Statistical Modelling

A conceptual, visual and practical introduction

true

23 April, 2025

Contents

Preface: what's the problem?

The world abounds with data - but scientific and investigative work does not *begin* with data. It is true that data which is already available may give ideas and suggest hypotheses, but a serious project will start by thinking carefully about well defined objectives. In other words, scientific work begins with a clearly stated *problem*.

Statistical modelling refers to the process by which we collect and use data to gain insight into the problem we have defined and to lead us towards a conclusion. In an interesting paper on scientific and statistical methods, MacKay and Oldford (2000) structured the process in the following broad steps, referred to by the acronym PPDAC. Some of the key questions which are likely to arise in each step have been highlighted.

- Problem
  - What are the key questions we would like to address?
  - What is the context in which these questions are framed?
- Plan
  - How should our experiment be designed?
  - What data should be collected and how much?
- Data
  - How should the data be checked for validity and consistency?
  - What methods of visual exploration should we use?
- Analysis
  - How should an appropriate model be constructed?
  - What does analysis of our model tell us about our problem?
- Conclusion
  - How should we report to others on our conclusions?
  - What limitations and caveats should we highlight?

This broad view of statistical modelling sets the agenda for the book.

There are many textbooks and resources available which discuss statistical modelling - why another one? This book aims takes a particular approach.

- The focus is on **conceptual** understanding of the main ideas behind statistical thinking and modelling. Some technical details are provided for those who are interested but engagement with this material is optional.
- There is a strong theme of **visual** communication of both data and the concepts behind statistical models.
- The approach is also **practical** with extensive reference to the widely used statistical computing environment R as a means of engaging with the concepts and implementing the methods discussed. In particular, there is extensive use of real datasets. The aim is to discuss interesting and scientifically important questions, where possible with the data used in published papers. As data become increasingly 'open', datsets are read from publicly available sources wherever possible.

The target audience is those who need statistical methods to understand data. A good example is PhD students who are well motivated to analyse their experimental data. The scientific contexts of the examples come from a wide range of application areas, including the life sciences, the social sciences and topics of general interest. The aim is to provide examples which are both interesting and accessible, without the need for detailed technical knowledge of a particular area.

Throughout the book there are frequent references to the widely used statistical computing system R. It is possible to read this book without using R but it is primarily intended that the reader will use this powerful system to engage with the examples and exercises and with the whole process of statistical modelling. A description of how to install R, and the popular system RStudio which helpfully manages some aspects of the environment, is available immediately after this preface. A gentle introduction to R is provided in Chapter **??**.

Acknowledgements

Please note that this is a work in progress. Please forgive some rough edges in the presentation here and there.

Chapter 1

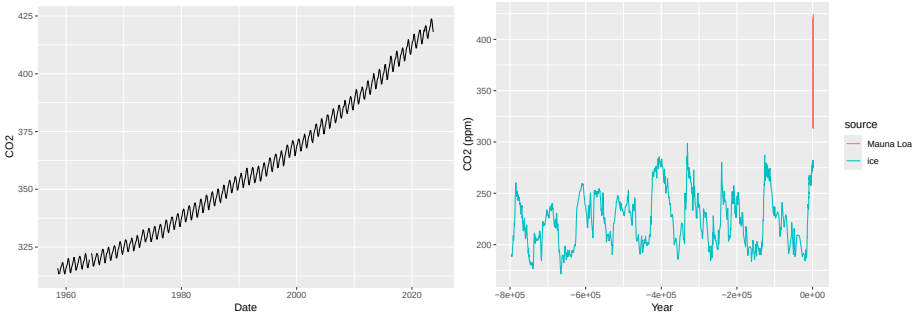
Where do data come from?

Data can take many forms. Sometimes we are dealing with simple numbers, such as blood pressure measurements or age. Some data may consist of categories, such as country of origin or presence/absence of a characteristic of interest. Modern data can be much more sophisticated, arising in the form of images, temperature curves, three-dimensional surfaces or networks. This chapter focuses on relatively simple data forms but uses these to discuss some important principles of how data are collected and what they are able to tell us.

1.1 Some examples of data

1.1.1 CO2 and global warming

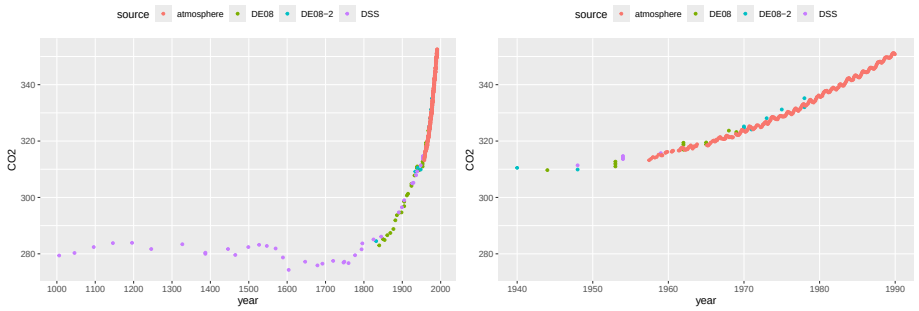
One of the world's most pressing issue is climate change as a result of global warming. The levels of CO2 in the atmosphere are clearly implicated in this as a result of the 'greenhouse' effect. Accurate monitoring of CO2 began in the late 1950's through the pioneering work of Charles David Keeling at Mauna Loa Observatory in Hawaii and measurements have been made continuously since then, providing an invaluable record of change. This 'time series' is displayed at monthly resolution in the left hand panel below.



This time series makes it clear that there has been a steady rise in atmospheric CO2 over the entire period of monitoring. The implications of this become startling when we compare the series to data which indicate the levels of CO2 which correspond to earlier historical times. This can be done by measuring the CO2 content of air trapped in ice cores. Careful analysis of the ice layers allow the identification of the time scale, which stretches back over hundreds of thousands of years. The CO2 measurements from ice cores has been superimposed on the plot of the Mauna Loa data in the right hand panel above. This makes it clear that the modern measurements of CO2 are 'off the scale' of the historical levels.

In order to assuage any concerns about the comparability between the modern atmospheric and historical ice core CO2 measurements, the plots below focus on the period where these overlap. The left hand plot highlights the sudden dramatic rise in CO2 as the industrial revolution gathered pace. The right hand panel zooms in on the years when both types of measurement are available, indicating the very strong level of agreement between the two.

We will look at corresponding temperature changes later, but these data already paint a stark picture of the nature and size of the challenge we face in addressing climate change. The Intergovernmental Panel on Climate Change synthesises our scientific understanding of the process and continues to report on the current situation, with urgent calls to action.



This example highlights that, in some settings, there is a need for an appropriate *control* to enable informative comparisons. Controls can sometimes be difficult to identify. In this case, considerable effort has been undertaken to ensure the validity of comparisons with the control data. The example also highlights the relevance of our scientific understanding of the context in which the experiment takes place. The physics of greenhouse gasses is well understood and this allows us to strengthen the interpretation of what we see in the data.

1.1.2 The first tuberculosis trial

One of the very earliest systematic evaluations of medical treatment, marking a significant step in what we now term 'clinical trials', was a study on the effects of *Streptomycin* on pulmonary tuberculosis by Marshall et al. (1948).

The statistician Austin Bradford Hill, whose picture is below, introduced very important methodology in this study. In his review of the development of clinical trials, Bhatt (2010) writes "This trial was a model of meticulousness in design and implementation, with systematic enrolment criteria and data collection compared with the ad hoc nature of other contemporary research. A key advantage of Dr Hill's randomization scheme over alternation procedure was "allocation concealment" at the time patients were enrolled in the trial. Another significant feature of the trial was the use of objective measures such as interpretation of x-rays by experts who were blinded to the patient's treatment assignment."

Figure 1.1: Sir Austin Bradford Hill, Wellcome Collection, CC BY.

The headline results reported in the scientific paper are shown in the table below.

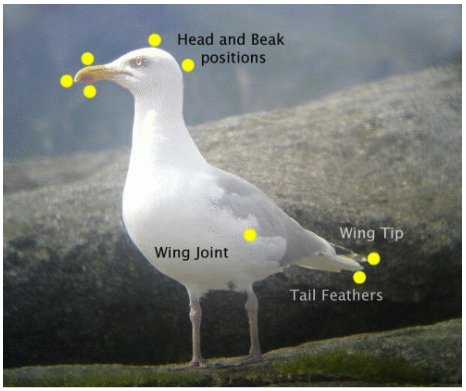
Radiological assessment	Streptomycin group	Control group
Considerable improvement	28	4
Moderate or slight improvement	10	13
No material change	2	3
Moderate or slight deterioration	5	12
Considerable deterioration	6	6
Deaths	4	14
Total	**55**	**52**

The beneficial effects of the treatment are clear and we can have confidence in the conclusion because of the careful conduct of the trial. Key features of the design include the presence of a *control* group and the use of randomisation.

1.1.3 The birds and the bees: how to tell the sex of a herring gull

Herring gulls are found across the coastal regions of North-Western Europe. When studying the behaviour of these birds, it is useful to be able to identify sex. With this species, this is not possible by visual examination of the obvious anatomical features as the appropriate organs are internal. It would therefore be very useful to be able to identify the sex of a bird by taking simple measurements of some kind, on the assumption that the sexes are lijkely to differn in size, as

happens with many animal species. The correct identification of the sex of a herring gull has to be carried out by dissection. The most suitable source of data for this purpose is therefore birds which have been found dead or have been culled for other reasons. Measurements from a sample of 100 male and 100 female birds, kindly provided by Prof. Pat Monaghan from thte University of Glasgow, are available for investigation.



Length measurements, based on the distance between two of the yellow landmarks in the picture above, could be useful in distinguishing between the sexes. as males and females tend to have different sizes in many species. There is an interactive application in R which can help in thinking this through. A gentle introduction to R is provided in Chapter **??** but, if you have followed the guidance at the start of the book on installing R and the add-on package `rpanel`, then type in the following instructions into the *console* window to launch the app.

```
library(rpanel)
rp.gulls()
```

Consider which pairs of landmarks might provide a suitable length measurement. Suitable criteria are:

- *reproducible*, by yourself and others;
- *valid*, in what they aim to measure;
- *informative*, as they are likely to be different for male and females;
- *well calibrated*, as they target a feature of interest;
- *practical*, as measurements can be made reasonably easily.

Click on your selected pairs of landmarks and some feedback will be given. If you are able to identify some suitable measurements, checkboxes and buttons will appear to allow you to see some plots of the data, separated out by sex.

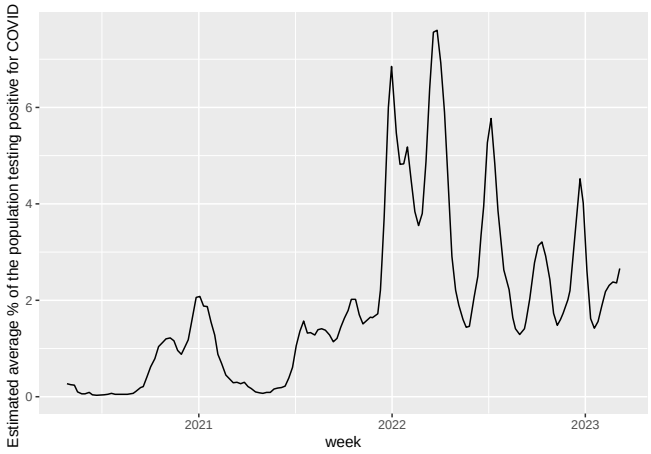
1.1.4 Tracking the covid pandemic

When the covid19 pandemic began there was immediate and urgent effort to track its progress. That happened in many ways, most immediately in the numbers

of deaths and hospital admissions. In the UK, when tests became available, the number of positive results was also regularly reported. The symptoms data collected by the app from the ZOE Health Study provided another source of information.

Although these sources provided useful information, the most reliable estimates of covid infection levels came from a large survey conducted by the Office of National Statistics (ONS), in partnership with the University of Oxford, the University of Manchester, the UK Health Security Agency (UKHSA) and the Wellcome Trust. Swab and blood samples were provided by thousands of people from across the UK who had been selected and random and who had agreed to participate by providing repeated samples over an extended time period. Statistical modeling was also undertaken to ensure that the results represented the population as a whole.

The ONS provided extensive information about the data, including details of the methods and study design. Scientific publication of the methods appeared in a *Lancet* paper. This is an example where very considerable effort was made to ensure that what we observe is an accurate representation of what is happening.



The plot above uses the survey data to track the pandemic in terms of the percentage of people in the UK who tested positive. The virulence of the infection changed over time so deaths and hospital admissions show rather different patterns. We will look at data on those later.

An account of the covid-19 tracking project is given in an article published by the Royal Statistical Society in its *Statistics Under Pressure* series. The hugely valuable nature of the information provided by the ONS survey is discussed in a *Conversation* article.

1.1.5 More complex data objects

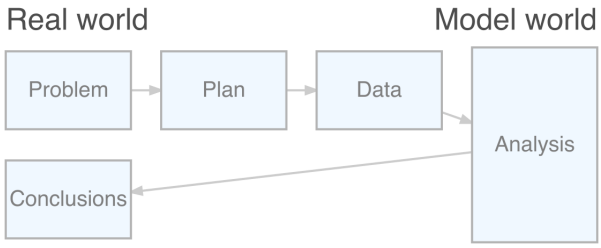
The examples above mostly involve measurements of a single quantity of interest, such as CO_2, the level of health improvement or the proportion of people with covid19. The herring gulls example might involve several different measurements on each bird. However, data can be much more complex. Some examples are:

- high-resolution images captured from a video camera monitoring wildlife movement;
- free-form text entered into the search box of a web browser;
- an extensive set of responses recorded from an individual in a survey, including later questions which are conditional on the responses to earlier questions;
- a network describing the interactions of one individual with others in a group;
- the life history of a hospital patient.

As the complexity of data increases, so the models required for analysis may also need to increase in sophistication. Nonetheless, some basic principles and concepts still apply and it is the aim of this book to discuss these. While the focus is on relatively straightforward types of data, the ideas will provide helpful building blocks for more complex situations.

1.2 Some broad issues

It is worthwhile reflecting on some of the broad issues which have already arisen in our review of these examples of data. To structure our reflections it will be helpful to recall the PPDAC framework discussed in the Preface, displayed below in graphical form. In this section we will focus on the *Problem*, *Plan* and *Data* stages.



1.2.1 What's the problem?

It is very important to define the objectives of an experiment clearly. Failure to do this will make subsequent analysis and interpretation very difficult. Pilot studies are a perfectly valid preliminary step but a clear objective is still needed. 'Fishing expeditions' are of rather limited use.

Our objectives will be informed by previous work so a review of the scientific literature is a wise start. Scientific knowledge of the subject domain will be very helpful not only in deciding what problem to tackle but also in informing later decisions in the planning process and in due course in interpreting the results of our analysis.

There are some general aspects of this process which are worth highlighting. The first is the distinction between an *observational study* where we simply observe and record variables of interest. We do not intervene or influence the situation in any way other by recording what we see. In an observational study we can only identify *association* between variables. A simple of example is the association between social or economic characteristics and political voting patterns.

In contrast, a *designed experiment* involves the specification of 'treatments' which are assigned by the experimenter. Rather than simply observe, we intervene. We are then interested in the way the treatments affect the outcome. If a designed experiment is conducted well it allows us to identify *causal* relationships between variables. A simple example is a clinical trial, such as the tuberculosis trial described in Section 1.1, where we compare the effects of different treatments on the recovery of patients suffering from a particular medical condition.

A second general issue is whether our aim is to *understand* the processes at work in our scientific context or whether we simply want to *predict* an outcome or design a system which will *classify* future observations into different groups. The classification of herring gulls into male and female groups discussed in Section 1.1 is an example of the latter. This is a case where we may not be primarily interested in which variables are involved in our model, simply in how successful our model is in prediction or classification. This may change our attitude to building a model, when the time comes, but it is also helpful to be aware of the distinction between these two aims from the start of the planning process.

1.2.2 What's the plan?

Acquiring data which are appropriate, informative and unbiase requires careful thought. The ONS covid-19 survey was described in Section 1. Take a look at the methodology guide for this survey. It is very extensive, covering all the major issues which had to be considered.

One of the exercises at the end of this chapter ask you to consider how you might design a simple experiment to investigate the operation of short term memory, known to psychologists as "working memory". How is a list of items recalled from memory? This is an experiment which can be caried out in a classroom or small group setting. It might surprise you just how extensive the list of detailed arrangements needs to be.

One of the important tasks is to identify which measurements should be taken. The list should include not only those which are mentioned in the definition of the problem being tackled but also those which we know or suspect from previous

work or our scientific understanding are likely to influence the process we are studying. This will help us to consider an appropriate experimental design - a topic we will revisit in a later chapter.

1.2.3 Where do the data come from?

The diagram at the start of this section includes two headings, "Real world" and "Model world". The question our experiment aims to tackle is about what is going on "out there in the wild", even when the "wild" refers to a laboratory setting. If we consider all the observations we might ever make then we can view this as the *population* we are studying. The process of collecting a particular dataset then delivers a *sample*. Our modelling process will use this sample to try to understand what is going on in the population. It is then crucially important to obtain a sample which properly represents the population and does not suffer from serious *bias*. If we make serious mistakes at this stage, it is unlikely that we will be able to retrieve the situation.

A good mechanism for avoiding bias is to use *randomisation*. This applies in two ways. The first is to the process of identifying which items from the population are recruited into our sample, ideally ensuring that all potential items have an equal chance of appearing. The second applies to the allocation of any treatments to sampled items. The tuberculosis trial outlined in Section 1.1 is one of the first occasions where this was used. It is now a standard component of clinical trials worldwide.

1.2.4 Who do you trust?

We live in a world where 'fake news' has become a commonly used term. Not everyone is careful in the way data are collected, analysed and interpreted. Sadly, data are sometimes used selectively to support a conclusion already adopted. It is very important that this is countered and that the analysis of data is conducted in an honest and professional manner.

The UK Statistics Authority is the body which oversees the production of official statistics in the UK. Its code of practice is based on the principles of:

- *trustworthiness*, confidence in the people and organisations that produce statistics and data;
- *quality*, data and methods that produce assured statistics;
- *value*, statistics that support society's needs for information.

The full code is well worth reading. Although it is couched in terms of official statistics, providing information for the public and for government, the principles it describes are very important. A Declaration on Professional Ethics is also provided by the *International Statistical Institute*.

In addition to the obligation on all who collect and analyse data to act with professionalism and integrity, specific ethical considerations arise in the planning

and design of experiments, particularly those involving humans. Indeed, the protocol for any experiment involving humans must be approved by an appropriate ethical committee before it can be put into practice. A simple example is in a clinical trial to compare two treatments where the size of the sample must be considered very carefully. The sample must be large enough to enable a clear conclusion about the treatments to be reached, but if it is too large then some patients may end up being given a treatment which might clearly have been shown by a smaller trial to be inferior. That would not be ethical.

We will often find ourselves using data from other sources as part of our investigations. Indeed, many different sources are used in this book. The issue of trustworthiness, raised in the UK Statistics Authority code of conduct, again arises. What sources can we trust as reliable?

Some guidance comes from the source organisations stated aims and code of conduct. Reputation also matters. Open documentation is important, so that the details of the data collection process can be reviewed. Accountability matters too, so that there is a mechanism for query and complaint if the need arises.

The RSS has provided a document, *Sound or suspicious? Ten tips to be statistically savvy* which offers advice on how to assess claims that are made.

1.3 What could possibly go wrong?

It is very instructive to think about studies which went wrong.

1.3.1 US presidential elections, 1936 and 1948

The *Literary Digest* was a magazine which surveyed 10 million people, beginning with its own readers, who they planned to vote for in the 1936 US presidential election. A massive 2.4 million people responded, leading to the prediction of a clear win for the candidate Alf Landon. In fact, Franklin Delano Roosevelt had a landslide victory. What went wrong?

For another occasion when things went wrong in the prediction of the outcome of a US presidential election, see the article in the Los Angeles Times about the 1948 election.

1.4 Further reading

The Royal Statistical Society has a very helpful Guide to UK official statistics on climate change.

The CEDA Archive, maintained by the National Environmental Research Council (NERC) in the UK, contains a very large collection of environmental data from atmospheric and earth observation research.

If data are to be collected through a survey, the nature and construction of the questions are very important, to avoid bias or leading the respondent. The Pew Research Centre, a high profile independent organisation, provides a helpful discussion of writing survey questions.

Copernicus: European environmental data. https://surfobs.climate.copernicus.eu/dataaccess/index.php

Diggle and Chetwynd (2011)

Rosling (2018)

The Tiger That Isn't. Andrew Dilnot & Michael Blastland.

How to lie with Statistics. Darrell Huff.

Damned Lies and Statistics. Joel BEst.

More Damned Lies and Statistics. Joel Best.

Innumeracy. John Allen Paulos.

Reckoning with Risk. Gerd Gegerenzer. (Some people object to technical errors?)

Dicing with Death. Stephen Senn.

Risk. John Adams.

Britain in Numbers. Simon Briscoe.

Why Do Buses Come in Threes. Rob Eastaway.

How Long is a Piece of String. Rob Eastaway.

How to Take a Penalty. Rob Eastaway.

1.5 Exercises

1.5.1 An investigation of short-term memory

Consider how you might design a simple experiment to investigate the operation of short term memory, known to psychologists as "working memory"? How is a list of items recalled from memory? This is an experiment which can be caried out in a classroom or small group setting. It might surprise you just how extensive the list of detailed arrangements needs to be.

Once you have spent some time considering this, you may like tyo consult Bowman (1994) which describes some of the issues which arose in a classroom setting.

1.5.2 A survey of dental health

Imagine you have been commissioned to conduct a survey of the dental health of five year old children in England. Write down some of the things on which you will need to make decisions and sketch out some possible answers. This should include how a suitable sample of children will be selected, what measurements will be made, how this will be done, and another other issues which you think are relevant.

In fact, a survey of exactly this type is regularly conducted under the National Dental Epidemiology Programme (NDEP) for England. Once you have spent some time considering the issues, you can see the detail of what was done in the documents available here for the 2016-2017 survey. The protocol document describes the planning of the survey in considerable detail. The results of a further survey in 2022, using the same protocol, are also available.

1.5.3 Hearings aid and dementia

A Lancet paper studied the association between hearing loss and dementia, in particular examining the role of hearing aids. The interpretation of the findings were:

> In people with hearing loss, hearing aid use is associated with a risk of dementia of a similar level to that of people without hearing loss. With the postulation that up to 8% of dementia cases could be prevented with proper hearing loss management, our findings highlight the urgent need to take measures to address hearing loss to improve cognitive decline.

Are there other possible interpretations? You may wish to look at this article published by the British Geriatrics Society which discusses the issue. The article also provides a link to a further scientific paper for the technical detail.

Part I

Statistical Inference

Chapter 2

Inference: turning data into evidence

So far we have considered how to read, organise and visualise datasets. We are now ready to discuss one of the main themes of statistics, namely weighing up the evidence for the presence or size of different features in the population lying behind the data. That process is referred to as *statistical inference*. The first step in doing this is to think about the uncertainty or variation involved in the data and how to quantify this. That will enable us to discuss some standard tools, such as confidence intervals and hypothesis test, which we will explore in this chapter in some simple settings.

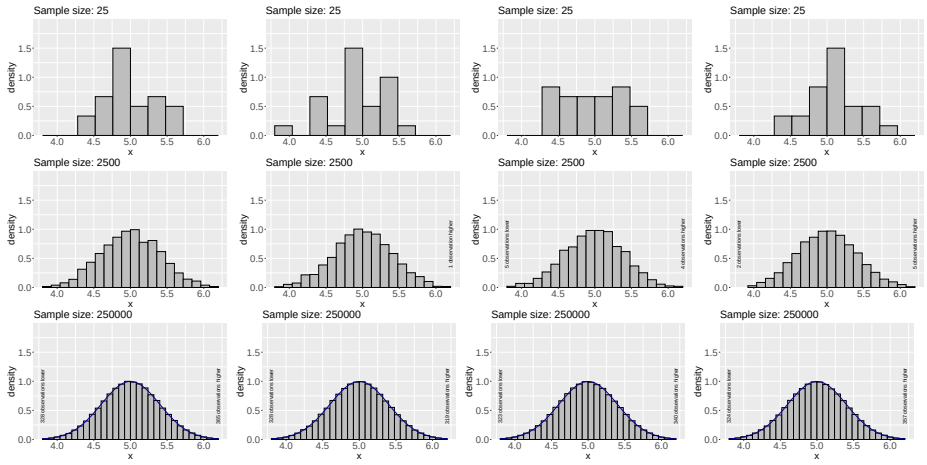
2.1 Samples and populations

The distinction between a *sample* of data and the *population* from which the data come was mentioned briefly in Section 1.2.3, where a population was defined rather informally as the collection of all the observations we could ever make of the process we are studying. We will now explore this further by carrying out a kind of 'thought experiment'. By sampling from a population whose characteristics are known, using R to do so, we can investigate how well the features of a sample guide us on the features of the underlying population.

To give our thought experiment a little context, let's assume we are taking water samples from a particular location on a river with a view to assessing the water quality. This is often quantified through the percentage of dissolved oxygen.

The `rp.sample` function in the `rpanel` package provides a convenient means of experimenting with this situation. The plots below give a flavour of what can be done but you may find it more instructive to run the function 'live' to create your own images and experiment with different settings. The simple instruction

`rp.sample()` will launch a window with a panel of interactive controls.



The plots above shows histograms of multiple sets of sampled data. In each case the population is the same but the sample is different every time. It is the nature of variability that the particular data points we see, and so the details of the patterns which are displayed, change with every sample.

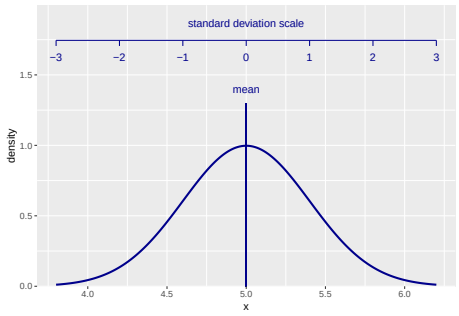
The first row of plots has sample size 25. Here there is a lot of variability in the shapes of the histograms. Sometimes there appear to be two clusters of data, or a rather skewed shape. We know that the underlying population is the same for all samples but the detailed shapes we see can be quite different. This tells us that we should be cautious in interpreting detailed patterns when the size of our sampled dataset is small.

The second row has sample size 2500 and here the variation in shape is much smaller. The third row of plots has sample size 250000. Now there is almost no variation in shape. The continuous curve shows the shape of the distribution from which the data were sampled and the histograms match this very closely every time. This illustrates the general principle that more data gives us more information, reducing the variability in the features of the population that the data express.

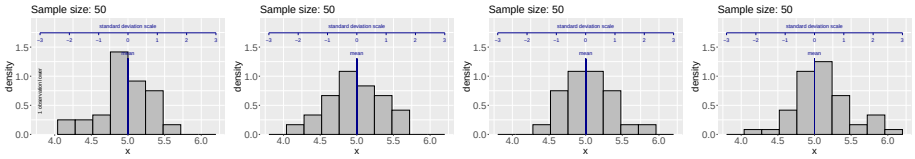
The continuous curve is what the histograms 'converge' to as the sample size gets higher and higher and variability gets smaller and smaller (and the histogram bin width also get smaller and smaller). This is known as the *density function* and it defines the population from which we are sampling. For any interval on the axis, the area under the density function gives the probability of an observation falling in this range so, in that sense, the density function describes how the probability of observing different values changes across the axis.

This particular population has a *normal distribution*. This has a characteristic symmetric shape which falls away smoothly as we move to values further from the centre of the distribution. Just as the mean of a sample can be characterised

as the 'centre of gravity' of the dataset, so the mean of the distribution can be defined as the 'centre of gravity' of the distribution. One consequence of the symmetry of the normal distribution is that the mean sits at its point of symmetry, where the density function reaches its highest value. In the present case the mean has the value 5. Similarly, the spread of the distribution around its mean value can be quantified by the standard deviation. Like the mean, this can be defined in terms of the density function but we can also simply regard it as the value to which the sample standard deviation converges as the sample size gets higher and higher. In the present case the population standard deviation is 0.4.



The plot above shows the normal distribution, with the mean highlighted and with a superimposed scale which measures distance along the axis in units of standard deviation, so '1' lies at a distance of 0.4 from the mean, '2' at 0.8 and so on. This shows that most of the distribution lies within 2 standard deviations of the mean. This happens with samples too so that, most of the time, the observations in a sample lie within 2 standard deviations of the mean. The plots below illustrate this.



This makes sense, as the standard deviation quantifies how far away observations are from the mean, on average. So, informally, one standard deviation away from the mean, on either side, takes us 'halfway' into the spread of the data. From that perspective, it then seems reasonable that two standard deviations away from the mean should cover most of the observations in our sample. This works rather neatly for the normal distribution but in fact the general principle holds for many distributions, even ones which are not symmetric, although the guideline will become less effective as the size of skewness increases. An exercise in Section 2.7 below will invite you to explore this.

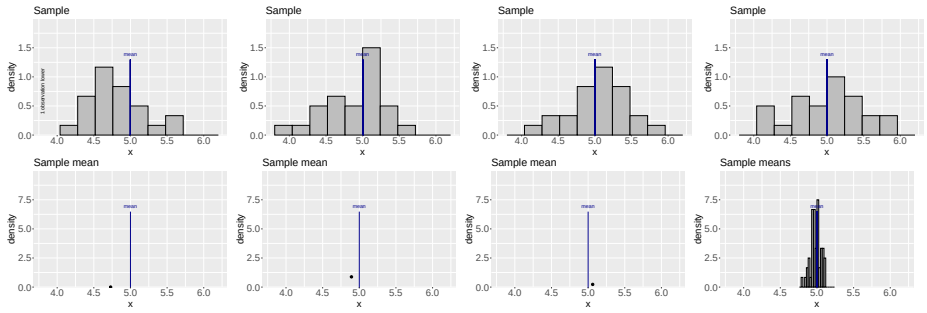
The normal distribution is a very important one for modelling measurements on a continuous scale but there are many other possible distributions. We will meet some of these in due course, as we encounter datasets with features which need

other distributional shapes to describe them.

2.2 Quantifying uncertainty: standard error

The 'thought experiment' of the previous section explored the relationship between a sample and a population but this was necessarily set up rather artificially, as we defined (through the settings in `rp.sample`) the population from which the samples were being drawn. In practice, of course, a sample is all we have and we seek to use this to learn about the population - perhaps simply its mean, or perhaps other features of interest. This is why the title of the current chapter is 'Inference', as we seek to infer features of the unknown population from the (usually rather limited) data in our sample.

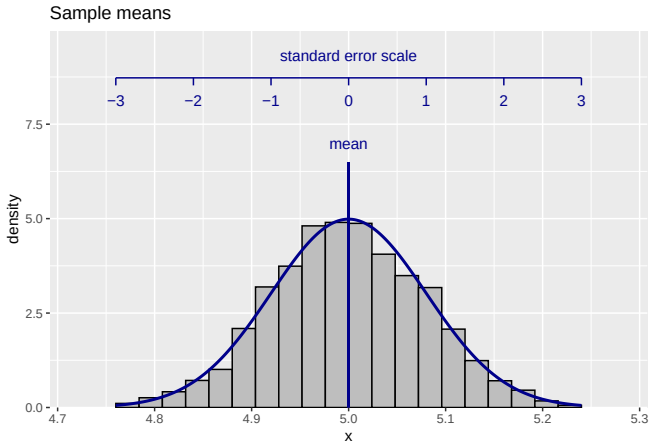
We will explore this by continuing our thought experiment, using the population mean as the focus, as this is often a parameter of major interest. How accurate is the sample mean as a guide to the population mean? The `rp.sample` function can again help us think this through. The plots below show, in the upper row, multiple samples, each of size 25. The second row of plots shows the mean of each sample, with the final plot in this row showing sample means accumulated over 50 different samples. The sample means are clustered round the true mean. There is variation in the sample means but the size of this is much smaller than the variation in the individual observations.



This is the point at which some theory can help us. This shows that the standard deviation of the sample mean, known as the *standard error*, can be easily calculated as

$$\text{se}(\bar{x}) = \sigma/\sqrt{n}.$$

This gives us a measure of uncertainty, or inaccuracy, of the sample mean $\bar{x}$ as an estimate of the true mean μ. In practice, we would estimate this as $s/\sqrt{n}$.



This is a very important fact which we should pause to highlight in general terms of a true *parameter*, here the mean μ, and an *estimate* of it, here the sample mean $\bar{x}$.

```
                    Most of the time,
the true parameter and the estimate are within 2 standard errors of one another.
```

The sample mean has a distribution. If we knew what this was then we could quantify the uncertainty of the mean of our sample. Theoretical calculations tell us that the sample mean also has a normal distribution, that it's mean is the mean of the original distribution, and we can derive an expression for the standard deviation of this distribution. Formula.

Estimated standard error. t-statistic. t-distribution.

This very important principle will be the basis of the tools we now develop.

2.3 Confidence intervals

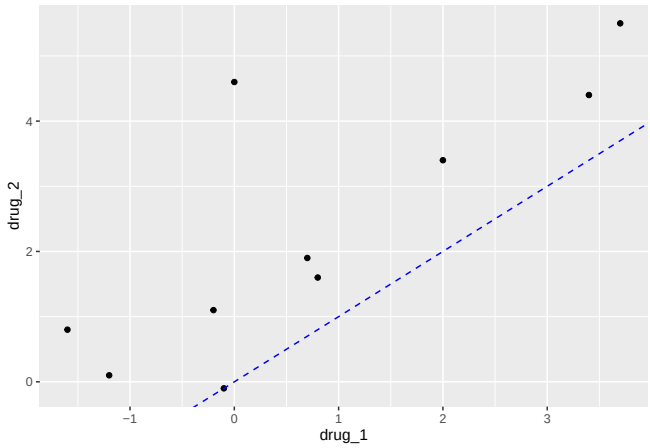
This statement describes how the sample means $\bar{x}$ varies around the true mean μ. In practice all we see from our observed sample of data is a single $\bar{x}$ and its estimated standard error. However, we can turn the statement around and expect that the true value of μ will lie somewhere within 2 standard errors of $\bar{x}$. The interval

$$(\bar{x} - 2\text{se}(\bar{x}), \bar{x} + 2\text{se}(\bar{x}))$$

is therefore a range of plausible values for the true mean μ. This is called a *confidence interval* because we can attach some useful properties of this kind of interval; these are discussed below. We can also carry out some more careful probability calculations which show that, while the multiplier 2 is a good approximation, a more precise value is the percentile of a t-distribution with $(n-1)$ *degrees of freedom*.

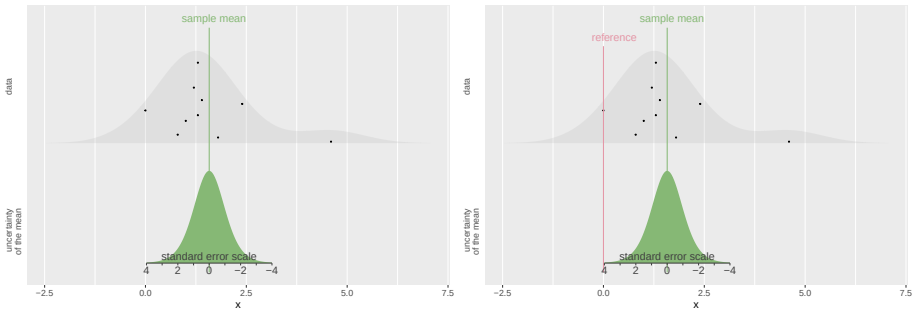
The first person to propose this idea was W.S.Gossett, in a famous paper in 1908. He published under the pseudonym *Student* because he was working for the *Guiness* brewery at the time. He used data from a paper by Cushny & Peebles (1905) which compared the effects of two different drugs on lengthening the hours of sleep in 10 different subjects. The experimental design involved each subject taking each of the two drugs, so this is *paired* data. A helpful first step is to plot the paired values against one another, with the line `y = x` providing a reference. If there is no systematic difference between the effects of the drugs then the observations should cluster around this line.

```
sleep_wide <- pivot_wider(sleep, values_from = extra, names_from = group,
                          names_prefix = 'drug_')
ggplot(sleep_wide, aes(drug_1, drug_2)) + geom_point() +
   geom_abline(slope = 1, col = 'blue', linetype = 2)
```



The extra times for drug 2 are higher than those for drug 1, apart from one patient where the times were identical. In one patient the difference between the two drugs is particularly large but otherwise the differences are broadly similar. In particular, there is no obvious change in the size of the difference as the size of the response to drug 1 changes. This allows us to focus on the differences of the responses from the two drugs as the key information to analyse.

```
sleep_diff <- with(sleep_wide, drug_2 - drug_1)
rp.t_test(sleep_diff)
rp.t_test(sleep_diff, mu = 0)
```



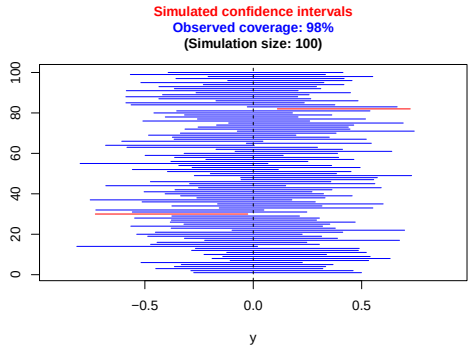
Comment that the areas of the data density and uncertainy distribution are not comparable. It is only relative sizes within in each distribution which matter.

Discuss the construction of a confidence interval.

Two sample example.

Relegate rp.ci to an exercise?

The meaning of a confidence interval sometimes causes a bit of confusion. The interpretation as a range of plausible values for the quantity we are estimating is a helpful but informal one. To explore the formal properties a little further we can experiment with the `rp.ci` function in the `rpanel` package for `R`. An example is shown below. Here we randomly sample 30 observations from a distribution with mean 0 and standard deviation 1 and compute a confidence interval for the mean. However, we do this 100 times. Each interval will be different because it is based on a different random sample of data. The *confidence* of the intervals is conventionally set at 95% and this is done by choosing percentiles in the confidence interval formula which capture 95% of the t-distribution. Then, on average over repeated random samples, the computed confidence intervals will capture the true value 95% of the time. Any particular set of 100 intervals will not have exactly 95 which capture the true value but if we keep repeating this and accumulate the tally, the proportion of intervals which capture the true value will settle down to 95%.



2.4 Hypothesis tests

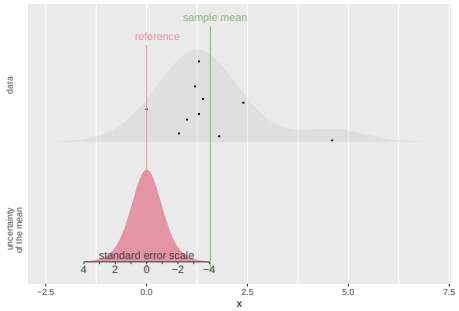
This way of thinking is based on weighing the available evidence to distinguish between two hypotheses. For example, if we have a sample of data there may be interest in examining whether it is plausible that the mean takes a particular value, say μ_0. If the measurements are differences in blood pressure between two time points for a set of patients, we may be interested in whether the mean of the measurements is 0, indicating no change in the mean, or not. The hypotheses are:

Null hypothesis: $\mu = \mu_0$
Alternative hypothesis: $\mu \neq \mu_0$

To distinguish between these hypothesis we can use again the helpful principle we identified in the discussion of samples and populations above.

```
                    Most of the time,
   the true parameter and the estimate are within 2 standard errors of one another.
```

```
rp.t_test(sleep_diff, uncertainty = 'reference')
```



We view things through the eyes of the null hypothesis. If this is true then, most of the time, $\bar{x}$ and μ_0 will be within 2 standard errors of one another. If we define the *test statistic* to be

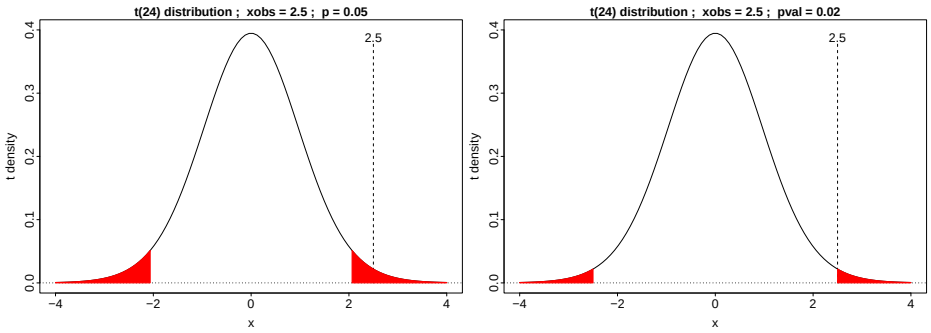
$$t = \frac{\bar{x} - \mu_0}{\text{s.e.}(\bar{x})}$$

then we should expect $-2 \leq t \leq 2$ most of the time. More precisely, $(\bar{x} - \mu_0)/\text{s.e.}(\bar{x})$ will follow a t-distribution with $(n-1)$ degrees of freedom.

Formal operation of a hypothesis test involves specifying the values of t which will lead us to reject the null hypothesis that $\mu = \mu_0$. The left hand plot below shows a t-distribution with the tail areas corresponding to 2.5% probability highlighted. This leaves 95% of the distribution on the central area. If the test statistic t falls in the highlighted region then this is taken as evidence that the null hypothesis is implausible. In other words, there is significant evidence that the true mean is different from μ_o.

The plots can be reproduced by using the `rp.tables` function from the `rpanel` package.

```
rp.tables(panel = FALSE, distribution = "t", degf1 = 24, observed.value = "2.5",
          tail.probability = "fixed probability")
rp.tables(panel = FALSE, distribution = "t", degf1 = 24, observed.value = "2.5",
          tail.probability = "from observed value")
```



There is an alternative way of quantifying the outcome of the test. We compare the observed value of the test statistic to the reference distribution by computing how much of the distribution is more extreme than the observed value. This is called the *p-value*. If we did not specify in advance that the alternative involved only values below μ_0 or only values above μ_0 then we should measure extremity in both tails of the reference distribution. This is illustrated in the right hand plot above, where we see the p-value is 0.02. As this is less than the conventional threshold of 0.05, we again have significant evidence that the true value of the mean is not μ_0.

2.5 Comparing two means

2.6 Further reading

M., D., and M. (2019)

Hubbard, Haig, and Parsa (2019)

2.7 Exercises

2.7.1 Standard errors

Use the `rp.sample` function to explore the phenomenon that *most of the time, the true parameter and the estimate are within 2 standard errors*. Try setting the true mean and standard deviation to different values to confirm that this priniciple holds.

Chapter 3

Other approaches to inference

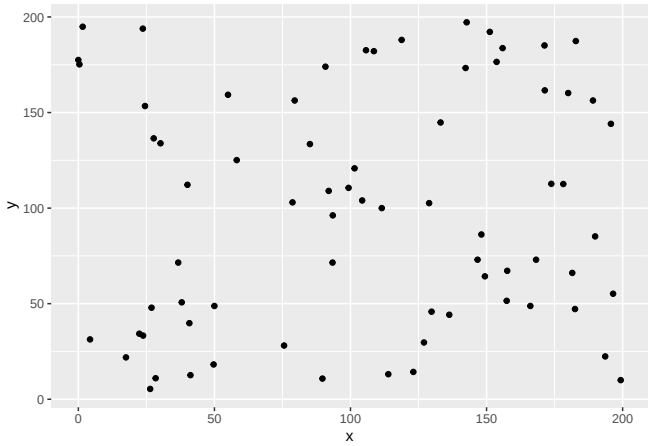
3.1 Computational inference

The problems we have discussed so far have been very simple once where theory can be worked out analytically to provide simple formulate and procedures. With more complex data structures, that can become much more difficult so this section will discuss some ways of approaching inference by computational means. The underlying principles are essentially the same but the method of implementation is different.

3.1.1 Simulation methods

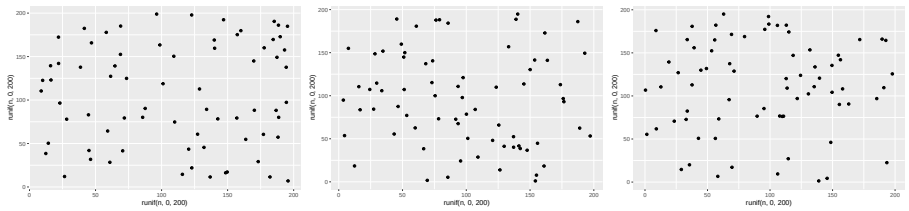
There are some circumstances where the model whose suitability we wish to assess is completely known, without any dependence on unknown parameters. A simple example arises from spatial point patterns where we may wish to assess the evidence that the point pattern is generated by a process which is not simply uniform. The plot below shows data on large (breast height diameter >= 50cm) longleaf pine trees in a 200m x 200m sqaure of forest.

```
library(spatstat)
x <- longleaf$x[longleaf$marks >=50]
y <- longleaf$y[longleaf$marks >=50]
library(ggplot2)
ggplot(mapping = aes(x, y)) + geom_point()
```



We can generate our own trees patterns in a uniform random manner simply by locating each tree at an x and y position each of which is generated uniformly in the range (0, 200). here are some examples of patterns generated in this way.

```
n <- length(x)
for (i in 1:3) {
   plt <- ggplot(mapping = aes(runif(n, 0, 200), runif(n, 0, 200))) +
            geom_point()
  print(plt)
}
```

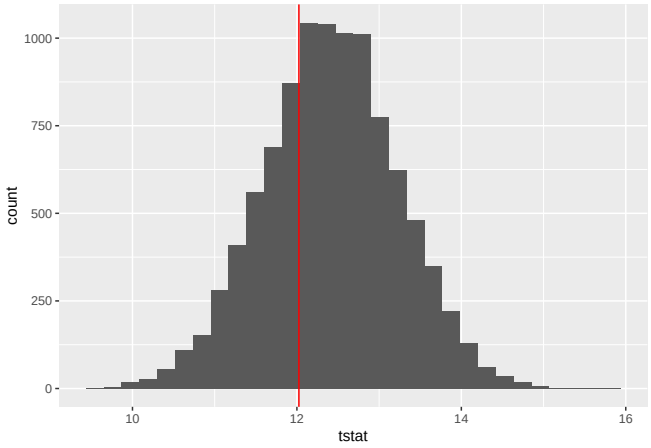


How should we compare the observed pattern with the randomly generated ones? A simple way of doing this is to measure for each tree the distance to its nearest neighbour. If the pattern exhibits clustering, then these distances should be small, while if the pattern is uniform they will tend to be larger. We might then use as our test statistic the average of the distances to each nearest neighbour. We can generate the distribution of this average distance in the case where the locations are uniformly distributed simply by repeatedly simulating locations and retaining the average nearest neighbour distance for each one. The `nndist` function from the `spatstat` package can help us with the nearest neighbour distances.

```
nsim  <- 10000
tstat <- rep(0, nsim)
for (i in 1:nsim) {
  pp <- cbind(runif(n, 0, 200), runif(n, 0, 200))
```

```
  tstat[i] <- mean(nndist(pp[ , 1], pp[ , 2]))
}
tstat.obs <- mean(nndist(x, y))
ggplot(mapping = aes(tstat)) + geom_histogram() +
   geom_vline(xintercept = tstat.obs, col = "red")
length(tstat[tstat < tstat.obs]) / nsim
```

```
## [1] 0.3141
```



The histogram shows the distribution (subject to the variation present through simulation) of the average nearest neighbour distance when of atoms really are uniform. The red line shows the average neighbour distance for the observed data. Informally, but clearly, we can see that from this perspective the observed data is entirely consistent with the assumption of uniformity. An empirical p-value can be computed simply from the proportion of simulated values which fall below the observed one (as we look for evidence against uniformity in small nearest neighbour distances).

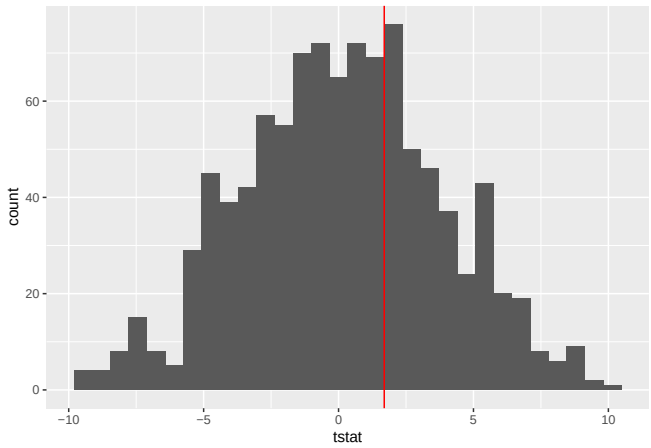
3.1.2 Randomisation tests

Randomisation is very important principle in designing experiments as it helps to overcome issues of bias. We will consider this in detail elsewhere. Randomisation can also be used as the basis for analysis when we are reluctant to make the normality and independence assumptions which underlie standard methods. The dataset below provides a very simple example of this it refers to an experiment to assess the possibly different effects of two fertilisers applied to tomato plants growing in a row. (This example comes from the book *Statistics for Experimenters* by Box, Hunter and Hunter.)

Row position	1	2	3	4	5	6	7	8	9	10	11
Fertiliser	A	A	B	B	A	B	B	B	A	A	B
Yield (pounds)	29.9	11.4	26.6	23.7	25.3	28.5	14.2	17.9	16.5	21.1	24.3

If the positions the fertilisers in the row have been randomised, and if the two fertilisers A and B are in fact equivalent in the effects, then the labels attached to the observations are immaterial. There is no connection between the labels and the outcomes. This allows us to calculate the distribution of the difference between the mean yields of A and B which is generated by the randomisation procedure. The histogram below shows this. For convenience (indeed, laziness!) it has been constructed from random permutations rather than the systematic list of all possible randomisations.

```
fertiliser <- c("A", "A", "B", "B", "A", "B", "B", "B", "A", "A", "B")
yield      <- c(29.9, 11.4, 26.6, 23.7, 25.3, 28.5, 14.2, 17.9, 16.5, 21.1, 24.3)
nperm      <- 1000
tstat      <- rep(0, nperm)
for (i in 1:nperm) {
  smp      <- sample(fertiliser)
  tstat[i] <- diff(tapply(yield, smp, mean))
}
tstat.obs <- diff(tapply(yield, fertiliser, mean))
ggplot(mapping = aes(tstat)) + geom_histogram() +
   geom_vline(xintercept = tstat.obs, col = "red")
```



There are several new functions used in the R code here (`c`, `diff`, `tapply`). You will hopefully feel ready to explore these through the help system by this stage. The end result is a histogram which represents the distribution, induced by randomisation, of the mean sample difference between A and B when the two treatments have identical effects. When we compare this with the observed mean difference it is clear that the two are compatible. So there is no evidence that A and B have different effects. If required, a p-value could be obtained by the means discussed in the simulation test section above. However, the conclusion is already clear.

3.1.3 The bootstrap

We generally regard observed data as having been generated by some underlying distribution. If we construct an estimator of some quantity of interest, such as a population mean, then we need to know how accurate estimate is as a guide to the true value. The simple but surprising idea behind the bootstrap is that we mimic the variation of the sample around the distribution by the variation of resampled datasets around our original sample. When the bootstrap was first proposed in the 1970s, it took people by surprise but in fact it can be shown to be well founded theoretically.

Suppose we are dealing with a distribution which has long tails. In other words, we may find that outliers (unusually large or small observations) are often present in sample datasets. Methods based on the assumption of a normal distribution will be strongly affected by this. A simple device is to compute a *trimmed mean*. This simply computes the main of the remaining data after a proportion of the highest and lowest observations has been removed. This is an attractive way of defending against the effect of long tails, but how do we quantify the uncertainty of our sample estimator as a guide to the true trimmed mean of the underlying distribution?

The bootstrap takes the observed sample and resamples the data with replacement. A trimmed mean is calculated on each resample. In the simplest approach to construction of a confidence interval, this is generated simply through percentiles of the trimmed means generated by resampling. The example below shows the mechanics of this, using a simple normal random sample to represent the observed data.

```
x      <- rnorm(25)
tm.obs <- mean(x, trim = 0.1)

nboot  <- 10000
tmdiff <- rep(0, nboot)
for (i in 1:nboot)
   tmdiff[i] <- mean(sample(x, replace = TRUE), trim = 0.1) - tm.obs
tm.obs - quantile(tmdiff, c(0.975, 0.025))
```

```
##      97.5%       2.5%
## -0.6245322  0.1898431
```

3.2 Bayesian inference

The methods which have been described in this course have a very long and well established history. They are in widespread use and have brought insight to an enormous array of applications across the board. From a philosophical point of view these methods are referred to as *frequentist*. The name comes from the idea of continued sampling of data, under the same conditions, for which it can be

shown that relative frequencies converge eventually to probabilities.

An alternative approach to inference is to consider the uncertainty associated with quantities of interest such as parameters, expressed not only through a model for the observed data but also through a prior distribution for the unknown parameters. Sometimes that prior information can be informative, based on knowledge or data about the situation being modelled. Sometimes the prior information can be vague or uninformative. Bayesian thinking allows uncertainty to be expressed as a combination of these two things, in a posterior distribution. A major step forward took place in the 1990s when computational methods of deriving posterior distributions were developed. For the first time this allowed complex problems to be handled by Bayesian models, with the attractive property that uncertainty at different levels can be propagated to generate posterior uncertainty about the quantities of interest, incorporating all the appropriate sources of uncertainty. This has led to extremely powerful tools for complex systems.

Treatment of this topic is beyond the scope of the current course.

3.3 Exercises

3.3.1 Coverage property of bootstrap confidence intervals

During the session an example of a confidence interval for a trimmed mean was discussed. Write some code to evaluate the *coverage* of this method of constructing a confidence interval. You can do this by repeatedly simulating sample from the true distribution (`rnorm(25)`), constructing a bootstrap confidence interval using the code discussed, and counting how many times the calculated confidence interval covers the true value (0).

Chapter 4

Inference with categorical data

So far, we have considered data measured on a continuous scale but there are, of course, many other types of data structure. Here we will deal with data in the form of categories.

4.1 Simple proportions

We often encounter problems where:

- the number of items in the sample, denoted by n, is fixed in advance;
- there are two possible outcomes for each item (yes/no, success/failure, etc.);
- each item has the same chance of producing a 'success', independently of all other items.

This leads to the *Binomial* model which describes the probabilities of the number of 'successes' out of the n items, when the probability of success for a single item is p.

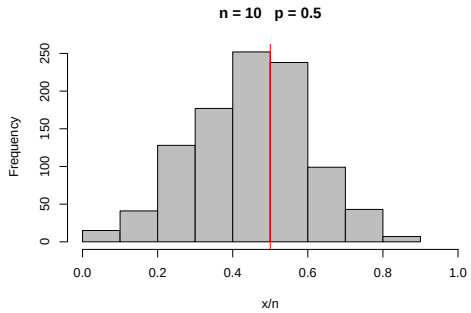
If the number of 'successes' in the sample of size n is x then a natural estimate of p is the *sample proportion*, x/n. We write:

$$\hat{p} = x/n$$

We need to quantify the uncertainty associated with estimating p by $\hat{p}$. As usual, the *standard error* does this for us.

The plot below uses simulation to show the variation in $\hat{p}$ when samples of size n are repeatedly drawn from populations where the true proportion if p. You might like to experiment with this code tp see the effectsd of changing `n` and `p`.

```
n <- 10
p <- 0.5
x <- rbinom(1000, n, p)
hist(x / n, col = 'grey', xlim = c(0, 1), main = paste('n =', n, '  p =', p))
abline(v = p, col = 'red')
```



There are two simple points to note:

- The sample proportion is subject to error but it is centred on the true proportion in the population.
- The size of the error in the sample proportion decreases as the size of the sample increases.

In fact it is possible to show that the variance of $\hat{p}$ is $p(1-p)/n$. We can estimate the unknown p in this expression to obtain the standard error

$$s.e.(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$$

As in other cases, we should find that most of the sample proportions lie within two standard errors of the true proportion. In this case we do not have an exact result for a confidence interval, as we do with the Normal model, but an approximate 95% confidence interval for p is easily obtained as

$$\hat{p} \pm 2\text{s.e.}(\hat{p})$$

In a briefing note for journalists, the UK Parliament provides a helpful guide to opinion polls. (The document was prepared by Peter Kellner for the British Polling Council.) This refers to an accuracy of 3% in the main percentages reported in a poll based on 1000 respondents. Where does 3% come from?

When the sample size is 1000, the standard error of a proportion is $\sqrt{\hat{p}(1-\hat{p})/1000}$. The largest possible value of $\hat{p}(1-\hat{p})$ is when $\hat{p} = 0.5$, so the upper limit on the standard error is $\sqrt{0.5 \times 0.5/1000} = 0.16$. Two standard errors is then 0.032, or around 3%.

This is a useful guide, but there are many other aspects of this to consider. We are often interested in comparing the proportions from different categories, such as support for the main parties in an election, and the uncertainty will be

increased when two proportions are involved. In addition, there are all the usual issues about the extent to which the sampling is genuinely random or subject to bias of various kinds. Nonetheless, the ability of standard error to help in quantifying uncertainty is helpful.

4.2 Comparing proportions

Example: Smoking and lung cancer
In a famous historical study of the association between smoking and lung cancer, Doll & Hill compared the numbers of smokers and non-smokers in samples of lung cancer patients and controls. The data for females are shown below.

	cases	controls
smokers	41	28
non-smokers	19	32

Is there evidence of a link between smoking and lung cancer?

Details of how these data were collected are given in the paper. There are interesting questions here about what constitutes an appropriate control group. In fact other hospital patients, not suffering from lung cancer, were used.

The principal question of interest is whether the proportion of smokers among the cases is different from the proportion of smokers among the controls. We denote the underlying true proportion among the cases and controls by p_1 and p_2 respectively, with corresponding sample sizes n_1 and n_2. We can estimate the true proportions by the sample proportions,

$$\hat{p}_1 = 41/60 = 0.683$$
$$\hat{p}_1 = 28/60 = 0.467$$

We can also calculate the standard error of each sample proportion as

$$\begin{aligned} \text{se}_1 &= \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1} = \sqrt{0.683 \times 0.317/60} = 0.060 \\ \text{se}_2 &= \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1} = \sqrt{0.467 \times 0.533/60} = 0.064 \end{aligned}$$

However, it is the *difference* between the two groups which is of interest to us. We have a natural estimate in the differences of the proportions $p_1 - p_2$ in the difference of the sample proportions

$$\hat{p}_1 - \hat{p}_2 = 0.683 - 0.467 = 0.216.$$

We can also calculate the standard error of this difference by combing the individual standard errors, as follows:

$$\text{se}_{\text{difference}} = \sqrt{\text{se}_1^2 + \text{se}_2^2}$$

Notice that the squared standard errors are added together, despite the fact that the estimates of the proportions are being subtracted. This is because we are measuring the uncertainty involved and so the uncertainty of the difference combines the uncertainties of the individual components. With the present data this gives

$$\text{se}_{\text{difference}} = \sqrt{0.060^2 + 0.064^2} = 0.088$$

A 95% confidence interval for the difference in proportions is then:

$$\begin{aligned} &0.216 \pm 2 \times 0.088 \\ \text{i.e.} &0.216 \pm 0.176 \\ \text{i.e.} &(0.040, 0.392) \end{aligned}$$

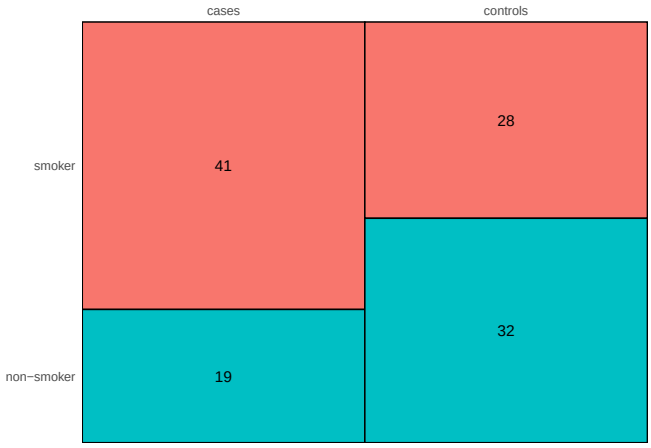
Since this confidence interval does not contain 0, we therefore have clear evidence that the proportions of smokers in the cases and control groups are different.

4.3 Contingency tables

The data on smoking and lung cancer can also be treated as a simple example of a *contingency table*, which cross-classifies counts by two different factors. In fact, this was how the data were viewed in the original paper by Doll & Hill. The method of analysis we will explore can be implemented in contingency tables with any number of rows or columns.

As ever, a helpful first step is to visualise the data, even when this consists of a very simple tabluation. The mosiacplot discussed earlier helps with this. The columns of the plot refer to the case and control groups. Here these are equal in size (60) but differences in in numbers would have been reflected in the width of the columns. This means that the height of each block now refers to proprtions of observations within each column.

```
x <- matrix(c(41, 19, 28, 32), ncol = 2,
            dimnames = list(c("smoker", "non-smoker"),
                            c("cases", "controls")))
rp.contingency(x)
```



If there is no association between smoking and lung cancer, then the proportions associated with each column will be identical. We can use this idea to calculate *expected values*, which describe the pattern we expect to see if the null hypothesis of no association is correct. Estimates of the common probabilities for each column are $(69/120, 51/120) = (0.575, 0.425)$. The expected values by row are therefore obtained by multiplying the column totals by these probabilities. It so happens that the column totals are identical in this dataset, namely 60.

```
60 * 0.575, 60 * 0.575 = 34.5, 34.5
60 * 0.425, 60 * 0.425 = 25.5, 25.5
```

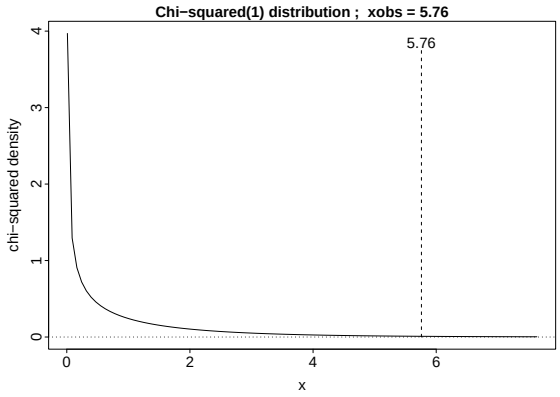
We can now compare this table of expected values (E_{ij}) with the table of observed values (O_{ij}) above. We do this through a quantity known as the chi-squared statistic, defined as

$$\sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where the subscripts i and j index the rows and columns. The chi-squared statistic for the current dataset is 5.76.

This value is meaningful only when we compare it to a reference distribution. The theory for this setting tells us that the relevant comparison is with a χ^2 distribution, which is plotted below. This distribution is indexed by a parameter known as the *degrees of freedom*. For contingency tables with r rows and c columns, the degrees of freedom should be set to $(r-1)(c-1)$, which in the current case is 1.

```
library(rpanel)
rp.tables(panel = FALSE, distribution = "chi-squared", degf1 = 1, observed.value = 5.76)
```



The observed value of the test statistic, which is also shown in the plot, is much higher than values we expect to see from this reference distribution. The upper 5% point of the distribution is 3.84, which gives us a specific benchmark. We therefore have significant evidence that the proportions of smokers are different in the cases and controls. This formn of analysis has, unsurprisingly, confirmed the conclusions of the comparison of proportions in the previous section.

The chi-squared test can be easily implemented in `R` through the `chisq.test` function. By default this includes a correction for 2x2 tables in order to improve the accuracy of the reference distribution. However, the conclusion is unchanged.

```
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  x
## X-squared = 4.9105, df = 1, p-value = 0.02669
```

4.4 Exercises

4.4.1 Opinion polls

Opionion polls reported in newspapers are sometimes treated with scepticism. This is particularly true of polls published before general elections! The *Independent* newspaper ran a series of articles on this some time ago. In particular, it was claimed that,in *'a sample of* 1000 *people, the main percentages should be accurate ... to within three percent, nineteen times out of twenty'.*

Is this correct? How can it be justified? Use the formula for the standard error of a proportion to think this through.

4.4.2 Respiratory disease in infancy

Holland, Bailey & Bland (1978, *Journal of Epidemiology and Community Health*) reported on a study of the effects of bronchitis in infancy on the occurrence of respiratory problems later in life. The following table reports the occurrence of these events in a sample which was studied by the authors.

	Cough at 14	No cough at 14
Bronchitis at 5	26	247
No bronchitis at 5	44	1002

Is there evidence of a link?

4.4.3 Smoking and lung cancer

The earlier analysis used a chi-squared test to assess the evidence that the proportion of smokers is different in cases and controls. Construct a confidence interval for the difference of these proportions as an alternative approach, making sure that you understand how to interpret the interval you produce.

4.4.4 Confidence intervals and p-values

In the two examples above, both confidence intervals and hypothesis tests have now been used. These should be equivalent in terms of the evidence for the presence of an effect. What are the relative merits of the confidence interval and hypothesis testing approaches in this context?

Bhatt, Arun. 2010. "Evolution of Clinical Research: A History Before and Beyond James Lind." *Perspectives in Clinical Research* 1 (1): 6.

Bowman, A. W. 1994. "Teaching by Design." *Teaching Statistics* 16: 2–4. https://doi.org/10.1111/j.1467-9639.1994.tb00670.x.

Diggle, Peter J, and Amanda Chetwynd. 2011. *Statistics and Scientific Method: An Introduction for Students and Researchers*. Oxford University Press.

Hubbard, Raymond, Brian D Haig, and Rahul A Parsa. 2019. "The Limited Role of Formal Statistical Inference in Scientific Inference." *The American Statistician* 73 (sup1): 91–98.

M., Diez D., Barr C. D., and Çetinkaya-Rundel M. 2019. *OpenIntro Statistics*. 4th edition. openintro.org/os.

MacKay, R Jock, and R Wayne Oldford. 2000. "Scientific Method, Statistical Method and the Speed of Light." *Statistical Science*, 254–78. https://www.jstor.org/stable/2676665.

Marshall, Geoffrey, J. W. S. Blacklock, C. Cameron, N. B. Capon, R. Cruickshank andJ. H. Gaddum, F. R. G. Heaf, A. Bradford Hill, et al. 1948. "Streptomycin Treatment of Pulmonary Tuberculosis." *British Medical Journal*, 769–82.

Rosling, Hans. 2018. *Factfulness*. London: Hodder & Stoughton.