 4.1, 4.1, 4.2, 4.3, 4.4, 4.5, 4.7, 4.7

(iii) We could obtain a sample of 500 adult males working full-time in Manchester. The following table summarizes a hypothetical data set of the annual incomes in thousands of pounds for such a sample.

| Interval | Frequency | Percentage |
|---|---|---|
| 5 to 15 | 83 | 16.6 |
| 15 to 25 | 142 | 28.4 |
| 25 to 35 | 90 | 18.0 |
| 35 to 45 | 79 | 15.8 |
| 45 to 55 | 46 | 9.2 |
| 55 to 65 | 28 | 5.6 |
| 65 to 75 | 13 | 2.6 |
| 75 to 85 | 6 | 1.2 |
| 85 to 95 | 4 | 0.8 |
| 95 to 105 | 3 | 0.6 |
| 105 to 115 | 0 | 0.0 |
| 115 to 125 | 2 | 0.4 |
| 125 to 135 | 0 | 0.0 |
| 135 to 145 | 0 | 0.0 |
| 145 to 155 | 1 | 0.2 |
| 155 to 165 | 0 | 0.0 |
| 165 to 175 | 1 | 0.2 |
| 175 to 185 | 1 | 0.2 |
| 185 to 195 | 1 | 0.2 |
| Totals | 500 | 100.0 |

The intervals in the table are open on the left and closed on the right, e.g. the first row gives the count of incomes in the range $(5, 15]$.

### 2.1 Finite population sampling

In modern Statistics, the most common way of guaranteeing representativeness is to use a random sample of size $n$ chosen according to a probabilistic sampling rule. This probabilistic sampling is objective and eliminates investigator bias. For a population of finite size $N$, the most common method is to use **simple random sampling**. This takes two main forms: **sampling without replacement** and **sampling with replacement**.

- *Sampling without replacement*: each of the $\binom{N}{n}$ possible samples of $n$ distinct individuals from the population has equal probability of selection, $\binom{N}{n}^{-1}$. No individual appears more than once in the sample.

  This can be implemented by choosing individuals sequentially, one at a time, as follows. For $i = 1, \ldots, n$:

  Step 1. Select an individual at random with equal probability from the remaining population of size $N-i+1$

  Step 2. Include the selected individual as the $i$th member of the sample, and remove the selected individual from the population, leaving $N - i$ individuals remaining.

  The above steps are repeated until a sample of size $n$ is obtained.

- *Sampling with replacement:* each individual may appear any number of times in the sample, leading to $N^n$ possible samples. The probability of selecting any particular sample is $N^{-n}$. This can be implemented using a similar sequential algorithm to before, where instead in Step 2 the selected individual is not removed from the population.

**Example.** Let $v_1, \ldots, v_N$ denote the values of the variable $X$ for the 1st, ..., $N$th individuals in the population.

Suppose that interest lies in estimating the population mean of $X$,

$$\mu = \frac{1}{N}\sum_{j=1}^{N} v_j \,.$$

Let $X_1, \ldots, X_n$ be the values of $X$ in a sample of size $n$ chosen by sampling without replacement. The population mean $\mu$ can be estimated by

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

The value of $\overline{X}$ will be different for different samples, and so $\overline{X}$ is a random variable because the sample is chosen randomly. Thus, $\overline{X}$ has its own probability distribution, which is known as its sampling distribution.

How can we measure the performance of the above method of estimating $\mu$? One way is to calculate the expectation and variance of the sampling distribution of $\overline{X}$. In particular, it can be shown that under sampling without replacement

$$\mathrm{E}(\overline{X}) = \mu \,.$$

As a result, $\overline{X}$ is said to be **unbiased**. We will study this unbiasedness property further in Chapter 5. Moreover it is possible to show that under sampling without replacement

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) , \tag{1}$$

where the **population variance** $\sigma^2$ is defined as

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(v_j - \mu)^2 \,.$$

Note that an alternative expression for the population variance is

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} v_j^2 - \mu^2 \,.$$

This is often more practical for calculations.

For illustration consider the highly simplified scenario where there are three individuals $A, B, C$, with corresponding $X$ values of 1, 2, 3 respectively, and a sample of size 2 is chosen using sampling without replacement. The table below shows all possible samples, together with the corresponding values of $\overline{X}$.

| Sample | Selection probability | $X$ values | $\overline{X}$ |
|---|---|---|---|
| {B,C} | $\frac{1}{3}$ | (2,3) | $\frac{5}{2}$ |
| {A,C} | $\frac{1}{3}$ | (1,3) | 2 |
| {A,B} | $\frac{1}{3}$ | (1,2) | $\frac{3}{2}$ |

Table: all possible samples in the illlustrative example

We can verify that $\mathrm{E}(\overline{X}) = \mu$ as follows. First note from the table that the p.m.f. of $\overline{X}$ is

| $\overline{x}$ | 5/2 | 2 | 3/2 |
|---|---|---|---|
| $p_{\overline{X}}(\overline{x})$ | 1/3 | 1/3 | 1/3 |

Hence

$$\mathrm{E}(\overline{X}) = \sum_{\overline{x} \in R_{\overline{X}}} \overline{x}\, p_{\overline{X}}(\overline{x}) = \frac{1}{3} \times \frac{5}{2} + \frac{1}{3} \times 2 + \frac{1}{3} \times \frac{3}{2} = 2\,.$$

Note also that the population mean is $\mu = \frac{1}{3}(1+2+3) = 2$, and so $\mathrm{E}(\overline{X}) = \mu$ as anticipated.

We can also compute the variance of the sampling distribution for $\overline{X}$. Recall that for any random variable $Y$, $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2$. Note further that

$$\mathrm{E}(\overline{X}^2) = \sum_{\overline{x} \in R_{\overline{X}}} \overline{x}^2 p_{\overline{X}}(\overline{x}) = \left(\frac{5}{2}\right)^2 \times \frac{1}{3} + 2^2 \times \frac{1}{3} + \left(\frac{3}{2}\right)^2 \times \frac{1}{3} = \frac{25}{6}\,,$$

and so $\mathrm{Var}(\overline{X}) = \mathrm{E}(\overline{X}^2) - (\mathrm{E}\,\overline{X})^2 = 25/6 - 2^2 = 1/6$.

## 2.2 Sampling from a general population

For a general (i.e. not necessarily finite) population, the value of a quantitative variable for a randomly selected individual can be described by a real-valued random variable $X$ with cumulative distribution function (c.d.f.)

$$F_X(x) = \mathrm{P}(X \leq x)\,.$$

If $X$ is a continuous random variable then there is also an associated probability density function (p.d.f.) $f_X(x)$, which satisfies

$$\frac{dF_X(x)}{dx} = f_X(x)\,.$$

If $X$ is a discrete random variable then there is instead a probability mass function (p.m.f.) $p_X(x)$ satisfying

$$\sum_{t \leq x} p_X(t) = F_X(x)\,.$$

We now recall several concepts from MATH10141 Probability I. For a continuous random variable, the population mean $\mu$ and variance $\sigma^2$ of $X$ are

$$\mu = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$
$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx\,.$$

For a discrete random variable, these quantities are instead defined in terms of the p.m.f.

$$\mu = \sum_{x \in R_X} x\, p_X(x)$$
$$\sigma^2 = \sum_{x \in R_X} (x-\mu)^2 p_X(x)\,,$$

where $R_X \subseteq \mathbb{R}$ denotes the range-space of $X$.

### 2.2.1 Independent events

Let $\Omega$ be a probability sample space. A pair of events $A, B \subseteq \Omega$ is said to be *independent* if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B)\,.$$

More generally, events $B_1, \ldots, B_n$ are *mutually independent* if for every subset $\{B_{i_1}, \ldots, B_{i_k}\}$, $(k \geq 2)$ of $\{B_1, \ldots, B_n\}$,

$$\mathrm{P}(B_{i_1} \cap \cdots \cap B_{i_k}) = \mathrm{P}(B_{i_1}) \times \cdots \times \mathrm{P}(B_{i_k}) \,.$$

### 2.2.2 Independent random variables

A collection of real-valued random variables $X_1, \ldots, X_n$ is said to be *independent* if for any subsets $B_1, \ldots, B_n \subseteq \mathbb{R}$ the events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are independent, i.e.

$$\mathrm{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathrm{P}(X_1 \in B_1) \times \cdots \times \mathrm{P}(X_n \in B_n) \,.$$

The following special cases have alternate equivalent definitions:

- If $X_1, \ldots, X_n$ are identically distributed with c.d.f. $F_X(x)$, then $X_1, \ldots, X_n$ are independent if and only if
$$\mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = F_X(x_1) \times \cdots \times F_X(x_n) \,.$$
- If $X_1, \ldots, X_n$ are discrete random variables with common p.m.f. $p_X(x)$, then $X_1, \ldots, X_n$ are independent if and only if the joint p.m.f satisfies
$$p_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = p_X(x_1) \times \cdots \times p_X(x_n) \,.$$
- If $X_1, \ldots, X_n$ are continuous random variables with common p.d.f. $f_X(x)$, then $X_1, \ldots, X_n$ are independent if and only if the joint p.d.f. satistifies
$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_X(x_1) \times \cdots \times f_X(x_n) \,.$$

The idea of independence is now used to define sampling from a general population. We say that $X_1, \ldots, X_n$ are a **random sample** from $X$ if $X_1, \ldots, X_n \sim F_X(x)$ independently. We may also say that $X_1, \ldots, X_n$ is a random sample from $F_X(x)$, $f_X(x)$ or $p_X(x)$.

**Example.** Simple random sampling of $n$ individuals with replacement from a finite population of size $N$ with $X$-values $v_1, \ldots, v_n$ corresponds to independent random sampling of $X_1, \ldots, X_n$ from the p.m.f.

$$p_X(x) = \frac{1}{N} \times \{\text{number of } j \text{ such that } v_j = x\} \,.$$

Similar to the previous section, we may use the characteristics of the sample to estimate the characteristics of the population. For example, suppose we are interested in the population mean $\mu$. This may again be estimated by the sample mean, i.e.

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

Once again, the value of $\overline{X}$ is random because $X_1, \ldots, X_n$ is a random sample from the population. Moreover, it is again true that

$$\mathrm{E}(\overline{X}) = \mu \,.$$

The variance of the sample mean is

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

For a finite population of size $N$, we can compare the properties of $\overline{X}$ under the two types of sampling: independent random sampling and random sampling without replacement. We see comparing equations (1) and (2) that when using sampling without replacement, $\mathrm{Var}(\overline{X})$ is smaller by a factor

$$\text{f.p.c.} = \frac{N-n}{N-1}\,,$$

which is called the **finite population correction** (f.p.c.). The difference in $\mathrm{Var}(\overline{X})$ occurs because under sampling without replacement the $X_i$ are not independent. However, the $X_i$ can be considered to be approximately independent when $N$ is large and the **sampling proportion** $n/N$ is small. In this case,

$$\text{f.p.c.} = \frac{1-n/N}{1-1/N} \approx 1\,.$$

In the remainder of this course we will always assume that $X_1, \ldots, X_n$ are sampled independently from a c.d.f. $F_X(x)$.

# 3 Probability models for data

Let $x_1, \ldots, x_n$ be the observed values in a particular random sample of the random variable $X$, whose distribution is unknown. We may wish to use these data to estimate the probability of an event $\{X \in A\}$, $A \subseteq R_X$. One way is to use the **empirical probability** of the event, in other words the proportion of the sample values that lie in $A$,

$$\widehat{\mathrm{P}}(X \in A) = \frac{\#\{i : x_i \in A\}}{n}\,.$$

An alternative approach is to assume that the data were generated as a random sample from a particular parametric probability model, e.g. $N(\mu, \sigma^2)$. Such models usually contain unknown parameters, e.g. in the previous example the parameters $\mu$ and $\sigma^2$ are unknown. We can use the sample to estimate the parameters of the distribution, thereby fitting the model to the data. A fitted model can be used to calculate probabilities of events of interest.

If the chosen model is a good fit then the empirical and model-based estimated probabilities of the event should be similar. Small differences between the empirical and model-based estimated probabilities will occur frequently due to the fact that we have only observed a random sample and not the entire population. Thus, both estimates exhibit random variation around the true population probability. However, large differences between empirical and model-based probabilities may be indicative that the chosen parametric model is a poor approximation of the true data generating process. This is best illustrated by studying some examples.

## 3.1 Continuous data

### 3.1.1 Component lifetime data

A sample of $n = 50$ components was taken from a production line, and their lifetimes (in hours) determined. A tabulation of the sample values is given overleaf. A possible parametric model for these data is to assume that they are a random sample from a normal distribution $N(\mu, \sigma^2)$. The parameters $\mu$ and $\sigma^2$ can be estimated from the sample by $\widehat{\mu} = \overline{x} = 334.6$, $\widehat{\sigma}^2 = s^2 = 15.288$.

We can informally investigate how well this distribution fits the data by superimposing the probability density function of a $N(334.6,\, 3.912^2)$ distribution onto a histogram of the data. This is illustrated in the figure overleaf, which shows the fit to be reasonably good, particularly for data greater than the mean.

| Intervals | Frequencies | Percents |
|---|---|---|
| 323.75 to 326.25 | 1 | 2 |
| 326.25 to 328.75 | 0 | 0 |
| 328.75 to 331.25 | 9 | 18 |
| 331.25 to 333.75 | 12 | 24 |
| 333.75 to 336.25 | 11 | 22 |
| 336.25 to 338.75 | 10 | 20 |
| 338.75 to 341.25 | 5 | 10 |
| 341.25 to 343.75 | 1 | 2 |
| 343.75 to 346.25 | 1 | 2 |
| Totals | 50 | 100 |

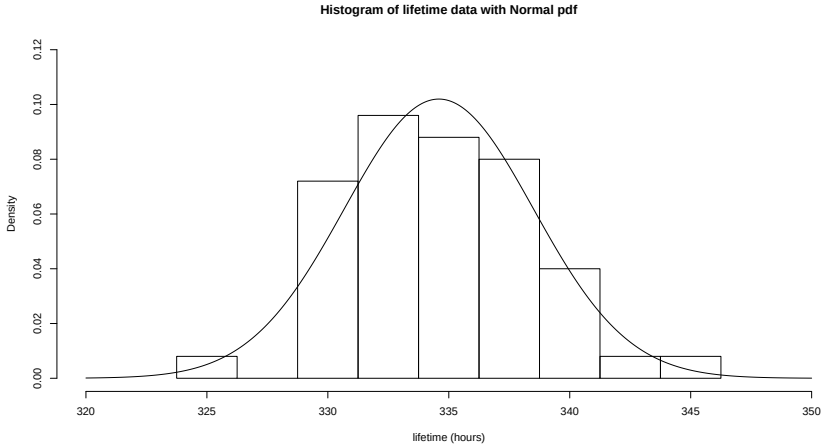


Figure 1: Histogram of the component lifetime data together with a $N(334.6,\, 3.912^2)$ p.d.f.

This figure can be obtained using the R code below. The `lines` command draws a curve through the $(x, y)$ co-ordinates provided.

```
xx <- comp_lifetime$lifetime
xv <- seq(320, 350, 0.1)
yv <- dnorm(xv, mean=mean(xx), sd=sd(xx))
hist(xx, freq=F, breaks=seq(from=323.75, to= 346.25, by=2.5),
  xlim=c(320, 350), ylim=c(0, 0.12), main="Histogram of
    lifetime data with Normal pdf", xlab="lifetime (hours)")
lines(xv, yv)
```

The fitted normal distribution appears to be a reasonably good fit to the observed data, thus we may use it to calculate estimated probabilities. For example, consider the question 'what is the estimated probability that a randomly selected component lasts between 330 and 340 hours?'. To answer this, let the random variable $X$ be the lifetime of a randomly selected component. We require $\mathrm{P}(330 < X < 340)$ under the fitted normal model, $X \sim N(334.6, 3.912^2)$:

$$\begin{aligned}
\mathrm{P}(330 < X < 340) &= \mathrm{P}\left(\frac{330.0 - 334.6}{3.912} < \frac{X - 334.6}{3.912} < \frac{340.0 - 334.6}{3.912}\right) \\
&= \mathrm{P}(-1.18 < Z < 1.38)\,, \quad \text{where } Z \sim N(0,1) \\
&= \Phi(1.38) - \Phi(-1.18) = 0.9162 - 0.1190 = 0.7972\,.
\end{aligned}$$

Hence, using the fitted normal model we estimate that 79.72% of randomly selected components will have lifetimes between 330 and 340 hours.

### 3.1.2 Manchester income data

If we superimpose a normal density curve onto the histogram for these data, then we see that the symmetric normal distribution is a poor fit, since the data are skewed. In particular, the normal density extends to negative income values despite the fact that all of the incomes in the sample are positive.

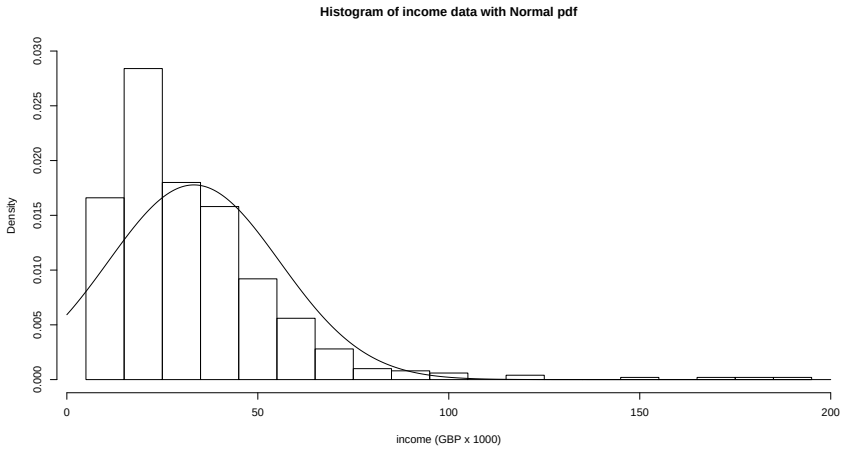


Figure 2: Histogram of the income data with the p.d.f. of the fitted normal distribution.

This figure can be obtained using the following R code:

```
xx <- income$income
xv <- seq(0, 200, 0.5)
yv <- dnorm(xv, mean=mean(xx), sd=sd(xx))
hist(xx, freq=F, breaks=seq(from=5, to=195, by=10),
  ylim=c(0, 0.030), xlab="income (GBP x 1000)",
    main="Histogram of income data with Normal pdf")
lines(xv, yv)
```

One way forward is to look for a transformation which will make the data appear to be more normally distributed. Because the data are strongly positively skewed on the positive real line one possibility is to take logarithms.

In the figure below, we see a histogram of the log transformed income data. The fit of the superimposed normal p.d.f. now looks reasonable, although there are perhaps slightly fewer sample observations than might be expected according to the normal model in the left-hand tail and centre. There are also some outliers in the right-hand tail.

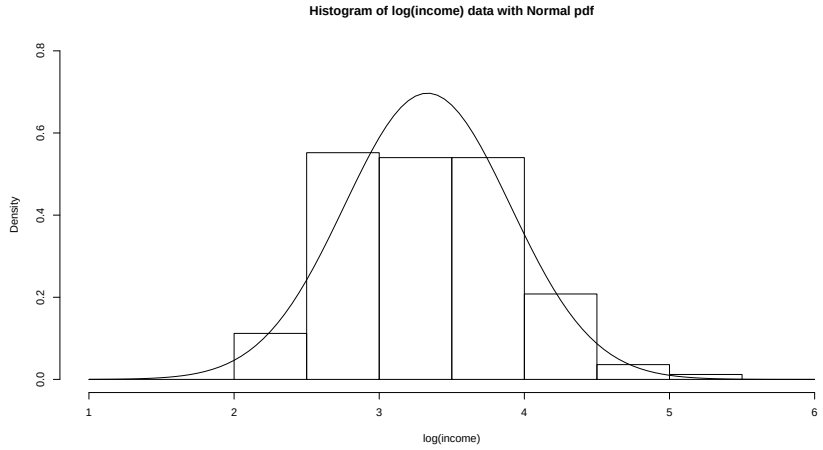


Figure 3: Histogram of log(income) with a normal p.d.f.

This figure can be obtained using the following R code:

```
lxx <- log(xx)
lxv <- seq(1, 6, 0.05)
lyv <- dnorm(lxv, mean=mean(lxx), sd=sd(lxx))
hist(lxx, freq=F, breaks=c(1, 1.5,2, 2.5, 3, 3.5, 4,
      4.5, 5, 5.5, 6), ylim=c(0, 0.80), xlab="log(income)",
      main="Histogram of log(income) data with Normal pdf")
lines(lxv, lyv)
```

Even if it is not clear whether or not we can find a completely satisfactory parametric model, we will see in a later section that we can still make approximate inferences about the mean income in the population by appealing to the central limit theorem.

## 3.2 Discrete data

### 3.2.1 Opinion poll data

Let $X$ be the party supported by a randomly selected voter,

$$X = \begin{cases} \text{Conservative} & \text{with probability } p_{\text{C}} \\ \text{Labour} & \text{with probability } p_{\text{L}} \\ \text{Liberal Democrats} & \text{with probability } p_{\text{LD}} \\ \text{UKIP} & \text{with probability } p_{\text{U}} \\ \text{Other} & \text{with probability } p_{\text{O}}\,, \end{cases}$$

where 'Other' includes all other parties. As suggested earlier, we can estimate the probabilities $p_C$, $p_L$, etc. by the proportions of sampled individuals supporting the corresponding party. Specifically we obtain the following

estimates:

$$\begin{aligned}
\widehat{p}_{\mathrm{C}} &= \widehat{\mathrm{P}}(X = \text{Conservatives}) = 369/1000 = 0.369, \\
\widehat{p}_{\mathrm{L}} &= \widehat{\mathrm{P}}(X = \text{Labour}) = 314/1000 = 0.314, \\
\widehat{p}_{\mathrm{LD}} &= \widehat{\mathrm{P}}(X = \text{Liberal Democrats}) = 75/1000 = 0.075, \\
\widehat{p}_{\mathrm{U}} &= \widehat{\mathrm{P}}(X = \text{UKIP}) = 118/1000 = 0.118\,, \\
\widehat{p}_{\mathrm{O}} &= \widehat{\mathrm{P}}(X = \text{Other party}) = 124/1000 = 0.124\,.
\end{aligned}$$

It is beyond the scope of this module to consider a joint probability model for the vector $(n_{\mathrm{C}}, n_{\mathrm{L}}, n_{\mathrm{LD}}, n_{\mathrm{U}}, n_{\mathrm{O}})$ containing the numbers of individuals supporting each of the five possible choices in a sample of size $n$. However we may slightly simplify the situation by focussing on whether or not a randomly chosen voter supports Labour.

Let the random variable $X_L$ denote the number of voters out of the 1000 who support Labour. An appropriate model may be

$$X_L \sim \mathrm{Bi}(n, p_L)\,,$$

with $n = 1000$, and $p_L$ is estimated by $\widehat{p}_L = 0.314$. We may use the fitted model to answer various questions, e.g. 'what is the estimated probability that in a random sample of 1000 voters at least 330 will support Labour?'.

We require $\mathrm{P}(X_L \geq 330)$ under the fitted model $\mathrm{Bi}(1000, 0.314)$. It is easiest to use a normal approximation to the binomial distribution, which gives

$$\begin{aligned}
\mathrm{P}(X_L \geq 330) &\approx 1 - \Phi\left(\frac{329.5 - 1000 \times 0.314}{\sqrt{1000 \times 0.314 \times 0.686}}\right) \\
&= 1 - \Phi(1.0561) = 0.1455\,,
\end{aligned}$$

using a continuity correction.

An interesting question is whether, in the population, voters are equally as likely to support Labour as they are to support the Conservatives, i.e. is it true that $p_{\mathrm{L}} = p_{\mathrm{C}}$? Even if it is true that the population proportions $p_{\mathrm{L}}$ and $p_{\mathrm{C}}$ are equal, the numbers supporting Labour and Conservative in the sample will usually be slightly different simply due to random variation in the sample selection. Thus, the sample only contains significant evidence that $p_L \neq p_C$ if the difference between the numbers of people in the sample supporting Labour and Conservative is 'large'. However, how do we decide how large the difference needs to be in order to support the conclusion $p_L \neq p_C$? This kind of question will be addressed in a later chapter on Hypothesis Testing.

# 4 Sampling distributions of sample statistics

Let $X_1, \ldots, X_n$ be a random sample from a distribution $F_X(x)$. A **statistic** is a function of the data,

$$h(X_1, \ldots, X_n)\,.$$

The value of this statistic will usually be different for different samples. As the sample data is random, the statistic is also a random variable. If we repeatedly drew samples of size $n$, calculating and recording the value of the sample statistic each time, then we would build up its probability distribution. The probability distribution of a sample statistic is referred to as its **sampling distribution**.

In this section we will see how to analytically determine the sampling distributions of some statistics, while with certain others we can appeal to the central limit theorem. Simulation techniques can also be used to

investigate sampling distributions of statistics empirically.

### 4.1 Sample mean

The mean and variance of the distribution $F_X(x)$ are denoted by $\mu$ and $\sigma^2$ respectively. In the case that the distribution is continuous with p.d.f. $f_X(x)$,

$$\begin{aligned}\mu = \mathrm{E}(X) &= \int_{-\infty}^{\infty} x\, f_X(x) dx \\ \sigma^2 = \mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu^2 \,.\end{aligned}$$

When the distribution is discrete with p.m.f. $p_X(x)$, $\mu$ and $\sigma^2$ are defined by:

$$\begin{aligned}\mu = \mathrm{E}(X) &= \sum_{x \in R_X} x p_X(x) \\ \sigma^2 = \mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2] \\ &= \sum_{x \in R_X} (x-\mu)^2 p(x) = \sum_{x \in R_X} x^2 p(x) - \mu^2 \,,\end{aligned}$$

where $R_X$ is the range space of $X$.

The random variables $X_1, \ldots, X_n$ are assumed to be independent and identically distributed (often abbreviated to i.i.d.) random variables, each being distributed as $F_X(x)$. This means that $\mathrm{E}(X_i) = \mu$ for $i = 1, \ldots, n$ and $\mathrm{Var}(X_i) = \sigma^2$ for $i = 1, \ldots, n$.

The sample mean of the $n$ sample variables is:

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

It is straightforward to calculate the mean of the sampling (probability) distribution of $\overline{X}$ as follows:

$$\begin{aligned}\mathrm{E}(\overline{X}) &= \mathrm{E}\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] \\ &= \frac{1}{n}[\mathrm{E}(X_1) + \ldots + \mathrm{E}(X_n)] \\ &= \frac{n\mu}{n} = \mu \,,\end{aligned}$$

while the variance is

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

Here we have used $\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$, which holds because the $X_i$ are independent.

These results tell us that the sampling distribution of the sample mean $\overline{X}$ is centered on the common mean $\mu$ of each of the sample variables $X_1, \ldots, X_n$ (i.e. the mean of the distribution from which the sample is

obtained) and has variance equal to the common variance of the $X_i$ divided by $n$. Thus, as the sample size $n$ increases, the sampling distribution of $\overline{X}$ becomes more concentrated around the true mean $\mu$.

In the above discussion nothing specific has been said regarding the actual distribution from which the $X_i$ have been sampled. All we are assuming is that the mean and variance of the underlying distribution are both finite.

### 4.1.1 Normally distributed data

In the special case that the $X_i$ are normally distributed then we can make use of some important results. Let the random variable $X \sim N(\mu_X, \sigma_X^2)$ and let the random variable $Y \sim N(\mu_Y, \sigma_Y^2)$, **independently** of $X$. Then we have the following results:

(i) $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

(ii) $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$

(iii) In general, $c_1X + c_2Y \sim N(c_1\mu_X + c_2\mu_Y, c_1^2\sigma_X^2 + c_2^2\sigma_Y^2)$; $c_1 \neq 0$, $c_2 \neq 0$.

These results extend in a straightforward manner to the linear combination of $n$ independent normal random variables. Let $X_1, \ldots X_n$ be $n$ independent normally distributed random variables with $\mathrm{E}(X_i) = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$ for $i = 1, \ldots, n$. Thus, here the normal distributions for different $X_i$ may have different means and variances. We then have that

$$\sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i\mu_i, \sum_{i=1}^{n} c_i^2\sigma_i^2\right)$$

where the $c_i \in \mathbb{R}$.

If now the $X_i$ in the sample are i.i.d. $N(\mu, \sigma^2)$ random variables then the sample mean, $\overline{X}$, is a linear combination of the $X_i$ (with $c_i = \frac{1}{n}$, $i = 1, \ldots, n$, using the notation above). Thus, $\overline{X}$ is normally distributed with mean $\mu$ and variance $\sigma^2/n$, i.e. $\overline{X}_n \sim N(\mu, \sigma^2/n)$. This result enables us to make probabilistic statements about the mean under the assumption of normality.

**Example 1.** (Component lifetime data).

In Chapter 3 we saw that the normal distribution is a reasonable probability model for the lifetime data and it seems sensible to estimate the two parameters ($\mu$ and $\sigma^2$) of this distribution by the corresponding sample quantities, $\overline{x}$ and $s^2$. For these data $\overline{x} = 334.59$ and $s^2 = 15.288$, and so our fitted model is $X \sim N(334.59, 15.288)$. Under this fitted model for $X$, the mean $\overline{X}$ of a new sample of size 50 from the population follows a $N(334.59, 15.288/50)$ distribution. We can then, for example, estimate the probability that the mean of such a sample exceeds 335,

$$\begin{aligned}\mathrm{P}(\overline{X} > 335.0) &= 1 - \Phi\left(\frac{335.0 - 334.59}{\sqrt{15.288/50}}\right) \\ &= 1 - \Phi(0.74) = 1 - 0.7704 = 0.2296\,.\end{aligned}$$

### 4.1.2 Using the central limit theorem

In the previous section, we saw that the random quantity $\overline{X}$ has a sampling distribution with mean $\mu$ and variance $\sigma^2/n$. In the special case when we are sampling from a normal distribution, $\overline{X}$ is also normally distributed. However, there are many situations when we cannot determine the exact form of the distribution of $\overline{X}$. In such circumstances, we may appeal to the central limit theorem and obtain an approximate distribution.

**The central limit theorem**: Let $X$ be a random variable with mean $\mu$ and variance $\sigma^2$. If $\overline{X}_n$ is the mean of a random sample of size $n$ drawn from the distribution of $X$, then the distribution of the statistic

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution as $n \to \infty$.

This means that, for a *large* random sample from a population with mean $\mu$ and variance $\sigma^2$, the sample mean $\overline{X}_n$ is *approximately* normally distributed with mean $\mu$ and variance $\sigma^2/n$. Since, for large $n$, $\overline{X}_n \sim N(\mu, \sigma^2/n)$ approximately we have that $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$ approximately.

There is no need to specify the form of the underlying distribution $F_X$, which may be either discrete or continuous, in order to use this result. As a consequence it is of tremendous practical importance.

A common question is 'how large does $n$ have to be before the normality of $\overline{X}$ is reasonable?' The answer depends on the degree of non-normality of the underlying distribution from which the sample has been drawn. The more non-normal $F_X$ is, the larger $n$ needs to be. A useful rule-of-thumb is that $n$ should be at least 30.

**Example 2.** (Income data). What is the approximate probability that the mean gross income based on a new random sample of size $n = 500$ lies between 33.0 and 33.5 thousand pounds?

The underlying distribution is not normal but we can appeal to the central limit theorem to say that

$$\overline{X}_{500} \sim N(\mu, \sigma^2/n) \quad \text{approximately.}$$

We may estimate $\mu$ and $\sigma^2$ from the data by $\widehat{\mu} = \overline{x} = 33.27$, $\widehat{\sigma}^2 = s^2 = 503.554$. Therefore, using the fitted values of the parameters we may estimate the probability as

$$\begin{aligned} \mathrm{P}(33.0 < \overline{X}_{500} < 33.5) &\approx \Phi\left(\frac{33.50 - 33.27}{22.44/\sqrt{500}}\right) - \Phi\left(\frac{33.00 - 33.27}{22.44/\sqrt{500}}\right) \\ &\approx \Phi(0.23) - \Phi(-0.27) = 0.5910 - 0.3936 \\ &\approx 0.1974\,. \end{aligned}$$

Hence we estimate the probability $\overline{X}$ lies between 33.0 and 33.5 to be 0.1974.

## 4.2 Sample proportion

Suppose now that we have a random sample $X_1, \ldots, X_n$ where the $X_i$ are i.i.d. $\mathrm{Bi}(1, p)$ random variables. Thus, $X_i = 1$ ('success') with probability $p$ and $X_i = 0$ ('failure') with probability $1 - p$. We know that $\mathrm{E}(X_i) = p$ for $i = 1, \ldots, n$ and $\mathrm{Var}(X_i) = p(1-p)$ for $i = 1, \ldots, n$.

The proportion of cases in the sample who have $X_i = 1$, in other words the proportion of 'successes', is given by

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

We have that $\mathrm{E}(\overline{X}_n) = p$ and $\mathrm{Var}(\overline{X}_n) = \frac{p(1-p)}{n}$. By the central limit theorem, for large $n$, $\overline{X}_n$ is approximately distributed as $N\left(p, \frac{p(1-p)}{n}\right)$ which enables us to easily make probabilistic statements about the proportion of 'successes' in a sample of size $n$.

We can also say that, for large $n$, the total number of 'successes' in the sample, given by $\sum_{i=1}^n X_i$, is approximately normally distributed with mean $np$ and variance $np(1-p)$.

Recall that, for the normal approximation to be reasonable in this context we require that

$$n \geq 9. \max\left\{\frac{1-p}{p}, \frac{p}{1-p}\right\}.$$

**Example 3.** Suppose that, in a particular country, the unemployment rate is 9.2%. Suppose that a random sample of 400 individuals is obtained. What are the approximate probabilities that:

(i) Forty or fewer were unemployed;

(ii) The proportion unemployed is greater than 0.125.

**Solution:**

(i) For $i = 1, \ldots, n$ let the random variable $X_i$ satisfy

$$X_i = \begin{cases} 1 & \text{if the } i\text{th worker is unemployed} \\ 0 & \text{otherwise}. \end{cases}$$

From the question, $\mathrm{P}(X_i = 1) = 0.092$ and $\mathrm{P}(X_i = 0) = 0.908$.

We have $n = 400 \geq \{0.9, 88.8\}$ so that the normal approximation will be valid. Note that $np = 400 \times 0.092 = 36.8$ and $np(1-p) = 400 \times 0.092 \times 0.908 = 33.414$, and $\sum_{i=1}^{n} X_i \sim N(np, np(1-p))$ approximately.

$$\begin{aligned} \mathrm{P}\left(\sum_{i=1}^{400} X_i \leq 40\right) &= \mathrm{P}\left(\frac{\sum_{i=1}^{400} X_i - 36.8}{\sqrt{33.414}} \leq \frac{40.5 - 36.8}{\sqrt{33.414}}\right) \\ &\approx \mathrm{P}\left(Z \leq 0.640\right), \quad \text{where } Z \sim N(0,1) \text{ approx.} \\ &= \Phi(0.640) \\ &= 0.7390. \end{aligned}$$

(ii) Here, $\frac{p(1-p)}{n} = \frac{0.092 \times 0.908}{400} = 0.0002088$. Thus,

$$\begin{aligned} \mathrm{P}\left(\overline{X}_{400} > 0.125\right) &= \mathrm{P}\left(\frac{\overline{X}_{400} - 0.092}{\sqrt{0.0002088}} > \frac{0.125 - 0.092}{\sqrt{0.0002088}}\right) \\ &\approx 1 - \Phi(2.284) \\ &= 1 - 0.9888 \\ &= 0.0112. \end{aligned}$$

### 4.3 Sample variance

In this section we will look at the sampling distribution of the sample variance, $S^2$, defined by

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

where $X_1, \ldots, X_n$ are a random sample from the distribution with c.d.f. $F_X(\cdot)$ with mean $\mu$ and variance $\sigma^2$.

If $F_X$ is *any* discrete or continuous distribution with a finite variance then

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

Hence, we can see that by using divisor $(n-1)$ in the definition of $S^2$, we obtain a statistic whose sampling distribution is centered on the true distribution value of $\sigma^2$. This would not be the case if we had used the perhaps more intuitively obvious value of $n$.

We will look more specifically at the case when the $X_i$ are sampled from the $N(\mu,\sigma^2)$ distribution. In order to do so, we first need to introduce a new continuous probability distribution, the chi-squared ($\chi^2$) distribution.

### 4.3.1 The chi-squared ($\chi^2$) distribution

The continuous random variable $Y$ is said to have $\chi^2$ distribution with $k$ degrees of freedom (d.f.), written as $\chi^2(k)$, iff its pdf is given by

$$f(y)=\begin{cases}\frac{1}{2^{k/2}\Gamma(k/2)}y^{(k/2)-1}e^{-y/2}, & y>0\\ 0, & \text{otherwise.}\end{cases}$$

Note that this is a special case of the Gamma distribution with parameters $\alpha=k/2$ and $\beta=1/2$. Note that when $k=2$, $Y\sim\mathrm{Exp}(1/2)$. The mean and variance are given by $\mathrm{E}(Y)=k$ and $\mathrm{Var}(Y)=2k$.

The p.d.f.s of chi-squared random variables with d.f. $=1,3,6$, and 12 are shown in Figure 1. Note that the p.d.f. becomes more symmetric as the number of degrees of freedom $k$ becomes larger.

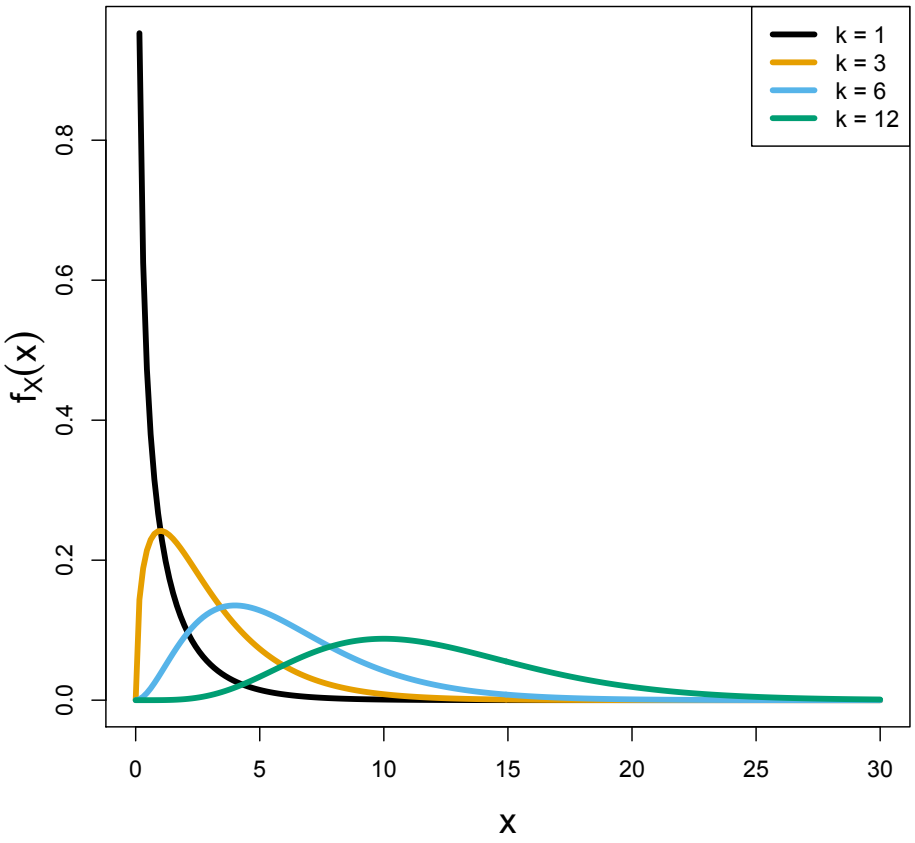


Figure 4: Chi-squared p.d.f.s with different degrees of freedom.

### 4.3.2 The connection with the normal distribution

Let $Z_1, \ldots, Z_k$ be $k$ i.i.d. standard normal random variables, i.e. each has a $N(0,1)$ distribution. Then, the random variable

$$Y = \sum_{i=1}^{k} Z_i^2$$

has a $\chi^2$ distribution with $k$ degrees of freedom.

We may use this fact to check that for $Y \sim \chi^2(k)$ we have $\mathrm{E}(Y) = k$, as follows. First note that if $Z_i \sim N(0,1)$ then

$$\begin{aligned} 1 &= \mathrm{Var}(Z_i) \\ &= \mathrm{E}(Z_i^2) - [\mathrm{E}(Z_i)]^2 \\ &= \mathrm{E}(Z_i^2), \quad \text{since } E(Z_i) = 0\,. \end{aligned}$$

Hence, $\mathrm{E}(Z_i^2) = 1$ $(i = 1, \ldots, n)$ and so

$$\mathrm{E}[Y] = \mathrm{E}\left[\sum_{i=1}^{k} Z_i^2\right] = \sum_{i=1}^{k} \mathrm{E}(Z_i^2) = k\,.$$

Suppose now the random variables $X_1, \ldots, X_n$ are a random sample from the $N(\mu, \sigma^2)$ distribution. We have that

$$\frac{X_i - \mu}{\sigma} \sim N(0,1)\,, \quad i = 1, \ldots, n\,,$$

so that

$$\sum_{i=1}^{n} \left[\frac{(X_i - \mu)}{\sigma}\right]^2 \sim \chi^2(n)\,.$$

If we modify the above by replacing the population mean $\mu$ by the sample estimate $\overline{X}$, the distribution changes and we obtain the following result.

**Theorem.** If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, then

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left[\frac{(X_i - \overline{X})}{\sigma}\right]^2 \sim \chi^2(n-1)\,.$$

(Proof of this result is outside the scope of the course).

By replacing $\mu$ with $\overline{X}$, the $\chi^2$ distribution of the sum of squares has lost one degree of freedom. This is because there is a single linear constraint on the variables $(X_i - \overline{X})/\sigma$, namely $\sum_{i=1}^{n}(X_i - \overline{X})/\sigma = 0$. Thus we are only summing $n-1$ independent sums of squares. **Important fact:** $\overline{X}$ and $S^2$ are independent random variables.

**Example 4.** Let $X_1, \ldots, X_{40}$ be a random sample of size $n = 40$ from the $N(25, 4^2)$ distribution. Find the probability that the sample variance, $S^2$, exceeds 20.

**Solution**. We need to calculate

$$\begin{aligned}
\mathrm{P}\left(S^2 > 20\right) &= \mathrm{P}\left(\frac{39 \times S^2}{16} > \frac{39 \times 20}{16}\right) \\
&= \mathrm{P}(Y > 48.75) \quad \text{where } Y \sim \chi^2(39) \\
&= 1 - \mathrm{P}(Y < 48.75) = 1 - 0.8638 = 0.1362\,,
\end{aligned}$$

where the probability calculation has been carried out using the `pchisq` command in R:

```
> 1-pchisq(q=48.75, df=39)
[1] 0.1362011
```

# 5 Point estimation

## 5.1 Introduction

The objective of a statistical analysis is to make inferences about a population based on a sample. Usually we begin by assuming that the data were generated by a probability model for the population. Such a model will typically contain one or more parameters $\theta$ whose value is unknown. The value of $\theta$ needs to be estimated using the sample data. For example, in previous chapters we have used the sample mean to estimate the population mean, and the sample proportion to estimate the population proportion.

A given estimation procedure will typically yield different results for different samples, thus under random sampling from the population the result of the estimation will be a random variable with its own sampling

distribution. In this chapter, we will discuss further the properties that we would like an estimation procedure to have. We begin to answer questions such as:

- Is my estimation procedure a good one or not?
- What properties would we like the sampling distribution to have?

## 5.2 General framework

Let $X_1, \ldots, X_n$ be a random sample from a distribution with c.d.f. $F_X(x;\theta)$, where $\theta$ is a parameter whose value is unknown. A **(point) estimator** of $\theta$, denoted by $\widehat{\theta}$ is a real, single-valued function of the sample, i.e.

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

As we have seen already, because the $X_i$ are random variables, the estimator $\widehat{\theta}$ is also a random variable whose probability distribution is called its sampling distribution.

The value $\widehat{\theta} = h(x_1, \ldots, x_n)$ assumed for a particular sample $x_1, \ldots, x_n$ of observed data is called a **(point) estimate** of $\theta$. Note the point estimate will almost never be exactly equal to the true value of $\theta$, because of sampling error.

Often $\theta$ may in fact be a vector of $p$ scalar parameters. In this case, we require $p$ separate estimators for each of the components of $\theta$. For example, the normal distribution has two scalar parameters $\mu$ and $\sigma^2$. These could be combined into a single parameter vector, $\theta = (\mu, \sigma^2)$, for which one possible estimator is $\widehat{\theta} = (\overline{X}, S^2)$.

## 5.3 Properties of estimators

We would like an estimator $\widehat{\theta}$ of $\theta$ to be such that:

(i) the sampling distribution of $\widehat{\theta}$ is centered about the target parameter, $\theta$.

(ii) the spread of the sampling distribution of $\widehat{\theta}$ is small.

If an estimator has properties (i) and (ii) above then we can expect estimates resulting from statistical experiments to be close to the true value of the population parameter we are trying to estimate.

We now define some mathematical concepts formalizing these notions. The **bias** of a point estimator $\widehat{\theta}$ is $\text{bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta$. The estimator is said to be **unbiased** if

$$\text{E}(\widehat{\theta}) = \theta\,,$$

i.e. if $\text{bias}(\widehat{\theta}) = 0$. Unbiasedness corresponds to property (i) above, and is generally seen as a desirable property for an estimator. Note that sometimes biased estimators can be modified to obtain unbiased estimators. For example, if $\text{E}(\widehat{\theta}) = k\theta$, where $k \neq 1$ a constant, then $\text{bias}(\widehat{\theta}) = (k-1)\theta$. However, $\widehat{\theta}/k$ is an unbiased estimator of $\theta$.

The spread of the sampling distribution can be measured by $\text{Var}(\widehat{\theta})$. In this context, the standard deviation of $\widehat{\theta}$, i.e. $\sqrt{\text{Var}(\widehat{\theta})}$, is called the **standard error**. Suppose that we have two different unbiased estimators of $\theta$, called $\widehat{\theta}_1$ and $\widehat{\theta}_2$, which are both based on samples of size $n$. By principle (ii) above, we would prefer to use the estimator with the smallest variance, i.e. choose $\widehat{\theta}_1$ if $\text{Var}(\widehat{\theta}_1) < \text{Var}(\widehat{\theta}_2)$, otherwise choose $\widehat{\theta}_2$.

**Example 5.** Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution where $\sigma^2$ is assumed **known**. Recall that the $X_i \sim N(\mu, \sigma^2)$ independently in this case. We can estimate $\mu$ by the sample mean, i.e.

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

We have already seen that $\mathrm{E}(\overline{X}) = \mu$, thus $\mathrm{bias}(\overline{X}) = 0$. Moreover, $\mathrm{Var}(\overline{X}) = \sigma^2/n$. Note that $\mathrm{Var}(\overline{X}) \to 0$ as $n \to \infty$. Thus, as the sample size increases, the sampling distribution of $\overline{X}$ becomes more concentrated about the true parameter value $\mu$. The standard error of $\overline{X}$ is

$$\mathrm{s.e.}(\overline{X}) = \sqrt{\mathrm{Var}(\overline{X})} = \frac{\sigma}{\sqrt{n}} \,.$$

Note that if $\sigma^2$ were in fact unknown, then this standard error would also need to be estimated from the data, via

$$\widehat{\mathrm{s.e.}}(\overline{X}) = \frac{s}{\sqrt{n}} \,.$$

Importantly, the results $\mathrm{E}(\overline{X}) = \mu$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$ also hold if $X_1, \ldots, X_n$ are sampled independently from any continuous or discrete distribution with mean $\mu$ and variance $\sigma^2$. Thus the sample mean is always an unbiased estimator of the population mean.

**Example 6.** Suppose now that $n = 5$, $X_1, \ldots, X_5 \sim N(\mu, \sigma^2)$, and an alternative estimator of $\mu$ is given by

$$\widetilde{\mu} = \frac{1}{9}X_1 + \frac{2}{9}X_2 + \frac{3}{9}X_3 + \frac{2}{9}X_4 + \frac{1}{9}X_5 \,.$$

We have that

$$\mathrm{E}[\widetilde{\mu}] = \frac{\mu}{9} + \frac{2\mu}{9} + \frac{3\mu}{9} + \frac{2\mu}{9} + \frac{\mu}{9} = \mu \,,$$

and

$$\mathrm{Var}[\widetilde{\mu}] = \frac{\sigma^2}{81} + \frac{4\sigma^2}{81} + \frac{9\sigma^2}{81} + \frac{4\sigma^2}{81} + \frac{\sigma^2}{81} = \frac{19\sigma^2}{81} \,.$$

Thus, $\widetilde{\mu}$ is an unbiased estimator of $\mu$ with variance $\frac{19\sigma^2}{81}$. The sample mean $\widehat{\mu} = \overline{X}$ is also unbiased for $\mu$ and has variance $\frac{\sigma^2}{5}$.

The two estimators $\widehat{\mu}$ and $\widetilde{\mu}$ both have normal sampling distributions centered on $\mu$ but the variance of the sampling distribution of $\widehat{\mu}$ is smaller than that of $\widetilde{\mu}$ because $\frac{\sigma^2}{5} < \frac{19\sigma^2}{81}$. Hence, in practice, we would prefer to use $\widehat{\mu}$.

**Example 7.** Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution where now both $\mu$ and $\sigma^2$ are assumed to be **unknown**. We can use $\overline{X}$ as an estimator of $\mu$ and $S^2$ as an estimator of $\sigma^2$. We have already seen that

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

is an unbiased estimator of $\sigma^2$, i.e. $\mathrm{E}[S^2] = \sigma^2$ and $\mathrm{bias}(S^2) = E[S^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0$.

If we instead consider the estimator

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \,,$$

we see that $\mathrm{E}[\widetilde{\sigma}^2] = \frac{(n-1)}{n}\sigma^2$. Thus $\widetilde{\sigma}^2$ is a biased estimator of $\sigma^2$ with bias $-\sigma^2/n$. Notice that $\mathrm{bias}(\widetilde{\sigma^2}) \to 0$

as $n \to \infty$. We say that $\widetilde{\sigma}^2$ is asymptotically unbiased. It is common practice to use $S^2$, with the denominator $n-1$ rather than $n$. This results in an unbiased estimator of $\sigma^2$ for all values of $n$.

Exactly the same argument as above could also be made for using $S^2$ as an estimator of the variance of the population distribution if the data were from another, non-normal, continuous distribution or even a discrete distribution. The only prerequisite is that $\sigma^2$ is finite in the population distribution. Therefore, calculations of the sample variance for any set of data should always be based on using divisor $(n-1)$.

**Example 8.** Let $X_1, \ldots, X_n$ be a random sample of Bernoulli random variables with parameter $p$ which is **unknown**. Thus, $X_i \sim \text{Bi}(1, p)$ for $i = 1, \ldots, n$ so that $\text{E}(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$, $i = 1, \ldots, n$.

If we consider estimating $p$ by the proportion of 'successes' in the sample then we have

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

so that

$$\begin{aligned} \text{E}(\widehat{p}) &= \frac{1}{n} \sum_{i=1}^{n} \text{E}(X_i) \\ &= \frac{1}{n} np \,, \end{aligned}$$

thus $\text{E}(\widehat{p}) = p$. Also,

$$\begin{aligned} \text{Var}(\widehat{p}) &= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) \quad \text{by independence} \\ &= \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n} \,, \end{aligned}$$

Hence, $\widehat{p}$ is an unbiased estimator of $p$ with variance $p(1-p)/n$. Notice that the variance of this estimator also tends towards zero as $n$ gets larger.

**Example 9.** Let $X_1, \ldots, X_n$ be a random sample from a $U[\theta, \theta+1]$ distribution where $\theta$ is **unknown**. Thus, the data are uniformly distributed on a unit interval but the location of that interval is unknown. Consider using the estimator $\widehat{\theta} = \overline{X}$.

Now,

$$\begin{aligned} \text{E}(\overline{X}) &= \frac{\theta + (\theta+1)}{2} \\ &= \frac{2\theta + 1}{2} \\ &= \theta + \frac{1}{2} \end{aligned}$$

Therefore, $\text{bias}(\overline{X}) = \theta + 1/2 - \theta = 1/2$ while $\text{Var}(\overline{X}) = \frac{1}{12n}$. However, if we instead define $\widehat{\theta} = \overline{X} - 1/2$ then $E(\widehat{\theta}) = \theta$ and $\text{Var}(\widehat{\theta}) = \frac{1}{12n}$.

### 5.3.1 Summary of point estimation

The key ingredients are:

- A probability model for the data.

- Unknown model parameter(s) to be estimated.
- An estimation procedure, or estimator.
- The sampling distribution of the estimator.

The main points are:

- Application of the estimation procedure, or estimator, to a particular observed data set results in an estimate of the unknown value of the parameter. The estimate will be different for different random data sets.
- The properties of the sampling distribution (bias, variance) tell us how good our estimator is, and hence how good our estimate is likely to be.
- Estimation procedures can occasionally give poor estimates due to random sampling error. For good estimators, the probability of obtaining a poor estimate is lower.

# 6 Likelihood for discrete data

## 6.1 The likelihood function

The parameter estimators we have considered so far have mostly been motivated by intuition. For example, the sample mean $\overline{X}$ is an intuitive estimator of the population mean. However in many situations, it is not obvious how to define an appropriate estimator for the parameter(s) of interest.

One method for deriving an estimator, which works for almost any parameter of interest, is the method of **maximum likelihood**. The estimators derived in this way typically have good properties. The method revolves around the **likelihood function**, which is of great importance throughout Statistics. The likelihood function is used extensively in estimation and also hypothesis testing, which we discuss in a later chapter.

Let $X_1, \ldots, X_n$ be an i.i.d. random sample from the discrete distribution with p.m.f. $p(x \,|\, \theta)$, where $\theta$ is a parameter whose value is unknown. Given observed data values $x_1, \ldots, x_n$ from this model, the likelihood function is defined as

$$L(\theta) = \mathrm{P}(X_1 = x_1,\, X_2 = x_2,\, \ldots,\, X_n = x_n \,|\, \theta)\,.$$

In other words,

> **the likelihood is the joint probability of the observed data considered as a function of the unknown parameter $\theta$.**

By independence, we can rewrite the likelihood as follows:

$$L(\theta) = p(x_1 \,|\, \theta) \times \cdots \times p(x_n \,|\, \theta)\,.$$

**Example 10.** Let $x_1, \ldots, x_n$ be a sample obtained from the Poisson($\lambda$) distribution with p.m.f.

$$p(x \,|\, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}\,, \quad x = 0, 1, 2, \ldots$$

The likelihood function for this sample is given by:

$$L(\lambda) = \prod_{i=1}^{n} p(x_i \,|\, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\,, \qquad \text{for } \lambda > 0\,.$$

## 6.2 Maximum likelihood estimation

In the discrete case, given sample data $x_1, \ldots, x_n$ the **maximum likelihood estimate** for $\theta$ is the value $\widehat{\theta}$ that maximizes the joint probability of the observed data, i.e. that maximizes the value of the likelihood function $L(\theta)$.

Maximization of $L(\theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta)$ leads to a numerical value $\widehat{\theta}$ for the estimate of $\theta$. The value of $\widehat{\theta}$ depends on the observed sample values $x_1, \ldots, x_n$, i.e. $\widehat{\theta}$ is a function of the data,

$$\widehat{\theta} = h(x_1, \ldots, x_n)\,.$$

We can also consider $\widehat{\theta}$ as a function of the random sample, $X_1, \ldots, X_n$,

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

in which case $\widehat{\theta}$ is a random variable called the **maximum likelihood estimator**. The maximum likelihood estimator possesses its own sampling distribution, which will be studied in later Statistics modules.

In simple cases, the maximum likelihood estimate can be found by standard calculus techniques, i.e. by solving

$$\frac{dL(\theta)}{d\theta} = 0\,. \tag{3}$$

However, it is usually much easier algebraically to find the maximum of the log-likelihood $l(\theta) = \log L(\theta)$ because for i.i.d. data,

$$\log L(\theta) = \log \left[ \prod_{i=1}^{n} p(x_i \,|\, \theta) \right] = \sum_{i=1}^{n} \log p(x_i \,|\, \theta)\,.$$

Hence, the log likelihood is additive as opposed to the likelihood which is multiplicative. This is advantageous because it is far easier to differentiate a sum of functions than to differentiate a product of functions.

To find the value of $\theta$ that maximizes $l(\theta)$ we instead find $\widehat{\theta}$ that solves:

$$\frac{dl(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{d \log p(x_i \,|\, \theta)}{d\theta} = 0\,. \tag{4}$$

The solution is a maximum if $\frac{d^2 l(\theta)}{d\theta^2} < 0$ at $\theta = \widehat{\theta}$. The estimate found by this method, i.e. by maximizing the log-likelihood, is identical to the one found by maximizing the likelihood directly, because the logarithm is a monotonically increasing function.

**Example 11.** Let $X_1, \ldots, X_n$ be a random sample from the Poisson($\lambda$) distribution. Find the maximum likelihood estimator of $\lambda$.

We have seen that

$$L(\lambda) = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}\,,$$

so that

$$l(\lambda) = -n\lambda + \left( \sum_{i=1}^{n} X_i \right) \log \lambda - \log \left( \prod_{i=1}^{n} X_i! \right).$$

Solving $\frac{dl(\lambda)}{d\lambda} = 0$, we obtain

$$\left. \frac{dl}{d\lambda} \right|_{\lambda = \widehat{\lambda}} = -n + \frac{\sum_{i=1}^{n} X_i}{\widehat{\lambda}} = 0\,, \qquad \text{which implies that } \widehat{\lambda} = \overline{X}\,.$$

Checking the second derivatives, we see that

$$\left.\frac{d^2l}{d\lambda^2}\right|_{\lambda=\widehat{\lambda}} = \frac{-\sum_{i=1}^n X_i}{\widehat{\lambda}^2} = \frac{-n}{\overline{X}} < 0\,.$$

Therefore, $\widehat{\lambda} = \overline{X}$ is indeed the maximum likelihood estimator of $\lambda$. If we have a set of data $x_1, \ldots, x_n$ then the maximum likelihood estimate of $\lambda$ is $\widehat{\lambda} = \overline{x}$, the sample mean. This is an intuitively sensible estimate, as the mean of the Poisson($\lambda$) distribution is equal to $\lambda$.

**Example 12.** Let $X_1, \ldots, X_n$ be a random sample from a Bi$(1, p)$ distribution. Find the maximum likelihood estimator of $p$.

In this example then the likelihood function is

$$L(p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{(n-\sum_{i=1}^n X_i)}\,,$$

so that the log-likelihood is given by

$$l(p) = \sum_{i=1}^n X_i \log p + \left(n - \sum_{i=1}^n X_i\right)\log(1-p)\,.$$

Solving $\left.\frac{dl}{dp}\right|_{p=\widehat{p}} = 0$, we obtain

$$\left.\frac{dl}{dp}\right|_{p=\widehat{p}} = \frac{\sum_{i=1}^n X_i}{\widehat{p}} - \frac{(n-\sum_{i=1}^n X_i)}{1-\widehat{p}} = 0\,,$$

Hence, multiplying all sides by $\widehat{p}(1-\widehat{p})$,

$$\sum_{i=1}^n X_i - \widehat{p}\sum_{i=1}^n X_i - \widehat{p}n + \widehat{p}\sum_{i=1}^n X_i = 0\,,$$

and so

$$\sum_{i=1}^n X_i = n\widehat{p}\,.$$

Thus, the maximum likelihood estimator of $p$ is $\widehat{p} = \frac{\sum_{i=1}^n X_i}{n} = \overline{X}$, i.e. the sample proportion. We have previously seen that this is unbiased for $p$.

Note that it is worth checking the second derivative at $p = \widehat{p}$,

$$\begin{aligned}\left.\frac{d^2l}{dp^2}\right|_{p=\widehat{p}} &= \frac{-\sum_{i=1}^n X_i}{\widehat{p}^2} - \frac{(n-\sum_{i=1}^n X_i)}{(1-\widehat{p})^2} \\ &= -\frac{n}{\widehat{p}} - \frac{n}{(1-\widehat{p})} \\ &= -\frac{n}{\widehat{p}(1-\widehat{p})}\,,\end{aligned}$$

which is negative, and so $\widehat{p} = \overline{X}$ does indeed maximize the likelihood.

### 6.3 Poisson likelihood examples

In this section we will look at two examples of the Poisson likelihood function. The first example is based on some simulated Poisson data while the second uses data on the numbers of hourly births over a 24 hour period in an Australian hospital.

The R function written and used to compute the Poisson likelihood and log-likelihood functions is as follows:

```
pois.lik <- function(x, lmin, lmax){
   nl <- 1000
   n <- length(x)
   lval <- numeric(nl)
   pl <- numeric(nl)
   lpl <- numeric(nl)
   lval <- seq(from=lmin, to=lmax, length.out=nl)
   for(k in 1:nl){
      pl[k] <- prod(dpois(x,lambda=lval[k]))
      lpl[k] <- sum(log(dpois(x,lambda=lval[k])))
    }
   pl.res <- data.frame(lval, pl, lpl)
   return(pl.res)
}
```

The data are in the argument `x` while the minimum and maximum $\lambda$ values to be considered are passed to the function in the arguments `lmin` and `lmax`.

The function returns a data frame called `pl.res` comprising three columns. The first contains the sequence of $\lambda$ values used, the second contains the corresponding likelihood values and the third the corresponding log-likelihood values.

**Example 13.** (Simulated data). The data in this example are a random sample of $n = 30$ simulated from the $\text{Po}(\lambda = 10)$ distribution. The data are simulated via:

```
> xp <- rpois(n=30, lambda=10)
```

The following code produces the likelihood and log-likelihood functions for these data:

```
> pl.res4 <- pois.lik(xp, lmin=7, lmax=13)
> names(pl.res4)
[1] "lval" "pl"   "lpl"
```

This can be plotted as follows:

```
> plot(pl.res4$lval, pl.res4$pl, type="l",
    xlab="lambda", ylab="L(lambda)",
    main="Poisson likelihood, simulated data")
> plot(pl.res4$lval, pl.res4$lpl, type="l",
    xlab="lambda", ylab="l(lambda)",
    main="Poisson log-likelihood, simulated data")
```

This gives the following plots.

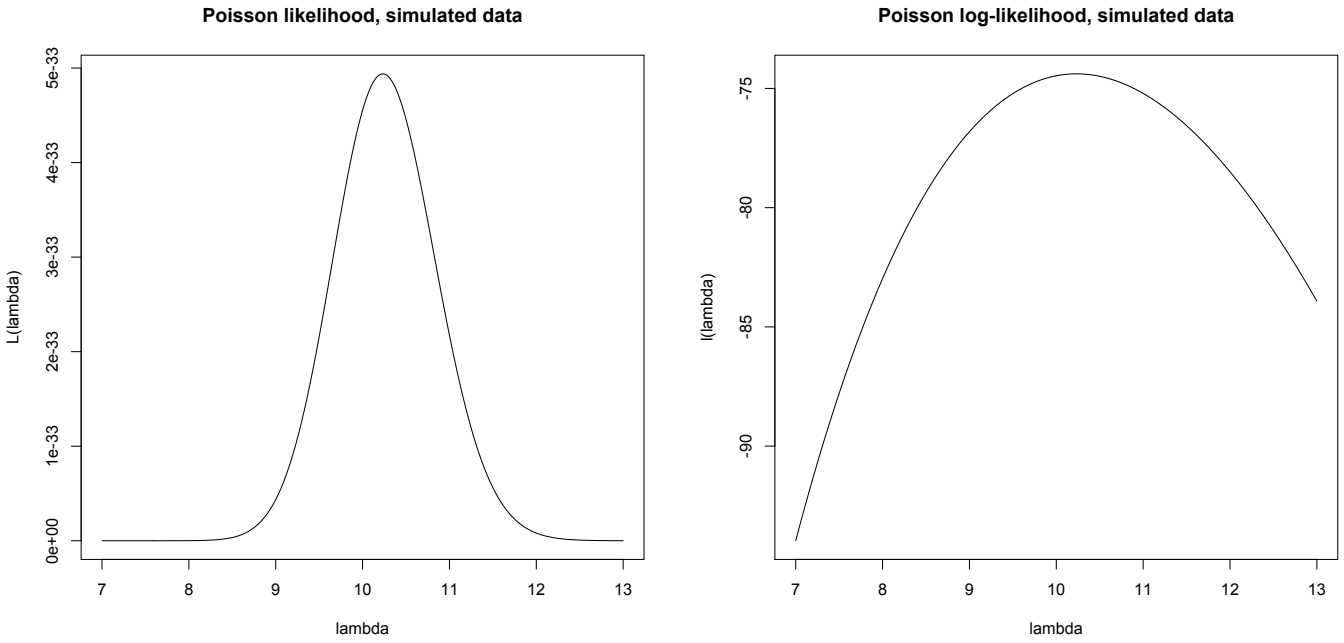


Figure 5: Likelihood (left) and log-likelihood (right) functions for the simulated Poisson data ($n = 30$, $\lambda = 10$).

The maximum likelihood estimate can be computed approximately via direct numerical maximization of the likelihood or log-likelihood:

```
> lopt1 <- pl.res4$lval[which.max(pl.res4$pl)]
> lopt1
[1] 10.23123
> lopt2 <- pl.res4$lval[which.max(pl.res4$lpl)]
> lopt2
[1] 10.23123
```

The maximum likelihood estimate of $\lambda$ from the two plots is calculated to be 10.23. We know that the maximum likelihood estimate can be determined analytically as the sample mean which is equal to 10.23.

```
> mean(xp)
[1] 10.23333
```

The reason for the slight discrepancy between the two results is the discretization error arising from the use of a discrete set of $\lambda$ values in the first method.

Please note that if you run the above code yourself, you will get slightly different results because you will have sampled a different set of data using the function `rpois`.

**Example 14.** (Australian birth data). The data give the number of births per hour over a 24-hour period on the 18 December 1997 at the Mater Mother's Hospital in Brisbane, Australia. There were a total of $n = 44$ births. At the time, this was a record number of births in one 24-hour period in this hospital. We denote the number of births in the $i$th hour by $X_i$ and fit the model

$$X_i \sim \text{Po}(\lambda)\,, \quad i = 1, \ldots, n\,,$$

with the $X_i$ assumed to be independent. The data can be read in to R as follows:

```
> birth <- read.table(file="https://minerva.it.manchester.ac.uk/~saralees/birth_freq.txt",
                                header=T)
> names(birth)
[1] "hour"   "number"

> birth
   hour number
1     1      1
2     2      3
3     3      1
4     4      0
5     5      4
6     6      0
7     7      0
8     8      2
9     9      2
10   10      1
11   11      3
12   12      1
13   13      2
14   14      1
15   15      4
16   16      1
17   17      2
18   18      1
19   19      3
20   20      4
21   21      3
22   22      2
23   23      1
24   24      2
```

The code to produce the likelihood plots is as follows:

```
> pl.res.birth <- pois.lik(birth$number, lmin=0, lmax=4)
> plot(pl.res.birth$lval, pl.res.birth$pl, type="l",
    xlab="lambda", ylab="L(lambda)", main="Poisson likelihood
      function for Australian birth data")
> plot(pl.res.birth$lval, pl.res.birth$lpl, type="l",
    xlab="lambda", ylab="l(lambda)", main="Poisson log-likelihood
      function for Australian birth data" )
```

This results in the following figures:

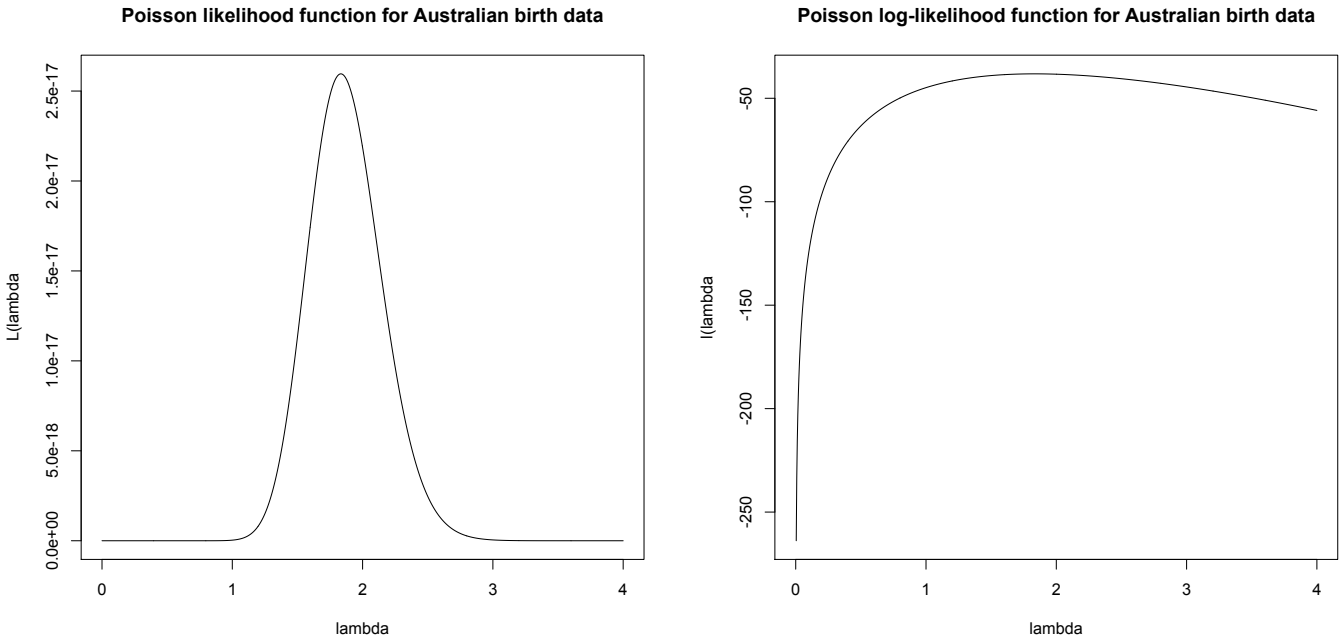


Figure 6: The likelihood (left) and log-likelihood (right) functions for the Australian births data ($n = 44$).

The maximum likelihood estimate is 1.83, which can be found by direct numerical maximization of the likelihood or log-likelihood function.

```
> lopt1 <- pl.res.birth$lval[which.max(pl.res.birth$pl)]
> lopt1
[1] 1.833834
> lopt2 <- pl.res.birth$lval[which.max(pl.res.birth$lpl)]
> lopt2
[1] 1.833834
```

The result can be compared back to the sample mean, $\overline{x}$, which gives the same result up to discretization error.

```
> mean(birth$number)
[1] 1.833333
```

# 7 Confidence intervals

## 7.1 Interval estimation

So far in this module, whenever we have fitted a probability model to a data set, we have done so by calculating point estimates of the values of any unknown parameters $\theta$. However, it is very rare for a point estimate to be exactly equal to the true parameter value. An alternative approach is to specify an interval, or range, of plausible parameter values. We would then expect the true parameter value to lie within this interval of plausible values. We call such an interval an **interval estimate** of the parameter.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an independent random sample from a distribution $F_X(x; \theta)$ with unknown parameter $\theta$. An interval estimator,

$$I(\mathbf{X}) = [l(\mathbf{X}), u(\mathbf{X})]$$

for $\theta$ is defined by two statistics, i.e. functions of the data. The statistic $u(\mathbf{X})$ defines the upper end-point of the interval, and the statistic $l(\mathbf{X})$ defines the lower end-point of the interval. We will see later how to choose appropriate statistics for the end-points.

The key property of an interval estimator for $\theta$ is its **coverage probability**. This defined as the probability that the interval contains, or 'covers', the true value of the parameter, i.e.

$$\mathrm{P}_\theta[\, l(\mathbf{X}) \leq \theta \leq u(\mathbf{X})\,],$$

or equivalently $\mathrm{P}_\theta[\, I(\mathbf{X}) \ni \theta\,]$. We use the notation $P_\theta$ for probabilities here to emphasize that the probability distributions of $l(\mathbf{X})$ and $u(\mathbf{X})$ depend on $\theta$.

Let $\alpha \in (0,1)$, and suppose that we have been able to find statistics $l$ and $u$ such that the coverage probability satisfies

$$\mathrm{P}_\theta[\, l(\mathbf{X}) \leq \theta \leq u(\mathbf{X})\,] = 1-\alpha\,, \quad \text{for all values of } \theta\,,$$

Then the interval *estimator* $I(\mathbf{X})$ and, for any particular data set $\mathbf{x} = (x_1, \ldots, x_n)$ the resulting interval *estimate* $I(\mathbf{x})$, is referred to as a $100(1-\alpha)\%$ **confidence interval for** $\theta$. The proportion $1-\alpha$ is referred to as the **confidence level**, and the interval end points $l(\mathbf{x})$, $u(\mathbf{x})$ are known as the **confidence limits**.

## 7.2 Single sample procedures

### 7.2.1 Confidence interval for the mean of a normal distribution with known variance

To illustrate the idea, let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, with $\mu$ unknown but $\sigma^2$ known. Recall that $\overline{X} \sim N(\mu, \sigma^2/n)$. Thus, if we standardize $\overline{X}$ then we obtain the random variable

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

A crucial property of $Z$ above is that the *distribution* of $Z$ does not depend on $\mu$ or $\sigma$, i.e. the right hand side of the above equation is the same no matter what the value of $\mu$ or $\sigma$.

Let $z_{1-\alpha/2}$ be such that $\mathrm{P}(Z \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha/2$. By symmetry of the normal distribution, it is also true that $\mathrm{P}(Z \leq -z_{1-\frac{\alpha}{2}}) = \alpha/2$, and furthermore $\mathrm{P}(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}) = \alpha$. We have therefore that

$$\mathrm{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha\,.$$

Moreover, the inequality inside the brackets can be rearranged to show that:

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(-\frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} - \overline{X} \leq -\mu \leq +\frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} - \overline{X}\right) \\ &= \mathrm{P}\left(\overline{X} - \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} \;\leq\; \mu \;\leq\; \overline{X} + \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\right). \end{aligned}$$

Hence, the interval estimator $I(\mathbf{X})$ for $\mu$ defined by

$$I(\mathbf{X}) = \left[\overline{X} - \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\,,\; \overline{X} + \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\right]$$

contains the true value of $\mu$ with probability $1-\alpha$.

The upshot of the above discussion is that for a particular set of data values $\mathbf{x} = (x_1, \ldots, x_n)$, the interval

estimate

$$I(\mathbf{x}) = \left[ \overline{x} - \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\,,\ \overline{x} + \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} \right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu$.

We must be careful how to interpret confidence intervals. Given a particular realised data set $\mathbf{x}$ with corresponding calculated interval $I(\mathbf{x})$, it is not true to say that the parameter $\theta$ lies within $I(\mathbf{x})$ with $100(1-\alpha)\%$ probability. The value of $\theta$ is a fixed unknown, and not a random variable. Moreover, once we have observed data $\mathbf{x}$, $I(\mathbf{x})$ is also fixed and no longer a random variable. Hence either $\theta$ is in $I(\mathbf{x})$ or it is not: there are no random variables remaining about which to make probability statements.

Instead, the correct interpretation is that *before the experiment* the probability that the interval estimator will ultimately contain the true value of $\theta$ is $100(1-\alpha)\%$. Alternatively, if we repeated the experiment a large number of times and calculated a confidence interval for each sample, then approximately $100(1-\alpha)\%$ of the confidence intervals would contain the true value of $\theta$.

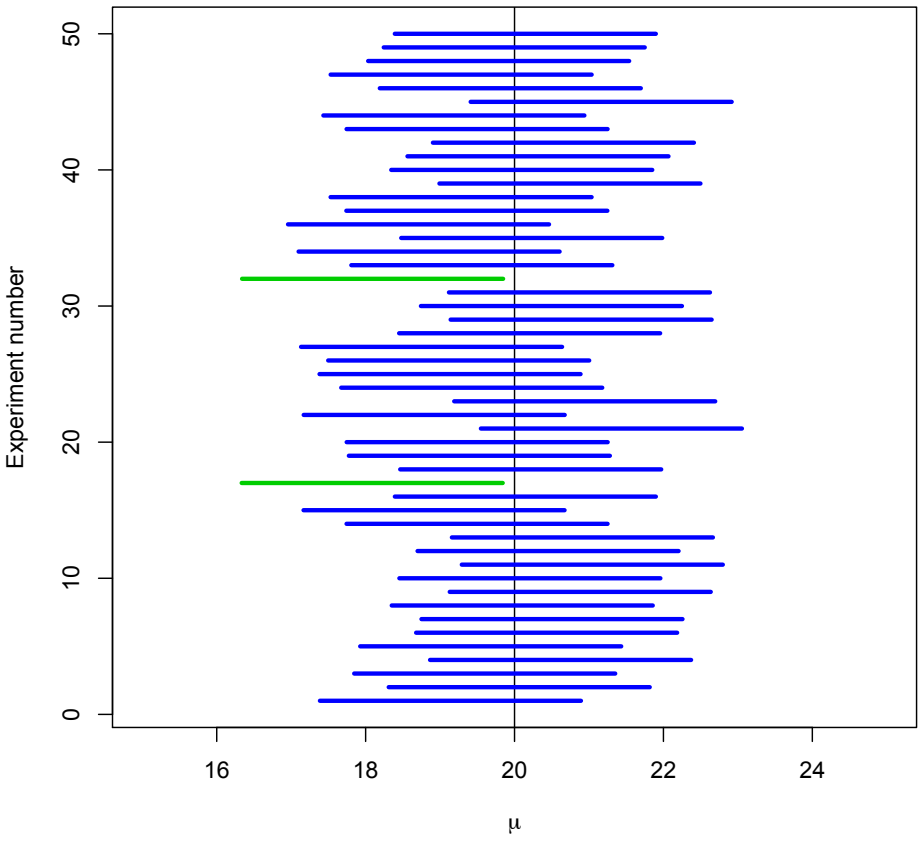


Figure 7: 95% confidence intervals computed for 50 different random samples.

In the figure above, each interval is coloured blue if it contains the true value of the parameter ($\mu = 20$) and green if it does not. The interval contains the true parameter value for $48/50 = 95\%$ of the samples.

**Example 15.** The following $n = 16$ observations are a random sample from a $N(\mu, 2^2)$ distribution, where $\mu$ is unknown:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 10.43 | 5.42 | 11.10 | 12.41 | 10.14 | 7.83 | 8.84 | 10.42 |
| 10.44 | 9.65 | 10.36 | 11.48 | 9.33 | 6.81 | 10.55 | 10.41 |

We want to use the data to construct a 95% confidence interval for $\mu$, i.e. here $\alpha = 0.05$. The sample mean is $\overline{x} = 9.73$ and $z_{1-\alpha/2} = z_{0.975} = 1.96$ so that the end-points of the 95% CI for $\mu$ are given by:

$$9.73 \pm 1.96 \times \sqrt{\frac{4.0}{16}}\,,$$

i.e. the interval is $(8.75, 10.71)$. These data were actually sampled (simulated) from a $N(10, 2^2)$ distribution. Thus the true value $\mu = 10$ is within the CI.

### 7.2.2 Confidence interval for the mean of a normal distribution, variance unknown

Suppose now that $X_1, \ldots, X_n$ are independent draws from a $N(\mu, \sigma^2)$ distribution where both $\mu$ and $\sigma^2$ are unknown. It is no longer possible to use the confidence interval $\left[\overline{x} - \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}},\, \overline{x} + \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\right]$, because $\sigma$ is unknown.

Instead of basing a confidence interval on the random variable

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

we plug in an estimate of the sample variance in the denominator, namely the sample variance (with divisor $n-1$), to obtain

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

Now, because both $\overline{X}$ and $S$ are random variables the distribution of $T$ is *not* $N(0,1)$. The fact that $S$ is also random induces extra variability into the distribution of $T$. Thus, for a given value of $n$, the distribution of $T$ has a longer tail than that of $Z$.

### 7.2.3 Student's t-distribution

We can show that the exact distribution of $T$ above is a Student's t-distribution with $(n-1)$ degrees of freedom, denoted $t(n-1)$ [or sometimes $t_{n-1}$ in the literature].

In general, if the random variable $T$ has a $t$-distribution with $\nu$ degrees of freedom then its probability density function is given by:

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

for $\nu > 0$ and $-\infty < x < \infty$. We have that $\mathrm{E}(T) = 0$ and $\mathrm{Var}(T) = \nu/(\nu-2)$, for $\nu > 2$. Moreover, the distribution is symmetric about the origin. As the parameter $\nu \to \infty$, the p.d.f. of $T$ approaches that of the $N(0,1)$ distribution.

As an exercise, produce a plot in R of the p.d.f. of the $N(0,1)$ distribution, together with the p.d.f.s of the $t(5)$ and $t(20)$ distributions. Use the `dt` function to compute the value of the $t$ p.d.f. for a given set of $x$-values.

Define $t_{1-\frac{\alpha}{2}}$ to be $1-\alpha/2$ point of the $t(n-1)$ distribution, i.e. if $T \sim t(n-1)$ then $\mathrm{P}(T \geq t_{1-\frac{\alpha}{2}}) = \alpha/2$. Then from the preceding discussion it follows that the random interval

$$I(\mathbf{X}) = \left[\overline{X} - \frac{t_{1-\frac{\alpha}{2}}\,S}{\sqrt{n}}\,,\; \overline{X} + \frac{t_{1-\frac{\alpha}{2}}\,S}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu$.

**Example 16.** Recall the electronic component failure time data introduced in Chapter 3. There are $n = 50$ observations and we found that $\overline{x} = 334.59$ and $s^2 = 15.288$. In Chapter 3 we saw that a normal distribution

with mean and variance equal to the sample values provides a good probability model for the data. As we do not know the true value of $\sigma^2$, we use the critical value $t_{0.975} = 2.0096$ for the $t(49)$ distribution. The 95% CI for $\mu$ has end-points:

$$334.59 \pm 2.0096 \times \sqrt{\frac{15.288}{50}},$$

i.e. $I(\mathbf{x}) = (333.48, 335.70)$ which gives a range of plausible values for $\mu$.

### 7.2.4 Confidence interval for the unknown mean of a non-normal distribution with either known or unknown variance

Suppose that we now have a 'large' random sample from a non-normal distribution, and that we wish to use the data to construct a confidence interval for the unknown distribution mean $\mu$. We can appeal to the central limit theorem and construct a $100(1-\alpha)\%$ CI as follows.

If the variance $\sigma^2$ is known then, by the central limit theorem, for large $n$ the statistic

$$Z_1 = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately distributed as $N(0,1)$. Thus an *approximate* $100(1-\alpha)\%$ confidence interval for $\mu$ is given by

$$\left[ \overline{X} - \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}},\ \overline{X} + \frac{z_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} \right].$$

If the variance is unknown, then we instead plug in the sample standard deviation $S$ for $\sigma$ to obtain the statistic

$$Z_2 = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

It can also be shown that $Z_2$ is also distributed approximately as $N(0,1)$ for large $n$. Thus an *approximate* $100(1-\alpha)\%$ confidence interval for $\mu$ is given by

$$\left[ \overline{X} - \frac{z_{1-\frac{\alpha}{2}}\,S}{\sqrt{n}},\ \overline{X} + \frac{z_{1-\frac{\alpha}{2}}\,S}{\sqrt{n}} \right].$$

**Example 17.** Recall the Manchester income data for adult males which we have clearly seen to be non-normally distributed. The data set contains $n = 500$ observations and we have that $\overline{x} = 33.27$ and $s^2 = 503.554$. By the above discussion, the end points

$$33.27 \pm 1.96 \times \sqrt{\frac{503.554}{500}}$$

define a 95% confidence interval for $\mu$, namely (31.30, 35.24). This gives a range of plausible values for the unknown value of $\mu$.

### 7.2.5 Confidence interval for the unknown variance of a normal distribution, mean also unknown

Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution where both $\mu$ and $\sigma^2$ are unknown. We would like to construct a $100(1-\alpha)\%$ confidence interval for $\sigma^2$.

We know that

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

is an unbiased estimator of $\sigma^2$. Also, we have the distributional result that

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

It then follows that

$$\mathrm{P}\left(\chi^2_{\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2}}\right) = 1-\alpha\,,$$

where $\chi^2_{1-\frac{\alpha}{2}}$ denotes the $(1-\alpha/2)$ point of a $\chi^2(n-1)$ distribution, i.e. if $Y \sim \chi^2(n-1)$ then $\mathrm{P}(Y \leq \chi^2_{1-\frac{\alpha}{2}}) = 1-\alpha/2$. We can re-arrange the inequalities to give bounds for the parameter $\sigma^2$, as follows

$$\mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}\right) = 1-\alpha\,.$$

Hence the $100(1-\alpha)\%$ confidence interval for $\sigma^2$, based on a sample of size $n$ from a normal population is given by

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\,,\ \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}\right].$$

The inference is that this random interval contains the true value of $\sigma^2$ with probability $1-\alpha$. A $100(1-\alpha)\%$ confidence interval for $\sigma$ can be obtained by taking the square roots of the confidence limits for $\sigma^2$.

**Example 18.** (Component lifetime data.) For these data $n = 50$ and $s^2 = 15.288$ so that a 95% confidence interval for $\sigma^2$, assuming normality, is given by

$$\left(\frac{49 \times 15.288}{\chi^2_{0.975}},\ \frac{49 \times 15.288}{\chi^2_{0.025}}\right),$$

where the $\chi^2$ values correspond to a $\chi^2$ distribution with 49 degrees of freedom. From tables of the $\chi^2(49)$ distribution we have $\chi^2_{0.025} = 31.5549$ and $\chi^2_{0.975} = 70.2224$ so that the required confidence interval is given by

$$\left(\frac{49 \times 15.288}{70.2224},\ \frac{49 \times 15.288}{31.5549}\right) = (10.668, 23.740)\,.$$

A 95% confidence interval for $\sigma$ is obtained by taking the square roots of these endpoints to give $(3.910, 4.872)$.

### 7.2.6 Confidence interval for an unknown population proportion

Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{Bi}(1, p)$, i.e. the Bernoulli distribution, where the value of $p$ is unknown. We have already seen that the estimator $\widehat{p} = \overline{X}$ is an unbiased estimator of $p$ with variance $p(1-p)/n$. By the central limit theorem, $\widehat{p} \sim N(p, p(1-p)/n)$ approximately for large $n$. Thus, for large $n$,

$$\mathrm{P}\left(-z_{1-\alpha/2} \leq \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha\,, \tag{5}$$

In fact it can also be shown that the above remains true even if $\sqrt{\mathrm{Var}\,\widehat{p}}$ in the denominator is estimated via $\sqrt{\widehat{p}(1-\widehat{p})/n}$, i.e. for large $n$,

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

Hence we have that for large $n$

$$\mathrm{P}\left(\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \approx 1-\alpha\,.$$

It then follows that

$$\left[\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\,,\ \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for the parameter $p$.

**Example 19.** Recall the opinion poll data collected from $n = 1000$ voters introduced in Chapter 1. We would like to use these data to obtain a 95% CI for the proportion in the population who support Labour, denoted by $p_L$. The proportion in the sample supporting Labour was found to be 0.314 which is our sample estimate of $p_L$, i.e. $\widehat{p}_L = 0.314$. From the above, our 95% CI has end points

$$0.314 \pm 1.96 \times \sqrt{\frac{0.314 \times 0.686}{1000}}\,,$$

i.e. the interval is $(0.285, 0.343)$.

Instead of substituting an estimate of $\sqrt{\mathrm{Var}(\widehat{p})}$ in the denominator of (5), we could adopt an alternative, more conservative approach. The value of $p$ which maximizes the function $p(1-p)$ for $0 < p < 1$ is 0.5. Thus, in our sample $\mathrm{Var}(\widehat{p}_L) = p(1-p)/n \leq 0.5 \times 1000 = 0.00025$. Using this value in the CI gives end points

$$0.314 \pm 1.96 \times \sqrt{0.00025} = 0.314 \pm 0.03\,,$$

i.e. the interval $(0.283, 0.345)$, which is a little wider than before. It is this approach which gives rise to the frequent comment that the proportions found in a poll based on 1000 voters are accurate to plus or minus 3%.

# 8 Hypothesis testing (Part I)

## 8.1 Introduction

As we have discussed earlier in the module, one of the main aims of a statistical analysis is to make inferences about the unknown values of population parameters based on a sample of data from the population. We previously considered both point and interval estimation of such parameters. Here we instead explore how to test hypotheses about the values of parameters.

**A statistical hypothesis** is a conjecture or proposition regarding the distribution of one or more random variables. In order to specify a statistical hypothesis we need to specify the family of the underlying distribution (e.g. normal, Poisson, or binomial) as well as the set of possible values of any parameters. A *simple* hypothesis specifies the distribution and the parameter values uniquely. In contrast, a *composite* hypothesis specifies several different possibilities for the distribution, most commonly corresponding to different possibilities for the parameter values.

An example of a simple hypothesis is 'the data arise from $N(5, 1^2)$'. An example of a composite hypothesis is 'the data arise from $N(\mu, 1^2)$, with $\mu > 5$'.

**The elements of a statistical test**:

(i) The **null hypothesis**, denoted by $H_0$, is the hypothesis to be tested. This is usually a 'conservative' or 'skeptical' hypothesis that we believe by default unless there is significant evidence to the contrary.

(ii) The **alternative hypothesis**, denoted by $H_1$, is a hypothesis about the population parameters which we will accept if there is evidence that $H_0$ should be rejected.

For example, when assessing a new medical treatment it is common for the null hypothesis to correspond to the statement that the new treatment is no better (or worse) than the old one. The alternative hypothesis would be that the new treatment is better.

In this module the null hypothesis will always be simple, while the alternative hypothesis may either be simple or composite. For example, consider the following hypotheses about the value of the mean $\mu$ of a normal distribution with known variance $\sigma^2$:

- $H_0$: $\mu = \mu_0$, where $\mu_0$ is a specific numerical value, is a simple null hypothesis.
- $H_1$: $\mu = \mu_1$ (with $\mu_1 \neq \mu_0$) is a simple alternative hypothesis.
- $H_1$: $\mu > \mu_0$ is a *one-sided* composite alternative hypothesis.
- $H_1$: $\mu < \mu_0$ is a *one-sided* composite alternative hypothesis.
- $H_1$: $\mu \neq \mu_0$ is a *two-sided* composite alternative hypothesis.

**How do we use the sample data to decide between $H_0$ and $H_1$?**

(iii) **Test statistic**. This is a function of the sample data whose value we will use to decide whether or not to reject $H_0$ in favour of $H_1$. Clearly, the test statistic will be a random variable.

(iv) **Acceptance and rejection regions**. We consider the set of all possible values that the test statistic may take, i.e. the range space of the statistic, and we examine the distribution of the test statistic under the assumption that $H_0$ is true. The range space is then divided into two disjoint subsets called the *acceptance region* and *rejection region*.

On observing data, if the calculated value of the test statistic falls into the rejection region then we reject $H_0$ in favour of $H_1$. If the value of the test statistic falls in the acceptance region then we do not reject $H_0$.

The rejection region is usually defined to be a set of extreme values of the test statistic which together have low probability of occuring if $H_0$ is true. Thus, if we observe such a value then this is taken as evidence that $H_0$ is in fact false.

(v) **Type I and type II errors**. The procedure described in (iv) above can lead to two types of possible errors:

(a) Type I error - this occurs if we reject $H_0$ when it is in fact true.

(b) Type II error - this occurs if we fail to reject $H_0$ when it is in fact false.

The probability of making a type I error is denoted by $\alpha$ and is also called the **significance level** or **size** of the test. The value of $\alpha$ is usually specified in advance; the rejection region is chosen in order to achieve this value. A common choice is $\alpha = 0.05$. Note that $\alpha = \mathrm{P}(\text{reject } H_0 \,|\, H_0)$.

The probability of making a type II error is $\beta = \mathrm{P}(\text{do not reject } H_0 \,|\, H_1)$. For a good testing procedure, $\beta$ should be small for all values of the parameter included in $H_1$.

**Example 20. Is a die biased or not?** It is claimed that a particular die used in a game is biased in favour of the six. To test this claim the die is rolled 60 times, and each time it is recorded whether or not a six is obtained. At the end of the experiment the total number of sixes is counted, and this information is used to decide whether or not the die is biased.

The null hypothesis to be tested is that the die is fair, i.e. P(rolling a six) = 1/6. The alternative hypothesis is that the die is biased in favour of the six so that P(rolling a six) > 1/6. Let the probability of rolling a six be denoted by $p$. We can write the above hypotheses as:

$$H_0 : p = 1/6$$
$$H_1 : p > 1/6\,.$$

Let $X$ denote the number of sixes thrown in 60 attempts. If $H_0$ is true then $X \sim \text{Bi}(60, 1/6)$, whereas if $H_1$ is true then $X \sim \text{Bi}(60, p)$, with $p > 1/6$. $H_0$ is a simple hypothesis, whereas $H_1$ is a composite hypothesis.

If $H_0$ were true, we would expect to see 10 sixes, since $\text{E}(X) = 10$ under $H_0$. However, the actual number observed will vary randomly around this value. If we observe a large number of sixes, then this will constitute evidence against $H_0$ in favour of $H_1$. The question is, how large does the number of sixes need to be so that we should reject $H_0$ in favour of $H_1$?

The test statistic here is $x$ and the rejection region is

$$\{x : x > k\}\,,$$

for some $k \in \mathbb{N}$. Above, we choose the smallest value of $k$ that ensures a significance level $\alpha < 0.05$, i.e. the smallest $k$ such that

$$\alpha = \text{P}(X > k \,|\, H_0) < 0.05\,.$$

Note that for $k = 14$, $\text{P}(X > k \,|\, H_0) = 0.0648$, while for $k = 15$, $\text{P}(X > k \,|\, H_0) = 0.0338$. Thus we select $k = 15$. In this case, the actual significance level of the test is 0.0338.

When, as in this case, the test statistic is a discrete random variable, for many choices of significance level there is no corresponding rejection region achieving that significance level exactly (e.g. $\alpha = 0.05$ above).

In summary, under $H_0$ the probability of observing more than 15 sixes in 60 rolls is 0.0338. This event is sufficiently unlikely under $H_0$ that if it occurs then we reject $H_0$ in favour of $H_1$. It is possible that by rejecting $H_0$ we may make a type I error, with probability 0.0338 if $H_0$ is true. If 15 or fewer sixes are obtained, then this is within the acceptable bounds of random variation under $H_0$. Thus, in this case we would not reject the null hypothesis that the die is unbiased. However in making this decision we may be making a type II error, if $H_1$ is in fact true.

### 8.1.1 Probability of correctly rejecting $H_0$ when it is false

The probability of correctly rejecting $H_0$ when it is false satisfies

$$\text{P}(\text{reject } H_0 \,|\, p) = 1 - \text{P}(\text{type II error})\,.$$

Ideally we would like the probability on the left to be high. It is straightforward to evaluate this probability for particular values of $p > 1/6$. Specifically, $\text{P}(\text{reject } H_0 \,|\, p) = \text{P}(X > 15 \,|\, p)$, where $X \sim \text{Bi}(60, p)$. For example, the following values have been computed using R:

| $p$ | P(reject $H_0 \mid p$) |
|---|---|
| 0.2 | 0.1306 |
| 0.25 | 0.4312 |
| 0.3 | 0.7562 |

Clearly, the larger the true value of $p$, the more likely we are to correctly reject $H_0$.

# 9 Hypothesis testing (Part 2) Single sample procedures

## 9.1 Introduction

In this chapter we will discuss specific applications of hypothesis testing where we have a single sample of data and wish to test hypotheses regarding the value of a population mean parameter.

We focus our main discussion on the scenario in which the random sample is from a $N(\mu, \sigma^2)$ distribution with $\mu$ unknown and $\sigma^2$ known. The ideas are then extended to develop hypothesis tests for (i) the mean of a normal distribution with unknown variance, (ii) the mean of a non-normal distribution, and (iii) a population proportion $p$. In cases (ii) and (iii) it is not possible to calculate the exact distribution of the test statistic under the null hypothesis, however we can appeal to the central limit theorem to find an approximate normal distribution.

## 9.2 Inference about the mean of a normal distribution when the variance is known

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where the value of $\mu$ is *unknown* but the value of $\sigma^2$ is *known*. We would like to use the data to make inferences about the value of $\mu$ and, in particular, we wish to test the following hypotheses:

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu > \mu_0 \,.$$

The null hypothesis $H_0$ posits that the data are sampled from $N(\mu_0, \sigma^2)$. In contrast, the alternative hypothesis $H_1$ posits that the data arise from $N(\mu_1, \sigma^2)$, where $\mu_1 > \mu_0$ is an unspecified value of $\mu$. This is a one-sided test.

We know that the sample mean, $\overline{X}$, is an unbiased estimator of $\mu$. Hence, if the true value of $\mu$ is $\mu_0$, then $\mathrm{E}[\overline{X} - \mu_0] = \mu_0 - \mu_0 = 0$. In contrast, if $H_1$ is true, we would have that $\mathrm{E}[\overline{X} - \mu_0] = \mu - \mu_0 > 0$. This suggests that we should reject $H_0$ in favour of $H_1$ if $\overline{X}$ is 'significantly' larger than $\mu_0$, i.e. if $\overline{X} > k$, for some $k > \mu_0$. The question is, *how much* greater than $\mu_0$ should $\overline{x}$ be before we reject $H_0$? In other words, what value should we choose for $k$?

One way to decide this is to fix the probability of rejecting $H_0$ if $H_0$ is true, i.e. the probability of making a Type I error; the critical value $k$ can then be determined on this basis. This is equivalent to fixing the significance level of the test. Suppose that we do indeed use $\overline{X}$ as the test statistic, with rejection region

$$C = \{\overline{x} > k\} \,,$$

and suppose we wish to find $k > \mu_0$ to ensure that

$$\mathrm{P}(\text{type I error}) = \mathrm{P}(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha \,.$$

Hence we have that

$$\begin{aligned}\alpha &= \mathrm{P}(\text{reject } H_0 \mid H_0 \text{ true}) = \mathrm{P}(\overline{X} > k \mid H_0 \text{ true}) \\ &= \mathrm{P}\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{k-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= \mathrm{P}\left(Z > \frac{k-\mu_0}{\sigma/\sqrt{n}}\right),\end{aligned}$$

where $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ under $H_0$. Let $z_{1-\alpha}$ denote the $\alpha$ point of $N(0,1)$, i.e. $\mathrm{P}(Z \leq z_{1-\alpha}) = 1-\alpha$. From this we see that $z_{1-\alpha} = \frac{k-\mu_0}{\sigma/\sqrt{n}}$ and so

$$k = \mu_0 + \frac{z_{1-\alpha}\,\sigma}{\sqrt{n}}\,.$$

Thus, $H_0$ is rejected in favour of $H_1$ if the sample mean is greater than $\mu_0$ by $z_{1-\alpha}$ standard errors.

Equivalently, we reject $H_0$ in favour of $H_1$ at the $100\alpha\%$ significance level if

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}\,.$$

The standardized version of $\overline{X}$ given by $Z$ is the most frequently used form of the test statistic in this scenario. The critical value $z_{1-\alpha}$ can be obtained from standard normal tables. In hypothesis testing it is common to use $\alpha = 0.05$, and in this case $z_{0.95} = 1.645$.

Suppose now that we wish to use our sample to test the hypotheses

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

This is again a one-sided test. In this case we will reject $H_0$ in favour of $H_1$ if $\overline{X} < k$ where $k < \mu_0$. Using analogous arguments to those used above, we will reject $H_0$ in favour of $H_1$ at the $100\alpha\%$ significance level if

$$\overline{X} < \mu_0 - \frac{z_{1-\alpha}\,\sigma}{\sqrt{n}}\,,$$

or, equivalently, if

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha}\,.$$

For a test having a 5% significance level the critical value is $-z_{0.95} = -1.645$.

If in fact our interest is in testing

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0\,,$$

then we now have a two-sided test. We will reject $H_0$ in favour of $H_1$ if $\overline{X}$ is either significantly greater or significantly less than $\mu_0$, i.e. if

$$\overline{X} < k_1 \;\text{ or }\; \overline{X} > k_2\,,$$

The critical values $k_1 < \mu_0$ and $k_2 > \mu_0$ are chosen so that the significance level is equal to $\alpha$, i.e.

$$\begin{aligned}\alpha &= \mathrm{P}(\overline{X} < k_1 \text{ or } \overline{X} > k_2 \mid H_0 \text{ true}) \\ &= \mathrm{P}(\overline{X} < k_1 \mid H_0) + \mathrm{P}(\overline{X} > k_2 \mid H_0)\,.\end{aligned}$$

It seems natural to choose the values of $k_1$ and $k_2$ so that the probability of rejecting $H_0$ is split equally between

the upper and lower parts of the rejection region. In other words, we choose $k_1$ and $k_2$ such that

$$\mathrm{P}(\overline{X} < k_1 \,|\, H_0) = \mathrm{P}(\overline{X} > k_2 \,|\, H_0) = \alpha/2\,.$$

For illustration, see the figure overleaf which shows the p.d.f. of $\overline{X}$, together with the rejection region.

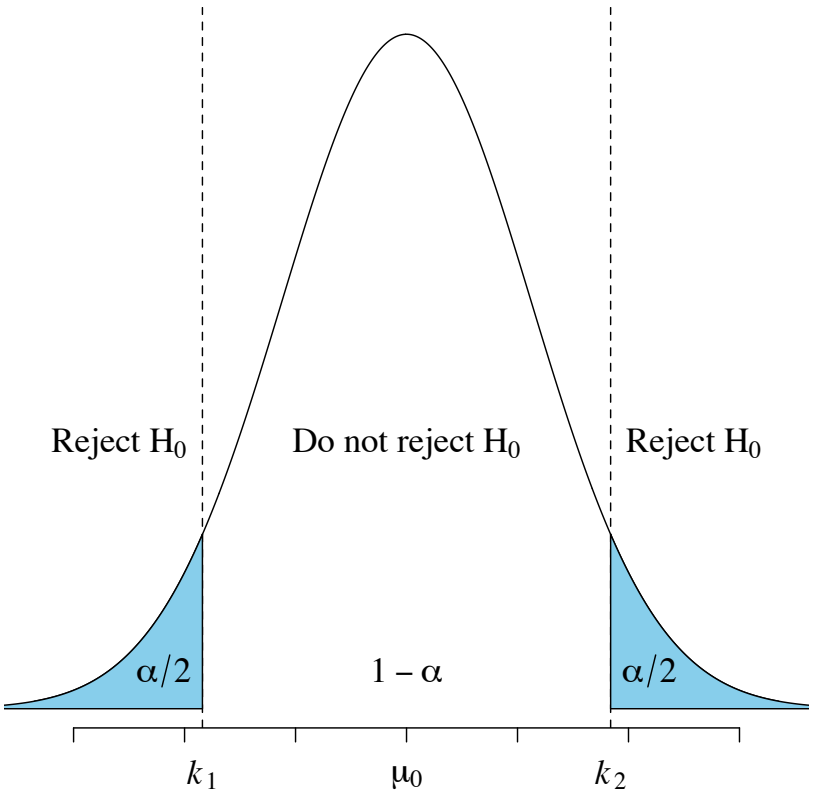


Figure 8: Illustration of a two-tailed test.

We now find appropriate values of $k_1$ and $k_2$ satisfying this property. We begin with $k_2$. Note that

$$\begin{aligned}\alpha/2 = \mathrm{P}(\overline{X} > k_2 \mid H_0 \text{ true }) &= \mathrm{P}\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{k_2-\mu_0}{\sigma/\sqrt{n}}\right)\\ &= \mathrm{P}\left(Z > \frac{k_2-\mu_0}{\sigma/\sqrt{n}}\right)\,,\ \text{with } Z \sim N(0,1)\,.\end{aligned}$$

However, we know that $z_{1-\alpha/2}$ satisfies $\mathrm{P}(Z \le z_{1-\alpha/2}) = 1-\alpha/2$. Hence,

$$\frac{k_2-\mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha/2}\,,$$

and so we have that

$$k_2 = \mu_0 + \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}}\,.$$

For $k_1$, observe that

$$\alpha/2 = \mathrm{P}(\overline{X} < k_1 \mid H_0 \text{ true }) = \mathrm{P}\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < \frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right)$$
$$= \mathrm{P}\left(Z < \frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right), \text{ with } Z \sim N(0,1).$$

We know that $\mathrm{P}(Z < -z_{1-\alpha/2}) = \alpha/2$ and so $\frac{k_1-\mu_0}{\sigma/\sqrt{n}} = -z_{1-\alpha/2}$. Hence

$$k_1 = \mu_0 - \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}}.$$

To summarize the two-tailed test here, we reject $H_0$ at significance level $\alpha$ if

$$\overline{X} > \mu_0 + \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}} \quad \text{or if}$$
$$\overline{X} < \mu_0 - \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}}.$$

Equivalently, we reject $H_0$ at significance level $\alpha$ if

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \quad \text{or if}$$
$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha/2}.$$

### 9.2.1 Connection between the two-tailed test and a confidence interval for the mean when the variance is known

Let $X_1, \ldots X_n$ be a random sample from $N(\mu, \sigma^2)$ with $\mu$ unknown and $\sigma^2$ known. Recall from Chapter 7 that a $100(1-\alpha)\%$ confidence interval for $\mu$ is given by

$$\left[\overline{X} - \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}},\ \overline{X} + \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}}\right].$$

From the preceding discussion, if we are testing the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0,$$

then we will 'accept' $H_0$ at the $100\alpha\%$ significance level if

$$\mu_0 - \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu_0 + \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{n}},$$

or, equivalently, if

$$\overline{X} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \mu_0 \leq \overline{X} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}.$$

Thus, the values of $\mu$ in the confidence interval correspond to values of $\mu_0$ for which the corresponding null hypothesis $H_0$ would not be rejected. In other words, informally, the $100(1-\alpha)\%$ confidence interval is a set of values of $\mu$ which would 'pass a hypothesis test at significance level $\alpha$'. It is in this sense that we can regard the confidence interval as a set of plausible values of $\mu$ given the data.

**Example 21.** (i) A random sample of $n = 25$ observations is taken from a normal distribution with unknown mean but known variance $\sigma^2 = 16$. The sample mean is found to be $\overline{x} = 18.2$. Test $H_0 : \mu = 20$ vs $H_1 : \mu < 20$ at the 5% significance level.

Solution: the test statistic is

$$Z = \frac{18.2 - 20.0}{\sqrt{16/25}} = -2.25$$

The appropriate 5% critical value is $-z_{0.95} = -1.645$. The observed value of $Z$ is less than $-1.645$. Hence, we reject $H_0$ at the 5% significance level and conclude that the true value of $\mu$ in the normal distribution from which the data are sampled satisfies $\mu < 20$.

(ii) Find the probability that we reject $H_0$ using this testing procedure when the true value of the mean $\mu$ is 19.0.

Solution: the null hypothesis is rejected if

$$\frac{\overline{X} - 20.0}{4/\sqrt{25}} < -1.645$$

or equivalently if

$$\overline{X} < 20.0 - 1.645 \times \frac{4}{\sqrt{25}}$$

The true distribution of $\overline{X}$ is $N(19.0, 16/25)$ and so the probability of rejecting $H_0$ is

$$\begin{aligned}
&\mathrm{P}\left(\overline{X} < 20.0 - 1.645 \times \frac{4}{\sqrt{25}}\right) \\
&= \mathrm{P}\left(\frac{\overline{X} - 19.0}{4/5} < \frac{20.0 - (1.645 \times \frac{4}{5}) - 19.0}{4/5}\right) \\
&= \mathrm{P}\left(\frac{\overline{X} - 19.0}{4/5} < -0.395\right) \\
&= \Phi(-0.395) = 0.3464\,, \\
&\qquad \text{since the true distribution of } \frac{\overline{X} - 19.0}{4/5} \text{ is } N(0,1)\,.
\end{aligned}$$

More generally, the probability of rejecting $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu < \mu_0$ can be written as

$$\Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha}\right)\,.$$

Clearly, the probability of rejecting $H_0$ will increase as the difference $\mu_0 - \mu$ becomes larger. Hence, the further the true mean from the hypothesized value, the more likely we are to reject $H_0$. When $\mu = \mu_0$ the above is the probability of rejecting $H_0$ when $H_0$ is true, i.e. the significance level. This can be verified by substituting in $\mu = \mu_0$ to obtain $\Phi(-z_{1-\alpha}) = \alpha$.

**Example 22.** Suppose now that we have a random sample of $n = 50$ observations from a normal distribution with unknown mean and known variance $\sigma^2 = 36$. It is found that $\overline{x} = 30.8$.

(i) Test $H_0 : \mu = 30$ vs $H_1 : \mu \neq 30$ at the 5% significance level.

Solution: here the test statistic is

$$Z = \frac{30.8 - 30.0}{\sqrt{36/50}} = 0.943\,.$$

As the alternative hypothesis is two-sided, we will now reject $H_0$ for either small or large values of $Z$. Using a 5% significance level the critical values are $-z_{0.975} = -1.96$ and $z_{0.975} = 1.96$. The observed value of $Z$ lies between the two critical values, thus $H_0$ is not rejected at the 5% significance level. We conclude that there is insufficient evidence to reject the claim that the normal distribution from which the data arise has mean 30.

(ii) Find the probability that we reject $H_0$ when the true value of the mean $\mu$ is 31.0.

**Solution:** here we require

$$\begin{aligned}
&1 - \mathrm{P}\left(-1.96 < \frac{\overline{X} - 30.0}{6/\sqrt{50}} < 1.96 \,\middle|\, \mu = 31.0\right) \\
&= 1 - \mathrm{P}\left(30 - 1.96 \times \frac{6}{\sqrt{50}} < \overline{X} < 30 + 1.96 \times \frac{6}{\sqrt{50}} \,\middle|\, \mu = 31\right) \\
&= 1 - \mathrm{P}\left(\frac{30 - (1.96 \times \frac{6}{\sqrt{50}}) - 31}{6/\sqrt{50}} < \frac{\overline{X} - 31}{6/\sqrt{50}} < \frac{30 + (1.96 \times \frac{6}{\sqrt{50}}) - 31}{6/\sqrt{50}}\right) \\
&= 1 - \left[\Phi\left(\frac{30 - 31}{6/\sqrt{50}} + 1.96\right) - \Phi\left(\frac{30 - 31}{6/\sqrt{50}} - 1.96\right)\right] \\
&= 1 - [\Phi(-1.179 + 1.96) - \Phi(-1.179 - 1.96)] = 0.218\,.
\end{aligned}$$

More generally, the probability of rejecting $H_0 : \mu = \mu_0$ in favour of $H_1 : \mu \neq \mu_0$ is

$$1 - \left[\Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right)\right].$$

This probability increases as $|\mu_0 - \mu|$ becomes larger. When $\mu = \mu_0$ it is equal to $\alpha$, the significance level.

## 9.3 Inference about the mean of a normal distribution when the variance is unknown

Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution, where the value of $\mu$ is *unknown* but that of $\sigma^2$ is also *unknown*. We want to test the following hypotheses:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

at significance level $\alpha$. Based on the discussion in the previous section, an appropriate test statistic which measures the discrepancy between $\mu_0$ and the sample estimator $\overline{X}$ is given by

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

where $S$ is the sample standard deviation. This is an estimate of the standardized difference between $\overline{X}$ and $\mu_0$. As we have discussed previously, because the statistic $T$ involves the random quantities $\overline{X}$ and $S$, its sampling distribution is no longer $N(0, 1)$. We have seen in Chapter 7 that $T \sim t(n-1)$, under the assumption that $H_0$ is true, i.e. $T$ has a Student $t$-distribution with $n - 1$ degrees of freedom.

Assuming that the significance level of the test is $\alpha$, we use one of the following rejection regions, depending on the alternative hypothesis:

- For the one-sided alternative hypothesis $H_1 : \mu > \mu_0$,

$$\text{reject } H_0 \text{ if } T > t_{1-\alpha}\,,$$

where $t_{1-\alpha}$ is the $1-\alpha$ point of a $t(n-1)$ distribution, i.e. $\mathrm{P}(T \leq t_{1-\alpha}) = 1-\alpha$.

- For the one-sided alternative hypothesis $H_1 : \mu < \mu_0$,

$$\text{reject } H_0 \text{ if } T < -t_{1-\alpha}\,.$$

- For the two-sided alternative hypothesis $H_1 : \mu \neq \mu_0$,

$$\text{reject } H_0 \text{ if } T < -t_{1-\alpha/2} \text{ or } T > t_{1-\alpha/2}\,.$$

**Example 23.** The drug 6-mP is used to treat leukaemia. A random sample of 21 patients using 6-mP were found to have an average remission time of $\overline{x} = 17.1$ weeks with a sample standard deviation of $s = 10.00$ weeks. A previously used drug treatment had a known mean remission time of $\mu_0 = 12.5$ weeks. Assuming that the remission times of patients taking 6-mP are normally distributed with both the mean $\mu$ and variance $\sigma^2$ being unknown, test at the 5% significance level whether the mean remission time of patients taking 6-mP is greater than $\mu_0 = 12.5$ weeks.

**Solution:** We want to test $H_0 : \mu = 12.5$ vs $H_1 : \mu > 12.5$ at the 5% significance level.

The test statistic is

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{17.1 - 12.5}{10/\sqrt{21}} = 2.108$$

Under $H_0$, $T \sim t(20)$. For a one-tailed test at the 5% significance level we will reject $H_0$ if $T > 1.725$ (from tables). Our observed value of $T$ is greater than 1.725 and so we reject the null hypothesis that $\mu = 12.5$ at the 5% significance level and conclude that $\mu > 12.5$, i.e. the drug 6-mP improves remission times compared to the previous drug treatment.

## 9.4 Using the central limit theorem

(i) **Inference about the mean of a non-normal distribution**.

Let $X_1, \ldots, X_n$ be a random sample from a non-normal distribution, where the value of the mean $\mu$ is *unknown* and that of the variance $\sigma^2$ is also *unknown*. We want to test the following hypotheses:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

at significance level $\alpha$. We can again use the test statistic

$$Y = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

defined above which, by asymptotic (large $n$) results, has an approximate $N(0,1)$ distribution when $H_0$ is true ($n \geq 30$). Aside from the choice of test statistic, the rejection regions for the various versions of $H_1$ are otherwise identical to those defined in the case of normal data with a known variance.

(ii) **Inference about the population proportion** $p$.

Let $X_1, \ldots, X_n$ be a random sample of Bi$(1, p)$ random variables, where the value of $p$ is *unknown*. We want to test the following hypotheses:

$$H_0 : p = p_0$$
$$H_1 : p > p_0$$

at significance level $\alpha$. As we have seen earlier in this module, an unbiased sample estimator of the parameter $p$ is given by

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \,.$$

By the central limit theorem, $\widehat{p} \sim N(p, p(1-p)/n)$ approximately for large $n$. As a rule of thumb, $n \geq 9\max\{p/(1-p), (1-p)/p\}$ guarantees this approximation has a good degree of accuracy. A suitable test statistic is

$$Y = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Here we have estimated the standard error of $\widehat{p}$ by $\sqrt{p_0(1-p_0)/n}$ which uses the value of $p$ specified under $H_0$. If $H_0$ is true then $Y$ has an approximate $N(0,1)$ distribution for large $n$. Thus, to achieve an approximate significance level of $\alpha$, we reject $H_0$ in favour of the above $H_1$ if $Y > z_{1-\alpha}$.

- For the one-sided alternative hypothesis $H_1 : p < p_0$, to achieve an approximate significance level of $\alpha$, we reject $H_0$ if $Y < -z_{1-\alpha}$.
- For the two-sided alternative hypothesis $H_1 : p \neq p_0$, to achieve an approximate significance level of $\alpha$, we reject $H_0$ if

$$Y < -z_{1-\alpha/2} \quad \text{or} \quad Y > z_{1-\alpha/2}\,.$$

**Example 24.** A team of eye surgeons has developed a new technique for an eye operation to restore the sight of patients blinded by a particular disease. It is known that 30% of patients who undergo an operation using the old method recover their eyesight.

A total of 225 operations are performed by surgeons in various hospitals using the new method and it is found that 88 of them are successful in that the patients recover their sight. Can we justify the claim that the new method is better than the old one? (Use a 1% level of significance).

**Solution:** Let $p$ be the probability that a patient recovers their eyesight following an operation using the new technique. We wish to test $H_0 : p = 0.30$ vs $H_1 : p > 0.30$ at the 1% significance level.

Our test statistic is

$$Y = \frac{\frac{88}{225} - 0.30}{\sqrt{\frac{0.30\times 0.70}{225}}} = 2.9823$$

As a check for the approximate normality of the distribution of $Y$ under $H_0$, we require $n > 9\max\{0.429, 2.333\} = 20.997$ which is true since $n = 225$.

The approximate 1% critical value, taken from standard normal tables, is 2.3263 which is less than the observed value of $Y$. Hence, we reject the null hypothesis at the 1% significance level and conclude that $p > 0.30$.

# 10 Hypothesis testing (Part 3)
# Procedures for two independent samples

## 10.1 Introduction

In this chapter we will extend hypothesis testing to the scenario in which there are two independent samples of data, and the aim is to make an inference about the difference in the means of the two populations from which the data have been sampled.

To this end, let $X_{11}, \ldots, X_{1n_1}$ be a random sample of size $n_1$ from a distribution with mean $\mu_1$ and variance $\sigma_1^2$. Also, let $X_{21}, \ldots, X_{2n_2}$ be a second random sample, independent from the first, from a distribution with mean $\mu_2$ and variance $\sigma_2^2$. Suppose that we wish to test

$$H_0 : \mu_1 - \mu_2 = \phi,$$

where $\phi$ is a constant (often $\phi = 0$), versus one of the following alternative hypotheses at the $100\alpha\%$ significance level:

(i) $H_1 : \mu_1 - \mu_2 > \phi$ (one-sided)

(ii) $H_1 : \mu_1 - \mu_2 < \phi$ (one-sided)

(iii) $H_1 : \mu_1 - \mu_2 \neq \phi$ (two-sided)

## 10.2 Both underlying distributions normal with known variances $\sigma_1^2$ and $\sigma_2^2$

An unbiased estimator of $\mu_1 - \mu_2 = \phi$ is given by $\overline{X}_1 - \overline{X}_2$ where

$$\overline{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}\,, \quad k = 1, 2\,.$$

This estimator satisfies

$$\mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\,.$$

We have seen in Chapter 4 that both $\overline{X}_1$ and $\overline{X}_2$ are normally distributed so their difference will also be normal. In fact

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2,\, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

and, when $H_0$ is true, $\mu_1 - \mu_2 = \phi$.

For a test statistic we will use the standardized distance between the sample estimate of $\phi$ and its hypothesized value, i.e.

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - \phi}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\,.$$

Under $H_0$, $Z \sim N(0, 1)$. We again find the critical value of our test by fixing the probability of a type I error to be $\alpha$, i.e. $\mathrm{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$. This idea was described in detail for single sample inference in Chapter 9. Below we list the rejection regions corresponding to the three possible alternative hypotheses introduced in Section 10.1.

(i) For $H_1 : \mu_1 - \mu_2 > \phi$, we reject $H_0$ at the $100\alpha\%$ significance level if $Z > z_{1-\alpha}$, where $z_{1-\alpha}$ satisfies

$\Phi(z_{1-\alpha}) = 1 - \alpha$. Equivalently, we reject $H_0$ if

$$\overline{X}_1 - \overline{X}_2 > \phi + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,.$$

E.g. if $\alpha = 0.05$ then $z_{0.95} = 1.645$.

(ii) For $H_1 : \mu_1 - \mu_2 < \phi$, we reject $H_0$ at the $100\alpha\%$ significance level if $Z < -z_{1-\alpha}$. Equivalently, we reject $H_0$ if

$$\overline{X}_1 - \overline{X}_2 < \phi - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,.$$

E.g. if $\alpha = 0.05$ then $-z_{0.95} = -1.645$.

(iii) For $H_1 : \mu_1 - \mu_2 \neq \phi$, we reject $H_0$ at the $100\alpha\%$ significance level if $|Z| > z_{1-\alpha/2}$. Equivalently, we reject $H_0$ if

$$|(\overline{X}_1 - \overline{X}_2) - \phi| > z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

E.g. if $\alpha = 0.05$ then $z_{0.975} = 1.96$.

## 10.3 Both distributions normal with unknown variances

### 10.3.1 Unequal variances (i.e. $\sigma_1^2 \neq \sigma_2^2$)

As the true values of $\sigma_1^2$ and $\sigma_2^2$ are unknown, we estimate them using the sample variances given by

$$S_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(X_{ki} - \overline{X}_k)^2, \qquad k = 1, 2\,.$$

Considering the estimated standardized difference between $\overline{X}_1 - \overline{X}_2$ and $\phi$ we have that, under $H_0$,

$$Y = \frac{\overline{X}_1 - \overline{X}_2 - \phi}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1) \text{ approximately}$$

when $n_1$ and $n_2$ are large, e.g. $n_1 > 30$ and $n_2 > 30$. To achieve an approximate significance level of $100\alpha\%$, the rejection regions for the three alternative hypotheses introduced in Section 10.1 are:

(i) For $H_1 : \mu_1 - \mu_2 > \phi$, reject $H_0$ if $Y > z_{1-\alpha}$

(ii) For $H_1 : \mu_1 - \mu_2 < \phi$, reject $H_0$ if $Y < -z_{1-\alpha}$

(iii) For $H_1 : \mu_1 - \mu_2 \neq \phi$, reject $H_0$ if $|Y| > z_{1-\frac{\alpha}{2}}$

### 10.3.2 Equal variances (i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$)

If we are prepared to assume that the unknown variances of the two normal distributions are equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the common variance $\sigma^2$ may be estimated using the estimator described in Chapter 7, i.e.

$$\widehat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\,.$$

The test statistic is then

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \phi}{\widehat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which can be shown to have a Student t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom when $H_0$ is true.

The rejection regions for the three alternative hypotheses in Section 9.1 are:

(i) For $H_1 : \mu_1 - \mu_2 > \phi$, we reject $H_0$ if $T > t_{1-\alpha}$, where $t_{1-\alpha}$ is the $1 - \alpha$ point of a $t$ distribution on $n_1 + n_2 - 2$ degrees of freedom.

(ii) For $H_1 : \mu_1 - \mu_2 < \phi$, we reject $H_0$ if $T < -t_{1-\alpha}$.

(iii) For $H_1 : \mu_1 - \mu_2 \neq \phi$, we reject $H_0$ if $|T| > t_{1-\alpha/2}$.

Each rejection region above defines a test with an exact significance level of $100\alpha\%$.

**Example 25.** An investigation was carried out comparing a new drug with a placebo. A random sample of $n_1 = 40$ patients was treated with the new drug, while an independent sample of $n_2 = 36$ patients was given the placebo. A response was measured for each patient. Under the new drug, the response had sample mean $\overline{x}_1 = 10.13$ and sample variance $s_1^2 = 4.721$. Under placebo, the response had sample mean $\overline{x}_2 = 12.16$ and sample variance $s_2^2 = 3.368$.

Supposing that the responses in both groups are normally distributed, test at the 5% significance level whether the population mean response under the new drug is the same as that under placebo. Conduct your analysis assuming that (i) $\sigma_1^2 \neq \sigma_2^2$ and (ii) $\sigma_1^2 = \sigma_2^2$.

Solution: we are required to test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$, where $\mu_1$ denotes the (population) mean response under the new drug, and $\mu_2$ denotes the (population) mean response under placebo.

(i) In the case where we assume that $\sigma_1^2 \neq \sigma_2^2$, the test statistic is

$$Y = \frac{10.13 - 12.16 - 0}{\sqrt{\frac{4.721}{40} + \frac{3.368}{36}}} = -4.413\,.$$

For a two-sided test at the approximate 5% significance level we will reject $H_0$ if $|Y| > z_{0.975} = 1.96$. The observed value of $|Y|$ is 4.413 and so we reject $H_0$ at the approximate 5% level. Hence, we conclude that the mean response for those receiving the new drug is not equal to the mean response for those receiving the placebo.

(ii) In the second case, where we assume that $\sigma_1^2 = \sigma_2^2$, we need to estimate the common variance $\sigma^2$ by

$$\widehat{\sigma}^2 = \frac{39 \times 4.721 + 35 \times 3.368}{40 + 36 - 2} = 4.081\,.$$

The test statistic is then

$$T = \frac{10.13 - 12.16 - 0}{\sqrt{4.081\left(\frac{1}{40} + \frac{1}{36}\right)}} = -4.374\,.$$

This time, for a two-sided test at the 5% significance level, we will reject $H_0$ if $|T| > t_{0.975} = 1.993$ on 74 degrees of freedom. We have $|T| = 4.374 > 1.993$ and so we reject $H_0$ at the 5% level and conclude that the two population means are not equal.

### 10.4 Both distributions non-normal with variances $\sigma_1^2$ and $\sigma_2^2$

If both distributions are non-normal then we can appeal to the central limit theorem. Provided $n_1 > 30$ and $n_2 > 30$, under $H_0$

$$Y = \frac{\overline{X}_1 - \overline{X}_2 - \phi}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ approximately}.$$

Below we give a rejection region resulting in an approximate significance level of $100\alpha\%$ for each of the three alternative hypotheses listed in Section 10.1:

(i) For $H_1 : \mu_1 - \mu_2 > \phi$, we reject $H_0$ at the approximate $100\alpha\%$ significance level if $Y > z_{1-\alpha}$.

(ii) For $H_1 : \mu_1 - \mu_2 < \phi$, we reject $H_0$ at the approximate $100\alpha\%$ significance level if $Y < -z_{1-\alpha}$.

(iii) For $H_1 : \mu_1 - \mu_2 \neq \phi$, we reject $H_0$ at the approximate $100\alpha\%$ significance level if $|Y| > z_{1-\frac{\alpha}{2}}$.

If the variances of the two distributions are unknown then we substitute the sample estimators $S_1^2$ and $S_2^2$ and proceed as just described for the case of known variances.

### 10.5 Bernoulli distributions $\mathrm{Bi}(1, p_1)$ and $\mathrm{Bi}(1, p_2)$

This time we have two independent samples of binary data with $\mathrm{E}(X_{1i}) = p_1$, $i = 1, \ldots, n_1$, and $\mathrm{E}(X_{2i}) = p_2$, $i = 1, \ldots, n_2$. We want to test the null hypothesis

$$H_0 : p_1 - p_2 = \phi,$$

where $\phi$ is a constant (often set equal to zero) against one of the three alternative hypotheses given by

(i) $H_1 : p_1 - p_2 > \phi$ (one-sided)

(ii) $H_1 : p_1 - p_2 < \phi$ (one-sided)

(iii) $H_1 : p_1 - p_2 \neq \phi$ (two-sided)

at the approximate $100\alpha\%$ significance level. Here we are making an inference about the difference in the proportions of 'successes' in the two underlying populations. When $n_1$ and $n_2$ are both large we have that

$$\widehat{p}_1 - \widehat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right) \text{ approximately},$$

and an appropriate test statistic is

$$Y = \frac{\widehat{p}_1 - \widehat{p}_2 - \phi}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}},$$

where in the denominator the following sample estimate of the standard error of $\widehat{p}_1 - \widehat{p}_2$ has been used:

$$\widehat{\text{s. e.}}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

Provided $n_1$ and $n_2$ are both reasonably large, under $H_0$ the test statistic $Y \sim N(0,1)$ approximately by asymptotic results. Note that

$$\widehat{p}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki} = \overline{X}_k, \quad k = 1, 2,$$

which can be expressed as

$$\widehat{p}_k = \frac{r_k}{n_k}\,, \quad k = 1, 2\,,$$

where $r_k = \sum_{i=1}^{n_k} X_{ki}$ denotes the number of successes observed in sample $k$, $k = 1, 2$.

The rejection regions for the three alternative hypotheses given above, using an approximate significance level of $100\alpha\%$, are:

(i) For $H_1 : p_1 - p_2 > \phi$, we reject $H_0$ at the approximate $100\alpha\%$ significance level if $Y > z_{1-\alpha}$

(ii) For $H_1 : p_1 - p_2 < \phi$, we reject $H_0$ at the approximate $100\alpha\%$ significance level if $Y < -z_{1-\alpha}$

(iii) For $H_1 : p_1 - p_2 \neq \phi$, we reject $H_0$ at the approximate $100\alpha\%$ significance level if $|Y| > z_{1-\frac{\alpha}{2}}$

**The case $H_0 : p_1 = p_2$**

If $\phi = 0$, then under $H_0$ we have $p_1 = p_2 = p$, say. An estimate of the common probability $p$ is given by the 'pooled estimate'

$$\overline{p} = \frac{r_1 + r_2}{n_1 + n_2}\,.$$

In this case it makes sense to use the estimate $\overline{p}$ when forming the estimated standard error of $\widehat{p}_1 - \widehat{p}_2$ that appears in the denominator of $Y$. The revised test statistic for the case when $H_0 : p_1 = p_2$ is thus

$$Y = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}}}\,.$$

The rejection regions are otherwise unchanged.

**Example 26.** In a random sample of $n_1 = 120$ voters from Town I, $r_1 = 56$ indicated that they would support Labour in a general election. In a second independent random sample of size $n_2 = 110$ from Town II, taken on the same day as the sample from Town I, $r_2 = 63$ indicated that they would support Labour in a general election. Carry out an appropriate test at the approximate 5% significance level to examine whether the proportions of voters supporting Labour are the same in the two towns.

**Solution**. Let $p_1$ denote the (population) proportion of Labour voters in Town I and $p_2$ denote the (population) proportion of Labour voters in Town II. We wish to test $H_0 : p_1 - p_2 = 0$ vs $H_1 : p_1 - p_2 \neq 0$ at the approximate 5% significance level. We have that $\widehat{p}_1 = r_1/n_1 = 56/120 = 0.467$ and $\widehat{p}_2 = r_2/n_2 = 63/110 = 0.573$.

Under $H_0$, we have that $p_1 = p_2$. An estimate of the common value of $p$ is given by

$$\overline{p} = \frac{r_1 + r_2}{n_1 + n_2} = \frac{56 + 63}{120 + 110} = \frac{119}{230} = 0.517\,.$$

This is used in the denominator of the test statistic to give

$$Y = \frac{0.467 - 0.573 - 0}{\sqrt{\frac{0.517 \times 0.483}{120} + \frac{0.517 \times 0.483}{110}}} = -1.607\,.$$

We would reject $H_0$ at the approximate 5% level if $|Y| > z_{0.975} = 1.96$. The observed value of $|Y| = 1.607 < 1.96$. Hence, there is insufficient evidence to reject $H_0$ at the approximate 5% level. In other words, there is insufficient evidence to reject the claim that the proportions supporting Labour in the two towns are equal.

(Note that both $n_1, n_2 > 9 \times \max\left\{\frac{0.517}{0.483}, \frac{0.483}{0.517}\right\} = 9.634$ which justifies the normal approximations for $\widehat{p}_1$ and $\widehat{p}_2$ under $H_0$.)