# MA22014 Statistics 2A

Simon Shaw, s.shaw@bath.ac.uk

2025/26 Semester I

# Contents

# Overview of MA22014 Statistics 2A

## Syllabus

This unit introduces classical estimation and hypothesis-testing principles. The unit is divided into six main topics:

- Point estimation.
- Confidence intervals.
- Hypothesis testing.
- Inference for relationships.
- One-way analysis of variance.
- Simple linear model.

Within each topic we will develop the theory behind common statistical inference procedures and apply the theory to real case studies.

**Learning outcomes:** After taking this unit, students should be able to:

- Perform **standard estimation procedures** and **hypothesis tests** for parameters in a variety of statistical models.
- Analyse a variety of models for **normally distributed data**.
- Carry out **goodness-of-fit tests** and analyse **contingency tables**.

**Pre-requisites:** Before taking this module you must (take MA12003 AND take MA12004) OR (take MA12011 AND take MA12008) OR take MA12013

I recommend reviewing the following topics from these units so that you have definitions and properties at your fingertips during lectures and tutorials:

- **Probability and Statistics 1A:** Properties of expectation and covariance. Independence. Properties of the binomial and Poisson distributions (mean, variance, mass functions).
- **Probability and Statistics 1B:** Properties of the normal and exponential distributions (mean, variance, density functions). Cumulative distribution

functions, central limit theorem. Random sampling from the distributions above, generating simple graphical and numeric summaries, and writing simple `for` loops in `R`.

Towards the end of Probability and Statistics 1B, you covered parameter estimation in simple cases via method of moments and maximum likelihood as well as sampling distributions of sample means. Our first major task in this unit will be to review these ideas and develop them further in Chapter 2.

## Timetable

**Lectures:**

- Monday 14:15-15:05 EB1.01
- Tuesday 17:15-18:05 CB1.10
- Thursday 12:15-13:05 CB1.10

**Problem classes:**

- Friday 16:15-17:05 via Zoom.
- https://bath-ac-uk.zoom.us/j/99555352535?pwd=WlRJKzM0QytBVXlqbUFPQmJodVJnQT09
- Meeting ID: 995 5535 2535
- Passcode: 122399

**Office hours:**

There will be a dedicated in-person office hour on Monday 15:15-16:05 in my office, 4W4.10. However, I am happy to discuss any matters relating to the course at any time, either via email or one-to-one. If you would like to meet then just send me an email, with a list of proposed times and whether you wish to meet in-person or on Teams.

**Tutorials:** You will be assigned to a small group which will meet weekly to go over a mixture of problems. You will hand in and receive marked work in your tutorial.

- Group 1: Thursday 14:15-15:05, 3WN3.8, tutored by Simon Shaw
- Group 2: Thursday 15:15-16:05, 4E3.5, tutored by Patrick Fahy
- Group 3: Thursday 16:15-17:05, CB3.16, tutored by Patrick Fahy
- Group 4: Thursday 17:15-18:05, CB3.16, tutored by Christian Rohrbeck
- Group 5: Thursday 17:15-18:05, 8W2.8, tutored by Merrilee Hurn
- Group 6: Friday 09:15-10:05 CB3.16, tutored by David Jones
- Group 7: Friday 09:15-11:05 8W2.8, tutored by Karim Anaya-Izquierdo

- Group 8: Friday 10:15-11:05 1WN3.11, tutored by Paddy O'Toole
- Group 9: Friday 11:15-12:05 8W2.8, tutored by Kirstin Strokorb
- Group 10: Friday 11:15-12:05 1WN3.11, tutored by Paddy O'Toole

# Moodle and Panopto Resources

**Recordings:** Recordings of the lectures and problem classes will be made available on Panopto.

**Moodle page:** The Moodle page will contain links for all lecture content and the question sheets.

**Lecture notes:** A full set of comprehensive lecture notes is available through the unit Moodle page. The notes are available in two formats (HTML and pdf) with identical content.

# Assessment and feedback

## Summative assessment

- Exam: 100% of unit mark.

The 2025/6 exam will be an in-person, closed-book assessment. What this means:

- **In person:** you will sit the exams at fixed times on fixed days in a venue at the University. The exams will be invigilated.
- **Closed-book:** You will not be allowed to have any revision materials with you or any access to the internet. Your exam papers will be tailored to this setting.

The exam will be designed to take 3 hours and there will be a total of 100 marks available. It will have two sections.

- **Section A** (worth 40 marks, corresponding to 40%) contains a number of short questions which cover the breadth of the syllabus.
- **Section B** (worth 60 marks, corresponding to 60%) contains longer questions that explore the depth of understanding. marks.
- All questions, in both Section A and Section B, should be answered and contribute to the assessment.

You will be permitted to use a calculator and the University Formula Book in the exam.

### Formative assessment

Homework sheets with 'pen and paper' exercises will be set in each tutorial session with submission of written exercises one week later. Any work submitted by the hand-in deadline will be marked and returned to you giving you personal feedback. Full solutions to all exercises and general feedback sheets will also be made available.

## Some useful books

This unit is self-contained in the sense that you will not strictly need to consult text books. However, the following books are possibly relevant.

**Background reading:**

- Peter Dalgaard, Introductory statistics with R. 2nd edition. Springer. The full text is available as an e-book, either by following the link from the library here or directly here. The first half of this book is a useful refresher on R. Chapters 5 and 8 cover the `R` implementation of the material in Chapter 5.

- John A Rice, Mathematical statistics and data analysis. 3rd edition. Duxbury. Multiple copies of eitherthis edition or earlier additions are available in the library. Chapters 1 – 5 were covered in Probability and Statistics 1A & 1B. Material relevant to this unit can be found primarily in Chapters 6, 8, 9, 11, and 13.

- Yudi Pawitan. In All Likelihood. Oxford University Press. The full text is available as an e-book, either by following the link from the library here. Chapters 2, 4, and 5 are most relevant, though some of the material in those chapters is more advanced.

**Sources for omitted proofs:** I will occasionally reference the following books for details of proofs omitted from these lecture notes:

- George Casella and Roger L Berger. Statistical Inference, 2nd edition. Brooks Cole/Cengage Learning. A very nice book, Chapters 7, 8, and 9 are most relevant.
- Erich L Lehmann. Elements of Large Sample Theory. Springer. The full text is available as an e-book, either by following the link from the library here or directly here. Chapters 2, 3, 4, 5, and 7 contain relevant material.
- Peter J Bickel and Kjell A Doksum. Mathematical statistics. Volume 1, Basic ideas and selected topics, 2nd edition. Pearson Prentice Hall. The appendices contain excellent detail on the probability limit results used in the course.

**O'Reilly Learning Online:** The University has a subscription to O'Reilly Learning Online, giving you access to a wide range of online courses and textbooks on topics including statistics and data science. Our research librarians have compiled a list of high quality courses for R and RStudio: https://library.bath.ac.uk/research-software/R-RStudio. I highly recommend the first two chapters in Learn R Programming course for a quick refresher on `R`.

## Acknowledgements

These notes have developed with input from past lecturers of the unit who include Drs Theresa Smith, Jonathan Bartlett, and Karim Anaya-Izquierdo. Please report any errors to Simon Shaw.

# Chapter 1

# Statistical models and inference

In this chapter we give an overview of statistical models and statistical inference, and through doing so, an overview of what this unit will cover. The purpose of a statistical model is to capture the stochastic (random) behaviour of some phenomenon of interest. Statistical models are not usually mechanistic models for the process under investigation. More commonly, they are empirical models that describe the variation in the data in a way that helps us answer particular substantive questions of interest.

## 1.1 Statistical models

A **statistical model** is a family of distributions which we believe includes – or at least approximates – the true, unknown distribution from which the data were generated. We shall proceed by assuming that our statistical model can be expressed as a ***parametric model***.

**Definition 1.1** (Parametric model)**.** A parametric model for a random variable $X$ is the triple $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$ where only the finite dimensional parameter $\theta \in \Theta$ is unknown.

Thus, the model specifies the sample space $\mathcal{X}$ of the quantity to be observed $X$, the parameter space $\Theta$, and a family of distributions, $\mathcal{F}$ say, where $f_X(x \,|\, \theta)$ is the distribution for $X$ when $\theta$ is the value of the parameter. In this general framework, both $X$ and $\theta$ may be multivariate and we use $f_X$ to represent the density function irrespective of whether $X$ is continuous or discrete. If it is discrete then $f_X(x \,|\, \theta)$ gives the probability of an individual value $x$. Where there is no confusion about the random variable we will suppress the explicit reference to it, writing $f(x \,|\, \theta)$. Typically, $\theta$ is continuous-valued.

Statistics has a somewhat confusing convention of sometimes using $\theta$ as an argument to a function and other times using it to denote the fixed but unknown true value of the parameter. We will stick to this convention, but where useful for clarity, we will denote the true value of the parameter by $\theta^*$.

Where $\theta$ is a vector, often only a subset of the parameters will be of interest, meaning that our substantive questions of interest can be translated into questions about the values of these parameters. The other parameters are called nuisance parameters, which are needed in order for the model to adequately capture the probabilistic structure of the data but otherwise are not related to our substantive questions.

### 1.1.1 Examples

**Example 1.1** (Bernoulli model)**.** Suppose that $X \sim$ Bernoulli($p$), so that $\mathcal{X} = \{0, 1\}$ and

$$P(X = x \,|\, p) \quad = \quad p^x (1-p)^{1-x}.$$

The parameter $p \in [0, 1]$ is continuous.

**Example 1.2** (Binomial model)**.** Suppose that, for $i = 1, \ldots, n$, $X_i \sim$ Bernoulli($p$) and that the $X_i$s are independent. Then the sum, $Y = \sum_{i=1}^n X_i$ is binomially distributed, which we denote as $Y \sim Bin(n, p)$, with

$$P(Y = y \,|\, p) \quad = \quad \binom{n}{y} p^y (1-p)^{n-y}$$

for $y \in \{0, 1, \ldots, n\}$.

**Example 1.3** (Normal model, unknown mean, known variance)**.** Suppose that $X \sim N(\mu, \sigma^2)$ where $\mu$ is the unknown mean and $\sigma^2 > 0$ is the known variance. Then

$$f(x \,|\, \mu) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

for all $x \in (-\infty, \infty)$.

**Example 1.4** (Normal model, unknown mean, unknown variance)**.** Suppose instead that $X \sim N(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown. In this case, the parameter $\theta = (\mu, \sigma^2)$ is bivariate. If, for example, we were interested only in the mean $\mu$ then $\sigma^2$ is a nuisance parameter.

**Example 1.5** (Exponential model)**.** This model is often used to measure lifetimes or waiting times and we write $X \sim \text{Exp}(\lambda)$ if

$$f(x \,|\, \lambda) \quad = \quad \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The exponential model thus has support on the non-negative real line. The parameter $\lambda > 0$ is called the rate parameter.

**Example 1.6** (Normal linear regression model)**.** For each unit in the population there exists a fixed vector of *covariates* $x_i$ and an outcome $Y_i$. The normal linear regression model stipulates that

$$Y_i = \alpha + \beta^T x_i + \epsilon_i$$

where $\alpha$ is an unknown parameter and $\beta$ is a column vector of unknown parameters (regression coefficients) and $\epsilon_i$ is a $N(0, \sigma^2)$ error term. Such models are the main focus of MA22015 Statistics 2B; we'll cover the simple case, where $\beta$ is univariate, in Chapter 7.

Learning about the parameter $\theta$, that is making **inferences** about it, will be the key issue of interest. Typically, we discuss how we can estimate $\theta$ based upon an observed random sample $x_1, \ldots, x_n$ of observations believed to come from the underlying distribution $f(x \,|\, \theta)$. Thus, the random variables $X_1, \ldots, X_n$ will be independent and identically distributed.

### 1.1.2 Motivating examples

We now give some motivating illustrative examples where some of these statistical models could potentially be used to draw inferences about the phenomenon in question. We will return to these examples later during the unit.

**Example 1.7** (Mean body mass index in NHANES)**.** The National Health and Nutrition Examination Survey (NHANES) is a long running program of studies whose aim is to assess the health and nutritional status of adults and children in the United States. The survey consists of a combination of interviews with participants, physical examinations and laboratory tests.

Here we focus on data from the 2003-2004 study. We focus specifically on data from those aged between 20 and 39 years of age when the survey was performed, for a total of $n = 1,753$ individuals. Figure 1.1 shows a histogram of body mass index (BMI) values for the 1,638 individuals between 20 and 39 years of age who had BMI recorded in the study.

A statistical model we might use for the BMI data is to assume that the sample of BMI values are an i.i.d. sample from a $N(\mu, \sigma^2)$ distribution, and we are primarily interested in the population mean BMI $\mu$. The histogram shows a

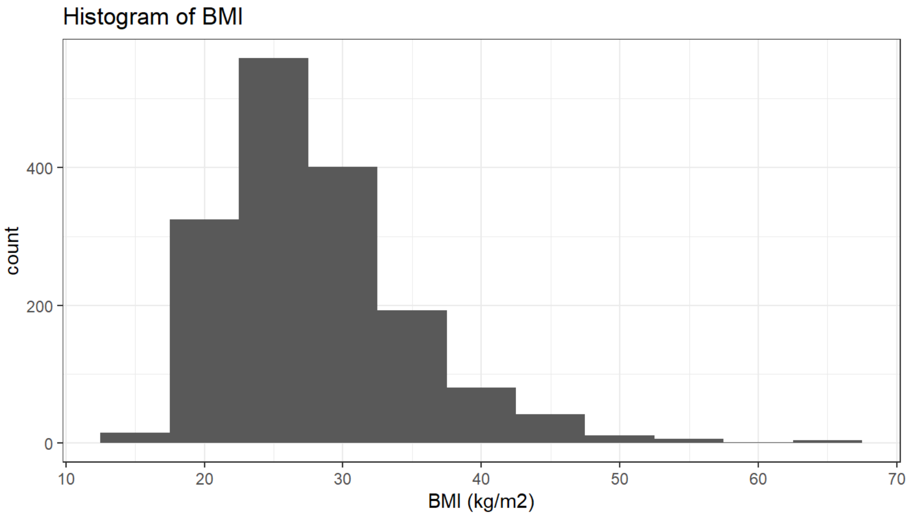


Figure 1.1: Histogram of 1,638 body mass index (kg/m2) values from NHANES 2003-2004.

skewed distribution however, and so an initial concern is thus whether using a model which assumes a normal distribution is appropriate.

Note that the use of a statistical model here is to capture the randomness induced by the random sampling process. The BMI for each individual in the population is fixed at the time of the survey. However, the BMI values in our sample are modelled as random variables because, prior to the survey, we do not know which individuals will be selected into the sample.

**Example 1.8** (Drinking in NHANES)**.** The NHANES data introduced in the previous example contains many additional variables. A subset of these were based on asking participants about their consumption of alcohol. One of the questions asked was “In the past 12 months, on those days that you drank alcoholic beverages, on the average, how many drinks did you have?”. In order to learn about (reported) alcohol drinking behaviour, we will consider a dichotomised version of this variable, defined as whether they answered that they drank one drink on average on days they drank alcohol, or more than one.

For this variable $n = 1,104$ responded to the question, and of these, 846 (76.6%) answered that they drank more than one alcoholic drink on average on days that they did drink alcohol. Since each response is binary, we can use the i.i.d. Bernoulli model here, with the probability $p$ representing the proportion of the population that would answer that they drink more than one alcoholic drink on average on days they drink.

**Example 1.9** (Randomised two arm clinical trial in AIDS)**.** The AIDS Clinical Trials Group Study 175 (ACTG175) was a randomised clinical trial conducted in the 1990s to compare different treatments for adults infected with HIV. We will focus on a comparison of those randomised to receive zidovudine treatment (532 patients) and those randomised to receive a combination treatment of zidovudine plus didanosine (522 patients). One of the outcomes is the patient's CD4 blood cell count. A larger CD4 value is indicative that the treatment is working. Thus one of the analyses of interest is to compare the CD4 count after treatment with zidovudine plus didanosine compared to treatment with zidovudine alone. Figure 1.2 shows histograms of the CD4 count 20 weeks after baseline in these two groups.

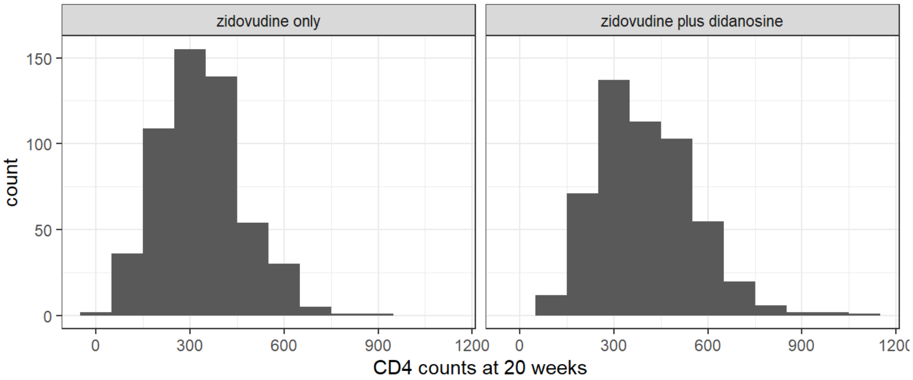


Figure 1.2: Histogram of CD4 T cell count at 20 weeks in zidovudine only versus zidovudine plus didanosine groups of ACTG175 trial.

The comparison of these two treatment groups can be represented as a two-sample problem. In particular, our model will assume that the CD4 data on the patients in these two groups represent independent samples from the two hypothetical populations that would exist were we to treat all the eligible patients in the population with these two treatment options.

We let $X_1, \ldots, X_n$ denote the patient's CD4 counts in the zidovudine treatment group and $Y_1, \ldots, Y_m$ the values in the zidovudine plus didanosine group. A possible parametric model is that $X_1, \ldots, X_n$ are i.i.d. $N(\mu_X, \sigma^2)$ and the $Y_1, \ldots, Y_m$ are i.i.d. $N(\mu_Y, \sigma^2)$. This assumes the variance of CD4 count is the same in the two groups. A less restrictive (but still parametric) model would allow potentially distinct variances in the two groups. Primary interest could be in the contrast of $\mu_Y$ with $\mu_X$. This is a so called two sample problem, which we will examine in Chapter 5.

**Example 1.10** (Randomised two arm trial in cardiovascular disease)**.** The Physician's Health Study was an important medical study was carried out in the 1980s to investigate the benefit and risks of aspirin for cardiovascular disease and cancer. The trial randomised 22,071 healthy physicians to receive either aspirin or placebo. They were then followed-up to see which experienced heart attacks during the following 5 years. The data are shown in Table 1.1.

Table 1.1: Number of heart attacks by treatment group in the Physician's Health Study

| Group | Heart attack | No heart attack | Total |
|---|---:|---:|---:|
| Aspirin | 139 | 10,898 | 11,037 |
| Placebo | 239 | 10,795 | 11,034 |

Primary interest is in whether receiving aspirin increases or decreases the probability of experiencing a heart attack. Let $X_i$ denote the heart attack outcome (1=yes, 0=no) for the $i$th participant in the aspirin group, and $Y_i$ denote the heart attack outcome in the $i$th participant in the placebo group. Then a natural model is to assume the $X_i$ are i.i.d. Bernoulli with 'success' probability $p_X$ and the $Y_i$ are i.i.d. Bernoulli with 'success' probability $p_Y$. Primary interest could then be in the contrast of $p_X$ with $p_Y$.

## 1.2 Statistical inference

The statistical inference problem is to make inferences about one or more parameters in the statistical model, based on the observed data. With a sample of infinite size, we would be able to empirically calculate the true parameter values of the model. In practice we observe data of finite size, meaning any estimates of the parameters will in general differ to some extent from their true values. The statistical inference problem then consists of a number of elements:

- What is our best guess or estimate of the unknown parameters based on the observed data? (point estimation, Chapter 2)
- How much uncertainty do we have in our parameter estimates? (confidence intervals, Chapter 3)
- Which of two competing hypotheses, specified in terms of model parameters, is true? (hypothesis testing, Chapter 4)

This unit will cover each of these steps in turn. At each step, we will also consider the **robustness** of our procedures to **model misspecification**. What guarantees, if any, can we have on the performance of our procedures if the true data generating distribution does not belong to the class of distributions characterised by our parametric model? Such robustness is desirable because, although the data can be used to some extent to check for model misspecification (see Section 5.4.2), it is impossible with a finite amount of data to prove a model is correctly specified.

In Chapter 5 we will apply our new skills from Chapters 2 – 4 to develop inferential techniques for investigating relationships between variables.

We extend these ideas to more than two independent samples in Chapter 6 whilst we conclude the unit with Chapter 7 which introduces the simple linear model, the extension of which is the topic of the Semester II unit MA22015 Statistics 2B.

# Chapter 2

# Point estimation

In this chapter we consider how to estimate the unknown parameter $\theta$ in a parametric model $f(x|\theta)$, $\theta \in \Theta$. Our generic setup is that we have $n$ independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ which are drawn from some common distribution. The respective realisations are $x_1, \ldots, x_n$ so that capital letters denote the random variables and lower case letter the realised observations.

**Definition 2.1** (Estimator/Estimate)**.** An **estimator** of a parameter $\theta$ is a function $T(X_1, \ldots, X_n)$ of the random variables, $X_1, \ldots, X_n$. For a specific set of observations $x_1, \ldots, x_n$, the value taken by the estimator, $\hat{\theta} = T(x_1, \ldots, x_n)$, is the (point) **estimate**.

Thus, an estimator is a random variable and the distribution of the estimator is often called its **sampling distribution**. The estimate is a real number calculated using the observed data $x_1, \ldots, x_n$.

In this unit we will follow notational convention in using $\hat{\theta}$ to sometimes refer to the estimator, for example when making statements about the random variable such as $E(\hat{\theta}) = E(\hat{\theta}(X_1, \ldots, X_n))$, and sometimes to the estimate, for example the evaluated number from the data such as $\hat{\theta} = 5.36$.

In many cases, there may be an intuitive candidate for a point estimator of a particular parameter. For example, we could estimate a parameter by its sample analogue. Thus, if we are interested in estimating the parameter $\mu$ of a $N(\mu, \sigma^2)$ distribution then the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a natural candidate. Equally, we could also consider the sample median $\text{med}\{X_1, \ldots, X_n\}$. In practice, we will need a more methodical approach, both to derive an estimator and to study properties of that estimator to justify the choice.

## 2.1 Method of moments

The first approach to finding estimators for parametric models we consider is the method of moments. For integer $p \geq 1$, the $p$th uncentred population/true moment is defined as $E(X^p)$. The corresponding centred moment is defined as $E\left[(X - E(X))^p\right]$. The uncentred sample moment is defined as

$$M'_p \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^p.$$

Thus $M'_1$ is the sample mean $\overline{X}$. The $p$th centred sample moment is

$$M_p \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^p.$$

Now suppose that $\theta = (\theta_1, \dots, \theta_p)$ is $p$ dimensional. The method of moments estimator for $\theta$ is obtained by equating the first $p$ (uncentred) sample moments with the corresponding population/true moments under the assumed model, and solving the resulting set of simultaneous equations.

**Example 2.1** (Method of moments, normal distribution)**.** Suppose $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Then $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$. Let $\hat{\mu}$, $\hat{\sigma}^2$ denote, respectively, the method of moments estimators of $\mu$ and $\sigma^2$ then

$$\overline{X} \;=\; \hat{\mu} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \;=\; \hat{\sigma}^2 + \hat{\mu}^2,$$

which solving for $\hat{\mu}$ and $\hat{\sigma}^2$ leads to

$$\hat{\mu} \;=\; \overline{X} \quad \text{and} \quad \hat{\sigma}^2 \;=\; \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \overline{X}^2.$$

The estimator $\hat{\sigma}^2$ can also be expressed as $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, since

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \;&=\; \frac{1}{n}\sum_{i=1}^{n}\left[X_i^2 - 2X_i\overline{X} + \overline{X}^2\right] \\
&=\; \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n} X_i + \overline{X}^2 \\
&=\; \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - 2\overline{X}^2 + \overline{X}^2 \\
&=\; \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \overline{X}^2.
\end{aligned}$$

Note that in the construction of these estimators, we have only used the properties of expectation and variance and so these estimators would apply for any model with expectation $\mu$ and finite variance $\sigma^2$.

The method of moments approach has however some drawbacks. One is that estimates are not in general guaranteed to lie in the parameter space $\Theta$ defined by the model.

## 2.2 Maximum likelihood estimation

In the model $\mathcal{E} = \{\mathcal{X}, \Theta, f(\mathbf{x} \,|\, \theta)\}$, $f(\cdot)$ is a function of $\mathbf{x}$ for known $\theta$. If we have instead observed $\mathbf{x}$ then we could consider viewing this as a function, termed the **likelihood**, of $\theta$ for known $\mathbf{x}$. This provides a means of comparing the plausibility of different values of $\theta$.

**Definition 2.2** (Likelihood function)**.** The likelihood for $\theta$ given observations $\mathbf{x}$ is

$$L(\theta \,|\, \mathbf{x}) \quad = \quad f(\mathbf{x} \,|\, \theta), \quad \theta \in \Theta$$

regarded as a function of $\theta$ for fixed $\mathbf{x}$.

Typically we will have $\mathbf{x} = (x_1, \ldots, x_n)$ and, in the special case where the data consist of $n$ i.i.d. samples from the univariate density $f(x|\theta)$, such that the joint density function $f(\mathbf{x} \,|\, \theta)$ is equal to the product of the univariate densities, the likelihood function reduces to

$$L(\theta \,|\, \mathbf{x}) \quad = \quad \prod_{i=1}^{n} f(x_i \,|\, \theta).$$

If $L(\theta_1 \,|\, \mathbf{x}) > L(\theta_2 \,|\, \mathbf{x})$ then the observed data $\mathbf{x}$ were more likely to occur under $\theta = \theta_1$ than $\theta_2$ so that $\theta_1$ can be viewed as more plausible, or more **likely**, than $\theta_2$. Thus, a natural approach is to choose the value of the unknown parameter as the one under which the data we have observed is most likely to occur; this is the maximum likelihood estimator/estimate.

**Definition 2.3** (Maximum likelihood estimator)**.** Given a sample of data $\mathbf{x}$ and parametric model $f(\mathbf{x} \,|\, \theta)$, a **maximum likelihood estimate** is a value of $\hat{\theta} \in \Theta$ which maximises the likelihood function $L(\theta \,|\, \mathbf{x})$. If each possible sample $\mathbf{x}$ leads to a unique value (estimate) of $\hat{\theta}$, then the procedure defines a function

$$\hat{\theta} \quad = \quad T(\mathbf{x}).$$

The corresponding random variable $T(\mathbf{X})$ is **the maximum likelihood estimator**.

We will use the abbreviation **MLE** to refer to either the maximum likelihood estimate or maximum likelihood estimator. Assuming the likelihood function is differentiable with respect to $\theta$, our usual approach for finding the maximum is to differentiate the likelihood function with respect to the parameter, set the derivative equal to zero and solve for $\theta$. We must also check that the resulting critical point is a maximum by looking at the value of the second derivative.

As we see in the following example, it is often easier to work with the **log-likelihood**, the log of the likelihood function,

$$l(\theta \,|\, \mathbf{x}) \quad = \quad \log(L(\theta \,|\, \mathbf{x})).$$

As the logarithm is a monotonically increasing function, maximising $l(\theta \,|\, \mathbf{x})$ is equivalent to maximising the likelihood function. In the case of i.i.d. data,

$$l(\theta \,|\, \mathbf{x}) \quad = \quad \log\left[\prod_{i=1}^{n} f(x_i \,|\, \theta)\right] = \sum_{i=1}^{n} \log\{f(x_i \,|\, \theta)\}.$$

**Example 2.2** (Normal mean, known variance)**.** Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Although typically this is usually not the case in reality, suppose that we knew the true value of the variance $\sigma^2$. Thus $\mu$ is the only unknown parameter. The likelihood function given data $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$\begin{aligned} L(\mu \,|\, \mathbf{x}) \quad &= \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &= \quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}, \end{aligned}$$

and the log likelihood is

$$l(\mu \,|\, \mathbf{x}) \quad = \quad -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Differentiating with respect to $\mu$ we have

$$l'(\mu \,|\, \mathbf{x}) \quad = \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

Solving $l'(\mu \,|\, \mathbf{x}) = 0$ we find that $\hat{\mu} = \overline{x}$. To confirm this maximises the log likelihood, note that the second derivative

$$l''(\mu \,|\, \mathbf{x}) \quad = \quad -\frac{n}{\sigma^2},$$

is always negative as $\sigma^2 > 0$. As this is the only candidate in $(-\infty, \infty)$ it thus follows that $\hat{\mu}$ is a global maximum.

**Example 2.3** (Normal mean, known variance equal to one)**.** Suppose that $\mathbf{x} = (0.4191, -1.4558, -1.4559, -1.0156, -0.5331, -0.6836, 0.3259, 0.9226, -0.0509, 0.2972)$ are $n = 10$ realisations from $N(\mu, 1)$. Consequently, $\overline{x} = -0.3230$. Figure 2.1 shows the corresponding likelihood and log-likelihood functions. On each plot the red line shows the value of the function at $\overline{x} = -0.3230$, illustrating the theoretical derivation of Example 2.2 that this is maximum likelihood estimate.

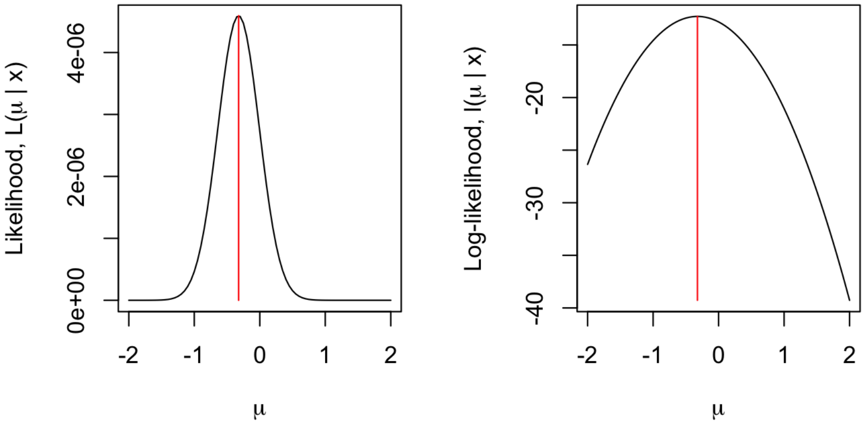


Figure 2.1: Likelihood and log-likelihood function for 10 observations from a normal model with unknown mean and known variance of 1.

**Example 2.4** (Exponential distribution)**.** Suppose that $X_1, \ldots, X_n$ are i.i.d. $\text{Exp}(\lambda)$, $\lambda > 0$. Recall that this means that each $X$ takes values in the non-negative reals, with density

$$f(x \mid \lambda) = \lambda \exp(-\lambda x).$$

The log likelihood function given data $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$l(\lambda \mid \mathbf{x}) \quad = \quad \sum_{i=1}^{n} \{\log(\lambda) - \lambda x_i\},$$

and the first derivative of this is

$$l'(\lambda \mid \mathbf{x}) \quad = \quad \sum_{i=1}^{n} \frac{1}{\lambda} - \sum_{i=1}^{n} x_i = \frac{n}{\lambda} - n\overline{x}.$$

Setting to zero and solving for the parameter, we have a candidate MLE

$$\hat{\lambda} \quad = \quad \frac{1}{\overline{x}}.$$

Since $l''(\lambda \,|\, \mathbf{x}) = -n\lambda^{-2}$ and $\lambda > 0$, $l''(\lambda \,|\, \mathbf{x}) < 0$. So $\hat{\lambda} = 1/\overline{x}$ is indeed the MLE.

In cases where the parameter space is closed, we should also be careful to check whether the likelihood is maximized at one of the boundary points of the parameter space.

**Example 2.5** (Binomial MLE)**.** Suppose that $Y \sim \text{Binomial}(n, p)$, with $0 \leq p \leq 1$. The likelihood function given an observation $y$ is

$$L(p \,|\, y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Then the log likelihood is

$$l(p \,|\, y) \quad = \quad \log \binom{n}{y} + y \log(p) + (n-y)\log(1-p)$$

for $0 < p < 1$. To find the MLE we differentiate with respect to $p$:

$$l'(p \,|\, y) \quad = \quad \frac{y}{p} - \frac{n-y}{1-p}.$$

Solving for $p$ in $l'(p \,|\, y) = 0$ leads to $\hat{p} = y/n$ provided that $0 < y/n < 1$, i.e. in the case that $0 < y < n$. To check it is a maximum, we can find the second derivative as

$$l''(p \,|\, y) = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2},$$

which is negative when $0 < y < n$. We can then check that $L(0 \,|\, y) = 0$ and $L(1 \,|\, y) = 0$, so that $\hat{p} = y/n$ is indeed the global maximum, and hence is the MLE, provided that $0 < y < n$.

We must now consider separately the cases when $y = 0$ or $y = n$. If $y = n$, then $L(p \,|\, y) = p^n$, which is clearly maximised by $\hat{p} = 1$ (note this is not a critical point of the function). If $y = 0$, then $L(p \,|\, y) = (1-p)^n$, which is maximised by $\hat{p} = 0$. Thus in all cases the MLE is $\hat{p} = y/n$. Note that if the parameter space had been restricted so that $0 < p < 1$, in the cases that $y = 0$ or $y = n$ then the MLE would not exist.

**Example 2.6** (Drinking in NHANES)**.** Recall the NHANES alcohol variable described in Example 1.8. $n = 1,104$ individuals responded to the alcohol question, and of these, 846 (76.6%) answered that they drank more than one alcoholic drink on average on days that they did drink alcohol. As noted earlier, a suitable model is to assume that the binary outcomes of the sampled individuals are a size $n$ i.i.d. sample from a Bernoulli distribution with unknown ‘success’ parameter $p$ corresponding to the probability that a randomly chosen individual says they drank more than one alcoholic drink on average. The MLE, $\hat{p}$, is thus the sample proportion, 0.766.

In the preceding examples we have differentiated the log likelihood, found unique roots and then confirmed these are maxima. This process provides us with values which could maximise the log likelihood (and hence likelihood), but it is not guaranteed to do so. One way in which this can occur is when the MLE falls on a boundary of the parameter space.

**Example 2.7** (MLE on the boundary)**.** Suppose $X_1, \ldots, X_n$ is $N(\mu, \sigma^2)$ with $\sigma^2$ known and the additional restriction that $\mu \geq 0$. If $\overline{x} \geq 0$, then the likelihood function is maximised at $\hat{\mu} = \overline{x}$. If however $\overline{x} < 0$, the MLE cannot be $\overline{x} < 0$ given the additional restriction. Figure 2.2 shows a plot of the log likelihood with $n = 10$, $\sigma^2 = 1$ and $\overline{x} = -0.5$.

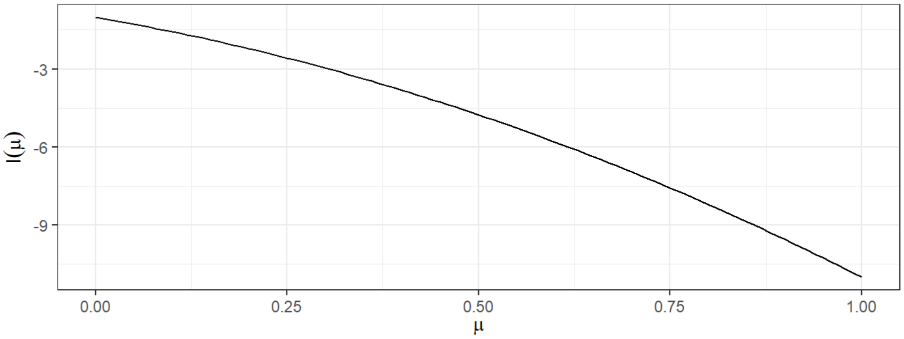


Figure 2.2: Log likelihood function for normal model with MLE on the boundary.

With $\overline{x} < 0$ the log likelihood is larger at $\mu = 0$ than for any $\mu > 0$, so that the MLE is then $\hat{\mu} = 0$, yet this is not a stationary point of the log likelihood function.

### 2.2.1 Multivariate case

We now consider the case where the parameter is multivariate $\theta = (\theta_1, \ldots, \theta_p)$. We seek the value of $\theta$ which maximises $L(\theta \,|\, \mathbf{x}) = L(\theta_1, \ldots, \theta_p \,|\, \mathbf{x})$, or equivalently the log likelihood function. Candidate values for the MLE are those values of $\theta$ for which the partial derivatives of the log likelihood are all zero:

$$\frac{\partial}{\partial \theta_r} l(\theta \,|\, \mathbf{x}) \quad = \quad 0$$

for each $r = 1, \ldots, p$.

For such values of $\theta$, a sufficient condition for them to be a (local) maximum is that the Hessian matrix of the log likelihood is a negative definite matrix. The Hessian matrix is given by

$$H(\theta) = \begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2} l(\theta \,|\, \mathbf{x}) & \frac{\partial^2}{\partial\theta_1\partial\theta_2} l(\theta \,|\, \mathbf{x}) & \dots & \frac{\partial^2}{\partial\theta_1\partial\theta_p} l(\theta \,|\, \mathbf{x}) \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} l(\theta \,|\, \mathbf{x}) & \frac{\partial^2}{\partial\theta_2^2} l(\theta \,|\, \mathbf{x}) & \dots & \frac{\partial^2}{\partial\theta_2\partial\theta_p} l(\theta \,|\, \mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_p\partial\theta_1} l(\theta \,|\, \mathbf{x}) & \frac{\partial^2}{\partial\theta_p\partial\theta_2} l(\theta \,|\, \mathbf{x}) & \dots & \frac{\partial^2}{\partial\theta_p^2} l(\theta \,|\, \mathbf{x}) \end{pmatrix}$$

The matrix $H(\theta)$ is negative definite at $\hat{\theta}$ if for all non-zero vectors $a = (a_1, \dots, a_p)^T$,

$$a^T H(\hat{\theta}) a < 0.$$

**Example 2.8** (Normal distribution, mean and variance unknown)**.** Suppose that $X_1, \dots, X_n$ are i.i.d. draws from $N(\mu, \sigma^2)$, with now both $\mu$ and $\sigma^2$ unknown. Thus now $\theta = (\mu, \sigma^2)$. We showed earlier (in Example 2.2) that the log likelihood is given by

$$l(\theta \,|\, \mathbf{x}) \quad = \quad -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The first order partial derivatives are

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu); \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned}$$

Solving $\frac{\partial l}{\partial \mu} = 0$ we find that, for $\sigma^2 > 0$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$. Substituting this into $\frac{\partial l}{\partial \sigma^2} = 0$ gives

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad = \quad 0,$$

so that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. The second order partial derivatives are

$$\begin{aligned} \frac{\partial^2 l}{\partial \mu^2} &= -\frac{n}{\sigma^2}; \\ \frac{\partial^2 l}{\partial (\sigma^2)^2} &= \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i - \mu)^2; \\ \frac{\partial^2 l}{\partial \mu \partial \sigma^2} &= -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu) \;=\; \frac{\partial^2 l}{\partial \sigma^2 \partial \mu}. \end{aligned}$$

Evaluating these at $\hat{\mu} = \overline{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ gives

$$\begin{aligned}
\left.\frac{\partial^2 l}{\partial \mu^2}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} &= -\frac{n}{\hat{\sigma}^2};\\
\left.\frac{\partial^2 l}{\partial (\sigma^2)^2}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} &= \frac{n}{2(\hat{\sigma}^2)^2} - \frac{1}{(\hat{\sigma}^2)^3}\sum_{i=1}^{n}(x_i - \overline{x})^2 \ = \ -\frac{n}{2(\hat{\sigma}^2)^2};\\
\left.\frac{\partial^2 l}{\partial \mu \partial \sigma^2}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} &= -\frac{1}{(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i - \overline{x}) \ = \ 0.
\end{aligned}$$

A sufficient condition for $L(\hat{\mu}, \hat{\sigma}^2 \,|\, \mathbf{x})$ to be a maximum is that the matrix

$$H(\hat{\theta}) \ = \ \left.\left( \begin{array}{cc} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l}{\partial (\sigma^2)^2} \end{array} \right)\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} \ = \ \left( \begin{array}{cc} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{array} \right)$$

is negative definite. For any non-zero vector $a = (a_1, a_2)^T$ we have

$$a^T H(\hat{\theta}) a \ = \ -\frac{n a_1^2}{\hat{\sigma}^2} - \frac{n a_2^2}{2(\hat{\sigma}^2)^2} < 0$$

so that $H(\hat{\theta})$ is negative definite. The maximum likelihood estimates are $\hat{\mu} = \overline{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. The corresponding maximum likelihood estimators are $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ for $\mu$ and $\sigma^2$ respectively.

**Example 2.9.** Recall the NHANES BMI data from Example 1.7. We have data on 1,638 individuals between 20 and 39 years of age who had BMI recorded in the study. We are interested in estimating the population mean BMI $\mu$, and we do not know the population variance $\sigma^2$. Assuming the population distribution of BMIs is normal $N(\mu, \sigma^2)$, we can obtain a point estimate for $\mu$ using the MLE $\hat{\mu} = \overline{x} = 27.9$. This is also the MOM estimate for $\mu$.

### 2.2.2 Invariance principle

Suppose that in addition to finding the MLE for $\theta$ we are also interested in finding the MLE for some function, $g(\theta)$, of $\theta$. For example, if $\theta$ is the variance then we might be interested in the precision, $1/\theta$, or the standard deviation, $\sqrt{\theta}$.

Maximum likelihood estimation enjoys the so called invariance principle in respect of transformation of the model parameters. Informally, if $\hat{\theta}$ is the MLE of $\theta$ then $g(\hat{\theta})$ is the MLE of $g(\theta)$. We now formalise this approach.

Let $\theta$ denote the model parameters and consider a transformation of $\theta$, given by $\psi = g(\theta)$. If the transformation is one-to-one (so each $\theta$ corresponds to a unique value of $g(\theta)$ and vice versa) then the likelihood function for $\psi$ is given by $L^*(\psi \,|\, \mathbf{x}) = L(g^{-1}(\psi) \,|\, \mathbf{x})$. In the more general case, we make the following definition of the **induced likelihood function** of the transformed parameter.

**Definition 2.4** (Induced likelihood of transformed parameter)**.** Let $\psi = g(\theta)$ be a transformation of the model parameter $\theta$. The induced likelihood function for the transformed parameter $\psi$ is defined as

$$L^*(\psi \,|\, \mathbf{x}) = \sup_{\{\theta : g(\theta)=\psi\}} L(\theta \,|\, \mathbf{x}).$$

**Theorem 2.1** (Invariance principle for maximum likelihood estimator)**.** *Suppose that $\hat{\theta}$ is the MLE of $\theta$. Then $g(\hat{\theta})$ is the MLE of $\psi = g(\theta)$.*

**Proof:** If the transformation $g(\theta)$ is one-to-one the result is clear. For the general case, see the proof of Theorem 7.2.10 in Casella and Berger. □

**Example 2.10** (MLE of the standard deviation)**.** In Example 2.8 we showed that the MLE of the variance in the normal model is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. By the invariance property, Theorem 2.1, we can immediately deduce that the MLE of the standard deviation is $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$.

## 2.3 Evaluating point estimators

We have seen two approaches for deriving estimators. We now consider ways in which we can judge the quality of competing estimators. We do so by examining their so called frequentist properties, that is how they perform in (hypothetical) repeated samples from the population or repeated experiments.

### 2.3.1 Bias and variance

**Definition 2.5** (Bias of an estimator)**.** The bias of an estimator is defined as the difference between its expected value in repeated samples and the true parameter value:

$$\text{Bias}(\hat{\theta}, \theta) \quad = \quad E(\hat{\theta}) - \theta.$$

Note that if there is no confusion about the parameter $\theta$ being estimated then we may suppress the direct reference to $\theta$ and write Bias($\hat{\theta}$). In general the bias could depend on the true parameter value $\theta$, although as we shall see sometimes this is not the case. As we shall also see, the bias may depend on the sample size $n$. If the bias is zero for all true parameter values, we say the estimator is unbiased.

**Example 2.11** (Bias of sample mean)**.** Suppose $X_1, \dots, X_n$ are i.i.d. from a distribution with mean $\mu$. Then the sample mean $\hat{\mu} = \frac{1}{n}\sum X_i$ is unbiased, since

$$\begin{aligned} E(\hat{\mu}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}n\mu = \mu. \end{aligned}$$

Assuming our estimator is unbiased or has small bias, we would like our estimator to have small variance, since small variance means that the estimator on average is close (in squared distance) to its mean.

**Example 2.12** (Variance of sample mean)**.** Suppose $X_1, \dots, X_n$ are i.i.d. from a distribution with mean $\mu$ and finite variance $\sigma^2$. Then the variance of the sample mean is given by

$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &= \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2. \end{aligned}$$

Putting Examples 2.11 and 2.12 together, we have that:

- The sample mean is unbiased for the true mean $\mu$.
- The variance of the sample mean decreases, like $1/n$, as the sample size $n$ increases.

Intuitively this makes sense - as the sample size $n$ increases, the estimate will tend to be closer on average to the true mean.

**Definition 2.6** (Standard error)**.** The standard error of an estimator $\hat{\theta}$ is defined as $\sqrt{\mathrm{Var}(\hat{\theta})}$.

**Example 2.13** (Standard error of sample mean)**.** Following on from Example 2.12, the standard error of the sample mean is given by

$$\sqrt{\mathrm{Var}(\hat{\mu})} = \frac{1}{\sqrt{n}}\sigma.$$

Thus, for example, reducing the error in the estimate of $\mu$ by a factor of two would take four times as many sample observations.

We now consider an estimator which, unlike the sample mean, is not unbiased (for finite $n$).

**Example 2.14** (Bias of method of moments variance estimator)**.** Consider again the method of moments estimator of the variance $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \hat{\mu})^2$ where $\hat{\mu} = \overline{X}$. We have

$$\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\text{Var}(X_i - \overline{X}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\text{Var}(X_i) - 2\text{Cov}(X_i, \overline{X}) + \text{Var}(\overline{X})\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma^2 - 2\text{Cov}(X_i, \overline{X}) + \frac{\sigma^2}{n}\right\}.
\end{aligned}$$

Now, noting that $\text{Var}(X_i) = \text{Cov}(X_i, X_i)$,

$$\begin{aligned}
\text{Cov}(X_i, \overline{X}) &= \frac{1}{n}\sum_{j=1}^{n}\text{Cov}(X_i, X_j) \\
&= \frac{1}{n}\left\{\text{Var}(X_i) + \sum_{j\neq i}^{n}\text{Cov}(X_i, X_j)\right\} \\
&= \frac{1}{n}\left\{\sigma^2 + \sum_{j\neq i}^{n} 0\right\} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

since the $X_i$s are independent. Consequently,

$$\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma^2 - 2\frac{\sigma^2}{n} + \frac{\sigma^2}{n}\right\} &(2.1)\\
&= \frac{(n-1)}{n}\sigma^2. &(2.2)
\end{aligned}$$

The method of moments variance estimator is thus biased downwards by $\sigma^2/n$. Note that, by multiplying Equation (2.2) by $n/(n-1)$, we have that the estimator

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

is an unbiased estimator of the population variance.

Note that the (absolute) bias of $\hat{\sigma}^2$ depends on the true parameter value of $\sigma^2$. From Equation (2.2), we can also see that the bias tends to zero as $n \to \infty$. It is thus asymptotically unbiased. It turns out that quite a lot of the most used statistical methods are biased for finite sample sizes and only unbiased asymptotically (in the limit).

### 2.3.2 Mean squared error

Bias is just one facet of an estimator. Consider the following two scenarios.

1. The estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ may be unbiased but the sampling distribution, $f(\hat{\theta} \mid \theta)$, may be quite disperse: there is a large probability of being far away from $\theta$, so for any $\epsilon > 0$ the probability that $P(\theta - \epsilon < \hat{\theta} < \theta + \epsilon)$ is small.

2. The estimator $\hat{\theta}$ may be biased but the sampling distribution, $f(\hat{\theta} \,|\, \theta)$, may be quite concentrated. So, for any $\epsilon > 0$ the probability that $P(\theta - \epsilon < \hat{\theta} < \theta + \epsilon)$ is large.

In these cases, the biased estimator may be preferable to the unbiased one. We would like to know more than whether or not an estimator is biased. In particular, we wish to capture some idea of how concentrated the sampling distribution of the estimator $\hat{\theta}$ is around $\theta$. Ideally we would like

$$P(|\hat{\theta} - \theta| < \epsilon) \quad = \quad P(\theta - \epsilon < \hat{\theta} < \theta + \epsilon)$$

to be large for all $\epsilon > 0$. This probability may be hard to evaluate, but we may make use of Chebyshev's inequality.

**Theorem 2.2** (Chebyshev's inequality)**.** *Let $Y$ be a random variable with finite mean $\mu$ and finite variance $\sigma^2$. For any $\epsilon > 0$ and $c \in \mathbb{R}$*

$$P(|Y - c| \geq \epsilon) \leq \frac{E\{(Y - c)^2\}}{\epsilon^2}$$

*so that $P(|Y - c| < \epsilon) \geq 1 - \frac{E\{(Y-c)^2\}}{\epsilon^2}$.*

**Proof:** (For completeness; non-examinable) Suppose that $Y$ is continuous (the result holds more generally than this), and let the density of $Z = Y - c$ be $f(z)$. Then

$$\begin{aligned} E(Z^2) &= \int z^2 f(z) dz \\ &= \int_{|z| \geq \epsilon} z^2 f(z) dz + \int_{|z| < \epsilon} z^2 f(z) dz \\ &\geq \int_{|z| \geq \epsilon} z^2 f(z) dz \\ &\geq \epsilon^2 \int_{|z| \geq \epsilon} f(z) dz = \epsilon^2 P(|Z| \geq \epsilon). \end{aligned}$$

The result then follows by rearrangement. □

A common choice of $c$ is $E(Y)$ so that

$$P(|Y - E(Y)| \geq \epsilon) \leq \frac{Var(Y)}{\epsilon^2}.$$

For an estimator $\hat{\theta}$ we are interested in $|\hat{\theta} - \theta|$ (which reduces to $|\hat{\theta} - E(\hat{\theta})|$ in the case when $\hat{\theta}$ is an unbiased estimator of $\theta$) and so applying Chebyshev's inequality with $Y = \hat{\theta}$ and $c = \theta$ gives

$$P(|\hat{\theta} - \theta| < \epsilon) \quad \geq \quad 1 - \frac{E\{(\hat{\theta} - \theta)^2\}}{\epsilon^2}.$$

Consequently, we will be interested in estimators for which $E\{(\hat{\theta} - \theta)^2\}$, the average distance between the estimator and its target parameter, is small. This quantity is formally known as the mean squared error (MSE).

**Definition 2.7** (Mean squared error)**.** The mean squared error of an estimator $\hat{\theta}$ is defined as

$$\text{MSE}(\hat{\theta}, \theta) \quad = \quad E\left\{(\hat{\theta} - \theta)^2\right\}.$$

Note that if there is no confusion about the parameter $\theta$ being estimated then we may suppress the direct reference to $\theta$ and write MSE$(\hat{\theta})$. We now show that the MSE is equal to the sum of the squared bias and variance.

**Theorem 2.3.** *The mean squared error of an estimator $\hat{\theta}$ can be expressed as*

$$MSE(\hat{\theta}, \theta) \quad = \quad Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2.$$

**Proof:**

$$\begin{aligned}
\text{MSE}(\hat{\theta}, \theta) &= E\left[(\hat{\theta} - \theta)^2\right] \\
&= E\left[\left(\hat{\theta} - E[\hat{\theta}] - (\theta - E[\hat{\theta}])\right)^2\right] \\
&= E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] + E\left[\left(\theta - E[\hat{\theta}]\right)^2\right] - 2\,(\theta - E[\hat{\theta}])E\left[\hat{\theta} - E[\hat{\theta}]\right] \\
&= E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] + E\left[\left(\theta - E[\hat{\theta}]\right)^2\right] \\
&= \text{Var}[\hat{\theta}] + \left(E[\hat{\theta}] - \theta\right)^2
\end{aligned}$$

since $E\left[\hat{\theta} - E[\hat{\theta}]\right]$ is zero and $E[\hat{\theta}] - \theta$ is a constant. □

If we have a choice between estimators then we might prefer to use the estimator with the smallest MSE.

**Definition 2.8** (Relative Efficiency)**.** Suppose that $T_1 = T_1(X_1, \ldots, X_n)$ and $T_2 = T_2(X_1, \ldots, X_n)$ are two estimators for $\theta$. The efficiency of $T_1$ relative to $T_2$ is

$$\text{RelEff}(T_1, T_2, \theta) \quad = \quad \frac{\text{MSE}(T_2, \theta)}{\text{MSE}(T_1, \theta)}.$$

Hopefully $\text{RelEff}(T_1, T_2, \theta)$ does not depend on the true unknown $\theta$, in which case we simply write $\text{RelEff}(T_1, T_2)$. Values of $\text{RelEff}(T_1, T_2)$ less than 1 then suggest a preference for the estimator $T_2$ over $T_1$ while values greater than 1 of $\text{RelEff}(T_1, T_2)$ suggest a preference for $T_1$. Notice that if $T_1$ and $T_2$ are unbiased estimators then

$$\text{RelEff}(T_1, T_2, \theta) \quad = \quad \frac{\text{Var}(T_2)}{\text{Var}(T_1)},$$

and we choose the estimator with the smallest variance.

We now consider two examples.

**Example 2.15** (Two estimators of the mean of a normal distribution)**.** In Examples 2.11 and 2.12 we showed that the sample mean, $\overline{X}$, is unbiased and has variance $\sigma^2/n$. Suppose that $X_1, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma^2)$. Then since for the normal the mean and median are identical, an alternative estimator for $\mu$ is the sample median, $\text{med}\{X_1, \ldots, X_n\}$. It turns out that for the normal distribution, and with large $n$, the sample median is approximately normally distributed with mean $\mu$ and variance $\pi\sigma^2/(2n)$. The relative efficiency of $T_1 = \overline{X}$ and $T_2 = \text{med}\{X_1, \ldots, X_n\}$ is thus

$$\text{RelEff}(T_1, T_2) \quad = \quad \frac{\pi\sigma^2/(2n)}{\sigma^2/n} = \pi/2.$$

Thus under the normal model, we would prefer $\overline{X}$ for estimating $\mu$ to the sample median.

In the next example we will consider the relative efficiency of two alternative estimators for the variance $\sigma^2$ in the normal $N(\mu, \sigma^2)$ model. These are:

- the unbiased sample variance $S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- the biased method of moments estimator $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

To proceed we will need the variance of both $S^2$ and $\hat{\sigma}^2$. In Section 3.2, as a consequence of Theorem 3.1, we will demonstrate that

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

Noting that

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

it follows that

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

We are now in a position to calculate the relative efficiency of the two estimators.

**Example 2.16** (Two estimators of the variance of a normal distribution)**.** Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. From Example 2.14, $T_1 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is a biased estimator of $\sigma^2$ with

$$\mathrm{Bias}(T_1) = (1 - n^{-1})\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

Hence,

$$\begin{aligned}\mathrm{MSE}(T_1) &= Var(T_1) + \mathrm{Bias}(T_1)^2 \\ &= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.\end{aligned}$$

An unbiased estimator of $\sigma^2$ is $T_2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Thus, $\mathrm{MSE}(T_2) = \mathrm{Var}(T_2) = 2\sigma^4/(n-1)$. Hence, the relative efficiency of $T_1$ to $T_2$ is

$$\begin{aligned}\mathrm{RelEff}(T_1, T_2) &= \frac{\mathrm{MSE}(T_2)}{\mathrm{MSE}(T_1)} \\ &= \frac{2\sigma^4/(n-1)}{(2n-1)\sigma^4/n^2} \\ &= \frac{2n^2}{(2n-1)(n-1)} > 1 \text{ if } n > 1/3.\end{aligned}$$

Although $T_1$ is biased, it is more concentrated around $\sigma^2$ than $T_2$.

### 2.3.3 Convergence in probability and consistency

Bias and MSE are criteria for a fixed sample size $n$. We might also be interested in large sample properties. Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator for $\theta$ based on a sample of size $n$, $X_1, \ldots, X_n$. What can we say about $T_n$ as $n \to \infty$? It might be desirable if, roughly speaking, the larger $n$ is, the 'closer' $T_n$ is to $\theta$.

The behaviour of estimators when the sample size increases to infinity is referred to as the asymptotic behaviour. Whilst we will never have an infinite amount of data in reality, we might consider asymptotic results as approximations to the finite $n$ behaviour, always mindful however that the approximation may not necessarily be good.

**Definition 2.9** (Convergence in probability)**.** A sequence of random variables $(A_n)_{n\in\mathbb{N}} = (A_1, A_2, \dots)$ is said to converge in probability to a random variable $A$, denoted $A_n \xrightarrow{P} A$, if for every $\epsilon > 0$

$$P(|A_n - A| < \epsilon) \to 1 \text{ as } n \to \infty. \tag{2.3}$$

Note that an equivalent condition to (2.3) is that

$$P(|A_n - A| \geq \epsilon) \to 0 \text{ as } n \to \infty.$$

By considering the random variable $A$ with distribution $P(A = c) = 1$, Definition 2.9 includes the case where the limiting random variable, $A$, is a constant, $c$. This will be useful for our purposes: in estimation problems it is desirable for the sequence of estimators of a parameter to converge in probability to the parameter.

**Definition 2.10** (Consistent estimator)**.** An estimator $T_n = T_n(X_1, \dots, X_n)$ is consistent for the parameter $\theta$ if the sequence $(T_n)_{n\in\mathbb{N}}$ satisfies $T_n \xrightarrow{P} \theta$ for every $\theta \in \Theta$.

Thus, an estimator is consistent if it is possible to get arbitrarily close to $\theta$ by taking the sample size $n$ sufficiently large. Now, from Chebyshev's inequality, Theorem 2.2, we have a lower bound for $P(|T_n - \theta| < \epsilon)$, while 1 is an upper bound, so that

$$1 \;\geq\; P(|T_n - \theta| < \epsilon) \;\geq 1 - \frac{MSE(T_n)}{\epsilon^2}.$$

Hence, a sufficient condition for consistency of the estimator $T_n$ is that

$$\lim_{n\to\infty} MSE(T_n) \;=\; 0.$$

As $MSE(T_n) = Var(T_n) + \text{Bias}(T_n)^2$ then a sufficient condition that $\lim_{n\to\infty} MSE(T_n) = 0$, and thus for consistency, is that both

$$\lim_{n\to\infty} \text{Bias}(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} \text{Var}(T_n) = 0.$$

**Example 2.17** (Consistency of method of moments and MLE variance estimators for the normal)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. We previously showed that for $T_1 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$:

$$\mathrm{Bias}(T_1) = -\frac{\sigma^2}{n}; \qquad \mathrm{Var}(T_1) = \frac{2(n-1)\sigma^4}{n^2}.$$

Since both of these tend to zero as $n \to \infty$, it follows that $T_1$ is consistent for $\sigma^2$. For the $S^2$, we previously showed $S^2$ is unbiased and also has variance which goes to zero as $n \to \infty$, so $S^2$ is also consistent for $\sigma^2$.

**Proposition 2.1** (Consistency of sample mean, weak law of large numbers)**.** *Suppose $X_1, \ldots, X_n$ are i.i.d. with mean $\mu$ and finite variance $\sigma^2$. Then $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is a consistent estimator for $\mu$.*

**Proof:** We showed in Example 2.11 that the sample mean $\overline{X}_n$ is unbiased for $\mu$. In Example 2.12 we showed it has variance $\sigma^2/n$. Let $(\overline{X}_n)_{n\in\mathbb{N}}$ denote the corresponding sequence of estimators. Since the bias is zero, and the variance tends to zero as $n \to \infty$, it follows that the mean squared error of $\overline{X}_n$ goes to zero as $n \to \infty$. Thus $\overline{X}_n \xrightarrow{P} \mu$. □

Next we will show that the MLE in the exponential model is consistent. To do this, we first state an important theorem concerning the limit preserving features of continuous maps.

**Theorem 2.4** (Continuous Mapping Theorem)**.** *Suppose the sequence of random variables $(A_n)_{n\in\mathbb{N}} = (A_1, A_2, \ldots)$ are such that $A_n \xrightarrow{P} A$, and that $g(\cdot)$ is a continuous function. Then $g(A_n) \xrightarrow{P} g(A)$.*

**Proof: (Non-examinable)** If $g(\cdot)$ is continuous then given $\epsilon > 0$ there exists a $\delta > 0$ such that $|a_n - a| < \delta$ implies $|g(a_n) - g(a)| < \epsilon$. Consequently,

$$P(|g(A_n) - g(A)| < \epsilon) \quad \geq \quad P(|A_n - A| < \delta).$$

From Definition 2.9, for every $\delta$, $P(|A_n - A| < \delta) \to 1$ as $n \to \infty$. Thus, $P(|g(A_n) - g(A)| < \epsilon) \to 1$ as $n \to \infty$. As the choice of $\epsilon$ was arbitrary, this must hold for every $\epsilon > 0$. □

**Example 2.18** (Consistency of MLE in the exponential model)**.** Recall from Example 2.4 that the maximum likelihood estimator for $\lambda$ in the exponential model is $1/\overline{x}$. For the exponential distribution, $E(X) = \lambda^{-1}$ and $\mathrm{Var}(X) = \lambda^{-2} < \infty$, so by the weak law of large numbers $\overline{X}_n \xrightarrow{P} \lambda^{-1}$. Since $g(x) = x^{-1}$ is continuous, by Theorem 2.4 it then follows that

$$\hat{\lambda} \quad = \quad \frac{1}{\overline{X}} \xrightarrow{P} \frac{1}{\lambda^{-1}} = \lambda.$$

Thus the MLE is consistent (although not unbiased) for $\lambda$.

## 2.4 Robustness to model misspecification

Up to now we have reasoned under the assumption that our posited parametric model is correctly specified. By this, we mean that the true data generating distribution belongs to the class of densities characterised by our model. As we will see in Chapter 5.4.2, it is possible to develop diagnostics and statistical tests to assess whether various model assumptions are reasonable in light of the observed data. Nevertheless, these have certain limitations, and so it is desirable that, if possible, our inferences are robust to certain model misspecifications.

A number of the estimators we have already seen are robust to misspecification, in the sense that properties such as unbiasedness and consistency are retained even if we relax some of the modelling assumptions. For example, it follows from Proposition 2.1 that the consistency of the method of moments and MLE $\hat{\mu}$ in the normal model is robust to violations of the normality assumption.

**Example 2.19.** Recall the histogram Figure 1.1 of BMI values from the NHANES study. The histogram suggests a normality asssumption for the population distribution of BMI values is questionable. Nevertheless, in terms of point estimation for the mean $\mu$, we know from the preceding results that the MLE for the mean, which for the normal model is simply the sample mean, remains consistent for the population mean even if the population distribution of BMI values is not normal.

In Example 2.17 we proved that both $T_1$ and $S^2$ are consistent for $\sigma^2$ under the normal model. The normal model was used to derive expressions for the variance of these two estimators, which we then saw tended to zero as $n \to \infty$. In practice the normality assumption might not hold. In this case, can we give similar guarantees about consistency of $T_1$ and $S^2$? Before pursuing this further, we describe a further theorem which strengthens our ability to prove consistency.

**Theorem 2.5.** *Let $A_n$ and $B_n$ denote sequences of random variables with $A_n \xrightarrow{P} a$ and $B_n \xrightarrow{P} b$. Then*

$$\begin{aligned} A_n + B_n &\xrightarrow{P} a + b; \\ A_n B_n &\xrightarrow{P} ab; \\ \text{and} \quad \frac{A_n}{B_n} &\xrightarrow{P} \frac{a}{b} \text{ provided } b \neq 0. \end{aligned}$$

**Proof:** Omitted. See Chapter 2 of Lehmann if you are interested. □

We are now in a position to prove that both $T_1$ and $S^2$ are consistent for $\sigma^2$ without the normality assumption.

**Proposition 2.2.** *Let $X_1, \dots, X_n$ be i.i.d. with $E(X^4) < \infty$. Then $T_1$ and $S^2$ are both consistent estimators of $\sigma^2$. Furthermore, $\sqrt{T_1}$ and $S$ are consistent estimators of $\sigma$.*

**Proof:** We first show that $T_1$ is consistent for $\sigma^2$, now without making any further distributional assumptions. We can express $T_1$ as

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \overline{X}^2.$$

By the weak law of large numbers (applied to $X^2$, requiring a finite 4th moment), the first term converges in probability to $E(X^2)$. By the Continuous Mapping Theorem, $\overline{X}^2 \xrightarrow{P} \mu^2$. By Theorem 2.5 it thus follows that $T_1 \xrightarrow{P} E(X^2) - \mu^2 = \mathrm{Var}(X)$. For $S^2$, since

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{1 - 1/n} T_1$$

and $\frac{1}{1-1/n} \to 1$ as $n \to \infty$, it follows that $S^2 \xrightarrow{P} 1 \times \sigma^2 = \sigma^2$. Moreover, by the Continuous Mapping Theorem, it follows that $\sqrt{T_1} \xrightarrow{P} \sigma$ and $S \xrightarrow{P} \sigma$ since the square root function is continuous. □

# Chapter 3

# Confidence intervals

A point estimate $\hat{\theta}$ gives our best guess of the unknown parameter $\theta$ based on the data. Unless we have very large samples, there will be some error in our estimates. In this chapter we introduce confidence intervals, which are used to communicate uncertainty in point estimates. Confidence intervals are random intervals that with a specified probability will include the true, unknown parameter value. They are useful to report along with point estimates to give an indication of how precise the point estimate is and to give a range of values of the unknown parameter that are plausible given the data.

## 3.1 Principles of confidence interval construction

We now consider construction of confidence intervals (CIs).

**Definition 3.1** (Confidence interval)**.** Let $X_1, \ldots, X_n$ be a sample of data, and $\theta$ denote an unknown parameter. Suppose that the random interval

$$(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n))$$

contains $\theta$ with probability $1 - \alpha$. Then we say that this interval is a $100(1-\alpha)\%$ confidence interval for $\theta$, and has coverage level $100(1-\alpha)\%$. With observed data $x_1, \ldots, x_n$, the realised value of this interval is

$$(g_1(x_1, \ldots, x_n), g_2(x_1, \ldots, x_n))\,.$$

The next question is, how can we construct confidence intervals? One approach is based on a quantity known as a pivot.

**Definition 3.2** (Pivot)**.** Let $X_1, \ldots, X_n$ be a random sample from $f(x|\theta)$ with $\theta$ unknown. A random variable $\phi(X_1, \ldots, X_n, \theta)$ is called a pivot if its distribution does not depend on $\theta$.

**Example 3.1.** Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. There are (at most) two parameters: $\mu$ and $\sigma^2$. Note that, given $\mu$ and $\sigma^2$,

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

and $N(0, 1)$ does not depend upon either $\mu$ or $\sigma^2$. Thus, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is a pivot.

If we have a pivot then based on its known distribution, we find values $c_1$ and $c_2$ such that

$$P(c_1 < \phi(X_1, \ldots, X_n, \theta) < c_2) = 1 - \alpha.$$

The key idea here is that since the distribution of $\phi(\cdot)$ does not depend on $\theta$, the value $c_1$ and $c_2$ also do not depend on $\theta$.

We then attempt to re-arrange the inequality so that

$$P\left(g_1(X_1, \ldots, X_n) < \theta < g_2(X_1, \ldots, X_n)\right) = 1 - \alpha$$

for some functions $g_1(\cdot)$ and $g_2(\cdot)$. In this case we have found a $100(1 - \alpha)\%$ confidence interval for $\theta$.

**Example 3.2** (Confidence interval for normal mean, variance known)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma^2)$ with $\sigma^2$ known. From Example 3.1 we have the pivot

$$\phi(X_1, \ldots, X_n, \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Let $c_1$ and $c_2$ be constants such that

$$P\left(c_1 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < c_2\right) = 1 - \alpha.$$

Re-arranging we have that

$$P\left(\overline{X} - c_2\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - c_1\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

and so $\left(\overline{X} - c_2\frac{\sigma}{\sqrt{n}}, \overline{X} - c_1\frac{\sigma}{\sqrt{n}}\right)$ is a $100(1-\alpha)$ confidence interval for $\mu$. Typically we choose $c_1$ and $c_2$ to form a symmetric interval around $\overline{X}$, so that $c_1 = -c_2$. Then we can write the confidence interval as

$$\overline{X} \pm c_2\frac{\sigma}{\sqrt{n}} \tag{3.1}$$

with $c_2 = z_{1-\alpha/2}$, where $z_p$ denotes the $p \times 100\%$ lower quantile of the standard normal distribution. This leads to the z-interval:

$$\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = \left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{3.2}$$

**Remark: How do we find the quantiles and probabilities of the normal distribution?**

If $Z \sim N(0,1)$, let $P(Z \leq z_p) = p$.

- In the University Formula Book, Section B2 tabulates the normal distribution function $\Phi(z) = P(Z \leq z) = P(Z < z)$. For a given $z$ find the row corresponding to the first decimal place of $z$ and then the column corresponding to the second decimal place and the intersection of these gives the corresponding $z$ value. To find $\Phi(z)$ for a given $z$, find the nearest value to $\Phi(z)$ in the inner table and the corresponding row gives the first decimal place of $z$ with the row giving the second decimal place. From the table you should be able to confirm, for example, that $P(Z \leq 1.25) = 0.8944$.
- In R, for a given $q$, $z_p$ may be found using the command `qnorm(p)`. For a given $z_p$, $p$ may be found using the command `pnorm(`$z_p$`)`

```
qnorm(0.8944) # finds z such that P(Z <= z) = 0.8944
```

```
## [1] 1.250273
```

```
pnorm(1.25)    # finds P(Z <= 1.25)
```

```
## [1] 0.8943502
```

```
pnorm(qnorm(0.8944))
```

```
## [1] 0.8944
```

- To compute an upper tail probability, say $P(Z > z_p)$, then you can either use the result that $P(Z > z_p) = 1 - P(Z \leq z_p)$ or change the default in `qnorm` or `pnorm` to calculate the upper tail.

```
1-pnorm(1.25)  # P(Z > 1.25) = 1 - P(Z <= 1.25)
```

```
## [1] 0.1056498
```

```
pnorm(1.25, lower.tail=F) # using the upper tail finds P(Z > 1.25)
```

```
## [1] 0.1056498
```

```
qnorm(0.8944, lower.tail=F) # finds z such that P(Z > z) = 0.8944
```

```
## [1] -1.250273
```

**How should we choose $\alpha$?** The statistician Sir Ronald Fisher arbitrarily suggested $\alpha = 0.05$, and empirical research in many disciplines across the world almost universally uses this by convention. For $\alpha = 0.05$ we have $z_{0.975} \approx 1.96$. Thus we have that a **95% confidence interval for** $\mu$ is given by

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

For a **90% confidence interval for** $\mu$ we have $\alpha = 0.1$ and $z_{0.95} = 1.645$ giving

$$\left(\overline{X} - 1.645\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.645\frac{\sigma}{\sqrt{n}}\right)$$

which is a **narrower** interval than the 95% confidence interval. We observe that:

- To increase confidence, we need to **widen** the interval.

Note that, as $\sigma^2$ is known, then we can compute a realisation of this interval by plugging-in the observed value $\overline{x}$ of $\overline{X}$.

There are some important points to remember when interpreting confidence intervals.

- A realised confidence interval either contains the true parameter value $\theta$ or it does not, but we do not know which.

- It is therefore incorrect to speak of the probability that the realised interval contains $\theta$. To do this one must adopt a Bayesian inferential approach, which we do not cover in this unit but is covered in MA32023 Bayesian statistics.

- In most situations it is reasonable to interpret the width of the confidence interval as measuring the precision of the estimate, and the range of values in the interval to show which parameter values are consistent with the data seen (under the assumed model). Such interpretations have though been the subject of criticism by some (if you are interested to read more, see The fallacy of placing confidence in confidence intervals, 2016).

## 3.2 Confidence intervals for the normal distribution

Suppose we have an i.i.d. sample $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$. In the preceding we considered construction of a confidence interval for $\mu$ when $\sigma^2$ was known.

### 3.2.1 Confidence interval for the variance with unknown mean

In this section we consider construction of a confidence interval for the variance $\sigma^2$, when this and $\mu$ are unknown. Recall the unbiased estimator of $\sigma^2$, $S^2$ :

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

**Definition 3.3** (Chi-squared distribution)**.** If $Z \sim N(0,1)$ then the distribution of $U = Z^2$ is called the chi-squared distribution with one degree of freedom, written $\chi^2_1$. If $U_1, \ldots, U_\nu$ are i.i.d. $\chi^2_1$ then $V = \sum_{i=1}^{\nu} U_i \sim \chi^2_\nu$, the chi-squared distribution with $\nu$ degrees of freedom.

Note that, see your Probability & Statistics 1B notes, that the $\chi^2$ distribution is a special case of the gamma distribution: $\chi^2_\nu = Ga(1/2, \nu/2)$. It follows from the properties of the gamma distribution that $E(\chi^2_\nu) = \nu$ and $\text{Var}(\chi^2_\nu) = 2\nu$.

**Theorem 3.1.** *Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Then $\overline{X}$ and $S^2$ are independent and*

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

**Proof:** Not examinable. See Chapter 8. □

We can now confirm the result used for the variance of $S^2$ used in Example 2.16. As

$$\text{Var}\left[\frac{(n-1)S^2}{\sigma^2}\right] \quad = \quad 2(n-1)$$

then it follows that

$$\text{Var}(S^2) \quad = \quad \frac{2\sigma^4}{n-1}.$$

Recall from Theorem 3.1 that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. It follows that $\frac{(n-1)S^2}{\sigma^2}$ is a **pivot** for $\sigma^2$. We can thus find values $c_1$ and $c_2$ such that

$$P\left(c_1 < \frac{(n-1)S^2}{\sigma^2} < c_2\right) = 1 - \alpha.$$

Rearranging gives

$$P\left(\frac{(n-1)S^2}{c_2} < \sigma^2 < \frac{(n-1)S^2}{c_1}\right) = 1 - \alpha$$

which allows us to form a $100(1-\alpha)\%$ confidence interval for $\sigma^2$. We must decide how to choose $c_1$ and $c_2$. The chi-squared distribution is not symmetric (see Figure 3.1), so we cannot take the same approach we did when constructing the confidence interval for the normal mean.

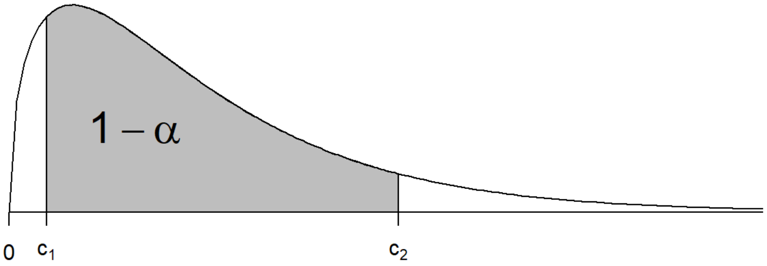


Figure 3.1: Probability density function of the chi-squared distribution.

The standard approach is to choose $c_1$ and $c_2$ so that

$$\begin{aligned} P\left(\frac{(n-1)S^2}{\sigma^2} < c_1\right) &= \alpha/2, \text{ and} \\ P\left(\frac{(n-1)S^2}{\sigma^2} > c_2\right) &= \alpha/2 \end{aligned}$$

Thus, we choose

$$c_1 = \chi^2_{n-1,\alpha/2}, \qquad c_2 = \chi^2_{n-1,1-\alpha/2}$$

where $P(V \leq \chi^2_{\nu,p}) = p$ for $V \sim \chi^2_\nu$. Thus a $100 \times (1-\alpha)\%$ confidence interval for $\sigma^2$ is

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right).$$

**Remark: How do we find the quantiles and probabilities of the chi-squared distribution?**

If $V \sim \chi^2_\nu$ let $P(V \leq \chi^2_{\nu,p}) = p$.

- In the University Formula Book, Section B4 tabulates the percentage points of the $\chi^2$-distribution. Note that it gives **upper-tail values** and so you will need to take care that you find the correct value. In the table, the rows correspond to the degrees of freedom, $\nu$ and the columns the upper-tail probability. Thus, if you want a lower tail probability of $p$ this corresponds to an upper tail probability of $1-p$. For example, $P(\chi^8 \leq 2.180) = 0.025$ since (using the table) $P(\chi^8 > 2.180) = 0.975$. Similarly, $P(\chi^8 \leq 17.535) = 0.975$ since (using the table) $P(\chi^8 > 17.535) = 0.025$.

- In `R`, for a given $\nu$ and $p$, $\chi^2_{\nu,p}$ may be found using the command `qchisq`$(p,\nu)$. For a given $\chi^2_{\nu,p}$, $p$ may be found using the command `pchisq`$(\chi^2_{\nu,p},\nu)$.

```
# Let V be a chi-squared distribution with 8 degrees of freedom
qchisq(0.025, 8) # finds v such that P(V <= v) = 0.025
```

```
## [1] 2.179731
```

```
pchisq(2.180, 8)   # finds P(V <= 2.180)
```

```
## [1] 0.02500977
```

- To directly compute an upper tail probability, or corresponding quantile, in `R` you need to change the default in `pchisq` or `qchisq` to calculate the upper tail.

```
# Let Y be a chi-squared distribution with 8 degrees of freedom
qchisq(0.025, 8, lower.tail=F) # finds y such that P(Y > y) = 0.025
```

```
## [1] 17.53455
```

```
pchisq(15.535, 8, lower.tail=F) # using the upper tail finds P(Y > 15.535)
```

```
## [1] 0.04954038
```

**Example 3.3** (Confidence interval for variance of BMI in NHANES)**.** Recall the NHANES BMI data from Example 1.7. BMI data are available on $n = 1638$ observations. The sample variance was $s^2 = 46.51$. Assuming the population distribution of BMI is normal, and using the function `qchisq` in R to find the required quantiles of the chi-squared distribution, we can construct a 95% confidence interval for the population variance of BMI as

$$\left(\frac{1637 \times 46.51}{1751.029}, \frac{1637 \times 46.51}{1526.759}\right) \quad = \quad (43.482, 49.869).$$

This confidence interval was constructed assuming the population distribution of BMI values is normal. Given Figure 1.1 we should therefore be cautious about our results given that the normality assumption is somewhat tenable.

### 3.2.2 Confidence interval for the mean with unknown variance

In Example 3.2, we constructed a $100(1-\alpha)\%$ confidence interval for a normal mean $\mu$ when the variance was assumed to be known. This result is not practically very useful because it would be rare that we would know the true value of $\sigma^2$. We now consider the more realistic case where both $\mu$ and $\sigma^2$ are unknown. To do this, we introduce a new distribution, the t-distribution.

**Definition 3.4** (t-distribution)**.** If $Z \sim N(0,1)$, $U \sim \chi^2_\nu$, and $Z$ and $U$ are independent, then the distribution of $\frac{Z}{\sqrt{U/\nu}}$ is called the t-distribution with $\nu$ degrees of freedom.

As shown in Figure 3.2, the t-distribution is symmetric around zero, and looks similar to the normal. It differs from the normal in that it has fatter tails, meaning there is more probability on values far away from the center than with the normal distribution. As the degrees of freedom increase, the t-distribution density converges to the standard normal density.

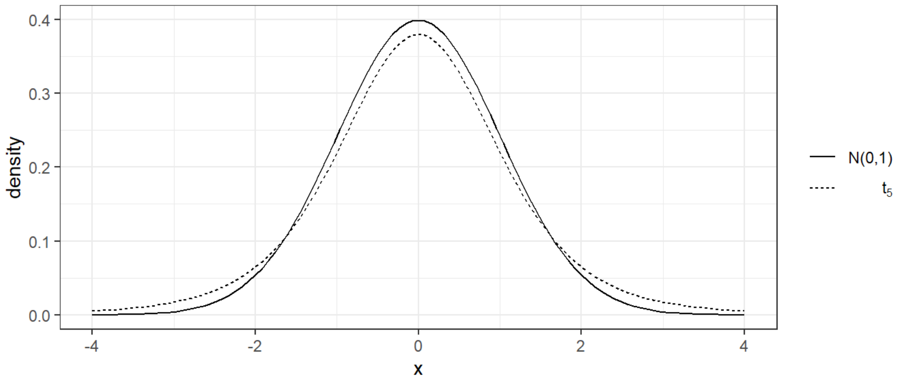


Figure 3.2: Standard normal and t-distribution (5 d.f.) density functions.

Recall that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Now we note that

$$\frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}.$$

Lastly, we recall that $\overline{X}$ and $S^2$ are independent. We have thus shown that

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

This quantity is thus a pivot for $\mu$, enabling us to form a confidence interval. Proceeding analogously to the case where $\sigma^2$ was known, suppose we have chosen constants $c_1$ and $c_2$ such that

$$P(c_1 < t_{n-1} < c_2) = 1 - \alpha.$$

Then we have that

$$P\left(c_1 < \frac{\overline{X} - \mu}{S/\sqrt{n}} < c_2\right) = 1 - \alpha$$

and as before re-arranging we obtain

$$P\left(\overline{X} - c_2\frac{S}{\sqrt{n}} < \mu < \overline{X} - c_1\frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$

thus allowing us to form a $100(1-\alpha)\%$ confidence interval for $\mu$. If we choose to construct a symmetric interval, we can choose $c_1 = -c_2$. If $t_\nu$ is a $t$-distribution with $\nu$ degrees of freedom, let $t_{\nu,p}$ be such that $P(t_\nu \leq t_{\nu,p}) = p$. We choose $c_2 = t_{n-1,1-\alpha/2}$ so that our $100(1-\alpha)\%$ confidence interval, sometimes called a t-interval, for $\mu$ is given by

$$\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right) \tag{3.3}$$

or equivalently $\overline{X} \pm t_{n-1,1-\alpha/2}S/\sqrt{n}$.

**Remark: How do we find the quantiles and probabilities of the t-distribution?**

If $t_\nu$ is a $t$-distribution with $\nu$ degrees of freedom, let $P(t_\nu \leq t_{\nu,p}) = p$.

- In the University Formula Book, Section B3 tabulates the percentage points of the $t$-distribution. Note that it gives **upper-tail values** and so you will need to take care that you find the correct value. In the table, the rows correspond to the degrees of freedom and the columns the upper-tail probability. Thus, if you want a lower tail probability of $p$ this corresponds to an upper tail probability of $1-p$. For example, $P(t_8 \leq 2.896) = 0.99$ since (using the table) $P(t_8 > 2.896) = 0.01$.
- In R, for a given $p$, $t_{\nu,p}$ may be found using the command `qt`$(p, \nu)$. For a given $t_{\nu,p}$, $p$ may be found using the command `pt`$(t_{\nu,p}, \nu)$

```r
# Let T be a t-distribution with 8 degrees of freedom
qt(0.99, 8) # finds t such that P(T <= t) = 0.99
```

```
## [1] 2.896459
```

```r
pt(2.896, 8)   # finds P(T <= 2.896)
```

```
## [1] 0.989993
```

- To directly compute an upper tail probability, or corresponding quantile, in `R` you need to change the default in `pt` or `qt` to calculate the upper tail.

```
# Let T be a t-distribution with 8 degrees of freedom
qt(0.01, 8, lower.tail=F) # finds t such that P(T > t) = 0.01
```

```
## [1] 2.896459
```

```
pt(2.896, 8, lower.tail=F) # using the upper tail finds P(T > 2.896)
```

```
## [1] 0.01000705
```

**Example 3.4** (Confidence interval for mean BMI in NHANES)**.** In Example 2.9, we estimated the population mean BMI from the NHANES data by its sample mean $\hat{\mu} = \overline{x} = 27.911$. This was based on $n = 1638$ observations. The sample standard deviation was $s = 6.82$.

The confidence interval limits for a 95% confidence interval for the mean are given by

$$\left(27.911 - t_{1637,0.975}\frac{6.82}{\sqrt{1638}}, 27.911 + t_{1637,0.975}\frac{6.82}{\sqrt{1638}}\right).$$

We calculate $t_{1637,0.975}$ in `R` using `qt(p = 0.975, df = 1637)`. Then the above expression evaluates to

$$(27.581, 28.242)$$

Again, given Figure 1.1 which suggests the distribution of BMI values is not normal, we should be concerned about whether the normality assumption made in the construction of this confidence interval for the mean is reasonable.

## 3.3 Confidence intervals for the exponential distribution

Consider an i.i.d. sample from the exponential distribution with rate parameter $\lambda$. We showed in Example 2.4 that the MLE for $\lambda$ is $\hat{\lambda} = 1/\overline{x}$. We now consider how to form a confidence interval for $\lambda$. We first note that the exponential is a special case of the Gamma distribution. Specifically, a Gamma distribution with shape 1 and rate $\lambda$, which we will denote by $Ga(\lambda, 1)$, is exponentially distributed with rate $\lambda$. We will need to use the following result from probability theory:

**Theorem 3.2.** *Let $X_1, \ldots, X_n$ be i.i.d. from $Ga(\lambda, k)$ with shape parameter $k > 0$. Then for $c > 0$,*

$$c \sum_{i=1}^{n} X_i \sim Ga(\lambda/c, nk)$$

**Proof:** See your Probability & Statistics 1B notes for the proof. □

It follows from Theorem 3.2 that if $X_1, \ldots, X_n$ are i.i.d. $Exp(\lambda) = Ga(\lambda, 1)$

$$\sum_{i=1}^{n} X_i \sim Ga(\lambda, n) \quad \text{and so} \quad \lambda \sum_{i=1}^{n} X_i \sim Ga(1, n)$$

and thus we have a pivot for $\lambda$. In particular

$$\alpha/2 = P\left(\lambda n\overline{X} < Ga_{1,n,\alpha/2}\right) = P\left(\lambda < \frac{Ga_{1,n,\alpha/2}}{n\overline{X}}\right)$$

where $Ga_{1,n,\alpha/2}$ denotes the $\alpha/2$ percentile of the Gamma distribution with shape $n$ and rate 1, and

$$\alpha/2 = P(\lambda n\overline{X} > Ga_{1,n,1-\alpha/2}) = P\left(\lambda > \frac{Ga_{1,n,1-\alpha/2}}{n\overline{X}}\right)$$

so we can construct an equitailed $100 \times (1 - \alpha)\%$ confidence interval for $\lambda$ as

$$\left(\frac{Ga_{1,n,\alpha/2}}{n\overline{X}}, \frac{Ga_{1,n,1-\alpha/2}}{n\overline{X}}\right).$$

## 3.4 Robustness to model misspecification

The confidence intervals we have just constructed relied on parametric assumptions that the individual observations (in the population) are normally or exponentially distributed. Note that this is an assumption about the data generating mechanism or the infinite population from which we have sampled – it is not a property of the sample of observed data. Thus although there exist various methods for checking modelling assumptions, we can never prove from the observed data that a particular modelling assumption is true. It is therefore extremely useful if we can derive methods which are valid (usually approximately) under weaker modelling assumptions.

To make progress in this direction, we recall the Central Limit Theorem (CLT) from Probability & Statistics 1B.

**Theorem 3.3** (Central Limit Theorem (CLT))**.** *Let $X_1, \dots, X_n$ be i.i.d. from a distribution with finite mean $\mu$ and finite variance $\sigma^2$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then as $n \to \infty$*

$$\sqrt{n}(\overline{X}_n - \mu)/\sigma \approx N(0,1)$$

**Proof:** A sketch proof can be found in Appendix A.5 of Lehmann, and references therein. □

The convergence being asserted by the CLT is known as convergence in distribution, or convergence in law.

**Definition 3.5** (Convergence in law)**.** Let $(A_n)_{n \in \mathbb{N}}$ be a sequence of random variables with corresponding cumulative distribution functions $F_n(\cdot)$, and let $A$ have cumulative distribution function $F(\cdot)$. Suppose that $F_n(a) \to F(a)$ at all continuity points $a$ of $F(\cdot)$. Then we say that $A_n$ converges in distribution or law to $A$, which we write as

$$A_n \xrightarrow{L} A$$

If the distribution of $A$ is a commonly-used distribution, e.g. $N(0,1)$, it is convention (although notationally not really correct) to write this as $A_n \xrightarrow{L} N(0,1)$. Now we can more precisely restate the final line in the CLT as

$$\sqrt{n}(\overline{X}_n - \mu)/\sigma \xrightarrow{L} N(0,1).$$

If $\sigma^2$ were known, the CLT enables us to use the same confidence interval that we constructed in Example 3.2 for the mean of a normal distribution and apply it to situations where the normality assumption may not hold, i.e. for large $n$

$$P\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

We will consider how good the approximation is shortly.

The usefulness of this result is limited by the fact that in practice the variance $\sigma^2$ is typically not known. We know that we can estimate $\sigma^2$ unbiasedly by $S^2$. One might therefore be tempted to use the preceding confidence interval, replacing the unknown $\sigma$ by the estimate $S$. The question is, can this be justified? As long as $S$ is consistent, the answer turns out to be ‘yes’, due to Slutsky’s Theorem.

**Theorem 3.4** (Slutsky’s Theorem)**.** *If $Y_n \xrightarrow{L} Y$, $A_n \xrightarrow{P} a$, $B_n \xrightarrow{P} b$, then*

$$A_n + B_n Y_n \xrightarrow{L} a + bY$$

**Proof:** Omitted. See Theorem A.14.9 of Bickel and Doksum. □

We can now use Slutsky's Theorem to justify replacing the unknown $\sigma$ by $\hat{\sigma}$ in the Central Limit Theorem.

**Proposition 3.1.** *Let $X_1, \dots, X_n$ be i.i.d. from a distribution with finite mean $\mu$ and finite variance $\sigma^2$. Let $\hat{\sigma}$ be a consistent estimator of $\sigma$. Then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\hat{\sigma}} \xrightarrow{L} N(0, 1).$$

**Sketch proof:** Apply the CLT to $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ and then multiply by $\sigma/\hat{\sigma}$ and invoke Slutsky. □

And finally we can justify replacing $\sigma$ by $S$ in the z-interval, assuming $S$ is a consistent estimator for $\sigma$ (see conditions in Proposition 2.2).

**Proposition 3.2.** *Let $X_1, \dots, X_n$ be i.i.d. from a distribution with mean $\mu$ and finite variance $\sigma^2$. Assume $S$ is a consistent estimator for $\sigma$. Then the interval*

$$\left(\overline{X} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) \tag{3.4}$$

*is asymptotically an $100(1-\alpha)\%$ confidence interval for $\mu$.*

**Sketch Proof:** Applying proposition 3.1 and the consistency of $S$, $\frac{\sqrt{n}(\overline{X}_n - \mu)}{S}$ is (approximately) a pivot for $\mu$, so

$$P\left(\overline{X}_n - z_{1-\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

as $n \to \infty$. □

For full proofs of these two propositions, see Chapter 8.

**Definition 3.6** (Coverage probability)**.** The coverage probability is the probability that a **procedure** for constructing confidence intervals will produce an interval containing, or covering, the true value.

The coverage probability is a property of the procedure itself rather than of a particular confidence interval produced based on a sample of data. There are no guarantees that confidence intervals constructed based on the preceding asymptotic derivations will have the nominal $100 \times (1 - \alpha)\%$ coverage for finite $n$.

Consider the confidence interval in Equation (3.4). The first thing to note is that if it were actually the case that $X_1, \dots, X_n$ were normally distributed and we estimate $\sigma$ by $S$, the t-interval

$$\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right)$$

has coverage $100(1-\alpha)\%$. Since the t-distribution has fatter tails than the normal, $t_{n-1,1-\alpha/2} > z_{1-\alpha/2}$, we can immediately deduce that were the population really normally distributed, the z-interval

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

must have coverage somewhat less than $100(1-\alpha)\%$ for finite $n$. One option is therefore to use the t-interval rather than the z-interval, because we know that in the case of normally distributed data this is the right thing to do, and asymptotically they are identical. When the normality assumption for $X_1, \dots, X_n$ does not hold, how good the approximation is will depend on how far from normality the true distribution is and the sample size $n$.

The confidence interval we have used for the mean uses $S$ to estimate $\sigma$, but the proof of Proposition 3.2 only relied on $S$ in so far as it was a consistent estimator for $\sigma$. It is thus easy to generalise this result by allowing any consistent estimator of $\sigma$ to be used.

**Example 3.5.** Let $X_1, \dots, X_n$ be i.i.d. Bernoulli($p$). Recall that for this model, $E(X) = p$ and $\text{Var}(X) = p(1-p)$. We showed in Example 2.5 that the MLE of $p$ is $\hat{p} = \overline{X}$, and (from a problems class) that $\hat{p}(1-\hat{p})$ is consistent for $\text{Var}(X) = p(1-p)$. By the Continuous Mapping Theorem (Theorem 2.4) we then have that

$$\sqrt{\hat{p}(1-\hat{p})} \xrightarrow{P} \sqrt{\text{Var}(X)}$$

Thus by Proposition 3.1 we have that

$$\frac{\overline{X}_n - p}{\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}} \xrightarrow{L} N(0,1)$$

For finite $n$, we can use this result to say that

$$P\left(-z_{1-\alpha/2} < \frac{\overline{X}_n - p}{\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}} < z_{1-\alpha/2}\right) \approx 1-\alpha$$

so that

$$P\left(\overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} < p < \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right) \approx 1-\alpha$$

An alternative to estimating $\text{Var}(X)$ by $\hat{p}(1-\hat{p})$ would be to estimate it by $S^2$. For large $n$ they will be very close.

**Example 3.6.** Recall from Example 1.8 the data on alcohol consumption from NHANES. For this variable $n = 1,104$ responded to the question, and of these, 846 (76.6%) answered that they drank more than one alcoholic drink on average on days that they did drink alcohol. We can use the i.i.d. Bernoulli model here, with $p$ representing the proportion of the population that would answer that they drink more than one alcoholic drink on average on days they drink. The MLE is $\hat{p} = 846/1104$. We can construct an asymptotic 95% confidence interval for $p$ using the limits in the inequality in Equation (3.5). This gives

$$\left(\frac{846}{1104} - 1.96\sqrt{\frac{\frac{846}{1104}(1-\frac{846}{1104})}{1104}}, \frac{846}{1104} + 1.96\sqrt{\frac{\frac{846}{1104}(1-\frac{846}{1104})}{1104}}\right),$$

which evaluates to (0.7413413,0.7912674).

## 3.5 Summary of confidence intervals

This section contains a summary of the key confidence intervals we have covered in this chapter.

### 3.5.1 Normal model

Assume that $X_1, \ldots, X_n$ are independent and identically distributed $N(\mu, \sigma^2)$ random variables.

- An unbiased estimator of $\mu$ is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ with $\overline{X} \sim N(\mu, \sigma^2/n)$.
- An unbiased estimator of $\sigma^2$ is $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ with $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

**Interval for $\mu$ when $\sigma^2$ is known (z-interval)**
Derived using the pivot $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, a $100(1-\alpha)\%$ confidence interval for $\mu$ is

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

where if $Z \sim N(0,1)$, $P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$.

**Interval for $\mu$ when $\sigma^2$ is unknown (t-interval)**
Derived using the pivot $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, a $100(1-\alpha)\%$ confidence interval for $\mu$ is

$$\left(\overline{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right)$$

where $P(t_{n-1} \leq t_{n-1,1-\alpha/2}) = 1-\alpha/2$ and $t_{n-1}$ is the $t$-distribution with $n-1$ degrees of freedom.

**Interval for $\sigma^2$ when $\mu$ is unknown**
Derived using the pivot $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, a $100(1-\alpha)\%$ confidence interval for $\sigma^2$ is

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

where $P(\chi^2_{n-1} \leq \chi^2_{n-1,\alpha/2}) = \alpha/2$, $P(\chi^2_{n-1} \leq \chi^2_{n-1,1-\alpha/2}) = 1-\alpha/2$ and $\chi^2_{n-1}$ is the chi-squared distribution with $n-1$ degrees of freedom.

### 3.5.2 Robustness of confidence intervals for the mean

If the normality assumption is removed, provided the mean and variance of the $X_i$s are finite, the z and t intervals above have the correct coverage asymptotically, and for finite $n$ will have approximately the correct coverage.

### 3.5.3 Asymptotic confidence intervals for means

Let $f(x|\theta)$ be some parametric model with finite mean $\mu$ and finite variance $\sigma^2$, each of which are functions of the model parameter(s) $\theta$. Let $\hat{\sigma}$ be a consistent estimator of $\sigma$. Then

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

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

# Chapter 4

# Hypothesis testing

## 4.1 Introduction to statistical hypothesis testing

In this chapter we introduce hypothesis testing for parameters in a statistical model. Suppose we have a model parameter $\theta \in \Theta$. A hypothesis is a statement that the true value of $\theta$ belongs in some subset of the parameter space $\Theta$. The null hypothesis, denoted $H_0$, specifies one such subset. The alternative hypothesis, denoted $H_1$, specifies another subset. In the Neyman-Pearson framework of hypothesis testing, we use the data to construct a test that will either result in our accepting the null hypothesis $H_0$ or rejecting this in favour of the alternative hypothesis $H_1$.

**Example 4.1** (Manufacturing defects)**.** Suppose a machine in a factory is known to produce defective items 20% of the time. The machine is sent for repairs and then re-introduced to the production line. The factory managers want to know if the proportion of defective products being produced by the machine has been reduced after the repairs. Letting $p$ denote the probability of an item being defective, we are interested in testing $H_0 : p = 0.2$ versus $H_1 : p < 0.2$.

**Example 4.2** (Two arm clinical trial)**.** In a two arm randomised clinical trial (c.f., Example 1.9) the true population mean outcome under control treatment is $\mu_0$ and is $\mu_1$ on the new treatment. The company developing the new treatment seeks a license to sell the new drug and needs to demonstrate to the regulatory agencies evidence in support of the hypothesis that $\mu_1 > \mu_0$. Thus the null hypothesis might be taken to be $H_0 : \mu_1 = \mu_0$ (the mean of the outcome is the same under treatment and control) and the alternative hypothesis $H_1 : \mu_1 > \mu_0$ (the mean of the outcome is greater in the treatment arm).

To construct a hypothesis test we choose a **test statistic**: a statistic

$T(X_1, \ldots, X_n)$ which is a function of the observed data. The basic idea is that the test statistic is measuring the relative compatibility of the observed data with the two hypotheses. Next we define the **critical region**.

**Definition 4.1** (Critical Region)**.** Let $\mathcal{T}$ denote the sample space of values of the test statistic $T$. The region $C \subseteq \mathcal{T}$ for which $H_0$ is rejected in favour of $H_1$ is termed the critical (or rejection) region while the region $\mathcal{T} \setminus C$, where we accept $H_0$, is called the acceptance region. The critical and acceptance regions can also equivalently be defined in terms of the subset of the sample space of the data for which we reject or accept $H_0$.

The substantive question of interest typically dictates the specification of the null and alternative hypotheses. The question is thus how we should choose the test statistic and the critical region.

## 4.2 Tests for simple hypotheses

If a hypothesis completely specifies the probability distribution of $X$ then it is said to be a **simple hypothesis**. If the hypothesis is not simple then it is said to be **composite**. In this section we will consider the case where $H_0$ and $H_1$ are both simple hypotheses. This means that $H_0$ specifies that the data are distributed according to a specific distribution, and similarly for $H_1$. The developments for the case where both $H_0$ and $H_1$ are simple are easier than the case where one or both of the hypotheses are composite, which we will consider in the next section.

### 4.2.1 Evaluating a test

In the paradigm of hypothesis testing developed by Neyman and Pearson, we reject or accept $H_0$ based on whether the value of the test statistic falls within the critical region. Unless we have an infinite amount of data, there is a non-zero probability that we may reject $H_0$ when it is in fact true or accept $H_0$ when it is actually false. Table 4.1 summarises the possible outcomes.

Table 4.1: Possible outcomes and types of error in a hypothesis test.

| | ACCEPT $H_0$ | REJECT $H_0$ |
|---|---|---|
| $H_0$ TRUE | CORRECT | ERROR (Type I) |
| $H_0$ FALSE | ERROR (Type II) | CORRECT |

**Definition 4.2** (Type I and Type II errors)**.** A Type I error occurs when $H_0$ is rejected when it is true. The probability of such an error is denoted by $\alpha$ so that

$$\alpha \ = \ P(\text{Type I error}) \ = \ P(\text{Reject } H_0 \,|\, H_0 \text{ true}).$$

A Type II error occurs when $H_0$ is accepted when it is false. The probability of such an error is denoted by $\beta$ so that

$$\beta \ = \ P(\text{Type II error}) \ = \ P(\text{Accept } H_0 \,|\, H_0 \text{ false}).$$

The **power** is defined as $1 - \beta$ and is the probability we reject $H_0$ when $H_0$ is false.

**Example 4.3.** Suppose $X_1, \ldots, X_n$ are i.i.d. normal with known variance $\sigma^2$ and mean $\mu$ either equal to $\mu_0$ or $\mu_1$ where $\mu_1 > \mu_0$. Assume in this case that $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$ and that the test statistic used is $T = \overline{X}$. Under $H_0$, $\overline{X} \sim N(\mu_0, \sigma^2/n)$ while under $H_1$, $\overline{X} \sim N(\mu_1, \sigma^2/n)$. A large observed value of $\overline{x}$ may indicate that $H_1$ rather than $H_0$ is true. Intuitively, we may consider a critical region of the form

$$C \ = \ \{\overline{x} : \overline{x} \geq c\}.$$

This critical region is shown in Figure 4.1

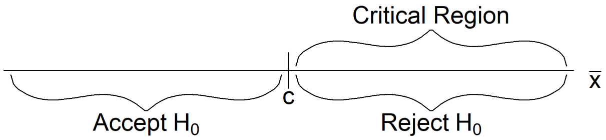


Figure 4.1: An illustration of the critical region $C = \{\overline{x} : \overline{x} \geq c\}$.

There are a number of immediate questions. How should we pick the value of $c$, and how will its choice affect the probability of making Type I and Type II errors? We would like to make the probability of either of these errors as small as possible.

Figure 4.2 shows schematically how the choice of $c$ can be expected to qualitatively alter these two probabilities.

It is clear from the figure that:

- If we INCREASE $c$ then the probability of a Type I error DECREASES but the probability of a Type II error INCREASES.

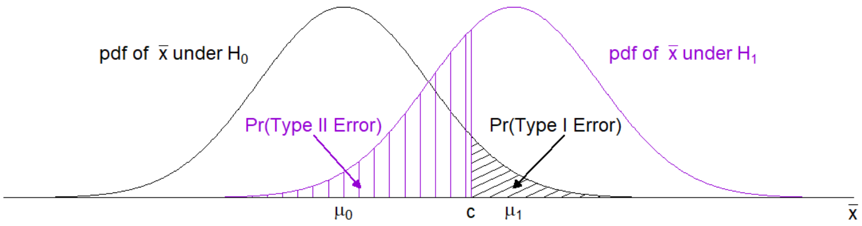


Figure 4.2: The errors resulting from the test with critical region $C = \{\bar{x} : \bar{x} \geq c\}$ of $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu_1 > \mu_0$.

- If we DECREASE $c$ then the probability of a Type I error INCREASES but the probability of a Type II error DECREASES.

It turns out, in practice, that this is always the case: in order to decrease the probability of a Type I error, we must increase the probability of a Type II error and vice versa. We deal with this issue by fixing the probability of Type I error $\alpha$ in advance.

**Definition 4.3** (Significance level/Size of the test)**.** When the probability of type I error $\alpha$ is fixed in advance of observing the data then $\alpha$ is known as the **significance level of the test**. In this context $\alpha$ is also known as the **size of the test**.

We usually fix $\alpha$ at some small value, say $\alpha = 0.1, 0.05, 0.01, \ldots$. Given a test statistic, fixing $\alpha$ determines the critical region. It is important to note that by doing this we introduce some asymmetry to the setup: controlling the probability of a Type I error implies that it is the null hypothesis (rather than the alternative) that we want to ensure is not rejected too often when it is in fact true. In most cases the choice of $\alpha$ level is arbitrary. The choice of $\alpha = 0.05$ has become essentially the world's default $\alpha$ level based on Sir Ronald Fisher's suggestion of it being a potentially reasonable value, rather than a value that should be automatically and universally used.

**Example 4.4.** In Example 4.3 we considered a critical region of the form

$$C = \{\bar{x} : \bar{x} \geq c\}.$$

Now,

$$\begin{aligned}\alpha = P(\text{Type I error}) &= P(\text{Reject } H_0 \,|\, H_0 \text{ true}) \\ &= P\left(\overline{X} \geq c \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\right) \\ &= P\left(Z \geq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right),\end{aligned}$$

where $Z \sim N(0,1)$. Suppose we choose $\alpha = 0.05$. We can then find the value of $c$ which gives us $\alpha = 0.05$:

$$\begin{aligned} P\left(Z \geq \frac{c-\mu_0}{\sigma/\sqrt{n}}\right) &= 0.05 \;\Rightarrow \\ \frac{c-\mu_0}{\sigma/\sqrt{n}} &= 1.645 \;\Rightarrow \\ c &= \mu_0 + 1.645\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Notice that as $n$ increases then $c$ tends towards $\mu_0$. Also notice how as the (assumed known) standard deviation $\sigma$ decreases, $c$ also gets closer to $\mu_0$.

Once the significance level $\alpha$ has been chosen, and hence the critical region, $\beta$ and the power $(1-\beta)$ are determined. It typically depends upon the sample size.

**Example 4.5.** Using the critical region given in Example 4.4 we find

$$\begin{aligned} \beta = P(\text{Type II error}) &= P(\text{Accept } H_0 \,|\, H_0 \text{ false}) \\ &= P\left(\overline{X} < c \,|\, \overline{X} \sim N(\mu_1, \sigma^2/n)\right) \\ &= P\left(Z < \frac{c-\mu_1}{\sigma/\sqrt{n}}\right). \end{aligned}$$

The corresponding value of $\beta$ is

$$\begin{aligned} \beta = P\left(Z < \frac{c-\mu_1}{\sigma/\sqrt{n}}\right) &= P\left(Z < \frac{(\mu_0-\mu_1) + 1.645\sigma/\sqrt{n}}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(1.645 - \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}\right) \end{aligned}$$

where $P(Z < z) = \Phi(z)$. Note that as $n$ increases then $\beta$ decreases to zero. Thus, the power is tending to 1 as $n \to \infty$.

### 4.2.2 Constructing tests

In the example above we chose the test statistic (the sample mean) solely on the basis of our intuition that large values of the sample mean are supportive of $H_1$ and not supportive of $H_0$ being true. One general approach is to construct tests based on the relative values of the likelihood function at the parameter values specified by the two competing hypotheses. Thus suppose we have a realization of data $\mathbf{x} = (x_1, \ldots, x_n)$. Then we can define the likelihood ratio as

$$\omega(\mathbf{x}; \theta_0, \theta_1) = \frac{L(\theta_0 \,|\, \mathbf{x})}{L(\theta_1 \,|\, \mathbf{x})}$$

where $L(\theta \,|\, \mathbf{x}) = f(\mathbf{x} \,|\, \theta)$ is the likelihood function given $\mathbf{x}$. In the case that the data are i.i.d. from $f(x \,|\, \theta)$, this statistic is

$$\omega(\mathbf{x};\theta_0,\theta_1) \quad = \quad \frac{\displaystyle\prod_{i=1}^{n} f(x_i \,|\, \theta_0)}{\displaystyle\prod_{i=1}^{n} f(x_i \,|\, \theta_1)}$$

We can now consider constructing tests based on $\omega(\mathbf{x};\theta_0,\theta_1)$. Small values of $\omega$ imply that the data are more likely under $H_1$ than $H_0$, and this is therefore evidence in favour of $H_1$ being true. Conversely large values of $\omega$ are consistent with $H_0$. Thus we might construct tests that reject $H_0$ when $\omega(\mathbf{x};\theta_0,\theta_1) \le k$ for a value $k$ which ensures the test has size $\alpha$.

**Example 4.6.** Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables with $\sigma^2$ known. We shall derive $\omega(\mathbf{x};\mu_0,\mu_1)$ for testing

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu = \mu_1$$

where $\mu_1 > \mu_0$. From Example 2.2 we have that

$$\begin{aligned} \omega(x_1,\ldots,x_n;\mu_0,\mu_1) &= \frac{L(\mu_0 \,|\, \mathbf{x})}{L(\mu_1 \,|\, \mathbf{x})} \\ &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\}} \\ &= \exp\left\{\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_1)^2-\sum_{i=1}^{n}(x_i-\mu_0)^2\right)\right\}. \end{aligned} \tag{4.1}$$

Now,

$$\begin{aligned} \sum_{i=1}^{n}(x_i-\mu_1)^2-\sum_{i=1}^{n}(x_i-\mu_0)^2 &= \sum_{i=1}^{n}(x_i^2-2\mu_1x_i+\mu_1^2)-\sum_{i=1}^{n}(x_i^2-2\mu_0x_i+\mu_0^2) \\ &= -2\mu_1 n\overline{x} + n\mu_1^2 + 2\mu_0 n\overline{x} - n\mu_0^2 \\ &= n(\mu_1^2-\mu_0^2) - 2n\overline{x}(\mu_1-\mu_0). \end{aligned} \tag{4.2}$$

Substituting equation (4.2) into (4.1) gives

$$\omega(x_1,\ldots,x_n;\mu_0,\mu_1) = \exp\left\{\frac{1}{2\sigma^2}\left(n(\mu_1^2-\mu_0^2)-2n\overline{x}(\mu_1-\mu_0)\right)\right\}.$$

The critical region for the corresponding test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$

$(\mu_1 > \mu_0)$ is thus

$$\begin{aligned} C &= \left\{(x_1,\ldots,x_n) : \exp\left\{\frac{1}{2\sigma^2}\left(n(\mu_1^2-\mu_0^2) - 2n\overline{x}(\mu_1-\mu_0)\right)\right\} \le k\right\} \\ &= \left\{(x_1,\ldots,x_n) : n(\mu_1^2-\mu_0^2) - 2n\overline{x}(\mu_1-\mu_0) \le 2\sigma^2 \log k\right\} \\ &= \left\{(x_1,\ldots,x_n) : -2n\overline{x}(\mu_1-\mu_0) \le 2\sigma^2 \log k + n(\mu_0^2-\mu_1^2)\right\} \\ &= \left\{(x_1,\ldots,x_n) : \overline{x} \ge \frac{-\sigma^2}{n(\mu_1-\mu_0)}\log k + \frac{(\mu_0+\mu_1)}{2}\right\} \qquad (4.3) \\ &= \{(x_1,\ldots,x_n) : \overline{x} \ge k^*\}. \qquad (4.4) \end{aligned}$$

Note that, in Equation (4.3), we have used the fact that $\mu_1 - \mu_0 > 0$. The critical region given in Equation (4.4) is identical to our intuitive interval derived in Example 4.3. For the test to be of significance $\alpha$ we choose (see Example 4.4)

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

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution (i.e. $P(Z \le z_{1-\alpha}) = 1-\alpha$.)

Constructing tests of two simple hypotheses based $\omega(\mathbf{x};\theta_0,\theta_1)$ is useful in that it gives us a general prescription for how to construct a test statistic for testing two simple hypotheses. A natural question however is whether such tests can be improved upon? Thus we ask can we find a different test statistic $T^*$ and critical region which, for the same sample size $n$, has the same value of $\alpha = P(\text{Type I error})$ but a smaller $\beta = P(\text{Type II error})$? Equivalently, can we find the test with the largest power?

**Definition 4.4** (Most powerful test)**.** Consider the test of the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1\,.$$

A test with statistic $T^*$ and critical region $C^*$ is **the most powerful test of significance level $\boldsymbol{\alpha}$** if and only if any other test with statistic $T'$ and critical region $C'$ and the same significance level $\alpha$, has smaller power. That is

$$\begin{gathered} \alpha = P(T^* \in C^* \,|\, \theta = \theta_0) = P(T' \in C' \,|\, \theta = \theta_0) \\ 1-\beta = P(T^* \in C^* \,|\, \theta = \theta_1) > P(T' \in C' \,|\, \theta = \theta_1) \end{gathered}$$

It turns out that when testing two simple hypotheses, tests based on $\omega(\mathbf{x};\theta_0,\theta_1)$ are indeed optimal.

**Lemma 4.1** (The Neyman-Pearson Lemma)**.** *Let $X_1, \ldots, X_n$ be i.i.d. with density/mass function $f(x|\theta)$. Consider the test of hypotheses*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1 \,.$$

*Then the test with critical region of the form $\{\mathbf{x} : \omega(\mathbf{x}; \theta_0, \theta_1) \leq k\}$ with $k$ chosen so that the test has size $\alpha$ is the most powerful test of significance level $\alpha$.*

**Proof:** See Theorem 8.3.12 of Casella and Berger for details. □

**Example 4.7.** Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli with success probability $p$, so that the sum $Y = \sum_{i=1}^n X_i$ is then binomially distributed $\text{Bin}(n, p)$. Consider testing $H_0 : p = p_0$ versus $H_1 : p = p_1$ where $p_1 > p_0$. The likelihood ratio is

$$\begin{aligned} \frac{L(p_0 \,|\, y)}{L(p_1 \,|\, y)} &= \frac{\binom{n}{y} p_0^y (1-p_0)^{n-y}}{\binom{n}{y} p_1^y (1-p_1)^{n-y}} \\ &= \left[\frac{p_0}{p_1} \frac{(1-p_1)}{(1-p_0)}\right]^y \left[\frac{1-p_0}{1-p_1}\right]^n \end{aligned}$$

The Neyman-Pearson Lemma then says we should reject when this is less than or equal to some $k$, or equivalently if

$$y \log\left(\frac{p_0}{p_1} \frac{(1-p_1)}{(1-p_0)}\right) + n \log\left(\frac{1-p_0}{1-p_1}\right) \leq \log(k)$$

Since $p_1 > p_0$, $\frac{p_0}{p_1} < 1$ and $\frac{(1-p_1)}{(1-p_0)} < 1$. Thus

$$\log\left(\frac{p_0}{p_1} \frac{(1-p_1)}{(1-p_0)}\right) < 0$$

and so we should reject $H_0$ when $y \geq k^*$ for $k^*$ chosen so that

$$P(Y \geq k^* | p = p_0) = \alpha$$

A difficulty now arises in that for the particular $n$ and $p_0$ value it will very likely be the case that there is no such $k^*$ value, due to the discreteness of the cumulative distribution function of the binomial. One approach then is to find the smallest value $\tilde{k}$ such that

$$P(Y \geq \tilde{k} | p = p_0) < \alpha$$

so that the size of the test is controlled at some maximal level no larger than $\alpha$. To illustrate, suppose $n = 10$ and $p_0 = 0.5$. Then $P(Y \geq 8) = 0.055$ and $P(Y \geq 9) = 0.011$, we can either choose a test with slightly more than 5% size, or one with size quite a bit less than 5%.

## 4.3 Composite hypotheses, one-sided and two-sided tests

So far, we have been concerned with testing "simple" hypotheses of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

Each of these hypotheses completely specifies the probability distribution of each $X_i$. In practice we are often interested instead in testing between two hypotheses, at least one of which does not completely specify the data distribution.

**Example 4.8.** In Example 4.6, we wished to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1$$

where $\mu_1 > \mu_0$ and $\sigma^2$ is known. Both of these hypotheses are simple. Under $H_0$ the distribution of each $X_i$ is completely specified as $N(\mu_0, \sigma^2)$ while under $H_1$ the distribution of each $X_i$ is completely specified as $N(\mu_1, \sigma^2)$.

**Example 4.9.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where $\sigma^2$ is known. Then the hypothesis $H : \mu > \mu_0$ is composite as if $H$ is true the distribution of each $X_i$ is not completely specified.

**Example 4.10.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where both parameters are unknown. Then the hypothesis $H_0 : \mu = \mu_0$ is again composite as the distribution of each $X_i$ is not completely specified since $\sigma^2$ is still unknown.

There are three particular tests of initial interest.

1. $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$
2. $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$
3. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$

In each case, the alternative hypothesis is composite. In the first two cases, we have a **one-sided alternative** whilst in the latter case we have a **two-sided alternative**. We cannot use the Neyman-Pearson Lemma as it only applies for a test of two simple hypotheses. In what follows we construct tests for the above three scenarios.

**Example 4.11.** In the normal case (Example 4.6, the critical region of the **most powerful** test of the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu_1 > \mu_0$ and $\sigma^2$ was known is

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \overline{x} \geq k^*\}$$

This region holds for **all** $\mu_1 > \mu_0$ and so is the most powerful test for **every** simple hypothesis of the form $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$. The value of $\mu_1$ only affects the power of the test. This can be seen from Example 4.5 where

$$\text{Power} = 1 - \beta = 1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}\right)$$

If $\mu_1$ is close to $\mu_0$ then we have a small power. The power increases as $\mu_1$ increases.

**Definition 4.5** (Uniformly Most Powerful Test)**.** Suppose that $H_1$ is composite. A test that is most powerful for every simple hypothesis in $H_1$ is said to be **uniformly most powerful**.

### 4.3.1 Hypothesis testing for one-sided alternatives

Uniformly most powerful tests exist for some common one-sided alternatives.

**Example 4.12.** If the $X_i$ are i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known, then the test

$$C^* \quad = \quad \{(x_1, \dots, x_n) : \overline{x} \geq k^*\}$$

is the most powerful for every $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$. For a test with significance $\alpha$ we choose $k^* = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. Thus, this test is uniformly most powerful for testing the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$, with significance level $\alpha$.

Now consider the other side ($\mu_1 < \mu_0$). The test

$$C^* \quad = \quad \{(x_1, \dots, x_n) : \overline{x} \leq k^*\}$$

is the most powerful for every $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ with $\mu_1 < \mu_0$. For a test with significance $\alpha$ we choose $k^* = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. This test is uniformly most powerful for testing the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$, with significance level $\alpha$.

**Example 4.13.** The existing standard drug to treat hypertension (high blood pressure) reduces blood pressure on average by an amount $\mu_0$. A pharmaceutical company has developed a new drug which they hope achieves a larger blood pressure reduction, on average. A random sample of $n$ patients are recruited into a study, have the new drug administered, and their reduction in blood pressure is measured. A potential model is that the blood pressure reductions $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$. Suppose that $\sigma^2$ is known. The company wishes to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$, The uniformly most powerful test for this rejects $H_0$ when $\overline{x} \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

Suppose the study is conducted with $\mu_0 = 10$, $n = 100$, $\sigma = 10$, and $\alpha = 0.025$. The critical value is thus

$$\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 10 + 1.96\frac{10}{\sqrt{100}} = 11.96$$

The observed sample mean is $\overline{x} = 12$, which is (just) greater than the critical value. Thus we reject $H_0 : \mu = \mu_0$ in favour of the alternative hypothesis $H_1 : \mu > 10$.

In practice it may well be impossible to rule out the possibility that $\mu < \mu_0$. In this case we can consider testing

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0,$$

Now the null hypothesis is composite, as well as the alternative. Suppose we used the same test as before, namely that we reject $H_0$ when $\overline{x} \geq \mu_0 + z_{1-\alpha} \cdot \sigma/\sqrt{n}$. This test controls the Type I error under this composite null hypothesis because for any point $\mu < \mu_0$ in the null hypothesis

$$P\left(\overline{X} \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\middle|\mu < \mu_0\right) < P\left(\overline{X} \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\middle|\mu = \mu_0\right) = \alpha$$

We now introduce the idea of a **power function** to characterise the behaviour of a test under a composite alternative. Recall that the power of a test is

$$1 - \beta = P(\text{Reject } H_0 \mid H_0 \text{ false}).$$

When the alternative hypothesis is composite, the power is a function of the possible values of $\theta$ (with particular interest in $\theta \in H_1$).

**Definition 4.6** (Power function)**.** The **power function** $\pi(\theta)$ of a test of $H_0 : \theta = \theta_0$ is

$$\begin{aligned} \pi(\theta) &= P(\text{Reject } H_0|\theta) \\ &= 1 - P(\text{Accept } H_0|\theta). \end{aligned}$$

**Example 4.14.** In the $\mu_1 > \mu_0$ case the Type II error rate for a particular value of $\mu^*$ in $H_1$ is (see Example 4.5)

$$\begin{aligned} \beta(\mu^*) &= P\left\{\overline{X} < \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\middle|\overline{X} \sim N(\mu^*, \sigma^2/n)\right\} \\ &= \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu^*}{\sigma/\sqrt{n}}\right) \end{aligned}$$

We can view this as a function of $\mu$ such that corresponding power is also a function of $\mu > \mu_0$. We have

$$\pi(\mu) = 1 - \beta(\mu) = 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right)$$

Figure 4.3 shows a sketch of $\pi(\mu)$.

Some observations:

- For $\mu = \mu_0$ we have

$$\pi(\mu_0) \;=\; 1 - \Phi(z_{1-\alpha}) \;=\; \alpha.$$

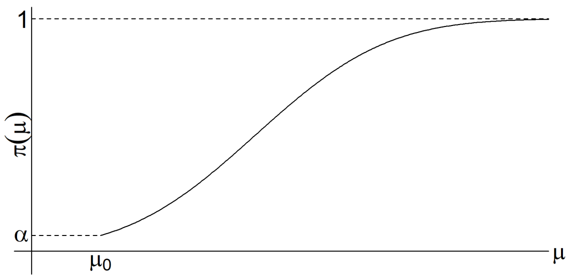


Figure 4.3: The power function, $\pi(\mu)$, for the uniformly most powerful test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.

- As $\mu$ increases, $\Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$ decreases so that $\pi(\mu)$ is an increasing function which tends to 1 as $\mu \to \infty$.
- As $\mu \to \mu_0$ it is very hard to distinguish between the two hypotheses.

### 4.3.2 Hypothesis testing for two-sided alternatives

We now consider testing for a so called two-sided alternative hypothesis, i.e. hypotheses of the form that $\theta \neq \theta_0$ for some $\theta_0$. One approach is to combine the critical regions for testing the two one-sided alternatives.

**Example 4.15.** We consider that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known. We wish to test the hypotheses

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

We combine the two one-sided tests to form a critical region of the form

$$C = \{(x_1, \ldots, x_n) : \overline{x} \leq k_1 \text{ or } \overline{x} \geq k_2\}.$$

For a test of size $\alpha$ we must have

$$\begin{aligned} \alpha &= P(\{\overline{X} \leq k_1\} \cup \{\overline{X} \geq k_2\} \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)) \\ &= P\{\overline{X} \leq k_1 \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\} + P\{\overline{X} \geq k_2 \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\} \end{aligned}$$

One way to select $k_1$ and $k_2$ is to place $\alpha/2$ into each tail as shown in Figure 4.4.

Then we have

$$k_1 = \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \quad \text{and} \quad k_2 = \mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Thus, the test rejects for

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{1-\frac{\alpha}{2}} \quad \text{or} \quad \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\frac{\alpha}{2}}$$

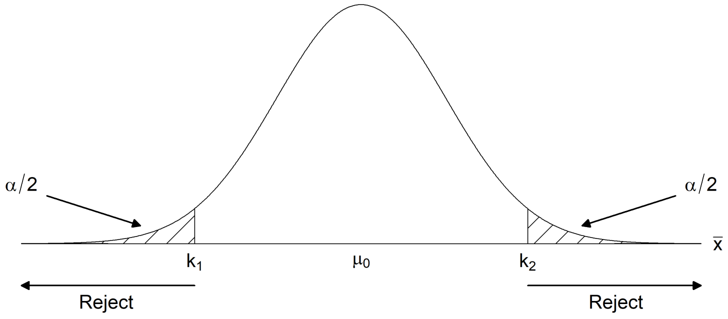


Figure 4.4: The critical region $C = \{(x_1, \ldots, x_n) : \overline{x} \le k_1 \text{ or } \overline{x} \ge k_2\}$ for testing the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$.

which is equivalent to

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} \ge z_{1-\frac{\alpha}{2}}.$$

Note that this test is not uniformly most powerful, since for any given simple hypothesis within the alternative, we can construct a one-sided test of size $\alpha$ that is more powerful.

The corresponding power function is

$$\begin{aligned}\pi(\mu) &= P(\text{Reject } H_0 \,|\, \mu \ne \mu_0) \\ &= P\{\overline{X} \le k_1 \,|\, \overline{X} \sim N(\mu, \sigma^2/n)\} + P\{\overline{X} \ge k_2 \,|\, \overline{X} \sim N(\mu, \sigma^2/n)\} \\ &= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\frac{\alpha}{2}}\right) + 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\frac{\alpha}{2}}\right).\end{aligned}$$

The power function is shown in Figure 4.5. Notice that the power function is symmetric about $\mu_0$.

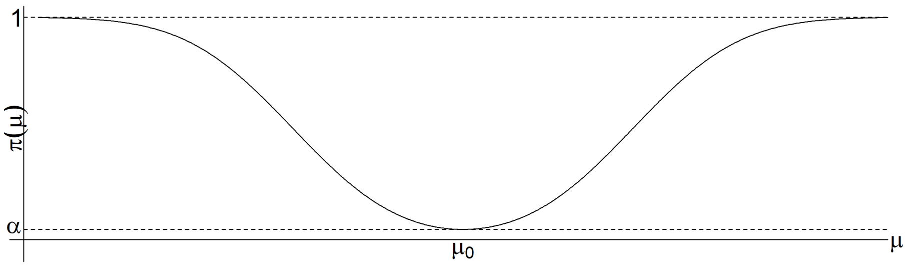


Figure 4.5: The power function, $\pi(\mu)$, for the test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$.

### 4.3.3 Tests with unknown variance

Thus far we have considered tests for the normal mean when the variance $\sigma^2$ is known. As we have noted previously, in practice it would typically not be known.

**Example 4.16.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where both $\mu$ and $\sigma$ are unknown. Consider the test of hypothesis

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

Because the variance $\sigma^2$ is unknown and not specified by the two hypotheses, both hypotheses are composite, and we cannot apply the Neyman-Pearson Lemma.

In this case we cannot use the intuitive critical region

$$C = \{(x_1, \ldots, x_n) : \overline{x} \geq k\}$$

since when trying to find the value of $k$, we realize that we cannot compute

$$P\left(\overline{X} \geq k | \mu = \mu_0\right)$$

based on $\overline{X} \sim N(\mu, \sigma^2/n)$ because $\sigma^2$ is now unknown. To construct a usable critical region we recall that under $H_0$

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1},$$

so that we can now define the critical region as

$$C = \left\{(x_1, \ldots, x_n) : \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \geq k\right\}.$$

We can compute $k$ since we can solve the following equation for $k$

$$P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq k | \mu = \mu_0\right) = \alpha$$

for a given significance level $\alpha$. In fact, we clearly have that $k = t_{n-1,1-\alpha}$ so that we reject when

$$\overline{x} \geq \mu_0 + t_{n-1,1-\alpha} \frac{s}{\sqrt{n}}$$

This test is known as the **one-sided, one-sample t-test**.

**Example 4.17.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown. Consider the test of hypothesis

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

Since the $t$-distribution is symmetrical around zero, we can define a critical region as

$$C = \left\{(x_1, \ldots, x_n) \,:\, \left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| \geq t_{n-1,1-\alpha/2}\right\} \tag{4.5}$$

and this region has significance level $\alpha$. This test is known as the **two-sided, one sample t-test**.

**Example 4.18.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown. Consider the test of hypothesis

$$H_0 \,:\, \sigma^2 = \sigma_0^2 \qquad \text{vs} \qquad H_1 \,:\, \sigma^2 \neq \sigma_0^2$$

Here we can use a result from Theorem 3.1 that

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

which is valid when $H_0$ is true. Then we can use the critical region

$$C = \left\{(x_1, \ldots, x_n) \,:\, \frac{(n-1)s^2}{\sigma_0^2} \leq k_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq k_2\right\}$$

where $k_1 < k_2$. Then for a given significance level $\alpha$, we can take $k_1 = \chi^2_{n-1,\alpha/2}$ and $k_2 = \chi^2_{n-1,1-\alpha/2}$ since

$$P\left(\left\{\frac{(n-1)S^2}{\sigma_0^2} \leq k_1\right\} \cup \left\{\frac{(n-1)S^2}{\sigma_0^2} \geq k_2\right\} \middle| \, \sigma^2 = \sigma_0^2\right) =$$

$$P\left(\frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{n-1,\alpha/2} \middle| \, \sigma^2 = \sigma_0^2\right) + P\left(\frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{n-1,1-\alpha/2} \middle| \, \sigma^2 = \sigma_0^2\right)$$

$$= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

### 4.3.4 Duality between Hypothesis Testing and Confidence Intervals

There is a close connection between hypothesis tests with two-sided alternatives and confidence intervals.

**Example 4.19.** Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known and we wish to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Earlier we derived a test where we accept $H_0 : \mu = \mu_0$ if

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} < z_{1-\frac{\alpha}{2}}.$$

Equivalently, accept for

$$\begin{aligned}
-z_{1-\frac{\alpha}{2}} < \tfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < z_{1-\frac{\alpha}{2}} & \quad \Leftrightarrow \\
\mu_0 - z_{1-\frac{\alpha}{2}} \tfrac{\sigma}{\sqrt{n}} < \overline{x} < \mu_0 + z_{1-\frac{\alpha}{2}} \tfrac{\sigma}{\sqrt{n}} & \quad \Leftrightarrow \\
\overline{x} - z_{1-\frac{\alpha}{2}} \tfrac{\sigma}{\sqrt{n}} < \mu_0 < \overline{x} + z_{1-\frac{\alpha}{2}} \tfrac{\sigma}{\sqrt{n}} &
\end{aligned}$$

Recall, from the previous chapter (Equation (3.2)) that $(\overline{x}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$ is a $100(1-\alpha)\%$ confidence interval for $\mu$. We see that $\mu_0$ lies in the $100(1-\alpha)\%$ confidence interval if and only if the hypothesis test with Type I error $\alpha$ accepts $H_0$. Conversely, the $100(1-\alpha)\%$ confidence interval contains exactly those values of $\mu_0$ for which we would accept $H_0 : \mu = \mu_0$ in the test with Type I error $\alpha$.

This shows a **duality** between the hypothesis test and the confidence interval: the latter may be obtained by inverting the former and vice versa. The duality in fact holds more generally.

**Example 4.20.** Suppose a 95% confidence interval of $(-1, 3)$ has been calculated for the mean $\mu$. We can immediately deduce that we would not reject a two-sided hypothesis test of the null that $\mu = 0$ at level $\alpha = 0.05$ because the confidence interval **includes zero**. But we would reject a test of the null that $\mu = 4$ a $\alpha = 0.05$ level test, since 4 lies outside the 95% confidence interval.

## 4.4 P-values

The hypothesis testing approach we have described so far results in a binary decision or outcome: either reject or accept the null hypothesis $H_0$. If the Type I error rate $\alpha$ was set very small, rejection of the null is more convincing than if $\alpha$ was set large. Apart from this however, the result of a hypothesis test is not very granular in regards to how strongly the null has been rejected, if it is rejected.

**Example 4.21.** Recall the drug testing example, Example 4.13. Suppose again that the study is conducted with $\mu_0 = 10$, $n = 100$, $\sigma = 10$, and $\alpha = 0.025$. The critical value for testing $\mu = \mu_0$ versus $\mu > \mu_0$ is thus

$$\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 10 + 1.96\frac{10}{\sqrt{100}} = 11.96$$

If the observed sample mean is $\overline{x} = 12$, we reject the null. If instead $\overline{x} = 15$, we again reject the null, but the binary decision of the test, on its own, gives no information about the strength of evidence against the null hypothesis. If in fact we had set $\alpha = 0.000001$, then the critical value would have been

$$\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 10 + 4.753\frac{10}{\sqrt{100}} = 14.753$$

and so with $\overline{x} = 15$ we would have rejected at this much smaller significance level.

In the Neyman-Pearson hypothesis testing framework, the significance level / Type I error rate is fixed in advance, and cannot be modified upon seeing the data. In Sir Ronald Fisher's testing framework, the **p-value** is a continuous measure of how compatible the observed data are with the null hypothesis.

The p-value can be viewed as the significance level at which the observed value of the test statistic would been on the accept/don't recept borderline.

**Definition 4.7** (P-value)**.** Suppose that $T(X_1, \dots, X_n)$ is our test statistic, such that the larger the value of $T$ the stronger the evidence against $H_0$ in favour of $H_1$. Let $t$ denote the observed value of $T$. The p-value is defined as

$$p(t) = P(T \geq t | H_0 \text{ true}).$$

The p-value measures the compatibility of the observed data (via the test statistic $T$) with the null hypothesis. It is therefore sometimes interpreted as measuring the evidence in the data against the null hypothesis: the smaller the p-value the stronger the evidence against the null hypothesis being true.

Simply put, the p-value can be viewed as the probability of observing a value at least as extreme as the observed data.

**Example 4.22.** Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known, and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. In Example 4.12 we derived a uniformly most powerful size $\alpha$ test for this setting which rejects $H_0$ when

$$\overline{x} \geq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}.$$

Large values of $\overline{x}$ are supportive of $H_1$. Having observed $\overline{x}$, the p-value is thus

$$\begin{aligned} P(\overline{X} \geq \overline{x} | H_0 \text{ true}) &= P\left(\overline{X} \geq \overline{x} \,\middle|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\right) \\ &= P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\right) \\ &= 1 - \Phi\left(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right). \end{aligned}$$

In the drug testing example (Example 4.21) with $\mu_0 = 10$, $n = 100$, $\sigma = 10$ and $\overline{x} = 12$ we have

$$p(\overline{x} \geq 12) = 1 - \Phi\left(\frac{12 - 10}{10/10}\right) = 1 - \Phi(2) = 0.023.$$

Since this value is less than 0.025, this agrees with our earlier finding that we reject $H_0$ (just) if we test at $\alpha = 0.025$. The p-value enables us to deduce the result of the test for every significance level: for all tests of significance greater than 0.023 we reject $H_0$. For all tests of significance less than 0.023 we do not reject $H_0$.

If instead we had observed $\overline{x} = 15$, the p-value would have been

$$p(\overline{x} \geq 15) = 1 - \Phi\left(\frac{15 - 10}{10/10}\right) = 1 - \Phi(5) < 0.000001.$$

In this case, either an incredibly rare event under $H_0$ has occurred or the data does not support $H_0$.

**Example 4.23.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ unknown, and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. In Example 4.16 we constructed a test for this situation which rejects when

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \geq t_{n-1,1-\alpha}.$$

Here large values of $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$ indicate support for the alternative hypothesis that $\mu > \mu_0$. Having observed $t$, the p-value is thus

$$P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \,\middle|\, H_0 \text{ true}\right).$$

Under $H_0$, $T$ is distributed $t_{n-1}$, and thus we have that the p-value is

$$P\left(T_{n-1} \geq \frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right) = 1 - P\left(T_{n-1} \leq \frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right)$$

where $T_{n-1}$ is a random variable t-distributed on $n-1$ degrees of freedom. This probability can thus be calculated using R or tables.

For two-sided tests we must be careful to calculate the p-value based on a test statistic for which large values indicate support for the alternative hypothesis:

**Example 4.24.** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ unknown, and consider testing $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu \neq \mu_0$. In Example 4.17 we derived a test for this setting which rejects when

$$\left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| \geq t_{n-1,1-\alpha/2}.$$

Large values of $\left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right|$ indicate support for the two-sided alternative. Since the t-distribution is symmetric, the p-value is

$$P\left(\left|\frac{\overline{X} - \mu_0}{S/\sqrt{n}}\right| \geq \left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| \,\middle|\, H_0 \text{ true}\right) \quad = \quad 2 \times P\left(T_{n-1} \geq \left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right|\right)$$

where $T_{n-1}$ is t-distributed with $n-1$ degrees of freedom.

If one has such a test statistic $T$, the p-value can be used to construct a valid size $\alpha$ test of $H_0$.

**Theorem 4.1.** *The test which rejects $H_0$ if and only if $p(t) \leq \alpha$ is a valid test of significance/size $\alpha$.*

**Proof:** First write the p-value as

$$p(t) = P(T \geq t | H_0 \text{ true}) = 1 - F_T(t)$$

where $F_T(.)$ denotes the cumulative distribution function of $T$ under $H_0$. Then

$$\begin{aligned} P(p(T) \leq \alpha | H_0 \text{ true}) &= P(1 - F_T(T) \leq \alpha | H_0 \text{ true}) \\ &= P(F_T(T) \geq 1 - \alpha | H_0 \text{ true}) \\ &= P(T \geq F_T^{-1}(1-\alpha) | H_0 \text{ true}) \\ &= 1 - F_T(F_T^{-1}(1-\alpha)) = \alpha. \end{aligned}$$

Thus the Type I error is controlled at $\alpha$ as desired. □

Hence, for all tests with significance level $\alpha \geq p(t)$ we reject $H_0$ and do not reject if $p(t) > \alpha$.

If my observation $t$ is such that I ject $H_0$ at $\alpha_1$ but do not reject at $\alpha_2$ then $\alpha_1 > p(t) > \alpha_2$.

Notice, from the proof, that we have $P(p(T) \leq \alpha | H_0 \text{ true}) = \alpha$ so that the p-value has a **uniform distribution** on $[0, 1]$ when $H_0$ is true.

### 4.4.1 Hypothesis testing and p-values - interpretation and recent controversies

The p-value is a continuous measure of strength of evidence against the null hypothesis, given the data, and assuming the model assumptions made are valid. A p-value less than 0.05 is conventionally reported as being ‘statistically significant’. Such phrasing has (particularly recently) been criticised, for many reasons, but partly because it is too easy to equate such a statement with the result being qualitatively/substantively significant.

The use of hypothesis testing and p-values in empirical research has in recent years come under quite a lot of criticism. In my view some of this is justified, but many of the problems discussed are more about the mis-interpretation or misuse by some of hypothesis tests and p-values, rather than intrinsic issues with the methodology. We shall not go through these here, except to mention perhaps the most striking **incorrect** interpretation of a p-value as being the probability that the null hypothesis is true.

If you are interested in reading more about this, see the following links, including the first with 25 misinterpretations of p-values, confidence intervals, and power:

- Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations (2016)
- A Psychology journal bans P-values (2015)
- The American Statistical Association releases statement on the use of P-values (2016)

## 4.5 Robustness to model misspecification

We previously constructed tests that had size $\alpha$ under the null hypothesis, for any sample size, under the assumed normal model. As we have discussed previously, a natural concern is whether our procedures are robust to certain misspecifications in the modelling assumptions. As for confidence intervals in Section 3.4, for certain tests we may be able to show that they would be robust to certain misspecifications as $n \to \infty$. By robust, here we mean that their size, or Type I error probability, is maintained asymptotically at $\alpha$ under the particular misspecification we are considering.

**Example 4.25** (Robustness of z-tests and t-tests)**.** We have previously considered the situation that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, and we are interested in tests for the mean $\mu$. When $\sigma^2$ is known, the test statistic for the z-test

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

is distributed $N(0,1)$ under the null $H_0 : \mu = \mu_0$. What if the normality assumption for $X_1, \ldots, X_n$ does not hold? Is the Type I error rate controlled at level $\alpha$? From the Central Limit Theorem (Theorem 3.3) we know that, provided $\text{Var}(X)$ is finite, without requiring the normality assumption, as $n \to \infty$

$$Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \xrightarrow{L} N(0,1).$$

under the null. As such, tests based on assuming $Z \sim N(0,1)$ will continue to have the correct $\alpha$ level asymptotically, even when the normality assumption for $X_1, \ldots, X_n$ does not hold. For finite $n$, how close the actual Type I error rate is to $\alpha$ will depend on $n$ and how far the true distribution is from the normal.

When $\sigma^2$ is unknown, we previously derived the t-test, whose test statistic is

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

which under the null $H_0 : \mu = \mu_0$ is t-distributed on $n-1$ degrees of freedom. Now suppose we remove the normality assumption. Under the null that $H_0 = \mu = \mu_0$, provided $\text{Var}(X) < \infty$ and $S \xrightarrow{P} \sigma$, Slutsky's Theorem gives

$$T_n = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \times \frac{\sigma}{S} \xrightarrow{L} N(0,1) \times 1 = N(0,1)$$

Thus again asymptotically the t-test will continue to have size $\alpha$ even when the normality assumption for $X_1, \dots, X_n$ does not hold.

## 4.6 Summary of test statistics in the normal case

Let $X_1, \dots, X_n$ be independent and identically distributed $N(\mu, \sigma^2)$ random variables.

- An unbiased estimator of $\mu$ is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ with $\overline{X} \sim N(\mu, \sigma^2/n)$.

- An unbiased estimator of $\sigma^2$ is $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ with $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

**Tests for $\mu$ when $\sigma^2$ is known (z-tests)**
Derived using the pivot $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. If $H_0 : \mu = \mu_0$ is true then

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

should be a realisation from $N(0,1)$.

Table 4.2: Critical regions for one-sided and two-sided z-tests (variance known) at significance level $\alpha$.

| $H_0$ | $H_1$ | Reject $H_0$ if |
|---|---|---|
| $\mu = \mu_0$ | $\mu > \mu_0$ | $z \geq z_{1-\alpha}$ |
| $\mu = \mu_0$ | $\mu < \mu_0$ | $z \leq -z_{1-\alpha} = z_\alpha$ |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | $\lvert z\rvert \geq z_{1-\alpha/2}$ |

where, for $p \in [0,1]$, if $Z \sim N(0,1)$ then $z_p$ satisfies $P(Z \leq z_p) = p$.

**Tests for $\mu$ when $\sigma^2$ is unknown (t-tests)**
Derived using the pivot $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$. If $H_0 : \mu = \mu_0$ is true then

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

should be a realisation from $t_{n-1}$, the t-distribution with $n-1$ degrees of freedom.

Table 4.3: Critical regions for one-sided and two-sided $t$-tests (variance unknown) at significance level $\alpha$.

| $H_0$ | $H_1$ | Reject $H_0$ if |
|---|---|---|
| $\mu = \mu_0$ | $\mu > \mu_0$ | $t \geq t_{n-1,1-\alpha}$ |
| $\mu = \mu_0$ | $\mu < \mu_0$ | $t \leq -t_{n-1,1-\alpha} = t_{n-1,\alpha}$ |
| $\mu = \mu_0$ | $\mu \neq \mu_0$ | $\lvert t\rvert \geq t_{n-1,1-\alpha/2}$ |

where, for $p \in [0,1]$, if $T \sim t_{n-1}$ then $t_{n-1,p}$ satisfies $P(T \leq t_{n-1,p}) = p$.

**Tests for $\sigma^2$ when $\mu$ is unknown ($\chi^2$-tests)**
Derived using the pivot $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. If $H_0 : \sigma^2 = \sigma_0^2$ is true then

$$w = \frac{(n-1)s^2}{\sigma_0^2}$$

should be a realisation from $\chi^2_{n-1}$, the $\chi^2$-distribution with $n-1$ degrees of freedom.

Table 4.4: Critical regions for one-sided and two-sided $\chi^2$-tests.

| $H_0$ | $H_1$ | Reject $H_0$ if |
|---|---|---|
| $\sigma^2 = \sigma_0^2$ | $\sigma^2 > \sigma_0^2$ | $w \geq \chi^2_{n-1,1-\alpha}$ |
| $\sigma^2 = \sigma_0^2$ | $\sigma^2 < \sigma_0^2$ | $w \leq \chi^2_{n-1,\alpha}$ |
| $\sigma^2 = \sigma_0^2$ | $\sigma^2 \neq \sigma_0^2$ | $w \geq \chi^2_{n-1,1-\alpha/2}$<br>or $w \leq \chi^2_{n-1,\alpha/2}$ |

where, for $p \in [0,1]$, if $V \sim \chi^2_{n-1}$ then $\chi^2_{n-1,p}$ satisfies $P(V \leq \chi^2_{n-1,p}) = p$.

# Chapter 5

# Inference for relationships

So far we have always considered one random sample $X_1, \dots, X_n$ from some density $f(x|\theta)$. Very often we are interested in the differences in distributions between two or more random samples.

We first turn to the case where two random samples of continuous data are available. In Chapter 6 we'll consider the case of more than two samples. Recall from Example 1.9 that the AIDS Clinical Trials Group Study 175 (ACTG175) was a randomised clinical trial conducted in the 1990s to compare different treatments for adults infected with HIV. The trial had four randomised groups, with each group receiving a different treatment. Within each treatment group, observations were made at different time-points. One of the outcomes is the patient's CD4 blood cell count. A larger CD4 value is indicative that the treatment is working.

**Example 5.1** (ACTG175 trial, paired sample)**.** Suppose that, in the ACTG175 trial, we are interested in comparing the CD4 count at baseline (before treatment) to the week 20 CD4 count (after treatment), within one of the randomised groups. In this case, the two samples are composed of exactly the same individuals, the only difference being that they are observed at two different time-points. Figure 5.1 shows the histogram of change in CD4 from baseline to 20 weeks (20 week value minus baseline value) in the zidovudine plus didanosine group.

A visual inspection of the histogram suggests there are more positive than negative counts and that the mean difference is also positive. We will perform a formal statistical test for this example in Example 5.3.

**Example 5.2** (ACTG175 trial, two sample comparison)**.** We now focus on a comparison of those individuals in the ACTG175 trial randomised to receive zidovudine (532 patients) and those randomised to receive a combination treatment of zidovudine plus didanosine (522 patients). An analyses of interest is to compare the CD4 count after treatment with zidovudine plus didanosine

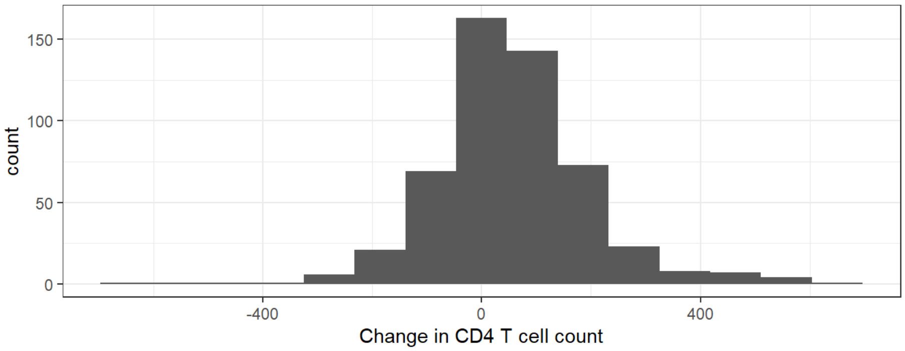


Figure 5.1: Histogram of change in CD4 T cell count from baseline to 20 weeks in zidovudine plus didanosine group of ACTG175 trial.

compared to treatment with zidovudine alone. Figure 5.2 shows overlaid density plots of the CD4 count 20 weeks after baseline in these two groups.

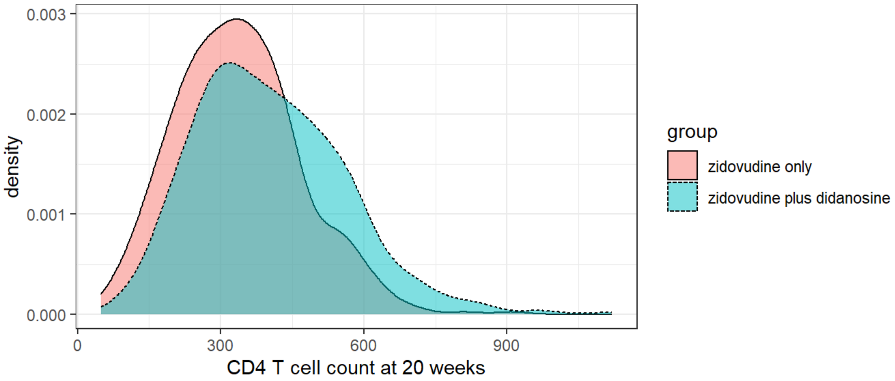


Figure 5.2: Distribution of CD4 T cell count at 20 weeks in zidovudine only versus zidovudine plus didanosine groups of ACTG175 trial.

Visually there is some suggestion that the CD4 distribution is shifted towards larger values in the zidovudine plus didanosine group. We will return to this example, performing a formal statistical test in Example 5.4.

In the first example, the observations are **paired** (and therefore dependent) since we can arrange the observations as pairs (before and after) corresponding to each individual. In the second example we usually assume that the random samples corresponding to two different groups of individuals can be considered to be **independent**. It would, however, be more accurate to describe the first example as a one-sample bivariate setting, rather than a two-sample setting.

As such, we treat the cases of inference for paired (dependent) samples and independent samples separately.

## 5.1 Paired samples - comparing means

We consider paired samples where we observe $n$ i.i.d. bivariate draws $(X_1, Y_1), \dots, (X_n, Y_n)$ from some distribution. The $n$ units are considered independent of each other, but $X$ and $Y$ for a given unit will generally be dependent. Example 5.1 introduced an example of this situation, where $X$ corresponds to the CD4 measurement on a patient at baseline (entry) to the clinical trial, and $Y$ corresponds to the CD4 measurement at week 20. We are interested in whether the distribution (e.g. the mean) of $X$ differs to $Y$. If it does this might be interpreted as an effect of the treatment.

A simple approach to modelling how the distributions of $X$ and $Y$ differ is to calculate the paired differences $D_i = X_i - Y_i$. The $D_i$ are independent by assumption with

$$\begin{aligned} E(D) &= \mu_X - \mu_Y \ =: \mu_D \\ \mathrm{Var}(D) &= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y) \\ &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} \ =: \sigma_D^2 \end{aligned}$$

where $\sigma_{XY}$ denotes the covariance between $X$ and $Y$. It follows from our assumptions that $D_i$ are i.i.d. with mean and variance as in the preceding equations. Thus if we confine ourselves to analysing the paired differences $D_i$, we are back in the one-sample situation which we have considered in all the preceding chapters. As such, all of the methods we have developed for model specification, point estimation, confidence intervals, and hypothesis testing can be applied. Because of the definition of $D_i$ as $X_i - Y_i$, a particular hypothesis of interest may be that $\mu_D = 0$, which is equivalent to $\mu_X = \mu_Y$.

**Example 5.3.** We now analyse the change in CD4 count data from the zidovudine plus didanosine group in the ACTG175 trial, as described in Example 5.1. The data for the first 6 patients in this group is shown below, where cd40 is the baseline measurement, cd420 is the 20 week measurement, and d is the 20 week measurement minus the baseline measurement:

```
##    pidnum cd40 cd420    d
## 6   10140  235   339  104
## 14  10361  212   190  -22
## 19  10389  180   200   20
## 23  10476  230    90 -140
## 26  10668  421   461   40
## 35  10896  444   468   24
```

For this data we have

$$n = 522; \sum_{i=1}^{522} d_i = 28422; \sum_{i=1}^{522} d_i^2 = 12392576.$$

We observe

$$\begin{aligned}
\overline{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{28422}{522} = 54.448, \\
s_d^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} d_i^2 - n\overline{d}^2\right) \\
&= \frac{1}{521}\left(12392576 - 522\left(\frac{28422}{522}\right)^2\right) = 20815.83.
\end{aligned}$$

A 95% confidence interval for $\mu_D$ can then be calculated just as we did earlier in Equation (3.3) as

$$\begin{aligned}
\overline{d} \pm t_{522-1,0.975} \times \frac{s_d}{\sqrt{n}} &= 54.448 \pm 1.965\frac{\sqrt{20815.83}}{\sqrt{522}} \\
&= (42.043, 66.854).
\end{aligned}$$

Since the 95% confidence interval excludes zero, we know immediately that a 5% two-sided test of the null hypothesis that the mean CD4 count is equal at the two times, or equivalently that the mean change in CD4 count is zero, would reject the null. We can calculate the p-value corresponding to the two-sided alternative. The t-statistic is calculated as in Equation (4.5)

$$\frac{\overline{d} - 0}{s_D/\sqrt{522}} = 8.622$$

and the p-value is

$$P(t_{522-1} \geq |8.622|) = 2 \times P(t_{522-1} \geq 8.622) = 7.918 \times 10^{-17}.$$

This is very small (much smaller than the conventional 5%), so we reject the null hypothesis that mean CD4 count is the same at baseline and 20 weeks. Based on the point estimate, we conclude that the treatment leads to an improvement in CD4 counts at 20 weeks.

## 5.2 Independent samples - comparing means

We now consider comparisons of the means $\mu_X$ and $\mu_Y$ in the two groups. Suppose we assume again that the distributions are normal, with $X_1, \ldots, X_n$ i.i.d. from $N(\mu_X, \sigma_X^2)$, and $Y_1, \ldots, Y_m$ i.i.d. from $N(\mu_Y, \sigma_Y^2)$. Note that in general $n$ need not equal $m$.

### 5.2.1 Variances known

Consider first the unrealistic case that both $\sigma_X^2$ and $\sigma_Y^2$ are known. We know that $\overline{X} \sim N(\mu_X, \sigma_X^2/n)$ and $\overline{Y} \sim N(\mu_Y, \sigma_Y^2/m)$. Then since the two groups are independent, it follows from the properties of linear combinations of normal random variables that

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right). \tag{5.1}$$

This result can readily be used to form confidence intervals for and perform hypothesis tests concerning the difference in means $\mu_X - \mu_Y$.

**CONFIDENCE INTERVAL**

To construct a $100 \times (1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$, first we note that:

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

is a pivot for $\mu_X - \mu_Y$. Thus

$$P\left(-z_{1-\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

Then re-arranging the inequality and multiplying through by $-1$, we have

$$P\left(\overline{X} - \overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < \overline{X} - \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$
$$= 1 - \alpha$$

Given some observed data, a $100 \times (1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is thus given by

$$\overline{x} - \overline{y} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

**HYPOTHESIS TESTING**

Now consider hypothesis tests concerning $\mu_X - \mu_Y$. A natural null hypothesis is that $\mu_X - \mu_Y = 0$, or equivalently that $\mu_X = \mu_Y$. Under this null:

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

Following the one-sample setting, we can construct significance level $\alpha$ tests for one-sided or two-sided alternatives are shown in Table 5.1

Table 5.1: Critical regions for one-sided and two-sided independent sample Normal tests (variances known).

| $H_0$ | vs. $H_1$ | Reject $H_0$ if |
|---|---|---|
| $\mu_X = \mu_Y$ | $\mu_X > \mu_Y$ | $z \geq z_{1-\alpha}$ |
| $\mu_X = \mu_Y$ | $\mu_X < \mu_Y$ | $z \leq -z_{1-\alpha} = z_\alpha$ |
| $\mu_X = \mu_Y$ | $\mu_X \neq \mu_Y$ | $\lvert z\rvert \geq z_{1-\alpha/2}$ |

### 5.2.2 Variances unknown, assumed equal

More typically, the two variances $\sigma_X^2$ and $\sigma_Y^2$ will be unknown. There are two routes for proceeding, based on whether we assume the two variances are equal. One could test the null hypothesis of equality of variances using the theory described later in Section 5.3 and proceed according to whether this test indicates evidence of inequality. However, there are limitations to this test in practice. For example, the power of this test may be low, such that we may accept the null of equal variances with high probability even when the variances differ.

One generally accepted rule of thumb is to assume equal variances if the ratio of the larger sample variance to the smaller is no more than two. But even in this case, some statisticians defer to methods that to allow them to be different. Having said all this, let us now proceed to examine comparison of the means under the assumption that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, again under the assumption that the true distribution in each group is normal.

We first have to estimate the common variance $\sigma^2$ using the data. We know that the usual $S^2$ estimator applied to each group is unbiased for $\sigma^2$. We therefore pool the $S^2$ estimates from the two groups to form a combined estimate of $\sigma^2$. We estimate $\sigma^2$ from each group's data and average them, weighting according to their respective sample sizes.

**Definition 5.1** (Pooled variance estimator)**.** The pooled estimator of the common variance $\sigma^2$ is defined as:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}. \tag{5.2}$$

Note that we cannot simply calculate $S^2$ in the combined sample, $\{X_1, \dots X_n, Y_1, \dots Y_m\}$, since this estimates the variability of the sample around the pooled sample mean, a different quantity.

Next we establish the sampling distribution of $S_p^2$ and of $\overline{X} - \overline{Y}$.

**Proposition 5.1.** *Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_m$ be i.i.d. $N(\mu_Y, \sigma^2)$, with the two samples independent. Then*

$$\frac{n+m-2}{\sigma^2} S_p^2 \sim \chi^2_{n+m-2}$$

*and*

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} \sim t_{n+m-2}.$$

**Proof:** From Theorem 3.1, we have that

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

As $S_X^2$ and $S_Y^2$ are independent, the sum

$$\frac{n-1}{\sigma^2} S_X^2 + \frac{m-1}{\sigma^2} S_Y^2 = \frac{n+m-2}{\sigma^2} S_p^2$$

is the sum of two independent $\chi^2$ random variables. Recalling from Probability & Statistics 1B that the $\chi^2$ distribution is a special case of the gamma distribution ($\chi^2_\nu = Ga(1/2, \nu/2)$) and that the sum of two independent gamma random variables with shared scale parameter is also gamma, we have

$$\frac{(n+m-2)S_p^2}{\sigma^2} \quad \sim \quad \chi^2_{n+m-2}.$$

Since the two samples are independent, we have that

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right),$$

and so

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1).$$

Lastly, from Theorem 3.1, we have that $S_X^2$ is independent of $\overline{X}$, and similarly that $S_Y^2$ is independent of $\overline{Y}$. From this it follows that $\overline{X} - \overline{Y}$ and $S_p^2$ are independent, so from the definition of the t-distribution (Definition 3.4):

$$\begin{aligned} & \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \Big/ (S_p/\sigma) \\ = \quad & \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} \sim t_{n+m-2} \quad \square \end{aligned}$$

This result can then be used to form confidence intervals for and test hypotheses concerning the difference in means $\mu_X - \mu_Y$.

### CONFIDENCE INTERVAL

Based on the conclusions of Proposition 5.1, analogously to the one sample situation considered in Section 3.2, we can form a $100 \times (1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ as

$$\overline{x} - \overline{y} \pm t_{n+m-2,1-\alpha/2} \times s_p \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

### HYPOTHESIS TESTING

Under $H_0 : \mu_X = \mu_Y$, we have from Proposition 5.1 that

$$T \quad = \quad \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

We can use this to construct one-sided and two-sided tests with significance level $\alpha$ as shown in Table 5.2.

Table 5.2: Critical regions for one-sided and two-sided independent sample $t$-tests (variances unknown but assumed equal).

| $H_0$ | vs. $H_1$ | Reject $H_0$ if |
|---|---|---|
| $\mu_X = \mu_Y$ | $\mu_X > \mu_Y$ | $t \geq t_{n+m-2,1-\alpha}$ |
| $\mu_X = \mu_Y$ | $\mu_X < \mu_Y$ | $t \leq -t_{n+m-2,1-\alpha}$ |
| $\mu_X = \mu_Y$ | $\mu_X \neq \mu_Y$ | $\lvert t \rvert \geq t_{n+m-2,1-\alpha/2}$ |

**Example 5.4.** We return to Example 5.2 where we consider comparing the mean CD4 count at week 20 in the ACTG175 study, assuming the two populations have a common unknown variance $\sigma^2$.

We let $x_1, \ldots, x_{532}$ denote the observed CD4 counts in the zidovudine treatment group. The data is summarised by

$$n = 532; \; \sum_{i=1}^{532} x_i \;\; = \;\; 178826; \; \sum_{i=1}^{532} x_i^2 = 69217556$$

so that

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{178826}{532} = 336.14, \\
s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \\
&= \frac{1}{531}\left(69217556 - 532\left(\frac{178826}{532}\right)^2\right) = 17150.934.
\end{aligned}$$

Similarly, let $y_1, \ldots, y_{522}$ denote the values in the zidovudine plus didanosine group. The data is summarised by

$$m = 522; \sum_{i=1}^{522} y_i = 210456; \sum_{i=1}^{522} y_i^2 = 97578584$$

so that

$$\begin{aligned}
\overline{y} &= \frac{1}{m}\sum_{i=1}^{m} y_i = \frac{210456}{522} = 403.17, \\
s_y^2 &= \frac{1}{m-1}\left(\sum_{i=1}^{m} y_i^2 - m\overline{y}^2\right) \\
&= \frac{1}{521}\left(97578584 - 522\left(\frac{210456}{522}\right)^2\right) = 24430.961.
\end{aligned}$$

Our point estimate for the difference in mean CD4 count, $\mu_X - \mu_Y$, is thus $\overline{x} - \overline{y} = -67.03$. The pooled estimate of the common variance is

$$\begin{aligned}
s_p^2 &= \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} \\
&= \frac{(532-1)(17150.934) + (522-1)(24430.961)}{532+522-2} = 20756.346.
\end{aligned}$$

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

$$\overline{x} - \overline{y} \pm t_{1052,1-\alpha/2} \times s_p\sqrt{\frac{1}{532} + \frac{1}{522}}$$

where $t_{1052,1-\alpha/2} = 1.962$ (using `qt(0.975,1052)`). This gives $(-84.45, -49.62)$.

We now perform a test of the hypothesis $H_0 : \mu_X = \mu_Y$ versus the two-sided alternative, at the 5% level. Our test statistic is

$$\frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} = \frac{-67.03}{144.071\sqrt{1/532 + 1/522}} = -7.552.$$

The corresponding $p$-value is

$$2 \times P(T_{1052} \geq |-7.552|) = 0$$

Given Figure 5.2 we should be concerned about the normality assumption. However, given the size of $n$ and $m$ here, we can be reasonably satisfied that the coverage of the confidence interval is close to 95% and the test's Type I error is close to the desired 5% level, **if** we believe the assumption of a common variance is correct. Given our earlier estimates of the variances in the two groups, we have reason to doubt this assumption.

### 5.2.3 Variances unknown, possibly unequal

We now consider the most realistic situation in practice: we allow for the possibility the variances in the two groups may well not be equal, and we do not know their true values. Recall from Equation (5.1) that

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right),$$

but now we do not know the values of $\sigma_X^2$ and $\sigma_Y^2$. From Proposition 2.2 we have that $S_X^2$ and $S_Y^2$ are consistent estimators of $\sigma_X^2$ and $\sigma_Y^2$ respectively, and so we can estimate $\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ by $\frac{S_X^2}{n} + \frac{S_Y^2}{m}$.

We then consider the distribution of

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

Unfortunately this distribution is not a standard one. One popular approach approximates the distribution with a t-distribution with a certain degrees of freedom, leading to Welch's test, which is available in `R`. For carrying out calculations by hand, we shall instead consider the asymptotic distribution of this test statistic, which as we shall see, is standard normal.

**Proposition 5.2.** *Let $X_1, \dots, X_n$ be i.i.d. $N(\mu_X, \sigma_X^2)$ and $Y_1, \dots, Y_m$ be i.i.d. $N(\mu_Y, \sigma_Y^2)$, with the two samples independent. Then assume that as $n \to \infty$ and $m \to \infty$, $\frac{m}{n+m} \to \rho$ for some $0 < \rho < 1$. Then*

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \xrightarrow{L} N(0, 1).$$

**Proof:** See the non-examinable proofs chapter, Chapter 8.

**Example 5.5.** Continuing with the ACTG175 data, we now calculate a 95% confidence interval for the difference in treatment group means, relaxing the assumption that they have a common population variance. Using Proposition 5.2 we have an asymptotically valid 95% confidence interval given by

$$\begin{aligned}\overline{x} - \overline{y} \pm 1.96\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} &= -67.033 \pm 1.96\sqrt{\frac{17150.934}{532} + \frac{24430.961}{522}} \\ &= (-84.46, -49.61)\end{aligned}$$

which compares with the confidence interval calculated assuming a common variance of $(-84.45, -49.62)$. They are similar despite the quite different sample variances in the two groups because the groups are of similar size, such that the two estimated standard errors for the difference in sample means are very similar.

### 5.2.4 Asymptotic confidence intervals and tests

We have developed confidence intervals and hypothesis tests for comparing the two group means based on normality assumptions for $X$ and $Y$. In fact these procedures are robust to non-normality provided $\sigma_X^2$ and $\sigma_Y^2$ are finite – so that we can apply the CLT – and that we have consistent estimators $\widehat{\sigma}_X^2$ and $\widehat{\sigma}_Y^2$ of these variances – so that we can apply Slutsky.

Then in the equal variance case we have

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \xrightarrow{L} N(0,1) \tag{5.3}$$

where

$$S_p^2 = \frac{(n-1)\widehat{\sigma}_X^2 + (m-1)\widehat{\sigma}_Y^2}{n+m-2}.$$

and in the un-equal variance case we have

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\widehat{\sigma}_X^2}{n} + \frac{\widehat{\sigma}_Y^2}{m}}} \xrightarrow{L} N(0,1) \tag{5.4}$$

More formal statements and proofs for the claims in (5.3) and (5.4) are given in Chapter 8. Note that we do not require $\widehat{\sigma}_X^2 = S_X^2$ or $\widehat{\sigma}_Y^2 = S_Y^2$, though often these estimators of the variance are used in practice. Review 2.2 to remind yourself when $S^2$ is a consistent estimator of the variance.

We can now exploit these results to construct confidence intervals and tests for particular (non-normal) distributions that asymptotically will have the correct coverage and size. How good the approximation is will depend on the magnitudes of $n$ and $m$ as well as how close the true data generating distributions in the two groups are to normals.

**Example 5.6** (Bernoulli model)**.** Let $X_1, \dots, X_n$ be i.i.d. Bernoulli with success probability $p_X$ and $Y_1, \dots, Y_m$ be i.i.d. Bernoulli with success probability $p_Y$. Since $E(X) = p_X$ and $E(Y) = p_Y$, we can use our preceding results for comparing means to compare $p_X$ with $p_Y$. We have $\mathrm{Var}(X) = p_X(1-p_X) < \infty$ and similarly $\mathrm{Var}(Y) < \infty$. Previously we considered the MLE of $\mathrm{Var}(X) = p_X(1-p_X)$ which is $\widehat{p_X}(1-\widehat{p_X})$, where $\widehat{p_X} = \overline{X}$. This is a consistent estimator by the Weak Law of Large Numbers and Continuous Mapping Theorem.

Thus we can use Equation (5.4) to construct confidence intervals for $p_X - p_Y$ and perform hypothesis tests. We have

$$\frac{\widehat{p_X} - \widehat{p_Y} - (p_X - p_Y)}{\sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{m}}} \xrightarrow{L} N(0,1)$$

A $100 \times (1-\alpha)\%$ confidence interval for $p_X - p_Y$ is then

$$\widehat{p_X} - \widehat{p_Y} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{m}}$$

To conduct a hypothesis test of the null hypothesis that $p_X = p_Y$, we can observe that under this hypothesis

$$\frac{\widehat{p_X} - \widehat{p_Y}}{\sqrt{\frac{\widehat{p_X}(1-\widehat{p_X})}{n} + \frac{\widehat{p_Y}(1-\widehat{p_Y})}{m}}} \xrightarrow{L} N(0,1)$$

This can be used to construct tests against one-sided or two-sided alternatives in the usual way.

The 'standard' large sample (asymptotic) test for this situation is actually based on a slightly different test statistic. Specifically, under the null that $p_X = p_Y = p$, the common variance $p(1-p)$ can be estimated by $\check{p}(1-\check{p})$, where $\check{p} = (\sum X_i + \sum Y_i)/(n+m)$. This leads to the test-statistic:

$$\frac{\widehat{p_X} - \widehat{p_Y}}{\sqrt{\check{p}(1-\check{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \tag{5.5}$$

which again is asymptotically $N(0,1)$ under the null.

**Example 5.7.** Recall from Example 1.10 that the Physician's Health Study was a medical study carried out in the 1980s to investigate the benefits and risks of aspirin for cardiovascular disease and cancer. The trial randomised 22,071 healthy doctors to receive either aspirin or placebo. They were then followed-up to see who experienced heart attacks during the following 5 years. The data are shown again in Table 5.3.

Table 5.3: Number of heart attacks by treatment group in the Physician's Health Study.

| Group | Heart attack | No heart attack | Total |
|---|---|---|---|
| Aspirin | 139 | 10898 | 11037 |
| Placebo | 239 | 10795 | 11034 |

Primary interest is in whether receiving aspirin increased or decreased the probability of experiencing a heart attack. We can model this using the two-sample Bernoulli model considered in the preceding example. Thus we let $X_i$ denote the heart attack outcome (1=yes, 0=no) for the $i$th participant in the aspirin group, and $Y_i$ denote the heart attack outcome in the $i$th participant in the placebo group.

The estimated 5-year risk of a heart attack for the aspirin group is $139/11037 = 0.013$. The estimated 5-year risk of a heart attack for the placebo group is $239/11034 = 0.022$. The estimated difference in probability of heart attack (aspirin minus placebo) is thus

$$\frac{139}{11037} - \frac{239}{11034} = -0.009$$

An asymptotically valid 95% confidence interval for $p_X - p_Y$ is

$$\frac{139}{11037} - \frac{239}{11034} \pm 1.96\sqrt{\frac{\frac{139}{11037}\left(1-\frac{139}{11037}\right)}{11037} + \frac{\frac{239}{11034}\left(1-\frac{239}{11034}\right)}{11034}}$$

which gives $(-0.012, -0.006)$.

Now consider a two-sided test of the null that $p_X = p_Y$. We shall use the second version of the test statistic, as given in Equation (5.5):

$$\frac{\frac{139}{11037} - \frac{239}{11034}}{\sqrt{\check{p}(1-\check{p})\left(\frac{1}{11037}+\frac{1}{11034}\right)}}$$

where $\check{p} = (239+139)/(11034+11037) = 0.017$. The test statistic evaluates to $-5.191$. The corresponding p-value is

$$2 \times P(Z > |-5.191|) = 2.09 \times 10^{-7}$$

(using `2*pnorm(5.191, lower.tail=FALSE)`). This is very small – we have strong evidence against the null hypothesis being true.

## 5.3 Independent samples - comparing variances

In this section we consider a setting where we have $X_1, \dots, X_n$ i.i.d. from some common distribution and $Y_1, \dots, Y_m$ from some other (in general different) distribution. We will initially construct estimators, confidence intervals and hypothesis tests based on normality assumptions for the two distributions. We will then discuss whether our procedures would be expected to be robust to deviations from the normality assumption.

Let $\sigma_X^2$ and $\sigma_Y^2$ denote the variances of the two distributions. We know we can unbiasedly estimate each by $S^2$. In the first group we have

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

Now suppose that the true distribution in the first group is normal, i.e. $X_1, \dots, X_n$ are i.i.d. $N(\mu_X, \sigma_X^2)$. Then

$$U_X = \frac{n-1}{\sigma_X^2} S_X^2 \sim \chi^2_{n-1}.$$

For the second group we have

$$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2,$$

and assuming normality for the outcomes $Y$,

$$U_Y = \frac{m-1}{\sigma_Y^2} S_Y^2 \sim \chi^2_{m-1}.$$

We will base our test for equality of $\sigma_X^2$ and $\sigma_Y^2$ on the ratio of these quantities:

$$\frac{S_X^2}{S_Y^2} = \left(\frac{\sigma_X^2}{\sigma_Y^2}\right)\frac{U_X/(n-1)}{U_Y/(m-1)} = \frac{\sigma_X^2}{\sigma_Y^2} W$$

where

$$W = \frac{U_X/(n-1)}{U_Y/(m-1)}.$$

The distribution of $W$ is the F-distribution.

**Definition 5.2** (F-distribution)**.** Let $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ be independent $\chi^2$ random variables. The distribution of the ratio:

$$W \quad = \quad \frac{U/\nu_1}{V/\nu_2}$$

is the $F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom.

Since $X_1, \dots, X_n$ are independent of $Y_1, \dots, Y_m$, it follows that $S_X^2$ and $S_Y^2$ are independent so from Definition 5.2,

$$W = \frac{U_X/(n-1)}{U_Y/(m-1)} \sim F_{n-1,m-1}. \tag{5.6}$$

Suppose we want to test hypotheses of the form

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad \begin{cases} H_1 : \sigma_X^2 < \sigma_Y^2 \\ H_1 : \sigma_X^2 > \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{cases}$$

Under $H_0$ we have that

$$W = \frac{U_X/(n-1)}{U_Y/(m-1)} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{S_X^2}{S_Y^2}$$

which is distributed as $F_{n-1,m-1}$. Thus we can use $S_X^2/S_Y^2$ as the test statistic.

**CASE 1:** $H_0 : \sigma_X^2 = \sigma_Y^2$ **versus** $H_1 : \sigma_X^2 < \sigma_Y^2$

If $H_1$ is true then $\sigma_X^2/\sigma_Y^2 < 1$ and so $s_X^2/s_Y^2$ being small supports the alternative hypothesis. We set a critical region of the form

$$C \quad = \quad \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : s_X^2/s_Y^2 \leq k_1\}$$

where $k_1$ is chosen such that

$$P(S_X^2/S_Y^2 \leq k_1 \mid H_0 \text{ true}) \ = \ P(W \leq k_1) \ = \ \alpha$$

with $W \sim F_{n-1,m-1}$ for a test with significance level $\alpha$. So we have $k_1 = F_{n-1,m-1,\alpha}$, where $F_{n-1,m-1,q}$ denotes the $q$ quantile of the F-distribution with $n-1$ and $m-1$ degrees of freedom.

**CASE 2:** $H_0 : \sigma_X^2 = \sigma_Y^2$ **versus** $H_1 : \sigma_X^2 > \sigma_Y^2$

If $H_1$ is true then $\sigma_X^2/\sigma_Y^2 > 1$ and so $s_X^2/s_Y^2$ being large supports the alternative hypothesis. We set a critical region of the form

$$C \quad = \quad \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : s_X^2/s_Y^2 \geq k_2\}$$

where $k_2 = F_{n-1,m-1,1-\alpha}$, for a test with significance level $\alpha$.

**CASE 3:** $H_0 : \sigma_X^2 = \sigma_Y^2$ **versus** $H_1 : \sigma_X^2 \neq \sigma_Y^2$

If $H_1$ is true then, then $s_X^2/s_Y^2$ being either small or large supports the alternative hypothesis. We set a critical region of the form

$$C \quad = \quad \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : s_X^2/s_Y^2 \leq k_1, \ s_X^2/s_Y^2 \geq k_2\}$$

where $k_1$ and $k_2$ are chosen such that

$$\begin{aligned} P(S_X^2/S_Y^2 \leq k_1 \mid H_0 \text{ true}) &= P(W \leq k_1) \ = \ \alpha/2 \\ P(S_X^2/S_Y^2 \geq k_2 \mid H_0 \text{ true}) &= P(W \geq k_2) \ = \ \alpha/2 \end{aligned}$$

Hence we choose $k_1 = F_{n-1,m-1,\alpha/2}$ and $k_2 = F_{n-1,m-1,1-\alpha/2}$.

For a given significance level $\alpha$, we can summarise the testing procedures in Table 5.4.

Table 5.4: Critical regions for one-sided and two-sided $F$-tests.

| $H_0$ | vs. $H_1$ | Reject $H_0$ if |
|---:|---:|---:|
| $\sigma_X^2 = \sigma_Y^2$ | $\sigma_X^2 < \sigma_Y^2$ | $s_X^2/s_Y^2 \leq F_{n-1,m-1,\alpha}$ |
| $\sigma_X^2 = \sigma_Y^2$ | $\sigma_X^2 > \sigma_Y^2$ | $s_X^2/s_Y^2 \geq F_{n-1,m-1,1-\alpha}$ |
| $\sigma_X^2 = \sigma_Y^2$ | $\sigma_X^2 \neq \sigma_Y^2$ | $s_X^2/s_Y^2 \geq F_{n-1,m-1,1-\alpha/2}$<br>or $s_X^2/s_Y^2 \leq F_{n-1,m-1,\alpha/2}$ |

**Remark: How do we find the quantiles and probabilities of the chi-squared distribution?**

If $W \sim F_{\nu_1,\nu_2}$ let $P(W \leq F_{\nu_1,\nu_2,p}) = p$.

- In the University Formula Book, Section B5 tabulates the percentage points of the F-distribution. Note that it gives **upper-tail values** and so you will need to take care that you find the correct value. Thus, if you want a lower tail probability of $p$ this corresponds to an upper tail probability of $1-p$. There are four separate tables corresponding to the upper 5%, 2.5%, 1%, and 0.5% values respectively. In the table, the columns correspond to $\nu_1$ and the rows to $\nu_2$. For example, $P(F_{10,12} \leq 2.75) = 0.95$ since (using the table) $P(F_{10,12} > 2.75) = 0.05$. Notice that $F_{\nu_1,\nu_2} = 1/F_{\nu_2,\nu_1}$ so that $P(F_{10,12} < 0.34) = 0.05$ since $P(F_{12,10} > 2.91) = 0.05$ and $1/2.91 = 0.34$.

- In R, for a given $\nu_1$, $\nu_2$, and $p$, $F_{\nu_1,\nu_2,p}$ may be found using the command qf($p, \nu_1, \nu_2$). For a given $F_{\nu_1,\nu_2,p}$, $p$ may be found using the command pf($F_{\nu_1,\nu_2,p}, \nu_1, \nu_2$).

```
# Let W be an F-distribution with 10, 12 degrees of freedom
qf(0.95,10, 12) # finds w such that P(W <= w) = 0.95
```

```
## [1] 2.753387
```

```
qf(0.05, 10, 12) # finds v such that P(W <= w) = 0.05
```

```
## [1] 0.3432914
```

```
pf(2.75, 10, 12)   # finds P(W <= 2.75)
```

```
## [1] 0.9498009
```

- To directly compute an upper tail probability, or corresponding quantile, in R you need to change the default in pf or qf to calculate the upper tail.

```
# Let W be an F-distribution with 10, 12 degrees of freedom
qf(0.05, 12, 10, lower.tail=F) # finds w such that P(W > w) = 0.05
```

```
## [1] 2.912977
```

```
pf(2.91, 12, 10, lower.tail=F) # using the upper tail finds P(W > 2.91)
```

```
## [1] 0.05015525
```

$W$ as given in Equation (5.6) has a distribution that does not depend on any unknown parameters. It is thus a pivot and can be used to find a confidence interval for the ratio $\sigma_X^2/\sigma_Y^2$, but we omit the details of this.

**Example 5.8.** Suppose we are interested in comparing the variability in patients' CD4 count at week 20 in the ACTG175 trial in those randomised to zidovudine and those randomised to zidovudine plus didanosine. We assume that the CD4 counts in the zidovudine group are i.i.d. $N(\mu_X, \sigma_X^2)$ and the CD4 counts in the zidovudine plus didanosine group are i.i.d. $N(\mu_Y, \sigma_Y^2)$.

Suppose we want to test the null of equal variances against the two-sided alternative that they differ, at the 5% significance level. The critical region is

$$C = \{(x_1, \ldots, x_{532}, y_1, \ldots, y_{522}) : s_X^2/s_Y^2 \leq k_1 \text{ or } s_X^2/s_Y^2 \geq k_2\}$$

where $k_1 = F_{531,521,0.025} = 0.843$ (using `qf(0.025,531,521)`) and $k_2 = F_{531,521,0.975} = 1.187$ (using `qf(0.975,531,521)`). The corresponding sample variances are $s_X^2 = 17150.93$ and $s_Y^2 = 24430.96$. The ratio is thus 0.702, and we reject the null hypothesis of equal variances as $0.702 < 0.843$.

What if the normality assumptions do not hold? It turns out that the F-test for comparing variances between two groups and corresponding confidence interval for the ratio of variances are **not robust** to violations of the normality assumption. Given Figure 5.2 for the CD4 data at 20 weeks in the ACTG175 trial, we ought to be somewhat concerned that the variance comparison test just performed may not be valid, in the sense that the actual Type I error may deviate from the desired level. We will not cover them here, but alternative approaches which are robust to non-normality exist.

## 5.4 Testing independence in contingency tables

So far we have considered scenarios where we observe two sets of numeric random variables. Now we consider the case where we have measured two factors (discrete random variables) on a set of units. This type of data can be displayed as a **contingency table** consisting of the counts of sampled units for each combination of factor levels. A common hypothesis of interest is whether the two factors measured are statistically independent.

**Example 5.9** (Physicians Health Study)**.** Consider again the Physician's Health Study data, which we show here again in Table 5.5 for convenience.

Table 5.5: Number of heart attacks by treatment group in the Physician's Health Study.

| Group | Heart attack | No heart attack | Total |
|---|---:|---:|---:|
| Aspirin | 139 | 10898 | 11037 |
| Placebo | 239 | 10795 | 11034 |

This table is a $2 \times 2$ contingency table. We now consider this data from the perspective of testing independence between the participant's randomised treatment group and their heart attack outcome (yes/no). If we now let $D$ and $H$ denote random variables indicating the drug a participant received ($D = 1$ for aspirin, $D = 0$ for placebo) and the heart attack outcome ($H = 1$ for heart attack, $H = 0$ for no heart attack) respectively. Then our null hypothesis is that $D$ and $H$ are independent random variables.

### 5.4.1 Pearson's $\chi^2$ test

We now introduce Pearson's $\chi^2$ test. Before doing so, we introduce the multinomial distribution.

**Definition 5.3** (Multinomial distribution)**.** Suppose that a random variable takes values in $\{1, \dots, k\}$, where it takes value $i$ with probability $p_i$, with $\sum_{i=1}^k p_i = 1$. Suppose we take $n$ i.i.d. draws from this distribution, and let $Y_i$ denote the number of observations which take value $i$. Thus each $Y_i$ takes a value between 0 and $n$. Then $(Y_1, \dots, Y_k)$ follows the **multinomial distribution**, with

$$P(Y_1 = y_1, \dots, Y_k = y_k) = \frac{n!}{y_1! \dots y_k!} p_1^{y_1} \dots p_k^{y_k}$$

with $\sum_{i=1}^k y_i = n$ and $\frac{n!}{y_1! \dots y_k!}$ is the number of ways $n$ objects can be grouped into $k$ classes, with $y_i$ objects in the $i$th class.

**Definition 5.4** (Pearson's chi squared statistic)**.** Suppose we have a vector of counts $(Y_1, \dots, Y_k)$, which follows a multinomial distribution. Suppose further that we have a model with parameter $\theta$, and an estimate of this from the data, denoted $\hat{\theta}$. Under this model we can calculate the expected number of values in each class if the true parameter value were equal to $\hat{\theta}$:

$$E_i = E(Y_i|\theta = \hat{\theta}) = np_i(\hat{\theta})$$

where $p_i(\hat{\theta})$ denotes the probability of being in the $i$th class when the model parameter is set to $\hat{\theta}$. The $E_i$, $i = 1, \dots, k$ are the expected values under the model. **Pearson's chi squared statistic** compares these expected values $(E_i)$ with the observed values $(O_i = Y_i)$:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The larger the value of $X^2$, the worse the fit of the model to the data. It can be shown that if the model is correctly specified, then $X^2$ is approximately $\chi^2_\nu$ distributed, with degrees of freedom equal to $\nu = k - p - 1$ where $k$ is the number of classes and $p$ is the number of parameters in $\theta$. The approximation works best when $n$ is large, with one commonly accepted rule of thumb being to check whether the expected counts are at least 5.

Given an observed value of $X^2$, denoted $x^2$, we can calculate a p-value. Larger values of $X^2$ correspond to the data supporting the alternative hypothesis that the model does not fit the data well, and so we have

$$p(x^2) = P(X^2 \geq x^2|\text{model is correct}) \approx P(\chi^2_\nu \geq x^2)$$

**Example 5.10** (Physicians Health Study continued)**.** Now we continue with Example 5.9. In this case we have $k = 4$ classes, consisting of the four combinations of treatment group and heart attack outcome. If we assume $D$ and $H$ for each individual are independent random variables, we can model $D$ and $H$ separately. Since each is a Bernoulli random variable, we have a 'success' probability $p_D$ and $p_H$ for each.

The MLEs are

$$\begin{aligned} \hat{p}_D &= 11037/(11037 + 11034) = 0.500, \\ \hat{p}_H &= (139 + 239)/(11037 + 11034) = 0.017. \end{aligned}$$

Under an assumption/model of independence between $D$ and $H$, we have $P(D = d, H = d) = P(D = d)P(H = h)$. The number of individuals we expect to be in the aspirin group and to have a heart attack is

$$n \times \hat{P}(D = 1)\hat{P}(H = 1) \quad = \quad (11037 + 11034) \times \hat{p}_D \times \hat{p}_H = 189.026$$

and similarly for the other three cells in the table. Table 5.6 gives the expected counts under the null hypothesis of independence.

Table 5.6: Expected number of heart attacks by treatment group in the Physician's Health Study, assuming independence.

| Group | Heart attack | No heart attack | Total |
|---|---|---|---|
| Aspirin | 189.026 | 10847.97 | 11037 |
| Placebo | 188.974 | 10845.03 | 11034 |

Pearson’s chi-squared statistic is then

$$\begin{aligned} & \frac{(139-189.026)^2}{189.026}+\frac{(239-188.974)^2}{188.974} \\ +&\frac{(10898-10847.974)^2}{10847.974}+\frac{(10795-10845.026)^2}{10845.026} = 26.944 \end{aligned}$$

Under the null hypothesis of independence this is a draw from a chi-squared distribution on $4-2-1=1$ degree of freedom. Using the code `pchisq(26.944, df=1, lower.tail=FALSE)`, the p-value is

$$P(\chi^2_1 \geq 26.944) = 0.0000002$$

Thus we reject the null hypothesis of independence between randomised treatment group and heart attack outcome. We conclude that the probability of having a heart attack does depend on your treatment group, with the probability being lower if you are assigned aspirin.

It in fact turns out that this test is identical to the two-sided test based on Equation (5.5). The chi-squared test statistic we have calculated here is precisely the square of the test statistic calculated in Example 5.7 using Equation (5.5).

We will not cover it here, but the chi-squared test for independence readily extends from the $2 \times 2$ case to the more general $r \times s$ case.

### 5.4.2 Goodness of fit tests

The chi-squared test for independence is an example of a **goodness of fit** test. As we have mentioned previously, a concern when using statistical models is that some of the assumptions made by the model may not hold. Or put another way, the true data generating distribution may not belong to the class of distributions encoded by the model. One response to this is to develop models and methods which make fewer distributional assumptions. We have pursued this route previously, in examining the robustness of some of our procedures to violations of their distributional assumptions. An alternative route is to examine how well the data conform to our chosen model. If we find that it doesn't fit well, we can try and change the model to fit it better.

The chi-squared test can be used for assessing goodness of fit for discrete random variables in general (and discretized/binned continuous random variables).

**Example 5.11.** Greenwood and Yule in 1920 published an analysis investigating the number of accidents over a five week period among 647 women working in a factory manufacturing shells. Table 5.7 shows the distribution of the number of accidents among the women.

Table 5.7: Greenwood and Yule (1920) data on number of accidents among 647 women working in a shell factory over 5 weeks.

| Number of accidents | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Number of women | 447 | 132 | 42 | 21 | 3 | 2 |

A possible model for this setting is the Poisson. Let $x_1, \dots, x_{647}$ denote the number of accidents experienced by the 647 women. Then we assume these are 647 i.i.d. realisations of a Poisson random variable with parameter $\lambda$. We showed previously that the maximum likelihood estimate for the Poisson model is $\hat{\lambda} = \overline{x} = n^{-1} \sum_{i=1}^{n} x_i$. This sum can be calculated from the table as

$$(0 \times 447) + (1 \times 132) + (2 \times 42) + (3 \times 21) + (4 \times 3) + (5 \times 2) = 301$$

so that $\hat{\lambda} = 301/647 = 0.465$. We can now use the same approach used above for contingency tables to test if the data are compatible with a Poisson model being correct. In principle there is no maximum value $k$ for the number of accidents each of the women could have experienced over a 5 week period. We will thus proceed by having classes $\{0, 1, 2, 3, 4, \geq 5\}$. Under the Poisson model we found $\hat{\lambda} = 0.465$, so the expected values of $Y_1, \dots, Y_6$ in the 6 classes are, under the Poisson model at the estimated parameter value:

$$\begin{aligned}
E_1 &= n \times P(X = 0) = n \times \frac{e^{-0.465} 0.465^0}{0!} = 406.403 \\
E_2 &= n \times P(X = 1) = n \times \frac{e^{-0.465} 0.465^1}{1!} = 188.978 \\
E_3 &= n \times P(X = 2) = n \times \frac{e^{-0.465} 0.465^2}{2!} = 43.937 \\
E_4 &= n \times P(X = 3) = n \times \frac{e^{-0.465} 0.465^3}{3!} = 6.810 \\
E_5 &= n \times P(X = 4) = n \times \frac{e^{-0.465} 0.465^4}{4!} = 0.792 \\
E_6 &= n \times P(X \geq 5) = n \times (1 - P(X < 5)) = 0.080
\end{aligned}$$

Table 5.8 shows the observed and expected counts.

Table 5.8: Observed and expected counts in the Greenwood and Yule (1920) data.

| Number of accidents | 0 | 1 | 2 | 3 | 4 | 5+ |
|---|---|---|---|---|---|---|
| Observed | 447 | 132 | 42 | 21 | 3 | 2 |
| Expected | 406.403 | 188.978 | 43.937 | 6.810 | 0.792 | 0.080 |

Based on our rule of thumb that the expected counts should be at least 5, we should not proceed with carrying out the chi-squared test. One solution is to pool the last three columns into a combined 'three or more' category as shown in Table 5.9 shows the observed and expected counts.

Table 5.9: Observed and expected counts in the Greenwood and Yule (1920) data, pooled so that the expected counts are at least five.

| Number of accidents | 0 | 1 | 2 | 3+ |
|---|---|---|---|---|
| Observed | 447 | 132 | 42 | 26 |
| Expected | 406.403 | 188.978 | 43.937 | 7.682 |

The chi-squared statistic is then

$$\begin{aligned} x^2 &= \frac{(447-406.403)^2}{406.403}+\frac{(132-188.978)^2}{188.978} \\ &\quad +\frac{(42-43.937)^2}{43.937}+\frac{(26-7.682)^2}{7.682} \\ &= 64.9999. \end{aligned}$$

Under the null hypothesis that a Poisson model is correctly specified, this is a random draw from a chi-squared distribution with $4-1-1$ degrees of freedom, since the Poisson model has one unknown parameter. The p-value is then

$$P(\chi^2_2 \geq 64.9999) = 7.681609 \times 10^{-15}$$

which is very small. We have strong evidence against a Poisson model being correct here. An explanation for this is that in fact the women varied with respect to their propensity to have accidents, perhaps because some worked in more dangerous parts of the factory than others. In this case, it would not be the case that the number of accidents for each woman was a draw from a Poisson with a common rate $\lambda$, and in this case our i.i.d. Poisson model would be incorrect.

# Chapter 6

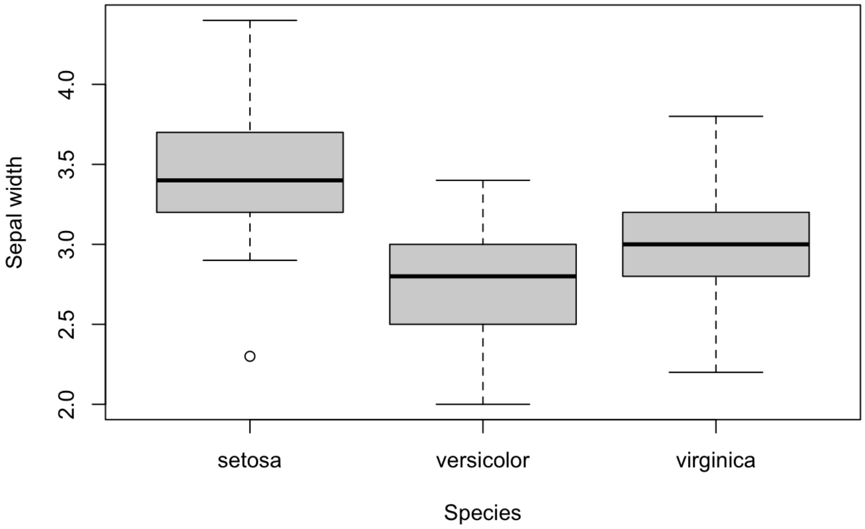
# One-way Analysis of Variance

In Section 5.2 we compared two normal means: we considered observations $X_1, \ldots, X_n$ i.i.d. from $N(\mu_X, \sigma_X^2)$, and $Y_1, \ldots, Y_m$ i.i.d. from $N(\mu_Y, \sigma_Y^2)$ where the $X_i$ are independent of the $Y_j$. Under the assumption that $\sigma_X = \sigma_Y$, we were able, in Section 5.2.2, to develop the independent sample t-test. We now consider the extension to comparing $k$ means for $k > 2$.

We consider the case where we have $k$ groups, or **treatments**, and we observe $n_i$ individuals in the $i$th treatment, $i = 1, \ldots, k$. The total number of observations is $n = \sum_{i=1}^{k} n_i$. Note that we do not assume that there are equal numbers of observations in each treatment. We let $X_{ij}$ denote the observation of the $j$th individual in the $i$th treatment. This is an example of a **one-way layout**.

**Example 6.1.** The Iris flower data set consists of 50 samples from each of three species of Iris: Iris setosa, Iris versicolor, and Iris virginica. For each sample, four features were measured: the length and width of the sepals and petals, in centimetres. We will restrict attention to the sepal width measurements. The dataset is available in `R` as `iris`. We create a boxplot of the sepal width measurements for the three species.

```r
boxplot(Sepal.Width~Species,data=iris,ylab="Sepal width")
```



We are interested in whether the mean sepal width depends upon the species. The boxplots suggest that an assumption of equal variance in each species may be reasonable (we could consider statistical ways to test this) and we will additionally assume that the measurements in each species are normally distributed. We will construct a hypothesis test to test whether the mean sepal length is the same for all species.

**Definition 6.1** (One-way analysis of variance model)**.** We assume that random variables $X_{ij}$ are given by the model

$$X_{ij} \quad = \quad \mu_i + \epsilon_{ij} \tag{6.1}$$

where the $\mu_i$ are unknown parameters, and the $\epsilon_{ij}$ are mutually independent zero mean normal random variables, each with the same finite variance $\sigma^2$.

It follows that $X_{ij} \sim N(\mu_i, \sigma^2)$ and that observations within the same treatment are independent and are independent across treatments. Thus, $E(X_{ij}) = \mu_i$ and the $\mu_i$s are typically known as **treatment means**. Thus, if $\mu_i = \mu$ for all $i$ then all treatments have the same expected response.

We will consider comparisons of the means using a method known as the **analysis of variance (ANOVA)**. The approach considers the way we measure the variance in the data and align it to different sources. In particular, we show that we can partition the total variance into the variance between the different treatments and to random error which is the variance within treatments.

Let $\overline{x}_{i\cdot}$ denote the mean of the observations in the $i$th treatment. Thus,

$$\overline{x}_{i\cdot} \quad = \quad \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}.$$

The mean of all of the observations is then

$$\overline{x}_{\cdot\cdot} \quad = \quad \frac{1}{n}\sum_{i=1}^{k} n_i \overline{x}_{i\cdot}.$$

Note the notation here: there are two suffixes $i$ and $j$ and we use a dot in place of a suffix to indicate averaging out over the suffix. Thus, $\overline{x}_{i\cdot}$ denotes averaging out over $j$ and $\overline{x}_{\cdot\cdot}$ averaging out over both $i$ and $j$.

**Definition 6.2** (Sums of squares)**.** For the observations $x_{ij}$, the total sum of squares is defined to be

$$\mathrm{SS_{Tot}} \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{\cdot\cdot})^2.$$

The sum of squares within treatments, also known as the residual sum of squares, is defined to be

$$\mathrm{SS_W} \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i\cdot})^2.$$

The sum of squares between treatments is defined to be

$$\mathrm{SS_B} \quad = \quad \sum_{i=1}^{k} n_i(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2.$$

**Theorem 6.1** (Partition of sum of squares)**.** *The total sum of squares may be expressed as the sum of squares within treatments and the sum of squares between treatments. That is*

$$SS_{Tot} \quad = \quad SS_W + SS_B. \tag{6.2}$$

**Proof:** From Definition 6.2

$$\begin{aligned}
\mathrm{SS}_{\mathrm{Tot}} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\{(x_{ij}-\overline{x}_{i\cdot})+(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot})\}^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot})^2 \\
&\quad +2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot}). \qquad (6.3)
\end{aligned}$$

Note that, as $(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot})$ does not depend on $j$,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot})^2 = \sum_{i=1}^{k}n_i(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot})^2 = \mathrm{SS}_{\mathrm{B}} \qquad (6.4)$$

whilst

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot}) = \sum_{i=1}^{k}(\overline{x}_{i\cdot}-\overline{x}_{\cdot\cdot})\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot}) = 0 \qquad (6.5)$$

as $\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot}) = n_i\overline{x}_{i\cdot}-n_i\overline{x}_{i\cdot}=0$. Substituting equations (6.4) and (6.5) into (6.3) and noting that $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_{i\cdot})^2=\mathrm{SS}_{\mathrm{W}}$ gives equation (6.2). □

## 6.1 Expectations of the sum of squares

We follow the usual notation that by capitalising the observations we denote the corresponding random variable. Using the standard properties of linear sums of independent normal random variables we can immediately deduce that, for each $i=1,\ldots,k$,

$$\overline{X}_{i\cdot} \sim N(\mu_i,\sigma^2/n_i) \qquad (6.6)$$

and

$$\overline{X}_{\cdot\cdot} \sim N(\overline{\mu},\sigma^2/n) \qquad (6.7)$$

where

$$\overline{\mu} = \frac{1}{n}\sum_{i=1}^{k}n_i\mu_i \qquad (6.8)$$

is a weighted average of the individual treatment means, weighted accorded to the corresponding sample sizes.

**Theorem 6.2.** *For the one-way analysis of variance model given in Definition 6.1, $\frac{1}{\sigma^2}SS_W \sim \chi^2_{n-k}$ so that $E(SS_W/(n-k)) = \sigma^2$.*

**Proof:** From single sample normal theory, for each $i = 1, \dots, k$, $\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2 \sim \chi^2_{n_i-1}$. As observations are independent between treatments then $\frac{1}{\sigma^2}\mathrm{SS_W}$ is thus the sum of $k$ independent $\chi^2_{n_i-1}$ random variables. As $\sum_{i=1}^{k}(n_i-1) = n-k$, it follows from the definition of the $\chi^2$-distribution, see Definition 3.3, that $\frac{1}{\sigma^2}\mathrm{SS_W} \sim \chi^2_{n-k}$ and thus $E(\mathrm{SS_W}/(n-k)) = \sigma^2$. □

Note that $\mathrm{SS_W}/(n-k)$ is the generalisation from two treatments to $k$ treatments of the pooled sample estimator $S_p^2$ given in Definition 5.1.

**Theorem 6.3.** *For the one-way analysis of variance model given in Definition 6.1,*

$$E(SS_B) = (k-1)\sigma^2 + \sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2.$$

*Thus, $E(SS_B/(k-1)) \geq \sigma^2$ with equality if and only if $\mu_i$ does not depend on $i$.*

**Proof:** Note that

$$\begin{aligned}\sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 &= \sum_{i=1}^{k} n_i(\overline{X}^2_{i\cdot} - 2\overline{X}_{i\cdot}\overline{X}_{\cdot\cdot} + \overline{X}^2_{\cdot\cdot}) \\ &= \sum_{i=1}^{k} n_i\overline{X}^2_{i\cdot} - n\overline{X}^2_{\cdot\cdot}\end{aligned}$$

so that

$$E(\mathrm{SS_B}) = \sum_{i=1}^{k} n_i E(\overline{X}^2_{i\cdot}) - nE(\overline{X}^2_{\cdot\cdot}). \tag{6.9}$$

Now, from equation (6.6),

$$E(\overline{X}^2_{i\cdot}) = Var(\overline{X}_{i\cdot}) + E^2(\overline{X}_{i\cdot}) = \frac{\sigma^2}{n_i} + \mu_i^2 \tag{6.10}$$

and, from equation (6.7),

$$E(\overline{X}^2_{\cdot\cdot}) = Var(\overline{X}_{\cdot\cdot}) + E^2(\overline{X}_{\cdot\cdot}) = \frac{\sigma^2}{n} + \overline{\mu}^2. \tag{6.11}$$

Substituting equations (6.10) and (6.11) into equation (6.9) gives

$$\begin{aligned} E(\mathrm{SS_B}) &= \sum_{i=1}^{k} n_i\left(\frac{\sigma^2}{n_i}+\mu_i^2\right)-n\left(\frac{\sigma^2}{n}+\overline{\mu}^2\right) \\ &= (k-1)\sigma^2+\sum_{i=1}^{k} n_i\mu_i^2-n\overline{\mu}^2 \\ &= (k-1)\sigma^2+\sum_{i=1}^{k} n_i(\mu_i-\overline{\mu})^2 \end{aligned}$$

It thus follows that $E(\mathrm{SS_B}/(k-1)) \geq \sigma^2$ with equality if and only if $\mu_i = \overline{\mu}$ for each $i$ which is that $\mu_i$ does not depend on $i$. □

## 6.2 Hypothesis testing for equality of means

We test the hypotheses

$$H_0 : \mu_i = \mu \ \forall i = 1, \ldots, k \quad \text{versus} \quad H_1 : \mu_i \neq \mu_{i'} \text{ for some } i, i'.$$

Note that if $H_0$ is rejected, we conclude that at least two of the $\mu_i$s are different but there no inference as to where the difference may be.

If $H_0$ is true then we have $n$ independent and identically distributed observations $X_{ij} \sim N(\mu, \sigma^2)$ so that

$$\frac{1}{\sigma^2}\mathrm{SS_{Tot}} \quad \sim \quad \chi^2_{n-1}.$$

Noting that $\frac{1}{\sigma^2}\mathrm{SS_W} \sim \chi^2_{n-k}$ and $\mathrm{SS_{Tot}} = \mathrm{SS_W} + \mathrm{SS_B}$ we may suspect that $\frac{1}{\sigma^2}\mathrm{SS_B} \sim \chi^2_{k-1}$ as $n-1 = (n-k)+(k-1)$ and the property of sums of independent $\chi^2$ random variables. The following theorem shows that this is indeed the case.

**Theorem 6.4.** *For the one-way analysis of variance model given in Definition 6.1, $SS_W$ and $SS_B$ are independent. If $H_0$ is true then $\frac{1}{\sigma^2}SS_B \sim \chi^2_{k-1}$.*

**Proof:** Note that $\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2$ is independent of $\overline{X}_{i\cdot}$ by the independence of the sample variance and sample mean of the $i$th sample, see Theorem 3.1. Equally, for $i \neq i'$, $\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2$ is independent of $\overline{X}_{i'\cdot}$ as these correspond to different samples. Thus, $\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2$ is independent of $\overline{X}_{i'\cdot}$ for all $i'$ and consequently is independent of $\overline{X}_{\cdot\cdot}$ as this is a linear sum of the $\overline{X}_{i'\cdot}$. It thus follows that $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent. If $H_0$ is true then, as $\mathrm{SS_{Tot}}$ is the sum of the independent random variables $\mathrm{SS_W}$ and $\mathrm{SS_B}$, with $\frac{1}{\sigma^2}\mathrm{SS_{Tot}} \sim \chi^2_{n-1}$

and $\frac{1}{\sigma^2}\mathrm{SS_W} \sim \chi^2_{n-k}$ if follows from properites of $\chi^2$ random variables that $\frac{1}{\sigma^2}\mathrm{SS_B} \sim \chi^2_{k-1}$. □

Noting the definition of an F-distribution, see Definition 5.2, we have the following corollary.

**Corollary 6.1.** *For the one-way analysis of variance model given in Definition 6.1, if $H_0$ is true then*

$$F \quad = \quad \frac{SS_B/(k-1)}{SS_W/(n-k)} \;\sim\; F_{k-1,n-k}.$$

We can thus use $F$ as our test statistic for our hypothesis test. Note that, from Theorem 6.3, $E(\mathrm{SS_B})$ is larger under $H_1$ and so large values of $F$ correspond to evidence against $H_0$. Thus, for a test of significance level $\alpha$, we have a critical region of the form

$$\mathcal{C} \quad = \quad \{(x_{ij}, i = 1, \dots k, j = 1, \dots n_i) : F \geq F_{k-1,n-k,1-\alpha}\}.$$

The construction of the test is typically displayed in an ANOVA table as shown below.

| Source | df | SS | MS | F-statistic |
|---|---|---|---|---|
| Between | $k-1$ | $\mathrm{SS_B} = \sum_{i=1}^{k} n_i(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2$ | $\mathrm{MS_B} = \mathrm{SS_B}/(k-1)$ | $F = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$ |
| Within | $n-k$ | $\mathrm{SS_W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i\cdot})^2$ | $\mathrm{MS_W} = \mathrm{SS_W}/(n-k)$ | |
| Total | $n-1$ | $\mathrm{SS_{Tot}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{\cdot\cdot})^2$ | | |

where df denotes the degrees of freedom, SS the sum of squares and MS the mean square. The quantities on the bottom row are equal to the sum of the corresponding quantities on the preceeding two rows.

Calculating the various sums of squares by hand, especially for large datasets, is time consuming. In `R` we can perform the ANOVA using the `aov` function and display the ANOVA table using the `anova` function. We now utilise this for the sepal length measurements.

**Example 6.2.** We perform the ANOVA for the sepal width data of Example 6.1.

```
anova_iris <- aov(Sepal.Width~Species, data=iris)
anova(anova_iris)
```

```
## Analysis of Variance Table
##
## Response: Sepal.Width
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 11.345  5.6725   49.16 < 2.2e-16 ***
## Residuals 147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If $H_0$ is true then $F = 49.16$ should be a realisation from an F-distribution with 2 and 147 degrees of freedom (recall that there are three species and 50 observations for each species). We verify the p-value calculation in the table. We can access the individual data points by their row/column entry. Thus, the F value is given by `anova(anova_iris)[1,4]`

```
(Fval <- anova(anova_iris)[1,4])
```

```
## [1] 49.16004
```

```
pf(Fval,2,147,lower.tail=FALSE)
```

```
## [1] 4.492017e-17
```

```
(pval <- anova(anova_iris)[1,5])
```

```
## [1] 4.492017e-17
```

The tiny p-value means that we reject $H_0$. The mean sepal widths are not all the same for each species.

The conclusion of the F-test is, perhaps, disappointing in that we conclude that the means are different but we do not know how they differ or which pairs are significantly different. To proceed further involves considering, for example, pairwise differences simultaneously. This leads to problem of **multiple comparisons** and preserving the type I error rate. Although outside the scope of this course, approaches include the **Bonferroni method**.

# Chapter 7

# Simple Linear Model

In many situations, measurements on one quantity could be used to infer a measurement on a related quantity. For example, given the height of an individual could we predict their weight? Alternatively, you might be interested in predicting the blood pressure of an individual based on observations of their age, gender, weight, exercise levels, alcohol consumption, smoking status, and so on. In the former case, we have a single predictor and in the latter multiple predictors.

The following plot shows the height and weight of 350 individuals from a population.

```r
heightweight = read.csv("http://people.bath.ac.uk/masss/ma22014/
height-weight.csv")
plot(Weight ~ Height, heightweight)
```

We see that `Height` is positively correlated with `Weight`. It is clear, from the figure, that the observed data do not follow an exact linear relationship, but that's not surprising. But there is some kind of relationship - how can we describe that? One way to proceed is to formulate a linear statistical model of the way that the data were generated, and to use this as the basis for inference. Suppose that the weight of the $i$th individual ($y_i$) is a linear function of the height of the individual ($x_i$) plus some variability ($\epsilon_i$). This is an example of a **simple linear model**.

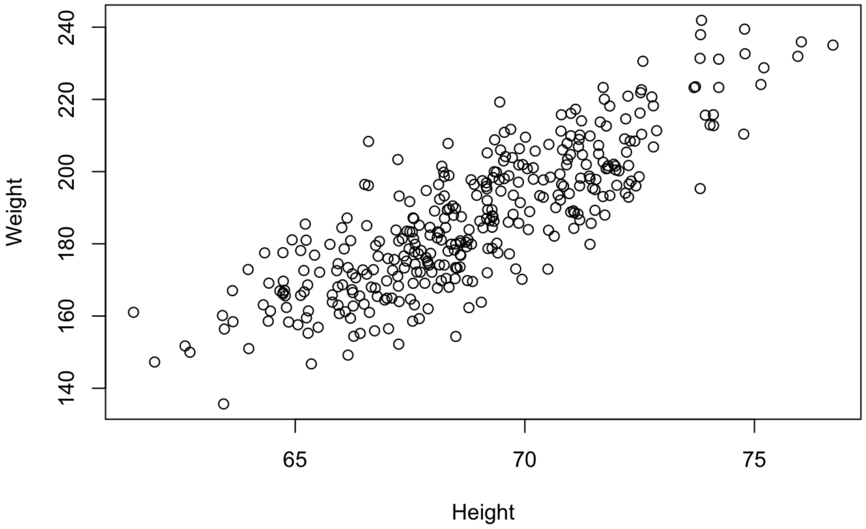


Figure 7.1: The weight of an individual, in pounds, is plotted against their height, in inches. There are 350 individuals in the sample.

**Definition 7.1** (Simple linear model)**.** Consider $n$ observations, $(x_i, y_i)$, where, for each $i = 1, \ldots, n$, $y_i$ is an observation on random variable, $Y_i$. The simple linear model is

$$Y_i \quad = \quad \alpha + \beta x_i + \epsilon_i \tag{7.1}$$

where $\alpha$, the intercept, and $\beta$, the slope, are unknown parameters and the $\epsilon_i$ are mutually independent zero mean random variables, each with the same finite variance $\sigma^2$.

Note that

$$E(Y_i) \quad = \quad \alpha + \beta x_i \tag{7.2}$$

and $\text{Var}(Y_i) = \sigma^2$. The random component of the model is intended to capture the fact that if we gathered a replicate set of data, for a new set of individuals, the underlying relationship would not change, but the random variation would be different. Notice that it is not implied that these errors are completely unpredictable: their mean and variance are assumed to be fixed, it is only their particular values, for any family, that are not known.

Given the variability, what can be inferred from these data? Let's focus on the slope $\beta$. In particular: (i) what value of $\beta$ is most consistent with the data?, (ii) what range of $\beta$ values is consistent with the data?, and (iii) are some particular, theoretically derived, values of $\beta$ consistent with the data? These cover our familiar statistical questions of point estimation, interval estimation and hypothesis testing.

The simple linear model states that $Y$ is given by a constant plus $x$ multiplied by a constant plus a random term. $Y$ is an example of a **response variable**, while $x$ is an example of a **predictor variable**. The model is called a *simple* linear model not just because it is simple - people use this term to refer to the situation where there's only one predictor $x$. This is **simple linear regression**: the term **linear regression** and **linear model** are used interchangeably by statisticians. Multiple linear regression is when you have more than one predictor and study of these models form the core of MA22015 Statistics 2B.

In this chapter we develop statistical methods for the simple linear model. This allows the key concepts of linear modelling to be introduced without the distraction of extra mathematical difficulty caused by more than one predictor.

## 7.1 Simple least squares estimation

How can $\alpha$ and $\beta$, in model (7.1), be estimated from the $x_i, y_i$ data? A sensible approach is to choose values of $\alpha$ and $\beta$ that makes the model fit closely to the data. To do this we need to define a measure of how well, or how badly, a model with a particular $\alpha$ and $\beta$ fits the data. One possible measure is the **residual sum of squares** (RSS) of the model.

**Definition 7.2** (Residual sum of squares)**.** The residual sum of squares (RSS) of the simple linear model is given by

$$\mathrm{RSS}(\alpha, \beta) \quad = \quad \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{7.3}$$

Thus, for each $i$, $y_i - \alpha - \beta x_i$ is the vertical distance of the observed $y_i$ from the model $E(Y_i) = \alpha + \beta x_i$. If we have chosen good values of $\alpha$ and $\beta$, close to the "true" values, then this distance should be small, so that $\mathrm{RSS}(\alpha, \beta)$ should be small. Conversely, poor choices will lead to $\alpha + \beta x_i$ far from their corresponding $y_i$, and high values of $\mathrm{RSS}(\alpha, \beta)$. Hence $\alpha$ and $\beta$ can be estimated by minimising RSS with respect to $\alpha$ and $\beta$ and this is known as the method of *least squares*.

**Theorem 7.1** (Least squares estimates)**.** *For the simple linear model the least squares estimates of $\alpha$ and $\beta$, $\hat{\alpha}$ and $\hat{\beta}$ respectively, are given by*

$$\hat{\alpha} \quad = \quad \overline{y} - \hat{\beta}\overline{x} \quad = \quad \sum_{i=1}^{n} \left\{ \frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}} \right\} y_i, \tag{7.4}$$

$$\hat{\beta} \quad = \quad \frac{S_{xy}}{S_{xx}} \quad = \quad \sum_{i=1}^{n} \left\{ \frac{(x_i - \overline{x})}{S_{xx}} \right\} y_i. \tag{7.5}$$

*whenever $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \neq 0$ where $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$.*

**Proof:** The first order partial derivatives are

$$\frac{\partial \mathrm{RSS}}{\partial \alpha} \quad = \quad -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i); \tag{7.6}$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta} \quad = \quad -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i). \tag{7.7}$$

Let $\hat{\alpha}$, $\hat{\beta}$ be, respectively, the values of $\alpha$ and $\beta$ for which the partial derivatives are simultaneously zero. Solving $\frac{\partial \mathrm{RSS}}{\partial \alpha} = 0$ gives, from equation (7.6),

$$\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) \ = \ 0 \quad \Rightarrow \quad \overline{y} - \hat{\alpha} - \hat{\beta}\overline{x} \ = \ 0.$$

Thus, $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$. Substituting this into equation (7.7) and solving equal to zero gives

$$\begin{aligned}\sum_{i=1}^{n} x_i\{(y_i-\overline{y})-\hat{\beta}(x_i-\overline{x})\} &= 0\\ \Rightarrow\ \hat{\beta}\sum_{i=1}^{n} x_i(x_i-\overline{x}) &= \sum_{i=1}^{n} x_i(y_i-\overline{y})\\ \Rightarrow\ \hat{\beta} &= \frac{S_{xy}}{S_{xx}}\end{aligned}$$

by noting that $\sum_{i=1}^{n} x_i(x_i-\overline{x}) = S_{xx}$ and $\sum_{i=1}^{n} x_i(y_i-\overline{y}) = S_{xy}$ and assuming that $S_{xx} > 0$ (for which having two distinct $x_i$ and $x_j$ is sufficient). Substituting $\hat{\beta}$ into $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$ gives the second equality in equation (7.5). Lemma 7.1 confirms that $\hat{\alpha}$ and $\hat{\beta}$ correspond to a minimum point of RSS. □

Note that $S_{xx} = \sum_{i=1}^{n}(x_i-\overline{x})^2 = 0$ if and only if all of the $x_i$s are identical. Consequently, so long as we have at least two distinct predictors then we can use least squares to estimate $\alpha$ and $\beta$. We shall assume that this is the case and explore the statistical properties of the corresponding estimators.

## 7.2 Sampling properties of the estimators

To evaluate the reliability of the least squares estimates we consider the sampling properties of the corresponding **least squares estimators** where we replace the observations $y_1,\ldots,y_n$ by the random variables $Y_1,\ldots,Y_n$. From equations (7.4) and (7.5) these are

$$\hat{\alpha} = \hat{\alpha}(Y_1,\ldots,Y_n) = \sum_{i=1}^{n}\left\{\frac{1}{n} - \frac{(x_i-\overline{x})\overline{x}}{S_{xx}}\right\}Y_i, \tag{7.8}$$

$$\hat{\beta} = \hat{\beta}(Y_1,\ldots,Y_n) = \sum_{i=1}^{n}\left\{\frac{(x_i-\overline{x})}{S_{xx}}\right\}Y_i. \tag{7.9}$$

For the random variables $Y_1,\ldots,Y_n$, an estimator $T(Y_1,\ldots,Y_n)$ is a **linear estimator** if it is of the form

$$T = \sum_{i=1}^{n} t_iY_i$$

for known constants $t_1,\ldots,t_n$. From equations (7.8) and (7.9), we observe that $\hat{\alpha}$ is a linear estimator of $\alpha$ whilst $\hat{\beta}$ is a linear estimator of $\beta$. This linear structure can be exploited to find expectations, variances and covariances of the estimators without having to make any further distributional asumptions than those given in Definition 7.1.

**Theorem 7.2.** *For the simple linear model, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ are, respectively, unbiased estimators of $\alpha$ and $\beta$.*

**Proof:** Note that, from equation (7.9),

$$\begin{aligned} E(\hat{\beta}) &= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})E(Y_i) \\ &= \frac{1}{S_{xx}}\left[\alpha\sum_{i=1}^{n}(x_i-\overline{x})+\beta\sum_{i=1}^{n}x_i(x_i-\overline{x})\right] \;=\; \beta \end{aligned}$$

since $\sum_{i=1}^{n}(x_i-\overline{x})=0$ and $\sum_{i=1}^{n}x_i(x_i-\overline{x})=S_{xx}$. Now, from equation (7.1),

$$\overline{Y} \;=\; \alpha+\beta\overline{x}+\frac{1}{n}\sum_{i=1}^{n}\epsilon_i$$

so that, as $E(\epsilon_i)=0$ for each $i=1,\ldots,n$, $E(\overline{Y})=\alpha+\beta\overline{x}$. Noting that, see equation (7.4), $\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{x}$, it follows that

$$\begin{aligned} E(\hat{\alpha}) &= E(\overline{Y})-E(\hat{\beta})\overline{x} \\ &= \alpha+\beta\overline{x}-\beta\overline{x} \;=\; \alpha \end{aligned}$$

since $E(\hat{\beta})=\beta$. □

**Theorem 7.3.** *For the simple linear model, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ satisfy*

$$\begin{aligned} Var(\hat{\alpha}) &= \left(\frac{\sum_{i=1}^{n}x_i^2}{nS_{xx}}\right)\sigma^2, && (7.10) \\ Var(\hat{\beta}) &= \frac{1}{S_{xx}}\sigma^2, && (7.11) \\ Cov(\hat{\alpha},\hat{\beta}) &= \frac{-\overline{x}}{S_{xx}}\sigma^2. && (7.12) \end{aligned}$$

**Proof:** As the $Y_i$s are independent then, from equation (7.9), equation (7.11) follows as

$$\text{Var}(\hat{\beta}) \;=\; \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})\text{Var}(Y_i) \;=\; \frac{1}{S_{xx}}\sigma^2.$$

Now, from equation (7.38) of Lemma 7.2, $\overline{Y}$ and $\hat{\beta}$ are uncorrelated. Hence,

$$\begin{aligned}\text{Var}(\hat{\alpha}) &= \text{Var}(\overline{Y} - \hat{\beta}\overline{x}) \;=\; Var(\overline{Y}) + \overline{x}^2\text{Var}(\hat{\beta}) \\ &= \frac{\sigma^2}{n} + \frac{\overline{x}^2}{S_{xx}}\sigma^2 \\ &= \frac{S_{xx} + n\overline{x}^2}{nS_{xx}}\sigma^2 \;=\; \left(\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\right)\sigma^2\end{aligned}$$

which is equation (7.10). Finally, once again using equation (7.38) of Lemma 7.2,

$$\begin{aligned}\text{Cov}(\hat{\alpha}, \hat{\beta}) &= \text{Cov}(\overline{Y} - \hat{\beta}\overline{x}, \hat{\beta}) \\ &= -\overline{x}\text{Var}(\hat{\beta}) \;=\; \frac{-\overline{x}}{S_{xx}}\sigma^2\end{aligned}$$

which is equation (7.12). □

It can be shown that, amongst the class of unbiased linear estimators for the respective parameters $\alpha$ and $\beta$, $\hat{\alpha}$ and $\hat{\beta}$ have the smallest variance. They are termed **best linear unbiased estimators (BLUE)**. The result is known as the **Gauss-Markov theorem**.

Note that the variances depend upon the $x_i$s so that, in principle, these could be chosen to minimise the variances. This does, however, depend on the model being correct and, in many data applications, this may not be known and it might be desirable to choose the "best" model amongst a collection of possible models.

### 7.2.1 Variance estimation

In most circumstances $\sigma^2$ itself is an unknown parameter and must also be estimated. Since $\sigma^2$ is the variance of the $\epsilon_i$, it makes sense to estimate it using the variance of the 'estimated' $\epsilon_i$, the model **residuals**. In general these are defined as $\hat{\epsilon}_i = y_i - \hat{y}_i$, where $\hat{y}_i$ are the **fitted values**, the model estimates of $\mu_i = E(Y_i)$. For the current simple model $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, so $\hat{\epsilon}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$.

Note that, from Theorem 7.2, as $\hat{\alpha}$ and $\hat{\beta}$ are, respectively, unbiased estimators of $\alpha$ and $\beta$ then $E(\hat{\epsilon}_i) = 0$ for each $i = 1, \ldots, n$. We now show that an unbiased estimator of $\sigma^2$ can be found from the residual sum of squares.

**Theorem 7.4.** *An unbiased estimator of $\sigma^2$ is*

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2. \tag{7.13}$$

**Proof:** From equation (7.39) of Lemma 7.3, $E(\hat{\epsilon}_i) = 0$ so that $\mathrm{Var}(\hat{\epsilon}_i) = E(\hat{\epsilon}_i^2)$. Thus,

$$E(\hat{\sigma}^2) \;=\; \frac{1}{n-2}\sum_{i=1}^{n} E(\hat{\epsilon}_i^2) \;=\; \frac{1}{n-2}\sum_{i=1}^{n} \mathrm{Var}(\hat{\epsilon}_i). \tag{7.14}$$

Substituting equation (7.40) of Lemma 7.3 into (7.14) gives

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

since $S_{xx} = \sum_{i=1}^{n}(x_i-\overline{x})^2$. Hence, $\hat{\sigma}^2$ is an unbiased estimator of $\sigma^2$. □

Note the divisor of $n-2$ and that we have estimated two parameters $\alpha$ and $\beta$ with $E(Y_i) = \alpha + \beta x_i$. Compare this to the case of $n$ independent and identically distributed observations with mean $\mu$ and variance $\sigma^2$. In that case, an unbiased estimator of $\sigma^2$, $S^2$, uses a divisor of $n-1$ and one parameter is estimator. In MA22015 Statistics 2B you will see that this property expands to linear models with $p$ parameters and that $n-p$ represents the **degrees of freedom**.

We can use Theorem 7.4 to obtain unbiased estimators of $\mathrm{Var}(\hat{\alpha})$ and $\mathrm{Var}(\hat{\beta})$ by plugging in $\hat{\sigma}^2$ for $\sigma^2$ in each case. This can be summarised in the following corollary.

**Corollary 7.1.** *An unbiased estimator of Var($\hat{\alpha}$) is*

$$\hat{\sigma}^2_{\hat{\alpha}} \;=\; \left(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}\right)\hat{\sigma}^2. \tag{7.15}$$

*An unbiased estimator of Var($\hat{\beta}$) is*

$$\hat{\sigma}^2_{\hat{\beta}} \;=\; \frac{1}{S_{xx}}\hat{\sigma}^2. \tag{7.16}$$

**Proof:** Follows immediately by taking expectations of $\hat{\sigma}^2_{\hat{\alpha}}$ and $\hat{\sigma}^2_{\hat{\beta}}$, noting that, from Theorem 7.4, $E(\hat{\sigma}^2) = \sigma^2$ and equating with equations (7.10) and (7.11) respectively. □.

## 7.3 Fitting the model

The least squares calculations derived above are available as part of R. The function `lm` fits linear models to data, including the simple example of heights and weights which are stored in data frame `heightweight`. The following R code fits the model and produces the output shown.

```r
lmod = lm(Weight ~ Height, heightweight)
summary(lmod)
```

```
##
## Call:
## lm(formula = Weight ~ Height, data = heightweight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.369  -6.909  -0.103   7.432  35.037
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -226.3039    14.7335  -15.36   <2e-16 ***
## Height         6.0005     0.2137   28.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.85 on 348 degrees of freedom
## Multiple R-squared:  0.6938, Adjusted R-squared:  0.6929
## F-statistic: 788.5 on 1 and 348 DF,  p-value: < 2.2e-16
```

The call to `lm` passed two arguments to the function. The first is a *model formula*, `Weight ~ Height`, specifying the model to be fitted: the name of the response variable is to the left of '`~`' while the predictor variable is specified on the right. The intercept is included by default. If you don't want an intercept term, you need to include a '`-1`' term in the formula. The second (optional) argument gives the name of the data frame in which the variables are to be found. `lm` takes this information and uses it to fit the model by least squares: the results are returned in a "fitted model object", which in this case has been assigned to an object called `lmod` for later examination.

The `summary` function is then used to examine the fitted model object. We'll explain all the items later but we can identify some now:

- $\hat{\beta} = 6.0005$ (labelled as the estimate in the Height row, reflecting the multiplication of heaight by $\beta$ in the model). For each one inch increase in height, we expect an increase of 6.0005 pounds in weight.

- $\hat{\alpha} = -226.3039$ (labelled as the estimate of the intercept). Literally, if your height was zero inches tall, you expect to weigh -226.309 pounds. Practically, this is nonsense - making predictions outside the range of the data is called an **extrapolation** and can produce silly answers. The minimum height in the dataset is 61.48 inches and the maximum 76.71 inches. Sometimes the intercept is just a numerical convenience and does not have a meaningful interpretation.

- $\hat{\sigma} = 10.85$ (labelled as the residual standard error). Our predictions of weight will be off by around 10.85 pounds (although we need to be more precise about what this means).

- $\hat{\sigma}_{\hat{\alpha}} = 14.7335$ (labelled as the standard error of the intercept) and $\hat{\sigma}_{\hat{\beta}} = 0.2137$ (labelled as the standard error of the height as $\beta$ is the multiplier of the height).

The other terms in the output are derived using additional distributional assumptions about the model. Before moving on to discuss these, it is important to check the current model assumptions. The best way to do this is to examine diagnostic plots generally using the residuals.

### 7.3.1 Residual-fitted plot

The 'fitted values' for this model are $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, while the residuals are $\hat{\epsilon}_i = y_i - \hat{y}_i$. We first consider the relationship between these.

**Theorem 7.5.** *For the simple linear model, $Cov(\hat{\epsilon}_i, \hat{Y}_j) = 0$ for all $i, j$.*

**Proof:** Noting that

$$\mathrm{Cov}(\hat{\epsilon}_i, \hat{Y}_j) \quad = \quad \mathrm{Cov}(\hat{\epsilon}_i, \hat{\alpha}) + x_j \mathrm{Cov}(\hat{\epsilon}_i, \hat{\beta}),$$

the result follows immediately from Lemma 7.4 which shows that the residuals are uncorrelated with the predictors. □

Consequently, if the model is supported by the data, a plot of the residuals against the fitted values should reflect this result. We consider the plot for our running example.

```r
plot(fitted(lmod), residuals(lmod), xlab="fitted values",
ylab="residuals")
abline(h=0)
```

What we would like to see, in such a plot, is an apparently random scatter of residuals around zero, with no trend in either the mean of the residuals, or their variability, as the fitted values increase. A trend in the mean violates the assumption about the linear form of the model, and is usually indicative of something missing in the model structure, while a trend in the variability violates the constant variance assumption. The plot is also useful for detecting outliers - points which have a poor fit to the model.

Note that, see Lemma 7.3, even if the errors $\epsilon_i$ are independent and have constant variance, the $\hat{\epsilon}_i$ are not actually independent and they don't generally have constant variance. As $n$ increases, we observe that these differences decrease.

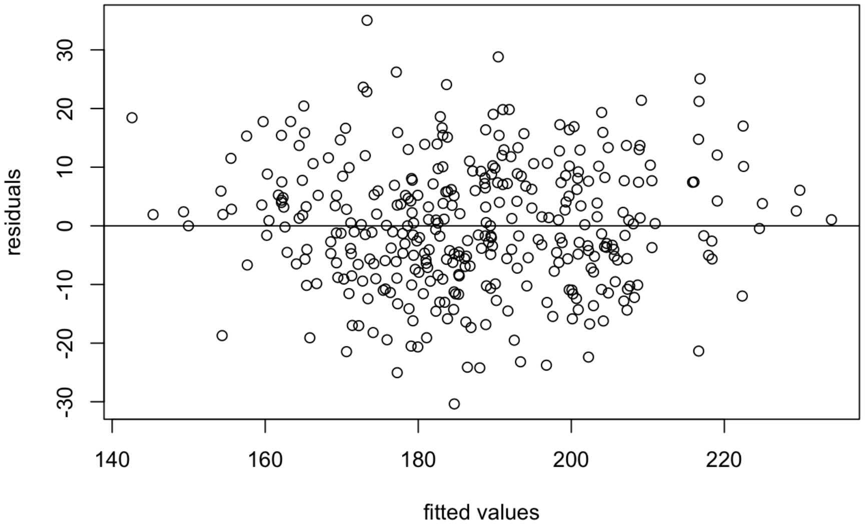


Figure 7.2: Residual-Fitted plot for the heights and weights of 350 individuals.

The plot in Figure 7.2 is fine. No problems are seen. Assessing these plots requires subjective judgement so some experience is valuable. But we are happy with this one and can proceed with conclusion or further modelling.

For example, we could use the model to predict the weight of an individual with given height $x$. The predicted weight would be $y = \hat{\alpha} + \hat{\beta}x$. In doing so, we are implicitly making a further assumption, a **qualitative extrapolation**, that the individual is from a similar population as those in the sample.

### 7.3.2 R-squared

In this section we will develop the idea of the $R^2$ statistic which provides a general measure of how closely a model fits the response data. The approach begins with a partitioning of the variance similar to that utilised in the one-way analysis of variance.

**Theorem 7.6.** *For the simple linear model,*

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 \quad = \quad \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \qquad (7.17)$$

*where* $\hat{y}_i=\hat{\alpha}+\hat{\beta}x_i$.

**Proof:** Noting that

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 \quad = \quad \sum_{i=1}^{n}\{(y_i-\hat{y}_i)+(\hat{y}_i-\overline{y})\}^2,$$

it's sufficient to show that

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y}) \quad = \quad \sum_{i=1}^{n}y_i(\hat{y}_i-\overline{y})-\sum_{i=1}^{n}\hat{y}_i(\hat{y}_i-\overline{y}) \quad = \quad 0. \qquad (7.18)$$

As $\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}$ then $\hat{y}_i-\overline{y}=\hat{\beta}(x_i-\overline{x})$. Thus, additionally using equation (7.5),

$$\sum_{i=1}^{n}y_i(\hat{y}_i-\overline{y}) \quad = \quad \hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})y_i \quad = \quad \hat{\beta}^2S_{xx}. \qquad (7.19)$$

Similarly,

$$\begin{aligned}\sum_{i=1}^{n}\hat{y}_i(\hat{y}_i-\overline{y}) \quad &= \quad \sum_{i=1}^{n}\{\overline{y}+\hat{\beta}(x_i-\overline{x})\}\hat{\beta}(x_i-\overline{x}) \\ &= \quad \overline{y}\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})+\hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2 \quad = \quad \hat{\beta}^2S_{xx}. \qquad (7.20)\end{aligned}$$

Subtracting equation (7.20) from (7.19) confirms equation (7.18) and thus (7.17) holds. □.

Equation (7.17) can be interpreted as

Total sum of squares = Regression sum of squares + Residual sum of squares.

The quantity

$$R^2 \quad = \quad \frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \quad = \quad 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \qquad (7.21)$$

is called the **coefficient of determination** and can be used to quantify how well the fitted model describes the data. From equation (7.17) we have that $0\leq R^2\leq 1$ and $R^2$ measures the proportion of the variance in the original data that is "explained" by the fitted model. The better the fit, the smaller the value of the residual sum of squares should be and so the larger the value of $R^2$. Note that a small value of $R^2$ does not necessarily mean the fitted model is: this depends upon how much random variability is contained in the response data about which nothing can be done.

Notice that, by multiplying equation (7.21) by $n/n$, we may write

$$R^2 \quad = \quad 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2/n}{\sum_{i=1}^n (y_i - \overline{y})^2/n}$$

Consequently, $R^2$ involves a ratio of biased variance estimators and this can lead to $R^2$ overestimating how well a model is doing. One adjustment to this is to consider the ratio of the unbiased variance estimators. This is known as the **adjusted** $R^2$ and, in this case where $\hat{y}_i$ involves two parameters, is given by

$$R^2_{\text{adj}} \quad = \quad 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^n (y_i - \overline{y})^2/(n-1)}.$$

Notice that $R^2_{\text{adj}}$ could now become negative.

**Example 7.1.** From the `R` output for our model fit of the height-weight data, we observe that `Multiple R-squared:  0.6938` and `Adjusted R-squared: 0.6929` Consequently, for this example,

$$\frac{\sum_{i=1}^{350} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{350} (y_i - \overline{y})^2} \quad = \quad 0.3062$$

so that

$$1 - \frac{\sum_{i=1}^{350} (y_i - \hat{y}_i)^2/348}{\sum_{i=1}^{350} (y_i - \overline{y})^2/349} \quad = \quad 1 - \frac{349}{348} 0.3062 \quad = \quad 0.6929.$$

Consequently, the model explains, or determines, about 69% of the total variation in $y_1, \ldots, y_n$.

Noting that $\hat{y}_i - \overline{y} = \hat{\beta}(x_i - \overline{x})$ then $\sum_{i=1}^n (\hat{y}_i - \overline{y})^2 = \hat{\beta}^2 S_{xx}$. Consequently, from equation (7.5), we can express equation (7.21) as

$$R^2 \quad = \quad \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

where $S_{yy} = \sum_{i=1}^n (y_i - \overline{y})^2$. In this formulation, $R^2$ can be aligned to the square of the correlation.

## 7.4 Adding a distributional assumption

So far everything done with the simple model has been based only on the model equations and the two assumptions of independence and equal variance, for the response variable. If we wish to go further, and find confidence intervals for the parameters or test hypotheses related to the model, then a further distributional assumption will be necessary.

The most common assumption, and the one that we shall assume, is that $\epsilon_i \sim N(0, \sigma^2)$ for all $i$, which is equivalent to assuming $Y_i \sim N(\alpha + x_i\beta, \sigma^2)$.

### 7.4.1 Maximum likelihood estimation

The distributional assumption provides the opportunity to calculate the maximum likelihood estimators of $\alpha$ and $\beta$. We will show that the least squares estimators of $\alpha$ and $\beta$ are also the respective MLEs.

We will work under the assumption that $\sigma^2$ is also an unknown parameter. The approach is similar to Example 2.8. The likelihood function is, for $\mathbf{y} = (y_1, \ldots, y_n)$,

$$\begin{aligned} L(\alpha, \beta, \sigma^2 \,|\, \mathbf{y}) &= \prod_{i=1}^{n} f(y_i \,|\, \alpha, \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2 \right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathrm{RSS}(\alpha, \beta) \right\} \end{aligned}$$

where the final equality follows from equation (7.3). The corresponding log likelihood is

$$l(\alpha, \beta, \sigma^2 \,|\, \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathrm{RSS}(\alpha, \beta).$$

Hence, due to the $-\frac{1}{2\sigma^2}$ multiplier, finding the $\alpha$ and $\beta$ values which maximise $l(\alpha, \beta, \sigma^2 \,|\, \mathbf{y})$ correspond to those which minimise $\mathrm{RSS}(\alpha, \beta)$. In analogy to Example 2.8, the corresponding MLE of $\sigma^2$ is

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

which, by comparison with equation (7.13), is biased.

### 7.4.2 Distribution of the estimators

From the properties of normal distributions, it follows that if $Y_1, \ldots, Y_n$ are independent normal random variables then $\sum_{i=1}^{n} t_i Y_i$ is also normally distributed.

As, see equations (7.8) and (7.9), $\hat{\alpha}$ and $\hat{\beta}$ are both linear sums of the $Y_i$s it follows that they are both normally distributed. In Theorems 7.2 and 7.3 we found the mean and variance of $\hat{\alpha}$ and $\hat{\beta}$. Putting these two results together gives the following theorem.

**Theorem 7.7.** *For the simple linear model with each $\epsilon_i \sim N(0, \sigma^2)$,*

$$\hat{\alpha} \sim N\left(\alpha, \left(\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\right)\sigma^2\right), \tag{7.22}$$

$$\hat{\beta} \sim N\left(\beta, \frac{1}{S_{xx}}\sigma^2\right). \tag{7.23}$$

In the unlikely case that $\sigma^2$ is known we can proceed using normal theory. In practice, this will not be the case and, as we will show, we can proceed using the $t$-distribution using $\hat{\sigma}^2$, as given by equation (7.13), as the estimator of $\sigma^2$. The first step is to generate the analogous theorem to Theorem 3.1.

**Theorem 7.8.** *For the simple linear model with each $\epsilon_i \sim N(0, \sigma^2)$, $(\hat{\alpha}, \hat{\beta})$ and $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ are independent and*

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}. \tag{7.24}$$

**Proof:** Independence of $(\hat{\alpha}, \hat{\beta})$ and $\hat{\sigma}^2$ follows as, from equations (7.43) and (7.44), $\mathrm{Cov}(\hat{\alpha}, \hat{\epsilon}_i) = 0$ and $\mathrm{Cov}(\hat{\beta}, \hat{\epsilon}_i) = 0$ and that this is sufficient for independence under the normal model. The derivation of the distribution of $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ follows a similar approach to Theorem 3.1 and is outside the scope of this course. □

Combining Theorems 7.7 and 7.8 we see that, by standardising the normal distributions of $\hat{\alpha}$ and $\hat{\beta}$ and replacing $\sigma$ by $\hat{\sigma}$ we will obtain $t_{n-2}$ distributions as the following corollary describes.

**Corollary 7.2.** *For the simple linear model with each $\epsilon_i \sim N(0, \sigma^2)$,*

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}\sqrt{(\sum_{i=1}^n x_i^2 / nS_{xx})}} = \frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} \sim t_{n-2}, \tag{7.25}$$

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-2}, \tag{7.26}$$

*where $\hat{\sigma}_{\hat{\alpha}}$ is as given in equation (7.15) and $\hat{\sigma}_{\hat{\beta}}$ is as given in equation (7.16).*

We can then use these pivots to construct confidence intervals for $\hat{\alpha}$ and $\hat{\beta}$ and test hypotheses using t-tests.

### 7.4.3 Confidence intervals

From equation (7.26), a $100(1-a)\%$ (we use $a$ here to avoid confusion with the parameter $\alpha$) confidence interval for the slope $\beta$ is given by

$$\hat{\beta} - t_{n-2,1-a/2}\hat{\sigma}_{\hat{\beta}} < \beta < \hat{\beta} + t_{n-2,1-a/2}\hat{\sigma}_{\hat{\beta}}$$

where $P(t_\nu \leq t_{\nu,p}) = p$ when $t_\nu$ is a t-distribution with $\nu$ degrees of freedom.

**Example 7.2.** We'll construct a 95% confidence interval for $\hat{\beta}$ for our height-weight example. From the R output of the model fit we have that $\hat{\beta} = 6.0005$ and $\hat{\sigma}_{\hat{\beta}} = 0.2137$ and noting that $t_{348,0.975} =$ `qt(0.975,348)` $= 1.966804$, the 95% confidence interval for $\beta$ is

$$(6.0005 - 1.966804(0.2137), 6.0005 + 1.966804(0.2137)) \quad = \quad (5.5802, 6.4208).$$

We can also do this all in R by extracting the estimate and its standard error. To do this we create a matrix `weightest` which contains the estimate details from the fit.

```
weightest = summary(lmod)$coef
weightest
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -226.303927 14.7335371 -15.35978 5.184975e-41
## Height         6.000539  0.2136913  28.08041 1.882645e-91
```

$\hat{\beta}$ can then be obtained as `weightest[2,1]` whilst $\hat{\sigma}_{\hat{\beta}}$ is `weightest[2,2]`. We can then construct the 95% confidence interval using

```
weightest[2,1] + qt(c(0.025,0.975), 348) * weightest[2,2]
```

```
## [1] 5.580250 6.420828
```

Recalling the duality between hypothesis testing and confidence interval, we can note that any null hypothesis value outside the interval would be rejected. In particular, we note that zero is not in this interval suggesting that the slope term $\hat{\beta}$ is relevant for the model (which is, in this case, not a surprise.)

A similar approach can be taken to construct a confidence interval for the intercept $\alpha$. However, as we have noted, there is usually more interest in $\beta$ than $\alpha$, especially when a predictor value of $x = 0$ is not a reasonable value as would be the case for our height-weight example.

### 7.4.4 Hypothesis testing

Similarly, we can use equation (7.26) to perform hypothesis tests. For the null hypothesis $H_0 : \beta = \beta_0$, if this is true then

$$t \quad = \quad \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

should be a realisation from a t-distribution with $n-2$ degrees of freedom. This leads to a critical region of the form

$$\begin{aligned} \mathcal{C} &= \{(y_1, \ldots, y_n) : t \leq -t_{n-2,1-a/2}, t \geq t_{n-2,1-a/2}\} \\ &= \{(y_1, \ldots, y_n) : |t| \geq t_{n-2,1-a/2}\} \end{aligned}$$

for a test of significance level $a\%$.

The most common hypotheses test is

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0 \tag{7.27}$$

as this states that a predictor $x$ has makes no difference to the response. Notice that, for an observed value of $t$, the corresponding $p$-value of this data for the test is $2P(t_{n-2} > |t|)$.

**Example 7.3.** We extend Example 7.2 to construct the corresponding hypothesis test for $\beta$ and compute the p-value. The observed test statistic is

```
(tstat = (weightest[2,1]-0)/weightest[2,2])
```

```
## [1] 28.08041
```

```
2*pt(abs(tstat), df=348,lower.tail=FALSE)
```

```
## [1] 1.882645e-91
```

The observed t value is very large, corresponding to the extremely small p-value. We are thus able to reject the null hypothesis that $\beta = 0$. Note that both of these figures are given as part of the summary output for `lm`. In particular, as the `t value` and `Pr(>|t|)` values on the `Height` row of the coefficients. These can be directly obtained:

```
weightest[2,3]   # access the observed t
```

```
## [1] 28.08041
```

```
weightest[2,4]   # access the p-value
```

```
## [1] 1.882645e-91
```

The test $H_0 : \beta = 0$ is similar to the test that all treatments are equal in the one-way analysis of variance model. Note that, from the efinition of the t-distribution, see Definition 3.4, that if $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$, and $Z$ and $U$ are independent, then the distribution of $T = \frac{Z}{\sqrt{U/\nu}}$ is a t-distribution with $\nu$ degrees of fredom. Noting that, from Definition 3.3, $Z^2 \sim \chi^2_1$ if follows immediately from Definition 5.2 that

$$T^2 \;=\; \frac{Z^2/1}{U/\nu} \;\sim\; F_{1,\nu}.$$

Thus, the test of the hypotheses

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_1 : \beta \neq 0$$

where we reject $H_0$ if

$$\left| \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} \right| \;\geq t_{n-2,1-a/2} \tag{7.28}$$

is equivalent to rejecting $H_0$ if

$$\frac{\hat{\beta}^2}{\hat{\sigma}^2_{\hat{\beta}}} \;\geq F_{1,n-2,1-a}. \tag{7.29}$$

Noting that $\hat{y}_i - \overline{y} = \hat{\beta}(x_i - \overline{x})$ then $\sum_{i=1}^n (\hat{y}_i - \overline{y})^2 = \hat{\beta}^2 S_{xx}$ and $\hat{\sigma}^2_{\hat{\beta}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2)S_{xx}$ it follows that

$$\frac{\hat{\beta}^2}{\hat{\sigma}^2_{\hat{\beta}}} \;=\; \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2)}$$

which is the "regression sum of squares" divided by the "(residual sum of squares)/(degrees of freedom)" linking once again to the one-way analysis of variance model.

**Example 7.4.** We continue Example 7.3 by noting that the final line in the summary from the `lm` application is `F-statistic: 788.5 on 1 and 348 DF, p-value: < 2.2e-16`. We confirm the relationship between equations (7.28) and (7.29) by noting that $788.5 = 28.08^2$ and that the corresponding p-value,

```
format(pf(weightest[2,3]^2,1,348,lower.tail=F), scientific=TRUE)
```

```
## [1] "1.882645e-91"
```

agrees with that for the t-test of $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

## 7.5 Meaning of regression

For a predictor $x$ we have a predicted response of $y = \hat{\alpha} + \hat{\beta}x$. As $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$ it follows that

$$y - \overline{y} \;=\; \hat{\beta}(x - \overline{x}). \tag{7.30}$$

As, see equation (7.5), $\hat{\beta} = S_{xy}/S_{xx}$ then substituting into equation (7.30) and dividing both sides by $\sqrt{S_{yy}}$ gives

$$\frac{(y - \overline{y})}{\sqrt{S_{yy}}} = \left( \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right) \frac{(x - \overline{x})}{\sqrt{S_{xx}}}. \tag{7.31}$$

Equation (7.31) can thus be summarised as

$$\frac{y - \bar{y}}{SD_y} = r \frac{(x - \bar{x})}{SD_x} \tag{7.32}$$

where $r$ is the correlation between $x$ and $y$: the response in standard units is the correlation times the predictor in standard units.

**Example 7.5.** In the height-weight example, one might naively expect that an individual whose weight is one standard deviation above the average weight might have a height which is one standard deviation above the average height, give or take. This would correspond to setting $r = 1$ in equation (7.32). This would give a slope equal to $\beta_1 = SD_y/SD_x$ and an intercept of $\alpha_1 = \overline{y} - \beta_1 \overline{x}$. For our observed values we find that

```
(beta1 = with(heightweight, sd(Weight)/sd(Height)))
```

```
## [1] 7.203994
```

```
(alpha1 = with(heightweight, mean(Weight)- beta1*mean(Height)))
```

```
## [1] -309.2152
```

We redisplay the data with the added least squares line and the line corresponding to $r = 1$:

```
plot(Weight ~ Height, heightweight)
abline(weightest[1,1], weightest[2,1]) # least squares line
abline(alpha1, beta1,lty=2) # line with r = 1
abline(h=with(heightweight, mean(Weight)),lty=4)
abline(v=with(heightweight, mean(Height)),lty=4)
```

The lines cross at the point of the averages. We can see that an individual of beyond the mean height is predicted by the least squares line to have a weight which is above average but not quite as heavy as the height of the dashed line would have you believe. Similarly individuals of below average height are predicted to have a weight which is still below average but not quite as light as the dashed line. This phenomenom is known as the regresion effect or **regression toward the mean**.

This effect applies to any $(x, y)$ situation like this. For example, in sports, an athlete may have a spectacular first season only to do not quite as well in the second season. Sports writers come up with all kinds of explanations for this

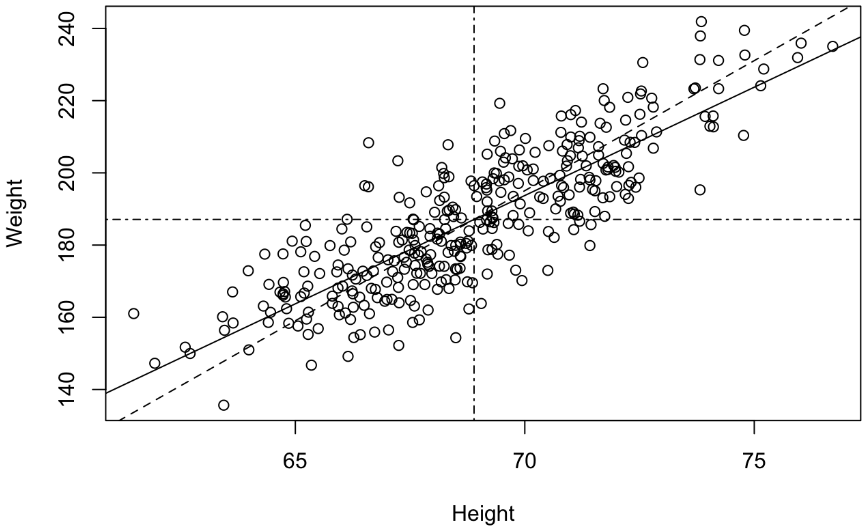


Figure 7.3: Least squares fit shown with a solid line and line with r=1 shown with a dashed line. The lines intersect at the point of the averages shown by the dashed-dot lines.

but the regression effect is likely to be the unexciting cause. In the USA, this has been called the *sophomore slump* and in the UK, *second year blues*.

Regression methodology developed rapidly with the advent of high-speed computing. Just fitting a regression model used to require extensive hand calculation. As computing hardware has improved, the scope for analysis has widened. This has led to an extensive development in the methodology and the scale of problems that can be tackled. You will see this development in future statistics units. In particular, in MA22015 Statistics 2B you will extend the simple linear model, with a single predictor, to a model with multiple predictors. Although the complexity increases, you will see that the underpinning ideas developed in the simple model, with one predictor, naturally extend to the case when we have $p$ predictors.

## 7.6 Non-examinable lemmas

In this section we state and prove four lemmas used to derive results in this chapter. These lemmas are non-examinable - the exam will not ask you to either to state or prove them.

The first lemma expresses the residual sum of squares of any $\alpha$ and $\beta$ in terms of the residual sum of squares for the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ and is used in Theorem 7.1 to demonstrate that the residual sum of squares is minimised at the least squares estimates.

**Lemma 7.1.** *For the simple linear model*

$$RSS(\alpha,\beta) \quad = \quad RSS(\hat{\alpha},\hat{\beta}) + \sum_{i=1}^{n}\{(\alpha-\hat{\alpha})+(\beta-\hat{\beta})x_i\}^2. \quad (7.33)$$

*Consequently, the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ correspond to a minimum point of RSS.*

**Proof:** Let $\hat{\alpha}$ and $\hat{\beta}$ be given as in Theorem 7.1. Then,

$$\begin{aligned}
\text{RSS}(\alpha,\beta) &= \sum_{i=1}^{n}\left[(y_i-\hat{\alpha}-\hat{\beta}x_i)-\{(\alpha-\hat{\alpha})+(\beta-\hat{\beta})x_i\}\right]^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2+\sum_{i=1}^{n}\{(\alpha-\hat{\alpha})+(\beta-\hat{\beta})x_i\}^2 \\
&\quad -2\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)\{(\alpha-\hat{\alpha})+(\beta-\hat{\beta})x_i\} \\
&= \text{RSS}(\hat{\alpha},\hat{\beta})+\sum_{i=1}^{n}\{(\alpha-\hat{\alpha})+(\beta-\hat{\beta})x_i\}^2 \\
&\quad -2(\alpha-\hat{\alpha})\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i) \\
&\quad -2(\beta-\hat{\beta})\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i). \qquad (7.34)
\end{aligned}$$

Now,

$$\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i) = n(\overline{y}-\hat{\alpha}-\hat{\beta}\overline{x}) = 0 \qquad (7.35)$$

as $\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}$. Noting that $\hat{\alpha}-\hat{\beta}x_i=\overline{y}-\hat{\beta}(x_i-\overline{x})$ we have

$$\begin{aligned}
\sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i) &= \sum_{i=1}^{n}x_i\{(y_i-\overline{y})-\hat{\beta}(x_i-\overline{x})\} \\
&= \sum_{i=1}^{n}x_iy_i-n\overline{xy}-\hat{\beta}\sum_{i=1}^{n}x_i(x_i-\overline{x}) \\
&= \sum_{i=1}^{n}x_iy_i-n\overline{xy}-\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})^2 = 0 \qquad (7.36)
\end{aligned}$$

by noting that $\overline{x}\sum_{i=1}^{n}(x_i-\overline{x})=0$ and using equation (7.5). Substituting equations (7.35) and (7.36) into equation (7.34) gives equation (7.33). It immediately follows from equation (7.33) that $\text{RSS}(\alpha,\beta)\geq\text{RSS}(\hat{\alpha},\hat{\beta})$. □

The following lemma shows that $\overline{Y}$ and $\hat{\beta}$ are uncorrelated and is utilised in the proof to Theorem 7.3.

**Lemma 7.2.** *For the simple linear model*

$$\begin{aligned}
Cov(Y_i,\hat{\beta}) &= \frac{x_i-\overline{x}}{S_{xx}}\sigma^2; \qquad (7.37) \\
Cov(\overline{Y},\hat{\beta}) &= 0. \qquad (7.38)
\end{aligned}$$

**Proof:** From Definition 7.1, the $Y_i$s are mutually independent with variance $\sigma^2$. Thus, from (7.9), as $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$,

$$\text{Cov}(Y_i, \hat{\beta}) \;=\; \frac{x_i - \overline{x}}{S_{xx}}\text{Var}(Y_i) \;=\; \frac{x_i - \overline{x}}{S_{xx}}\sigma^2$$

which is equation (7.37). Similarly, equation (7.38) follows as

$$\text{Cov}(\overline{Y}, \hat{\beta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\text{Cov}(Y_i, \hat{\beta}) \;=\; 0$$

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

The following lemma derives the mean, variance and covariance structure of the residuals $\hat{\epsilon}_i$ which can be used, in Theorem 7.4, to derive an unbiased estimator of $\sigma^2$ based on the residual sum of squares.

**Lemma 7.3.** *For each $i = 1, \ldots, n$, the residuals $\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i$ satisfy*

$$E(\hat{\epsilon}_i) \;=\; 0; \tag{7.39}$$

$$Var(\hat{\epsilon}_i) \;=\; \left(\frac{n-1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\sigma^2. \tag{7.40}$$

$$Cov(\hat{\epsilon}_i, \hat{\epsilon}_j) \;=\; \left(\frac{n-1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}}\right)\sigma^2. \tag{7.41}$$

**Proof:** As $\hat{\alpha}$ and $\hat{\beta}$ are, respectively, unbiased estimators of $\alpha$ and $\beta$ then

$$E(\hat{\epsilon}_i) \;=\; E(Y_i - \hat{\alpha} - \hat{\beta}x_i) \;=\; E(Y_i) - \alpha - \beta x_i \;=\; 0$$

where the final equality follows from equation (7.2). We thus have equation (7.39). Noting that $\hat{\alpha} + \hat{\beta}x_i = \overline{Y} + \hat{\beta}(x_i - \overline{x})$ then

$$\hat{\epsilon}_i \;=\; Y_i - \hat{\alpha} - \hat{\beta}x_i \;=\; (Y_i - \overline{Y}) - \hat{\beta}(x_i - \overline{x})$$

so that

$$\begin{aligned}\text{Var}(\hat{\epsilon}_i) \;=\;& \text{Var}(Y_i - \overline{Y}) + (x_i - \overline{x})^2\text{Var}(\hat{\beta}) \\ & -2(x_i - \overline{x})\text{Cov}(Y_i - \overline{Y}, \hat{\beta})\end{aligned} \tag{7.42}$$

Now,

$$\begin{aligned}\text{Var}(Y_i - \overline{Y}) \;&=\; \text{Var}(Y_i) + \text{Var}(\overline{Y}) - 2\text{Cov}(Y_i, \overline{Y}) \\ &=\; \sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} \;=\; \frac{n-1}{n}\sigma^2; \\ \text{Cov}(Y_i - \overline{Y}, \hat{\beta}) \;&=\; \text{Cov}(Y_i, \hat{\beta}) - \text{Cov}(\overline{Y}, \hat{\beta}) \;=\; \frac{x_i - \overline{x}}{S_{xx}}\sigma^2,\end{aligned}$$

where the final equality follows from Lemma 7.2. Substituting these and equation (7.11) into equation (7.42) gives

$$\begin{aligned}\mathrm{Var}(\hat{\epsilon}_i) &= \frac{n-1}{n}\sigma^2 + (x_i - \overline{x})^2 \frac{\sigma^2}{S_{xx}} - 2(x_i - \overline{x})\frac{x_i - \overline{x}}{S_{xx}}\sigma^2 \\ &= \left(\frac{n-1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\sigma^2\end{aligned}$$

which is equation (7.40). Equation (7.41) follows in a similar way by expanding $\mathrm{Cov}(\hat{\epsilon}_i, \hat{\epsilon}_j)$ analogously to equation (7.42). □

Theorem 7.5 shows that the fitted values $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$ are uncorrelated with the residuals are $\hat{\epsilon}_i = Y_i - \hat{Y}_i$. This result follows immediately from the following lemma.

**Lemma 7.4.** *For the simple linear model, the residuals are uncorrelated with the predictors. That is, for all $i$,*

$$\begin{aligned} Cov(\hat{\epsilon}_i, \hat{\alpha}) &= 0, & (7.43) \\ Cov(\hat{\epsilon}_i, \hat{\beta}) &= 0. & (7.44)\end{aligned}$$

**Proof:** Noting that $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \overline{Y} - \hat{\beta}(x_i - \overline{x})$ then

$$\mathrm{Cov}(\hat{\epsilon}_i, \hat{\beta}) = \mathrm{Cov}(Y_i, \hat{\beta}) - \mathrm{Cov}(\overline{Y}, \hat{\beta}) - (x_i - \overline{x})\mathrm{Var}(\hat{\beta}) = 0$$

from equations (7.37), (7.38), and (7.11). Using $\hat{\alpha} = \overline{Y} - \hat{\beta}(x_i - \overline{x})$, then

$$\begin{aligned}\mathrm{Cov}(\hat{\epsilon}_i, \hat{\alpha}) &= \mathrm{Cov}(\hat{\epsilon}_i, \overline{Y}) - (x_i - \overline{x})\mathrm{Cov}(\hat{\epsilon}_i, \hat{\beta}) \\ &= \mathrm{Cov}(Y_i, \overline{Y}) - \mathrm{Var}(\overline{Y}) - (x_i - \overline{x})\mathrm{Cov}(\hat{\beta}, \overline{Y}) = 0\end{aligned}$$

using equation (7.38) and that $\mathrm{Cov}(Y_i, \overline{Y}) = \mathrm{Var}(\overline{Y})$. □

# Chapter 8

# Non-examinable proofs

Proofs (non-examinable) for some of the results in the notes are provided below.

## 8.1 Proof of Theorem 3.1

Reminder: Theorem 3.1 states that for i.i.d. normal samples, $\overline{X}$ and $S^2$ are independent and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

**Proof:** We first show that $\overline{X}$ and $S^2$ are independent random variables. We first express $S^2$ are a sum of $n-1$ deviations from the mean:

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \\ &= \frac{1}{n-1}\left[(X_1-\overline{X})^2+\sum_{i=2}^{n}(X_i-\overline{X})^2\right] \qquad (8.1) \end{aligned}$$

Since $\sum_{i=1}^{n}(X_i-\overline{X})=0$, we have

$$X_1-\overline{X} = -\sum_{i=2}^{n}(X_i-\overline{X})$$

and so

$$\begin{aligned} (X_1-\overline{X})^2 &= \left[-\sum_{i=2}^{n}(X_i-\overline{X})\right]^2 \\ &= \left[\sum_{i=2}^{n}(X_i-\overline{X})\right]^2 \end{aligned}$$

Substituting this into Equation (8.1) we have

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

Thus $S^2$ can be expressed as a function of $(X_2 - \overline{X}, .., X_n - \overline{X})$. We will show that this $n-1$ multivariate random vector is independent of $\overline{X}$, and hence that $S^2$ is independent of $\overline{X}$. For simplicity we will assume that $\mu = 0$ and $\sigma^2 = 1$ – it being (hopefully) clear that these choices will not affect the dependence (if any) between $\overline{X}$ and $S^2$. The joint density of $X_1, .., X_n$ in this case is

$$f(x_1, .., x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right]$$

We now define a transformation of the vector $(\overline{X}, X_2 - \overline{X}, .., X_n - \overline{X})$ to $Y_1, .., Y_n$ via:

$$\begin{aligned} Y_1 &= \overline{X} \\ Y_2 &= X_2 - \overline{X} \\ &\vdots \\ Y_n &= X_n - \overline{X} \end{aligned}$$

The theorem that gives the density of a random variable defined as a transformation of another extends to the vector case, which can be found in (for example) Casella and Berger (page 185). Application of this result implies that the joint density function of $Y_1, .., Y_n$ is given by

$$\begin{aligned} f(y_1, .., y_n) &= \frac{n}{(2\pi)^{n/2}} e^{-\frac{1}{2}(y_1 - \sum_{i=2}^{n} y_i)^2} e^{\frac{1}{2}\sum_{i=2}^{n}(y_i + y_1)^2} \\ &= \left[\left(\frac{n}{2\pi}\right)^{1/2} e^{(-ny_1^2)/2}\right]\left[\frac{n^{1/2}}{(2\pi)^{(n-1)/2}} e^{-(1/2)[\sum_{i=2}^{n} y_i^2 + (\sum_{i=2}^{n} y_i)^2]}\right] \end{aligned}$$

This shows that the joint density of $Y_1, .., Y_n$ can be factorized into the product of the density of $Y_1$ and the density of $(Y_2, .., Y_n)$, from which it follows that $Y_1$ is independent of $(Y_2, .., Y_n)$. Recalling the definition of the random variables $Y_1, .., Y_n$, this means that $\overline{X}$ is independent of the $n-1$ deviations which determine $S^2$, and hence $\overline{X}$ and $S^2$ are independent. We now show that

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

where the earlier simplifying assumption that $\mu = 0$ and $\sigma^2 = 1$ is now not used. We use an induction argument. Let $\overline{X}_n$ and $S_n^2$ denote the corresponding statistics based on the ‘first’ $n$ observations. Note the ordering is arbitrary and simply needed to make the argument. After some tedious algebra, one can show that

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}(X_n - \overline{X}_{n-1})^2 \tag{8.2}$$

Consider $n = 2$. Defining $0 \times S_1^2 = 0$, we have

$$\begin{aligned} \frac{S_2^2}{\sigma^2} &= \frac{1}{2\sigma^2}(X_2 - X_1)^2 \\ &= \left[\frac{1}{\sqrt{2}\sigma}(X_2 - X_1)\right]^2 \end{aligned}$$

where we use that $\overline{X}_1 = X_1$. Due to independence of the $X_i$, the difference $X_2 - X_1 \sim N(0, 2\sigma^2)$. Therefore

$$\frac{1}{\sqrt{2}\sigma}(X_2 - X_1) \sim N(0,1)$$

By the definition of the $\chi^2$ distribution, we then have that

$$\frac{S_2^2}{\sigma^2} \sim \chi_1^2$$

Next, suppose that for $n = k$, $\frac{(k-1)S_k^2}{\sigma^2} \sim \chi_{k-1}^2$. For $n = k+1$ we have using Equation (8.2) that

$$kS_{k+1}^2 = (k-1)S_k^2 + \frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2$$

and dividing through by $\sigma^2$ that

$$\begin{aligned} \frac{kS_{k+1}^2}{\sigma^2} &= \frac{(k-1)S_k^2}{\sigma^2} + \frac{1}{\sigma^2}\frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2 \\ &= \frac{(k-1)S_k^2}{\sigma^2} + \left[\frac{1}{\sigma}\sqrt{\frac{k}{k+1}}(X_{k+1} - \overline{X}_k)\right]^2 \end{aligned} \tag{8.3}$$

The first term on the right hand side is distributed $\chi_{k-1}^2$ by the induction assumption. Thus to complete the proof we need to show that the second term on the right hand side is the square of an independent standard normal. The difference $X_{k+1} - \overline{X}_k$ is normal since it is the difference of two normals, and it has mean zero since each term has mean zero. Since $X_{k+1}$ is independent of $\overline{X}_k$ its variance is

$$\begin{aligned} \mathrm{Var}(X_{k+1} - \overline{X}_k) &= \mathrm{Var}(X_{k+1}) + \mathrm{Var}(\overline{X}_k) \\ &= \sigma^2 + \frac{\sigma^2}{k} = \sigma^2(1 + k^{-1}) \end{aligned}$$

and thus

$$\text{Var}\left[\frac{1}{\sigma}\sqrt{\frac{k}{k+1}}(X_{k+1}-\overline{X}_k)\right] \quad = \quad \frac{k}{\sigma^2(k+1)}\sigma^2(1+k^{-1}) = 1$$

The second term on the right hand side of Equation (8.3) is thus indeed distributed $\chi^2_1$, as it is the square of a standard normal. All that remains is to show that it is independent of the first term on the right hand side of equation (8.3). First, from the first part of the proof we have that $S_k^2$ is independent of $\overline{X}_k$. Then since $X_{k+1}$ is independent of both of these, it follows that $S_k^2$ is independent of the difference $X_{k+1} - \overline{X}_k$. The proof is thus complete. □

## 8.2 Proof of Propositions 3.1 and 3.2

Throughout this section, assume we have Let $X_1, .., X_n$ be i.i.d. from a distribution with finite mean $\mu$ and finite variance $\sigma^2$.

Proposition 3.1 states that, for a consistent estimator $\hat{\sigma}$ of $\sigma$

$$\frac{(\overline{X}_n - \mu)}{\hat{\sigma}/\sqrt{n}} \xrightarrow{L} N(0,1).$$

**Proof:** First we express the statistic as

$$\frac{\overline{X}_n - \mu}{\hat{\sigma}/\sqrt{n}} \quad = \quad \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{\hat{\sigma}}$$

The first term in the product converges in law to $N(0,1)$ by the Central Limit Theorem (Theorem 3.3). By assumption $\hat{\sigma} \xrightarrow{P} \sigma$. Since the function $g(x) = \sigma/x$ is continuous, by the Continuous Mapping Theorem (Theorem 2.4),

$$\frac{\sigma}{\hat{\sigma}} \xrightarrow{P} \sigma/\sigma = 1$$

Finally, by Slutsky's Theorem (Theorem 3.4), we have that

$$\begin{aligned} \frac{\overline{X}_n - \mu}{\hat{\sigma}/\sqrt{n}} \quad &= \quad \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{\hat{\sigma}} \\ &\xrightarrow{L} \quad Z \times 1 \end{aligned}$$

where $Z \sim N(0,1)$. Hence

$$\frac{(\overline{X}_n - \mu)}{\hat{\sigma}/\sqrt{n}} \xrightarrow{L} N(0,1). \quad \square$$

Proposition 3.2 states that if $S$ is a consistent estimator, then

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

is asymptotically an $100 \times (1-\alpha)\%$ confidence interval for $\mu$.

**Proof:** It thus follows from Proposition 3.1 and the consistency of $S$, that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \xrightarrow{L} N(0,1)$$

As such, $\frac{\sqrt{n}(\overline{X}_n - \mu)}{S}$ is (approximately) a pivot for $\mu$, and

$$P\left(\overline{X}_n - z_{1-\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

as $n \to \infty$. □

## 8.3 Proof of Equations 5.3 and 5.4

Throughout this section we assume we have two independent samples: $X_1, \ldots, X_n$ are independent and identically distributed draws from some $f_X(x \mid \theta_x)$ with $E(X)$, $\text{Var}(X) < \infty$, and $Y_1, \ldots, Y_m$ are independent and identically distributed draws from some $f_Y(y \mid \theta_y)$ with $E(Y)$, $\text{Var}(Y) < \infty$. We will additionally assume that as $n \to \infty$ and $m \to \infty$, $\frac{m}{n+m} \to \rho$ for some $0 < \rho < 1$.

First we prove a lemma.

**Lemma 8.1.** *Suppose* $Var(X) = \sigma_X^2$ *and* $Var(Y) = \sigma_Y^2$ *are known. Then*

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \quad \xrightarrow{L} \quad N(0,1).$$

**Proof:** Let $N = n + m$, so that $m/N \to \rho$ and $n/N \to 1-\rho$ as $m$ and $n$ tend to infinity. Then we can express the asymptotic behavior of $\overline{X}_n$ and $\overline{Y}_m$ in terms of the overall sample size $N = n + m$ by

$$\sqrt{N}(\overline{X}_n - \mu_X) \quad = \quad \sqrt{n}(\overline{X}_n - \mu_X) \times \sqrt{\frac{N}{n}} \xrightarrow{L} A \times (1-\rho)^{-1/2}$$

where $A \sim N(0, \sigma_X^2)$. Then $\text{Var}(A(1-\rho)^{-1/2}) = \sigma_X^2(1-\rho)^{-1}$ and thus we have

$$\sqrt{N}(\overline{X}_n - \mu_X) \xrightarrow{L} N(0, \sigma_X^2/(1-\rho))$$

and similarly

$$\sqrt{N}(\overline{Y}_m - \mu_Y) \xrightarrow{L} N(0, \sigma_Y^2/\rho)$$

We now want to examine the asymptotic distribution of the difference of these two quantities. A result from probability theory (Lemma 3.1.1. of Lehmann) says that if $U_N$ and $V_N$ are independent sequences of random variables, and $U$ and $V$ are independent random variables for which $U_N \xrightarrow{L} U$ and $V_N \xrightarrow{L} V$, then $U_N \pm V_N \xrightarrow{L} U \pm V$. From this result it follows that

$$\sqrt{N}\left[(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)\right] \quad \xrightarrow{L} \quad N\left(0, \frac{\sigma_X^2}{1-\rho} + \frac{\sigma_Y^2}{\rho}\right)$$

or equivalently that

$$\frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \quad \xrightarrow{L} \quad N\left(0, 1\right). \square$$

Now we formally restate the claim in Equation (5.3)

**Proposition 8.1.** *Suppose that Var$(X) =$ Var$(Y) = \sigma^2$ and let $\widehat{\sigma}_X^2$ and $\widehat{\sigma}_Y^2$ be consistent estimators of these variances*
*Then*

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} \xrightarrow{L} N(0,1)$$

*where*

$$S_p^2 = \frac{(n-1)\widehat{\sigma}_X^2 + (m-1)\widehat{\sigma}_Y^2}{n+m-2}.$$

**Proof:** From Lemma 8.1 we have that

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \xrightarrow{L} \quad N\left(0, 1\right)$$

Now we can express

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}} = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}\frac{\sigma}{S_p}$$

Thus by Slutsky's Theorem (Theorem 3.4), this will converge in distribution to $N(0,1)$ provided $\sigma/S_p$ converges in probability to one. To show this, we have by

Theorem 2.5 that

$$S_p^2 = \frac{(n-1)\widehat{\sigma}_X^2 + (m-1)\widehat{\sigma}_X^2}{n+m-2} \xrightarrow{P} (1-\rho)\sigma^2 + \rho\sigma^2 = \sigma^2.$$

Thus $S_p$ is consistent for $\sigma$ and the by the Continuous Mapping Theorem (Theorem 2.4) $\sigma/S_p$ converges in probability to one. □

Now we formally restate the claim in Equation (5.4)

**Proposition 8.2.** *Suppose that* $Var(X) = \sigma_X^2 \neq \sigma_Y^2 = Var(Y) = \sigma^2$ *and let* $\widehat{\sigma}_X^2$ *be consistent estimator of* $Var(X)$ *and* $\widehat{\sigma}_Y^2$ *be consistent estimator of* $Var(Y)$.

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\widehat{\sigma}_X^2}{n} + \frac{\widehat{\sigma}_Y^2}{m}}} \xrightarrow{L} N(0,1)$$

**Proof:** We rewrite the quantity whose asymptotic distribution we are considering as

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \frac{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}{\sqrt{\frac{\widehat{\sigma}_X^2}{n} + \frac{\widehat{\sigma}_Y^2}{m}}}$$

By Lemma 8.1, the first term converges in distribution to $N(0,1)$. The second term in this expression can be written as

$$\sqrt{\frac{m\sigma_X^2 + n\sigma_Y^2}{m\widehat{\sigma}_X^2 + n\widehat{\sigma}_Y^2}} = \sqrt{\frac{(m/N)\sigma_X^2 + (n/N)\sigma_Y^2}{(m/N)\widehat{\sigma}_X^2 + (n/N)\widehat{\sigma}_Y^2}}$$

where $N = m + n$. If $m/N \to \rho$ as $n$ and $m$ go to infinity, then the numerator converges to $\rho\sigma_X^2 + (1-\rho)\sigma_Y^2$, and the denominator converges in probability to the same quantity by Theorem 2.5. Through another application of Theorem 2.5, the term overall converges in probability to one, and then by Slutsky's Theorem (Theorem 3.4), we have that

$$\frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\widehat{\sigma}_X^2}{n} + \frac{\widehat{\sigma}_Y^2}{m}}} \xrightarrow{L} N(0,1). \quad \square$$

## 8.4 Proof of Proposition 5.2

Reminder: Proposition 5.2 states that for comparing the means of two independent normal samples with unequal variances,

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \xrightarrow{L} N(0,1).$$

**Proof:** This is a special case of Proposition 8.2 with $\widehat{\sigma}^2_X = S^2_X$ and $\widehat{\sigma}^2_Y = S^2_Y$. We need to establish that $S^2_X$ and $S^2_Y$ are consistent estimators of $\sigma^2_X$ and $\sigma^2_Y$. We showed this is true for the normal distribution case in Example 2.17. □