Clinical Trials 4H

Rachel Oughton

2024-03-05

Contents

Welcome to Clinical Trials 4H!

This page contains the notes for Clinical Trials IV. As we progress through the course, more will appear. You can also download the PDF version (see the icon at the top left). If you notice any typos, mistakes or places that are unclear, please do let me know!

Practical details

Lectures

Our lectures are 12 noon on Mondays and 9am on Wednesdays, all in ES231. This is in the Earth Sciences / Arthur Holmes building, which is behind (almost surrounding) the Calman Learning Centre. The door of this room is actually labelled 'Teaching Room 4'.

Computer classes

We have two 2-hour practicals for this module. They are 11am - 1pm on the Fridays of weeks 14 and 19 (2nd February and 8th March). These classes are in RH-0003. This is Rowan House, which is in Upper Mountjoy, across the pond from the MCS building.

Office Hour

The office hour will be every Monday, 3-4pm, in MCS 2091 (this is a meeting room, not my office). If you walk up the stairs, keep straight on past the main maths office and computer room (MCS 2094), turn right and it's on your right through the double doors. Alternatively there are maps in the main areas!

Assessment

This module is assessed through two equally weighted pieces of coursework. The first will be assigned on Wednesday 7th February, the second on Wednesday 13th March.

There will also be some formative assignments throughout the course. More details on these to follow.

Books

The main reference for the first half of the course is Matthews (2006). There are a couple of copies in the Bill Bryson Library. Some other books we will make use of are Hulley et al. (2013), Hayes and Moulton (2017). You shouldn't need to use any of these books, but of course you're welcome to if you want to read further.

What to expect from this module

Clinical Trials IV is somewhat different from the majority of statistics modules, because

- It is more focussed on application than on methodology
- It is assessed purely through coursework.

This means that your experience of it might be different from what you're used to

- We will cover quite a lot of different statistical methods (drawing on most of the 1H and 2H courses, and some 3H!) but not in great depth
- There is no pressure to memorize anything - indeed, if you really were a trial statistician, you would definitely have access to the internet, various textbooks and even these notes (should they prove useful!).
- There is an emphasis on understanding which method we use and why, and what it means. Hopefully this has been the case in some of your other modules too!

What I expect from you

Because we will be covering quite a lot of different areas within statistics, there may be some things that you haven't seen before (or can't remember very well). I will try my best to explain them as clearly as I can, but there isn't time to go into the nuts and bolts of everything we come across. Therefore, if you do feel a bit rusty on some area, you may need to read up on that a bit, so that you're happy with it. I am very happy to suggest resources from time to time, and you're welcome to come to the office hour to talk about such things.

This is the first year this course is running, and so I would also really appreciate your feedback. I may not be able to address everything (or I may only be able to implement things for following years), but if I can act on it quickly then I will!

Chapter 1

Introduction to Clinical Trials

A clinical trial is an experiment, usually performed on human subjects, to test the effect of some sort of treatment or intervention. We may also use the term **Randomised controlled trial** (RCT). These are not fully the same thing; a clinical trial may not have been randomised, for example if it follows a pre-determined cohort through some sort of process. Likewise, an RCT may not be clinical, but instead may be about an intervention in some other setting like agriculture or education. For this module, we are really focussing on RCTs, and almost all of our examples will be clinical.

For the purposes of this module, a clinical trial will have two groups:

1. The **treatment group** or **intervention group**: this group of people will be subject to the new treatment.
2. The **control group**: this group of people will be subject to the status quo - the 'standard' or most widely used treatment path for their cohort (sometimes this is no treatment).

These groups are usually, though not always, of the same size. Which group each patient is assigned to is usually decided by randomization, which is something we will go on to explore in later lectures. In reality, trials can have more than two groups, and many statistical methods extend quite naturally to this.

The goal of the trial is to estimate the **treatment effect**: is the treatment better than the control, and if so, how much? This short description raises lots of statistical issues, which will take up the next few weeks!

Before we get into the theory, we'll think about some of the background to clinical trials, and introduce some key ideas.

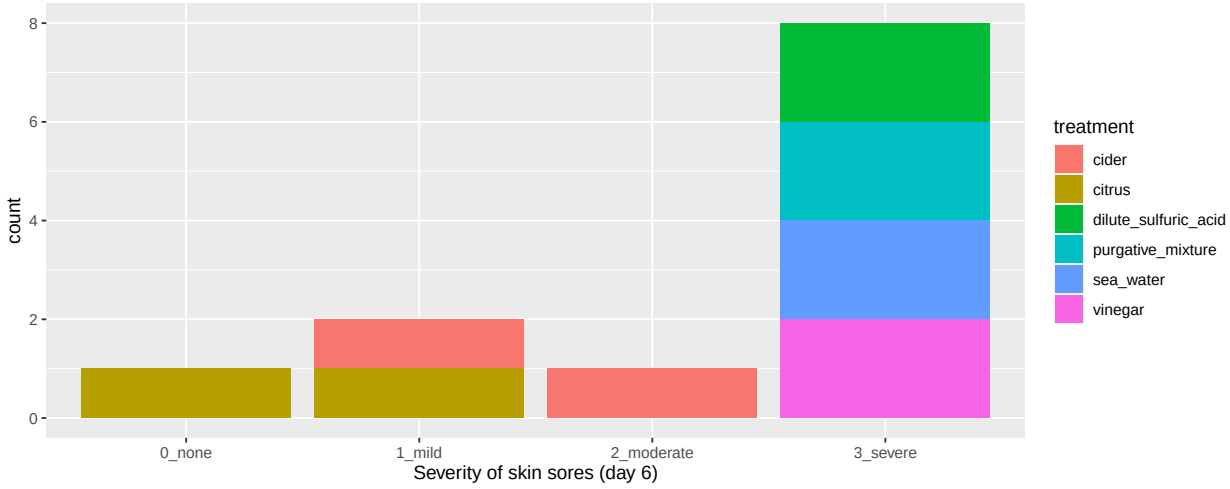
Put (very!) simply, the goal of a clinical trial is to determine what works to make people better. Although clinical trials as we know them now have only been around since the Second World War, similar sorts of experiments can be seen from much longer ago. If you're interested in learning about the evolution of clinical trials from Biblical times to now, the James Lind Library has some fascinating resources and articles.

Example 1.1. Scurvy (James Lind, 1757) Scurvy was a serious disease, particularly affecting seamen on long voyages. Symptoms were unpleasant (mouth sores, skin lesions etc.) and it could often be fatal. Lind was the ship's surgeon on board the HMS Salisbury, and had several patients with scurvy. Many remedies were proposed and in popular use at the time (with only anecdotal evidence, if any, to support them). In 1757 Lind decided to test six such treatments, on two patients each:

- cider
- dilute sulfuric acid
- vinegar
- sea water
- citrus (oranges and lemons)

- purgative mixture (a paste of garlic, mustard seed, horseradish, balsam of Peru, and gum myrrh)

Lind chose twelve seamen with similar severity of symptoms, and subjected them to their assigned treatment for 6 days. They were kept in the same quarters, and fed the same diet apart from their treatment. Unsurprisingly (to us!) "The most sudden and visible good effects were perceived from the use of oranges and lemons,"

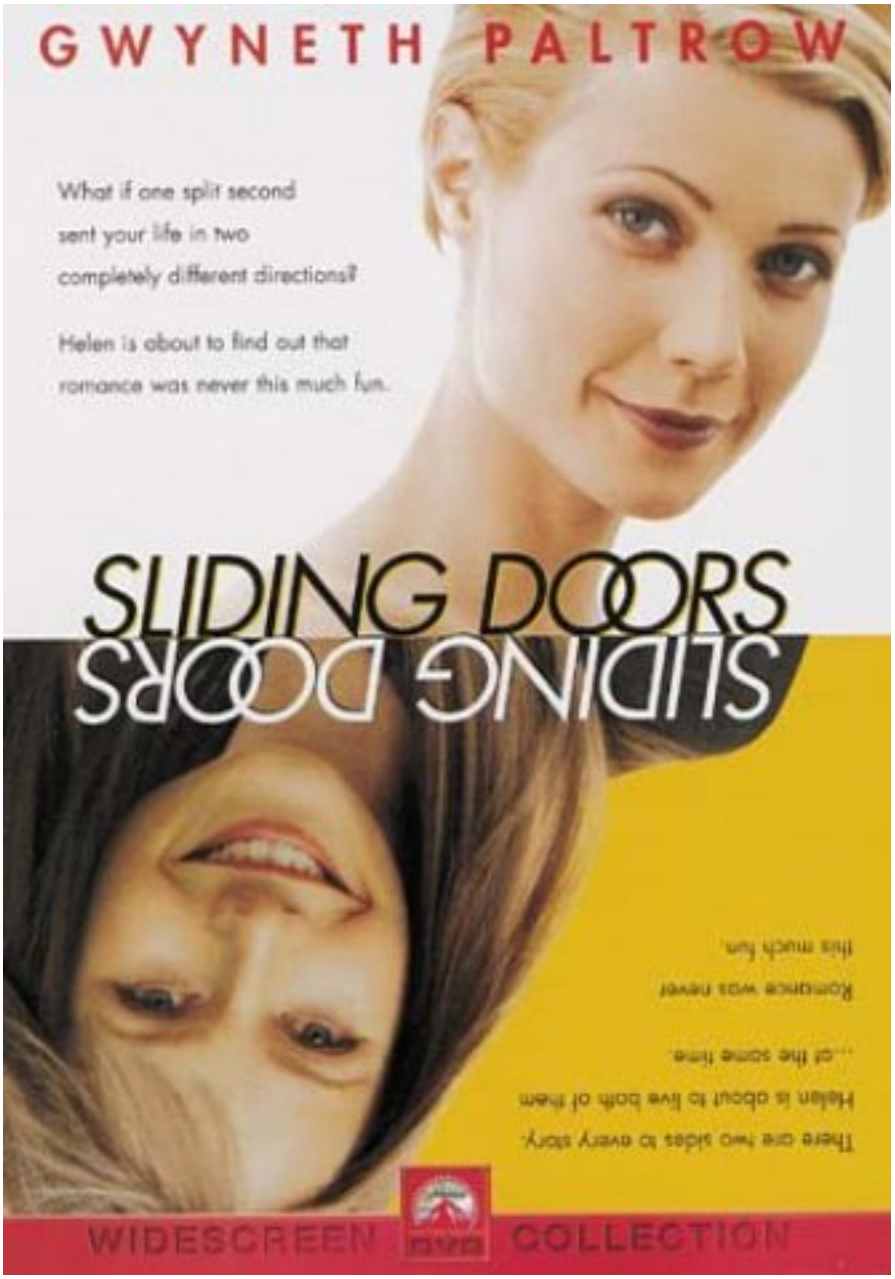


A key thing to notice about the Scurvy example is that Lind went to great lengths to ensure that the treatment was the only thing affecting these 12 sailors: they all started with a similar severity of symptoms, they were kept in the same place and their diet was identical apart from their treatment. This links to one of the foundational principles of clinical trials: causal inference.

1.1 Causal inference and clinical trials

You're probably familiar with the mantra that **"correlation does not imply causation"**: just because two things are correlated, it doesn't mean we can conclude that one causes the other. If you're not convinced, here are some humorous (and slightly macabre) examples. Causal inference is concerned with the design and analysis of data for uncovering causal relationships.

This is important for us, because we really want to be able to conclude that a treatment works (or doesn't) - that it *causes* recovery, or a reduction in symptoms, or helps the patient in some way. If we were experimental scientists in some laboratory, we could conduct some controlled experiment in which everything was kept under very specific conditions, and could fairly easily make conclusions about the treatment we were testing, and how it behaved in a range of conditions. However, testing treatments on real people is different: we don't have several identical versions of the same person to test the treatment on, and even if we did, as Gwyneth Paltrow shows us it doesn't take very much to completely alter the conditions of someone's existence!



Neither can we just base our conclusion of whether a treatment works on lab-based tests or theory (although undoubtedly these will both play a part in developing the treatment in the first place). The treatment needs to be tested on actual people.

Because, as we noted, people are all different, and living different lives (and unlike James Lind we can't force them all to live in the same part of a ship and eat the same food!) we will need to test the treatment on lots of people in order to gather empirical evidence. This is why statistics is so important in the design and analysis of clinical trials. The results of the trial must concluded beyond reasonable doubt, and must be able to be generalized to as-yet-untreated patients. We want to avoid any spurious correlations that are down to chance, or to associations we haven't taken into account. For example, what if the two seamen given citrus were also much younger and generally healthier than the other ten? Maybe they would have recovered quickly anyway? Or what if another treatment was actually much better than citrus, but just happened to have been given to two sailors who had some other pre-existing illness, causing them to suffer much worse with scurvy?

Clinical trials are therefore crucial for modern medicine, and statistics is crucial to clinical trials. But why exactly are clinical trials given this position of importance? Do we really have to do things this way?

1.2 The structure of a clinical trial

In a clinical trial, people are grouped and subdivided in various ways.

The population of eligible patients

One of the first steps in conducting a trial is to specify exactly what sort of person you want to test the treatment on, and where these people will be found. They may be of a certain sex and/or age range, they may have (or definitely not have) certain conditions. They may suffer from some particular symptom, or be at a particular stage of an illness.

A clear set of criteria is key to consistency. Patients are usually recruited as they present (eg. to hospital or a GP centre) and may be being recruited over several years, or by several different clinicians, so it is important that everyone is sticking to the same plan.

Example 1.2. In a study by Hjalmas and Hellstrom (1998) of the use of desmopressin in children with nocturnal enuresis (bed-wetting), children had to be aged 6 - 12 with a history of PMNE (primary monosymptomatic nocturnal enuresis) and no organic pathology (no disease that alters the structure or function of organs). The children had to be free of other urinary problems (such as frequency, urgency or daytime incontinence) and not to have received any treatment for nocturnal enuresis during the 2 months before entering the trial. Children with clinically significant endocrine, metabolic, hepatic, psychiatric, neurological, musculoskeletal, cardiovascular, haematological, renal or genitourinary disease were excluded from the trial.

Knowing exactly what type of patients were recruited into the trial is also key when generalizing the results to the population. If the trial recruited males aged 55-70, we cannot confidently conclude that the results will apply to a female aged 26.

Entry to the trial

The group of patients recruited will be some subset of the possible population. Patients are allowed to refuse consent to take part, or individual patients may be judged unsuitable despite meeting the criteria. Knowing how many patients to recruit is a statistical question, which we will deal with soon.

Allocation to groups

These patients are then allocated to receive either the treatment, or to be part of the control group (or more, if there are more than two groups). These groups are often referred to as the **trial arms** - the treatment arm and the control arm. Deciding which patients should be allocated to which group is another statistical question. Once the patients have been allocated, they will receive the treatment (or not) and important measurements will be taken during the trial period.

Comparing results

Now that the trial has been run, we have two sets of measurements: one for the treatment group and one for the control group. But guess what?! Comparing these and coming to a conclusion about the effect of the treatment is a statistical question.

Why bother with a control group?

Surely if we want to see whether a treatment works, we should just give it to a patient and see if they get better? Why do we need to also have a group of people not receiving the treatment?

In rare and extreme cases, this is a decent strategy: if a disease has always been fatal, but we start giving patients the treatment and some live, that is pretty solid evidence that the treatment works. This was the case with tuberculous meningitis, until the introduction of Streptomycin in 1944.

This was also the case when Edward Jenner tested cowpox as a vaccination for the fatal disease smallpox. After observing that milkmaids, once they had suffered from the mild condition cowpox (which they did often), seemed to be immune to smallpox, Jenner tested his theory by injecting an 8 year old boy called James Phipps with fluid from a milkmaid's cowpox lesions (yum). Once the boy's cowpox infection had run its course, he injected him again, this time with matter from a fresh smallpox lesion. Thankfully, James Phipps did not contract smallpox. After several more successful such tests, and a gradual shift in attitudes to the idea of vaccination (a word coined by Jenner, from the latin 'vaccinia', meaning cowpox) Jenner's results were published and vaccination became commonplace. Clearly, injecting people with smallpox who had not been given the cowpox innoculation would be very cruel (they would almost certainly die) and would prove nothing; there was already plenty of evidence for the fatality of smallpox.

However, most diseases have a fairly uncertain and variable trajectory. If we give a group of patients the treatment, we can't know what would have happened to them if they hadn't received the treatment, or had received a different treatment. Comparing them to patients the past is dodgy because lots of other things may have changed since even the recent past. This is why we have a *concurrent control group* (usually known as just the *control group*). These patients do not receive the new treatment, but instead carry on as usual. The aim is to make the control and treatment groups as similar as possible in all other respects (especially those we deem important) so that at the end we can attribute the difference between the two groups to the treatment.

1.3 The primary outcome

In a clinical trial, there are usually many measurements performed on patients, and possibly at various different points throughout the trial. However, for the sake of the analysis, we usually determine one to be the **primary outcome variable**. The research questions should be phrased in terms of this variable, and the goal of our design should be to be able to answer questions about this variable.

Example 1.3. In a trial by Villar et al. (2020) investigating the use of Dexamethasone treatment for acute respiratory distress syndrome , the primary outcome was the 'number of ventilator free days up to 28 days', while other outcomes included 'all-cause mortality after 60 days' and 'incidence of infections in ICU'.

1.4 Ethical issues

Clinical trials differ from most scientific experiments in that they are experimenting on people. This means that the team designing, conducting and analysing the trial have various ethical responsibilities. This is a huge area; we will touch on it from time to time but will not go into anywhere near enough detail! Some key things to note though are....

- A patient must never be given a treatment that is known to be inferior.
- Patients must be fully informed about the trial, the treatments used, possible adverse effects and side-effects and so on. Patients should only be recruited into the trial if, after being given all this information (and having had it communicated clearly and at an appropriate level) they give their consent.

- After entering a trial, a patient has the right to withdraw at any point, and should then receive whatever treatment is most appropriate for them. They should not face any negative repurcussions for withdrawing.

The patients' interests are safeguarded by the Declaration of Helsinki. This statement is implemented differently by different countries. In the UK, health authorities each have their own ethics committee, by which proposals for experiments involving human subjects must be approved.

You might think that these ethical issues largely concern the clinicians, and that we statisticians don't need to worry too much about the ethics of clinical trials. After all, we are likely never to meet any patients or to get our hands dirty in any way! But as we will see, at each stage the choices made by the statistician can in fact have serious ethical implications.

1.5 Phases of clinical trials

If you read about clinical trials (or hear about them in the news), you'll hear talk of a 'phase 3 trial' or similar. Broadly speaking, clinical trials follow a progression from phase one (or sometimes zero) to phase four. These phases apply to most countries, and any for any drug to be licensed multinationally it must get through phase III.

Phase zero

The first step is to test a low dose of the treatment on a small number of people, to check that it isn't harmful. The dose is too low to have any medicinal effect, but is designed to verify that the drug behaves as expected from laboratory studies, and doesn't have any harmful effects. There may only be 10-20 participants, and there is no randomisation (or control group).

Phase one

Phase one trials are also quite small (around 20-50 participants) and are designed to find the best dose of the treatment and what any side effects are. Phase one trials tend to be recruited very slowly: a small group will be recruited onto a low dose and monitored closely. If all goes well, another small group will be recruited on a slightly higher dose, and so on. This is known as a *dose escalation study*. Participants at this phase are monitored very closely, for example through regular blood tests and recording daily symptoms.

Phase two

If the drug makes it through the phase one trial, it can progress to phase two (often written as 'phase II'). These involve more people than phase one, possibly up to 100. The cohort may now be restricted to people with a particular version of a condition (eg. a particular type of cancer), but it may still be broader than the sorts of trials we will be looking at. Now the aim is to find out if the new treatment works well enough to progress to a large phase three (phase III) trial:

- Exactly what conditions (or versions of a condition) does this treatment work for?
- What are the side effects and can they be managed?
- What is the best dose to administer?

Phase II trials sometimes compare the treatment to a placebo, and sometimes use randomisation to group participants.

This is the stage at which most drugs fail, for a multitude of reasons (cost, safety, efficacy,...).

Phase three

Phase III trials are much bigger, often involving hundreds or thousands of participants, and aim to compare the new treatment to the best currently available treatment (which may be another treatment, or may be nothing). Side effects are still monitored, as some of the rarer ones may not show themselves at the smaller phases, because there are fewer participants.

In phase III trials, the aim is to find out if the new treatment is better, and if so by how much. Phase III trials almost always use randomisation to allocate participants to groups, and go to great lengths to make the trial as reliable as possible, for example using a placebo for the control group (who aren't getting the real treatment) that looks identical to the real drug. These are the sorts of trials we will mainly be concerned with in this course.

To be licensed, a treatment has to get through phase III and be found to be effective (and of course safe).

Phase four

Phase IV trials happen after the treatment has been found to work, has been licensed and is in use. The aims of phase IV trials are

- to find out more about the rarer side effects
- to investigate the long term risks and benefits
- to find out how well the treatment works when given to a broader group of people than in phase III.

Part I

Part I: Continuous outcome variables

Chapter 2

Sample size for a normally distributed primary outcome variable

For most of this module, we'll focus on randomised controlled trials (RCTs). These have mainly been used for clinical applications (for example, to test a particular drug), but have also recently become popular ways to test interventions in areas such as education and policing.

Having laid the groundwork in Chapter 1, we now go on to some more technical details. In this Chapter, we focus on the 'vanilla' scenario, where we have a trial with two arms, and our unit of randomization is individuals. At first we will focus only on continuous outcomes, but in later weeks we will go on to think about binary and time-to-event data.

Broadly speaking, the topics we cover fall into the categories of 'before the trial' (design and planning) or 'after the trial' (analysis), although as we'll see there is some interaction between these stages.

The first big question asked of a trial statistician is usually how many participants does the trial need in order to be viable: the sample size. We will clarify what is meant by 'viable' later in this section.

Broadly speaking, there are two (opposing) ethical issues around sample size:

1. If we don't recruit enough patients, then we may not gather enough evidence to draw any conclusion about the research question (eg. whether there is a treatment effect). As well as being scientifically disappointing, this is unethical. To conduct the trial, some of the patients will have been subject to an inferior treatment (assuming one treatment was actually better), and if there is no conclusion then this was effectively for no purpose.
2. If we recruit too many patients (ie. we would be sufficiently likely to reach a conclusion with many fewer) then we have subjected more patients than necessary to an inferior treatment, and possibly also taken up more time and resources than was necessary.

Therefore it is important to think about this issue carefully. We've framed the question in quite a woolly way so far, but now we'll start to think more carefully.

2.1 The treatment effect

In Section 1.3 we discussed the need to settle on a **primary outcome variable**. One reason this is important is that we base our sample size calculations on the primary outcome variable.

Definition 2.1. Suppose our primary outcome variable is X, which has mean μ in the control group and mean $\mu + \tau$ in the treatment group. The variable τ is the **treatment effect**. The goal of our RCT is to learn about τ. The larger τ is (in magnitude), the more pronounced the effect of the intervention.

This problem is usually framed as a **hypothesis test**, where the null hypothesis is that $\tau = 0$.

2.2 Reminder: hypothesis tests (with a focus on RCTs)

When performing a hypothesis test, what we are aiming to find is the **P-value**.

Definition 2.2. The **P-value** is the probability of obtaining a result as extreme or more extreme (ie. further away from the null hypothesis value) than the one obtained *given that the null hypothesis is true.*

Put simply, the p-value is a measure of the probability of obtaining whatever result (eg. treatment effect) we have have found simply by random chance, when in fact there is no treatment effect (ie. $\tau = 0$). Generally, a P-value of $\alpha = 0.05$ is accepted as sufficient evidence to reject the null hypothesis, although in clinical settings it can often be smaller (eg. $\alpha = 0.01$). It is conventional to present the P-value by simply saying whether it is smaller than some threshold (often 0.05), rather than giving the exact value.

Definition 2.3. The threshold for the p-value below which the results are considered 'significant' is known as the **significance level** of the test, and is generally written α (as above).

This use of a significance level is (in part) a legacy from early days when computers were rare and values were looked up in t-tables (or similar). Now that it is very simple to find the exact P-value, it is becoming more and more common to report the actual number. Indeed, there is a big difference between $p = 0.049$ and $p = 0.000049$.

2.2.1 Insignificant results

If our P-value is relatively large, say 0.3 or 0.5 (or really, greater than α), then our result is not at all unlikely (or sufficiently unlikely) under the null hypothesis, and provides insufficient evidence to reject H_0. However, it is not inconsistent with the existence of a treatment effect, so we don't say there is evidence to accept H_0. One can imagine that if the true treatment effect τ were tiny, many trials would fail to find evidence to reject H_0. However, if our sample size were sufficiently large, we should be able to detect it. Conversely, if τ is very large, even a relatively small sample size is likely to provide enough evidence to reject H_0.

A non-significant P-value means that our results are consistent with the null hypothesis $\tau = 0$, but they are also consistent with some small treatment effect, and therefore we can't conclude very much. The key issue is, what size of treatment effect do we care about? We must ensure that our sample size is sufficiently large to be sufficiently likely to detect a clinically meaningful treatment effect.

We are being vague for now, but this is a key issue in determining an appropriate sample size.

2.2.2 One-tailed or two-tailed?

It is highly likely that the scientists running the trial will have a strong idea of the likely 'direction' of the treatment effect. Assuming that a larger value of the primary outcome variable X is good, they will expect a positive value of the treatment effect τ (or be prepared to accept a possible value of zero for no effect).

It would therefore be tempting to perform a one-sided test, with

$$\begin{aligned} H_0 &: \tau = 0 \\ H_1 &: \tau > 0. \end{aligned}$$

For example, suppose our test statistic t has a t distribution with 31 degrees of freedom and we obtain a value of 2, as shown in Figure 2.1. In this case our P-value is $1 - F_t\left(2, df = 31\right) = 0.0272$ (where $F_t\left(\cdot\right)$ is

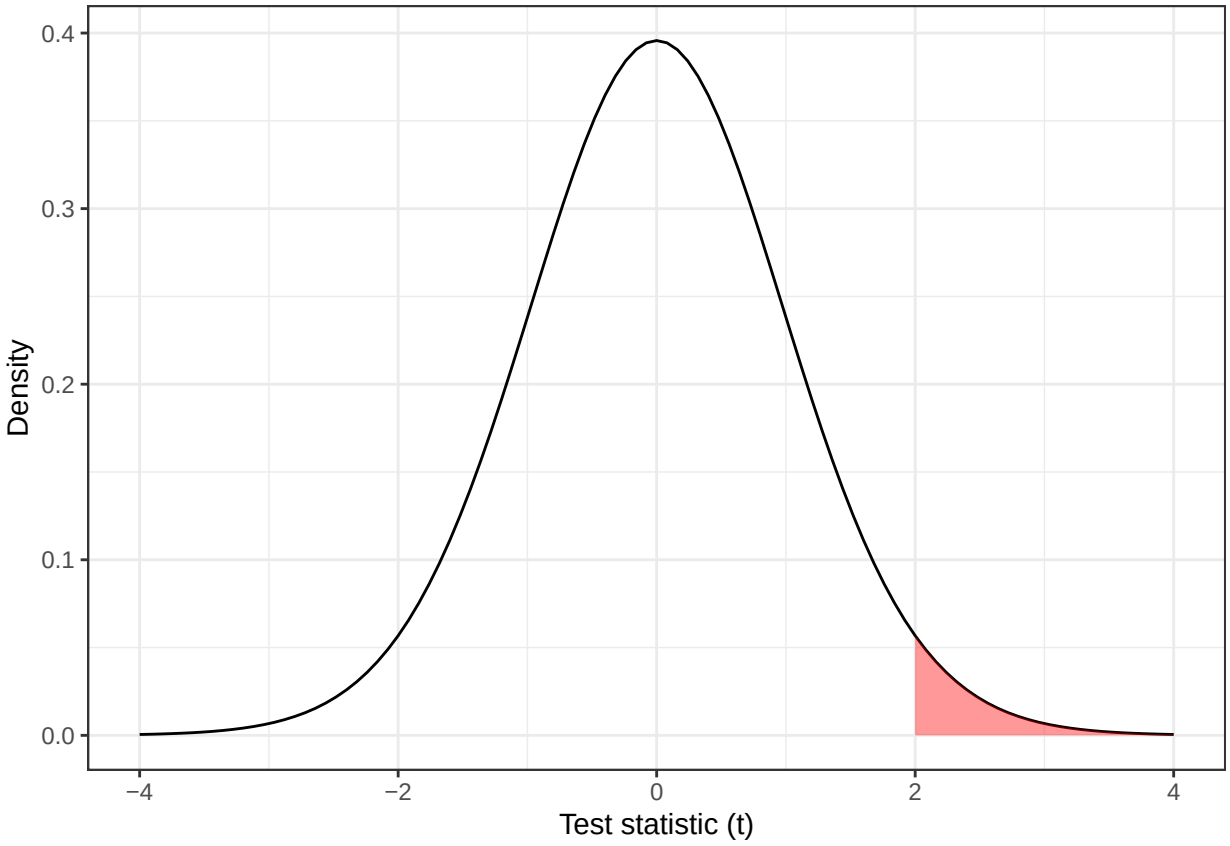


Figure 2.1: The distribution t_{31}, with the area corresponding to $t > 2$ shaded.

the cumulative distribution function of the t distribution) , and the result would be considered significant at the 0.05 level.

For a large positive value of t, we obtain a small P-value, and reject H_0, concluding that the intervention is effective (in a good way). However, what if we obtain a large negative value of t? In this one-sided set-up, there is no value of $t < 0$ that would give a significant result; negative values of t are simply considered consistent with H_0, and there is no mechanism to conclude that an intervention has a significantly negative effect.

For this reason, we always conduct two sided hypothesis tests, with

$$H_0 : \tau = 0$$
$$H_1 : \tau \neq 0.$$

In this scenario, Figure 2.1 is replaced by the plot shown in Figure 2.2, where values of t with $t < -2$ are considered 'equivalent' to those with $t > 2$, in the sense of how unlikely they are under H_0.

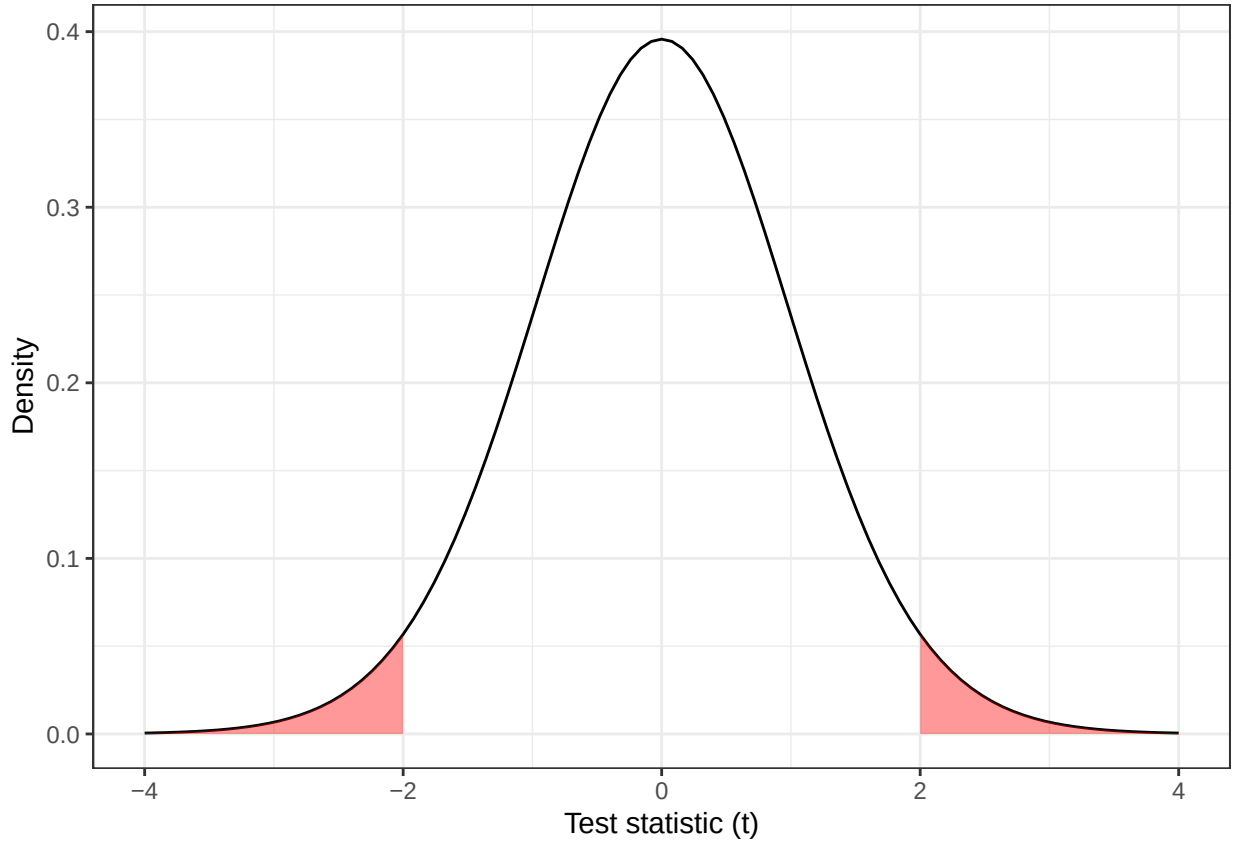


Figure 2.2: The distribution t_{31}, with the area corresponding to $|t| > 2$ shaded.

The P-value for the two-sided test as shown in Figure 2.2 is

$$F(-2, df = 31) + [1 - F(2, df = 31)] = 2 \times 0.0272 = 0.0543$$

and the result is no longer significant at the 0.05 level. Throughout this course, we will always assume two-tailed tests.

Figure 2.3: A rather ubiquitous two-tailed mermaid

2.3 Constructing a measure of effect size

Let's say we are recruiting participants into two groups: group T will be given the new treatment (they will sometimes be referred to as the *treatment group* or *treatment arm*) and group C will be given the control (they are the *control group* or *control arm*).

Suppose that we have n patients in group C, and m in group T. The primary outcome variable X is normally distributed with mean μ in group C (the control group) and mean $\mu+\tau$ in group T (the intervention group), and common standard deviation σ. So

$$X \sim N\left(\mu, \sigma^2\right) \text{ in group } C$$
$$X \sim N\left(\mu+\tau, \sigma^2\right) \text{ in group } T.$$

We are testing the null hypothesis $H_0 : \tau = 0$ against the alternative hypothesis $H_1 : \tau \neq 0$.

Using the data obtained in the trial, we will be able to obtain sample means $\bar{x}_C$ and $\bar{x}_T$ from each group, and a pooled estimate of the standard deviation

$$s = \sqrt{\frac{(n-1)\, s_C^2 + (m-1)s_T^2}{n+m-2}},$$

where s_C and s_T are the sample standard deviations for groups C and T respectively, for example

$$s_C = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}_C\right)^2}{n-1}}.$$

Using these values we can compute

$$D = \frac{\bar{x}_T - \bar{x}_C}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

as a standardised measure of the effect τ.

Theorem 2.1. *Under H_0, D has a t-distribution with $n+m-2$ degrees of freedom.*

Proof. Under H_0 the x_i are iid $N\left(\mu,\, \sigma^2\right)$, and so

$$\bar{x}_C \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$
$$\bar{x}_T \sim N\left(\mu, \frac{\sigma^2}{m}\right)$$

and therefore

$$\bar{x}_T - \bar{x}_C \sim N\left(0, \sigma^2\left[\frac{1}{n}+\frac{1}{m}\right]\right)$$

and

$$\frac{\bar{x}_T - \bar{x}_C}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1).$$

We know that for $x_1, \ldots, x_n, \sim N\left(\mu, \sigma^2\right)$ for some arbitrary μ and σ^2,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sim \chi^2_{n-1},$$

and so we have

$$\begin{aligned}\frac{n-1}{\sigma^2}s_C^2 &\sim \chi^2_{n-1}\\ \frac{m-1}{\sigma^2}s_T^2 &\sim \chi^2_{m-1}\\ &\text{and}\\ \frac{1}{\sigma^2}\left[\left(n-1\right)s_C^2 + \left(m-1\right)s_T^2\right] &= \frac{n+m-2}{\sigma^2}s^2\\ &\sim \chi^2_{n+m-2}.\end{aligned}$$

The definition of a t-distribution is that if $Z \sim N\left(0,1\right)$ and $Y \sim \chi^2_n$ then

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}} \sim t_n,$$

that is X has a t distribution with n degrees of freedom.

Plugging in our $N\left(0,1\right)$ variable for Z and our χ^2_{n+m-2} variable for Y, we have

$$\begin{aligned}\frac{\frac{\bar{x}_T - \bar{x}_C}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\left(\frac{n+m-2}{\sigma^2}s^2\right)\Big/\left(n+m-2\right)}} &= \frac{\bar{x}_T - \bar{x}_C}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\Bigg/\frac{s}{\sigma}\\ &= \frac{\bar{x}_T - \bar{x}_C}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}\\ &= D\end{aligned}$$

and therefore D has a t distribution with $n+m-2$ degrees of freedom. □

We can therefore use D as our test statistic; if D is such that

$$|D| > t_{n+m-2}\left(\alpha/2\right)$$

where $t_{n+m-2}\left(\cdot\right)$ is the function such that $P\left(T > t_{df}\left(\xi\right)\right) = \xi$ when $T \sim t_{df}$ then we can reject H_0.

In practical terms, for more than around 40 degrees of freedom, the t distribution is indistinguishable from the normal distribution, and since it is rare to have fewer than 40 participants in an RCT, we use a normal approximation in what follows, and a difference is significant at the $100\left(1-\alpha\right)\%$ level if $|D| > z_{\alpha/2}$, where z are standard normal quantile values. For example, for $\alpha = 0.05$ have $z_{\alpha/2} = 1.960$, since the probability of a standard normal variable exceeding this value is 0.025.

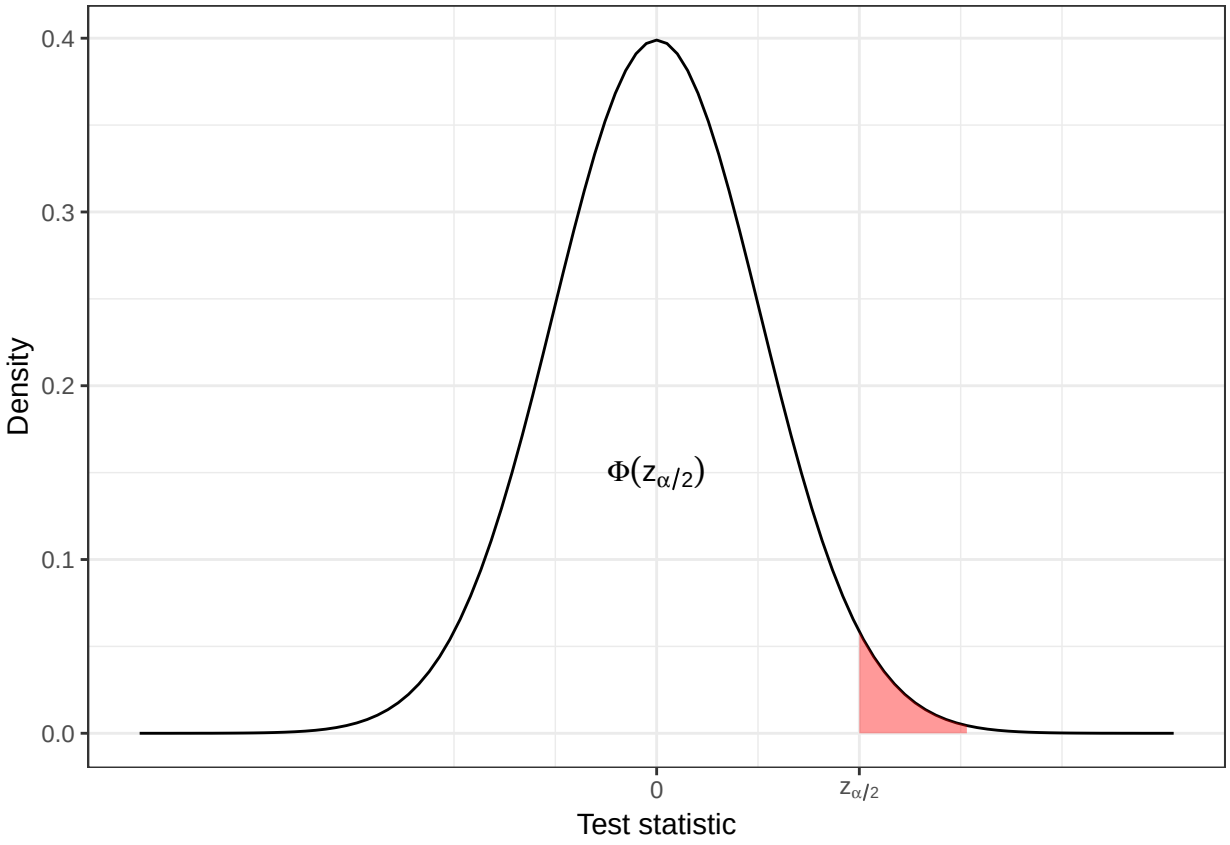
So, if we have run a trial, and have obtained n values of X from group C and m values of X from group T, we can compute D. If D lies outside the interval $\left[-z_{\alpha/2}, z_{\alpha/2}\right]$ then we reject H_0.

This is equivalent to $\bar{x}_T - \bar{x}_C$ falling outside the interval

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

Brief aside on notation

We'll see a lot of the notation $z_{\alpha/2}$ and similar, so to clarify:



In R, we have $\Phi\left(z_{\alpha/2}\right) = \texttt{pnorm}\left(z_{\alpha/2}\right)$ and $z_{\alpha/2} = \texttt{qnorm}\left(\Phi\left(z_{\alpha/2}\right)\right)$. `qnorm` is the quantile and `pnorm` is the cumulative distribution function. So, for example

$$\frac{\alpha}{2} = 1 - \Phi\left(z_{\alpha/2}\right)$$

We have constructed our whole argument under the assumption that H_0 is true, and that the probability of such a value is therefore α. We want this probability to be small, since it constitutes an error; H_0 is true, but our value of D (or the difference in means) leads us to reject H_0. This is sometimes called the 'type I' error rate. But what if H_0 is false?

2.4 Power: If H_0 is false

We have constructed things so that if H_0 is true, we have a small probability of rejecting H_0. But if H_0 is false, and $\tau \neq 0$, we want our test to have a high probability of rejecting H_0.

Definition 2.4. The **power** of a test is the probability that we reject H_0, given that H_0 is false. The **power function** depends on the value of τ and is

$$\Psi(\tau) = \Pr\left(\text{Reject } H_0 \mid \tau \neq 0\right) = 1 - \beta.$$

The quantity β therefore represents $\Pr\left(\text{Fail to reject } H_0 \mid \tau \neq 0\right)$, which is the **type II error rate**.

If you find the notation confusing (as I do!) then it might be helpful to remember that both α and β are **error rates** - probabilities of coming to the wrong conclusion. It is common to talk in terms of α, the significance level, (which will be a low number, often 0.05) and of $1-\beta$, the power (which will be a high number, often 0.8). I've found though that it is not uncommon to find people refer to β (rather than $1-\beta$) as the power. If in doubt, keep in mind that we require $\alpha,\ \beta \ll 0.5$. It is also common to use percentages: a significance level of $\alpha = 0.05$ can also be referred to as "the 95% level", and $\beta = 0.2$ is the same as a "power of 80%". When using percentages, we talk in terms of the amount of time we expect the test to come to the correct conclusion.

If you notice any mistakes in these notes along these (or other!) lines, please point them out.

Under H_1, we have (approximately)

$$D \sim N\left(\frac{\tau}{\sigma\lambda(n,m)}, 1\right),$$

where $\lambda(n,m) = \sqrt{\frac{1}{n}+\frac{1}{m}}$ and

$$D = \frac{\bar{x}_T - \bar{x}_C}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

Figure 2.4 shows the distribution of D under H_0 and H_1 for some arbitrary (non-zero) effect size τ. The turquoise bar shows the acceptance region of H_0, ie. the range of observed values of D for which we will fail to reject H_0. We see that this contains 95% of the area of the H_0 distribution (we have set $\alpha = 0.05$ here), so under H_0, we have a 0.95 probability of observing a value of D that is consistent with H_0.

However, if H_1 is true, and $\tau \neq 0$, there is a non-zero probability of observing a value of D that would lead us to fail to reject H_0. This is shown by the area shaded in red, and it has area β. One minus this area (ie. the area under H_1 that leads us to accept H_1) is the power, $1-\beta$.

We can see that if the distributions have better separation, as in Figure 2.5, the power becomes greater. This can be as a result of a larger τ, a smaller σ or a smaller λ (therefore larger m and/or n).

For given values of α, σ and $\lambda(n,m)$, we can calculate the power function in terms of τ by finding the area of the distribution of D under H_1 for which we accept H_1.

$$\Psi(\tau) = 1-\beta = \left[1-\Phi\left(z_{\frac{\alpha}{2}} - \frac{\tau}{\sigma\lambda}\right)\right] + \Phi\left(-z_{\frac{\alpha}{2}} - \frac{\tau}{\sigma\lambda}\right) \tag{2.1}$$

The first term in Equation (2.1) is the area in the direction of τ. In Figures 2.4 and 2.5 this is the region to the right of the interval for which we fail to reject H_0, ie. where

$$D > z_{\frac{\alpha}{2}}.$$

The second term in Equation (2.1) represents the area away from the direction of τ, ie. a value of D such that

$$D < -z_{\frac{\alpha}{2}},$$

assuming without loss of generality that $\tau > 0$.

Figure 2.6 shows the power function $\Psi(\tau)$ for τ in units of σ (or you could think of this as for $\sigma = 1$), for three different pairs of values of n and m (remember that these enter the power function via λ) with $\alpha = 0.05$. We see that in general the power is higher for larger sample sizes, and that of the two designs where $n+m = 200$, the balanced one with $n = m = 100$ achieves the greatest power.

In general, the probability of rejecting H_0 increases as τ moves away from zero.

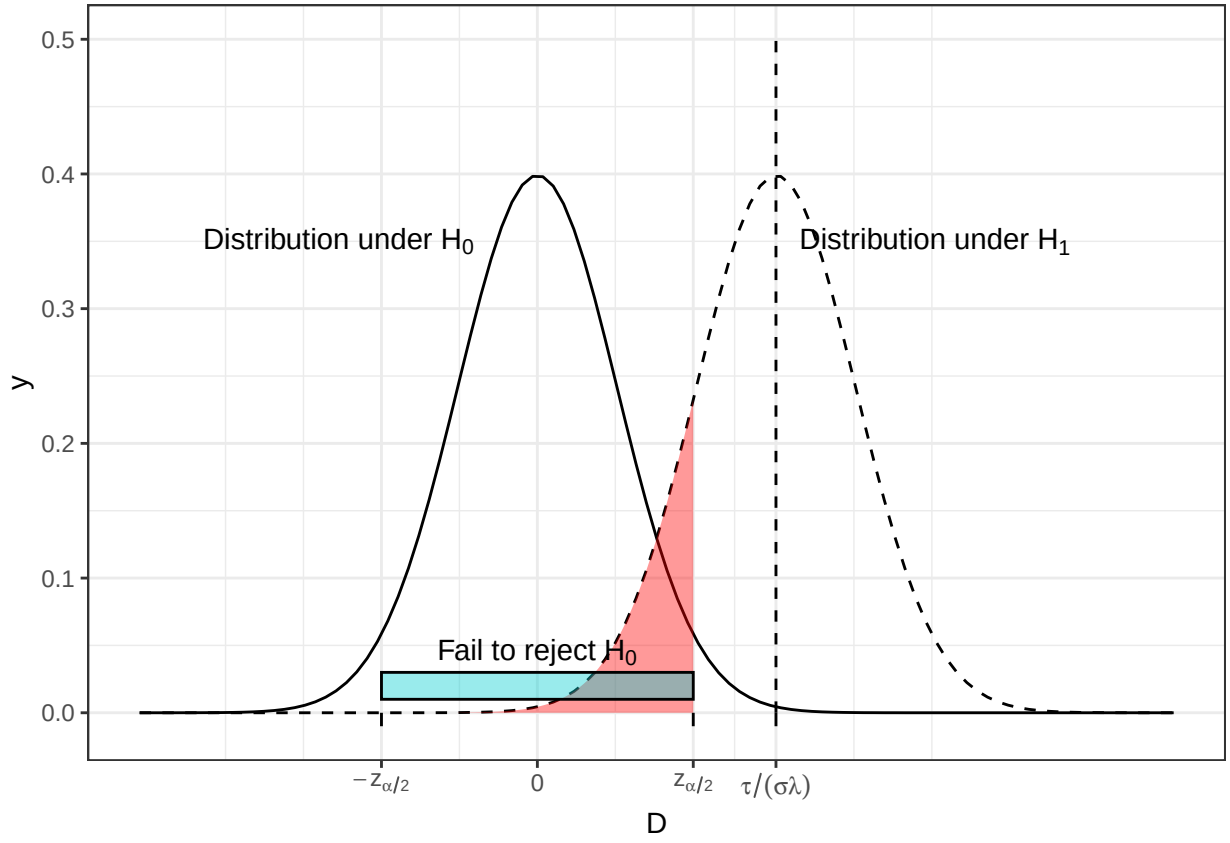


Figure 2.4: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

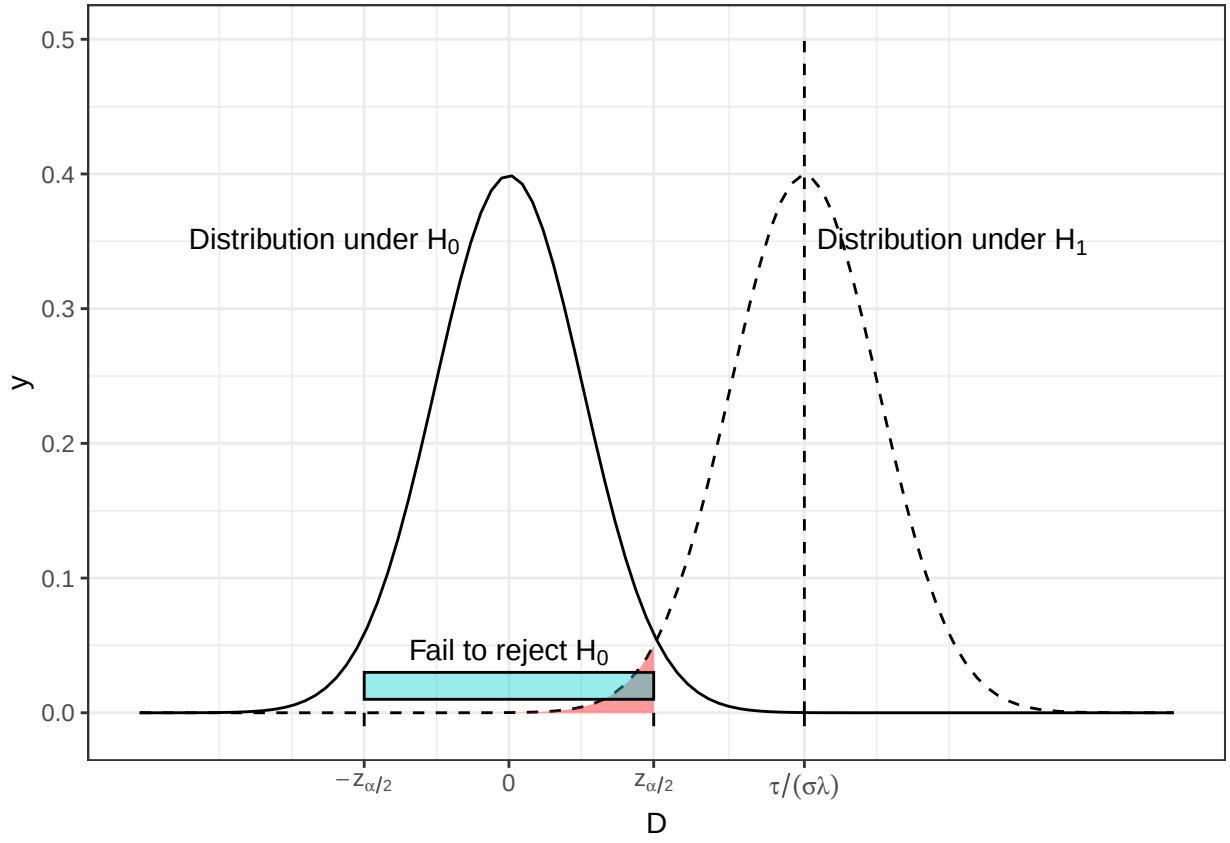


Figure 2.5: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

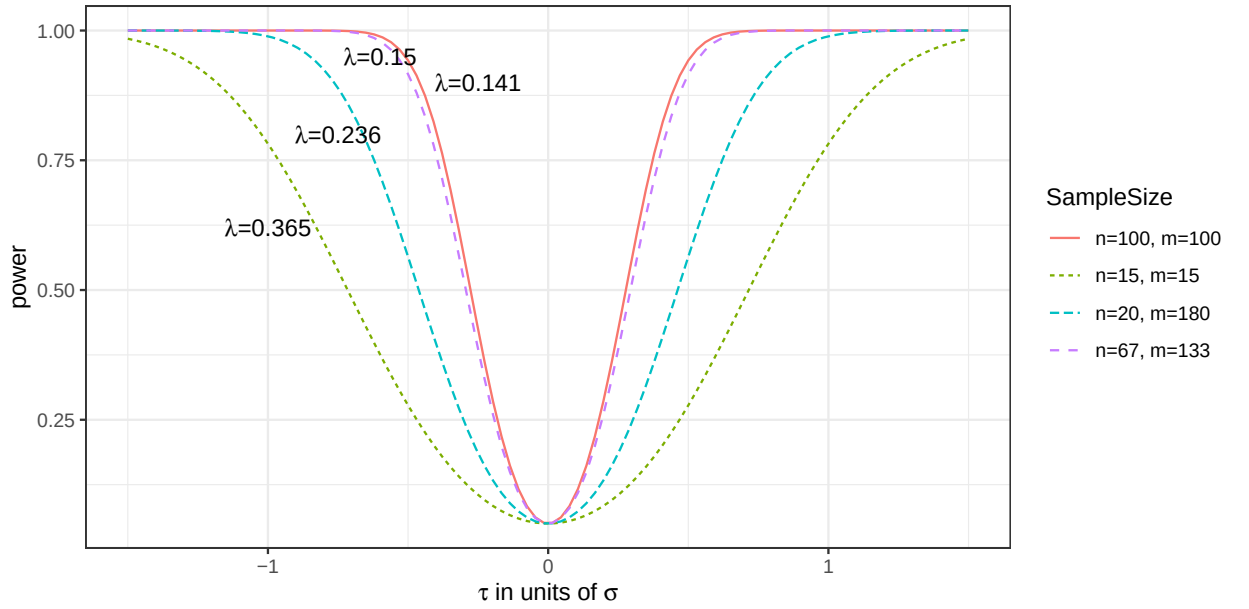


Figure 2.6: Power curves for various values of n and m, with effect size in units of standard deviation, given a type I error rate of 0.05.

Notice also that all the curves pass through the point $\tau = 0$, $\beta = 0.05$. Since $\tau = 0$ corresponds to H_0 being true, it makes sense that the probability of rejecting the H_0 is the significance level α.

It is common to think of the effect size in units of σ, as we have done here. This makes results more intuitive, since we don't need to have a good knowledge of the actual outcome variable to know what is a small or large effect size. It is also helpful in situations where the population standard deviation is not well understood, since the trial can be planned with this sort of effect size in mind. To denote the effect size in units of σ, we will write τ_σ, although in practice it is more usual to give both the same notation.

2.5 A sample size formula

Equation (2.1) allows us to find any one of τ_σ, α, β and $\lambda(n, m)$ given values for the others. Values for α and β are often specified by those planning the trial as around $\alpha \in [0.01, 0.05]$, $1 - \beta \in [0.8, 0.9]$.

The remaining two variables, τ_σ and $\lambda(n, m)$ are generally settled using one or both of the following questions:

- Given our budget constraints, and their implications for n and m, what is the smallest value of τ_σ we can achieve?
- What is the smallest value of τ_σ that would be clinically useful to detect, and what value of $\lambda(n, m)$ do we need in order to achieve it?

In a medical setting, an estimate of σ is usually available, and so we will return to thinking in terms of τ and σ. In this equation, the value we use (or find) for τ is the **minimum detectable effect size**, which we will denote τ_M.

Definition 2.5. The **minimum detectable effect size** τ_M for a particular trial is the smallest value of effect size that is able to be detected with power $1 - \beta$ and at significance level α (for some specified values of α, β).

Note that we will not *definitely* detect an effect of size τ_M, if it exists; by construction, we will detect it with probability $1-\beta$. If $|\tau| > |\tau_M|$ (ie. the true effect size is further from zero than τ_M is) then the probability of detecting it will be greater than $1-\beta$. If $|\tau| < |\tau_M|$ then the probability of detecting it will be less than $1-\beta$.

Although we could solve Equation (2.1) numerically, in practice we use an approximation. The second term, representing observed values of D that are far enough away from 0 *in the opposite direction from the true* τ to lead us to reject H_0 is so negligible as to be able to be discounted entirely. Indeed, if we were to observe such a value of D, we would come to the wrong conclusion about τ.

Therefore, Equation (2.1) becomes

$$\Psi(\tau) = 1-\beta = \left[1 - \Phi\left(z_{\frac{\alpha}{2}} - \frac{\tau_M}{\sigma\lambda}\right)\right]. \tag{2.2}$$

Because $\Phi(z_\beta) = 1-\beta$ (by definition) and $\Phi(-z) = 1-\Phi(z)$ we can write this as

$$\Phi(z_\beta) = \Phi\left(\frac{\tau_M}{\sigma\lambda} - z_{\frac{\alpha}{2}}\right),$$

where τ_M is our minimum detectable effect size. Because of the monotonicity of $\Phi(\cdot)$, this becomes

$$\begin{aligned} z_\beta &= \frac{\tau_M}{\sigma\lambda} - z_{\frac{\alpha}{2}} \\ z_\beta + z_{\frac{\alpha}{2}} &= \frac{\tau_M}{\sigma\lambda}. \end{aligned} \tag{2.3}$$

Because we want to think about sample sizes, we rewrite this further. It is most common to perform trials with $n = m = N$ participants in each group, in which case

$$\lambda(n,m) = \sqrt{\frac{2}{N}}$$

and Equation (2.3) rearranges to

$$N = \frac{2\sigma^2\left(z_\beta + z_{\frac{\alpha}{2}}\right)^2}{\tau_M^2}. \tag{2.4}$$

Example 2.1. (from Zhong, 2009) A trial is being planned to test whether there is a difference in the efficacy of ACEII antagonist (a new drug) and ACE inhibitor (the standard drug) for the treatment of primary hypertension (high blood pressure). The primary outcome variable is change in sitting diastolic blood pressure (SDBP, mmHg) compared to a baseline measurement taken at the start of the trial. The trial should have a significance level of $\alpha = 0.05$ and a power of $1-\beta = 0.8$, with the same number of participants in each group. The minimum clinically important difference is $\tau_M = 3$ mmHg and the pooled standard deviation is $s = 8$ mmHg. Therefore, using equation (2.4) the sample size should be at least

$$\begin{aligned} N &= \frac{2\times 8^2\left(0.842+1.96\right)^2}{3^2} \\ &= 111.6, \end{aligned}$$

and therefore we need at least 112 participants in each trial arm.

Chapter 3

Allocation

Once we've decided how many participants we need in our trial, and they've been recruited, we next need to determine which participants should be assigned to which trial arm. This is process is known as **allocation** (or sometimes as **randomization**). Before we think about methods for allocation, we are going to spend some time talking about bias.

3.1 Bias

In statistics, *bias* is a systematic tendency for the results of our analysis to be different from the true value. We see this particularly when we are using sample data to estimate a parameter. We will revisit what we have learned in previous courses about bias before going on to see how it affects RCTs.

Definition 3.1 (Bias of an estimate)**.** Suppose that T is a statistic calculated to estimate a parameter θ. The **bias** of T is

$$E(T) - \theta.$$

If the bias of T is zero, we say that T is an **unbiased estimator** of θ.

An example you will have seen before is the standard deviation. If we have some data $x_1, \ldots, x_n$ that are IID $N(\mu, \sigma^2)$, we can calculate the sample variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

In this case, $E(s^2) \neq \sigma^2$ (you've probably seen this proved so we're not going to prove it now), and s^2 is a biased estimator of σ^2. However, we know that

$$E\left(\frac{n}{n-1}s^2\right) = \sigma^2,$$

and therefore we can apply this correction to the sample variance s^2 to produce an unbiased estimate of the population variance σ^2.

Now, suppose our sample $x_1, \ldots, x_n$ were drawn from $N(\mu, \sigma^2)$, but were **not** independent of one another. Then, neither our estimator s^2, nor our bias-corrected estimator $\frac{n}{n-1}s^2$ would have expected value σ^2. Furthermore, we cannot use our sample $x_1, \ldots, x_n$ to produce an unbiased estimator of σ^2, or even of the mean μ.

This scenario is much closer to what we mean when we talk about *bias* in a clinical trial setting. Suppose we are testing some new treatment T against the standard C. We measure some outcome X for each patient, and our hypothesis is that X behaves differently for those in the treatment group than for those in the control group. It is common practice to express this additively,

$$E(X) = \mu + \tau,$$

where τ is our treatment effect, which we can estimate using the difference in the groups' means, $\bar{X}_T - \bar{X}_C$. Our null hypothesis is that $\tau = 0$, and our alternative hypothesis is that $\tau \neq 0$, and therefore an estimate of τ from our data is very important! Put equivalently, it is important that there is no bias in our estimates of $\bar{X}_C$ and $\bar{X}_T$.

Usually, what this comes down to is that the assumption that the data are independent, identically distributed random variables from the relevant distributions (which we have already relied on a lot for our sample size calculations) has been violated in some way.

Example 3.1. Historically, women and the elderly are underrepresented in clinical trials (Cottingham and Fisher (2022)) and results are often translated from young or middle aged healthy men to these other groups (Vitale et al. (2017)). This isn't reasonable, since women have very different hormonal activity from men, causing them to often react differently to drugs compared to men involved in the trial. The standard dose (based on trials with mostly male participants) can also be too high for many women. The complicated nature of women's hormones is sometimes even given as a reason for not including them in the trial. Women and elderly people are also both more likely to have adverse effects to drugs in some fields.

There are also ethical reasons behind the low numbers of women in trials, especially phase I and phase II trials. If a woman is possibly pregnant (and trials tend to be extremely cautious in deciding who might be pregnant!) then they are quite often excluded, in order to protect the (actual or hypothetical) fetus. Indeed, in 1977 the Food and Drug Administration (FDA) in the US recommended that women be excluded from phase I and II trials (of Health (2023)) as a result of some severe cases of fetuses being harmed by drugs (especially Thalidamide) . This means that even some very mainstream drugs, for example antihistamines (Kar et al. (2012)), haven't been tested for safety/efficacy during pregnancy, as well as some (for example HIV treatments) that would be of huge benefit to many many pregnant women. This article is an interesting read if you would like to know more.

3.1.1 Where does bias come from?

Having established that bias is a serious issue in clinical trials, we will think about several sources of bias. Some of these we will elaborate on as we get to the relevant part of methodology. Most sources of bias creep in during the allocation or selection phase.

Selection bias

Selection bias occurs when certain patients or subjects are systematically more (or less) likely be entered into the trial because of the treatment they will receive. In a properly run trial this isn't possible, because it is only after a participant has been recruited that their treatment is chosen. If a medical professional is not comfortable with a particular patient potentially receiving one of the possible treatments, then that patient should not be entered into the trial at all. If there are many such [technically eligible] patients, then this might cause the estimated treatment effect to be worryingly far from the true population treatment effect, since the recruited group of participants would not be very representative of the true population (this is not technically selection bias, but it comes from the same problem).

It may happen that the doctor knows which treatment a patient would be given, for example if the allocation follows some deterministic pattern, or is fully known to the doctor in advance. Consciously or subconsciously this knowledge may influence the description they give to potential participants, and this in turn may affect which patients sign up, and the balance of the groups. In practice there should be various safeguards against this situation.

Example 3.2. Suppose we run a trial comparing a surgical (S) and a non-surgical (N) treatment for some condition. Patients who are eligible are given the opportunity to join the trial by a single doctor.

The severity of the disease is graded as 1 (less serious) or 2 (more serious) for each patient. Across the full group of patients, proportion λ have severity 1 and proportion $1-\lambda$ have severity 2.

Our primary outcome is survival time, X, which depends on the severity of disease:

$$\begin{aligned} E(X \mid 1) &= \mu_1 \\ E(X \mid 2) &= \mu_2 \end{aligned}$$

and we assume $\mu_1 > \mu_2$.

For the overall trial group, for untreated patients we have

$$E(X) = \mu = \lambda\mu_1 + (1-\lambda)\mu_2.$$

Suppose that for treatment group N, the expected survival time increase by τ_N, and similarly for group S, so that we have

$$\begin{aligned} E(X \mid N, 1) &= \mu_1 + \tau_N \\ E(X \mid N, 2) &= \mu_2 + \tau_N \\ E(X \mid S, 1) &= \mu_1 + \tau_S \\ E(X \mid S, 2) &= \mu_2 + \tau_S. \end{aligned}$$

If all patients were admitted with equal probability to the trial (ie. independent of the severity of their disease) then the expected survival time for group N, $E(X \mid N)$, would be

$$\begin{aligned} E(X \mid 1, N)\, P(1 \mid N) + E(X \mid 2, N)\, P(2 \mid N) &= (\mu_1 + \tau_N)\,\lambda + (\mu_2 + \tau_N)\,(1-\lambda) \\ &= \mu + \tau_N. \end{aligned}$$

Similarly, the expected survival time in group S would be $\mu + \tau_S$, and the treatment effect difference between the two would be $\tau = \tau_N - \tau_S$ and the trial is unbiased.

Suppose that although all eligible patients are willing to enter the trial, the doctor is reticent to subject patients with more severe disease (severity 2) to the surgical procedure. This is reflected in the way they explain the trial to each patient, particularly those with severity 2 whom the doctor knows will be assigned to group S. In turn this leads to a reduced proportion $q = 1-p$ of those with severity 2 assigned to surgery entering the trial (event A):

$$\begin{aligned} P(A \mid N, 1) = P(A \mid S, 1) = P(A \mid N, 2) &= 1 \\ P(A \mid S, 2) &= 1 - p = q. \end{aligned}$$

Since our analysis is based only on those who enter the trial, our estimated treatment effect will be

$$E(X \mid A, N) - E(X \mid A, S).$$

We can split these according to disease severity, so that

$$E(X \mid A, N) = E(X \mid A, N, 1) P(1 \mid A, N) + E(X \mid A, N, 2) P(2 \mid A, N)$$

and similarly for group S.

We can calculate $P(1 \mid A, N)$ using Bayes' theorem,

$$\begin{aligned} P(1 \mid A, N) &= \frac{P(A \mid 1, N) P(1 \mid N)}{P(A \mid N)} \\ &= \frac{P(A \mid 1, N) P(1 \mid N)}{P(A \mid N, 1) P(1 \mid N) + P(A \mid N, 2) P(2 \mid N)} \\ &= \frac{1 \times \lambda}{1 \times \lambda + 1 \times (1 - \lambda)} \\ &= \lambda. \end{aligned}$$

Therefore we also have $P(2 \mid A, N) = 1 - P(1 \mid A, N) = 1 - \lambda$.

Following the same process for group S, we arrive at

$$\begin{aligned} P(1 \mid A, S) &= \frac{P(A \mid 1, S) P(1 \mid S)}{P(A \mid S)} \\ &= \frac{P(A \mid 1, S) P(1 \mid S)}{P(A \mid S, 1) P(1 \mid S) + P(A \mid S, 2) P(2 \mid S)} \\ &= \frac{\lambda}{\lambda + q(1 - \lambda)}, \end{aligned}$$

which we will call b.

Notice that $P(2 \mid S) = 1 - \lambda$, since it is not conditional on actually participating in the trial. Therefore,

$$\begin{aligned} E(X \mid A, N) &= E(X \mid N, 1) P(1 \mid A, N) + E(X \mid N, 2) P(2 \mid A, N) \\ &= (\mu_1 + \tau_N) \lambda + (\mu_2 + \tau_N)(1 - \lambda) \\ &= \lambda \mu_1 + (1 - \lambda) \mu_2 + \tau_N \end{aligned}$$

and

$$\begin{aligned} E(X \mid A, S) &= E(X \mid S, 1) P(1 \mid A, S) + E(X \mid S, 2) P(2 \mid A, S) \\ &= (\mu_1 + \tau_S) b + (\mu_2 + \tau_S)(1 - b) \\ &= b \mu_1 + (1 - b) \mu_2 + \tau_S. \end{aligned}$$

From here, we can calculate the expected value of the treatment effect τ as (substituting our equation for b and rearranging):

$$\begin{aligned} E(X \mid A, N) - E(X \mid A, S) &= \tau_N - \tau_S + (\lambda - b)(\mu_1 - \mu_2) \\ &= \tau_N - \tau_S - \frac{p \lambda (1 - \lambda)(\mu_1 - \mu_2)}{\lambda + q(1 - \lambda)}, \end{aligned}$$

where the third term represents the bias.

Notice that if $q = 1 - p = 1$, then there is no bias. There is also no bias if $\mu_1 = \mu_2$, ie. if there is no difference between the disease severity groups in terms of survival time.

Assuming $\mu_1 - \mu_2 > 0$, then the bias term is positive and

$$E(X \mid A, N) - E(X \mid A, S) < \tau_N - \tau_S.$$

If N is the better treatment, then $\tau_N - \tau_S > 0$ and the bias will cause the trial to underplay the treatment effect. Conversely, if S is better, then $\tau_N - \tau_S < 0$ and the trial will exaggerate the treatment effect. Essentially, this is because more severely ill patients have been assigned to N than to S, which reduces the average survival time for those in group N.

Allocation bias

Mathematically, allocation bias is similar to selection bias, but instead of coming from human 'error', it arises from the random process of allocation.

Suppose a trial investigates a drug that is likely to have a much stronger effect on male patients than on female patients. The cohort of recruited participants are randomised into treatment and control groups, and it happens that there is a much smaller proportion of female patients in the treatment group than in the control group. This will distort the estimated treatment effect.

We will investigate various strategies for randomization designed to address this issue for known factors.

Assessment bias

Measurements are made on participants throughout (and often during) the trial. These measurements will often be objective, for example the patients' weight, or concentration of blood sugar. However, some types of measurement are much more subject to the individual practitioner assessing the patient. For example, many skin conditions are assessed visually, for example estimating the proportion of the body affected. Measuring quantities such as quality of life or psychological well-being involve many subjective judgements on the part of both patient and clinician.

Clearly it is ideal for both the patient and the clinician not to know which arm of the trial the patient was part of (this is known as a **double blind trial**). For treatments involving drugs, this is usually straightforward. However, for surgical interventions it is often impossible to keep a trial 'blind', and for interventions involving therapy (for example cognitive behavioural therapy) it is impossible for the patient to be unaware.

Slight aside: publication bias

In most areas of science, including clinical trials, the ultimate aim is to affect practice. This is usually done by publishing a write-up of the trial, including its design, methods, analysis and results, and publishing that in a [medical] journal. These are peer-reviewed, which means that experts from the relevant field are ask to review submitted papers, and either reject or accept them (usually conditional on some revision). These reviewers advise the editor of the journal, who ultimately decides whether or not the paper will be published.

It seems that papers reporting positive / conclusive results are more likely to be published than papers about [viable] trials that ultimately fail to reject the null hypothesis. As we know, in most cases if the null hypothesis is rejected this is indicative that there is a true treatment difference. However, sometimes by random chance a trial will detect a difference even when there isn't one (approximately 5% of the time if $\alpha = 0.05$). If these papers are disproportionately likely to be published, the body of literature will not reflect the truth, and there may be serious implications for impact on practice.

Measures are being taken to prevent this: for example, leading medical journal *The Lancet* insists that any clinical trial related paper is registered with them before the first participant has been recruited, with details of the design and statistical analysis plan. This is then reviewed before the trial begins.

3.1.2 Implications for allocation

Historically (and probably still, to an extent), clinical trials have not necessarily used random allocation to assign participants to groups. Altman and Bland (1999) gives an overview of why this has led to bias, and gives some examples.

Sometimes analyses compare groups in serial, so that N_A patients one year (say) form the control group, and N_B patients in a subsequent year, who are given treatment B, form the intervention group. In this scenario it is impossible to control for all other changes that have occurred with time, and this leads to a systematic bias, usually in favour of treatment B.

Given the need for contemporary control participants, the question becomes how to assign participants to each group. If the clinician is able to choose who receives which treatment, or if each patient is allowed to choose or refuse certain treatments, this is almost certain to introduce bias. This is avoided by using random allocation.

There are two important aspects to the allocation being *random* that we will draw attention to.

1. Every patient should have the same probability of being assigned to each treatment group.
2. The treatment group for a particular patient should not be able to be predicted.

Point 1 is important because, as we have already mentioned, the statistical theory we use to plan and analyse the trial is based on the groups being random samples from the population.

Point 2 is important to avoid biases that come through the assignment of a particular patient being known either in advance or after the fact. There are some approaches that 'pass' the first point, but fail at the second. As well as strict alternation ($ABABAB\ldots$), some such methods use patient characteristics such as date of birth or first letter of surname, which is not related to the trial outcome, but which enables allocations to be predicted.

We will now explore some commonly used methods of allocation. We will usually assume two equally sized groups, A and B, but it is simple to generalize to three or more groups, or to unequal allocation.

3.2 Allocation methods

3.2.1 Simple random allocation

Perhaps intuitively the most simple method is a 'toin coss', where each participant has a probability 0.5 of being placed in each group. As participants arrive, assignment C or T is generated (with equal probability). Statistically, this scheme is ideal, since it generates the random sample we need, and the assignment of each participant is statistically independent of that of all other participants. It also doesn't require a 'master' randomisation; several clinicians can individually assign participants to treatment groups in parallel and the statistical properties are maintained.

This method is, effectively, used in many large trials, but for small trials it can be statistically problematic. The main reason for this is chance imbalance of group sizes.

Suppose we have two groups, T of size N_T and C of size N_C, with $N_T + N_C = 2n$. Patients are allocated independently with equal probability, which means

$$N_C \sim \text{Bi}\left(2n, \frac{1}{2}\right),$$

and similar for N_T. If the two groups are of unequal size, the larger will be of some size N_{max} between n and $2n$, such that for $r = n+1, \ldots, 2n$,

$$\begin{aligned} P\left(N_{max} = r\right) &= P\left(N_C = r\right) + P\left(N_T = r\right) \\ &= 2\binom{2n}{r}\left(\frac{1}{2}\right)^{2n}. \end{aligned}$$

The probability that $N_C = N_T = n$ is

$$P\left(N_T = N_C = n\right) = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n}.$$

These probabilities are shown in Figure 3.1. We can see that this method leads to very unequal groups relatively easily; with $n = 15$, $P\left(N_{max} \geq 20\right) = 0.099$, so there is around a one in ten chance that one group will be double or more the size of the other.

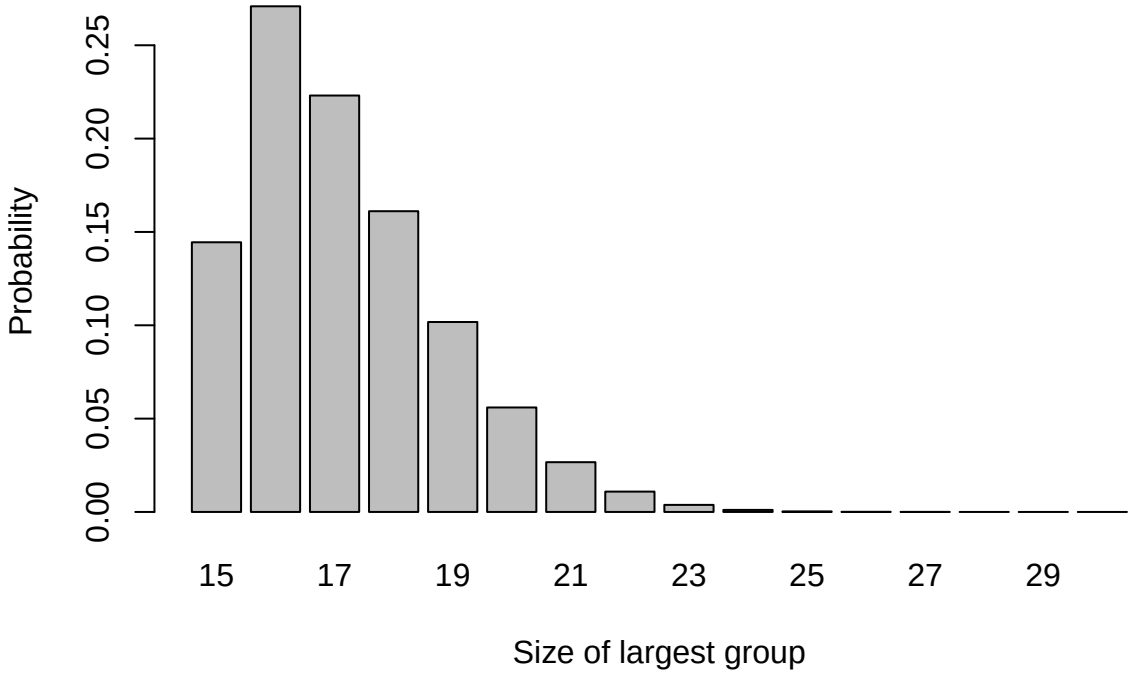


Figure 3.1: The probability distribution of largest group size for n=15.

As we have seen when thinking about sample sizes in Section 2.4, this will reduce the power Ψ of the trial, since it depends on $\lambda\left(N_C, N_T\right) = \sqrt{\frac{1}{N_C} + \frac{1}{N_T}}$.

For larger trials, this imbalance will be less pronounced, for example Figure 3.2 shows the same for $n = 200$.

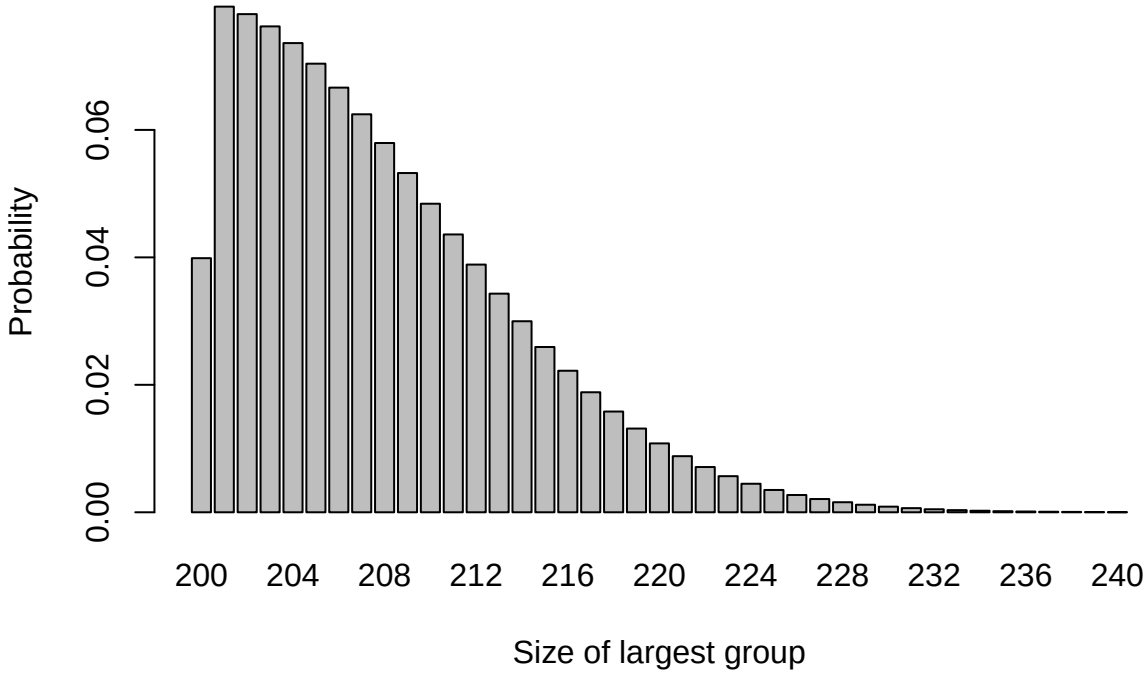


Figure 3.2: The probability distribution of largest group size for n=200.

In this case the $P(N_{max} \geq 220) = 0.051$, so the chance of highly imbalanced groups is much lower. However, we may want to achieve balance on some factor thought to be important, for example sex, age group or disease state, and in this case there may be small numbers even in a large trial.

We saw in the sample size section that the greatest power is achieved when group sizes are equal, since this minimises the function

$$\lambda(n, m) = \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

However, with simple random sampling we can't guarantee equal group sizes.

Example 3.3. Suppose we are designing a trial to have $\alpha = 0.05$, and our minimum detectable effect size is such that $\frac{\tau_M}{\sigma} = 1$. If 30 participants are recruited, then we can calculate the power of the study using methods from Chapter 2:

$$1 - \beta = \Phi\left(\sqrt{\frac{n_T\, n_C}{30}} - 1.96\right).$$

The first term in the standard normal CDF comes from the fact that

$$[\lambda(n, m)]^{-1} = \sqrt{\frac{nm}{n+m}}.$$

If we have equal group sizes $n_T = n_C = 15$, then the power achieved is 78%. If the group sizes are 10 and 20, we have a power of 73%. If the group sizes are 6 and 24, the power goes down to 59%.

So, as we saw when looking at power, we don't lose too much if the group sizes are 2:1, but a more pronounced imbalace has resulted in a much more noticeable loss. There may be other disadvantages to having such imbalance, for example increased costs, or a reduction in the amount of information gained about side effects. If this imbalance can be avoided, it should be.

3.2.2 Random permuted blocks

One commonly used method to randomly allocate participants while avoiding too much imbalance is to use *random permuted blocks* (RPBs). If the blocks have size $2m$, and there are two groups then there are

$$\binom{2m}{m},$$

but this method can be adapted to more than two groups and to unequal group size.

If we have two groups, A and B, then there are six *blocks* of length 4 containing two As and two Bs

1. $AABB$
2. $ABAB$
3. $ABBA$
4. $BAAB$
5. $BABA$
6. $BBAA$.

We can also randomly generate a sequence of numbers from $\{1, 2, 3, 4, 5, 6\}$, where each number has equal probability. This sequence will correspond to a sequence in A and B with four times the length. In this method, each patient is equally likely to receive A and B, but there will never be a difference of more than two between the size of the two groups.

For example, suppose the sequence begins $2, 1, 3, 6, \ldots$. Replacing each number by its block, we have $ABAB\ AABB\ ABBA\ BBAA\ \ldots$.

One serious disadvantage of this method is that if the block size is fixed, and the doctors involved in the trial know which participants have received which treatments (which is unavoidable in cases such as surgery), then the allocation for some patients can be perfectly predicted. This is true for the fourth in every block, and for the third and fourth if the first two were the same. This means that selection bias may be a problem in more than 25% of participants, which is deemed unacceptable; indeed, it fails our second point about randomization.

3.2.2.1 RPBs with random block length

The issue above can be circumvented by not only randomly choosing from a selection of blocks, but also randomly choosing the length of the block. For example, there are

$$\binom{6}{3} = 20$$

possible blocks of size 6. Instead of always selecting from the six possible 4-blocks, a sampling scheme can be as follows.

1. A random number X is drawn from $\{4, 6\}$ to select the block length.
2. A second random number Y is drawn from 1 to 6 (if the block length is four) or 1 to 20 (if the block length is 6).
3. The block corresponding to Y is chosen and participants assigned accordingly.
4. If more participants are needed, go back to step 1.

As well as ensuring that patients are equally likely to receive treatments A and B, and that N_A and N_B can never differ by more than three, this method hugely reduces the possibility of enabling selection bias. The assignment of a patient can only be perfectly predicted if the difference is three, and this happens only for two of the twenty blocks of length six.

3.2.3 Biased coin designs and urn schemes

It may be that we prefer a method which achieves balance while retaining the pure stochasticity of simple random sampling. An advantage of RPBs was that once the sequence was generated, no computing power was needed. However, it is safe now to assume that any hospital pharmacy, nurse's station, GP office or other medical facility will have a computer with access to the internet (or some internal database), and therefore more sophisticated methods are available. It is also very likely that all trial data may be stored on some central database, and so methods that rely on knowing the allocation so far (albeit in some encrypted form) should be possible even if there are multiple clinicians and sites involved.

Biased coin designs and urn schemes both work by adjusting the probabilities of allocation according to balance of the design so far, such that a participant is less likely to be assigned to an over-represented group.

3.2.3.1 Biased coin designs

The biased coin design was introduced by Efron (1971), with the aim of ensuring balance whilst not becoming vulnerable to various forms of experimental bias. Efron (1971) suggested the biased coin design be used within categories (eg. age group, sex, disease state etc.), but in this section we will think about the whole cohort (the maths for a subgroup would be the same).

Suppose we are using a biased coin design for a trial to compare two treatments, T and C. At the point where some number n (not the total trial cohort) have been allocated, we can use the notation $N_T(n)$ for

the number of participants allocated to treatment T, and $N_C(n)$ for the number of participants allocated to treatment C. Using these, we can denote the *imbalance* in treatment numbers by

$$D(n) = N_T(n) - N_C(n) = 2N_T(n) - n.$$

We use the imbalance $D(n)$ to alter the probability of allocation to each treatment in order to restore (or maintain) balance in the following way:

- If $D(n) = 0$, allocate patient $n+1$ to treatment T with probability $\frac{1}{2}$.
- If $D(n) < 0$, allocate patient $n+1$ to treatment T with probability P.
- If $D(n) > 0$, allocate patient $n+1$ to treatment T with probability $1-P$.

where $P \in \left(\frac{1}{2}, 1\right)$.

Question: What would happen if $P = \frac{1}{2}$ or $P = 1$?

If, at some point in the trial, we have $|D(n)| = j$, for some $j > 0$, then we must have either

$$|D(n+1)| = j+1$$

or

$$|D(n+1)| = j-1.$$

Because of the way we have set up the scheme,

$$p\big(|D(n+1)| = j+1\big) = 1-P$$

and

$$p\big(|D(n+1)| = j-1\big) = P.$$

If $|D(n)| = 0$, ie. the scheme is in exact balance after n allocations, then we must have $|D(n)| = 1$.

The absolute imbalances therefore form a simple random walk on the non-negative integers, with transition probabilities

$$\begin{aligned} P\Big(|D(n+1)| = 1 \;\Big|\; |D(n)| = 0\Big) &= 1 \\ P\Big(|D(n+1)| = j+1 \;\Big|\; |D(n)| = j\Big) &= 1-P \\ P\Big(|D(n+1)| = j-1 \;\Big|\; |D(n)| = j\Big) &= P \end{aligned}$$

Figure 3.3 shows four realisations of this random walk with $P = 0.667$ (Efron's preferred value). We see that sometimes the imbalance gets quite high, but in general it isn't too far from 0.

Figure 3.4 shows four realisations of the random walk with $P = 0.55$. Here, the imbalance is able to get very high (note the change in y-axis); for example in the first plot, if we stopped the trial at $n = 50$ we would have 34 participants in one arm and only 16 in the other.

By contrast, with $P = 0.9$ as in Figure 3.5, there is much less imbalance. However, this brings with it greater predictability. Although allocation is always random, given some degree of imbalance (likely to be known about by those executing the trial), the probability of guessing the next allocation correctly is high (0.9). This invites the biases we have been trying to avoid, albeit in an imperfect form.

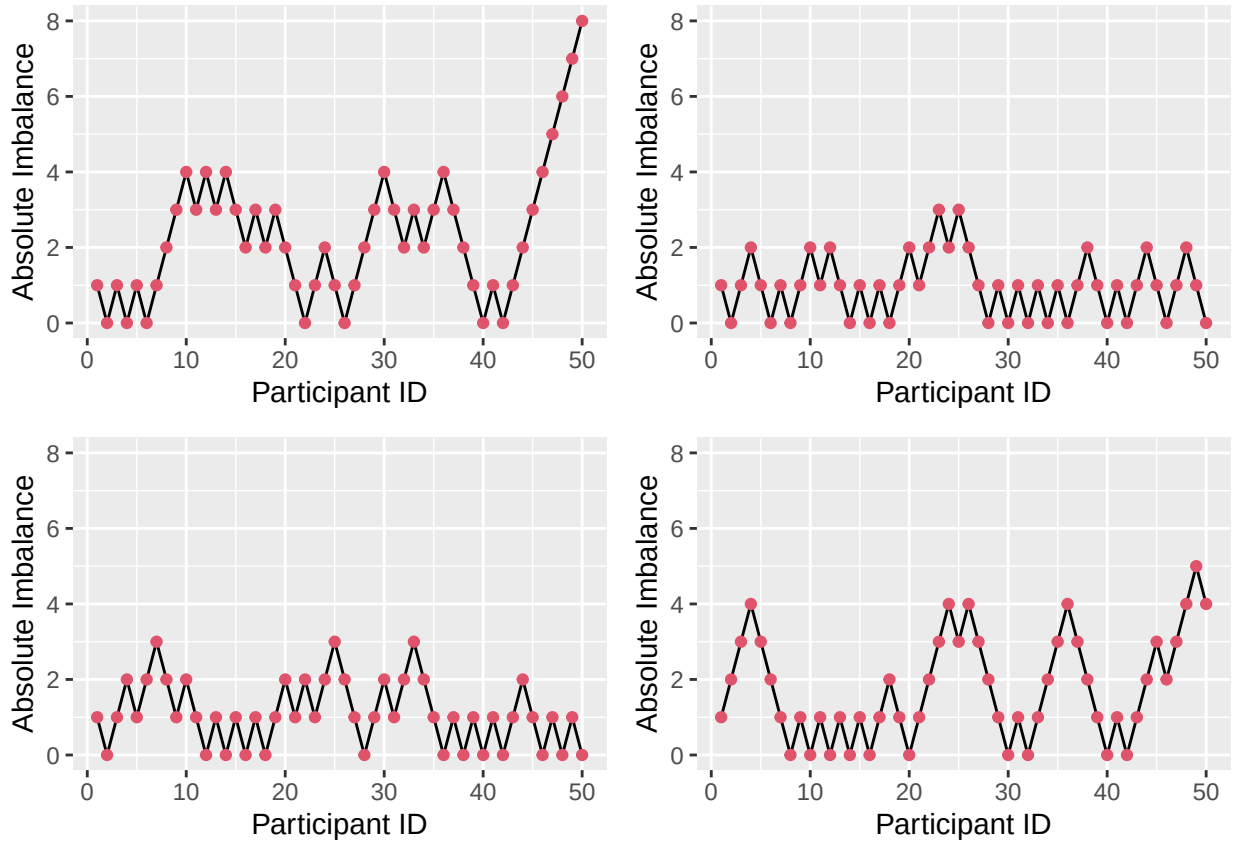


Figure 3.3: Absolute imbalance for a biased-coin scheme with $P = 0.667$.

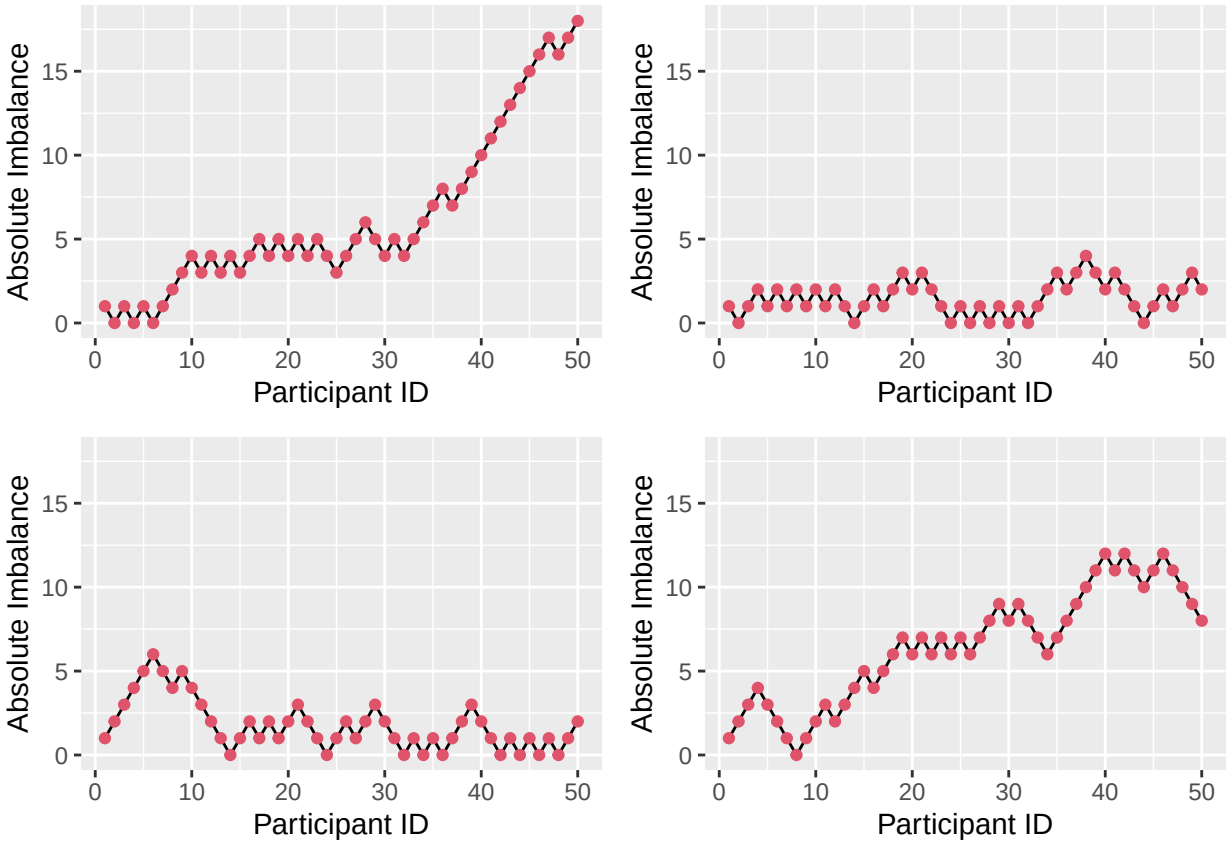


Figure 3.4: Absolute imbalance for a biased-coin scheme with $P = 0.55$.

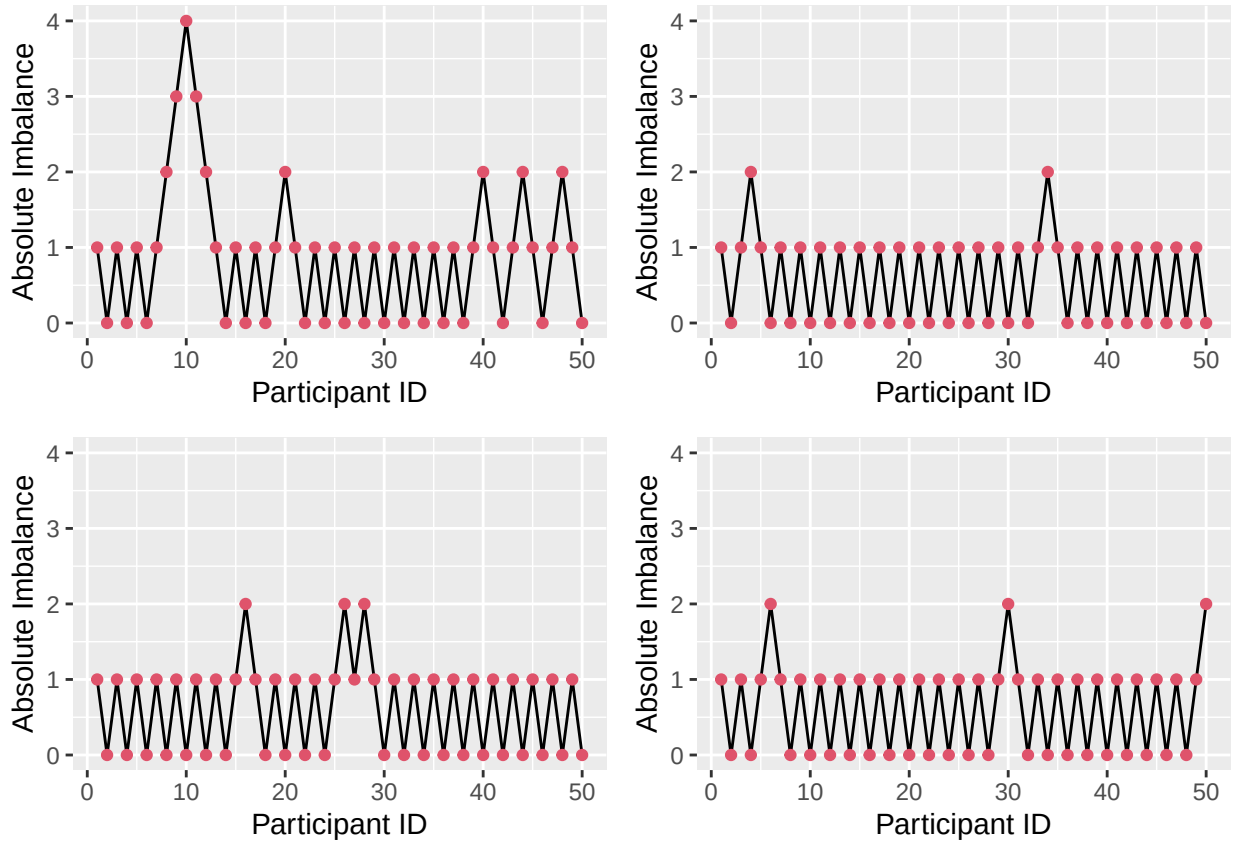


Figure 3.5: Absolute imbalance for a biased-coin scheme with $P = 0.9$.

Efron's suggestion for implementation was that each clinician would received an unordered stack of envelopes. Each would contain three more envelopes, each with instructions covering one of the three possible cases ($|D(n)| < 0$, $|D(n)| = 0$ and $|D(n)| > 0$). The clinician would open the appropriate envelope and implement the instruction. Remember this was 1971!

A big disadvantage to the biased coin scheme is that the same probability is used regardless of the size of the imbalance (assuming it isn't zero). In the next section, we introduce a method where the probability of allocating the next patient to the underrepresented treatment gets larger as the imbalance grows.

3.2.3.2 Urn models

Urn models for treatment allocation use urns in the way that you might well remember from school probability (or indeed often we had drawers of socks). They were first applied to clinical trials by Wei (1978). In this setting, the urn starts off with a ball for each treatment, and a ball is added to the urn each time a participant is allocated. The ball is labelled according to the treatment allocation that participant **did not** receive.

To allocate the next participant, a ball is drawn from the urn. If the allocations at this point are balanced, then the participant has equal probability of being allocated to each treatment. If there is imbalance, there will be more balls labelled by the underrepresented treatment, and so the participant is more likely to be allocated to that one. The greater the imbalance, the higher the probability of reducing it.

The process described so far is a $UD(1,1)$; there is one ball for each treatment to start with, and one ball is added to the urn after each allocation. To be more general, we can assume a $UD(r,s)$ scheme. Now, there are r balls for each treatment in the urn to begin with, and s are added after each allocation.

Near the start of the allocation, the probabilities are likely to change a lot to address imbalance, but once a 'reasonable number' of allocations have been made it is likely to settle into simple random sampling (or very close).

Once again, we can find the transition probabilities by considering the absolute imbalance $|D(n)|$.

Suppose that after participant n, $N_T(n)$ participants have been allocated to group T, and $N_C(n) = n - N_T(n)$ to group C. The imbalance is therefore

$$D(n) = N_T(n) - N_C(n) = 2N_T(n) - n.$$

After n allocations there will be $2r + ns$ balls in the urn: r for each treatment at the start, and s added after each allocation. Of these, $r + N_C(n)s$ will be labelled by treatment T and $r + N_T(n)s$ by treatment C.

To think about the probabilities for the absolute imbalance $|D(n)|$, we have to be careful now about which direction it is in. If the trial currently (after allocation n) has an imbalance of participants in favour of treatment C, then the probability that it becomes less imbalanced at the next allocation is the probability of the next allocation being to treatment T, which is

$$\begin{aligned} p(|D(n+1)| = j-1 \mid D(n) = j, j > 0) &= \frac{r + N_C(n)s}{2r+ns} \\ &= \frac{r + \frac{1}{2}(n + D(n))s}{2r+ns} \\ &= \frac{1}{2} + \frac{D(n)s}{2(2r+ns)} \\ &= \frac{1}{2} + \frac{|D(n)|s}{2(2r+ns)}. \end{aligned}$$

Similarly, if there is currently an excess of patients allocated to treatment T, then the imbalance will be reduced if the next allocation is to treatment C, and so the conditional probability is

$$\begin{aligned} p\left(|D\left(n+1\right)|=j-1 \mid D\left(n\right)=j, j<0\right) &= \frac{r+N_T\left(n\right)s}{2r+ns} \\ &= \frac{r+\frac{1}{2}\left(n-D\left(n\right)\right)s}{2r+ns} \\ &= \frac{1}{2}-\frac{D\left(n\right)s}{2\left(2r+ns\right)} \\ &= \frac{1}{2}+\frac{|D\left(n\right)|s}{2\left(2r+ns\right)}. \end{aligned}$$

Because the process is symmetrical, an imbalance of a given magnitude (say $|D\left(n\right)|=j$) is equally likely to be in either direction. That is

$$p\big(D\left(n\right)<0 \mid |D\left(n\right)|=j\big)=p\big(D\left(n\right)>0 \mid |D\left(n\right)|=j\big)=\frac{1}{2}.$$

Therefore we can use the law of total probability (or partition theorem) to find that

$$p\big(|D\left(n+1\right)|=j-1 \mid |D\left(n\right)|=j\big)=\frac{1}{2}+\frac{|D\left(n\right)|s}{2\left(2r+ns\right)}.$$

Since the two probabilities are equal this is trivial. Since the only other possibility is that the imbalance is increased by one, we also have

$$p\big(|D\left(n+1\right)|=j+1 \mid |D\left(n\right)|=j\big)=\frac{1}{2}-\frac{|D\left(n\right)|s}{2\left(2r+ns\right)}.$$

As with the biased coin design, we also have the possibility that the imbalance after n allocations is zero, in which case the absolute imbalance after the next allocation will definitely be one. This gives us another simple random walk, with

$$\begin{aligned} P\big(|D\left(n+1\right)|=1 \mid |D\left(n\right)=0\big) &= 1 \\ P\big(|D\left(n+1\right)|=j+1 \mid |D\left(n\right)=j\big) &= \frac{1}{2}-\frac{|D\left(n\right)|s}{2\left(2r+ns\right)} \\ P\big(|D\left(n+1\right)|=j-1 \mid |D\left(n\right)=j\big) &= \frac{1}{2}+\frac{|D\left(n\right)|s}{2\left(2r+ns\right)} \end{aligned}$$

We see that imbalance is reduced, particularly for small n. A small r and large s enhance this, since the large number (s) of balls added to the urn with each allocation weight the probabilities more heavily, as in Figure 3.7. By contrast, if r is large and s is small, as in Figure 3.8, the probabilities stay closer to $\left(\frac{1}{2},\frac{1}{2}\right)$ and so more imbalance occurs early on.

3.3 Incorporating baseline measurements

At the start of the trial (ideally before allocation) various baseline measurements are usually taken. If the primary outcome variable is a continuous measurement (eg. blood pressure, weight,...) this same quantity will often be included, so that there is some measure of each participant's condition/symptoms at the start of the trial. Factors such as age, sex, level of symptoms, things to do with treatment history and many others are included. Essentially, we include any variable we can that may lead to bias if not properly dealt with. The crucial thing is that none of these measurements (taken when they are) should be affected by the trial.

Such baseline measurements can be used in allocation.

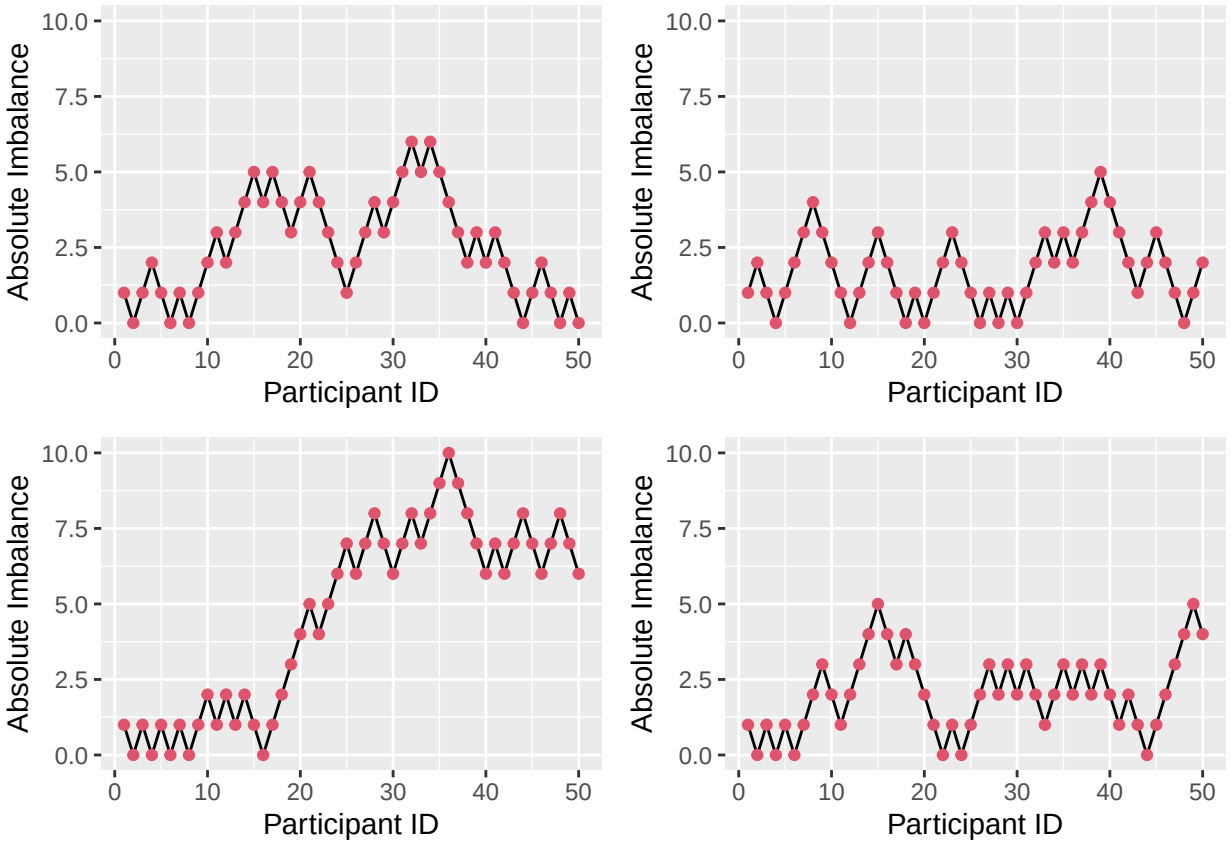


Figure 3.6: Four realisations of absolute imbalance for r=1, s=1, N=50.

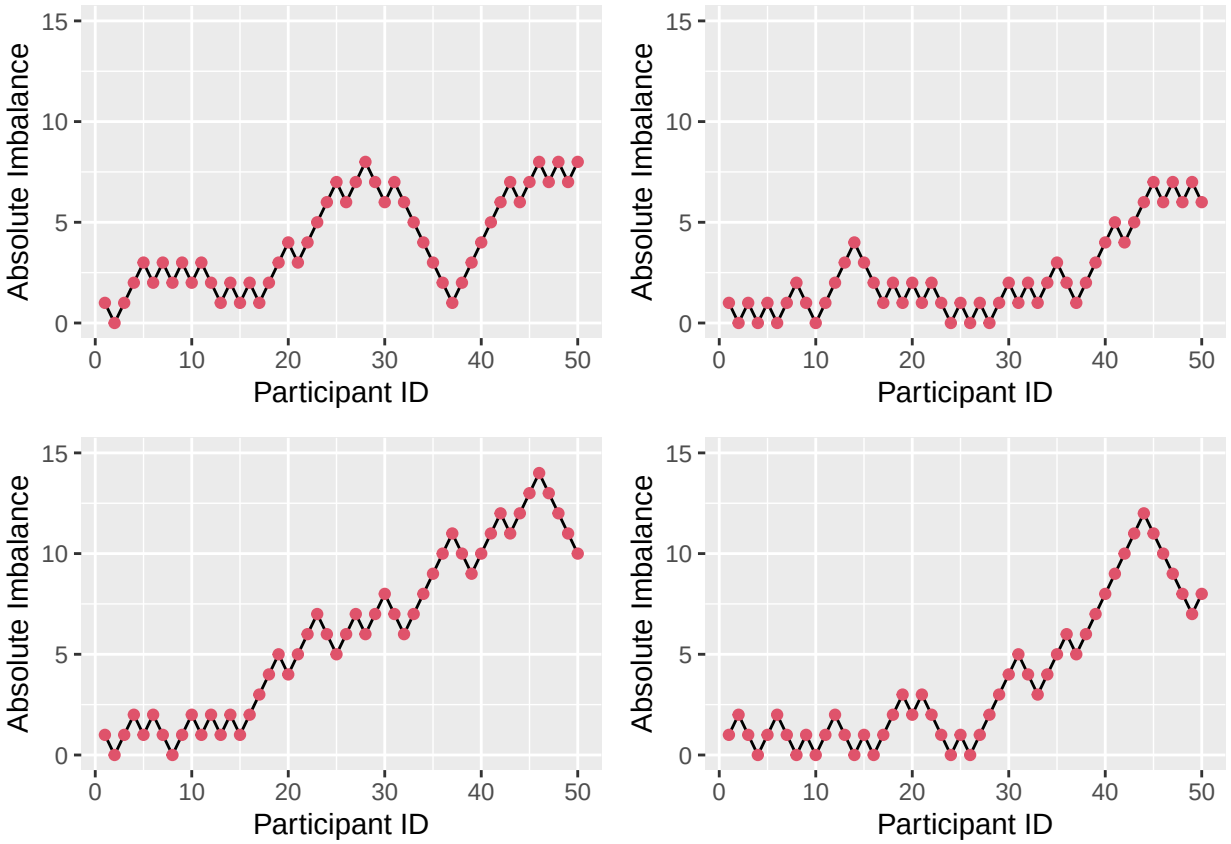


Figure 3.7: Four realisations of absolute imbalance for r=1, s=8, N=50.

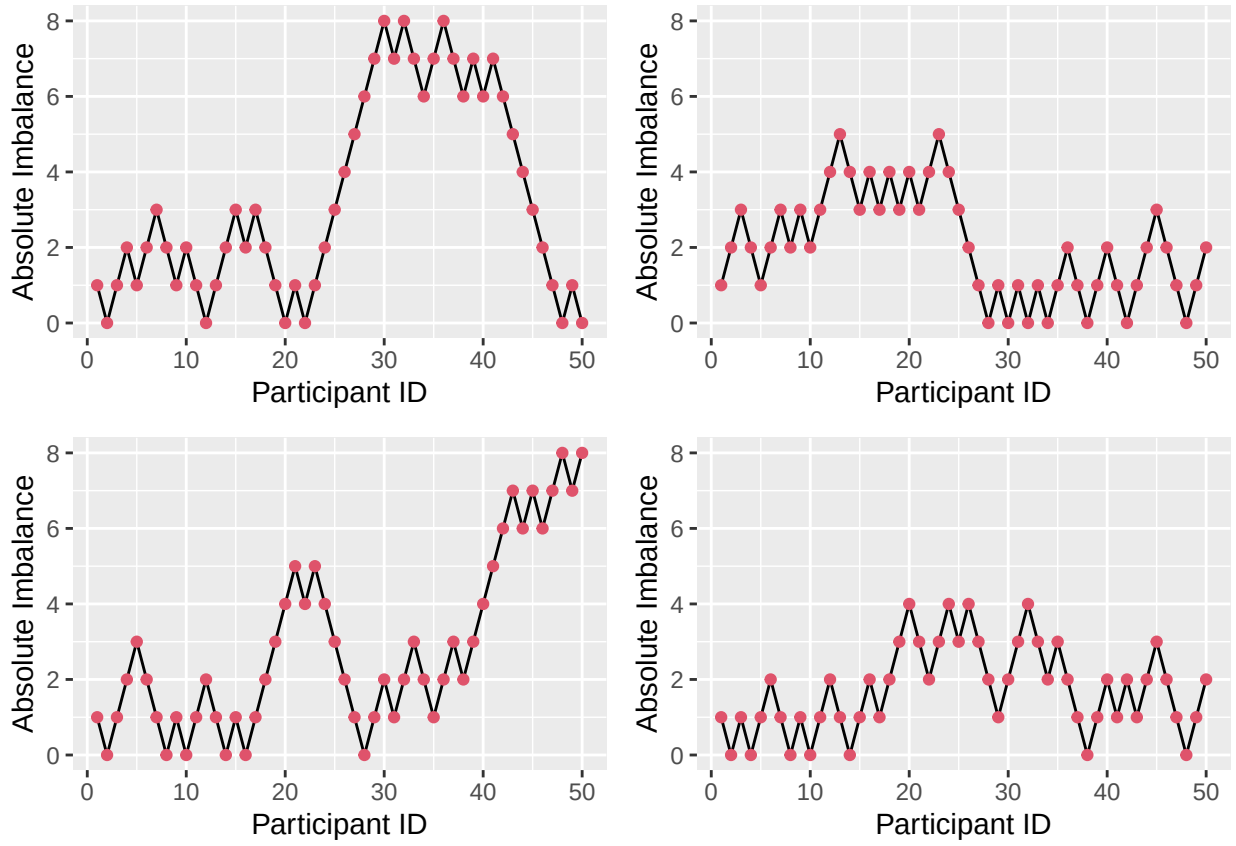


Figure 3.8: Four realisations of absolute imbalance for r=8, s=1, N=50.

3.4 Stratified sampling

The usual method of achieving balance with respect to prognostic factors is to divide each factor into several levels and to consider treatment assignment separately for patients having each particular combination of such factor levels. Such groups of patients are commonly referred to as randomization groups or strata. Treatment assignment is performed entirely separately for each stratum, a permuted block design of the type mentioned above often being used. In fact, using purely random treatment assignment for each stratum is equivalent to simple random assignment, so that some equalization of treatment numbers within each stratum is essential. Both the biased coin design and the urn design were intended for use in this way, adjusting in relation to imbalance within each stratum independently. This whole procedure is analogous to performing a factorial experiment, without being able to control the factor levels of the experimental units.

Example 3.4. Suppose we are planning a trial involving people over the age of 50, and we anticipate that age and sex might both play an important role in how participants respond to the treatment.

For sex, we use the levels 'male' and 'female', and for age we split the range into 50-65, 66-80 and 81 or over. We therefore have six strata, and we use an allocation strategy independently in each stratum. For example, below we have used randomly permuted blocks of length four.

	Male	Female
50-65	ABAB BBAA ...	ABBA BBAA ...
66-80	BAAB AABB ...	BABA BAAB ...
81 and over	ABAB ABBA ...	ABBA BAAB ...

Each time a new participant arrives, we follow the randomization pattern for their stratum. We could use another allocation scheme within each stratum, for example an urn model or a biased coin. It is important that we use one that aims to conserve balance, or else the benefits of stratification are lost.

A difficulty with stratified sampling is that the number of strata can quickly become large as the number of factors (or the number of levels within some factors) increases. For example, if we have four prognostic factors each with three levels, there are $3^4 = 81$ strata. This creates a situation that is at best unwieldy, and at worst completely unworkable; in a small trial (with say 100 patients in each arm) there may be some strata with no patients in (this is actually not a problem), and probably many more with only one (this is much more problematic).

3.5 Minimization

Minimization was first proposed by Taves (1974), then shortly after by Pocock and Simon (1975) and Freedman and White (1976). The aim of minimization is to minimize the difference between the two groups. It was developed for use with strata, as an alternative to randomly permuted blocks. Although the method was developed in the seventies, it has only gained popularity relatively recently, mainly as computers have become widely available.

To form the strata, the people running the trial must first specify all of the factors they would like to be balanced between the two groups. These should be any variables that are thought to possibly affect the outcome.

Example 3.5. The study by Kallis et al. (1994) investigates the effect of giving aspirin to patients before coronary artery surgery with giving them a placebo. Interestingly, the effects of aspirin were found to be both positive (decreases platelet aggregation to arachidonic acid and to collagen) and negative (increased likelihood of post-operative excessive blood loss). For their prognostic factors, Kallis et al. (1994) chose age (≤ 50 or 50), sex (M or F), operating surgeon (3 possibilities) and number of coronary arteries affected (1 or 2). This creates 24 strata. The trial had 100 participants, meaning an average of 4.17 in each stratum.

When a patient enters the trial, their level of each factor is listed. The patient is then allocated in such a way as to minimise any difference in these factors between the two groups. The minimization method has evolved since its conception, and exists in several forms. Two areas in which methods vary are

- Whether continuous variables have to be binned
- Whether there is any randomness

It is generally agreed that if the risk of selection bias cannot be avoided, there should be an element of randomness. It is also usually accepted that if a variable is included in the minimization, it should also be included in the statistical analysis.

3.5.1 Minimization algorithm

Suppose we have a trial in which patients are recruited sequentially and need to be allocated to a trial arm (of which there are two). Pocock and Simon (1975) give an algorithm in the general case of N treatment arms, but we will not do that here.

Suppose there are several prognostic factors over which we require balance, and that these factors have $I, J, K, ...$ levels. In our example above, there would be $I = 2,\ J = 2,\ K = 3,\ L = 2$. Note that this equates to 24 strata.

At some point in the trial, suppose we have recruited n_{ijkl} patients with levels i, j, k, l of the factors. For example, this may be males, aged over 50, assigned to the second surgeon, with both coronary arteries affected. Within these, n^A_{ijkl} have been assigned to treatment arm A, and n^B_{ijkl} to arm B. So we have

$$n^A_{ijkl} + n^B_{ijkl} = n_{ijkl}.$$

If we were to use random permuted blocks within each stratum, then we would be assured that

$$|n^A_{ijkl} - n^B_{ijkl}| \leq \frac{1}{2}b,$$

where b is the block length. However, there are two issues with this:

- There may be very few patients in some strata, in which case RPBs will fail to provide adequate balance.
- It is unlikely that we actually need this level of balance.

The first point is a pragmatic one - the method usually guaranteed to achieve good balance is likely to fail, at least for some strata. The second is more theoretical. In general, we require that groups be balanced according to each individual prognostic factor, but not to interactions. For example, it is often believed that younger patients would have generally better outcomes, but that other factors do not systematically affect this difference.

Therefore, it is enough to make sure that the following are all small:

$$\begin{gathered} |n^A_{i+++} - n^B_{i+++}| \text{ for each } i = 1, \ldots, I \\ |n^A_{+j++} - n^B_{+j++}| \text{ for each } j = 1, \ldots, J \\ \ldots \end{gathered}$$

where $+$ represents summation over the other factors, so that for example

$$n^A_{++k+} = \sum_{i,j,l} n^A_{ijkl}$$

Table 3.1: Allocations of first 15 patients, divided by diagnostic factor

factor	level	Mustine (A)	Talc (B)
Age	1. 50 or younger	3	4
Age	2. >50	4	4
Stage	1. I or II	1	2
Stage	2. III or IV	6	6
Time interval	1. 30 months or less	4	2
Time interval	2. >30 months	4	5
Menopausal status	1. Pre	4	3
Menopausal status	2. Post	5	3

is the total number of patients with level k of that factor assigned to treatment arm A.

Therefore, instead of having $IJKL$ constraints constraints, as we would with using randomly permuted blocks within each stratum, we have $I + J + K + L$ constraints, one for each level of each factor. In our example this is 9 constraints rather than 24.

In order to implement minimisation, we follow these steps:

1. Allocate the first patient by simple randomisation.
2. Suppose that at some point in the trial we have recruited n_{ijkl} patients with prognostic factors i, j, k, l. Of these n^A_{ijkl} are allocated to treatment arm A and n^B_{ijkl} to arm B.
3. A new patient enters the trial. They have prognostic factors at levels w, x, y, z.
4. We form the sum

$$\left(n^A_{w+++} - n^B_{w+++}\right) + \left(n^A_{+x++} - n^B_{+x++}\right) + \left(n^A_{++y+} - n^B_{++y+}\right) + \left(n^A_{+++z} - n^B_{+++z}\right). \tag{3.1}$$

5. If the sum from step 4 is negative (that is, allocation to arm B as dominated up to now) then we allocate the new patient to arm A with probability P, with $P > 0.5$. If the sum is positive, they are allocated to arm B with probability P. If the sum is zero, they are allocated to arm A with probability $\frac{1}{2}$.

Some people set $P = 1$, whereas others would set $\frac{1}{2} < P < 1$ to retain some randomness. Although setting $P = 1$ makes the system deterministic, to predict the next allocation a doctor (or whoever) would need to know n^A_{i+++} and so on. This is very unlikely unless they are deliberately seeking to disrupt the trial. However, generally the accepted approach is becoming to set $P < 1$.

Example 3.6. From Altman (1990) (citing Fentiman et al. (1983)). In this trial, 46 patients with breast cancer were allocated to receive either Mustine (arm A) or Talc (arm B) as treatment for pleural effusions (fluid between the walls of the lung). They used four prognostic factors: age (≤ 50 or > 50), stage of disease (I or II, III or IV), time in months between diagnosis of breast cancer and diagnosis of pleural effusions (≤ 30 or > 30) and menopausal status (Pre or post).

Let's suppose that 15 patients have already been allocated. The totals of patients in each treatment arm in terms of each level of each prognostic factor are shown in Table 3.1.

Suppose our sixteenth patient is under 50, has disease at stage III, has less than 30 months between diagnoses and is pre-menopausal. Our calculation from step 4 of the minimisation algorithm is therefore

$$\begin{aligned}\left(n^A_{1+++} - n^B_{1+++}\right) + \left(n^A_{+2++} - n^B_{+2++}\right) + \left(n^A_{++1+} - n^B_{++1+}\right) + \left(n^A_{+++1} - n^B_{+++1}\right) &= (3-4) + (6-6) + (4-2) + (4-3) \\ &= -1 + 0 + 2 + 1 \\ &= 2.\end{aligned}$$

Since our sum is greater than zero, we allocate the new patient to arm B (talc) with some probability $P \in (0.5, 1)$ and update the table before allocating patient 17.

One shortcoming of minimisation is that the factors are equally weighted in the algorithm, regardless of the number of patients with that particular factor level. For example, suppose at some later stage of allocation in Example 3.6, only four patients with stage I or II disease had been recruited, and that one of these had been allocated to group A and three to group B. At the same point, 18 of the recruited number were post-menopausal, and of these 10 had been allocated to group A and 8 to group B. The values contributed to the sum in Equation (3.1) are $+2$ and -2, so these imbalances effectively cancel one another out, but intuitively it would feel sensible to prioritise equal distribution within the stage I or II women, since proportionally this stratum is less balanced. Wei (1978) proposed an extension of the Urn Design that does exactly this, but we won't cover this method in our course.

3.6 Problems around allocation

In clinical trials papers, the allocation groups are usually summarised in tables giving summary statistics (eg. mean and SD) of each characteristic for the control group and the intervention group. The aim of these is to show that the groups are similar enough for any difference in outcome to be attributed to the intervention itself. Figure 3.9 shows an example, taken from Ruetzler et al. (2013).

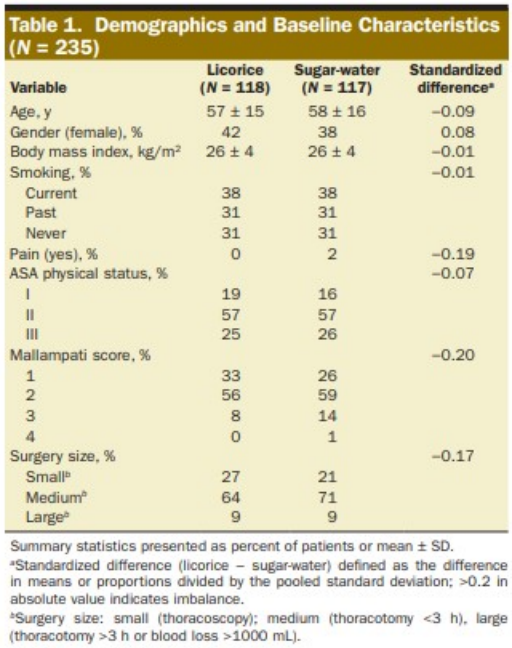


Table 1. Demographics and Baseline Characteristics (*N* = 235)

Variable	Licorice (*N* = 118)	Sugar-water (*N* = 117)	Standardized difference[a]
Age, y	57 ± 15	58 ± 16	−0.09
Gender (female), %	42	38	0.08
Body mass index, kg/m²	26 ± 4	26 ± 4	−0.01
Smoking, %			−0.01
Current	38	38	
Past	31	31	
Never	31	31	
Pain (yes), %	0	2	−0.19
ASA physical status, %			−0.07
I	19	16	
II	57	57	
III	25	26	
Mallampati score, %			−0.20
1	33	26	
2	56	59	
3	8	14	
4	0	1	
Surgery size, %			−0.17
Small[b]	27	21	
Medium[b]	64	71	
Large[b]	9	9	

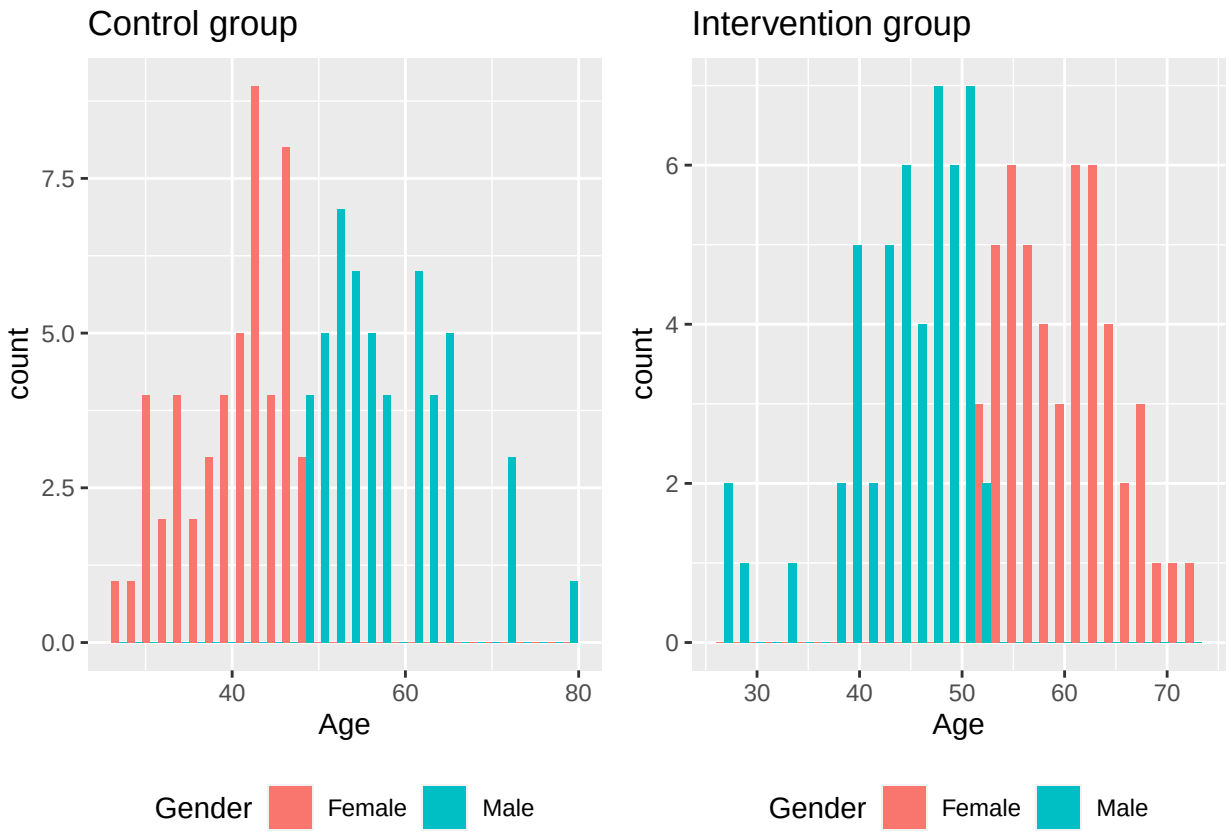
Summary statistics presented as percent of patients or mean ± SD.
[a]Standardized difference (licorice – sugar-water) defined as the difference in means or proportions divided by the pooled standard deviation; >0.2 in absolute value indicates imbalance.
[b]Surgery size: small (thoracoscopy); medium (thoracotomy <3 h), large (thoracotomy >3 h or blood loss >1000 mL).

Figure 3.9: Summary statistics for an RCT comparing a licorice gargle (the intervention) to a sugar-water gargle (the standard). From @ruetzler2013randomized

The problem here is that only the marginal distributions are compared for similarity. Consider the following (somewhat extreme and minimalistic) scenario. A study aims to investigate the effect of some treatment, and to balance for gender and age in their allocation, resulting in the following summary table.

	Male	Female
Control	57.51 (7.09)	40.31 (5.83)
Intervention	44.19 (5.96)	60.03 (5.27)

This appears to be a reasonably balanced design. However, if we look at the joint distribution, we see that there are problems.



If the intervention is particularly effective in older men, our trial will not notice. Likewise, if older women generally have a more positive outcome than older men, our trial may erroneously find the intervention to be effective.

Although this example is highly manufactured and [hopefully!] unlikely to take place in real life, for clinical trials there are often many demographic variables and prognostic factors being taken into account. Achieving joint balance across all them is very difficult, and extremely unlikely to happen if it isn't aimed for. Treasure and MacRae (1998) give an example in relation to a hypothetical study on heart disease

> Supposing one group has more elderly women with diabetes and symptoms of heart failure. It would then be impossible to attribute a better outcome in the other group to the beneficial effects of treatment since poor left ventricular function and age at outset are major determinants of survival in any longitudinal study of heart disease, and women with diabetes, as a group, are likely to do worse. At this point the primary objective of randomisation—exclusion of confounding factors—has failed. ... If a very big trial fails, because, for example, the play of chance put more hypertensive smokers in one group than the other, the tragedy for the trialists, and all involved, is even greater.

However, this issue is rarely addressed in clinical trials: a lot of faith is placed (with reasonable justification) in the likely balance achieved by random sampling, whatever method is used. We will also see in the next Chapter that we can account for some degree of imbalance at the analysis stage.

Chapter 4

Analyzing RCT data

We're now in the post-trial stage. The trial has been run, and we have lots of data to analyze to try to assess what effect the treatment or intervention has had. In general we will use the notation τ to denote the treatment effect.

In this chapter we'll keep our focus on the scenario where the trial outcome is measured on a continuous scale, but in later weeks we'll go on to look at other types of data.

Example 4.1. To illustrate the theory and methods, we'll use an example dataset from Hommel et al. (1986) (this example is also used by Matthews (2006)). The data involves a trial of 16 diabetes patients, and focusses on a drug (Captopril) that may reduce blood pressure. This is important, since for those with diabetes, high blood pressure can exacerbate kidney disease (specifically diabetic nephropathy, a complication of diabetes). To participate in the trial, people had to be insulin-dependent and already affected by diabetic nephropathy. In the trial, systolic blood pressure was measured before participants were allocated to each trial arm, and then measured again after one week on treatment. A placebo was given to the control group, so that all participants were blinded.

The baseline and outcome blood pressure measurements (in mmHg) are shown in Table 4.1. We see that nine participants were assigned to the treatment arm (Captopril) and the remaining seven to the placebo group. Hommel et al. (1986) say that the patients were 'randomly allocated' to their group.

This is very small dataset, and so in that respect it is quite unusual, but its structure is similar to many other trials.

We will build up from the simplest type of analysis to some more complicated / sophisticated approaches.

4.1 Confidence intervals and P-values

Because the randomization process should produce groups that are comparable, we should in principle be able to compare the primary outcome (often referred to as X) between the groups.

Example 4.2. Summary statistics of the outcome for each group are shown below.

We see that the difference in mean outcome (systolic blood pressure) between the two groups is $141.86 - 135.33 = 6.53$mmHg. Clearly overall there has been some reduction in systolic blood pressure for those in the Captopril arm, but how statistically sound is this as evidence? It could be that really (for the hypothetical population) there is no reduction, and we have just been 'lucky'.

The variances within the two groups are fairly close, so we can use the pooled estimate of standard deviation:

Table 4.1: Data for the Captopril trial from @hommel1986effect.

Patient (ID)	Baseline (B)	Outcome at 1 week (X)	Trial Arm
1	147	137	Captopril
2	129	120	Captopril
3	158	141	Captopril
4	164	137	Captopril
5	134	140	Captopril
6	155	144	Captopril
7	151	134	Captopril
8	141	123	Captopril
9	153	142	Captopril
1	133	139	Placebo
2	129	134	Placebo
3	152	136	Placebo
4	161	151	Placebo
5	154	147	Placebo
6	141	137	Placebo
7	156	149	Placebo

Table 4.2: Summary statistics for each group.

	Sample Size	Mean (mmHg)	SD (mmHg)	SE of mean (mmHg)
Captopril	9	135.33	8.43	2.81
Placebo	7	141.86	6.94	2.62

$$s_p = \sqrt{\frac{\sum_{i=1}^{N} (n_i - 1)\, s_i^2}{\sum_{i-1}^{N} (n_i - 1)}}.$$

In our case

$$\begin{aligned} s_p &= \sqrt{\frac{8 \times 8.43^2 + 6 \times 6.94^2}{8 + 6}} \\ &= 7.82 \text{ mmHg}. \end{aligned}$$

This enables us to do an independent two-sample t-test, and we can find the t statistic

$$\begin{aligned} t &= \frac{\bar{X_C} - \bar{X_T}}{s_p \sqrt{\frac{1}{n_C} + \frac{1}{n_T}}} \\ &= \frac{6.53}{7.82 \sqrt{\frac{1}{7} + \frac{1}{9}}} \\ &= 1.65. \end{aligned}$$

Note that here the placebo group is group C, and the Captopril group is group T.

Under the null hypothesis that the mean systolic blood pressure at the end of the week of treatment/placebo is the same in both groups, this value should have a t distribution with 14 degrees of freedom ($n_i - 1$ for each group).

The dashed line in Figure 4.1 is at $t = 1.65$, and the red shaded areas show anywhere 'at least as extreme'. We can find the area (ie. the probability of anything at least as extreme as our found value) in R by

```
2*(1-pt(1.65, df=14))
```

```
## [1] 0.1211902
```

This is the value we know as 'the P value'. We see that in this case our results are not statistically significant (at the 0.10 level), under this model.

4.1.1 What do we do with this outcome?

The outcome of this Captopril study is in some ways the worst case scenario. The difference in means is large enough to be compelling, but our dataset is too small for it to be statistically significant, and so we can't confidently conclude that Captopril has any effect on blood pressure. However, we also can't say that there is no effect. This is exactly the sort of scenario we hoped to avoid when planning our study.

One way to reframe the question is to consider the range of treatment effects that are compatible with our trial data. That is, we find the set

$$\left\{ \tau \mid \frac{|\bar{x}_C - \bar{x}_T - \tau|}{s\sqrt{n_C^{-1} + n_T^{-1}}} \leq t_{n_C + n_T - 2;\, 0.975} \right\},$$

which contains all possible values of treatment effect τ that are compatible with our data. That is, suppose the true treatment effect is τ^*, and we test the hypothesis that $\tau = \tau^*$. For all values of τ^* inside this range,

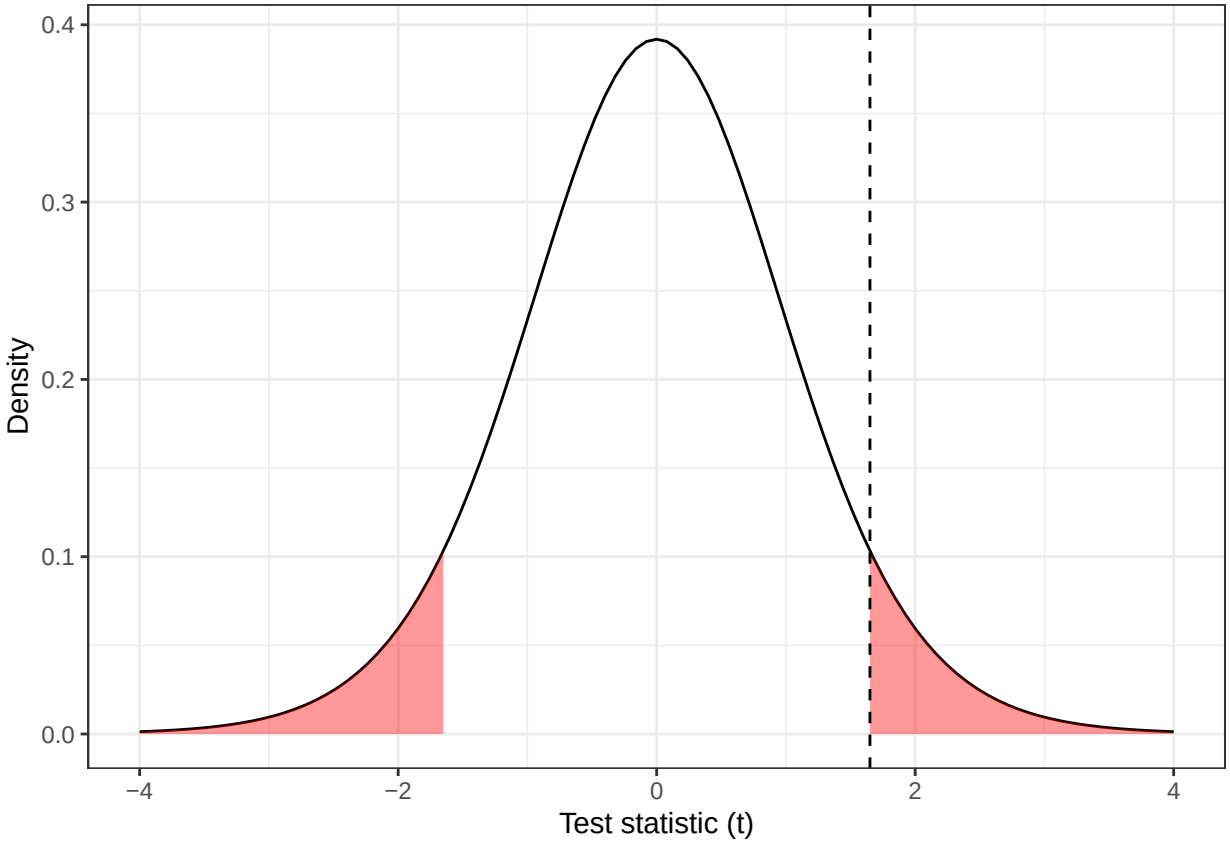


Figure 4.1: The distribution t_{14}, with $t = 1.65$ shown by the dashed line and the 'more extreme' areas shaded.

our data are not sufficiently unlikely to reject the hypothesis at the 0.05 level. However, for all values of τ^* outside this range, our data are sufficiently unlikely to reject that hypothesis. We can rearrange this to give a 95% confidence interval for τ,

$$\left\{\tau \mid \bar{x}_C - \bar{x}_T - t_{n_C+n_T-2;\,0.975}\, s\sqrt{n_C^{-1} + n_T^{-1}} \leq \tau \leq \bar{x}_C - \bar{x}_T + t_{n_C+n_T-2;\,0.975}\, s\sqrt{n_C^{-1} + n_T^{-1}}\right\}$$

Example 4.3. Continuing our example, we have

$$\left\{\tau \mid \frac{|6.53 - \tau|}{7.82\sqrt{\frac{1}{7} + \frac{1}{9}}} \leq t_{14;0.975} = 2.145\right\}$$

Here, $t_{14;0.975} = 2.145$ is the t-value for a significance level of 0.05, so if we were working to a different significance level we would change this.

Rearranging as above, this works out to be the interval

$$-1.92 \leq \tau \leq 14.98.$$

Notice that zero is in this interval, consistent with the fact that we failed to reject the null hypothesis.

Some things to note

- We can compute this confidence interval whether or not we failed to reject the null hypothesis that $\tau = 0$, and for significance levels other than 0.05.
- In most cases, reporting the confidence interval is much more informative than simply reporting the P-value. In our Captopril example, we found that a negative treatment effect (ie. Captopril reducing blood pressure less than the placebo) of more than 2 mmHg was very unlikely, whereas a positive effective (Captopril reducing blood pressure) of up to 15 mmHg was plausible. If Captopril were inexpensive and had very limited side effects (sadly neither of which is true) it may still be an attractive drug.
- These confidence intervals are exactly the same as you have learned before, but we emphasise them because they are very informative in randomised controlled trials (but not so often used!).

At the post trial stage, when we have data, the confidence interval is the most useful link to the concept of *power*, which we thought about at the planning stage. Remember that the power function is defined as

$$\psi(\tau) = P(\text{Reject } H_0 \mid \tau \neq 0),$$

that is, the probability that we successfully reject H_0 (that $\tau = 0$) given that there is a non-zero treatment effect $\tau \neq 0$. This was calculated in terms of the theoretical model of the trial, and in terms of some minimum detectable effect size τ_M that we wanted to be able to correctly detect with probability $1 - \beta$ (the power). Sometimes people attempt to re-calculate the power after the trial, to detect whether the trial was underpowered. However, now we have actual data. If we failed to reject H_0 and τ_M is in the confidence interval for τ, then that is a good indication that our trial was indeed underpowered.

4.2 Using baseline values

In our example above, our primary outcome variable X was the systolic blood pressure of each participant at the end of the intervention period. However, we see in Table 4.1 that we also have *baseline* measurements: measurements of systolic blood pressure for each patient from before the intervention period. Baseline measurements are useful primarily for two reasons:

Table 4.3: Data for the Captopril trial, with differences shown.

Patient (ID)	Baseline (B)	Outcome at 1 week (X)	Trial Arm	Difference
1	147	137	Captopril	-10
2	129	120	Captopril	-9
3	158	141	Captopril	-17
4	164	137	Captopril	-27
5	134	140	Captopril	6
6	155	144	Captopril	-11
7	151	134	Captopril	-17
8	141	123	Captopril	-18
9	153	142	Captopril	-11
1	133	139	Placebo	6
2	129	134	Placebo	5
3	152	136	Placebo	-16
4	161	151	Placebo	-10
5	154	147	Placebo	-7
6	141	137	Placebo	-4
7	156	149	Placebo	-7

Table 4.4: Summary statistics for each group.

	Sample Size	Mean (mmHg)	SD (mmHg)	SE of mean (mmHg)
Captopril	9	-12.67	8.99	3.00
Placebo	7	-4.71	7.91	2.99

1. They can be used to assess the balance of the design.
2. They can be used in the analysis.

We will demonstrate these by returning to our Captopril example.

Example 4.4. Firstly, we use the baseline systolic blood pressure to assess balance. The placebo group has a mean of 146.6 mmHg and an SD of 12.3 mmHg, whereas the Captopril group has mean 148.0 mmHg, SD 11.4 mmHg. While these aren't identical, they are sufficiently similar not to suspect any systematic imbalance. In a study this small there is likely to be some difference.

Secondly, since we are interested in whether the use of Captopril has reduced blood pressure for each individual, and these individuals had different baseline values, it makes sense to compare not just the outcome but the difference from baseline to outcome for each individual. We can see individual data in Table 4.3 and summary statistics in Table 4.4.

Now we can perform our test as before, in which case we find

$$t = \frac{-4.71 - (-12.67)}{8.54\sqrt{\frac{1}{7} + \frac{1}{9}}} = 1.850$$

where 8.54 is the pooled standard deviation (as before). Under the null distribution of no difference, this has a t-distribution with 14 degrees of freedom, and so we have a P-value of 0.086. Our 0.95 confidence interval is

$$-4.71 - (-12.67) \pm t_{14;\,0.975} \times 8.54\sqrt{\frac{1}{7} + \frac{1}{9}} = [-1.3,\, 17.2]\,.$$

We see that taking into account the baseline values in this way has slightly reduced the P-value and shifted the confidence interval slightly higher. Though at the $\alpha = 0.05$ level we still don't have significance.

We will now look into why the confidence interval and P-value changed in this way, before going on to another way of taking into account the baseline value.

Let's label the baseline measurement for each group B_C and B_T, and the outcome measurements X_C, X_T, where we will take group C to be the placebo/control group and group T to be the treatment group. Because all participants have been randomised from the same population, we have

$$\mathrm{E}(B_C) = \mathrm{E}(B_T) = \mu_B.$$

Assuming some treatment effect τ (which could still be zero) we have

$$\begin{aligned}\mathrm{E}(X_C) &= \mu \\ \mathrm{E}(X_T) &= \mu + \tau.\end{aligned}$$

Usually we will assume that

$$\mathrm{Var}(X_C) = \mathrm{Var}(X_T) = \mathrm{Var}(B_C) = \mathrm{Var}(B_T) = \sigma^2,$$

and this is generally fairly reasonable in practice.

Notice that for the two analyses we have performed so far (comparing outcomes and comparing differences) we have

$$\begin{aligned}\mathrm{E}(X_T) - \mathrm{E}(X_C) &= (\mu + \tau) - \mu = \tau \\ \mathrm{E}(X_T - B_T) - \mathrm{E}(X_C - B_C) &= (\mu - \mu_B + \tau) - (\mu - \mu_B) = \tau,\end{aligned}$$

that is, both are unbiased estimators of τ.

However, whereas the first is based on data with variance σ^2, the second has

$$\begin{aligned}\mathrm{Var}(X_T - B_T) &= \mathrm{Var}(X_T) + \mathrm{Var}(B_T) - 2\,\mathrm{cov}(X_T, B_T) \\ &= \sigma^2 + \sigma^2 - 2\rho\sigma^2 \\ &= 2\sigma^2(1-\rho),\end{aligned}$$

where ρ is the true correlation between X and B, and is assumed to be the same in either group. Similarly,

$$\mathrm{var}(X_C - B_C) = 2\sigma^2(1-\rho).$$

Using this to work out the variance of the estimator $\hat{\tau}$ we find that for comparing means, assuming two equally sized groups of size N, we have

$$\mathrm{var}(\hat{\tau}) = \mathrm{var}(\bar{x}_T - \bar{x}_C) = \frac{2\sigma^2}{N}.$$

whereas for comparing differences from baseline

$$\mathrm{var}(\hat{\tau}) = \mathrm{var}\left[\left(\overline{X_T - B_T}\right) - \left(\overline{X_C - B_C}\right)\right] = 2(1-\rho)\left(\frac{2\sigma^2}{N}\right).$$

Therefore, if $\frac{1}{2} < \rho \leq 1$ there will be a smaller variance when comparing differences. However, if $0 \leq \rho < \frac{1}{2}$, the variance will be smaller when comparing outcome variables.

Table 4.5: Summary statistics for the dodgy analysis

	T statistic	Deg of freedom	P-value
Captopril	4.23	8	0.003
Placebo	1.58	6	0.170

Intuitively, this seems reasonable: if the correlation between baseline and outcome measurements is very strong, then we can remove some of the variability between participants by taking into account their baseline measurement. However, if the correlation is weak, then by including the baseline in the analysis we are essentially just introducing noise.

For our Captopril example, the sample correlation between baseline and outcome is 0.63 in the Captopril group and 0.80 in the Placebo group. This fits with the P-value having reduced slightly.

4.2.1 A dodgy way to use baseline variables

Sometimes the analysis performed on a dataset is rather spurious, but it isn't always immediately obvious why. We'll look at one example now, because it is done sometimes.

This approach involves looking at each group separately, and determining whether there has been a significant change in the outcome variable (note that this only 'works' if $\mu_B = \mu$).

For example, with our Captopril data, we could perform a paired t-test on the difference between baseline B and outcome X for each patient, for each group.

If we do this, we find the summary statistics in Table 4.5.

From this we see that there is strong evidence for a change in blood pressure for the Captopril patients (group T), which isn't surprising, and no such evidence for the placebo patients. Can we therefore conclude that Captopril is significantly better than the placebo? No! The analysis is flawed:

- The p-value of 0.17 in the control group doesn't show that the null hypothesis (no treatment effect for the control group) is true, just that we can't reject the null hypothesis. It is quite possible that there is a difference in the control group, and that numerically it could even be comparable to that in the treatment group, so although we can say that there is a significant reduction in blood pressure for the captopril group, we can't conclude that Captopril is better than the placebo.
- Having set up the experiment as a randomised controlled trial, with a view to comparing the two groups, it seems strange to then deal with them separately.

4.3 Analysis of covariance (ANCOVA)

In the previous section we based our analysis on the baseline values being statistically identical draws from the underlying distribution, and therefore having the same expectation and variance.

However, although this is theoretically true, in real life trials there will be some imbalance in the baseline measurements for the different treatment arms. We can see this in our Captopril example, in Figure 4.2.

The baseline measurements are not identical in each group. Indeed, we saw earlier that the means differ by 1.4 mmHg. Although this isn't a clinically significant difference, or a large enough difference to make us doubt the randomisation procedure, it is still a difference.

The basic principle of ANCOVA is that if there is some correlation between the baseline and outcome measurements, then if the baseline measurements differ, one would expect the outcome measurements to differ, even if there is no treatment effect (ie. if $\tau = 0$). Indeed, how do we decide how much of the difference

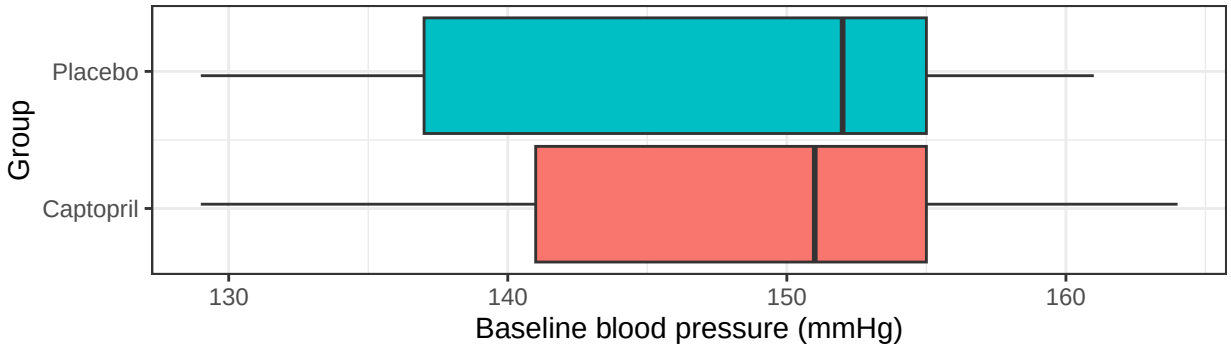


Figure 4.2: Baseline measurements from the Captopril trial.

in outcome is down to the treatment itself, and how much is simply the difference arising from different samples?

This issue arises in many trials, particularly where there is a strong correlation between baseline and outcome measurements.

4.3.1 The theory

Suppose the outcome for a clinical trial is X and the baseline is B. X has mean μ in the control group (C) and mean $\mu + \tau$ in the test group (T), and as usual our aim is to determine the extent of τ, the treatment effect. We suppose also that X has variance σ^2 in both groups.

The same quantity is measured at the start of the trial, and this is the baseline B, which we can assume to have true mean μ_B in both groups (because of randomisation) and variance σ^2. We also assume that the true correlation between B and X is ρ in each group. Finally, we assume that both treatment groups are of size N.

We therefore have $2N$ patients, and so we observe baseline measurements $b_1, b_2, \ldots, b_{2N}$. Given these values, we have

$$\begin{aligned} \mathrm{E}\left(X_i \mid b_i\right) &= \mu + \rho\left(b_i - \mu_B\right) \text{ in the control group} \\ \mathrm{E}\left(X_i \mid b_i\right) &= \mu + \tau + \rho\left(b_i - \mu_B\right) \text{ in the test group.} \end{aligned}$$

From this, we find that

$$\mathrm{E}\left(\bar{X}_T - \bar{X}_C \mid \bar{b}_T, \bar{b}_C\right) = \tau + \rho\left(\bar{b}_T - \bar{b}_C\right). \tag{4.1}$$

That is, if there is a difference in the baseline mean between the control and test groups, then the difference in outcome means is not an unbiased estimator of the treatment effect τ. Assuming $\rho > 0$ (which is almost always the case) then if $\bar{b}_T > \bar{b}_C$ the difference in outcome means overestimates τ. Conversely, if $\bar{b}_T < \bar{b}_C$, the difference in outcome means underestimates τ. The only situation in which the difference in outcome means is an unbiased estimator is when $\rho = 0$, however this is not common in practice.

Comparing the difference between outcome and baseline, as we did in 4.2, does not solve this problem, since we have

$$\mathrm{E}\left[\left(\bar{X}_T - \bar{b}_T\right) - \left(\bar{X}_C - \bar{b}_C\right) \mid \bar{b}_T, \bar{b}_C\right] = \tau + (\rho - 1)\left(\bar{b}_T - \bar{b}_C\right),$$

which is similarly biased (unless $\rho = 1$, which is never the case).

Notice, however, that if we use as our estimator

$$\hat{\tau} = \left(\bar{X}_T - \bar{X}_C\right) - \rho\left(\bar{b}_T - \bar{b}_C\right) \tag{4.2}$$

then, following from Equation (4.1) we have

$$\mathrm{E}\left[\left(\bar{X}_T - \bar{X}_C\right) - \rho\left(\bar{b}_T - \bar{b}_C\right) \mid \bar{b}_T, \bar{b}_C\right] = \tau + \rho\left(\bar{b}_T - \bar{b}_C\right) - \rho\left(\bar{b}_T - \bar{b}_C\right) = \tau.$$

4.3.1.1 What's the variance of this estimator?

To work out the variance of $\hat{\tau}$ in Equation (4.2) we need to think about bivariate normal variables.

Let's suppose that random variables X and Y are jointly normally distributed with correlation ρ

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right). \tag{4.3}$$

From Equation (4.3), we know that $\mathrm{E}(Y) = \mu_Y$.

But, if we have observed $X = x$, this gives us some information about likely values of Y: if $\rho > 0$ then a lower value of x should lead us to expect a lower value of Y, for example. Figure 4.3 shows

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1.5 \\ 1.5 & 3 \end{pmatrix} \right). \tag{4.4}$$

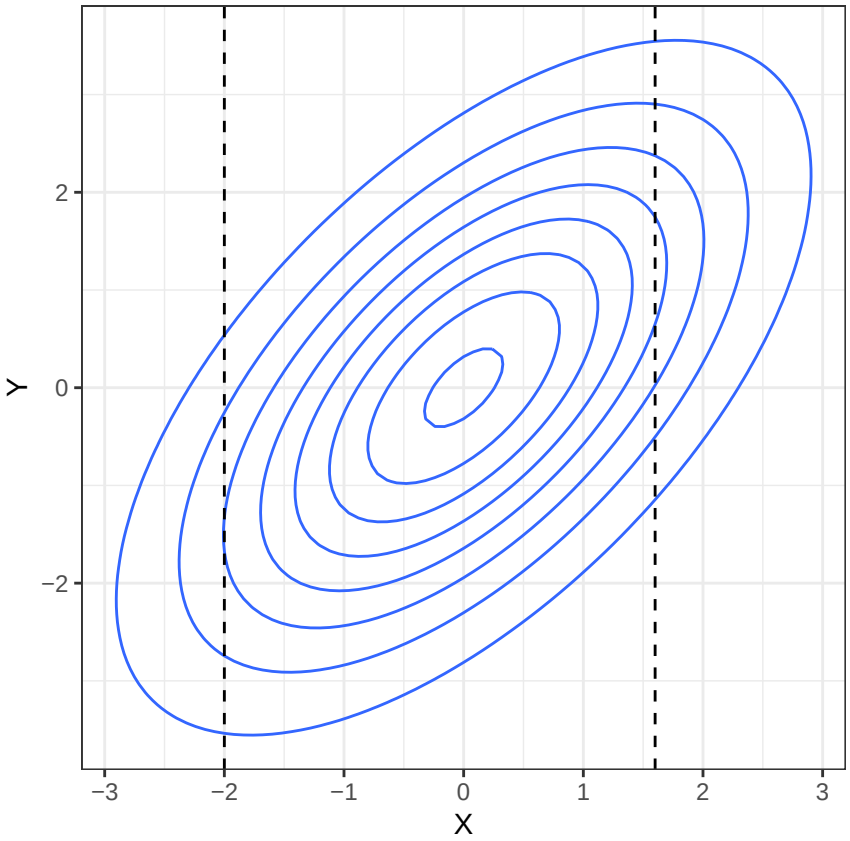


Figure 4.3: A bivariate normal density.

The higher the value of ρ (in magnitude), the more the conditional distribution of Y given an observed value of x deviates from the marginal distribution of Y (in our example, $N(0,\ 3)$). In particular,

$$\mathrm{E}(Y \mid X = x) \neq \mathrm{E}(Y).$$

If we have another random variable, W, that is independent of Y (and note that if two normally distributed variables are uncorrelated, they are also independent), then observing $W = w$ doens't give us any information about the distribution of Y, so we have

$$\mathrm{E}(Y \mid W = w) = \mathrm{E}(Y).$$

We can combine this information to work out $\mathrm{E}(Y \mid X = x)$. Firstly, we'll calculate the covariance of X and $Y - kX$, for some constant k. We can find this by

$$\begin{aligned}\operatorname{cov}(X,\, Y-kX) &= \mathrm{E}\left[(X-\mu_X)\left(Y-kX-\mu_Y+k\mu_X\right)\right] \\ &\quad \text{(using that } \operatorname{cov}(\mathrm{X},\mathrm{Y}) = \mathrm{E}\left[(\mathrm{X}\text{- }\mathrm{E}(\mathrm{X}))\,(\mathrm{Y}\text{- }\mathrm{E}(\mathrm{Y}))\right] \\ &= \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y) - k(X-\mu_X)^2\right] \\ &= \rho\sigma_X\sigma_Y - k\sigma_X^2.\end{aligned}$$

If we set

$$k = \beta = \frac{\rho\sigma_Y}{\sigma_X}$$

then $\operatorname{cov}(X,\, Y-\beta X) = 0$, and since $Y-\beta X$ is also normally distributed, this means that X and $Y-\beta X$ are independent. Therefore we have

$$\mathrm{E}(Y-\beta X \mid X=x) = \mathrm{E}(Y-\beta X) = \mu_Y - \beta\mu_X.$$

However, since we're conditioning on an observed value of $X=x$ we can take X to be fixed at this value, and so $\mathrm{E}(\beta X \mid X=x) = \beta x$. Finally, this allows us to calculate

$$\begin{aligned}\mathrm{E}(Y \mid X=x) &= \mathrm{E}(\beta X \mid X=x) + \mu_Y - \beta\mu_X \\ &= \mu_Y + \beta(x-\mu_X).\end{aligned}$$

We can use the same idea to find $\operatorname{var}(Y \mid X=x)$.

Recall that $\operatorname{var}(Y) = \mathrm{E}\left[Y^2\right] - [\mathrm{E}(Y)]^2$, and so

$$\operatorname{var}(Y \mid X=x) = \mathrm{E}\left(Y^2 \mid X=x\right) - [\mathrm{E}(Y \mid X=x)]^2. \tag{4.5}$$

We already know the second term, and we can find the first term using the same idea as before, this time noting that X and $(Y-\beta X)^2$ are independent.

From this, and using the fact that (for example)

$$\begin{aligned}\operatorname{var}(X) &= \mathrm{E}\left(X^2\right) - [\mathrm{E}(X)]^2 \\ \text{and therefore} \\ \mathrm{E}\left(X^2\right) &= \sigma_X^2 + \mu_X^2,\end{aligned}$$

we find that

$$\mathrm{E}\left[(Y-\beta X)^2 \mid X=x\right] = \mathrm{E}\left[(Y-\beta X)^2\right] = S^2 + (\mu_Y - \beta\mu_X)^2, \tag{4.6}$$

where $S^2 = \sigma_Y^2 + \beta^2\sigma_X^2 - 2\beta\rho\sigma_X\sigma_Y = \sigma_Y^2\left(1-\rho^2\right)$ (by plugging in $\beta = \frac{\rho\sigma_Y}{\sigma_X}$).

If we multiply out the left-hand side of Equation (4.6), we find that this is the same as

$$\mathrm{E}\left[Y^2 \mid X=x\right] - 2\beta\,\mathrm{E}(Y \mid X=x) + \beta^2x^2 = \mathrm{E}\left[Y^2 \mid X=x\right] - 2\beta x(\mu_Y - \beta\mu_X) - \beta^2x^2.$$

Equating this with Equation (4.6) and rearranging, we find

$$\mathrm{E}\left[Y^2 \mid X=x\right] = S^2 + (\mu_Y - \beta\mu_X)^2 + 2\beta x(\mu_Y - \beta\mu_X) + \beta^2x^2.$$

Now we can expand out

$$\mathrm{E}(Y \mid X=x) = \mu_Y + \beta(x-\mu_X) = (\mu_Y - \beta\mu_X) + \beta x$$

to find

$$[\mathrm{E}(Y \mid X = x)]^2 = (\mu_Y - \beta\mu_X)^2 + 2\beta x(\mu_Y - \beta\mu_X) + \beta^2 x^2.$$

Finally (!) we can use these two expressions to find

$$\begin{aligned}\mathrm{var}(Y \mid X = x) &= \mathrm{E}\left[Y^2 \mid X = x\right] - [\mathrm{E}(Y \mid X = x)]^2 \\ &= S^2 \\ &= \sigma_Y^2\left(1 - \rho^2\right).\end{aligned}$$

One thing to notice is that this conditional variance of doesn't depend on the observed value of $X = x$. It can also never exceed σ_Y^2, and is only equal to σ_Y^2 if X and Y are uncorrelated.

Back to our estimator!

Recall that in ANCOVA our estimator of the treatment effect τ is

$$\hat{\tau} = \left(\bar{X}_T - \bar{X}_C\right) - \rho\left(\bar{b}_T - \bar{b}_C\right)$$

and that we have

$$\mathrm{cor}\left(\bar{X}_T - \bar{X}_C,\ \bar{b}_T - \bar{b}_C\right) = \rho.$$

Therefore, using the result we just found,

$$\begin{aligned}\mathrm{var}(\hat{\tau}) = \mathrm{var}\left[\left(\bar{X}_T - \bar{X}_C\right) - \rho\left(\bar{b}_T - \bar{b}_C\right) \mid \bar{b}_T, \bar{b}_C\right] &= \mathrm{var}\left[\left(\bar{X}_T - \bar{X}_C\right) \mid \bar{b}_T, \bar{b}_C\right] \\ &= \mathrm{var}\left(\bar{X}_T - \bar{X}_C\right)\left(1 - \rho^2\right) \\ &= \frac{2\sigma^2}{N}\left(1 - \rho^2\right).\end{aligned}$$

Notice that unlike our first estimator that used baseline values, in Section 4.2, the variance of the ANCOVA estimate can never exceed $\frac{2\sigma^2}{N}$; if the baseline and outcome are uncorrelated, ANCOVA will perform as well as the t-tests we covered in Sections 4.1 and 4.2.

Borm et al. (2007) discuss how this reduction in $\mathrm{var}(\hat{\tau})$ can impact our sample size calculations.

4.3.2 The practice

In the previous section we established an unbiased estimate of the treatment effect that takes into account the baseline measurements. However, we can't use it as a model, because there are a few practical barriers:

- Our estimate for τ relies on the correlation ρ, which is unknown
- In real life, the groups are unlikely to have equal size and variance, so ideally we'd lose these constraints

We can solve both of these by fitting the following statistical model to the observed outcomes x_i:

$$\begin{aligned}x_i &= \mu + \gamma b_i + \epsilon_i && \text{in group C} \\ x_i &= \mu + \tau + \gamma b_i + \epsilon_i && \text{in group T.}\end{aligned}$$

Here, the ϵ_i are independent errors with distribution $N\left(0, \sigma_\epsilon^2\right)$, the b_i are the baseline measurements for $i = 1, \ldots, N_T + N_C$, for groups T and C with sizes N_T and N_C respectively. Sometimes this is written instead in the form

$$x_i = \mu + \tau G_i + \gamma b_i + \epsilon_i$$

where G_i is 1 if participant i is in group T and 0 if they're in group C. This is a factor variable, which you may remember from Stats Modelling II (if you took it). If $G_i = 1$ (ie. participant i is in group T) then τ is added. If $G_i = 0$ (ie. participant i is in group C) then it isn't. Notice that μ is no longer the mean outcome in any meaningful sense, but is the intercept of the linear model.

We now have four parameters to estimate: μ, τ, γ and σ_ϵ^2. For the first three we can use least squares (as you have probably seen for linear regression). Our aim is to minimise the sum of squares

$$S(\mu, \tau, \gamma) = \sum_{i \text{ in } T} (x_i - \mu - \tau - \gamma b_i)^2 + \sum_{i \text{ in } C} (x_i - \mu - \gamma b_i)^2 .$$

This leads to estimates $\hat{\mu}$, $\hat{\tau}$ and $\hat{\gamma}$. We won't worry about how this sum is minimised, since we'll always be using pre-written R functions. We can use the estimates $\hat{\mu}$, $\hat{\tau}$ and $\hat{\gamma}$ to estimate σ_ϵ^2, using

$$\hat{\sigma_\epsilon}^2 = \frac{S(\hat{\mu}, \hat{\tau}, \hat{\gamma})}{N_T + N_C - 3}.$$

The general form for this is

$$\hat{\sigma_\epsilon}^2 = \frac{SSE}{n - p},$$

where SSE is the residual sum of squares, n is the number of data points and p the number of parameters (apart from σ_ϵ^2) being estimated. If you want to know why that is, you can find out here (look particularly at page 62), but we will just take it as given!

As well as generating a fitted value $\hat{\tau}$, we (or rather R!) will also find the standard error of $\hat{\tau}$, and we can use this to generate a confidence interval for the treatment effect τ.

The technique described above is a well-established statistical method known as **ANCOVA** (short for the **An**alysis of **Cova**riance), which can be implemented in R and many other statistical software packages. Notice that it is really just a linear model (the like of which you have seen many times) with at least one factor variable, and with a particular focus (application-wise) on the coefficient of the treatment group variable.

Example 4.5. Let's now implement ANCOVA on our Captopril data in R. We do this by first fitting a linear model using 'lm', with baseline measurement and arm as predictor variables and outcome as the predictand.

```
lm_capt = lm(outcome ~ baseline + arm, data = df_hommel)
summary(lm_capt)
```

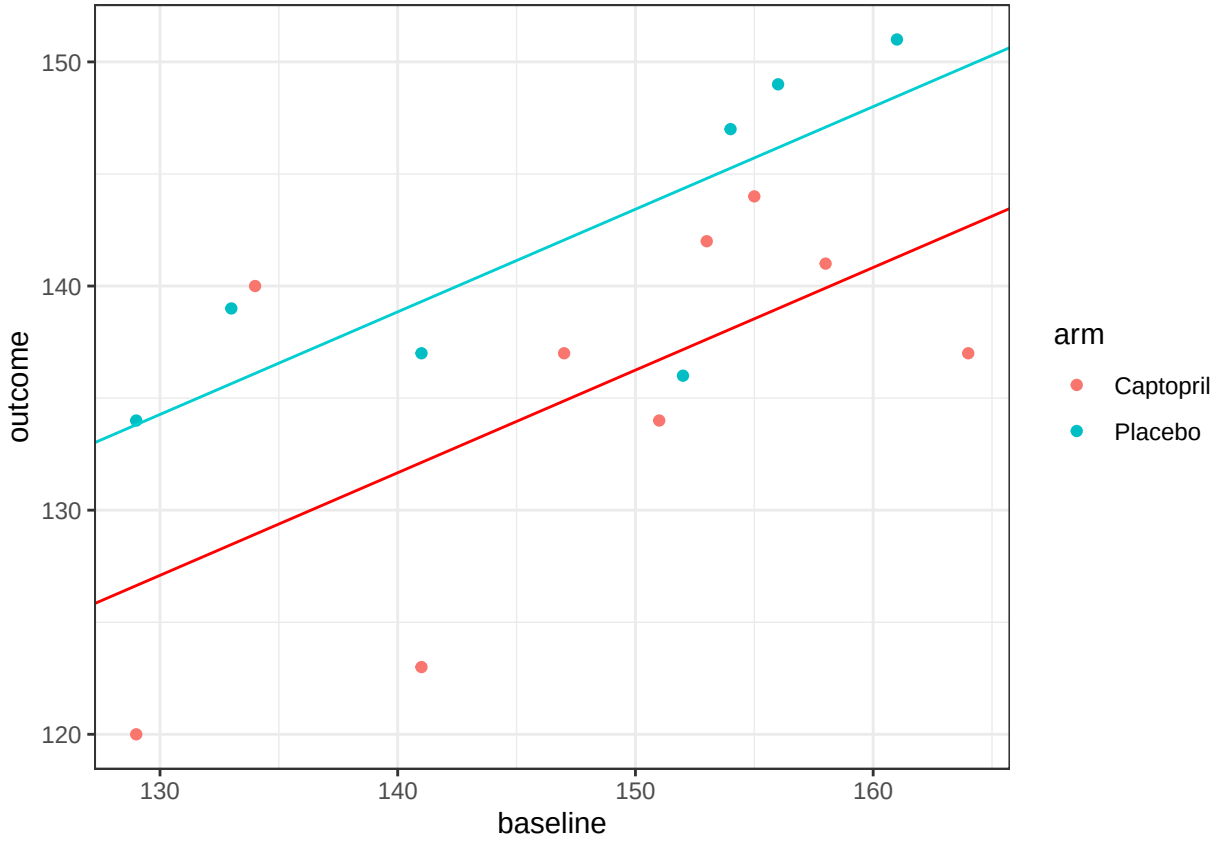
```
##
## Call:
## lm(formula = outcome ~ baseline + arm, data = df_hommel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.129 -3.445  1.415  2.959 11.076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.5731    19.7577   3.420  0.00456 **
## baseline      0.4578     0.1328   3.446  0.00434 **
```

```
## armPlacebo    7.1779     2.9636   2.422  0.03079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.869 on 13 degrees of freedom
## Multiple R-squared:  0.5629, Adjusted R-squared:  0.4957
## F-statistic: 8.372 on 2 and 13 DF,  p-value: 0.004608
```

The variable 'arm' here is being included as a factor variable, so it behaves like

$$\text{arm}_i = \begin{cases} 0 & \text{if participant } i \text{ is assigned Captopril} \\ 1 & \text{if participant } i \text{ is assigned Placebo.} \end{cases}$$

Therefore, for a patient assigned Placebo, a value of 7.1779 is added, as well as the intercept and baseline term. This results in a model with two parallel fitted lines.



For our previous methods we have calculated a confidence interval for the treatment effect τ, and we will do that here too. The second column of the linear model summary (above) gives the standard errors of each estimated parameter, and we see that the standard error of $\hat{\tau}$ is 2.9636. Therefore, to construct a 95/% confidence interval for $\hat{\tau}$, we use (to 3 decimal places)

$$7.178 \pm t_{0.975;13} \times 2.964 = (0.775,\ 13.580)\,.$$

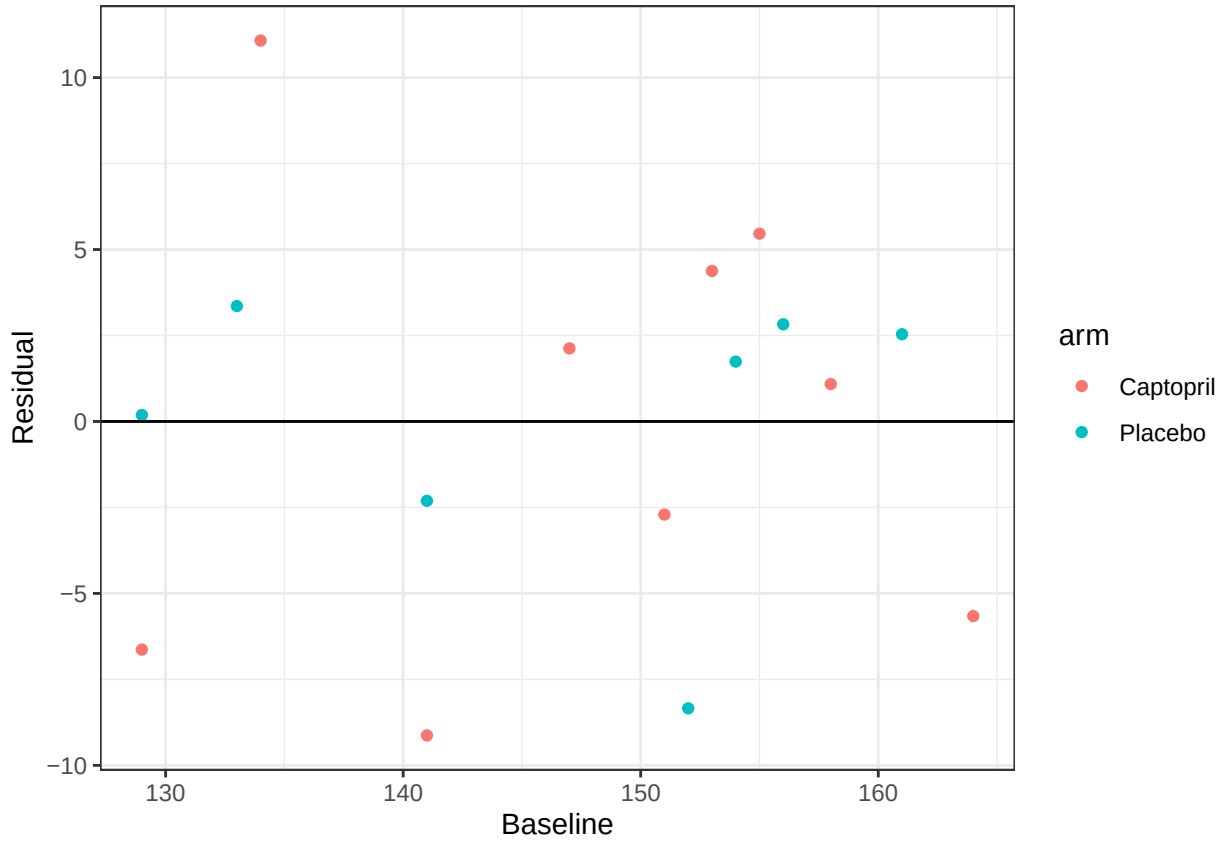
The model has $n - p = 13$ degrees of freedom because there are $n = 16$ data points and we are estimating $p = 3$ parameters. Notice that unlike our previous confidence intervals, this doesn't contain zero, and so our analysis has enabled us to conclude that there is a significant reduction in blood pressure with Captopril. You can also see this in that $p = 0.03079 < 0.05$. However, you can tell from the width of the interval (and the fact that p is still quite close to 0.05) that there is still a lot of uncertainty about τ.

The 'Residual standard error' term near the bottom of the linear model summary is the estimate of $\hat{\sigma_\epsilon}$, so here we have $\hat{\sigma_\epsilon}^2 = 5.869^2 = 34.44$.

As with any fitted model, we should check the residuals.

```
resid_capt = resid(lm_capt)
df_hommel$resid= resid_capt

ggplot(data = df_hommel, aes(x=baseline, y=resid, col=arm)) +
  geom_point() +
  geom_hline(yintercept=0)+
  xlab("Baseline")+
  ylab("Residual")+theme_bw()
```



These look pretty good; there are no clear patterns and the distribution appears to be similar for each treatment group. Though, with such a small sample it's difficult really to assess the fit of the model.

4.4 Some follow-up questions. . . .

This might have raised a few questions, so we will address those now.

4.4.1 Didn't we say that $X_T - X_C$ was an unbiased estimator of τ?

In Sections 4.1 and 4.2 we used both $\bar{X}_T - \bar{X}_C$ and $\left(\bar{X}_T - \bar{B}_T\right) - \left(\bar{X}_C - \bar{B}_C\right)$ as unbiased estimators of τ. Then, in Section 4.3.1 we showed that

$$\mathrm{E}\left(\bar{X}_T - \bar{X}_C \mid \bar{b}_T,\, \bar{b}_C\right) = \tau + \rho\left(\bar{b}_T - \bar{b}_C\right)$$
$$\mathrm{E}\left[\left(\bar{X}_T - \bar{b}_T\right) - \left(\bar{X}_C - \bar{b}_C\right) \mid \bar{b}_T,\, \bar{b}_C\right] = \tau + (\rho - 1)\left(\bar{b}_T - \bar{b}_C\right),$$

that is, neither of these quantities are unbiased estimators of τ (except in very specific circumstances).

Is this a contradiction?

You'll be relieved to hear (and may already have realised) that it isn't; the first pair of equations are blind to the baseline values B_T and B_C, and are using their statistical properties. Because of the randomisation procedure, a priori they can be treated the same. However, once we have observed values for the baseline, b_T and b_C, they are very unlikely to be exactly the same. They are also (along with all other baseline measurements, often things like age, sex, height etc.) definitely not affected by the trial, since they are taken before any placebo or treatment has been administered, and often even before allocation. However, conditioning on their observed values can reduce the variance of our estimate of τ, as we have seen.

In this sense, the observed baseline means $\bar{b}_T$ and $\bar{b}_C$ are known as **ancillary statistics**; they contain no direct information about the parameter we are interested in (in this case the treatment effect τ), but our inferences can be improved by conditioning on the observed values of the ancillary statistics.

4.4.2 What if the lines shouldn't be parallel? The unequal slopes model

In the analysis above, we have assumed that the coefficient γ of baseline (the estimate of the correlation between outcome and baseline) is the same in both groups; we have fitted an **equal slopes model**. It isn't obvious that this should be the case, and indeed we can test for it.

Allowing each group to have a different slope means including an interaction term between baseline and treatment group,

$$x_i = \mu + \tau G_i + \gamma b_i + \lambda b_i G_i + \epsilon_i.$$

The term $\lambda b_i G_i$ is 0 if participant i is in group C and λb_i if participant i is in group T. Therefore, for participants in group C, the gradient is still γ, but for participants in group T it is now $\gamma + \lambda$. We can test whether this interaction term should be included (that is, whether we should fit an unequal slopes model) by including it in a model and analysing the results.

Example 4.6. Continuing once again with the Captopril dataset, we now fit the model

```
lm_capt_int = lm(outcome ~ arm + baseline + baseline:arm, data = df_hommel)
summary(lm_capt_int)
```

```
##
## Call:
## lm(formula = outcome ~ arm + baseline + baseline:arm, data = df_hommel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.094 -3.475  1.412  2.979 11.145
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          66.85150   28.02488   2.385   0.0344 *
## armPlacebo            8.72484   40.93465   0.213   0.8348
## baseline              0.46272    0.18886   2.450   0.0306 *
## armPlacebo:baseline -0.01051    0.27723  -0.038   0.9704
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.108 on 12 degrees of freedom
## Multiple R-squared:  0.563,  Adjusted R-squared:  0.4537
## F-statistic: 5.153 on 3 and 12 DF,  p-value: 0.01614
```

We see that the p-value for the coefficient λ (seen in the `arm:baseline` row) is not at all significant (0.97). Therefore we can be confident that there is no need to fit unequal slopes for this dataset. This fits with our earlier conclusion (from inspecting the residuals) that just including first order terms is fine.

4.4.3 Can we include any other baseline covariates?

In Section 4.2 when our estimated treatment effect was $\hat{\tau} = \left(\bar{x}_T - \bar{b}_T\right) - \left(\bar{x}_C - \bar{b}_C\right)$, the only other variable we could take into account was the baseline measurement, because it is on the same scale as the outcome X. However, in ANCOVA, our treatment effect is

$$\hat{\tau} = (\bar{x}_T - \bar{x}_C) - \hat{\gamma}\left(\bar{b}_T - \bar{b}_C\right),$$

and the inclusion of the coefficient γ means that we can include other covariates on different scales too. The key issue is that we can only include as covariates things that were already known before allocation (hence they are sometimes known as *baseline covariates*, not to be confused with 'the baseline', which would generally mean the same measurement as the primary outcome, but before treatment). This is because they cannot, at that point, have been affected by the treatment, or have had an influence on the post-trial outcome measurement. Indeed, as a rule, any variable that was used in the randomisation procedure (this particularly applies to minimisation and stratified sampling) should be included in the analysis.

Example 4.7. The data for this example is taken from Kassambara (2019). In this study, 60 patients take part in a trial investigating the effect of a new treatment and exercise on their stress score, after adjusting for age. There are two treatment levels (yes or no) and three exercise levels (low, moderate and high) and 10 participants for each combination of treatment and exercise levels. Because in ANCOVA we fit a coefficient to every covariate, we can include exercise (another factor variable) and age (a continuous variable) in this analysis.

id	score	treatment	exercise	age
1	95.6	yes	low	59
2	82.2	yes	low	65
3	97.2	yes	low	70
4	96.4	yes	low	66
5	81.4	yes	low	61
6	83.6	yes	low	65
7	89.4	yes	low	57
8	83.8	yes	low	61
9	83.3	yes	low	58
10	85.7	yes	low	55
11	97.2	yes	moderate	62
12	78.2	yes	moderate	61
13	78.9	yes	moderate	60
14	91.8	yes	moderate	59
15	86.9	yes	moderate	55
16	84.1	yes	moderate	57
17	88.6	yes	moderate	60
18	89.8	yes	moderate	63
19	87.3	yes	moderate	62
20	85.4	yes	moderate	57
21	81.8	yes	high	58
22	65.8	yes	high	56
23	68.1	yes	high	57
24	70.0	yes	high	59
25	69.9	yes	high	59
26	75.1	yes	high	60
27	72.3	yes	high	55
28	70.9	yes	high	53
29	71.5	yes	high	55
30	72.5	yes	high	58
31	84.9	no	low	68
32	96.1	no	low	62
33	94.6	no	low	61
34	82.5	no	low	54
35	90.7	no	low	59
36	87.0	no	low	63
37	86.8	no	low	60
38	93.3	no	low	67
39	87.6	no	low	60
40	92.4	no	low	67
41	100.0	no	moderate	75
42	80.5	no	moderate	54
43	92.9	no	moderate	57
44	84.0	no	moderate	62
45	88.4	no	moderate	65
46	91.1	no	moderate	60
47	85.7	no	moderate	58
48	91.3	no	moderate	61
49	92.3	no	moderate	65
50	87.9	no	moderate	57
51	91.7	no	high	56
52	88.6	no	high	58
53	75.8	no	high	58
54	75.7	no	high	58
55	75.3	no	high	52
56	82.4	no	high	53
57	80.1	no	high	60

Table 4.6: Summary of the stress dataset

Treatment	Exercise	Mean age	SD age
yes	low	61.7	4.691600
yes	moderate	59.6	2.590581
yes	high	57.0	2.211083
no	low	62.1	4.332051
no	moderate	61.4	5.947922
no	high	57.9	3.381321

Table 4.6 shows the mean and standard deviation of age for each combination of treatment and exercise level. If we were being picky / thorough, we might note that (perhaps unsurprisingly!) the mean and standard deviation of age are both lower in the high exercise groups. This might well affect our analysis, but we won't go into this now.

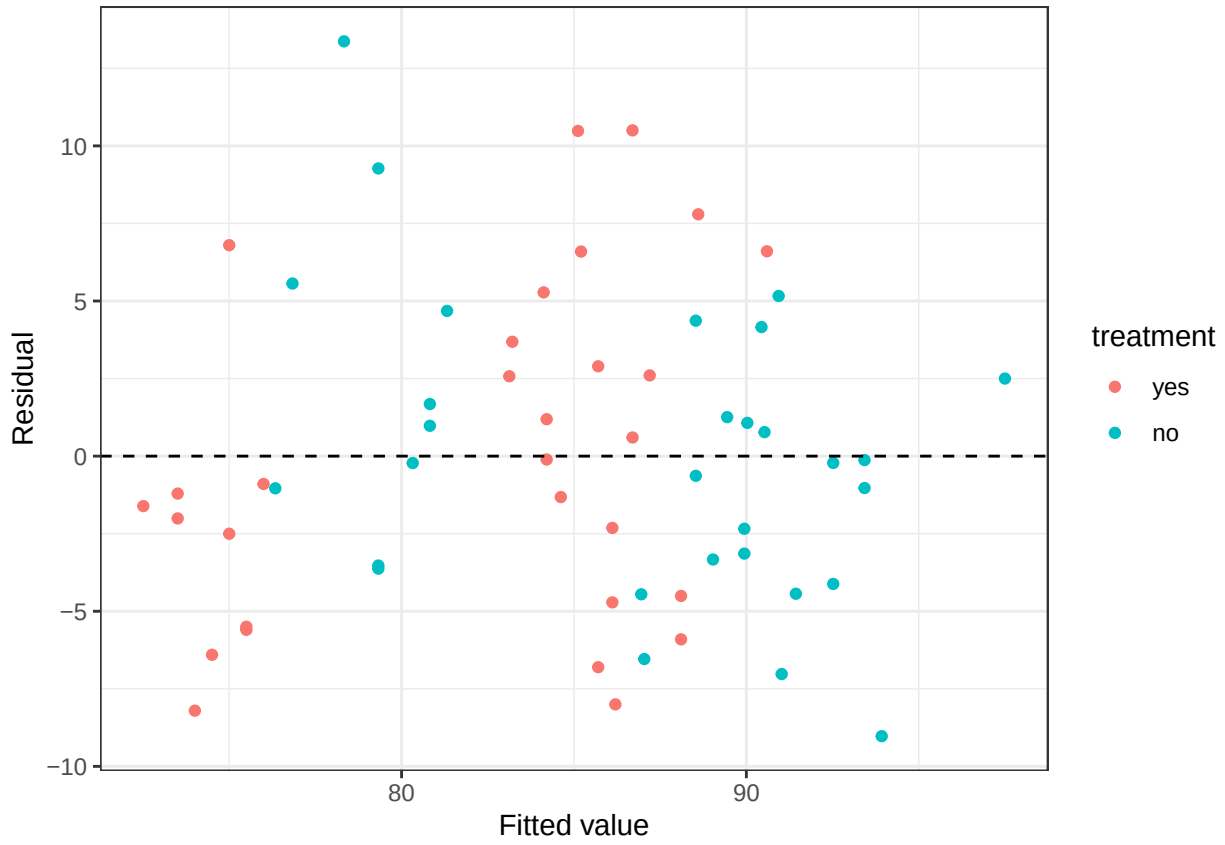
Fitting a linear model, we see that treatment, high levels of exercise and age each have a significant effect on stress.

```
lm_stresslin = lm(score ~ treatment + exercise + age, data = stress)
summary(lm_stresslin)
```

```
##
## Call:
## lm(formula = score ~ treatment + exercise + age, data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0261 -3.7497 -0.4285  3.0943 13.3696
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      55.72934   10.91888   5.104 4.27e-06 ***
## treatmentno       4.32529    1.37744   3.140  0.00272 **
## exercisemoderate  0.08735    1.69032   0.052  0.95897
## exercisehigh     -9.61841    1.84741  -5.206 2.96e-06 ***
## age               0.49811    0.17648   2.822  0.00662 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.288 on 55 degrees of freedom
## Multiple R-squared:  0.6045, Adjusted R-squared:  0.5757
## F-statistic: 21.01 on 4 and 55 DF,  p-value: 1.473e-10
```

In particular, taking a high level of exercise reduced participants' stress scores by around 9.6, and the treatment reduced stress scores by around 4.3. Participants' stress scores increased slightly with age (just under half a point per year!).

We can plot the residuals to check that the model is a reasonable fit



The first thing we notice is that the data are sort of 'clumped'. This is common in factor models, especially where one or more factors is highly influential. Working from right to left (higher fitted stress score to lower), the highest clump (in blue) are those who do moderate or low levels of exercise and didn't receive the treatment. The next clump (in red) are those who do moderate or low levels of exercise and did receive the treatment (their stress score is around 4.3 points lower, for the same age). The next clump, in blue, are those who do high levels of exercise and didn't receive the treatment. Their scores are around 9.6 points lower than the low/moderate exercise groups who didn't receive treatment. Finally, the lowest scoring group are those who do high levels of exercise and did receive the treatment. We see that some clumps aren't really centred around zero, and this should raise alarm bells.

We could also test for interactions, firstly across all factors:

```
##
## Call:
## lm(formula = score ~ (treatment + exercise + age):(treatment +
##     exercise + age), data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5637 -3.3982  0.4173  2.3827 10.3907
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  61.25416   19.86949   3.083  0.00333 **
## treatmentno                   3.89781   24.16324   0.161  0.87250
## exercisemoderate            -14.60897   24.31690  -0.601  0.55070
## exercisehigh                -12.03441   29.53812  -0.407  0.68544
```

```
## age                            0.43121    0.32097   1.343  0.18518
## treatmentno:exercisemoderate  -0.20723    3.35949  -0.062  0.95106
## treatmentno:exercisehigh       8.12783    3.72077   2.184  0.03365 *
## treatmentno:age               -0.03769    0.38851  -0.097  0.92311
## exercisemoderate:age           0.24286    0.40215   0.604  0.54864
## exercisehigh:age              -0.03524    0.50722  -0.069  0.94488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.106 on 50 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6043
## F-statistic: 11.01 on 9 and 50 DF,  p-value: 3.181e-09
```

and then restricted to the interactions that seem important:

```
##
## Call:
## lm(formula = score ~ treatment + exercise + age + treatment:exercise,
##     data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3250 -3.0192  0.2745  2.4650 10.6667
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  56.79090   10.41383   5.453 1.32e-06 ***
## treatmentno                   1.52858    2.23026   0.685   0.4961
## exercisemoderate              0.01746    2.25662   0.008   0.9939
## exercisehigh                -13.70331    2.36314  -5.799 3.78e-07 ***
## age                           0.50355    0.16684   3.018   0.0039 **
## treatmentno:exercisemoderate  0.15503    3.16129   0.049   0.9611
## treatmentno:exercisehigh      8.21822    3.15375   2.606   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.985 on 53 degrees of freedom
## Multiple R-squared:  0.6613, Adjusted R-squared:  0.623
## F-statistic: 17.25 on 6 and 53 DF,  p-value: 6.167e-11
```

Notice that now, the effect of the treatment on its own is not significant. Also notice that for both the linear exercise terms and the interactions between the exercise and treatment, the effects of moderate and low exercise are very similar. Combining the coefficients, someone who does a high level of exercise:

- is likely to reduce their stress score by 13.7 if they receive the treatment
- is likely to reduce their stress score by $13.7 - 8.2 = 5.5$ if they don't receive the treatment

Returning to our initial look at the dataset, the fact that age is a factor, and high levels of exercise are clearly very important should worry us slightly, since there are very few older people doing high levels of exercise. This may mean our model is inaccurate.

An important caution!

As you'll have seen if you read Kendall (2003) (for formative assignment 1), we should have everything in place, including a statistical analysis plan, **before** the trial. We should already know which covariates we plan to include in our model, and how. 'Trawling' for the best possible model by trying lots of different things (and inevitably settling on the one that leads to the most significant conclusion) is poor practice, and can increase the type I error rate (α).

I realise that is sort of what we've done in this Section on Analysis, but that was to demonstrate and compare the different methods. Proceeding in the way we have, trying lots of different models, when analysing and writing up a trial would be very poor practice!

There's another excellent episode of the JAMA Evidence podcast, with a focus on adjusting for covariates, that talks about this issue (you can find it here and linked from Ultra).

That draws to a close our work with continuous outcome variables. In the next lecture, we'll start thinking about binary outcome variables.

Part II

Part II: Binary outcome variable

Chapter 5

Sample size for a binary variable

So far almost everything we've covered has related to continuous outcome variables, which we assumed to be normally distributed. This allowed us to use familiar techniques such as the t-test, and to take baseline information into account in an accessible way (the linear model / ANCOVA). However, very often clinical trials do not have a continuous, normally distributed output, and in the next two sections we will look at two other common possibilities: binary data (this section) and survival data (next section).

A binary outcome might be something like 'the patient was alive 2 years after the procedure' or not, or 'the patient was clear of eczema within a month' or not. Such variables are often coded as 'success' or 'failure', or 1 or 0.

For a trial whose primary outcome variables are binary, the sample size calculations we derived in Chapter 2 will not work, so in this section we'll work through a similar method developed for binary variables.

Suppose we conduct a trial with a binary primary outcome variable and two groups, T and C, containing n_T and n_C participants respectively. The number of successes in each group, R_T and R_C, will be Binomially distributed,

$$\begin{aligned} R_T &\sim Bi\left(n_T, \pi_t\right) \\ R_C &\sim Bi\left(n_C, \pi_C\right). \end{aligned}$$

Our null hypothesis now is therefore that $\pi_T = \pi_C$, ie. that the probability of success is the same in each group, and we will need enough participants to test this hypothesis with sufficient power. With the trial data we will be able to produce estimates

$$\begin{aligned} p_T &= \frac{R_T}{n_T} \\ p_C &= \frac{R_C}{n_C}. \end{aligned}$$

Recall that the variance of p_X (where X is T or C) is $\frac{\pi_X(1-\pi_X)}{n_X}$, such that the variance depends on the mean. This means there is no free parameter equivalent to σ in the binary situation, and the number of participants required will depend on the approximate value of π_T and π_C. This makes the derivation of a sample size formula somewhat more complicated, and so we first of all make a transformation to remove the dependence of mean and variance. To do this we use an approximation technique called *the delta method*.

5.1 The Delta Method

We start with a random variable X that has mean μ and variance $\sigma^2 = \sigma^2(\mu)$, ie. its variance depends on its mean. If we have a 'well-behaved' (infinitely differentiable etc.) function $f(X)$, what are its mean and variance? To find this exactly requires us to evaluate a sum or integral, and this may be analytically intractable, so we use instead a crude approximation.

First, we expand $f(X)$ in a first-order Taylor series about μ, which gives us

$$f(X) \approx f(\mu) + (X - \mu) f'(\mu) \tag{5.1}$$

and therefore

$$(f(X) - f(\mu))^2 \approx (X - \mu)^2 [f'(\mu)]^2. \tag{5.2}$$

If we take expectations of Equation (5.1) we find $E(f(X)) \approx f(\mu)$. We can use this in the left-hand side of Equation (5.2) so that when we take expectations of Equation (5.2) we find

$$\text{var}(f(X)) = \sigma^2(\mu) [f'(\mu)]^2, \tag{5.3}$$

where both sides come from

$$\text{var}(X) = \text{E}\left[(X - \mu)^2\right].$$

This series of approximations, which generally works well, is the Delta method.

One way in which it is often used, and the way in which we will use it now, is to find a transformation $f(X)$ for which (at least approximately) the variance is unrelated to the mean. To do this, we solve the differential equation

$$\text{var}[f(X)] = \sigma^2(\mu) [f'(\mu)]^2 = \text{constant}.$$

In the case of proportions for a binary variable, this becomes

$$\frac{\pi(1-\pi)}{n} [f'(\pi)]^2 = K$$

for some constant K. We can rearrange this to

$$f'(\pi) = \sqrt{\frac{Kn}{\pi(1-\pi)}} \propto \sqrt{\frac{1}{\pi(1-\pi)}}.$$

So we need

$$\int^{\pi} \sqrt{\frac{1}{u(1-u)}} du,$$

where the notation indicates that we want the anti-derivative, evaluated at π. By substituting $u = w^2$ we find

$$\begin{aligned} f(\pi) &\propto \int^{\pi} \frac{1}{\sqrt{w^2(1-w^2)}} 2w\, dw \\ &\propto \int \frac{1}{\sqrt{1-w^2}} dw \\ &\propto \arcsin\left(\sqrt{\pi}\right). \end{aligned}$$

Setting $f(\pi) = \arcsin(\sqrt{\pi})$ and using the chain rule, we find

$$\left[f'(\pi)\right]^2 = \frac{1}{4\pi(1-\pi)}.$$

Finally, we can substitute this into Equation (5.3), with $f(X) = \arcsin\left(\sqrt{X}\right)$ to find

$$\begin{aligned} \operatorname{var}\left[f(X)\right] &\approx \sigma^2(\pi)\left[f'(\pi)\right]^2 \\ &\approx \frac{\pi(1-\pi)}{n} \cdot \frac{1}{4\pi(1-\pi)} \\ &\approx \frac{1}{4n}, \end{aligned}$$

and we have achieved our aim of finding a transformation of X whose variance is not related to the mean. This is sometimes called the *angular transformation*.

5.2 A sample size formula

For a binary variable, our estimate p_X (the proportion of successes in group X) is approximately normally distributed, since the central limit theorem applies. This is not true for small values of n (less than around 30, which is very small for a clinical trial) or for values of π close to 0 or 1, say $\pi < 0.15$ or $\pi > 0.85$ (this is more likely to be an issue for some trials).

The linear approximation in Equation (5.1) shows us that if p_X is normally distributed then $f(p_X) = \arcsin\left(\sqrt{p_X}\right)$ will be [approximately] normally distributed too. In fact, $\arcsin\left(\sqrt{p_X}\right)$ is approximately normally distributed with mean $\arcsin\left(\sqrt{\pi_X}\right)$ and variance $1/(4n_X)$. Using this information, we can test $H_0 : \pi_T = \pi_C$ at the $100\alpha\%$ confidence level by using the variable

$$D = \frac{\arcsin\left(\sqrt{p_T}\right) - \arcsin\left(\sqrt{p_C}\right)}{\sqrt{\frac{1}{4n_T} + \frac{1}{4n_C}}} = \frac{\arcsin\left(\sqrt{p_T}\right) - \arcsin\left(\sqrt{p_C}\right)}{\frac{1}{2}\lambda(n_T, n_C)},$$

which is analogous to the variable D constructed in Section 2.3; the difference in $f(p_T)$ and $f(p_C)$ divided by the standard error of the difference.

Using the same logic as in Sections 2.4 and 2.5, the starting place for a sample size formula to achieve significance level α and power $1-\beta$ is

$$\frac{2\left(\arcsin\left(\sqrt{\pi_T}\right) - \arcsin\left(\sqrt{\pi_C}\right)\right)}{\lambda(n_T, n_C)} = z_\beta + z_{\frac{\alpha}{2}}.$$

For two groups of equal size N, this leads us to

$$N = \frac{\left(z_\beta + z_{\frac{\alpha}{2}}\right)^2}{2\left(\arcsin\left(\sqrt{\pi_T}\right) - \arcsin\left(\sqrt{\pi_C}\right)\right)^2}. \tag{5.4}$$

Because $\arcsin\left(\sqrt{\pi_T}\right) - \arcsin\left(\sqrt{\pi_C}\right)$ is not a function of $\pi_T - \pi_C$, we cannot express this in terms of the difference itself, but instead need to specify the expected probabilities of success in each group. In practice, it is likely that the success rate for the control group (π_C) is well understood, and the probability for the intervention group (π_T) can be specified by using the nearest clinically important value of π_T.

Example 5.1. (From Smith et al., 1994) This trial compares two approaches to managing malignent low bile duct obstruction: surgical biliary bypass and endoscopic insertion of a stent. The primary outcome variable was 'Did the patient die within 30d of the procedure?', and the trial was designed to have $\alpha = 0.05$, $1 - \beta = 0.95$, which gives $z_{\frac{\alpha}{2}} = 1.96$, $z_\beta = 1.65$. The trial wanted to be able to determine a change in 30 day mortality rate from 0.2 to at most 0.05. Plugging these numbers into Equation (5.4)) gives us

$$N = \frac{(1.65 + 1.96)^2}{2\left(\arcsin\left(\sqrt{0.2}\right) - \arcsin\left(\sqrt{0.05}\right)\right)^2} = 114.9,$$

and so each group in our trial should contain 115 patients.

If instead our aim had been to detect a change from around 0.5 to 0.35 (the same in terms of $\pi_A - \pi_B$), we would instead have needed

$$N = \frac{(1.65 + 1.96)^2}{2\left(\arcsin\left(\sqrt{0.5}\right) - \arcsin\left(\sqrt{0.35}\right)\right)^2} = 280.8,$$

that is 281 patients per trial arm.

Chapter 6

Analysis for binary outcomes

For a group of $2n$ participants, we will have allocated n_C to the control group (group C), and n_T to the treatment group (group T). The natural statistical model to apply to this situation is therefore a binomial distribution, for example in group C the number of 'successes' would be modelled by

$$R_C \sim \text{Bi}\left(n_C, \pi_C\right).$$

Similarly the number of successes in the treatment group can be modelled as

$$R_T \sim \text{Bi}\left(n_T, \pi_T\right),$$

and the focus of our analysis is on comparing π_C and π_T. To do this we will require point estimates of both quantities and interval estimates for some measure of the discrepancy between them. We will also need ways to test the null hypothesis that $\pi_C = \pi_T$.

6.1 Point estimates and Hypothesis tests

First of all, we can tabulate the results of a trial with a binary outcome like this:

	Successes	Failures	Total
Treatment	r_T	$n_T - r_T$	n_T
Control	r_C	$n_C - r_C$	n_C
Total	r	$n - r$	n

Note that because this is a table of observed values, they are now all in lower case.

We can estimate π_C and π_T by the sample proportions

$$\begin{aligned} p_C &= \frac{r_C}{n_C} \\ p_T &= \frac{r_T}{n_T} \end{aligned}.$$

We know from the properties of the binomial distribtion that $\text{E}\left(p_C\right) = \pi_C$ and

$$\text{Var}\left(p_C\right) = \frac{\pi_C\left(1 - \pi_C\right)}{n_C},$$

and similarly for $\mathrm{E}(p_T)$ and $\mathrm{Var}(p_T)$.

If we think in terms of individual participants, we have the variable Y_{iC} for the outcome of the i-th patient in group C, with $Y_{iC} = 1$ if the participant's outcome is 'success' and $Y_{iC} = 0$ otherwise. Then we have

$$r_C = \sum_{i=1}^{n_C} y_{iC},$$

and similarly for group T. Since p_C and p_T are therefore sample means, we can apply the Central Limit Theorem to conclude that p_C and p_P can be approximated by normal distributions:

$$p_C \sim N\left(\pi_C, \frac{\pi_C(1-\pi_C)}{n_C}\right)$$
$$p_T \sim N\left(\pi_T, \frac{\pi_T(1-\pi_T)}{n_T}\right).$$

This means we can test the null hypothesis that $\pi_C = \pi_T$ by referring our observed value of $p_T - p_C$ to a normal distribution with mean 0 and variance

$$\frac{\pi_T(1-\pi_T)}{n_T} + \frac{\pi_C(1-\pi_C)}{n_C},$$

which we can approximate by substituting in p_C and p_T.

However, since under the null hypothesis $\pi_C = \pi_T = \pi$, it would be more appropriate to use this as the common variance. In this case, the variance of $p_T - p_C$ becomes

$$\pi(1-\pi)\left(\frac{1}{n_C} + \frac{1}{n_T}\right),$$

and in calculations we replace π with $p = r/n$.

Putting all this together, our test statistic is

$$Z = \frac{p_T - p_C}{\sqrt{p(1-p)\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}}.$$

Example 6.1. The data in this example comes from Marshall (1948), in which 109 patients with tuberculosis were assigned to either receive Streptomycin, or the control group. The primary outcome variable is whether or not the patient was improved after the treatment period. The data include several other covariates, including gender, baseline condition (good, fair or poor) and whether the patient had developed resistance to streptomycin after 6 months.

```
##               improved
## arm            FALSE TRUE
##   Streptomycin    17   38
##   Control         35   17
```

We therefore have

$$\begin{aligned} n_C &= 52 \\ n_T &= 55 \\ p_C &= \frac{17}{17+35} &= 0.327 \\ p_T &= \frac{38}{38+17} &= 0.691 \\ p &= \frac{38+17}{107} &= 0.514. \end{aligned}$$

and can calculate our Z statistic to be

$$\begin{aligned} Z &= \frac{0.691 - 0.327}{\sqrt{0.514\,(1-0.514)\left(\frac{1}{52}+\frac{1}{55}\right)}} \\ &= 3.765. \end{aligned}$$

Finally, we can find the p-value of this test statistic (making sure to have two tails!)

```
2*(1-pnorm(3.765, mean=0, sd=1))
```

```
## [1] 0.0001665491
```

So we can reject the hypothesis that streptomycin has no effect on tuberculosis at the $\alpha = 0.05$ level (and indeed many lower levels).

6.1.1 An alternative approach: chi-squared

Another way to approach this would be to conduct a **chi-squared** test.

In a chi-squared test, we first calculate the **expected** values (E_i) in each box of the summary table, and compare them to the **observed** values (O_i) by finding the summary statistic

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i}.$$

Under the null hypothesis (that $\pi_C = \pi_T$) this has a χ^2 distribution with one degree of freedom. We see that the larger the differences between the observed and expected values, relative to the expected values, the larger the test statistic, and therefore the less probably under the χ^2_1 distribution.

Example 6.2. Continuing our streptomycin example, we can calculate a table of expected values by observing that proportion $p = 0.514$ of the total number of patients were improved. There are 52 in the control group, therefore we expect $0.514 \times 52 = 26.73$ improved patients in the control group, and by the same logic $0.514 \times 55 = 28.27$ in the treatment group. Our expected table is therefore

```
##              improved
## arm            FALSE   TRUE
##   Streptomycin 26.730 28.270
##   Control      25.272 26.728
```

We can therefore calculate the χ^2 statistic by looping through the elements of the tables:

```
sum_chi_sq = 0 # set a running total going
# in the following, tab_obs is the table of observed values and
# tab_exp is the table of expected values
for (i in 1:2){
  for (j in 1:2){
    tmp = ((tab_obs[i,j] - tab_exp[i,j])^2)/tab_exp[i,j]
    sum_chi_sq = sum_chi_sq + tmp
  }
}
sum_chi_sq
```

```
## [1] 14.17595
```

```
1-pchisq(sum_chi_sq, df=1)
```

```
## [1] 0.0001664847
```

and again we have a very significant result.

In fact, these two tests are almost equivalent, and we have that $\sqrt{X^2} = Z$:

```
sqrt(sum_chi_sq)
```

```
## [1] 3.765097
```

6.1.2 Likelihood: A more rigorous way

Our method above was quite informal, and also made heavy use of the central limit theorem. We can use maximum likelihood to derive a more formally justified test for binary outcomes. This also lays a good foundation for more complex situations.

Earlier we set up notation y_{iC} to be outcome variable (0 or 1, in this case) of the i-th participant in the control group (and so on), and we will use that here.

The contribution of the i-th patient in group C to the likelihood is

$$\pi_C^{y_{iC}} \left(1 - \pi_C\right)^{1-y_{iC}}$$

(remember we can ignore multiplicative constant terms). Combining all n_C patients in group C, their contribution will be

$$\pi_C^{r_C} \left(1 - \pi_C\right)^{n_C - r_C},$$

where r_C is the number of 'successes' in group C. Similarly for the treatment group we will have

$$\pi_T^{r_T} \left(1 - \pi_T\right)^{n_T - r_T}.$$

Gathering these terms together we can find the complete likelihood function

$$\begin{aligned} L\left(\pi_C, \pi_T \mid \{y_{iC}\}, \{y_{iT}\}\right) & L\left(\pi_C, \pi_T \mid n_C, n_T, r_C, r_T\right) \\ & = \pi_C^{r_C} \left(1 - \pi_C\right)^{n_C - r_C} \pi_T^{r_T} \left(1 - \pi_T\right)^{n_T - r_T}. \end{aligned}$$

The log-likelihood is therefore

$$l\left(\pi_C, \pi_T \mid n_C, n_T, r_C, r_T\right) = r_C \log \pi_C + \left(n_C - r_C\right) \log\left(1 - \pi_C\right) + r_T \log \pi_T + \left(n_T - r_T\right) \log\left(1 - \pi_T\right).$$

If we differentiate with respect to π_C, we find

$$\frac{\mathrm{d}l\left(\pi_C, \pi_T \mid n_C, n_T, r_C, r_T\right)}{\mathrm{d}\pi_C} = \frac{r_C}{\pi_C} - \frac{n_C - r_C}{1 - \pi_C}.$$

Setting this to zero we find (reassuringly!) that $\hat{\pi}_C = \frac{r_C}{n_C}$. We can repeat this exercise for π_T. If we assume that there is one common probability π of success, we can find $\hat{\pi}$ by maximising $l\left(\pi, \pi \mid n_C, n_T, r_C, r_T\right)$ with respect to π, and again this works out to be $\frac{r_C + r_T}{n}$ as before.

We can use these to construct a **likelihood ratio test**, by calculating

$$\begin{aligned}
\lambda_{LR} =& -2\left[l\left(\hat{\pi}, \hat{\pi} \mid n_C, n_T, r_C, r_T\right) - l\left(\hat{\pi}_C, \hat{\pi}_T \mid n_C, n_T, r_C, r_T\right)\right] \\
=& 2\left[\underbrace{r_C \log \frac{r_C}{n_C} + \left(n_C - r_C\right) \log\left(1 - \frac{r_C}{n_C}\right) + r_T \log \frac{r_T}{n_T} + \left(n_T - r_T\right) \log\left(1 - \frac{r_T}{n_T}\right)}_{l\left(\hat{\pi}_C, \hat{\pi}_T \mid n_C, n_T, r_C, r_T\right)}\right. \\
& \left. - \underbrace{\left(r \log\left(p\right) + \left(n - r\right) \log\left(1 - p\right)\right)}_{l\left(\hat{\pi}, \hat{\pi} \mid n_C, n_T, r_C, r_T\right)}\right] \\
=& 2\left[\underbrace{r_C \log\left(\frac{r_C}{n_C p}\right)}_{\text{Group } C \text{ success}} + \underbrace{\left(n_C - r_C\right) \log\left(\frac{n_C - r_C}{n_C\left(1 - p\right)}\right)}_{\text{Group } C \text{ fail}}\right. \\
& \left. + \underbrace{r_T \log\left(\frac{r_T}{n_T p}\right)}_{\text{Group } T \text{ success}} + \underbrace{\left(n_T - r_T\right) \log\left(\frac{n_T - r_T}{n_T\left(1 - p\right)}\right)}_{\text{Group } T \text{ fail}}\right]
\end{aligned}$$

where we use p, r, n to denote the pooled values ($n = n_C + n_T$ etc.).

Each term in the final line corresponds to a subgroup of the participants, as labelled, and if we rearrange them slightly we see that this can be re-written as

$$\lambda_{LR} = 2 \sum_{i \in G} o_i \log\left(\frac{o_i}{e_i}\right),$$

where G is the set of subgroups (group C success etc.). Under the null hypothesis that $\pi_C = \pi_T = \pi$, and for sufficiently large n_C, n_T, λ_{LR} has a χ^2 distribution with one degree of freedom.

Example 6.3. Continuing with the streptomycin example, we can calculate this new test statistic in R by looping through the subgroups.

```
sum_LR = 0 # set a running total going
# in the following, tab_obs is the table of observed values and
# tab_exp is the table of expected values
for (i in 1:2){
```

```
  for (j in 1:2){
    tmp = tab_obs[i,j] * log(tab_obs[i,j]/tab_exp[i,j])
    sum_LR = sum_LR + tmp
  }
}
teststat_LR = 2*sum_LR
teststat_LR
```

```
## [1] 14.5028
```

```
1-pchisq(teststat_LR, df=1)
```

```
## [1] 0.0001399516
```

Not surprisingly, this value is quite close to the one we obtained earlier!

6.2 Measures of difference for binary data

An important note: we're treating $\pi_T > \pi_C$ as good here, as we would when the primary outcome is something positive, like a patient being cured. All the methods can easily be adapted to a situation where $\pi_C > \pi_T$ is desirable.

In the above example the question we were interested in was 'is what we've observed statistically significant?' and in our streptomycin example the answer was a resounding 'Yes!'. However, if we then ask questions like 'How big is the difference between the effects of each treatment?' or 'What is the treatment effect?', things get a bit less clear.

In the continuous case, it made sense to simply think about the treatment effect as the difference $\mu_T - \mu_C$ between outcomes. However, in the binary case there are a few different ways we can think of the difference between two proportions π_C and π_T, and each of them requires a different approach.

6.2.1 Absolute risk difference and Number Needed to Treat

The **absolute risk difference** is

$$\text{ARD} = \pi_T - \pi_C,$$

and is sometimes used. However, it loses a lot of information that we'd probably like to keep in some how. For example, suppose a treatment increases the proportion cured from $\pi_C = 0.01$ to $\pi_T = 0.03$. The absolute risk difference is 0.02 here. For some other treatment that results in an increase from $\pi_C = 0.55$ to $\pi_T = 0.57$ we have the same absolute risk difference, even though it feels (and is!) a much less significant reduction.

It is useful though to remember that usually these numbers are about people. If the outcome is 'cured' or 'not cured', then for some cohort of N patients, $N \times \text{ARD}$ is the number of extra patients you would expect to cure if you used treatment T instead of treatment C (which may be nothing or may some usual course of treatment).

Linked to this is the **number needed to treat** (NNT), which is defined as

$$\text{NNT} = \frac{1}{\pi_T - \pi_C} = \frac{1}{\text{ARD}}.$$

The NNT is the number of patients you'd need to treat (with treatment T rather than C) before you would bring benefit to one extra patient. The website TheNNT collects together results from many clinical trials and uses the NNT as a summary. Some of the results are quite surprising, compared to how effective we think medicines are!

Note that the NNT doesn't tell us how many patients are cured: it is just a measure of how much more effective the treatment is than the control. In both the examples above, with ARD=0.02, we have NNT=50. However, in the first example we would expect around 0-1 patients out of 50 cured under the control and 1-2 out of 50 cured under the treatment. In the second case, we would expect 27-28 cured out of 50 under the control, and 28-29 out of 50 cured under the treatment.

The NNT is popular as a clinical benchmark, and provides useful intuition in terms of the number of people it will help. For example, if $\pi_T = 0.25$, $\pi_C = 0.2$, then ARD $= 0.05$ and NNT $= 20$. After treating 20 patients with treatment C we expect to cure (say) 4, whereas treating 20 patients with treatment T it is expected that we will cure 5. For very small proportions, the NNT can be large even for what appears to be an important difference. For example, if $\pi_C = 0.005$ and $\pi_T = 0.015$ then ARD $= 0.01$ and NNT $= 100$. It might be decided that the necessary changes and costs are not worth it for such a small difference. That said, the NNT is not the easiest statistic to work with, as we shall see!

6.2.1.1 Confidence intervals for ARD and NNT

Let's suppose we want to work with the ARD, and to make a confidence interval for the treatment difference $\tau_{ARD} = \pi_T - \pi_C$. Using the same normal approximation as before, we can estimate τ_{ARD} by $p_T - p_C$, and $\text{var}(p_T - p_C)$ by

$$\frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C}.$$

Our $100(1-\alpha)\%$ confidence interval is therefore given by

$$\left(p_T - p_C - z_{\frac{\alpha}{2}}\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C}},\ p_T - p_C + z_{\frac{\alpha}{2}}\sqrt{\frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C}}\right)$$

Example 6.4. Back to our streptomycin example, we can now construct a $100(1-\alpha)\%$ confidence interval for the ARD.

Our estimated treatment effect is (to 3 decimal places)

$$\hat{\tau}_{ARD} = p_T - p_C = \frac{38}{55} - \frac{17}{52} = 0.364.$$

Our estimate of the standard error of $\hat{\tau}_{ARD}$ is

$$\begin{aligned}\frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C} &= \frac{\frac{38}{55}\times\frac{17}{55}}{55} + \frac{\frac{17}{52}\times\frac{35}{52}}{52} \\ &= 0.00811\end{aligned}$$

and therefore a 95% confidence interval for τ_{ARD} is

$$\left(0.364 - z_{0.975}\sqrt{0.0811},\ 0.364 + z_{0.975}\sqrt{0.0811}\right) = (0.187,\ 0.541).$$

As we should expect from the very low p-value we saw, the 95% confidence interval does not contain zero.

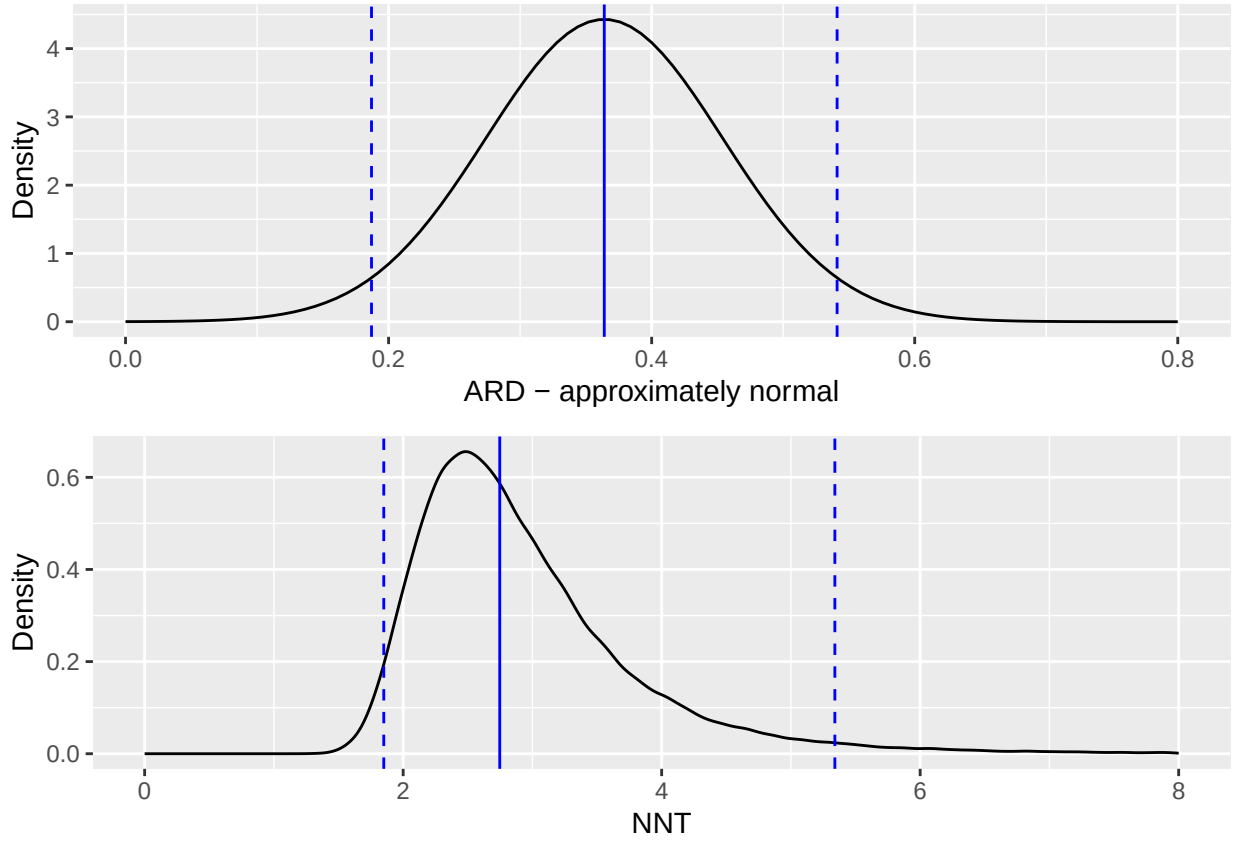
If we want to think instead in terms of NNT (the number needed to treat), then we need to find the reciprocal of our estimate of τ_{ARD}:

$$\text{NNT} = \frac{1}{\tau_{ARD}} = \frac{1}{0.364} = 2.75.$$

That is, we would expect to treat nearly three patients before one is improved (in terms of their tuberculosis symptoms). We can use the limits of the 95% CI for τ_{ARD} to form a 95% CI for NNT, simply by taking the reciprocals of the limits to get

$$\left(\frac{1}{0.541}, \frac{1}{0.187}\right) = (1.85, 5.33).$$

Because the NNT is the reciprocal of something approximately normally distributed, it has a distribution with a long tail, and we see that the confidence interval is therefore skewed.



6.2.1.2 What if the difference is not significant?

In the above section you might have already wondered what happens if the confidence interval for the absolute risk difference (ARD) contains zero. To illustrate this, we will make up some data for a small trial, loosely based on the Streptomycin data we've been using.

The dataset for our made-up trial is

	Successes	Failures	Total
Treatment	9	5	14
Control	4	8	12
Total	13	13	26

The ARD is now

$$\frac{9}{14} - \frac{4}{12} = \frac{3}{13} \approx 0.310$$

and our 95% confidence interval for τ_{ARD} is $(-0.0567,\ 0.676)$.

Clearly because of the small size of the trial our confidence interval is very wide (this is not a very good trial!), but the important thing to note is that it now contains zero. It looks very likely that the treatment is effective (the interval only just contains zero) but how many patients might we need to treat before we expect to see an extra success? The expected value of NNT is

$$\frac{1}{0.310} = 3.23,$$

which does not pose a problem. However, our 95% confidence interval now contains the possibility that the ARD is zero, and in this case the NNT is in some sense infinite: no matter how many patients we treat, we don't expect to see any extra improvements. Therefore, since our confidence interval for ARD contains zero it feels appropriate that our confidence interval for NNT should contain infinity.

When thinking about a confidence interval for the NNT, we need to think about signs, and what negative and positive values mean. If both the lower and upper limits of the confidence interval for ARD are positive, there is no issue - the treatment is effective, and our NNT confidence interval is another entirely positive interval. If the confidence interval for ARD is entirely negative, we have an entirely negative interval for NNT. A negative value of NNT can be thought of as the 'number needed to treat to harm one extra person'.

The tricky situation is when the confidence interval for the ARD is $(-L, U)$ with $L, U > 0$, ie. an interval containing zero. As we approach zero from U, the upper limit of the CI for $\pi_T - \pi_C$, the number of patients we need to treat increases, since the treatment effect is getting smaller, until at $\pi_T - \pi_C = 0$ the NNT is infinite. Therefore, the part of the CI for NNT corresponding to the positive part of the CI for ARD is

$$\left(\frac{1}{U},\ \infty\right)$$

As we approach zero from the left in the interval (ie. from $-L$), the treatment gets less and less effective (and right now we mean effective in a bad way, likely doing harm to the patients compared to the control), and so we need to treat more and more patients to harm one extra patient compared to the control. In this region the NNT is negative, since if we deny some patients the treatment we will benefit a few. Therefore the CI for the NNT corresponding to the negative part of the CI for ARD is

$$\left(-\infty,\ -\frac{1}{L}\right),$$

and altogether the confidence interval for the number needed to treat (NNT) is the union of these two intervals

$$\left(-\infty,\ -\frac{1}{L}\right) \cup \left(\frac{1}{U},\ \infty\right).$$

The plot below shows relationship between ARD and NNT, with the intervals for our toy example shown in bold on the respective axis (the NNT interval should continue infinitely in both directions so for obvious reasons this is not all shown!).

Altman (1998) (available here) makes a compelling push for the use of confidence intervals for the number needed to treat. You can decide for yourself whether what you think of it!

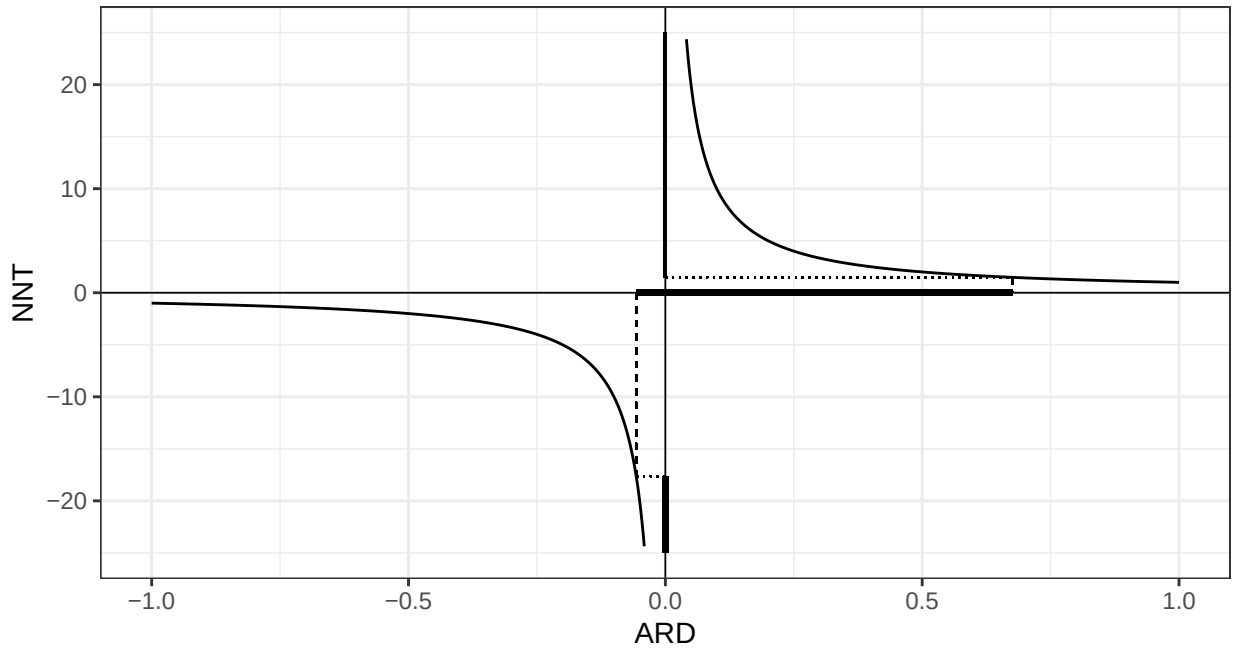


Figure 6.1: The relationship between the confidence interval for the ARD and the NNT, when the ARD interval contains zero.

Problems with the confidence interval for the ARD

You may well remember from the dim and distant past that the method we have been using so far (which in this section we'll be calling the 'standard' method) is not so reliable if the proportion is close to zero or one. Newcombe (1998) compared eleven different methods for finding confidence intervals for the difference in proportions (as we are doing when we work with the ARD) and found the standard method to be the worst! The coverage probability turns out to be much lower than the nominal value, with a so-called 95% confidence interval being closer to 90% or even 85%. A further problem with this method (although it will rarely affect us in practice in this setting) is that the limits of the confidence interval aren't forced to be in $[-1, 1]$.

We will give a sketch of the favourite method of Newcombe (1998), chosen for its ease of implementation and its accuracy, now.

The first step is to find an interval estimate for a single proportion π. As before, this can be written

$$\left\{ \pi \mid \frac{|p-\pi|}{\sqrt{\pi\left(1-\pi\right)/n}} \leq z_{\frac{\alpha}{2}} \right\} = \left\{ \pi \mid \left(p-\pi\right)^2 \leq z_{\frac{\alpha}{2}}^2 \frac{\pi\left(1-\pi\right)}{n} \right\}.$$

We can find the limits of the $100\left(1-\alpha\right)\%$ level confidence interval by changing the right hand side to an equality

$$\left(p-\pi\right)^2 = z_{\frac{\alpha}{2}}^2 \frac{\pi\left(1-\pi\right)}{n}. \tag{6.1}$$

In the standard method, we substitute p (the estimated value of π from our sample) into the right hand side of Equation (6.1) for π, to get

$$\left(p-\pi\right)^2 = z_{\frac{\alpha}{2}}^2 \frac{p\left(1-p\right)}{n}$$

which we solve to get the limits

$$\pi = p \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p\,(1-p)}{n}}.$$

In Newcombe's proposed method, we instead keep π in the right hand side and solve the quadratic in Equation (6.1) in terms of π.

The benefit of this new method will be most obvious for a probability that is close to 0 or 1.

Example 6.5. Suppose we have 1 success out of 50 patients, so $p = 0.02,\ n = 50$.

The limits of a standard 95% confidence interval will be

$$\left(0.02 - z_{0.975}\sqrt{\frac{0.02 \times 0.98}{50}},\ 0.02 + z_{0.975}\sqrt{\frac{0.02 \times 0.98}{50}}\right) = (-0.0188,\ 0.0588)\,,$$

whereas the limits to the Newcombe 95% CI will be the roots of

$$(0.02 - \pi)^2 = z^2_{\alpha/2}\frac{\pi\,(1-\pi)}{50}$$

which work out to be

```
## [1] 0.003539259 0.104954436
```

Visually, we can represent this as in Figure 6.2 by plotting the LHS (solid) and RHS (dashed for new method, dotted for standard method). The thick solid red line shows p_T, the estimated proportion, the thinner dashed red lines show the Newcombe 95% CI and the dotted red lines show the standard 95% CI. Notice that the limits of each confidence interval are formed by the points at which the solid line (LHS) crosses the dashed / dotted lines (RHS).

Example 6.6. Returning to our streptomycin example, our estimate of the probability of success for the treatment group is $p_T = \frac{38}{55},\ n_T = 55$, and therefore our equation becomes

$$\left(\frac{38}{55} - \pi\right)^2 = z^2_{\frac{\alpha}{2}}\frac{\pi\,(1-\pi)}{55}.$$

Solving this equation in the usual way (using the quadratic formula) we find the limits

```
## [1] 0.5597141 0.7971771
```

By contrast, in our standard method we have

$$\left(\frac{38}{55} - \pi\right)^2 = z^2_{\frac{\alpha}{2}}\frac{\frac{38}{55}\left(1 - \frac{38}{55}\right)}{55}$$

which is

```
## [1] 0.5687797 0.8130385
```

We can see this graphically

Notice that the interval with the new method is now asymmetrical, which is more realistic.

Similarly for the control proportion π_C, we have $p_C = \frac{17}{52},\ n_C = 52$, and our Newcombe interval is

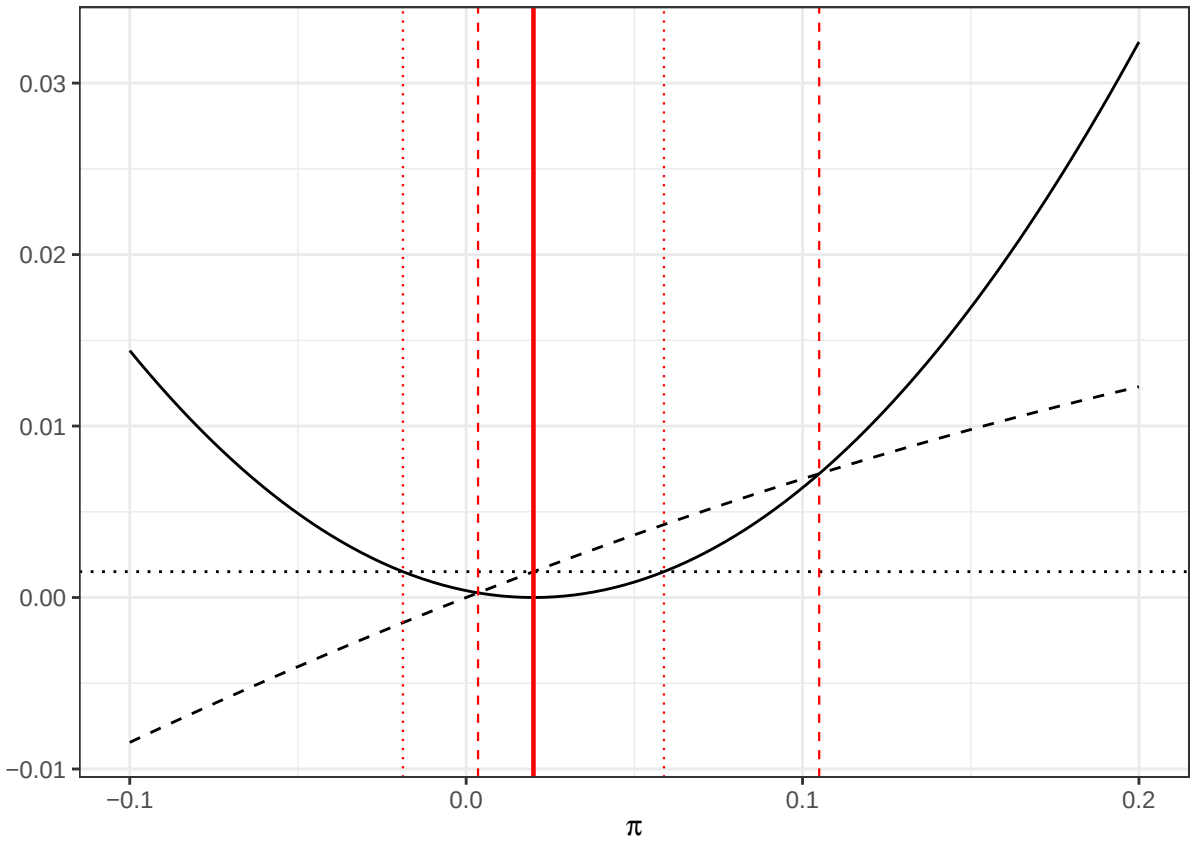


Figure 6.2: LHS of Equation 6.1 (solid black) with RHS of (black dashed) and RHS with estimate p subbed in (black dotted). The limits of the confidence intervals are where the curves cross, shown by red lines: dashed for Newcombe, dotted for standard.

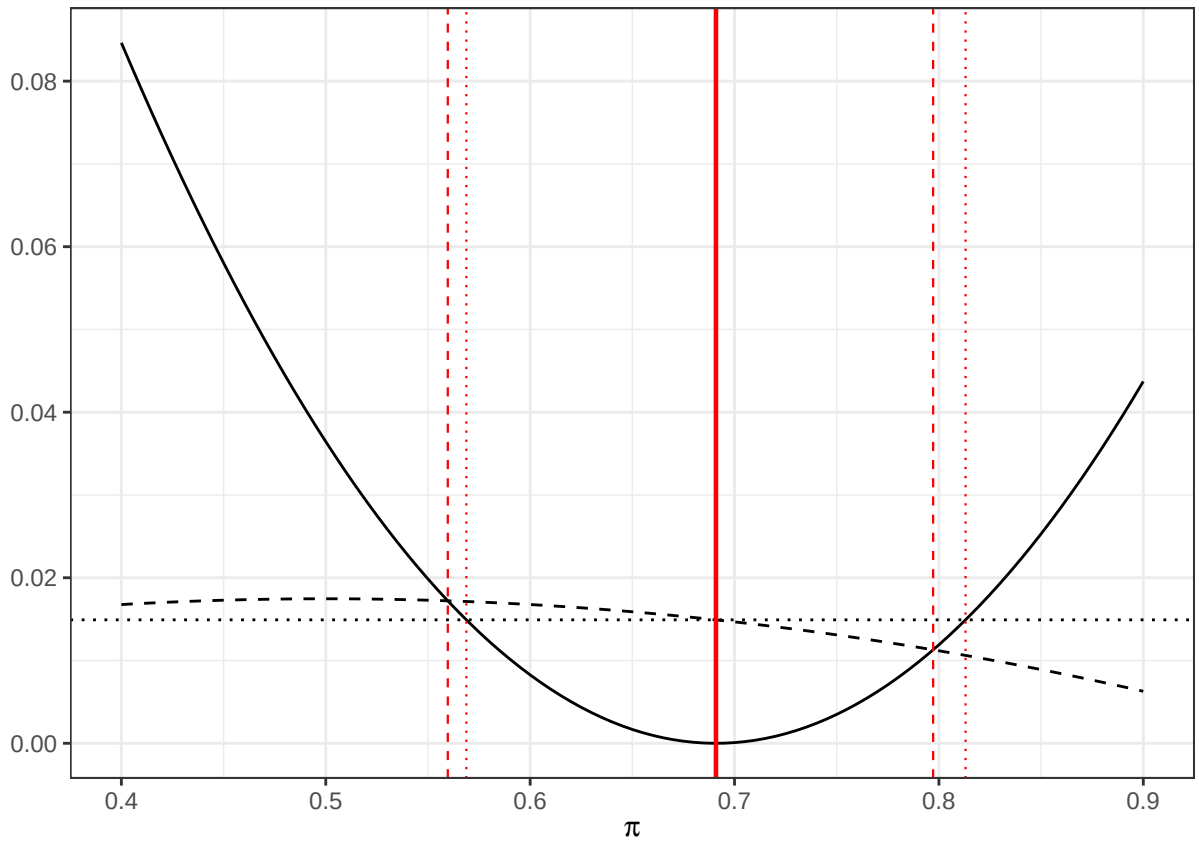


Figure 6.3: As before, dashed for Newcombe, dotted for standard

```
## [1] 0.2152207 0.4624381
```

compared to the standard confidence interval

```
## [1] 0.1994256 0.4544205
```

Again, we can see this graphically.

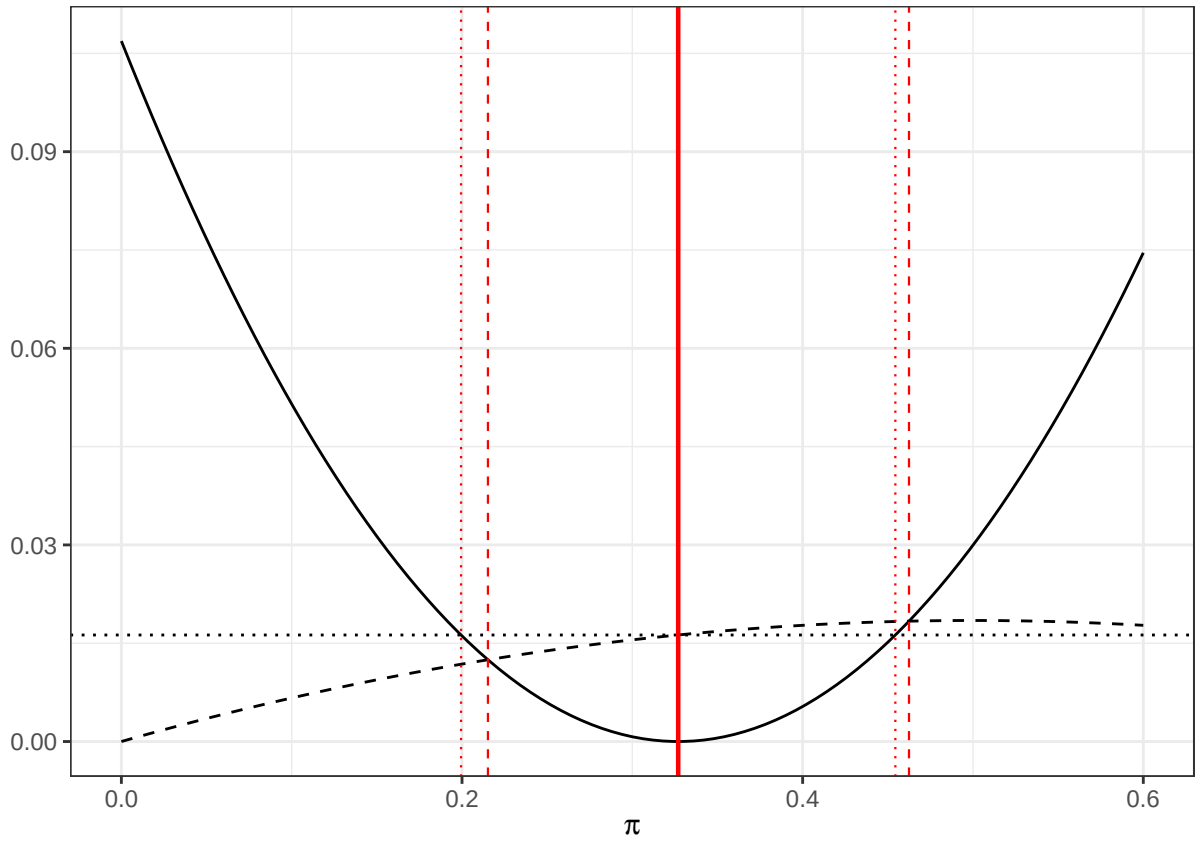


Figure 6.4: As before, dashed for Newcombe, dotted for standard

6.2.1.3 Extending this to $\pi_T - \pi_C$

What the Newcombe interval has given us is a superior method for creating confidence intervals for proportions. But, what we would like is a method for calculating a confidence interval for the difference in two proportions. You'll be relieved to hear that there is such a method, and we'll give a sketch here of how it works.

The limits of the 'standard method' confidence interval at significance level α are given by

$$\left(p_T - p_C - z_{\frac{\alpha}{2}} \sqrt{\frac{p_T\left(1-p_T\right)}{n_T} + \frac{p_C\left(1-p_C\right)}{n_C}},\ p_T - p_C + z_{\frac{\alpha}{2}} \sqrt{\frac{p_T\left(1-p_T\right)}{n_T} + \frac{p_C\left(1-p_C\right)}{n_C}} \right). \tag{6.2}$$

We can rewrite this as

$$\left(p_T - p_C - \sqrt{\omega_T^2 + \omega_C^2},\ p_T - p_C + \sqrt{\omega_T^2 + \omega_C^2}\right) \tag{6.3}$$

where ω_T and ω_C are the widths of the separate single-sample 'standard' confidence intervals for p_T and p_C. In Newcombe's method, we proceed in the same way, but instead use the widths of the Newcombe confidence intervals for the individual probabilities p_T and p_C. This is obviously a little more complicated, since the widths (eg. $p_T - l_T$ and $u_T - p_T$) will now not be the same, since the Newcombe CI is not symmetrical. So, we have

$$\left(p_T - p_C - \sqrt{(p_T - l_T)^2 + (u_C - p_C)^2},\ p_T - p_C + \sqrt{(u_T - p_T)^2 + (p_C - l_C)^2}\right).$$

These differences must be calculated using the individual sample confidence interval method.

Example 6.7. Applying this Newcombe method to our Streptomycin example, recall that we have

$$\begin{aligned}
p_T &= \frac{38}{55} \\
p_T - l_T &= \frac{38}{55} - 0.5597 = 0.1312 \\
u_T - p_T &= 0.7972 - \frac{38}{55} = 0.1064 \\
p_C &= \frac{17}{52} \\
p_C - l_C &= \frac{17}{52} - 0.2152 = 0.1117 \\
u_C - p_C &= 0.4624 - \frac{17}{52} = 0.1355.
\end{aligned}$$

Our 95% confidence interval is therefore

$$\begin{gathered}
\left(p_T - p_C - \sqrt{(p_T - l_T)^2 + (u_C - p_C)^2},\ p_T - p_C + \sqrt{(u_T - p_T)^2 + (p_C - l_C)^2}\right) \\
\left(\frac{38}{55} - \frac{17}{52} - \sqrt{0.1312^2 + 0.1355^2},\ \frac{38}{55} - \frac{17}{52} + \sqrt{0.1064^2 + 0.1117^2}\right) \\
(0.3640 - 0.1886,\ 0.3640 + 0.1543) \\
(0.157,\ 0.500).
\end{gathered}$$

This is skewed somewhat lower than our standard CI of $(0.187,\ 0.541)$.

6.2.2 Risk Ratio (RR) and Odds ratio (OR)

The measures we have looked at so far, particularly the ARD, are quite analagous to the continuous normally distributed case. However, there are yet more commonly used measures of difference for proportions, which need to be dealt with differently, but also afford more opportunities for modelling.

The **risk ratio** is defined as

$$\text{RR} = \frac{\pi_T}{\pi_C}$$

The **odds ratio** is defined as

$$\text{OR} = \frac{\pi_T / (1 - \pi_T)}{\pi_C / (1 - \pi_C)}$$

The first thing to note is that for both the risk ratio and the odds ratio, the null value is one (not zero, as for the ARD), and both values must always be positive. We think about things multiplicatively, so for example if $RR = 3$ we can say that the event is "3 times more likely" in group T than in group C.

Odds

Odds and odds ratios are a bit trickier to think about (this article explains them really well - it's aimed at 'kids and teens' but don't let that put you off!). The odds of an event are the probability of it happening over the probability of it not happening. So, if (for some event A), $p(A) = 0.2$, the odds of A are

$$\frac{p(A)}{p(A')} = \frac{0.2}{0.8} = \frac{1}{4},$$

which we say as "1 to 4" or 1:4. For every one time A occurs, we expect it not to occur four times.

The **odds ratio** compares the odds of the outcome of interest in the Treament group with the odds of that event in the Control group. It tells us how the odds of the event are affected by the treatment (vs control).

With the ARD, we knew that our confidence interval should always be in $[-1, 1]$, and that if we compare treatments in one direction (say $p_T - p_C$) we would obtain the negative of the interval for the other way ($p_C - p_T$). With the RR and OR, the discrepancy between two proportions is given by a ratio, and so comparing them in one direction (p_T/p_C) will give the reciprocal of the other direction (p_C/p_T).

Example 6.8. For our Streptomycin example, we estimated the ARD by

$$\hat{\tau}_{ARD} = p_T - p_C = \frac{38}{55} - \frac{17}{52} = 0.364,$$

or could have alternatively had

$$\hat{\tau}_{ARD} = p_C - p_T = \frac{17}{52} - \frac{38}{55} = -0.364.$$

For the risk ratio, we have

$$\hat{\tau}_{RR} = \frac{p_T}{p_C} = \frac{38/55}{17/52} = 2.113,$$

or could alternatively have

$$\hat{\tau}_{RR} = \frac{p_C}{p_T} = \frac{17/52}{38/55} = 0.473 = \frac{1}{2.113}.$$

We could say that a patient is "more than twice as likely to be cured with streptomycin than by the control".

For the odds ratio, we have

$$\hat{\tau}_{OR} = \frac{p_T/(1-p_T)}{p_C/(1-p_C)} = \frac{(38/55)/(17/55)}{(17/52)/(35/52)} = 4.602,$$

and therefore the odds of recovery are around 4.6 greater for Streptomycin than for the control. Similarly, we could reframe this as

$$\hat{\tau}_{OR} = \frac{p_C/(1-p_C)}{p_T/(1-p_T)} = \frac{(17/52)/(35/52)}{(38/55)/(17/55)} = 0.217 = \frac{1}{4.602}.$$

6.2.2.1 Confidence intervals for RR and OR

One thing to notice is that symmetry works differently on the RR and OR scale from on the ARD scale. There is an equivalence between an interval $(l,\, u)$ (with $l, u > 1$) and $\left(\frac{1}{u}, \frac{1}{l}\right)$, since these intervals would equate to comparing the same two treatments in different directions (assuming the difference was significant and neither interval contains 1). Similarly, on this scale the interval

$$\left(\frac{1}{k},\, k\right) \text{ for some } k > 1$$

can be thought of as symmetric, in that one treatment may be up to k times more effective than the other, in either direction. Therefore, to build a confidence interval for OR or RR, we will not be following the usual formula

$$\text{point estimate } \pm z \times SE.$$

You may have already been thinking that a log transformation would be useful here, and you'd be correct! The *sort-of* symmetric intervals we've been discussing here actually are symmetric (about zero) on the log scale.

Firstly we'll consider the risk ratio. Let's define

$$\phi = \log\left(\frac{\pi_T}{\pi_C}\right).$$

The natural way to estimate this is with the sample proportions

$$\log\left(\frac{p_T}{p_C}\right) = \log\left(p_T\right) - \log\left(p_C\right).$$

These estimated proportions should be approximately normal and independent of one another, and so $\log\left(\frac{p_T}{p_C}\right)$ is approximately normal with mean ϕ (the true value) and variance

$$\text{var}\left(\log\left(p_T\right)\right) + \text{var}\left(\log\left(p_C\right)\right).$$

We can now apply the Delta method (see section 5.1) to find that (using Equation (5.3))

$$\text{var}\left[\log\left(p_T\right)\right] = \text{var}\left[\log\left(\frac{r_T}{n_T}\right)\right] \approx \frac{\pi_T\left(1-\pi_T\right)}{n_T} \times \left(\frac{1}{\pi_T}\right)^2 = \frac{1}{n_T \pi_T} - \frac{1}{n_T}.$$

Since we estimate π_T by r_T/n_T this can be estimated by $r_T^{-1} - n_T^{-1}$. Notice that we are relying on the derivative of $\log(x)$ being x^{-1}, so we must always use natural logarithms.

This leads us to the result that, approximately

$$\log\left(\frac{p_T}{p_C}\right) \sim N\Big(\phi,\, \left(r_T^{-1} - n_T^{-1}\right) + \left(r_C^{-1} - n_C^{-1}\right)\Big)$$

and so we can generate $100\left(1-\alpha\right)\%$ confidence intervals for ϕ as $(l_{RR},\, u_{RR})$, where the limits are

$$\log\left(\frac{p_T}{p_C}\right) \pm z_{\frac{\alpha}{2}} \sqrt{\left(r_T^{-1} - n_T^{-1}\right) + \left(r_C^{-1} - n_C^{-1}\right)}.$$

This then translates to an interval for the risk ratio itself of $\left(e^{l_{RR}}, e^{u_{RR}}\right)$.

Example 6.9. Returning once again to our streptomycin example, recall that we have

$$\begin{aligned} r_T &= 38 \\ n_T &= 55 \\ r_C &= 17 \\ n_C &= 52 \end{aligned}$$

and so the limits of the confidence interval (with $\alpha = 0.05$) on the log scale are

$$\log\left(\frac{38/55}{17/52}\right) \pm 1.96\sqrt{\left(\frac{1}{38} - \frac{1}{55}\right) + \left(\frac{1}{17} - \frac{1}{52}\right)} = \log(2.11) \pm 1.96 \times 0.218$$

which gives us (0.320, 1.176) on the log scale, and a 95% CI for the risk ratio of (1.377, 3.243).

So, we've seen that we can find confidence intervals for each of our four measures of difference. But we probably want to also be able to incorporate baseline measurements, as we did for continuous outcome variables.

6.3 Accounting for baseline observations: logistic regression

We saw with the continuous outcomes that it is often advantageous to include baseline measurements of the outcome (if they are known) in our analysis, and this is the same for binary outcomes.

In this section we use the term 'baseline observations' to mean any measurement that was known before the trial started. Unlike with continuous measurements, with a binary outcome, there is not usually a pre-trial value of the primary outcome. A binary outcome is often already relative to pre-trial (for example 'Have the patient's symptoms improved?') or refers to an event that definitely wouldn't have happened pre-trial (for example 'Did the patient die within the next 6 months?' or 'Was the patient cured?'). However, as we saw with ANCOVA, we can include other sorts of covariates in a linear model, so this is fine.

The general form of model that we would like for patient i is

$$\text{outcome}_i = \mu + \tau G_i + \beta_1 \times \text{baseline}_{1i} + \ldots + \beta_p \times \text{baseline}_{pi} + \text{error}_i,$$

where G_i is an indicator function taking values 1 if patient i was in group T and 0 if they were in group C, and baseline_1, ..., baseline_p are p baseline measurements that we would like to take into account.

However, this actually creates quite a few problems with binary variables. The outcome for patient i will be either 0 or 1, but the terms in the model above do not guarantee this at all. Adding a normally distributed error term doesn't really make sense in this context, so we will remove it. We can also make the LHS more continuous by thinking of the mean outcome rather than a single outcome. This makes sense, since if several patients were identical to patient i (in the sense of having the same baseline covariate values and being allocated to the same treatment), we probably wouldn't expect them all to have exactly the same outcome. Therefore we might instead think in terms of mean outcome, in which case our model becomes

$$\text{mean outcome}_i = \mu + \tau G_i + \beta_1 \times \text{baseline}_{1i} + \ldots + \beta_p \times \text{baseline}_{pi}.$$

There is one final problem to overcome, which is that the LHS will certainly be in $[0,\ 1]$, but the RHS could take any value. To address this we need to use a transformation, to take the mean outcome from $[0, 1]$ to $\mathbb{R}$.

The transformation that is usually used for a binary variable is the **logit** function, which is the log of the odds,

$$\text{logit}(\pi) = \log \frac{\pi}{1-\pi}.$$

As π tends to zero, $\text{logit}(\pi)$ tends to $-\infty$, and as π tends to one, $\text{logit}(\pi)$ tends to ∞. The derivative of the logit function is

$$\frac{d\,\text{logit}(\pi)}{d\pi} = \frac{1}{\pi(1-\pi)}$$

which is always positive for $\pi \in [0,1]$. This means that we can use it to transform our mean outcome (which we will now call π, since the mean outcome is the estimate of the probability of success) in the model

$$\text{logit}(\pi) = \mu + \tau G + \beta_1 \times \text{baseline}_1 + \ldots + \beta_p \times \text{baseline}_p \tag{6.4}$$

and any value in $\mathbb{R}$ is allowed on both sides. This model is known as **logistic regression**, and belongs to a class of models called **Generalized Linear Models**. If you did Advanced Statistical Modelling III you'll have seen these before. If you haven't seen them, and want to know more, this article gives a nice introduction (and some useful R tips!).

6.3.1 What does this model tell us?

We now have an equation for a model that makes sense, but what is it actually modelling? And what does it tell us about the effect of the treatment? Consider the difference between two patients who are the same in every respect except one is assigned to group C (so $G = 0$) and the other to group T (so $G = 1$). The model gives:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log(\text{Odds of success}) = \mu + \tau + \beta_1 x_1 + \ldots + \beta_p x_p \quad \text{(group T)}$$
$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log(\text{Odds of success}) = \mu + \beta_1 x_1 + \ldots + \beta_p x_p \quad \text{(group C)}$$

Subtracting one from the other, we find

$$\begin{aligned}\log(\text{Odds of success for group T}) &- \log(\text{Odds of success for group C}) \\ &= \log\left(\frac{\text{Odds of success for group T}}{\text{Odds of success for group C}}\right) = \log(OR) \\ &= \tau.\end{aligned}$$

That is, τ is the log of the odds ratio, or e^τ is the odds ratio adjusted for variables $x_1, \ldots, x_p$. Put another way, while the baseline covariates $x_1, \ldots, x_p$ affect the probability of 'success' (or whatever our binary outcome's one means), τ is a measure of the effect of the treatment compared to control given some set of baseline covariate values.

6.3.2 Fitting a logistic regression model

Logistic regression models are generally fitted using *maximum likelihood*. In the notation of Equation (6.4), the parameters we need to fit are the coefficients μ, τ and $\beta_1, \ldots, \beta_p$. To ease notation, we will collect these into a vector $\boldsymbol{\beta}$, with $\beta_0 = \mu$, $\beta_1 = \tau$ and $\beta_2, \ldots, \beta_{p+1}$ the original $\beta_1, \ldots, \beta_p$. Sorry this is confusing - we won't really use the vector $\boldsymbol{\beta}$ after this, or think about the parameters individually (apart from τ).

This notation allows us to write the linear function on the RHS of Equation (6.4) for participant i as

$$x_i^T \boldsymbol{\beta} = \sum_{j=0}^{q} x_{ij}\beta_j,$$

where

- $x_{i0} = 1$ (so that β_0 is the intercept μ)
- $x_{i1} = \begin{cases} 0 \text{ if participant } i \text{ is in group } C \\ 1 \text{ if participant } i \text{ is in group } T \end{cases}$
- $x_{i2}, \ldots, x_{iq}$ are the baseline covariates.

If π_i is the probability that the outcome for participant i is 1, where $i = 1, \ldots, n$, then the logistic model specifies these n parameters through the $q + 1$ parameters β_j, via the n expressions

$$\text{logit}\,(\pi_i) = x_i^T \boldsymbol{\beta}. \tag{6.5}$$

Using the Bernoulli distribution, the log-likelihood given data $y_1, \ldots, y_n$ is

$$\begin{aligned} \ell\left(\{\pi_i\} \mid \{y_i\}\right) &= \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1 - y_i) \log\left(1 - \pi_i\right)\right] \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log\left(1 - \pi_i\right)\right], \end{aligned}$$

where $y_i = 0$ or 1 is the outcome for participant i. Using Equation (6.5) we can rewrite this in terms of $\boldsymbol{\beta}$ as

$$\ell\left(\{\beta_j\} \mid \text{data}\right) = \sum_{i=1}^{n} \left[y_i x_i^T \boldsymbol{\beta} - \log\left(1 + e^{x_i^T \boldsymbol{\beta}}\right)\right].$$

The fitted model is then the one with the values β_j, $j = 0, \ldots, q$, that maximise this expression (and hence maximise the likelihood itself), which we will label the $\left\{\hat{\beta}_j\right\}$.

This is generally done some via some numerical method, and we won't go into that here. The method used by R will generate the MLE $\hat{\beta}_j$ for each β_j, and also an estimate of the standard error of each $\hat{\beta}_j$. In particular there will be an estimate of the standard error of $\hat{\beta}_1$, better known as $\hat{\tau}$, the estimate of the treatment effect. This is important, because it means we can test the hypothesis that $\tau = 0$, and can form a confidence interval for the adjusted log odds ratio.

Example 6.10. This study is detailed in Elmunzer et al. (2012). ERCP, or endoscopic retrograde cholangiopancreatogram, is a procedure performed by threading an endoscope through the mouth to the opening in the duodenum where bile and pancreatic digestive juices are released into the intestine. ERCP is helpful for treating blockages of flow of bile (gallstones, cancer), or diagnosing cancers of the pancreas, but has a high rate of complications (15-25%). The occurrence of post-ERCP pancreatitis is a common and feared complication, as pancreatitis can result in multisystem organ failure and death, and can occur in ~ 16% of ERCP procedures. This study tests whether the use of anti-inflammatory NSAID therapies at the time of ERCP reduce the rate of this complication. The study had 602 participants.

The dataset contains 33 variables, but we will focus on a small number:

- X: (primary outcome) - incidence of post-ercp pancreatitis 0 (no), 1 (yes).
- Treatment arm `rx`: 0 (placebo), 1 (treatment)
- Site: 1, 2, 3, 4

- Risk: Risk score (1 to 5). Should be factor but treated as continuous.
- Age: from 19 to 90, mean 45.27, SD 13.30.

The correlation between `risk` and `age` is -0.216, suggesting no problems of collinearity between those two variables.

Note: an obvious one to include would be `gender`, but I tried it and it is not at all significant, so I have pre-whittled it down for [even more] simplicity.

```
data("indo_rct")
summary(indo_rct[ ,c(1,2,3,4,6,32)])
```

```
##        id           site          age              risk        outcome                  rx
##  Min.   :1001   1_UM  :164   Min.   :19.00   Min.   :1.000   0_no :523   0_placebo     :307
##  1st Qu.:1152   2_IU  :413   1st Qu.:35.00   1st Qu.:1.500   1_yes: 79   1_indomethacin:295
##  Median :2138   3_UK  : 22   Median :45.00   Median :2.500
##  Mean   :1939   4_Case:  3   Mean   :45.27   Mean   :2.381
##  3rd Qu.:2289                3rd Qu.:54.00   3rd Qu.:3.000
##  Max.   :4003                Max.   :90.00   Max.   :5.500
```

```
## Some things to note:
# There are very few patients in group 4, and not many in group 3
# The age range goes from 19 to 90
# 'rx' is the group variable

## Checking for collinearity with factor variables

# No consistent patterns between age and site or risk and site
indo_rct%>%
  group_by(site) %>%
  summarise(
    meanage=mean(age), sdage=sd(age),
    meanrisk = mean(risk), sdrisk=sd(risk)
    )
```

site	meanage	sdage	meanrisk	sdrisk
1_UM	47.21951	14.19930	2.064024	0.8881557
2_IU	44.44552	12.90537	2.520581	0.8455696
3_UK	45.90909	11.59191	2.227273	0.8961196
4_Case	47.33333	22.89833	1.666667	0.2886751

```
## We will try models with age and age^2

glm_indo_agelin = glm(outcome ~ age + site + risk + rx, data=indo_rct,
                      family = binomial(link = "logit"))
glm_indo_agesq = glm(outcome ~ I(age^2) + site + risk + rx, data=indo_rct,
                     family = binomial(link = "logit"))

summary(glm_indo_agelin)
```

```
##
## Call:
## glm(formula = outcome ~ age + site + risk + rx, family = binomial(link = "logit"),
```

```
##     data = indo_rct)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.786293   0.641354  -2.785  0.00535 **
## age               -0.008458   0.009921  -0.853  0.39390
## site2_IU          -1.229290   0.269258  -4.565 4.98e-06 ***
## site3_UK          -1.127935   0.775917  -1.454  0.14603
## site4_Case       -13.864394 827.921132  -0.017  0.98664
## risk               0.561880   0.142342   3.947 7.90e-05 ***
## rx1_indomethacin  -0.763269   0.261538  -2.918  0.00352 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.07  on 595  degrees of freedom
## AIC: 441.07
##
## Number of Fisher Scoring iterations: 14
```

```
summary(glm_indo_agesq)
```

```
##
## Call:
## glm(formula = outcome ~ I(age^2) + site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.954e+00  4.930e-01  -3.963 7.39e-05 ***
## I(age^2)          -9.388e-05  1.081e-04  -0.869  0.38498
## site2_IU          -1.231e+00  2.693e-01  -4.571 4.87e-06 ***
## site3_UK          -1.135e+00  7.759e-01  -1.463  0.14355
## site4_Case        -1.385e+01  8.275e+02  -0.017  0.98664
## risk               5.593e-01  1.427e-01   3.919 8.88e-05 ***
## rx1_indomethacin -7.617e-01  2.614e-01  -2.914  0.00357 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.03  on 595  degrees of freedom
## AIC: 441.03
##
## Number of Fisher Scoring iterations: 14
```

Since neither `age` nor `age^2` appear influential, we'll remove it and keep the other covariates.

```
glm_indo = glm(outcome ~ site + risk + rx, data=indo_rct, family = binomial(link = "logit"))
summary(glm_indo)
```

```
##
## Call:
## glm(formula = outcome ~ site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.2307     0.3814  -5.848 4.97e-09 ***
## site2_IU          -1.2204     0.2689  -4.539 5.66e-06 ***
## site3_UK          -1.1289     0.7755  -1.456  0.14546
## site4_Case       -13.8400   833.2426  -0.017  0.98675
## risk               0.5846     0.1395   4.191 2.78e-05 ***
## rx1_indomethacin  -0.7523     0.2610  -2.883  0.00395 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.81  on 596  degrees of freedom
## AIC: 439.81
##
## Number of Fisher Scoring iterations: 14
```

From the summary we see that $\hat{\tau} = -0.752$, with a standard error of 0.261. A 95% CI for τ is therefore

$$-0.752 \pm 1.96 \times 0.261 = (-1.26,\ -0.240)\,.$$

This model supports the hypothesis that the treatment difference isn't zero. We do see however from the Null deviance and the Residual deviance that the model isn't explaining a huge proportion of the variation.

We can also use the model to estimate the odds of 'success' (the outcome 1) for different groups of patients, by fixing the values of the covariates. The linear expression $x^T\hat{\boldsymbol{\beta}}$ for given values of x gives us as estimate of

$$\log\left(\frac{p(X=1)}{1-p(X=1)}\right),$$

where X here is the primary outcome. The exponent of this therefore gives the odds, and this can be rearranged to find the probability,

$$p\left(X_i = 1\right) = \frac{\exp(\text{logit}_i)}{1+\exp(\text{logit}_i)},$$

where logit_i is the fitted value of the linear model (on the logit scale) given all the baseline characteristics of some patient i. This will be the probability, according to the model, that a patient with this particular combination of baseline characteristics will have outcome 1.

Example 6.11. Continuing with Example 6.10, we can find estimates of the log odds (and therefore the odds) of post-ECRP pancreatitis for various categories of patient.

For this we will make use of the summary table

```
summary(glm_indo)
```

```
##
## Call:
## glm(formula = outcome ~ site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.2307     0.3814  -5.848 4.97e-09 ***
## site2_IU          -1.2204     0.2689  -4.539 5.66e-06 ***
## site3_UK          -1.1289     0.7755  -1.456  0.14546
## site4_Case       -13.8400   833.2426  -0.017  0.98675
## risk               0.5846     0.1395   4.191 2.78e-05 ***
## rx1_indomethacin  -0.7523     0.2610  -2.883  0.00395 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.81  on 596  degrees of freedom
## AIC: 439.81
##
## Number of Fisher Scoring iterations: 14
```

For example, a patient from site 1, with risk level 3, in the control group would have odds

$$\exp\left(-2.2307 + 3 \times 0.5846\right) = 0.6207,$$

which translates to a probability of post-ECRP pancreatitis of

$$\frac{0.6207}{1 + 0.6207} = 0.383.$$

By contrast, a patient in group T, from site 2, at risk level 1, would have odds

$$\exp\left(-2.2307 - 1.2204 + 1 \times 0.5846 - 0.7523\right) = 0.0268,$$

which is equivalent to a probability of post-ECRP pancreatitis of

$$\frac{0.0268}{1 + 0.0268} = 0.0261.$$

Being more methodical we can collect these into a table, and use `predict.glm`. Since the site 3 and 4 coefficents are not significant (mainly due to a lack of data), we will focus only on the site 1 and 2 participants.

site	risk	rx	logodds	odds	prob
2_IU	1	0_placebo	-2.87	0.06	0.05
2_IU	2	0_placebo	-2.28	0.10	0.09
2_IU	3	0_placebo	-1.70	0.18	0.15
2_IU	4	0_placebo	-1.11	0.33	0.25
2_IU	5	0_placebo	-0.53	0.59	0.37
1_UM	1	0_placebo	-1.65	0.19	0.16
1_UM	2	0_placebo	-1.06	0.35	0.26
1_UM	3	0_placebo	-0.48	0.62	0.38
1_UM	4	0_placebo	0.11	1.11	0.53
1_UM	5	0_placebo	0.69	2.00	0.67
2_IU	1	1_indomethacin	-3.62	0.03	0.03
2_IU	2	1_indomethacin	-3.03	0.05	0.05
2_IU	3	1_indomethacin	-2.45	0.09	0.08
2_IU	4	1_indomethacin	-1.86	0.15	0.13
2_IU	5	1_indomethacin	-1.28	0.28	0.22
1_UM	1	1_indomethacin	-2.40	0.09	0.08
1_UM	2	1_indomethacin	-1.81	0.16	0.14
1_UM	3	1_indomethacin	-1.23	0.29	0.23
1_UM	4	1_indomethacin	-0.64	0.52	0.34
1_UM	5	1_indomethacin	-0.06	0.94	0.49

Some cautions

As with any linear model, we need to ensure that it is appropriate for our dataset. Two key things we need to check for are:

- **Collinearity**: we should make sure that none of the independent variables are highly correlated. This is not uncommon in clinical datasets, since measurements are sometimes strongly related. Sometimes therefore, this can mean choosing only one out of a collection of two or more strongly related variables.
- **linear effect across the range of the dataset**: a linear model is based on the assumption that the effect of the independent variables is the same across the whole range of the data. This is not always the case. For example, the rate of deterioration with age can be more at older ages. This can be dealt with either by binning age into categories, or by using a transformation, eg. age^2. Note that this would still be a linear model, because it is linear in the coefficients.

6.4 Diagnostics for logistic regression

There are many diagnostic techniques for binomial data (see eg. Collett (2003a)) but we will only touch on a small number. Unlike with a linear regression model, we don't have residuals to analyse, because our model output is fundamentally different from our data: our model outputs are probabilities, but our data is all either 0 or 1. Just because a particular patient had an outcome of `1`, we can't conclude that their probability should have been high. If the 'true' probability of $X = 1$ for some group of similar (in the baseline covariates sense) patients is 0.9, this means we should expect 1 in 10 of these patients to have $X = 0$.

This makes diagnostics somewhat trickier.

Diagnostics for logistic regression fall into two categories: **discrimination** and **calibration**. We will look at each of these in turn, though by no means exhaustively.

6.4.1 Discrimination

Here we are thinking of the logistic regression model as a classifier: for each participant the model outputs some value, on the logit (p) scale. If that value is below some threshold, we classify that participant as 0 If the value is above the threshold, we classify them as 1. Here, we are slightly abandoning the notion that the model is predicting probabilities, and instead testing whether the model can successfully order the patients correctly. Can we set some threshold on the model output that (almost) separates the cohort into its ones and zeros?

A classic way to asses this is by using Receiver Operating Characteric (ROC) analysis. ROC analysis was developed during the second world war, as radar operators analysed their classification accuracy in distinguishing signal (eg. an enemy plane) from noise. It is still widely used in the field of statistical classification, including in medical diagnostics. ROC analysis can be applied to any binary classifier, not just logistic regression.

6.4.1.1 ROC analysis

To understand ROC analysis, we need to revisit two concepts relating to tests or classifiers that you might not have seen since Stats I, and we will introduce (or remind ourselves of) some notation to do this:

- $\hat{p}_i \in (0, 1)$ is the fitted value of the logistic regression model for patient i
- $X_i = 0$ or 1 is the true outcome for patient i
- $t \in (0, 1)$ is the threshold value.

If $\hat{p}_i < t$ we classify patient i as 0, if $\hat{p} \geq t$ we classify them as 1. The language of ROC analysis is so entrenched in diagnostic/screening tests that I have kept it here for consistency. A 'positive' result for us is $X = 1$, and a 'negative' result is $X = 0$.

Definition 6.1. The **sensitivity** of a test (or classifier) is the probability that it will output positive (or 1) if the true value is positive (or 1):

$$p\left(\hat{p}_i \geq t \mid X_i = 1\right).$$

Definition 6.2. The **specificity** of a test (or classifier) is the probability that it will output negative (or 0) if the true value is negative (or 0):

$$p\left(\hat{p}_i < t \mid X_i = 0\right)$$

We estimate these by the proportions within the dataset.

These are very commonly used for thinking about diagnostic tests and screening tests, and in these contexts a 'success' or 'positive' is almost always the presence of some condition or disease. In our context, we need to be mindful that a 1 could be good or bad, depending on the trial.

The core part of a ROC analysis is to plot **sensitivity** against **1-specificity** for every possible value of the threshold. In a logistic regression context, the lowest the threshold can be is zero. If we set the $t = 0$, the model will predict everyone to have an outcome of 1. The sensitivity will be 1 and the specificity will be 0. At the other extreme, if we set $t = 0$, we will classify everyone as a 0, and have sensitivity 0 and specificity 1. If we vary the threshold from 0 to 1 the number of people classified in each group will change, and so will the sensitivity and specificity. This forms a **ROC curve**.

The dashboard below shows the distributions of fitted values for patients with $X = 0$ and $X = 1$, with options for good, moderate and poor separation, and the corresponding ROC curve. You can move the threshold to see the sensitivity and specificity at that value. Also note the AUC (area under the curve) which is an overall summary of the model's predictive efficacy. If AUC=1, the model is perfect. If AUC

is 0.5, the model is no better than random guessing. Generally it is thought that AUC around 0.8 is quite good, and AUC around 0.9 is excellent.

If you're viewing this in PDF you'll just have a static image, but you can find the dashboard at https://racheloughton.shinyapps.io/ROCplots/.

Note that I've used beta distributions for some hypothetical distributions of fitted values for the different groups, but this is just for convenience: ROC analysis makes no distributional assumptions.

Example 6.12. Let's look at the model we fitted in Example 6.10. To draw the ROC curve of this data, we will use the R package `pROC`.

```
fit_indo = fitted(glm_indo)    # Fitted values from glm_indo
out_indo = indo_rct$outcome    # outcome values (0 or 1)
roc_indo_df = data.frame(fit = fit_indo, out = out_indo)
```

The main function in the package `pROC` is `roc`, which creates a `roc` object. and `ggroc` that sort and plot the data for us:

```
roc_indo = roc(data=roc_indo_df, response = out, predictor=fit)
```

With that object we can do various things, such as plot the ROC curve:

```
ggroc(roc_indo, legacy.axes=T) + geom_abline(slope=1, intercept=0, type=2)
```

and find the area under the curve for the model

```
auc(roc_indo)
```

```
## Area under the curve: 0.7
```

So we see that our model is better than random guessing, but really not all that good! In particular, wherever we put a threshold (if we use the model that way), many people will be mis-classified. It's also worth noting that here we're performing the diagnostics on the data we used to fit the model: if we were to use the model on a new set of patients, the fit would likely be slightly worse.

6.4.2 Calibration

Now we are thinking of the model as actually predicting probabilities, and therefore we want to determine whether these probabilities are, in some sense, 'correct' or 'accurate'. One intuitive way to do this is to work through different 'types' of patient (by which we mean different combinations of baseline covariate values) and see whether the proportions of ones in the data broadly match the probability given by the model.

If the explanatory variables are factors, and we have repeated observations for the different combinations of factor levels, then for each combination we can estimate the probability of success (or whatever our outcome variable is) using the data, and compare this to the fitted model value.

Example 6.13. This example uses the model fitted in Example 6.10.

The trial has 602 participants and there are many fewer than 602 combinations of the above factor variables, so for many such combinations we will have estimates. Since we are in three dimensions, plotting the data is moderately problematic. We will have a plot for each site (or for the two main ones), use risk score for the x axis and colour points by treatment group. The circles show the proportions of ones in the data, and are sized by the number of observations used to calculate that estimate, and the crosses and lines show the mean and 95% CI of the fitted value.

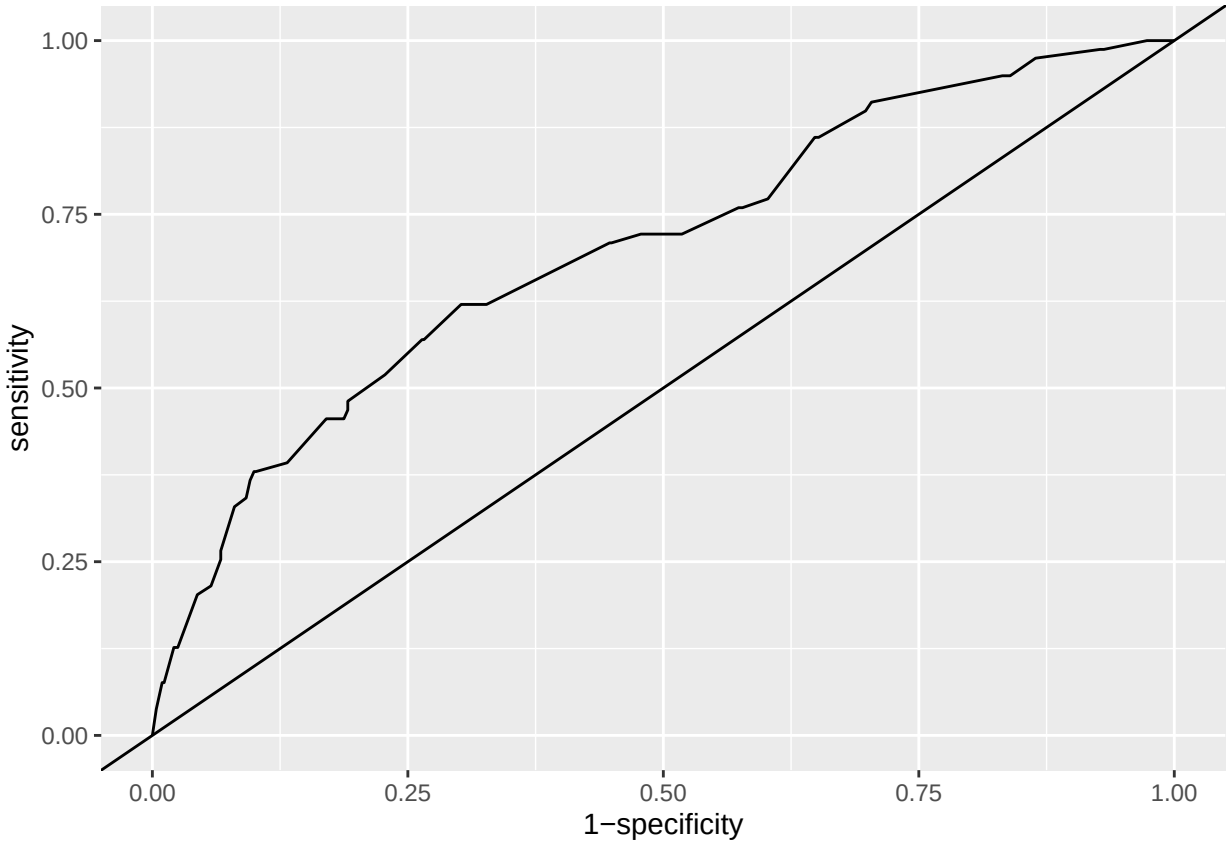


Figure 6.5: ROC curve for our logistic regression model of the indo RCT data (solid line). The dotted line shows the ROC curve we'd expect with random guessing.

```
## what columns do I want?
# site, risk, rx
# minus one for mean because outcome is 1 or 2
indo_sum = indo_rct %>%
  group_by(site, risk, rx) %>%
  summarise(est = mean(as.numeric(outcome))-1, size=length(id), .groups="keep")
indo_sum = indo_sum[indo_sum$site %in% c("1_UM", "2_IU"),]
fit_sum = predict(glm_indo, newdata=indo_sum[,1:3], se.fit=T, type="response")

indo_sum$fit = fit_sum$fit
indo_sum$fit_se = fit_sum$se.fit

ggplot(data=indo_sum, aes(x=risk, col=rx)) +
  geom_point(aes(y=est, size=size), pch=16) +
  geom_point(aes(y=fit), pch=4) +
  geom_segment(aes(x=risk, xend=risk, y=fit-1.96*fit_se, yend=fit+1.96*fit_se))+
  theme_bw()+
  facet_wrap("site") + theme(legend.position = "bottom")
```

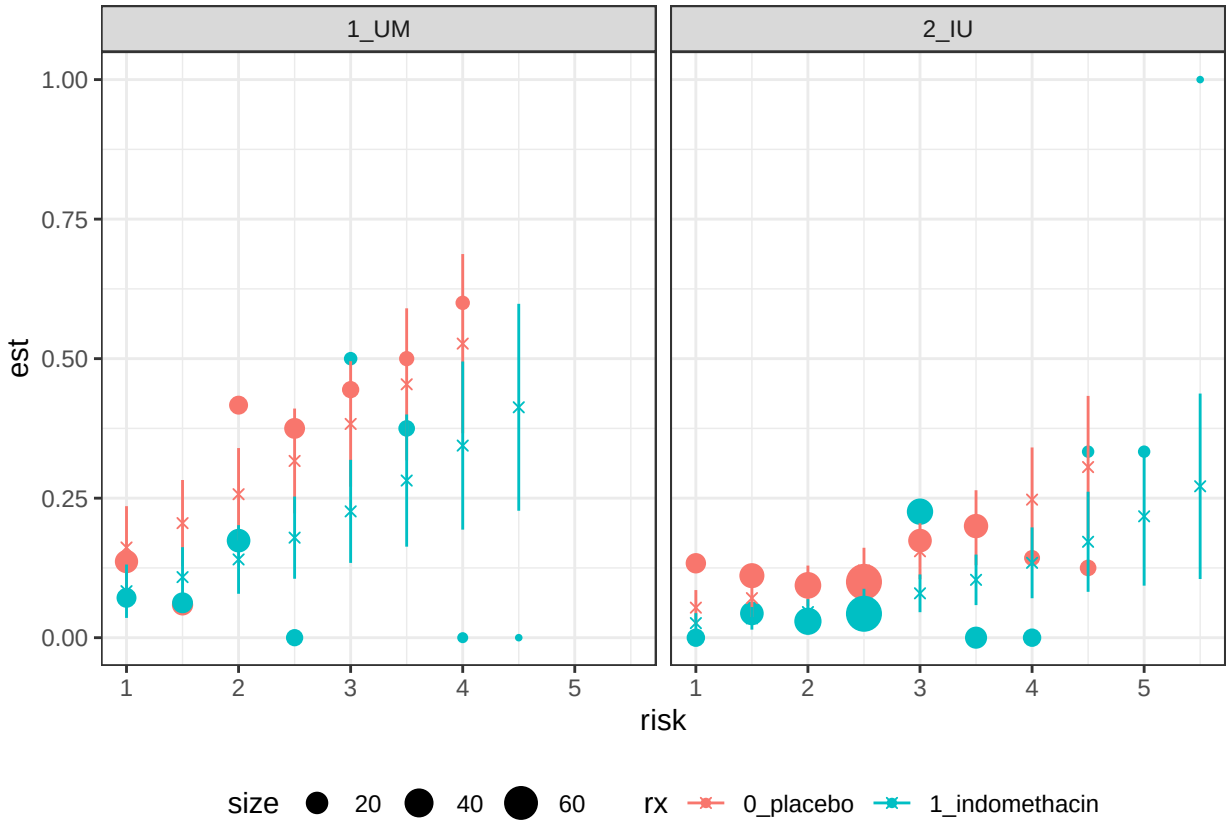


Figure 6.6: Calibration-based plots for indo RCT data for sites 1 (left) and 2 (right).

These plots are not the easiest to interpret, but there seems to be no evidence of systematic trends away from the model.

We will look some more at this in the upcoming practical class, as well as some further principles of model

validation.

For now, we're done with Binary data, and in our next few lectures we'll think about survival, or time-to-event data.

Part III

Part III: Survival data

Chapter 7

Working with time-to-event data

A data type that is commonly found in clinical trials is **time to event data**. This type of data deals with the amount of time that elapses before a particular event happens. As a sub-field of statistics, survival analysis has been around for a long time, as people have thought about and worked data like mortality records (most notably John Graunt, who used the 'Bills of Mortality' during the 1600s to better understand the plague and other causes of death). However, it developed rapidly during the many cancer related clinical trials of the 1960s and 1970s. In these cases, the event in question was very often death, and which is why this branch of statistics came to be known as **survival analysis**. However, the event can be many other things, and indeed can be a positive outcome (for example being cured of some condition). Time-to-event data also appears in other applications, such as engineering (eg. monitoring the reliability of a machine) and marketing (eg. thinking of the time-to-purchase). As well as the books already mentioned, this chapter makes use of Collett (2003b).

Usually, survival data is given in terms of time, but it can also be the number of times something happens (for example, the number of clinic appointments attended) before the event in question occurs.

Survival data is trickier to handle than the data types we have seen so far, for two main reasons. Firstly (and simply) survival data is very often skewed, so even though it is (usually) continuous, we can't just treat it as normally distributed. Secondly (and more complicatedly, if that's a word) with time-to-event data we don't usually observe the full dataset.

7.1 Censored times

If a trial monitors a sample of participants for some length of time, many will experience the event before the trial is finished. However, for some of the sample we likely won't observe the event. This could be because it doesn't happen within the lifetime of the trial, or it could be because the participant exits the trial prematurely for some reason (eg. withdrawal), or simply stops attending follow-up appointments after a certain times. For these participants, we do know that they had not experienced the event up to some time t, but we don't know what happened next. All we know is that their time-to-event or **survival time** is greater than that time t. These partial observations are known as **censored** times, and in particular as **right-censored** times, because the event happens *after* the censored time. It is possible (but less common) to have *left-censored* or *interval-censored* data, but in this course we will deal only with right-censoring.

If we were to treat censored times as observations, ie. as though the event had happened at time t, we would bias the results of the trial very seriously. The survival times reported would be systematically shorter than the true ones. For example, in the dataset shown in Figure 7.1, we would estimate the survival probability at time 10 as 0.2, since only two of the 10 participants were still in the trial after time 10. But it may well be that some of the participants whose observations were censored before $t = 10$ were still alive at $t = 10$.

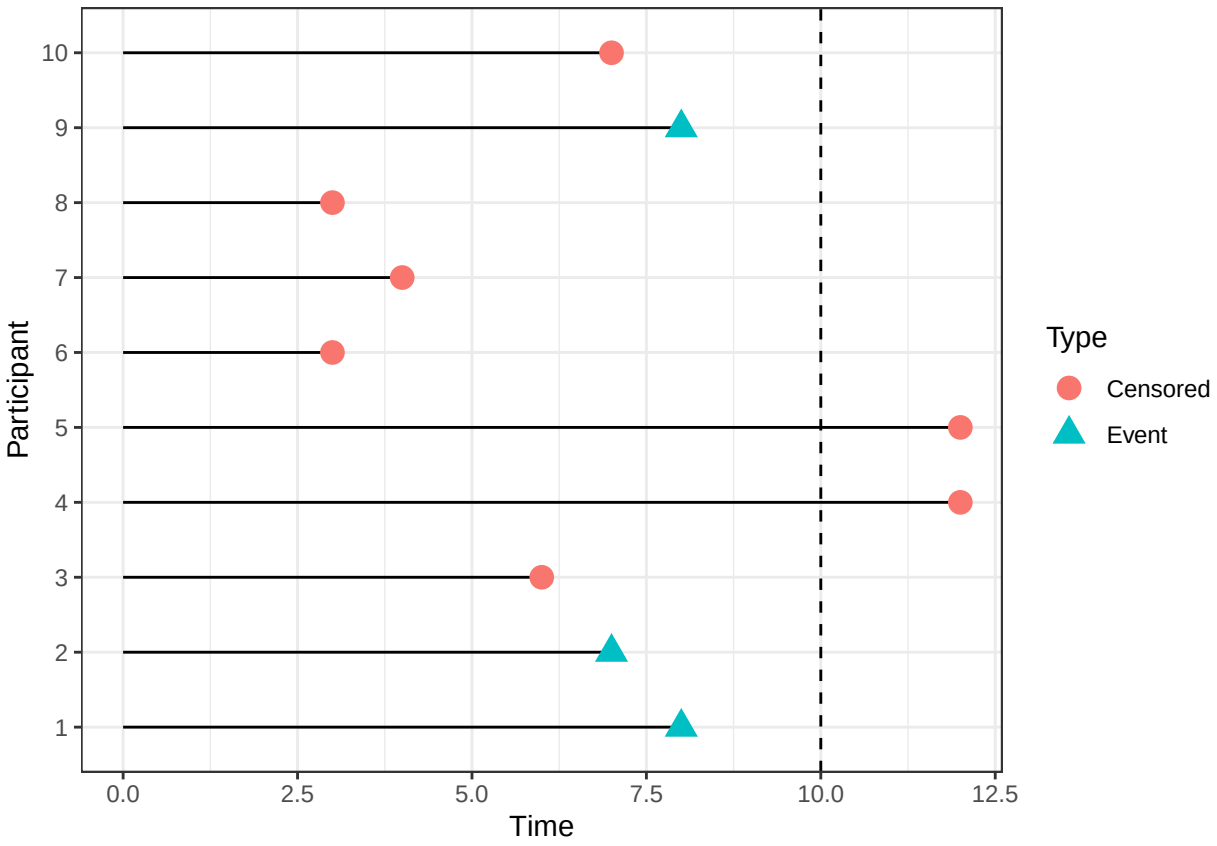


Figure 7.1: An example of some censored data.

If we were to remove the censored times, and only analyse the data in which the event was observed during the linespan of the trial, we would be losing data and therefore valuable information. This approach may well also lead to bias, for example if some subset of patients died quite soon into the trial, but the remainder lived a long time (past the end of the trial). If our analysis ignores the survivors, we are likely to underestimate the general survival time. In the dataset in Figure 7.1 there are five participants (3,6,7,8,10) whom we are no longer able to observe at time 10, but of whom none had experienced the event by the point at which they were censored.

So we know that we need to somehow include these censored times in our analysis. How we do so will depend on our approach.

7.2 The Survival Curve and the Hazard function

The field of survival analysis is relatively unusual in statistics, in that it isn't treated predominantly parametrically. For most continuous data, it is overwhelmingly common to work with the normal distribution and its friends (eg. the student's t distribution). Similarly binary data is dominated by the binomial distribution. Inference is therefore often focussed on the parameters μ, σ or p, as an adequate summary of the truth given whatever parameteric assumption has been made.

However, in survival analysis, it is often the case that we focus on the whole shape of the data; there isn't an accepted dominating probability distribution. In order to be able to deal with time-to-event data, we need to introduce some key ways of working with such data.

The **survival time** (or time-to-event) t for a particular individual can be thought of as the value of a random variable T, which can take any non-negative value. We can think in terms of a probability distribution over the range of T. If T has a probability distribution with underlying *probability density function* $f(t)$, then the *cumulative distribution function* is given by

$$F(t) = P(T < t) = \int_0^t f(u)\, du,$$

and this gives us the probability that the survival time is less than t.

Definition 7.1. The **survival function**, $\mathrm{S}(t)$, is the probability that some individual (in our context a participant) survives longer than time t. Therefore $\mathrm{S}(t) = 1 - F(t)$. Conventionally we plot $\mathrm{S}(t)$ against t and this gives us a **survival curve**.

We can immediately say two things about survival curves:

1. Since all participants must be alive (or equivalent) at the start of the trial, $\mathrm{S}(0) = 1$.
2. Since it's impossible to survive past $t_2 > t_1$ but not past time t_1, we must have $\frac{d\,\mathrm{S}(t)}{dt} \leq 0$, ie. $\mathrm{S}(t)$ is non-increasing.

Figure 7.2 shows two survival curves, comparing different therapies. We see that the hormonal therapy reduces the survival time slightly compared to no hormonal therapy.

Following on from the survival function, we have another (slightly less intuitive) quantity: the **Hazard function** $h(t)$.

Definition 7.2. The **Hazard function** $h(t)$ is the probability that an individual who has survived up to time t fails just after time t; in other words, the instantaneous probability of death (or *experiencing the event*) at time t.

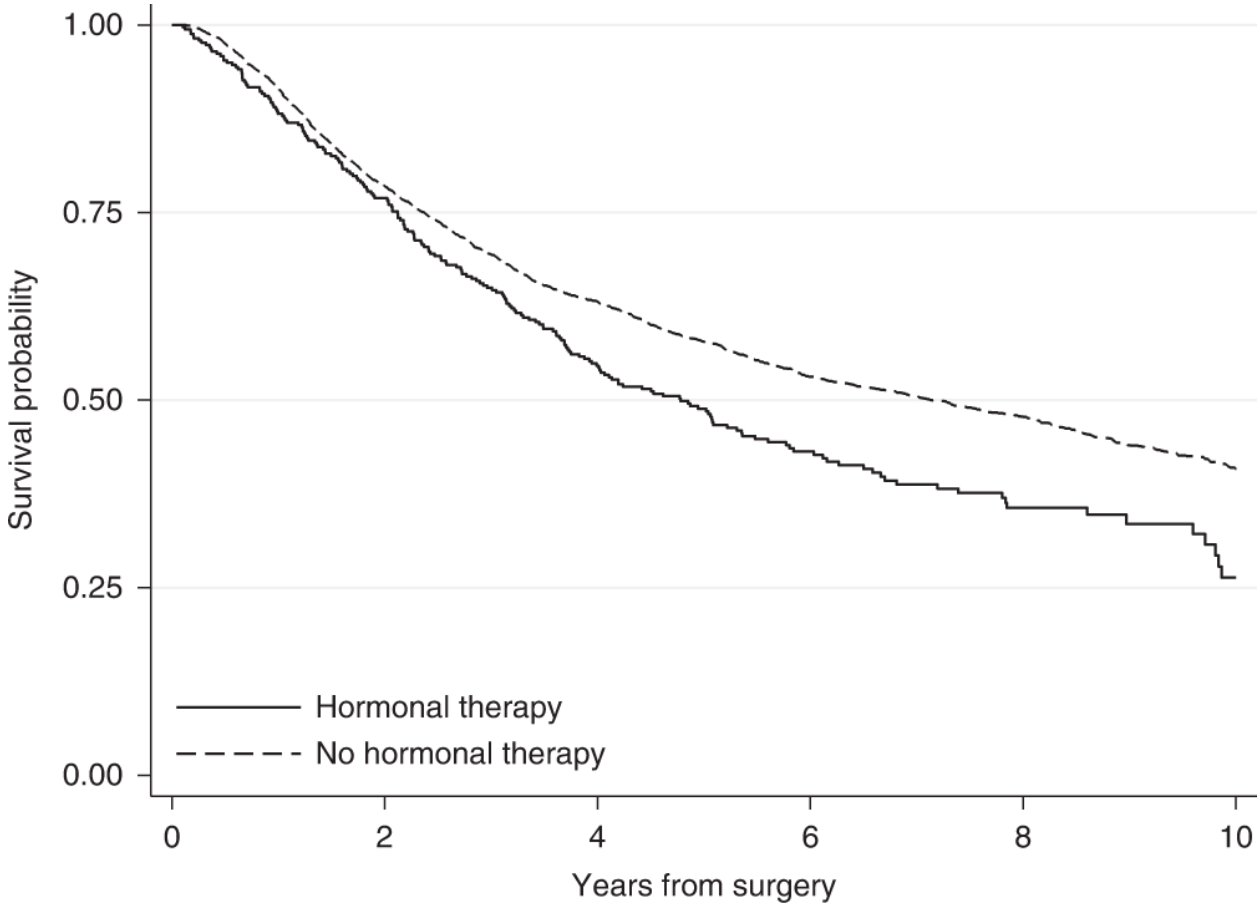


Figure 7.2: An example of two survival curves, taken from Syriopoulou et al. (2022).

If we use T to denote the random variable of survival time (or time-to-event) then $\mathrm{S}(t)$ and $h(t)$ are defined by

$$\begin{aligned}\mathrm{S}(t) &= \Pr(T > t)\\ h(t) &= \lim_{s\to 0+} \frac{\Pr(t < T < t+s \mid T > t)}{s}.\end{aligned}$$

Using the definition of conditional probability, we can rewrite $h(t)$ as

$$\begin{aligned}h(t) &= \lim_{s\to 0+} \frac{\Pr(t < T < t+s \mid T > t)}{s}\\ &= \lim_{s\to 0+} \left[\frac{1}{\Pr(T > t)} \cdot \frac{\Pr((t < T < t+s) \cap (T > t))}{s}\right]\\ &= \lim_{s\to 0+} \left[\frac{1}{\Pr(T > t)} \cdot \frac{\Pr(t < T < t+s)}{s}\right]\\ &= \frac{f(t)}{\mathrm{S}(t)},\end{aligned}$$

where $f(\cdot)$ is the probability density of T. The hazard function can take any positive value (unlike the survival function), and for this reason $\log(h(t))$ is often used to transform it to the real line. The hazard function can also be called the 'hazard rate', the 'instantaneous death rate', the 'intensity rate' or the 'force of mortality'.

As we hinted before, there are fundamentally two ways to deal with survival data: we can go about things either parametrically or non-parametrically. Unusually for statistics in general, the non-parametric paradigm is prevalent in survival analysis. We will consider some methods from both paradigms.

7.2.1 The Kaplan-Meier estimator

The **Kaplan-Meier estimator** is a non-parametric estimate of $\mathrm{S}(t)$, originally presented in Kaplan and Meier (1958). The idea behind it is to divide the interval $[0,\ t]$ into many short consecutive intervals,

$$[0,\ t] = \bigcup_{k=0}^{K} [s_k,\ s_{k+1}],$$

where $s_k < s_{k+1}\ \forall k$, $s_0 = 0$ and $s_{K+1} = t$. We then estimate the probability of surviving past some time t by multiplying together the probabilities of surviving the successive intervals up to time t. No distributional assumptions are made, and the probability of surviving interval $[s_k,\ s_{k+1}]$ is estimated by $1 - Q$, where

$$Q = \frac{\text{Number who die in that interval}}{\text{Number at risk of death in that interval}}.$$

More precisely, let's say that deaths are observed at times $t_1 < t_2 < \ldots < t_n$, and that the number of deaths at time t_i is d_i out of a possible n_i. Then for some time $t \in [t_J,\ t_{J+1})$, the Kaplan-Meier estimate of $\mathrm{S}(t)$ is

$$\hat{\mathrm{S}}(t) = \prod_{j=0}^{J} \frac{(n_j - d_j)}{n_j}.$$

Notice that the number of people at risk at time t_{j+1}, n_{j+1}, will be the number of people at risk at time t_j (which was n_j), minus any who died at time t_j (which we write as d_j) and any who were censored in the interval $[t_j,\ t_{j+1})$. In this way, the Kaplan-Meier estimator incorporates information from individuals with censored survival times up to the point they were censored.

Table 7.1: Ovarian cancer data. FU time gives the survival or censoring time, and FU status the type: 0 for a censored observation, 1 for death.

	FU_time	FU_status
1	59	1
2	115	1
3	156	1
22	268	1
23	329	1
24	353	1
25	365	1
26	377	0
4	421	0
5	431	1
6	448	0
7	464	1
8	475	1
9	477	0
10	563	1
11	638	1
12	744	0
13	769	0
14	770	0
15	803	0
16	855	0
17	1040	0
18	1106	0
19	1129	0
20	1206	0
21	1227	0

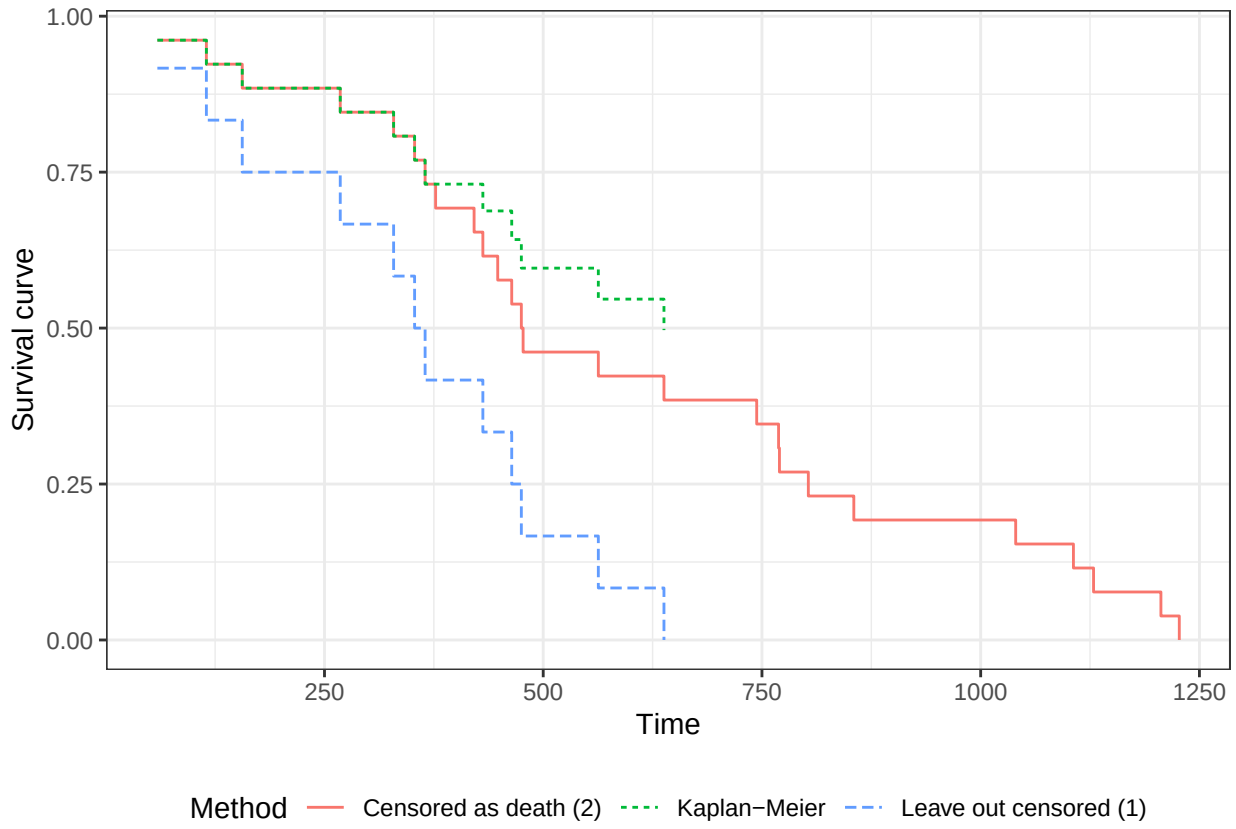
Example 7.1. Edmonson et al. (1979) conducted a trial on patients with advanced ovarian cancer, comparing cyclophosphamide (group C) with a mixture of cyclophosphamide and adriamycin (group T). Patients were monitored, and their time of death was recorded, or a censoring time if they were alive at their last observation. The data are shown in Table 7.1.

We see that there are 26 individuals, and we have the time of death for 12 of them. The remaining 14 observations are censored. We can use this data to calculate the Kaplan-Meier estimator of the survival curve, as shown in Table 7.2. The columns are (from left to right): time t_j; number at risk n_j; number of events/deaths d_j; number of censorings in $[t_{j-1}, t_j)$; estimate of survival curve.

To demonstrate the effect of including (and correctly treating) the censored data, we can do the same thing, but this time (1) leaving out all censored data and (2) treating all censored data times as deaths. Figure **??** shows the resulting survival curve estimates.

Table 7.2: Kaplan-Meier estimator calculations for ovarian cancer dataset.

time	n_risk	n_event	n_cens	survival
59	26	1	0	**0.9615385**
115	25	1	0	**0.9230769**
156	24	1	0	**0.8846154**
268	23	1	0	**0.8461538**
329	22	1	0	**0.8076923**
353	21	1	0	**0.7692308**
365	20	1	0	**0.7307692**
431	17	1	2	**0.6877828**
464	15	1	1	**0.6419306**
475	14	1	0	**0.5960784**
563	12	1	1	**0.5464052**
638	11	1	0	**0.4967320**



Leaving out the censored observations entirely is the most problematic approach, since it causes a serious underestimate, and in this case behaves as though there are no survivors after $t = 638$, which is untrue. Treating the censored data as deaths also leads to an underestimate of the survival probability, and notably creates a rather spurious curve past the last real death observation.

The 'correct' Kaplan-Meier estimate may seem a bit disatisfying, since it stops at $t = 638$ with a probabiilty of 0.497. However, this is really (in a non-parametric setting) all we can say with the data available; 10 of the participants were definitely still alive at $t = 638$, and some of the other censored participants may also have been.

For a clinical trial, we want to plot the survival curves separately for the different treatment groups. This will give a first, visual, idea of whether there might be a difference, and also of the suitability of certain models (we'll talk about this later).

Example 7.2. Figure 7.3 shows the Kaplan Meier plots for the ovarian cancer data from Figure **??**, this time split by treatment group.

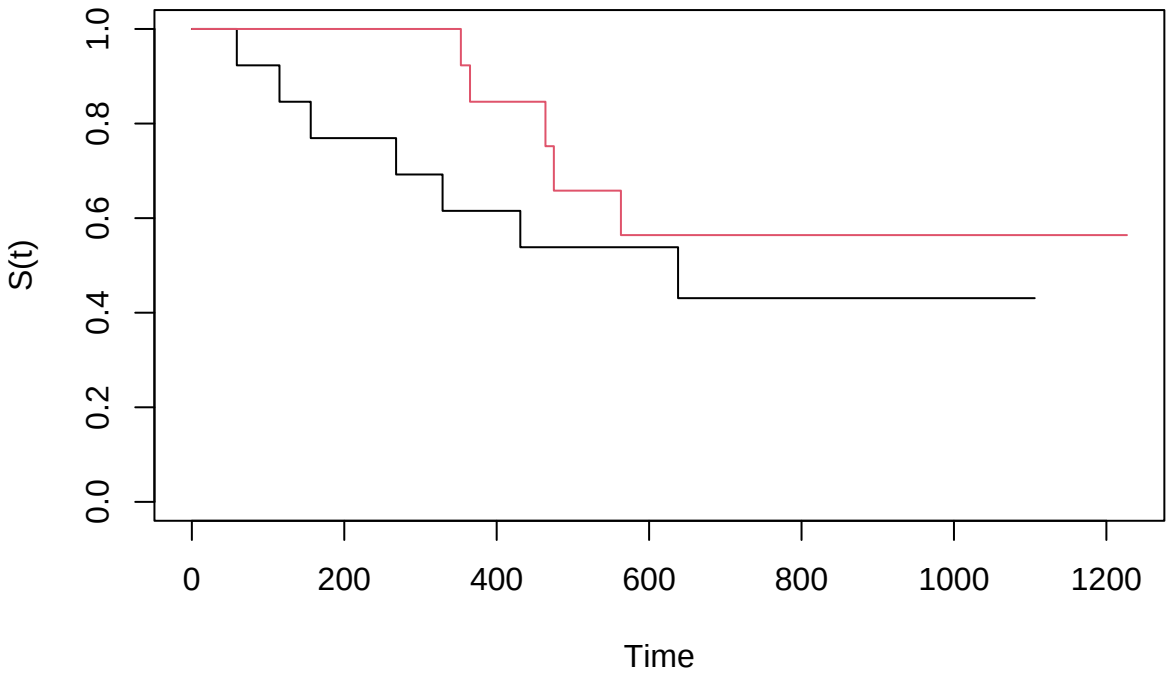


Figure 7.3: Kaplan Meier curves for the ovarian cancer data, split by treatment. Red is for group T (mixture of cyclophosphamide and adriamycin), black for group C (cyclophosphamide only).

The second dataset we will use throughout this chapter has been simulated based on a trial of acute myeloid leukemia (Le-Rademacher et al. (2018)) and is from the `survival` package Therneau (2024).

7.2.2 A parametric approach

In a parametric approach, we'll assume that the survival time T follows some probability distribution, up to unknown parameters which we will estimate from the data. The simplest distribution for time-to-event data is the *exponential distribution*, which has density

$$f(t) = \lambda e^{-\lambda t} \text{ for } t > 0,$$

survival function

$$\mathrm{S}(t) = 1 - \int_0^t \lambda e^{-\lambda t} = e^{-\lambda t},$$

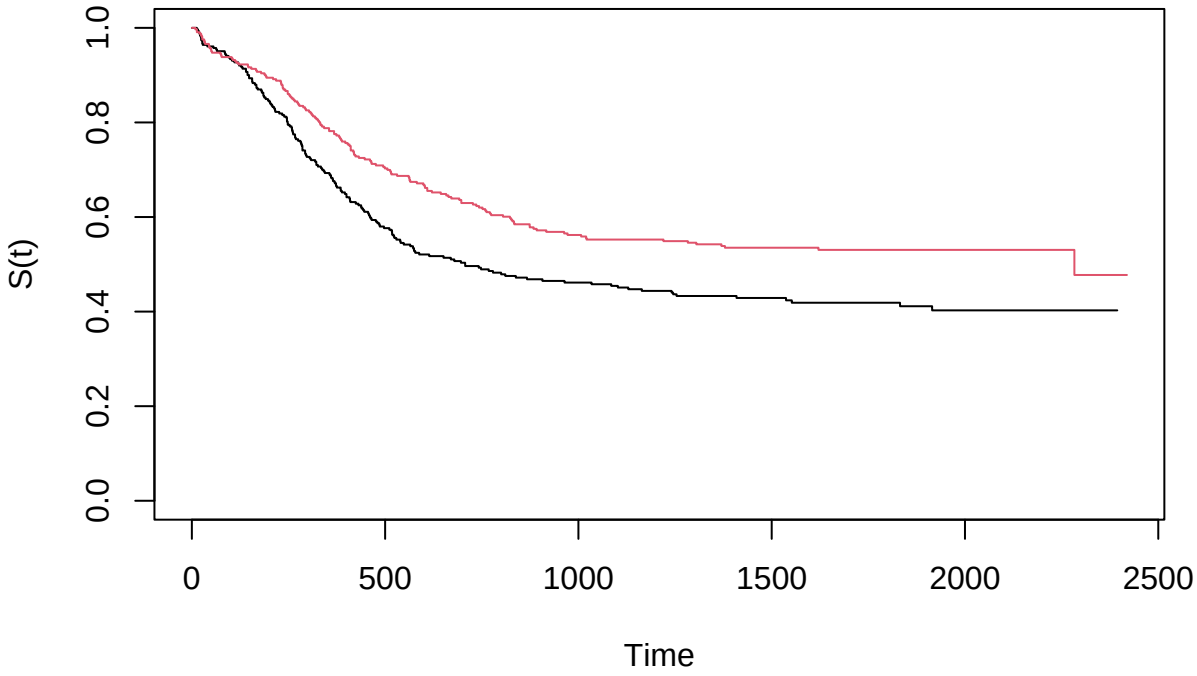


Figure 7.4: Kaplan Meier curves for the Myeloid data, split by treatment. Red is for group T, black for group C (placebo).

and mean survival time $\frac{1}{\lambda}$. The hazard function is therefore

$$h(t) = \frac{f(t)}{S(t)} = \lambda,$$

that is, the hazard is constant.

Given some dataset, we want to be able to find an estimate for λ (or the parameters of our distribution of choice).

7.2.2.1 Maximum likelihood for time-to-event data

Suppose our dataset has n times $t_1, t_2, \ldots, t_n$. Of these, m are fully observed and $n - m$ are censored. We can create a set of indicators $\delta_1, \ldots, \delta_n$, where $\delta_i = 1$ if observation i is fully observed and $\delta_i = 0$ if it is censored.

Usually, the likelihood function is computed by multiplying the density function evaluated at each data point, $f(t_i \mid \text{params})$. However, this won't work for survival data, because for our censored times (those for which $\delta_i = 0$) we only know that the time-to-event is greater than t_i. For these observations, it is the survival function (remember that this is $p(T > t)$) that contributes what we need to the likelihood function.

Therefore (for any probability distribution) we have

$$L = \prod_{i=1}^{n} f(t_i)_i^{\delta} S(t_i)^{1-\delta_i}. \tag{7.1}$$

If we have $T \sim Exp(\lambda)$ then the log-likelihood is

$$\begin{aligned}\ell &= \sum_{i=1}^{n} \delta_i (\log \lambda - \lambda t_i) - \sum_{i=1}^{n} (1 - \delta_i) \lambda t_i \\ &= m \log \lambda - \lambda \sum_{i=1}^{n} t_i.\end{aligned}$$

From this we can find the maximum likelihood estimator (MLE)

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i}.$$

The variance of the MLE is

$$\text{var}\left(\hat{\lambda}\right) = \frac{\lambda^2}{m}, \tag{7.2}$$

which we can approximate by

$$\text{var}\left(\hat{\lambda}\right) \approx \frac{m}{\left(\sum_{i=1}^{n} t_i\right)^2}.$$

Notice that the numerator in Equation (7.2) is m, the number of complete observations (rather than n the total number including censored observations). This shows that there is a limit to the amount we can learn if a lot of the data is censored.

Example 7.3. Returning to the dataset from Example 7.1, we can fit an exponential distribution to the data simply by estimating the MLE

$$\begin{aligned}\hat{\lambda}_C &= \frac{m_C}{\sum_{i=1}^{n_C} t_i} \\ &= \frac{7}{6725} \\ &= 0.00104\end{aligned}$$

and

$$\begin{aligned}\hat{\lambda}_T &= \frac{m_T}{\sum_{i=1}^{n_T} t_i} \\ &= \frac{5}{8863} \\ &= 0.00056\end{aligned}$$

```
mC_ov = sum((ovarian$fustat==1)&(ovarian$rx==1))
mT_ov = sum((ovarian$fustat==1)&(ovarian$rx==2))
tsum_ov_C = sum(ovarian$futime[ovarian$rx==1])
tsum_ov_T = sum(ovarian$futime[ovarian$rx==2])
m_ov = mT_ov + mC_ov
tsum_ov = tsum_ov_C + tsum_ov_T
lamhat_ov_C = mC_ov / tsum_ov_C
lamhat_ov_T = mT_ov / tsum_ov_T
```

We can do the same for the `myeloid` data. Figure 7.6 shows the fitted curves, using $S(t) = \exp\left[-\hat{\lambda}_X t\right]$ for group X.

7.2.3 The Weibull distribution

Having only one parameter, the exponential distribution is not very flexible, and often doesn't fit data at all well. A related, but more suitable distribution is the **Weibull distribution**.

Definition 7.3. The probability density function of a **Weibull** random variable is

$$f(x \mid \lambda, k) = \begin{cases} \lambda\gamma t^{\gamma-1} \exp\left[-\lambda t^{\gamma}\right] & \text{for } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Here, γ is the *shape* parameter, and λ is the *scale* parameter. If $\gamma = 1$ then this reduces to an exponential distribution. You can read more about it, should you choose to, in Collett (2003b).

For the Weibull distribution, we have

$$\mathrm{S}(t) = \exp\left(-\lambda t^{\gamma}\right).$$

As with the exponential distribution, we can use Equation (7.1) for the likelihood. For the Weibull distribution this becomes

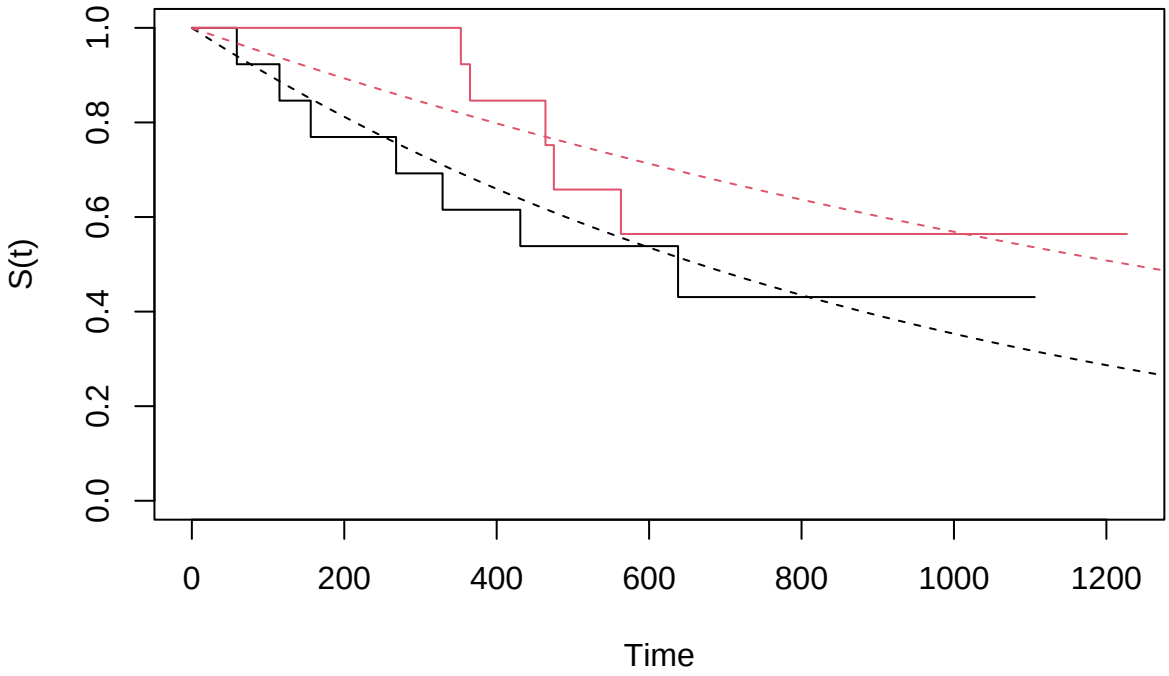


Figure 7.5: Kaplan Meier estimates of survival curves for the ovarian data (solid lines), with the fitted exponential S(t) shown in dashed lines (red = group T, black = group C).

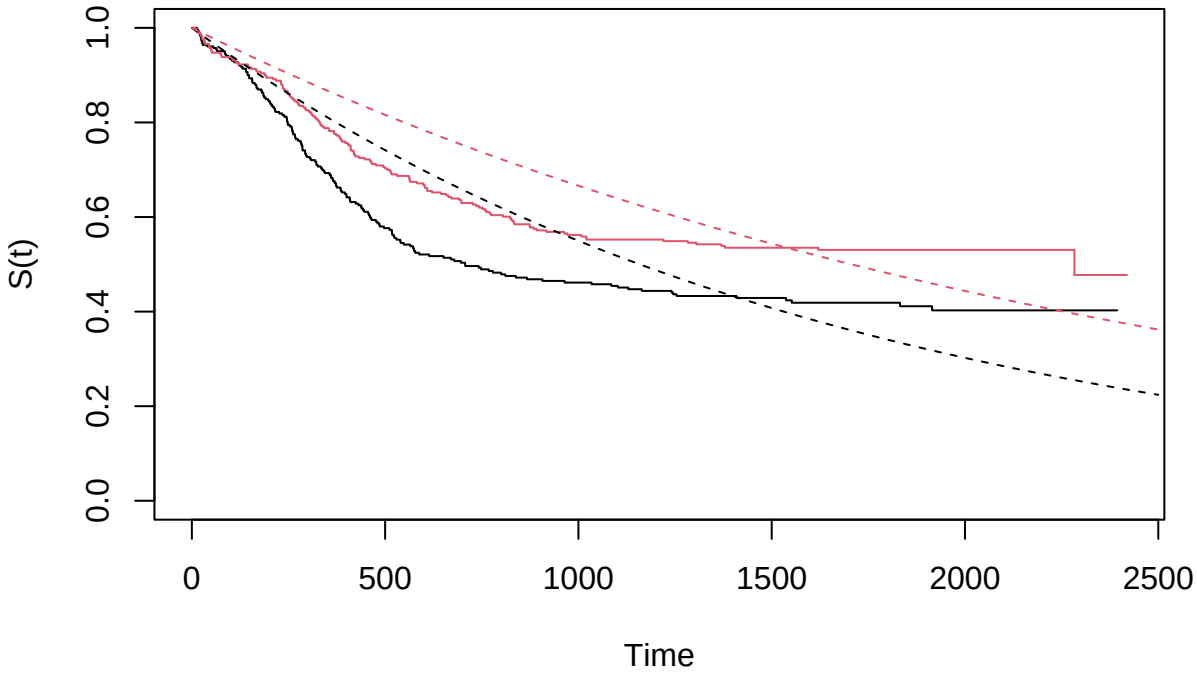


Figure 7.6: Kaplan Meier estimates of survival curves for the Myeloid data (solid lines), with the fitted exponential S(t) shown in dashed lines (red = group T, black = group C).

$$\begin{aligned} L(\lambda, \gamma \mid \text{data}) &= \prod_{i=1}^{n} \left\{ \lambda\gamma t_i^{\gamma-1} \exp\left(-\lambda t_i^{\gamma}\right) \right\}^{\delta_i} \left\{ \exp\left[-\lambda t_i^{\gamma}\right] \right\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left\{ \lambda\gamma t_i^{\gamma-1} \right\}^{\delta_i} \exp\left(-\lambda t_i^{\gamma}\right) \end{aligned}$$

and therefore

$$\begin{aligned} \ell(\lambda, \gamma \mid \text{data}) &= \sum_{i=1}^{n} \delta_i \log(\lambda\gamma) + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma} \\ &= r \log(\lambda\gamma) + (\gamma - 1) \sum_{i=1}^{n} \delta_i \log t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma}. \end{aligned}$$

For the maximum likelihood estimators, we differentiate (separately) with respect to λ and γ and equate to zero, to solve for the estimators $\hat{\lambda}$ and $\hat{\gamma}$.

The equations we end up with are

$$\frac{r}{\hat{\lambda}} - \sum_{i=1}^{n} t_i^{\hat{\gamma}} = 0 \tag{7.3}$$

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^{n} \delta_i \log t_i - \hat{\lambda} \sum_{i=1}^{n} t_i^{\hat{\gamma}} \log t_i = 0. \tag{7.4}$$

We can rearrange Equation (7.3) to

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\gamma}}},$$

and substitute this into Equation (7.4) to find

$$\frac{r}{\hat{\gamma}} + \sum_{i=1}^{n} \delta_i \log t_i - \frac{r}{\sum_{i=1}^{n} t_i^{\gamma}} \sum_{i=1}^{n} t_i^{\hat{\gamma}} \log t_i = 0.$$

This second equation is analytically intractable, so numerical methods are used to find $\hat{\gamma}$, and then this value can be used to find $\hat{\lambda}$.

Example 7.4. We can fit Weibull distributions to our `myeloid` dataset, as shown in Figure 7.7.

$$S(t) = \exp\left(-\lambda t^{\gamma}\right).$$

We see that there is some improvement compared to the exponential fit in Figure 7.6

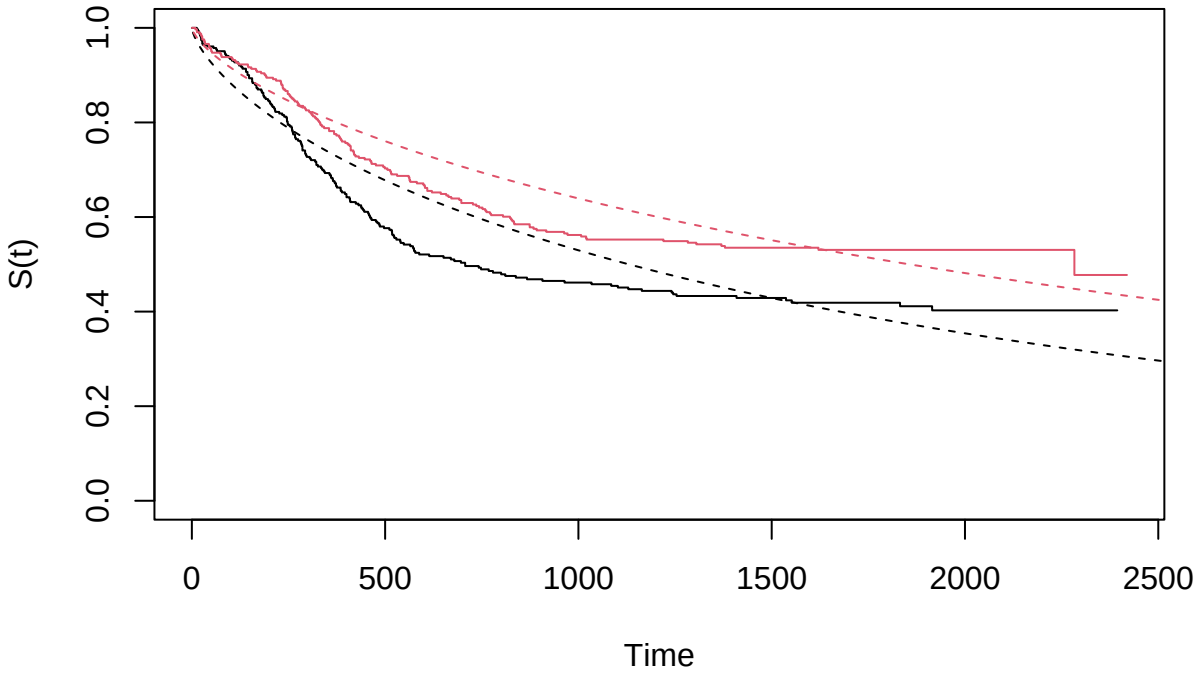


Figure 7.7: Weibull fit to survival curve of Myeloid data, dashed lines (Kaplan Meier estimate also shown in solid lines). Red for group T, black for group C.

Aside: Sample size calculations for time-to-event data

There are implications here for sample size calculations, which must take into account the duration of a trial; it is important that trials monitor patients until a sufficient proportion have experienced the event (whatever it is). Sample size calculations for time-to-event data therefore have two components:

1. The power of the trial can first be expressed in terms of m, the number of complete observations.
2. A separate calculation is needed to estimate the number of participants needing to be recruited, and length of trial, to be sufficiently likely to achieve that value of m.

Both of these calculations rely on a number of modelling assumptions, and on previous scientific/clinical data (if available).

We will think more about how this can be used in the next section, when we come to compare treatment effects.

Chapter 8

Comparing survival curves

Really what we would like to be able to do is to compare two survival curves (showing, for example, the results from different treatments), so that we can say whether one is significantly different from the other. In most cases, this boils down to constructing a hypothesis test along the lines of

$$H_0 : \text{the treatments are the same}$$
$$H_1 : \text{the treatments are different.}$$

There are various ways to do this, and we will look at some now.

8.1 Parametric: likelihood ratio test

For a parametric analysis, our null hypothesis that the two treatments are the same can be reduced to a test of whether the parameter(s) for each group are the same. We can do this using a likelihood ratio test. We've already calculated the log-likelihood for the exponential distribution in Section 7.2.2.1, and found the MLE.

$$\ell(\lambda) = m \log \lambda - \lambda \sum_{i=1}^{n} t_i$$
$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i}.$$

Working with the exponential distribution, we can model the survival function as

$$S(t) = \begin{cases} e^{-\lambda_C t} & \text{for participants in group C} \\ e^{-\lambda_T t} & \text{for participants in group T} \end{cases}$$

and the null hypothesis boils down to

$$H_0 : \ \lambda_C = \lambda_T = \lambda.$$

We can adapt the log-likelihood we found in Section 7.2.2.1 in light of the separate groups, and we find

$$\ell(\lambda_C, \lambda_T) = m_C \log \lambda_C - \lambda_C \sum_{i=1}^{n_C} t_{iC} + m_T \log \lambda_T - \lambda_T \sum_{i=1}^{n_T} t_{iT} \tag{8.1}$$

and

$$\hat{\lambda}_X = \frac{m_X}{\sum_{i=1}^{n_X} t_{iX}}$$

where $X = C$ or T. In these equations m_X is the number of non-censored observations in group X, n_X is the total number of participants in group X and t_{iX} is the time for participant i in group X. To simplify notation, we will write

$$t_X^+ = \sum_{i=1}^{n_X} t_{iX},$$

and t^+ for the sum over both groups.

Substituting the MLEs into Equation (8.1) gives

$$\ell\left(\hat{\lambda}_C, \hat{\lambda}_T\right) = m_C \log\left(\frac{m_C}{t_C^+}\right) - m_C + m_T \log\left(\frac{m_T}{t_T^+}\right) - m_T$$

and

$$\ell\left(\hat{\lambda}, \hat{\lambda}\right) = m \log\left(\frac{m}{t^+}\right) - m,$$

where n, m are the corresponding totals over both groups.

We can therefore perform a maximum likelihood test by finding

$$\begin{aligned}
\lambda_{LR} &= -2\left[\ell\left(\hat{\lambda}, \hat{\lambda}\right) - \ell\left(\hat{\lambda}_C, \hat{\lambda}_T\right)\right] \\
&= 2\left[\left(m_C \log\left(\frac{m_C}{t_C^+}\right) - m_C + m_T \log\left(\frac{m_T}{t_T^+}\right) - m_T\right) - \left(m \log\left(\frac{m}{t^+}\right)\right)\right] \\
&= 2\left(m_C \log\left(\frac{m_C}{t_C^+}\right) + m_T \log\left(\frac{m_T}{t_T^+}\right) - m \log\left(\frac{m}{t^+}\right)\right)
\end{aligned}$$

and referring this value to a χ_1^2 distribution.

We can also find a confidence interval for the difference between λ_T and λ_C, by using the asymptotic variances of the MLEs, which are $\frac{\lambda_C^2}{m_C}$ and $\frac{\lambda_T^2}{m_T}$. Therefore, the limits of a $100\,(1-\alpha)\%$ CI for $\lambda_T - \lambda_C$ is given by

$$\frac{m_T}{t_T^+} - \frac{m_C}{t_C^+} \pm z_{\alpha/2} \sqrt{\frac{m_T}{\left(t_T^+\right)^2} + \frac{m_C}{\left(t_C^+\right)^2}}.$$

Example 8.1. In this example we'll conduct a likelihood ratio test for each of the datasets in Example 7.1. For each dataset, the quantities we need are:

- m_C, m_T: the number of complete observations in each group
- t_C^+, t_T^+ the sum of all observation times (including censored times) in each group

Note that $m = m_C + m_T$ and $t^+ = t_C^+ + t_T^+$.

For the `ovarian` data we have

```
mC_ov = sum((ovarian$fustat==1)&(ovarian$rx==1))
mT_ov = sum((ovarian$fustat==1)&(ovarian$rx==2))
tsum_ov_C = sum(ovarian$futime[ovarian$rx==1])
tsum_ov_T = sum(ovarian$futime[ovarian$rx==2])
m_ov = mT_ov + mC_ov
tsum_ov = tsum_ov_C + tsum_ov_T

## Can now plug these into LR test stat
LRstat_ov =  2*(mC_ov*log(mC_ov/tsum_ov_C) + mT_ov*log(mT_ov/tsum_ov_T) - m_ov*log(m_ov/tsum_ov))
LRstat_ov
```

```
## [1] 1.114895
```

We can find the p-value of this test by

```
1-pchisq(LRstat_ov, df=1)
```

```
## [1] 0.2910204
```

and we find that it isn't significant. A 95% confidence interval for the difference is given by

```
## [1] -0.0013927697  0.0004392714
```

Figure 7.5 shows the fitted curves, using $S(t) = \exp\left[-\hat{\lambda}_X t\right]$ for group X, along with the Kaplan Meier estimate of the survival curve. Although there isn't much data, the exponential distribution looks to be an OK fit.

For the Myeloid data we can do the same thing

```
mC_my = sum((myeloid$death==1)&(myeloid$trt=="A"))
mT_my = sum((myeloid$death==1)&(myeloid$trt=="B"))
tsum_my_C = sum(myeloid$futime[myeloid$trt=="A"])
tsum_my_T = sum(myeloid$futime[myeloid$trt == "B"])
m_my = mT_my + mC_my
tsum_my = tsum_my_C + tsum_my_T

## Can now plug these into LR test stat
LRstat_my =  2*(mC_my*log(mC_my/tsum_my_C) + mT_my*log(mT_my/tsum_my_T) - m_my*log(m_my/tsum_my))
LRstat_my
```

```
## [1] 11.95293
```

Again, we refer this to χ_1^2:

```
1-pchisq(LRstat_my, df=1)
```

```
## [1] 0.0005456153
```

This time we find that the difference is significant at even a very low level, and the 95% CI is given by

```
## [1] -3.028814e-04 -8.108237e-05
```

Although the confidence around $\hat{\lambda}_X$ is high (ie. small standard error of the estimate), because of the large amount of data, the fit appears to actually be rather poor (recall Figure 7.6), mainly because of the inflexibility of the exponential distribution.

We could also perform LR tests with the fitted Weibull distributions, but instead we will continue on through some more commonly used methods.

8.2 Non-parametric: the log-rank test

The log-rank test is performed by creating a series of tables, and combining the information to find a test statistic.

We work through each time t_j at which an event is observed (by which we mean a death or equivalent, not a censoring) in either of the groups.

For notation, we will say that at time t_j,

- n_j patients are 'at risk' of the event
- d_j events are observed (often the 'event' is death, so we will sometimes say this)

For groups C and T we would therefore have a table representing the state of things at time t_j, with this general form:

Group	No. surviving	No. events	No. at risk
Treatment	$n_{Tj} - d_{Tj}$	d_{Cj}	n_{Cj}
Control	$n_{Cj} - d_{Cj}$	d_{Tj}	n_{Tj}
Total	$n_j - d_j$	d_j	n_j

Under H_0, we expect the deaths (or events) to be distributed proportionally between groups C and T, and so the expected number of events in group X (C or T) at time t_j is

$$e_{Xj} = n_{Xj} \times \frac{d_j}{n_j}.$$

This means that $e_{Cj} + e_{Tj} = d_{Cj} + d_{Tj} = d_j$.

If we take the margins of the table (by which we mean n_j, d_j, n_{Cj} and n_{Tj}) as fixed, then d_{Cj} has a **hypergeometric distribution**.

Definition 8.1. The **hypergeometric distribution** is a discrete probability distribution describing the probability of k successes in n draws (without replacement), taken from a finite population of size N that has exactly K objects with the desired feature. The probability mass function for a variable X following a hypergeometric function is

$$p\left(X = k \mid K, N, n\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

An example would be an urn containing 50 (N) balls, of which 16 (K) are green and the rest (34, $N - K$) are red. If we draw 10 (n) balls **without replacement**, X is the random variable whose outcome is k, the number of green balls drawn.

In the notation of the definition, the mean is

$$\mathrm{E}(X) = n\frac{K}{N}$$

and the variance is

$$\mathrm{var}(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}.$$

In the notation of our table at time t_j, we have

$$\mathrm{E}(d_{Cj}) = e_{Cj} = n_{Cj} \times \frac{d_j}{n_j}$$
$$\mathrm{var}(d_{Cj}) = v_{Cj} = \frac{d_j n_{Cj} n_{Tj} (n_j - d_j)}{n_j^2 (n_j - 1)}$$

With the marginal totals fixed, the value of d_{Cj} fixes the other three elements of the table, so considering this one variable is enough.

Under H_0, the numbers dying at successive times are independent, so

$$U = \sum_j (d_{Cj} - e_{Cj})$$

will (asymptotically) have a normal distribution, with

$$U \sim N\left(0, \sum_j v_{Cj}\right).$$

We label $V = \sum_j v_{Cj}$, and in the log-rank test we refer $\frac{U^2}{V}$ to χ^2_1.

A somewhat simpler, and more commonly used, version of the log-rank test uses the fact that under H_0, the expected number of events (eg. deaths) in group X is $E_X = \sum_j e_{Xj}$, and the observed number is $O_X = \sum_j d_{Xj}$. The standard χ^2 test formula can then be applied, and the test-statistic is

$$\frac{(O_C - E_C)^2}{E_C} + \frac{(O_T - E_T)^2}{E_T}.$$

It turns out that this test statistic is always smaller than $\frac{U^2}{V}$, so this test is slightly more conservative (ie. it has a larger p-value).

Notice that for both of these test statistics, the actual difference between observed and expected is used, not the absolute difference. Therefore if the differences change in sign over time, the values are likely to cancel out (at least to some extent) and the log-rank test is not appropriate.

Example 8.2. Let's now perform a log-rank test on our data from Example 8.1.

First, the ovarian cancer dataset. To do this, we can tabulate the key values at each time step.

Time	n_Cj	d_Cj	e_Cj	n_Tj	d_Tj	e_Tj	n_j	d_j
59	13	1	0.5000000	13	0	0.5000000	26	1
115	12	1	0.4800000	13	0	0.5200000	25	1
156	11	1	0.4583333	13	0	0.5416667	24	1
268	10	1	0.4347826	13	0	0.5652174	23	1
329	9	1	0.4090909	13	0	0.5909091	22	1
353	8	0	0.3809524	13	1	0.6190476	21	1
365	8	0	0.4000000	12	1	0.6000000	20	1
431	8	1	0.4705882	9	0	0.5294118	17	1
464	6	0	0.4000000	9	1	0.6000000	15	1
475	6	0	0.4285714	8	1	0.5714286	14	1
563	5	0	0.4166667	7	1	0.5833333	12	1
638	5	1	0.4545455	6	0	0.5454545	11	1

From this, we can find the v_j and the test statistic $\frac{U^2}{V}$:

```
# Add up the differences
UC = sum(logrank_df$d_Cj - logrank_df$e_Cj)
vCj_vec = sapply(
  1:n_event,
  function(j){
    nCj = logrank_df$n_Cj[j]
    nTj = logrank_df$n_Tj[j]
    dj = logrank_df$d_j[j]
    nj = logrank_df$n_j[j]

    (nCj*nTj*dj*(nj-1))/((nj^2)*(nj-1))
})
VC = sum(vCj_vec)
cs_ov_stat = (UC^2)/VC
1-pchisq(cs_ov_stat, df=1)
```

```
## [1] 0.3025911
```

For the simpler, more conservative, version of the log-rank test, we have

```
EC = sum(logrank_df$e_Cj)
ET = sum(logrank_df$e_Tj)
OC = sum(logrank_df$d_Cj)
OT = sum(logrank_df$d_Tj)

test_stat = ((EC-OC)^2)/EC + ((ET-OT)^2)/ET

test_stat
```

```
## [1] 1.057393
```

and we can find the p-value by

```
1-pchisq(test_stat, df=1)
```

```
## [1] 0.3038106
```

As we expected, slightly larger, but not much different from the first version. These values are also pretty close to the results of our LR test in Example 8.1, where we had $p = 0.291$.

Since the Myeloid dataset is much bigger, we won't go through the rigmarole of making the table, but will instead use an inbuilt R function from the `survival` package (more on this in practicals).

```
myeloid$trt = as.factor(myeloid$trt)
survdiff(Surv(futime, death) ~ trt, data = myeloid, rho=0)
```

```
## Call:
## survdiff(formula = Surv(futime, death) ~ trt, data = myeloid,
##     rho = 0)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=A 317      171      143      5.28      9.59
## trt=B 329      149      177      4.29      9.59
##
##  Chisq= 9.6  on 1 degrees of freedom, p= 0.002
```

This time the p-value is quite far from the one we found using the likelihood ratio test (p=0.00055), further supporting the view that the likelihood ratio test was not appropriate because of the poor fit of the exponential distribution.

8.3 Semi-parametric: the proportional hazards model

As with continuous and binary outcome variables, what we would really like to be able to do is to adjust our model for baseline covariates. It seems intuitively reasonable to suppose that factors like age, sex, disease status etc. might affect someone's chances of survival (or whatever event we're concerned with).

The conventional way to do this is using a **proportional hazards model**, where we assume that

$$h_T(t) = \psi h_C(t)$$

for any $t > 0$ and for some constant $\psi > 0$. We call ψ the **relative hazard** or **hazard ratio**. If $\psi < 1$ then the hazard at time t under treatment T is smaller than under control C. If $\psi > 1$ then the hazard at time t is greater in greater in group T than in group C. The important point is that ψ doesn't depend on t. The hazard for a particular patient might be greater than for another, due to things like their age, disease history, treatment group and so on, but the extent of this difference doesn't change over time.

We can adopt the concept of a **baseline hazard function** $h_0(t)$, where for someone in group C (for now), their hazard at time t is $h_0(t)$, and for someone in group T it is $\psi h_0(t)$. Since we must have $\psi > 0$, it makes sense to set

$$\psi = e^{\beta},$$

so that $\beta = \log \psi$ and $\psi > 0 \; \forall \beta \in \mathbb{R}$. Note that $\beta > 0 \iff \psi > 1$.

We can now (re)-introduce our usual indicator variable G_i, where

$$G_i = \begin{cases} 0 \text{ if participant } i \text{ is in group } C \\ 1 \text{ if participant } i \text{ is in group } T \end{cases}$$

and model the hazard function for participant i as

$$h_i(t) = \exp[\beta G_i] h_0(t).$$

This is the proportional hazards model for the comparison of two groups. Now, the relative hazard is a function of the participant's characteristics. Naturally, we can extend it to include other baseline covariates, as we have with linear models in ANCOVA, and with logistic regression.

8.3.1 General proportional hazards model

Extending the model to include baseline covariates $B_1, \ldots, B_p$, we have

$$\psi(\mathbf{x}_i) = \exp(\beta_0 G_i + \beta_1 b_{1i} + \ldots + \beta_p b_{pi}) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right),$$

where the first element of the vector $\mathbf{x}_i$ is G_i, and the hazard function for participant i is

$$h_i(t) = \psi(\mathbf{x}_i) h_0(t).$$

Now, our baseline hazard function $h_0(t)$ is the hazard function for a participant in group C for whom all baseline coviariates are either zero (if continuous) or the reference level (if a factor variable). For factor covariates this makes sense, since all levels are realistic values, but for continuous variables zero is likely to be unrealistic (for example you'd never expect zero for age, weight, height, blood pressure etc.). So, if any continuous variables are present, the baseline will always need to be adjusted, but if all covariates are factors, it is likely that the baseline hazard function will be applicable for some set of participants.

The linear component $\mathbf{x}_i^T \boldsymbol{\beta}$ is often called the **risk score** or **prognostic index** for participant i.

The general form of the model is therefore

$$h_i(t) = \exp\left[\mathbf{x}_i^T \boldsymbol{\beta}\right] h_0(t), \tag{8.2}$$

and we can rewrite it as

$$\log\left(\frac{h_1(t)}{h_0(t)}\right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Notice that there is no constant in the linear term - if there was, it could just be absorbed into the baseline hazard function.

There are ways of fitting this model that rely on specifying the hazard function using parametric methods, but the method we will study (and the most widely used) is one developed by Cox (1972).

8.3.2 Cox regression

The beauty of Cox regression is that it avoids specifying a form for $h_0(t)$ altogether.

To fit the model in Equation (8.2) we must estimate the coefficients $\beta_0, \ldots, \beta_p$. It also appears like we should estimate the baseline hazard $h_0(t)$ somehow too, but the great advance made by Cox was to develop a method where this isn't necessary. We don't need to estimate $h_0(t)$ to make inferences about the hazard ratio

$$\frac{h_i(t)}{h_0(t)}.$$

We will estimate the coefficients $\boldsymbol{\beta}$ using maximum likelihood, and so we'll need to specify a likelihood function for the $\boldsymbol{\beta}$, which will be a function of $\mathbf{x}^T\boldsymbol{\beta}$ and our observed data, the survival times t_i.

Suppose we have data for n participants, and that these include m complete observations (often referred to as deaths) and $n-m$ right-censored survival times. Suppose also that all the complete observation times are distinct. Since time itself is continuous, this is always technically true, but in data the time will be rounded and so there may be multiple observations at one time.

We can order these m event times

$$t_{(1)} < t_{(2)} < \ldots < t_{(m)},$$

such that $t_{(j)}$ is the time of the j^{th} event to be observed.

At time $t_{(j)}$, there will be some number of individuals who are 'at risk' of the event, because either their observation time or their censored survival time is greater than $t_{(j)}$. The set of these individuals is the **risk set**, denoted $R\left(t_{(j)}\right)$.

Cox (1972) shows that the relevant likelihood function for the proportional hazards model in Equation (8.2) is

$$L\left(\boldsymbol{\beta}\right) = \prod_{j=1}^{m} \frac{\exp\left[\mathbf{x}_{(j)}^T\boldsymbol{\beta}\right]}{\sum\limits_{l \in R\left(t_{(j)}\right)} \exp\left[\mathbf{x}_l^T\boldsymbol{\beta}\right]} \tag{8.3}$$

where $\mathbf{x}_{(j)}$ is the vector of covariates for the individual who dies (or equivalent) at time $t_{(j)}$. Notice that the product is over only those individuals with complete observations, but individuals with censored data do contribute to the sum in the denominator.

The numerator of the fraction inside the product in Equation (8.3) is the relative hazard for the person who actually did die at time $t_{(j)}$. The denominator is the sum of the relative hazards for all those who possibly could have died at time $t_{(j)}$ (the risk set $R\left(t_{(j)}\right)$). Thus, in very loose terms, maximizing the likelihood means finding values for $\boldsymbol{\beta}$ that mean the people who did die were 'the most likely' to die at the time they did.

Notice that this is not a true likelihood, since it depends only on the ordering of the data (the observation and censoring times) and not the data itself. This makes it a **partial likelihood**. The argument given to justify this is that because the baseline hazard $h_0\left(t\right)$ has an arbitrary form, it's possible that except for at these observed times, $h_0\left(t\right) = 0$, and therefore $h_i\left(t\right) = 0$. This means the intervals between successive observations convey no information about the effect of the covariates on hazard, and therefore about the $\boldsymbol{\beta}$ parameters.

If you want to know more detail about how this likelihood was derived, you can find in in Section 3.3 of Collett (2003b), or in Cox's original paper (Cox (1972)).

Moving on, if we set

$$\delta_i = \begin{cases} 0 & \text{if individual } i \text{ is censored} \\ 1 & \text{if individual } i \text{ is observed} \end{cases}$$

then we can write Equation (8.3) as

$$L\left(\boldsymbol{\beta} \mid \text{data}\right) = \prod_{i=1}^{n} \left(\frac{\exp\left[\mathbf{x}_i^T\boldsymbol{\beta}\right]}{\sum\limits_{l \in R(t_i)} \exp\left[\mathbf{x}_l^T\boldsymbol{\beta}\right]} \right)^{\delta_i},$$

where $R(t_i)$ is the risk set at time t_i.

From this we can find the log-likelihood

$$\ell(\boldsymbol{\beta} \mid \text{data}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^T \boldsymbol{\beta} - \log \sum_{l \in R(t_i)} \exp\left(\mathbf{x}_l^T \boldsymbol{\beta}\right) \right].$$

The MLE $\hat{\boldsymbol{\beta}}$ is found using numerical methods (often Newton-Raphson, which you'll have seen if you did Numerical Analysis II).

How can we tell if a proportional hazards model is appropriate?

We can't easily visualise the hazard function for a dataset, and instead would plot the survival curve. So can we tell if the proportional hazards assumption is met by looking at the survival curve?

Thankfully, it turns out that if two hazard functions are proportional, their survival functions won't cross one another.

Suppose $h_C(t)$ is the hazard at time t for an individual in group C, and $h_T(t)$ is the hazard for that same individual in group T. If the two hazards are proportional then we have

$$h_C(t) = \psi h_T(t)$$

for some constant ψ.

Recall from Section 7.2 that

$$h(t) = \frac{f(t)}{S(t)},$$

where $S(t)$ is the survival function and $f(t)$ is the probability density of T. We can therefore write

$$h(t) = -\frac{d}{dt}\left[\log\left(S(t)\right)\right]$$

and rearrange this to

$$S(t) = \exp\left(-H(t)\right) \tag{8.4}$$

where

$$H(t) = \int_0^t h(u)\, du.$$

Therefore for our two hazard functions, we have

$$\exp\left\{-\int_0^t h_C(u)\, du\right\} = \exp\left\{-\int_0^t \psi h_T(u)\, du\right\}$$

From Equation (8.4) we see that therefore

$$S_C(t) = \left[S_T(t)\right]^{\psi}.$$

Since the survival function is always between 0 and 1, we can see that the value of ψ determines whether $S_C(t) < S_T(t)$ (if $\psi > 1$) or $S_C(t) > S_T(t)$ (if $0 < \psi < 1$). The important thing is that **the survival**

curves will not cross. This is an informal conclusion, and lines not crossing is a necessary condition but not a sufficient one. There are some more formal tests that can be conducted to assess the proportional hazards assumption, but we won't go into them here.

Example 8.3. First of all, we can use Cox regression adjusted only for the Group (or treatment arm) of the participants.

For the `ovarian` dataset

```
coxph(formula = Surv(futime, fustat)~rx, data=ovarian)
```

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ rx, data = ovarian)
##
##       coef exp(coef) se(coef)      z    p
## rx -0.5964    0.5508   0.5870 -1.016 0.31
##
## Likelihood ratio test=1.05  on 1 df, p=0.3052
## n= 26, number of events= 12
```

and for the `myeloid1` dataset

```
coxph(formula = Surv(futime, death)~trt, data=myeloid)
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ trt, data = myeloid)
##
##         coef exp(coef) se(coef)      z       p
## trtB -0.3457    0.7077   0.1122 -3.081 0.00206
##
## Likelihood ratio test=9.52  on 1 df, p=0.002029
## n= 646, number of events= 320
```

We see that for both results, our p-values are close to what we have found with the log rank test.

We can now account for more baseline covariates. For the `ovarian` data we can include `age` and `resid.ds` (whether residual disease is present):

```
coxph(formula = Surv(futime, fustat)~rx+age+resid.ds, data=ovarian)
```

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ rx + age + resid.ds, data = ovarian)
##
##              coef exp(coef) se(coef)      z       p
## rx       -0.8489    0.4279   0.6392 -1.328 0.18416
## age       0.1285    1.1372   0.0473  2.718 0.00657
## resid.ds  0.6964    2.0065   0.7585  0.918 0.35858
##
## Likelihood ratio test=16.77  on 3 df, p=0.0007889
## n= 26, number of events= 12
```

What this shows is that the most significant factor by far is the particpant's age, with the hazard function increasing as age increases. The coefficient for treatment group (`rx`) has increased in magnitude and the p-value has decreased now that age is being adjusted for (although it is still not significant).

We can do the same for the `myeloid` data:

```
coxph(formula = Surv(futime, death)~trt+sex, data=myeloid)
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ trt + sex, data = myeloid)
##
##         coef exp(coef) se(coef)      z       p
## trtB -0.3582    0.6989   0.1129 -3.174 0.00151
## sexm  0.1150    1.1219   0.1128  1.020 0.30782
##
## Likelihood ratio test=10.56  on 2 df, p=0.005093
## n= 646, number of events= 320
```

We see that the only covariate we have, `sex` has very little effect.

This concludes our section on survival data, but we will revisit the topic in the next computer practical.

Part IV

Part III: Further designs

Chapter 9

Cluster randomised trials

For the trials we've been studying so far, the intervention is applied at an individual level. For many treatments this is realistic, for example a medicine, injection or operation. However, for some treatments this is not practical. One example would be implementing a new cleaning regime in operating theatres. It would be almost impossible to implement this if different patients within the same hospital would be allocated to different cleaning styles. Logistically it would be very difficult, and there would likely be contamination as staff may be reluctant to clean an operating theatre in what might now seem an inferior way, for a control participant. In general it is very difficult (if not impossible) to implement changes in practice across healthcare systems at an individual level.

The solution to this is to work at the group level, rather than the individual level.

9.1 What is a cluster RCT?

In a cluster RCT, participants within the same natural group (eg. doctor's surgery, hospital, school, classroom,...) are all allocated to the same group together. This means that, in the cleaning example above, the staff at a hospital in the treatment group can be trained in the new practice, all patients at that hospital will 'receive the new treatment', and contamination between groups is minimised.

The main issue that makes cluster RCTs different is that participants within the same group are often likely to be more similar than those in a different group. We expect that each group has its own 'true' mean μ_k , which is different from the underlying population mean μ, and that the cluster means are distributed with mean μ and variance σ_b^2 (more on σ_b^2 soon). This violates one of the key assumptions we've held so far, that the data are independent, and leads us to a very important quantity called the **intracluster correlation** (ICC).

9.1.1 Intracluster correlation

The ICC quantifies the relatedness of data that are clustered in groups by comparing the variance within groups by the variance between groups. The ICC is given by

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2},$$

where σ_b^2 is the variance **between** groups and σ_w^2 is the variance **within** groups. At one extreme, where $\sigma_w^2 = 0$, we have $ICC = 1$ and all measurements within each group are the same. At the other extreme, where $\sigma_b^2 = 0$, $ICC = 0$ and in fact all groups are independent and identically distributed.

We can estimate σ_w^2 and σ_b^2 using s_w^2 and s_b^2, which we find by decomposing the pooled variance. Here, g is the number of groups, n_j is the number of participants in group j and n is the total number of participants.

$$s_{Tot}^2 = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2}{n-g}$$

We can split this up as

$$\begin{aligned}
s_{Tot}^2 &= \frac{1}{n-g}\sum_{j=1}^{g}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j+\bar{x}_j-\bar{x})^2 \\
&= \frac{1}{n-g}\sum_{j=1}^{g}\sum_{i=1}^{n_j}\left[(x_{ij}-\bar{x}_j)^2+(\bar{x}_j-\bar{x})^2+2\,(x_{ij}-\bar{x}_j)\,(\bar{x}_j-\bar{x})\right] \\
&= \frac{1}{n-g}\sum_{j=1}^{g}\sum_{i=1}^{n_j}\left[(x_{ij}-\bar{x}_j)^2+(\bar{x}_j-\bar{x})^2\right] \\
&= \underbrace{\frac{1}{n-g}\sum_{j=1}^{g}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2}_{\text{Within groups}} + \underbrace{\frac{1}{n-g}\sum_{j=1}^{g}n_j\,(\bar{x}_j-\bar{x})^2}_{\text{Between groups}}
\end{aligned}$$

Example 9.1. We will demonstrate the ICC using a dataset that has nothing to do with clinical trials. The `cheese` dataset contains the price per unit and volume of sales of cheese (who knows what kind) at many Kroger stores in the US. We also know which city the Krogers are in, and we have data for 706 stores across 11 cities. It might be reasonable to expect that if we have information about the price and volume for several stores within a particular city, this gives us more information about the price and volume for another store in that same city than for a store in another city. Figure 9.1 shows the price and volume for all stores, coloured by city.

```
cheese = read.csv("kroger.csv", header=T)
cheese$city = as.factor(cheese$city)
ggplot(data=cheese, aes(x=price, y=vol, col=city)) + geom_point()
```

To calculate the ICC we define two functions, `between.var` and `within.var`, to calculate the between group and within group variance, as explained above.

Click to show R functions

```
# Firstly we define functions for the estimates
between.var = function(
    data,
    groupvec
){
  groups = levels(as.factor(groupvec))
  ng = length(groups)
  ntot = length(data)


  means = sapply(1:ng, function(i){mean(data[groupvec == groups[i]])})
  njvec = sapply(1:ng, function(i){length(data[groupvec == groups[i]])})
  mean = mean(data)
  ssqvec = sapply(1:ng, function(i){(njvec[i]*(means[i]-mean)^2)})
```

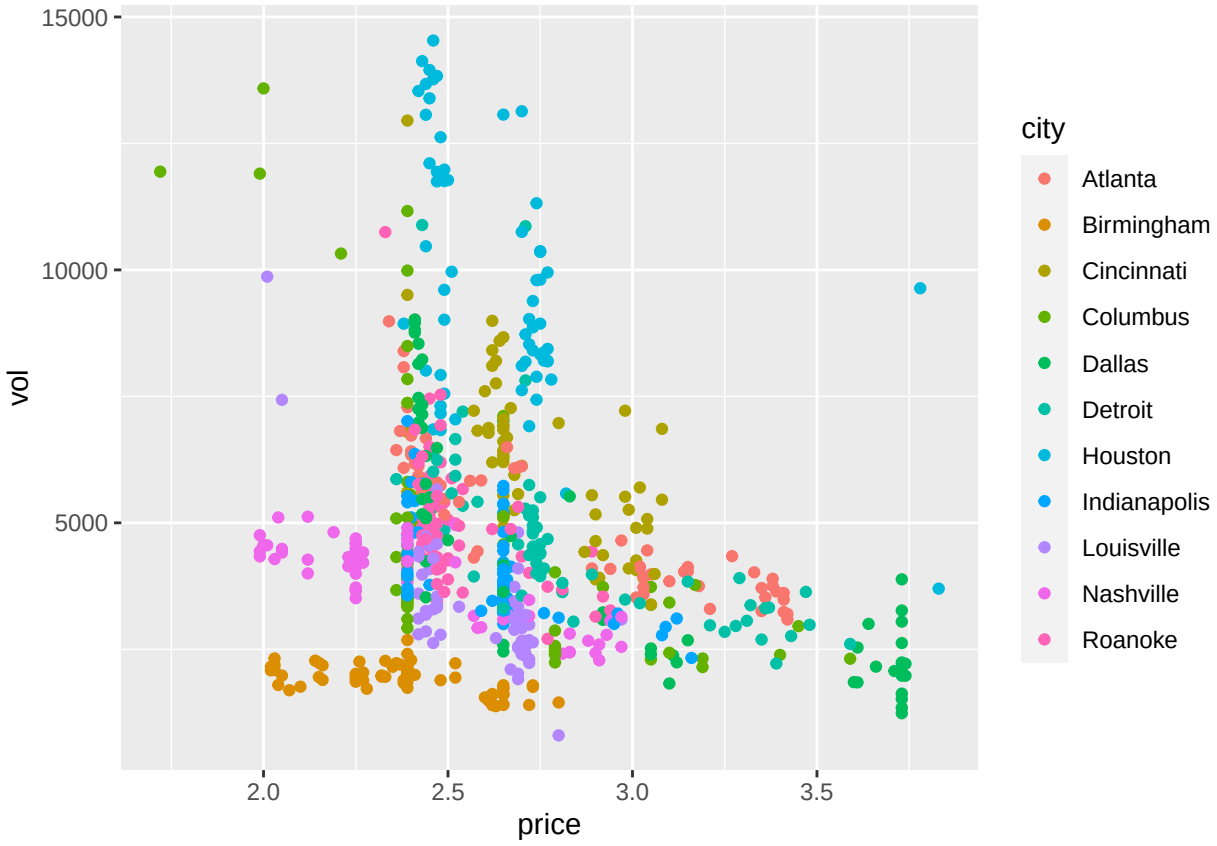


Figure 9.1: Price per unit and volume of sales of cheese for 706 Kroger stores.

```
  sum(ssqvec)/(ntot-ng)
}

within.var = function(
    data,
    groupvec
){
  groups = levels(as.factor(groupvec))
  ng = length(groups)
  ntot = length(data)

  means = sapply(1:ng, function(i){mean(data[groupvec == groups[i]])})
  njvec = sapply(1:ng, function(i){length(data[groupvec == groups[i]])})


  g_sums = rep(NA, ng)

  for (j in 1:ng){
    data_j = data[groupvec == groups[j]]
    ssqvec = rep(NA, njvec[j])
    for (i in 1:njvec[j]){
      ssqvec[i] = (data_j[i] - means[j])^2
    }
    g_sums[j] = sum(ssqvec)/(ntot - ng)
  }
  sum(g_sums)
}

## Now we can calculate them

bvar = between.var(iris$Sepal.Length, iris$Species)
bvar
```

```
## [1] 0.4300145
```

```
wvar = within.var(iris$Sepal.Length, iris$Species)
wvar
```

```
## [1] 0.2650082
```

```
## And find the ICC:

icc = bvar/(bvar+wvar)
icc
```

```
## [1] 0.6187057
```

```
wv_price = within.var(cheese$price, cheese$city)
bv_price = between.var(cheese$price, cheese$city)
icc = bv_price / (bv_price + wv_price)
icc
```

```
## [1] 0.253104
```

If we had to predict the price of cheese in a new city, all we can say is we expect the mean μ_{g+1} to come from $N\left(\mu,\ \sigma_b^2\right)$, where μ is the mean price of cheese in the overall population and σ_b^2 is the between group variance, and the individual cheese prices to come from $N\left(\mu_{g+1},\ \sigma_w^2\right)$.

Estimating the ICC when planning a study is an important step, but isn't always easy. For a well-understood (or at least well-documented) condition, it can often be estimated from existing data, which is likely to cover many sites. In non-medical studies like education or social interventions (where cluster RCTs are very common), it can be much more difficult because there is generally less data. Statistical studies are much newer in these areas, though they are becoming increasingly common, and even mandated by some organisations (for example the Educational Endowment Foundation).

9.2 Sample size

The upshot of the non-independence of the sample is that we have less information from n participants in a cluster RCT than we would do for an individual-based RCT where all the participants were independent (at least conditional on some covariates).

At one extreme, where ICC=0, there is in fact no intracluster correlation, all the groups have the same mean, and this is the same as a normal RCT. At the other extreme, where ICC=1, all measurements within a cluster are identical, and to achieve the same power as with n particpants in a standard RCT, we would need n clusters (and their size would be irrelevant). Obviously neither of these is ever true! In most studies, the ICC is in $(0,\ 0.15)$.

We will consider the sample size (and indeed most other things) for a cluster RCT in which the outcome is continuous (as in Chapter 2), but you can equally do a cluster RCT with a binary or time-to-event outcome.

The first step is to think about how the clustering affects the variance of the sample mean for either group. An estimate of the outcome variance in the control group, ignoring the clustering, is

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}\right)^2}{n-1}, \tag{9.1}$$

where as before there are g clusters, cluster j has size n_j and $\sum_{j=1}^{g} n_j = n$ is the total sample size.

The mean is also calculated without reference to the clustering, so

$$\bar{X} = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j} X_{ij}}{n}.$$

It can be shown that the variance of the overall mean in either group is inflated by a factor of

$$1 + \rho\left[\frac{\sum_j n_j^2}{N} - 1\right].$$

This quantity is known as the **design effect**. Notice that if all the groups are the same size $n_g = \frac{n}{g}$ then the design effect simplifies to

$$1 + \rho\left(n_g - 1\right).$$

We will assume from now on that this is the case.

9.2.0.1 A formula for sample size

Now that we know $\mathrm{E}\left(\hat{\sigma}^2\right)$ we can adapt our sample size formula from Section 2.5. For an individual-level RCT with a continuous outcome, we had

$$n = \frac{2\sigma^2\left(z_\beta + z_{\alpha/2}\right)^2}{\tau_M^2}, \tag{9.2}$$

and the reason we were able to do this was because the variance of the treatment effect estimate was σ^2/n. For a cluster RCT, the variance of the treatment effect is

$$\frac{\sigma^2}{n}\left[1 + \rho\left(\frac{\sum_{j=1}^{g} n_j}{n} - 1\right)\right] \tag{9.3}$$

At the planning stage of a cluster RCT we are unlikely to know the size of each cluster; each individual involved will usually need to give their consent, so knowing the size of the hospital / GP surgery / class is not enough. Instead, usually a [conservative] average cluster size n_g is specified, and this is used. In this case, the variance of the treatment effect in Equation (9.3) becomes

$$\frac{\sigma^2}{n}\left[1 + \rho\left(n_g - 1\right)\right]. \tag{9.4}$$

Equation (9.4) can be combined with Equation (9.2) to give the sample size formula for a cluster RCT:

$$n = \frac{2\sigma^2\left[1 + \rho\left(n_g - 1\right)\right]\left(z_\beta + z_{\alpha/2}\right)^2}{\tau_M^2}.$$

Since $n = n_g g$, this can be rearranged to find the number of clusters of a given size needed, or the size of cluster if a given number of clusters is to be used.

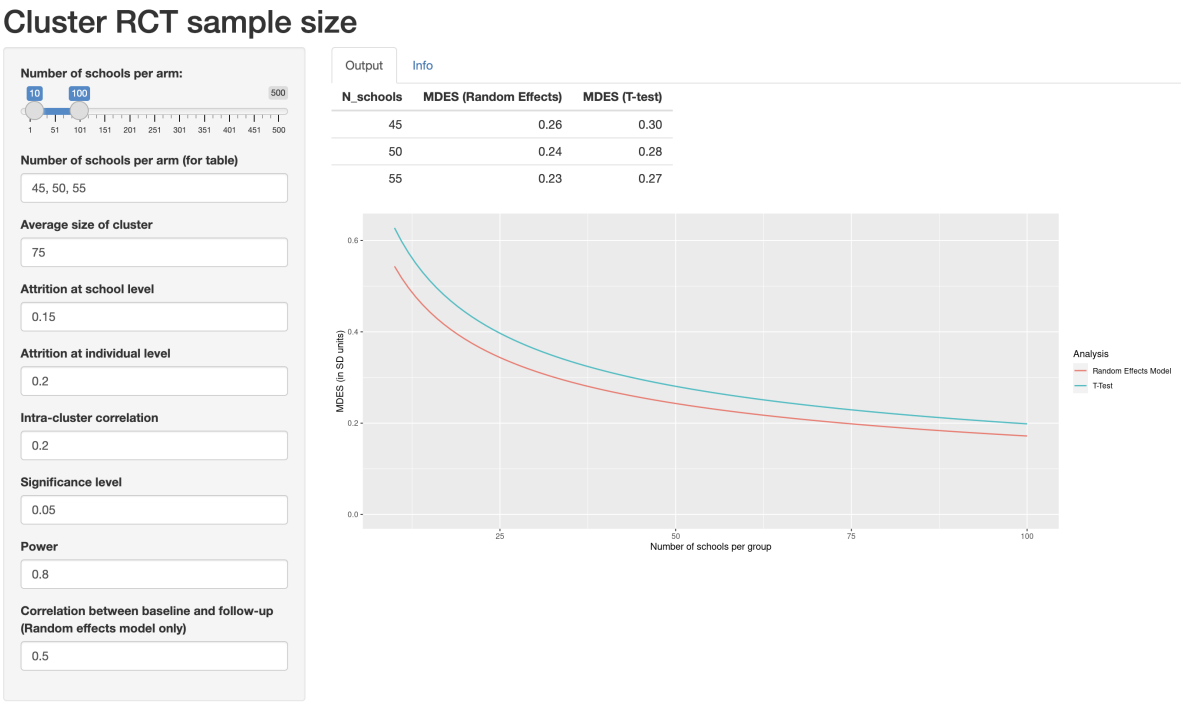
The sample size (and therefore the real power of the study) depends on two additional quantities that are generally beyond our control, and possibly knowledge: ICC and n_g. It is therefore sensible to conduct some sensitivity analysis, with several scenarios of ICC and n_g, to see what the implications are for the power of the study if things don't quite go to plan.

Example 9.2. This is something I wrote for an education study I'm involved in (funded by the EEF) where the treatment is a particular way of engaging 2 year olds in conversation. The outcome variable is each child's score on the British Picture Vocabulary Scale (BPVS), a test aimed at 3 - 16 year olds designed to assess each child's vocabulary. We needed to recruit some number of nurseries, but had very little information about the IIC (notice that the age in our study is outside the intended range of the BPVS test!).

To help the rest of the evaluation team understand the sensitivity of the power of the study to various quantities, I designed this dashboard, so that they could play around with the variables and see the effect.

As well as giving the sample size for a simple t-test (as we've done above) it also shows the size for a random effects model (similar to ANCOVA, more on this soon), which is why the baseline-outcome correlation (about which we also know very little!) is included.

The plot shows the minimum detectable effect size (MDES, or τ_M, in SD units), since the evaluation team wanted to know what size of effect we could find with our desired power.



9.3 Allocation

In a cluster RCT, everyone within a particular cluster will be in the same group (T or C). Therefore, the allocation needs to be performed at the cluster level, rather than at the individual level as we did in Chapter 3.

In theory we could use any of the first few methods we learned (simple random sampling, random permuted blocks, biased coin, urn design) to allocate the clusters. However, there are often relatively few clusters, and so the potential for imbalance in terms of the nature of the clusters would be rather high. This means we are more likely to use a stratified method or minimisation.

In terms of prognostic factors, there are now two levels: cluster level and individual level. For example, in a study with GP practices as clusters, some cluster-level covariates could be the size of the practice, whether it was rural or urban, the IMB (index of mass deprivation) of the area it was in. It would be sensible to make sure there was balance in each of these in the allocation. One might also include aggregates of individual-level characteristics, for example the mean age, or the proportion of people with a particular condition (especially if the study relates to a particular condition).

However, a key feature of cluster RCTs means that in fact some different, and perhaps more effective, allocation methods are open to us.

9.3.1 Allocating everyone at once

The methods we've covered so far assume that participants are recruited sequentially, and begin the intervention at different points in time. In this scenario, when a particular participant (participant n) is allocated we only know the allocation for the previous $n-1$ participants. It is very likely that we don't know the details of the following participants, in particular their values of any prognostic variables. This makes sense

in many medical settings, where a patient would want to begin treatment as soon as possible, and there may be a limited number of patients with particular criteria at any one time.

However, cluster RCTs rarely deal with urgent conditions (at least in the sense of administering a direct treatment), and so the procedure is usually that the settings (the clusters) are recruited over some recruitment period and all begin the intervention at the same time. This means that at the point of allocation, the details of all settings involved are known. There are a couple of proposals for how to deal with allocation in this scenario, and we will look at one now.

9.3.2 Covariate constrained randomization

This method is proposed in Dickinson et al. (2015), and implemented in the R package `cvcrand`. We'll review the key points of the method, but if you're interested you can find the details in the article.

Baseline information must be available for all settings, for any covariate thought to be potentially important. These can be setting-level variables or aggregates of individual-level variables. Once all this data has been collected, the randomisation procedure is as follows.

Firstly, generate all possible allocations of the clusters into two arms (T and C).

Secondly, rule out all allocations that don't achieve the desired balance criteria.

For a categorical covariate, the procedure is very simple. For example, we may stipulate that we want groups T and C to have the same number of rural GP practices as one another. In this case, we would remove from our set of possible allocations any where the numnber was different, or perhaps where it differed by more than some number d_{rural}. We continue for all the covariates we want to balance, setting rules for each one.

Continous covariates are standardized and used to calculate a 'balance score' B for each of the remaining allocations. A cut-off is used to rule out all allocations that don't achieve the desired level of balance. This leaves an 'optimal set' of allocations.

Finally, an allocation is chosen at random from the optimal set, and this is the allocation that is used.

9.4 Analysing a cluster RCT

As with the other stages of a cluster RCT, to conduct an effective and accurate analysis we need to take into account the clustered nature of the data. There are several ways to do this, and we will whizz through the main ones now.

Example 9.3. The data we use will be from an educational trial, contained in `crtData` in the package `eefAnalytics`, shown in Figure 9.2 . The dataset contains 22 schools and 265 pupils in total. Each school was assigned to either `1` (group T) or `0` (group C). Each pupil took a test before the trial, and again at the end of the trial. We also know the percentage attendance for each pupil. We will use this data to demonstrate each method.

```
## 'data.frame':    265 obs. of  4 variables:
##  $ School      : Factor w/ 22 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Posttest    : num  16 13 18 14 25 13 23 26 16 8 ...
##  $ Intervention: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Prettest    : num  1 4 5 4 5 2 5 5 2 2 ...
```

```
ggplot(data=crt_df, aes(y=Posttest, fill=Intervention, group = School)) + geom_boxplot()
```

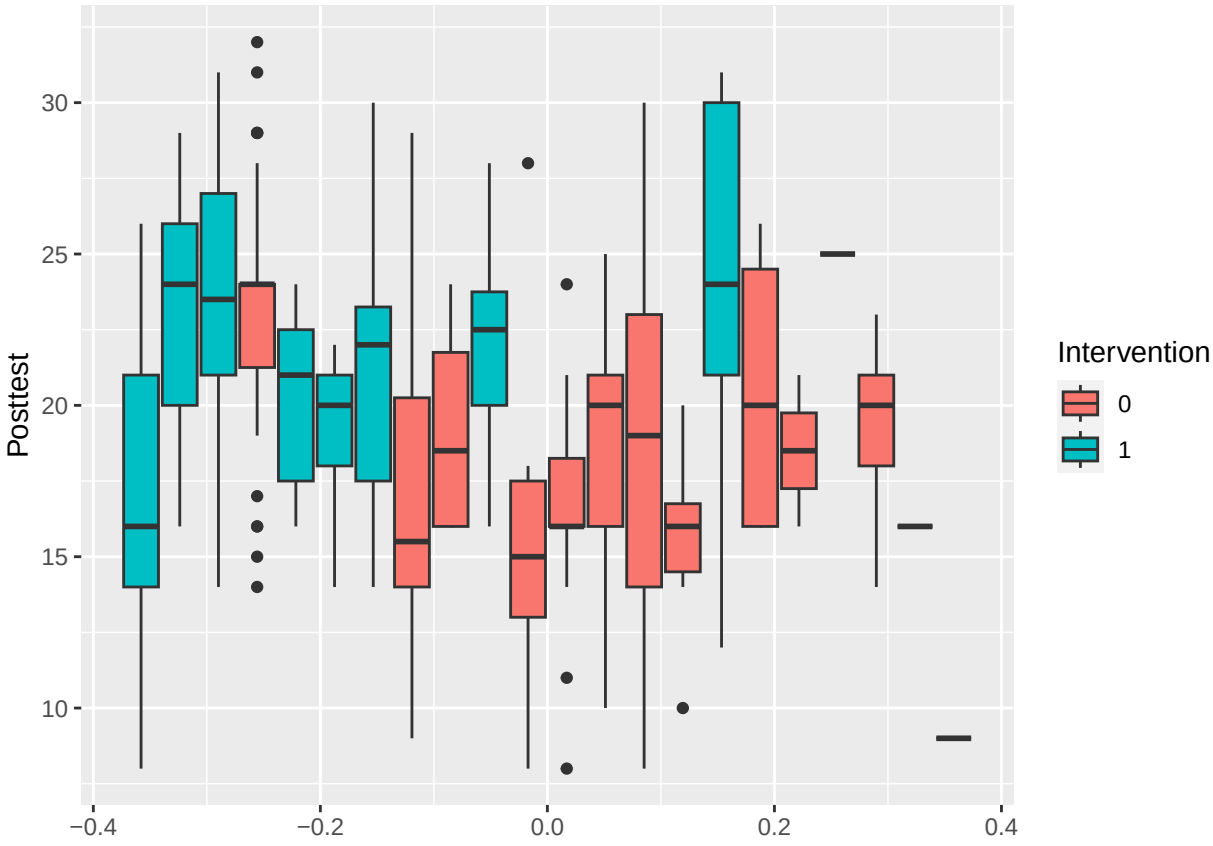


Figure 9.2: Box plots of the outcome Posttest for each school, coloured by Intervention.

Table 9.1: EEF data with difference calculated (first 10 rows).

School	MeanDiff	ng	Group
1	13.76923	13	1
2	19.09091	33	1
3	20.43333	30	1
4	19.00000	30	0
5	16.66667	15	1
6	14.40000	5	1
7	16.75000	24	1
8	13.83333	12	0
9	15.00000	4	0
10	18.28571	14	1

9.4.1 At the cluster level

In cluster level analysis, the data is aggregated to the cluster level, so that for each cluster (school, in our case), there is effectively one data point. The advantage of this approach is that, because the design is conducted at the cluster level, the statistical methodology is relatively simple - for example, a t-test. However, if cluster sizes vary a lot, a cluster-level analysis is often not appropriate because they usually rely on the assumption that the variance of the outcome in each cluster is approximately the same. As we have seen, the larger a group, the smaller the variance of its sample mean. There are methods designed to account for this, such as the weighted t-test, but these methods are generally inefficient and less robust. These methods are generally thought to be appropriate for fewer than 15-20 clusters per treatment arm.

One possibility for our trial would be to collect the mean and SD of scores within each school, and perform a t-test to find out if there is a significant difference between the intervention and control arms. If we wanted to find out whether this depended on, say, gender, we could split the data set and perform seperate t-tests for the different gender groups. This has the advantage that it is simple to implement, but the disadvantage that it is difficult to take into account covariates (apart from in the simple way discussed for eg. gender). With a small study, it is likely that there is some imbalance in the design in terms of covariates.

The required sample size for this option would be

$$g = \frac{2\sigma^2\left[1+(n_g-1)\,\rho_{icc}\right]}{n_g\tau_M^2}\left(z_\beta + z_{\frac{1}{2}\alpha}\right)^2,$$

where n_g is the average cluster size and g is the number of clusters per treatment arm. This is the value we worked out in Section 9.2.0.1

Example 9.4. We can perform this analysis on our schools data. The first step is to calculate the difference between `posttest` and `pretest` for each pupil.

```
crt_df$diff = crt_df$Posttest - crt_df$Prettest
```

We can then aggregate this to find the mean of `diff` for each school, shown in Table 9.1:

We can also visualise the mean differences by group

From Figure 9.3 it certainly looks likely that a significant difference will be found.

```
t.test(
  x=crt_summ$MeanDiff[crt_summ$Group==0],
  y=crt_summ$MeanDiff[crt_summ$Group==1],
```

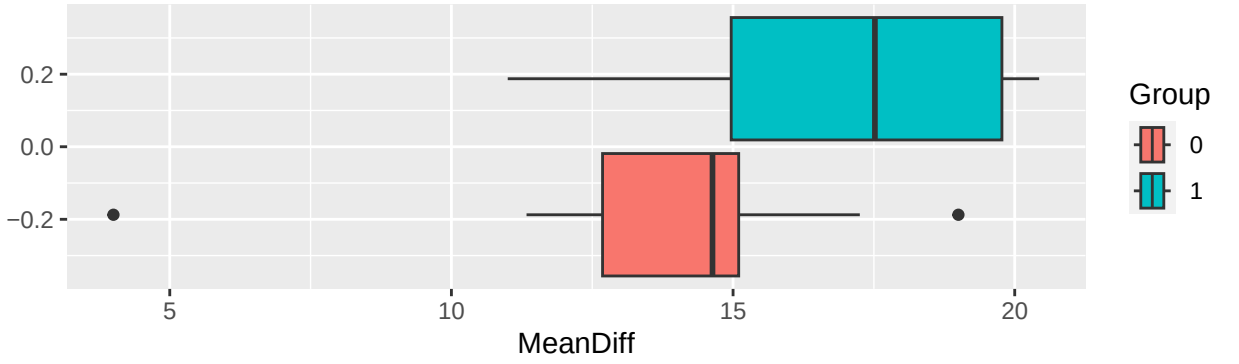


Figure 9.3: Boxplots of the mean differences for each trial arm

```
  alternative = "two.sided",
  paired = F,
  var.equal=F
)
```

```
##
##  Welch Two Sample t-test
##
## data:  crt_summ$MeanDiff[crt_summ$Group == 0] and crt_summ$MeanDiff[crt_summ$Group == 1]
## t = -2.2671, df = 19.983, p-value = 0.03463
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.4133879 -0.2667059
## sample estimates:
## mean of x mean of y
##  13.72454  17.06459
```

This is fairly easy to implement, but it seems rather unsatisfactory. It's also probably not entirely appropriate because the group sizes vary from one (!) to 33. What we need is a linear (in the continuous outcome case at least) model that takes into account the covariates, and makes the most of the available data.

9.4.2 At the individual level: mixed effects models

Perhaps the prevalent way of analysing the data from cluster randomized trials is the **mixed effects model** or **random effects model**, or **multilevel models**. In basic terms, mixed effects models are used whenever it cannot be assumed that the outputs are all independent of one another. In the cluster randomized trial setting, this is because outcomes for participants within the same cluster can be expected to be more highly correlated than outcomes for patients from different clusters, but we will see examples of other designs in the next lecture.

To understand mixed effects models, we need to think about the difference between **fixed effects** and **random effects**. These are two different types of factor variables.

Fixed effects

These are the sorts of factor variables we're used to dealing with in linear models: we don't assume any relationship between the levels, and we are generally interested in comparing the groups or categories repre-

sented by the fixed effect factors. We've seen these throughout the course, most notably with the treatment group variable G_i, but also with things like smoking history, sex, disease details. When conducting a clinical trial, we're likely to try to include participants with all levels of the fixed effects we're interested in, so that we can make inferences about the effect of the different levels of those effects. For example, we might want a good balance of male and female participants, so that we can understand the effect of the treatment on both groups.

Random effects

Random effects are probably just as common in real life, but we haven't seen them yet. These are factor variables that we think of as being drawn from some underlying model or distribution. For example, this could be GP surgery or school class, or an individual. We expect each GP surgery / school class / individual to behave slightly differently (depending on the extent of the intracluster correlation) but to behave as though from some overall distribution. Unlike fixed effects, random effects are generally things we aren't specifically interested in understanding the effect of, but we want to account for the variation they bring. We're also unable to include all levels of the random effect in our study - for example, a study looking at the effect of an intervention in schools will involve perhaps 50 schools, but we want to apply to results to all schools in the UK (say). We therefore assume that the schools are drawn from some normal distribution (in terms of the outcome we're interested in), and that therefore all the schools we haven't included also belong to this distribution. In Example 9.1 we aren't trying to compare different cities, and we certainly don't have data for Kroger stores in all cities, but we're assuming that the mean cheese price μ_{city_i} in the different cities is drawn from $N\left(\mu,\, \sigma_B^2\right)$, and that within each city the cheese price is drawn from $N\left(\mu_{\text{city}_i},\, \sigma_W^2\right)$.

Including random effects allows us to account for the fact that some schools might be in general a bit better / worse performing, or some individuals might be a bit more / less healthy, because of natural variation. We will see more examples of the use of random effects in the next lecture.

The mixed effects model

Mixed effects models allow us to combine fixed effects and random effects. We'll look at them next lecture too, because they are useful for many more situations than cluster RCTs, but this is as good a place as any to start!

The mixed effects model takes the form

$$x_{ijk} = \alpha + \beta G_i + \sum_l \gamma_l z_{ijkl} + u_{ij} + v_{ijk} \tag{9.5}$$

where

- x_{ijk} is the outcome for the k-th individual in the j-th cluster in the i-th treatment arm (usually $i = 0$ is the control arm and $i = 1$ is the intervention arm)
- α is the intercept of the model
- β is the intervention effect, and G_i the group indicator variable (0 for group C, 1 for group T). Our null hypothesis is that β is also zero)
- The z_{ijkl} are L different individual level covariates that we wish to take into account, and the γ_l are the estimated coefficients.
- u_{ij} is a random effect relating to the j-th cluster in the i-th treatment arm. This is the term that accounts for the between-cluster variation. We assume u_{ij} is normally distributed with mean 0 and variance σ_B^2 (the between-cluster variance).
- v_{ijk} is a random effect relating to the k-th individual in the cluster (ie. an individual level random error term), assumed normally distributed with mean 0 and variance σ_W^2 (the within-cluster variance).

The part of the model that makes this particularly suitable to a cluster randomized trial is u_{ij}. Notice that this has no k index, and is therefore the same for all participants within a particular cluster.

With a random effects model we can take into account the effects of individual-level covariates and also the clustered design of the data. Approximately, our sample size requirements are

$$k = \frac{2\sigma^2 \left[1 + (m-1)\,\rho_{icc}\right]\left(1-\rho^2\right)}{m\tau_M^2}\left(z_\beta + z_{\frac{1}{2}\alpha}\right)^2.$$

Broadly this follows on from the logic we used to show the reduction in variance from the ANCOVA model in Section 4.3.1.1, and you'll notice that the factor of $1-\rho^2$ is the same. The details for cluster randomized trials are given in Teerenstra et al. (2012).

This is the 'Random effects model' line in the shiny dashboard.)

The random effects model is more suitable when there are more than around 15-20 clusters in each arm.

Example 9.5. We'll now fit a random effects model to the `crtData` dataset from `eefAnalytics`. A good starting point is to plot the data with this in mind, to see what we might expect.

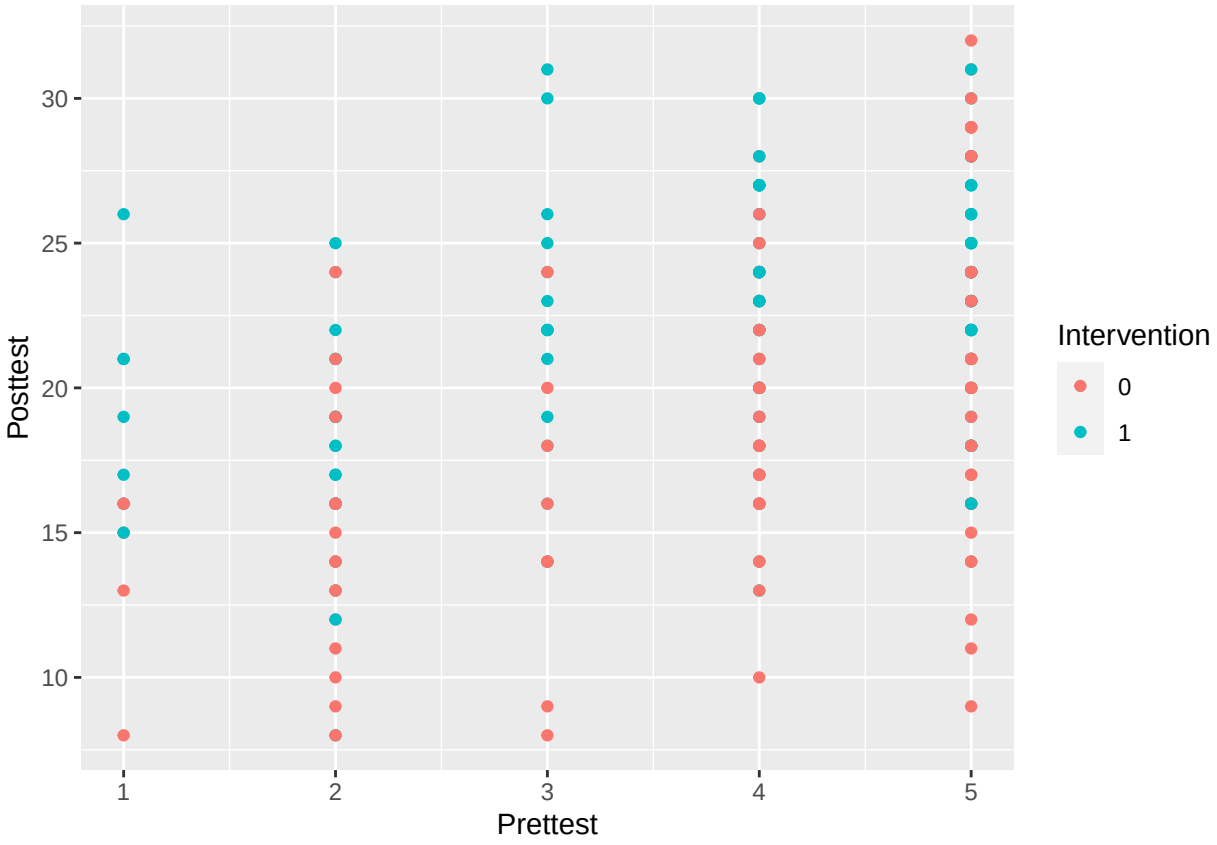


Figure 9.4: Posttest against Prettest, coloured by Intervention

From Figure 9.4 it appears there might be a positive relationship between `Prettest` and `Posttest`, and also that the `Posttest` scores might be higher in the intervention group.

We'll do this using the R package `lme4`. The function to specify a linear mixed effects model is called `lmer`, and works very similarly to `lm`. The term `(1|School)` tells R that the variable `School` should be treated as a random effect, not a fixed effect.

```
library(lme4)
library(sjPlot)
lmer_eef1 = lmer(Posttest ~ Prettest + Intervention + (1|School), data=crt_df )
```

The package `sjPlot` contains functions to work with mixed effect model objects, for example `plot_model`

```
sjPlot::plot_model(lmer_eef1)
```

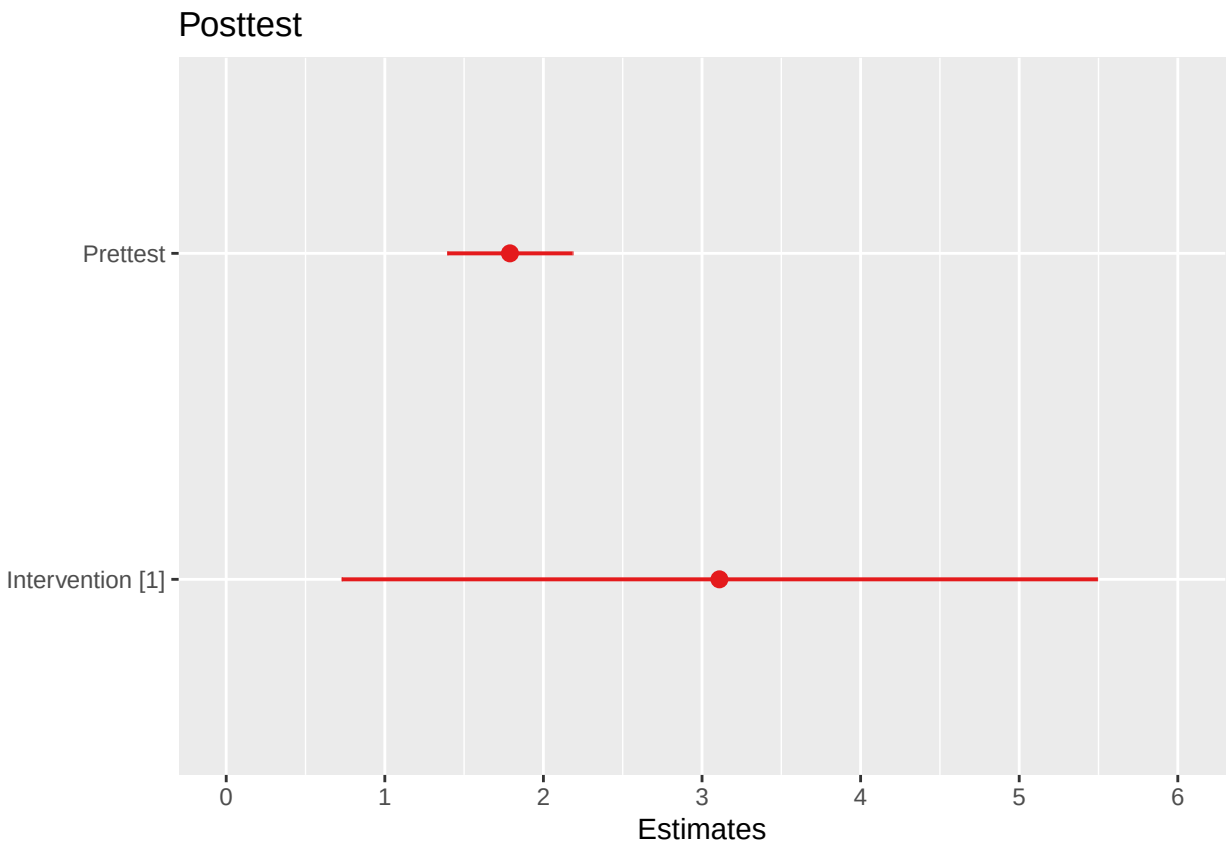


Figure 9.5: CIs for the model coefficients

and `tab_model`

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Posttest ~ Prettest + Intervention + (1 | School)
##    Data: crt_df
##
## REML criterion at convergence: 1493.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.46249 -0.64166  0.01626  0.59655  2.60277
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
```

```
##  School   (Intercept)  5.674   2.382
##  Residual              14.779   3.844
## Number of obs: 265, groups:  School, 22
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)    11.2286     1.1250   9.981
## Prettest        1.7889     0.2004   8.928
## Intervention1   3.1097     1.2094   2.571
##
## Correlation of Fixed Effects:
##             (Intr) Prttst
## Prettest    -0.681
## Interventn1 -0.493 -0.010
```

Perhaps unsurprisingly, the coefficient of the baseline test score is very significant, and the intervention also has a significant effect. This function also estimates the intracluster correlation.

There are various different ways we can include random effects in the model, as shown in Figure 9.6. In our EEF data example we have used a fixed slope and fixed intercept.

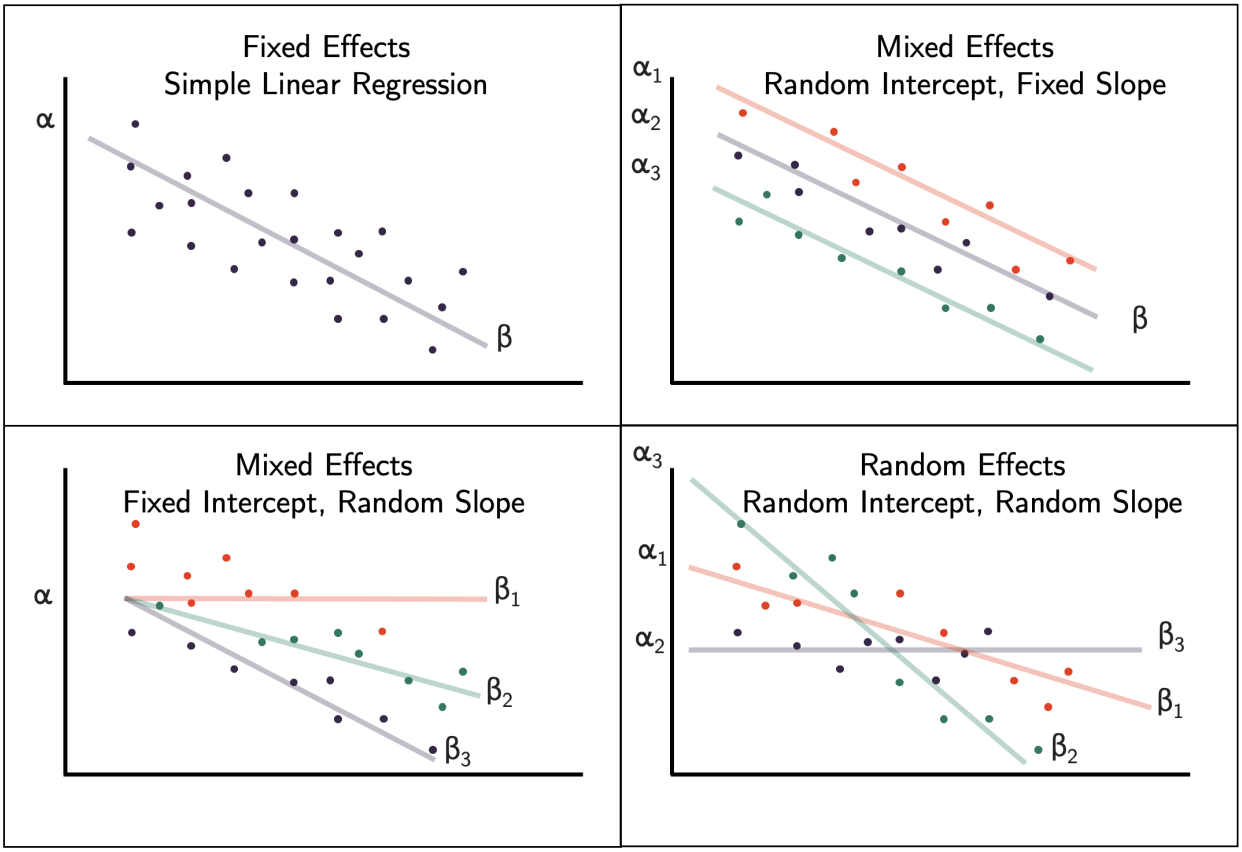


Figure 9.6: Different ways of including random effects in a mixed effects model.

The mixed effects model can be extended to a **generalized mixed effects model**, which is akin to a generalized linear model. For example, with a binary outcome X_{ijk} we can adopt the model

$$\text{logit}\left(\pi_{ijk}\right) = \alpha + \beta G_i + \sum_l \gamma_l z_{ijkl} + u_{ij} + v_{ijk}.$$

We will look in the next lecture at some more trial designs for which mixed effects models are useful.

References

This sections lists the references used in the course - it will be updated as the notes are updated. Some of the more accessible (dare I say 'interesting') resources are linked from the notes. If you want to read any of these articles, the easiest way is to copy the title into Google scholar.

Bibliography

Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.

Altman, D. G. (1998). Confidence intervals for the number needed to treat. *Bmj*, 317(7168):1309–1312.

Altman, D. G. and Bland, J. M. (1999). Treatment allocation in controlled trials: why randomise? *Bmj*, 318(7192):1209–1209.

Borm, G. F., Fransen, J., and Lemmens, W. A. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of clinical epidemiology*, 60(12):1234–1238.

Collett, D. (2003a). *Modelling Binary Data*. Texts in Statistical Science. Chapman & Hall, 2nd edition.

Collett, D. (2003b). *Modelling Survival Data in Medical Research*. Texts in Statistical Science. Chapman & Hall, 2nd edition.

Cottingham, M. D. and Fisher, J. A. (2022). Gendered logics of biomedical research: Women in us phase i clinical trials. *Social Problems*, 69(2):492–509.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dickinson, L. M., Beaty, B., Fox, C., Pace, W., Dickinson, W. P., Emsermann, C., and Kempe, A. (2015). Pragmatic cluster randomized trials using covariate constrained randomization: a method for practice-based research networks (pbrns). *The Journal of the American Board of Family Medicine*, 28(5):663–672.

Edmonson, J. H., Fleming, T. R., Decker, D. G., Malkasian, G. D., Jorgensen, E. O., Jefferies, J. A., Webb, M. J., and Kvols, L. K. (1979). Prognosis in advanced ovarian carcinoma versus minimal residual. *Cancer treatment reports*, 63(2):241–247.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417.

Elmunzer, B. J., Scheiman, J. M., Lehman, G. A., Chak, A., Mosler, P., Higgins, P. D., Hayward, R. A., Romagnuolo, J., Elta, G. H., Sherman, S., et al. (2012). A randomized trial of rectal indomethacin to prevent post-ercp pancreatitis. *New England Journal of Medicine*, 366(15):1414–1422.

Fentiman, I. S., Rubens, R. D., and Hayward, J. L. (1983). Control of pleural effusions in patients with breast cancer a randomized trial. *Cancer*, 52(4):737–739.

Freedman, L. and White, S. J. (1976). On the use of pocock and simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics*, pages 691–694.

Hayes, R. J. and Moulton, L. H. (2017). *Cluster randomised trials*. CRC press.

Hjalmas, H. and Hellstrom, K. (1998). Long-term treatment with desmopressin in children with primary monosymptomatic nocturnal enuresis: an open multicentre study. *British Journal of Urology*.

Hommel, E., Parving, H.-H., Mathiesen, E., Edsberg, B., Nielsen, M. D., and Giese, J. (1986). Effect of captopril on kidney function in insulin-dependent diabetic patients with nephropathy. *Br Med J (Clin Res Ed)*, 293(6545):467–470.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., and Newman, T. B. (2013). *Designing clinical research, fourth edition*. Lippincott Williams & Wilkins.

Kallis, P., Tooze, J., Talbot, S., Cowans, D., Bevan, D., and Treasure, T. (1994). Pre-operative aspirin decreases platelet aggregation and increases post-operative blood loss–a prospective, randomised, placebo controlled, double-blind clinical trial in 100 patients with chronic stable angina. *European journal of cardio-thoracic surgery: official journal of the European Association for Cardio-thoracic Surgery*, 8(8):404–409.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kar, S., Krishnan, A., K, P., and Mohankar, A. (2012). A review of antihistamines used during pregnancy. *Journal of Pharmacology and Pharmacotherapeutics*, 3(2):105–108.

Kassambara, A. (2019). *datarium: Data Bank for Statistical Analysis and Visualization*. R package version 0.1.0.

Kendall, J. (2003). Designing a research project: randomised controlled trials and their principles. *Emergency medicine journal: EMJ*, 20(2):164.

Le-Rademacher, J. G., Peterson, R. A., Therneau, T. M., Sanford, B. L., Stone, R. M., and Mandrekar, S. J. (2018). Application of multi-state models in cancer clinical trials. *Clinical Trials*, 15(5):489–498.

Marshall, G. (1948). Streptomycin treatment of pulmonary tuberculosis a medical research council investigation. *British Medical Journal*.

Matthews, J. N. (2006). *Introduction to randomized controlled clinical trials*. CRC Press.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in medicine*, 17(8):873–890.

of Health, N. I. (2023). History of women's participation in clinical research.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115.

Ruetzler, K., Fleck, M., Nabecker, S., Pinter, K., Landskron, G., Lassnigg, A., You, J., and Sessler, D. I. (2013). A randomized, double-blind comparison of licorice versus sugar-water gargle for prevention of postoperative sore throat and postextubation coughing. *Anesthesia & Analgesia*, 117(3):614–621.

Smith, A., Dowsett, J., Russell, R., Hatfield, A., and Cotton, P. (1994). Randomised trial of endoscopic steriting versus surgical bypass in malignant low bileduct obstruction. *The Lancet*, 344(8938):1655–1660.

Syriopoulou, E., Wästerlid, T., Lambert, P. C., and Andersson, T. M.-L. (2022). Standardised survival probabilities: a useful and informative tool for reporting regression models for survival data. *British journal of cancer*, 127(10):1808–1815.

Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, 15(5):443–453.

Teerenstra, S., Eldridge, S., Graff, M., de Hoop, E., and Borm, G. F. (2012). A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in medicine*, 31(20):2169–2178.

Therneau, T. M. (2024). *A Package for Survival Analysis in R*. R package version 3.5-8.

Treasure, T. and MacRae, K. D. (1998). Minimisation: the platinum standard for trials?: Randomisation doesn't guarantee similarity of groups; minimisation does.

Villar, J., Ferrando, C., Martínez, D., Ambrós, A., Muñoz, T., Soler, J. A., Aguilar, G., Alba, F., González-Higueras, E., Conesa, L. A., et al. (2020). Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. *The Lancet Respiratory Medicine*, 8(3):267–276.

Vitale, C., Fini, M., Spoletini, I., Lainscak, M., Seferovic, P., and Rosano, G. M. (2017). Under-representation of elderly and women in clinical trials. *International journal of cardiology*, 232:216–221.

Wei, L. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, 73(363):559–563.

Zhong, B. (2009). How to calculate sample size in randomized controlled trial? *Journal of thoracic disease*, 1(1):51.