Contents

Welcome to Statistics for Health Data Science

These notes provide the core material for the MSc module, Statistics for Health Data Science.

This is a compulsory module for the programme MSc Health Data Science. The module provides an introduction to the key statistical concepts and methods for health data science. Topics covered include probability, initial data description and exploration, frequentist and Bayesian approaches to statistical inference and regression modelling. These topics provide the framework needed for subsequent modules. The module places a focus on learning through practical examples and incorporates directed learning, lectures, group discussion, and computer practical exercises.

1.1 Overall aim of the module

The overall module aims are to introduce:

- the motivation and critical thinking towards solving a question in health science through interrogation of data and drawing conclusions from evidence;
- the principles of probability, regression modelling and statistical inference within frequentist and Bayesian frameworks.

1.2 Module Intended Learning Outcomes

Intended learning outcomes

Upon successful completion of the module you will be able to:

- evaluate the application of different probability distributions to model health data (including Poisson, Binomial and Normal);
- critically analyse frameworks for frequentist and Bayesian inference and evaluate their strengths, limitations and differences;
- examine the concepts of sampling variability, estimators, bias, confidence intervals and credible intervals;
- examine the theoretical basis of linear regression and generalized linear models;
- assess the application of regression modelling to address specific health data science questions;
- critically evaluate strengths and limitations of different statistical methods, including regression models, within a health data science project;
- draw conclusions from the results of a data analysis and justify those conclusions, appropriately acknowledging uncertainty in the results.

1.3 Module Content

The module is split into 16 taught sessions, each building statistical knowledge for health data science. The sessions are:

1. Introduction
2. Probability and Discrete Probability Distributions
3. Continuous Probability Distribution
4. Populations and Sampling
5. Likelihood
6. Maximum Likelihood Estimation
7. Frequentist Inference I
8. Frequentist Inference II
9. Bayesian Statistics I
10. Bayesian Statistics II
11. Types of Investigation
12. Linear Regresion I
13. Linear Regresion II
14. Linear Regresion III
15. Logistic Regression

16. GLMs and Poisson Regression

A final short section (17) connects the regression models to the session regarding types of investigation. This is optional reading and does not have an accompanying taught session.

Acknowledgements

Many people have contributed to this document over time, including a large number of previous and current members of the Department of Medical Statistics at the London School of Hygiene and Tropical Medicine. In particular, we would like to acknowledge contributions from Corentin Segalas, Elizabeth Williamson, Emily Granger, Emily Nightingale, Kathleen O'Reilly, Linda Sharples, Melanie Smuk, Mia Tackney, Nicholas Jewell and Ruth Keogh.

We thank Jennifer Nicholas, whose notes were particularly useful in the development of the linear regression sessions. We thank Katy Morgan for notes which helped inform the development of the section about inference for maximum likelihood models.

The notes for the Bayesian Inference sessions are heavily based on the Foundations course material created by Alex Lewin and Alexina Mason, which was previously developed by James Carpenter, Marcel Zwahlen and Beat Neuenschwander. Some sections are inspired also by notes from Michail Papathomas. We are grateful for their work and permission to re-use.

Version

This document was last updated: September 2021

Inevitably there will be some typos in these notes. Please do let us know any you spot (at: mailto:mscHDS@lshtm.ac.uk) and we will correct them.

How to use this book

This book contains the core content for the module. It is designed to be read in conjunction with the practical sessions and accompanying videos.

Each numbered session has an accompanying lecture and practical session. Each section of the book additionally has a short summary of the contents of that section to help you build an overview of the material. The final sections of the book do not have accompanying taught sessions but have some brief comments to help you pull all the material together.

R code

Analyses are illustrated using R code. Each page which contains R code should be self-contained. You should be able to copy and paste each chunk of R code (using the little icon to the top-right of the code cell) into R and run all the code in that page.

A number of packages will be required. These are loaded using the `library` function, at the appropriate points. If you are using your local version of R you may need to install these packages. Below, you will find code to install all packages used in these notes.

```
# Graphics package
install.packages("ggplot2")

# Will be used for fitting some generalised models:
install.packages("VGAM")

# Contains some useful goodness-of-fit diagnostics:
install.packages("pscl")

install.packages("sandwich")
```

1 Introduction

Statistics for Health Data Science is the scientific approach behind investigating health. The organisers of this module have been specific in the wording. Of course, we will learn techniques to interogate data using statistics! But also, the focus on the approach is to think about a problem scientifically, for example to consider a research question, or a hypothesis. The scientific approach is important, and is described in more detail in this Introduction and the associated lecture. The scientific inquiry is applied to health data, which can take a number of forms, including 'found data'. We think this is what makes data science for health unique; found data presents great advantages as there may be a lot of found data avaialble, but also challenges as the origin of the data and the potential for biases in the data can make analysis more challenging.

Intended learning outcomes

- consider the concept of Statistics for Health Data Science and the bigger picture of scientific inquiry
- understand the data by identifying broad issues of structure, type, provenance and design
- describe variable types
- understand the concept of selection bias
- think about data summaries using exploratory data analysis and visualizations (simple examples)
- consider what to measure and why in a scientific study
- have a basic understanding of the difference between frequentist (Fisher) and Bayesian inference

1.1 What you will learn

By engaging with module you will acquire skills that a data scientist will need to interogate data to answer a health related question. Much of the focus is on the statistical tools that are most often used, as shown above in the *Intended Learning Objectives*. The module is designed towards using statistics within a problem solving cycle (more in Section 1.3).

Consider this book to be a *practical guide* in using statistics. Every session provides some statistical theory and examples so you can see the theory in action. Especially in the earlier sessions, some of the examples can be done without using a computer. As the module progresses many of the calculations are carried out using R, and we will increasingly apply the concepts to real data. We provide the code for each example.

As you work through the sessions, your ability to use statistics in health data science should improve. We will begin with relatively simple questions that we want to answer. As the module progresses, the questions will be more relevant to a health, and the steps involved will require more statistical inference and scientific inquiry.

Module layout

The module is roughly divided into three sections: **Basic Probability**, **Statistical inference** and **Statistical modelling**. Each sections build upon the previous. Probability is perhaps an obvious underpinning of statistics; it gives us the *building blocks* for both statistical inference and modelling. In probability (sessions 2 and 3) we cover discrete and continuous distributions, and make use of the *maths refresher* in several sessions. Fundamentally, when dealing with data we often need to make assumptions about what *distribution* the data is drawn from, and knowing the properties of these distributions then enables us to carry out statistical inference. When we move to statistical inference (sessions 4 to 10) this gives us the understanding of how statistical theory can be used to make statements during our investigations. An important consideration is thinking about the statistical theory that enables us to invesitgate our data, but then use this knowledge to then make statements about the wider population (or target population). With this knowledge we then move to Bayesian statistics, where we apply our knowledge to specific health questions and make use of prior knowledge. We then move onto applications that are more likely to be encountered in health applications; and consider the process of investigation and statistical modelling (sessions 11 to 17). These sessions are linked together because as well as being able to run a model and generate results, it is very important to articulate why it was done, ie. identify the purpose of investigation. In the sessions on regression modelling several classes of model will be described and illustrated in detail.

At the end of the module there will be a revision session, and an assessment (more detail is provided on Moodle).

1.2 What you won't learn

We provide important aspects of statistical theory in order for you to understand the reasoning behind the approaches, but there are aspects of statistical theory that are outside of the module scope. In this case, we may provide further reading. Other modules within the LSHTM may cover this in more detail, such as *Foundations of Medical Statistics*.

This module provides the basics that further modules in the MSc may require, such as *Data Challege* and *Machine Learning*. As the name might imply, the statistical techniques used in machine learning are covered in the other module.

The programming associated with this module is carried out in R. The statistical analysis can be carried out in other software, such as Python. In some of the later sessions, we will provide the equivalent Python code, but we do not expect you to use this (and you will not be assessed on this).

1.3 The Data Science Project

Once a health related question is posed, this could be the start of a Data Science Project. Typically, we try to frame the question around a scientific hypothesis, identify some data that can be used to address this question/hypothesis, carry out some anlaysis and draw a conclusion. In David Spiegelhalter's Book *The Art of Statistics* this process is referred to as the PPDAC cycle (fig. 1.1). If doing data science is new to you, this might be considered a linear process. But, in many circumstances the problem solving process is a cycle, where a problem may be solved using a iterative process. The iterative process doesn't mean that the first attempt was *wrong*, but instead this iterative way of thinking enables the data scientist to think critically about each stage of the cycle and identify strengths, weaknesses and opportunities for improvement.

Note that there are many ways to describe the cycle of a Data Science Project, and more examples are given in the lecture by Prof. Nick Jewell. Some might chime with you more (or less) than the PPDAC presented here.

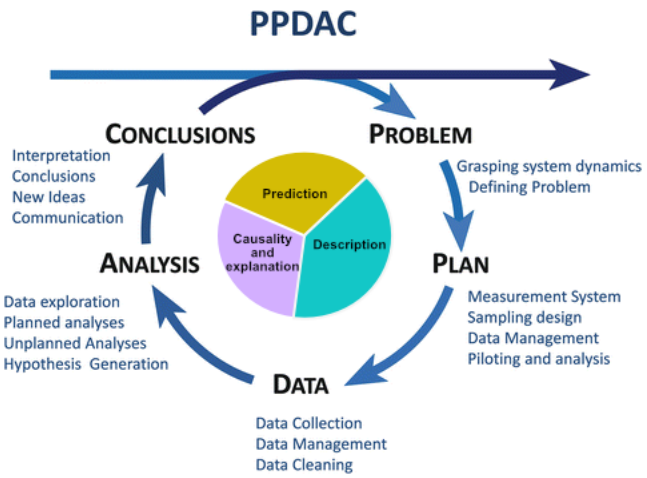


Fig. 1.1 The PPDAC cycle (From *The Art of Statistics* by David Spiegelhalter)

1.3.1 Identify the problem and generate a hypothesis

Typically Health Data Science projects start with a question. The question may be framed around one of three investigation types: **description**, **prediction** and **causality and explanation**. The specific of this is explored in more detail in Session 11 (Types of Investigation).

When developing a Data Science Project there is a need to create a question that is answerable within the timeframe available, and sufficiently precise. It is also preferable to frame a question around a hypothesis that can be a testable prediction, and this is where statistics can be used, because a lot of statistics are framed using a hypothesis (this is especialy true of frequentist statistics). However, it is not always necessary to have a hypothesis, for example if the question is exploratory.

1.3.2 Develop a plan, consider the data design

The plan to answer the question/hypothesis will involve some data. For data science it is likely that the data has not been collected specifically for the purposes of answering the question. Examples may include surveillance data for infectious diseases (eg. self-reported cases of influenza-like illness to a public website) or internet searches for "sore throat". In this case, it is important to recognise specific attributes of the data:

- Where did the data come from? What is its provenance?
- How and why was the data collected?
- What kind of individuals provided data, and why were they selected?

These are important questions because they relate to the principles of *statistical inference*, which is covered in sessions 4-10. Central to using data to draw a conclusion is that your *sample* data is representive of the *population*. Consequently, we can carry out an analysis on the data and make statements about the wider population. This is covered in more detail in session 4 (*Populations and Samples*).

At this stage it is important to identify the "outcome" variable and the "explanatory variables" present in the dataset, and whether we know already that some explanatory variables are associated with the outcome variable. It is also a good idea to idenitfy what type of data each variable corresponds to: continuous, ordinal, categorical.

The design of the dataset is also important, as this helps us understand the structure of the data, and a framework for analyses on the data. Commonly encountered designs are (note that these are also covered in the *Epidemiology for Health Data Science* module):

- A cross-sectional design
- A cohort design
- An outcome-based design
- A longitudinal design

At this stage, you may start to consider the appropriate analysis to make considering the data. As the module (and others, for example the *Data Challenge* module) develops you will identify the analysis steps that can be undertaken according to the question.

1.3.3 The data

There are several aspects of the data that need to be considered, and some of which are covered in the module *Health Data Management*, such as entering the data, managing the data, and cleaning the data.

Here we will focus on aspects which might affect the analysis and conclusions that you make later in the PPDAC cycle.

The first is the presence of potential data filtering, ie. is there any reason to suspect that data are missing or censored, in reference to to wider population? If so, this could result in potential bias. The most commonly encountered bias is *selection bias*, where extrapolation to the wider population may be challenging. Additionally, *collider bias* may result in inappropriate conclusions being made on the effect of explanaotry variables on the outcome.

The second consideration is confounding, where there may be a common cause for both a explanatory variable and the outcome. The result is that an assoication between the explanatory variable and the outcome may be identified, but the relationship is not causal.

1.3.4 Data analysis

Exploratory data anlaysis, and especially **plotting your data** is a really important part of the Data Science Project. As you progress through the module, this will become more and more familiar. Plotting your data is important to *sense check* the data and identify any errors, outliers or omissions (this is especially important with found data). Further to this, many statistical anlaysis benefit from plotting the results, for example by plotting the residuals of a linear model against the outcome to check that the model is correctly specified. Often, suitable plots may carry with them *parameter estimates* from the data, for example the mean number of influenza-like illnesses reported per week when the data are available daily.

It is at this stage that you do the analysis. This is where the concepts covered in this *statistics module* become useful. What we want to emphasize here is that this is done while considering all the others factors within the PPDAC cycle.

1.3.5 Conclusions

So you've gotten this far! An ideal conclusion has brought all the other aspects together, and at most stages some form of statistical inference is considered. The conclusion then needs to consider the statistical result *in the context of the other considerations*, such as wanting to make inference about the population from the sample of data.

For example, let's say the influenza-like illness data from the internet reported 40 cases per 100,000 of the population from November to January. Reporting symptoms might be skewed towards people who regularly use the internet, which might exclude elderly individuals. Consequently, this mean estimate may be an under-estimate of the population incidence due to selection bias.

1.6 Why we teach both frequentist and Bayesian statistics

A majority of the module covers statistical inference from the *frequentist* perspective. Much of frequentist statistical inference was developed by Ronald Fisher, who has been described as the founder of modern statistics, and much of his focus was on experimental design in agriculture. A simple explanation of the philosophy behind frequentist statistics is that a *fact* is either true or not true, and data can be used to assess which of these outcomes can be accepted. In contrast, Bayesian statistics suggests that a probability can be assigned to whether the fact is true. The field of Bayesian statistics is named after Reverend Thomas Bayes, who developed the theory almost 200 years before Fisher was alive (and owing to improvements in computation the theory is now much more accessible and has overtaken frequentist approaches in some scientific fields). In addition, frequentist statistics makes use of the data available, and there is little (or any) ability to incorporate additional knowledge. Within a Bayesian framework, inclusion of prior knowledge is inherent, and this prior knowledge can be combined with data.

Some argue that the philosophies are diametrically opposed to each other, and statisticians should choose a side. This is a strong view (and perhaps not the majority view?), but first it is important to understand the principles behind each approach. We have opted to teach both because of this reason, and leave it up to you to consider the advantages and disadvantages of each approach. Ultimately, both have data central to the approach in making statistical inference, so it is likely that both should be considered as perspectives in a Data Science Project.

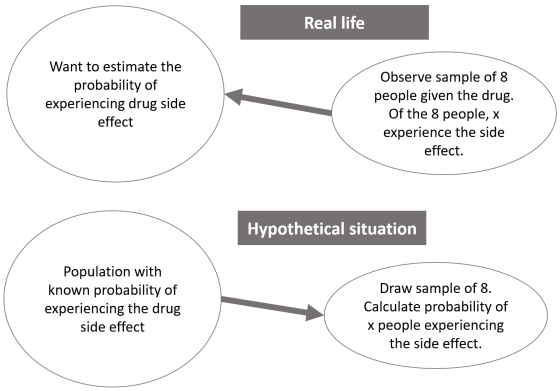
Probability and statistics

We assume that students are already familiar with the basic probability concepts covered in the pre-course Refresher.

How does probability relate to statistics?

Probability theory describes the chance of an event occurring while statistics concerns the collection, organization, analysis and interpretation of data. However, probability theory and statistics are intrinsically linked.

A typical **probability** problem is as follows. We are planning to run a small clinical study, which involves giving 8 patients a particular drug. We are told that the probability that a single patient experiences a side effect from a particular drug is 0.23. From this information, we can calculate the probability of various complex events occurring. For example, we might want to know the probability that more than 6 of the 8 patients will experience a side effect. Or we might wish to know the probability that none of the 8 patients experience a side effect. Here, we are assuming that a characteristic (parameter) of the population is known. Specifically, we are assuming that we know the true probability of a single patient experiencing a side effect.



This is not how real life works! Typically, in health data science studies, we have observed some data which we believe can be modelled using a particular distribution (such as the binomial distribution that we will soon meet), but the parameters of that distribution are unknown. For the small clinical study, for example, in real life we would run the study and observe how many of the 8 patients did in fact experience a side effect. But the probability of a patient experiencing a side effect would be unknown. The study aim would be to use the observed data to make statements - **inferences** - about this unknown probability. So in some senses, the problem is the opposite way round.

It turns out that the process of statistical inference relies very heavily on probability calculations. Suppose we conduct our small clinical study and, for example, observe that in fact 2 of the 8 patients experience the side effect. Loosely speaking, the process of inference involves the following steps. We use probability theory to calculate the probability that 2 people within our sample of 8 experience the side effect, for every possible value of the unknown probability of

experiencing a side effect. We then use these probabilities to make statements about plausible values of the unknown probability. As we will see, we can take various approaches to this inference, in particular using the frequentist or Bayesian frameworks.

Therefore, the next two lectures, concerning probability theory, comprise building blocks that you will need when you subsequently meet ideas about likelihood, inference and regression modelling.

2. Discrete Distributions

This session is the first of two sessions covering useful elements and applications of basic probability. In this first session, we focus on variables which have a **discrete** distribution. In the next session, we extend these ideas to variables which have a **continuous** distribution.

Intended learning outcomes

- apply Bayes' Theorem to obtain useful properties of screening tests
- derive the binomial and Poisson probability distribution functions
- apply the binomial and Poisson distributions to health settings
- evaluate the appropriateness of the binomial and Poisson distributions to health settings

The first part of this session explores Bayes' Theorem. This is a crucial probability theorem, which underlies the Bayesian approach to inference. Another important use is its application to quantify how well a screening test or prognostic classification tool is performing. In this session, we focus on this latter application but you will return to Bayes' Theorem in the later sessions about Bayesian statistics.

The second and third parts of this lecture explore the binomial distribution and the Poisson distribution. For these sessions, we will assume we know characteristics of the population (e.g. the true prevalence of a disease, or incidence rate of a disease) and will explore how to calculate probabilities of various outcomes occurring. In subsequent sessions we will see how these sorts of calculations are used within the important area of statistical inference.

2.1 Application of Bayes' Theorem

Bayes' Theorem has important and powerful applications in medical statistics. One important link, which you will return to later, is its connection to Bayesian statistics. In this session, we focus on another common application of Bayes' theorem, in the area of assessing the accuracy of screening tests and prognostic scores.

2.1.1 Screening tests and prognostic scores

Screening tests are tests that attempt to identify people with a particular condition or disease of interest. Babies are often screened for cystic fibrosis at birth, for example. Sometimes, screening tests attempt to identify high risk people rather than those who already have the condition of interest. Cervical screening, for example, which is offered to all women and people with a cervix aged 25 to 64 in the UK by the National Health Service, is a test to help prevent cancer. It doesn't look for existing cancer, but instead looks for certain viruses which can increase the subsequent risk of cancer. Similarly, prognostic tests or prognostic scores are used to identify a high risk group.

Screening tests or prognostic scores can be based on one genetic marker, as in our example below, or many. Or they might incorporate information from other sources (e.g. biomarkers, family history of the disease). These processes typically result in a binary classification of "positive" or "negative".

2.1.2 Bayes' Theorem

Suppose we have an event A and a set of events $B_1, B_2, ..., B_n$ that partition the sample space. Suppose that we have information about the conditional probability of A conditional on event B_j, i.e. we know $P(A \mid B_j)$, for each j. However, what we actually want to know about is $P(B_j \mid A)$.

Bayes' Theorem provides a way of reversing the conditioning.

Bayes' Theorem:
$$ P(B_{j}|A) = \frac{P(A|B_{j}) P(B_{j})}{P(A)} = \frac{P(A|B_{j}) P(B_{j})}{\sum^{n}_{k=1} P(A|B_{k}) P(B_{k})}. $$

2.1.3 Example: Genetic marker in childhood cancer

In a population, 10% of people develop a particular childhood cancer. Of those who develop the cancer (C), 1 in 4 carry a genetic marker, M, whereas of those who don't develop the cancer, 1 in 10 carry M. A newly born infant is tested for the genetic marker and is found to carry it. What is the probability that this infant will develop cancer?

The first couple of sentences tell us that $P(C) = 0.1$, $P(M|C) = 0.25$ and $P(M \mid \bar{C})=0.1.$ Our interest lies in $P(C|M)$. So we wish to reverse the conditioning. We can obtain this by applying Bayes' Theorem:
$$ P(C|M) = \frac{P(M|C)P(C)}{P(M|C)P(C) + P(M|\bar{C})P(\bar{C})} $$
Substituting in the values above gives
$$ P(C|M) = \frac{0.25\times 0.1}{0.25\times 0.1 + 0.1\times 0.9}=0.22 $$
$P(C|M)$ is called the **positive predictive value** (PPV) of the test. It is the probability, given a positive test result, that the individual actually will develop the disease. i.e. in this case there is a 22% chance that the infant will develop the disease if they tested positive.

2.1.4 The confusion matrix

More generally, suppose we have a procedure that results in a binary classification (a binary prediction). This might be a screening test, which could be based on one or more genetic markers or biomarkers. It could be based on the output from a prognostic risk score or a label derived from an algorithm. Whatever the procedure, suppose we end up with a binary classification: "Positive" or "Negative". In the health context, this terminology (positive/negative) might represent pairs such as: "Diseased" and "Undiseased"; "Dead" and "Alive" or "Hospitalised" and "Not hospitalised". We can contrast the binary classification with the (binary) true status. In the general discussion below, we will also use the terms positive and negative to denote the two possible true statuses.

The following table is often called a **confusion matrix**, or sometimes error matrix. The name confusion matrix stems from the fact that the matrix allows you to see whether the classification is *confusing* two classes. The values A, B, C and D are the numbers in each category. The name comes from the fact that the table allows you to see if the classification procedure is "confusing" two categories.

classification Truth:	**Positive**	**Negative**
Positive	A	B
Negative	C	D

Two groups of people were correctly classified:

- *True Positives*. The A individuals are people who are, in truth, positive (for the disease or outcome of interest) and were classified as positive. So they are often called true positives.
- *True Negatives*. The D individuals are people who are, in truth, negative and and were classified as negative.

Two groups of people were incorrectly classified:

- *False Positives*. The B individuals are people who are, in truth, negative but were incorrectly classified as positive. These are sometimes called Type I errors.
- *False Negatives*. The C individuals are people who are, in truth, positive but were incorrectly classified as negative. These are sometimes called Type II errors.

Now let us imagine the same table but with joint probabilities rather than numbers from a sample. So, for instance, p_A is the joint probability of being classified as positive *and* being truly positive for the outcome.

Prediction Truth:	**Positive**	**Negative**
Positive	p_A	p_B
Negative	p_C	p_D

We can obtain estimates of various useful quantities from this matrix. We will use the following notation: O represents being, in truth, positive for the outcome of interest, and $\bar{O}$ represents being truly negative. P represents being classified as positive and $\bar{P}$ being classified as negative.

The tabs below show various useful quantities.

Outcome prevalence | Sensitivity | Specificity | PPV | Accuracy

The prevalence of the outcome is:
$$ P(O) = \frac{p_A+p_C}{p_A+p_B+p_C+p_D} $$
Prevalence is another word for risk or proportion. It tells us the fraction of the population of interest who have the outcome.

2.2 The binomial distribution

The binomial distribution is used to model the number of successes out of a fixed number of trials.

In the following calculations, we will assume that we know the true probability of success within each trial. In practice, of course, this probability is often the unknown quantity that we are trying to estimate. Later sessions will revisit this example, under the more realistic scenario where this probability is unknown and we are using the sample of data to *make inferences* about the probability. The calculations in the current session will form important building blocks for those later sessions.

> Note on terminology:
>
> - Do not confuse the word "trial" here with the idea of a clinical trial or randomised controlled trial. In our discussion of the binomial distribution, we simply mean a Bernoulli trial, which is a statistical experiment which results in a binary outcome. So the trial in question could be whether or not a baby is a male; whether or not someone is alive in 30 days time; whether or not someone experiences a side effect.
>
> - Similarly, the word "success" can be confusing. We use the word success to denote having the event of interest. It does not imply that this is a good event. In fact, the event we are interested in, in health applications, is often a bad one. It might be diagnosis of cancer or death, in which case a success would refer to someone having cancer or dying. Conversely, if our study was looking at treatments for improving pregnancy rates, our event, and thus the definition of success, might be a couple becoming pregnant. So the word success, in this context, does not necessarily refer to a good event (although sometimes it does!).

2.2.1 Example of a binomial distribution

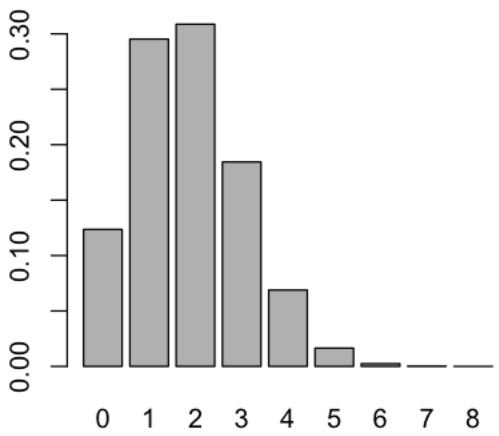
A small study of 8 participants is being run. All 8 participants will be given an experimental drug. The aim of this study is to obtain data about how many people will experience a side-effect of the drug.

From previous data, the clinical researcher running the trial believes that the probability of the side-effect is 0.23.

Let X be the number of people in the study (i.e. among the 8 participants) who experience a side-effect. Suppose we are happy to assume that X follows a binomial distribution. Then, using the formula for the probability distribution function that we derive below, we can calculate the probability that $P(X=x)$ for all possible values $x=0,1,\ldots,8$.

The code below (in R) does that calculation and displays a bar chart of the probability distribution function.

```
# Obtain the probability distribution function (for values x=0,1,...,8)
x <- seq(0,8)
pi <- 0.23
px <- dbinom(x, 8, pi)

# Create bar chart of PDF
options(repr.plot.width=4, repr.plot.height=4)
barplot(height=px, names=x)
```



Deriving the binomial distribution

Suppose we are conducting research on quadruplets (sets of four siblings born within the same pregnancy). In this session we will consider the number of boys among a set of quadruplets.

Let X be the number of boys within a particular set of quadruplets. The sample space for X (the set of possible values X could take) is: $\{0, 1, 2, 3, 4 \}$. We will now derive the full probability distribution function for X. In our calculations, we will assume that the proportion of males at birth is 0.51 and that the gender of each birth is an independent event.

Consider one set of quadruplets. We will start by calculating the probability of no boys i.e. the probability of four girls. By applying the multiplication rule (using the assumption of independence between sex of the children) we obtain:
$$ P(X=0) = P(\mbox{four girls}) = P(GGGG) = 0.49^4, $$
where $GGGG$ is shorthand for the event that the first child is a girl, *and* the second is a girl, *and* the third is a girl, *and* the fourth is a girl,

Consider now the probability of one boy and three girls. This may occur in one of four ways: BGGG, GBGG, GGBG and GGGB, each of which has probability $0.49^3\times 0.51$. Thus
$$ P(X=1) = P(\mbox{one boy}) = 4 \times 0.49^3 \times 0.51. $$
A family of 2 boys and 2 girls will arise in one of the following 6 ways: BBGG, BGBG, BGGB, GBBG, GBGB, GGBB each with a probability $0.49^2 \times 0.51^2$ and a total probability of
$$ P(X=2) = P(\mbox{two boys}) = 6 \times 0.49^2 \times 0.51^2. $$
With similar reasoning we have that
$$\begin{split} P(X=3) = P(\mbox{three boys}) = 4 \times 0.49 \times 0.51^3 \\ \end{split}$$
and
$$ P(X=4) = P(\mbox{four boys}) = 0.51^4. $$
We now let X be the random variables which records the number of boys in a randomly selected family of size four. This random variable takes four possible values: 0, 1, 2, 3 or 4. Its probability distribution is given by the following table:

x P(X=x)

0 0.49^4 = 0.0576

1 4 $\times$ 0.49^3 $\times$ 0.51 = 0.2400

2 6 $\times$ 0.49^2 $\times$ 0.51^2 = 0.3747

3 4 $\times$ 0.49 $\times$ 0.51^3 = 0.2600

4 0.51^4=0.0677

More generally, consider a sequence of n independent observations/trials (in the example above it was four). Each observation results in a binary outcome, e.g. each trial is a success or a failure. In fact, a Binomial sequence is the sum of n independent Bernoulli trials (i.e. n independent Bernoulli variables). Let π denote the probability of an individual success (or the defined binary feature, e.g. boy vs. girl).

How do we obtain the probability distribution for the random variable X which records the number of successes in a sequence of n trials? The possible values for the random variable are $0,1,..,n-1,n$. We saw from the previous example that the probability of x successes and $n-x$ failures is
$$ P(X=x) = \pi^{x} (1-\pi)^{n-x} \times \mbox{number of ways of obtaining } x \mbox{ successes}. $$
The multiplying factor on the right above is the binomial coefficient, i.e. the number of combinations of x objects chosen from n. The number of ways x successes can be obtained from n observations is equal to ${}^n C_x$ as we are not interested in the order of the successes, only the number of combinations in which such a number of successes could have occurred, and a "success" can be considered the same as "choosing" an object: we are "choosing" x successes and $n-x$ failures out of a "bag" of n successes and failures.

So we have that
$$\begin{split} P(X=x) = \begin{pmatrix}n \\ x \end{pmatrix} \pi^{x} (1-\pi)^{n-x} \end{split}$$

General form of the binomial distribution

Suppose we have a sequence of n independent Bernoulli trials (i.e. n independent Bernoulli variables). Let π denote the probability of an individual "success". To write that X follows a binomial distribution with these features, we write $X\sim binomial(n,\pi)$, (where $\sim$ means "follows").
$$\begin{split} P(X=x) = \begin{pmatrix}n \\ x \end{pmatrix} \pi^{x} (1-\pi)^{n-x}, \mbox{ for } x=0,1,2,..,n. \end{split}$$
Expectation and variance

The expected value of a binomial variable is $E(X) = n\pi$

The variance of a binomial variable is $Var(X) = n\pi (1-\pi)$

Applications of the Binomial distribution

Assumptions

In order for a variable to follow a binomial distribution, some "structural" things need to be true.

1. There must be a fixed number of Bernoulli trials
2. Each trial must result in a binary outcome (success or failure)
3. The outcome we are interested in must be defined as the total number of successes.

There are also two key *statistical assumptions*, implied by our derivation above:

1. The Bernoulli trials must be *independent* of one another
2. The probability of success must be the same across Bernoulli trials

Applications

Suppose we are interested in a particular disease within a large population of N individuals. If, in the population, M is the number of individuals with the disease of interest, then the probability of "success" (i.e. an individual having the disease) is $\pi = M/N$.

Suppose we take a random sample of n individuals from the large population. We will use X to be the random variable for the number of "successes" out of the n individuals. Then we might be happy to assume that X follows a Binomial distribution.

Notes

- In order for the probability π to remain constant, if we took another sample of n we would have to "replace" the original n individuals, so there would be some small possibility of picking the same person twice. In practice, people are not sampled twice. But populations are usually so large that we can ignore this.

- We also need to assume that individual outcomes (here, having the disease or not) are independent. There are many ways in which this could be violated. People within the same family have shared genetics, shared environments, etc. all of which might lead to outcomes that are more similar between family members than between individuals from different families.

2.3 The Poisson distribution

The Poisson distribution is used to model the *number of events* occurring in a fixed time interval.

Similarly to our approach with the Binomial distribution, in the following calculations we will assume that we know the true rate at which events occur. In practice, of course, this rate is often the unknown quantity that we are trying to estimate. Later sessions will revisit this example, under the more realistic scenario where this rate is unknown and we are using the sample of data to *make inferences* about the rate. The calculations in the current session will form important building blocks for those later sessions.

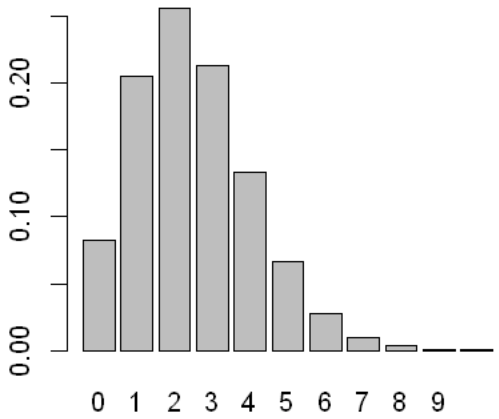
2.3.1 Example of the Poisson distribution

A clinical research is interested in modelling the number of asthma attacks that people with asthma experience in one year. Based on a large sample the researcher has estimated that the average number of attacks in a year is 2.5.

If we let X be the variable for the number of attacks a randomly selected person with asthma will experience in a year and we are happy to assume that X follows a Poisson distribution, then we can calculate $P(X=x)$ for any given value of x.

The code below (in R) does this calculation and plots the probability distribution function of the number of asthma attacks in a year.

```
# Obtain the probability distribution function (for values x=0,1,...,10)
x <- seq(0,10)
lambda <- 2.5
px <- dpois(x, lambda)

# Create bar chart of PDF
options(repr.plot.width=4, repr.plot.height=4)
barplot(height=px, names=x)
```



2.3.2 Deriving the Poisson distribution

To give a heuristic derivation of the probability distribution function of the Poisson, we divide the total time T into a very large number of small intervals (see Figure below). As the number of intervals we divide T into increases, at most one event will occur in each interval, and so X will equal the number of intervals in which an event occurs. Since the occurrence of events in each interval are assumed independent of each other, $X \sim Bin(n,\pi)$, where n is the number of intervals and π is the probability of an event occurring in any given interval.

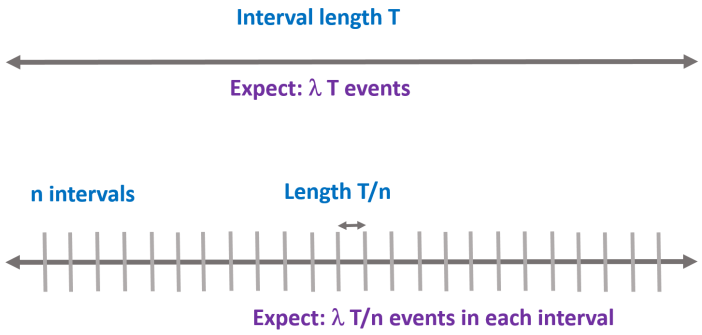


Fig. 1 Derivation of Poisson distribution by dividing time into small intervals

With a rate of λ events per unit of time, we expect $\mu=\lambda T$ events in the whole period, and therefore we expect $\lambda T / n = \mu/n$ events in each interval. Thus $\pi=\mu/n$. Therefore, using the probability distribution function for the binomial we have that

$$ P(X=x) = {n \choose x} \pi^{x} (1-\pi)^{n-x} = {n \choose x} \left(\frac{\mu}{n}\right)^{x} \left(1-\frac{\mu}{n}\right)^{n-x} $$

Then we have that

$$ P(X=x) = {n \choose x} \left(\frac{\mu}{n}\right)^{x} \left(1-\frac{\mu}{n}\right)^{n-x} $$

$$ = \frac{n!}{x! (n-x)!} \left(\frac{\mu^{x}}{n^{x}}\right) \left(1-\frac{\mu}{n}\right)^{n-x} $$

$$ = \frac{n!}{n^{x} (n-x)!} \frac{\mu^{x}}{x!} \left(1-\frac{\mu}{n}\right)^{n-x} $$

Now to simplify the first term, we note that:

$$ \frac{n!}{n^{x} (n-x)!} = \frac{n(n-1)...(n-x+1)}{n^{x}} \rightarrow 1 \mbox{ as } n \rightarrow \infty, $$

and to simplify the third term, we note that:

$$ \left(1-\frac{\mu}{n}\right)^{n-x} \rightarrow \left(1-\frac{\mu}{n}\right)^{n} \rightarrow e^{-\mu} $$

Replacing the first and third terms by these limits gives

$$ P(X=x) \rightarrow \frac{\mu^{x}}{x!} e^{-\mu} \mbox{ as } n \rightarrow \infty. $$

2.3.3 General form of the Poisson distribution

We can now define a Poisson distribution for the number of events occurring in a fixed interval T at a constant rate λ with parameter $\mu=\lambda T$, which we write as

$$ X \sim \mbox{Poisson}(\mu=\lambda T) $$

as the distribution which has probability distribution function

$$ P(X=x) = \frac{\mu^{x}}{x!} e^{-\mu}, \ \mbox{ for } x=0,1,2,... $$

Expectation and variance

The derivation of the expectation and variance of a Poisson random variable X with parameter μ will be set as a practical question.

2.3.4 Applications of the Poisson distribution

Assumptions

The Poisson distribution is used to model the *number of events* occurring in a fixed time interval T when:

- events occur randomly in time,
- they occur at a constant rate λ per unit time,
- they occur independently of each other.

Applications

A random variable X which follows a Poisson distribution can take any non-negative integer value. Examples where the Poisson distribution might be appropriate include:

- Emissions from a radioactive source,
- The number of deaths in a large cohort of people over a year,
- The number of accidental deaths occurring in a city over a year.

2.3.5 Approximating the binomial by a Poisson

When n is large relative to π, the binomial distribution can be approximated by a Poisson with a mean $n\pi$. That this approximation is reasonable follows directly from our earlier heuristic derivation of how a Poisson distribution arises as an approximation to a binomial distribution when the number of trials tends to infinity.

There are many such approximations. Nowadays, we may not need to use them because we have enormous computing power at our disposal. In earlier times, in contrast, calculations could take a long time so any simplification that could be reasonably applied could provide meaningful extra calculation speed.

2.4 Summary

In this session, we have met a number of useful applications of discrete probability theory.

- Bayes' Theorem can be used to quantify properties of screening and prognostic tests and forms the basis of Bayesian statistics. We will return to Bayes' Theorem when we explore Bayesian inference.
- The binomial distribution is used to model the number of successes out of a fixed number of trials. We will revisit the binomial distribution in the sessions about likelihood. Logistic regression, a very commonly used regression model, is based on an underlying Bernoulli distribution (remember that the binomial distribution can be obtained by summing multiple identical independent Bernoulli distributions). As such, our work with the binomial distribution is closely connected to the later logistic regression sessions.
- The Poisson distribution is used to model the number of events occurring in a fixed time interval. It forms the basis for Poisson regression, which you will meet later in this module.

3. Continuous distributions

This session is the second of two sessions covering basic probability. In this session, we extend the ideas from discrete distributions, from the previous session, to variables which have a **continuous** distribution.

Intended learning outcomes
By the end of this session you will be able to:

- explain the concept of a continuous random variable
- define several continuous probability distributions, and relationships between parameters, expectations and variance
- understand the relationship between normally distributed data and standard scores
- evaluate the appropriateness of assuming normality in data and other options
- understand properties of joint distributions, such as the multivariate normal distribution

The five sub-sections describe properties of continuous random variables, explore a number of useful continuous distributions, consider direct applications for the standard Normal (Gaussian) distribution, consider how to assess whether a variable follows a normal distribution and, finally, describe joint distributions and correlations.

3.1 Continuous random variables

We have previously seen several discrete probability distributions (including the binomial and the Poisson). We now extend random variables to those that are continuous. A continuous random variable is one that can take a value in continuous space; this may vary from $-\infty$ to $+\infty$ (like the normal distribution) or have limits set on the lower (eg. the log-normal) or upper bound (eg. the uniform).

3.1.1. The probability density function

Previously we characterised the distribution of a variable by assigning a probability to each specific value. However, because athere are infinitely many values that could be taken by a continuous variable, paradoxically, the probability of a continuous random variable taking any specific value is zero. Therefore, we cannot use a probability distribution function

to characterise the distribution of a continuous variable.

Instead, we turn to something called a **probability density function**. Instead of attaching a probability to each value the variable could take, the probability density function tells us the probability that a continuous variable lies within each possible interval (range of values). Specifically, the area under the curve (of the probability density function) between two limits tells us the probability that the continuous variable takes a value between those two limits.

Generally, a random variable X has density f_X where

- $f(x) \geq 0$ for all of x
- $\int_{-\infty}^{\infty} f(x) \hspace{0.2cm} dx = 1.00$

which states that the "sum" of all probabilities of $f(x)$ from the minimum to the maximum is equal to 1.

We can obtain various useful probabilities from this density function. We can calculate the probability that the variable takes a value within a given interval, the probability that it is below or above a given value. For example:
$$Pr(X>b) = \int_{b}^{max} f(X) \hspace{0.2cm} dx$$
Further information about continuous probability distributions are given in the Refresher.

3.2 Useful continuous distributions

Below are several useful probability distributions for data science in health. Some of the information below is a repeat of the **Maths refresher**, but we include some practical applications of each distribution.

3.2.1 The normal distribution

The normal distribution is defined with the following probability density function:
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}exp \Big[-\Big(\frac{1}{2}\Big)\Big(\frac{x-\mu}{\sigma}\Big)^2\Big]$$
for values x in $(-\infty, +\infty)$. If we have a random variable X that is normally distributed we can specify this using $X {\sim} N(\mu, \sigma^2)$. The expected value is given by $E[X]=\mu$ and the variance is given by $Var[X] = \sigma^2$.

A **standard normal** distribution has a mean of 0 and a variance of 1. A standard normal random variable is usually represented by $Z {\sim} N(0,1)$ and is sometimes called the *Z-score*.

So much of statistics relies on the normal distribution, so it is an important distribution to be familiar with. We will see that the normal distribution has an important role to play in statistical inference. It is also sometimes a good distribution for directly modelling continuous variables, for example blood pressure.

3.2.2 The log-normal distribution

The log-normal distribution is essentialy a transformed version of the normal distribution, and has its own probability density function;
$$f(x) = \frac{1}{x \sigma \sqrt{2\pi}}exp\Big(-\frac{(ln(x)-\mu)^2}{2\sigma^2}\Big)$$
for values x in $([0,+\infty)$. If a random variable X is log-normally distributed, $Y=ln(X)$ has a normal distribution, and if Y is a normal distribution then $X=exp(Y)$ has a log-normal distribution. These simple transformations mean that calculations using transformed data is the standard approach. The parameters μ and σ refer to the mean and standard deviation on the *normal scale*. Consequently, the median of a log-normally distributed sample is $exp(\mu)$.

Many biological datasets are log-normally distributed, for example most measurements (height, weight, speed) will be above 0, and will often be right-skewed. A good approach to take with these sorts of data is to log the data, and work on the *log scale*. Any inference should be converted back to the *natural scale*. Sometimes measurements are sufficiently greater than 0 that they become more centered. In this case, it may not be necessary to assume that are log-normal, and assuming normality may be acceptible.

3.2.3 The χ^2 distribution

The χ^2 distribution is here because we will use the properties of this distribution later in hypothesis testing. Its origins come from a random sample of the *standard normal*, where the χ^2 distribution is the distribution of the sum of squared standard normals. The degrees of freedom come from the number of standard normal random variables

being summed. It is not necessary to know the parameters or estimates of the χ^2 parameters. A variable which follows the chi-squared distribution can only take positive values (i.e. greater than zero).

3.2.4 The t-distribution

Student's t-distribution arises as the ratio of the sample mean to its standard error. The t-distribution has a complex density function which we shall not state here.

For now we note that the t-distribution has an additional parameter of sorts, known as the degrees of freedom (d.f.). The density function is similar to that of the standard normal, but the t-distribution has heavier tails. If X follows a t-distribution with ν degrees of freedom, we write
$$X \sim t_\nu$$
The expectation and variance of a variable X which follows a t-distribution with ν degrees of freedom are given by:

- $E[X] = 0$
- $Var[X] = \frac{\nu}{\nu-2}$ if $\nu>2$; ∞ for $1<\nu<2$; undefined otherwise

As the number of degrees of freedom increases the t-distribution gets closer and closer to the standard normal distribution.

3.2.5 The F distribution

The F distribution doesn't have a simple mathematical formula, but is used extensively to compare equality of variances of two normal populations (*think anova*), and is used in linear regression.

For two normal populations with variances σ_1^2 and σ_2^2, the two random samples of size n_1 and n_2 with corresponding sample variance(s) s_1^2 and s_2^2 has the variable
$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$
with n_1-1 and n_2-1 degrees of freedom.

3.2.6 The exponential distribution

The exponential distribution is defined with the probability density function:
$$f(x)=\lambda e^{-\lambda x}$$
with parameter λ, which is usually described as the rate. The limits of the distribution are $[0,\infty)$, which means values of x are always greater than 0 (and not including it).

The expected value is given by $E[X]=\frac{1}{\lambda}$ and variance $Var[X]=\frac{1}{\lambda^2}$.

The exponential distribution is really useful in statistics because its distribution nicely describes *the time to which something occurs*, if the event happens at a roughly constant rate in time. Health related examples include injuries, births and deaths (although in reality not all occur at a constant rate). The exponential distribution is important in methods such as *survival analysis*.

3.2.7 The uniform distribution

The uniform distribution is in some ways the simplest to conceptualise. A random variable that is uniformly distributed can have any value between the parameters a (min) and b (max) with equal probability;
$$f(x)= \frac{1}{b-a}$$
Outside of these limits, the probability density is 0. The expected value is $E[X] = \frac{(a+b)}{2}$ and variance $Var[X] = \frac{(b-a)^2}{12}$.

The uniform distribution is very commonly used when randomly allocating outcomes. An example in statistical modelling includes stochastic infectious disease modelling; here several different events (transmission, death) may have a corresponding probability and one event needs to be selected from the two options. A uniform distribution (where the maximum is the total probability of all events) is used to select

3.3 Uses of the standard Normal distribution

Suppose we wanted to answer the following question:

> What is the probability of having a 'healthy' weight?

A healthy weight is often is often measured using the Body Mass Index (BMI - although see here and here for a discussion on why this may be too simplistic a measure). An individual's BMI can be calculated using their height and weight, using the formula BMI \(= \frac{mass(kg)}{height(m)^2}\). Then people can be classified as:

Classification	BMI
Underweight	<18.5
Normal	18.5-24.9
Overweight	25-29.9
Obese	30-39.9
Extremely obese	>40

To address our question, we will use data taken from a study undertaken among a group of 76 cleaners, that investigated whether telling the cleaners they had an active lifestyle influenced their BMI. We will assume that values of BMI approximately follow a normal distribution. We do not know the true values of \(\mu\) and \(\sigma\) so we will replace these with the sample mean and standard deviation. This gives us values of \(\mu=\)26.5 and \(\sigma^2=\) 18.1, as demonstrated in the snippet of code below.

```
# BMI dataset
dat <- read.csv("Practicals/Datasets/BMI/MindsetMatters.csv")
head(dat)
#remove observations with no BMI data
dat <- dat[!is.na(dat$BMI),]
#estimate mu and sigma
mu <- mean(dat$BMI)
print(paste0("value of mu is ",round(mu,2)))
sig <- sd(dat$BMI)
print(paste0("value of sigma is ",round(sig,2)))
```

	Cond	Age	Wt	Wt2	BMI	BMI2	Fat	Fat2	WHR	WHR2	Syst	Syst
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int
1	0	43	137	137.4	25.1	25.1	31.9	32.8	0.79	0.79	124	11
2	0	42	150	147.0	29.3	28.7	35.5	NA	0.81	0.81	119	11
3	0	41	124	124.8	26.9	27.0	35.1	NA	0.84	0.84	108	10
4	0	40	173	171.4	32.8	32.4	41.9	42.4	1.00	1.00	116	12
5	0	33	163	160.2	37.9	37.2	41.7	NA	0.86	0.84	113	11
6	0	24	90	91.8	16.5	16.8	NA	NA	0.73	0.73	NA	N

A data.frame: 6 × 14

```
[1] "value of mu is 26.46"
```

```
[1] "value of sigma is 4.25"
```

So what is the probability a randomly selected person in this sample has a normal BMI?

Approach 1: Manual calculation

One option is to make use of pre-calculated probabilities of the standard normal distribution. If we write \(X\) to represent the value of a person's BMI, then we are assuming that
\[X \sim N(\mu=26.5, \sigma^2=18.1) \]
To make use of the pre-calculated probabilities for the standard normal distribution, we must first transform our normally distributed variable to have a standard normal distribution. We know that the transformed variable \(Z\) (the *Z score*) has a standard normal distribution, where
\[Z = \frac{X - \mu}{\sigma} \]

Given values for \(\mu\) and \(\sigma\) we can go from the *X scale* to the *Z scale* and *vice versa*. The important point about describing a distribution on the Z scale is that this opens the ability to calculate specific probabilities. So back to answering the question...

From the table above we can see that a normal weight is classified as a BMI between 18.5 and 24.9, and we want to know what the probability is that a randomly selected person falls between these limits. We write this as;
\[P(18.5 < X < 24.9) \]
On the Z-scale, this is equivalent to saying that:
\[P(-1.87 < Z < -0.37) = P(Z < -0.37) - P(Z < -1.87) \]
Tables exist containing a range of pre-calculated probabilities that a variable following a standard normal distribution takes a value of less than \(k\), for a range of possible values of \(k\). These are often called *z-tables* (found online or at the back of most stats books). From these tables, we can look up the corresponding probability for each z-score, giving:
\[0.3557 - 0.0307 = 0.325 \]

Approach 2: Using R to do the same calculation

Using this approach, R is ultimately using the same pre-calculated probability tables. However, it is considerably quicker and easier to ask R to look up the values rather than finding them in tables.

```
# a) if we were to use Z tables within R (to illustrate the point)

z_min <- (18.5-mu)/sig
z_max <- (24.9-mu)/sig

# note when using pnorm we don't need to specify mu and sigma as the
# function assumes mu=0 and sigma=1 unless specified.
print(paste0("z_max is ",round(z_max,2)," and z_min is ",round(z_min,2)))
print(paste0("Probability of having a healthy BMI is (z-score) ",round(pnorm(z_max)-
pnorm(z_min),3)))
```

```
[1] "z_max is -0.37 and z_min is -1.87"
```

```
[1] "Probability of having a healthy BMI is (z-score) 0.326"
```

Approach 3: Using R to do the calculation on the untransformed scale

```
# b) if we were to directly estimate

print(paste0("Probability of having a healthy BMI is (direct)
",round(pnorm(24.9,mu,sig)-pnorm(18.5,mu,sig),3)))
```

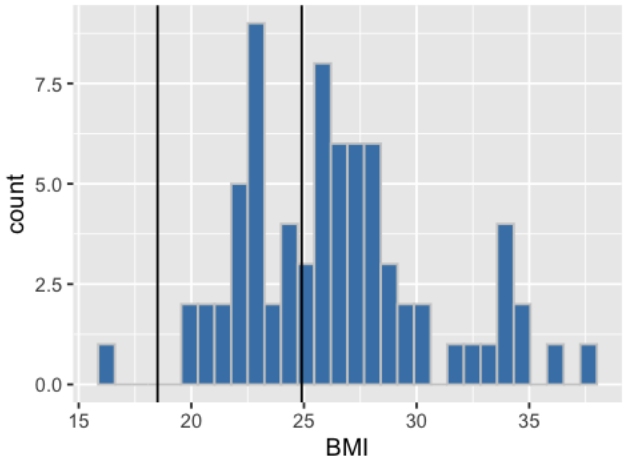
```
[1] "Probability of having a healthy BMI is (direct) 0.326"
```

Calculating directly gives the same result as using a z-score in R, and this returns the same information as using z-tables.

In answer to our question, we estimate that the probability of having a 'healthy' weight is 32.6%. We can compare this to the observed proportion of our sample of data with a 'healthy' BMI.

```
# c) provide a sanity check against the data
options(repr.plot.width=4, repr.plot.height=3)
library(ggplot2)

ggplot(dat,aes(x=BMI)) + geom_histogram(bins = 30,fill="steelblue",col="grey80") +
    geom_vline(xintercept = c(18.5,24.9))
#hist(dat$BMI,col="steelblue")
#abline(v=c(18.5,24.9),lty=2)
print(paste0("Within the data a healthy BMI is seen ",
             round(100*((sum(dat$BMI<24.9)-
sum(dat$BMI<18.5))/length(dat$BMI)),1),"%"))
```

```
[1] "Within the data a healthy BMI is seen 35.1%"
```



So we can see that the sample estimate (35.1%) is roughly similar to the population estimate of 32.6%.

3.4 Are the data normally distributed?

3.4.1 Data and their relationship with statistical distributions

We often have data on a particular characteristic and want to make general statements about it, such as: the probability of it being greater than or less than something, provide a range in which "most" observations will lie, what is the central value (e.g. mean/median), etc.

> However,
>
> - we rarely *know* the true distribution that a variable follows
> - a distribution will not quite fit the data but will form a sufficiently good approximation to address the questions above with sufficient accuracy.

So we often want to find a distribution which fits our data well enough. How do we make a decision? Some of this comes with experience, but there are some useful steps to go through when confronted with data (this is covered in more detail in the lecture). Think back to the **PPDAC** cycle in the first session;

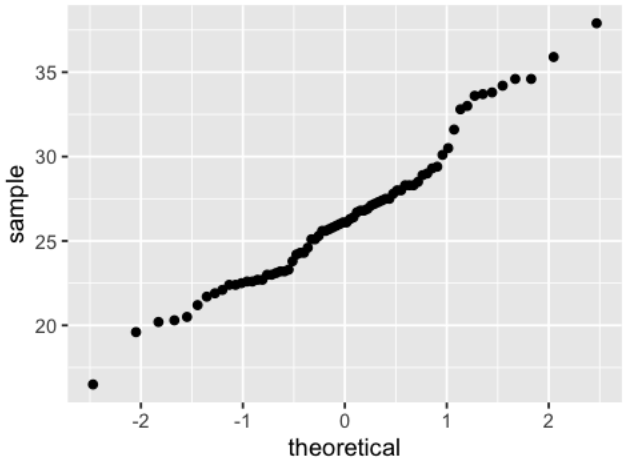
- plot your data. What does the data look like? Consider the lower and upper bounds, the most common number, and evidence of symmetry
- summarise your data. Report the minimum, maximum, mean and mode. This should aid with thinking about the criteria of specific distributions
- depending upon the application and what the data looks like, you may want to consider using the empirical distribution function rather than assumption a specific form. However, this gives you fewer options for inference

3.4.2 Are the data normally distributed?

Many analyses and tests of data start with the assumption that the data are normally distributed. A simple example would be using a t-test to check whether the mean of 2 groups are different, more complex examples would include linear regression analysis. If the outcome being analysed is a qualitive outcome, or successes and failures, it should be obvious that the data aren't normally distributed. But what if the data are continuous or count values, and they look like they are centered, but have some skewness? Is it safe to proceed as if they are normal?

The first step, as always, is to plot the data to see what they look like. A histogram, as above, or density plot is a good step forward. Additionally, a *quantile-quantile* plot calculates the correlation between a sample and the equivalent normal distribution with the same mean μ and standard deviation σ. If a variable follows a normal distribution, the quantile-quantile plot will follow the diagonal line.

```
# BMI dataset
library(ggplot2)
dat <- read.csv("Practicals/Datasets/BMI/MindsetMatters.csv")
dat <- dat[!is.na(dat$BMI),]
options(repr.plot.width=4, repr.plot.height=3)
ggplot(dat,aes(sample=BMI)) + stat_qq()
```



From the figure you can see that the theoretical quantiles follow the diagonals reasonably well, and especially at the extremes do not move away much from the diagonal. A plot like this would be enough to show that the data approximately follows normally distribution. Looking at plots such as this to assess normality is a *judgement* which you will build up during this module.

To formally test for normality we can use the Shapiro-Wilk test, described briefly below.

```
shapiro.test(dat$BMI)
```

```
        Shapiro-Wilk normality test

data:  dat$BMI
W = 0.9692, p-value = 0.06756
```

Although we haven't yet covered hypothesis testing (see *Session 8*), this is testing the null hypothesis that the data follow a normal distribution. In this case, the test returns a p-value of 0.067. This p-value suggests some, but not strong, evidence against normality of the data.

3.4.3 Approaches to non-normally distributed data

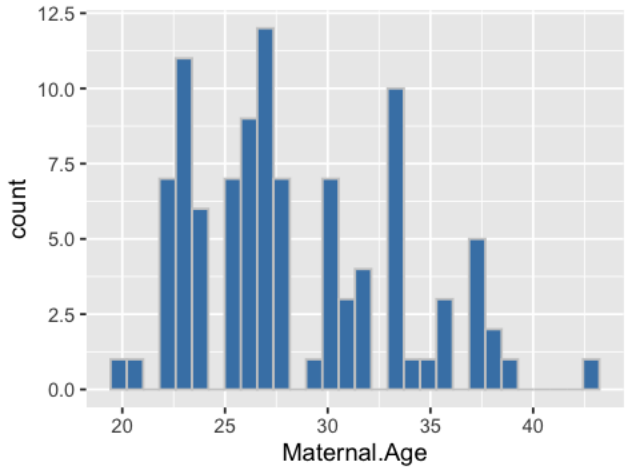
A really useful approach to dealing with non-normally distributed data is transformations. The most often used approach is to apply a log-transformation, either on the *natural* (\(Y = log_e(X)\)) or *log10* (\(Y = log_{10}(X)\)) scale. The transformed data may behave more like normally distributed data.

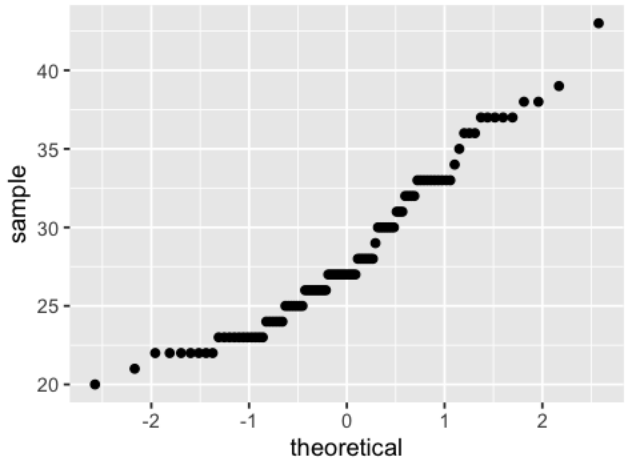
An example is given below for weights of 1174 babies. First we will look at the distribution of (untransformed) birth weights.

```
options(repr.plot.width=4, repr.plot.height=3)
library(ggplot2)

# mother-baby dataset
dat <- read.csv("Practicals/Datasets/MotherBaby/baby.csv")
head(dat)
dat <- dat[1:100,] # we will use just the first 100 observations

# plot the data on maternal age
ggplot(dat,aes(x=Maternal.Age)) + geom_histogram(bins=30,fill="steelblue",col="grey80")
# plot a quantile plot of this log-normally distributed data
ggplot(dat,aes(sample=Maternal.Age)) + stat_qq()
```

	Birth.Weight	Gestational.Days	Maternal.Age	Maternal.Height	Maternal.Pregnancy.We
	<int>	<int>	<int>	<int>	<
1	120	284	27	62	
2	113	282	33	64	
3	128	279	28	64	
4	108	282	23	67	
5	136	286	25	62	
6	138	244	33	62	

A data.frame: 6 × 6

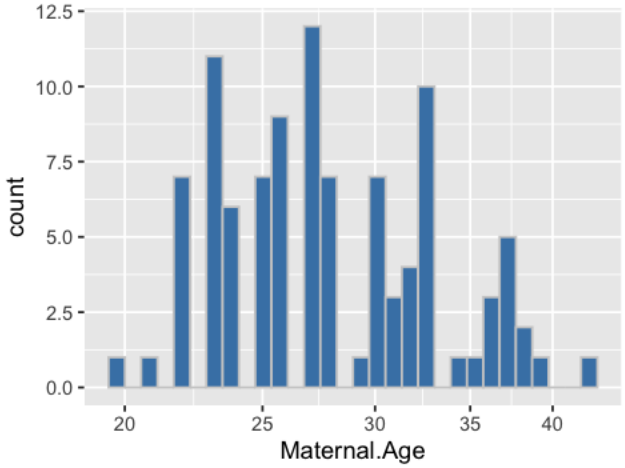


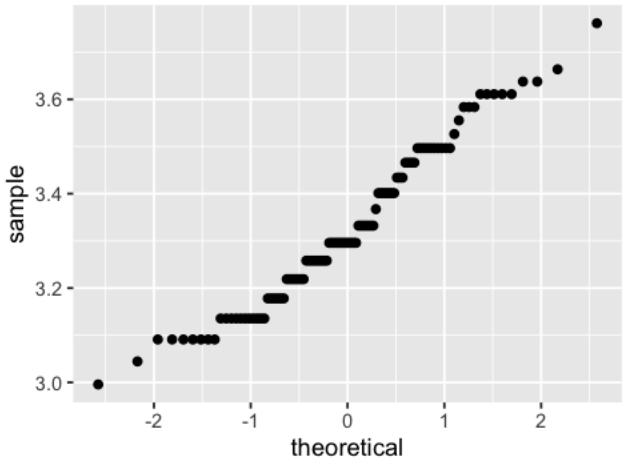


We can see clearly that maternal age is right skewed. This is a classic log-normal distribution. The quantile plot is not straight along the diagonal but forms an *s-shape*. This confirms that the data does not conform to a normally distribution.

A sensible next step would be to log-transform the data using the natural logarithm. The distribution of the transformed birth weights is shown below.

```
options(repr.plot.width=4, repr.plot.height=3)

# plot the data on maternal age
ggplot(dat,aes(x=Maternal.Age)) +
    geom_histogram(bins=30,fill="steelblue",col="grey80") +
    scale_x_continuous(trans = "log")

# but note that any analysis should be carried out on the transformed variable
y <- data.frame(age_log=log(dat$Maternal.Age))
# and here we should check whether this is normally distributed using a qqplot
ggplot(y,aes(sample=age_log)) + stat_qq()
```





The log-transformed data now looks more symmetrical in the histogram. And the quantile plot is much less *s-shaped*. While it's not perfectly straight, it's probably *good enough* for further analysis which relies on the assumption of normality.

3.5 Joint distributions and correlations

We are often interested not in the distribution of a single variable but in the relationship between two or more variables. This requires us to understand the concepts of **joint distributions** and **correlation**.

Returning to the BMI dataset, a high BMI is indicative of being overweight and this is likely to mean that an individual may have a high percentage of body fat. Typically, those individuals with high BMI may also be at risk of health conditions such as heart disease, which may be indicated by high blood pressure.

If we wish to address questions relating to two or more variables, we need to understand their joint distribution.

3.5.1. Joint distributions

If we have two random variables X and Y, the cumulative joint distribution function (CDF) is,
$$F(x,y) = P(X \leq x,Y \leq y)$$
regardless of whether X and Y are continuous or discrete. For continuous random variables the joint density function will be $f(x,y)$ and will be non-negative and
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\: dy\: dx = 1.$$

3.5.2 Marginal distributions

We might sometimes want to think about the marginal density of, say, X. This means we want to know the probability of X irrespective of Y, and consequently we will need to integrate over all possible values of Y. The marginal cdf of X, or F_X is
$$F_X(x) = P(X \leq x)$$
$$= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u,y)\: dy\: du$$
From this, it follows that the density function of X alone, known as the **marginal density** of X, is

$$f_x (x) = F_{X}'(x) = \int_{-\infty}^{\infty} f(x,y)\: dy$$
Note that this is different to assuming that X is independent of Y.

So what does this mean in practical terms? Returning to the BMI data we can report that the average BMI (μ_X) is 26.46 and the average body fat percentage (μ_Y) is 35.31. If BMI and body fat were independent variables knowing BMI would tell us nothing about body fat and *vice versa*. But plotting the data (and some common sense) tells us that this is not the case; if we know one we can say quite a lot about the other. We could explore the correlation between the data (more about this later), but we can also describe these variables together using a joint distribution. By defining them using a joint distribution we are saying nothing about *cause and effect*, just that they are dependent variables.

```
options(repr.plot.width=4, repr.plot.height=3)

# BMI dataset

dat <- read.csv("Practicals/Datasets/BMI/MindsetMatters.csv")
head(dat)
#remove observations with no BMI data
dat <- dat[!is.na(dat$BMI),]
# scatter plot of BMI and body fat
ggplot(dat,aes(x=BMI,y=Fat)) + geom_point()

# report the mean of each variable
# note that some values of Y are missing...we need to add na.rm otherwise the estimate
will be NA
mux <- mean(dat$BMI)
print(paste0("value of mu_x is ",round(mux,2)))
muy <- mean(dat$Fat,na.rm=T)
print(paste0("value of mu_y is ",round(muy,2)))
```

	Cond	Age	Wt	Wt2	BMI	BMI2	Fat	Fat2	WHR	WHR2	Syst	Syst
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int
1	0	43	137	137.4	25.1	25.1	31.9	32.8	0.79	0.79	124	11
2	0	42	150	147.0	29.3	28.7	35.5	NA	0.81	0.81	119	11
3	0	41	124	124.8	26.9	27.0	35.1	NA	0.84	0.84	108	10
4	0	40	173	171.4	32.8	32.4	41.9	42.4	1.00	1.00	116	12
5	0	33	163	160.2	37.9	37.2	41.7	NA	0.86	0.84	113	11
6	0	24	90	91.8	16.5	16.8	NA	NA	0.73	0.73	NA	N

A data.frame: 6 × 14

```
Error in ggplot(dat, aes(x = BMI, y = Fat)): could not find function "ggplot"
Traceback:
```

So this joint distribution has a joint cdf, $F(x,y)$ and a continuous piecewise density function $f(x,y)$. The joint mean is defined as μ_x,μ_y What about the variance? Here we need to consider the variance and covaraince between X and Y.

```
# correlation between variables
dat2 <- dat[!is.na(dat$Fat),]
round(cov(x=cbind(dat2$BMI,dat2$Fat)),3)
paste0("variance of BMI = ",round(var(dat2$BMI),3))
paste0("variance of fat = ",round(var(dat2$Fat),3))
```

15.850	20.696
20.696	36.282

A matrix: 2 × 2 of
type dbl

'variance of BMI = 15.85'
'variance of fat = 36.282'

The *covariance matrix* is returned. The diagnoals return the variance of each parameter, and the off-diagnoals the covariance, indicating a positive correlation.

3.5.3 Correlation

Correlation and covariance are closely related. Pearson's correlation coefficient is defined as:
$$ \rho(X,Y) = Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} $$
So this helps us define BMI from body fat and *vice versa*. Examples of when this might be useful include;

- Inputing missing data
- Summarising many variables with one metric (more about this in the Machine learning module)
- Efficient sampling of distributions, which is used in Monte Carlo Markov Chain (MCMC) estimation

3.5.4 Connections to regression modelling

Later sessions exploring regression modelling will provide a powerful and flexible approach to exploring and quantifying *dependencies* between variables.

Statistical Inference

This section of the notes concerns a really important part of statistics: *statistical inference*.

Statistical analysis is often separated into two types: descriptive and inferential. Descriptive statistics attempt to describe the data at hand (the sample). Inferential statistics go further - they attempt to use the data at hand to make statements about a wider population.

There is more than one framework for statistical inference. The traditional and most widely used is the frequentist or "classical" approach. An important alternative, the Bayesian approach, is increasingly influential.

Overview of the statistical inference sessions

This section of the notes comprises 7 sessions:

- Population and samples
- Likelihood (x2)
- Frequentist inference (x2)
- Bayesian inference (x2)

The first session introduces the concept of **statistical inference**, defining populations, samples and estimators. The second half of the session introduces the idea of **sampling distributions**, a fundamental building block for frequentist inference. The sampling distribution gives us information about how different our estimate of the unknown population quantity of interest might have been, had we selected a different sample. In other words, the sampling distribution describes how our estimate behaves under **repeated sampling**. One particular feature of the sampling distribution, the **standard error**, gives us information about the amount we might expect our estimate to change if we took a different sample (i.e. it describes the variability of the estimate between different samples).

The idea of sampling distributions is crucial to frequentist inference, but while it gives us important information about how our estimate behaves under repeated sampling, it does not provide a recipe for choosing an estimator. **Maximum likelihood estimation** (MLE), the subject of the following two sessions, does exactly this. Given a statistical model for the data, MLE provides a method for choosing an estimator with desirable statistical properties.

Having explored MLE to obtain our estimator, we return to the idea of sampling distributions in the following two frequentist inference sessions. We see how the idea of sampling distributions allows us to create **confidence intervals**, which are ranges of values of the population quantity which we believe are consistent with the observed data. A complementary frequentist inference tool, hypothesis testing, allows us to assess the evidence against a **null hypothesis**, which proposes a specific value (or range of values) for the unknown population parameter.

Thus far, our attention has been largely on the frequentist paradigm. The last two sessions focus instead on an important alternative approach, **Bayesian inference**. In this paradigm we do not base our inference on the idea of repeated sampling. Instead, we use the likelihood to update prior information (in the form of a probability distribution) about the unknown parameter, to provide a **posterior distribution** for the unknown parameter. The posterior can be summarised by obtaining its mean, or a **credible interval** (interval within which the unknown parameter falls with a particular probability).

4. Populations and Samples

In this session we will begin thinking about statistical inference. Loosely, this describes the process of using a sample of data to make statements about a wider population. There is more than one framework for statistical inference. The traditional and most widely used approach is termed the "classical" or frequentist, and this is the one pursued in this and the next few sessions. An important alternative, the Bayesian approach, is increasingly influential. You will meet Bayesian inference later in this module.

Intended learning outcomes

By the end of this session you will be able to:

- describe the process of frequentist statistical inference
- define the terms population, sample, estimator and estimate
- explain what a sampling distribution is and how it relates to the idea of repeated sampling
- understand that sample estimates will vary as defined by the standard error
- appreciate that the uncertainty in estimates can be described using central limit theorem and resampling (bootstrapping)

The five sub-sections in this session explore the concept of sampling from a population and define parameters and estimators. They formally define the concept of a statistical model. Sampling distributions are described, and the standard error defined, along with key ways of obtaining the approximate sampling distribution.

4.1 Sampling from a population

Much of statistical inference is concerned with making statements about properties of populations, based on properties of samples from the populations. A helpful mental picture of the population and the sample to have in mind is as follows. Imagine that the **population** is a very large number of "objects" contained in a large urn, from which we can randomly sample a relatively small number of the "objects" at a time to provide our **sample**.

The objects in our population are often called **sampling units**. For many health research questions these sampling units are individual patients. If we were collecting information about different hospitals in order to make comparison between different providers, then the sampling unit might be hospitals.

Making statements about a population using information contained in a sample of data relies critically on the process of how the sample was drawn from the population (i.e. the sampling process). A common example of a sample process is random sampling. Under random sampling, each object in the population has the same chance of appearing in the selected sample. Inference procedures tend to be most straightforward in this setting.

In many cases, sampling is not random. For example, many populations have intrinsic structure that might facilitate sampling. If our population is people in rural Gambia, for example, then the easiest way to sample individual people might be to choose 5 villages and then go and survey the people in those villages. While it is not necessarily difficult to modify the process of statistical inference for such situations, statistically invalid conclusions can be reached if such modifications are not undertaken.

4.1.1 Parameters and estimators

In statistical inference, the aim is to make statements about certain features of the population, using the information contained in the sample data. Typically, we quantify the features of interest in terms of unknown population quantities (some examples might be a population mean, standard deviation, proportion, or risk ratio) and attempt to **estimate** these population quantities. We call these unknown population quantities population **parameters**. Parameters are typically denoted using Greek letters. Often, certain letters tend to be used for certain types of quantities. For instance, \(\mu\) will often denote a population mean and \(\pi\) will often denote a population proportion. When we are talking about a general "parameter of interest", we often use the letter \(\theta\).

A **statistic**, is any quantity that can be calculated from the known measurements on the sample data. It can be any function or combination of random variables that does not depend on unknown parameters for its calculation. As for population parameters, certain letters tend to be used for certain sample statistics. For instance, \(\bar{x}\) ("x bar") will often denote a sample mean and \(p\) a sample proportion.

We often want to estimate a population parameter from the sample. We do this by using sample statistics to estimate population parameters. For example, the obvious statistic to use to estimate a population mean is the sample mean. When a sample statistic is used for the purpose of estimating a population parameter it is known as an **estimator**. So an

estimator is a statistic that is designed to be a "guess" at a particular parameter of a population. When we use a sample statistic to esitmate a population parameter we use a "hat" to denote the estimator, e.g. $\hat{\mu}$ is an estimator for the population quantity μ and $\hat{\theta}$ is an estimator for the population quantity θ.

Once we have drawn our sample of data and calculated the value of the estimator in that sample, we refer to this as the **estimate**. In other words, the term estimate is used for the value obtained by substituting sample data values into the formula for the estimator.

The basic structure of frequentist inference can be represented diagrammatically as follows:

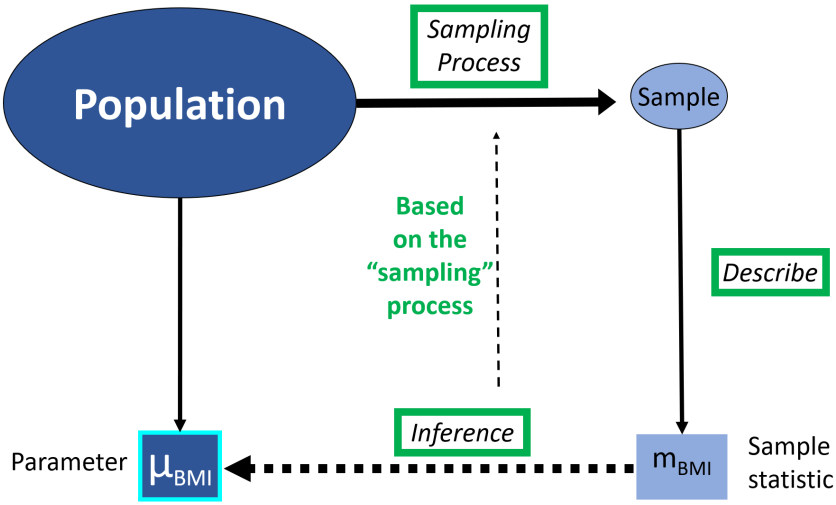


Fig. 2 Statistical inference

4.1.2 Example

To explore these issues further, we will consider a study question investigated by a recent MSc student at LSHTM as part of their summer project. The student explored the question of whether people who engage with victims of violence themselves suffer from emotional distress. This question was assessed using a sample of 53 violence researchers in Uganda. Subsequently, these violence researchers took part in a randomised trial, but we will focus on the initial description of the sample. The full article can be found here https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5455179/

For the purposes of illustration, in this session we will take a smaller sample of 10 violence researchers. Among our 10 sampled violence researchers, the sample mean age and the sample proportion suffering from emotional distress are

- Sample mean age $\bar{x}= 29.75$; sample standard deviation of age $SD = 4.49$
- Proportion suffering from emotional distress $p = 26\%$ (14 out of 53)

Let $X_1, ..., X_{10}$ be random variables representing the ages of 53 sampled researchers. In other words, $X_1, ..., X_{10}$ represent the random process by which the eventual 10 values of age are obtained. We call the realisation of these random variables (i.e. the observed data) $x_1, ..., x_{10}$.

We will initially focus on the population mean age (μ) and its estimator. The obvious estimator for the population mean age is the sample mean age. In terms of the general random variables $X_1, ..., X_{10}$, we can write this estimator as,
$$ \bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i $$
This is a random variable representing the mean of the random variables X_1,...,X_{10}. The sample statistic estimate is,
$$ \bar{x} = \frac{1}{10} \sum_{i=1}^n x_i = \mbox{sample mean age} = 29.75 $$
The estimate is the sample mean age, which is the realisation of $\bar{X}$ in the observed data. Since we are now viewing this as an estimate of μ, we can also write $\hat{\mu} = \bar{x}$.

> **Discussion question**
>
> - From above, our "best guess" at (our estimate of) the population mean age is 29.75.
> - Is this estimate a "good" estimate of the population mean? Do we think it is close to being correct?
> - If we sampled a different 10 researchers would we be likely to see a similar sample mean age? Or could we see a very different sample mean? Is it possible that, just by chance, this is a particularly old (or young) group of researchers?

4.2 Statistical models

To extract information about population quantities from sample statistics we need a precise and formal description of the whole sampling process from population to sample. This description is called the **statistical model**. Relevant features of the population are represented by parameters, such as the mean, variance, or correlation. The structure of the population, together with the sampling process, allows a model to be formulated that describes the statistical behaviour of the sample.

The crucial importance of the statistical model is that, given a certain value of the population parameter (in the simple case where there is only one parameter of interest), it allows us to calculate the probability of drawing a sample with the properties we observe: this will allow us to quantify the compatibility between the observed data and possible values of the population parameter.

4.2.1 Example: a statistical model

We will now write down a formal statistical model for the (sub-sample from the) emotional distress study. Remember that $X_1, ...,X_{10}$ are random variables representing the ages of 10 sampled researchers and $x_1, ..., x_{10}$ are the realised values of these random variables (i.e. the observed ages).

We will assume that each random variable is drawn from the same population distribution, and that the observations are independent of each other. We use the term **independent and identically distributed** as a succinct way of describing these assumptions, often abbreviated as **iid**.

Finally, we will assume that ages of violence researchers in the wider population follows a normal distribution with population mean μ and population variance σ^2.

This model can be compactly written as follows
$$ X_i \overset{\small{iid}}{\sim} N(\mu, \sigma^2), \qquad i=1,2,...,10 $$

4.3 Sampling distributions

In order to use our estimate (the value of the estimator in our sample of data) to make any sort of statement about the true but unknown value of the population parameter, we need to consider questions such as:

> How precise do we believe our estimate is?
>
> Are we fairly certain that the true parameter is close to the estimate, or do we believe the estimate may well be far from the true value?

The following thought experiment might help to develop these ideas. Suppose our population is a large bucket full of identical marbles. We want to know the population mean weight of a marble (our population parameter of interest). To estimate this population mean, we can simply sample a single marble from the bucket. So our estimator is the weight of the single sampled marble. Now suppose we took two samples: we sample a single marble, weigh it, put it back in the bucket, sample another marble and weight that one. In this case, our estimate (the weight of the sampled marble) would be exactly the same as the estimate from the first sample. No matter how many different samples we took, the sample estimate would be identical. In this case, because all possible samples would give us an identical estimate of the mean, we can confidently say what the population mean is using a single sample of one marble.

Now consider a bucket full of different marbles. In this case, randomly sampling a single marble and using the weight of that marble as an estimate of the population mean weight could give us a weight far too large (if we just happened to sample one of the very large marbles) or far too small (if we happened to pick a very small marble). However, if we were to pick 100 marbles and take the sample mean of those 100 marbles as our estimator, we would expect our estimate to

be closer to the population mean. If we were to resample another 100 marbles we would expect the sample mean weight to be fairly close to the mean weight of the previous 100 marbles. Conversely, if we took two samples containing one marble each, we might expect those two weights to be quite different from one-another.

This thought experiment makes it clear that in order to use our single sample of data to make statements about a wider population, we need to think about what would happen if we repeated our sampling: if we re-did our study many times, each time calculating the sample estimate, what values would those different sample estimates take? In fact, this is exactly what the **sampling distribution** is. It is the distribution of the **estimator** (the statistic we have chosen to use to estimate the population parameter of interest) under repeated sampling.

4.3.2 Simulated data: sampling distribution of a mean

We will return again to the emotional distress study. In reality, we do not know the true population mean and standard deviation. However, for the purposes of illustration, for the rest of the session we will imagine that we do know these values. Suppose that, in truth, the population mean age (\(\mu\)) is 30 and the population standard deviation (which will will call \(\sigma\)) is 4.8. Further, suppose that age follows a normal distribution in the population.

Under this scenario, the following code draws many (10,000) different samples from this population, with each sample containing the ages of 10 people. Note the line `set.seed(1042)` is coded to keep the same pseudo random number starting point.

```
# Population parameters
mu <- 30
sd <- 4.8
n_in_each_study <- 10

# Draw samples and ages for sampled individuals, for 100 different studies
# in this example we're going to have a list which generates study_measurements_age
repeatedly
different_studies <- 10000
set.seed(1042)
study_measurements_age <- list()
for (i in 1:different_studies) {
  study_measurements_age[[i]] <- round(rnorm(n_in_each_study, mu, sd),3)
}

# Print the sample data for two of the studies
print("Ages of the 10 participants selected in study 1:")
study_measurements_age[[1]]

print("Ages of the 10 participants selected in study 5:")
study_measurements_age[[5]]
```

```
[1] "Ages of the 10 participants selected in study 1:"
```

17.897 · 30.27 · 26.896 · 35.448 · 28.514 · 33.891 · 42.021 · 25.994 · 31.061 · 28.756

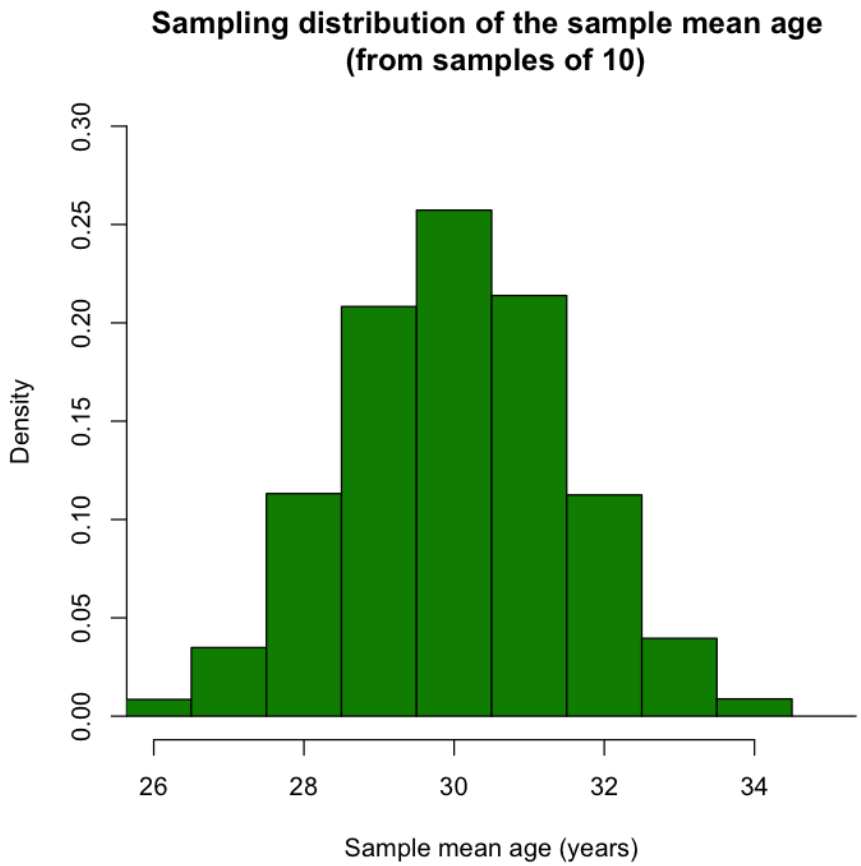
```
[1] "Ages of the 10 participants selected in study 5:"
```

28.502 · 33.725 · 35.155 · 26.544 · 31.147 · 31.732 · 39.582 · 31.223 · 25.802 · 16.168

Now we will calculate the sample mean for each sample of 10 people and plot them on a histogram. In the graphs below, we see a graph of sample means from 10,000 different studies (i.e. 10,000 different samples). This only gives us an approximation to the true sampling distribution, because it is based on a finite number of samples (10,000 samples). However, this is a large number so it will give us a fairly good approxiation to the sampling distribution of the sample mean.

For this estimator and this population, we can see that the sampling distribution follows a normal distribution. Note that the sampling distribution is centred around the true population value of 30. We also see that almost all sample means lie within 4 or so years of the mean either way (i.e. most sample means are between 26 and 34).

```
options(repr.plot.width=6, repr.plot.height=6)
sample.means   <- sapply(study_measurements_age, mean)

# Draw graphs using base R
hist(sample.means[1:10000], freq=FALSE,
     breaks=c(0, 25.5, 26.5, 27.5, 28.5, 29.5, 30.5, 31.5, 32.5, 33.5, 34.5, 100),
     xlim=c(26, 35), ylim=c(0, 0.3), col="green4",
     xlab="Sample mean age (years)",
     main="Sampling distribution of the sample mean age \n (from samples of 10)")
```



4.3.3 The standard error of an estimate

When we are talking about the sampling distribution (i.e. the distribution of an *estimator*), we call the standard deviation the **standard error**. The standard error refers to the variability we might expect in estimates of the parameter, because we are inferring the estimates from a sample. When we have two different estimators for the population parameter of interest, we would typically choose the one with the lower standard error.

The standard error
If an independently distributed random variable X has population mean (μ) and population variance (σ^2), the sampling distribution of sample means (of samples of size n) has population mean μ and population variance σ^2/n. This irrespective of the population distribution; it does not need to be *normal*. In other words the standard error is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

4.4 Obtaining the sampling distribution

The sampling distribution is hypothetical: in reality, we are not going to repeat our study many times to see how much our estimates differ from sample to sample.

In many cases we can describe the sampling distribution of our estimator well enough to do statistical inference, i.e. well enough to make useful statements about the population parameter. There are three main approaches to obtaining the (approximate) sampling distribution of an estimator:

1. **Algebraic calculation**. Sometimes we can algebraically obtain the distribution of the estimator from our statistical model. An example is given below for the sampling distribution for the sample mean age in the emotional distress example.
2. **The Central Limit Theorem**. If we have an estimator which can be written as the sum of independent random variables, then for large samples, the estimator will have an approximately normal distribution. This is described in more detail shortly.
3. **Resampling**. In many situations, we can use a resampling principle to obtain an approximate sampling distribution.

Returning to the question posed at the start of the previous section, those questions can be answered using the sampling distribution:

> How precise do we believe our estimate is?
>
> Are we fairly certain that the true parameter is close to the estimate, or do we believe the estimate may well be far from the true value?

The first question asks whether the sampling distribution is centred at the right value (i.e. the population parameter being estimated). If it is, we say the estimator is **unbiased**. In the epidemiology module and earlier in this module, you have already come across how the *sample* can be biased. Here, we are referring to whether the estimator is biased, which is sometimes referred to as *statistical bias*. Most estimators are unbiased, and this can be shown using statistical theory. For a small number of estimators it can be shown that they are in fact biased, and sometimes a correction can be applied to account for this. An example of exploring whether estimators are biased is given in the Appendix.

The second question will be examined when looking at the standard error and forms the basis for constructing 95% confidence intervals.

4.4.1 Algebraic calculation

We will illustrate the idea of obtaining a sampling distribution via algebraic calculation by revisiting the sub-sample from the emotional distress study.

For the moment, we will assume that in truth, ages follow a normal distribution with population mean $\mu=30$ and population standard deviation $\sigma=4.8$. Of course, in real life we would not have this information.

We now imagine that the population value of μ is unknown to the investigators undertaking the study; indeed, making inferences about μ is the aim of the study. We further imagine the rather unrealistic (but simplifying) situation that the investigators know the true value of the population standard deviation, $\sigma=4.8$.

Our model for the emotional distress study states that:
$$ X_i \overset{\small{iid}}{\sim} N(\mu, 4.8^2), \qquad i=1,2,...,10 $$
Under this statistical model, we want to know the distribution of our estimator for μ:
$$ \hat{\mu} = \frac{1}{10} \sum_{i=1}^n X_i $$
In this case, it's quite easy to derive the sampling distribution algebraically. We use the following fact:

> The mean of independent normally distributed variables also follows a normal distribution

It is then easy to calculate the expectation and variance of $\hat{\mu}$ using techniques covered in the maths refresher. So we know that the sampling distribution of $\hat{\mu}$ is:
$$ \hat{\mu} \sim N\left(\mu, 1.52^2 \right) $$
where the variance of this normal distribution was obtained as
$$ Var(\hat{\mu}) = \frac{1}{10^2} \times 10 \times Var(X_i) = \frac{4.8^2}{10}. $$
This gives us a lot of useful information about how the sampling distribution in relation to the unknown parameter μ:

- It follows a normal distribution (has a symmetric bell-shape)
- It is centred around the true (unknown) population value
- The standard error of the sample mean (the standard deviation of the estimator) is 1.52.

In many situations, this sort of algebraic calculation is possible. If not, we often rely on the central limit theorem to obtain the approximate sampling distribution in large samples.

4.4.2 The Central Limit Theorem

The Central Limit Theorem (CLT) is a key concept in statistics and in estimation. When we use the mean from a sample to estimate a parameter, we already acknowledge that there will be some error around this estimate, as described above. The CLT takes this further;

> **The Central Limit Theorem**

> If a random variable X has population mean μ and population variance σ^2 the sample mean $\bar{X}$, based on n observations, is *approximately* normally distributed with mean μ and variance σ^2/n, for sufficiently large n.

So even for situations where X follows a distribution that is not even close to being normal (e.g. X might be Poisson, or Binomial, or some wacky distribution), for sufficiently large samples the *mean* will follow a normal distribution. An example where X is a binary variable is given in the Appendix to this session.

This theorem is hugely powerful. We will see that this allows us to conduct large-sample inference fairly easily on any type of data.

4.4.3 Resampling

An alternative approach, which is computationally intensive but very flexible, is to use a resampling approach.

For a population of interest, we want to estimate a parameter θ using a sample S of n individuals (for our example $n=10$) from the population. We have an estimator $\hat{\theta}$ from our sample S. We want to know about the sampling distribution of the estimator $\hat{\theta}$.

We discussed above the idea that we could obtain the sampling distribution by repeatedly sampling from the population and calculating our estimate in each sample. Then a histogram of those many estimates would give us (approximately) the sampling distribution. In practice, it is logistically impossible to repeat the study a large number of times. However, we can mimic this process by using resampling.

The basic idea is to pretend that the observed data are the population and repeatedly sample from the data to learn about the relationship between $\hat{\theta}$ and the estimates obtained from the re-sampled data, which we will call $\hat{\theta}^*$.

Suppose we sample with replacement from the sample S to obtain "sub-samples" also of size n. These "sub-samples" are called **bootstrap samples**.

For example, suppose we have a sample S of size 10 ($n=10$):
$$ S = \{ x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} \} $$
And suppose our estimate is the sample mean,
$$ \hat{\theta} = \frac{(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10})}{10} $$
Then a bootstrap sample might be:
$$ S^*_1 = \{x_{10}, x_3, x_2, x_8, x_6, x_2, x_4, x_1, x_8, x_1 \} $$
Another bootstrap sample could be:
$$ S^*_2 = \{ x_5, x_9, x_4, x_7, x_{10}, x_9, x_3, x_4, x_6, x_2 \} $$
In each bootstrap sample, we obtain a new estimate (the sample mean in the bootstrap sample):
$$ \hat{\theta}^*_1 = \frac{(x_{10} + x_3 + x_2 + x_8 + x_6 + x_2 + x_4 + x_1 + x_8 + x_1)}{10} $$
and
$$ \hat{\theta}^*_2 = \frac{(x_5 + x_9 + x_4 + x_7 + x_{10} + x_9 + x_3 + x_4 + x_6 + x_2)}{10} $$
We do this a very large number of times to obtain lots of estimates from different bootstrap samples. Then we can draw a histogram of these many bootstrap estimates to see the shape and dispersion of the distribution.

The **bootstrap principle** says that the distribution of $\hat{\theta}$ given θ (i.e. the sampling distribution) is approximated by the distribution of $\hat{\theta}^*$ given $\hat{\theta}$. For example, if we find that our values of $\hat{\theta}^*$ are approximately normally distributed and centred around $\hat{\theta}$ then the bootstrap principle tells us that $\hat{\theta}$ follows a normal distribution centred around θ.

4.4.3.1 Example: resampling

We illustrate the idea of resampling using the sub-sample of the emotional distress study. Suppose our data - the 10 sampled ages - are the set: $\{ 28.1, 27.5, 25, 29.9, 29.7, 29.9, 39.9, 33.6, 21.3, 30.8 \}$. Our estimate of the population age is the sample mean age, which is: $\hat{\mu} = 29.57$.

To obtain an approximation to the sampling distribution for the sample mean age, using a resampling approach, we first take a large number of bootstrap samples from the data. The code below does this.

```r
# Our sample of data (ages for 10 sampled researchers)
ages <- c(28.1,27.5,25,29.9,29.7,29.9,39.9,33.6,21.3,30.8)

# Randomly select 10,000 bootstrap samples (each of size 10)
set.seed(532)
bootstrap_samples <- lapply(1:10000, function(i) sample(ages, replace = T))

# List some of the bootstrap samples
print("First bootstrap sample:")
bootstrap_samples[1]
print("Third bootstrap sample:")
bootstrap_samples[3]
print("500th bootstrap sample:")
bootstrap_samples[500]
```

```
[1] "First bootstrap sample:"
```

1. 27.5 · 29.9 · 28.1 · 27.5 · 29.9 · 21.3 · 27.5 · 28.1 · 29.9 · 30.8

```
[1] "Third bootstrap sample:"
```

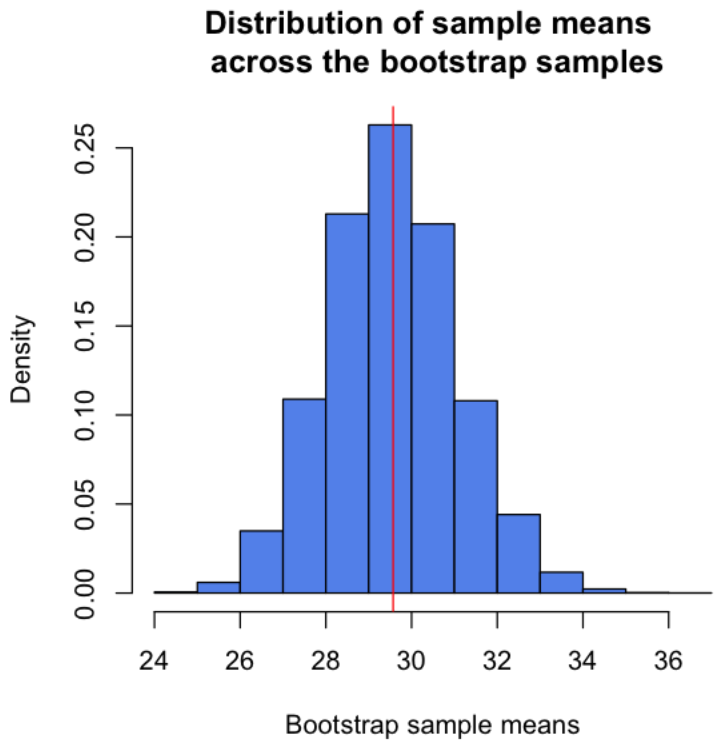
1. 30.8 · 28.1 · 21.3 · 29.9 · 39.9 · 21.3 · 27.5 · 29.9 · 29.9 · 29.7

```
[1] "500th bootstrap sample:"
```

1. 21.3 · 29.9 · 39.9 · 29.9 · 29.9 · 27.5 · 29.9 · 29.9 · 29.9 · 21.3

The next step is to calculate the estimate (the sample mean, in our case) in each bootstrap sample. These estimates are called $\hat{\mu}^*_1, \hat{\mu}^*_2, .., \hat{\mu}^*_{10,000}$. Then we can plot the histogram of all the estimates across the bootstrap samples.

```r
# Calculate the sample mean in each of the bootstrap samples
r.mean <- sapply(bootstrap_samples, mean)

# Draw a histogram with a red vertical line indicating the original sample mean age
options(repr.plot.width=5, repr.plot.height=5)
hist(r.mean, freq=FALSE, main="Distribution of sample means \n across the bootstrap
samples",
     xlab="Bootstrap sample means",col="cornflowerblue")
abline(v=mean(ages),col="red")
```



We see a number of features from the graph above;

- The histogram follows an approximately normal distribution (has a symmetric bell-shape)
- It is centred around the sample mean age (from the original sample, $\hat{\mu} = 29.57$)
- The code below tells us that the standard deviation of this distribution is 1.51.

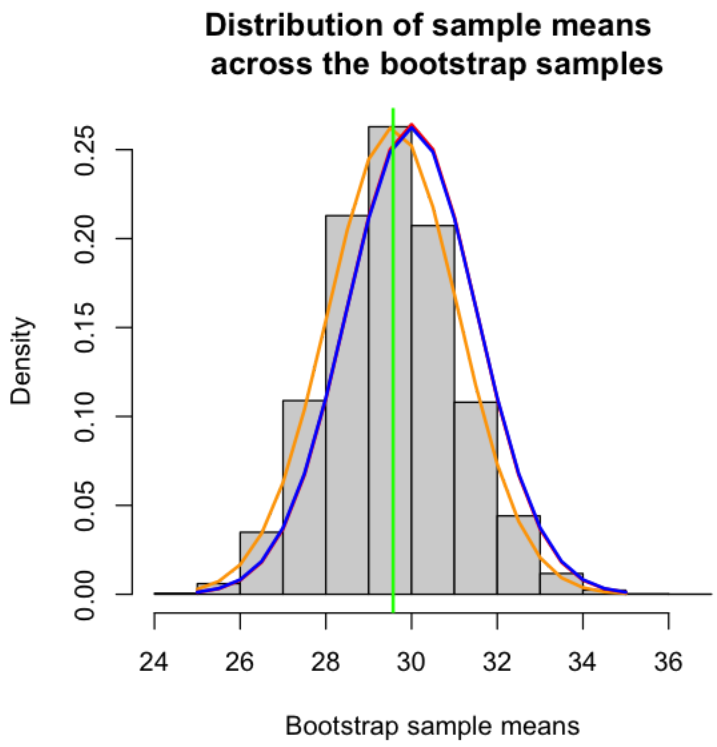
```
sqrt(var(r.mean))
```

1.50868637530863

So we have seen that the bootstrap approximation of the distribution of $\hat{\mu}^*$ given $\hat{\mu}$ is a normal distribution centred around $\hat{\mu}$ with standard deviation of 1.51. The bootstrap principle tells us that the distribution of $\hat{\mu}$ given μ is approximately the same. In other words, approximately:
$$ \hat{\mu} \sim N(\mu, 1.51^2) $$
Remember, that we obtained the true distribution algebraically above and found that
$$ \hat{\mu} \sim N(\mu, 1.52^2) $$
So the resampling (bootstrap) approach has given us a very good approximation to the true sampling distribution. The code below redraws the histogram above, with the approximate (bootstrap) sampling distribution and the algebraically-calculated one.

We see that the bootstrap sampling distribution (shown in red) is simply a shift of the normal distribution which best follows the histogram (shown in orange), and that the bootstrap and true (algebraic, shown in blue) sampling distributions are very similar.

```
# Histogram of estimates (sample means) in bootstrap samples
options(repr.plot.width=5, repr.plot.height=5)
hist(r.mean, freq=FALSE, main="Distribution of sample means \n across the bootstrap
samples", xlab="Bootstrap sample means")

# Add the normal distribution which most closely follows the histogram
lines(seq(25, 35, by=.5), dnorm(seq(25, 35, by=.5), mean(ages), 1.52),
col="orange",lwd=2)

# Add the bootstrap approximation to the sampling distribution: normal distribution
with mean mu=30 SD=1.51
lines(seq(25, 35, by=.5), dnorm(seq(25, 35, by=.5), 30, 1.51), col="red",lwd=2)

# Add the algebraic sampling distribution: normal distribution with mean mu=30 SD=1.52
lines(seq(25, 35, by=.5), dnorm(seq(25, 35, by=.5), 30, 1.52), col="blue",lwd=2)

# Add a vertical line at original sample mean
abline(v=mean(ages),col="green",lwd=2)
```



4.4.5 What do we use a sampling distribution for?

In practice, we have a single sample of data and a single estimate from that sample of the population parameter of interest. When we present our single estimate of a population quantity, we need to be able to say something about how precise it is. Is it likely to be close to the true value? Can we provide a range of values within which we believe the true value lies?

We can only answer these questions, within the framework of frequentist statistical inference, by thinking about what estimates we might have got had we chosen a different sample. This leads us to the sampling distribution - the distribution of the estimator across samples.

In subsequent sessions we will see how the sampling distribution allows us to

- construct confidence intervals for population parameters (intervals within which we believe the true value is likely to lie)
- conduct hypothesis tests for population parameters

4.5 Summary

- In this session we have introduced the concept of populations and samples, and how we use statistical inference to make statements about the unknown population parameters.
- These unknown popultion parameters are estimated by sample statistics. Variability in the sample and any associated statistics are to be expected.
- The variability in the sample statistics can be quantified by estimating the standard error.

Appendix: Additional Reading

This appendix section contains additional information which will deepen your understanding. However, it is not examinable and is completely optional reading.

A1 More on populations

There are additional issues related to the definition of the population, that should be considered.

- Is the population well defined?

Loosely speaking, we think about the population as being the wider group (often of people or patients) who we can generalise the results to. For some research questions the population of interest is well defined. For instance, suppose we undertake a study where we are attempting to estimate the proportion of adults (18 years and above) in the UK with hypertension in 2020. The population is well defined. Conversely, suppose we undertake a study to estimate the effect of a blood-pressure-lowering treatment among a sample of 50 patients in the UK in 2020. In this case, the population of interest can be difficult to pin down. Who can we generalize our results to? Is the population restricted in time and space? Can we generalise to patients in other countries? Can we generalise to future patients?

- Is the sample representative of the population?

Clearly a sample can be chosen in may ways, and the way in which we are able to make inferences about the population depends critically on the way in which the sample is selected: it is hard to over-emphasize the importance and relevance of the sampling process to the meaning and validity of the subsequent inferences. In this module, we will assume that sampling units (in this case, people) are randomly sampled from the population.

- Is the population finite, or (effectively or potentially) infinite?

For example, a study of a new treatment for a disease may wish to generalise to all potential patients.

- Have we sampled all the population?

For example, a study of leukemia in the years following a leak from a nuclear power station may sample all subjects developing leukemia within the relevant time period in the vicinity of the power station. In such an example it is not clear how to define a wider population from which the sample can be considered to have been drawn. In these and other cases one approach is to consider a notional or counterfactual population, which can only have a conceptual existence.

In general the issues can be complex and will not be considered further here.

A2 Bias of estimators

Using statistical theory it is possible to show that the sample mean, $\bar{X}$, is an unbiased estimator of the population mean, μ. One of the simplest examples is when our random variables follow the *Bernoulli* distribution.

Example Let $X_1, X_2,.., X_n$ be Bernoulli trials with success parameter p.

Our estimate of p is the sample mean,
$$ \hat{p} = \bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} $$
We will now show that the expected value of this estimator is equal to the population mean, p.
$$\begin{split} \begin{align*} E(\bar{X}) & = E\left[\frac{X_1 + X_2 + ... + X_n}{n}\right] \\ & = \frac{1}{n} E[X_1 + X_2 + ... + X_n] \qquad \mbox{(we can take constants out of expectations)} \\ &= \frac{1}{n} (E(X_1) + E(X_2) + ... + E(X_n)) \\ &= \frac{1}{n} (p + p + ... + p) = p \\ \end{align*} \end{split}$$
This simple use of algebra illustrates that $\bar{X}$ is an unbiased estimator for p.

Exercise: You can use similar logic to demonstrate that the sample mean is an unbiased estimator for the population mean when the random variables X follow a normal distribution with known variance.

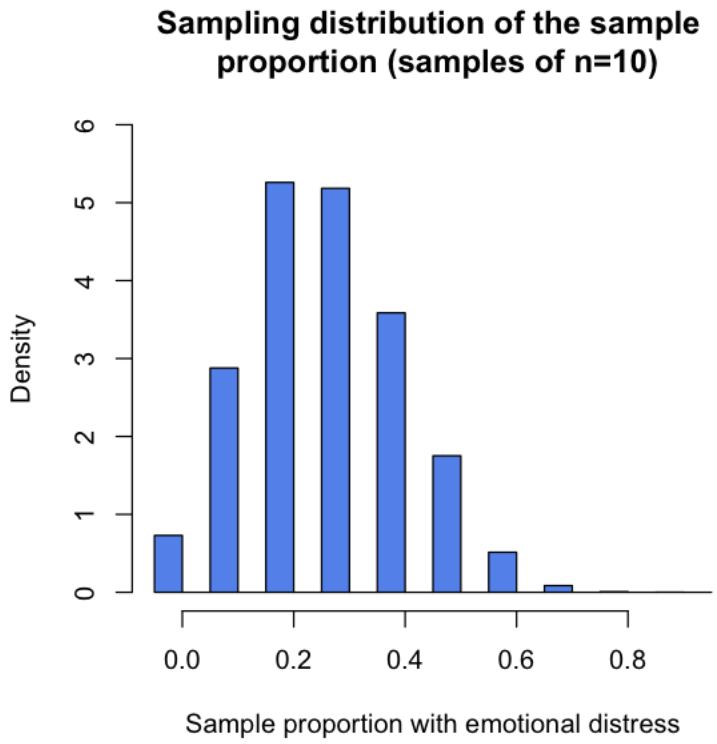
A3 CLT for binary data

We will return to the emotional distress study again, using the sub-sample of 10 people, but this time measure a binary characteristic for each person - the presence of emotional distress.

We suppose that, in the population, the true proportion is 28%. Under this assumption, we can simulate (draw) different samples, each containing 10 people. If we do this a very large number of times, say 10,000, then the distribution of the different sample proportions we obtain will give us a very good picture of the true sampling distribution of the proportion. (Of course, remember that in practice we cannot do this because we won't know the true population proportion).

The code below obtains the approximate sampling distribution.

```
# Population parameters
pi <- 0.28
n_in_study <- 10

# Simulate data from multiple studies
different_studies <- 10000
set.seed(1042)
study_measurements_ed <- list()
for (i in 1:different_studies) {
  study_measurements_ed[[i]] <- rbinom(n_in_study, 1, pi)
}

# Calculate the proportion in each study
sample.props   <- sapply(study_measurements_ed, mean)

# Draw graphs
options(repr.plot.width=5, repr.plot.height=5)
hist(sample.props[1:10000],
     freq=FALSE, breaks=seq(-0.05, 0.95, 0.05), col="cornflowerblue",
     ylim=c(0, 6), xlab="Sample proportion with emotional distress",
     main="Sampling distribution of the sample \n proportion (samples of n=10)")  # the
"\n" makes a newline
```



The graph above shows us a reasonably accurate picture of the sampling distribution. Unlike the sample mean, the sampling distribution of the sample proportion is not quite symmetric. It is also not continuous - the sample statistic can only take 10 different values, with a sample size of \(n=10\).

Below, the code shows that the mean of the sample means is approximately 0.28. (The discrepancy is just random error due to the fact that our "sampling distribution" does not come from an infinite number of samples. If we simulated a sufficiently large number of samples, this number would become closer to the true value of 0.28.)

The final line of code below lists the (approximate) probability density function, which gives us the whole sampling distribution for the sample proportion in this example.

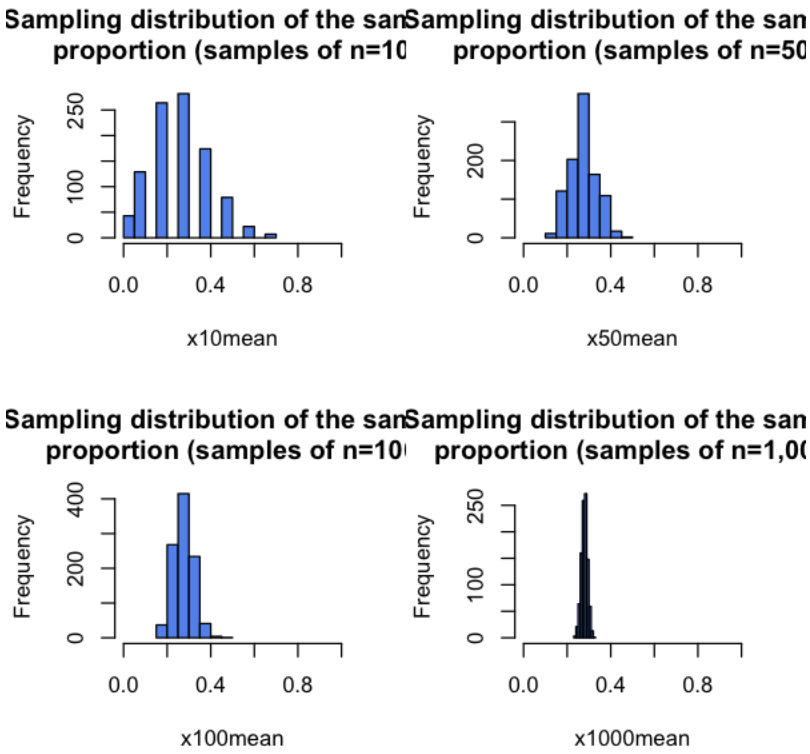
```
### Summarise the approximate sampling distribution

# The mean value of the different sample means
mean(sample.props)

# The whole sampling distribution (i.e. the PDF)
table(sample.props)/different_studies
```

0.27911

```
sample.props
     0    0.1    0.2    0.3    0.4    0.5    0.6    0.7    0.8    0.9
0.0365 0.1439 0.2630 0.2592 0.1793 0.0876 0.0257 0.0043 0.0004 0.0001
```

Now we will explore what happens to the sampling distribution as we take larger samples. So instead of 10 people per sample, suppose we had 100 or 10,000 people in each sample. The central limit theorem tells us we expect the distribution to become more normal as we increase the sample size.

```r
x10n <- sapply(rep(10,1000),function(x) rbinom(x, 1, p = 0.28))
x10mean <- colMeans(x10n)

x50n <- sapply(rep(50,1000),function(x) rbinom(x, 1, p = 0.28))
x50mean <- colMeans(x50n)

x100n <- sapply(rep(100,1000),function(x) rbinom(x, 1, p = 0.28))
x100mean <- colMeans(x100n)

x1000n <- sapply(rep(1000,1000),function(x) rbinom(x, 1, p = 0.28))
x1000mean <- colMeans(x1000n)

par(mfrow=c(2,2))
hist(x10mean,col="cornflowerblue",xlim=c(0,1), main="Sampling distribution of the
sample \n proportion (samples of n=10)")
hist(x50mean,col="cornflowerblue",xlim=c(0,1), main="Sampling distribution of the
sample \n proportion (samples of n=50)")
hist(x100mean,col="cornflowerblue",xlim=c(0,1), main="Sampling distribution of the
sample \n proportion (samples of n=100)")
hist(x1000mean,col="cornflowerblue",xlim=c(0,1), main="Sampling distribution of the
sample \n proportion (samples of n=1,000)")
```



As predicted by the CLT, even though the original data are binary (each person has emotional distress or does not), the sample proportion (which is the mean of the binary outcomes) has a distribution which becomes approximately normal with sufficiently large samples.

5. Likelihood

In statistical inference, our task is to make statements about the underlying parameter(s) of our proposed model given the observed data. In particular, we typically wish to obtain the best estimate of the unknown parameters. We also wish to know how well we have estimated the unknown parameter(s). The concept of likelihood provides the best single framework for this task. We will see that the likelihood function, often simply called the likelihood, plays a fundamental role in both frequentist and Bayesian inference.

Intended learning outcomes

By the end of this session you will be able to:

- explain the concepts of likelihood and maximum likelihood estimation
- derive a likelihood in a simple situation
- explain the connection between maximising the likelihood and maximising the log-likelihood
- describe and apply the process of obtaining a maximum likelihood estimator

The next sub-sections introduce the idea of maximum likelihood estimation, define the likelihood and log-likelihood functions and illustrate the process of obtaining the maximum likelihood estimator.

5.1 Maximum likelihood estimation

Suppose we are interested in the probability that a single patient will experience a particular side effect from a particular drug. We decide to run a small clinical study including 8 patients. The observed data consist of the number, of those 8 patients, who experience a side effect. Suppose that we conduct the study and observe that 2 patients experience a side effect. We wish to use these observed data to make statements - inferences - about the unknown probability of experiencing a side effect from that drug.

Statistical model: We begin by defining a model for the data. Here, we define X as the random variable representing the total number of the 8 patients who experience a side effect. Our model is that
$$ X \sim binomial(8, \pi) $$
which - we remember from the probability sessions - involves the assumptions that each Bernoulli event (whether or not each individual patient experiences a side effect) is independent and has the same probability of occurring.

This model involves the unknown parameter π.

Data: We have observed a realisation from this model, $X=2$. These are often called our observed data.

Under our proposed statistical model, the probability that 2 out of 8 patients experience a side effect is:
$$ P(X=2) = {8 \choose 2} \pi^2 (1-\pi)^6 $$
Since π is unknown, it is natural to consider how the probability of observing these data varies with different values of π:

π	P(X)=(2)
0	0
0.25	0.311
0.5	0.109
0.75	0.004
1	0

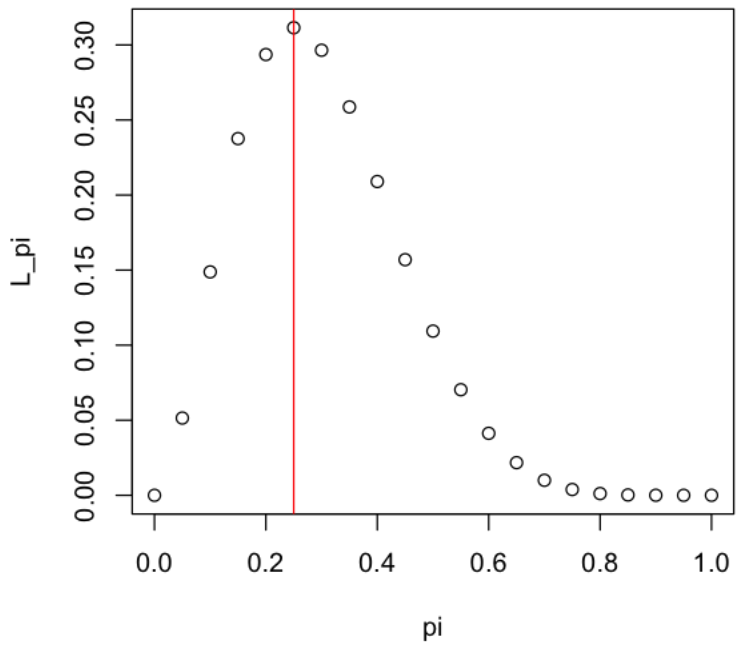
Suppose that, in truth, the unknown probability of a patient experiencing a side effect from this drug was 0.75. The probability of then observing 2 from 8 patients experiencing a side effect is 0.004. This is a very low probability, so this would be an unusual or perhaps unexpected event, although not strictly impossible.

Suppose that, conversely, the unknown probability of a patient experiencing a side effect from this drug was actually 0.25. Then the probability of observing 2 from 8 patients experiencing a side effect would be 0.31 (31%). If this were the case, there would be nothing unusual or unexpected about our observed data.

We do not know which value of π is the true value. But a sensible strategy to obtain a 'best guess', or estimate, of π, might be to pick the value which maximises the probability of observing the data that we observed. We will see below that this probability is in fact the likelihood, leading to the concept of maximising the likelihood or maximum likelihood. This is a term that you will encounter frequently in statistics.

Following these ideas, we can extend the table above by considering a finer range of possible values for π between 0 and 1, and plot the probability of observing $X=2$, assuming that that value of π were true. This gives the graph below.

```r
# Define a range of values for pi
pi = seq(0,1,by = 0.05)

# Calculate the likelihood for each value, given n=8 and x=2
L_pi <- choose(8,2)*pi^2*(1-pi)^(8-2)

# Plot the output
options(repr.plot.width=5, repr.plot.height=5)
plot(x = pi, y = L_pi)

# Add a line to indicate the value which yields the highest likelihood
abline(v = pi[which.max(L_pi)], col = "red")
```



We see that $\pi=0.25$ is the value that leads to the highest probability of observing the data that we did indeed observe (i.e $X=2$) so we choose this value as our best guess for π. We will see that this value is called the maximum likelihood estimator. We write $\hat{\pi} = 0.25$, where we have added a hat to indicate that this is being viewed as an estimate of an unknown parameter.

The likelihood when $\pi = 0$ is exactly zero, as is the likelihood when $\pi = 1$. This makes sense because these two probabilities would make the observed data impossible - they imply that patients would either *never* or *always* experience side effects. Informally, we could say that these values are *inconsistent* with the data.

Note that, our estimate of the probability of a patient experiencing a side effect is intuitively a sensible one: it is the sample proportion, $\frac{2}{8}$.

We will see later on that estimators obtained in this way (by maximising a likelihood) have very nice statistical properties.

5.2 The likelihood

The function that we maximised above to find our estimate for the unknown parameter π took the same algebraic appearance as the probability distribution function, evaluated at the value of the observed data. We will see below that this function is called the likelihood. The likelihood looks like a probability distribution function. It has a probabilistic interpretation for any particular value of π: it's the probability of seeing the observed data assuming that is the true value of π. However, in contrast to the probability distribution function, which is a function of x and sums to 1 over all possible values of x, the likelihood function is a function of π. So, for example, this does not sum to 1 over all possible values of π.

A general definition of the likelihood is as follows.

For a probability model with parameter θ, the likelihood of the parameter θ given the observed data x is defined as
$$ L(\theta | x) = P(x | \theta) $$
On the right hand side of this equation:

- This is either a probability distribution function or a density function
- If our distribution is discrete, as above, this is: $P(x | \theta) = P(X=x)$
- If our distribution is continuous, this becomes: $P(x | \theta) = f(x)$
- $P(x | \theta)$ is a probability statement. It is the probability of seeing the observed data, under the assumed model, assuming that the true parameter value is equal to θ.

And on the left hand side of this equation:

- $L(\theta | x)$ is the likelihood function, often just called the likelihood.

In an informal sense the likelihood conveys the *consistency* of different values of the parameter with the observed data.

We often just write the likelihood as $L(\theta)$. The additional notation (writing "$| x$") is merely to remind ourselves that the likelihood function involves the observed data, but it is not a function of these: x is treated as a fixed quantity in the likelihood.

5.2.1 Example: the Binomial model

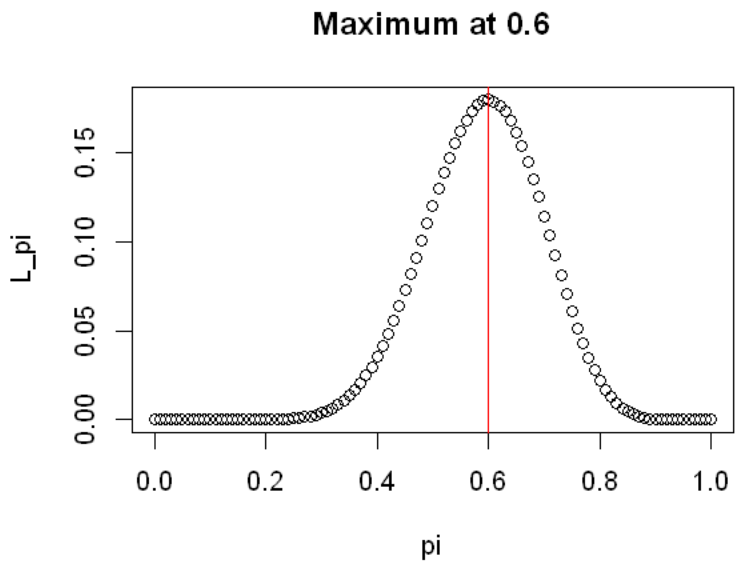
Consider a diabetes clinic at which patients present following initial diagnosis. The first line of intervention for diabetes is lifestyle change, and the clinician wants to determine what proportion of patients will respond to this intervention. She decides to conduct a study by following up twenty patients who present to the clinic in one day.

Statistical model: We assume that a binomial model is appropriate for the number of patients who will respond to lifestyle changes out of the twenty patients in total.
$$ X \sim binomial(20, \pi) $$
Data: Out of the twenty patients sampled, she found that twelve of them had responded well after six weeks of recommended lifestyle changes. Our observed data are $x = 12$.

Probability distribution function: As we described before, the likelihood of π given these data is the probability of observing the data for different values for π. Remember the probability distribution function for a binomial distribution of size 20 is
$$ P(X = x|\pi) = {20 \choose x} \pi^{x} (1-\pi)^{20 - x} $$
for a given value of π.

Likelihood: The likelihood has this same form but is viewed as a function of π, rather than a function of x. For our observed data of 12 out of 20 patients,
$$ L(\pi | x = 12) = {20 \choose 12} \pi^{20} (1-\pi)^{20 - 12} $$
As before, we can identify the value of π which gives the maximum likelihood by plotting the likelihood for a range of values of π.

```
options(repr.plot.width=5, repr.plot.height=4)

# Define a range of values for pi
pi = seq(0,1,by = 0.01)

# Calculate the likelihood for each value, this time given n=20 and x=12
L_pi <- choose(20,12)*pi^12*(1-pi)^(20-12)

# Plot the output
plot(x = pi, y = L_pi)

# Find the value of pi for which L_pi is highest
pi_max <- pi[which.max(L_pi)]

# Add a line to the plot at pi_max
abline(v = pi_max, col = "red")

# Add a title specifying the value of pi_max
title(paste("Maximum at", pi_max))
```



The value which maximises this function is 0.6, the observed sample proportion; we'll call this value $\hat{\pi}$ to indicate that it is an estimate of π. Notice that the likelihood for values of π smaller than 0.3 or greater than 0.9 is very small - much smaller than that of values around 0.6 - suggesting that these values are inconsistent with the observed data.

5.2.2 Example: the Exponential model

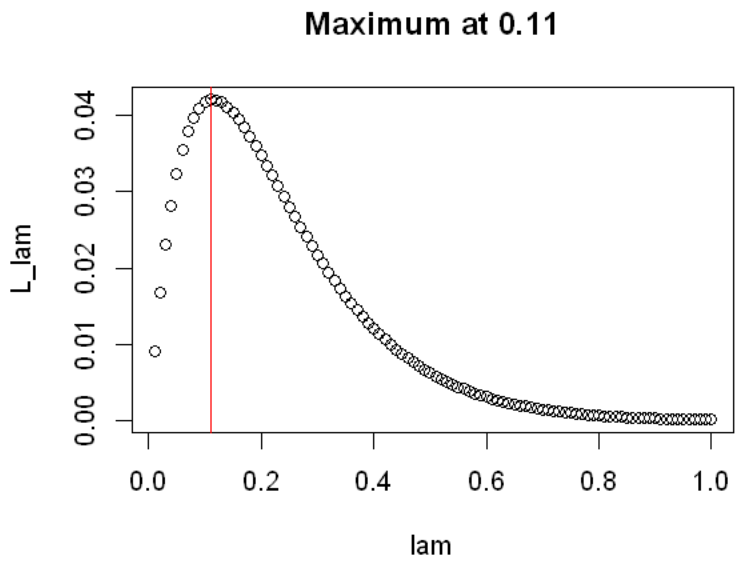
Suppose we wish to estimate how long patients usually wait in reception before their GP appointment. At one practice, a patient walks through the door and the receptionist records the time until they get called through.

Statistical model: The waiting time in minutes, Y, is a continuous random variable which must be non-negative. It is common to use an exponential distribution to model waiting times, so we will assume it's a reasonable choice for this example.
$$ Y \sim Exp(\lambda) $$
Remember that the mean of this distribution is equal to one over the rate parameter (λ), i.e. $E(Y) = \frac{1}{\lambda}$.

Data: The receptionist observes that the patient waits for eight minutes and forty-five seconds, so $y = 8.75$.

Probability density function: The PDF for an exponential distribution is
$$ f_Y(y|\lambda) = \lambda e^{-y\lambda} $$
Likelihood: We write down the likelihood for λ based on the exponential PDF above.
$$ L(\lambda | y = 8.75) = \lambda e^{-8.75\lambda} $$

```
options(repr.plot.width=5, repr.plot.height=4)

# Define a range of values for lambda, equating to mean waiting times from 1 to 100
minutes
lam = seq(0.01,1,by = 0.01)

# Calculate the likelihood for each value, given y=8.75
L_lam <- lam*exp(-8.75*lam)

# Find the value of lambda for which L_lam is highest
lam_max <- lam[which.max(L_lam)]

# Plot the likelihood and indicate the maximum value
plot(x = lam, y = L_lam)
abline(v = lam_max, col = "red")
title(paste("Maximum at", round(lam_max,2)))
```



If we evaluate over a fine enough range of values for \(\lambda\), we find that the value which maximises this exponential likelihood is equal to \(\frac{1}{8.75}\), i.e. one over the observed waiting time. This defines an exponential distribution with mean equal to the observed waiting time.

As with the binomial example, the estimate obtained by maximising the likelihood is intuitively sensible based on the data we've observed.

5.3 Log likelihood

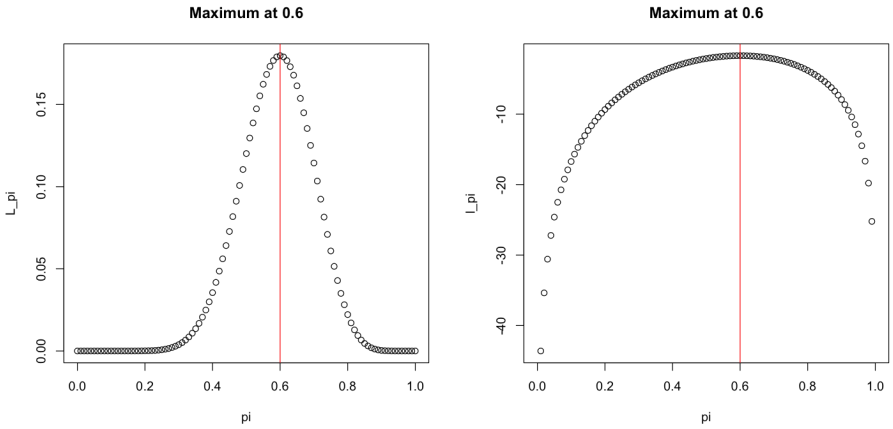
We have discussed the idea that finding the maximum value of a likelihood gives us sensible estimates for the unknown parameters. For the examples above it is relatively clear from calculating a few values of the likelihood where the maximum lies, but this will not always be the case.

A theoretical result which will come in handy is that a value which maximises the likelihood also maximises the log-transform of the likelihood, or the *log-likelihood*. This is because the log is a *concave* function, so when we use it to transform the likelihood, any maximum or minimum stays in the same place on the x-axis. We will denote the log-likelihood by \(l(\theta) = \log(L(\theta))\).

This result is evident when plotting the transformation of the likelihoods for the binomial and exponential distributions.

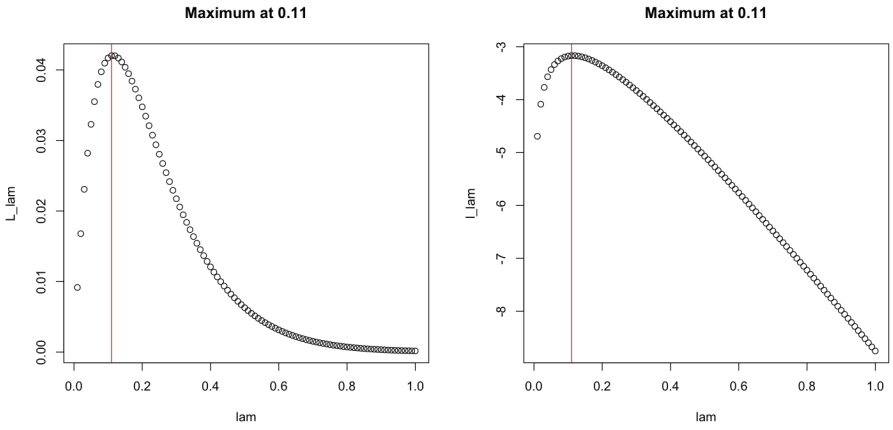
For the binomial example:

```
options(repr.plot.width=12, repr.plot.height=6)
par(mfrow = c(1,2))

# likelihood L(pi)
pi = seq(0,1,by = 0.01)
L_pi <- choose(20,12)*pi^12*(1-pi)^(20-12)
pi_max <- pi[which.max(L_pi)]

plot(x = pi, y = L_pi)
abline(v = pi_max, col = "red")
title(paste("Maximum at", pi_max))

# log-likelihood l(pi)
l_pi <- log(L_pi)

plot(x = pi, y = l_pi)
abline(v = pi[which.max(l_pi)], col = "red")
title(paste("Maximum at", pi[which.max(l_pi)]))
```



For the exponential example:

```
options(repr.plot.width=12, repr.plot.height=6)
par(mfrow = c(1,2))

# likelihood L(beta)
lam = seq(0.01,1,by = 0.01)
L_lam <- lam*exp(-8.75*lam)
lam_max <- lam[which.max(L_lam)]

plot(x = lam, y = L_lam)
abline(v = lam_max, col = "red")
title(paste("Maximum at", round(lam_max,2)))

# log-likelihood l(beta)
l_lam <- log(L_lam)

plot(x = lam, y = l_lam)
abline(v = lam[which.max(l_lam)], col = "red")
title(paste("Maximum at", round(lam[which.max(l_lam)],2)))
```



5.3.1 Why use the log likelihood?

Log-transformed likelihoods are generally “better-behaved” and easier to work with than the original form. Remember the rules of logs for products and powers - we’ll see in the next section how these make computation with the log-likelihood very convenient.

5.4 Finding the MLE

So far we have obtained the maximum likelihood estimate (MLE) by plotting the likelihood for different parameter values and looking for the value which yields the maximum. Of course, the estimate obtained in this way depends on how many parameter values we evaluate.

A more formal way is to determine the location of that maximal point algebraically, from the likelihood function itself. In this way, we can directly obtain the general form for the MLE in terms of the data.

This is where the log-likelihood comes into its own; we know that a value which maximises the log-likelihood also maximises the likelihood, and the impact of logs on products and powers make the algebra much simpler.

We find the maximum likelihood estimator of a parameter from the log-likelihood function through the following steps:

Method for finding MLEs:

1. Obtain the derivative of the log-likelihood: $\frac{d l(\theta \mid {x})}{d \theta}$
2. Set $\frac{d l(\theta \mid {x})}{d \theta}=0$ and solve for θ
3. Verify that it is a maximum by showing that the second derivative $\frac{d ^2 l(\theta \mid {x})}{d \theta ^2 }$ is negative when the MLE is substituted for θ.

5.4.1 Binomial model

We will derive the MLE for the binomial example described earlier. In general, the likelihood given observed data of x responders out of n patients is
$$ L(\pi|x) = {n \choose x} \pi^{x} (1-\pi)^{n - x} $$
and so the log-likelihood is
$$\begin{split} \begin{align*} l(\pi|x) & = \log\left({n \choose x} \pi^{x} (1-\pi)^{n - x}\right) \\ & = \log {n \choose x} + x \log \pi + (n-x)\log (1-\pi) \end{align*} \end{split}$$
We can now obtain the maximum likelihood estimate from this function.

Step 1: Differentiate the log-likelihood with respect to our parameter π
$$ \begin{equation} \frac{d l \left(\pi \mid x \right)}{d \pi} = \frac{x}{\pi} -\frac{(n-x)}{(1-\pi)} \end{equation} $$
Step 2: We set the derivative equal to zero and solve for π
$$\begin{split} \begin{align*} 0 &= \frac{x}{\hat{\pi}} -\frac{(n-x)}{(1-\hat{\pi})} \\ \frac{(n-x)}{(1-\hat{\pi})} &= \frac{x}{\hat{\pi}} \\ (n-x)\hat{\pi} &= x(1-\hat{\pi}) \\ n\hat{\pi}-x\hat{\pi} &= x-x\hat{\pi} \\ \hat{\pi} &= \frac{x}{n} \\ \end{align*} \end{split}$$
Having solved the equation, we get that the maximum likelihood estimator for π is $\hat{\pi} = \frac{x}{n}$ (note that we put a hat to indicate that it is an estimator).

There is one thing left for us to check: we have found that $\frac{x}{n}$ is the point where the derivative of the log-likelihood is zero, but that could mean that it is a maximum or a minimum of the log-likelihood function. To verify that this is indeed a maximum, we need to compute the second derivative of the log-likelihood and check that it takes a negative value when $({\pi} = \frac{x}{n}$.

Step 3: Find the second derivative:
$$ \begin{equation} \frac{d l^2 \left(\pi \mid x \right)}{d \pi ^2} = -\frac{x}{\pi^2} -\frac{(n-x)}{(1-\pi)^2} \end{equation} $$
This second derivative must be negative if we plug in $\frac{x}{n}$ for $({\pi}$. Both fractions on the right hand side have a squared number in the denominator which is positive, so we only have to think about the numerators. We have that $-x \leq 0 $ since $x \geq 0$ and also $-(n-x) \leq 0 $ since $n \geq x \geq 0$. Therefore, the value $\frac{x}{n}$ is indeed a maximum.

Thus the MLE for π is $\frac{x}{n}$.

5.4.2 Exponential model

The likelihood function for the exponential example is
$$ L(\lambda | y) = \lambda e^{-\lambda y} $$
Therefore we have
$$ l(\lambda | y) = \log \left(\lambda e^{-\lambda y} \right) = \log \lambda - y \lambda $$
Step 1: Differentiate the log-likelihood with respect to our parameter λ
$$ \frac{d l \left(\lambda \mid y \right)}{d \lambda} = \frac{1}{\lambda} - y $$
Step 2: Set the derivative to zero and solve for λ
$$\begin{split} \begin{align*} \frac{1}{\lambda} - y &= 0 \\ \hat{\lambda} & = \frac{1}{y} \end{align*} \end{split}$$
Step 3: Verify that this is a maximum rather than a minimum by considering the second derivative
$$ \frac{d l^2 \left(\lambda \mid y \right)}{d \lambda ^2} = -\frac{1}{\lambda^2} $$
This is negative for any value of λ. So the MLE for λ is $\hat{\lambda} = \frac{1}{y}$, one over the observed waiting time.

5.5 Summary

Likelihood is a fundamental concept in statistical inference. In this session we have introduced the definition of likelihood and demonstrated how it can be used to estimate an unknown parameter, through maximum likelihood estimation. For two examples, we have seen that estimates obtained by MLE are intuitively sensible for the parameter of interest and have derived them algebraically via the log-likelihood.

In the next session, we will find out about the specific mathematical properties which make the MLE a "good" estimator, and extend to the situation where our data consist of more than one observation.

6. Maximum Likelihood

In inferential statistics, the problem we are often faced is this: we have collected some data, and we have a statistical model for how this data was generated. However, we do not know what the values of the parameters of this model are. We need to find a way to estimate these parameters. In the previous session, we were introduced to the likelihood function, which measures how consistent different values of the parameter are with the data that we have observed. Extending this concept, we used calculus to obtain the maximum likelihood estimator for the parameter.

So far, we have only looked at examples where our data consists of one observation - surely, this is not a sufficient sample size!

We will now consider the more realistic scenario, where we have a random sample of observations from a particular distribution - in this case, we say that the sample is independently and identically distributed (i.i.d).

Intended learning outcomes
By the end of this session you will be able to:

- Derive the likelihood and log-likelihood functions given an i.i.d. sample
- Derive maximum likelihood estimator from single and multi-parameter distributions given an i.i.d. sample
- Describe the main properties of MLEs

The following subsections define the likelihood function for n i.i.d observations and describe the process of obtaining the maximum likelihood estimator in this setting. The session ends with a demonstration of some important properties of maximum likelihood estimators.

6.1 Likelihood with independent observations

Suppose that the observed data consiste of a sample of n observations. If these observations are independent, then the joint likelihood function from these n observations has a very convenient form; it is the product of the likelihood from each observation.

Suppose that the random variables $X_1,..., X_n$ are i.i.d., and that our observed data are $\mathbf{x} = \left\{ x_1, x_2, ..., x_n \right\}$. Then the likelihood function is given by:
$$\begin{split} \begin{align*} L \left(\theta \mid \mathbf{x} \right) &= L\left(\theta \mid x_1 \right) L\left(\theta \mid x_2 \right) ... L \left(\theta \mid x_n \right) \\ &= \prod_{i=1}^n L\left(\theta \mid x_i \right). \end{align*} \end{split}$$
Recall that we often prefer to work with the log-likelihood function, as it simplifies the algebra when it comes to finding the MLE. The log-likelihood function for n independent observations is given by:
$$\begin{split} \begin{align*} l \left(\theta \mid \mathbf{x} \right) &= log \prod_{i=1}^n L\left(\theta \mid x_i \right) \\ &= \sum_{i=1}^n log L\left(\theta \mid x_i \right) \\ &= \sum_{i=1}^n l\left(\theta \mid x_i \right) \end{align*} \end{split}$$
Finding the MLE involves the same three steps as we saw in the previous session, but the log-likelihood function is now a joint function for the n observations:

Method for finding MLEs:

1. Obtain the derivative of the log-likelihood: $\frac{d l(\theta \mid \mathbf{x})}{d \theta}$
2. Set $\frac{d l(\theta \mid \mathbf{x})}{d \theta}=0$ and solve for θ
3. Verify that it is a maximum by showing that the second derivative $\frac{d ^2 l(\theta \mid \mathbf{x})}{d \theta ^2 }$ is negative when the MLE is substituted for θ.

6.1.1 Example: Exponential distribution

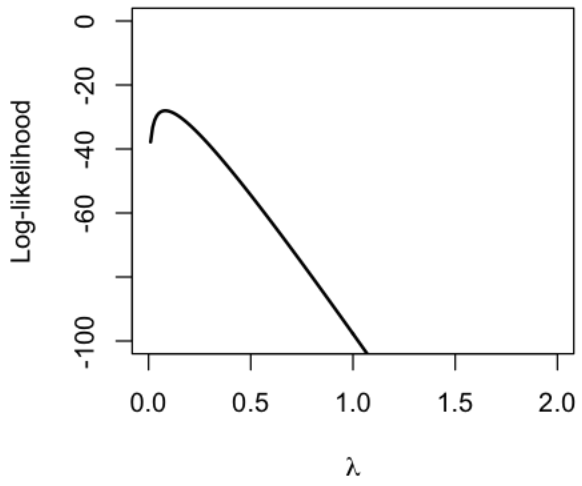
Recall the example from the previous session, investigating the time that patients wait until their GP appointment in a particular practice. The receptionist records the time that elapses between when a patient walks through the door, and when they are called through for their appointment for a random sample of 8 people. These times (in minutes) are: 8.75, 10.20, 15.29, 7.89, 7.04, 12.04, 19.04, 17.50.

As a reminder, we can model the waiting time until a specific event using the exponential distribution with parameter \(\lambda\), which has a probability density function given by:
\[\begin{equation} f _X\left(x \mid \lambda \right)=\lambda e^{-x\lambda} , x > 0, \lambda > 0 \end{equation} \]

> Recall that the mean of this distribution is equal to one over the rate parameter \(\lambda\), i.e. \(E(X) = \frac{1}{\lambda}\).

We have that the log-likelihood is:
\[\begin{split} \begin{align} \log L\left(\lambda \mid \mathbf{x} \right) &= \sum_{i=1}^n \log L(\lambda \mid x_i) \\ &= \sum_{i=1}^n \log \left(\lambda e^{-x_i \lambda } \right) \\ &= \sum_{i=1}^n \log \lambda -x_i\lambda \\ &= n \log \lambda - \lambda \sum_{i=1}^n x_i \end{align} \end{split}\]
We can make a plot of this log-likelihood, using the data from our example with eight observations.

```
options(repr.plot.width=4, repr.plot.height=4)

#six independent observations for waiting times
obs <- c(8.75, 10.20, 15.29, 7.89, 7.04, 12.04, 19.04, 17.50)
n <- length(obs)

#possible values for the parameter lambda
lambda <- seq(0, 2, 0.01)

#plot the log-likelihood
plot(lambda, n*log(lambda) - lambda*sum(obs), type="l",lwd=2,
     xlab= expression(lambda), ylim=c(-100,0),
     ylab="Log-likelihood")
```



Graphically, we observe that the maximum is between 0 and 0.25. We will use the three steps, as before, to derive the MLE algebraically:

Step1: Taking the derivative of the log-likelihood with respect to \(\lambda\):
\[\begin{equation} \frac{d log L\left(\lambda \mid x_1 ,..., x_n \right) }{d \lambda} = \frac{n}{\lambda}- \sum_{i=1}^n x_i \end{equation} \]
Step2: Set the derivative equal to zero and solve for \(\lambda\):
\[\begin{split} \begin{align*} 0 &= \frac{n}{\lambda}- \sum_{i=1}^n x_i \\ \hat{\lambda} &= \frac{n }{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}} \end{align*} \end{split}\]
The MLE is \(\hat{\lambda}= \frac{1}{\bar{x}}\). And to check that this provides a maximum, we go on to the next step:

Step3: Find the second derivative:

$$ \begin{equation} \frac{d l^2 \left(\lambda \mid \boldsymbol{x} \right)}{d \lambda ^2} = - \frac{n}{\lambda^2} \end{equation} $$
When $\hat{\lambda}=\frac{1}{\bar{x}}$, we have:
$$ \begin{align} \frac{d l^2 \left(\lambda \mid \boldsymbol{x} \right)}{d \lambda ^2} &=-n \bar{x}^2 \end{align} $$
which is negative. This verifies that we found the maximum likelihood estimate.

Going back to our example of eight patients waiting for their GP appointment, the maximum likelihood estimate λ is given by one over the average of the eight waiting times:

```
1/mean(obs)
```

```
0.0818414322250639
```

We have that $\hat{\lambda}=0.0818$ minutes.

6.1.2 Example: Normal distribution

We will now consider the normal distribution. Remember that the normal distribution has two parameters, μ and σ^2. We will first obtain the MLE for μ (treating σ^2 as a constant), and in the practical, we will obtain the MLE for σ^2 (treating μ as a constant).

Recall that normal distribution has probability density function given by*:
$$ \begin{equation} f_X \left(x \mid \mu, \sigma^2 \right)= (2 \pi \sigma^2)^{-\frac{1}{2}} \exp \left\{-\frac{(x-\mu)^2}{2\sigma^2} \right\} \end{equation} $$
(* note that the notation here is slightly different to section 3. Here we are more prescriptive; on the left hand side the notation says that the random variable X is sampled from parameters μ and σ^2, where the distribution is defined on the right hand side. Both versions of notation are acceptible. Another notation style is to use a semi-colon instead, ie. $f_X(x ; \mu, \sigma^2)$).

We have that the log-likelihood given an i.i.d. sample of size n is:
$$\begin{split} \begin{align} l \left(\mu, \sigma^2 \mid \boldsymbol{x} \right) &= \sum_{i=1}^n \log \left\{ (2 \pi \sigma^2)^{-\frac{1}{2}} \exp \left\{-\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \right\} \\ &= \sum_{i=1}^n \left\{ \log (2 \pi \sigma^2)^{-\frac{1}{2}}+ \log \exp \left\{-\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \right\} \\ &= \sum_{i=1}^n \left\{ -\frac{1}{2} \log (2 \pi \sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2} \right\} \\ &= {-\frac{n}{2}}\log (2 \pi \sigma^2) -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2 \end{align} \end{split}$$
We will first find the MLE for the parameter μ.

Step1: Take the derivative of the log-likelihood with respect to μ. Note that this requires use of the chain rule:
$$\begin{split} \begin{align} \frac{d l \left(\mu, \sigma^2 \mid \mathbf{x} \right) }{d \mu} &= -\frac{2}{2\sigma^2}(-1) \sum_{i=1}^n (x_i-\mu) \\ &= \frac{ \sum_{i=1}^n (x_i-\mu)}{\sigma^2} \\ &= \frac{ \sum_{i=1}^n x_i-n\mu}{\sigma^2} \end{align} \end{split}$$
Step2: Setting the derivative equal to zero and solving for μ:
$$\begin{split} \begin{align} 0 &= \frac{ \sum_{i=1}^n x_i-n\mu}{\sigma^2} \\ \end{align} \end{split}$$
Since $\sigma^2 > 0$, we have that:
$$ \begin{equation} 0 = \sum_{i=1}^n x_i-n\mu \end{equation} $$
$$ \begin{equation} \hat{\mu} =\frac{ \sum_{i=1}^n x_i}{n} = \bar{x} \end{equation} $$
We have that the MLE for μ is the sample mean, $\bar{x}$.

Step3: Find the second derivative:
$$ \begin{align} \frac{d^2 l \left(\mu, \sigma^2 \mid \mathbf{x} \right) }{d \mu^2 } &= -\frac{n}{\sigma^2}, \end{align} $$
since both $n>0$ and $\sigma^2 >0$, we have that the second derivative is negative, verifying that we have found the maximum.

In the practical, we will find the MLE for σ^2.

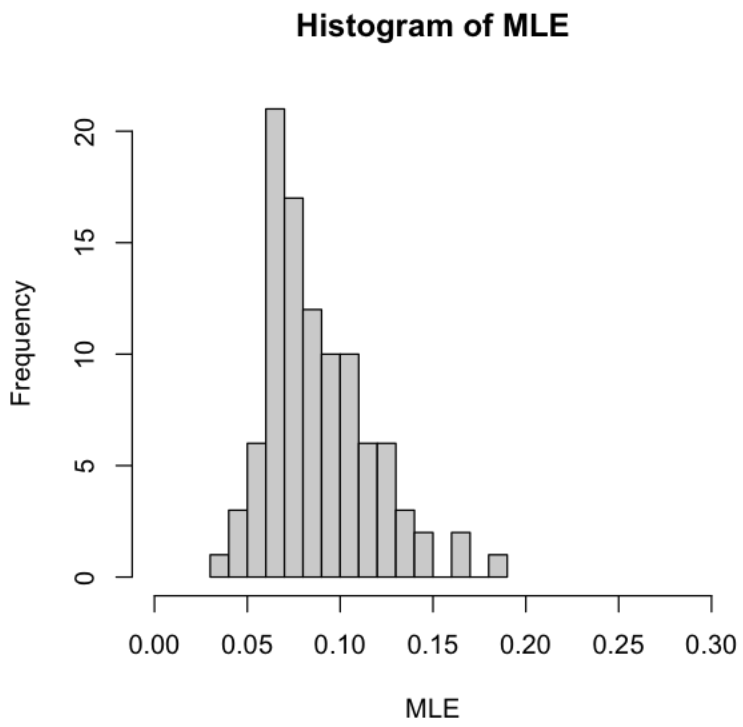
6.2 Properties of maximum likelihood estimators

Maximum likelihood estimators can be shown to have some very useful properties. In particular, there are some very important asymptotic properties (properties that we observe as the sample size of our data gets very very large).

To explore these properties, have a look at the simulation below. We generate a sample of size 8 from the exponential distribution where $\lambda=12.22$. The MLE is calculated from this the observed mean of the sample. We repeat this 100 times, and we plot a histogram of the 100 MLEs that we obtain.

Change the sample size, n, to larger numbers and see what you notice about the histogram.

```r
n <- 8  #  make this sample size bigger, and see what happens to the histogram!

# MLEs will be stored in this vector
mle <- rep(0, 100)

for (i in 1:100){
  # Generate a sample of size n from an exponential distribution with lambda=0.0818
  sample <- rexp(n, rate=0.0818)
  # Calculate the MLE (the reciprocal mean of the sample) and store it
  mle[i] <- 1/mean(sample)
}

# Plot a histogram of the 100 MLEs
options(repr.plot.width=5, repr.plot.height=5)
hist(mle, breaks=20,
     xlim=c(0, 0.3),
     main="Histogram of MLE",
     xlab="MLE")
# Add red line to indicate true lambda
abline(v=12.22, col="red")
```



You may notice that, as n becomes large, the distribution of the MLE becomes more and more concentrated around the true value, and the histrogram appears to look more bell-shaped.

Suppose we denote the parameter of interest as θ and its MLE as $\hat{\theta}$. The tabs below show some important properties of MLEs.

Bias | Consistency | Normality | Efficiency | Transformation invariance

The MLE is **asymptotically unbiased**, i.e. on average we obtain the correct answer as samples become large.

$\mathbb{E}(\hat{\theta}) \rightarrow \theta$ as $n \rightarrow \infty$.

You might question to what extent these asymptotic properties are useful in practical examples where the sample size is relatively small.

Further, in the cases that we have covered so far, it is fairly straightforward to compute the likelihood function and to find the value that maximizes it, but in many situations, this will be a complex task that requires numerical approaches.

In the subsequent sessions on Bayesian Statistics, we will see a different paradigm for making inference which can address some of these issues.

6.3 Summary

We now know how to obtain the likelihood and log-likelihood functions when you have an i.i.d. sample of observations. We can then obtain the maximum likelihood estimators of the parameters of the distribution. The MLE is an important tool as it has a number of important asymptotic properties, as we demonstrated using a simulation in R. Finally, we introduced the idea of a log-likelihood ratio, which is a way of comparing estimates of a parameter with the maximum likelihood estimate.

You may be wondering how you might measure the precision of your estimator. We will return to this question in our session about confidence intervals.

Note that the maximum likelihood estimator, and confidence intervals, are tools from the "frequentist" or "classical" approach to statistics. In later sessions, you will meet the Bayesian approach to statistics, where the Likelihood will also play an important role.

Appendix: Additional Reading

This appendix section contains additional information which will deepen your understanding. However, it is not examinable and is completely optional reading.

A1: Log-likelihood ratios

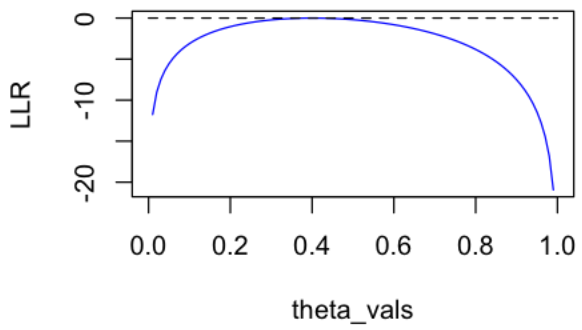
So far we have used the MLE to find an estimate of a parameter. Typically, the estimate is computed from a sample, so if we were to *sample again* we would expect the estimate to vary a little. But what about others values; what steps are involved to compare other estimates to the MLE? How much would the sample estimates vary? The **log-likelihood ratio (LLR)** is a useful approach. The LLR gives a measure of consistency of a value of θ relative to the most likely value.

The LLR is defined as,
$$ log\frac{L(\theta)}{L(\hat\theta)} $$
where $L(\theta)$ is the likelihood evaluated at any value, and $L(\hat\theta)$ is the likelihood evaluated at the MLE.

Alternatively, and especially when evaluating in software, the following is used;
$$ LLR(\theta) = l(\theta)-l(\hat\theta) $$
Let's explore the LLR and its properties with an small example.

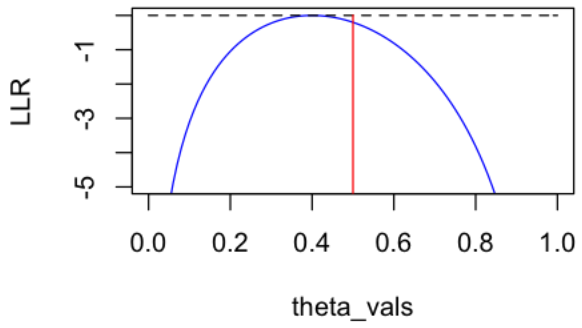
For a simple coin-flipping example, from a trial of 10 coin-flips, 4 were heads ($X=4$) and the remainder were tails. From this experiment, we know that the MLE (ie. $\hat\theta$) is 0.4, but we also want to use the LLR to compare other estimates of θ. Looking at the LLR graphically we note that the LLR is a negative value, the further away from zero the less consistent the parameter value with to the MLE.

```
options(repr.plot.width=4, repr.plot.height=3)
 x <- 4; n<-10
 theta_vals <- seq(0,1,0.01)
 LLR <- dbinom(x,n,theta_vals,log=T)-dbinom(x,n,x/n,log=T)
 plot(theta_vals,LLR,type='l',col='blue')
# add additional things
lines(x=theta_vals,y=rep(0,length(theta_vals)),lty=2)
```



In the above experiment, for a *fair* coin it would not be unusual to observe 4 heads from 10 trials. The MLE is 0.4 but we *know* for a fair coin that the true parameter π will be 0.5. The MLE is a *sample* of the distribution for π. So let's zoom in on the figure previously generated;

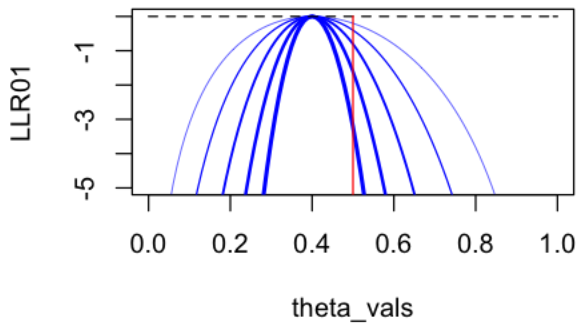
```
options(repr.plot.width=4, repr.plot.height=3)
 x <- 4; n<-10
 theta_vals <- seq(0,1,0.01)
 LLR <- dbinom(x,n,theta_vals,log=T)-dbinom(x,n,x/n,log=T)
 plot(theta_vals,LLR,type='l',col='blue',ylim=c(-5,0.01))
# add additional things
lines(x=theta_vals,y=rep(0,length(theta_vals)),lty=2)
lines(x=rep(0.5,2),y=c(-10,0.01),col='red',lty=1)
```



We can see in the figure (red line) that when $\theta=0.5$ the LLR is very close to 0. Values of θ further away from the MLE than 0.5 will have a even lower LLR. So we can make qualitative statements using the LLR in relation to the MLE.

Let's increase the sample size and observe what happens to the LLR.

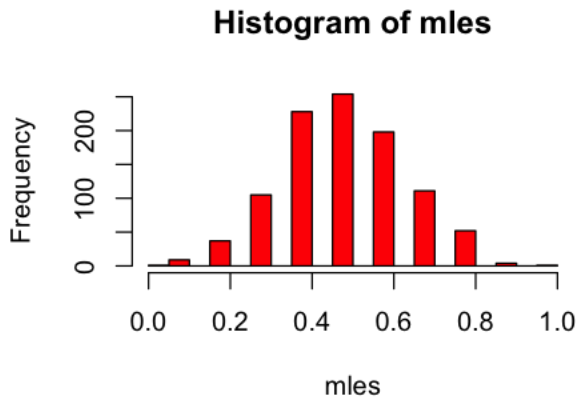
```r
options(repr.plot.width=4, repr.plot.height=3)
 ns<-c(10,20,40,80,160)
# assume we have an 'unfair coin' with heads being more likely than tails
# rather than taking a random sample, assume a sample consistent with the null
hypothesis - this is so the MLE remains 0.5
 x1<-round(ns[1]*0.4,0)
 x2<-round(ns[2]*0.4,0)
 x3<-round(ns[3]*0.4,0)
 x4<-round(ns[4]*0.4,0)
 x5<-round(ns[5]*0.4,0)
 theta_vals <- seq(0,1,0.01)
 LLR01 <- dbinom(x1,ns[1],theta_vals,log=T)-dbinom(x1,ns[1],x1/ns[1],log=T)
 LLR02 <- dbinom(x2,ns[2],theta_vals,log=T)-dbinom(x2,ns[2],x2/ns[2],log=T)
 LLR03 <- dbinom(x3,ns[3],theta_vals,log=T)-dbinom(x3,ns[3],x3/ns[3],log=T)
 LLR04 <- dbinom(x4,ns[4],theta_vals,log=T)-dbinom(x4,ns[4],x4/ns[4],log=T)
 LLR05 <- dbinom(x5,ns[5],theta_vals,log=T)-dbinom(x5,ns[5],x5/ns[5],log=T)
 plot(theta_vals,LLR01,type='l',col='blue',lwd=0.5,ylim=c(-5,0.01))
# compare to large sample sizes
 lines(theta_vals,LLR02,col='blue',lwd=1)
 lines(theta_vals,LLR03,col='blue',lwd=1.5)
lines(theta_vals,LLR04,col='blue',lwd=2)
lines(theta_vals,LLR05,col='blue',lwd=2.5)
lines(x=theta_vals,y=rep(0,length(theta_vals)),lty=2)
lines(x=rep(0.5,2),y=c(-10,0.01),col='red',lty=1)
```



At smaller sample sizes the LLR is slightly left skewed when the sample mean is 0.4. As the sample size increases you can see that the LLR becomes more symmetrical about the sample mean, and that the slope of the LLR at values away from the sample mean is steeper. We can start to see the relationship between sample size and the precision of the sample mean. Qualitatively, if we wanted to test whether the coin was fair, it is clear that a larger sample would enable us to have more confidence in our assessment.

Returning to the fact that the data are a sample from a population distribution, we can explore what happens when multiple samples of the same size are drawn. MLE is a sample of the true parameter we can perform the above experiment multiple times and identify the parameters space where the LLR will be zero.

```r
options(repr.plot.width=4, repr.plot.height=3)
 x <- 4; n<-10
 sampl <- rbinom(n=1000,size=n,prob=0.5)
mles <- sampl/n
 hist(mles,col="red",breaks=30)
```



From the histogram above you can see that repeating the experiment (different samples) will return different values of the MLE and corresponding LLR. The LLR ratio can be used to assess how consistent different values of the parameter are with the MLE.

The principles behind the LLR also relate to construction of confidence intervals, an issue which we will return to when we meet logistic regression and other models estimated using maximum likelihood estimation.

7. Frequentist I: Confidence Intervals

In previous sessions we considered the concept of estimating population parameters using information from a sample from the population. When we present an estimate of a population quantity, it is important to also provide a measure of how precise that estimate is. Do we believe it is close to the true value? Can we provide a range of values within which we believe the true value lies?

This is the purpose of a confidence interval, often abbreviated by CI. Loosely speaking, a confidence interval provides a range of values for the population parameter which our observed data are consistent with.

Intended learning outcomes
By the end of this session you will be able to:

- correctly interpret a 95% confidence interval interval
- describe properties of a 95% confidence interval over repeated sampling
- calculate a 95% confidence interval for the mean
- use resampling (bootstrapping) approaches to obtain percentile confidence intervals

7.1 Confidence intervals

To explore the concept of confidence intervals, we will return to the example of emotional distress among violence researchers.

We will again consider the smaller subsample of 10 researchers and focus on estimating the population mean age, \(\mu\). Among our 10 sampled violence researchers, the sample mean age and the sample proportion suffering from emotional distress are:

> Sample mean age \(\bar{x}= 29.57\); sample standard deviation of age \(SD = 4.95\)

Statistical model: As before, we will let \(X_1, ...,X_{10}\) be random variables representing the ages of 10 sampled researchers . For simplicity, we will assume that we know the true value of the population standard deviation, \(\sigma = 4.8\). We assume the following model
\[X_i \overset{\small{iid}}{\sim} N(\mu, 4.8^2), \qquad i=1,2,...,10 \]
Data: The realised values of the random variables are \(x_1, ..., x_{10}\) (i.e. the observed ages).

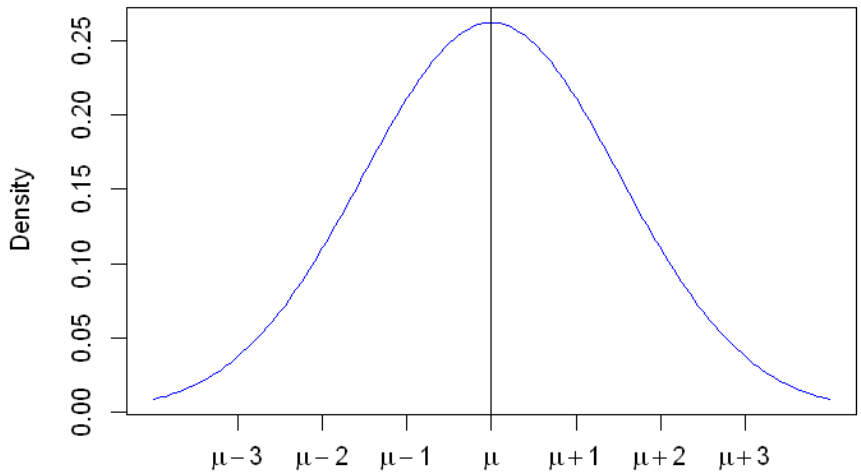
Estimator and estimate: The best estimator of the population mean age is the sample mean age.

From our sample of data, the estimate is $\hat{\mu} = 29.57$. But how good an estimate is this? In order to answer that question, we will construct a 95% confidence interval around the estimate.

Sampling distribution of the estimator: Recall that the sampling distribution of the sample mean is the distribution we would see if we repeatedly sampled 10 researchers a very large number of times, each time calculating the sample mean age, and drew a histogram of the sample means. We obtained the sampling distribution algebraically:
$$ \hat{\mu} \sim N(\mu, 1.52^2) $$
Recall that when we are talking about the sampling distribution (i.e. the distribution of an *estimator*), we call the standard deviation the **standard error**. So the sample mean age follows a normal distribution, under repeated sampling, centred around the population mean μ with standard error given by $SE(\hat{\mu}) = 1.52$.

We do not quite have sufficient information to plot the sampling distribution, because we still do not know where the central value μ is. However, otherwise we can draw the exact shape. The graph below draws the sampling distribution around an unknown population mean μ.

[The code used to generate the graph is suppressed, since it is not our focus here, but if you wish to see it you can click the button to the right.]



7.1.1 Confidence interval for the mean

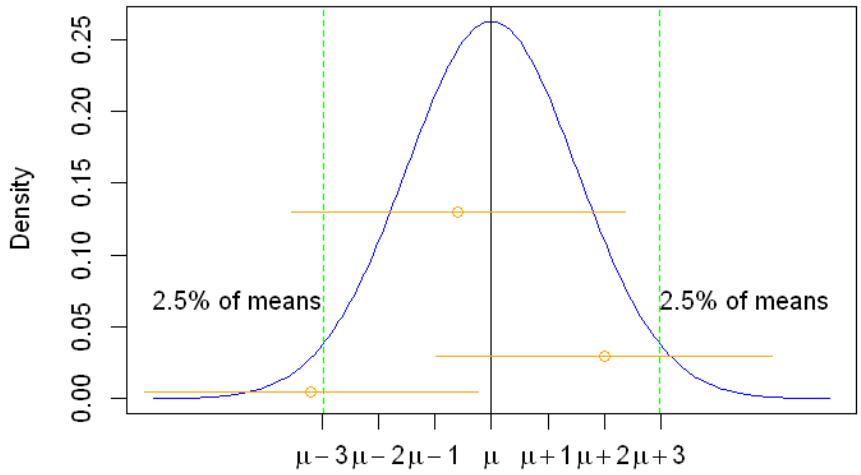
We now use a general fact about normal distributions:

> For a normal distribution, 95% of the observations lie within 1.96 standard deviations of the mean.

For the sampling distribution above, the "observations" are the different sample means we would see under (hypothetical) repeated sampling. Recall that when the we talk about a distribution of an estimator, we call the standard deviation the standard error. Thus the standard deviation of these observations (sample means) is the standard error of the mean, which takes a value of 1.52 here.

Therefore, 95% of the sample means lie within $1.52 \times 1.96 = 2.98$ of the population mean μ.

Imagine taking each (hypothetical) sample mean and "stretching out" a distance of 2.98 either way to give a range of values around that sample mean.



What proportion of such intervals would we expect to contain the true value \(\mu\)? Have a think about it and then click the button to the right.

These intervals are called **95% confidence intervals**.

7.2 Confidence intervals for the mean

7.2.1 Example

In the sample of 10 researchers, the estimate of the population mean age is \(\hat{\mu} = 29.75\), the sample mean age.
The standard error of the mean is
\[SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} = \frac{4.8}{\sqrt{10}} = 1.52 \]
We have seen that the 95% confidence interval for the mean is calculated as
\[\hat{\mu} \pm 1.96 \times SE(\hat{\mu}) \]
Substituting in the sample mean and the standard error gives
\[29.57 \pm 1.96 \times 1.52 \]
This gives the 95% confidence interval for the population mean age: \((26.6, 32.5)\).

The code below reads in the data, prints the sample mean age and then calculates the 95% confidence interval for the population mean age.

```
# Our sample of data (ages for 10 sampled researchers)
ages <- c(28.1,27.5,25,29.9,29.7,29.9,39.9,33.6,21.3,30.8)

# Sample mean (estimate of the population mean)
mean(ages)

# Display the lower and upper limits of the confidence interval
mean(ages) - 1.96*1.52
mean(ages) + 1.96*1.52
```

29.57
26.5908
32.5492

7.2.2 95% confidence interval for a mean

For random variables \(X_1,...,X_n\), with \(X_i \overset{\small{iid}}{\sim} N (\mu, \sigma^2)\) for \(i=1,...,n\) and \(\sigma\) is a known value, a 95% confidence interval for \(\mu\) is given by:
\[\hat{\mu} \pm 1.96 \ SE(\hat{\mu}) \]
where the standard error of \(\hat{\mu}\) is given by
\[SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \]
The calculation of this confidence interval relies on the assumptions that

- the original random variables follow a normal distribution
- the value of \(\sigma\) is known

However, if these assumptions are not true, we can still obtain valid confidence intervals:

- If the original random variables do not follow a normal distribution but the sample size is large, then the Central Limit Theorem tells us that the sampling distribution of the mean is approximately normal. So this formula for the confidence interval is still valid.
- If \(\sigma\) is unknown (which is typically the case), there is a modified confidence interval based on the t-distribution which provides a correct interval. Essentially, we replace the number 1.96 above by a slightly larger number to compensate for the estimation of the standard deviation. For large sample sizes (\(n>30\) or so), the substitution of the estimated standard deviation makes little difference. More detail is provided later in this session.

7.3 Interpretation of confidence intervals

For our emotional distress sub-sample, our estimated mean age is \(\hat{\mu}= 29.75\), with a 95% confidence interval of \((26.6, 32.5)\). Having calculated this confidence interval for our unknown population age, \(\mu\), how do we interpret it?

7.3.1 Operational definition

For the 95% confidence interval calculated above \((26.6, 32.5)\), it is tempting to say that the probability that the population mean age, \(\mu\), is between 26.6 and 32.5 is 95%. However, this is incorrect, because this implies that \(\mu\) has a probability distribution, rather than being a fixed unknown number. Either the true value of \(\mu\) lies within the interval 26.6 to 32.5, or it does not.

Strictly, the interpretation of a confidence interval has to be with respect to the process of repeated sampling: if we repeated the study an infinite number of times, 95% of the 95% confidence intervals calculated would include the true population mean \(\mu\).

This operational definition is long-winded and can be confusing. In practice, we often use looser interpretations to aid communication of results, as described below.

7.3.2 Looser interpretation (practical)

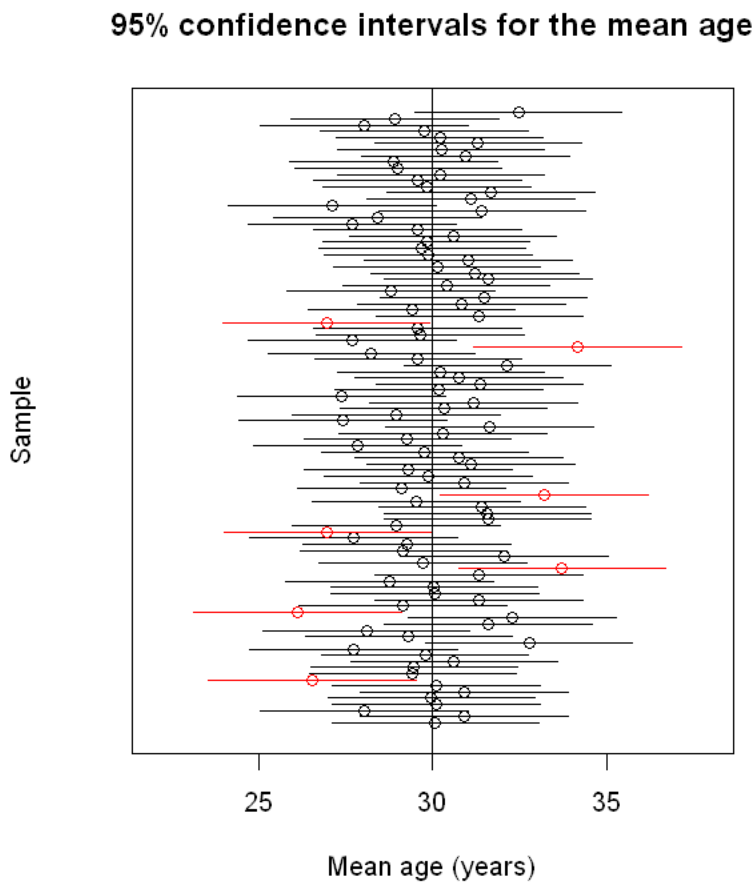
In practice, we often loosely interpret a 95% confidence interval by saying

- that we are 95% confident that the true population mean lies within the 95% confidence interval calculated.
- that our data are consistent with values of the population mean within the 95% confidence interval calculated.

It is important, however, to bear the strict operational definition of the confidence interval in mind when we use these types of interpretations.

7.3.3 Confidence intervals under repeated sampling

In order to see how confidence intervals behave under repeated sampling, we will now randomly draw 100 different samples of 10 people. Within each sample, we calculate the sample mean and the 95% confidence interval (assuming the population value \(\sigma\) is known). The graph below shows the 95% confidence intervals from the 100 samples. [Click to view the code that generates the graph].



In the figure above, what do you notice? Approximately what proportion of intervals include the true value of μ? Have a think about this and then click to see some comments about the graph.

7.4 Approximate confidence intervals for parameters estimated using large samples

You will encounter many different confidence intervals during your studies. Many of these rely on an asymptotic normal distribution, as described below.

There are many different confidence intervals and many approaches to calculating confidence intervals. We do not aim to give you a comprehensive list here. Below, we describe a few commonly used confidence intervals to give you a flavour. **Please note:** We do not expect you to memorise these formulae.

7.4.1 Normal-based confidence intervals

The Central Limit Theorem tells us that the mean of independent identically distributed random variables, with finite expectation and variance, tends to a normal distribution as the sample size tends to infinity.

In fact, the Central Limit Theorem means that most typically encountered parameter estimators tends to normal as the sample sizes tend to infinity. So we can follow a very similar approach to the one above to construct confidence intervals for any parameter estimators that follow an approximate normal distribution when sample sizes are large, giving a confidence interval of the form
$$ \mbox{Estimate} \pm 1.96 \times SE(\mbox{Estimator}) $$

7.4.2 Proportions and rates

First, we need some notation.

Proportion	Rate	Logarithm of rate
We are estimating a population proportion from a single observation from a binomial distribution. Our observed data consist of one observation from $X \sim binomial(n, \pi)$, with the realised (observed) value being $X=k$.	We are estimating a population rate (per person-year), from the total number of events out of P person-years of observation. Our observed data consist of one observation from $X \sim Poisson(\lambda P)$. The realised (observed) value is $X=d$.	For the rate, we may wish to perform our calculations on the log scale. These confidence intervals are approximate; the approximation can work better following a transformation (e.g. the log). This is one example of that approach. To do this, we need to define the log-rate, $\nu = log(\lambda)$

Using this notation, we can write down the estimate of the parameter of interest, it's standard error and an approximate 95% confidence interval. These are shown in the table below.

	Estimate of parameter	**Standard Error**	**Approximate 95% Confidence Interval**
Proportion	$\hat{\pi} = \frac{k}{n}$	$\sqrt{\frac{\pi (1-\pi)}{n}}$	$\hat{\pi} \pm 1.96 \times \sqrt{\frac{\hat{\pi} (1-\hat{\pi})}{n}}$
Rate	$\hat{\lambda} = \frac{d}{P}$	$\frac{\lambda}{\sqrt{d}}$	$\hat{\lambda} \pm 1.96 \times \frac{\hat{\lambda}}{\sqrt{d}}$
Log Rate	$\hat{\nu} = log\left(\frac{d}{P}\right)$	$\sqrt{\frac{1}{\lambda P}}$	$\hat{\nu} \pm 1.96 \times \sqrt{\frac{1}{e^{\hat{\nu}} P}}$

The three tabs below provide examples of using the formulae above to obtain approximate 95% confidence intervals for proportions and rates.

Proportion | Rate | Logarithm of the rate

- We want to estimate the population proportion of patients who experience a side effect from a particular drug.
- In a clinical study of 80 patients given the drug, $X=20$ experience a side effect.
- Our estimate of the population proportion experiencing a side effect is $\hat{\pi} = 0.25$
- Our 95% confidence interval for this proportion is:

$$ 0.25 \pm 1.96 \times \sqrt{\frac{0.25 (1-0.25)}{80}} $$

- This gives a range of 0.155 to 0.349.
- So our estimate of the proportion of patients who experience a side effect is: 0.25 (95% CI 0.155 to 0.349). Our best guess is that 25% of patients experience a side-effect from this drug. We are 95% confident that the true proportion lies between 15.5% and 34.9%.

7.4.2 The mean

In this subsection we consider estimating a population mean. Our observed data comprise n independent observations, $x_1, x_2, ..., x_n$. We consider two possibilities:

- Data are normally distributed, $X_i \sim Normal(\mu, \sigma^2),$ for $i=1,..,n$
- Data are not normally distributed.

In each case, the population mean and variance (the square of the population standard deviation) are:
$$ E[X] = \mu, \ \ Var(X) = \sigma^2 $$
The sample mean is $\bar{x}$ and the sample standard deviation is s. Our estimate of the population mean is just the sample mean: $\hat{\mu} = \bar{x}$. And, as we have seen, the standard error is given by $\frac{\sigma}{\sqrt{n}}$.

There are various ways of constructing a 95% confidence interval, depending on the situation. These are shown in the table below.

	Approximate 95% Confidence Interval
Small samples	
- Normal distribution, known σ	$\hat{\mu} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$
- Normal distribution, unknown σ	$\hat{\mu} \pm t_{n-1} \times \frac{s}{\sqrt{n}}$
Large samples	
- Normal or not, known σ	$\hat{\mu} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$
- Normal or not, unknown σ	$\hat{\mu} \pm 1.96 \times \frac{s}{\sqrt{n}}$

where t_{n-1} is a number obtained from the t-distribution, which is similar to the standard normal distribution. This number is around 2 for small n and becomes very close to 1.96 for large samples.

For small samples, where the data are not normally distributed, the confidence interval assuming data are normally distributed often has reasonable performance, but other methods (e.g. bootstrap confidence intervals) may be advisable.

7.4.3 Comparing two groups

There are many ways of comparing outcomes between two groups. Two popular options are the difference in proportions for binary outcomes and the difference in means for continuous outcomes. Confidence intervals for other measures (e.g. the risk ratio, the odds ratio, the difference in medians, etc.) also can be obtained.

Difference in proportions | Difference in means

We are interested in the population difference in proportions from two observations from two binomial distributions. Suppose our observed data consist of two observations from $X_1 \sim binomial(n_1, \pi_1)$, with the realised values being $X_1=k_1$ and $X_2 = k_2$. We want to estimate the difference $\delta = \pi_1 - \pi_2$.

We estimate the proportion in the first group by $\hat{\pi}_1 = \frac{k_1}{n_1}$. Similarly, we estimate the proportion in the second group by $\hat{\pi}_2 = \frac{k_2}{n_2}$. Then our estimate of the difference in proportions is $\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_2$.

The standard error for the difference in proportions is:
$$ \sqrt{\frac{\pi_1 (1-\pi_1)}{n_1} + \frac{\pi_2 (1-\pi_2)}{n_2}} $$
And we can obtain an approximate 95% confidence interval as:
$$\hat{\delta} \pm 1.96 \times \sqrt{\frac{\hat{\pi}_1 (1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2 (1-\hat{\pi}_2)}{n_2}}$$

7.5 Confidence Intervals using resampling

We saw that we can often create an approximate sampling distribution by resampling from our sample data. This is particularly useful in situations where there is no algebraic derivation for the sampling distribution.

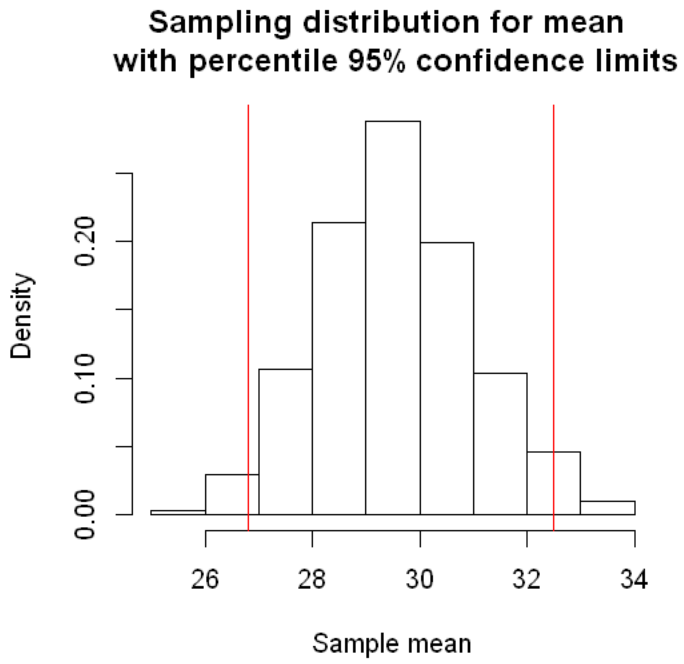
We have seen that the important connection between sampling distributions and confidence intervals. So we would intuitively expect to be able to construct a confidence interval from the approximate sampling distribution we obtained using resampling. This is indeed possible. There are many ways of doing this, but the simplest and most intuitive method is the **bootstrap percentile confidence interval**.

The basic idea is very simple. We construct an approximate sampling distribution using bootstrap samples, as we did previously. Then we take the 2.5th and 97.5th percentiles of that distribution (the value such that 2.5% of the observations - the estimates across bootstrap samples - lie below the value; and the value such that 2.5% of observations lie above the value, respectively). These form the limits of our 95% confidence interval.

```r
set.seed(78234)

# Read in the sample of 10 ages
ages <- c(28.1,27.5,25,29.9,29.7,29.9,39.9,33.6,21.3,30.8)

# Draw bootstrap samples
bootstrap_samples <- lapply(1:1039, function(i) sample(ages, replace = T))

# Calculate sample means in each bootstrap sample
r.mean <- sapply(bootstrap_samples, mean)

# Obtain the 2.5th and 97.5th percentiles of the sample means across bootstrap samples
(q<-quantile(r.mean, c(0.025, 0.975)))

# Draw the approximate sampling distribution with the percentile confidence limits
marked in red
options(repr.plot.width=4.5, repr.plot.height=4.5)
hist(r.mean, freq=FALSE, main="Sampling distribution for mean \n with percentile 95%
confidence limits", xlab="Sample mean")
abline(v=q, col="red")
```

2.5%
26.798

97.5%
32.501



The approximate 95% confidence interval for the mean age obtained by using the algebric approximation to the sampling distribution was: 26.6 to 32.5. The bootstrap percentile 95% confidence interval is: 26.8 to 32.5. We see that these intervals are very similar to one another, as we would expect.

7.6 Summary: Use of confidence intervals

In this session we have discussed the concepts underlying confidence intervals and different interpretations of 95% confidence intervals.

We can construct a confidence interval for any estimate, including:

- means
- proportions
- differences in means
- risk ratios
- regression coefficients
- etc.

The way we construct confidence intervals can vary but the basic interpretation of the confidence interval remains the same.

While we have focused on 95% confidence intervals, we can construct other intervals, e.g. 99% confidence intervals. The use of 95% confidence intervals is largely convention.

Further resources

Note: further resources are for you to deepen your understanding of the subject if you wish to do so. This is entirely optional. All examinable material is contained within the notes.

Ashley I Naimi, Brian W Whitcomb, Can Confidence Intervals Be Interpreted?, American Journal of Epidemiology, Volume 189, Issue 7, July 2020, Pages 631–633, https://doi.org/10.1093/aje/kwaa004

8. Frequentist II: Hypothesis tests

In this session we continue with frequentist inference, exploring the concept of hypothesis testing and p-values. We discuss the general principle underlying a p-value, connections to confidence intervals and common misinterpretations and misuses of p-values.

Intended learning outcomes

By the end of this session you will be able to:

- describe the meaning of the terms null and alternative hypotheses
- describe what a p-value is
- correctly interpret a p-value
- explain the connection between 95% confidence intervals and p-values
- describe the calculation of a p-value

This session does not cover the mathematical derivation underlying hypothesis tests and common test statistics. Our intention is not to equip you with the ability to construct novel hypothesis tests. The purpose of this session is rather to convey an understanding of what a hypothesis test is, what a p-value is and how to interpret p-values correctly.

8.1 Evidence against hypotheses

This session considers the concept of testing hypotheses.

8.1.1 Proving and disproving hypotheses

Let's consider a very simple example. Suppose we believe that all men are over 120cm tall. We could:

- *Prove* the hypothesis by finding every man and showing they are more than 120cm tall
- *Disprove* the hypothesis by finding a single man less than 120cm tall

> It is easier to find evidence against a hypothesis than to prove it to be correct.

The general approach we will take is as follows. We start with a **null hypothesis**, which is typically a statement about the population value of parameters. This will often be a statement of "no difference". Some examples might be:

- Exposure to passive smoking is not associated with subsequent risk of lung cancer.
- Treatment A does not improve survival compared with placebo
- The mean body mass index (BMI) in England is the same as the mean BMI in Scotland.

We assume that our null hypothesis holds, i.e. that our sample of data came from a population in which our null hypothesis is true. We then look for evidence, in our sample data, against the null hypothesis (i.e. to falsify the hypothesis).

For example, suppose our null hypothesis is that the mean BMI is the same in England and Scotland and that we have a random sample of adults from England and Scotland. If we assume our null hypothesis is true (the two populations have the same mean BMI), then we would expect our two samples to have similar means. If, in fact, we observed very different sample means in the two sample groups then we would take this as *evidence against our null hypothesis*.

8.1.2 Example

To explore the concept of hypothesis testing, we will return to the example of emotional distress among violence researchers. The researchers were randomly assigned to receive an intervention (group debriefing aimed at reducing emotional distress) or control (nothing). At the end of the intervention, 22 researchers in the intervention group and 26 researchers in the control group filled in a questionnaire measuring emotional distress. The score gives a value of 0-20, with higher scores indicating higher distress.

The sample mean scores and their standard deviations in the two groups are:

- Control group ($n_0=26$), sample mean emotional distress score (sample standard deviation): $\bar{x}_0 = 6.35$, (SD = 1.87)
- Intervention group ($n_1=22$), sample mean emotional distress score (sample standard deviation): $\bar{x}_1 = 5.45$, (SD = 1.87)

The research question we consider in this session is:

> Is the true mean emotional distress score is different in the intervention and control group?

The population parameter of interest is therefore the difference between the population mean emotional distress score in the intervention and control groups.

Aside: as is often the case, the population is a bit hard to define here. We can think about it as being the wider population of people who could be given the intervention (or not).

The code below reads in the data, obtains the sample means and SDs and draws histograms of the scores in each group.

```
# Read in data (emotional distress scores in control and intervention group)
dist0 <- c(5, 2, 5,  7,  6,  7,  7,  5,  8,  6,  6,  9,  4,  5,  9,  7,  9,  5,  6,
10,  9,  4,  6,  6,  5,  7)
dist1<- c(5,  5,  6,  6, 1,  5, 10,  7,  3,  6,  7,  8,  6,  7,  5,  4,  5,  6,  4,
6,  3,  5)

# Calculate sample means
print("Sample means: ")
mean(dist0)
mean(dist1)

# Calculate sample standard deviations
print("Sample SDs: ")
sqrt(var(dist0))
sqrt(var(dist1))

# Sample difference in means
print("Difference in sample means:")
(delta.hat <- mean(dist1) - mean(dist0))

# Draw histograms of the scores in each group
options(repr.plot.width=6, repr.plot.height=4)
par(mfrow=c(1,2))
hist(dist0, main="Distress score, control", xlim=c(0, 12))
hist(dist1, main="Distress score, intervention", xlim=c(0, 12))
```

```
[1] "Sample means: "
```

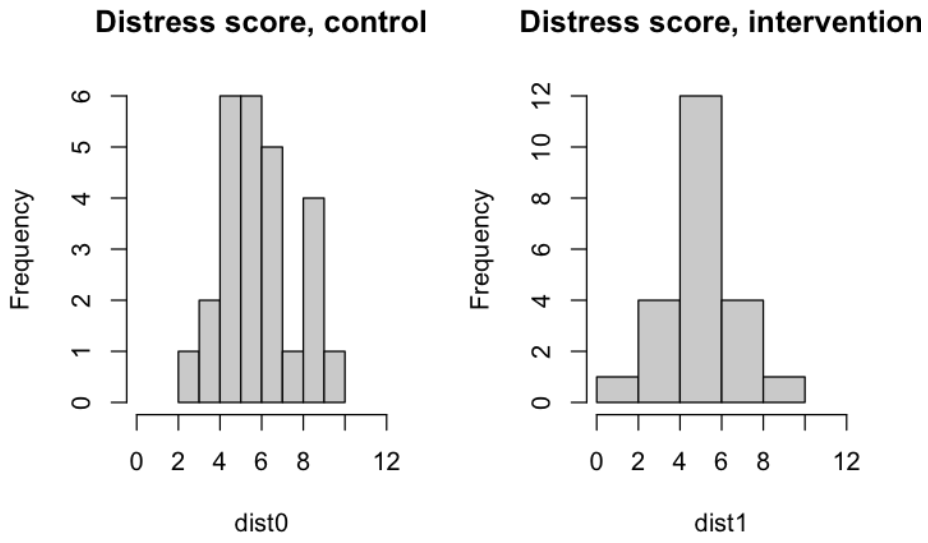
6.34615384615385
5.45454545454545

```
[1] "Sample SDs: "
```

1.87493589634009
1.8702501163843

```
[1] "Difference in sample means:"
```

-0.891608391608392



Statistical model: We will let $Y_{0,1}, ...,Y_{0,26}$ be random variables representing the emotional distress scores of the 26 sampled researchers in the control group and $Y_{1,1}, ...,Y_{1,22}$ be random variables representing the emotional distress scores of the 22 sampled researchers in the intervention group. So the first subscript denotes the group (0=control, 1=intervention) and the second is an index for the person (i=1, ..., 26 in the control group; i=1,...,22 in the intervention group).

We will assume that all random variables are independent of each other. The emotional distress scores in the control group are all drawn from the same normal distribution, with population mean μ_0 and population standard deviation σ. For now, we suppose σ is a known value, with $\sigma = 1.75$. The emotional distress scores in the intervention group are assumed to be drawn from a normal distribution with population mean μ_1 and the same population standard deviation.

This model can be compactly written as follows
$$ Y_{j,i} \overset{\small{iid}}{\sim} N(\mu_j, 1.75^2), \qquad i=1,2,...,n_j $$
Data: We will let $y_{0,1}, ...,y_{0,26}$ and $y_{1,1}, ...,y_{1,22}$ represent the realised values of these random variables (i.e. the observed emotional distress scores).

Estimand, estimator and estimate: The population parameter (estimand) we are interested in is:
$$ \delta = \mu_1 - \mu_0 $$
The obvious estimator for this is the sample difference in means:
$$ \hat{\delta} = \bar{Y_1} - \bar{Y_0} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1,i} - \frac{1}{n_0} \sum_{i=1}^{n_0} Y_{0,i} $$
And the sample estimate is:
$$ \hat{\delta} = \bar{y_1} - \bar{y_0} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1,i} - \frac{1}{n_0} \sum_{i=1}^{n_0} y_{0,i} = 5.4545 - 6.3462 = -0.892 $$
Null and alternative hypotheses: The null hypothesis is that there is no difference in the population mean emotional distress score in the intervention and control groups. Formally, we write:
$$ H_0: \delta = 0 $$
The alternative hypothesis (sometimes written H_1 or H_A) is that the null hypothesis is not true:
$$ H_1: \delta \neq 0 $$
In our sample, we have seen that $\hat{\delta} = -0.892$. So the sample mean emotional distress score is lower in the intervention group (which is the direction we might be hoping for, since this group have received a form of counselling to reduce their emotional distress). However, the two sample means are very unlikely to be exactly equal, even if the true

mean emotional distress score is the same in the two groups, due to sampling variability (i.e. due to random chance). So how should we interpret this sample difference in means? Does it constitute evidence against our null hypothesis?

In order to answer this question, we need to consider the sampling distribution of the difference in means. Unlike in previous sessions, where we used the sampling distribution to obtain confidence intervals, we are now interested in a subtly different sampling distribution: the sampling distribution that we would see *if the null hypothesis were true*.

8.1.3 Sampling distribution for the difference in sample means

Under the statistical models above, if σ is a known value it is straightforward to derive the sampling distribution of the estimator (the difference in sample means between groups).

> Linear combinations of independent normal distributions are also normal

Thus the distribution of $\hat{\delta}$ is normal. We can then calculate its expectation and variance using techniques from the Refresher to obtain:
$$ \hat{\delta} \sim N\left(\delta, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_0} \right) \right) $$
Substituting in the values $\sigma = 1.75$, $n_0 = 26$ and $n_1 = 22$, we have
$$ \hat{\delta} \sim N\left(\delta, 1.75^2 \left(\frac{1}{22} + \frac{1}{26} \right) \right) = N(\delta, 0.507^2) $$
So the expectation of the sampling distribution is the true population value δ and the standard error is 0.507. Remember that because we are considering the distribution of an estimator, we call the standard deviation of the estimator (the square root of the variance) the standard error.

Sampling distribution under the null hypothesis

We are interested in the distribution of the difference in sampling means would look like under repeated sampling *if the null hypothesis were true*. The null hypothesis states that $\delta = 0$. Therefore, under the null hypothesis,
$$ \hat{\delta} \sim N(0, 0.507^2) $$
The graph below shows this distribution. The code to draw the graph is suppressed, since it is not the focus here, but can be viewed by clicking the button.

```
# Sample difference in means
delta.hat <- mean(dist1) - mean(dist0)

# Randomly generate 10000 sample differences in means (following the sampling
distribution under the null hypothesis)
sample.diff.means <- rnorm(10000, 0, 0.507)

# Draw the approximate sampling distribution with the percentile confidence limits
marked in red
options(repr.plot.width=6, repr.plot.height=5)
hist(sample.diff.means, freq=FALSE, main="Sampling distribution for \n difference in
sample means, \n UNDER THE NULL HYPOTHESIS", xlab="Difference in sample means",
xlim=c(-2.5, 2.5), ylim=c(0, 0.8))
lines(seq(-2.5, 2.5, 0.025), dnorm(seq(-2.5, 2.5, 0.025), 0, 0.507))
abline(v=0, col="red")
abline(v=delta.hat, col="green", lty=2)

text(1.5, 0.2, "Null value")
text(-1.7, 0.4, "Observed value")
lines(c(0, 1.5), c(0.005, 0.18))
lines(c(-1.8, -0.9), c(0.38, 0.28))
```

The sampling distribution above shows us the distribution of the differences in sample means that we could have seen under repeated sampling, i.e. if we had done the same study a very large number of times. The question we must ask now is: is the value we have seen consistent with this sampling distribution? Or is it surprising? A "surprising" result is taken as evidence against the null hypothesis. In order to clarify these ideas, consider two scenarios that could have happened.

Scenario 1: Suppose we had done exactly the same study, but had seen a difference in sample means of $\hat{\delta} = -3.5$ (i.e. the intervention group sample mean score was 3.5 units lower than the control group mean).

Scenario 2: Suppose we had done this study, but had actually seen a difference in sample means of $\hat{\delta} = 0.02$.

What would we conclude in these scenarios? The graph below superimposes these two scenarios on the sampling distribution.

```
# Draw the approximate sampling distribution with the percentile confidence limits
marked in red
options(repr.plot.width=6, repr.plot.height=5)
hist(sample.diff.means, freq=FALSE, main="Sampling distribution for \ndifference in
sample means, \nUNDER THE NULL HYPOTHESIS", xlab="Difference in sample means",
ylim=c(0, 0.8), xlim=c(-3.5, 3.5))
lines(seq(-3.5, 3.5, 0.025), dnorm(seq(-3.5, 3.5, 0.025), 0, 0.507))
abline(v=0, col="red")
abline(v=delta.hat, col="green", lty=2)
abline(v=-3.5, col="orange", lty=2)
abline(v=0.02, col="blue", lty=2)

text(-2.5, 0.42, "Scenario 1: \nObserved value")
text(2, 0.22,  "Scenario 2: \nObserved value")
text(-1.9, 0.62, "Actual data: \nObserved value")
lines(c(0.05, 1.5), c(0.005, 0.15))
lines(c(-2.5, -3.4), c(0.35, 0.2))
lines(c(-1.8, -0.9), c(0.55, 0.48))
```

Scenario 1 We can see from the histogram that, under the null hypothesis, the probability of seeing a difference in sample means of -3.5 or less is incredibly low. In fact, this probability is less than 1 in 10,000. So if we did 10,000 studies we would expect only one of them to have a difference in sample means of -3.5 or less.

- So, have we been very unlucky and picked a very very unusual sample by chance?
- Or is our initial premise incorrect? Is the null hypothesis wrong?

This particular sample mean difference appears to be inconsistent with our null hypothesis. We interpret these "surprising" sample statistics as evidence against the null hypothesis.

Scenario 2 Again, the histogram shows quite clearly that, under the null hypothesis, many of the samples that we could have obtained would give us a sample mean difference close to zero. So this sample difference is completely consistent with the null hypothesis.

In this case, we would conclude that there is no evidence against the null hypothesis.

Our actual observed data Our observed sample mean difference (-0.892) is somewhere in between. In fact, we can calculate the probability of observing a sample mean difference of -0.892 or lower (i.e. the proportion of the area of the histogram that lies to the left of -0.89): this turns out to be 4%. So under repeated sampling, if our null hypothesis is true and there is truly no difference between the mean emotional distress score in the intervention and control groups, then we would expect to see a difference at least this big 4% of the time.

In fact, we typically consider the proportion of samples in which we would get an estimate at least as extreme as the one we did get *in either direction*. In our case, this is the probability of seeing a sample mean difference of less than -0.892 or greater than +0.892. Under the null hypothesis, approximately 8% of samples would produce a sample mean difference at least as extreme as the one we have seen in our sample.

So we had around a 1 in 13 chance of ending up with this result, if the null hypothesis is true. We interpret this as weak evidence against the null hypothesis.

8.2 The p-value

The p-value is defined as the probability of observing the sample estimate or a more extreme one (in either direction) given that the null hypothesis is true.

The smaller the p-value, the lower the chance of getting a difference as big as the one observed if the null hypothesis is true.

Therefore, the smaller the p-value, the stronger the evidence against the null hypothesis.

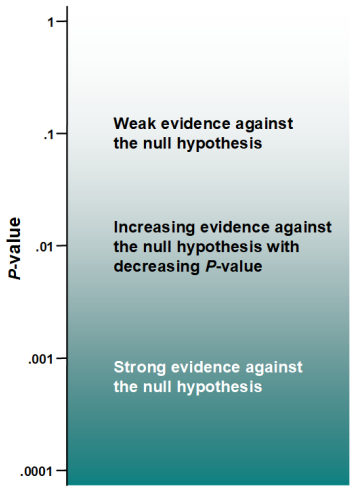


Fig. 3 Interpretation of p-values (taken from Sterne & Davey-Smith)

The value of 0.05 has historically been used as a cut-off, with values of $p<0.05$ deemed "statistically significant" and values of $p\geq 0.05$ "not significant". As discussed further in a later sub-section, we do not recommend dichotomising p-values in this way.

> Note that:
>
> - We have described what is called a *two-sided test*. Occasionally, a *one-sided test* might be used, where the p-value is the probability of results that are as extreme, or more extreme, *in the same direction* as the observed result. However, it is rare that it is justifiable to ignore sample statistics in one direction, so we will focus on two-sided tests.
>
> - When the sampling distribution is not symmetric, it can be hard to define what is *as extreme as* the estimate we have seen. In this case, there are various ways of obtaining the two-sided p-value. We do not pursue this further.

8.3 Connection between p-values and confidence intervals

Recall that we previously used the following fact:

> For a normal distribution, approximately 95% of observations are contained within 1.96 standard deviations of the mean.

Which, applied to sampling distributions, tells us that:

> For a normally distributed sampling distribution that is centred around the true population value, 95% of the estimates obtained under repeated sampling would be contained within 1.96 standard errors of the true population value

Applying this to the estimator $\hat{\delta}$, this leads to a 95% confidence interval of
$$\hat{\delta} \pm 1.96 \times SE(\delta)$$
The graph below shows some possible values of $\hat{\delta}$, along with their 95% confidence intervals. We see that:

- if $\hat{\delta}$ is exactly equal to the number $1.96 \times SE(\delta)$ then the 95% confidence interval just touches zero.

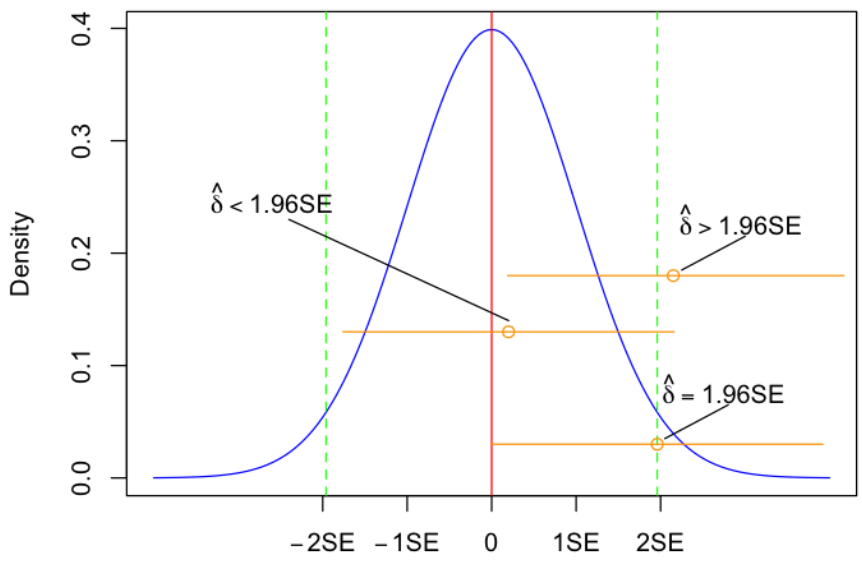
- if $\hat{\delta} > 1.96 \times SE(\delta)$ then the 95% confidence interval does not include zero - the whole interval lies above zero.
- if $0 < \hat{\delta} < 1.96$ then the 95% confidence interval does include zero.

So what p-values would these values of $\hat{\delta}$ result in?

- if $\hat{\delta} = 1.96 \times SE(\delta)$ then we know that 2.5% of the estimates lie above that point, so p=0.05.
- if $\hat{\delta} > 1.96 \times SE(\delta)$ then fewer than 2.5% of estimates lie above $\hat{\delta}$, so p<0.05
- if $0 < \hat{\delta} < 1.96 \times SE(\delta)$ then more than 2.5% of estimates lie above $\hat{\delta}$, so p>0.05

This leads us to the connection between 95% confidence intervals and p-values. When a 95% confidence interval and p-value are obtained from the same sampling distribution (which is typically the case when both are presented),

P-value	**95% confidence interval**
<0.05	Excludes the null value
≥ 0.05	Contains the null value



8.4 Other (mis-)interpretations of p-values

8.4.1 P-values as decision rules

Traditionally, hypothesis tests have been thought of as a means to make decisions. In this paradigm, a cut-off (typically p<0.05) is chosen. If the p-value is smaller than the chosen cut-off, the null hypothesis is rejected. If the p-value is above the cut-off then the null hypothesis is accepted. This leads to the terminology of:

- "Type I error", rejecting the null hypothesis when it is true
- "Type II error", accepting the null hypothesis when it is false

Linked to this approach is the habit of labelling p-values < 0.05 as "significant" and those larger as "non-significant".

There are some instances where this decision-making paradigm seems appropriate. Some health data science research is indeed concerned with decision making. For example, we may wish to carry out a trial to assess whether a particular clinical decision support system improves the clinicians' ability to detect malignant tumours. However much health data science research is not, at least directly, concerned with decision making. For example if we carry out an epidemiological study in which we relate risk of a particular disease to gender, we do this because we are interested in understanding

the aetiology of the disease, not because we want to assess whether to modify gender! For this reason many researchers regard p-values as a measure of strength of evidence against the null hypothesis, rather than as an aid to decision making.

In general, we do not advocate any approach which dichotomises p-values. There is very little difference, in terms of the information contained about the population parameter, between the two p-values of $p=0.049$ and $p=0.051$. Therefore it seems counter-intuitive to make very different decisions based on these p-values.

P-values represent an area of substantial philosophical controversy in statistics. We choose to interpret the p-value as a measure of strength of evidence against the null hypothesis. It should, however, be pointed out that some statisticians advocate strongly against this interpretation.

In much health data science research, we are interested in knowing more about a particular population parameter. Many health data scientists, therefore, choose to focus on obtaining and interpreting estimates and confidence intervals rather than calculating p-values.

8.4.2 Misinterpretations of p-values

The p-value is the subject of a lot of argument, debate and controversy, both within the statistical world and beyond. The following warn against some common misinterpretations and mis-uses of p-values:

> Do not:
> - believe that an association or effect exists just because it was statistically significant.
> - conclude that an association or effect is absent just because it was not statistically significant.
> - base conclusions solely on whether an association or effect was statistically significant or not.
> - conclude anything about scientific or practical importance based on statistical significance (or lack thereof).
> - interpret a p-value as the probability that chance alone produced the observed association or effect or the probability that the null hypothesis is true.

Importantly, statistical significance was never meant to imply scientific or clinical importance. As well as the p-value, always consider the estimated effect of the population parameter of interest and its confidence interval. These will often provide more insight than the p-value alone.

8.5 Calculating p-values

8.5.1 Example: Calculation of the p-value

In the emotional distress example, our difference in sample means is $\hat{\delta} = -0.892$. We are interested in the distribution of the difference in sampling means would look like under repeated sampling *if the null hypothesis were true*. The null hypothesis states that $\delta = 0$. Therefore, under the null hypothesis,
$$ \hat{\delta} \sim N(0, 0.507^2) $$
The easiest way to do this calculation is to standardise the estimator to follow a standard normal distribution, i.e.
$$ Z = \frac{\hat{\delta}}{0.507} \sim N(0, 1) $$
In our sample, we get a value of $Z=-0.892/0.507 = -1.76$. The p-value is defined as
$$ p = Pr(| \hat{\delta} | \geq -0.892) = Pr(| Z | \geq 1.76) $$
The standard normal distribution is symmetric, so this is equal to $2 \times P(Z \geq 1.76)$. This probability can be looked up using pre-calculated tables stored in all standard statistical software.

```
# Manual calculation of p-value:
2*(1-pnorm(1.76))
```

0.0784078065749654

8.5.2 Approximate tests in large samples

More generally, suppose that the random variable used to calculate our p-value (above, the random variable was the difference in sample means) is denoted by R and that it has an expected value and variance (under the null hypothesis) denoted by $E(R)$ and $Var(R)$. Then define:
$$ Z = \frac{R - E[R]}{\sqrt{Var(R)}} = \frac{R - E[R]}{SE(R)} $$

where $SE(R)$ is the standard error of R (the standard deviation of the sampling distribution; alternatively the square root of the variance of R). To simplify this even further, in many cases, as for the difference in sample means, $E(R) = 0$.

Thanks to the Central Limit Theorem, in almost all situations, as the sample size n becomes large, the distribution of Z tends towards a standard normal distribution.
$$ lim_{n\rightarrow \infty} \ \ Z \sim N(0, 1). $$
The standard normal distribution can then be used to calculate the two-sided p-value, as above.

8.5.3 The two-sample t-test

Let us return to the comparison in population means between two groups. When, as is more typical, we do not know the value of σ, we need to replace it with an estimate from our sample, $\hat{\sigma}$. Typically we use an estimate based on the sample standard deviations in the two groups, s_1 and s_0:
$$ \hat{\sigma}^2 = \frac{(n_1 - 1) s_1^2 + (n_0 - 1) s_0^2}{n_1 + n_0 - 2} $$
For our sample of data, $\hat{\sigma} = 1.873$. The sampling distribution we used above involves the true population standard deviation
$$ \hat{\delta} \sim N\left(\delta, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_0} \right) \right) $$
Similarly, the equivalent version of the sampling distribution (which we will find it easier to modify for our current purposes), is also no longer exactly true:
$$ \frac{\hat{\delta} - \delta}{\sigma \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_0}\right) }}\sim N(0,1) $$
This is only approximately true if we substitute the sample estimate $\hat{\sigma}$ into the equation. A little more algebra (not shown here), however, gives us an exact distribution.
$$ \frac{\hat{\delta} -\delta}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \sim t_{n_1 + n_0 - 2} $$
Under the null hypothesis, $\delta = 0$, giving
$$ T = \frac{\hat{\delta}}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \sim t_{n_1 + n_0 - 2} $$
Substituting in the numbers from our sample of data,
$$ T = \frac{-0.892}{1.873 \sqrt{\frac{1}{22} + \frac{1}{26}}} $$
gives $t = -1.644$ (remembering that T is the random variable and t here is the realised (observed) value of that statistic). T-distributions are symmetric around zero, so we take *at least as extreme as* to mean less than -1.64 or greater than +1.64, which in turn is twice the probability of being less than -1.64. We simply need to calculate this probability for a t-distribution with 46 degrees of freedom (where we obtained 46 as $n_1 + n_0 - 2$).

The code below performs this calculation and then uses an inbuilt R package to obtain the same p-value.

```
# Manual calculation of p-value (two equivalent calculations)
2*pt(-1.644, 46)

# Read in data (emotional distress scores in control and intervention group)
dist0 <- c(5, 2, 5,  7,  6,  7,  7,  5,  8,  6,  6,  9,  4,  5,  9,  7,  9,  5,  6,
10,  9,  4,  6,  6,  5,  7)
dist1<- c(5,  5,  6,  6, 1,  5, 10,  7,  3,  6,  7,  8,  6,  7,  5,  4,  5,  6,  4,
6,  3,  5)

# T-test using inbuilt R package
dist <- c(dist0, dist1)
gp <- c(rep(0, 26), rep(1, 22))

t.test(dist~gp, var.equal=TRUE)
```

0.106994541315052

```
        Two Sample t-test

data:  dist by gp
t = 1.6435, df = 46, p-value = 0.1071
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2004223  1.9836391
sample estimates:
mean in group 0 mean in group 1
       6.346154        5.454545
```

Rounding to 2 decimal places, the p-value is 0.11.

In the output from the R package, the line

```
t = 1.6435, df = 46, p-value = 0.1071
```

tells us that the value of the statistic T above is $t=1.64$ in this sample, the degrees of freedom tell us that we are looking at a t-distribution on 46 degrees of freedom. We are also given a 95% confidence interval for the population difference in means: (-0.20 to 1.98). As we noted above, when the p-value is >0.05 then the null value (here, zero) will be included in the 95% confidence interval.

8.5.4 Other hypothesis tests

You will meet many types of hypothesis tests over your statistical studies. Many, like the t-test above, are constructed around a particular estimator and so there is a nice connection between the estimate, the 95% confidence interval and the p-value from the hypothesis test. Where this is the case, it is good practice to present the estimate and confidence interval alongside the p-value, since they contain much more information than the p-value alone.

In other cases, tests can be constructed without a specific parameter being estimated. The chi-squared test is a very commonly-used test. It tests the null hypothesis of no association between two unordered categorical variables. This test does not directly invoke the sampling distribution of an estimator, so typically only the p-value is presented, rather than also presenting an estimate and confidence interval.

In general, hypothesis testing is a controversial and widely misunderstood area of frequentist statistics. Where possible, focusing on estimating parameters along with confidence intervals can avoid some of the more damaging misuses of p-values.

Further resources

Note: further resources are for you to deepen your understanding of the subject if you wish to do so. This is entirely optional. All examinable material is contained within the notes.

Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25(4):225-230. doi:10.1007/s10654-010-9440-x

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “p < 0.05”, The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

Sterne JA, Davey Smith G. Sifting the evidence-what’s wrong with significance tests? BMJ. 2001;322(7280):226-231. doi:10.1136/bmj.322.7280.226

9. Bayesian Statistics I

So far in this module, we have looked at frequentist or classical statistical ideas, such as maximum likelihood estimation, hypothesis testing and p-values. Underlying the frequentist approach is the belief that there is a true state of reality, and that parameters have a fixed and true value. Probabilities are long-run frequencies; for example, the probability of a driver in London having a car accident is a fixed value between 0 and 1. A typical way of estimating this value is to take a sample of observations, construct a likelihood function for these observations, and to obtain the parameter value that maximizes the likelihood. When we take a Bayesian approach, the parameter we wish to estimate is considered to be a random variable, and probabilities may represent a subjective belief about the state of uncertainty, or there may be a data generating distribution underlying the random parameter. For example, you may have a prior belief about the probability of a driver in London having a car accident, and after collecting a sample of data, you combine your prior beliefs with the likelihood for those observations to construct an updated belief - the posterior. Your belief may change in light of the data.

Bayesian methods sometimes require numerical integration, and cheaper computing has made Bayesian approaches more feasible in the last 20 years. Bayesian approaches are likely to be an important part of working in Health Data Science. In the next two sessions, we introduce the fundamental principles.

The current session introduces the basic concepts underlying Bayesian inference and then applies the basic principles to a simple example using proportions.

Intended learning outcomes

By the end of this session you will be able to:

- compare the notions of probability and likelihood in Bayesian and Frequentist paradigms
- explain the notions of prior and posterior distributions
- apply Bayes Theorem in the discrete case
- Understand and apply the basic principles of Bayesian analysis using proportions, specifically:
    - use the beta distribution as a prior and derive the posterior distribution
    - obtain credible HPD intervals for the parameter
    - obtain prior and posterior predictive distributions
    - explain the concept of conjugate priors

9.1 Introduction to Bayesian Inference

9.1.1 Probability

In Session 2, we learned about probability in the frequentist sense: the proportion of times an event occurs in the long-run. Let's have a look at the following two scenarios:

1. A research group wishes to know the probability that a baby who is born in a particular hospital ward has cystic fibrosis. They look at the records on screening tests done at birth to investigate.
2. A 34 year old woman attends her GP practice, worried that she has cancer because she has had feelings of "fullness" and "bloating" as well as mild nausea for the last 2 weeks. The patient mentions ovarian, bowel and pancreatic cancer as concerns having read about her symptoms on the internet. The rest of the history as well as physical examination are unremarkable. If the GP's assessment of the risk were above a certain level, the GP might refer the patient for tests (collect more data). In this case, the GP concludes that the current information about the patient suggests there is a very low risk that the patient has cancer.

> What is the quantity that we trying to estimate in each scenario?
> What is the frequentist definition of probability in each of these settings? Does it make sense?

A key problem with the frequentist paradigm is that the "long-run" frequency definition is not always relevant, or even appropriate, as we see in the second example above. Further, notice that the GP uses information from different sources to draw his/her conclusion about the probability that the patient has cancer. This synthesis of information can be incorporated into a Bayesian framework. A frequentist, in contrast, would tackle this problem by thinking about:

> a) the probability of the patient having these symptoms, given that she has cancer;
> b) the probability of the patient having these symptoms, given that she does not have cancer;

and comparing the two probabilities. Note that this does not take into account the extra information about the context.

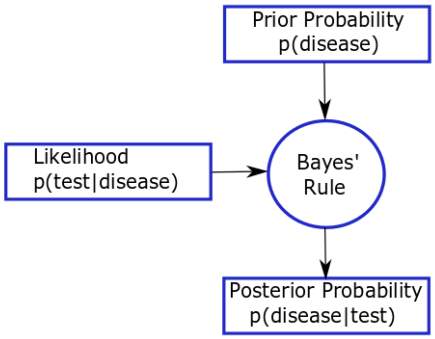
9.1.2 Bayesian Inference

The underlying concept for Bayesian inference essentially works as follows. We have some population parameter θ which we wish to make inference on, and the likelihood $p(y|\theta)$ which tells us how likely different values of y are, conditional on different parameter values θ. In the frequentist approach, θ is considered to be a fixed, but unknown, constant. Inference is then based on the likelihood $p(\mathbf{y}|\theta)$, where $\mathbf{y} = \left\{y_1, \ldots ,y_n\right\}$ is a sample of observations from the population. The frequentist approach looks at the distribution of the data given θ to estimate θ by using, for example, the maximum likelihood approach which we covered in Session 6.

In the Bayesian paradigm, we no longer assume that the parameters have a fixed true value, but consider θ to be a random quantity with an unknown distribution, which we wish to estimate. This distribution is denoted by $p(\theta|y)$, and so we look at the distribution of the parameter, having seen data y. To achieve this, we will have to specify a prior probability distribution, denoted $p(\theta)$, which represents our initial beliefs about the distribution of θ prior to observing any data. In some situations, when we are trying to estimate a parameter θ we have some knowledge, about the possible value of θ before we take into account the data that we observe.

For example, consider the way a physician makes diagnostic decisions. A patient presents with a set of symptoms, concerned that they might have a certain disease. The physician assesses the probability that this patient has this disease, based on symptoms, family history, alternative explanations of symptoms and prevalence of the disease (their prior view that the patient has the disease). The physician might send the patient for a diagnostic test (collects some data) if her prior assessment of risk is above some threshold. Then the physician re-assesses the chance that the

patient has this disease, taking account of the results and reliability of the diagnostic test (updates their prior in light of the data to get a posterior view on whether the patient has the disease). Depending on their certainty, the physician may then send the patient for further diagnostic tests. This thought process can be represented by the figure below and is analogous to Bayesian thinking.



In this example, the physician is assessing the probability that the patient has the disease. It is the physician's prior probability based on their own training, knowledge and experience; a colleague may have a different prior probability. Here, prior probability is being defined subjectively. The size of the probability represents the physician's degree of belief about the occurrence of an event, i.e. their own personal assessment of how likely an event is, based on the evidence available to them before the test results are given. This definition corresponds more closely to the everyday, intuitive usage of probability than a frequentist interpretation (where the probability of a particular event occurring can be interpreted as the proportion of times the event would/does occur in a large number of similar trials or situations). The prior probability of the event might come from direct data, known prevalance of disease in a population, or data from related populations. If such prior information does not exist, then it can be formally elicited from experts, but we would want to acknowledge the uncertainty in the experts' knowledge.

9.2 Bayes Theorem (recap)

Let's remind ourselves of Bayes theorem for discrete events, which we met in Session 2 (probability):

If A and B are events, then
$$ P(A|B) = \frac{ P(B|A) P(A) } {P(B)} \propto P(B|A) P(A), $$
or in words:
$$ \mbox{posterior probability of A given B} \propto \mbox{the likelihood of B given A} \times \mbox{the prior probability of A}. $$
Also, if A_i is a set of mutually exclusive and exhaustive events, i.e. $ p(\bigcup\limits_i A_i) = \sum\limits_i p(A_i) = 1$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then
$$ p(A_i|B) = \frac{ p(B|A_i) p(A_i) } {\sum\limits_j p(B|A_j) p(A_j) }. $$
The calculation of the denominator is more difficult if we have continuous parameters as it requires integration over A; we will discuss this in the next section.

We will illustrate Bayes Theorem further with the diagnostic test example for Covid-19 below. We see Bayesian reasoning is purely probabilistic. Bayes theorem gives us a principled way to update prior probabilities on the basis of new data.

9.2.1 Example

Watson (2020) discusses some interesting issues around the interpretation of Covid-19 diagnostic tests. Typically, a clinician estimates a pre-test probability (a prior probability) of having Covid-19 for a particular area, which is derived from knowledge about local rates of Covid-19. Then, given a patient's test result, the post-test probability (the posterior probability) of having Covid-19 is obtained. The posterior probability depends on the pre-test probability, as well as the sensitivity and specificity of the test, which are difficult to estimate; often, sensitivity is over-estimated. The article discusses how one can be fairly confident about a positive test result, but more caution is needed for a negative test result, as there may still be quite a high chance that a person has Covid-19. We illustrate this with Bayes' theorem.

Suppose that, in a student hall of residence, the prevalence of Covid-19 if you have a persistent cough is 75%. Suppose we assume that the test will be positive in Covid-19 patients 70% of the time (sensitivity is 0.7), and it will be negative in non-Covid-19 patients 95% of the time (specificity is 0.95). Given that a student in this hall with a

persistent cough tests negative, what is the probability that they have Covid-19? In other words, what is the probability of a false negative?

Let us denote by $C+$ the event that a person has Covid-19, and $C-$ the event that a person does not have Covid-19. Further we denote by $T+$ and $T-$ the events that a person has a positive and a negative test, respectively. The information we are given is that:
$$ p(C+)=0.75, \qquad p(T+|C+)=0.70, \qquad p(T-|C-)=0.95 $$
Now, what we want is:
$$\begin{split} \begin{align} p(\mbox{false negative}) &= p(C+|T-) = \frac{p(T-|C+)p(C+)}{p(T-)} \\ &= \frac{p(T-|C+)p(C+)}{p(T-|C+)p(C+) + p(T-|C-)p(C-)} \\ &= \frac{(1-0.7) \times 0.75}{(1-0.7) \times 0.75 + 0.95 \times 0.25} \\ &= \frac{0.225}{0.4625} \\ &= 0.4864 \end{align} \end{split}$$
You can see that, despite the negative test result, due to the very high prevalence of Covid-19 in the hall of residence and the relatively low sensitivity rate, there is still a 48.64% chance that a person has Covid-19.

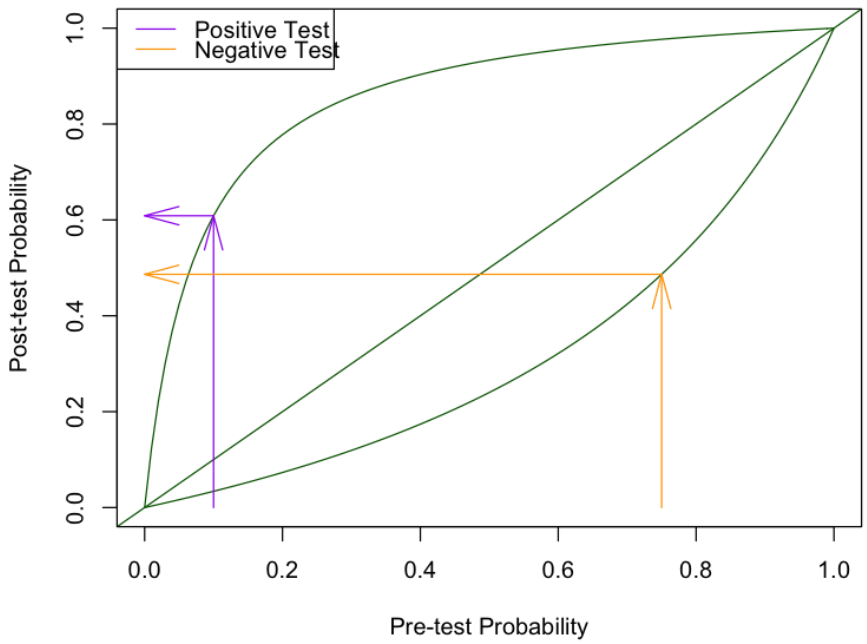
Suppose a different student has no symptoms. The prevalence of Covid-19 in asymptomatic people is 0.1. They use the same diagnostic test and the test result is positive. What is the probability that this student with a positive test result has Covid-19? In other words, what is $p(C+|T+)$?

Solution:
$$\begin{split} \begin{align} p(C+|T+) &= \frac{p(T+|C+)p(C+)}{p(T+)} \\ & = \frac{p(T+|C+)p(C+)}{p(T+|C+)p(H+) + p(T+|C-)p(C-)} \\ &= \frac{0.7 \times 0.1}{0.7 \times 0.1 + (1-0.95) \times 0.9} \\ &= \frac{0.07}{0.115} \\ &= 0.609 \end{align} \end{split}$$
This means that, amongst all the people who test positive, 60.9% will actually have the disease. After a positive result from a test, the probability that you have Covid-19 increase from 10% to 61%.

Note that these results are specific to the the prevalence of Covid-19 in the area, as well as the sensitivity and specificity of the diagnostic test. The code below reproduces the leaf-plot from Watson (2020). The x-axis is the pre-test probability of having Covid-19. The corresponding y-values on the lower curve (lower leaf) are the post-test probabilities of having Covid-19, following a negative test result. The corresponding y-values on the upper curve (upper leaf) are the post-test probabilities of having Covid-19, following a positive test result. The correponding values on the diagonal ($y=x$) line represent probabilities if no test is carried out.

In our first example, the prevalence in symptomatic people is 0.75, so we follow the orange arrows to find that the post-test probability after a negative result 0.4864. In the second example, the prevalence in asymptomatic people is 0.1. We follow the purple arrows to find that the post-test probability after a positive result is 0.609. How do you think the shape of the lower and upper leaves would change, if sensitivity was higher? If specificity was lower? Re-run the code with different values to check.

```r
# Function takes as arguments the sensitivitiy of the test (sensi)
# and the specificity (speci)

leafplot <- function(sensi, speci){

  pretest <- seq(0, 1, 0.01) #possible pre-test probabilities

  #probability of having Covid-19 after a positive test result
  pos.test <- sensi*pretest/(sensi*pretest+(1-speci)*(1-pretest))

  #probability of having Covid-19 after a negative test result
  neg.test <- ((1-sensi)*(pretest))/((1-sensi)*pretest+speci*(1-pretest))

  #plot leaves
  plot(pretest, pos.test, type="l", col="darkgreen",
     xlab="Pre-test Probability", ylab="Post-test Probability")
  points(pretest, neg.test, type="l", col="darkgreen")
  abline(a=0, b=1, col="darkgreen")
  legend("topleft", legend=c("Positive Test", "Negative Test"),
        col=c("Purple", "Orange"), lty=1, bg="transparent")

  #plot arrows
    #we use pretest[11] to get the prevalence value of 0.1, and
    #pretest[76] to get the prevalence value of 0.75 in the vector "pretest"

  arrows(pretest[11], 0, pretest[11], pos.test[11], angle=15, col="purple")
  arrows(pretest[11], pos.test[11], 0, pos.test[11], angle=15, col="purple")
  arrows(pretest[76], 0, pretest[76], neg.test[76], angle=15, col="orange")
  arrows(pretest[76], neg.test[76], 0, neg.test[76], angle=15, col="orange")

  }

options(repr.plot.width=6.5, repr.plot.height=5.5)
leafplot(sensi=0.7, speci=0.95)


#See what happens to the plot when you change sensitivity and specificity!
#leafplot(0.95, 0.8)
```



9.3 The Bayesian paradigm in Health data science problems.

In this section we discuss the Bayesian approach in Health data science problems. Some features of the Bayesian paradigm are particularly useful in this context:

1. Bayes theorem provides a statistically principled method for combining data. Thus, we can take into account the context within which the data are generated. For example, results of a diagnostic test may have a different interpretation/consequence if used in a symptomatic patient than in a general screening programme. The prior

probability of disease would be higher in the former than the latter. Priors can then be updated by the test result to give an assessment of disease risk specific to the local prevalence.

2. For problems where there are multiple or diverse sources of data which must be combined, the Bayesian framework provides a natural environment for doing so. Examples where Bayesian synthesis of information is common are:
   • models of biological systems, for example genetic and genomic pathways,
   • models of the natural history of diseases over time and relationships with clinical events,
   • economic models of disease trajectories and cost-effect trade-offs for interventions that interrupt the trajectories,
   • ecological studies of pollutant emissions and effects on population health,
   • demographic studies, for example to study migration,
   • speech recognition software,
   • other pattern recognition models such as medical imaging or search engines,
   • epidemic modelling.
   In all these examples complex data is synthesised and/or used to update outputs.
3. Bayesian models fit well into decision theory methodology, providing we can also specify consequences of model outputs.
4. In many examples, especially those that aim to model complicated processes, some of the data inputs are very sparse, or even non-existent. In such cases, prior data may be formally elicited from an expert panel and incorporated in a Bayesian analysis. Examples include multiple evidence synthesis and identification of latent groups.
5. Bayesians are allowed to make direct probability statements about unknown quantities. Frequentists cannot make these direct probability statements because the unknown model parameters are assumed fixed.
6. In recent years the resources available to complete Bayesian analysis have increased, including bespoke software and packages within commercial statistical software.

But Bayesian methods are not that widely used in statistics compared with more classical approaches because they have some limitations.

1. Sometimes the need for a prior distribution is a barrier if little is known about a parameter and researchers fall back on priors that are weakly informative. In that case, it is not easy to see how much benefit comes from a Bayesian analysis.
2. Because of the need to use Bayesian updating via a prior distribution, the analysis almost always requires a parametric approach. This limits the structure of the analysis models. Although non-parametric Bayesian methods are available for some situations, they often have underlying parametric assumptions.
3. The numerical integration methods usually required for realistic problems are often computationally expensive. This is especially true if there are multiple sources of evidence to be combined.
4. Many statisticians are unfamiliar with the methods and associated software.

9.4 Bayes thorem for discrete and continous data

So far this session, we have looked at Bayes theorem in the discrete case. We turn to the more general case of Bayes thorem to make inference about an unknown parameter θ, which could be discrete or continuous.

The probability distribution for θ reflects our uncertainty about it before seeing the data, **prior distribution**, $p(\theta)$. Once the data data y is known, we condition on it. Using Bayes theorem we obtain a conditional probability distribution for unobserved quantities of interest given the data. If θ is continuous, we have:
$$ p(\theta \mid y)= \frac{ p(\theta)\, p(y \mid \theta)}{\int p(\theta)\,p(y \mid \theta)\,d\theta}, $$
and θ is discrete and takes values in the set Θ, we have:
$$ p(\theta \mid y)= \frac{ p(\theta)\, p(y \mid \theta)}{\sum_{\theta \in \Theta} p(\theta) p(y \mid \theta) }. $$
We call $p(\theta \mid y)$ the **posterior distribution**.

Note that the Bayesian approach is naturally synthetic in that it allows data from different sources to be combined, according to Bayes principles. This approach is most useful when there is informative prior information. We note that the Bayesian approach can be recursive, so $p(\theta \mid y)$ may be used as a prior when calculating $p(\theta \mid y, z)$ for a second data set z.

The denominator, ${\int p(\theta)\,p(y \mid \theta)\,d\theta}$ or $\sum_{\theta \in \Theta} p(\theta) p(y \mid \theta)$, is a constant with respect to θ. One of the challenges of using Bayesian approaches is that the integration can be analytically intractable, so that numerical methods are needed (for example, numerical integration or Markov Chain Monte Carlo methods). These methods are beyond the scope of the current module. In this introductory course, we will only look at examples where this constant need not be calculated, since the form of the posterior can be inferred by inspection once observing that the posterior is proportional to the product of the prior and likelihood:

$$p(\theta \mid y) \propto p(\theta)\,p(y \mid \theta).$$
We will see how this works for the inference of proportions.

9.5 Bayesian inference on proportions

Consider a new drug being developed for the relief of chronic pain. To find out about its efficacy, we propose to run a single-arm early-phase clinical trial in which we give this drug to a number n of randomly selected patients. Because patients are independent of each other, so it seems reasonable to model the data using the Binomial distribution, $Y\sim Bin(n,\theta).$ We have that $\theta\in [0,1]$ is the probability of pain relief (success) in each patient, and this is unknown. We then make the observation that there are y successes out of n independent trials. As a reminder, the probability distribution function of the Binomial distribution is:
$$ p \left(y \mid \theta \right) = {n \choose y} \theta^y (1-\theta)^{n-y}. $$
To proceed, we need to have a prior distribution for θ. Let us consider three possible prior distributions:

1. An uninformative prior, where all values of θ are equally probable.
   You essentially have no prior information about the effectiveness of the drug.
2. A symmetrical, concave prior that is centered at 0.5.
   You think that the drug is likely to be effective for patients around half of the time.
3. An asymmetrical prior with a spike at 0.1.
   You think that the drug is generally ineffective, and feel quite strongly about it.

Now, the Beta distribution is a flexible distribution that can represent each of these prior beliefs by appropriate choice of its parameters. It is also convenient because it has a similar form to the Binomial distribution.

9.5.1 The Beta prior

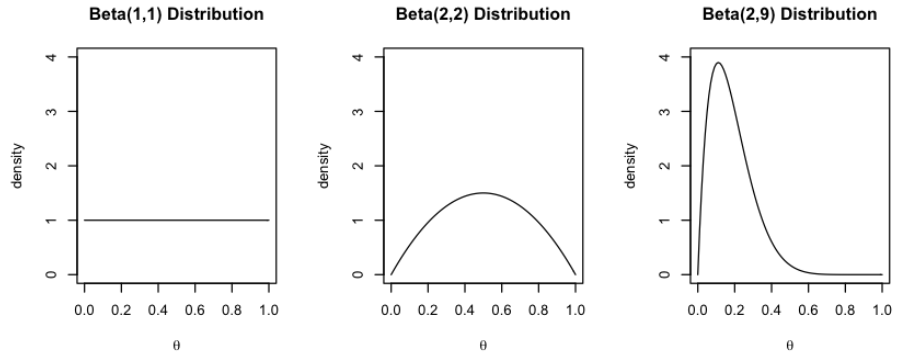
The Beta distribution is a flexible two parameter distribution that is restricted to the interval between 0 and 1, and so it is a reasonable form for a probability distribution for a proportion. The two parameters, a and b, are often called "shape" parameters. Given $\theta \sim \hbox{Beta}(a,b)$, the probability density function, expectation and variance of the distribution are as follows:
$$\begin{split} \begin{align} p(\theta|a, b) &= \frac{\Gamma(a+ b)}{ \Gamma(a)\Gamma(b) } \,\, \theta^{a-1} \; (1-\theta)^{b-1} \mbox{ where } \theta \in (0,1) \\ {E}(\theta |a, b) &= \frac{a}{a+ b } \\ {Var}(\theta |a, b) &= \frac{a b}{ (a+ b)^2 (a+ b+1)} \end{align} \end{split}$$
The *Gamma function* $\Gamma(x)$ is defined for positive integers as $\Gamma (x)=(x-1)!$, and has a more complex form for real numbers.

This prior distribution is very flexible. For example:

1. $a=1, b=1$ results in the uniform distribution
2. $a=2, b=2$ results in a symmetrical distribution centered on $p=0.5$
3. $a=2, b=9$ results in an asymmetrical distribution with a spike at $p=0.1$.

These are the priors we specified earlier; they are plotted below. Note that the higher the values of $a, b,$ the smaller the variance of the distribution.

```
options(repr.plot.width=7, repr.plot.height=3)
theta <- seq(0, 1, 0.01)
par(mfrow=c(1,3))
plot(theta, dbeta(theta, 1, 1), type="l", main="Beta(1,1) Distribution", ylim=c(0,4),
xlab=expression(theta), ylab="density")
plot(theta, dbeta(theta, 2, 2), type="l", main="Beta(2,2) Distribution", ylim=c(0,4),
xlab=expression(theta), ylab="density")
plot(theta, dbeta(theta, 2, 9), type="l", main="Beta(2,9) Distribution", ylim=c(0,4),
xlab=expression(theta), ylab="density")
```



9.5.2 Posterior

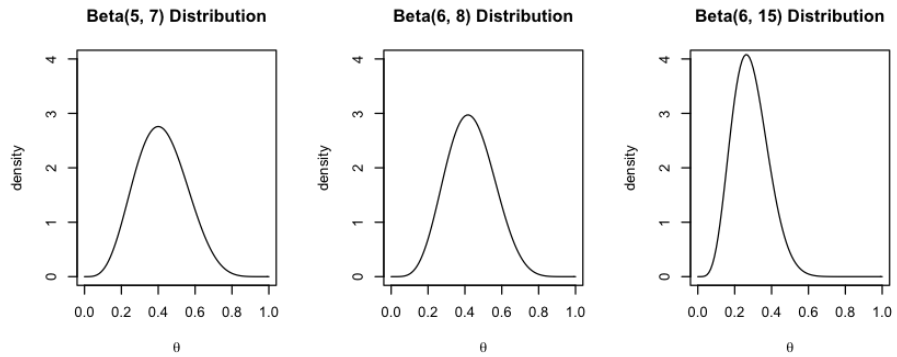
Now, we apply Bayes theorem to obtain the posterior distribution using a $Beta(a,b)$ distribution for the prior:
$$\begin{split} \begin{align} p(\theta \mid y) &= \frac{ p(\theta)\, p(y \mid \theta)} {\int p(\theta)\,p(y \mid \theta)\,d\theta}\\ &\propto p(\theta)\, p(y \mid \theta) \end{align} \end{split}$$
Substituting in the appropriate distributions gives
$$\begin{split} \begin{align} p(\theta \mid y) &= \frac{\Gamma(a+ b)}{ \Gamma(a)\Gamma(b) } \theta^{a-1} (1-\theta)^{b-1} {n \choose y} \theta^y (1-\theta)^{n-y} \\ &\propto \theta^{a-1} (1-\theta)^{b-1} \theta^y (1-\theta)^{n-y} \\ &\propto \theta^{a+y-1} (1-\theta)^{b+n-y-1} \end{align} \end{split}$$
Now by inspection, we can see that this is in the form of a Beta distribution: we have that the posterior is proportional to $\theta^{a+y-1} (1-\theta)^{b+n-y-1}$. In other words, the posterior is $Beta(a+y, b+n-y).$ This distribution has mean given by: $\frac{a+y}{a+b+n}$ and variance $\frac{(a+y)(b+n-y)}{(a+b+n)^2(a+b+n+1)}.$

Suppose the data we observe is $y=4$ successes out of a total of 10 patients. Then:

1. With the uniform $Beta(1,1)$ prior, our posterior is $Beta(5, 7)$.
2. With the symmetrical $Beta(2, 2)$ prior, our posterior is $Beta(6, 8)$.
3. With the asymmetrical $Beta(2, 9)$ prior, our posterior is $Beta(6, 15)$.

We plot the possible distibutions below:

```
p <- seq(0, 1, 0.01)
par(mfrow=c(1,3))
plot(p, dbeta(p, 5, 7), type="l",  main="Beta(5, 7) Distribution", ylim=c(0,4),
xlab=expression(theta), ylab="density")
plot(p, dbeta(p, 6, 8), type="l", main="Beta(6, 8) Distribution", ylim=c(0,4),
xlab=expression(theta), ylab="density")
plot(p, dbeta(p, 6, 15), type="l", main="Beta(6, 15) Distribution", ylim=c(0,4),
xlab=expression(theta), ylab="density")
```



We can see that the uninformative prior leads to the posterior with the highest variance amongst the three. The narrow prior in the third example shifts the posterior distribution to the right. We can see that different choices of prior lead to different results. For this reason, it is often recommended to repeat analyses with different priors to see how much the results change: this is called *sensitivity analysis*.

9.6 Summarising Posteriors

We often display the posterior distribution graphically to get a sense of the information that we have about the parameter. However, other ways to summarize the distribution can be helpful. We may also wish to summarise the posterior distribution by a credible interval.

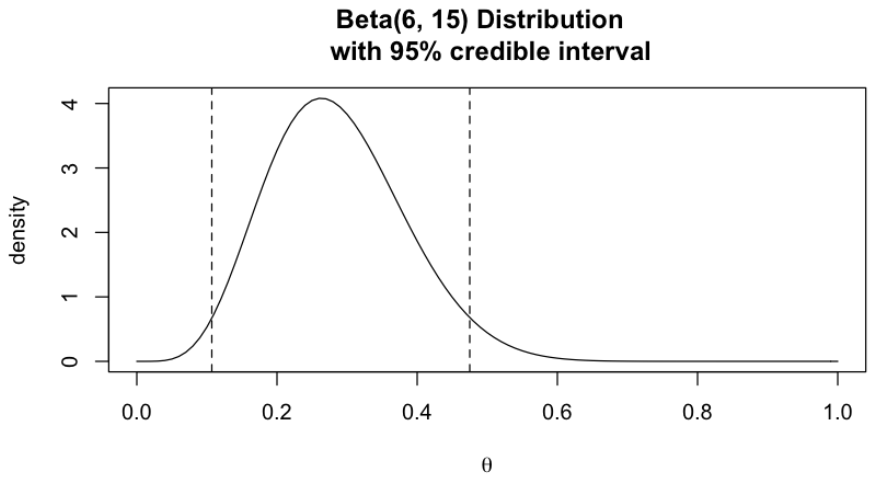
Remember that a classical 95% confidence interval is defined such that, if the data collection process is repeated again and again, then in the long run, 95% of the confidence intervals formed would contain the true parameter value.

A Bayesian 95% **credible interval** is an interval which contains 95% of the posterior distribution of the parameter.

There may be several different credible intervals such that the interval contains 95% of the distribution. The 95% **Highest Posterior Density (HPD)** interval is the credible interval with the smallest range of values for θ (providing the posterior is concave). Algebraically, this is the region $[\theta_L, \theta_U]$ that contains 95% of the probability, such that:
$$ P(\theta \in [\theta_L,\theta_U])= 0.95 \mbox{ such that for all } \theta_O \notin [\theta_L,\theta_U] \mbox{ and all } \theta_I\in[\theta_L,\theta_U], p(\theta_O|y) < p(\theta_I|y). $$
In our previous example, when we used the asymmetrical $Beta(2, 9)$ prior, our posterior was $Beta(6, 15)$. The posterior mean is $\frac{6}{6+15}=0.286$. The 95% HPDI is (0.107,0.475). We plot the distribution below and check that the area between these two values gives us 0.95. Now, note that the interval (0.09, 0.465) also gives us an area of 0.95, but this interval is wider. In a sense, the HPDI is the "tightest" interval so that the area under the posterior distribution is 0.95.

```
p <- seq(0, 1, 0.01)
options(repr.plot.width=7, repr.plot.height=4)
plot(p, dbeta(p, 6, 15), type="l", main="Beta(6, 15) Distribution  \n with 95% credible
interval",  xlab=expression(theta), ylab="density")
abline(v=0.475, lty="dashed")
abline(v=0.107,  lty="dashed")

#Area under the 95% HDPI
pbeta(0.475, 6, 15)-pbeta(0.107, 6, 15)

#The interval (0.09, 0.465) also a 95% credible interval
pbeta(0.465, 6, 15)-pbeta(0.09, 6, 15)
```

```
0.949975144544822
0.951266598161814
```



Note:

We have phrased the above discussion in terms of 95% confidence and credible intervals. However, there is nothing special about the level 95%. We can make the discussion more general by talking about $100(1-\alpha)\%$ confidence or credible intervals instead, with $\alpha \in (0,1)$ (where $\alpha = 0.05$ for 95% confidence or credible intervals but e.g. $\alpha = 0.01$ for 99% intervals).

9.7 Prior Predictions

Before observing a quantity y, we can provide its predictive distribution by integrating out the unknown parameter,
$$ p(y) = \int p(y|\theta) p(\theta) d\theta. $$

Predictions are useful in many settings, for example forecasting, cost-effectiveness models and design of studies. In the trial described earlier in this section, we had 10 patients. Suppose we are interested in predicting the number of patients who will have a positive response. Recall that the Beta distribution is a suitable prior distribution for θ, the proportion of positive responses. We have:
$$\begin{split} \begin{align*} \theta &\sim \hbox{Beta}(a,b) \\ y &\sim \hbox{Binomial}(\theta,n) \end{align*} \end{split}$$
The exact predictive distribution $p(y)$ can be computed analytically and is known as the *Beta-Binomial* distribution. It has the complex form with three parameters, number of trials n and shape parameters, a and b:
$$\begin{split} \begin{align*} p(y) &= \frac{ \Gamma (a+ b)}{ \Gamma (a) \Gamma (b) } {n \choose y} \frac{\Gamma (a+ y) \Gamma (b+n-y)}{\Gamma (a+b+n)} \\ E(y) &= n \frac{a}{a+b} \end{align*} \end{split}$$
Given that we use the asymmetrical $Beta(2, 9)$ prior, our predictive distribution would be:
$$ p(y) = \frac{ \Gamma (11)}{ \Gamma (2) \Gamma (9) } {10 \choose y} \frac{\Gamma (2+ y) \Gamma (19-y)}{\Gamma (21)}, $$
with $E(y) = 10 \, \frac{2}{11} = 1.81$. So, before observing any data, we would predict around 2 patients to have a positive response out of 10.

9.7.1 Posterior Prediction

Suppose that have observed y, and we want to predict future observations z, assuming that z and y are independent, conditional on θ. The posterior predictive distribution for z is given by,
$$\begin{split} \begin{align*} p(z|y) &= \int p(z, \theta | y) d \theta \\ &= \int p(z |y, \theta) p(\theta |y) d \theta \\ &= \int p(z | \theta) p(\theta |y) d \theta \end{align*} \end{split}$$
We are now weighting the probability distribution function for z with our posterior belief after having observed y.

For our example, we found that the posterior distribution $p(\theta |y) \) is a Beta(\(a+y, b+n-y\)) distribution. Thus our posterior predictive distribution is a Beta-binomial distribution with the number of trials n_p and shape parameters \(a+y, b+n-y\).

Now, given that we use the asymmetrical $Beta(2, 9)$ prior, and then observe that $y=4$ patients out of $n=10$ had a successful result, and we wish to predict how many sucesses z out of $n_p=20$ to expect, our posterior predictive distribution is a Beta-binomial with parameters 20 and shape parameters 6 and 15. The expectation of this distribution is $E(y) = 20 \frac{6}{21} \approx 6$ patients.

9.8 Conjugacy

In the example with the Beta-Binomial model, we found that using the Beta distribution for the prior lead us to a posterior distribution that is also a Beta distribution. This is not a coincidence. Often, a particular distributional family is chosen for the prior, so that the resulting posterior distribution belongs to the same family. This is called a conjugate prior. Below are the conjugate priors for some common likelihood models.

Likelihood	**Conjugate Prior**
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Geometric	Beta
Normal	Normal, Gamma and a few others
Exponential	Gamma
Gamma	Gamma

9.8.1 Exercise

Suppose that there is an experiment where n patients are asked to try different treatments each time they get a headache. We are interested in the number of different treatments a patient takes before they find one that is successful. For patient i, for $1 \leq i \leq n$, we denote by y_i the number of treatments tried before the first success. Note that $\left\{ y_1, y_2, ..., y_n \right\}$ are a sample from a Geometric distribution: $y_i \sim Geom(\theta)$. The probability density function of a geometric distribution is:

\[p(y | \theta) = \theta (\theta -1)^{y-1}\]
Suppose we wish to make inference on \(\theta\). By specifying a Beta prior for \(\theta\): \(\theta \sim Beta(a, b)\), derive the posterior distribution of \(\theta\).

Try the exercise and then click the button to reveal the solution.

10. Bayesian Statistics II: Normal data

In the previous session, we looked at Bayesian inference for proportions. We now consider continuous data and explore Bayesian inference for data when they are assumed to follow a Normal distribution.

Intended learning outcomes
By the end of this session you will be able to:

- Find the posterior for a Normally distributed mean when the variance of the data is known
- Find credible and HPD intervals for a Normally distributed mean
- Find the Bayesian predictive distributions for Normal data and data summaries.

The next sessions calculate the posterior for the mean of a Normal distribution, obtain HPD credible intervals and use the posterior to make predictions.

10.1 Example: CD4 cell counts

In this session, we will use a dataset on CD4 cell counts which is available in R through the *boot* package. CD4 cells are in our blood as part of our immune system. Since these cells die in people who have HIV, CD4 cell counts are used in HIV patients to determine the health of their immune system and susceptibility to opportunistic infections.

In this dataset, there are 20 patients with HIV. Their CD4 cell counts are recorded before and after they were put on treatment. We wish to investigate whether this treatment increased their CD4 cell counts.

We install the *boot* package where the data is stored and we look at the data. Note that the unit of CD4 cell count is 100 \(cells/mm^3\). We are interested in the difference in CD4 cell counts before and after treatment. We look at the summary statistics of the difference.

```
library(boot)
ydata <- cd4$oneyear - cd4$baseline
data <- cbind(cd4, y=ydata)
data
summary(ydata)
```

	baseline	oneyear	y
	<dbl>	<dbl>	<dbl>
1	2.12	2.47	0.35
2	4.35	4.61	0.26
3	3.39	5.26	1.87
4	2.51	3.02	0.51
5	4.04	6.36	2.32
6	5.10	5.93	0.83
7	3.77	3.93	0.16
8	3.35	4.09	0.74
9	4.10	4.88	0.78
10	3.35	3.81	0.46
11	4.15	4.74	0.59
12	3.56	3.29	-0.27
13	3.39	5.55	2.16
14	1.88	2.82	0.94
15	2.56	4.23	1.67
16	2.96	3.23	0.27
17	2.49	2.56	0.07
18	3.03	4.31	1.28
19	2.66	4.37	1.71
20	3.00	2.40	-0.60

A data.frame: 20 × 3

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.6000  0.2675  0.6650  0.8050  1.3775  2.3200 
```

In the classical framework, we could use a paired t-test to see if the mean change in CD4 cell counts is significantly different from the null hypothesis value of zero (\(H_0: \mu = E[Y]=0\)).

For our Bayesian analysis, we will assume these measurements come from a Normal distribution with an unknown mean \(\mu\), which represents the mean change in CD4 counts. We will assume that the variance is known to be \(\sigma^2 = 0.7\). This is slightly artificial as, in a real example, we may not know what the true variance is; however, we might be able to infer the variability of CD4 counts from earlier studies. Having both \(\mu\) and \(\sigma^2\) unknown requires a more complicated analysis which we will not cover in this course.

The Bayesian analysis involves constructing a likelihood for the data, specifying an appropriate prior distribution and combining them to obtain a posterior distribution. We will then describe how credible intervals for \(\mu\), and prior and posterior predictive distributions can be found.

10.2 Calculating the posterior for the mean of a Normal distribution

In this section, we obtain the posterior for the mean of a Normal distribution with known variance, \(\sigma^2\).

Suppose we have \(n\) observed independent data points, each assumed to come from the Normal distribution: \(y_1,\dots,y_n \sim N(\mu,\sigma^2)\). Recall that the Normal distribution has probability density function given by \[p(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}. \] Note that some authors will parameterize the Normal distribution with the *precision* instead of the variance: \(\eta=\frac{1}{\sigma^2}\).

10.2.1 Likelihood

For convenience, we will drop the conditioning on \(\sigma^2\), since we are assuming this is a known number. Since we assume all observations are independent, the likelihood is the product of the \(n\) individual p.d.f.s:

$$\begin{split} \begin{align*} p(y_1,\dots,y_n \mid \mu) &= p(y_1 \mid \mu)\, p(y_2 \mid \mu) \dots p(y_n \mid \mu) \\ &= \prod_{i=1}^n p(y_i \mid \mu) \\ &= \left(\frac{1}{2\pi \sigma^2}\right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2 \right\} \end{align*} \end{split}$$

Notice that

$$\begin{split} \begin{align*} \sum_{i=1}^n (y_i - \mu)^2 &= \sum_{i=1}^n (y_i - \bar y + \bar y - \mu)^2 \\ & = \sum_{i=1}^n (y_i-\bar y)^2 + n(\bar y - \mu)^2, \mbox{ since } \sum_{i=1}^n (y_i-\bar y)=0,\\ &= (n-1)s^2 + n(\bar y - \mu)^2, \end{align*} \end{split}$$

where (as usual) $s^2 = \sum_{i=1}^n (y_i - \bar y)^2 /(n-1).$

Thus the Likelihood can be written:

$$ p(y_1,\dots,y_n \mid \mu) = \left(\frac{1}{2\pi \sigma^2}\right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar y - \mu)^2 \right] \right\}. $$

Since we are interested in the posterior for $\mu,$ we can drop all terms not involving $\mu,$ so the likelihood is proportional to

$$ p(y_1,\dots,y_n \mid \mu) \propto \exp\left\{ -\frac{n}{2\sigma^2} (\bar y - \mu)^2 \right\}. $$

Notice that this also has the same form of a Normal distribution for the mean $\bar{y}$, specifically, $\bar{y} \sim N(\mu, \frac{\sigma^2}{n})$.

10.2.2 Prior

We noted in the previous session that the Normal distribution is a conjugate prior when the likelihood is a Normal distribution. Thus, for convenience, we will use a Normal distribution as a prior for μ:

$$ \mu \sim N(\phi, \tau^{2}), $$

as the posterior distribution will conveniently be a Normal distribution as well. The prior parameters ϕ and τ^2 should be specified based on prior knowledge of μ and the uncertainty around this prior knowledge. It may come from previous research or formally elicited from investigators. If no prior evidence is available, we assign an appriopriately large value to τ.

10.2.3 Posterior

To derive the posterior for the mean μ, we need to find the distribution of that parameter conditional on the data (both the empirical data and prior distribution). In the following calculation, we are only interested in the parts of the p.d.f. that depend on μ. Any terms not involving μ are part of the *normalisation constant*. This is part of the p.d.f., but does not affect the shape of the density.

The posterior is given by

$$\begin{split} \begin{align*} p(\mu \mid y_1,\dots,y_n) &\propto p(y_1,\dots,y_n \mid \mu)\, p(\mu) \\ &\propto \exp\left\{ -\frac{n}{2\sigma^2} (\bar y - \mu)^2 \right\} \exp\left\{-\frac{1}{2\tau^2}(\mu-\phi)^2\right\} \\ & = \exp\left\{ -\frac{n}{2\sigma^2}(\bar{y}-\mu)^2 -\frac{1}{2\tau^2}(\mu-\phi)^2\right\} \\ \end{align*} \end{split}$$

Expanding the brackets and retaining only terms containing μ:

$$ p(\mu \mid y_1,\dots,y_n) \propto \exp\left\{ -\frac{n}{2\sigma^2 \tau^2} (-2n\bar{y}\mu\tau^2 - n\mu^2\tau^2+\mu^2\sigma^2-2\mu\phi\sigma^2) \right\} $$

Completing the squared term for μ:

$$ p(\mu \mid y_1,\dots,y_n) \propto \exp\left\{ -\frac{\tau^2 n + \sigma^2}{2 \sigma^2 \tau^2}\left(\mu - \frac{ \tau^2 n\bar{y} -\sigma^2\phi}{\tau^2n+\sigma^2}\right)^2\right\} $$

We can recognise this has the form the p.d.f. of the Normal distribution, therefore we see that

$$ \mu \vert y_1,\dots,y_n \sim N\left\{ \frac{ \tau^2 n\bar{y} + \sigma^2\phi }{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2n+\sigma^2} \right\}. $$

We see that:

1. the Normal prior is *conjugate* for a Normal Likelihood, as the posterior is also Normal.
2. The posterior mean, $\frac{ \tau^2 n\bar{y} + \sigma^2\phi }{\tau^2 n + \sigma^2}$ is a weighted average of the data $\bar y$ and the prior mean ϕ: we can write it as $w \bar{y} + (1-w) \phi$, where $w= \frac{\tau^2 n}{\tau^2 n + \sigma^2}$. Hence the posterior combines the information from the likelihood (data) and prior (a priori belief).
3. The variance of the posterior is $\frac{\sigma^2\tau^2}{\tau^2n+\sigma^2}$. In a larger study, since n becomes very large, we have $\tau^2 >> \frac{\sigma^2}{n}$, so the posterior variance tends to zero.
4. In smaller studies, $\tau^2 << \frac{\sigma^2}{n}$, the posterior mean is closer to ϕ and the posterior variance depends both on the prior and sampling variance $\frac{\sigma^2\tau^2}{\tau^2n+\sigma^2}$.

10.3 Credible Intervals

We saw in the previous session that a Bayesian \(95\%\) credible interval is an interval which contains \(95\% \) of the posterior distribution of the parameter, and the \(95 \%\) Highest Posterior Density (HPD) interval is the credible interval with the smallest range of values for \(\theta\).

Given that the posterior distribution has mean \(\psi\) and variance \(\gamma^{2}\), the \(95\%\) HPD interval is given by \(\psi \pm 1.96 \times \gamma\). Thus, for a standard Normal posterior, the 95% HPD interval is \((-1.96,1.96).\)

10.3.1 CD4 cell counts example:

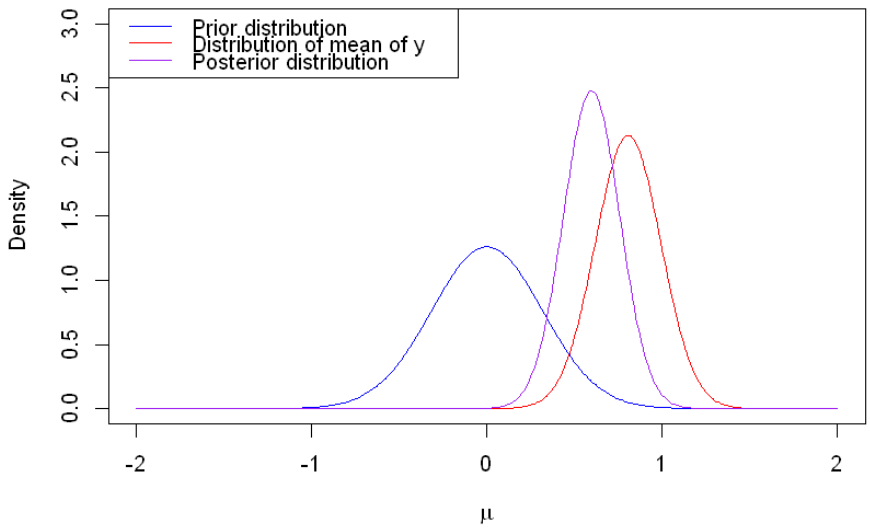
In the CD4 cell count example, suppose that we have very strong prior information that suggests the treatment is not effective, and we expect that the difference in cell counts is approximately zero. Let us denote by \(y\) the difference in CD4 cell counts. We set \(\mu \sim N(0, 0.1)\) to reflect that there is only about \(2.5\%\) chance that the treatment increases mean CD4 counts by more than 0.62 (1.96 \(\times \sqrt{0.1}\)) and a \(50\%\) chance that it will actually decrease the mean CD4 count).

Summarizing the information we have:

> sample size \(n = 20\)
> mean of data \(\bar{y} = 0.805\)
> variance of data (assumed known) \(\sigma^2 = 0.7\)
> prior mean \(\phi = 0\)
> prior variance \(\tau^2= 0.1\)

We find the posterior distribution:
\[\begin{split} \begin{align*} \mu \vert y_1,\dots,y_n &\sim N\left\{ \frac{ \tau^2 n\bar{y} + \sigma^2\phi }{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2n+\sigma^2} \right\} \\ &\sim N\left\{ \frac{ 0.1 \times 20 \times 0.805 + 0 }{0.1 \times 20 + 0.7}, \frac{0.7 \times 0.1}{0.1 \times 20 +0.7 } \right\} \\ &\sim N\left\{ 0.596, 0.0259 \right\} \end{align*} \end{split}\]
We plot below the prior distribution (in blue), the distribution of \(\bar{y}\) (red) and the posterior distribution (purple). We observe that the mean of the posterior distribution is in between the mean of the prior and that of the likelihood. Note that in R, the Normal distribution is parameterized by the standard deviation rather than the variance.

```r
options(repr.plot.width=7, repr.plot.height=5)
x <- seq(-2, 2, 0.01)
#plot the prior
y1 <- dnorm(x, mean=0, sd=sqrt(0.1))
plot(x, y1, type="l", lwd=1, col="blue", ylim=c(0,3), ylab="Density",
xlab=expression(mu))
legend("topleft", legend=c("Prior distribution", "Distribution of mean of y",
"Posterior distribution"),
       col=c("blue", "red", "purple"), lty=1)
#plot the observed distribution
y2 <- dnorm(x, mean=0.805, sd=sqrt(0.7/20))
lines(x, y2, type="l", lwd=1, col="red")
y3 <- dnorm(x, mean=0.596, sd=sqrt(0.0259))
lines(x, y3, type="l", lwd=1, col="purple")
```



The \(95\%\) HPD interval can be calculated as \(0.596 \pm 1.96 \times \sqrt{0.0259} = (0.281, 0.911)\). This interval lies wholly above zero, so we can state that we have a strong posterior belief that there is an increase in CD4 cell counts.

10.4 Predictions

10.4.1 Prior predictive distributions

Finding the predictive distribution for a new patient \(y\) before making any observations involves finding the following distribution:
\[\begin{split} p(y | \sigma^2, \phi, \tau^2) = \int p(y, \mu | \sigma^2, \phi, \tau^2) d \mu\\ = \int p(y | \mu, \sigma^2, \phi, \tau^2) p(\mu | \phi, \tau^2) d \mu \end{split}\]
This calculation involves a lot of algebra. We instead use a different approach: note that we can write the observation as \(y = \mu + \epsilon\), where \(\mu \sim N(\phi, \tau^2)\) and \(\epsilon \sim N(0, \sigma^2)\). Then, since \(\mu\) and \(\epsilon\) are independent, we can use this result:

> If X and Y be independent random variables that are Normally distributed, \(X\sim N(\mu _{X},\sigma _{X}^{2})\) and \(Y\sim N(\mu _{Y},\sigma _{Y}^{2})\), then their sum is also Normally distributed: \(X + Y \sim N(\mu _{X}+\mu _{Y},\sigma _{X}^{2}+\sigma _{Y}^{2})\).

Thus we have that \(y \sim N(\phi, \tau^2 + \sigma^2)\).

In our example, before collecting any data, suppose we wish to predict the probability that the difference in cell counts is greater than 0.3 (30 \(cells/mm^3\)). We have that \(y \sim N(0, 0.1 + 0.7)\). We compute \(p(y > 0.3)\):

```
1-pnorm(0.3, 0, sqrt(0.8))
```

0.368657838608209

Given our prior distribution alone, the probability that the change in CD4 count for a new patient will exceed 0.3 (30 \(cells/mm^3\)) is approximately 0.369.

10.4.2 Posterior predictive distributions

Suppose that have observed \(y_1, ..., y_n \), and we want to predict future observations \(z\), assuming that \(z\) and \(y_i\) are independent for all \(1 \leq i \leq n\), conditional on \(\mu\). The posterior predictive distribution for \(z\) is given by,
\[\begin{split} \begin{align*} p(z| y_1, ..., y_n, \sigma^2, \phi, \tau^2) &= \int p(z, \mu | y_1, ..., y_n, \sigma^2, \phi, \tau^2) d \mu \\ &= \int p(z | y_1, ..., y_n,\mu, \sigma^2) p(\mu |y_1, ..., y_n,\sigma^2, \phi, \tau^2) d \mu. \\ \end{align*} \end{split}\]

Again, this involves some fiddly algebra but we can use a similar method to that we used for the prior predictive distribution. We wish to know what the predictive distribution of a new patient z is, given the previous observations $y_1, ..., y_n$. We can write $z = \mu + \epsilon$. We have that $\mu \vert y_1,\dots,y_n \sim N\left\{ \frac{ \tau^2 n\bar{y} + \sigma^2\phi }{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2n+\sigma^2} \right\}$, and $\epsilon \sim N(0, \sigma^2)$.

Using the result for the sum of two independent Normal distributions, the posterior predictive distribution has the form $N\left\{ \frac{ \tau^2 n\bar{y} + \sigma^2\phi }{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2n+\sigma^2} + \sigma^2\right\}$

In our example, based on both prior and observed data, the predictive distribution for cell counts in a new patient being greater than 0.3 (30 $cells/mm^3$) is $N(0.596, 0.0259 + 0.7)$. We can compute $f(z \mid y_1, ..., y_n > 0.3)$:

```
1- pnorm(0.3, 0.596, sqrt(0.7259))
```

0.635861643314828

After having observed the data, the predictive probability that the next patient will have a difference in CD4 cell counts of greater than 0.3 (30 $cells/mm^3$) has increased substantially to 0.636.

10.5 Multiparameter models

Suppose now that our likelihood has two unknown parameters, $(\mu,\sigma^2).$ In this case, we would need a prior distribution for both parameters, and our posterior distribution will now be bivariate. If desired we can summarise this by the mean and covariance matrix or by HPD contour maps. However, often in applications, interest focusses only on one parameter, say $\mu;$ the other parameter is usually referred to as a nuisance parameter. In Bayesian inference, we typically use simulation to draw from the posterior distribution of $(\mu,\sigma^2).$ For marginal inference for $\mu,$ we summarise the draws from μ in the usual way, across all simulated values of σ^2. Analytically, this is equivalent to integrating the posterior over σ^2:
$$ p(\mu | y) = \int p(\mu,\sigma^2 |y) \, d\sigma^2, $$
where we have used Bayes' theorem to obtain the posterior, i.e.$p(\mu,\sigma^2 |y).$ This integral may be intractable (hence the preference for simulation approaches).

Further Resources

Note: further resources are for you to deepen your understanding of the subject if you wish to do so. This is entirely optional. All examinable material is contained within the notes.

Resources for learning

These textbooks are recommended for further learning and examples:

- Bayesian data analysis by Gelman et. al can be downloaded in PDF format.
- The Bugs Book by Lunn et. al. is available at the LSHTM library.
- an introductory book by Jim Stone with nice examples The first chapter is freely available online.

Examples of applications

- Article on Nature providing guidelines Bayesian analyses for genetic association studies.
- The potential benefits of incorporating prior information in the context of health care evaluation is discussed by David Spiegelhalter in this article.
- We mentioned earlier that Bayesian approaches can be helpful for overcoming challenges with small sample sizes in clinical trials for rare diseases; you can read more about this in an article by Lilford et. al.

Investigations and the role of regression modelling

Data scientists don't do statistics just for fun (although, clearly statistics is indeed fun!). At the heart of each data science project is a question to be answered.

This section of the notes begins by thinking about the different types of investigation that might be carried out within a data science project. We consider three important classes of investigation type: **description**, **prediction** and **causal**.

We then move on to consider a commonly used family of statistical analysis: **regression modelling**. Three sessions introduce linear regression, beginning with the simplest type which we call **simple linear regression** involving a single explanatory variable. We then extend this to incorporate multiple explanatory variables, through **multivariable linear regression** modelling. We explore how to model various types of explanatory variables, including continuous, binary and categorical covariates and discover how to include interactions and higher-order terms (which are need to model non-linear relationships) in the regression model. The last of the linear regression sessions explores diagnostics to assess whether the underlying assumptions of the linear model hold in a particular dataset.

These ideas are then extended to other settings in the remaining two sessions. First, we meet **logistic regression**, an extension of linear regression modelling to settings where the outcome variable is binary. Finally, we define the **Generalised Linear Model (GLM)**, which is a generalisation of linear regression to a wide range of settings and can be seen as a way of unifying linear, logistic and Poisson regression models, as well as many other types of regression model. We explore **Poisson regression** as an important example of a GLM.

We conclude this section of the notes by returning to the idea of investigations and – armed with our new knowledge about regression modelling – consider the role of regression modelling in different types of investigations.

11. Types of Investigation

This session introduces how to set up a research question, explores the different types of investigation and brings in key concepts like prediction, causality and confounding. This session demonstrates how methods learnt in previous sessions are applied in research.

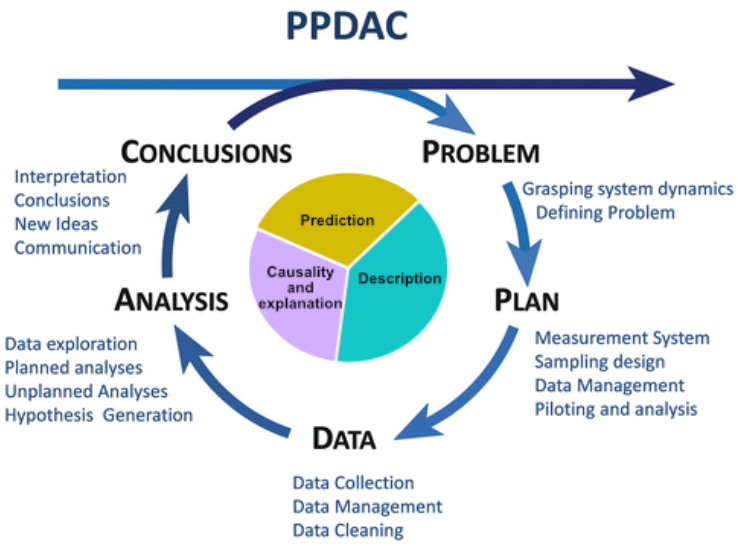
Intended learning outcomes

By the end of this session you will be able to:

- Describe three different types of investigations that arise in medical statistics and health data science.
- Link a research question to an investigation type and compare the properties of different investigation types.
- Explain how and why explanatory variables are used differently in prediction studies and in causal investigations.

11.1 Specifying research questions

Specifying the research question or questions is a crucial starting point for an investigation. In some cases the research question will be highly specific, and in others could be more wide ranging with several components. The research question then informs the subsequent stages of the investigation, ranging from choice of study population; study design; data collection; monitoring and quality control; data analysis; presentation of conclusions; interpretation. Figure 11.1 illustrates one way of representing the whole process of an investigation (which we saw earlier in the Introduction).



The statistician/data scientist plays an important role at all stages of an investigation, not just at the data analysis phase. It is perhaps most usual for collaborators who are subject-matter experts (e.g. clinicians) to pose the initial research question. However, the statistician very often plays a key part in refining these initial ideas in order to translate them into something formal and clearly specified.

11.2 Different types of investigation

11.2.1 Classification

The research question informs what type of investigation is required. Investigations can be divided broadly into the following types:

1. Description
2. Prediction
3. Causality and explanation

Hernan, Hsu & Healy (Chance, 2019) set out to classify data science tasks and used three classifications: *Description*, *Prediction*, and *Counterfactual prediction* (meaning causality). Schmueli (Statistical Science, 2010) also described similar classifications: *Descriptive modelling*, *Predictive modelling*, and *Explanatory modelling*. See also Hand (Harvard Data Science Review, 2019) for a nice discussion on this topic.

11.2.2 Implications of investigation type

The distinction between the different types of investigation is crucial because it has a fundamental impact on the steps of the analysis and beyond. For example, the investigation type influences:

- How we decide what variables are to be included in the analysis
- What analysis methods to use
- How we assess the fit/performance of the model or oter analysis approach used
- How we present the results from the analysis
- How the findings might be used in practice
- How we need to work with other experts at different stages

11.2.3 The role of study design

The different types of investigation may be performed using data from studies of different design. Having posed a research question, we can consider (with input from collaborators) what data are required to answer it robustly, including whether new data collection is needed, or whether there are existing data that could be used to address the question. This process needs to take into account considerations of cost, timeliness, feasibility and ethics. For example, for some questions our ideal study could be a randomized controlled trial, but to perform one would require such long followup that it would be infeasible and unethical, and so we would turn to observational data to address the research question. There is a major emphasis in the recent biostatistical and epidemiological literature on the use of 'found' data from sources such as electronic health records, which present great opportunities to answer research questions using data

on a large number of individuals, but also present challenges for analysis and interpretation. All three types of investigation may make use of observational data. Randomized controlled trials are designed to estimate treatment effects (i.e. for causal investigations), but secondary analyses of trial data can be used in other types of investigation, such as to develop a prediction model.

11.3 Properties of different types of investigation

11.3.1 Description

In a descriptive investigation the data are used to provide a quantitative summary of features of the population of interest, or in other words the data are summarised in a compact way.

Simple descriptive analyses involve calculating proportions of individuals with a particular characteristic (e.g. males and females; smokers and non-smokers), or estimating features of the distribution of continuous variables (e.g. mean and variance of weight or blood pressure). The resulting information is then presented using tables and data visualisation.

Some descriptive analyses may extend to use of more complex methods of analysis. For example, the research question may concern how individuals within a population cluster together interms of their dietary habits, requiring clustering methods. It may be of interest to estimate theexpected survival time post-disease diagnosis in the presence of censored survival times, which would require survival analysis techniques.

All investigations should start with some basic descriptive analysis to gain understanding of the features of the data at hand. It is at this stage that we can uncover challenges such as missing data, gain insights into how certain variables are distributed, and, where relevant, gain understanding of correlations between key variables, including to identify collinearities. Some investigations then go on to the main research question, which goes beyond description, and others may be entirely descriptive and not proceed onto other questions.

Huebner et al. (2019) provide useful guidance on 'initial data analysis'. See also Spiegelhalter (2019) for an accessible discussion of summarising and communicating descriptions of data.

11.3.2 Prediction

Prediction is about using data on some features of individuals to predict other features with the aim of predicting the outcome for new or future observations. More formally, prediction is concerned with mapping data on variables X_{1}, X_{2}, ... , X_{p} to an outcome Y . The prediction model could be developed using statistical models such as regression, or approaches that would be described as machine learning algorithms.

Results from prediction investigations are used for a range of purposes: to inform people of their risk or prognosis; to identify people at high risk of an adverse event and hence take action such as more frequent screening (though the investigation will not tell us whether such screening would be effective).

Prediction models are typically developed using observational data. A well known example is the Framingham Risk Score, which provides predictions of a person's 10-year of developing coronary heart disease (D'Agostino et al 2008).

There is a huge literature on prediction in the medical setting. See for example the books by Riley et al. (2019) and Steyerberg (2019).

11.3.3 Causality and explanation

In causal investigations we seek to understand the causal effect of one or more variables on an outcome. Hernan et al. (2019) describe this as "Using data to predict certain features of the world as if the world had been different". For a simple example of a causal investigation, consider a continuous outcome Y (e.g. blood pressure) and a binary treatment variable X, where $X = 1$ denotes treated and $X = 0$ denotes untreated. A causal investigation asks how the mean of Y would be different if all individuals had $X = 1$ compared with if all individuals had $X = 0$. In other words, if we could change X what would be the expected change in Y ?

Questions such as this can be arguably simple to answer using a randomized controlled trial, where there is no confounding of the treatment-outcome association. However, issues of drop-out and non-compliance are important to consider. Historically, some have considered answering causal questions to lie only in the domain of randomized experiments. However, randomized experiments are not feasible or ethical to address many important questions. It is now recognised that causality is often the goal of investigations using observational data. See for example the paper of

Hernan (2018), who wrote "being explicit about the causal objective of a study reduces ambiguity in the scientific question, errors in the data analysis, and excesses in the interpretation of the results". The field of 'causal inference' has developed in recent decades, with particular advances in recent years, to enable this.

Schmeuli (2010) equates causality with 'explanation', meaning explanation of mechanisms of how one (or more) variable affects another. However, Hernan et al. (2019) make the point that we may be able to say that X causes Y without understanding the underlying mechanism. For example we may find strong evidence from a trial that a drug is effective for a given outcome, but the precise biological mechanisms through which the effect is transmitted are not well understood.

The variable of interest in a causal investigation could be use of a medical treatment (a drug) or application of a procedure. More generally it could be an 'exposure' such as 'smoking' or 'exercising for at least 30 minutes per day'. The 'hypothetical intervention' of interest should be (reasonably) well defined, even if we could never in reality intervene on it in the real world (e.g. it would be impractical, not to say unethical, to intervene on smoking status). See Hernan (2016) for a discussion of related issues.

11.3.4 Is there a fourth investigation type?

There is arguably a fourth investigation type which is concerned with exploring how several explanatory variables X_{1}, ... , X_{p} are associated with an outcome Y. This might be described as an "exploration of risk factors" investigation. It may involve univariable analyses, looking at the association of each explanatory variable ("risk factor") individually with the outcome, and multivariable analyses which look at association of several variables with the outcome in a single model. These types of analysis are typically carried out using observational data, and many (or perhaps most) epidemiological studies are investigations of this type, at least historically.

These types of investigation can be useful for understanding associations between variables in the population of interest and, as such, some may consider these analyses to be descriptive. However, as we all know, association is not causation! These types of investigation often do not consider the relative temporal ordering of explanatory variables, which means that interpretation of estimated associations as causal effects can be misleading. There is recent emphasis in the epidemiological literature on more principled investigations which are more explicit about the aim of the investigation.

Like in a prediction investigation, the interest is in several explanatory variables. However, unlike in a prediction investigation, the aim is to actually explore quantitatively the unconditional and conditional associations of the explanatory variables with Y, rather than being purely on predicting Y. Unlike in a causal investigation, there is not a particular focus on a single variable. However, there is often an attempt to discuss the associations as though they may be causal even though an explicit causal question has not been posed.

Investigators should be wary of over-interpreting findings from "exploration of risk factors" investigations. And if we are really interested in addressing a causal question we should be explicit about that and carry out our analysis and interpretations accordingly.

11.4 An example: stroke in women

Table 1 provides an example of the features of different investigation types. The overall topic is stroke in women. The table (taken from Hernan et al. 2019) provides an example research question, the features of data that would be required to answer it, and the types of analysis that could be used for investigations of three types: Description, Prediction and Causal inference.

Table 1: From Hernan, Hsu & Healy 2019. Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Description	Prediction	Causal inference
Example of scientific question	How can women aged 60-80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	- Eligibility criteria - Features (symptoms, clinical parameters ...)	- Eligibility criteria - Output (diagnosis of stroke over the next year) - Inputs (age, blood pressure, history of stroke, diabetes at baseline)	- Eligibility criteria - Outcome (diagnosis of stroke over the next year) -Treatment (initiation of statins at baseline) - Confounders - Effect modifiers (optional)
Example of analytics	Cluster analysis	Regression Decision trees Random forests Support vector machines Neural networks	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation

11.5 Role of explanatory variables in different types of investigation

The role of explanatory variables in different types of investigation differs. We focus here on prediction investigations and causal investigations.

11.5.1 Prediction

In prediction investigations the aim is to use X_{1}, ... , X_{p} to predict Y . In this setting the X_{1}, ... , X_{p} are often referred to as the 'predictors' for obvious reasons. For a prediction problem we may well use all of the explanatory variables X_{1}, ... , X_{p} in the prediction model or algorithm. Crucially, in prediction we are not interested in the inter-relationships between the explanatory variables X_{1}, ... , X_{p} and their temporal ordering. The only aim is to achieve a good prediction of the outcome Y. It may be desirable to reduce the number of explanatory variables, particularly in settings where the number of potential predictors p is very large. Various principled procedures are available for reducing the number of predictor variables.

11.5.2 Causality and explanation

In investigations of causality, one of the explanatory variables is designated as the treatment or exposure of interest. Let's suppose this is variable X_{1} and the research question is about how X_{1} affects Y. Or, in other words, if X_{1} had been different, how would Y have been different? Let's consider the setting of an Randomized Controlled Trials and an observational study separately and think of the situation where X_{1} is a binary treatment variable

Randomized controlled trials (RCT)

Suppose individuals are randomized to receive treatment $(X_{1} = 1)$ or not $(X_{1} = 0)$, and the outcome Y is observed after some period of follow-up. It is straightforward to estimate the treatment effect in this setting because of the randomization. For a continuous outcome, we would quantify the treatment effect using a difference in the mean outcome in the two treatment groups $(E(Y |X_{1} = 1) − E(Y |X_{1} = 0))$. For a binary outcome we could quantify the treatment effect in terms of a risk difference $((Pr(Y = 1|X_{1} = 1) − Pr(Y = 1|X_{1} = 0))$, risk ratio $((Pr(Y = 1|X_{1} = 1)/Pr(Y = 1|X_{1} = 0))$ or odds ratio $(((Pr(Y = 1|X_{1} = 1)/Pr(Y = 0|X_{1} = 1))/(Pr(Y = 1|X_{1} = 0)/Pr(Y = 0|X_{1} = 0)))$, for example.

Some of the other explanatory variables X_{2}, ... , X_{p} are likely to be associated with Y, but we do not need to use them to estimate the treatment effect due to the study design. Sometimes investigators will adjust for baseline variables, measured at the start of the trial prior to treatment. By the study design, baseline variables are not associated

with the treatment. There can be advantages of adjusting for baseline variables that are predictors of the outcome. Though there are particular nuances to the interpretation of the resulting estimates depending on the types of outcome (continuous, binary, etc) and on how the treatment effect is quantified.

Of course, there are many important considerations surrounding the validity and interpretation of treatment effects estimated using RCTs, such as whether the effect is a 'per-protocol' or 'intentionto-treat' effect, whether there is drop-out, non-adherence or treatment switching.

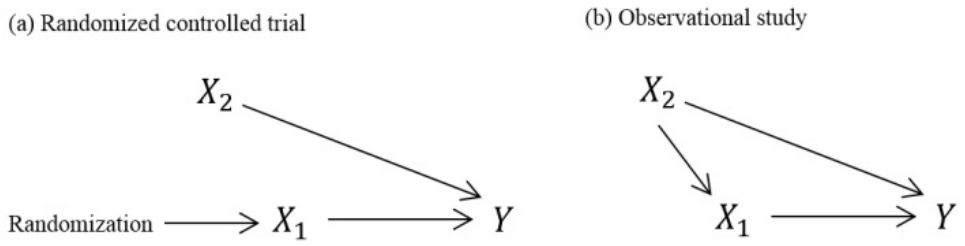
Observational studies

Suppose we have available observational data on the treatment variable X_{1} and the outcome Y, for example from electronic health records. In this setting the treatment in non-randomized, and there are very likely to be confounders of the association between the treatment and the outcome.

A confounder is a variable that affects both the treatment and the outcome. Confounding variables occur prior in time to both the treatment/exposure and the outcome. See VanderWeele and Schpitser (2013) for a formal statistical discussion of confounding.

To estimate the causal effect of X_{1} on Y requires us to control for confounding. Consider a simple setting in which there is only one other variable at play, X_{2}, which in the observational setting affects whether a person gets the treatment X_{1} and also affects their outcome Y. For example, if X_{1} is a blood pressure-lowering medication and Y is blood pressure 1 year later, then X_{2} could be the person's blood pressure at the time origin. The assumed relationships between the three variables X_{1}, X_{2} and Y are illustrated in Figure 2 using directed acyclic graphs (DAGs), contrasting the relationships in an RCT and in an observational study.

Figure 2: Directed acyclic graphs (DAGs) illustrating relationships between a treatment X_1, outcome Y and third variable X_2 in a randomized controlled trial and in an observational study.



DAGs, also called 'causal diagrams', are used to graphically describe mechanistic relationships between variable using uni-directional arrows. An arrow connecting two variables indicates (potential) causation in the direction of the arrow and the absence of an arrow indicates an assumption that there is no direct causal effect of the first variable on the second. See Greenland et al. (Epidemiology, 1999) and Shrier and Platt (2008) for introductions to causal diagrams. Some other useful more recent articles on this are from Etminan et al. (2020) and Tenant et al. (2019). In simple situation such as this example, we don't need a DAG to tell us that we need to account for the confounding by X_{2} in our analysis in order to estimate the effect of X_{1} on Y . However, when there are lots of variables at play DAGs become very useful, and have formal theory attached.

In summary, in a causal investigation the variables on which the research question focuses are X_{1} and Y . However, depending on the study design, we may need to account for other variables in the analysis, though those other variables are not our main focus. The concept of confounding is not relevant in prediction investigations.

11.6 Summary

We have placed some emphasis on how the investigation type affects what variables should be included in the analysis and on how the results might be interpreted. There are naturally many other things to consider which are beyond the scope of this session. The above example focused on regression. The next few sessions in this module will focus on regression models of different types. They are a fundamental part of the statistician's toolbox and are used in investigations of different types. However, there are many other specialised methods available for specific tasks. For example, in descriptive analyses we may use clustering methods and principal components analysis. In prediction tasks, machine learning methods not based on regression are increasingly used. In studies of causal effects many specialised methods have been developed over recent years. Some of these involve regression and others not.

The type of investigation affects how we should assess the performance and assumptions of a model/analysis. For example, in prediction tasks we should assess how well the prediction model performs in terms of predicting the outcome for a new individual. This requires tools such as cross validation, and measures of predictive performance such

as \(R^2\) , area under the curve, sensitivity and specificity. In causal analyses we are concerned with whether the assumptions of the models used are valid and whether the model is correctly specified, alongside the validity of untestable assumptions such as whether there are any important confounders that have not been accounted for in the analysis.

This session aimed to provide a broad overview of different types of investigation used in medical statistics/health data science, and which you are likely to encounter in your future careers. This topic has seen some recent emphasis in the literature. The statistical and epidemiological community is increasingly emphasising the need for researchers to ensure they conduct meaningful studies and interpret findings appropriately, particularly relating to the use of observational data. It is a wide topic, and we have only touched on some aspects here.

References

NOTE: You are not expected to read all of these references! It is intended as a list of resources that you may find useful in the future or if you wish to follow-up on some of the topics discussed in more detail.

Bandoli G., Palmsten K., Chambers C.D., et al. Revisiting the Table 2 fallacy: A motivating example examining preeclampsia and preterm birth. Pediatric and Perinatal Epidemiology 2018; 32: 390-397.

D'Agostino R.B., Vasan R.S., Pencina M.J., et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation 2008; 117: 743–753.

Etminan M, Collins GS, Mansournia MA. Using Causal Diagrams to Improve the Design and Interpretation of Medical Research. CHEST 2020; 158: Supplement S21-S28.

Greenland S., Pearl J., Robins J.M. Causal diagrams for epidemiological research. Epidemiology 1999; 10:37–48.

Hand D. What is the Purpose of Statistical Modelling? Harvard Data Science Review 2019 https://doi.org/10.1162/99608f92.4a85af74

Hernan M.A. Does water kill? A call for less casual causal inferences. Annals of Epidemiology 2016; 26: 674-680.

Hernan M.A. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health. 2018;108: 616–619.

Hernan M.A., Hsu J., Healy B.. A second chance to get causal inference right: a classification of data science tasks. Chance 2019; 32: 42-49.

Huebner M., le Cessie S., Schmidt C., Wach W. A Contemporary Conceptual Framework for Initial Data Analysis. Observational Studies 2019; 4: 171-192.

Riley R.D. et al. Prognosis Research in Healthcare: Concepts, Methods, and Impact. 2019. Oxford University Press.

Schmueli. To explain or to predict? Statistical Science 2010; 25: 289-310.

Schooling CM, Jones H. Clarifying questions about "risk factors": predictors versus explanation. Emerging Themes in Epidemiology 2018; 15: 10.

Schrier I., Platt R.W. Reducing bias through directed acyclic graphs. BMC Medical Research Methodology 2008; 8: 70.

Spiegelhalter D. The Art of Statistics: Learning from Data. 2019. Penguin.

Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd Edition. 2019. Springer.

Tennant PWG, Harrison WJ, Murray EJ, et al. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. MedRxiv 2019. https://www.medrxiv.org/content/10.1101/2019.12.20.19015511v1

VanderWeele T.J., Shpitser I. On the definition of a confounder. Annals of Statistics 2013; 41: 196-220.

Westreich D., Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of Epidemiology 2013; 177: 292-298.

12. Linear Regression I

This is the first of three sessions that explore linear regression modelling. These are models where the outcome of interest is a continuous variable.

Intended learning outcomes

By the end of this session, you will be able to:

- explain, in general, the rationale behind parametric statistical models;
- fit and interpret a linear regression model;
- describe the main properties of ordinary least squares estimators;
- explain confidence intervals and hypothesis testing for regression coefficients

Acknowledgements: Thank you to Jennifer Nicholas and Chris Frost whose notes on linear regression were particularly useful in the development of the current lesson.

12.1 Introduction

A parametric statistical model is an algebraic description of how one or more **outcome** variables are influenced by **covariates**. Such models are widely used in medical research. Some examples of questions that we can investigate using statistical models include:

- Does birthweight increase with length of pregnancy?
- Does taking drug A reduce inflammation more than taking drug B in patients with arthritis?
- Can we predict the risk of heart disease for our patients?

In the above examples, the outcome variables are birthweight, inflammation and heart disease. In the first two examples, the length of pregnancy and drug use are covariates. In the third example, no covariates are explicitly mentioned. However, when answering the third question, researchers may want to consider a range of patient characteristics that are associated with the risk of heart disease as covariates in their model, for example: diet, exercise, comorbodities, medications etc.

Recall that statistical models contain **population parameters** and representations of **uncertainty**. The population parameters are unknown quantities that we want to estimate from our sample and the uncertainty is a measure of the variability in the outcome variable that is not explained by the covariates.

This is the first ofthe sessions on linear regression. In this session, we will learn how to define linear regression models, how to estimate their population parameters and how to estimate measures of uncertainty. We begin by introducing the **simple linear regression model** which includes one outcome and one covariate. In the second session, we introduce the **multivariable linear regression model**, which is an extension of the simple linear regression model to situations with multiple covariates. We explore linear regression models with categorical variables, interactions and non-linear terms. In the third session, we discuss the key assumptions underlying linear regression models and important model diagnostics. The optional material to the session explores how to conduct an **analysis of variance** of statistical models.

Before delving in, it is worth making a note of the different terminologies that you may come across in the medical literature. Here, I have already used the terms: outcome and covariates. Table 1 summarises alternatives terms that may be used to describe the same concepts.

Outcome	**Covariates**
Y-variable	x-variables
Dependent variable	Independent variables
Response variable	Regressors
Output variable	Input variables
(no direct analogy)	Explanatory variables
(no direct analogy)	Predictor variables

Table 1: Different terminology used for outcome and covariates

Finally, it is important to understand that statistical models make **assumptions** about the form of relationships between outcomes and covariates. Although we can examine our data to investigate the validity of these assumptions (using methods covered in the next session), we can never be certain that the model is correct.

12.2 Data used in our examples

For our examples we will use data on babies and their mothers. The data contains a random sample of 1,174 mothers and their newborn babies. The column Birth Weight contains the birth weight of the baby, in ounces; Gestational Days is the number of gestational days, that is, the number of days the baby was in the womb. There is also data on maternal age, maternal height, maternal pregnancy weight, and whether or not the mother was a smoker.

The following code can be used to download and look at the data:

```
#Load data
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')

#Look at the first 10 rows of the data
head(data)
```

	Birth.Weight	Gestational.Days	Maternal.Age	Maternal.Height	Maternal.Pregnancy.We
	<int>	<int>	<int>	<int>	<
1	120	284	27	62	
2	113	282	33	64	
3	128	279	28	64	
4	108	282	23	67	
5	136	286	25	62	
6	138	244	33	62	

A data.frame: 6 × 6

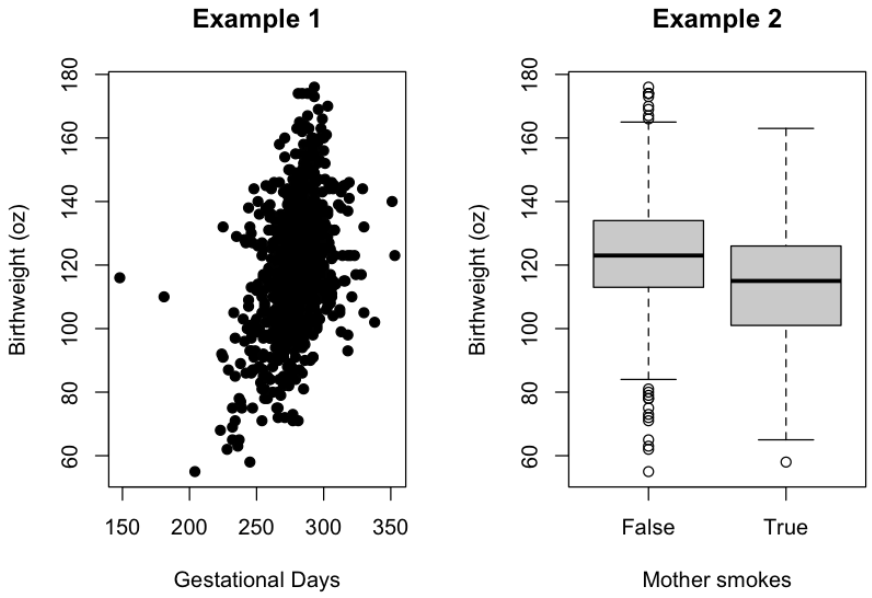
12.2.1 Exploratory analyses

The simple linear regression model is used to model the relationship between one single variable (\(X\)) and a single outcome (\(Y\)). For example, suppose we are interested in investigating the following relationships in our birthweight data:

1. Association between the length of pregnancy (i.e. number of gestational days) and birthweight.
2. Association between mother's smoking status and birthweight.

An important first step in an analysis is to summarise and display the data. Below is a scatterplot and boxplot displaying the relevant data for Examples 1 and 2 respectively.

```
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')

# Set the plot area into a 1x2 array
par(mfrow=c(1,2))
options(repr.plot.height=5)

# Example 1: Scatter Plot
plot(data$Gestational.Days, data$Birth.Weight, main="Example 1",
     xlab="Gestational Days", ylab="Birthweight (oz)", pch=19)

# Example 2: Box plot
boxplot(data$Birth.Weight~data$Maternal.Smoker, main="Example 2", xlab="Mother smokes",
ylab="Birthweight (oz)")
```



Example 1: Birthweight and gestational days appear to be highly correlated, where an increase in gestational days is associated with increased birthweight.

Example 2: It appears that mothers who do not smoke give birth to heavier babies, on average, than mothers who do smoke.

12.2.2 Determining the dependent and independent variables

Before defining a regression model, we have to decide which is the independent variable and which is the outcome (i.e. the dependent variable). In this context, it is natural to consider birthweight as the outcome: conceptually, it makes little sense to investigate how birthweight influences length of pregnancy or the mother's smoking status. However, it is not necessarily always as straightforward. Suppose we were investigating the association between age and weight. It is possible that we might be interested in age as a predictor of weight, or in weight as a predictor of age. The aim of the analysis will guide the choice of outcome.

While the outcome is the same in our two examples, an important difference is the type of independent variable. In Example 1, the independent variable (length of pregnancy) is a continuous variable, whereas in Example 2, the independent variable (mother's smoking status) is binary (yes or no). Using these examples, we will later see how the two different types of variables are modelled differently in linear regression.

12.3 The simple linear regression model

The equation for the simple linear regression model, relating X and Y is:
$$ Y = \beta_0 + \beta_1 X + \epsilon $$
There are two components of this model: the **linear predictor** and the **error term**. The linear predictor represents the variation in Y that can be predicted using the model: $\beta_0 + \beta_1 X$. The error term, denoted by ϵ, represents the variation in Y that cannot be predicted (by a linear relationship with X). This variation is sometimes referred to as the **random error** or **noise**.

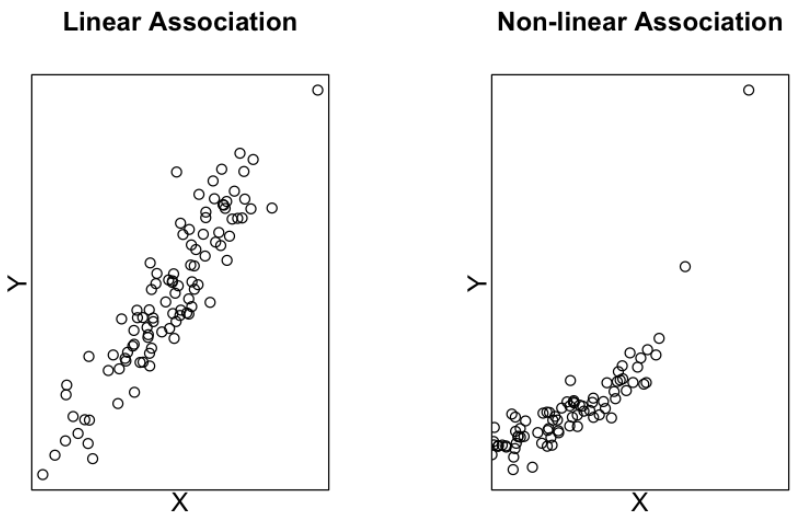
The subsequent two sections take a closer look at the linear predictor and error term, respectively.

12.3.1 The linear predictor

The linear predictor is an additive function of the independent variables. With a single variable, it is simply:
$$ \beta_0 + \beta_1 X $$
In linear regression, the linear predictor represents the algebraic relationship between the mean of the outcome and the independent variable. When X takes a particular value, $X=x$, the value of the linear predictor, $\beta_0+\beta_1x$, is interpreted as the expected value of Y when X takes the value x:
$$ E(Y|X=x) = \beta_0 + \beta_1 x. $$
The specification of the linear predictor has two parameters: β_0 and β_1. These are interpreted as follows:

- β_0 is the intercept. It is the expected value of Y when X takes the value 0.
- β_1 is the slope (or gradient). It is the expected change in Y per one unit increase in X.

It is worth emphasising that this model assumes that **the relationship between X and Y is linear**. It is important to note that it is possible to have more complex relationships between variables that do not meet this assumption (see examples in the plots below). When this is the case, simple linear regression would not be an appropriate method to use but we might be able to model the relationship well by including non-linear terms. We will pursue these ideas further in the next session.



12.3.2 The error term

The error term, ϵ represents the variance in Y that cannot be predicted by the model. Individual values of the errors can be written as (Y - ($\beta_0 + \beta_1X$)). These errors cannot be observed, since they involve the unknown population parameters β_0 and β_1.

We assume that ϵ has a normal distribution with mean 0 and variance σ^2, where σ^2 is termed the **residual variance** (i.e. the variance of the residuals):
$$ \epsilon \sim N(0,\sigma^2) $$
Importantly, note that the errors must be independent of the independent variable X.

12.3.3 Different ways of expressing the simple linear regression model

Suppose we have a sample size of n and we let y_i and x_i $(i=1,...,n)$ denote the observed outcome and value of X for the i^{th} observation, respectively. Then, we can write the simple linear regression model as:
$$ y_i = \beta_0 + \beta_1 x_{i}+ \epsilon_i \text{ where } \epsilon_i \overset{\small{iid}}{\sim} N(0, \sigma^2). $$
Recall that iid means "identically, independently distributed". A key assumption of linear regression model is that all of the observations are independent.

This relationship can equivalently be expressed using matrix algebra:
$$\mathbf{Y}=\mathbf{X}\mathbf{\beta}+\mathbf{\epsilon} \text{ where } \epsilon \sim N(0,\mathbf{I}\sigma^2)$$
$$\begin{split} \begin{vmatrix}y_1\\y_2 \\. \\. \\y_n \end{vmatrix}=\begin{vmatrix}1 & x_1 \\1 & x_2 \\1 & . \\1 & . \\1& . \\1 & x_n \end{vmatrix}\begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix}+\begin{vmatrix}\epsilon_1\\ \epsilon_2 \\ . \\ . \\. \\ \epsilon_n \end{vmatrix} \end{split}$$
In this formulation, $\mathbf{X}$ is an $n \times 2$ matrix, Y and ϵ are vectors of length n whilst β is a vector of length 2.

12.3.4 Assumptions

It is worth emphasising the four key assumptions that we have made in the simple linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Normality**: The errors follow a normal distribution.
3. **Homoscedasticity**: The variance of error terms are constant across all values of X.
4. **Independence**: All observations are independent of each other.

12.4 Estimation of the population parameters

In the specification of the simple linear regression model there are three population parameters (β_0, β_1, and σ). Since we do not know these parameters, we need to estimate them based on a sample from our population. We will use the symbols $\hat{\beta}_0$ $\hat{\beta}_1$, and $\hat{\sigma}$ to represent the sample estimates of the true population parameters.

There are many different methods available for obtaining estimates of the parameters β_0 and β_1. In this section, we focus on an approach that works by minimising the amount of error in the model. These estimates are called the **ordinary least squares estimates** (the reason for this name will become clear in the next section).

12.4.1 Fitted values and residuals

Fitted values: Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the fitted value for the $i{th}$ observation (in other words, the predicted value of the outcome for that individual) is:
$$ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i $$
Residuals: The residual for the $i{th}$ observation is
$$ \hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i $$
Sometimes, the word residual is used to refer to both the residual we have defined here and the error term, in which case it is necessary to distinguish between the true and fitted/estimated residual. Here, we use the term error to refer to the deviation of the observed value from the true mean outcome and we use the term residual to refer to the deviation of the observed value from the fitted value, as defined above.

12.4.2 Ordinary least squares estimates

The ordinary least square (OLS) estimates are those which minimise the sum of squared deviations from the fitted regression line. The residuals, $\hat{\epsilon}$, measure deviations of the observed outcomes from the fitted regression line. Therefore This sum is sometimes called the **residual sum of squares**. It is often denoted by SS_{RES} (where "SS" stands for Sum of Squares and "RES" is shorthand for RESiduals).

Formally, the OLS estimators are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise:
$$ SS_{RES} = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta_0} -\hat{\beta_1}x_i)^2. $$
The ordinary least squares estimates of β_0 and β_1 are given by the following:
$$\begin{split} \begin{align} \hat{\beta_0} &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta_1} &= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} \end{align} \end{split}$$
where $\bar{y}=\frac{\sum_{i=1}^n y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$. A proof of this result is given at the end of this session.

12.4.3 Estimation of the error variance

The residual sum of squares can be thought of as the remaining unexplained variation in the outcome. Therefore, an intuitively appealing estimator of σ^2 is given by dividing the residual sum of squares by the number of observations:
$$ \hat{\sigma}^2 = \sum_{i=1}^n \frac{\hat{\epsilon}_i^2}{n} = \sum_{i=1}^n \frac{(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{n} $$
However, this is a biased estimator. The bias arises because the observed values tend, on average, to lie closer to the fitted line (defined by $\hat{\beta}_0$ and $\hat{\beta}_1$) than they do to the true regression line (defined by β_0 and β_1). This is an exact parallel to the way the variablility of a sample around its mean underestimates the variability around the population mean.

It can be shown that an unbiased estimator of the residual variance in the simple linear regression model is given by:
$$ \hat{\sigma}^2 = \sum_{i=1}^n \frac{\hat{\epsilon_i}^2}{n-2}=\sum_{i=1}^n \frac{(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{n-2} $$
This quantity is referred to as the residual mean square. It is often denoted by MS_{RES}, where "MS" stands for Mean Square and "RES" is shorthand for residual. The denominator is $(n-2)$ because fitting the model first requires the estimation of two parameters (β_0 and β_1) and the estimation of these parameters is said to reduce

the information about the variance by two degrees of freedom.

12.4.4 Maximum likelihood estimation

An alternative approach to estimating the model parameters is maximum likelihood estimation. This approach selects the estimates which maximise the likelihood (or equivalently, the log-likelihood) of the parameter values. It can be shown that the ordinary least square estimates for \(\beta_0\) and \(\beta_1\) are also the maximum likelihood estimates (a proof of this result is at the end of the session).

The maximum likelihood estimate of \(\sigma^2\) is equal to the biased estimate given above, obtained by dividing the residual sum of squares by the number of observations.

12.5 Example: continuous independent variable

We now return to our first example, where we are interested in investigating the association between birthweight and length of pregnancy. We will fit a linear model to explore this association.

12.5.1 The model

The outcome is birthweight, which is measured in ounces (oz). The independent variable is length of pregnancy, \(L\) (i.e. number of gestational days).The following model defines our assumed relationship between the length of pregnancy (\(L\)) and a baby's birthweight (\(Y\)):
\[\text{Model 1: }y_i = \beta_0 + \beta_1 l_i + \epsilon_i \]
We will use the `lm()` to perform simple linear regressions in R. Click here for details of how this command works.

The following code can be used to perform this linear regression in R:

```
# Model 1: Investigating the relationship between birthweight and length of pregancy
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
model1<-lm(Birth.Weight~Gestational.Days, data=data)
summary(model1)
```

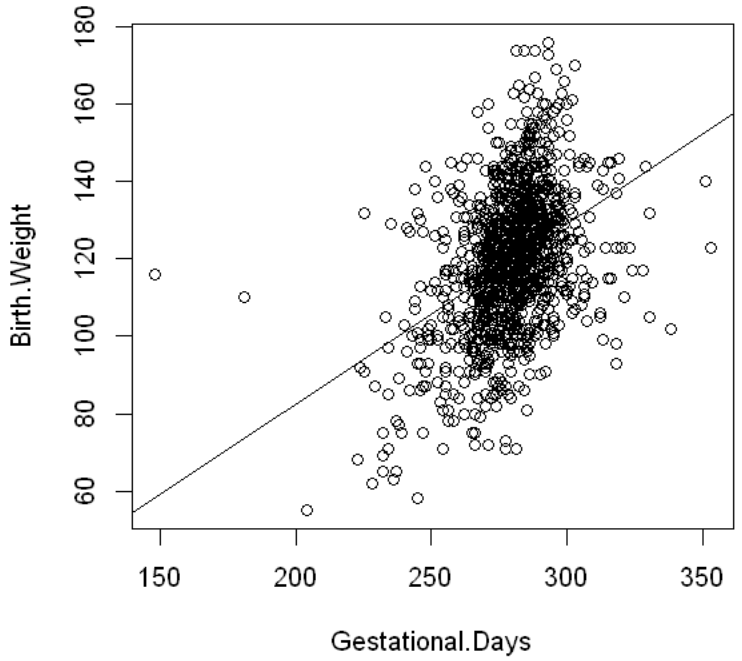
```
Call:
lm(formula = Birth.Weight ~ Gestational.Days, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.348 -11.065   0.218  10.101  57.704

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -10.75414    8.53693   -1.26    0.208
Gestational.Days   0.46656    0.03054   15.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.74 on 1172 degrees of freedom
Multiple R-squared:  0.1661,    Adjusted R-squared:  0.1654
F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```

There is a lot of information contained in this output. For the moment, we will focus on the estimates of the intercept and slope. These can be found under the column heading `Estimate`.

- The estimated intercept, \(\hat{\beta}_0\) is equal to -10.75. This is interpreted as: the estimated mean birthweight of a child born after 0 gestational days is -10.75oz. Since there are no observations with 0 gestational days in the study, this is an extrapolation based on the observed data and an assumption of linearity. Estimates based on extrapolation should be interpreted with caution and in this case, the results make little sense because a negative birthweight is estimated. Moreover, no child is born after 0 gestational days and so this intercept is of little interest. Later on, we will discuss a technique called **centering** which is often used to make the intercept term more interpretable.
- The estimated slope, \(\hat{\beta}_1\) is equal to 0.47. This is interpreted as: the mean birthweight of a baby is estimated to increase by 0.47oz for each daily increase in the gestational period.
- The estimated residual standard error, \(\hat{\sigma}\) is equal to 16.74 (the residual variance is equal to \(16.74^2\)). This means that the observed outcomes are scattered around the fitted regression line with a standard deviation of 16.74oz.

It is always useful to look at the data. The code below graphs the data and superimposes the fitted regression line.

```
options(repr.plot.width=5, repr.plot.height=5)
with(data, plot(Gestational.Days, Birth.Weight))
abline(model1)
```



12.6 Inference for the slope

Most commonly, we wish to conduct statistical inference on the estimated slope. Consequently, we focus our attention here on β_1, but it is possible to apply the same methods to the intercept, β_0.

For our statistical inference, we will view the values of the independent variable ($x_1, x_2, ..., x_n$) as fixed quantities.

12.6.1 Sampling distribution of estimated slope

Knowing the sampling distribution of $\hat{\beta_1}$ allows us to perform hypothesis tests and construct confidence intervals for β_1. Therefore, we now obtain that sampling distribution. The estimated slope, $\hat{\beta}_1$, is a linear combination of the observed outcome values:

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^n \left(\frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}(y_i-\bar{y})\right) \end{aligned}$$

We can simplify this by substituting in $((y_i - \bar{y})=\beta_1(x_i-\bar{x})+(\epsilon_i-\bar{\epsilon}))$, allowing us to write the estimated slope as a function of the random error (ϵ):

$$\begin{aligned} \hat{\beta}_1 &=\beta_1 + \sum_{i=1}^n \left(\frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}(\epsilon_i-\bar{\epsilon})\right) \end{aligned}$$

Since the estimated parameter is a linear combination of the ϵ_i, and $\epsilon_i \overset{\small{iid}}{\sim} N(0, \sigma^2)$, the estimated parameter itself is normally distributed, with a distribution centred around the true value, β_1. More specifically,

$$\begin{aligned} \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{xx}}\right), \qquad \mbox{with} \qquad SS_{xx} = \sum_{i=1}^n(x_i-\bar{x})^2 \end{aligned}$$

So the standard error of the slope is $SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{SS_{xx}}}$. The standard error (and therefore also the variance) of $\hat{\beta}_1$:

- increases with the size of the error variance (as might be expected intuitively),
- decreases with increasing sample size (larger sample sizes give more precise estimates) and
- decreases as ${SS}_{xx}^2$ increases (a wider range of x values leads to more precision in the slope estimate).

We have seen that the sampling distrbution of $\hat{\beta}_1$ follows a normal distribution. We can convert this to a standard normal distribution:

$$ \begin{align} \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/SS_{xx}}} \sim N(0,1) \end{align} $$
Of course, we do not know the true value of the error variance, σ. When we replace this with our sample estimate, this changes the distribution above from a normal to a t-distribution. For large samples, these two distributions are very similar but for smaller samples the t-distribution has larger tails (suggesting that in smaller samples needing to estimate σ leads to more variability in the estimated slope, which makes sense intuitively).

Therefore, replacing σ by the estimate $\hat{\sigma}$, we have:
$$ \begin{align} \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2/SS_{xx}}} \sim t_{n-2} \end{align} $$
This t-distribution is the one we will use to obtain p-values and confidence intervals for the slope parameter.

12.6.2 Hypothesis testing

Typically, we are interested in assessing whether there is a relationship, or association, between the independent variable X and outcome Y. Recall our model: $E[Y | X=x] = \beta_0 + \beta_1 X$. Within the framework of this linear model, no association between X and Y would be reflected by a value of $\beta_1 = 0$.

Therefore, we are typically interested in testing the **null hypothesis** $H_0: \beta_1=0$ against the alternative $H_1: \beta_1 \neq 0$.

Under our null hypothesis, the following test statistic follows a t-distribution, as we saw above:
$$ T = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2/SS_{xx}}} \sim t_{n-2} $$
We now follow the familiar process from the session in hypothesis testing. We evaluate T in our particular sample and then calculate the probability of obtaining that value or one more extreme for a t-distribution with $n-2$ degrees of freedom.

Notes

> T is simply the estimate divided by its standard error. This is a familiar form of test statistic which we saw in the session about hypothesis testing.
>
> Although typically we are interested in the null hypothesis $H_0: \beta_1=0$ we could use the same approach to test other null hypotheses, such as $H_0: \beta_1=5$. However, it is rare that we are interested in values other than 0. Therefore, p-values outputted by statistical software arise from testing the null hypothesis value of 0 by default.

Example

Returning to our linear model relating birthweight to gestational days in the baby dataset, we will now test the null hypothesis that there is no association between gestational days and birthweight, i.e. $H_0: \beta_1 = 0$ in Model 1.

First, we rerun the code to reproduce the linear model output.

```
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
model1<-lm(Birth.Weight~Gestational.Days, data=data)
summary(model1)
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.348 -11.065   0.218  10.101  57.704

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -10.75414    8.53693   -1.26    0.208
Gestational.Days   0.46656    0.03054   15.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.74 on 1172 degrees of freedom
Multiple R-squared:  0.1661,    Adjusted R-squared:  0.1654
F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```

In the above output, the column `Std.Error` gives the standard errors of the estimated intercept and slope.

The columns `t value` and `Pr(>|t|)` give the test statistic and associated p-value for a hypothesis test, testing the null hypothesis that $H_0: \beta_0 = 0$ in the top row and $H_0: \beta_1 = 0$ in the bottom row.

Note that the t-value for the slope (which we call T in the discussion above) is 15.28. You can check that this is equal to the estimate divided by the standard error (0.46656/0.03054 = 15.277014).

To test the null hypothesis that $H_0: \beta_1=0$ against the alternative $H_1:\beta_1 \neq 0$, the test statistic is 15.28 and the associated p-value is $<2\times10^{-16}$. This is a very small p-value and therefore the data provide strong evidence against the null hypothesis. Based on these results, we can conclude that birthweight is associated with length of pregnancy.

12.6.3 Confidence intervals for the regression coefficients

We saw above that:
$$ \begin{align} \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1) } \sim t_{n-2} \end{align} $$
with $\hat{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2/SS_{xx}}$. This leads to the following 95% confidence interval for β_1:
$$ \hat{\beta}_1 \pm t_{0.025, n-2} \hat{SE}(\hat{\beta}_1) $$
where $t_{0.025, n-2}$ is the 97.5th percentile of a t-distribution with $n-2$ degrees of freedom. For large samples, this value will be approximately 1.96. For smaller samples it will be a slightly larger number (reflecting additional imprecision in our estimate).

If n is sufficiently large, the t-distribution is well approximated by a normal distribution. In this case, a 95% confidence interval can be found by:
$$ \hat{\beta}_1 \pm 1.96 \times \hat{SE}(\hat{\beta}_1) $$
Example: Calculate a 95% for $\hat{\beta}_1$ (using the values given in the R output above):

Click the button to reveal the solution.

Alternatively, we can obtain confidence intervals using `confint` in R. The option `parm` tells R which regression coefficients to provide confidence intervals for. Try omitting this option or changing it to value 1 to see what happens.

```
# Confidence intervals for the slope, beta_1
confint(model1, parm=2, level=0.95)
```

	2.5 %	97.5 %
Gestational.Days	0.4066435	0.5264702

A matrix: 1 × 2 of type dbl

12.7 Example: binary independent variable

We now return to our second example, where we are interested in the association between birthweight and the mother's smoking status. In exploratory analyses, we saw that mothers who do not smoke give birth to heavier babies, on average, than mothers who do smoke. We will now use a simple linear regression model to further explore this association.

The outcome variable is the same as for our previous example. A key difference, however, is that the independent variable is a binary variable.

12.7.1 The model

We have a continuous outcome variable and a binary independent variable. To include this binary variable in the model, we create a **dummy** variable that takes the value 1 if the mother smokes and 0 if the mother doesn't smoke:
$$\begin{split} s_{i} = \begin{cases} 1 & \text{ if the i^{th} baby's mother smokes} \\ 0 & \text{ if the i^{th} baby's mother does not smoke} \end{cases} \end{split}$$
We then define the following linear regression model:
$$ \text{Model 2: } y_i = \beta_0 + \beta_1 s_i + \epsilon_i $$
When including binary (or categorical) variables in a linear regression in R, we can tell R to treat it as a factor variable using `factor()`:

```r
# Example 2: Investigating the relationship between birthweight and mother's smoking status.
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
model2<-lm(Birth.Weight~factor(Maternal.Smoker), data=data)
summary(model2)
```

```
Call:
lm(formula = Birth.Weight ~ factor(Maternal.Smoker), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-68.085 -11.085   0.915  11.181  52.915

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                123.0853     0.6645 185.221   <2e-16 ***
factor(Maternal.Smoker)True  -9.2661     1.0628  -8.719   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.77 on 1172 degrees of freedom
Multiple R-squared:  0.06091,   Adjusted R-squared:  0.06011
F-statistic: 76.02 on 1 and 1172 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_0 = 123.09$. This is interpreted as the estimated mean birthweight (in oz) of a baby with “dummy” variable equal to 0, i.e. it is the estimated mean birthweight of babies whose mothers do not smoke.
- $\hat{\beta}_1=-9.23$. The mean birthweight is estimated to decrease by 9.23oz per unit increase in the “dummy” variable. A unit increase in the dummy variable equates to moving from the non-smoking group to the smoking group, so we can interpret this as the difference in mean birthweights between the two groups.
- $\hat{\sigma}=17.77$. The observed outcomes are scattered around the fitted regression line with a standard deviation of 17.77oz.

Exercises

1. Perform a hypothesis test of the null hypothesis that there is no association between maternal smoking and birthweight. Write down the null hypothesis, the test statistic and the p-value. interpret your p-value.
2. Calculate (manually or using R) a 95% confidence interval for the difference in mean birthweight between the group whose mothers smoke and those who don’t.

Try the exercise and then click the button to reveal the solution.

12.8 Additional material

This section contains additional material concerning confidence intervals for a fitted value and reference ranges. These are useful topics in regression, but will not be examinable.

12.8.1 Confidence intervals for a fitted value

So far we have only discussed conducting inference on the estimated regression coefficients. However, it may also be of interest to determine **confidence intervals for the fitted outcomes**, or **prediction intervals**. The subsequent two sections describe and illustrate these two concepts, respectively.

Rather than focusing on associations between variables and the outcome, we are sometimes interested in the expected value of the outcome at particular values of X, i.e. $y_x = E[Y \mid X=x] = \beta_0+\beta_1 x$.

The fitted value is our estimate of this quantity,
$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x.$$
The variance of the fitted value is given by:
$$V(\hat{y}_x) = \sigma^2 \left(\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_{xx}}\right)$$
where $SS_{xx}=\sum_{i=1}^n(x_i-\bar{x})^2$, i.e. the sum of squares of X.

The 95% confidence interval for the fitted value is given by:
$$\hat{y_x} \pm t_{n-2, 0.975}\hat{\sigma} \sqrt{\frac{1}{n}+ \frac{(x-\bar{x})^2}{SS_{xx}}}$$

- 95% confidence intervals can be obtained for values of the independent variable that do not arise in the data. However, the width of the confidence interval increases with the distance from the mean (as can be seem from the formula and figure given below).
- Care must be taken when extrapolating outside the range of the observed data as this makes an un-testable assumption that linearity continues outside the observed data range.

Example. The R code below calculates a 95% confidence interval for the fitted value of birthweight of a baby born after 280 gestational days.

Example

We return to our first example, exploring the association between birthweight and length of pregnancy (gestational days).Suppose we are interested in the expected birthweight for a baby who is born at 280 days' gestation.

The code below refits our model and uses it to estimate the expected birthweight for this gestational age. It also provides a 95% confidence interval around that estimate.

```
# Refit model
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
model1<-lm(Birth.Weight~Gestational.Days, data=data)

# Confidence interval for a fitted value
new.data<-data.frame(Gestational.Days=280)
predict(model1, newdata=new.data, interval="confidence", level=0.95)
```

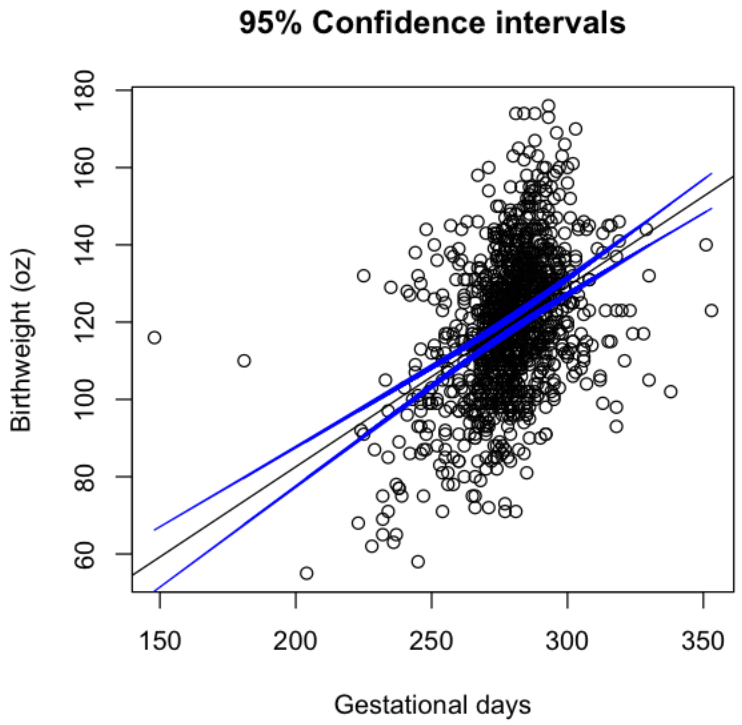
	fit	lwr	upr
1	119.8818	118.9215	120.8421

A matrix: 1 × 3 of type dbl

We estimate that the expected (average) birthweight for babies born at 280 days' gestation is 119.9oz. The 95% confidence interval for this estimate is is (118.9, 120.8). Informally, we can interpret this as: it is plausible that the true value of the expected birthweight, for babies born at 280 days' gestation, lies between 118.9oz and 120.8oz.

We can extend this idea to graph the fitted values - estimated expected birthweight - and their confidence intervals across the range of gestational days. The code below does this.

```
options(repr.plot.height=5, repr.plot.width=5)

# Plot the fitted regression line (the fitted values)
plot(data$Gestational.Days, data$Birth.Weight, xlab="Gestational days",
ylab="Birthweight (oz)", main="95% Confidence intervals")
abline(model1)

# Add the confidence intervals for the fitted regression line
conf_interval<-predict(model1, newdata=data, interval="confidence", level=0.95)
lines(data$Gestational.Days, conf_interval[,2], col="blue")
lines(data$Gestational.Days, conf_interval[,3], col="blue")
```



Notice how the confidence interval around the fitted line is narrowest in the centre of the x-axis, where most of our data are concentrated, and widest at the extremes.

12.8.2 Prediction intervals

The 95% confidence interval around the fitted line describes our certainty about where the fitted line is (i.e. where the expected value is).

Sometimes, we are not interested in the average outcome at a particular point, but the likely spread of values arount the average. In this case, we are interested in obtaining what is called a **prediction interval**, or **reference range**.

A 95% prediction interval, or 95% reference range, is an interval within which 95% of future observations are expected to lie.

The predicted value for an individual with $X=x$ is the fitted value, as above. However, there are now two sourcces of uncertainty to take into account. (1) There is uncertainty about the fitted value (the expected value), as above. (2) There is random error around that point (σ^2). Thus, the variance in the individual prediction is given by:
$$\begin{split} \begin{align*} V(\hat{y_x}) + \sigma^2 &= \sigma^2 \left(\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_{xx}}\right)+ \sigma^2 \\ & = \sigma^2\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right) \end{align*} \end{split}$$
A 95% prediction interval is then given by:
$$ \hat{y_x} \pm t_{n-2, 0.975} \hat{\sigma} \sqrt{1+ \frac{1}{n}+ \frac{(x-\bar{x})^2}{S_{xx}}} $$

Example

The R code below calculates a 95% prediction interval for the birthweight of babies who are born at 280 days' gestation.

```
# Prediction interval
predict(model1, newdata=new.data, interval="prediction", level=0.95)
```

	fit	lwr	upr
1	119.8818	87.01496	152.7486

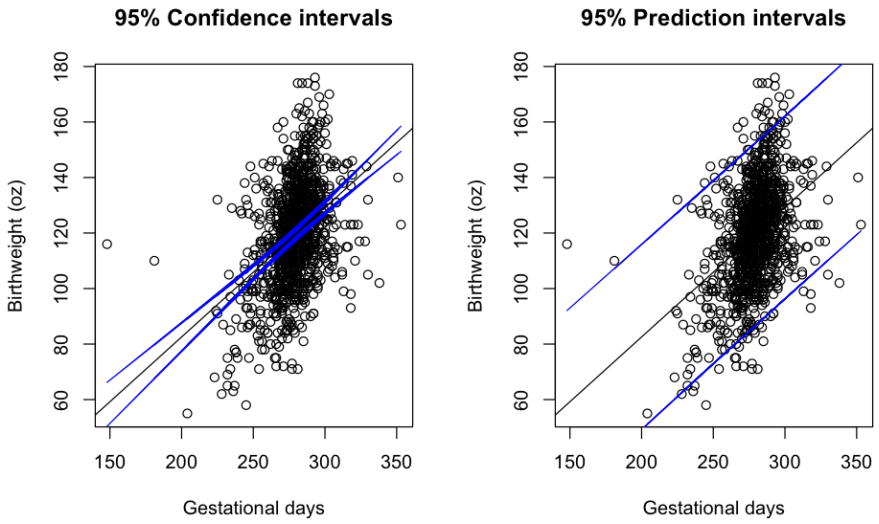
A matrix: 1 × 3 of type dbl

The 95% prediction interval for babies born at 280 days' gestation is (87.0, 152.7). This means that we would expect 95% of babies born after 280 gestational days to weigh between 87 and 152.7 ounces.

12.8.3 Comparing intervals

The code below produces two scatterplots of gestational days against birthweight with the linear regression line of best fit (obtained from Model 1) superimposed. The blue lines on the left-hand side plot represent the 95% confidence intervals for the fitted values across the entire range of gestational days. The blue lines on the right-hand side plot represent the 95% prediction intervals.

```
#Set the graphical space so that two plots are shown side-by-side in one row
par(mfrow=c(1,2))
options(repr.plot.height=5, repr.plot.width=8)

#Confidence intervals for predicted values
plot(data$Gestational.Days, data$Birth.Weight, xlab="Gestational days",
ylab="Birthweight (oz)", main="95% Confidence intervals")
abline(model1)

conf_interval<-predict(model1, newdata=data, interval="confidence", level=0.95)
lines(data$Gestational.Days, conf_interval[,2], col="blue")
lines(data$Gestational.Days, conf_interval[,3], col="blue")

#Reference ranges
plot(data$Gestational.Days, data$Birth.Weight, xlab="Gestational days",
ylab="Birthweight (oz)", main="95% Prediction intervals")
abline(model1)

conf_interval<-predict(model1, newdata=data, interval="prediction", level=0.95)
lines(data$Gestational.Days, conf_interval[,2], col="blue")
lines(data$Gestational.Days, conf_interval[,3], col="blue")
```



As expected, we see that the prediction interval is much wider. Loosely speaking, the plot on the left shows a range of uncertainty about where the *average* line is. The plot on the right shows a range of uncertainty about where individual measurements will lie.

13. Linear Regression II

This is the second of three sessions that explore linear regression modelling. These are models where the outcome of interest is a continuous variable.

Intended learning outcomes

By the end of this session, you will be able to:

- explain the difference between a univariable and multivariable linear regression model
- fit and interpret a multivariable linear regression
- describe the principles of centering
- interpret categorical variables, quadratic terms and interaction terms included in a linear regression model

Acknowledgements: Thank you to Jennifer Nicholas and Chris Frost whose notes on linear regression were particularly useful in the development of the current lesson.

13.1 Categorical independent variables

We have explored simple linear regression with a continuous independent variable and with a binary variable. We now extend these ideas to include a categorical independent variable.

We will return to the baby example and use linear regression to explore the association between maternal Body Mass Index (BMI) category and the baby's birthweight.

13.1.1 Dummy variables

We have height measured in inches and weight measured in pounds. BMI is obtained using the formula $BMI = 703 \times weight\ (lb) / height\ (in)^2$. We will then categorise BMI according to the World Health Organisation's classification.

We will define a categorical variable c_{i} denoting the i^{th} mother's BMI category, defined as follows:
$$\begin{split} c_{i} = \begin{cases} 1 & \text{ if the mother's BMI is less than 18.5 (underweight)} \\ 2 & \text{ if the mother's BMI is at least 18.5 and less than 25 (normal)} \\ 3 & \text{ if the mother's BMI is at least 25 and less than 30 (overweight)} \\ 4 & \text{ if the mother's BMI is 30 or more (obese)} \end{cases} \end{split}$$
Our BMI categorical variable has four categories. To distinguish between all four categories we need *three* dummy variables. We choose a *baseline* or *reference* group, which for us will be the underweight category. For each of the other categories, we create a dummy variable which indicates that the woman is in (or not) that category. Specifically, we define our dummy variables as:
$$\begin{split} w_{1i} = \begin{cases} 1 & \text{ if } c_{i}=2\\ 0 & \text{ if } c_{i} \neq 2 \end{cases} \end{split}$$
and
$$\begin{split} w_{2i} = \begin{cases} 1 & \text{ if } c_{i}=3\\ 0 & \text{ if } c_{i} \neq 3 \end{cases} \end{split}$$
and
$$\begin{split} w_{3i} = \begin{cases} 1 & \text{ if } c_{i}=4\\ 0 & \text{ if } c_{i} \neq 4 \end{cases} \end{split}$$
The R code below read in the baby data and create variables containing the mother's BMI (`Maternal.BMI`) and the mother's BMI category (`Maternal.BMIcat`).

```
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
# Calculate maternal BMI (with conversion factor due to measurement in lb and in)
data$Maternal.BMI <- 703*data$Maternal.Pregnancy.Weight/(data$Maternal.Height)**2

# Categorise the BMI values
data$Maternal.BMIcat <-1
data$Maternal.BMIcat[data$Maternal.BMI>=18.5 & data$Maternal.BMI<25]<-2
data$Maternal.BMIcat[data$Maternal.BMI>=25   & data$Maternal.BMI<30]<-3
data$Maternal.BMIcat[data$Maternal.BMI>=30]<-4

# Tabulate the BMI categories
table(data$Maternal.BMIcat)
```

```
  1   2   3   4
 84 932 124  34
```

13.1.1 The model

A linear regression model relating birthweight (Y, the outcome) to the three dummy variables (W_1, W_2, W_3) representing the mother's BMI category (C, the categorical independent variable) is defined as:
$$\text{Model 3: } y_i = \beta_0 + \beta_{1}w_{1,i} + \beta_{2}w_{2,i} + \beta_{3}w_{3,i} + \epsilon_i$$
The equation above can also be written as follows:
$$\begin{split} \begin{aligned} y_i &= \beta_0 + \epsilon_i \text{ if } c_i=1 \text{ (underweight mothers) } \\ y_i &= \beta_0 + \beta_1 + \epsilon_i \text{ if } c_i=2 \text{ (normal weight mothers) } \\ y_i &= \beta_0 + \beta_2 + \epsilon_i \text{ if } c_i=3 \text{ (overweight mothers) } \\ y_i &= \beta_0 + \beta_3 + \epsilon_i \text{ if } c_i=4 \text{ (obese mothers) } \\ \end{aligned} \end{split}$$
This makes explicit the interpretation of the parameters in the model.

- β_0 is the expectation of Y when $C=1$
- $\beta_0 + \beta_1$ is the expectation of Y when $C=2$. Hence β_1 is the difference in the expectation of Y between groups defined by $C=1$ and $C=2$.
- $\beta_0 + \beta_2$ is the expectation of Y when $C=3$. Hence β_2 is the difference in the expectation of Y between groups defined by $C=1$ and $C=3$.
- $\beta_0 + \beta_3$ is the expectation of Y when $C=4$. Hence β_3 is the difference in the expectation of Y between groups defined by $C=1$ and $C=4$.

Notes

> In this parameterisation of the model, the group defined by $C=0$ is often referred to as the baseline group. There is no statistical reason why one group rather than another should be chosen as the baseline group. It can sometimes be desirable to re-parameterise a model of this type to estimate parameters representating differences in mean levels from a particular baseline group. In this example, for instance, we might instead want the group with normal weight to be our baseline group, in which case we would need to redefine our first dummy variable.
>
> Note that, in contrast to the models that we have met so far, this has more than one variable in it (even though the three dummy variables together measure a single characteristic). Therefore, this is no longer strictly a simple linear regression model. It is an example of a multivariable linear regression model. We will discuss general theory for this model later. Broadly speaking, the ideas we have met in the context of simple linear regression extend to this more general model very naturally.

```
# Model 3: Relating birthweight to length of pregnancy and mother's height group.
model3<-lm(Birth.Weight~factor(Maternal.BMIcat), data=data)
summary(model3)
```

```
Call:
lm(formula = Birth.Weight ~ factor(Maternal.BMIcat), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-64.828 -10.828   0.172  11.172  57.677

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              115.286      1.996  57.752   <2e-16 ***
factor(Maternal.BMIcat)2    4.543      2.084   2.180   0.0295 *
factor(Maternal.BMIcat)3    3.037      2.585   1.175   0.2404
factor(Maternal.BMIcat)4    8.626      3.719   2.320   0.0205 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.3 on 1170 degrees of freedom
Multiple R-squared:  0.006152,  Adjusted R-squared:  0.003604
F-statistic: 2.414 on 3 and 1170 DF,  p-value: 0.06511
```

- $\hat{\beta}_0 = 115.29$. This is interpreted as the estimated mean birthweight (in oz) of a baby with all “dummy” variables equal to 0, i.e. it is the estimated mean birthweight of babies in our baseline category (those with mothers who are underweight).
- $\hat{\beta}_1= 4.543$. The mean birthweight is estimated to increase by 4.5 oz per unit increase in the first “dummy” variable. A unit increase in the first dummy variable equates to moving from the underweight group to the normal weight group. So we can interpret this as the difference in mean birthweights between the group whose mothers have normal BMI and those whose mothers are underweight.
- $\hat{\beta}_2= 3.037$. The mean birthweight is estimated to increase by 3.0 oz per unit increase in the second “dummy” variable. A unit increase in the second dummy variable equates to moving from the underweight group to the overweight group. So we can interpret this as the difference in mean birthweights between the group whose mothers are overweight and those whose mothers are underweight.
- $\hat{\beta}_3= 8.626$. The mean birthweight is estimated to increase by 8.6 oz per unit increase in the third “dummy” variable. A unit increase in the first dummy variable equates to moving from the underweight group to the obese group. So we can interpret this as the difference in mean birthweights between the group whose mothers are obese and those whose mothers are underweight.

Overall, we see a pattern of higher maternal BMI being associated with higher birthweights, particularly for the group with obese mothers.

- $\hat{\sigma}$=18.3. The observed outcomes are scattered around the fitted regression line with a standard deviation of 18.3oz.

We can obtain confidence intervals around the three BMI estimates as follows:

```
confint(model3, level=0.95)
```

	2.5 %	97.5 %
(Intercept)	111.3691529	119.202276
factor(Maternal.BMIcat)2	0.4533602	8.631864
factor(Maternal.BMIcat)3	-2.0356770	8.109410
factor(Maternal.BMIcat)4	1.3296865	15.922414

A matrix: 4 × 2 of type dbl

The R output from the model provides p-values for each of the three coefficients relating maternal BMI to birth weight. However, we are typically interested in the broad question of whether maternal BMI is related to birth weight, rather than whether an individual dummy variable is related to the outcome.

Therefore, when we have a categorical variable in a regression model, the hypothesis of interest usually relates to the combination of all dummy variables representing the categorical variable. An appropriate hypothesis test jointly tests the hypothesis that all coefficients for dummy variables are zero, i.e.

- $H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$
- $H_1: \text{at least one of } \beta_1, \beta_2, \beta_3 \neq 0$

We use a partial F-test to test this hypothesis. Details are beyond the scope of the course, but are outlined in the appendix to this session. The R code to obtain the joint p-value is shown below.

```
# Remove maternal BMI from the model (i.e. a constant-only model)
model3_without<-lm(Birth.Weight~1, data=data)

anova(model3, model3_without)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1170	391633.5	NA	NA	NA	NA
2	1173	394057.9	-3	-2424.344	2.414232	0.06510651

A anova: 2 × 6

In this case we have a p-value of p=0.065, indicating some evidence against the null hypothesis of no association bewteen maternal BMI category and the baby's birthweight.

13.1.2 Categorising continuous variables

In the above example, we have categorised an continuous variable (BMI) in order to demonstrate how a categorical variable should be included in a linear regression model. This is important to know, since there are many variables that are categorical by definition and may be required for a statistical analysis. For example: cancer stage, ethnicity, education level, etc. While these examples should be included as a categorical variable in a linear regression model, it is not, in general, recommended to categorise a continuous variable in a linear model. We did so above, purely for pedagogical reasons.

One of the problems with categorising continuous variables is that it is difficult to decide what the cut-off for each category should be. In the example above, however, there are widely used categorisations.

We often lose information by categorising continous variables. We can often obtain a better and more parsimonious fit (using fewer parameters to describe the relationship) by modelling the continuous variable without categorisation. We will return to these ideas later.

13.2 Multivariable linear regression

Multivariable linear regression extends the simple linear regression model to situations in which we wish to relate two or more independent variables to one outcome. Where there are multiple independent variables, we will refer to them as **covariates**.

There can be a number of different reasons why we would want to add more covariates in our linear regression model. Recall these two examples of questions we might want to answer using statistical models (given at the beginning of this lesson):

- Does taking drug A reduce inflammation more than taking drug B in patients with arthritis?
- Can we predict the risk of heart disease for our patients?

In the first example, we could use a statistical model with inflammation as the outcome and drug use as the independent variable of interest, but we may need to **control** (or **adjust**) for the **confounding** effects of other patient characteristics (age, gender, other medication use, etc.). In the second example, there are many different factors that could be associated with risk of heart disease (age, gender, lifestyle choices, etc.) and we may wish to include all such factors in a statistical model to predict heart disease.

Here, we introduce the multivariable linear regression model and describe how to estimate and interpret the parameters in the model using an example from the birthweight data.

Note

> *A note on notation.* There can be some confusion between the terms **multivariable** models and **multivariate** models. Multivariate models are those which have more than one outcome variable. Such models are beyond the scope of this module; we focus our attention on **univariate** models which have only one outcome. Both simple linear regression models and multivariable linear regression models are considered as univariate.

13.2.1 The multivariable linear regression model

Suppose we wish to relate an outcome (\(Y\)) to \(p\) predictor variables \((X_1, X_2, ..., X_p)\). The appropriate multivariable linear regression model is a straightforward extension of the simple linear regression model:
\[y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ,..., \beta_p x_{ip}+\epsilon_i \text{ with } \epsilon_i \overset{iid}{\sim} N(0,\sigma^2). \]
where, \(y_i\) is the value of the dependent variable for the ith participant and \(x_{ji}\) is the value of the jth predictor variable for the ith participant.

The parameters in the model are interpreted as follows:

- \(\beta_0\) is the intercept. It is the expectation of \(Y\) when all the \(X_j\)'s are zero.
- \(\beta_j\) is the expected change in \(Y\) for a 1 unit increase in \(X_j\) *with all the other covariates held constant*.

The \(\beta_j\)'s are the **regression coefficients** (otherwise known as **partial regression coefficients**). Each one measures the effect of one covariate controlled (or adjusted) for all of the others.

13.2.2 The multivariable linear regression model in matrix notation

Similarly to the simple linear regression model, the multivariable linear regression model can be expressed using matrix algebra.
\[\mathbf{Y}=\mathbf{X}\mathbf{\beta}+\mathbf{\epsilon} \text{ where } \epsilon \sim N(0,\mathbf{I}\sigma^2) \]
\[\begin{split} \begin{vmatrix}y_1\\y_2 \\. \\. \\y_n \end{vmatrix}=\begin{vmatrix}1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\1 & . \\1 & . \\ 1& . \\1 & x_{p1} & x_{p} & ... & x_{pn} \end{vmatrix}\begin{vmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{vmatrix}+\begin{vmatrix}\epsilon_1\\ \epsilon_2 \\ . \\ . \\. \\ \epsilon_n \end{vmatrix} \end{split}\]
In this formulation, \(\mathbf{X}\) is an \(n \times (p+1)\) matrix, \(Y\) and \(\epsilon\) are vectors of length \(n\) whilst \(\mathbf{\beta}\) is a vector of length \((p+1)\).

13.2.3 Estimation of the parameters

The regression coefficients in multivariable linear regression can be estimated by minimising the residual sum of squares:
\[\begin{split} \begin{align} SS_{RES} &= \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i-\hat{y})^2 \\ &= \sum_{i=1}^n (y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i}-...-\hat{\beta}_px_{pi})^2 \end{align} \end{split}\]
The closed form solution, obtained by solving the \((p+1)\) simultaneous equations that result from setting the partial derivatives of the above equation with respect to each parameter estimate to zero, can be written succinctly using matrix notation:
\[\mathbf{\hat{\beta}}= (\mathbf{X'X})^{-1}X'Y \]

$\mathbf{\hat{\beta}}$ is an unbiased estimator of $\mathbf{\beta}$. Its distribution is as follows:
$$ \mathbf{\hat{\beta}} \sim \mathbf{N(\beta, (X'X)^{-1}\sigma^2)}. $$
This expresses the fact that the elements of $\mathbf{\hat{\beta}}$ follow a multivariate normal distribution whose variances and covariances are given by $\mathbf{(X'X)^{-1}\sigma^2}$.

It can also be shown that the following is an unbiased estimator for σ^2:
$$\begin{split} \begin{align} \hat{\sigma}^2 &= \sum_{i=1}^n \frac{\hat{\epsilon}_i^2}{(n-(p+1))}\\ &=\sum_{i=1}^n \frac{(y_i - \hat{\beta}_0} - \hat{\beta}_1x_{1i}- ... - \hat{\beta}_px_{ip})^2}{(n-(p+1))} \end{align} \end{split}$$
While it is useful to know how these parameters are estimated, in practice they are often obtained using statistical software. Next, we demonstrate how to perform multivariable regression in R using the birthweight data and discuss the interpretation of the estimated regression coefficients.

13.3 Including multiple covariates

We are interested in investigating a model that relates birthweight to length of pregnancy and mother's height. We will use the following multivariable linear regression model:
$$ \text{Model 4: } y_i = \beta_0 + \beta_1 l_i + \beta_2h_i + \epsilon_i $$
The outcome y_i denotes the birthweight (in oz) for the i^{th} baby. The predictors l_i and h_i denote the length of pregnancy (i.e. number of gestational days), and the height of the mother (in inches), for the i^{th} baby, respectively.

The linear regression can be conducted in R using the `lm()` command:

```
# Model 4: Relating birthweight to length of pregnancy and mother's height
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
model4<-lm(Birth.Weight~Gestational.Days+Maternal.Height, data=data)
summary(model4)
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days + Maternal.Height,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-53.829 -10.589   0.246  10.254  54.403

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -88.51993   14.31910  -6.182 8.73e-10 ***
Gestational.Days   0.45237    0.03006  15.051  < 2e-16 ***
Maternal.Height    1.27598    0.19049   6.698 3.27e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.44 on 1171 degrees of freedom
Multiple R-squared:  0.1969,    Adjusted R-squared:  0.1955
F-statistic: 143.5 on 2 and 1171 DF,  p-value: < 2.2e-16
```

Interpretation of the regression coefficients

- $\hat{\beta}_1=0.45$. This is the estimated regression coefficient for number of gestational days. It is interpreted as: the expected increase in a baby's birthweight for each gestational day, *amongst babies whose mothers were of the same height*, is 0.45 ounces.

It may be tempting to make causal inference from regression models such as Model 4, i.e. "longer pregnancies **cause** an increase in birthweight". However, this is far from straightforward. Based on the results presented above, it would be reasonable to say that "birthweight increases with length of pregnancy". However, it is much less reasonable to claim that higher birthweight is caused by longer pregnancies (based on these results alone), because there may be an unobserved third variable that is the "real" cause of both increased length of pregancy and birthweight. Causal statements require more than just the results of a statistical model to make them plausible; this is a topic that we return to in the next lesson.

Excerise: What is the interpretation of $\hat{\beta}_2$?

Interpretation of the intercept

- $\hat{\beta}_0=-88.52$. The interpretation is that the estimated mean birthweight for a child who was born after 0 gestastional days and whose mother's height is 0 inches is -88.52 ouces. Clearly this is an absurd value to estimate because no babies are born that quickly and no mothers are that short. If we wish to obtain a more

reasonable intercept, we can use a technique called **centering**.

13.4 Centering

In many analyses, interpreting the intercept is not as important as interpreting the estimated regression coefficients and so it does not matter if our intercept is an absurd value (as in the example above). However, if we do wish to obtain an interpretable intercept, we can **center** the independent variables.

Centering a variable means subtracting a constant from every value of the variable. This essentially shifts the scale of the predictor (the point 0 is shifted to the chosen constant), but does not affect the units of the variable. Consequently, the new interpretation of the intercept would be the mean of \(Y\) when the independent variable is equal to the constant. The estimated regression coefficient of the independent variable is not affected.

As as example, we will repeat the analysis above, but center each of the covariates on their mean value.

```
# What are the mean gestational days and mothers height in our data?
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
summary(data$Gestational.Days)
summary(data$Maternal.Height)

# Create new (centered) variables in our data
data$Gestational.Days.Centered<-data$Gestational.Days-mean(data$Gestational.Days)
data$Maternal.Height.Centered<-data$Maternal.Height-mean(data$Maternal.Height)

# Redefine Model 4 using the centered variables
model4<-lm(Birth.Weight~Gestational.Days.Centered+Maternal.Height.Centered, data=data)
summary(model4)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  148.0   272.0   280.0   279.1   288.0   353.0 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  53.00   62.00   64.00   64.05   66.00   72.00 
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days.Centered + Maternal.Height.Centered, 
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.829 -10.589   0.246  10.254  54.403 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               119.46252    0.47980 248.983  < 2e-16 ***
Gestational.Days.Centered   0.45237    0.03006  15.051  < 2e-16 ***
Maternal.Height.Centered    1.27598    0.19049   6.698 3.27e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.44 on 1171 degrees of freedom
Multiple R-squared:  0.1969,    Adjusted R-squared:  0.1955 
F-statistic: 143.5 on 2 and 1171 DF,  p-value: < 2.2e-16
```

Now the intercept (\(\hat{\beta}_0\)) is equal to 119.46. This is interpreted as: the estimated mean birthweight for a child who was born after 279.1 gestastional days and whose mother's height is 64.05 inches is 119.46 ouces. Additionally, notice that the estimated regression coefficients for gestational days and mother's height, and their associated standard errors have not changed.

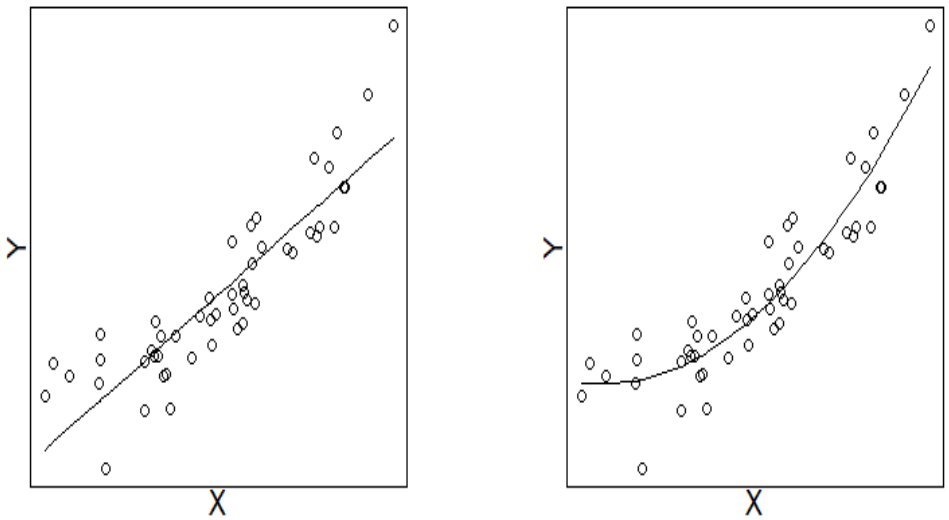
13.6 Including higher-order terms

As we have already discussed, linear regression assumes that the relationship between the outcome and the independent variables is linear. As we already know, this is not always the case in real data. For example, suppose we are interested in the association between weight and age. On average, the weight of young adults will increase with age. However, at a certain age, the average weight may start to decrease. In this case, the association between weight and age would follow a non-linear (upside-down) \(u\)-shape. It could still be possible to model this relationship within the linear regression framework, by adding a **second-order term** to the model. This procedure is known as **quadratic regression**.

13.6.1 The quadratic regression model

The quadratic regression model is a multivariable regression model with two independent variables where the second variable is the square of the first variable. Algebraically:
$$ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \text{ where } \epsilon_i \overset{\small{iid}}{\sim} N(0, \sigma^2). $$
Despite the fact that one of the variables is the square of the other, this is still a linear regression model because the expectation of the outcome is a linear function of both parameters.

The figure below shows two scatter plots of the data used in Scenario A above. The plot on the left-hand side includes the fitted values of a linear regression model (with no higher-order terms included) and the right-hand side plot includes the fitted values of a quadratic regression model. By comparing the plots, we can see that the quadratic regression model does have a better fit, particularly at the extreme values of X.



Unfortunately, interpreting β_1 and β_2 is not as straightforward as in most linear models. The reason for this is that it is not possible to change X^2 by 1 unit whilst holding X constant.

13.6.2 Example

Suppose the outcome, birthweight, is denoted Y and length of pregnancy (gestational days) is denoted by L. The original linear model we considered was:
$$ \text{Model 1: } y_i = \beta_0 + \beta_1 l_i + \epsilon_i $$
We now extend this to allow a quadratic relationship between L and Y:
$$ \text{Model 5: } y_i = \beta_0 + \beta_1 l_i + \beta_2 l^2_i + \epsilon_i $$

```r
# What are the mean gestational days and mothers height in our data?
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')

# Add quadratic term:
model5<-lm(Birth.Weight~Gestational.Days+I(Gestational.Days**2), data=data)
summary(model5)
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days + I(Gestational.Days^2),
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.527 -10.980   0.190   9.973  69.655

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -6.901e+01  5.043e+01  -1.368    0.171
Gestational.Days      8.962e-01  3.678e-01   2.436    0.015 *
I(Gestational.Days^2) -7.890e-04  6.731e-04  -1.172    0.241
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.74 on 1171 degrees of freedom
Multiple R-squared:  0.1671,    Adjusted R-squared:  0.1656
F-statistic: 117.4 on 2 and 1171 DF,  p-value: < 2.2e-16
```

The estimated coefficient for the quadratic term is very small (-0.000789). The p-value, testing the null hypothesis that \(H_0: \beta_2 = 0\) is p=0.241, suggesting no evidence against the null hypothesis. Therefore, we conclude there is no evidence of a quadratic relationship.

13.6.3 More complex models

Quadratic regression models are limited in terms of describing relationships between variables in most medical applications. Quadratic functions either increase to a maximum and the decline, or fall to a minimum and then increase. Further, the behaviour of a quadratic is symmetric about the turning point. Such relationships in medical research are rarely plausible.

There are a number of alternative approaches that can be used to model complex relationships between continuous independent variables and the outcome within a linear regression model. Some are discussed below.

The quadratic regression model belongs to a family of **polynomial regression models** and is the simplest model in that family. Further power terms can be added to the regression model in order to increase complexity. For example, a cubic regression model is one which includes a cubic term as well as a squared term.

An even more flexible approach is to use a **piecewise polynomial model**, which allows for a different polynomial function in different ranges of the observed values of \(X\), defined according to specified **knots**. The flexibility of the model (and therefore its ability to model more complex relationships) can be increased by increasing the degree of polynomial and/or the number of knots. However, highly flexible models may overfit the data and make the model difficult to interpret. In general, it is a good idea to consider an appropriate trade-off between flexibility and interpretability

A related idea is to use **splines** to flexibly model the relationship. These are a type of piecewise polynomial model, where the adjacent polynomials are constrained to meet at the join points (the knots).

13.7 Modelling interaction terms

Suppose we fit a multivariable linear regression model relating the outcome of weight to the covariates age, sex and height, for adults in the general population. In this case, the estimated regression coefficient for height represents the effect of a unit increase in height on weight in people of the same age and sex. The model assumes that the coefficient relating weight to height is the same for all people of all ages and sexes. For example, that it is the same for twenty year old men as in ninety-three year old women. But this is not necessarily true! It could be that the slope of the association between weight and height differs by sex and by age. If this is the case, we say there is an **interaction** between height and sex and between height and age.

The term **interaction** is used to describe situations in which the relationship between \(Y\) and \(X\) differs according to the level of one or more other covariates.

13.7.1 Linear regression with an interaction term

Suppose we wish to relate an outcome (\(Y\)) to two covariates (\(X_1\) and \(X_2\)), but we want to allow the association between \(Y\) and \(X_1\) to differ according to the value of \(X_2\). To allow for this we fit an interaction model that contains an additional variable (\(X_3\)) that is the product of \(X_1\) and \(X_2\):

$$ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \text{ where } \epsilon_i \overset{\small{iid}}{\sim} N(0, \sigma^2). $$

- y_i = value of the outcome for the i^{th} observation
- x_{1i} = value of the first covariate for the i^{th} observation
- x_{2i} = value of the second covariate for the i^{th} observation
- x_{3i} = $x_{1i} \times x_{2i}$

To understand why this model allows the association between Y and X_1 to vary according to X_2, we can consider the form of the equation when we fix X_2 to have a particular value, say $X_2=k$. In this situation the relationship between Y and X_1 is as follows:
$$ y_i = (\beta_0 + \beta_2k) + (\beta_1 + \beta_3k)x_{1i} + \epsilon_i. $$
In other words, when $x_2=k$ the relationship between Y and X_1 is a linear one with both slope and intercept dependent upon k. The intercept is $\beta_0+\beta_2k$ and the slope is $\beta_1 + \beta_3k$.

By allowing the association between Y and X_1 to vary according to X_2, we have also allowed the slope for the association between Y and X_2 to vary according to X_1. If we look at the form of the model when X_1 takes particular value, say $X_1=m$, we find:
$$ y_i = (\beta_0+\beta_1m) + (\beta_2+\beta_3m)x_{2i} + \epsilon_i. $$
Again, the relationship between Y and X_2 is a linear one with both slope and intercept dependent upon m.

13.7.2 Interaction between a continuous variable and a binary variable

The interaction model is particularly easy to interpret when one of the covariates (say X_2) is a binary, taking the values 0 and 1 (i.e. a dummy variable). The linear regression model then becomes:
$$\begin{split} \begin{aligned} &y_i = \beta_0 + \beta_1x_{1i} + \epsilon_i \text{ when } x_2=0\\ &y_i = (\beta_0 + \beta_2) + (\beta_1+\beta_3)x_{1i} + \epsilon_i \text{ when } x_2=1 \end{aligned} \end{split}$$
The interpretation of each of the parameters is as follows.

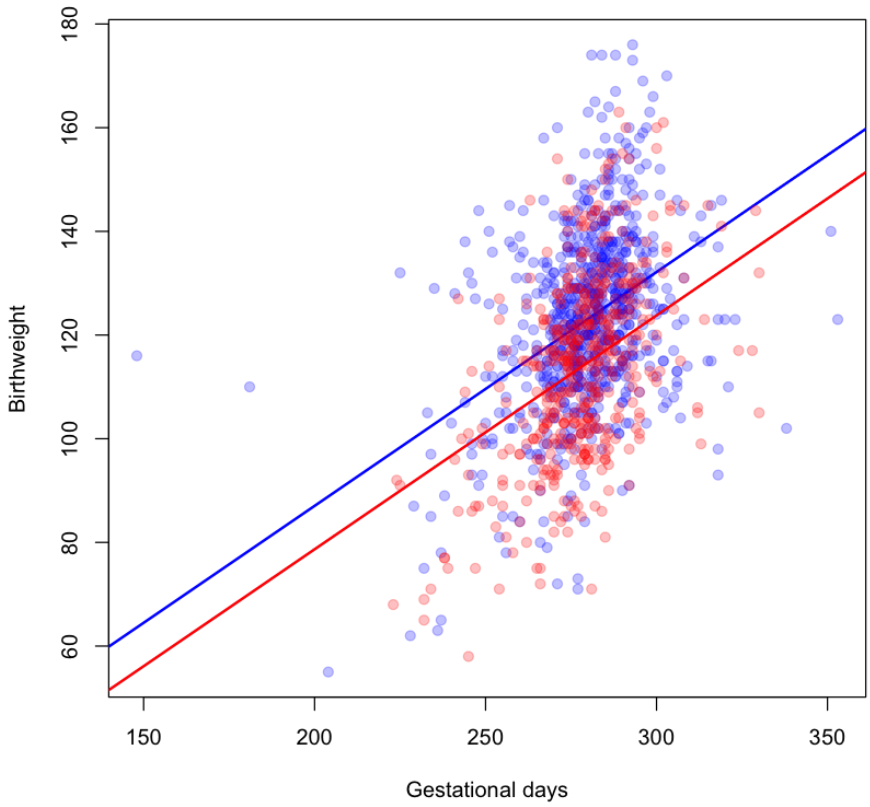
- β_0 is the intercept when $X_2=0$
- $\beta_0 + \beta_2$ is the intercept when $X_2=1$
- β_2 is the difference in intercepts between the two groups defined by X_2.
- β_1 is the slope when $X_2=0$
- $\beta_1+\beta_3$ is the slope when $X_2=1$
- β_3 is the difference in slopes between the two groups defined by X_2.

Example. To demonstrate the impact of adding an interaction term, we will consider two models: (1) relating birthweight (Y) to length of pregnancy (X_1) and mother's smoking status (X_2) and (2) relating birthweight (Y) to length of pregnancy (X_1), mother's smoking status (X_2) and their interaction (X_3). In these models, $X_2=1$ indicates that the mother smokes and $X_2=0$ indicates that the mother does not smoke.

We first consider the model with no interaction term. The code below defines the model in R, summarises the results and produces a scatter plot of birthweight against gestational days, with the fitted values superimposed. The blue points (and line) on the scatter plot are observations in the group of babies whose mothers do not smoke and the red points (and line) are observations in the group of babies whose mothers do smoke.

```
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
#Define a dummy variable for maternal.smoker
data$Maternal.Smoker2<-0
data$Maternal.Smoker2[data$Maternal.Smoker=="True"]<-1

#Model without the interaction term
no_int_model<-lm(data$Birth.Weight~data$Gestational.Days+factor(data$Maternal.Smoker2))
summary(no_int_model)

#Scatterplot
plot(x = data[data$Maternal.Smoker2 == 0, ]$Gestational.Days, y =
data[data$Maternal.Smoker2 == 0, ]$Birth.Weight,
     pch = 19, xlab = "Gestational days", ylab = "Birthweight", col = rgb(red = 0,
green = 0, blue = 1, alpha = 0.25))
abline(a = no_int_model$coefficients[1], b = no_int_model$coefficients[2], col =
"blue", lwd = 2)
points(x = data[data$Maternal.Smoker2 == 1, ]$Gestational.Days, y =
data[data$Maternal.Smoker2 == 1, ]$Birth.Weight,
       col = rgb(red = 1, green = 0, blue = 0, alpha = 0.25), pch = 19)
abline(a = coef(no_int_model)[1] + coef(no_int_model)[3], b = coef(no_int_model)[2],
       col = "red", lwd = 2)
```

```
Call:
lm(formula = data$Birth.Weight ~ data$Gestational.Days +
factor(data$Maternal.Smoker2))

Residuals:
    Min      1Q  Median      3Q     Max
-50.789 -11.035  -0.211  10.053  52.412

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3.18492    8.32945  -0.382    0.702
data$Gestational.Days            0.45117    0.02968  15.200   <2e-16 ***
factor(data$Maternal.Smoker2)1 -8.37440    0.97346  -8.603   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 1171 degrees of freedom
Multiple R-squared:  0.2157,    Adjusted R-squared:  0.2143
F-statistic:   161 on 2 and 1171 DF,  p-value: < 2.2e-16
```

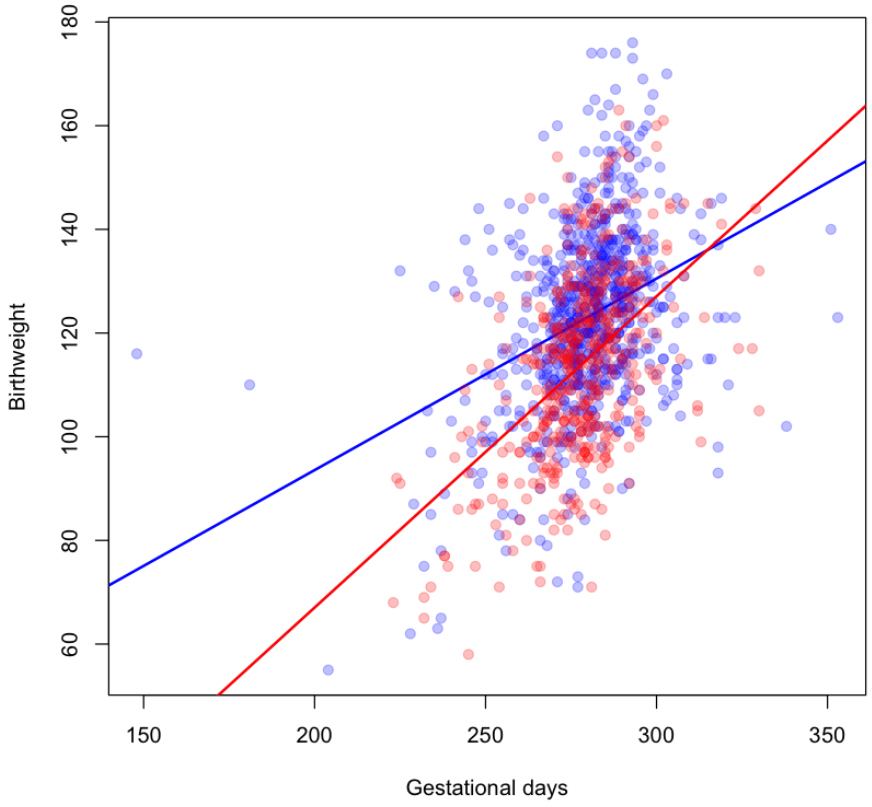


As can be seen from the above figure, the fitted values from the model with no interaction term form two straight lines with a common slope 0.45 ounces and intercepts -3.18 ounces for the non-smoking group and -3.18-8.37=-11.55 ounces for the smoking group. This type of model (no interactions) is sometimes known as a **parallel lines** regression model, because it restricts the lines to be parallel. It permits adjustment of the effect of one covariate for the effects of others, but forces the effects of a unit change in each covariate to be constant, whatever the level of the other covariate. This restriction is not appropriate if the slope effect of one covariate depends on the value of another covariate. Adding an interaction term removes this restriction.

Below, we fit the second model which includes an interaction term, and produce a second scatter plot with the fitted values from our new model superimposed.

```r
#Create the interaction term
data$int1<-data$Gestational.Days*data$Maternal.Smoker2

#Include the interaction term in our model
int_model1<-
lm(data$Birth.Weight~data$Gestational.Days+factor(data$Maternal.Smoker2)+data$int1)
summary(int_model1)

#Scatter plot
plot(x = data[data$Maternal.Smoker2 == 0, ]$Gestational.Days, y =
data[data$Maternal.Smoker2 == 0, ]$Birth.Weight,
     pch = 19, xlab = "Gestational days", ylab = "Birthweight", col = rgb(red = 0,
green = 0, blue = 1, alpha = 0.25))
abline(a = int_model1$coefficients[1], b = int_model1$coefficients[2], col = "blue",
lwd = 2)
points(x = data[data$Maternal.Smoker2 == 1, ]$Gestational.Days, y =
data[data$Maternal.Smoker2 == 1, ]$Birth.Weight,
       col = rgb(red = 1, green = 0, blue = 0, alpha = 0.25), pch = 19)
abline(a = coef(int_model1)[1] + coef(int_model1)[3], b = coef(int_model1)[2] +
coef(int_model1)[4],
       col = "red", lwd = 2)
```

```
Call:
lm(formula = data$Birth.Weight ~ data$Gestational.Days +
factor(data$Maternal.Smoker2) +
    data$int1)

Residuals:
    Min      1Q  Median      3Q     Max
-51.023 -11.078  -0.084   9.995  50.499

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    19.63964   10.29098   1.908 0.056580 .
data$Gestational.Days           0.36962    0.03671  10.069  < 2e-16 ***
factor(data$Maternal.Smoker2)1 -72.68713   17.23243  -4.218 2.65e-05 ***
data$int1                       0.23085    0.06176   3.738 0.000194 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.16 on 1170 degrees of freedom
Multiple R-squared:  0.2249,    Adjusted R-squared:  0.2229
F-statistic: 113.2 on 3 and 1170 DF,  p-value: < 2.2e-16
```



In our new model, the intercept and slope among the non-smoking group are 19.64 ounces and 0.37 ounces respectively. The intercept and slope among the smoking group are 19.64-72.69=-53.05 ounces and 0.37+0.23=0.60 ounces respectively. The interaction term has p=0.0001, so there is evidence that the slopes are different.

14. Linear Regression III

This is the final session on Linear Regression. In this session we focus on the assumptions underlying this model.

Intended learning outcomes

By the end of this session you will be able to:

- explain the assumptions underlying multivariable linear regression
- apply a range of graphical techniques to investigate the assumptions of the linear regression model;

Acknowledgements: Thank you to Jennifer Nicholas, Chris Frost and Ruth Keogh whose notes on linear regression and generalised linear models were particularly useful in the development of the current session.

14.1 Assumptions

The linear regression model makes a number of assumptions. All inferences made from a model are contingent on these assumptions being correct. It is therefore important that we have statistical techniques (or **diagnostic tools**) to investigate these assumptions.

In practice, it is rare for all the assumptions of a statistical procedure to hold exactly. We may have evidence in the data, or prior knowledge about the data, that lead us to believe that the assumptions made by the model do not hold. This does not necessarily mean that the results from the model should be disregarded, since statistical procedures are **robust** to departures from assumptions in many settings. When conducting statistical analyses, it is a good idea to first try to establish to what extent assumptions hold and then consider whether the methods used can be adapted to improve the extent to which assumptions hold. If adaptations cannot be made, it is necessary to consider to what extent the results of an analysis can be trusted.

In this section we largely focus on diagnostic tools that can be used to identify assumption violations. Some pointers are given to possible adaptations and alternative techniques that can be used when assumptions are violated, however issues of robustness are not considered in great detail. It is worth noting that, broadly speaking, the central limit theorem implies that departures from assumptions are less important for large datasets than for small ones, and so assumption violations are less of a concern when working with big data.

14.1.1 Assumptions of the linear regression model

The assumptions made by the linear regression model are as follows:

1. **Linearity:** There is a linear relationship between the dependent variable Y and each of the independent variables. Here we are contrasting a linear relationship with a non-linear relationship, not with no relationship. A model in which one of the regression coefficients is zero can satisfy the assumptions of linear regression.
2. **Normality:** The error terms follow a normal distribution.
3. **Homoscedasticity:** The error variance is constant i.e. the scatter of points around the true regression line has the same variance, irrespective of the value of x_i. The converse of this feature is termed **heteroscedasticity**.
4. **Independence:** The observations of y_i are independent.

In this session we will focus on the first three assumptions. Violations of the independence assumption are often more apparant from the context of a study than from the data itself. For example, if we carry out a study in which the blood pressure of 100 people are each measured twice, and then treat the 200 measurements as independent in the statistical analysis it is clear that the assumption of independence is violated.

Notice that the normality and homoscedasticity assumptions concern the error terms, which can be thought of as the *true* residuals, defined in terms of deviations from the model defined by population parameters. Since these errors or true residuals can never be observed in practice, we have to use the observed residuals (obtained by replacing the population parameters with their estimates). In fact, observed residuals are neither independent nor do they have constant variance, but in most settings the departures from independence and homoscedasticity are very small. Consequently, we can proceed as if the observed residuals were the true residuals when investigating assumptions.

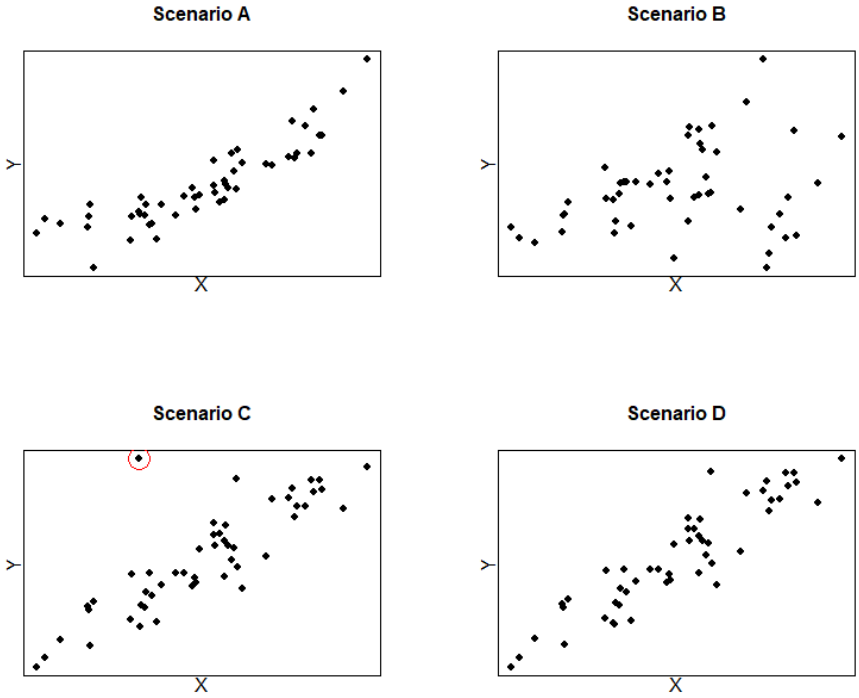
14.2 Investigating assumptions using plots

It is a good idea to explore your data using a number of simple plots. Here we will introduce the most useful plots for both simple and multivariable linear regression models.

14.2.1 Scatter plots of the outcome against independent variables

For simple linear regression models, a scatter plot of the outcome against the independent variable can usually make serious violations of assumptions apparent. Such plots are particularly good for identfiying non-linearity, heteroscedasticity and **outliers** (points which lie atypically far from the regression line).

Let our outcome and independent variable be denoted by \(Y\) and \(X\), respectively. The figure below depicts four different scenarios where various assumptions are violated. In Scenario A, there is a slight curvature in the scatter of points between \(Y\) and \(X\), suggesting a non-linear relationship which violates the linearity assumption. In Scenario B, the variance of \(Y\) is larger for larger values of \(X\), violating the homoscedasticity assumption. In Scenario C, the linearity and homoscedasticity assumptions appear to hold, but there is a possible outlier (circled in red). Scenario D depicts an ideal situation for simple linear regression, where there appears to be no violations.

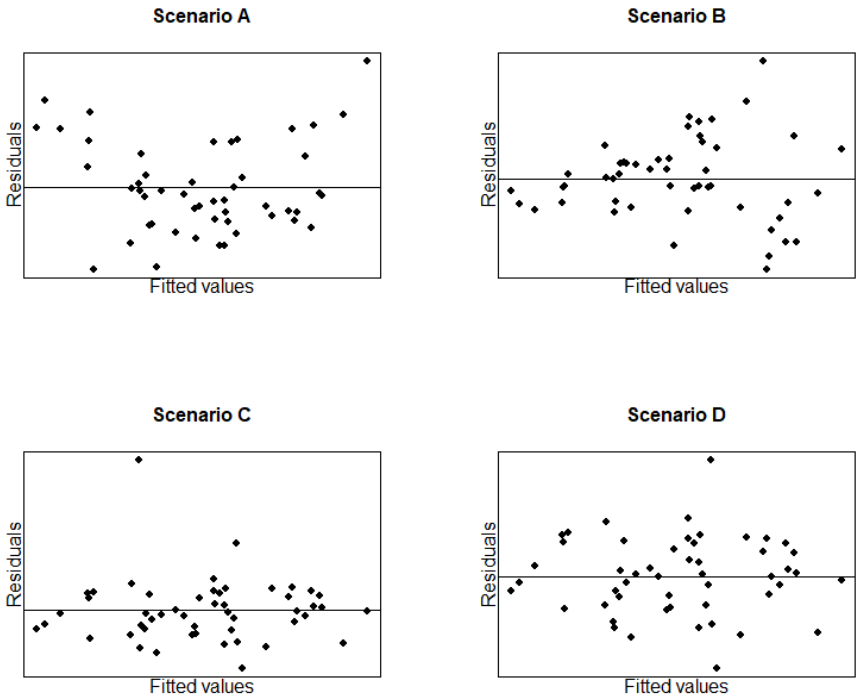


For multivariable linear regression, the linearity assumption requires that the relationship between the outcome and each independent variable is linear *conditional on the other covariates in the model*. So, there is no requirement that the relationship between the outcome and each individual covariate is linear when other covariates are ignored. This means that assessment of the fit of a multivariable linear regression cannot be inferred from a series of scatter plots relating the outcome to each covariate. Such plots can be useful for detecting points with extreme values, but the residual plots considered next are more useful for multivariable models.

14.2.2 Plots of residuals against fitted values or covariates

Plots of the observed residuals against the fitted values are useful for investigating the assumptions of linearity and homoscedasticity. For linearity: if a non-linear relationship is present, then the residuals will not be equally distributed above and below zero across the range of fitted values. For homoscedasticity: if there is heterogeneity in the residuals, the variance of residuals will not be constant across the range of fitted values.

The figure below uses the same data from Scenarios A-D above, but displays the observed residuals against fitted values. We can see that linearity is violated in Scenario A, since the scatter points are not equally distributed above and below the line at \(\epsilon=0\). Furthermore, in Scenario B we can see that the variance of residuals increase with increasing \(\hat{y}\), indicating a violation of homoscedasticity.



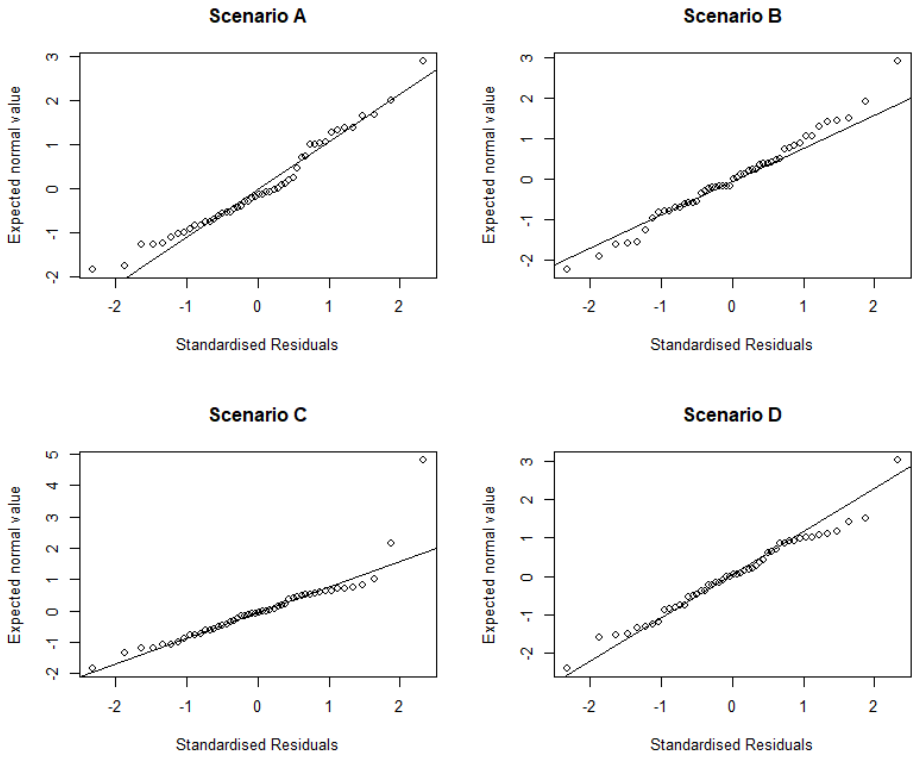
It can also be useful to plot residuals against each covariate, as a futher check for a linear relationship between \(Y\) and each of the independent variables (conditional on the other covariates in the model). If there are only a small number of covariates in the model, then these plots can be done for all variables. However, if the model is very complex, it may be judged sufficient to only plot residuals against fitted values and residuals against the most important covariates.

14.2.3 Normal plots of residuals

Normal plots (such as the **Q-Q plot**) provide the best means of visually detecting departures from normality. The normal Q-Q plot plots observed values against a standard normal distribution with the same number of points. If the data are perfectly normally distributed, the points on a Q-Q plot would lie on the line \(Y=X\). Deviations from this line indicate deviations from normality. Q-Q plots of residuals can be used to investigate the normality assumption.

As previously mentioned, the observed residuals do not have constant variance even when true residuals do. Therefore, some authors suggest using **standardised** residuals in the normal plots (since standardised residuals do have constant variance). On the other hand, some prefer to work with the observed residuals since these have the same units as the outcome. In practice, the differences between the two approaches are minor.

The figure below depicts normal Q-Q plots for the standardised residuals in Scenarios A-D. In such plots we might expect to see some deviation from the straight line in the extreme values of the residuals and so the variation in the tails are not of great concern. In Scenario A however, there is deviation away from the line \(Y=X\) towards the middle, indicating a violation of the normality assumption. Furthermore, the outlier in Scenario C may need further investigation (we discuss outliers further in a subsequent section).

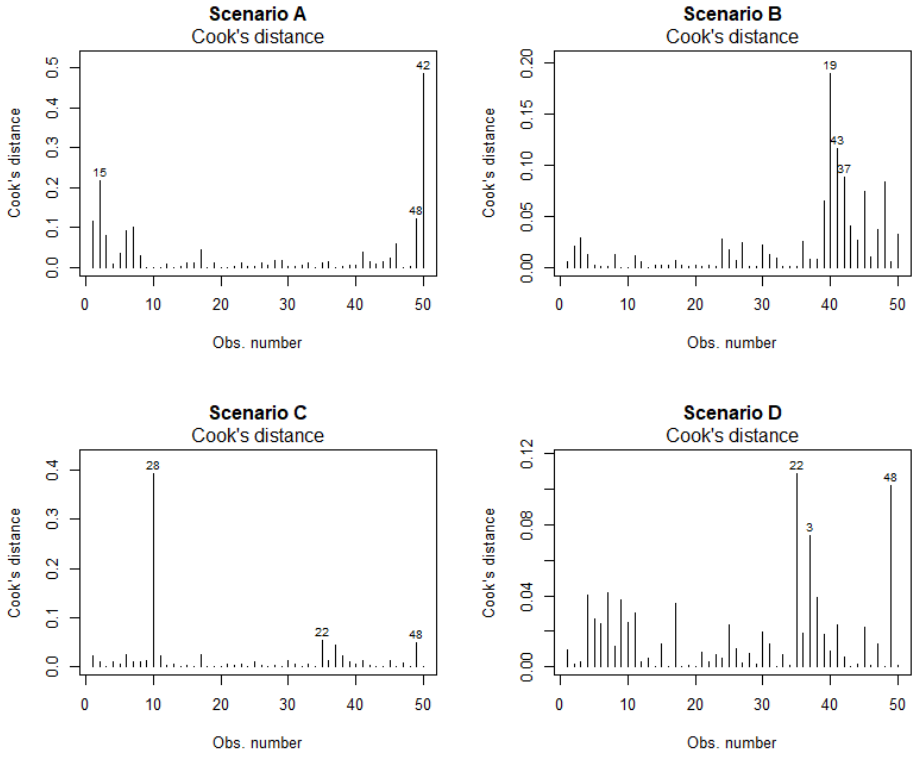


14.2.4 Plots based on Cook's distance

Cook's distance is a measure of the **influence** of an observation. An influential observation is one that has a large impact on the model parameter estimates. It is worth checking the influence of observations, particularly potential outliers, to see if they are having a much larger impact on model fit than we would expect.

For a model with p parameters (with estimated residual variance σ^2), the Cook's distance for the i^{th} observation (D_i) is obtained by refitting the model excluding this observation and obtaining new fitted values ($\hat{y_{j(i)}}$) for all n observations (including the omitted one). D_i is then defined as:
$$D_i = \frac{\sum_{i=1}^n(\hat{y}_{j(i)}-\hat{y}_i)^2}{(p+1)\hat{\sigma}^2}$$
The higher the value of D_i, the more influential the observation.

It can be informative to display Cook's distances graphically. The figure below plots the Cook's distances for each observation in Scenarios A-D.



In Scenario C the outlier identified in the previous plots and potentially problematic has a much higher Cook's distance than the other observations, indicating that it is highly influential and worth further investigation.

14.2 Statistical tests of assumptions

It might be anticipated that the assumptions of the linear regression model can be investigated using formal hypothesis tests. Indeed there exist a number of statistical tests for normalilty including the Kolmorogorov-Smirnov test and the Shapiro-Wilk test. Further, there exist statistical tests for heteroscedasticity of rediduals.

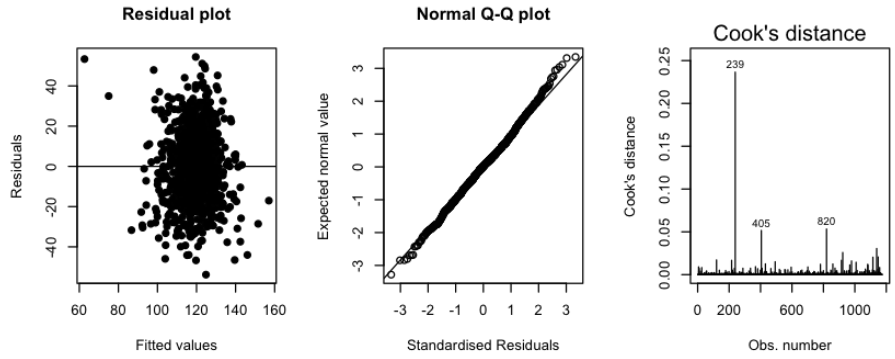
However, these tests suffer from the drawback that they tend to only have statistical power to detect model violations when datasets are large and when datasets are large the central limit theorem means that the consequences of these violations of are less important than in small datasets. With large datasets, tests of normality and heteroscedasticity can often be statistically significant, but the impact of these violations may be practically unimportant. For these reasons, the tests are considered by many statisticians to be of limited practical use and so details of these procedures will not be given here.

14.2.1 Examples using the birthweight data

We will use some of the graphical tools discussed above to assess the validity of assumptions in the multivariable model defined in the previous session (Model 4). Recall Model 4 was defined as:
$$ \text{Model 4: } y_i = \beta_0 + \beta_1 l_i + \beta_2h_i + \epsilon_i $$
The outcome y_i denotes the birthweight (in oz) for the i^{th} baby. The covariates l_i and h_i denote the length of pregnancy (i.e. number of gestational days), and the height of the mother (in inches) for the i^{th} baby, respectively.

The code below fits Model 4 to the birthweight data, and then produces (1) a plot of residuals against fitted values (2) a Q-Q plot of the standardised residuals and (3) a plot of Cook's distances by observation.

```
#Load the data
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')

#Fit Model 3 to the data
model4<-lm(Birth.Weight~Gestational.Days+Maternal.Height, data=data)

#Set the graphical space so that two plots are shown side-by-side in one row
par(mfrow=c(1,3))
options(repr.plot.height=3)

#Plot the residuals against the fitted values
plot(model4$fitted.values, model4$residuals, main = "Residual plot", xlab="Fitted
values", ylab="Residuals", pch=19)
abline(h=0)

#Obtain the standardised residuals
Standardised.Residuals<-rstandard(model4)

#Normal Q-Q plot of the standardised residuals
qqnorm(Standardised.Residuals, main="Normal Q-Q plot", ylab="Expected normal value",
xlab="Standardised Residuals")
qqline(Standardised.Residuals)

#Plot of Cook's distance
plot(model4, which=4)
```



We make the following observations:

1. **Linearity:** The residuals are equally distributed above and below zero in the "Residual plot"
2. **Normality:** There do not appear to be any serious departures from normality, based on the "Normal Q-Q plot"
3. **Homoscedasticity:** The variance of residuals are constant across the fitted values (based on the "Residual plot")

However, the Cook's distance plot reveals that observation 239 is highly influential, compared to the remaining observations. Observations 405 and 820 also have a relatively high Cook's distance. Sensitivity analyses may be required to assess model fit with and without these observations (this is discussed in Section 3.5.3).

Finally, we can assume that the independence assumption holds, since the birthweight of a baby from one mother is not expected to be associated with the birthweight of a baby from a different mother. Therefore, we can reasonably conclude that all the assumptions are met in this model (but there are some potentially problematic observations in terms of influence).

Next, we briefly introduce some of the statisical solutions available for when assumptions are not met.

14.3 Dealing with violations of assumptions

So far, we have discussed diagnostic tools that are useful for identifying possible violations of the assumptions of a linear model. Identification of potential violations of concern is only the first, and arguably the easiest, aspect of an exploration of the robustness of the results of fitting a model. Here, we briefly describe some approaches that can be used to deal with violations. When these approaches do not work, then more complex methods (beyond the scope of this lesson) may be needed to analyse the data.

14.3.1 Checking the data

Clearly, it is important that errors in data are eliminated as far as possible. In practice, ensuring that a large dataset is 100% error free may be impossible. Observations with large standardised residuals can potentially arise through data entry or coding errors and so a useful first step is to check such values with the data provider or original source of data,

if available.

14.3.2 Transformations

Sometimes it can be useful to transform either the outcome variable and/or one or more of the covariates. The transformed variables are then used in the analysis in replacement of the original variables. There are a number of possible motivations for this:

1. Transformations can be used to convert a non-linear relationship into a linear one. For example:
$$ y_i = \alpha(x_i)^{\beta} \Rightarrow log(y_i) = log (\alpha)+\beta log (x_i). $$
1. Transformations can be used to improve the normality of residuals. For example, the Box-Cox transformation is a power transformation for this purpose.
2. Transformations can help stabilise the variance of residuals. For example, if $\hat{\sigma}^2$ is proportional to $([E(Y)]^2$ then $y^*=log(y)$ is a useful variance-stabilising transformation. Alternatively, if $\hat{\sigma}^2$ is proportional to $[E(Y)]^3$ then $y^*=1/\sqrt{y}$ can be used.

14.3.3 Sensitivity analyses

If we observe potentially problematic outliers, sensitivity analyses can be used to assess how problematic they are. This involves repeating the analysis after omitting the outlier (or group of outliers) and considering the extent to which the results are altered.

However, even if the outlier affects the results (and/or assumptions) it is not a good idea to simply drop the data point. If it is not a data error, then it is a legitimate observation that should be included and understanding the reasons why it is an outlier could be important. In most cases, it is preferable to report the results including all data points, but discuss the impact removing the outlier had on the results.

14.5 Collinearity

A potential issue that can arise from including higher-order terms or interaction terms in a linear regression model is collinearity. Collinearity occurs when there is correlation between one or more of the independent covariates. If the degree of correlation between covariates is high enough, it can cause the following problems:

1. Estimated coefficients can swing in either direction: known important variables may have surprisingly small coefficient estimates and known less important variables may have surprisingly large coefficient estimates.
2. Increased variance of estimated coefficients, therefore reducing the statistical power of the model.

Including higher-order terms or interaction terms can result in collinearity due to the inevitable correlation between variables, their powers and the interaction terms involving them. Having said that, collinearity is only a concern in particular situations.

Collinearity only affects the specific independent variables that are correlated. Therefore, if the aim of the analysis is to estimate the way X influences Y after adjusting for W and V, and there is only correlation between W and V, then collinearity is not a concern. Moreover, collinearity rarely affects the predicted outcomes, so if the aim of the analysis is to predict Y using data from X, W and V, then collinearity between any of the covariates is not a concern. Finally, the severity of the problems caused by collinearity increases with the degree of correlation. Therefore, if only moderate or weak correlation is present, then collinearity is not a concern.

If we are in a situation where collinearity is causing problems, then we could either remove some of the highly correlated variables, or transform one of them. Examples of transformations include:

1. Instead of using systolic and diastolic blood pressure as collinear predictor variables, use diastolic blood pressure and (systolic-diastolic blood pressure).
2. Instead of using height and weight as predictor variables, use height and body mass index (weight/height^2).
3. When fitting a quadratic regression model, use X and $(X-\bar{X})^2$, rather than X and X^2 as covariates.

14.6 Optional Reading: Analysis of Variance

14.6.1 Partitioning variance

The total variation in Y is equal to:
$$ SS_{TOT} = \sum_{i=1}^n (Y_i-\bar{Y})^2 $$
This is often referred to as the sum of squares of the Y's. This represents all of the variation in Y about its overall mean value. We can think about two components of this total variation:

- the predictable variation in Y (predicted by the variables included in the model) and
- the unpredictable variation in Y (the remaining "noise").

The predictable variation represents the variation of the predicted values $\hat{Y}$ about the mean. We can measure this as:
$$ \begin{align} SS_{REG}= \sum_{i=1}^n (\hat{Y_i}-\bar{Y})^2 \end{align} $$
The unpredictable variation represents the variation of the observed values about their predicted values:
$$ \begin{align} SS_{RES} = \sum_{i=1}^n(Y_i-\hat{Y_i})^2 \end{align} $$
Note that in the equations above, SS stands for sums of squares, REG for regression and RES for residual.

A key result for ANOVA is that the total variation in Y can be partitioned into the predictable variation, explained by the regression model, and the unpredictable (residual) variation:
$$\begin{split} \begin{align} SS_{TOT} &= SS_{REG}+SS_{RES} \\ \rightarrow \sum_{i=1}^n (Y_i-\bar{Y})^2 &= \sum_{i=1}^n (\hat{Y_i}-\bar{Y})^2 + \sum_{i=1}^n(Y_i-\hat{Y_i})^2 \end{align} \end{split}$$

14.6.1.1 The coefficient of determination

Using the sums of squares defined above, we can calculate the proportion of variance explained by the statistical model, known as the **coefficient of determination**.

The proportion of variation which is explained by a statistical model is denoted by R^2 and is given by:
$$ R^2 = \frac{SS_{REG}}{SS_{TOT}}. $$

Example

The coefficient of determination is given in the `summary()` output for a linear regression in R. In Model 1, $R^2=0.1661$ (see output below). This means that Model 1 explains 16.6% of the total variation in Y.

```
#The coefficient of determination
data<- read.csv('https://www.inferentialthinking.com/data/baby.csv')
model1<-lm(Birth.Weight~Gestational.Days, data=data)
summary(model1)
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-49.348 -11.065   0.218  10.101  57.704

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -10.75414    8.53693   -1.26    0.208
Gestational.Days   0.46656    0.03054   15.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.74 on 1172 degrees of freedom
Multiple R-squared:  0.1661,    Adjusted R-squared:  0.1654
F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```

While R^2 is sometimes used an overall measure of goodness-of-fit (or predictive performance), it isn't used to formally compare models. This is because R^2 will never decrease when new covariates are added to a model (provided that the number and identity of observations remains the same). Therefore, using R^2 for model comparisons, we would always conclude that the more complex model is at least as good a fit as the simpler model, even if this is not true.

An adjusted R^2 has been proposed to account for this issue. Above, for example, the R^2 is 0.166 but the adjusted R-squared is a little smaller $R^2 = 0.165$. However, we would not recommend using the R^2 for formal model comparison.

14.6.2 Comparing models

When fitting statistical models, we may wish to compare how well two models fit the data, to see which is most appropriate. Consider the following three models:

- Model 1 (birthweight~length of pregnancy)
- Model 2 (birthweight~mother's smoking status)
- Model 4 (birthweight~length of pregnancy+mothers height)

We might want to make the following comparisons between models:

- Comparison 1: Model 1 vs Model 4
- Comparison 2: Model 2 vs Model 4

In these examples, Comparison 1 is much simpler than Comparison 2, because the models in Comparison 1 are **nested**.

Statistical models are said to be **nested** when one model (the simpler model) contains a subset of the covariates in the other one (the complex model) and no other additional variables. In Comparison 2, the models are not nested because the simpler model (Model 2) contains mother's smoking status as a variable, which is not included in Model 4.

Nested models can be compared using **Analysis of Variance (ANOVA)** (the comparison of non-nested models is much more complicated and is beyond the scope of this module).

The main idea of ANOVA is that: if the complex model better describes the data than the simpler model, then we would expect a reasonably large amount of the residual variation that is unexplained by the simpler model to be explained by the complex one. ANOVA provides a statistical framework that can formally test this.

We will first consider ANOVA in the context of simple linear regression, where the simpler model assumes no association between the outcome and the independent variable (the **null** model). We will then consider ANOVA in the context of multivariable linear regression and we end by learning how ANOVA can be used to test for differences between groups in a categorical variable.

14.6.3 The ANOVA table

Sums of squares (SS): The first step is to partition the total variation into the regression (predictable) and residual (unpredictable) components. Variation is measured by sums of squares (SS). So we partition the total sum of squares (SS_{TOT}) into (SS_{REG} and SS_{RES})

Degrees of freedom: Each of these sum of squares have an associated degrees of freedom (d.f.). The d.f. for the total sum of squares is $(n-1)$, since the variance of Y is $\sum_{i=1}^n (Y_i-\bar{Y})^2/(n-1)$. The d.f. for the regression sum of squares in the number of covariates in the regression model (when a simple linear regression model is used this is equal to 1). The residual d.f. is found by subtracting the regression d.f. from the total d.f.

Mean squares (MS) The sums of squares also have associated mean squares, which are obtained by dividing each sum of squares by its associated degrees of freedom (note that the residual mean square is then equal to $\hat{\sigma}^2$).

These statistics are typically summarised in an ANOVA table. The table for simple linear regression is shown below.

Source	**d.f.**	**SS**	**Mean Square**
Regression	1	SS_{REG}	$MS_{REG}=\frac{SS_{REG}}{1}$
Residual	n-2	SS_{RES}	$MS_{RES}=\frac{SS_{RES}}{n-2}$
Total	n-1	SS_{TOT}	$MS_{TOT}=\frac{SS_{TOT}}{n-1}$

Notes

> Very loosely speaking, degrees of freedom are "bits of information". We start with n bits of information. Every time we estimate something we "use" a bit of information (and so lose a degree of freedom. Therefore, when we calculate the overall variation in Y, we lose one of the n bits of information because we need to calculate the overall mean to obtain the sum of squares of Y. Therefore, we have $n-1$ bits of information overall. In a simple linear regression model, we estimate two parameters, so we're using two bits of information, but one of these is essentially the same bit we lost from calculating the overall mean, so we say that the regression model is using 1 degree of freedom. We started with n-1 df and the regression model used 1 of them so there are n-2 left for the remaining component, the residual SS.

14.6.3.1 Hypothesis testing using ANOVA

The values in the ANOVA table can be used to conduct formal hypothesis tests.

ANOVA is used to test the null hypothesis that the simpler of the two nested models better fits the data. In simple linear regression, the simpler model is the null model, in which case:

- $H_0:$ The null model is a better fit
- $H_1:$ The simple linear regression model is a better fit

To test the null hypothesis defined above, we use an F statistic, defined as:
$$ F = \frac{MS_{REG}}{MS_{RES}} $$
This ratio measures how much more variation in Y is explained by the model than would be expected by chance. If the model does not fit the data well, then we would expect this ratio to be equal to 1. The larger the value of F, the stronger the evidence that the complex model is a better fit. To obtain a p-value for a formal hypothesis test, F can be compared to the $F_{1,(n-2)}$ distribution (where 1 and (n-2) are the relevant degrees of freedom for the mean squares).

```
# F-test using anova()
anova(model1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Gestational.Days	1	65449.51	65449.5131	233.4293	3.395226e-48
Residuals	1172	328608.34	280.3825	NA	NA

A anova: 2 × 5

The F-statistic is equal to 233.4 with p-value $<2.2\times10^{-16}$. With such a small p-value there is strong evidence against the null hypothesis. Therefore we conclude that the model which includes gesational days is a better fit.

14.6.3.2 Connection between F tests and t-tests in simple linear regression

Above, we used a F-test to compare the model for birthweight (Y) including gestational days (L):
$$ y_i = \beta_0 + \beta_1 l_i + \epsilon_i $$
with a model including just a constant
$$ y_i = \alpha_0 + \epsilon_i $$
In other words, we have just used a F-test to test the null hypothesis $H_0: \beta_1 = 0$. This is exactly the same hypothesis test we tested previously using a t-test.

In other words, the F-test for a simple linear regression model is the same as the t-test of the null hypothesis that the slope parameter is equal to 0. Below, we perform the t-test again.

```
# F-test
summary(model1)
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.348 -11.065   0.218  10.101  57.704 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -10.75414    8.53693   -1.26    0.208    
Gestational.Days   0.46656    0.03054   15.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.74 on 1172 degrees of freedom
Multiple R-squared:  0.1661,    Adjusted R-squared:  0.1654 
F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```

We see that the t-statistic for the slope is $t=15.28$, with p-value $p<2e-16$. Previously, we had $F=233.4=15.24^2=t^2$. The p-value for the F-test was identical to the t-test.

The two tests are equivalent, with $F=t^2$, providing identical p-values. Consequently, it is not particularly common to use F-tests in the simple linear regression model, they are more useful for assessing more complex models with multiple covariates.

14.6.4 ANOVA for multivariable linear regression

In the context of multivariable linear regression, ANOVA can be used to test whether a more complex model is a better fit than the null model (**the Global F test**), or whether a more complex model is a better fit than a simpler model that includes a subset of the covariates in the complex model (**the partial F test**). Each test requires slight modifications to the ANOVA table defined above and we will discuss these in turn.

14.6.4.1 The Global F test

The general formulation of the ANOVA table (suitable for simple and multivariable linear regression models) is given below. p is the number of covariates in the model.

Source	**d.f.**	**SS**	**Mean Square**
Regression	p	SS_{REG}	$MS_{REG}=\frac{SS_{REG}}{p}$
Residual	$n-(p+1)$	SS_{RES}	$MS_{RES}=\frac{SS_{RES}}{n-p-1}$
Total	$n-1$	SS_{TOT}	$MS_{TOT}=\frac{SS_{TOT}}{n-1}$

Note that this is equivalent to the previous table (for simple linear regression) when $p=1$.

The Global F test tests the null hypothesis (H_0) that the null model is a better fit than the more complex model against the alternative hypothesis (H_1) that the complex model is a better fit. Or, equivalently:

- $H_0:$ All slope parameters in the complex model are equal to 0.
- $H_1:$ At least one of the slope parameters in the complex model is not equal to 0.

The appropriate F statistic is the ratio
$$ F = \frac{MS_{REG}}{MS_{RES}} $$
Under the null hypothesis, F follows an $F_{p,(n-(p+1))}$ distribution.

Example

We can use `summary()` to conduct a global F-test for Model 4, our model relating birthweight to both length of pregnancy and maternal height.

The null and alternative hypotheses, for the global F-test, are defined as:

- H_0: the regression coefficients for both gestational days and mother's height are equal to 0.
- H_1: the regression coefficient for either gestational days or mother's height (or both) is not equal to 0.

```
# ANOVA for Model 4
data$Gestational.Days.Centered<-data$Gestational.Days-mean(data$Gestational.Days)
data$Maternal.Height.Centered<-data$Maternal.Height-mean(data$Maternal.Height)

model4<-lm(Birth.Weight~Gestational.Days.Centered+Maternal.Height.Centered, data=data)
summary(model4)
```

```
Call:
lm(formula = Birth.Weight ~ Gestational.Days.Centered + Maternal.Height.Centered,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-53.829 -10.589   0.246  10.254  54.403

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              119.46252    0.47980 248.983  < 2e-16 ***
Gestational.Days.Centered   0.45237    0.03006  15.051  < 2e-16 ***
Maternal.Height.Centered    1.27598    0.19049   6.698 3.27e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.44 on 1171 degrees of freedom
Multiple R-squared:  0.1969,    Adjusted R-squared:  0.1955
F-statistic: 143.5 on 2 and 1171 DF,  p-value: < 2.2e-16
```

The F statistic is 143.5 with a p-value $<2.2 \times 10^{-16}$. Therefore, there is strong evidence against the null and we can conclude that at least one of the estimated regression coefficients is non-zero (i.e Model 4 is a better fit than the null model).

14.6.4.2 The partial F test

The global F-test is a joint test of the statistical signficance of all the slope parameters in a linear regression model. On the other hand, the partial F-test compares the fit of:

- Model A (complex model, with p predictors)
- Model B (simpler nested model with $p-k$ predictors).

The key to the partial F-test is the construction of an Analysis of Variance table that partitions the sum of squares explained by the complex model into that explained by the simple model and the extra sum of squares only explained by the complex model. Using the notation that SS_{REG_A} denotes the sum of squares explained by the complex model, whilst SS_{REG_B} denotes the sum of squares explained by the simpler model, the ANOVA table is as shown below.

Source	**d.f.**	**SS**	**Mean Square**
Explained by Model B	$p-k$	SS_{REG_B}	$MS_{REG_B}=\frac{SS_{REG_B}}{p-k}$
Explained by Model A	p	SS_{REG_A}	$MS_{REG_A}=\frac{SS_{REG_A}}{p}$
Extra explained by Model A	k	$SS_{REG_A}-SS_{REG_B}$	$MS_{REG_X}=\frac{(SS_{REG_A}-SS_{REG_B})}{k}$
Residual from Model A	$(n-(p+1))$	SS_{RES_A}	$MS_{RES}=\frac{SS_{RES_A}}{n-(p+1)}$
Total	$(n-1)$	SS_{TOT}	$MS_{TOT}=\frac{SS_{TOT}}{n-1}$

The partial F-test tests the following null hypothesis:

- $H_0:$ all of the slope parameters included in Model A but omitted from Model B are equal to zero.
- $H_1:$ at least one of the additional parameters in Model A is not equal to 0.

The appropriate test statistic (F) is the ratio of extra mean sum of squares in Model A to the mean residual sum of squares from Model A. Under the null hypothesis, this test statistic follows an F-distribution:
$$ \text{Under } H_0: F = \frac{MS_{REG_X}}{MS_{RES}} \sim F_{k,(n-(p+1))} $$
Example: We can use `anova()` to conduct a partial F-test to compare Models 1 and 3:

- $H_0:$ Model 1 is the better fit
- $H_0:$ Model 3 is the better fit

Example

We can use `anova()` to conduct a partial F-test to compare Models 1 and 4:

- Model 1 (birthweight~length of pregnancy)
- Model 4 (birthweight~length of pregnancy+mothers height)

Model 1 is nested within Model 4. Model 4 is our complex model.

In this case, the two models only differ by one variable (mother's height) and so the hypotheses being tested within the partial F-test could be written as:

- $H_0: \beta_2=0$, where β_2 is the regression coefficient for mother's height.
- $H_0: \beta_2 \neq 0$.

```
anova(model1, model4)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1172	328608.3	NA	NA	NA	NA
2	1171	316482.2	1	12126.13	44.86728	3.266475e-11

A anova: 2 × 6

The F-statistic is 44.87 with a p-value of 3.23×10^{-11}. This is strong evidence against the null hypothesis. Hence the data indicates that Model 4 (the more complex model) is the better fit.

When the two models being compared only differ by one variable, the partial F test is equivalent to the t test of the null hypothesis that the regression coefficient for that variable is equal to 0. Notice in our example that the results of the partial F test are the same as the t-test for $\beta_2=0$, with $F=t^2$.

For this reason, partial F tests are more useful in situations where we wish to compare models that differ by more than one variable. The approach is identical to that shown above.

14.6.5 ANOVA for models with categorical independent variables

Another useful application of ANOVA is to test for differences in means between categories of a categorical variable.

Suppose we are interested in the association between an outcome Y and a categorical variable X with K groups. We have already seen how to define a multivariable linear regression model using dummy variables for this situation. An alternative model, often termed the **ANOVA model**, is as follows:

Let y_{ki} be the value of the outcome for the i^{th} observation in the k^{th} group ($i=1,...,n_k$ and $k=1,...,K$). The ANOVA model is then defined as:
$$ y_{ki}=\mu_k + \epsilon_{ki} \text{, where } \epsilon_{ki} \overset{iid}{\sim} N(0,\sigma^2) $$
Here, μ_k is the mean of the outcome in the k^{th} group. With this representation, the null and alternative hypothesis are:

- $H_0: \mu_k= \mu$ (i.e. the means in all groups defined by the categorical variables are equal to a common value).
- $H_1: \mu_k \neq \mu$ (i.e. the group means are not all equal).

14.6.5.1 Sum of squares for models with categorical variables

For models with a single independent categorical variable the fitted values are simply the group means ($\bar{y_k}$). Under the null hypothesis that the group means are all equal, the fitted values are all equal to the overall mean ($\bar{y}$). This leads to new terminology for the residual sum of squares (SS_{RES}) and the sum of squares explained by the model (SS_{REG}):

- $SS_{RES} = \sum_{k=1}^K \sum_{i=1}^{n_k} (y_{ki} - \bar{y}_k)^2$
- $SS_{REG} = \sum_{k=1}^K \sum_{i=1}^{n_k} (\bar{y}_k - \bar{y})^2 = \sum_{k=1}^K n_k (\bar{y}_k - \bar{y})^2 $

In this case, the residual sum of squares is often termed the **within group sum of squares (SS_{Within})** and the regression sum of squares is often termed the **between group sum of squares $(SS_{Between})$**.

14.6.5.2 The ANOVA table

When there are K groups, the degrees of freedom for the within groups sum of squares is $n-K$ (because the model includes K parameters) and the degrees of freedom for the between groups sum of squares is $K-1$ (because the null model contains a single parameter, the overall mean). Hence the ANOVA table is as follows:

Source	d.f.	SS	Mean Square
Between groups	$K-1$	$SS_{Between}$	$MS_{Between}=\frac{SS_{Between}}{(K-1)}$
WIthin Groups	$n-K$	SS_{Within}	$MS_{Within}=\frac{SS_{RES}}{n-K}$
Total	$n-1$	SS_{TOT}	$MS_{TOT}=\frac{SS_{TOT}}{n-1}$

14.6.5.3 The F-test

To test the null hypothesis that the means in all groups are equal to a common value, the appropriate F-statistic is:
$$ F = \frac{MS_{Between}}{MS_{Within}} \sim F_{(K-1), n-K} \text{ under } H_0. $$
If this test obtains a small p-value, then we have evidence that the means in the groups are not all the same. However, it does not tell us which of the group means differed from which other group means. For this reason, if we do find evidence of difference in means on an F-test, we may want to follow up with further analysis. Such further analysis may include pair-wise comparisons of means through analysis restricted to two groups.

Example

We conduct an F-test to compare the average birthweights between babies whose mothers smoke and whose mothers don't smoke using the birthweight data.

Let μ_1 and μ_0 denote the mean birthweight for babies whose mothers do smoke and don't smoke, respectively. Then, the relevant hypotheses are:

- $H_0: \mu_1= \mu_0$ (i.e. the birthweight of a baby does not depend on whether the mother smoked)
- $H_1: \mu_1\neq \mu_0$

Recall that we previously defined Model 2 to related birthweight and mother's smoking status:
$$ y_i = \alpha_0 + \alpha_1 s_i + \epsilon_i $$
Where Y denotes the birthweight and
$$\begin{split} s_{i} = \begin{cases} 1 & \text{ if the mother smokes} \\ 0 & \text{ if the mother does not smoke} \end{cases} \end{split}$$
We can rewrite this equation using the ANOVA model as follows:
$$\begin{split} \begin{align} y_{1i} &=\mu_1 + \epsilon_{1i} \text{ if the mother smokes} \\ y_{0i} &=\mu_0 + \epsilon_{0i} \text{ if the mother does not smoke} \end{align} \end{split}$$
Where y_{ki} is the mean birthweight in the k^{th} group (groups are defined by mother's smoking status), $\mu_1 = \beta_0 + \beta_1$ and $\mu_0=\beta_0$ (in other words, our null hypothesis can be rewritten as: $\beta_1=0$).

We can use either `anova()` or `summary()` to conduct the test in R:

```
model2<-lm(Birth.Weight~factor(Maternal.Smoker), data=data)
anova(model2)
summary(model2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
factor(Maternal.Smoker)	1	24002.06	24002.0638	76.0167	9.461068e-18
Residuals	1172	370055.79	315.7473	NA	NA

A anova: 2 × 5

```
Call:
lm(formula = Birth.Weight ~ factor(Maternal.Smoker), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-68.085 -11.085   0.915  11.181  52.915

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               123.0853     0.6645 185.221   <2e-16 ***
factor(Maternal.Smoker)True  -9.2661     1.0628  -8.719   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.77 on 1172 degrees of freedom
Multiple R-squared:  0.06091,   Adjusted R-squared:  0.06011
F-statistic: 76.02 on 1 and 1172 DF,  p-value: < 2.2e-16
```

In the ANOVA table, the `factor(Maternal.Smoker)` row gives the between groups results and the `Residuals` row gives the within group results.

The F-statistic is equal to 76.02 with a p-value equal to 9.46×10^{-18}. This evidence suggests that there is a difference in the mean birthweight between the two groups defined by mother's smoking status.

14.7 Proofs

This section contains two important proofs. These are not examinable.

14.7.1 Proof for the ordinary least squares estimates in simple linear regression

Recall the ordinary least square (OLS) estimates of the intercept (β_0) and slope (β_1) in simple linear regression are:
$$\begin{split} \begin{align} \hat{\beta_0} &= \bar{y} - \hat{\beta_0}\bar{x} \\ \hat{\beta_1} &= \frac{\sum_{i=1}^2 (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x_i})^2} \end{align} \end{split}$$
Proof:

To solve for the value of β_0 that minimises SS_{RES}, we differentiate SS_{RES} with respect to $\hat{\beta}_0$ and set the derivative to zero:
$$ \frac{d(SS_{RES})}{d(\hat{\beta}_0)} = \sum_{i=1}^n -2(y_i-\hat{\beta_0}-\hat{\beta_1}x_i)=0 $$
Since $\sum_{i=1}^n (y_i)=n\bar{y}$ and $\sum_{i=1}^n(x_i)=n\bar{x}$, we can simplify to:
$$ -n\bar{y}+n\hat{\beta}_0+n\hat{\beta}_1\bar{x}=0 $$
Rearranging the above and divide by n to give:
$$ \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}. $$
To solve for the value of β_1 that minimises SS_{RES}, we have to differentiate with respect to $\hat{\beta}_1$. First, we substitute in our solution for $\hat{\beta}_0$ as follows:
$$ SS_{RES}=\sum_{i=1}^n(y_i-(\bar{y}-\hat{\beta}_1\bar{x})-\hat{\beta}_1x_i)^2=\sum_{i=1}^n ((y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x}))^2 $$
Now differentiating the above with respect to $\hat{\beta}_1$ and setting the differential to zero gives:
$$ \frac{d(SS_{RES})}{d(\hat{\beta}_1)} = \sum_{i=1}^n -2(x_i-\bar{x})(y_i-\bar{y})+2\hat{\beta}_1(x_i-\bar{x})^2=0 $$
Rearranging gives:
$$ \hat{\beta}_1\sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y}) $$
$$ \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n(x_i-\bar{x})^2} $$

14.7.2 Proof that the OLS estimates are also the maximum likelihood estimates

If $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, the log likelihood function is:
$$ l(\mu | y_1,...,y_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2 $$
So, for the simple linear regression model:

$$ l(\beta_0, \beta_1 | y_1,...,y_n) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1x_i-\beta_0)^2. $$
Therefore, for any fixed positive value for σ^2, maximising the log likelihood function is equivalent to minimising SS_{RES} and so the OLS estimates are also maximum likelihood estimates of β_0 and β_1.

15 Logistic Regression

In this session, we continue exploring regression modelling. We now extend the ideas encountered in the context of linear regression models and apply them to a setting where the outcome of interest is a binary variable.

Intended learning outcomes

By the end of this session, you will be able to:

- explain the rationale and the structure behind the logistic regression
- apply a logistic regression model to real data
- interpret the results of a logistic regression
- interpret output from key diagnosis tools used for logistic regression

15.1 Regression modelling for binary outcomes

In many health applications, the outcome of interest is a binary outcome. Examples are 30-day mortality following surgery, 5-year survival following diagnosis of breast cancer, whether or not a particular side effect was experienced after taking a particular.

In this session, we explore a very commonly used technique in health data science: logistic regression. This can be seen as an extension of the linear regression model we have been exploring in the last few sessions. We will see that many aspects of linear regression modelling extend quite naturally to logistic regression modelling.

15.1.2 Why can't we just use Linear Regression?

We developed our linear regression model for a continuous outcome. As we have seen, the linear regression model relies on a number of assumptions. Two of them, normality of residuals and homoscedasticity (constant residual variance) are particularly relevant to this discussion.

Normality of errors: The usual inference procedures (calculating confidence intervals and p-values) for linear regression assume that the errors follow a normal distribution. This assumption will be violated with binary outcome data.

Homoscedasticity: The error variance is constant. With binary data the variance is a function of the mean. Therefore, the variance would change as the mean changes. Therefore, this assumption is unlikely to hold.

An additional problem arises when we attempt to use linear regression to model binary outcome data, which relates to the fitted values.

Fitted values can be impossible: The mean of a binary outcome is the probability that it is a "success", using the terminology of Bernoulli trials. This mean, because it is a probability, must lie between 0 and 1. If we tried to use linear regression to model our binary outcome we may find that some of the fitted values are below 0 or greater than 1. This problem arises largely because of the assumption that the mean is modelled by the linear predictor, an additive function of the covariates. This assumption is unlikely to be valid.

Logistic regression avoids these problems by relating the covariates (through the linear predictor) to a *function of* the mean, instead of the mean.

15.2 Data used in our examples

We will use a dataset that is simulated to represent data from electronic health records for 200,000 patients. The outcome we will consider is whether or not a patient is diagnosed with dementia. In this example, there is an additional complexity because patients were followed up for different amounts of time. A longer follow-up will naturally lead to a higher probability of being diagnosed with dementia. In later modules, we will encounter survival analysis which allows the aspect of time to be accounted for. For now, we will ignore this aspect.

The code below reads in the dataset and displays the first few rows.

```
# we load the dataset and display its first lines
dementia <- read.csv("Practicals/Datasets/Dementia/dementia2.csv")
head(dementia)
```

	id	prac	pr_lcd	sex	age	bmi	bmi_category	consultations	agegp
	<int>	<int>	<chr>	<int>	<int>	<dbl>	<chr>	<int>	<int>
1	23189	142	08dec2009	1	53	20.4	Normal (18.5-<25)	12	50
2	92186	132	03feb2003	0	73	21.5	Normal (18.5-<25)	4	70
3	187963	43	06jul2001	0	40	27.1	Overweight (25-<30)	0	40
4	148379	215	08mar2012	1	40	20.9	Normal (18.5-<25)	3	40
5	44194	225	02feb2011	1	92	32.5	Obese class I (30-<35)	10	90
6	169915	175	02nov2011	1	55	26.3	Overweight (25-<30)	3	55

A data.frame: 6 × 30

15.2.1 Exploratory analyses

The variables we will use during this session are:

- id: a variable that identifies a patient
- sex: a factor variable that gives the sex of the patient (\(0\) for men, \(1\) for women)
- age: age in years of the patient at study baseline
- bmi: Body Mass Index of the patient at study baseline
- dementia: an indicator variable that equals \(1\) if the patient is diagnosed with dementia during follow-up, \(0\) if not.

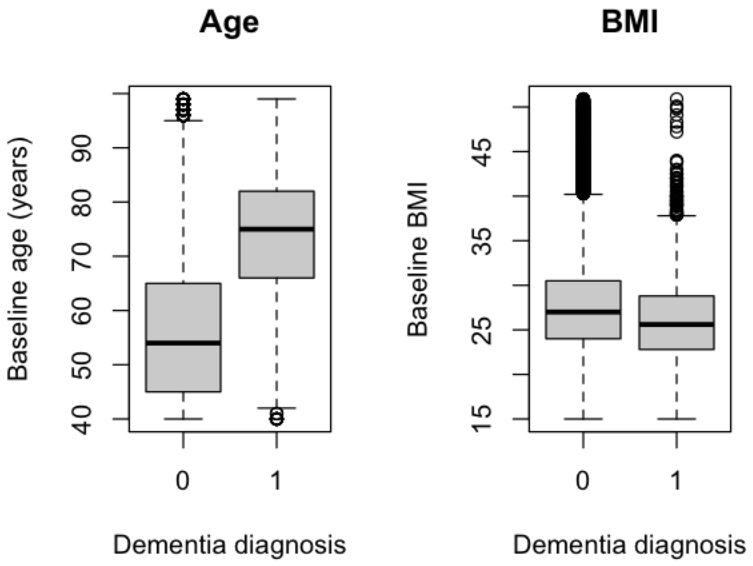
In this session the outcome of interest is dementia diagnosis, which we will treat as a binary variable. We are interested in modelling the relationship between dementia diagnosis and age, sex and BMI. Generally, we would expect older people to have a higher risk of being diagnosed with dementia. Females typically have higher risk. The relationship with BMI is less well understood.

The code below tabulates dementia and sex and draws box-plots of age and BMI, separately by dementia diagnosis status.

```
# Tabulate dementia diagnosis versus sex (dementia = right-hand column)
(table<-table(dementia$sex, dementia$dementia))
prop.table(table, 1)

# Box plot of age by dementia diagnosis
par(mfrow=c(1,2))
options(repr.plot.height=4, repr.plot.width=5)
boxplot(dementia$age ~ dementia$dementia, main="Age", xlab="Dementia diagnosis",
ylab="Baseline age (years)")
boxplot(dementia$bmi ~ dementia$dementia, main="BMI", xlab="Dementia diagnosis",
ylab="Baseline BMI")
```

```
        0      1
0 107981   1707
1  88132   2180
```

```
           0         1
0 0.98443768 0.01556232
1 0.97586146 0.02413854
```



From the output above, we see that dementia is fairly rare in this study population, with 1.6% of males receiving a dementia diagnosis during follow-up compared to a slightly higher 2.4% among females.

The box-plots show that patients who received a dementia diagnosis during follow-up generally had a much higher age at baseline, as expected. The second box-plot perhaps hints at a slightly lower BMI among those diagnosed with dementia, but there is a less evident relationship than for age.

15.3 The logistic regression model

Throughout this session we will assume that the outcome Y is binary. Further, we assume that Y takes a value of 0 ("failure") or 1 ("success"). As we discussed earlier, the terminology of success and failure does not imply success is a good thing; in health applications "success" often refers to a bad outcome such as death.

We will initially consider the simple situation with a single independent variable of interest, X. We assume that conditional on X, the outcome Y follows a Bernoulli distribution:
$$ Y | X=x \sim Bernoulli(\pi_x) $$
Then π_x is the conditional probability of sucess, given $X=x$. It also represents the conditional expectation of the outcome, given $X=x$.
$$ \pi_x = E[Y | X=x] = P(Y=1 | X=x) $$
Typically, our research question involves relating this probability to the covariate(s).

15.3.1 Components of the model

The logit function

As we have discussed, we do not wish to directly model π, because fitted values can lie outside the possible range of values. Instead, we will first transform π. In other words, we will model a function of π. We want a one-to-one function (so we can back-transform to the original scale, if we wish) that maps a probability π to the whole real line.

The function that we use in logistic regression is called the **logit function**. Specifically,
$$ logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) $$
The probability π lies in the interval $[0,1]$ but the transformed value, $logit(\pi)$ lies in the range $(-\infty, \infty)$.

It will also be useful to know how to back-transform. If $logit(\pi) = L$ then
$$ \pi = \frac{exp(L)}{1 + exp(L)} $$
This relationship will allow us to obtain fitted probabilities from our logistic regression model.

Odds

Suppose we have a binary outcome, where the probability of success is π, i.e. $P(Y=1) = \pi$. Then the **odds** of success are given by
$$ \frac{\pi}{1-\pi} $$
Therefore, $logit(\pi)$ is the logarithm of the odds, or the log-odds. We will see below that using the logit function leads to the parameters of the regression model being interpreted in terms of odds and odds ratios.

The linear predictor

Just as for linear regression models, the linear predictor is an additive function of the independent variables. With a single covariate, it is simply:
$$ \beta_0 + \beta_1 X $$

15.3.2 The basic logistic regression model

The equation for a logistic regression model with, relating X to a binary outcome Y is:
$$ logit(\pi_x) = \beta_0 + \beta_1 X $$
Note that, unlike linear regression, there is no explicit error term in the logistic regression model.

Interpreting the parameters

Suppose that our single covariate X is binary, taking values 1 (exposed, say) and 0 (unexposed). Our model is then:
$$\begin{split} logit(\pi_x) = \begin{cases} \beta_0 &\text{when X=0 (unexposed group)} \\ \beta_0 + \beta_1 &\text{when X=1 (exposed group)} \end{cases} \end{split}$$
In other words, we have:
$$\begin{split} \begin{align*} \beta_0 &\qquad \text{is the log-odds of the outcome in the unexposed group} \\ \beta_0 + \beta_1 &\qquad \text{is the log-odds of the outcome in the exposed group} \end{align*} \end{split}$$
Taking the exponential, we have
$$\begin{split} \begin{align*} e^{\beta_0} &\qquad \text{is the odds of the outcome in the unexposed group} \\ e^{\beta_0 + \beta_1} &\qquad \text{is the odds of the outcome in the exposed group} \end{align*} \end{split}$$
Now we have that $e^{\beta_0 + \beta_1} = e^{\beta_0} \times e^{\beta_1}$. Therefore, e^{β_1} also represents the multiplicative increase in the odds, going from the unexposed group to the exposed group. This multiplicative increase is known as the **odds ratio**. Therefore, we can also write:
$$\begin{split} \begin{align*} e^{\beta_0} &\qquad \text{is the odds of the outcome in the unexposed group} \\ e^{\beta_1} &\qquad \text{is the odds ratio of the outcome, comparing the exposed group to the unexposed group} \end{align*} \end{split}$$

General interpretation

This leads us to the following interpretation of the model:
$$ logit(\pi_x) = \beta_0 + \beta_1 X $$

- The intercept, β_0 is the log-odds among those with $X=0$. This is often called the **baseline log-odds**. Alternatively, the exponential e^{β_0} is the odds among those with $X=0$.
- The slope, β_1, is the difference in the log-odds associated with a one-unit increase in X. Equivalently, e^{β_1} is the odds ratio associated with a one-unit increase in X.

15.4 Estimating the parameters

Having specified our model, we now want to use a sample of data to obtain estimates of the model parameters.

15.4.1 Statistical model and observed data

Data: Suppose we have a sample of n people. Person i has an observed X value of x_i and an observed outcome y_i. Therefore, our sample of data consists of: $\{ (x_i, y_i); i=1,2,...,n\}$.

Statistical model: Our statistical model assumes that these observations are independent (between people) and are drawn from the distribution:
$$ Y_i | X=x_i \sim Bernoulli(\pi_i) $$
where
$$ \pi_i = P(Y_i=1 | X=x_i) $$
We further have a model relating the outcome to the independent variable:
$$ logit(\pi_i) = \beta_0 + \beta_1 x_i \qquad \text{or, equivalently:} \qquad \pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} $$
We could put these aspects all together to write our statistical model concisely as:
$$ Y_i | X=x_i \sim Bernoulli\left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) $$
Note

> In the previous section, we used the notation π_x in order to emphasise that this probability is conditional on the value of X. Now we are applying the distribution to a sample of people so we have changed to π_i to emphasise that the probability is conditional on whatever value X takes for person i.

15.4.2 Maximum likelihood estimation

We first need to derive the likelihood of the model. we assume that observations (people) are independent of each other, thus:
$$ \begin{align*} L(\beta_0, \beta_1) &= \prod_{i=1}^n Pr(Y_i = y_i|X_i=x_i) \end{align*} $$
We can write this as
$$ \begin{align*} L(\beta_0, \beta_1) &= \prod_{i=1}^n Pr(Y_i = 1 | X_i = x_i)^{y_i}Pr(Y_i = 0 | X_i = x_i)^{1-y_i} \end{align*} $$
Because when the observed outcome is 1, the second term above is 1 (recall $x^0=1$ for any x) so we just have $Pr(Y_i = 1|X_i=x_i)$, which is equal to $Pr(Y_i = y_i|X_i=x_i)$ when $y_i = 1$. Conversely, when $y_i = 0$, the first term becomes 1 and we are left with just the second term.

Now $Pr(Y_i = 1 | X_i = x_i)$ is just the fraction within the Bernoulli distribution above, so we can substitute this in to get:
$$\begin{split} \begin{align*} L(\beta_0, \beta_1) &= \prod_{i=1}^n \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i} \\ & = \prod_{i=1}^n (e^{\beta_0 + \beta_1 x_i})^{y_i} \times \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \end{align*} \end{split}$$
Taking the log of the above likelihood, we derive the following log-likelihood
$$\begin{split} \begin{align*} l(\beta_0, \beta_1) &= \sum_{i=1}^n \{ y_i log(e^{\beta_0 + \beta_1 x_i}) - log \left(1 + e^{\beta_0 + \beta_1 x_i} \right) \} \\ &= \sum_{i=1}^n \{ y_i (\beta_0 + \beta_1 x_i) - log \left(1 + e^{\beta_0 + \beta_1 x_i} \right) \} \end{align*} \end{split}$$
By maximizing this log-likelihood over the parameters $((\beta_0, \beta_1)$, we can obtain the maximum likelihood estimates of the parameters: $((\hat{\beta}_0, \hat{\beta}_1)$. There is no closed-form solution to this optimisation problem. Therefore, the maximisation over the parameters is done numerically.

15.5 Examples

15.5.1 Dementia and sex

We now return to the dementia dataset and explore the relationship between sex and diagnosis of dementia during the study period. In this is example, our outcome Y is the binary variable of whether the patient was diagnosed with dementia during follow-up (1=yes, 0=no). Our single independent variable S is sex (0=male, 1=female). The logistic regression model we will fit is:
$$ \mathrm{logit}(\pi_i) = \beta_0 + \beta_1 s_i $$
where $\pi_i=E(Y| S=s_i)$.

We will use the `glm()` to perform simple linear regressions in R. Click here for details of how this command works.

The following code can be used to perform this logistic regression in R. We need to specify the formula for the model, which is very similar to the syntax used in linear regression modelling. In addition, we now need to tell R that we are using the `logit` function and that we are assuming that the data are assumed to follow a Bernoulli distribution (which, recall is a special case of the Binomial distribution).

```
dementia <- read.csv("Practicals/Datasets/Dementia/dementia2.csv")
dementia1 <- glm(dementia ~ sex, data = dementia, family = binomial(link="logit"))
summary(dementia1)
```

```
Call:
glm(formula = dementia ~ sex, family = binomial(link = "logit"),
    data = dementia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2211  -0.2211  -0.1771  -0.1771   2.8855

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.14722    0.02439 -170.01   <2e-16 ***
sex          0.44771    0.03264   13.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38333  on 199999  degrees of freedom
Residual deviance: 38143  on 199998  degrees of freedom
AIC: 38147

Number of Fisher Scoring iterations: 7
```

We interpret the two estimated coefficients as follows:

- The estimated log-odds of dementia diagnosis among males (the “baseline” group, with \(S=0\)) is -4.147.
- The estimated log odds ratio for females, compared with males, is 0.4477.

For a slightly more intuitive interpretation, we will take the exponential transformation.

```
exp(coefficients(dementia1))
```

(Intercept): 0.0158083366516896 **sex:** 1.5647202094567

Now we can equivalently, and perhaps more intuitively, interpret the coefficients as follows:

- The estimated odds of dementia diagnosis among males is 0.0158.
- The estimated odds ratio for females, compared with males, is 1.576. In other words, the odds of dementia diagnosis among females is estimated to be 1.576 times higher than among males.

15.5.2 Dementia and age

We now explore the relationship of dementia diagnosis and age, measured in years. In this is example, our outcome \(Y\) remains dementia diagnosis, as above, but our single independent variable \(A\) is age, measured in years. The logistic regression model we will fit is:
\[\mathrm{logit}(\pi_i) = \beta_0 + \beta_1 a_i \]
where \(\pi_i=E(Y| A=a_i)\).

```
dementia2 <- glm(dementia ~ age, data = dementia, family = binomial(link="logit"))
summary(dementia2)
```

```
Call:
glm(formula = dementia ~ age, family = binomial(link = "logit"),
    data = dementia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9935  -0.1989  -0.1140  -0.0721   3.5947

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.533958   0.103139 -102.13   <2e-16 ***
age           0.101865   0.001402   72.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38333  on 199999  degrees of freedom
Residual deviance: 31876  on 199998  degrees of freedom
AIC: 31880

Number of Fisher Scoring iterations: 8
```

We interpret the two estimated coefficients as follows:

- The estimated log-odds of dementia diagnosis among people aged 0 is is -10.53. Of course, this is not a meaningful quantity. As for linear regression, we could center the age variable to provide an interpretable intercept.
- The estimated log odds ratio for each increase of one year in age is 0.101.

For a slightly more intuitive interpretation, we will take the exponential transformation.

```
exp(coefficients(dementia2))
```

(Intercept): 2.66170781376369e-05 **age:** 1.10723429559233

Now we can interpret the two estimated coefficients as follows:

- The estimated odds of dementia diagnosis among people aged 0 is is 2.66.
- The estimated odds ratio for each increase of one year in age is 1.107. In other words, the estimated odds of dementia diagnosis is multiplied by 1.11 (or, increased by 11%) with each increase in year of age at study baseline.

15.6 Inference

We have fitted the following logistic regression model:
$$ logit(\pi_i) = \beta_0 + \beta_1 x_i $$
Having estimated the parameters of the logistic regression model using maximum likelihood estimation, we would like to obtain 95% confidence intervals for the parameters and perform hypothesis testing. We will now explore options available to do those things.

A sketch of the relevant statistical theory is provided in the optional reading in the appendix to this session.

15.6.1 Confidence intervals

A number of approximate confidence intervals can be obtained. Two commonly used confidence intervals are the Wald-type confidence intervals and profile confidence intervals.

Wald-type confidence interval: This confidence interval takes a familiar form. For slope parameter β_1, an approximate 95% confidence inteval is given by
$$ \hat{\beta}_1 \pm 1.96 \hat{SE}(\hat{\beta}_1) $$
where $\hat{\beta}_1$ is the maximum likelihood estimate for β_1 and $\hat{SE}(\hat{\beta}_1)$ is its standard error.

Profile likelihood confidence intervals These intervals are based on the log-likelihood-ratio. For each parameter of interest, a **profile** likelihood is constructed, which treats all other parameters as nuisances and removes them from the likelihood (by setting to their values which maximise the likelihood for each value of the parameter of interest). Then confidence intervals are constructed based on the profile likelihood. The Wald-type confidence intervals provide an approximation to this process. Profile likelihod confidence intervals are provided in R using the command `confint`.

15.6.2 Hypothesis tests

Often, the hypothesis we are interested in testing is that the independent variable X is *not associated with* the outcome. Therefore, the null and alternative hypotheses are:

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

This is the null hypothesis tested, by default, in regression output provided in R.

There are three important type of tests available. These are all approximate tests and are asymptotically equivalent to one another. So in large samples, we would expect to see similar p-values from each test.

Likelihood ratio test This test is based directly on the approximate distribution of the log-likelihood-ratio.

Wald test This test is based on a quadratic approximation to the log-likelihood-ratio. As such, it can be less accurate than the likelihood ratio test, particularly if the null value is a long way from the maximum likelihood estimate. However, in this case all tests are likely to provide small p-values and similar qualitative conclusions.

The Wald test is used to obtain the p-values automatically displayed in regression output for GLMs in R and many other software platforms. This is because Wald tests are computationally less intensive than likelihood ratio tests.

Score test These tests are based on a slightly different quadratic approximation to the log-likelihood-ratio. This type of test is much less used than the other types, so we do not pursue this further here. Early tests used in epidemiology tended to be score tests, since they are less computationally intensive than the other approaches.

15.6.3 Example

We return to our model exploring the association between sex and diagnosis of dementia. We first perform a hypothesis test investigating the null hypothesis that sex is not associated with dementia diagnosis. Then we obtain 95% confidence intervals for our two parameters of interest.

```
dementia <- read.csv("Practicals/Datasets/Dementia/dementia2.csv")
dementia1 <- glm(dementia ~ sex, data = dementia, family = binomial(link="logit"))
summary(dementia1)
```

```
Call:
glm(formula = dementia ~ sex, family = binomial(link = "logit"),
    data = dementia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2211  -0.2211  -0.1771  -0.1771   2.8855

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.14722    0.02439 -170.01   <2e-16 ***
sex          0.44771    0.03264   13.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38333  on 199999  degrees of freedom
Residual deviance: 38143  on 199998  degrees of freedom
AIC: 38147

Number of Fisher Scoring iterations: 7
```

The p-value for sex is $p<0.001$, providing strong evidence against the null hypothesis that sex is not associated with the odds of being diagnosed with dementia.

Now we will obtain the profile confidence intervals for the two estimated regression coefficients:

```
confint(dementia1)
```

```
Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	-4.1954026	-4.0997726
sex	0.3838153	0.5117587

A matrix: 2 × 2 of type dbl

In fact, these are more easily interpreted on the exponentiated scale, as below.

```
cbind(exp(coefficients(dementia1)), exp(confint(dementia1)))
```

```
Waiting for profiling to be done...
```

		2.5 %	97.5 %
(Intercept)	0.01580834	0.01506468	0.01657644
sex	1.56472021	1.46787427	1.66822252

A matrix: 2 × 3 of type dbl

- The estimated odds in males is 0.0158 (95% CI 0.01506, 0.01657). We are 95% confident that thee odds of dementia diagnosis among males lies within this range.
- The estimated odds ratio for females, compared with males, is 1.56 (95% CI 1.47, 1.67). We estimate that the odds of dementia diagnosis is 1.56 times higher among females than among males. The data are consistent with this value being as low as 1.47 or as high as 1.67.

Below is the code to obtain Wald test confidence intervals. Comparing these with the (unexponentiated) confidence intervals above, we see these are very similar, as we would expect.

```
confint.default(dementia1)
```

	2.5 %	97.5 %
(Intercept)	-4.1950299	-4.0994058
sex	0.3837405	0.5116735

A matrix: 2 × 2 of type dbl

15.7 Multivariable logistic regression

Suppose we wish to relate a binary outcome (\(Y\)) to \(p\) predictor variables \((X_1, X_2, ..., X_p)\). The appropriate multivariable logistic regression model is a straightforward extension of the simple logistic regression model:
\[\mathrm{logit}(\pi_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ,..., \beta_p x_{ip} \]
where, \(x_{ji}\) is the value of the jth predictor variable for the ith participant and \(\pi_i = P(Y_i = 1 | X_1=x_1, ..., X_p=x_p)\).

The parameters in the model are interpreted as follows:

- \(\beta_0\) is the intercept. It is the estimated log-odds of \(Y\) when all the \(X_j\)'s are zero.
- \(\beta_j\) is the expected change in the log-odds of \(Y\) for a 1 unit increase in \(X_j\) *with all the other covariates held constant*.

The \(\beta_j\)'s are the **regression coefficients** (otherwise known as **partial regression coefficients**). Each one measures the effect of one covariate controlled (or adjusted) for all of the others.

The maximum likelihood estimation process outlined earlier can be naturally extended to the multivariable model above.

15.7.1 Example

We consider an example using the dementia dataset. This time, our interest lies in modeling the relationship between the odds of being diagnosed with dementia during study follow-up and to sex (\(S\)), age (\(A\)) and BMI (\(B\)) at study baseline.

Our multivariable logistic regression model is:
\[\mathrm{logit}(\pi_i) = \beta_0 + \beta_1 s_i + \beta_2 a_i + \beta_3 b_i \]
This model can be estimated in `R`using the `glm` function

```
dementia <- read.csv("Practicals/Datasets/Dementia/dementia2.csv")
dementia2 <- glm(dementia ~ sex + age + bmi, data = dementia, family =
binomial(link="logit"))
summary(dementia2)
```

```
Call:
glm(formula = dementia ~ sex + age + bmi, family = binomial(link = "logit"),
    data = dementia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1067  -0.1959  -0.1134  -0.0732   3.6917

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.783837   0.152138 -64.309  < 2e-16 ***
sex          0.306798   0.033773   9.084  < 2e-16 ***
age          0.098682   0.001413  69.826  < 2e-16 ***
bmi         -0.025619   0.003596  -7.124 1.05e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38333  on 199999  degrees of freedom
Residual deviance: 31732  on 199996  degrees of freedom
AIC: 31740

Number of Fisher Scoring iterations: 8
```

```
cbind(exp(coefficients(dementia2)), exp(confint(dementia2)))
```

```
Waiting for profiling to be done...
```

		2.5 %	97.5 %
(Intercept)	5.635516e-05	4.179134e-05	7.587428e-05
sex	1.359066e+00	1.272090e+00	1.452170e+00
age	1.103716e+00	1.100675e+00	1.106790e+00
bmi	9.747061e-01	9.678335e-01	9.815740e-01

A matrix: 4 × 3 of type dbl

We can interpret the parameters as follows:

- *sex*: Females are estimated to have 1.36 times higher odds of being diagnosed with dementia than men *who have the same age and BMI at study baseline*. The data are consistent with the true odds ratio lying between 1.27 and 1.45. The p-value, \(p<0.001\), provides strong evidence against the null hypothesis of no association between sex and dementia *after adjusting for age and BMI*.

- *age*: The odds of being diagnosed with dementia is estimated to increase 1.1-fold for each increase in year of age at study baseline. The data are consistent with the true odds ratio lying between 1.1006 and 1.107. The p-value, \(p<0.001\), provides strong evidence against the null hypothesis of no association between age and dementia *after adjusting for sex and BMI*.

- *bmi*: The odds of being diagnosed with dementia is estimated to reduce by 0.97 times for each increase in unit of BMI, suggesting an inverse association between BMI and odds of dementia diagnosis. The p-value, \(p<0.001\), provides strong evidence against the null hypothesis of no association between BMI and dementia *after adjusting for sex and age*.

15.8 Interactions and higher-order terms

We are often interested in exploring whether associations between an independent variable and our outcome differ depending on the value taken by another independent variable, i.e. we are interested in *interactions*.

Similarly, we may be modelling the relationship between the odds of the outcome and a continuous covariate, and wish to explore whether the relationshp is linear on the log-odds scale.

The good news is that the same techniques that we met in linear regression modelling can be applied here. We can add interactions, quadratic or higher order terms or splines to our logistic regression model.

15.9 Model diagnostics

This material is not examinable and is provided for your information.

Many model diagnostics are available for logistic regression models. We touch on a few very briefly here.

15.9.1 Goodness-of-fit

Deviance

The deviance of a model M is a measure of the goodness-of-fit of the model. It is defined as
$$ D = -2(l_M - l_S) $$
where l_M is the log-likelihood of model M and l_S is the log-likelihood of the saturated model (one which uses the maximum possible number of parameters without redundancies; this is the model with the best possible fit).

In general, higher values of deviance indicate worse model fit to the data. Two deviance statistics are often produced in output following logistic regression:

- Null deviance: the deviance computed for the null model, i.e. the minimal model containing only an intercept.
- Residual deviance: the deviance computed for the model that has just been estimated.

Note

> When computing deviances of different models for the same dataset, the log-likelihood of the saturated model l_S is constant. Therefore, statistical software (including the output from `glm`) often provides the deviance in a simplified form: as $-2\, l_M$.

Akaike information criterion

The Akaike information criterion (AIC) quantifies model fit as a function of the likelihood and the number of parameters being estimated. It is defined as $AIC = 2k - 2\, l(\hat{\beta})$ where k is the number of parameter in the model and $l(\hat{\beta})$ the log-likelihood of the model computed at the estimated parameter values $\hat{\beta}$.

The AIC is mainly used as a way to compare different models. The best model, in the scale of the AIC, is the one with the lowest AIC. (Note that sometimes, contrarily to the `glm` package, the AIC is computed as $AIC = -2k + 2l(\hat{\beta})$ in which case the best model would be the one with the highest AIC value.)

The AIC is actually minus the sum of the deviance and twice the number of the parameters. By including the number of parameters, the AIC penalizes models that have too many parameters, thus avoiding the selection of overfitted models.

McFadden pseudo-R^2

For the linear regression model, the coefficient of determination (R^2) measures how much variability is explained by the model.

For the logistic regression model, several generalization of the R^2 measure have been proposed. Here, we will focus on the McFadden's pseudo-R^2. The McFadden R^2 is defined as follow:
$$ R^2_{McFadden} = 1 - \frac{l_M}{l_0} $$
where l_M is the log-likelihood of the estimated model and l_0 is the log-likelihood of the null model (containing an intercept only).

The rationale behind this measure is that when the estimated model does not explain correctly the variability, its log-likelihood will be close to the null log-likelihood so that the ratio will be close to 1 and the McFadden's pseudo-R^2 close to 0. Conversely, when the model correctly explains the variability of the model, the likelihood will be close to 1 and therefore l_M will be close to 0 so that the McFadden's pseudo-R^2 will be close to 1. However, when applied to a classic linear regression model, the McFadden's pseudo-R^2 is not equivalent to the classic R_2.

The Hosmer-Lemeshow test

The Hosmer-Lemeshow test is a classic approach to assess the goodness-of-fit of a logistic regression model. The rationale of this test is to divide the vector of predicted probabilites $\hat{\pi} = (\hat{\pi}_i)$ with $i=1,\dots,n$ into G groups, e.g. based on the quantiles, with n_g subjects. In each group, the mean of the predicted probabilites

($\bar{\pi}_g$) is compared to the proportion of observed success. Formally, for the group $g=1,\dots,G$, we have that

- the observed values are
  - for Y = 1: y_g
  - for Y = 0: $n_g - y_g$
- the predicted values are
  - for Y = 1: $\bar{\pi}_gn_g$
  - for Y = 0: $n_g(1 - \bar{\pi}_g)$

The Hosmer-Lemeshow test statistics is based on the chi-square statistics computed over all groups and all possible values for Y
$$\sum_{g=1}^G\sum_{l=0}^1 \frac{(o_{gl} - e_{gl})^2}{e_{gl}} = \sum_{g=1}^G \frac{(n_g\bar{\pi}_g - y_g)^2}{n_g\bar{\pi}_g(1-\bar{\pi}_g)}$$
and has been shown to follow asymptotically a χ^2 distribution with $g-2$ degrees of freedom under the null hypothesis of a correctly specified model. However, we insist on the fact that this test if often criticized for several reasons. First it is known to have low power. Secondly, its results can be sensible to the choice of the number of groups G and this is even worst for small sample sizes.

The Hesmer-Lemeshow test statistics has not been implemented into the `glm` package but is available on the `ResourceSelection` package through the `hoslem.test` function.

15.12 Common pitfalls

15.12.1 Perfect separation

Perfect separation happens when the outcome can be directly predicted from one of the predictor variables. For example, let say that we model an outcome Y using one explanatory standard gaussian variable X_1 and that Y is such that $Y=0$ whenever $X_1\leq0$ and $Y=1$ whenever $X_1>0$.

```
x1 <- rnorm(1000, 0, 1)
y <- (x1 <= 0)*1

data_sep <- data.frame(y,x1)
```

Let us try to estimate this logistic regression model

```
model_sep <- glm(y ~ x1, data = data_sep, family = binomial(link="logit"))
```

```
Warning message:
“glm.fit: algorithm did not converge”
```

```
Warning message:
“glm.fit: fitted probabilities numerically 0 or 1 occurred”
```

`R` detects the perfect separation and prompts an error that states that `fitted probabilities numerically 0 or 1 occurred`. The reason of this error is that, due to the perfect separation, the maximum likelihood of the parameter β_1 for the variable X_1 cannot be estimated as its value is actually infinite. Options to consider when facing this issue include:

- removing the problematic variable from the model
- setting β_1 at an arbitrary high value and estimate the model
- changing the model or manipulating the data

Note that, in practice, perfect separation is not very likely to happen. However, *quasi_perfect* separation is totally possible and needs to be tackled. For more details about how to handle separation, one can read the following articles:

> *Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine*
>
> *Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. Biometrika*

15.12.2 Low events per variable

A common issue when estimating logistic regression model is the problem of the ratio between the number of events and the number of predictive variables. This ratio is known as *Events Per Variable*. When this ratio is low, it can lead to biased estimation and models with poor predictive abilities.

In the biomedical literature, the so-called *ten events per variable rule* is commonly used. However, we emphasize here the absence of theoretical justification and even the lack of actual evidence that this rule gives good results. If you want more information about the issues raised by this commonly used rule, you can read the following article:

> *Smeden, M., de Groot, J.A., Moons, K.G. et al. (2016) No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol.*

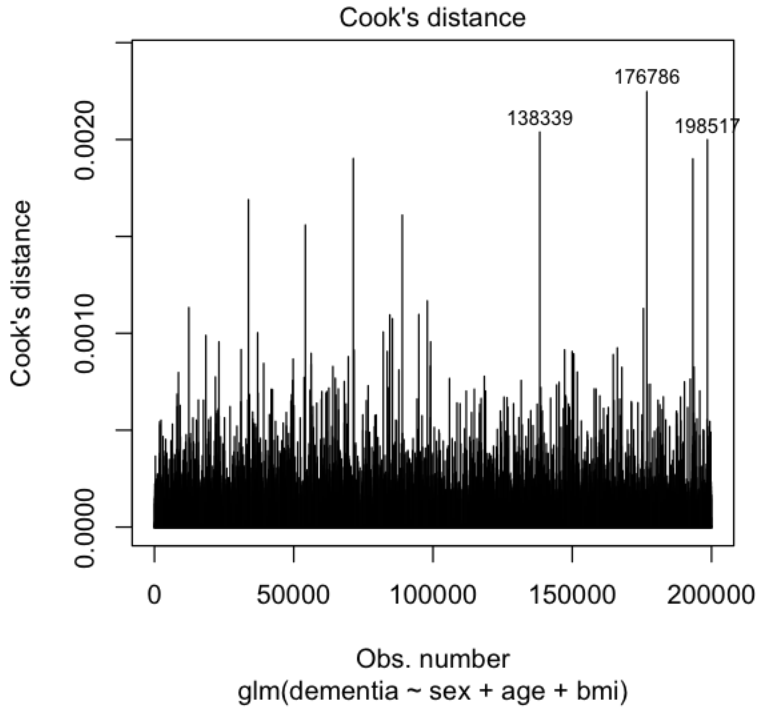
15.12.3 Influential values

Another aspect to take into account when estimating a logistic regression model is the presence of influential values among the observations which, as their names indicates, might have a huge impact on the estimation of the model. The Cook's distance is a useful measure to assess how influential an observation is. It measures how much the outcome would be modifier by removing this observation from the data.

In, `R`, the Cook's distance can be easily plotted and directly plotted by specifying `which = 4` as an argument to the `plot` function.

```
dementia <- read.csv("Practicals/Datasets/Dementia/dementia2.csv")
dementia2 <- glm(dementia ~ sex + age + bmi, data = dementia, family =
binomial(link="logit"))
summary(dementia2)
options(repr.plot.height=5, repr.plot.width=5)
plot(dementia2, which = 4)
```

```
Call:
glm(formula = dementia ~ sex + age + bmi, family = binomial(link = "logit"),
    data = dementia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1067  -0.1959  -0.1134  -0.0732   3.6917

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.783837   0.152138 -64.309  < 2e-16 ***
sex          0.306798   0.033773   9.084  < 2e-16 ***
age          0.098682   0.001413  69.826  < 2e-16 ***
bmi         -0.025619   0.003596  -7.124 1.05e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38333  on 199999  degrees of freedom
Residual deviance: 31732  on 199996  degrees of freedom
AIC: 31740

Number of Fisher Scoring iterations: 8
```



As you can see from the above example, some observations seems to have higher influence than the other. However, if we look at the y-axis scale, this difference is not huge.

If some observations appear to have a lot of influence on the estimated regression coefficients, it is important to assess the robustness of your conclusions to these observations. This is typically done using *sensitivity analysis*, i.e. performing the analysis including and excluding the problematic observations.

Note that we would not recommend excluding observations from an analysis entirely just because they are influential or outlying.

15.13 Further resources

Note: resources below are for you to deepen your understanding of the subject if you wish to do so. This is entirely optional. The extensions below (conditional logistic regression, multinomial regression, ordinal logistic regression and neural networks) are not examinable. These comments are provided for your interest.

The following book contains a much more detailed presentation of logistic regression. Chapter 7, 8 and 11 are particularly useful.

> Agresti, A., Categorical Data Analysis, 3rd edition, 2012

The following also provide more detail on the subjects we have covered:

> Hosmer D. W., Lemeshow S. and Rodney X. S. Applied logistic regression. Wiley, 2013.

> *Hosmer, D. W., Lemeshow, S. (1980) Goodness of fit tests for the multiple logistic regression model. Communications in Statistics – Theory and Methods*

15.13.1 Conditional logistic regression

Conditional logistic regression is specifically designed for grouped data and in particular for pair-matched studies where each group is a pair of two matched subjects.

To analyse a pair-matched study, a naïve idea would be to include into a logistic regression model a parameter specific to each data stratum. However, for matched studies, it would mean having as many parameters as pairs. This raised two main issues. First, each of this stratum specific parameter would be estimated only from the information contained in a unique pair, i.e. very little information. Secondly, as the number of pairs increases, the number of parameters would also increase and maximum likelihood theory would fail to provide valid estimation.

The idea of conditional logistic regression is to remove the dependence upon the stratum specific parameters by conditioning the probability on sufficient exposure information. This way, because the stratum specific parameters vanish from the equation, there is no violation of the assumptions underlying maximum likelihood theory and the other model parameters can be estimated consistently using classic techniques.

We note that in some cases (e.g. where the matching was on measurable characteristics, such as age and sex only), standard logistic regression adjusting for the matching variables is valid.

15.13.2 Multinomial logistic regression

The multinomial logistic regression model is a generalization of the logistic regression model for outcomes that have more than 2 categories, $Y\in\{1,\dots,J\}$ with $J\geq 2$ a natural number. In this case, the conditional distribution of Y_i given the covariates X_i is the multinomial distribution. Among the J categories, a reference one is chosen, e.g. the first category, and for $j=2,\dots,J$
$$\log\left(\frac{P(Y_i=j|X_i)}{P(Y_i=1|X_i)} \right) = \beta_{0,j} + \sum_{k=1}^p \beta_{k,j}X_{i,k}$$
The model is estimated simultaneously for all values of j. For a fixed j, the interpretation of the parameters is similar to the logistic regression model.

15.13.3 Ordinal logistic regression

The ordinal logistic regression is designed for outcomes that have more than 2 categories, $Y\in\{1,\dots,J\}$ with $J\geq 2$, and whose categories have an explicit ordering. In the ordinal logistic regression, the modelled quantity is $\mathrm{logit}(P(Y_i\geq j|X_i))$ for $j\geq 2$. Indeed, when $j=1$, t $P(Y\geq 1)=1$ and does not need to be modelled. As the categories are ordered, a fundamental assumption made by ordinal logistic regression is that the effect of the covariates are homogenous between the different categories. Therefore, the model is written
$$\mathrm{logit}(P(Y_i\geq j|X_i)) = \beta_{0j} + \sum_{k=1}^n \beta_{k}X_{i,k}$$
where only the intercept term depends upon the category. However, it is important to carefully check for violations of this assumption.

15.13.4 Neural networks

Artificial neural networks are a class of model widely used for data classification in machine learning. Artificial neural networks are at the heart of *deep learning* methods used to develop computer vision, speech recognition, audio recognition, etc. Actually, the basic logistic regression model happens to be a special case of artificial neural network. If you are interested in this subject and want to have more insight on the relation between these two models, you might want to read the following article:

> Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics

15.14 Additional reading

The following notes sketch out the statistical theory underlying confidence interval construction and hypothesis testing using the log-likelihood ratio. This material is not examinable.

These notes are intended to draw connections between the previous material surrounding maximum likelihood and the material concerning frequentist inference.

15.14.1 Likelihood for simple logistic regression

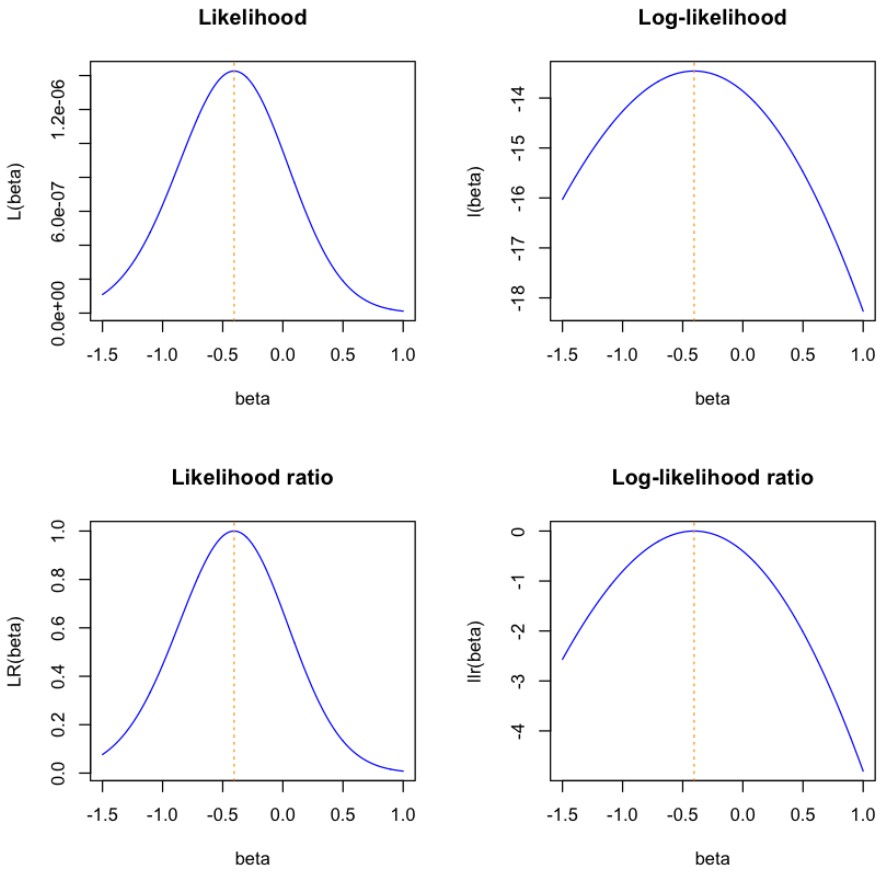
Data: We have a sample of n binary observations $y_1, ..., y_n$. We will consider a very simple situation, with no covariates of interest.

Statistical model: We assume our sample arises from n independent variables $Y_1, ..., Y_n$, with $Y_i \sim Bernoulli(\pi)$. Our logistic regression model is:
$$ logit(\pi) = \beta $$
Likelihood: Following the notes in the main text, we can obtain the likelihood function:
$$ \begin{align*} L(\beta) = e^{k \beta} \times \left(\frac{1}{1 + e^{\beta}} \right)^n \end{align*} $$
where $k = \sum_i y_i$.

Log-likelihood: Taking the log of the above likelihood, we derive the following log-likelihood
$$ \begin{align*} l(\beta) = k \beta - n log \left(1 + e^{\beta} \right) \end{align*} $$
Likelihood ratio: This is the likelihood function divided through by the likelihood function evaluated at its maximum point (i.e. at the maximum likelihood estimator, $\hat{\beta}$). Therefore, this is simply the likelihood scaled to have a maximum of 1:
$$ \begin{align*} LR = \frac{L(\beta)}{L(\hat{\beta})} = \frac{e^{k \beta} \left/ \left(1 + e^{\beta} \right. \right)^{n}}{e^{k \hat{\beta}} \left/ \left(1 + e^{\hat{\beta}} \right. \right)^{n}} \end{align*} $$
Log likelihood ratio: The log of the likelihood ratio is:
$$ \begin{align*} llr(\beta) = k \beta - n log \left(1 + e^{\beta} \right) - \left\{ k \hat{\beta} - n log \left(1 + e^{\hat{\beta}} \right) \right\} \end{align*} $$
Maximum likelihood estimate: We take the derivative of the log-likelihood and evaluate it at zero to obtain the maximum likelihood estimator, $\hat{\beta}$:
$$ \begin{align*} \frac{d l(\beta)}{d \beta} = \frac{d}{d \beta} \left\{ k \beta - n log \left(1 + e^{\beta} \right) \right\} = k - n \frac{e^{\beta}}{\left(1 + e^{\beta} \right)} \end{align*} $$
Setting this to zero gives
$$ \begin{align*} \hat{\beta} = log\left(\frac{\bar{y}}{1 - \bar{y}} \right) \end{align*} $$
where $\bar{y} = k/n$ is the sample proportion of successes.

The figure below shows the Likelihood function and the log-likelihood at the top and the likelihood ratio and log-likelihood ratio on the bottom. We see that all four have a maximum at the same value of β. The two graphs on the log scale (on the right hand side) are flatter and more symmetric. The likelihood ratio is simply the likelihood but scaled so the maximum value is 1. The log-likelihood ratio is scaled so the maximum value is 0.

The code is suppressed to focus on the output but you can click to see the code.

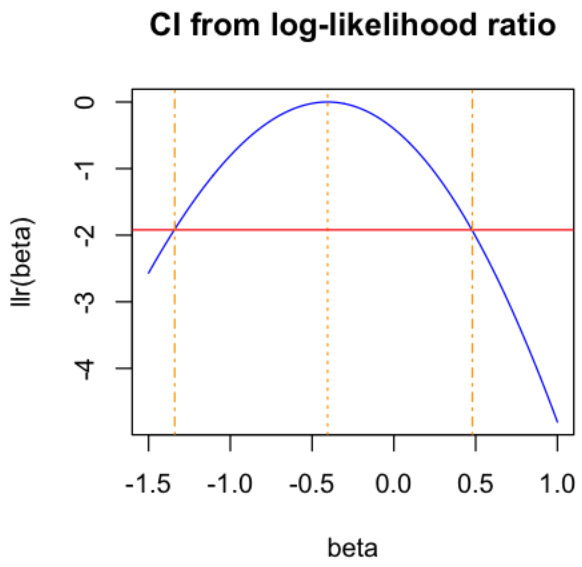


15.14.2 Confidence intervals based on the likelihood

Remember that the likelihood is a measure of how consistent the different values of the parameter are with the observed data. The most consistent value is at the maximum, i.e. the maximum likelihood estimator. We can also see that values with a much lower likelihood are much less consistent with the data.

This suggests the idea of obtaining a confidence interval by taking all values that have a likelihood within a certain range of the maximum.

In fact, when we have a single parameter of interest (which we will call β_0) then it turns out that for an independent sample (under a number of "regularity" conditions not stated here), we have the following asymptotic distribution:
$$ -2 llr(\beta_0) = -2 (l(\beta_0) - l(\hat{\beta})) \sim \chi^2_1 \qquad \text{as} \ n \rightarrow \infty $$
A χ^2_1 distribution has 5% of the distribution above the value 3.84. Therefore, this means that
$$ P(-2 llr(\beta_0) \geq 3.84) = 0.95 \rightarrow P(llr(\beta_0) \geq -1.92) = 0.95 $$
leading to the 95% confidence interval of all values of β that have a log-likelihood ratio at most 1.92 units lower than the maximum:
$$ \{ \beta \text{ s.t. } l(\beta) - l(\hat{\beta}) \geq -1.92\} $$
The plot below shows the line -1.92. Our 95% confidence interval is formed by all values of β which have a log-likelihood falling above this line. This is approximately the iterval: (-1.34, 0.48). The MLE and confidnence limits are shown in orange dashed lines.



15.14.2 Quadratic approximation

There is often no closed form solution to obtain the exact values at which the log-likelihood ratio takes value -1.92. An often simpler approach is to work instead with a quadratic approximation to the log-likelihood ratio, for which there is a simple closed-form solution.

We will now make a quadratic approximation to the log likelihood ratio. In the plot above, we see that this graph is not quite symmetric but looks fairly quadratic near the maximum.

To obtain our quadratic approximation, we will look for a function of the (quadratic) form:
$$ f(\beta) = -\frac{1}{2} \left(\frac{\beta - M}{S} \right)^2 $$
We want our quadratic approximation to

- have the same maximum
- have the same curvature near the maximum

The first condition above means that we need $f(\hat{\beta}) = 0$. This fixes $M = \hat{\beta}$.

The second condition means that we need the second derivatives of $f(\beta)$ and $llr(\beta)$ to be equal at the MLE, $\hat{\beta}$, since curvature is measured by the second derivative. In fact, we will consider making the curvature (second derivatives) of $f(\beta)$ and $l(\beta)$ to be equal at the MLE, since this is algebraically a little simpler. From the plots above we can see that the curvature of $l(\beta)$ and $llr(\beta)$ are identical.

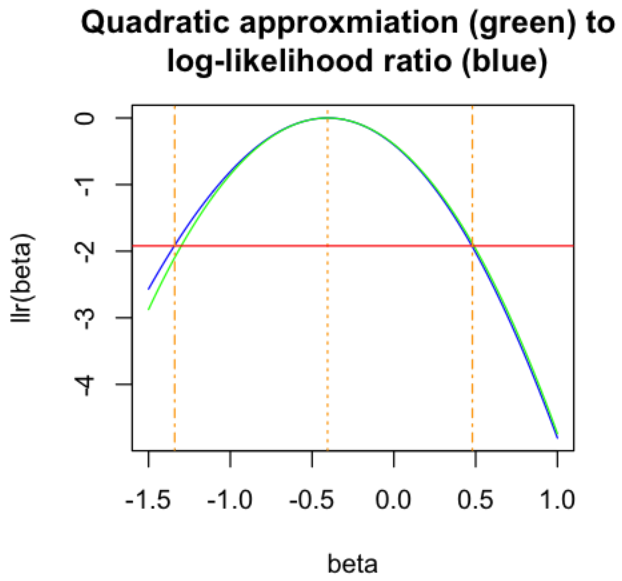
Differentiating $f(\beta)$ twice shows that $f''(\beta) = -1/S^2$ for any value of β. Thus we set
$$ S^2 = -\frac{1}{l''(\hat{\beta})} $$
It also turns out that the resulting value for S is also the standard error of $\hat{\beta}$, i.e. $S = SE(\hat{\beta})$.

This gives us our required quadratic approximation to the log-likelihood ratio:
$$ f(\beta) = -\frac{1}{2} \left(\frac{\beta - \hat{\beta}}{SE(\hat{\beta})} \right)^2 \qquad \text{with SE obtained as: } SE^2 = -\frac{1}{l''(\hat{\beta})} $$

Example

Returning to our example, we take the second derivative of the log-likelihood to obtain S. First, we have already taken the first derivative to obtain our MLE:
$$ \begin{align*} \frac{d l(\beta)}{d \beta} = \frac{d}{d \beta} \left\{ k \beta - n log \left(1 + e^{\beta} \right) \right\} = k - n \frac{e^{\beta}}{\left(1 + e^{\beta} \right)} \end{align*} $$
Taking the derivative of this, we get:
$$ \begin{align*} \frac{d^2 l(\beta)}{d \beta^2} = \frac{d}{d \beta} \left\{ k - n \frac{e^{\beta}}{\left(1 + e^{\beta} \right)} \right\} = -\frac{n e^\beta}{\left(1 + e^{\beta} \right)^2} \end{align*} $$
Thus we set
$$ S^2 = -\frac{1}{l''(\hat{\beta})} = \frac{\left(1 + e^{\hat{\beta}} \right)^2}{n e^\hat{\beta}} $$

The figure below shows the log likelihood ratio and the quadratic approximation.

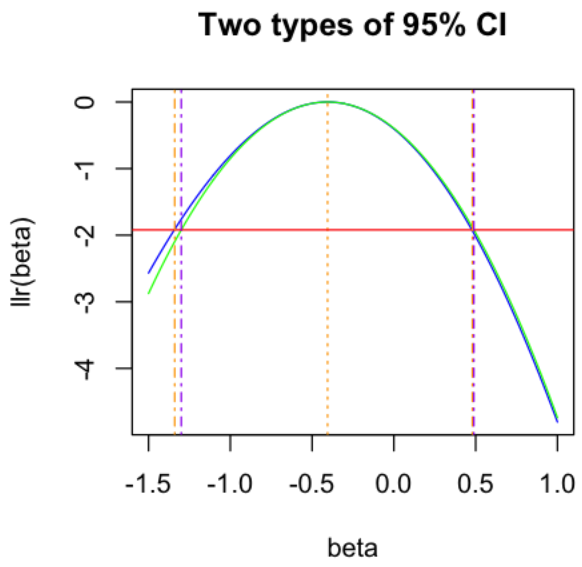


The quadratic approximation is very good near to the maximum. The horizontal red line indicates the 95% confidence interval obtained using the log likelihood ratio. The quadratic approximation starts to deviate from the log likelihood ratio at that point, but not by much. overall, this plot suggests that the quadratic approximation will provide us with a 95% confidence interval very close to the one obtained directly from the log likelihood ratio.

15.14.3 Quadratic approximation

If our quadratic approximation is a good approximation to the log-likelihood ratio then we will have, approximately
$$ -2 f(\beta) \sim \chi^2_1 $$
Thus
$$ -2 f(\beta) = -2 \times -\frac{1}{2} \left(\frac{\beta - \hat{\beta}}{SE(\hat{\beta})} \right)^2 \sim \chi^2_1 $$
In other words,
$$ \left(\frac{\beta - \hat{\beta}}{SE(\hat{\beta})} \right)^2 \sim \chi^2_1 $$
A χ^2_1 distribution has 5% of the distribution above the value 3.84, so
$$ P\left(\left(\frac{\beta - \hat{\beta}}{SE(\hat{\beta})} \right)^2 \geq 3.84\right) = 0.95 $$
Noticing that $\sqrt{3.84} = 1.96$, this gives the 95% confidence interval:
$$ \hat{\beta} \pm 1.96 \times SE(\hat{\beta}) $$
This is sometimes called a **Wald-type** confidence interval.

The plot below graphs the interval constructed in this way (indicated by the purple dashed lines), along with the previous confidence interval calculated from the log-likelihood ratio.



In the plot above, the two confidence intervals are very similar.

15.14.4 Hypothesis testing

Suppose we wish to test the null hypothesis:

- $H_0: \beta = \beta_0$
- $H_1: \beta \neq \beta_0$

Likelihood ratio test

Under the null hypothesis,
$$ -2 llr(\beta_0) \sim \chi^2_1 $$
Tests can be based on this distribution, giving a p-value.

Wald test

Using the quadratic approximation,
$$ \left(\frac{\beta - \hat{\beta}}{SE(\hat{\beta})} \right)^2 \sim \chi^2_1 $$
Or, equivalently:
$$ \left(\frac{\beta - \hat{\beta}}{SE(\hat{\beta})} \right) \sim N(0,1) $$
This is exactly the form of hypothesis tests we encountered in the session about hypothesis testing.

15.14.5 Additional comments

Notes

> We often construct confidence intervals on the log scale (e.g. the log odds ratios). Confidence intervals based on the log likelihood ratio are transformation invariant, but Wald-type intervals are not. Often, basing calculations on a log scale improves the approximations made above.
>
> We have focused on situations with a single unknown parameter. With more than one unknown parameter, things are a little more complex. The profile likelihood, which treats some parameters as "nuisance" parameters and *removes* them from the likelihood, using a process called profiling, is beyond the scope of these notes. The fundamental principles remain the same.

16. Generalised Linear Models (GLMs)

Intended learning outcomes
By the end of this session, you will be able to:

- explain what Generalised Linear Model is;
- describe the role of a link function;
- apply a GLM to Poisson distributed data and evaluate the findings;
- demonstrate how to explore the goodness of fit.

16.1 Introduction to Generalised Linear Models (GLMs)

The term Generalised Linear Model (GLM) refers to a large class of models popularised by McCullagh and Nelder in 1982. It should not be confused with the similarly named method, General Linear Model (which was covered in sessions 12 to 14).

GLMs can be seen as a extension to the familiar regression models you have already been introduced to in previous chapters. GLM allows for the outcome variable to have an error distribution other than the normal distribution. The name comes from the method which generalises linear regression by allowing the linear model to be related to the outcome variable via something called a link function. This means that GLMs can model outcomes with distributions in the exponential family with a link function which varies linearity with the predictors (covariates) rather than assuming the outcome itself must vary linearly.

16.2 Generalised Linear Model Components

A generalised linear model consists of three components:

A random component

This refers to the probability distribution of the outcome variable Y_{i} (for the ith of n independently sampled observations). It specifies the conditional distribution of the outcome given the values of the predictors (covariates) in the model. Y_{i} is generally formulated as distribution from the exponential family, however subsequent work has extended GLMs to multivariate exponential families, to certain non-exponential families and to also to situations where the distribution of Y_{i} is not completely specified. Within this chapter we will only explore the application to distributions from the exponential family (i.e. Normal, Gamma, Poisson, Bernoulli etc.)

A systematic component (the linear predictor)

This is the linear function of the predictors (covariates) in the model
$$ \eta_{i} = \beta_{0} + \beta_{1}X_{i1} + \beta_{2}X_{i2} + ... + \beta_{k}X_{ik} $$
Just as in linear and logistic regression models, the predictors (covariates) X_{ij} may be continuous and/or categorical.

A link function

This function transforms the expectation of the predictors (covariates) to be linear with the outcome variable.

Suppose we let $\mu_{i} = E[Y_{i}]$. Then the *link function* is a function $g(.)$ with
$$ g(\mu_{i}) = \eta_{i} $$
Or in other words,
$$ g(\mu_{i}) = \beta_{0} + \beta_{1}X_{i1} + \beta_{2}X_{i2} + ... + \beta_{k}X_{ik} $$
This can be inverted so
$$ \mu_{i} = g^{-1}(\eta_{i}) = g^{-1}(\beta_{0} + \beta_{1}X_{i1} + \beta_{2}X_{i2} + ... + \beta_{k}X_{ik}). $$
The inverse link $g^{-1}(.)$ is often called the *mean function* as it gives the expected value of the outcome.

16.2.1 An example - logistic regression

Suppose we wish to fit a logistic regression model (which we will see is one particular type of GLM) for a binary outcome Y and a single covariate X. Normally we write π_i for the expected outcome but here we will use μ_i instead, so you can see how the model we had previously connects with the more general notation above. Thus, we let $\mu_i = E[Y_i]$ (=π_i in previous sessions). We have:
$$ Y_i \overset{iid}{\sim} Bernouilli(\mu_i), \qquad \text{for} \ i=1, 2, ..., n $$
The linear predictor is given by:
$$ \eta_i = \beta_0 + \beta_1 X_i $$

The link function can be defined generically. In the equation below, z has no intrinsic meaning; it is just used here to enable us to define a function. The link function for logistic regression is the logit function:
$$ g(z) = \log \left\{ \frac{z}{1-z} \right\} $$
Setting this equal to η_i, as per the definition above, we get:
$$ g(\mu_i) = \log \left\{ \frac{\mu_i}{1-\mu_i} \right\} = \beta_0 + \beta_1 X_i $$
which is the logistic regression model we met previously.

16.3 GLM Assumptions

To successful apply a GLM, a number of assumptions about the data must be met.

1. The data must be independently distributed.
2. The outcome variable Y_{i} does not have to be normally distributed but should typically form a distribution from the exponential family
3. GLM's must assume a linear relationship between the transformed outcome in terms of the link function and the predictor (covariate) variables.
4. The homogeneity of variance does not need to be satisfied. Generally the model structure, and overdispersion (when the observed variance is larger than what the model assumes) can be present.
5. Errors need to be independent but not normally distributed.
6. GLM's use maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

16.4 Link Functions

The link function provides the relationship between the systematic component and the mean of the distribution. There are many commonly used link functions, the table below lists only three examples with their distributions and mean functions. Here we use matrix notation where $\mu_{i} = \beta_{0} + \beta_{1}X_{i1} + \beta_{2}X_{i2} + ... + \beta_{k}X_{ik}$ is represented by $\mathbf{X}\mathbf{\beta}$.

Distribution	Data	Link Name	Link function	Mean function
Normal	real: (-∞ , + ∞)	Identity	$\mathbf{X}\mathbf{\beta} =\mu$	$ \mu = \mathbf{X}\mathbf{\beta} $
Poisson	integer: 0,1,2,...	Log	$\mathbf{X}\mathbf{\beta} =\ln(\mu)$	$ \mu = exp(\mathbf{X}\mathbf{\beta})$
Binomial	integer: 0,1,2,... N	Logit	$\mathbf{X}\mathbf{\beta}=\ln(\frac{\mu}{n-\mu})$	$ \mu =\frac{exp(\mathbf{X}\mathbf{\beta})}{1 + exp(\mathbf{X}\mathbf{\beta}} $
Gamma	real: (0, + ∞)	Negative Inverse	$\mathbf{X}\mathbf{\beta} = -\mu^{-1}$	$ \mu = - (\mathbf{X}\mathbf{\beta})^{-1}$

It is important to note that both linear regression which is covered in sessions 12 to 14 and logistic regression in sessions 15 can be reproduced through a GLM.

Recall that a linear regression assumes data is normal distributed so using the identity link function for a normal distribution within the GLM framework will give the same estimated regression coefficients. However, the inference (p-values and confidence intervals) is slightly better using ordinary least squares compared to than maximum likelihood estimation thus we prefer to fit linear regression models using OLS.

In logistic regression if you use the logit function for a binomial family (recalling that Bernoulli is a special type of binomial distribution) you will be able to reproduce the same results as obtained through standard logistic regression modelling. For binary outcomes, the GLM has the extra flexibility compared to the logistic regression model. You can also use other link functions, for example the Probit, the Log-Log and the Complementary log-log functions. These will give similar results but adjust for slight differences from data collection situations to improve the transformation of the expectation of the outcome to the systematic component. In this module we only focus on the logit link, however if you wish to explore further, more information can be found here: https://aip.scitation.org/doi/pdf/10.1063/1.5139815

16.5 Programming GLM's in R

To fit a GLM in R you will need to use the glm() function where we tell R what the distribution of the errors and linear predictor should be.

The function syntax is as followed:

```
glm(formula, family = gaussian, data, weights, subset,
na.action, start = NULL, etastart, mustart, offset,
control = list(…), model = TRUE, method = "glm.fit",
x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, …)
```

The minimal inputted parameters required is the *formula* and the *family*. The *formula* tells R in a symbolic description what model is required to be fitted and the *family* describes what error distribution and link function are required.

16.6 Introduction to Poisson Generalised Linear Modelling (Poisson Regression)

In the previous session we met an important type of GLM - logistic regression. Now we will now consider another important GLM - Poisson regression.

16.6.1 Poisson Distribution Recap

The Poisson distribution was first published by Siméon Denis Poisson in 1838. Poisson was a French mathematician, engineer, and physicist, his name is one of 72 engraved on the Eiffel Tower in Paris. The Poisson distribution is a skewed, discrete distribution restricted to non-negative numbers. The shape of the distribution is defined by the shape parameter λ which represents the average number of events in the given time interval. As λ increases the distribution looks more and more like the normal distribution. When λ is about 10 or greater, then a normal distribution is a good approximation.

16.6.2 Why can't we just use Ordinary Linear Regression?

One of the main assumptions required for fitting an ordinary linear regression (OLR) is that the residual errors must follow a normal distribution. For this to be achieved with data from a skewed distribution, a transformation must be applied however with discrete data this can be very problematic (making the interpretation of the findings unfeasibly difficult) or impossible (for example, a high number of 0's could prevent normality from being achieved). Another issue is that an OLR has the ability to create negative predicted values which would be theoretically impossible. For these reasons it is better to apply a method which actually reflects the natural distribution instead of trying to make the distribution reflect the method. This is why a Poisson regression is generally more suited to count data than OLR.

16.6.3 Poisson Regression

A GLM for Poisson distributed outcome is commonly known as Poisson regression but is sometimes referred to as a log-linear model.

Random component

Supppose we have an outcome Y representing counts of events over a fixed time period T. For simplicity, we will let $T=1$, i.e. we have followed our individuals up for a single unit of time (e.g. one year).

Suppose we are happy to assume that Y follows a Poisson distribution. We wish to model the relationship between a vector of p covariates $\mathbf{X}$ and the expectation of Y (and therefore also the variance of Y, since the mean is equal to the variance for a Poisson variable).

We let $\mu = E[Y | X]$ and assume:
$$ \mathbf{Y} \overset{iid}{\sim} Poisson(\mu) $$

Systematic component (linear predictor)

The linear predictor is a linear function of the covariates $\mathbf{X}$:
$$ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p $$
Using vector notation to simplify the maths, this is equivalent to:
$$ \mathbf{X}^T\mathbf{\beta} $$

(Note that we assume the vector $\mathbf{X}$ contains a constant, so $\mathbf{X}^T = (1, X_1, X_2, ..., X_p)$.

Link function

We want an equation that connects the expected outcome μ (which must be positive) to the linear predictor $\mathbf{X}^T\mathbf{\beta}$ (which can be any real value, positive or negative). The link function is applied to the expected outcome, therefore in this case an appropriate link function must map the positive real values to all real values. An obvious candidate is the natural logarithm. This is the default (often called the *canonical*) link function for Poisson variables. Then, applying the link function to the expected value of the outcome, we have:
$$ ln(\mu) = \mathbf{X}^T\mathbf{\beta} $$
This is the **Poisson regression** model.

Regression coefficients

In the model above, β is a vector of regression coefficients. An element of β represents the expected change in the natural log of the mean per unit change of one explanatory variable in X (constraining the other elements to not change).

We can interpret μ as the expected rate of the outcome (remember we assume the fixed time period being considered is $T=1$). Consider a simpler example with a single binary covariate X. Then the model above becomes
$$ ln(\mu) = \beta_0 + \beta_1 X $$
This says that
$$\begin{split} \begin{align*} ln(\mu) &= \begin{cases} \beta_0 &\mbox{if } X=0 \\ \beta_0 + \beta_1 &\mbox{if } X=1 \\ \end{cases} \end{align*} \end{split}$$
The expeted rates in the two groups $(X=0)$ and $(X=1)$ are then:
$$\begin{split} \begin{align*} \mu &= \begin{cases} e^{\beta_0} &\mbox{if } X=0 \\ e^{\beta_0 + \beta_1} &\mbox{if } X=1 \\ \end{cases} \end{align*} \end{split}$$
So the rate ratio comparing the rate in the exposed $(X=1)$ with the rate in the unexposed $(X=0)$ is:
$$ \begin{align*} RR = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} \end{align*} $$
Therefore, e^{β_1} can be interpreted as the (incidence) rate ratio comparing the exposed with the unexposed groups. Or β_1 can be interpreted as the log rate ratio.

More generally, a regression coefficient can be intepreted as the log rate ratio associated with a unit change of that covariate. Its exponential is the analogous rate ratio.

Eestimating the regression coefficients

As for all GLMs, the regression coefficients, $\mathbf{\beta}$, are estimated by maximum likelihood estimation.

The mean and variance

To obtain an equation for the mean of the outcome, we need to apply the inverse link function:
$$ E[\mathbf{Y}|\mathbf{X}] = \mu = g^{-1} (\mathbf{X}^T\mathbf{\beta}) $$
Which here is equal to
$$ E[\mathbf{Y}|\mathbf{X}] = \mu = e^{\mathbf{X^T\beta}}, $$
Similarly the variance of $\mathbf{Y}$ is written:
$$ Var[\mathbf{Y}|\mathbf{X}] = Var[\mu] =Var[g^{-1} (\mathbf{X}^T\mathbf{\beta})]. $$

16.6.4 Offsets [optional]

In this short sub-section, we explore an extension of the Poisson regression model above which allows us to take into account the fact that the observations in our data may represent counts from different lengths of observation time. This is a common occurrence in practice. We handle this through something called an **offset term**.

Above, we simplified the model by assuming each individual was observed for the same period of time. But suppose that was not he case. Suppose, for example, we are counting the number of asthma attacks experienced by school-aged children over time. Some children are followed up for one year and others are followed up for up to five years. Naturally, we would expect those followed up for longer to experience more asthma attacks, on average.

Suppose individual i is followed up for T_i years and experiences Y_i asthma attacks. We have:
$$ \mathbf{Y_i} \overset{iid}{\sim} Poisson(\lambda_i T_i) $$

where \(\lambda_i\) is the annual rate of asthma attacks for individual \(i\). The expected number of attacks for individual \(i\) is \(\mu_i = \lambda_i T_i\). We might wish to propose the following model for the rate (based on the Poisson regression model we met previously):
\[ln(\lambda_i) = \mathbf{X}_i^T\mathbf{\beta} \]
This implies the following model for the expected number of attacks:
\[ln(\mu_i) = ln(\lambda_i T_i) = \mathbf{X}_i^T\mathbf{\beta} + ln(T_i) \]
This is very similar to our previous model. The only difference is that we now have a covariate (\(ln(T_i)\)) appearing in the model without a regression coefficient. In other words, the regression coefficient for \(ln(T_i)\) is constrained to be equal to 1.

This is exactly what an offset term is: a covariate in a regression model with its regression coefficient constrained to be equal to 1.

16.7 Poisson Regression Example

16.7.1 Example data set

For the purpose of illustration, we will simulate some data and pretend it comes from a clinical trial. We generate 100 participants (n) and three variables. The first is a count variable representing the number of hospital admissions (*counts*) a participant has had in a year and it is created from a Poisson distribution with $lambda$=2 . The second is a categorical variable (*country*) with 4 groups representing the country a participant lives in (England, Northern Ireland, Scotland, Wales) and the last is a binary variable (*treatment*) representing which treatment arm the participant was randomised to. Let's start with simulating the data and looking at some descriptive statistics.

```
## Simulate Data
set.seed(42)
n<-100
lambda<-6
counts <- rpois(n, lambda)
country <-  factor(sample(1:4, n, replace=T), levels=1:4, labels=c("England","Northern
Ireland","Scotland","Wales"))
treatment <- factor(gl(2,n/2), levels=1:2, labels=c("Active Arm", "Placebo Arm"))
df <- data.frame(treatment, country, counts)
```

Assume we wish to model *counts* using a GLM with *treatment* and *country* as predictors. We already know the admissions count variable follows a Poisson distribution as we have simulated the data directly from the distribution without adding noise, therefore we know a Poisson regression is suitable. To fit the model, we call the glm() function with the family set to "poisson" and use the summary command to look at the output.

Note: We have used the option `family=poisson`. We could be more explicit and state the link function we want R to use by replacing this with `family=poisson(link=log)`. Try re-running the command using `family=poisson(link=identity)`. What is this doing? Is this a sensible/useful model?

```
set.seed(42)
summary(m1 <- glm(counts ~ treatment + country, family=poisson, data=df))
```

```
Call:
glm(formula = counts ~ treatment + country, family = poisson,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2467  -0.5783   0.0477   0.6381   2.2260

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.82733    0.08448  21.631  < 2e-16 ***
treatmentPlacebo Arm   -0.23187    0.08226  -2.819  0.00482 **
countryNorthern Ireland  0.17001    0.10822   1.571  0.11620
countryScotland          0.14936    0.12396   1.205  0.22822
countryWales             0.06669    0.11640   0.573  0.56670
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 130.13  on 99  degrees of freedom
Residual deviance: 120.53  on 95  degrees of freedom
AIC: 483.69

Number of Fisher Scoring iterations: 5
```

16.7.2 Example GLM output

The first part of the output gives information on deviance residuals. We would expect to see the deviance residuals to be approximately normally distributed if the model is correctly specified. Here we can see the median is close to 0 (0.05) and there does not appear to be any skewness as Q1 (quartile 1 = -0.58) and Q3 (quartile 3 = 0.64) have a similar distance from the median and so are the minimum and maximum.

The second part of the output gives the Poisson regression coefficients for each variable with their standard errors, z values, p-values. We interpret Poisson regression coefficients as if there was a one unit change in the predictor variable (if a continuous variable otherwise change from the reference category to the category listed) the regression coefficient tells us the effect on the logs of the expected counts (admission counts in our example), given the other variables in the model are held constant. The coefficient for treatment is -0.23 which tells use the expected log admissions count for being randomised to the active arm compared to the placebo arm is -0.23. The expected log admissions count for the other countries compared to England are all positive.

We can also see the regression estimate when all the variables in the model are evaluated at zero (or categorical reference group) and this is called the constant and labelled "(Intercept)". In our model this would represent the expected log admissions count for participants in the placebo arm who live in England.

The standard errors are given which are used to calculate the z-value which in turn is used to calculate the p value. The null hypothesis for each p value is that the corresponding regression coefficient is zero given the rest of the variables in the model. The z value here is just the ratio of the coefficient to the standard error for example treatment we can see the estimate/standard error equals the z value: -0.23187/0.08226=-2.819. The z value follows a normal distribution and is tested against a two-sided alternative hypothesis that the coefficient is not equal to zero. We can see for treatment the p value is 0.005 and if we set out alpha significant level at α=0.05 we would reject the null hypothesis and conclude the Poisson regression coefficient for treatment is statistically different from zero, given country is in the model.

Lastly, at the bottom of the output, we have information on the residual deviance which can be used to perform a goodness of fit test for the overall model.

16.7.3 Poisson Regression Goodness of Fit Example

At the bottom of the output we see the null deviance and residual deviance from the model. The residual deviance is 120.53 on 95 degrees of freedom (df). There are 100 observations in our model and 5 estimates which gives us 95 df (100-1df for treatment- 3df for each country – 1df for the constant) . To calculate the p-value for the deviance goodness of fit test we simply calculate the probability to the right of the deviance value for the chi-squared distribution on 95 df

```r
pchisq(m1$deviance, df=m1$df.residual, lower.tail=FALSE)
```

0.0395607905595946

The null hypothesis is that our model is correctly specified. Here we can see the p value is 0.0396 which is significant if we set our level of significant at 0.05. We therefore have strong evidence to reject the null hypothesis. This result is expected as when creating the simulated data we made no relationship between any of the variables in the model, so we would expect a poor fit.

16.8 Common Problems in Poisson Regression

16.8.1 Problems

There are two frequent common problems when applying Poisson Regression to count data and both are caused by the deviations from the Poisson distribution assumptions. The first problem is overdispersion and the second is zero inflation.

16.8.2 Overdispersion

Overdispersion happens with then the variance is no longer equal to the mean but larger which violates the Poisson distribution principle. There are two main ways to handle overdispersion, the first is through using a negative binomial distribution (not covered here) instead and the second is to implement something called a quasi-likelihood through a GLM also called a Quasi-Poisson regression.

16.8.3 Quasi-Poisson regression

A Quasi-Poisson regression is often fitted to handle over-dispersion, it uses the same mean regression function and variance function from Poisson regression but allows the dispersion parameter ϕ to be unrestriced from 1. In Poisson regression ϕ is assumed to be fixed at 1 to make the mean and variance equal, in Quasi-Poisson regression ϕ is not fixed and is estimated from the data. Quasi- Poisson regression leads to the same coefficient estimates as the Poisson regression model but inference are adjusted for the over-dispersion through the standard errors. To run a Quasi-Poisson regression in R we just tell the glm() function that the family is "quasipoisson"

15.3.5 Zero inflation

Zero inflation happens when the distribution contains a large number of zero's. For example, if you were to count how many occasions people drank alcohol in a month but included a large number of non-drinkers you will expect to have multiple counts of 0. A Zero-Inflated Poisson (ZIP) distribution can be thought of being generated by two processes, the first generates zeros and the second is generated by the Poisson distribution (which will contain zeros). The two processes look like this:

$P[\mathbf{Y}=0] = \pi (1-\pi)e^{- \lambda }$,

$P[\mathbf{Y}=k] = (1-\pi)\frac{\lambda^{k}e^{-\lambda}}{k!}$,

Where k is a non-negative integer value, λ is the expected Poisson count and π is the probability of extra zeros. The mean of a ZIP is $(1-\pi)\lambda$ and the variance is $\lambda (1-\pi) (1+\pi \lambda)$.

Unfortunately the glm() function is incapable of running a ZIP regression to run, you will need to use the "pscl" package which fits a GLM with a binomial logit link to predict the excess zeros and a GLM with a Poisson log link to model the rest of the distribution.

17. The role of regression in different types of investigation

At the start of this section of the notes, we considered different types of investigation that might be of interest within a health data science project. We grouped these investigations into three classes: descriptive, predictive and causal.

Having explored various types of regression modelling, we now revisit the idea of the underlying investigation and consider the role of the regression model in different types of investigation.

We often use the same statistical tools to address research questions for investigations of different types. Regression is a key tool for analyses in all the types of investigation. In this short session we illustrate how the same regression model could be used in prediction and causal investigations, but that the output from the regression should be used and interpreted differently.

17.1 Simple example

We focus on a simple (fictitious) observational study involving three variables: two binary explanatory variables 'maternal smoking status' (X_{1} = 1: smoker, X_{1} = 0: non-smoker) and maternal socioeconomic status (X_{2} = 1: low, X_{2} = 0: high), and a continuous outcome 'birth weight' (measured in grams). The assumed relationships between the three variables are summarised in the causal diagram in the figure below.

For the purposes of a simple illustration, we suppose that these are the only three variables at play in this 'system'. In reality of course there are many other maternal and other characteristics that affect a baby's birthweight, such as genetics, maternal diet and alcohol consumption, mother's access to prenatal care, and other features of the environment.

Consider a linear regression of Y on X_{1}, X_{2}, i.e.

$$ Y = \beta_{0} + \beta_{1}X_{1} + \beta_{2}X_{2} + \epsilon \tag{1}$$

The estimated regression coefficients and corresponding 95% confidence intervals for this fictitious example are:

$$\begin{split} \begin{align*} \hat{\beta_{0}} &= 3227 \ \ (95\% \text{CI}: \ 1603, 4851) \\ \hat{\beta_{1}} &= -341 \ \ (95\% \text{CI}: \ -513, -169) \\ \hat{\beta_{2}} &= -214 \ \ (95\% \text{CI}: \ -410, -18) \end{align*} \end{split}$$

We now consider how the output from this regression could be used in different investigation types.

17.2 Prediction

If the aim is to predict birth weight based on the two characteristics of the mother, this model allows us to do this. We could obtain the expected value of Y given X_{1} and X_{2} (in this very simple example there are only 4 possible combinations).

In this prediction setting we do not, however, particularly care about the estimates of the regression coefficients. We should instead be concerned with the predictive performance of the model. This could be measured, for example, using R^2, which measures the proportion of variability in the outcome that is explained by the statistical model.

There are many details about how to appropriately assess and quantify the predictive performance of a prediction model which we do not discuss here.

17.3 Causality and explanation

Suppose instead that the aim is to assess the causal effect of maternal smoking (X_{1}) on birth weight (Y). In the simple setting shown in the causal diagram above, maternal socioeconomic status (X_{2}) is the only confounder of the association between X_{1} and Y.

The regression model in equation (1) adjusts for X_{2} and hence the coefficient for X_{1} can be interpreted as the conditional causal effect of X_{1} on Y.

We can make the interpretation that if all mothers in the study population had smoked, the mean birthweight would have been 341 grams lower than had all mothers in the study population not smoked. This is referred to as an 'average causal effect'. The 95% confidence interval can be used to provide information about how precisely we believe we have estimated this causal effect. In this case, the 95% confidence interval excludes 0 and includes only negative numbers, running from -513 to -169.

Here, we have not given any interpretation of the estimate of β_{2} because it wasn't relevant for our research question, even though it was important to adjust for X_{2} to adjust for confounding. In a more realistic setting, there will be many other variables that confound the association between X_{1} and Y and which would need to be accounted for to enable a causal interpretation of β_{1}.

17.4 The "Table 2 Fallacy"

After adjusting for maternal socioeconomic status, maternal smoking was associated with a lowering of 341 grams in mean birthweight. After adjusting for maternal smoking, low maternal socioeconomic status was associated with a lowering of 214 grams in mean birthweight. However, β_{1} and β_{2} in model (1) do not have the same type of interpretation. This is due to the relationships between the three variables.

- *Interpreting |(|beta_2|)*: According to the causal diagram above, maternal smoking status is on the causal pathway from socioeconomic status to birth weight. Hence the parameter \(\beta_2\) represents the effect of socioeconomic status on birth weight that does not go through smoking status - this is a 'direct effect' rather than a 'total effect'.

- *Interpreting |(|beta_1|)*: By contrast, \(\beta_{1}\) represents the total effect of smoking status on birth weight. We do not go into details about definitions of different types of effect. The aim here is simply to point out that the correct interpretation of the coefficients in the regression model in (1) depends on assumptions about the inter-relationships between the three variables, including how they are ordered in time.

In some (or perhaps many) epidemiological investigations that involve exploration of risk factors, estimates of regression coefficients from multivariable models such as that in (1) (and versions with many more explanatory variables) are presented alongside one another in a table, together with confidence intervals and p-values. They may then be interpreted as though all coefficients had the same meaning, ignoring possible inter-relationships between the variables and temporal ordering. As we have seen from the above example, this could be misleading. This problem has been referred to in the literature the 'Table 2 fallacy', because the estimates of regression coefficients are often presented in 'Table 2' in a paper (where 'Table 1' is usually a table of descriptive statistics). See Westreich and Greenland (2013) for a description of the Table 2 fallacy. Bandoli et al. (2018) provide an example in the context of preeclampsia and preterm birth.

17.5 Other analysis approaches

Regression modelling is a fundamental part of the statistician's toolbox and is used in many investigations of different types. We have used regression modelling to illustrate the connection between the analysis method and the underlying aim of the investigation.

However, regression models are not the only tool available. You will come across many other types of analysis method, such as clustering or neural networks, which can also be used in various types of investigation.

Statistics and Health Data Science

We end with some brief remarks about the application of statistics in health data science.

Focus on the research question

The more complicated the statistical analysis becomes, the easier it is to get lost in the technical details. As a data scientist, it is always important to be able to take a step back and re-focus on the underlying research question.

Ask yourself:

- What is the research question?
- What assumptions can I reasonably make, taking into account where and when the data were collected and how they were collected?
- Does the proposed statistical analysis answer the research question?
- How can I assess the robustness of the conclusions of my analysis to the key assumptions I have made?

Know your data

We cannot stress too much the importance of being familiar with your data. Where does it come from? How was it collected? How accurate are measurements? Do similar biases affect measurements from different units/places/times?

A hugely important step in any data science project is to look at your data. The most sophisticated analysis will produce invalid results if based on data that contains substantial errors or incorrectly assembled datasets.

Continue to learn

This module has introduced some key building blocks, concepts and statistical tools that will be very useful for data science projects. However, there are many more statistical techniques that we have not touched on. In your career as a health data scientist, you will continue to learn new methods and approaches.

We hope that this module has provided a solid foundation to build on!

By MSc Health Data Science, LSHTM