Statistical Modelling

A conceptual, visual and practical introduction

true

20 August, 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.

Part I

Inference: turning data into evidence

Chapter 1

Inference on means

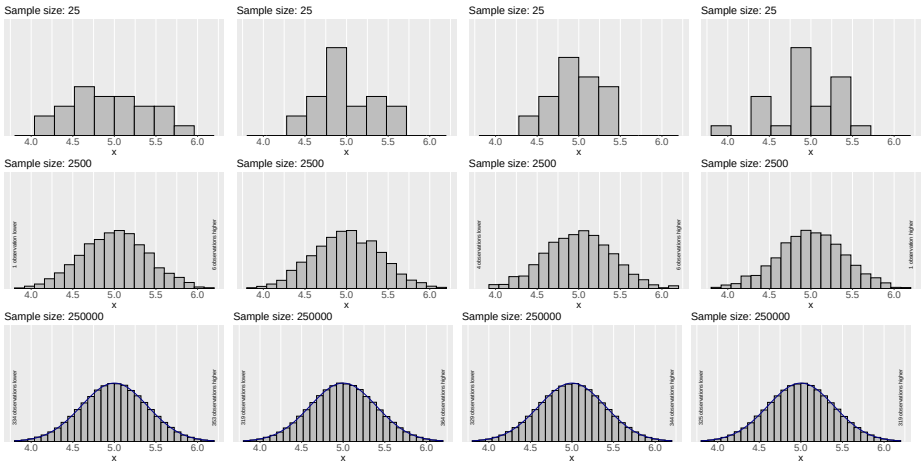
So far we have considered how to read, organise and visualise datasets. We are now ready to discuss one of the main themes of statistics, namely weighing up the evidence for the presence, characteristics or size of different features in the population lying behind the data. That process is referred to 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.

1.1 Samples and populations

The distinction between a *sample* of data and the *population* from which the data come was mentioned briefly in Section **??**, 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. So we are dealing with measurements on a continuous scale.

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.

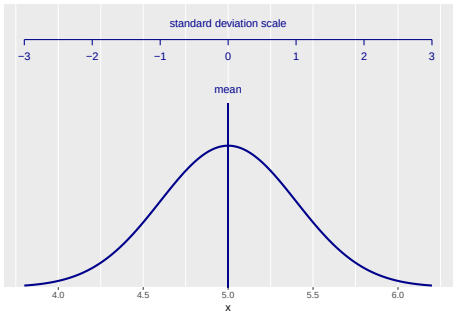
The first row of plots has sample size 25. Here there is a lot of variability in the shapes of the histograms. Sometimes there appear to be two clusters of data, or a rather skewed shape. We know that the underlying population is the same for all samples but the detailed shapes we see can be quite different.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

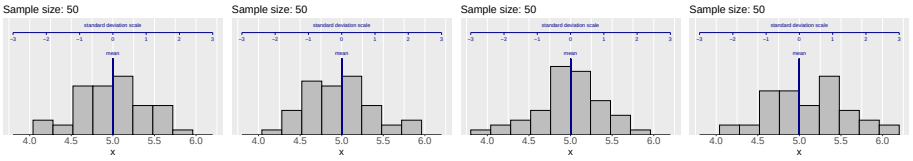
The continuous curve is what the histograms 'converge' to as the sample size gets higher and higher and variability gets smaller and smaller (and the histogram bin width also get smaller and smaller). This is known as the *density function* and it defines the population from which we are sampling. For 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.



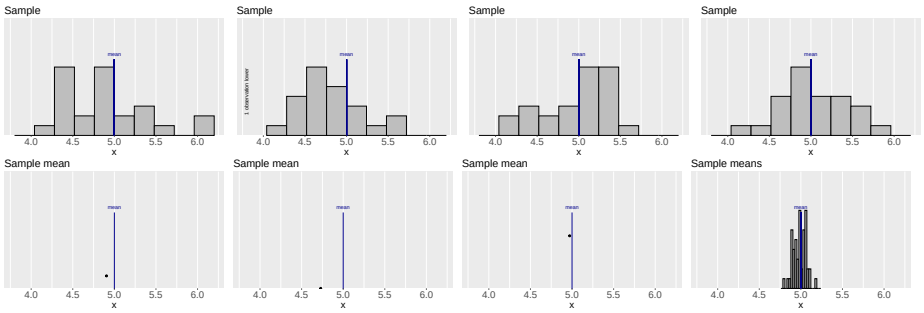
This makes sense, as the standard deviation quantifies how far away observations are from the mean, on average. So, informally, one standard deviation away from the mean, on either side, takes us 'halfway' into the spread of the data. From that perspective, it then seems reasonable that two standard 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 1.7 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.

1.2 Quantifying uncertainty: standard error

The setup of the 'thought experiment' of the previous section was necessarily rather artificial, 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 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. If we denote the population standard deviation by σ and the size of the sample by n, then a theoretical calculation tells that the standard deviation of the sample mean is

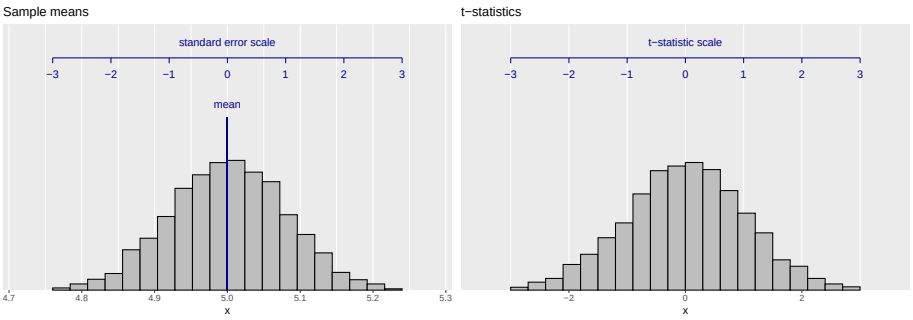
$$\text{se}(\bar{x}) = \sigma/\sqrt{n}.$$

This gives us a measure of uncertainty, or inaccuracy, of the sample mean $\bar{x}$ as an estimate of the true mean μ. In a change of terminology which reflects the fact that we are no longer dealing with the uncertainty of individual observations but of a feature of the underlying population, we refer to this as the *standard error*.

The left hand plot below shows a large collection of sample means along with a standard error scale. We can see that the sample means mostly lie within two standard errors of the true mean. This is a very important phenomenon which we should pause to highlight.

```
                                    Most of the time,
the sample mean is within 2 standard errors of the population mean.
```

We will exploit to this to create some very useful inferential tools. In fact, as we will see later, this principle is a much more general one which applies in many different settings involving a population *parameter* and an *estimate* of it.

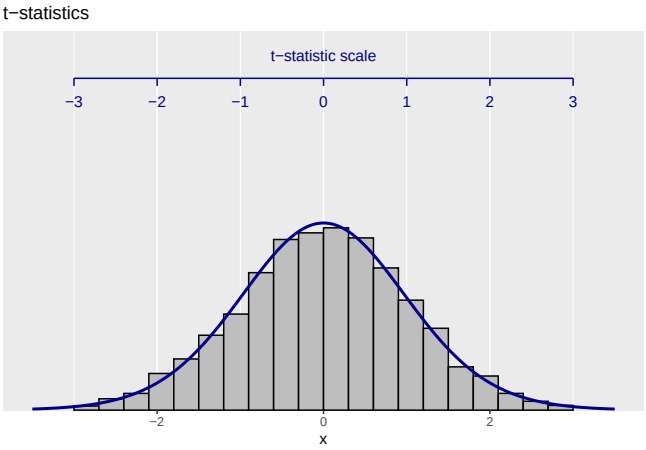


One issue remains, however, as the standard error involves the standard deviation of the population which, usually, we do not know. It is natural to estimate this by the sample standard deviation, s. We can use `rp.sample` to investigate the effect of this. As the standard deviation of the data changes with every sample, it is easier to investigate this in terms of the *t-statistic*

$$\frac{\bar{x} - \mu}{se(\bar{x})} = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

where $\bar{x}$ and s denote the sample mean and standard deviation and μ denotes the population mean. We can think of this as measuring the distance between the population and sample means in units of (estimated) standard error. The right hand plot below shows the t-statistics produced by `rp.sample`. Reassuringly, these generally lie within ± 2 which confirms that that principle highlighted above still applies. (From now on, we will generally assume that the term 'standard error' implies the estimation of any unknown parameters in the standard error formula.)

This gives a very helpful guide to the accuracy of the sample mean as a guide to the population mean but a more precise guide is available and so we should use this. Theoretical calculations can tell us about the complete distribution of the sample mean. When we are sampling from a normal population, it turns out that the sample mean also has a normal distribution, with mean μ and the standard error shown above. When we use the estimated standard error, the t-statistic then becomes the focus and it turns out to have a *t-distribution*. The plot below superimposes the density function for this distribution on the t-statistics generated by `rp.sample`. The shape is very similar to the normal distribution with the main difference being that the t-distribution usually has slightly thicker tails. (This is due to the additional variation which arises from using the sample standard deviation in the expression for the standard error.) The t-distribution has a parameter called the *degrees of freedom* and in this setting that takes the value $n - 1$.

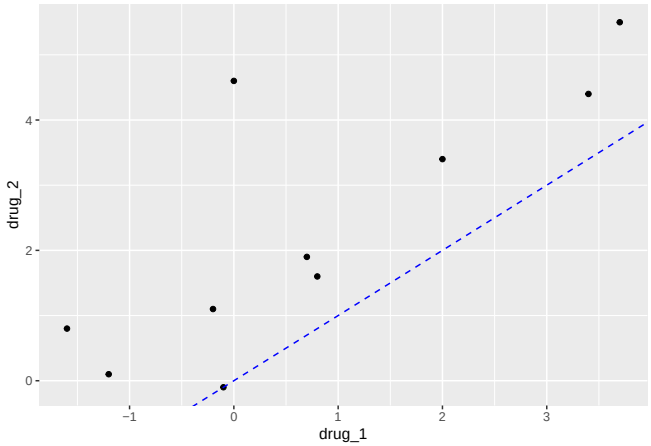


These 'thought experiments' have enabled us to consider some of the concepts involved when we aim to learn about a population parameter, such as a mean, from a sample of data. This places us in an good position to interpret what individual samples are able to tell us, using tools developed in the following sections.

1.3 Confidence intervals

One of the earliest people to think about the issue of quantifying the uncertainty of a sample mean as a guide to the true mean 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)
```



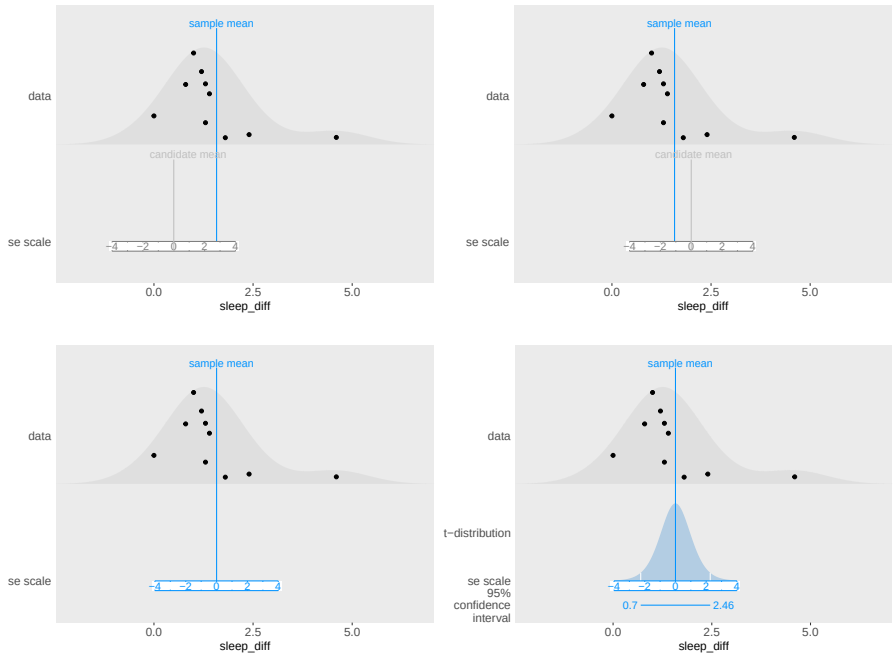
The extra times for drug 2 are higher than those for drug 1, apart from one patient where the times were identical. In one patient the difference between the two drugs is particularly large but otherwise the differences are broadly similar. In particular, there is no obvious change in the size of the 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.

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 proporties 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*.

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



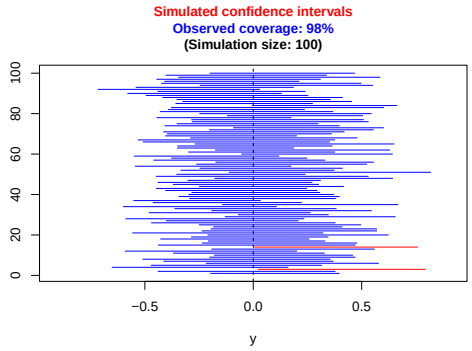
Emphasise that the true mean is fixed while the sample mean is subject to sampling variation.

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

Discuss the construction of a more accurate confidence interval.

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%.



1.4 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')
```

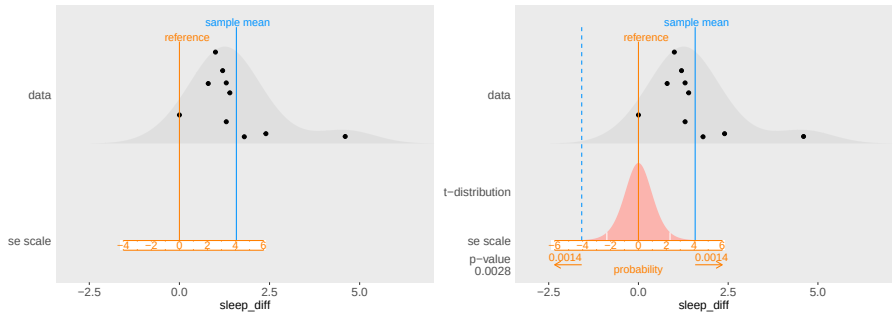
We view things through the eyes of the null hypothesis. If this is true then, most of the time, $\bar{x}$ and μ_0 will be within 2 standard errors of one another. If we define the *test statistic* to be

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

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.

```
rp.t_test(sleep_diff, mu = 0)
```



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.

1.5 Comparing two means

Fisher et al. (1925) - see books and papers - used a small agricultural experiment to illustrate the two-sample t-test. Number of tillers (independent lateral branches) which grow on barley in two expereimental groups - electrified and caged.

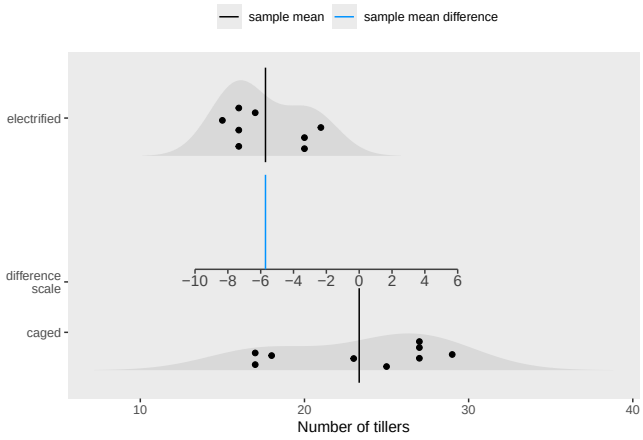
Comment on small sample sizes - so msall that it can easily be typed into the code!

Agriculture, and Fisher, were key historical settings.

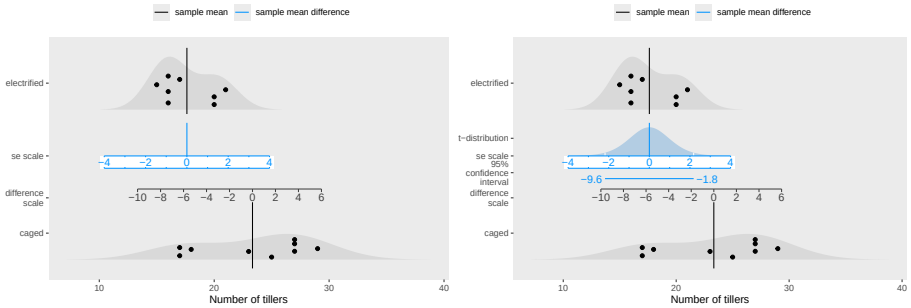
Discuss the standard error as the combination of the se for each group, as well as the pooled variance case.

```
electrified <- c(16, 16, 20, 16, 20, 17, 15, 21)
caged       <- c(17, 27, 18, 25, 27, 29, 27, 23, 17)
```

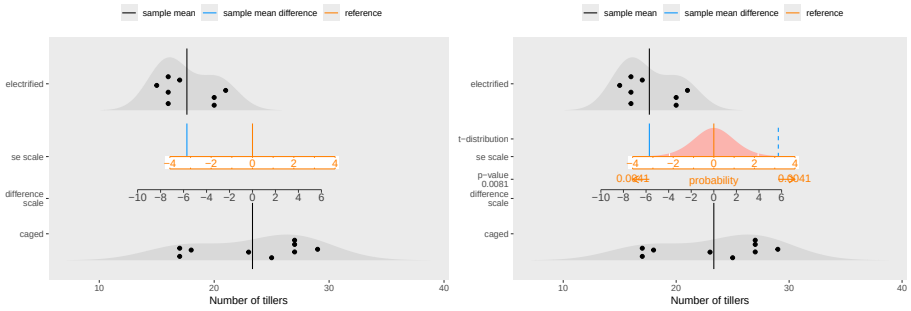
```
set.seed(64318)
rp.t_test(electrified, caged, vlab = 'Number of tillers')
```



The more accurate distribution.



Hypothesis test perspective.




```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -3.1753, df = 11.847, p-value = 0.008109
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.630924 -1.785743
```

```
## sample estimates:
## mean of x mean of y
##  17.62500  23.33333
```

1.6 Further reading

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

Hubbard, Haig, and Parsa (2019)

1.7 Exercises

1.7.1 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 2

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.

2.1 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.

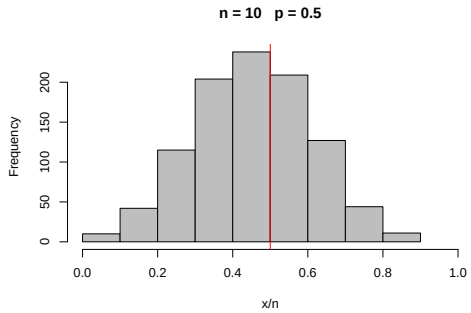
If the number of 'successes' in the sample of size n is x then a natural estimate of p is the *sample proportion*, x/n. We write:

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

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.

2.2 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:

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

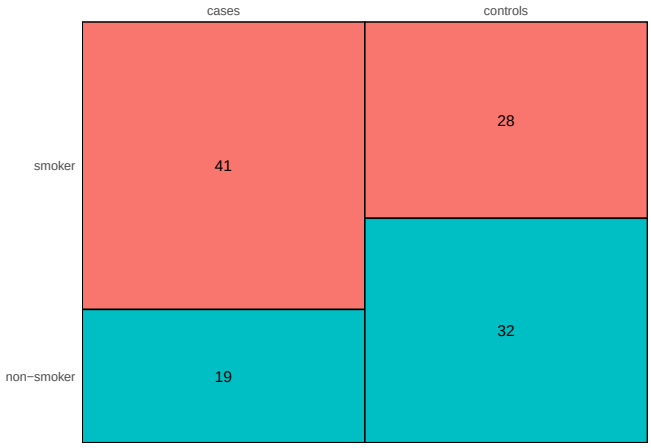
Since this confidence interval does not contain 0, we therefore have clear evidence that the proportions of smokers in the cases and control groups are different.

2.3 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
```

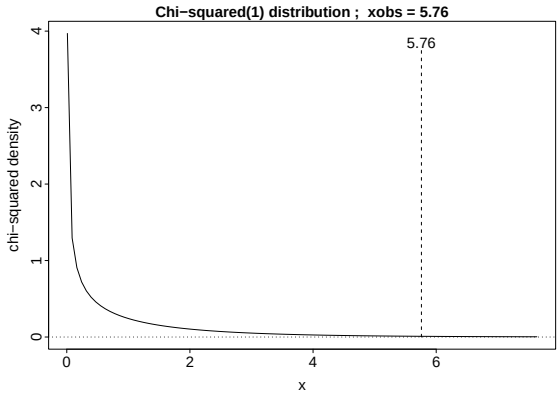
We can now compare this table of expected values (E_{ij}) with the table of observed values (O_{ij}) above. We do this through a quantity known as the chi-squared statistic, defined as

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

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
```

2.4 Exercises

2.4.1 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.

2.4.2 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?

2.4.3 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.

2.4.4 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?

Chapter 3

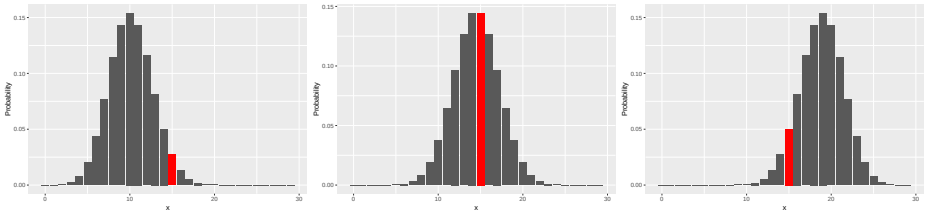
Likelihood: a powerful principle

3.1 The idea

Explain that this principle shows us how to construct estimators.

The idea of likelihood is a very powerful one. To start gently, consider the single observation at ethylene chloride concentration 35.55 which killed 15 beetles out of a group of 29, giving the proportion 0.517. The binomial distribution describes all the possible values that we might observe, and the probabilities of observing them, for a particular `size` of group and with a particular probability (`prob`) of success (death in the present case) for each beetle. A formula for the binomial distribution can be derived by the rules of probability but we can use R to compute these conveniently, using the `dbinom(x, size, prob)` function. The plots below show all the outcomes (0 to 29) and their probabilities when `size = 29` and with `prob` set to 0.35, 0.50 and 0.65. The code is shown only for the case when `prob` is 0.35.

```
ggplot() + geom_col(aes(0:29, dbinom(0:29, 29, 0.35))) +
   geom_col(aes(15, dbinom(15, 29, 0.35)), fill = 'red') +
   labs(x = 'x', y = 'Probability')
```

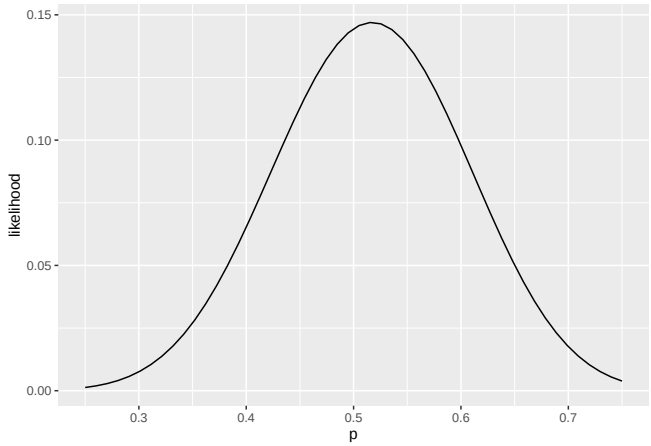


The idea of likelihood turns this perspective around. Instead of specifying the

parameter of the model, namely the probability of success, and then considering the consequences in terms of the observations we may see, likelihood starts with the observed data and considers what parameter values might be consistent with that. In the plots above, the probability associated with the observed value of 15 is highlighted in red. If we assume that the true parameter is 0.35 then the probability of observing an outcome of 15 is very low and the same is true if we assume the true parameter is 0.65. The probability of observing 15 is much higher when we assume the true parameter to be 0.5. This gives a means of quantifying the relative plausibility of different parameter values.

To use more scientific notation, the binomial probability for any outcome k for the number killed can be written as $pr(k; m, p)$, where pr is a (mathematical) function which gives the probability of outcome k when the group size is m and the probability of success for an individual beetle is p. So pr is a function of k for fixed values of m and p. The likelihood function is simply $pr(k; m, p)$ viewed as a function of p, with k fixed at the observed value. (The parameter m is fixed by the design of the experiment.) To emphasise this change in perspective, the likelihood function is written as $L(p; k, m)$ to indicate that k is fixed and the likelihood is a function of p.

```
p    <- seq(0.25, 0.75, length = 50)
dfrm <- data.frame(p, likelihood = dbinom(15, 29, p))
ggplot(dfrm, aes(p, likelihood)) + geom_line()
```



The plot above shows the likelihood function. The *maximum likelihood estimate* of p is, as the name suggests, the value of p which maximises the likelihood function. Inspection of the plot shows this to be just above 0.5. In fact it is 0.517, the sample proportion of beetles killed. It is not a surprise that a good estimate of the proportion of beetles killed in the population at this concentration is the proportion of beetles killed in the sample. However, it is impressive that this has resulted simply by employing the concept of the likelihood function, without building in any further information. This strategy has supplied an obvious

estimate in this simple case but it can also provide estimates in much more complex cases where it is not at all clear what might constitute a good estimate.

3.2 Inference

3.3 An example

Chapter 4

Other approaches to inference

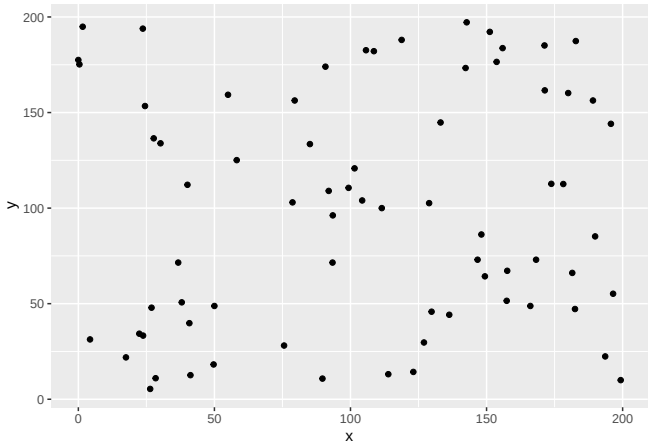
4.1 Computational inference

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.

4.1.1 Simulation methods

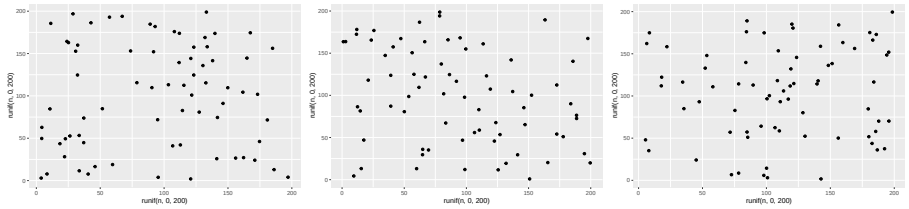
There are some circumstances where the model whose suitability we wish to assess is completely known, without any dependence on unknown parameters. A simple example arises from spatial point patterns where we may wish to assess the evidence that the point pattern is generated by a process which is not 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)
}
```

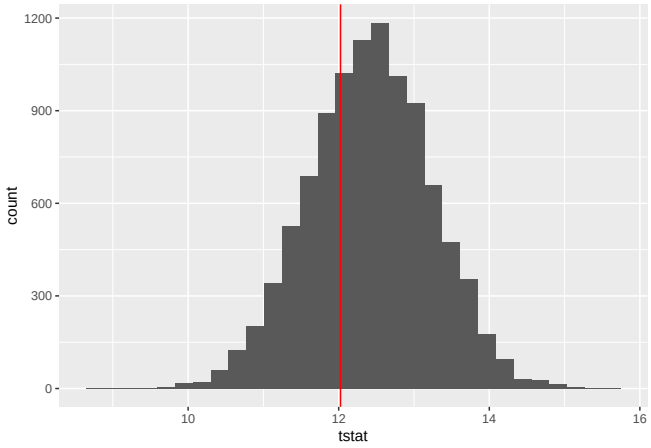


How should we compare the observed pattern with the randomly generated ones? A simple way of doing this is to measure for each tree the distance to its nearest neighbour. If the pattern exhibits clustering, then these distances should be small, while if the pattern is uniform they will tend to be larger. 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] 0.319
```



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).

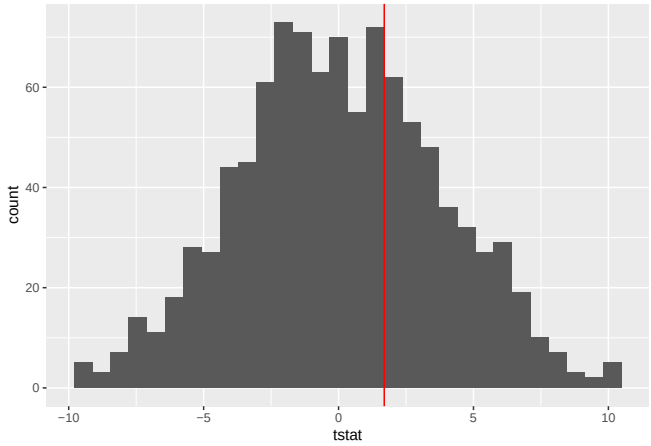
4.1.2 Randomisation tests

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.

4.1.3 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%
## -0.4127958  0.4288787
```

4.2 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.

4.3 Exercises

4.3.1 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).

Fisher, Ronald Aylmer et al. 1925. "Applications of 'Student's' Distribution." *Metron* 5 (3): 90–104.

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.