

# Clinical Trials 4H

Rachel Oughton

2025-02-26

# Contents

|  |           |
|--|-----------|
| <b>Welcome to Clinical Trials 4H!</b>                                    | <b>4</b>  |
| Practical details . . . . .  | 4         |
| What to expect from this module . . . . .                                | 5         |
| <b>1 Introduction to Clinical Trials</b>                                 | <b>6</b>  |
| 1.1 Causal inference and clinical trials . . . . .                       | 7         |
| 1.2 The structure of a clinical trial . . . . .                          | 8         |
| 1.3 The primary outcome . . . . .  | 9         |
| 1.4 Ethical issues . . . . .   | 9         |
| 1.5 Phases of clinical trials . . . . .                                  | 10        |
| <b>I Part I: Continuous outcome variables</b>                            | <b>12</b> |
| <b>2 Sample size for a normally distributed primary outcome variable</b> | <b>13</b> |
| 2.1 The treatment effect . . . . .                                       | 13        |
| 2.2 Reminder: hypothesis tests (with a focus on RCTs) . . . . .          | 14        |
| 2.3 Constructing a measure of effect size . . . . .                      | 17        |
| 2.4 Power: If $H_0$ is false . . . . .                                   | 20        |
| 2.5 A sample size formula . . . . .                                      | 24        |
| <b>3 Allocation</b>  | <b>26</b> |
| 3.1 Bias . . . . .   | 26        |
| 3.2 Allocation methods . . . . .   | 31        |
| 3.3 Incorporating baseline measurements . . . . .                        | 45        |
| 3.4 Problems around allocation . . . . .                                 | 48        |
| <b>4 Analyzing RCT data</b>  | <b>51</b> |
| 4.1 Confidence intervals and P-values . . . . .                          | 51        |
| 4.2 Using baseline values . . . . .                                      | 55        |
| 4.3 Analysis of covariance (ANCOVA) . . . . .                            | 58        |
| 4.4 Some follow-up questions.... . . . .                                 | 67        |

|   |            |
|---|------------|
| <i>CONTENTS</i>   | 3          |
| <b>II Part II: Binary outcome variable</b>                              | <b>74</b>  |
| <b>5 Sample size for a binary variable</b>                              | <b>75</b>  |
| 5.1 The Delta Method . . . . .  | 76         |
| 5.2 A sample size formula . . . . .                                     | 77         |
| <b>6 Analysis for binary outcomes</b>                                   | <b>79</b>  |
| 6.1 Simple hypothesis tests . . . . .                                   | 79         |
| 6.2 Measures of difference for binary data . . . . .                    | 83         |
| 6.3 Accounting for baseline observations: logistic regression . . . . . | 94         |
| 6.4 Diagnostics for logistic regression . . . . .                       | 101        |
| <b>References</b>   | <b>108</b> |

# Welcome to Clinical Trials 4H!

This page contains the notes for Clinical Trials IV 2024-25. 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!

In case you're reading the PDF, the notes are best viewed in HTML (the link is <https://racheloughton.github.io/ct4/>). The PDF will be up-to-date, but the formatting is designed for HTML.

## Practical details

### Lectures

Our lectures are 9am on Wednesdays in W007 (this is the Appleby Building, between Pysics and the main library), and 3pm on Fridays in PCL 048 (this is in the ground floor of the Palatine Centre).

### Computer classes

We have four computer practicals for this module. They are 1-2pm on the Mondays of weeks 13, 15, 17 and 19. These classes are in MCS 3098.

### Office Hour

The office hour will be every Wednesday, 10-11am, in the Calman Learning Centre cafe.

### Assessment

This module is assessed through two equally weighted pieces of coursework. The first will be assigned on Wednesday 7th February (due 3rd March), the second on Wednesday 19th March (due 5th May).

There will also be two formative assignments during the course. More details on these to follow.

### Books and resources

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.

I will share links along the way for relevant resources like podcast episodes and articles, which I hope will help you to get a feel for the topic more generally.

## What to expect from this module

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

- Its main focus is on application
- 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 course is in its infancy, 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.

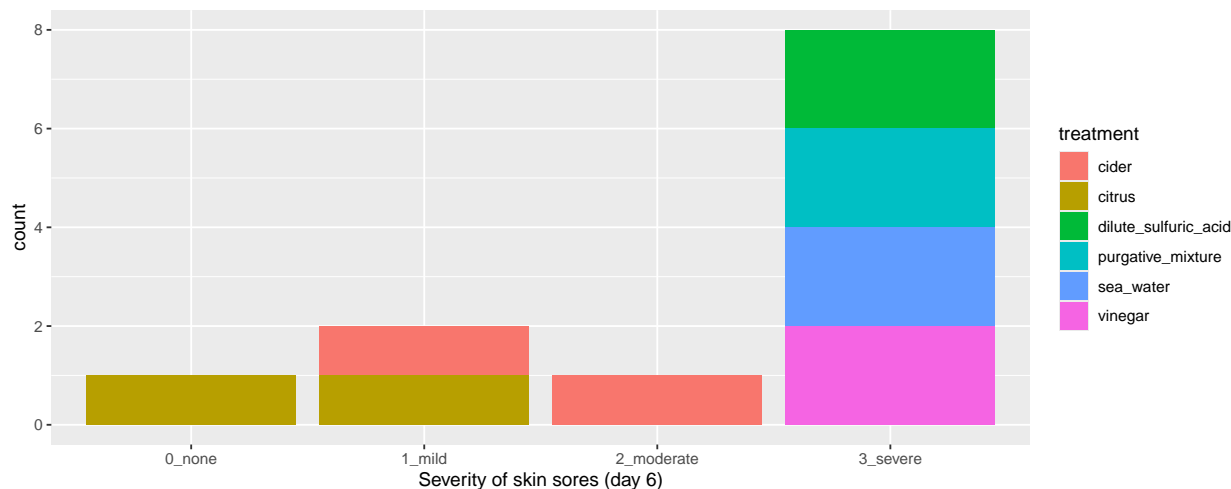
Put (very!) simply, the goal of a clinical trial is to determine what works to make people better. Although clinical trials as we know them now have only been around since the Second World War, similar sorts of experiments can be seen from much longer ago. If you’re interested in learning about the evolution 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 be 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 inoculation 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 repercussions 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.

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.

It is therefore 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  $\mu$  for the control group and mean  $\mu + \tau$  for the treatment group. The variable  $\tau$  is the **treatment effect**. The goal of our RCT is to learn about  $\tau$ . The larger  $\tau$  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 as 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 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  $\alpha$  (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 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  $\tau$  (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  $1 - F_t(2, df = 31) = 0.0272$  (where  $F_t(\cdot)$  is 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

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

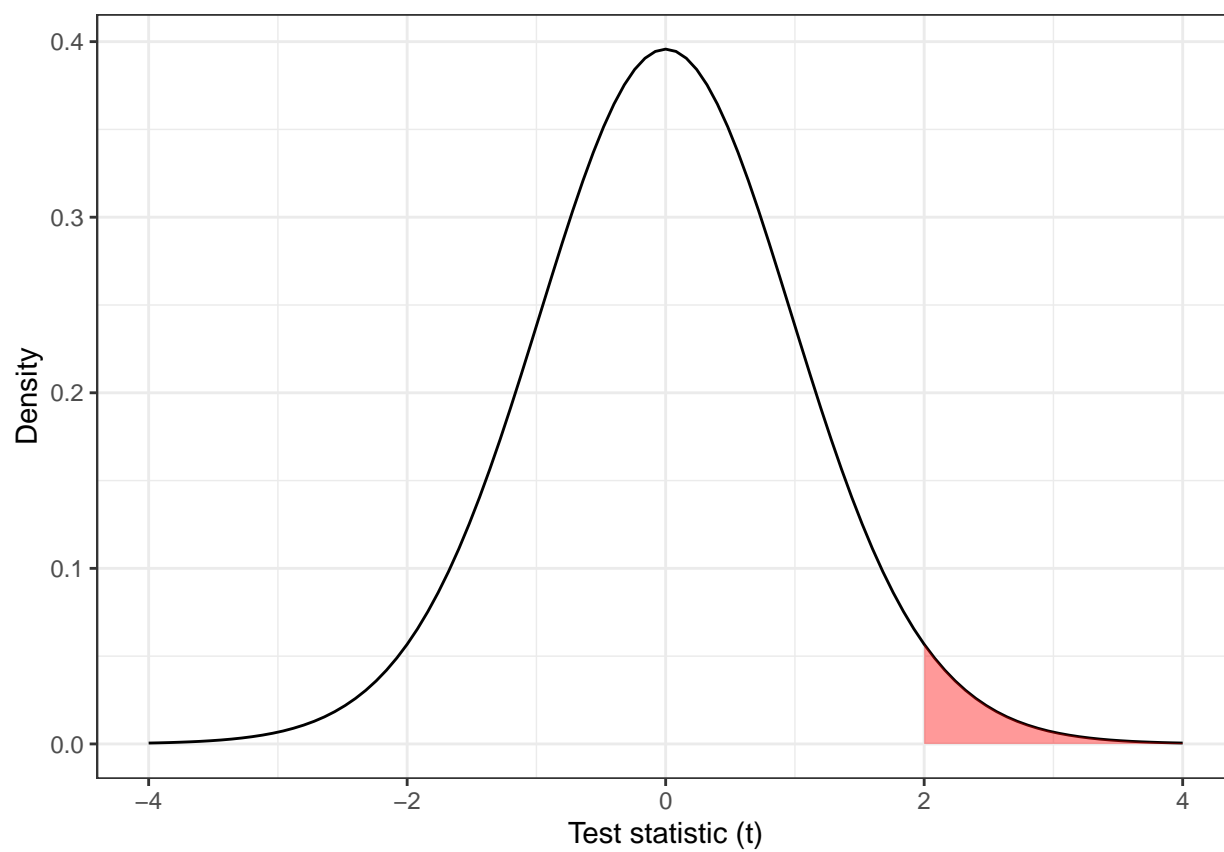


Figure 2.1: The distribution  $t_{31}$ , with the area corresponding to  $t > 2$  shaded.

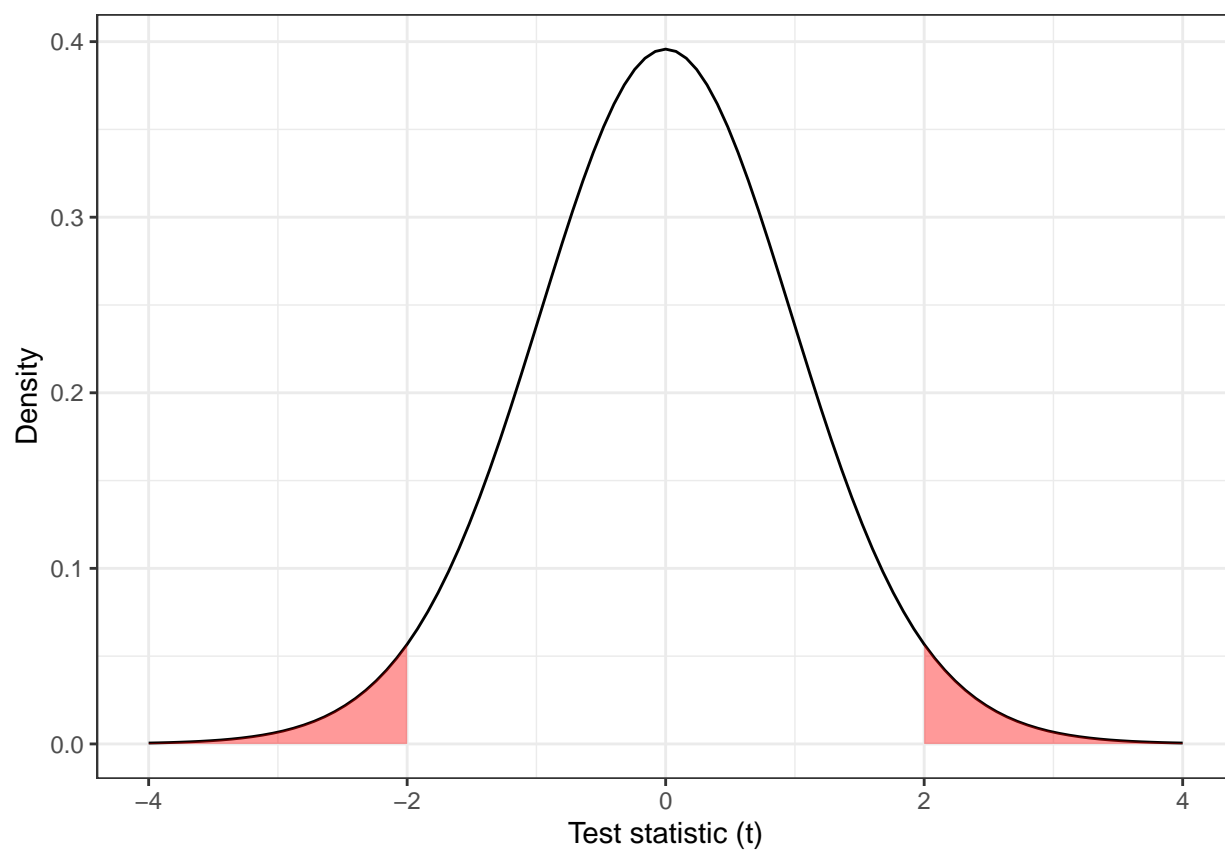


Figure 2.2: The distribution  $t_{31}$ , with the area corresponding to  $|t| > 2$  shaded.



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  $H_0$ .

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.

### 2.2.2 Insignificant results

If our P-value is relatively large, say 0.3 or 0.5 (or really, greater than  $\alpha$ ), 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  $\tau$  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  $\tau$  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 large enough 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.

## 2.3 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*, or alternatively the *intervention group*) 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  $\mu$  in group  $C$  (the control group) and mean  $\mu + \tau$  in group  $T$  (the intervention group), and common standard deviation  $\sigma$ . So

$$\begin{aligned} X &\sim N(\mu, \sigma^2) \text{ in group } C \\ X &\sim N(\mu + \tau, \sigma^2) \text{ in group } T. \end{aligned}$$

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

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

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  $\tau$ .

**Theorem 2.1.** *Under  $H_0$ , and the model we've specified,  $D$  has a  $t$ -distribution with  $n + m - 2$  degrees of freedom.*

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

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

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  $x_1, \dots, x_n, \sim N(\mu, \sigma^2)$  for some arbitrary  $\mu$  and  $\sigma^2$ ,

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

and so we have

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

The definition of a  $t$ -distribution is that if  $Z \sim N(0, 1)$  and  $Y \sim \chi_n^2$  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  $N(0, 1)$  variable for  $Z$  and our  $\chi_{n+m-2}^2$  variable for  $Y$ , we have

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

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

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

where  $t_{n+m-2}(\cdot)$  is the function such that  $P(T > t_{df}(\xi)) = \xi$  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  $100(1 - \alpha)\%$  level if  $|D| > z_{\alpha/2}$ , where  $z$  are standard normal quantile values. For example, for  $\alpha = 0.05$  we have  $z_{\alpha/2} = 1.960$ , since the probability of a standard normal variable exceeding this value is 0.025.

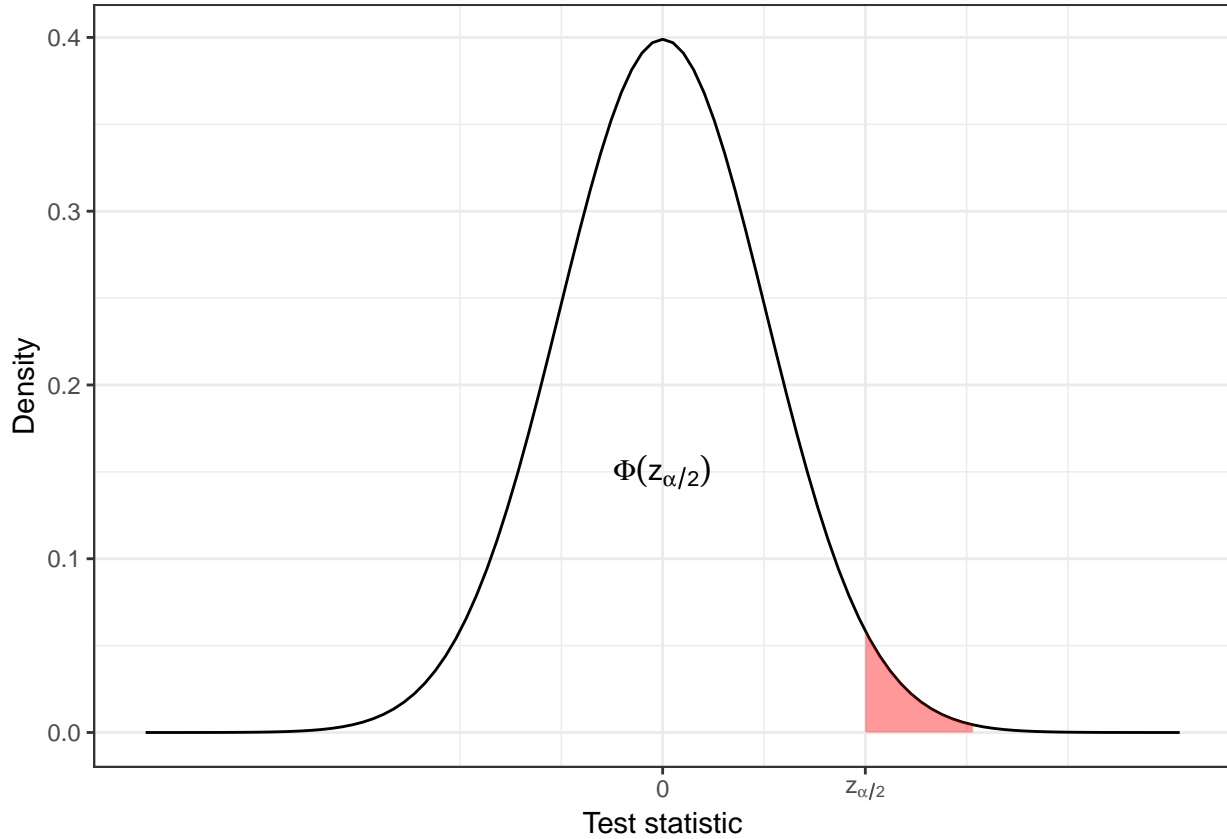
So, if we have run a trial, and have obtained  $n$  values of  $X$  from group  $C$  and  $m$  values of  $X$  from group  $T$ , we can compute  $D$ . If  $D$  lies outside the interval  $[-z_{\alpha/2}, z_{\alpha/2}]$  then we reject  $H_0$ .

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  $\Phi(z_{\alpha/2}) = \text{pnorm}(z_{\alpha/2})$  and  $z_{\alpha/2} = \text{qnorm}(\Phi(z_{\alpha/2}))$ . `qnorm` is the quantile and `pnorm` is the cumulative distribution function. So, for example

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

We have constructed our whole argument under the assumption that  $H_0$  is true, and that the probability of such a value is therefore  $\alpha$ . 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  $\tau$  and is

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

We therefore also have

$$\beta = \Pr(\text{Fail to reject } H_0 \mid \tau \neq 0).$$

We call  $\beta$  the **type II error rate**.

If you find the notation confusing (as I do!) then it might be helpful to remember that both  $\alpha$  and  $\beta$  are **error rates** - probabilities of coming to the wrong conclusion. It is common to talk in terms of  $\alpha$ , 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  $\beta$  (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 2.3 shows the distribution of  $D$  under  $H_0$  and  $H_1$  for some arbitrary (non-zero) effect size  $\tau$ . 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  $H_0$ . This could easily generalise to another value of  $\alpha$ .

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  $\beta$ . One minus this area (ie. the area under  $H_1$  that leads us to reject  $H_0$ ) is the power,  $1 - \beta$ .

We can see that if the distributions have better separation, as in Figure 2.4, the power becomes greater. This can be as a result of a larger  $\tau$ , a smaller  $\sigma$  or a smaller  $\lambda$  (therefore larger  $m$  and/or  $n$ ).

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

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

The first term in Equation (2.1) is the area in the direction of  $\tau$ . In Figures 2.3 and 2.4 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  $\tau$ , ie. a value of  $D$  such that

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

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

Figure 2.5 shows the power function  $\Psi(\tau)$  for  $\tau$  in units of  $\sigma$  (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  $\lambda$ ) 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  $\tau$  moves away from zero.

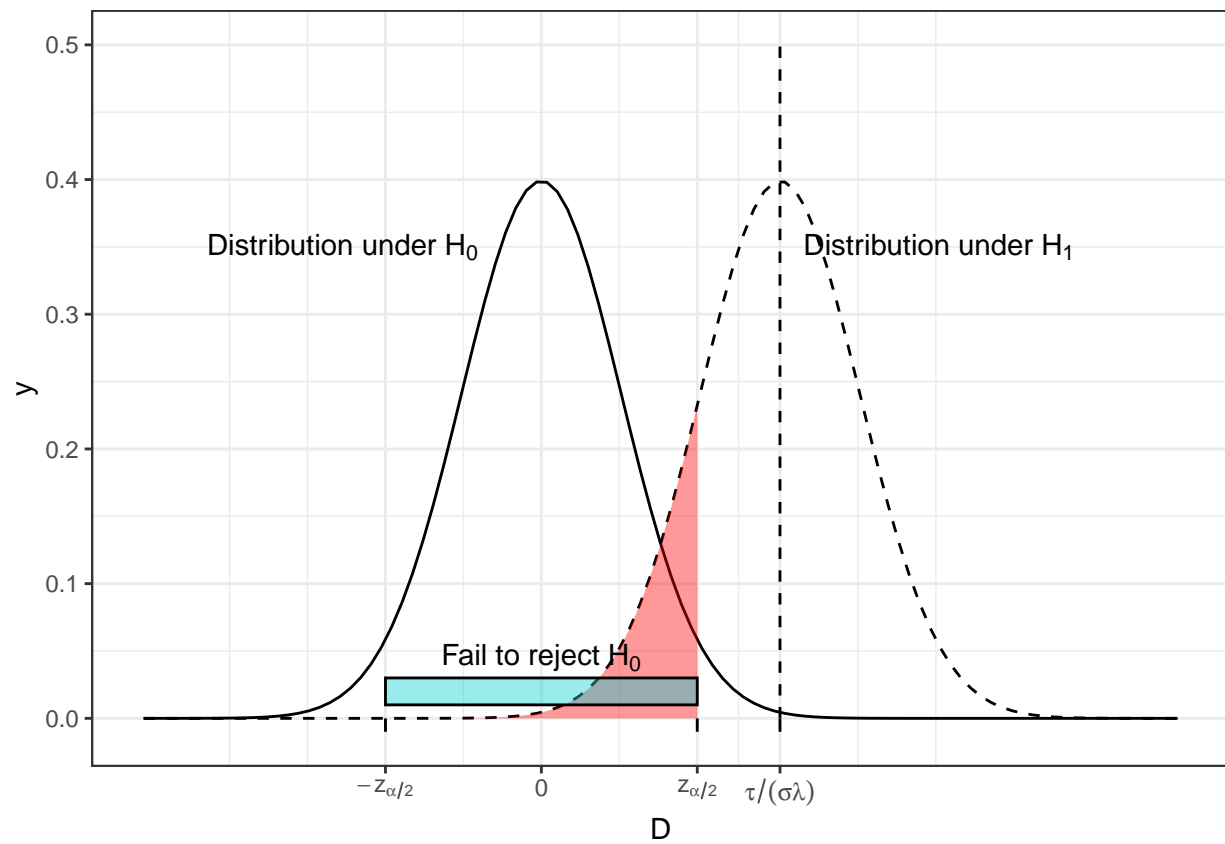


Figure 2.3: The distribution of  $D$  under both  $H_0$  and  $H_1$  for some arbitrary values of effect size, population variance,  $n$  and  $m$ , with the region in which we fail to reject  $H_0$  shown by the turquoise bar and the red shading.

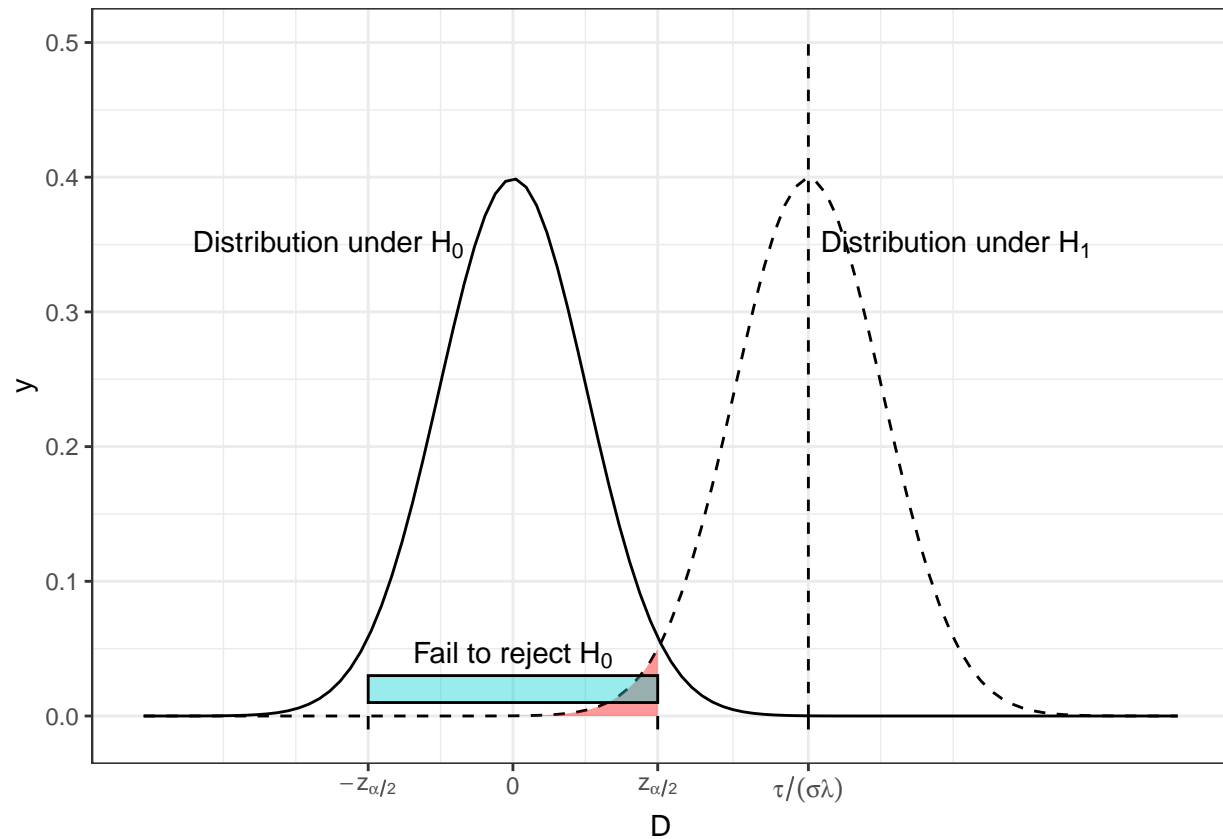


Figure 2.4: The distribution of  $D$  under both  $H_0$  and  $H_1$  for some arbitrary values of effect size, population variance,  $n$  and  $m$ , with the region in which we fail to reject  $H_0$  shown by the turquoise bar and the red shading.

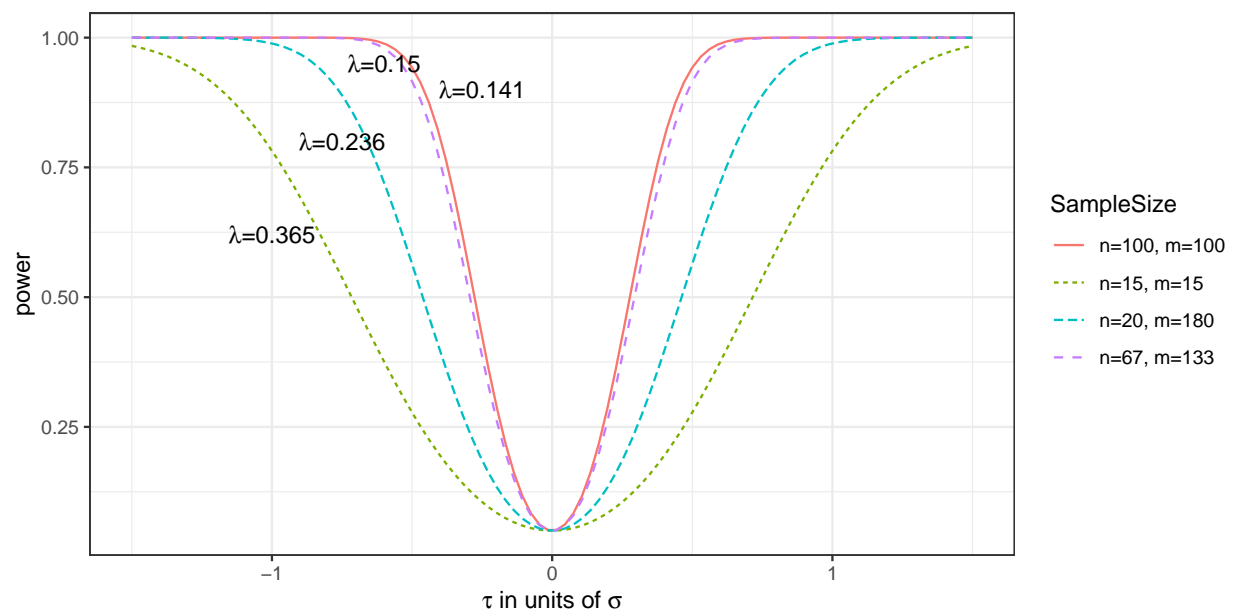


Figure 2.5: Power curves for various values of  $n$  and  $m$ , with effect size in units of standard deviation, given a type I error rate of 0.05.

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  $\alpha$ .

It is common to think of the effect size in units of  $\sigma$ , 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  $\sigma$ , we will write  $\tau_\sigma$ , although in practice it is more usual to give both the same notation. In a medical setting we will often have an estimate for  $\sigma$  (for example from previous studies).

## 2.5 A sample size formula

Equation (2.1) allows us to find any one of  $\tau_\sigma$ ,  $\alpha$ ,  $\beta$  and  $\lambda(n, m)$  given values for the others.

Values for  $\alpha$  and  $\beta$  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 values,  $\tau_\sigma$  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  $\tau_\sigma$  we can achieve?
- What is the smallest value of  $\tau_\sigma$  that would be clinically useful to detect, and what value of  $\lambda(n, m)$  do we need in order to achieve it?

An important quantity is therefore the **minimum detectable effect size**, which we will denote  $\tau_M$ .

**Definition 2.5.** The **minimum detectable effect size**  $\tau_M$  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  $\alpha$  (for some specified values of  $\alpha$ ,  $\beta$ ).

Note that we will not *definitely* detect an effect of size  $\tau_M$ , 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  $\tau_M$  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  $\tau$*  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  $\tau$ .

Therefore, Equation (2.1) becomes

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

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

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



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

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

**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

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

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, we need to determine how they will each be assigned to a trial arm. This 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  $\theta$ . The **bias** of  $T$  is

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

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

An example you will have seen before is the standard deviation. If we have some data  $x_1, \dots, x_n$  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  $\sigma^2$ . 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  $\sigma^2$ . Here, the sample are representative of the population, but the size of the sample leads to some bias.

Now, suppose our sample  $x_1, \dots, x_n$  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  $\sigma^2$ . Furthermore, we cannot use our sample  $x_1, \dots, x_n$  to produce an unbiased estimator of  $\sigma^2$ , or even of the mean  $\mu$ .

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  $\tau$  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  $\tau$  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.

In the above example, the trial cohort is very deliberately (and arguably sometimes defensibly) not representative of the target population. However, bias can creep into trials in a number of ways.

### 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.

### 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.

**Example 3.2.** *This example is framed in terms of selection bias, but applies equally to allocation bias*

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  $\lambda$  have severity 1 and proportion  $1 - \lambda$  have severity 2.

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

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

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  $\tau_N$ , and similarly for group  $S$ , so that we have

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

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  $N$ ,  $E(X | N)$ , would be

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

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  $A$ ):

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

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

$$E(X | A, N) - E(X | A, S).$$

We can split these according to disease severity, so that

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

and similarly for group  $S$ .

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

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

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

Following the same process for group  $S$ , we arrive at

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

which we will call  $b$ .

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

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

and

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

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

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

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

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

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  $N$ .

### Assessment bias

Measurements are made on participants throughout 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. In this situation, it is possible for the patient or clinician's judgement to be affected by their knowledge of the allocation, thus affecting the estimated treatment effect.

### 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 asked 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.

There is compelling evidence 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 practice. This is a huge issue in the pharmaceutical industry, and one from which we can't escape. If you'd like to learn more about it the book 'Bad Pharma' by Goldacre (2012) is an excellent source.

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 ( $ABABAB\dots$ ), 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 used effectively 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  $r = n + 1, \dots, 2n$ ,

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

The probability that  $N_C = N_T = n$  is

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

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

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

In this chapter, and in the computer practical on allocation, we will study the behaviour of various allocation methods by implementing them many times. This gives us an idea of how variable the results are, especially in terms of being likely to introduce bias or reduce power. In almost all real trials, patients are allocated and begin treatment as they arrive, and this may happen over the course of weeks or months. The allocation that is generated is the final allocation.

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

In this case  $P(N_{max} \geq 220) = 0.051$ , 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:

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



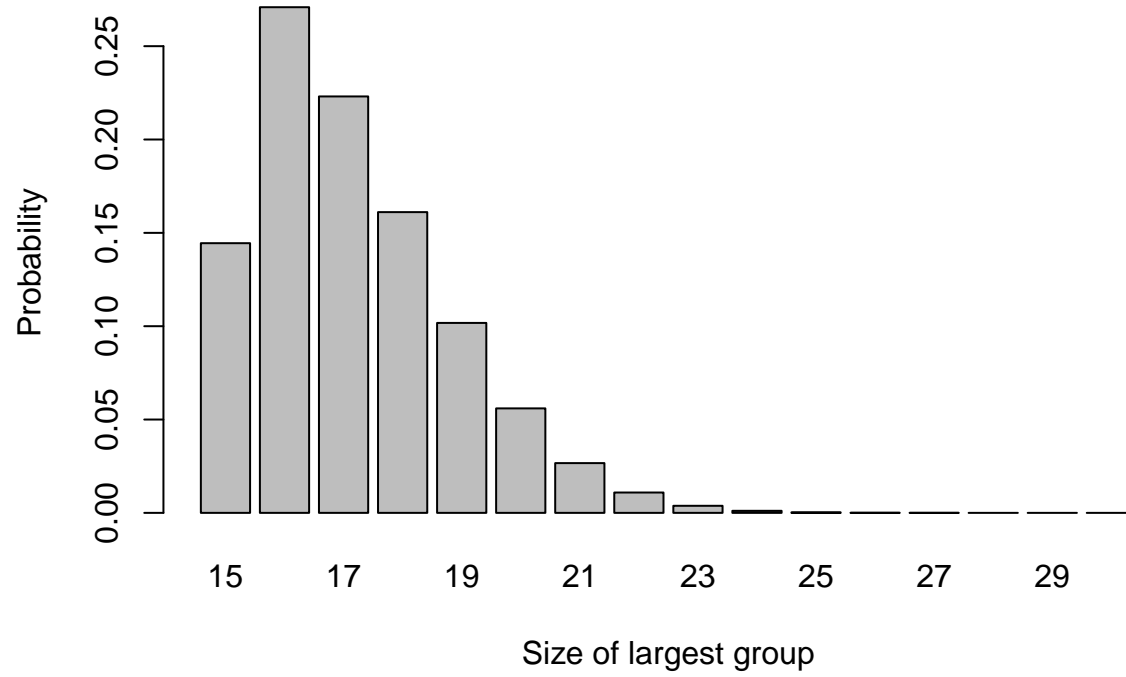


Figure 3.1: The probability distribution of largest group size for  $n=15$ .

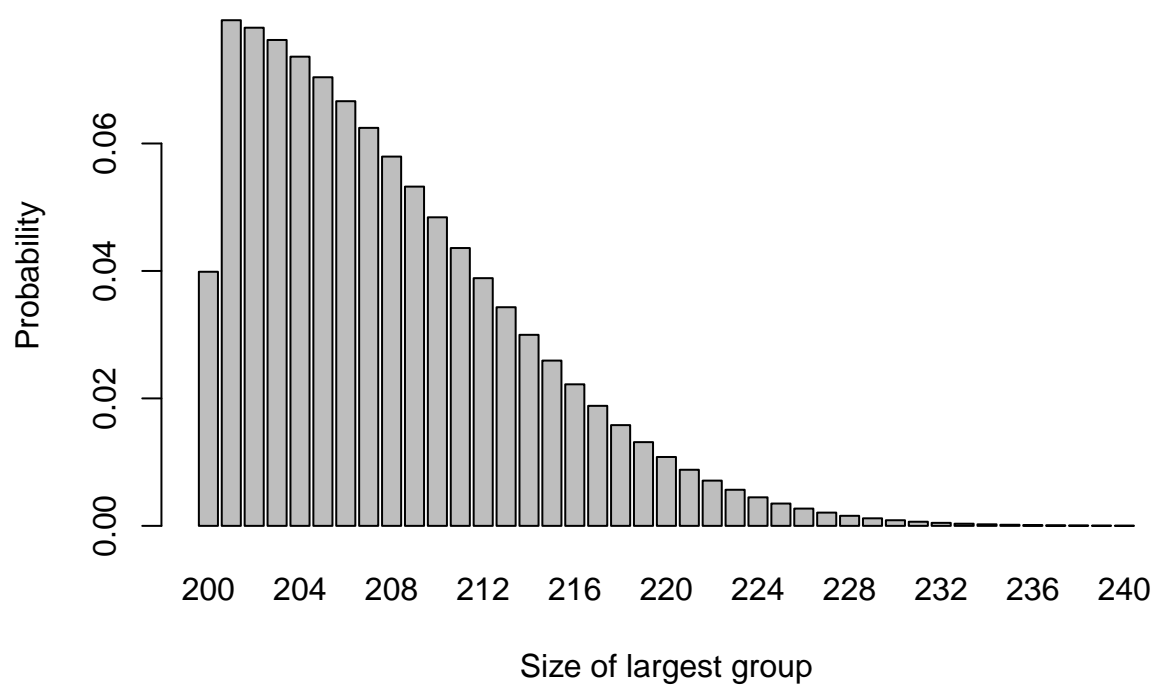


Figure 3.2: The probability distribution of largest group size for  $n=200$ .

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 imbalance 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 imbalance is to use *random permuted blocks* (RPBs). With RPBs we randomly generate the allocation in *blocks* of some predetermined size. 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.

**Example 3.4.** If we have two groups,  $A$  and  $B$ , then there are six *blocks* of length 4 containing two  $A$ s and two  $B$ s

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, .... Replacing each number by its block, we have  $ABAB AABB ABBA BBAA \dots$

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. For blocks of size four 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 that 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  $P \in (\frac{1}{2}, 1)$ .

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,

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

and

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

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

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

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.

Efron's suggestion for implementation was that each clinician would receive 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

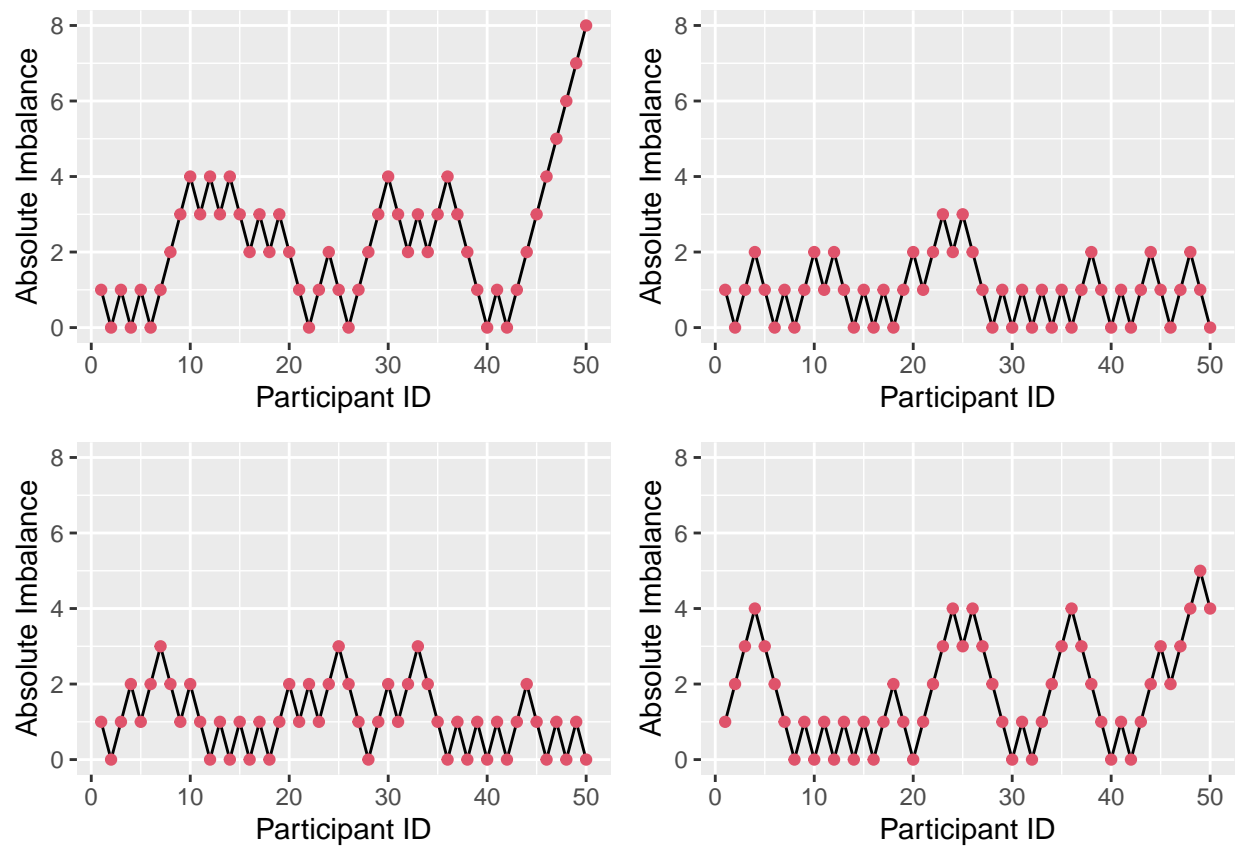


Figure 3.3: Absolute imbalance for a biased-coin scheme with  $P = 0.667$ .

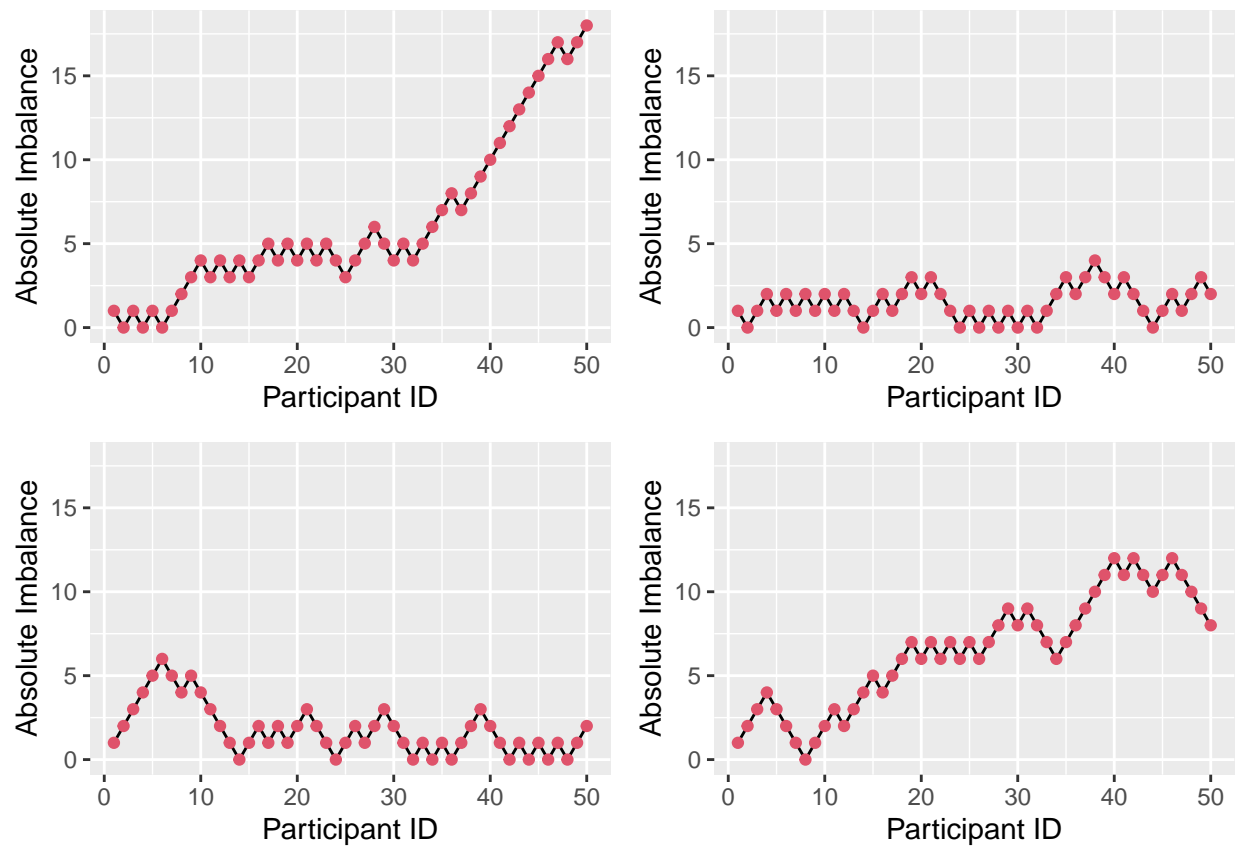


Figure 3.4: Absolute imbalance for a biased-coin scheme with  $P = 0.55$ .

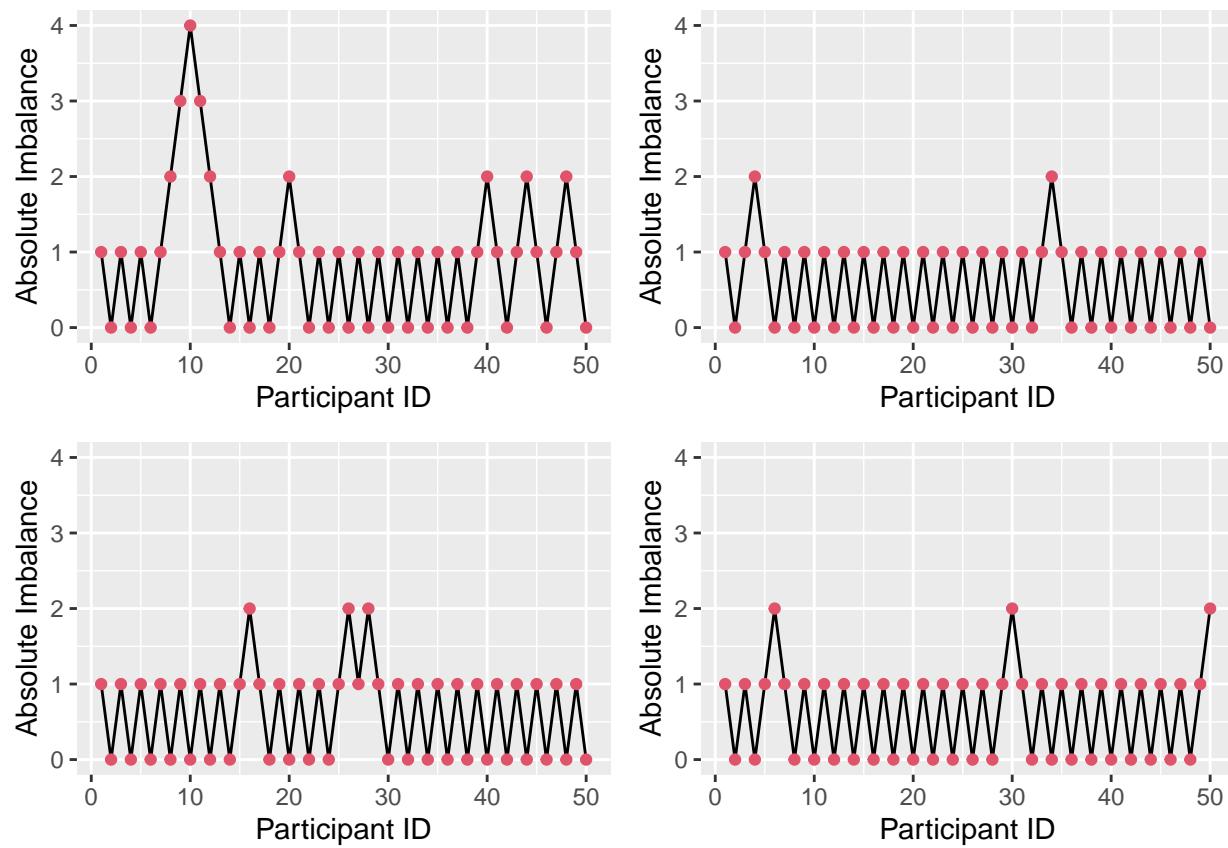


Figure 3.5: Absolute imbalance for a biased-coin scheme with  $P = 0.9$ .



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

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

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

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

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

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

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

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

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

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

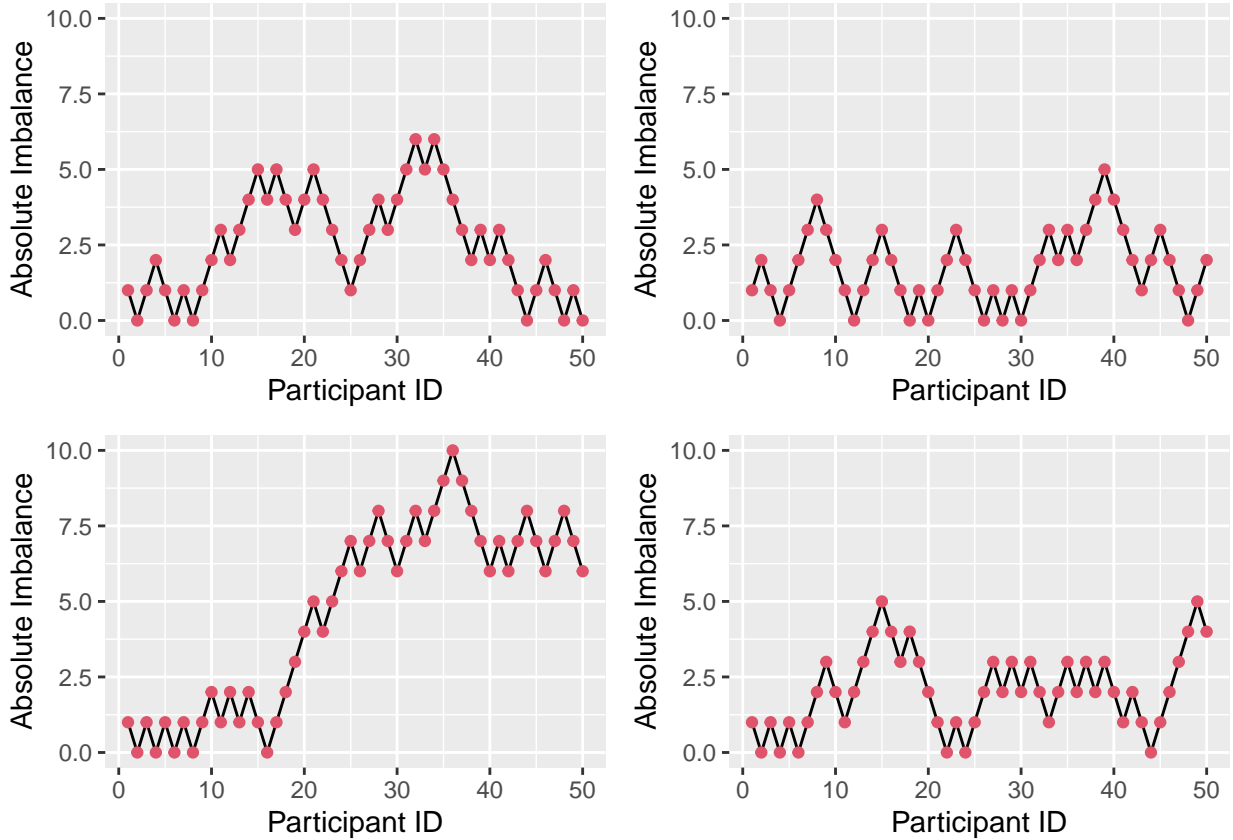


Figure 3.6: Four realisations of absolute imbalance for  $r=1$ ,  $s=1$ ,  $N=50$ .

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  $(\frac{1}{2}, \frac{1}{2})$  and so more imbalance occurs early on.

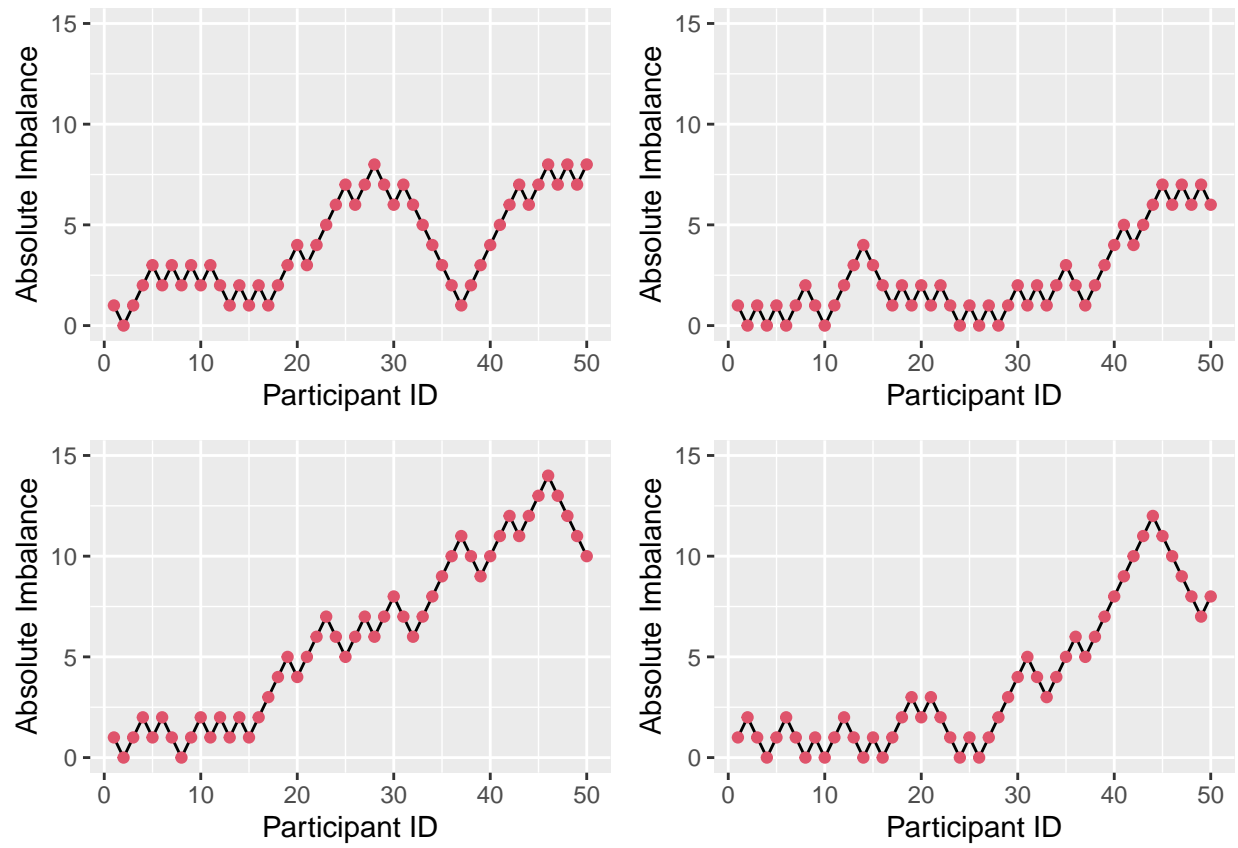


Figure 3.7: Four realisations of absolute imbalance for  $r=1$ ,  $s=8$ ,  $N=50$ .

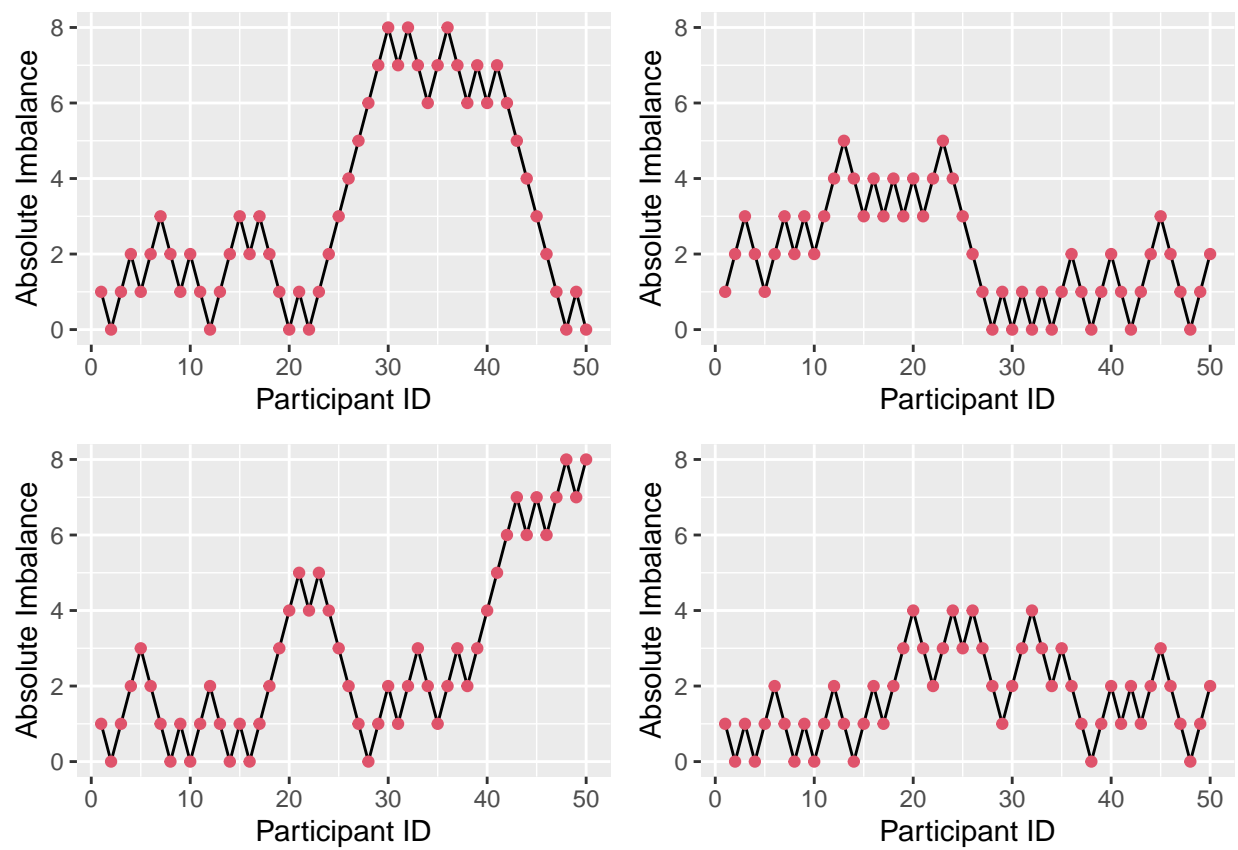


Figure 3.8: Four realisations of absolute imbalance for  $r=8$ ,  $s=1$ ,  $N=50$ .

### 3.3 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.

#### 3.3.1 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 3.5.** 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        |
|--------------------|---------------|---------------|
| <b>50-65</b>       | ABAB BBAA ... | ABBA BBAA ... |
| <b>66-80</b>       | BAAB AABB ... | BABA BAAB ... |
| <b>81 and over</b> | 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).

#### 3.3.2 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 3.6.** The study by Kallis et al. (1994) compares 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 ( $\leq 50$  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.

### 3.3.2.1 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  $I, J, K, \dots$  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,  $n_{ijkl}^A$  have been assigned to treatment arm  $A$ , and  $n_{ijkl}^B$  to arm  $B$ . So we have

$$n_{ijkl}^A + n_{ijkl}^B = n_{ijkl}.$$

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

$$|n_{ijkl}^A - n_{ijkl}^B| \leq \frac{1}{2}b,$$

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:

$$\begin{aligned} &|n_{i+++}^A - n_{i+++}^B| \text{ for each } i = 1, \dots, I \\ &|n_{+j++}^A - n_{+j++}^B| \text{ for each } j = 1, \dots, J \\ &\dots \end{aligned}$$

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

$$n_{++k+}^A = \sum_{i,j,l} n_{ijkl}^A$$

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

Therefore, instead of having  $IJKL$  constraints, as we would with using randomly permuted blocks (or some other randomization method) 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  $n_{ijkl}^A$  are allocated to treatment arm  $A$  and  $n_{ijkl}^B$  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

$$(n_{w+++}^A - n_{w+++}^B) + (n_{+x++}^A - n_{+x++}^B) + (n_{++y+}^A - n_{++y+}^B) + (n_{++++}^A - n_{++++}^B). \quad (3.1)$$

5. If the sum from step 4 is negative (that is, allocation to arm  $B$  has 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  $n_{i+++}^A$  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 3.7.** 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 ( $\leq 50$  or  $> 50$ ), stage of disease (I or II, III or IV), time in months between diagnosis of breast cancer and diagnosis of pleural effusions ( $\leq 30$  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 the table below.

| 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        |

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

$$\begin{aligned}
 (n_{1+++}^A - n_{1+++}^B) + (n_{+2++}^A - n_{+2++}^B) + (n_{++1+}^A - n_{++1+}^B) + (n_{+++1}^A - n_{+++1}^B) \\
 = (3 - 4) + (6 - 6) + (4 - 2) + (4 - 3) \\
 = -1 + 0 + 2 + 1 \\
 = 2.
 \end{aligned}$$

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 our Mustine example, 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.

### 3.4 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).

| <b>Table 1. Demographics and Baseline Characteristics (N = 235)</b> |                           |                              |  |
|---|---------------------------|------------------------------|--|
| <b>Variable</b>   | <b>Licorice (N = 118)</b> | <b>Sugar-water (N = 117)</b> | <b>Standardized difference<sup>a</sup></b> |
| Age, y  | 57 ± 15                   | 58 ± 16                      | -0.09                                      |
| Gender (female), %  | 42                        | 38                           | 0.08                                       |
| Body mass index, kg/m <sup>2</sup>                                  | 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 <sup>b</sup>  | 27                        | 21                           |  |
| Medium <sup>b</sup>   | 64                        | 71                           |  |
| Large <sup>b</sup>  | 9                         | 9                            |  |

Summary statistics presented as percent of patients or mean ± SD.  
<sup>a</sup>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.  
<sup>b</sup>Surgery size: small (thoracoscopy); medium (thoracotomy <3 h), large (thoracotomy >3 h or blood loss >1000 mL).

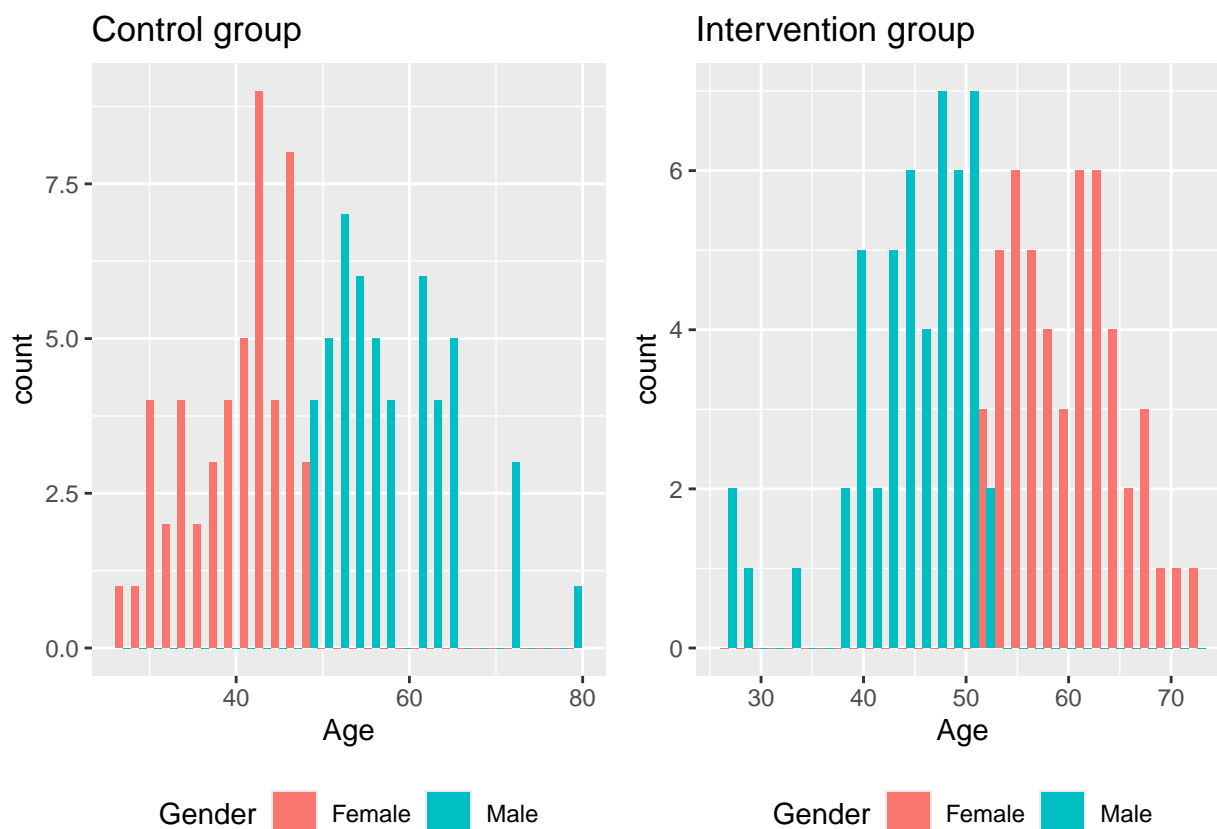
Figure 3.9: Summary statistics for an RCT comparing a licorice gargle (the intervention) to a sugar-water gargle (the standard). From @rueztler2013randomized



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       |
|---------------------|--------------|--------------|
| <b>Control</b>      | 57.51 (7.09) | 40.31 (5.83) |
| <b>Intervention</b> | 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  $\tau$  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 (which we'll still denote  $X$ ) between the groups.

**Example 4.2.** Summary statistics of the outcome for each group are shown in Table 4.2.

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 wider 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, by computing the standardized measure of effect size we found in Section 2.3. We 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).

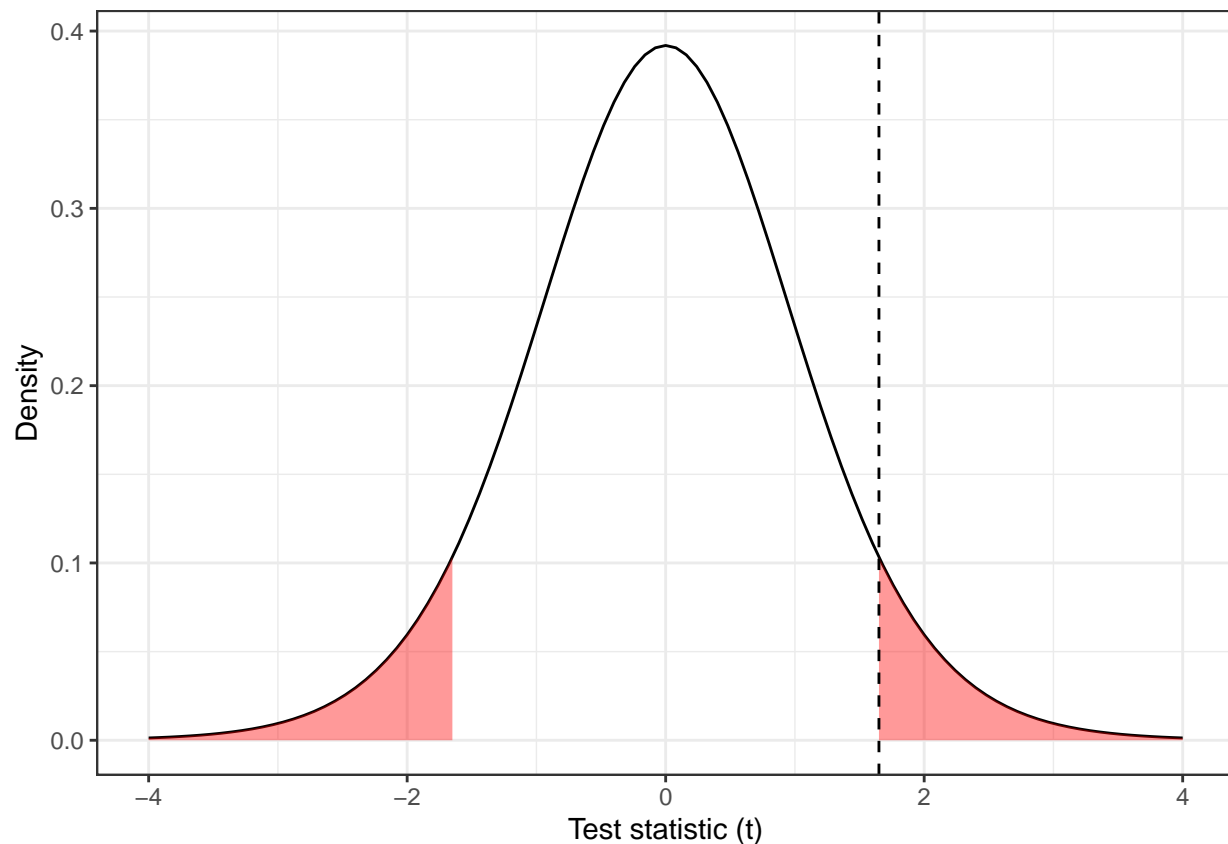


Figure 4.1: The distribution  $t_{14}$ , with  $t = 1.65$  shown by the dashed line and the ‘more extreme’ areas shaded.

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 (even 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  $\tau$  that are compatible with our data. That is, suppose the true treatment effect is  $\tau^*$ , and we test the hypothesis that  $\tau = \tau^*$ . For all values of  $\tau^*$  inside this range, our data are not sufficiently unlikely to reject the hypothesis at the 0.05 level. However, for all values of  $\tau^*$  outside this range, our data are sufficiently unlikely to reject that hypothesis. We can rearrange this to give a 95% confidence interval for  $\tau$ ,

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

**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  $\tau_M$  that we wanted to be able to correctly detect with probability  $1 - \beta$  (the power). It is difficult to say anything meaningful about the power of a trial after the fact. That said, if we failed to reject  $H_0$  and  $\tau_M$  is in the confidence interval for  $\tau$ , then that is a good indication that our trial was indeed underpowered.

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 of the differences 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              |

## 4.2 Using baseline values

In the 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 allocation (and therefore before they received the treatment or placebo). Baseline measurements are useful primarily for two reasons:

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.

We can perform an unpaired t-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. Under the null hypothesis of no difference, this test statistic has a  $t$ -distribution with 14 degrees of freedom, and so we have a  $P$ -value of 0.086.

```
2*(1-pt(1.85, df=14))
```

```
## [1] 0.08553164
```

Our 0.95 confidence interval is

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

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.

### 4.2.1 The variance of the treatment effect estimate

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

$$E(B_C) = E(B_T) = \mu_B.$$

Assuming some treatment effect  $\tau$  (which could still be zero) we have

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

Usually we will assume that

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

and this is generally fairly reasonable in practice.

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

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

that is, both are unbiased estimators of  $\tau$ .

However, whereas the first is based on data with variance  $\sigma^2$ , the second has

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

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



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   |

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

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

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

whereas for comparing differences from baseline

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

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

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. If the correlation is negative (very unusual in practice) then the  $\text{var}(\hat{\tau})$  is much larger - this also makes sense intuitively because we aren't allowing the coefficient of the baseline measurement to reflect the relationship between baseline and outcome.

**Example 4.5.** 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.

### 4.2.2 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  $\mu_B = \mu$ , ie. if you wouldn't expect a change in  $X$  over time anyway). One thing that makes this analysis attractive is that it lends itself to a paired  $t$ -test, which is more powerful (in general).

**Example 4.6.** Returning to 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.

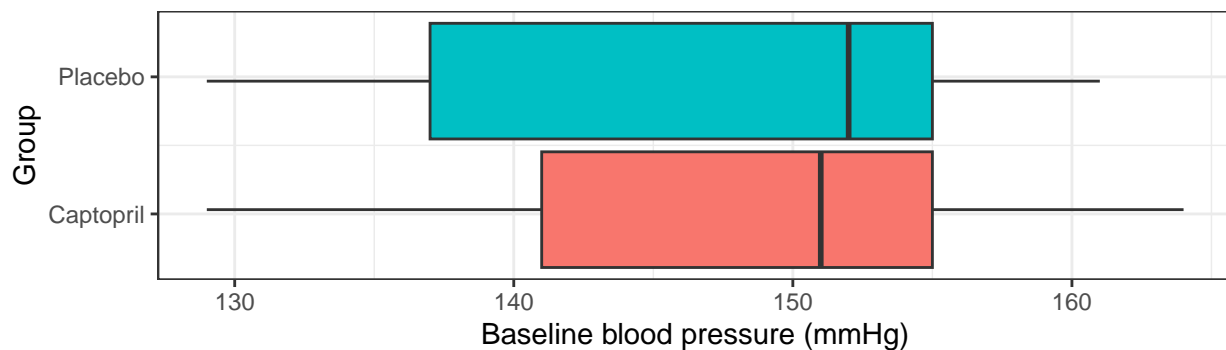


Figure 4.2: Baseline measurements from the Captopril trial.

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 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  $X$ , which has mean  $\mu$  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  $\tau$ , the treatment effect. We suppose also that  $X$  has variance  $\sigma^2$  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  $\mu_B$  in both groups (because of randomisation) and variance  $\sigma^2$ . We also assume that the true correlation between  $B$  and  $X$  is  $\rho$  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  $b_1, b_2, \dots, b_{2N}$ .

Our first step is to find an estimator for  $\tau$  that uses this information.

#### 4.3.1.1 Finding an estimator

Let's suppose that random variables  $Z$  and  $Y$  are jointly normally distributed with correlation  $\rho$

$$\begin{pmatrix} Z \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_Z \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_Z^2 & \rho\sigma_Z\sigma_Y \\ \rho\sigma_Z\sigma_Y & \sigma_Y^2 \end{pmatrix} \right). \quad (4.1)$$

From Equation (4.1), we know that  $E(Y) = \mu_Y$ .

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

$$\begin{pmatrix} Z \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1.5 \\ 1.5 & 3 \end{pmatrix} \right). \quad (4.2)$$

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

$$E(Y | Z = z) \neq E(Y).$$

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  $W = w$  doesn't give us any information about the distribution of  $Y$ , so we have

$$E(Y | W = w) = E(Y).$$

We can combine this information to work out  $E(Y | Z = z)$ . Firstly, we'll calculate the covariance of  $Z$  and  $Y - kZ$ , for some constant  $k$ . We can find this by

$$\begin{aligned} \text{cov}(Z, Y - kZ) &= E[(Z - \mu_Z)(Y - kZ - \mu_Y + k\mu_Z)] \\ &\quad (\text{using that } \text{cov}(Z, Y) = E[(Z - E(Z))(Y - E(Y))]) \\ &= E[(Z - \mu_Z)(Y - \mu_Y) - k(Z - \mu_Z)^2] \\ &= \rho\sigma_Z\sigma_Y - k\sigma_Z^2. \end{aligned}$$

If we set

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

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

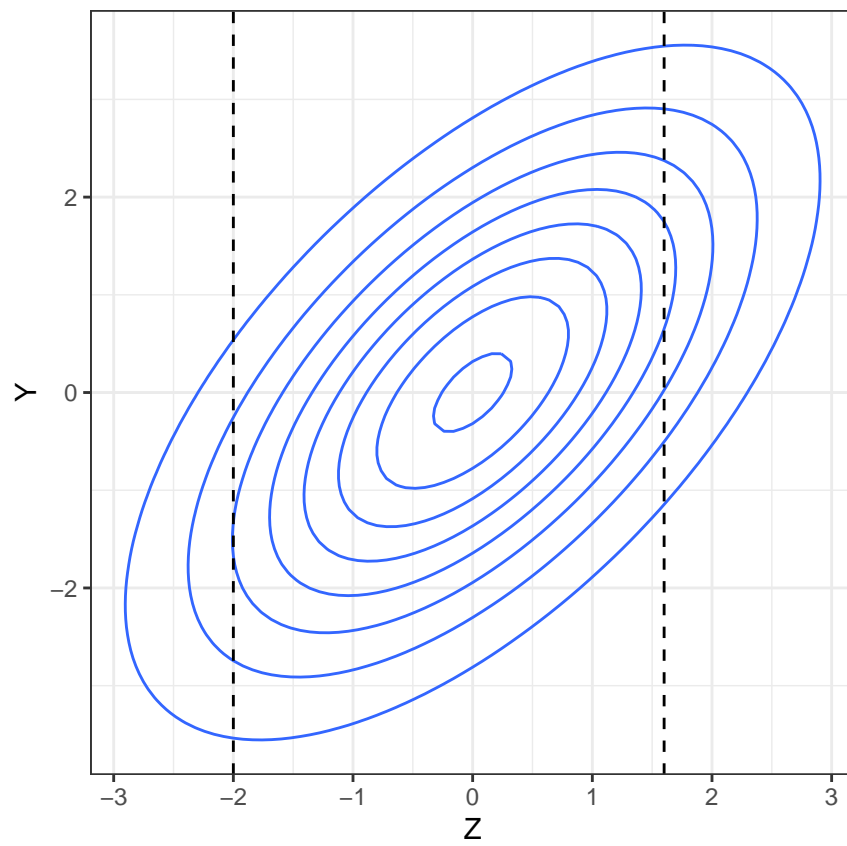


Figure 4.3: A bivariate normal density.

$$E(Y - \beta Z \mid Z = z) = E(Y - \beta Z) = \mu_Y - \beta \mu_Z.$$

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

$$\begin{aligned} E(Y \mid Z = z) &= E(\beta Z \mid Z = z) + \mu_Y - \beta \mu_Z \\ &= \mu_Y + \beta(z - \mu_Z). \end{aligned}$$

We can apply this to our clinical trials setting to find

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

Notice that because in our model  $\sigma_T = \sigma_C = \sigma$ , we have  $\beta = \rho$ .

From this, we find that

$$E(\bar{X}_T - \bar{X}_C \mid \bar{b}_T, \bar{b}_C) = \tau + \rho(\bar{b}_T - \bar{b}_C). \quad (4.3)$$

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  $\tau$ . Assuming  $\rho > 0$  (which is almost always the case) then if  $\bar{b}_T > \bar{b}_C$  the difference in outcome means overestimates  $\tau$ . Conversely, if  $\bar{b}_T < \bar{b}_C$ , the difference in outcome means underestimates  $\tau$ . 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

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

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

Notice, however, that if we use as our estimator

$$\hat{\tau} = (\bar{X}_T - \bar{X}_C) - \rho(\bar{b}_T - \bar{b}_C) \quad (4.4)$$

then, following from Equation (4.3) we have

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

We now return to our general bivariate normal variables  $Y$  and  $Z$  to find the variance of this estimator.

#### 4.3.1.2 What's the variance of this estimator?

To work out the variance of  $\hat{\tau}$  in Equation (4.4) we need to go back to our bivariate normal variables.

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

$$\text{var}(Y \mid Z = z) = E(Y^2 \mid Z = z) - [E(Y \mid Z = z)]^2. \quad (4.5)$$

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

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

$$\begin{aligned}\text{var}(Z) &= \text{E}(Z^2) - [\text{E}(Z)]^2 \\ \text{and therefore} \\ \text{E}(Z^2) &= \sigma_Z^2 + \mu_Z^2,\end{aligned}$$

we find that

$$\text{E}[(Y - \beta Z)^2 \mid Z = z] = \text{E}[(Y - \beta Z)^2] = S^2 + (\mu_Y - \beta\mu_Z)^2, \quad (4.6)$$

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

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

$$\text{E}[Y^2 \mid Z = z] - 2\beta\text{E}(Y \mid Z = z) + \beta^2 z^2 = \text{E}[Y^2 \mid Z = z] - 2\beta z(\mu_Y - \beta\mu_Z) - \beta^2 z^2.$$

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

$$\text{E}[Y^2 \mid Z = z] = S^2 + (\mu_Y - \beta\mu_Z)^2 + 2\beta z(\mu_Y - \beta\mu_Z) + \beta^2 z^2.$$

Now we can expand out

$$\text{E}(Y \mid Z = z) = \mu_Y + \beta(z - \mu_Z) = (\mu_Y - \beta\mu_Z) + \beta z$$

to find

$$[\text{E}(Y \mid Z = z)]^2 = (\mu_Y - \beta\mu_Z)^2 + 2\beta z(\mu_Y - \beta\mu_Z) + \beta^2 z^2.$$

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

$$\begin{aligned}\text{var}(Y \mid Z = z) &= \text{E}[Y^2 \mid Z = z] - [\text{E}(Y \mid Z = z)]^2 \\ &= S^2 \\ &= \sigma_Y^2(1 - \rho^2).\end{aligned}$$

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

### Back to our estimator!

Recall that in ANCOVA our estimator of the treatment effect  $\tau$  is

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

and that we have

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

Therefore, using the result we just found,

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

Notice that unlike our first estimator that used baseline values, in Section 4.2, the variance of the ANCOVA estimate can never exceed  $\frac{2\sigma_\epsilon^2}{N}$ ; 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  $\text{var}(\hat{\tau})$  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  $\tau$  relies on the correlation  $\rho$ , 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  $x_i$ :

$$\begin{aligned} x_i &= a_0 + \gamma b_i + \epsilon_i && \text{in group C} \\ x_i &= a_0 + \tau + \gamma b_i + \epsilon_i && \text{in group T.} \end{aligned}$$

Here, the  $\epsilon_i$  are independent errors with distribution  $N(0, \sigma_\epsilon^2)$ , the  $b_i$  are the baseline measurements for  $i = 1, \dots, N_T + N_C$ , for groups  $T$  and  $C$  with sizes  $N_T$  and  $N_C$  respectively, and  $a_0, \gamma$  are coefficients. Sometimes this is written instead in the form

$$x_i = a_0 + \tau G_i + \gamma b_i + \epsilon_i$$

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  $\tau$  is added. If  $G_i = 0$  (ie. participant  $i$  is in group  $C$ ) then it isn't.

Thinking back to the expectations of  $X, B$ , we expect that

$$a_0 = \mu - \gamma \mu_B,$$

though of course we won't know the true values of  $\mu$  or  $\mu_B$ .

We now have four parameters to estimate:  $a_0, \tau, \gamma$  and  $\sigma_\epsilon^2$ . 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

$$S(a_0, \tau, \gamma) = \sum_{i \text{ in } T} (x_i - a_0 - \tau - \gamma b_i)^2 + \sum_{i \text{ in } C} (x_i - a_0 - \gamma b_i)^2.$$

This leads to estimates  $\hat{a}_0, \hat{\tau}$  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  $\hat{a}_0, \hat{\tau}$  and  $\hat{\gamma}$  to estimate  $\sigma_\epsilon^2$ , using

$$\hat{\sigma}_\epsilon^2 = \frac{S(\hat{a}_0, \hat{\tau}, \hat{\gamma})}{N_T + N_C - 3}.$$

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  $\sigma_\epsilon^2$ ) 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  $\tau$ .

The technique described above is a well-established statistical method known as **ANCOVA** (short for the **A**nalysis of **C**ovariance), 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.

### 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:

- **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.  $\text{age}^2$ . Note that this would still be a linear model, because it is linear in the coefficients.
- **Multicollinearity:** 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. We can check this by finding the correlation matrix of the data (excluding the output). If there is multicollinearity in the data, we would expect to see a very high standard error for the coefficients of the affected variables.

**Definition 4.1.** In a linear regression model the variance of the fitted coefficient  $\hat{\beta}_j$  is

$$\text{var}(\hat{\beta}_j) = \frac{s^2}{1 - \mathbf{R}_j^2},$$

where  $s$  is the root mean-square error (RMSE) and  $\mathbf{R}_j^2$  is the squared multiple correlation for the regression of  $x_j$  on all the other covariates. The **variance inflation factor** is given by

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - \mathbf{R}_j^2}.$$

If  $\text{VIF} = 1$  then  $x_j$  is orthogonal to all other covariates. Large values of VIF indicate potential problems.

**Example 4.7.** 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
```

We can find the variance inflation factor and see that there appear to be no problems with collinearity:

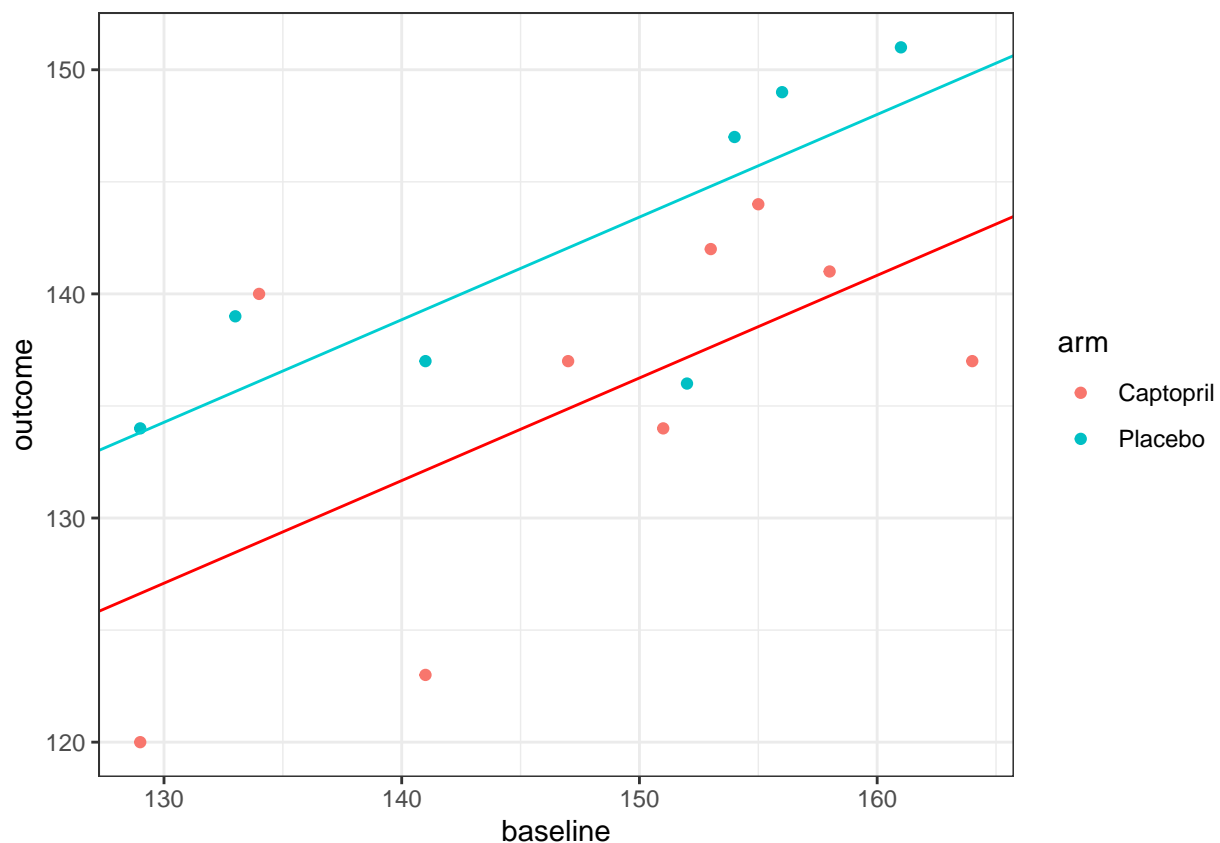
```
vif(lm_capt)
```

```
## baseline      arm
## 1.004117 1.004117
```

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 7.18 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  $\tau$ , 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 2.96. Therefore, to construct a 95% confidence interval for  $\hat{\tau}$ , we use (to 2 decimal places)

$$7.18 \pm t_{0.975;13} \times 2.96 = (0.78, 13.57).$$

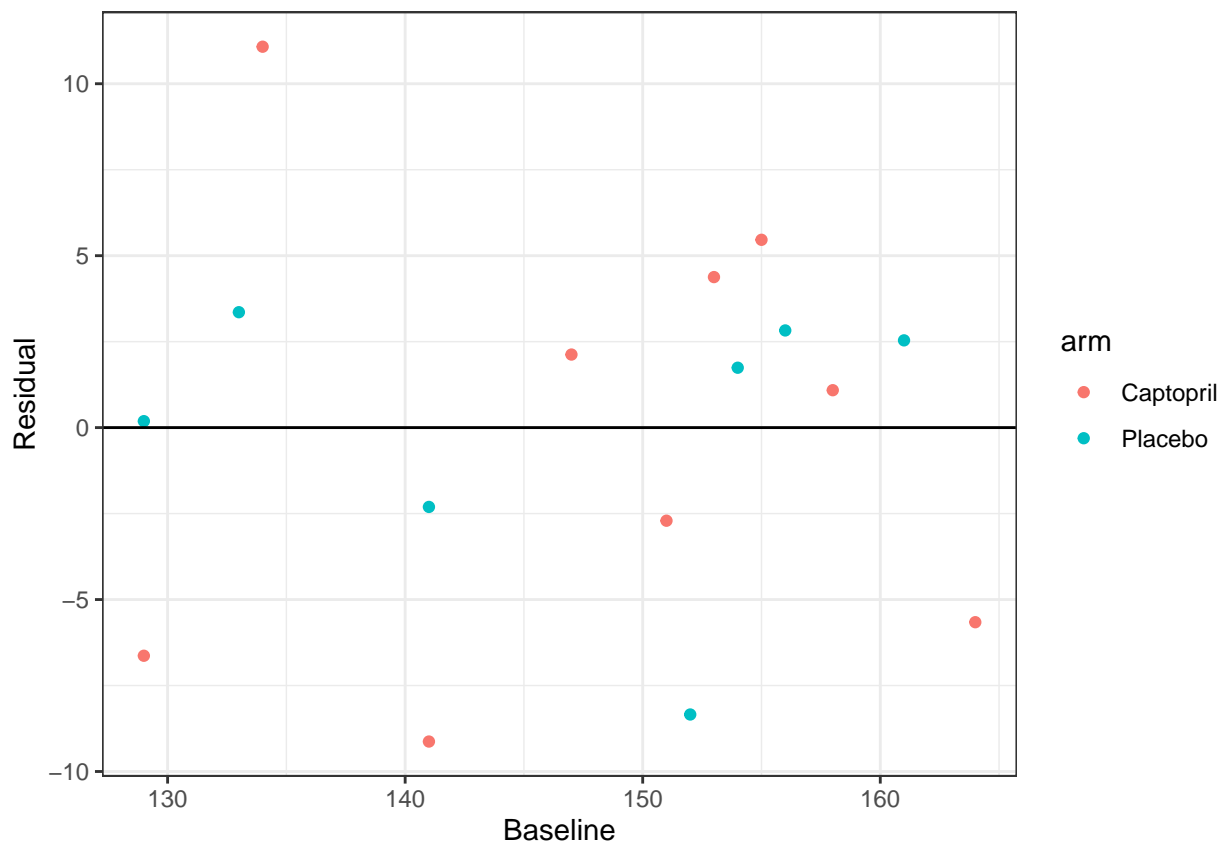
The model has  $n - p = 13$  degrees of freedom because there are  $n = 16$  data points and we are estimating  $p = 3$  parameters. Notice that unlike our previous confidence intervals, this doesn't contain zero, and so our analysis has enabled us to conclude that there is a significant reduction in 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  $\tau$ .

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 $\tau$ ?

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

$$\begin{aligned} E(\bar{X}_T - \bar{X}_C \mid \bar{b}_T, \bar{b}_C) &= \tau + \rho(\bar{b}_T - \bar{b}_C) \\ E[(\bar{X}_T - \bar{b}_T) - (\bar{X}_C - \bar{b}_C) \mid \bar{b}_T, \bar{b}_C] &= \tau + (\rho - 1)(\bar{b}_T - \bar{b}_C), \end{aligned}$$

that is, neither of these quantities are unbiased estimators of  $\tau$  (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  $\tau$ , 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  $\tau$ ), 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  $\gamma$  of 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  $\lambda b_i$  if participant  $i$  is in group  $T$ . Therefore, for participants in group  $C$ , the gradient is still  $\gamma$ , 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 4.8.** Continuing once again with the Captopril dataset, we now fit the model

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

We see that the  $p$ -value for the coefficient  $\lambda$  (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.

| Characteristic     | Beta  | 95% CI <sup>†</sup> | p-value |
|--------------------|-------|---------------------|---------|
| arm                |       |                     |         |
| Captopril          | —     | —                   |         |
| Placebo            | 8.7   | -80, 98             | 0.8     |
| baseline           | 0.46  | 0.05, 0.87          | 0.031   |
| arm * baseline     |       |                     |         |
| Placebo * baseline | -0.01 | -0.61, 0.59         | >0.9    |

<sup>†</sup>CI = Confidence Interval

#### 4.4.3 Can we include any other baseline covariates?

In Section 4.2 when our estimated treatment effect was  $\hat{\tau} = (\bar{x}_T - \bar{b}_T) - (\bar{x}_C - \bar{b}_C)$ , 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

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

and the inclusion of the coefficient  $\gamma$  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 4.9.** 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  |
| 58 | 83.8  | no        | high     | 62  |

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
```

We can calculate the variance inflation factor for this model

```
vif(lm_stresslin)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## treatment 1.017841  1      1.008881
## exercise  1.230692  2      1.053264
## age       1.248533  1      1.117378
```

and see that values are fairly close to one.

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 (Figure 4.4) to check that the model is a reasonable fit

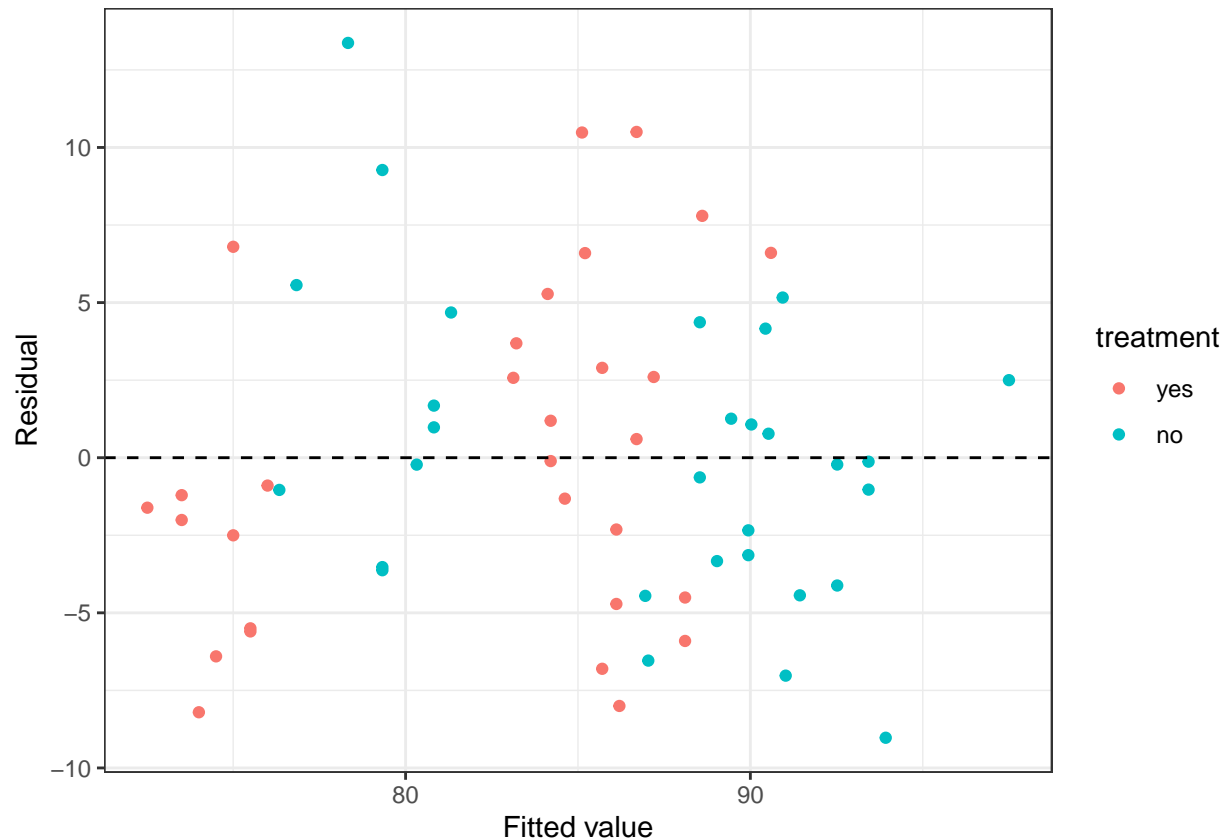


Figure 4.4: Residuals from the stress data model with only linear terms.

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.

The data description says that the researchers want to test for the effect of ‘the intervention and exercise’ so it seems reasonable to include the interaction of these two terms:

```
##
## 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.

This has somewhat improved the homogeneity of our residuals, as shown in Figure 4.5.

Combining the coefficients, someone who does a high level of exercise:

- is likely to reduce their stress score by around 13.7 if they receive the treatment
- is likely to reduce their stress score by around  $13.7 - 8.2 = 5.5$  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 ( $\alpha$ ).

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.



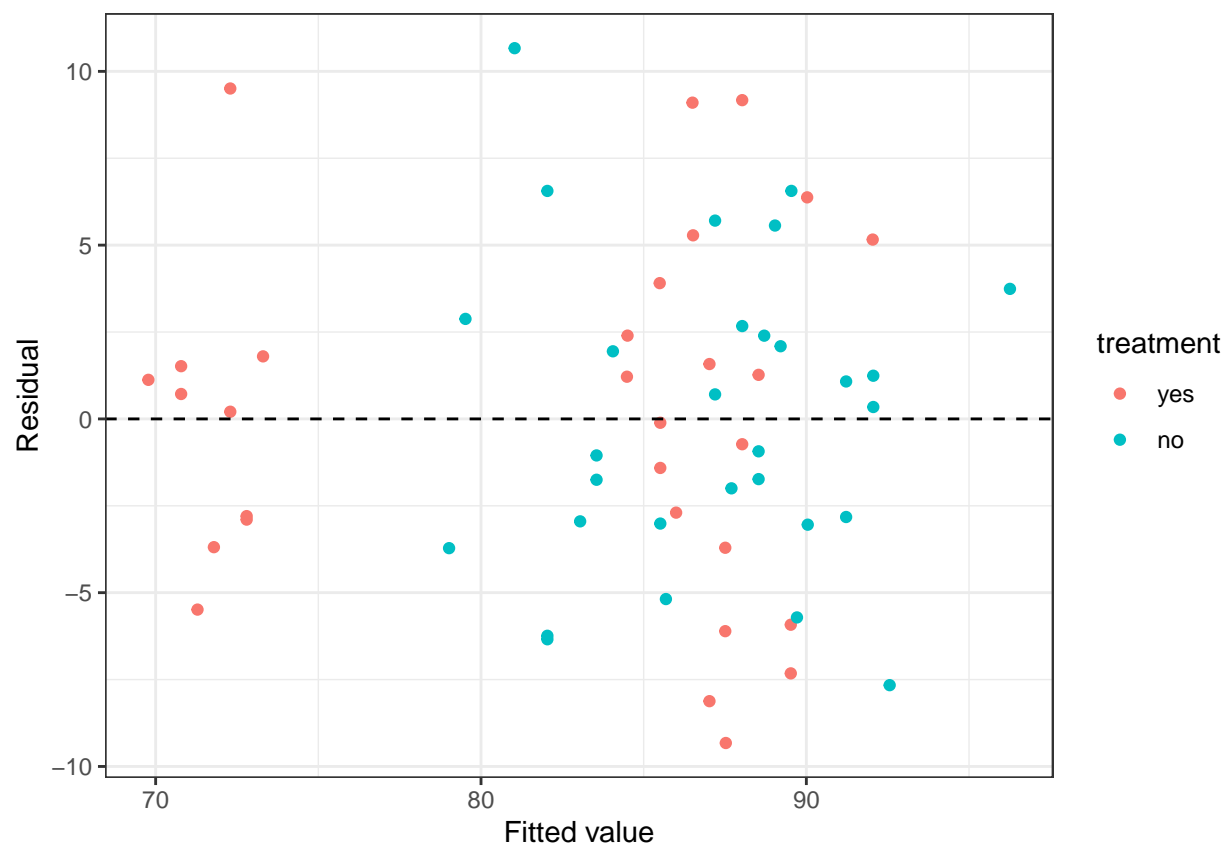


Figure 4.5: Residuals from the stress data model with the interaction between intervention and exercise included.

## 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,

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

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

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

where  $r_T, r_C$  are the observed values of  $R_T, R_C$ .

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  $\sigma$  in the binary situation, and the number of participants required will depend on the approximate values of  $\pi_T$  and  $\pi_C$ . 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  $\mu$  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  $\mu$ , which gives us

$$f(X) \approx f(\mu) + (X - \mu) f'(\mu) \quad (5.1)$$

and therefore

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

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

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

where both sides come from

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

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  $\pi$ . By substituting  $u = w^2$  we find

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

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

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

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

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

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  $\pi$  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  $f(p_X) = \arcsin(\sqrt{p_X})$  will be [approximately] normally distributed too. In fact,  $\arcsin(\sqrt{p_X})$  is approximately normally distributed with mean  $\arcsin(\sqrt{\pi_X})$  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

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

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  $\alpha$  and power  $1 - \beta$  is

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

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

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

Because

$$\arcsin(\sqrt{\pi_T}) - \arcsin(\sqrt{\pi_C})$$

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 ( $\pi_C$ ) is well understood, and the probability for the intervention group ( $\pi_T$ ) can be specified by using the nearest clinically important value of  $\pi_T$ .

**Example 5.1.** (From Smith et al., 1994) This trial compares two approaches to managing malignant 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

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

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

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

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

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

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

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

and the focus of our analysis is on comparing  $\pi_C$  and  $\pi_T$ . 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 Simple hypothesis tests

We can tabulate the results of a trial with a binary outcome like this:

|                  | Successes | Failures    | Total |
|------------------|-----------|-------------|-------|
| <b>Treatment</b> | $r_T$     | $n_T - r_T$ | $n_T$ |
| <b>Control</b>   | $r_C$     | $n_C - r_C$ | $n_C$ |
| <b>Total</b>     | $r$       | $n - r$     | $n$   |

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

We can estimate  $\pi_C$  and  $\pi_T$  by the sample proportions

$$p_C = \frac{r_C}{n_C}$$
$$p_T = \frac{r_T}{n_T}.$$

and can use this data to test our null hypothesis that  $\pi_T = \pi_C = \pi$ .

**Example 6.1.** The data for the examples in this section 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

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

### 6.1.1 A simple method: chi-squared

One 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 (ie. the values we'd expect under  $H_0$ ), 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  $\chi^2$  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 probable it is under the  $\chi^2_1$  distribution.

**Example 6.2.** Applying this to 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  $\chi^2$  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 we have sufficient evidence to reject  $H_0$  for  $\alpha = 0.05$  (or indeed for many much lower values of  $\alpha$ ).

### 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.

We will use  $y_{iC}$  to denote the outcome variable (0 or 1, in this case) of the  $i$ -th participant in the control group (and so on).

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

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

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

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

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

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

Gathering these terms together we can find the complete likelihood function

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

The log-likelihood is therefore

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

If we differentiate with respect to  $\pi_C$ , we find

$$\frac{dl(\pi_C, \pi_T | n_C, n_T, r_C, r_T)}{d\pi_C} = \frac{r_C}{\pi_C} - \frac{n_C - r_C}{1 - \pi_C}.$$

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

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

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

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  $n_C, n_T$ ,  $\lambda_{LR}$  has a  $\chi^2$  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 section the question we were interested in was 'is there sufficient evidence to reject  $H_0 : \pi_C = \pi_T$ ?' 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  $\pi_C$  and  $\pi_T$ , 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 be 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  $\text{ARD}=0.02$ , we have  $\text{NNT}=50$ . 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  $\text{ARD} = 0.05$  and  $\text{NNT} = 20$ . 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  $\text{ARD} = 0.01$  and  $\text{NNT} = 100$ . It might be decided that the necessary changes and costs are not worth it for such a small difference.

Although the NNT is popular with medics, it 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  $\tau_{ARD}$  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  $\tau_{ARD}$  is

$$\left( 0.364 - z_{0.975} \sqrt{0.00811}, 0.364 + z_{0.975} \sqrt{0.00811} \right) = (0.187, 0.541).$$

As we should expect from the very low  $p$ -values we saw in the examples in Section 6.1, the 95% confidence interval does not contain zero.

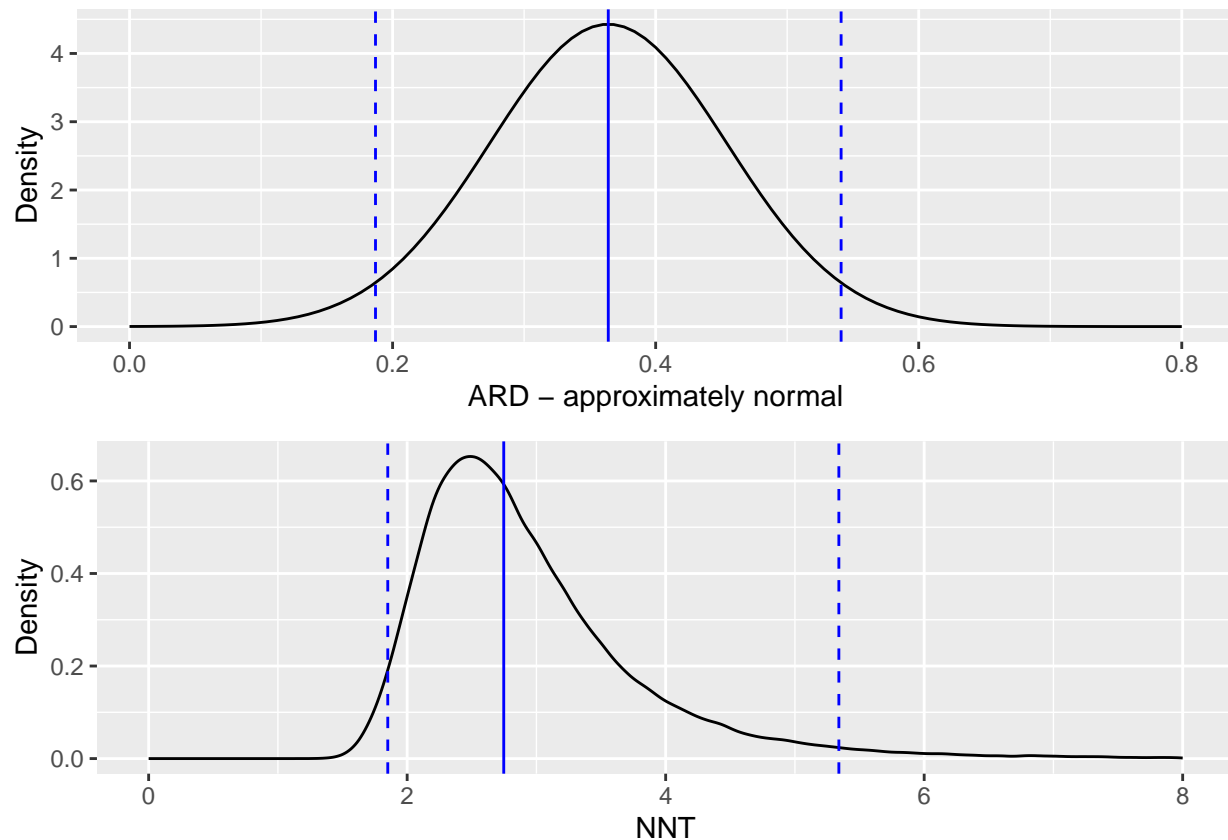
If we want to think instead in terms of NNT (the number needed to treat), then we need to find the reciprocal of our estimate of  $\tau_{ARD}$ :

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

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  $\tau_{ARD}$  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 |
|------------------|-----------|----------|-------|
| <b>Treatment</b> | 9         | 5        | 14    |
| <b>Control</b>   | 4         | 8        | 12    |
| <b>Total</b>     | 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  $\tau_{ARD}$  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 you agree!

### 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. This is the Wilson method, first presented in Wilson (1927).

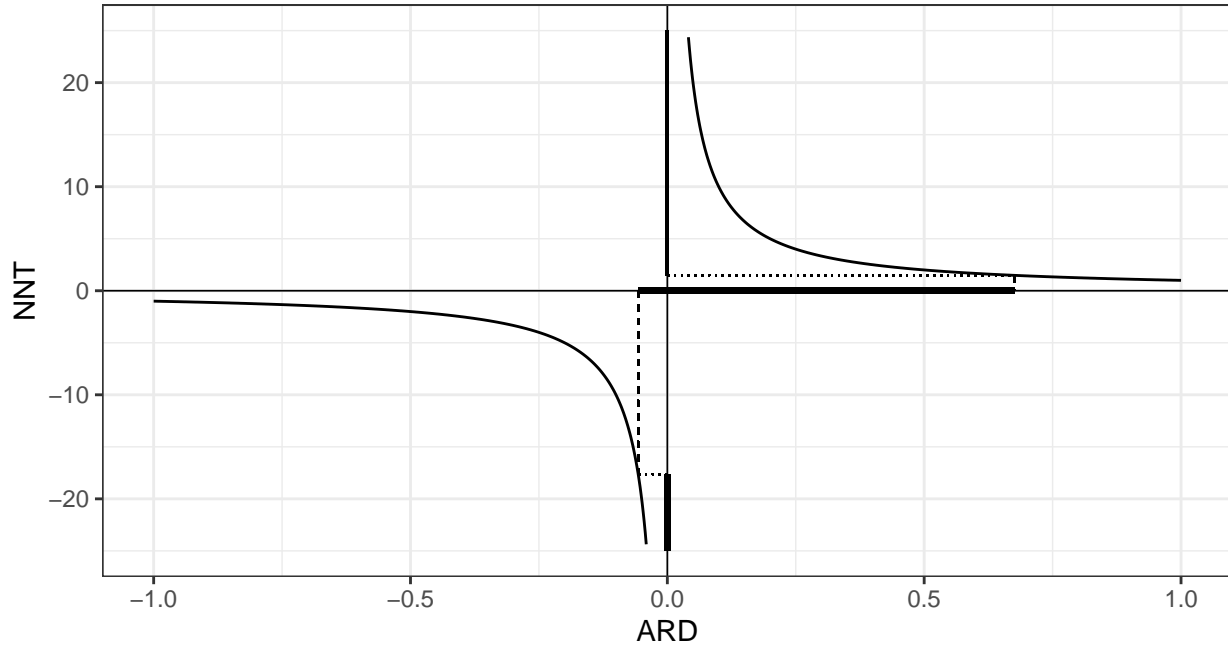


Figure 6.1: The relationship between the confidence interval for the ARD and the NNT, when the ARD interval contains zero.

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

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

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

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

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

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

which we solve to get the limits

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

In Wilson's method, we instead keep  $\pi$  in the right hand side and solve the quadratic in Equation (6.1) in terms of  $\pi$ .

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 Wilson 95% CI will be the roots of

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

which work out to be

## [1] 0.004 0.105

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 Wilson 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).

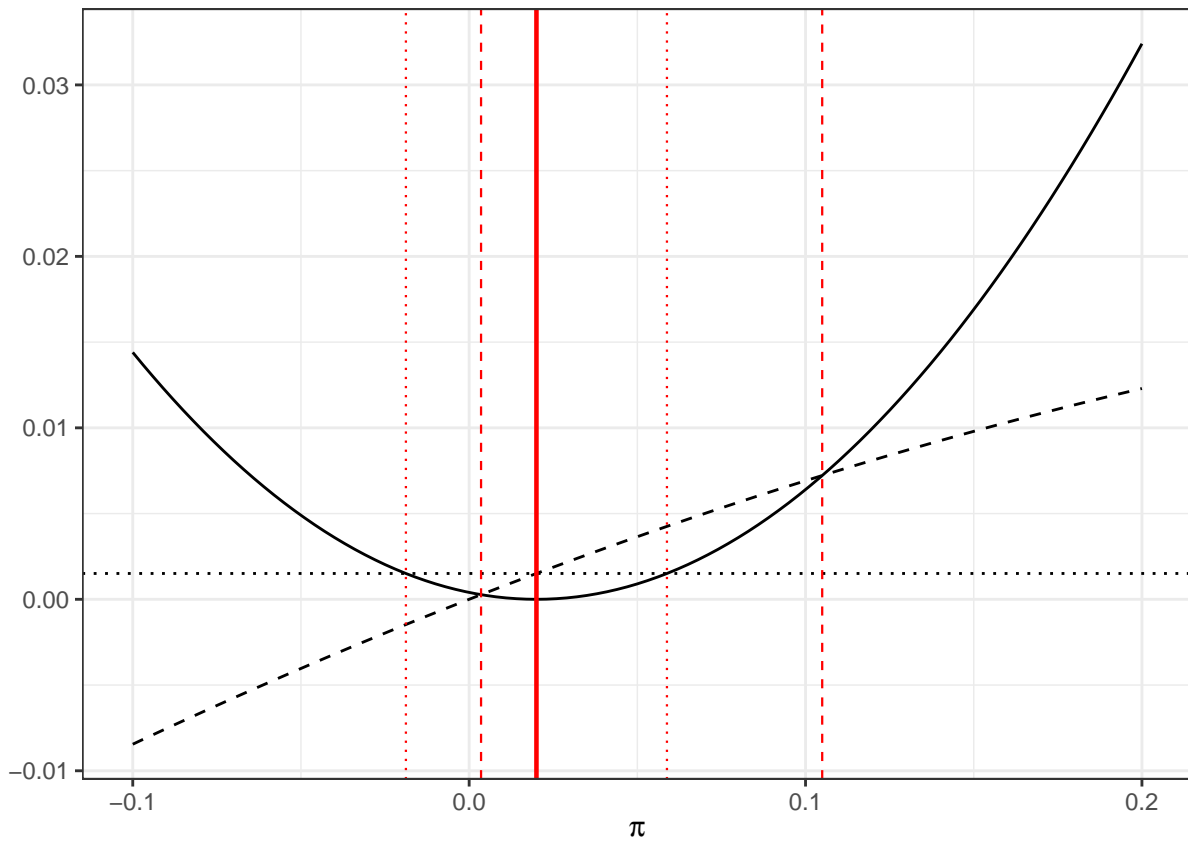


Figure 6.2: LHS of Equation 6.1 (solid black) with RHS of (black dashed) and RHS with estimate  $p$  subbed in (black dotted). The limits of the confidence intervals are where the curves cross, shown by red lines: dashed for Wilson, dotted for standard.

**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



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

Solving this equation in the usual way (using the quadratic formula) we find the Wilson 95% CI to be

```
## [1] 0.56 0.80
```

By contrast, in our standard method we have

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

which is

```
## [1] 0.57 0.81
```

We can see this graphically

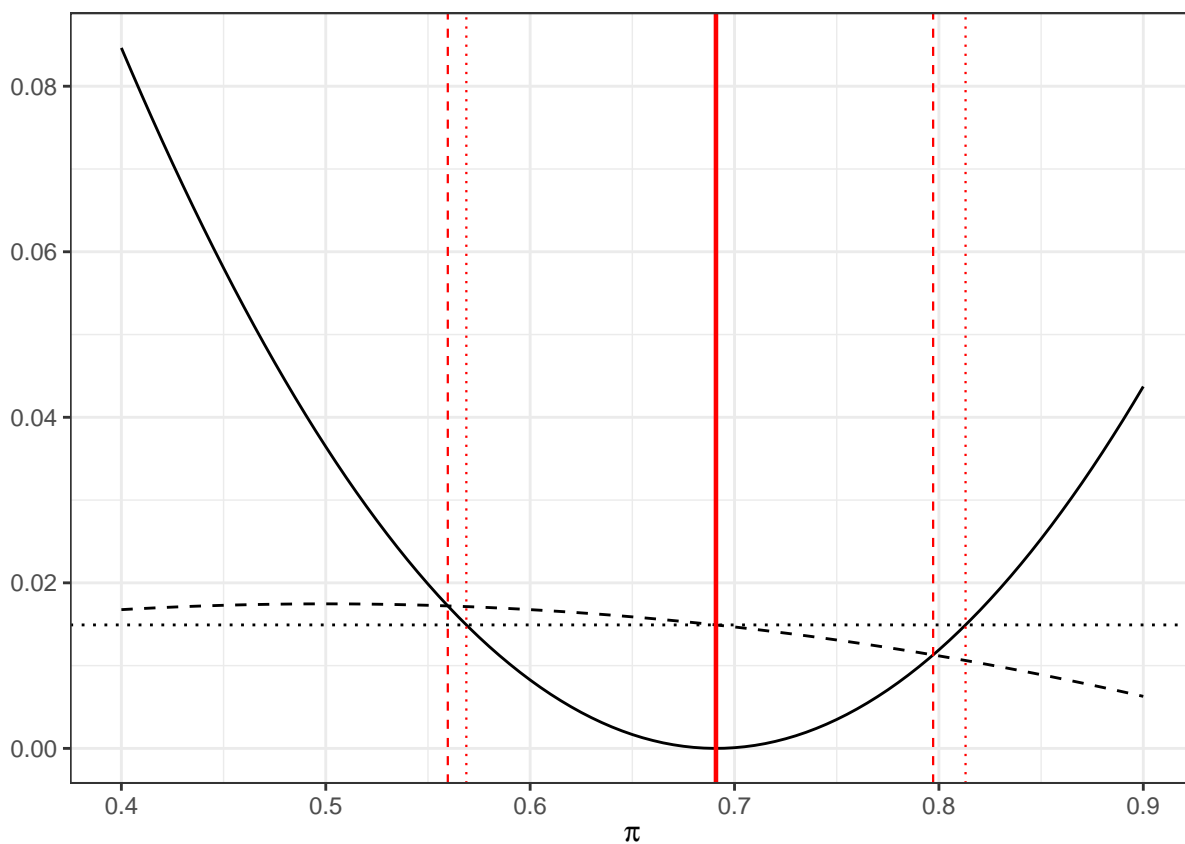


Figure 6.3: As before, dashed for Wilson, dotted for standard

Notice that the Wilson interval is asymmetric, which is more realistic.

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

```
## [1] 0.22 0.46
```

compared to the standard confidence interval

```
## [1] 0.20 0.45
```

Again, we can see this graphically.

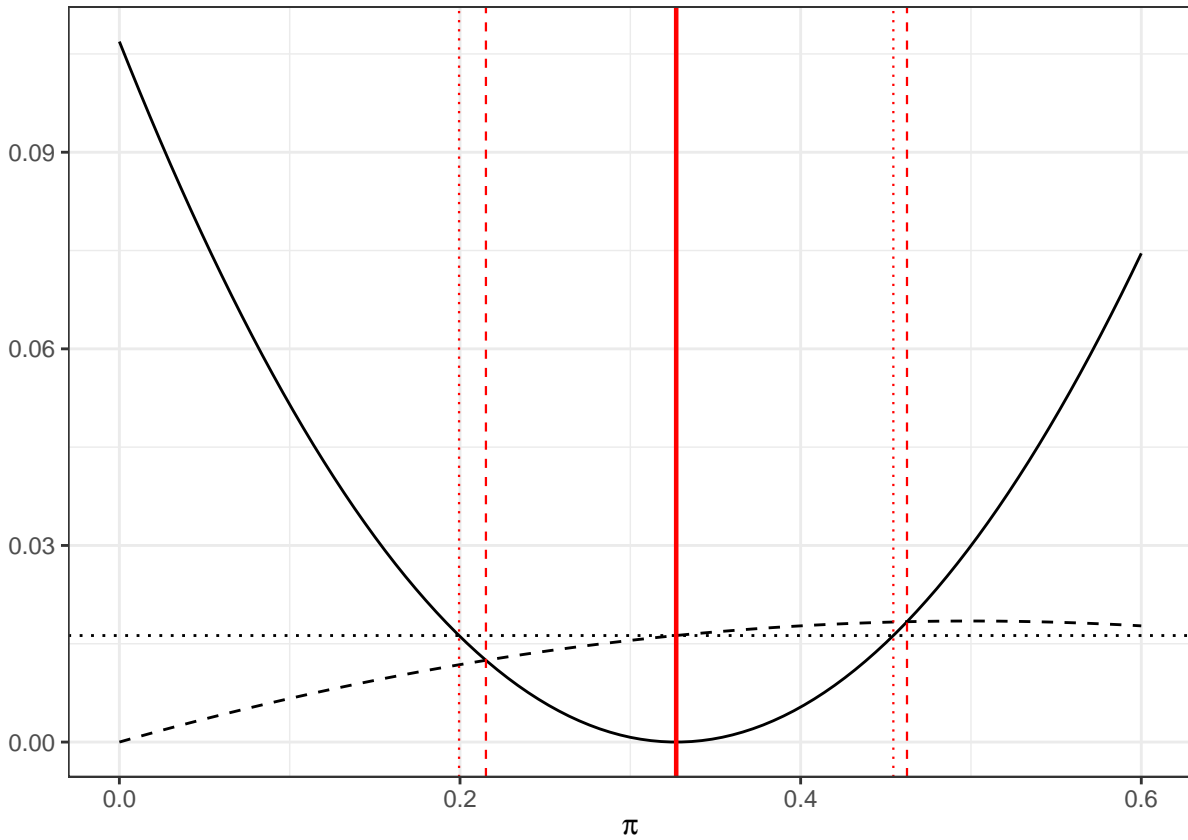


Figure 6.4: As before, dashed for Wilson, dotted for standard

### 6.2.1.3 Extending this to $\pi_T - \pi_C$

What the Wilson 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  $\alpha$  are given by

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

We can rewrite this as

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

where  $\omega_T$  and  $\omega_C$  are the widths of the separate single-sample ‘standard’ confidence intervals for  $p_T$  and  $p_C$ . In Wilson’s method, we proceed in the same way, but instead use the widths of the Wilson 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 Wilson 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 Wilson 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

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

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

Returning to our earlier material, we could invert this interval to find a confidence interval for the number needed to treat (NNT).

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

The measures we have looked at so far, particularly the ARD, are quite analogous 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

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

The **odds ratio** is defined as

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

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 Treatment 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, 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  $(\frac{1}{u}, \frac{1}{l})$ , 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

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

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  $\phi$  (the true value) and variance

$$\text{var}(\log(p_T)) + \text{var}(\log(p_C)).$$

We can now apply the Delta method (see section 5.1) to find that (using Equation (5.3))

$$\text{var}[\log(p_T)] = \text{var} \left[ \log \left( \frac{r_T}{n_T} \right) \right] \approx \frac{\pi_T(1 - \pi_T)}{n_T} \times \left( \frac{1}{\pi_T} \right)^2 = \frac{1}{n_T \pi_T} - \frac{1}{n_T}.$$

Since we estimate  $\pi_T$  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

$$\log \left( \frac{p_T}{p_C} \right) \sim N \left( \phi, (r_T^{-1} - n_T^{-1}) + (r_C^{-1} - n_C^{-1}) \right)$$

and so we can generate  $100(1 - \alpha)\%$  confidence intervals for  $\phi$  as  $(l_{RR}, u_{RR})$ , where the limits are

$$\log \left( \frac{p_T}{p_C} \right) \pm z_{\frac{\alpha}{2}} \sqrt{(r_T^{-1} - n_T^{-1}) + (r_C^{-1} - n_C^{-1})}.$$

This then translates to an interval for the risk ratio itself of  $(e^{l_{RR}}, e^{u_{RR}})$ .

**Example 6.9.** Returning once again to our streptomycin example, recall that we have

$$r_T = 38$$

$$n_T = 55$$

$$r_C = 17$$

$$n_C = 52$$

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 no less true 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

$$\text{outcome}_i = \mu + \tau G_i + \beta_1 \times \text{baseline}_{1i} + \dots + \beta_p \times \text{baseline}_{pi} + \text{error}_i,$$

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  $\text{baseline}_1, \dots, \text{baseline}_p$  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

$$\text{mean outcome}_i = \mu + \tau G_i + \beta_1 \times \text{baseline}_{1i} + \dots + \beta_p \times \text{baseline}_{pi}.$$

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  $\pi$  tends to zero,  $\text{logit}(\pi)$  tends to  $-\infty$ , and as  $\pi$  tends to one,  $\text{logit}(\pi)$  tends to  $\infty$ . 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  $\pi$ , since the mean outcome is the estimate of the probability of success) in the model

$$\text{logit}(\pi) = \mu + \tau G + \beta_1 \times \text{baseline}_1 + \dots + \beta_p \times \text{baseline}_p \quad (6.4)$$

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:

$$\begin{aligned} \text{logit}(\pi) &= \log \left( \frac{\pi}{1 - \pi} \right) = \log(\text{Odds of success}) = \mu + \tau + \beta_1 x_1 + \dots + \beta_p x_p & (\text{group T}) \\ \text{logit}(\pi) &= \log \left( \frac{\pi}{1 - \pi} \right) = \log(\text{Odds of success}) = \mu + \beta_1 x_1 + \dots + \beta_p x_p & (\text{group C}) \end{aligned}$$

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,  $\tau$  is the log of the odds ratio, or  $e^\tau$  is the odds ratio adjusted for variables  $x_1, \dots, x_p$ . Put another way, while the baseline covariates  $x_1, \dots, x_p$  affect the probability of 'success' (or whatever our binary outcome's one means),  $\tau$  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  $\mu$ ,  $\tau$  and  $\beta_1, \dots, \beta_p$ . To ease notation for this section we will collect these into a vector  $\beta$ , with  $\beta_0 = \mu$  and  $\beta_{p+1} = \tau$ , so that our  $G$  variable is now  $x_{p+1}$ .

This notation allows us to write the linear function on the RHS of Equation (6.4) for participant  $i$  as

$$x_i^T \beta = \sum_{j=0}^{p+2} x_{ij} \beta_j,$$

where

- $x_{i0} = 1$  (so that  $\beta_0$  is the intercept  $\mu$ )
- $x_{i1}, \dots, x_{ip}$  are the baseline covariates.
- $x_{i,p+1} = \begin{cases} 0 & \text{if participant } i \text{ is in group } C \\ 1 & \text{if participant } i \text{ is in group } T \end{cases}$

If  $\pi_i$  is the probability that the outcome for participant  $i$  is 1, where  $i = 1, \dots, n$ , then the logistic model specifies these  $n$  parameters through the  $p + 2$  parameters  $\beta_j$ , via the  $n$  expressions

$$\text{logit}(\pi_i) = x_i^T \beta. \quad (6.5)$$

Using the Bernoulli distribution, the log-likelihood given data  $y_1, \dots, y_n$  is

$$\begin{aligned} \ell(\{\pi_i\} \mid \{y_i\}) &= \sum_{i=1}^n [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] \\ &= \sum_{i=1}^n \left[ y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i) \right], \end{aligned}$$

where  $y_i = 0$  or  $1$  is the outcome for participant  $i$ . Using Equation (6.5) we can rewrite this in terms of  $\beta$  as

$$\ell(\{\beta_j\} \mid \text{data}) = \sum_{i=1}^n \left[ y_i x_i^T \beta - \log(1 + e^{x_i^T \beta}) \right].$$

The fitted model is then the one with the values  $\beta_j$ ,  $j = 0, \dots, q$ , that maximise this expression (and hence maximise the likelihood itself), which we will label the  $\{\hat{\beta}_j\}$ .

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  $\beta_j$ , 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 cholangio-pancreatogram, 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  $\sim 16\%$  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  $\mathbf{rx}$ : 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
##  Min.    :1001    1_UM  :164  Min.    :19.00  Min.    :1.000  0_no :523
##  1st Qu.:1152    2_IU  :413  1st Qu.:35.00  1st Qu.:1.500  1_yes: 79
##  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
##
##          rx
##  0_placebo    :307
##  1_indomethacin:295
##
##
##
##
```

```
## Some things to note:
## There are very few patients from site 4, and not many from site 3
## The age range goes from 19 to 90
## 'rx' is the group variable

## 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"))

print(summary(glm_indo_agelin), digits=2)
```

```
##
## Call:
## glm(formula = outcome ~ age + site + risk + rx, family = binomial(link = "logit"),
##      data = indo_rct)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.7863     0.6414   -2.8    0.005 **
## age           -0.0085     0.0099   -0.9    0.394
## site2_IU      -1.2293     0.2693   -4.6 5e-06 ***
## site3_UK      -1.1279     0.7759   -1.5    0.146
```

```
## site4_Case      -13.8644   827.9211    0.0    0.987
## risk            0.5619    0.1423     3.9    8e-05 ***
## rx1_indomethacin -0.7633    0.2615    -2.9    0.004 **
## ---
## 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: 441.1
##
## Number of Fisher Scoring iterations: 14
```

```
vif(glm_indo_agelin)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## age  1.046789  1          1.023127
## site 1.089701  3          1.014420
## risk 1.127145  1          1.061671
## rx   1.008980  1          1.004480
```

```
print(summary(glm_indo_agesq), digits=2)
```

```
##
## 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)   -2.0e+00   4.9e-01   -4.0    7e-05 ***
## I(age^2)       -9.4e-05   1.1e-04   -0.9    0.385
## site2_IU       -1.2e+00   2.7e-01   -4.6    5e-06 ***
## site3_UK       -1.1e+00   7.8e-01   -1.5    0.144
## site4_Case     -1.4e+01   8.3e+02    0.0    0.987
## risk           5.6e-01   1.4e-01    3.9    9e-05 ***
## rx1_indomethacin -7.6e-01   2.6e-01   -2.9    0.004 **
## ---
## 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
##
## Number of Fisher Scoring iterations: 14
```

```
vif(glm_indo_agesq)
```

```
##           GVIF Df GVIF^(1/(2*Df))
```

```
## I(age^2) 1.052813 1 1.026067
## site 1.089772 3 1.014431
## risk 1.132658 1 1.064264
## rx 1.007972 1 1.003978
```

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"))
print(summary(glm_indo), digits=2)
```

```
##
## Call:
## glm(formula = outcome ~ site + risk + rx, family = binomial(link = "logit"),
##      data = indo_rct)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.23      0.38   -5.8    5e-09 ***
## site2_IU          -1.22      0.27   -4.5    6e-06 ***
## site3_UK          -1.13      0.78   -1.5    0.145
## site4_Case       -13.84    833.24    0.0    0.987
## risk               0.58      0.14    4.2    3e-05 ***
## rx1_indomethacin  -0.75      0.26   -2.9    0.004 **
## ---
## 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: 439.8
##
## Number of Fisher Scoring iterations: 14
```

```
vif(glm_indo)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## site 1.088963 3 1.014306
## risk 1.094106 1 1.045995
## rx 1.005852 1 1.002922
```

Notice that

- The estimate and standard error of  $\hat{\tau}$  (the coefficient of `rx`) haven't changed much at all (mainly because, as the VIF shows us, there are no issues with multicollinearity)
- The standard error for the site 4 cases is enormous - this is because there are so few observations.

From the summary we see that  $\hat{\tau} = -0.752$ , with a standard error of 0.261. A 95% CI for  $\tau$  is therefore

$$-0.752 \pm 1.96 \times 0.261 = (-1.26, -0.240).$$

This model supports the hypothesis that the log of the odds ratio isn't zero, and therefore we can reject  $H_0$ .

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  $x^T \hat{\beta}$  for given values of  $x$  gives us an 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,

$$p(X_i = 1) = \frac{\exp(\text{logit}_i)}{1 + \exp(\text{logit}_i)},$$

where  $\text{logit}_i$  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

```
print(summary(glm_indo), digits=2)

##
## Call:
## glm(formula = outcome ~ site + risk + rx, family = binomial(link = "logit"),
##      data = indo_rct)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.23      0.38   -5.8    5e-09 ***
## site2_IU          -1.22      0.27   -4.5    6e-06 ***
## site3_UK          -1.13      0.78   -1.5    0.145
## site4_Case     -13.84    833.24    0.0    0.987
## risk              0.58      0.14    4.2    3e-05 ***
## rx1_indomethacin  -0.75      0.26   -2.9    0.004 **
## ---
## 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: 439.8
##
## Number of Fisher Scoring iterations: 14
```

For example, a patient from site 1, with risk level 3, in the control group would have odds

$$\exp(-2.2307 + 3 \times 0.5846) = 0.6207,$$

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

$$\exp(-2.2307 - 1.2204 + 1 \times 0.5846 - 0.7523) = 0.0268,$$

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 coefficients 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 |

## 6.4 Diagnostics for logistic regression

There are many diagnostic techniques for binomial data (see eg. Collett (2003)) 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 assessing logistic regression models somewhat tricky.

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. 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 assess this is by using Receiver Operating Characteristic (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  $\hat{p}_i \geq t$  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):

$$p(\hat{p}_i \geq t \mid X_i = 1).$$

**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):

$$p(\hat{p}_i < t \mid X_i = 0)$$

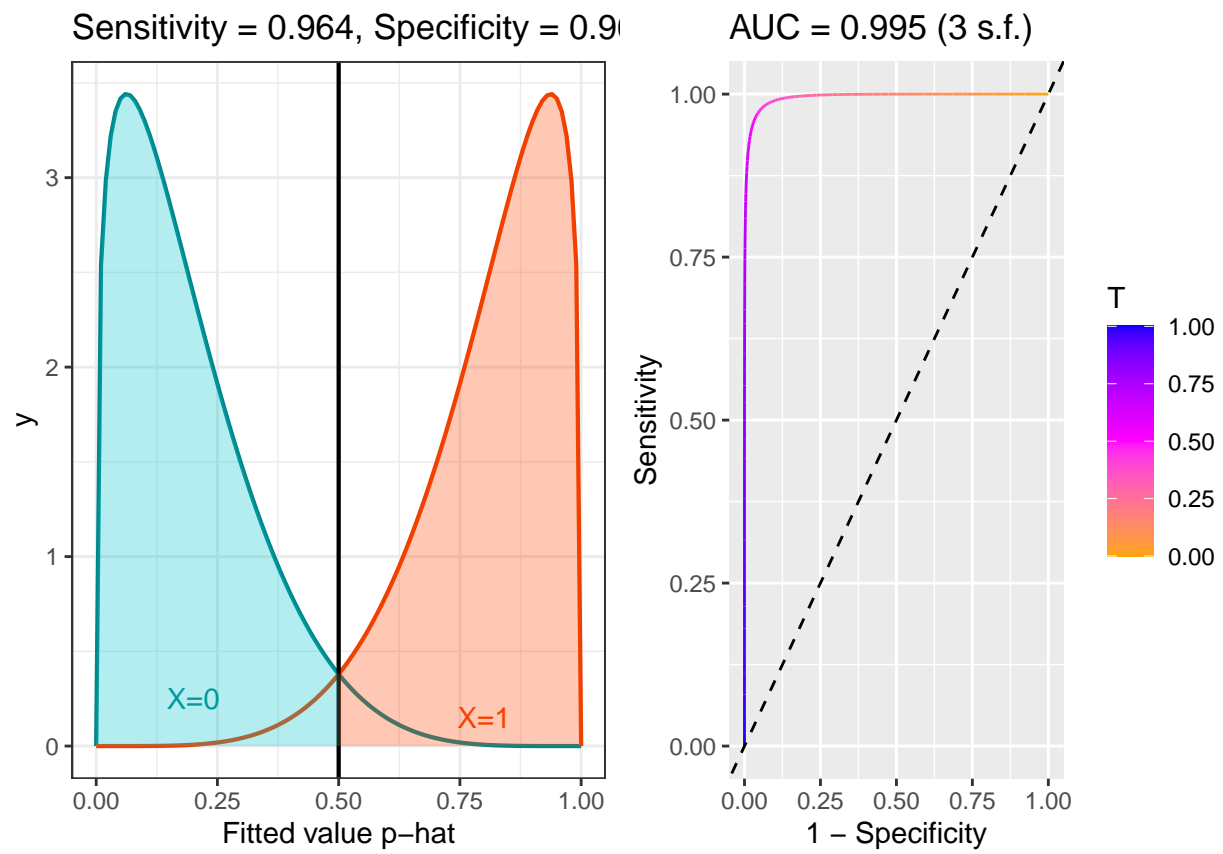
We estimate these by the proportions within the dataset.

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  $t = 1$ , 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 plots below shows the distributions of fitted values for patients with  $X = 0$  and  $X = 1$ , for models with good and poor separation, and the corresponding ROC curves. The threshold is (arbitrarily) set to 0.5, and so the Sensitivity and Specificity reflect that. The AUC (area under the curve) 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.

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.



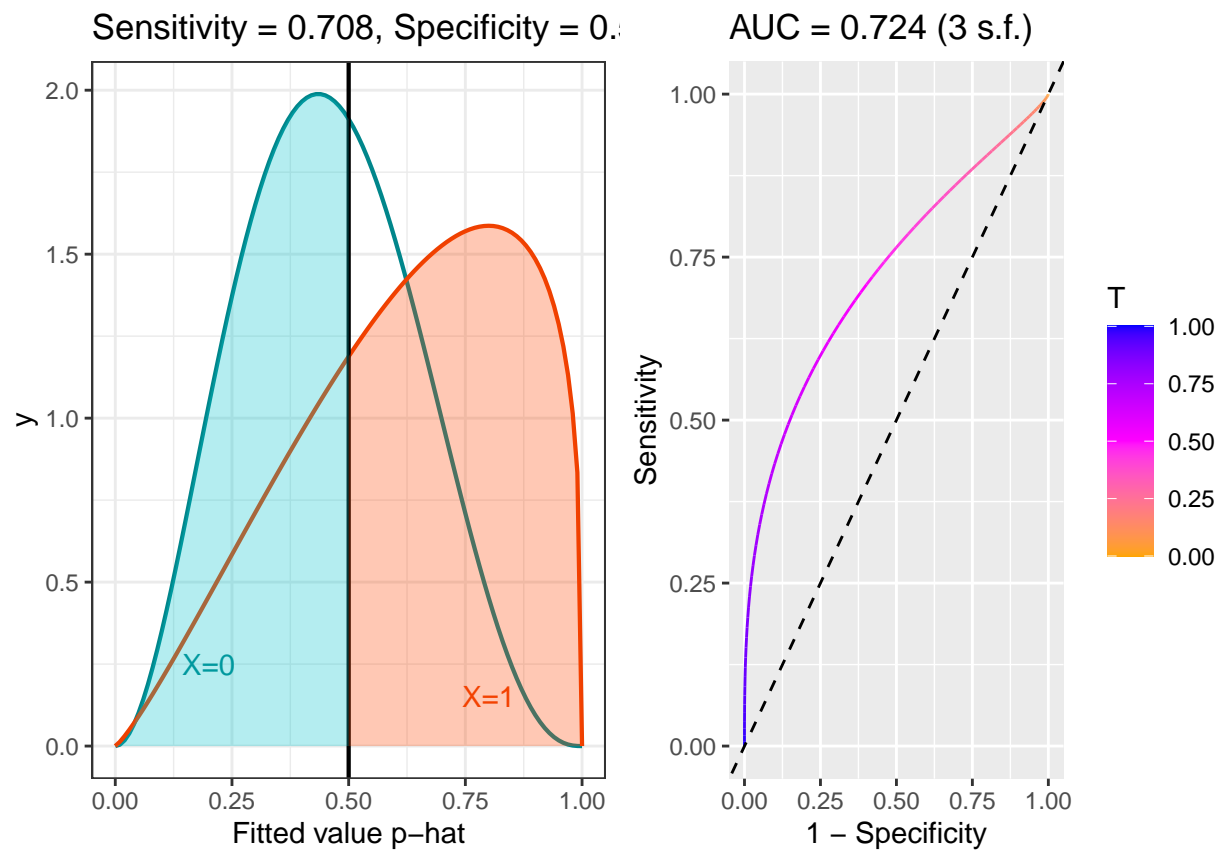


Figure 6.6: ROC example with poor separation.



**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)
```

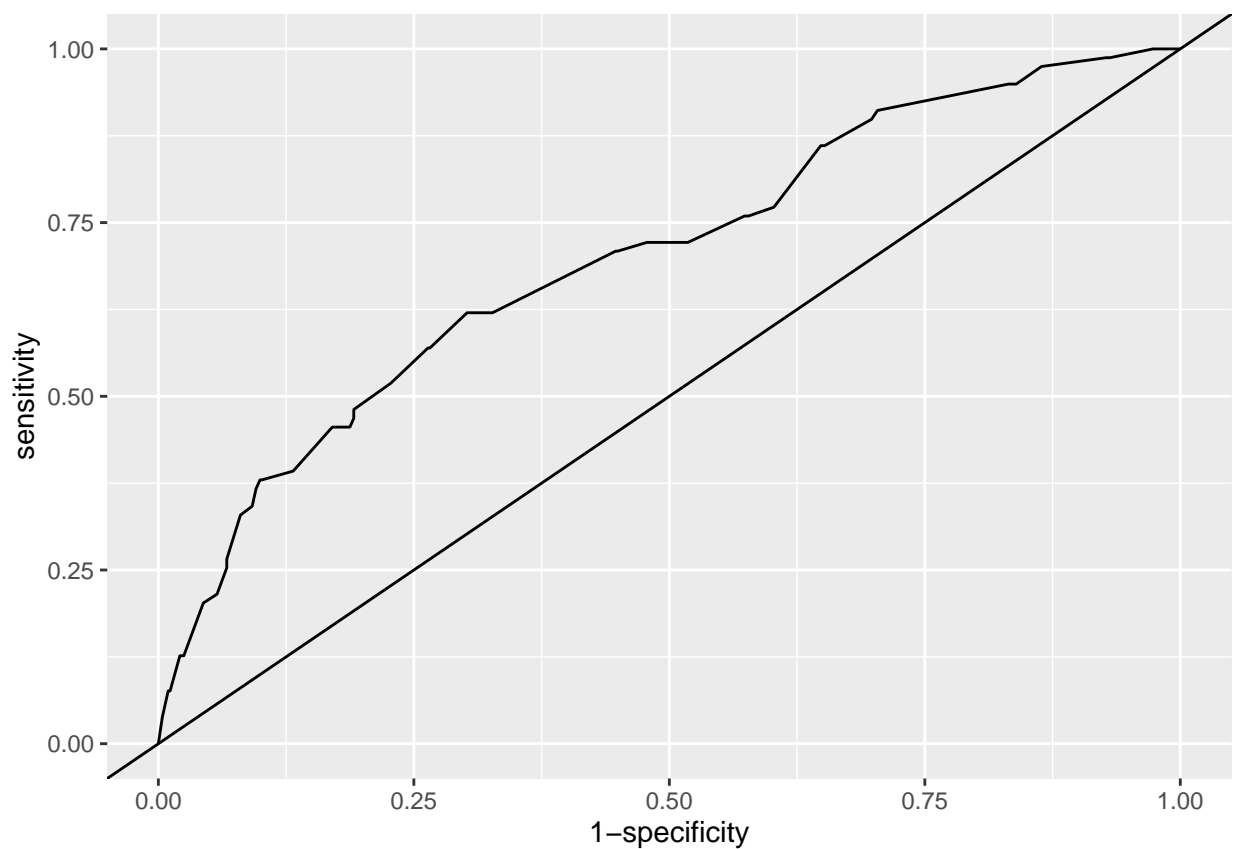


Figure 6.7: ROC curve for our logistic regression model of the indo RCT data (solid line). The line  $y=x$  shows the ROC curve we'd expect with random guessing.

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

In this section 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.

```
## 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")
```

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.

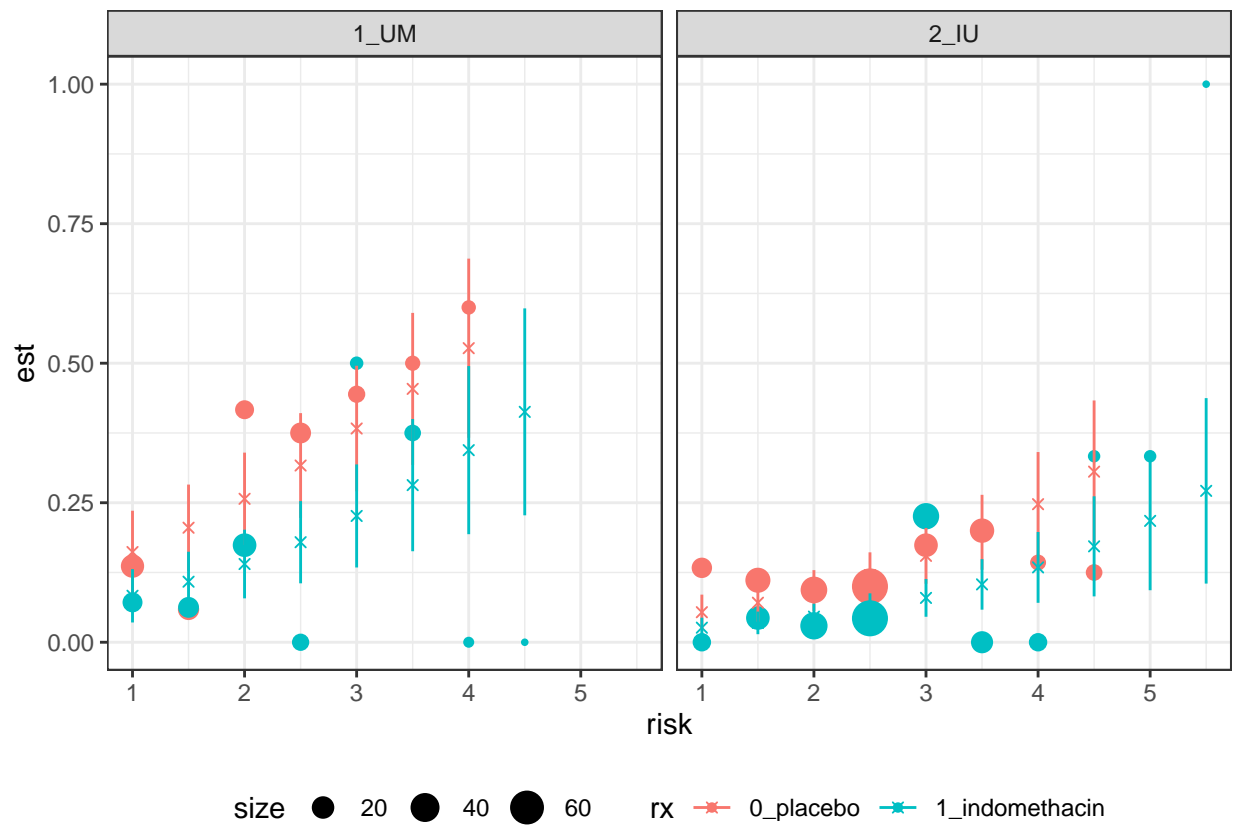


Figure 6.8: Calibration-based plots for indo RCT data for sites 1 (left) and 2 (right).

# 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. (2003). *Modelling Binary Data*. 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.
- 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.
- Goldacre, B. (2012). *Bad Pharma: How Medicine is Broken, And How We Can Fix It*. HarperCollins Publishers.
- 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.
- 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.
- 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.
- Taves, D. R. (1974). Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, 15(5):443–453.
- 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.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.
- Zhong, B. (2009). How to calculate sample size in randomized controlled trial? *Journal of thoracic disease*, 1(1):51.