

# Clinical Trials 4H

Rachel Oughton

2024-01-22

# Contents

<b>Welcome to Clinical Trials 4H!</b>	<b>3</b>
Practical details . . . . .	3
What to expect from this module . . . . .	4
<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 . . . . .	9
1.3 The primary outcome . . . . .	10
1.4 Ethical issues . . . . .	10
1.5 Phases of clinical trials . . . . .	11
<b>I Part I: Continuous outcome variables</b>	<b>13</b>
<b>2 Sample size for a normally distributed primary outcome variable</b>	<b>14</b>
2.1 The treatment effect . . . . .	14
2.2 Reminder: hypothesis tests (with a focus on RCTs) . . . . .	15
2.3 Constructing a measure of effect size . . . . .	19
2.4 Power: If $H_0$ is false . . . . .	21
2.5 A sample size formula . . . . .	25
<b>3 Allocation</b>	<b>27</b>
3.1 Bias . . . . .	27
3.2 Allocation methods . . . . .	32
3.3 Incorporating baseline measurements . . . . .	42
3.4 Stratified sampling . . . . .	46
3.5 Minimization . . . . .	46
3.6 Problems around allocation . . . . .	49

# Welcome to Clinical Trials 4H!

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

## Practical details

### Lectures

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

### Computer classes

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



## Assessment

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

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

## Books

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

## What to expect from this module

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

- It is more focussed on application than on methodology

- It is assessed purely through coursework.

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

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

## **What I expect from you**

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

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

# Chapter 1

## Introduction to Clinical Trials

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

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

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

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

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

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

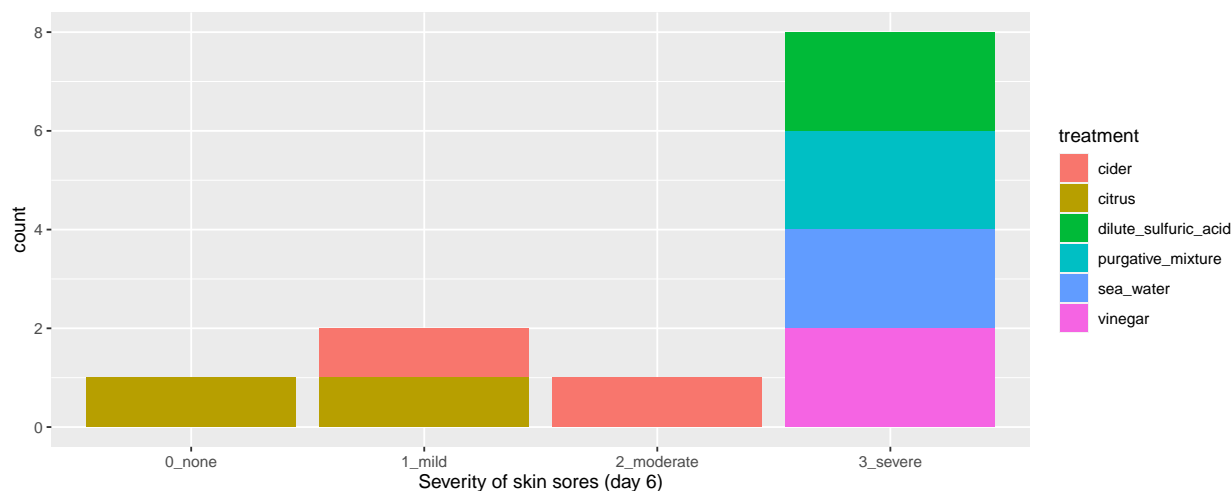
Put (very!) simply, the goal of a clinical trial is to determine what works to make people better. Although clinical trials as we know them now have only been around since the Second World War, similar sorts of experiments can be seen from much longer ago. If you’re interested in learning about the evolution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

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

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

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

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

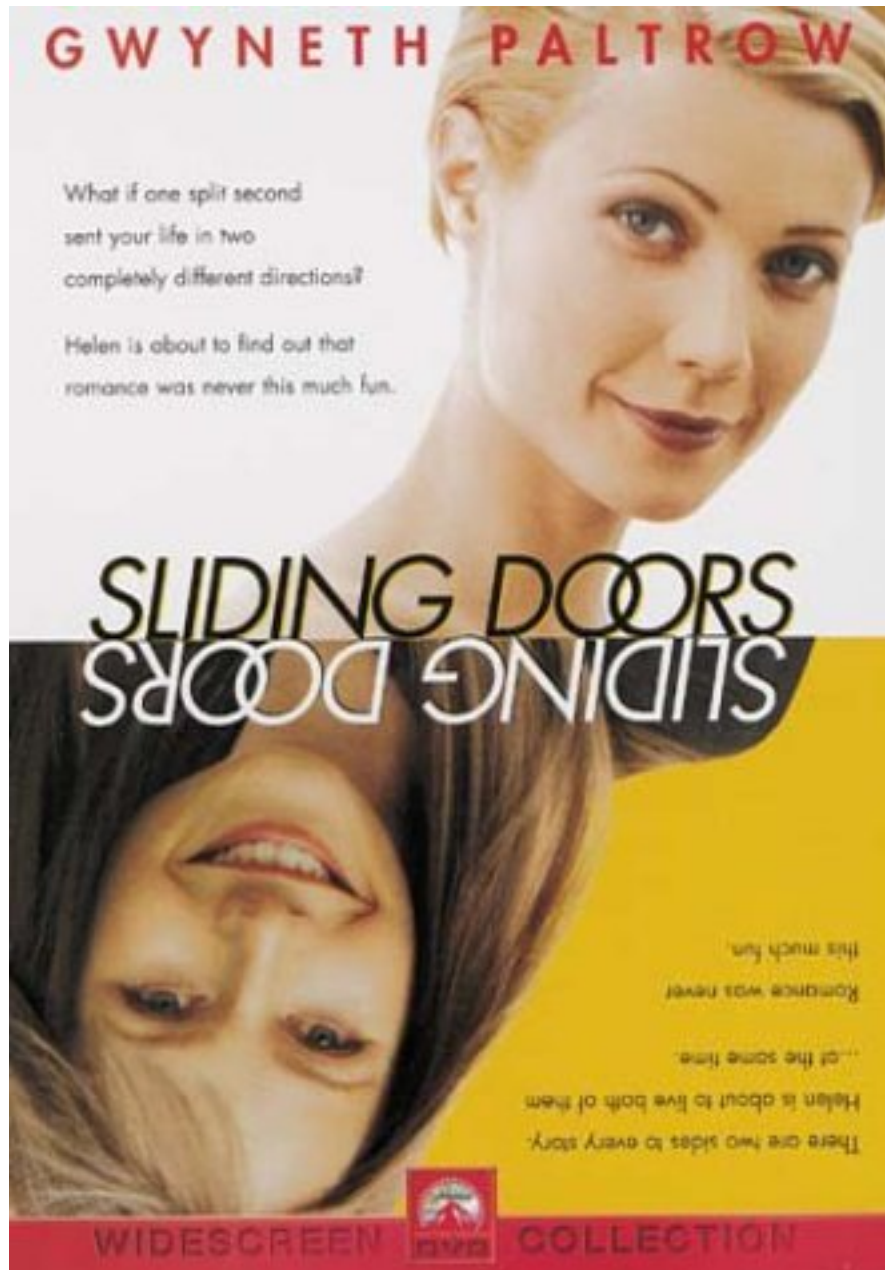


A key thing to notice about the Scurvy example is that Lind went to great lengths to ensure that the treatment was the only thing affecting these 12 sailors: they all started with a similar severity of symptoms, they were kept in the same place and their diet was identical apart from their treatment.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

## 1.1 Causal inference and clinical trials

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

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



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

Because, as we noted, people are all different, and living different lives (and unlike James Lind we can't force them all to live in the same part of a ship and eat the same food!) we will need to test the treatment on lots of people in order to gather empirical evidence. This is why statistics is so important in the design and analysis of clinical trials. The results of the trial must 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.

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

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

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

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

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

### 2.1 The treatment effect

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

**Definition 2.1.** Suppose our primary outcome variable is  $X$ , which has mean  $\mu$  in the control group and mean  $\mu + \tau$  in 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 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 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 sufficiently large to be sufficiently likely to detect a clinically meaningful treatment effect.

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

### 2.2.2 One-tailed or two-tailed?

It is highly likely that the scientists running the trial will have a strong idea of the likely ‘direction’ of the treatment effect. Assuming that a larger value of the primary outcome variable  $X$  is good, they will expect a positive value of the treatment effect  $\tau$  (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

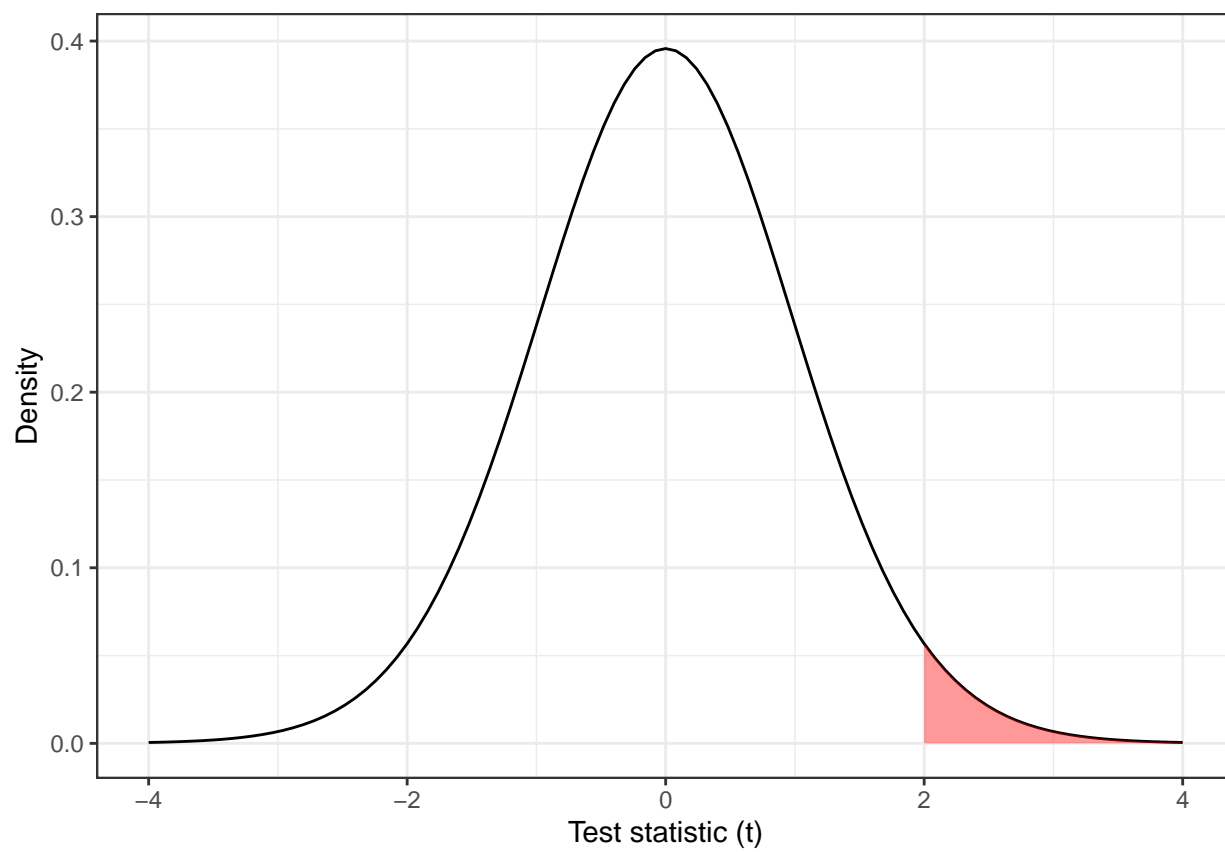


Figure 2.1: The distribution  $t_{31}$ , with the area corresponding to  $t > 2$  shaded.



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

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

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

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

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

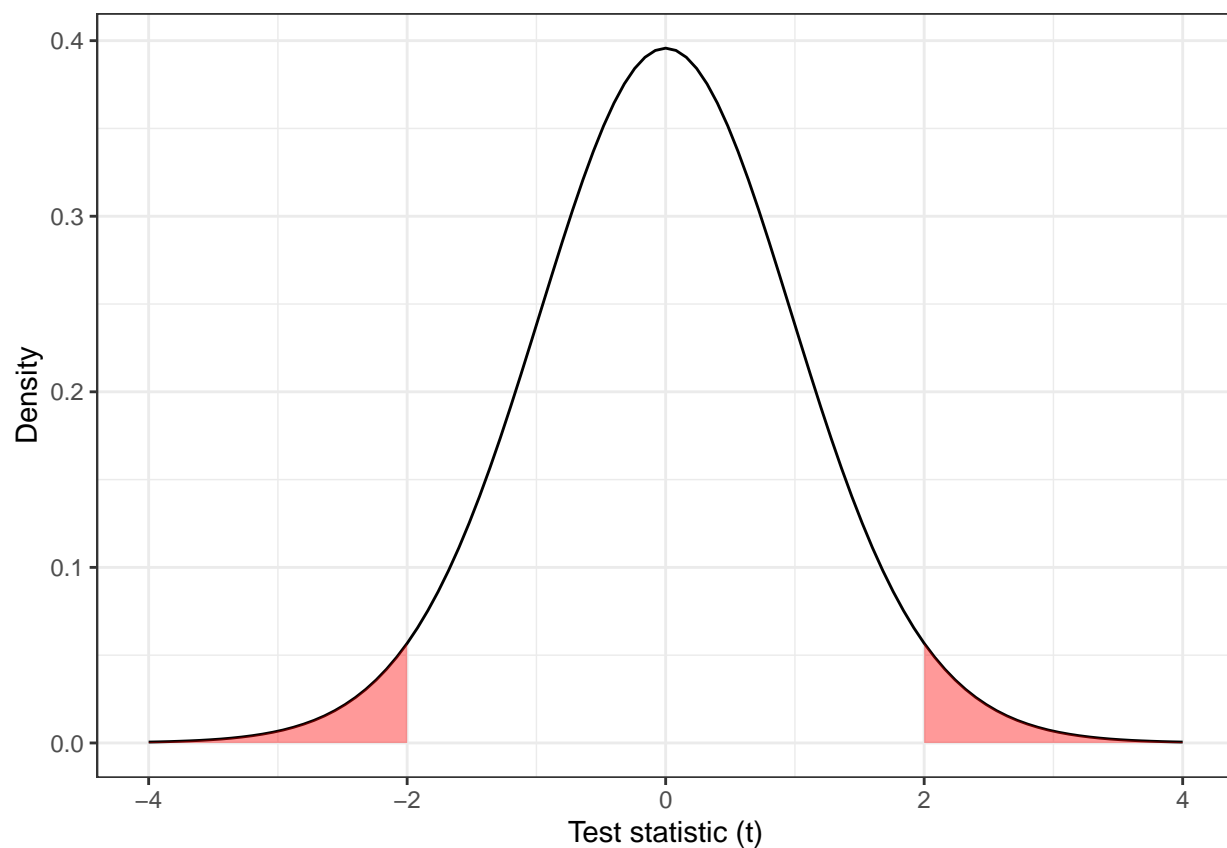


Figure 2.2: The distribution  $t_{31}$ , with the area corresponding to  $|t| > 2$  shaded.

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

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

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



Figure 2.3: A rather ubiquitous two-tailed mermaid

## 2.3 Constructing a measure of effect size

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

Suppose that we have  $n$  patients in group  $C$ , and  $m$  in group  $T$ . The primary outcome variable  $X$  is normally distributed with mean  $\mu$  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  $H_0$ ,  $D$  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  $\alpha = 0.05$  have  $z_{\alpha/2} = 1.960$ , since the probability of a standard normal variable exceeding this value is 0.025.

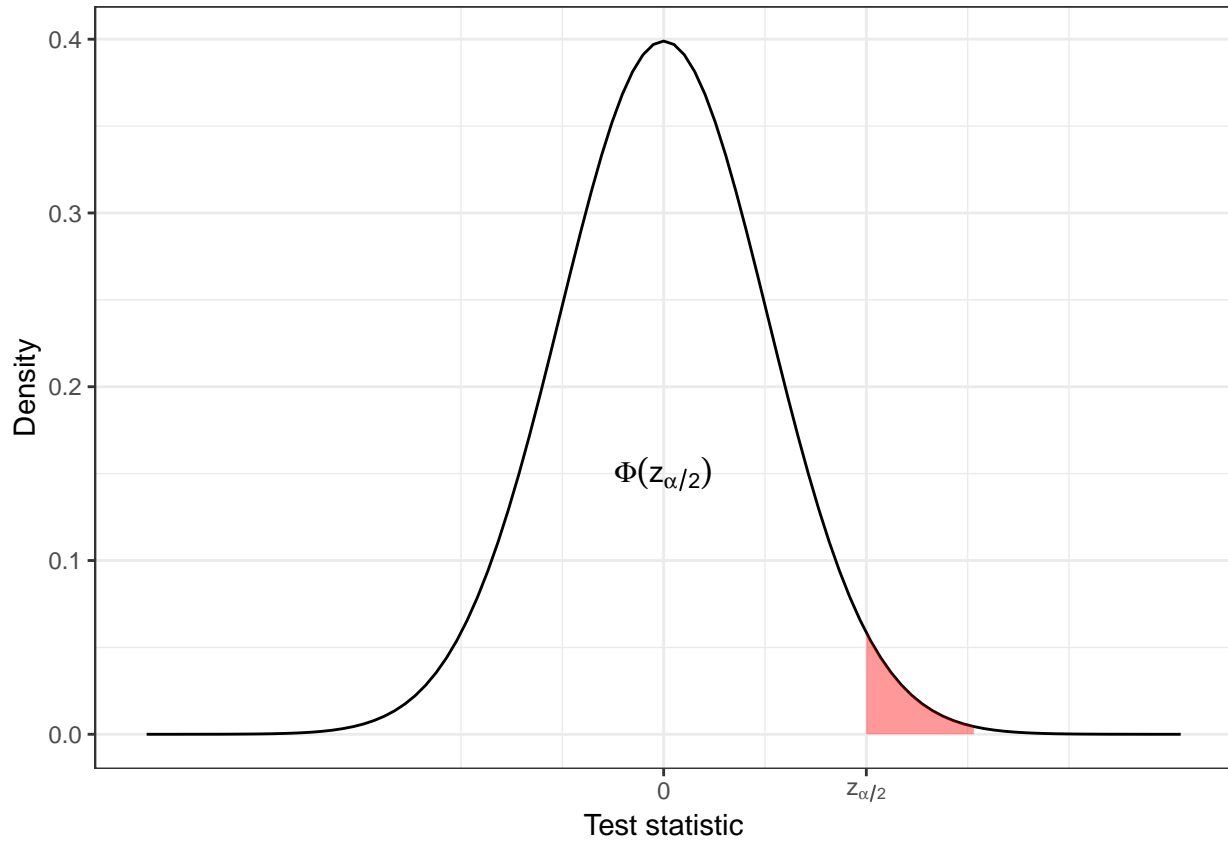
So, if we have run a trial, and have obtained  $n$  values of  $X$  from group  $C$  and  $m$  values of  $X$  from group  $T$ , we can compute  $D$ . If  $D$  lies outside the interval  $[-z_{\alpha/2}, z_{\alpha/2}]$  then we reject  $H_0$ .

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

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

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

If you find the notation confusing (as I do!) then it might be helpful to remember that both  $\alpha$  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 2.4 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  $H_0$ .

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 accept  $H_1$ ) is the power,  $1 - \beta$ .

We can see that if the distributions have better separation, as in Figure 2.5, the power becomes greater. This can be as a result of a larger  $\tau$ , 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 2.4 and 2.5 this is the region to the right of the interval for which we fail to reject  $H_0$ , ie. where

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

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

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

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

Figure 2.6 shows the power function  $\Psi(\tau)$  for  $\tau$  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.

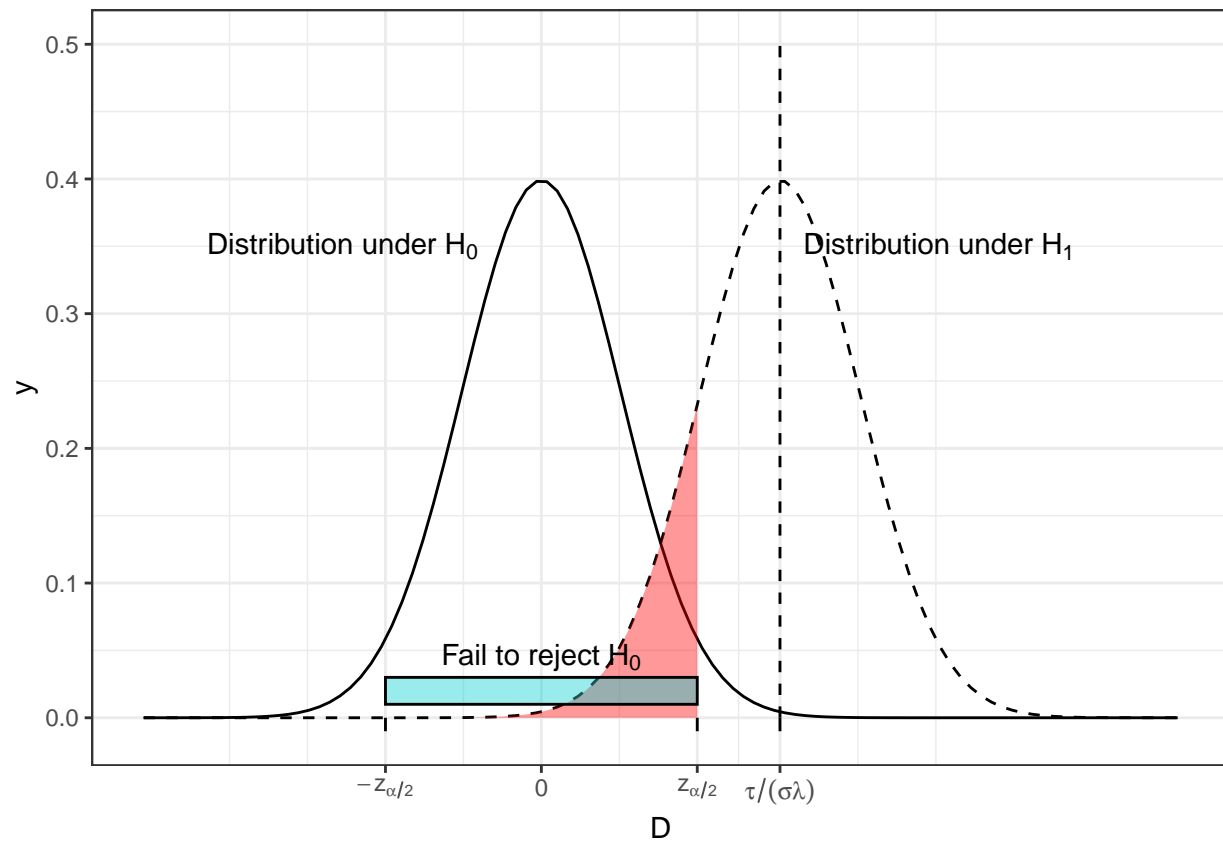


Figure 2.4: The distribution of  $D$  under both  $H_0$  and  $H_1$  for some arbitrary values of effect size, population variance,  $n$  and  $m$ , with the region in which we fail to reject  $H_0$  shown by the turquoise bar and the red shading.





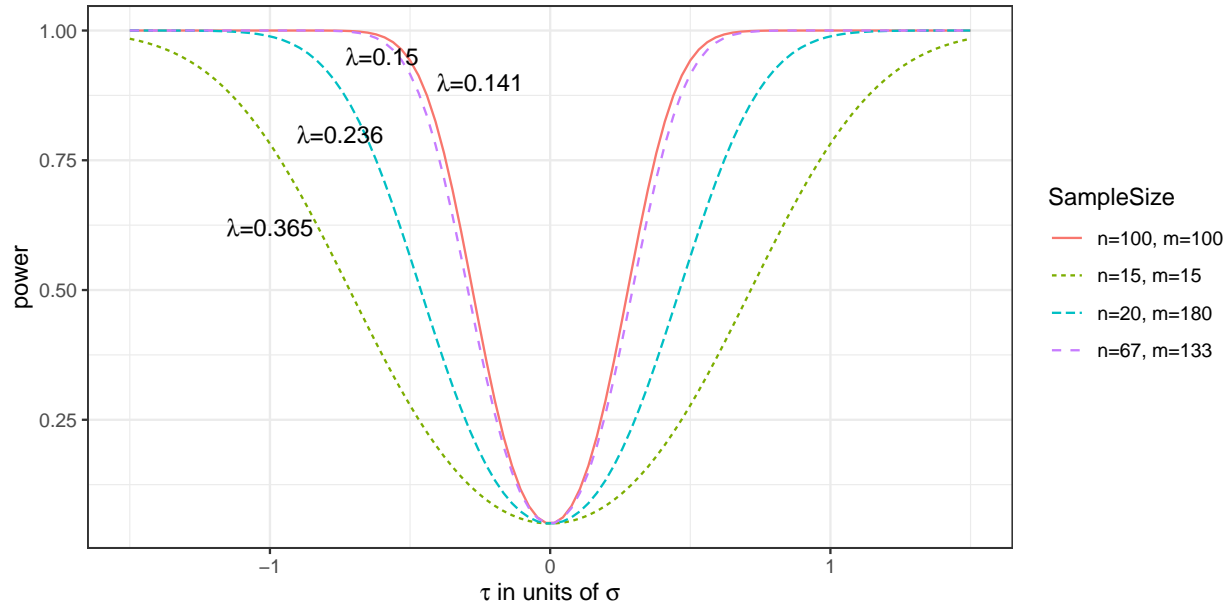


Figure 2.6: Power curves for various values of  $n$  and  $m$ , with effect size in units of standard deviation, given a type I error rate of 0.05.

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

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.

## 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 variables,  $\tau_\sigma$  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?

In a medical setting, an estimate of  $\sigma$  is usually available, and so we will return to thinking in terms of  $\tau$  and  $\sigma$ . In this equation, the value we use (or find) for  $\tau$  is 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, and they've been recruited, we next need to determine which participants should be assigned to which 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  $\sigma^2$ .

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.

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

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

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

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

### Allocation bias

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

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

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

### Assessment bias

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

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

### Slight aside: publication bias

In most areas of science, including clinical trials, the ultimate aim is to affect practice. This is usually done by publishing a write-up of the trial, including its design, methods, analysis and results, and publishing that in a [medical] journal. These are peer-reviewed, which means that experts from the relevant field are 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.

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

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

### 3.1.2 Implications for allocation

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

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

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

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

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

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

Point 2 is important to avoid biases that come through the assignment of a particular patient being known either in advance or after the fact. There are some approaches that ‘pass’ the first point, but fail at the second. As well as strict alternation ( $ABABAB\dots$ ), 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, effectively, used in many large trials, but for small trials it can be statistically problematic. The main reason for this is chance imbalance of group sizes.

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



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

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

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

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.

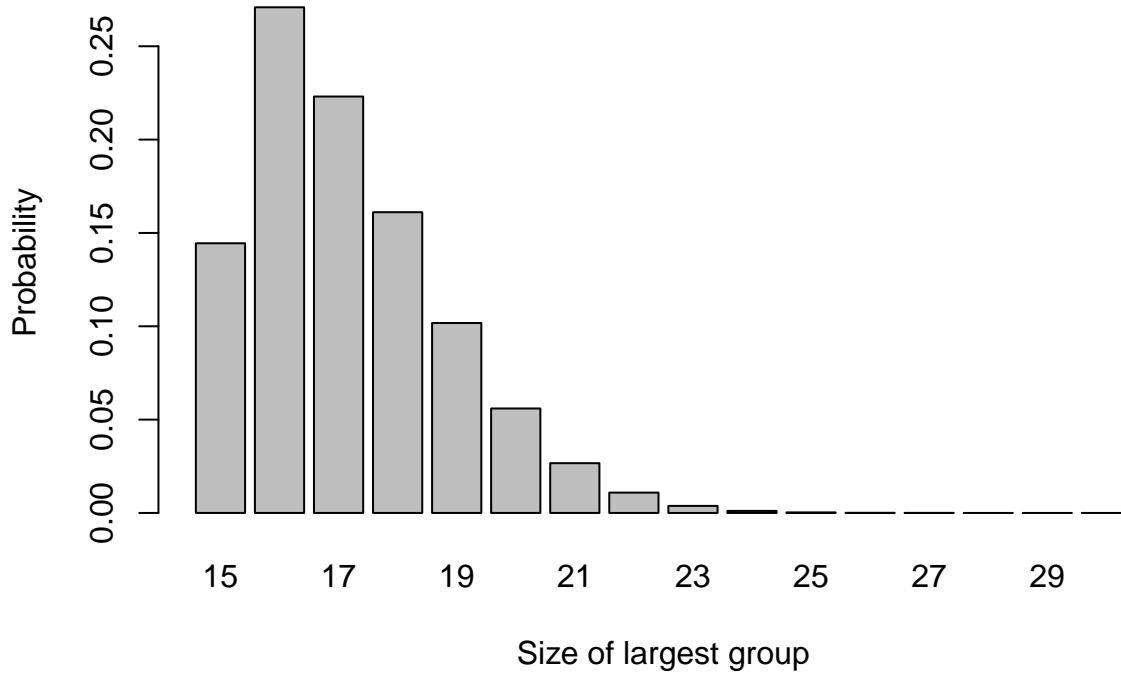


Figure 3.1: The probability distribution of largest group size for  $n=15$ .

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

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

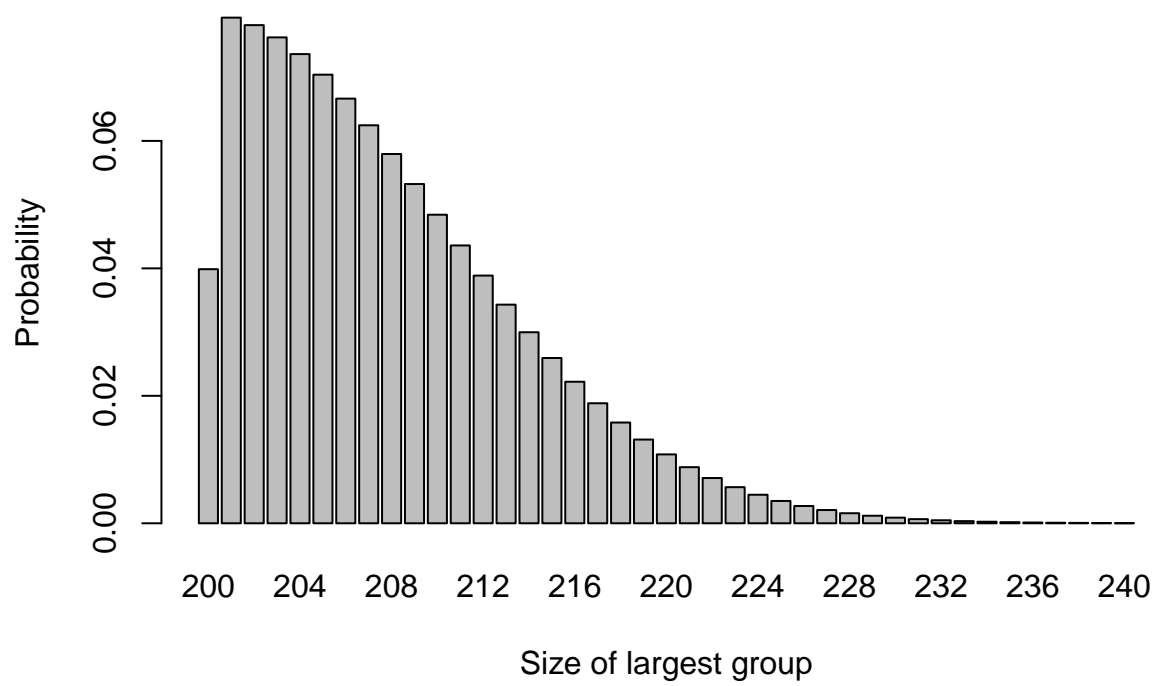


Figure 3.2: The probability distribution of largest group size for  $n=200$ .

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

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

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

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

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

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

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

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

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

So, as we saw when looking at power, we don't lose too much if the group sizes are 2:1, but a more pronounced 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 too much imbalance is to use *random permuted blocks* (RPBs). If the blocks have size  $2m$ , and there are two groups then there are

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

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

If we have two groups,  $A$  and  $B$ , then there are six *blocks* of length 4 containing two  $A$ 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 AAB B ABBA BBAA* ...

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

### 3.2.2.1 RPBs with random block length

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

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

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

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

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

### 3.2.3 Biased coin designs and urn schemes

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

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

#### 3.2.3.1 Biased coin designs

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

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

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

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

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

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

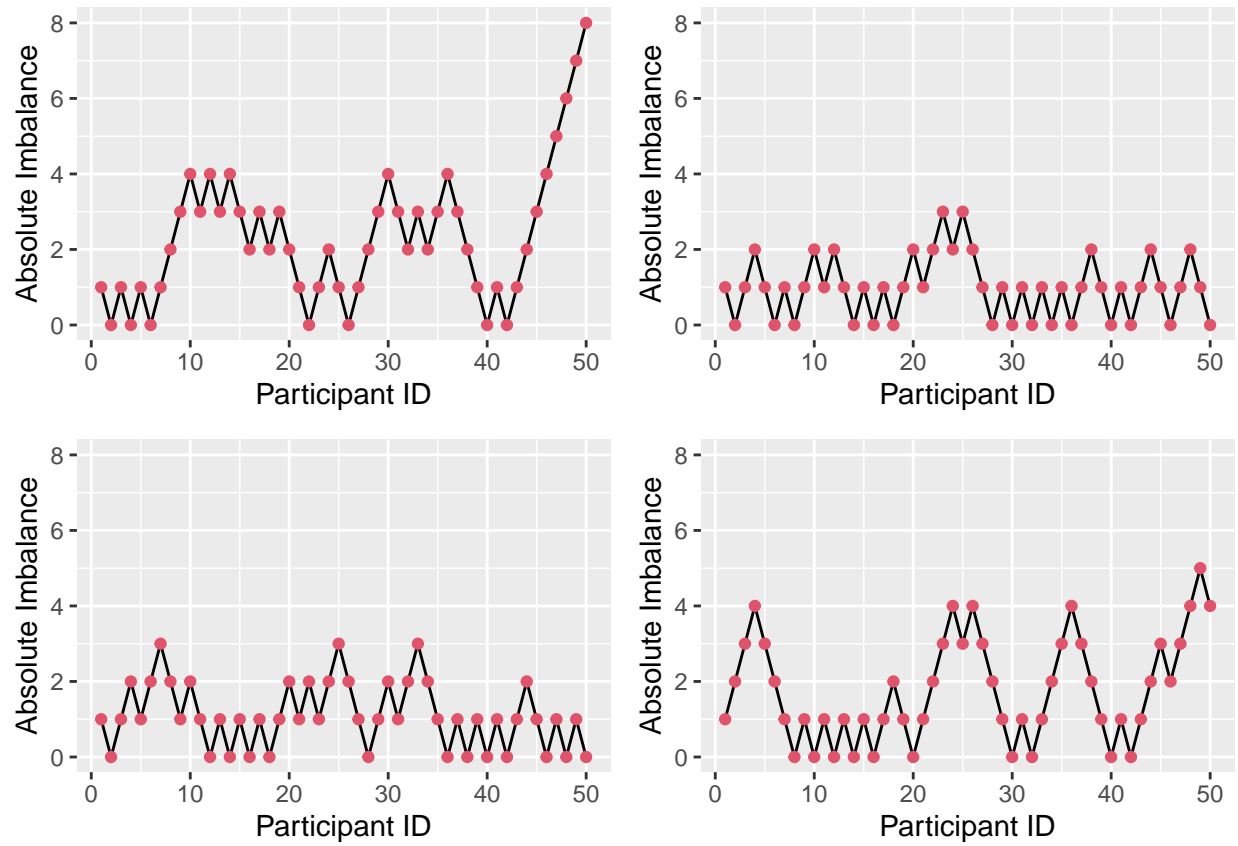


Figure 3.3: Absolute imbalance for a biased-coin scheme with  $P = 0.667$ .

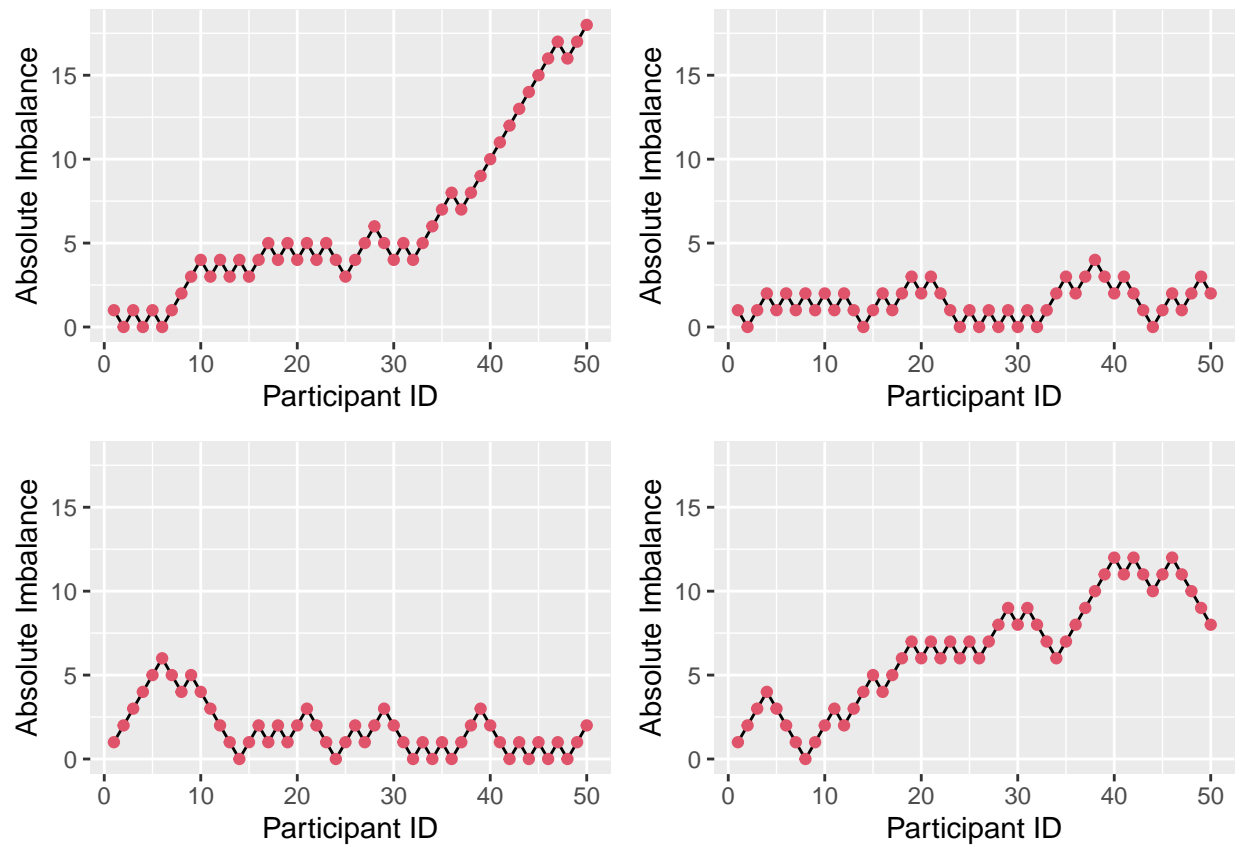


Figure 3.4: Absolute imbalance for a biased-coin scheme with  $P = 0.55$ .

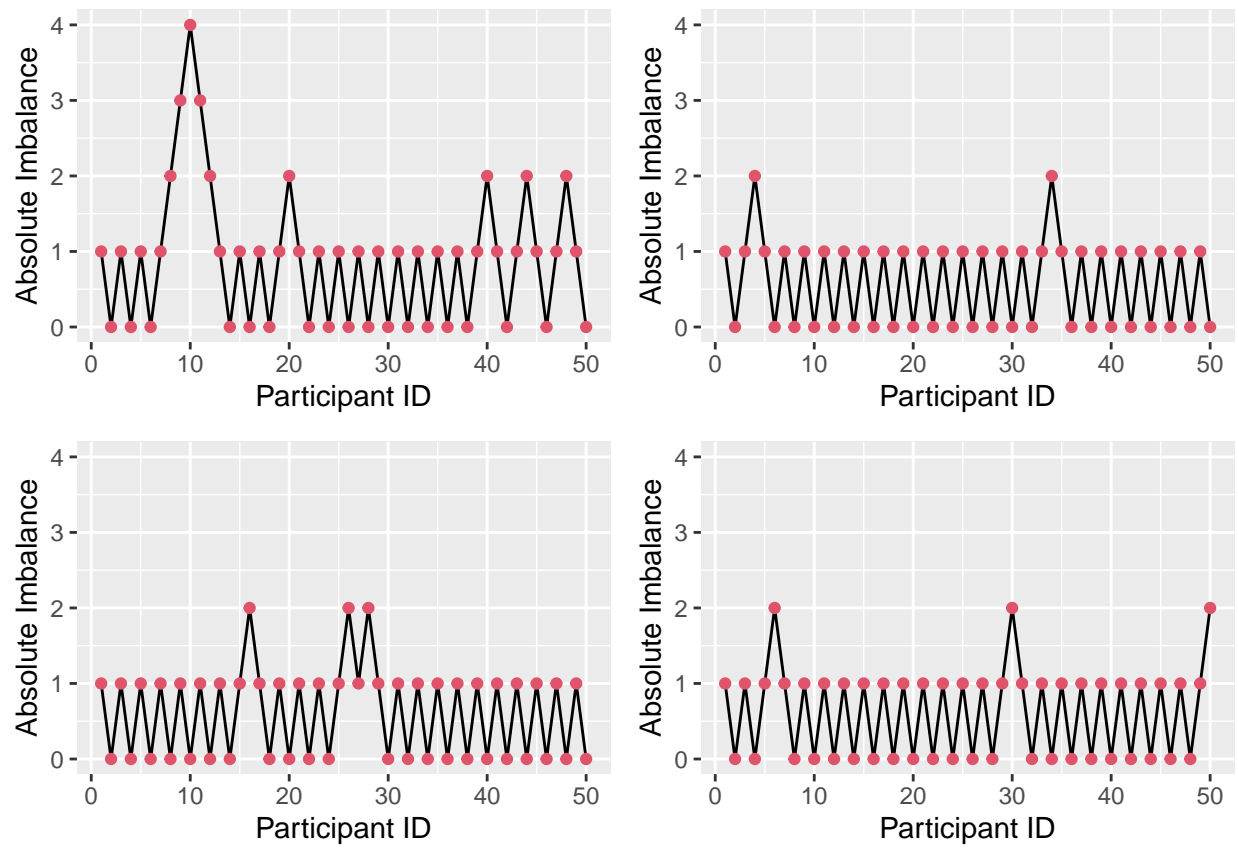


Figure 3.5: Absolute imbalance for a biased-coin scheme with  $P = 0.9$ .



Efron's suggestion for implementation was that each clinician would 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 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

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

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.

### 3.3 Incorporating baseline measurements

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

Such baseline measurements can be used in allocation.

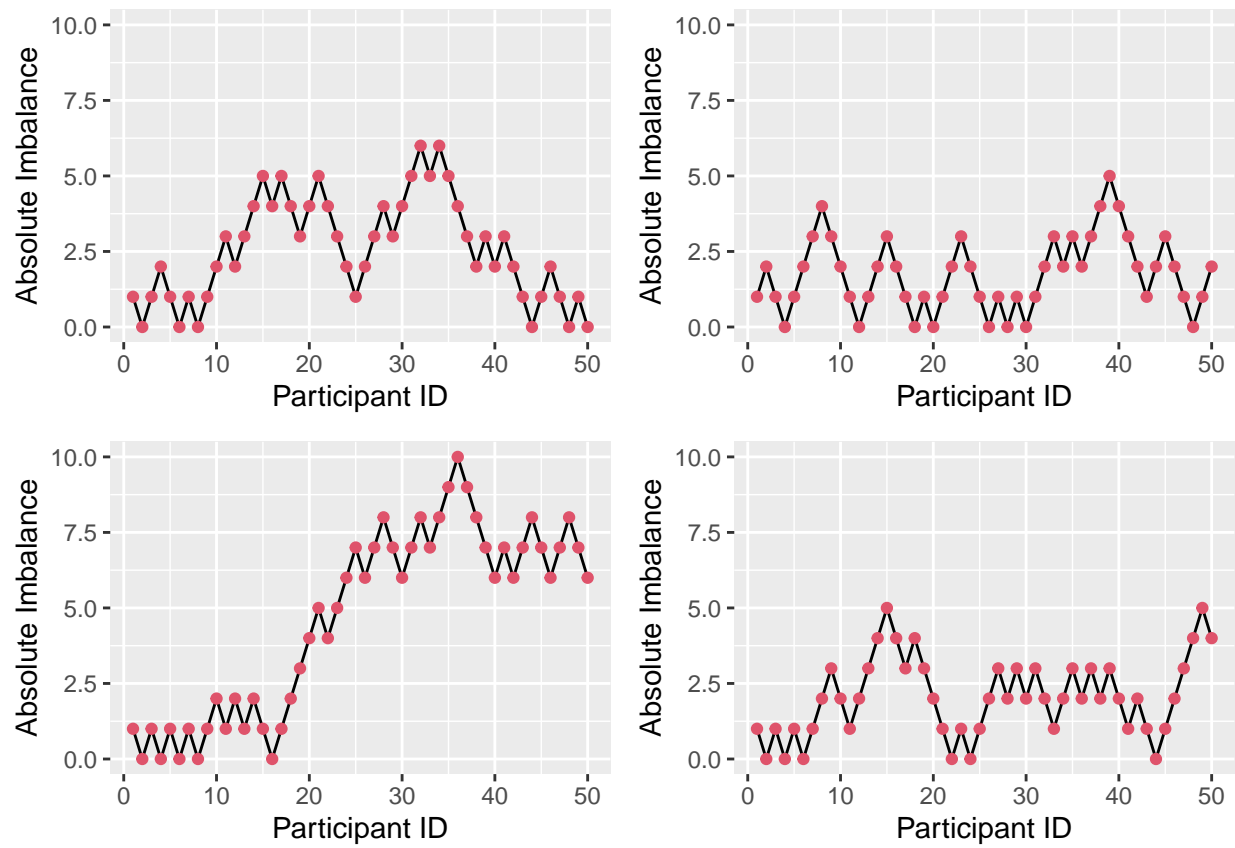


Figure 3.6: Four realisations of absolute imbalance for  $r=1$ ,  $s=1$ ,  $N=50$ .

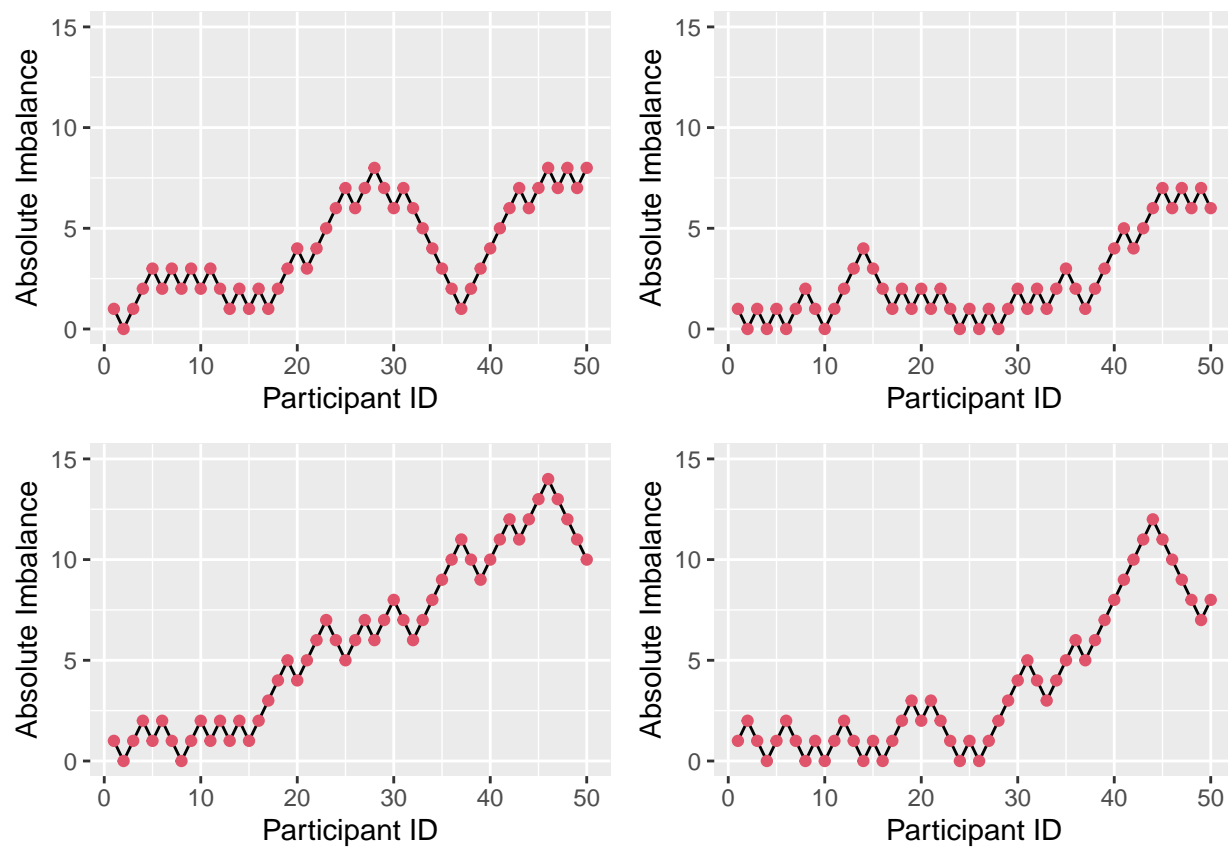


Figure 3.7: Four realisations of absolute imbalance for  $r=1$ ,  $s=8$ ,  $N=50$ .

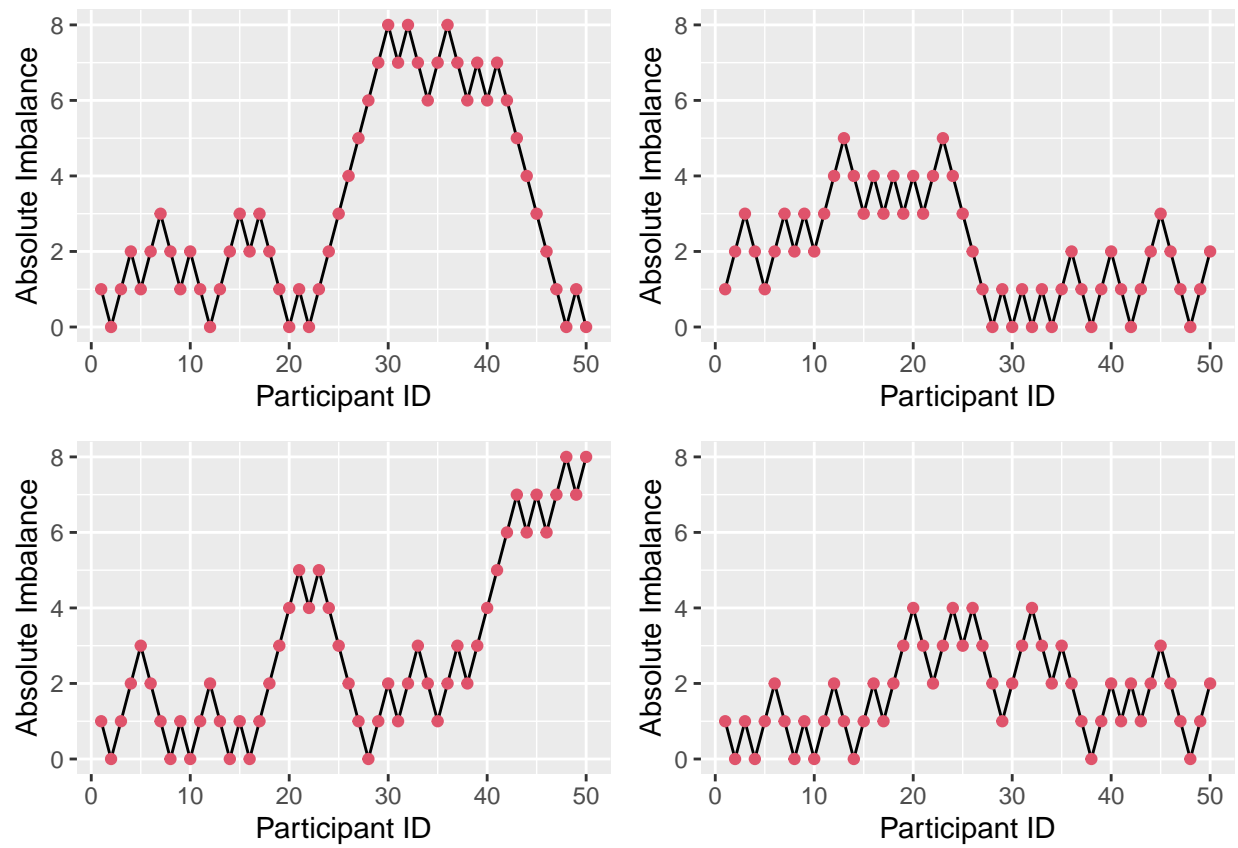


Figure 3.8: Four realisations of absolute imbalance for  $r=8$ ,  $s=1$ ,  $N=50$ .

### 3.4 Stratified sampling

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

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

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

	Male	Female
<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).

### 3.5 Minimization

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

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

**Example 3.5.** The study by Kallis et al. (1994) investigates the effect of giving aspirin to patients before coronary artery surgery with giving them a placebo. Interestingly, the effects of aspirin were found to be both positive (decreases platelet aggregation to arachidonic acid and to collagen) and negative (increased likelihood of post-operative excessive blood loss). For their prognostic factors, Kallis et al. (1994) chose age ( $\leq 50$  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.

### 3.5.1 Minimization algorithm

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

Suppose there are several prognostic factors over which we require balance, and that these factors have  $I, J, K, \dots$  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

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

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

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

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

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

In order to implement minimisation, we follow these steps:

1. Allocate the first patient by simple randomisation.
2. Suppose that at some point in the trial we have recruited  $n_{ijkl}$  patients with prognostic factors  $i, j, k, l$ . Of these  $n_{ijkl}^A$  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

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

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

Some people set  $P = 1$ , whereas others would set  $\frac{1}{2} < P < 1$  to retain some randomness. Although setting  $P = 1$  makes the system deterministic, to predict the next allocation a doctor (or whoever) would need to know  $n_{i+++}^A$  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 3.6.** From Altman (1990) (citing Fentiman et al. (1983)). In this trial, 46 patients with breast cancer were allocated to receive either Mustine (arm A) or Talc (arm B) as treatment for pleural effusions (fluid between the walls of the lung). They used four prognostic factors: age ( $\leq 50$  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 Table 3.1.

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

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

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.



## 3.6 Problems around allocation

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

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

Variable	Licorice (N = 118)	Sugar-water (N = 117)	Standardized difference <sup>a</sup>
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	

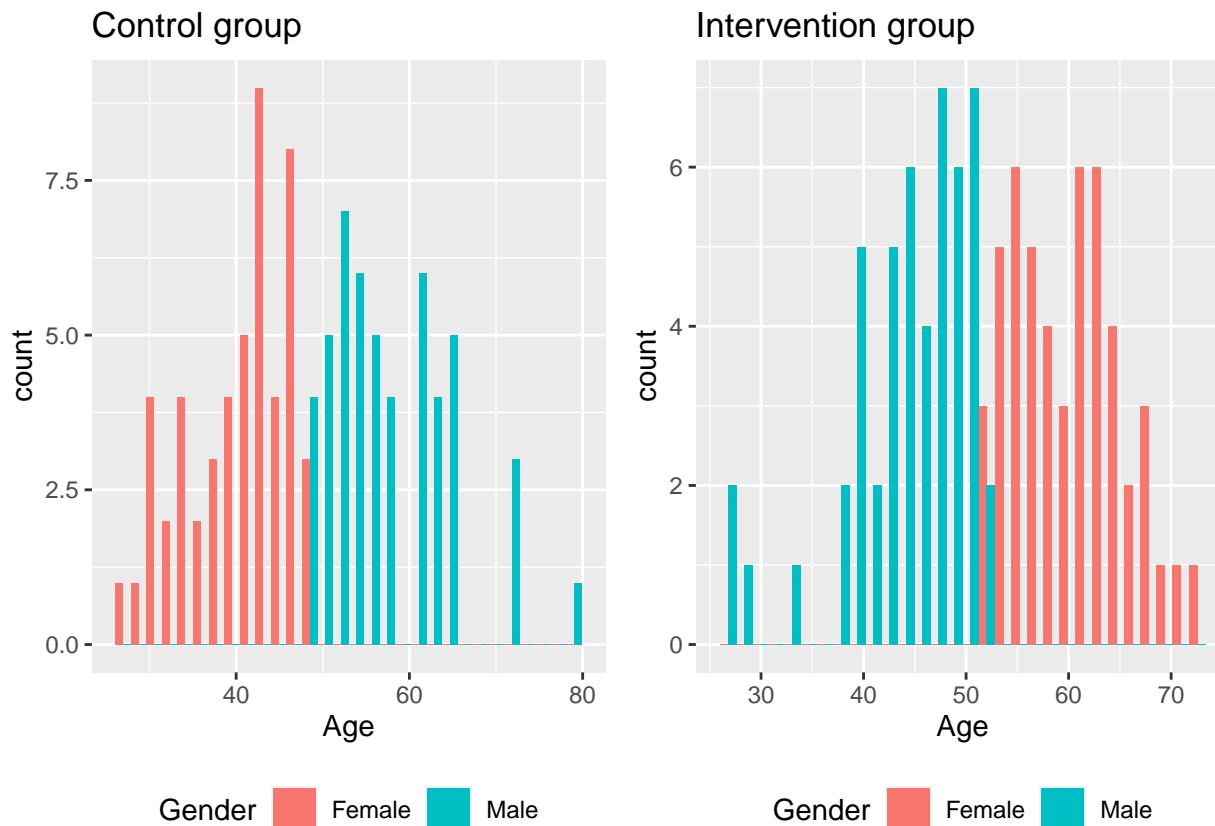
Summary statistics presented as percent of patients or mean ± SD.  
<sup>a</sup>Standardized difference (licorice – sugar-water) defined as the difference in means or proportions divided by the pooled standard deviation; >0.2 in absolute value indicates imbalance.  
<sup>b</sup>Surgery size: small (thoracoscopy); 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.

# Bibliography

- Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.
- Altman, D. G. and Bland, J. M. (1999). Treatment allocation in controlled trials: why randomise? *Bmj*, 318(7192):1209–1209.
- 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.
- Fentiman, I. S., Rubens, R. D., and Hayward, J. L. (1983). Control of pleural effusions in patients with breast cancer a randomized trial. *Cancer*, 52(4):737–739.
- Freedman, L. and White, S. J. (1976). On the use of pocock and simon’s method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics*, pages 691–694.
- Hayes, R. J. and Moulton, L. H. (2017). *Cluster randomised trials*. CRC press.
- Hjalmas, H. and Hellstrom, K. (1998). Long-term treatment with desmopressin in children with primary monosymptomatic nocturnal enuresis: an open multicentre study. *British Journal of Urology*.
- 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.
- Matthews, J. N. (2006). *Introduction to randomized controlled clinical trials*. CRC Press.
- 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.
- 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.
- Zhong, B. (2009). How to calculate sample size in randomized controlled trial? *Journal of thoracic disease*, 1(1):51.