

# Clinical Trials 4H

Rachel Oughton

2024-01-08

# 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 . . . . .	24

# 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.

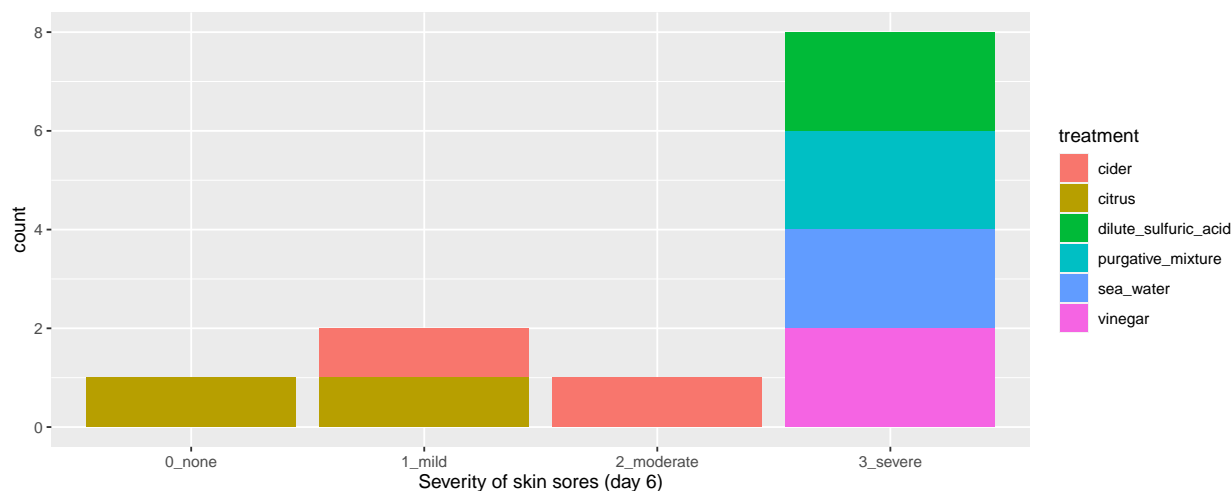
Put (very!) simply, the goal of a clinical trial is to determine what works to make people better. Although clinical trials as we know them now have only been around since the Second World War, similar sorts of experiments can be seen from much longer ago. If you’re interested in learning about the evolution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

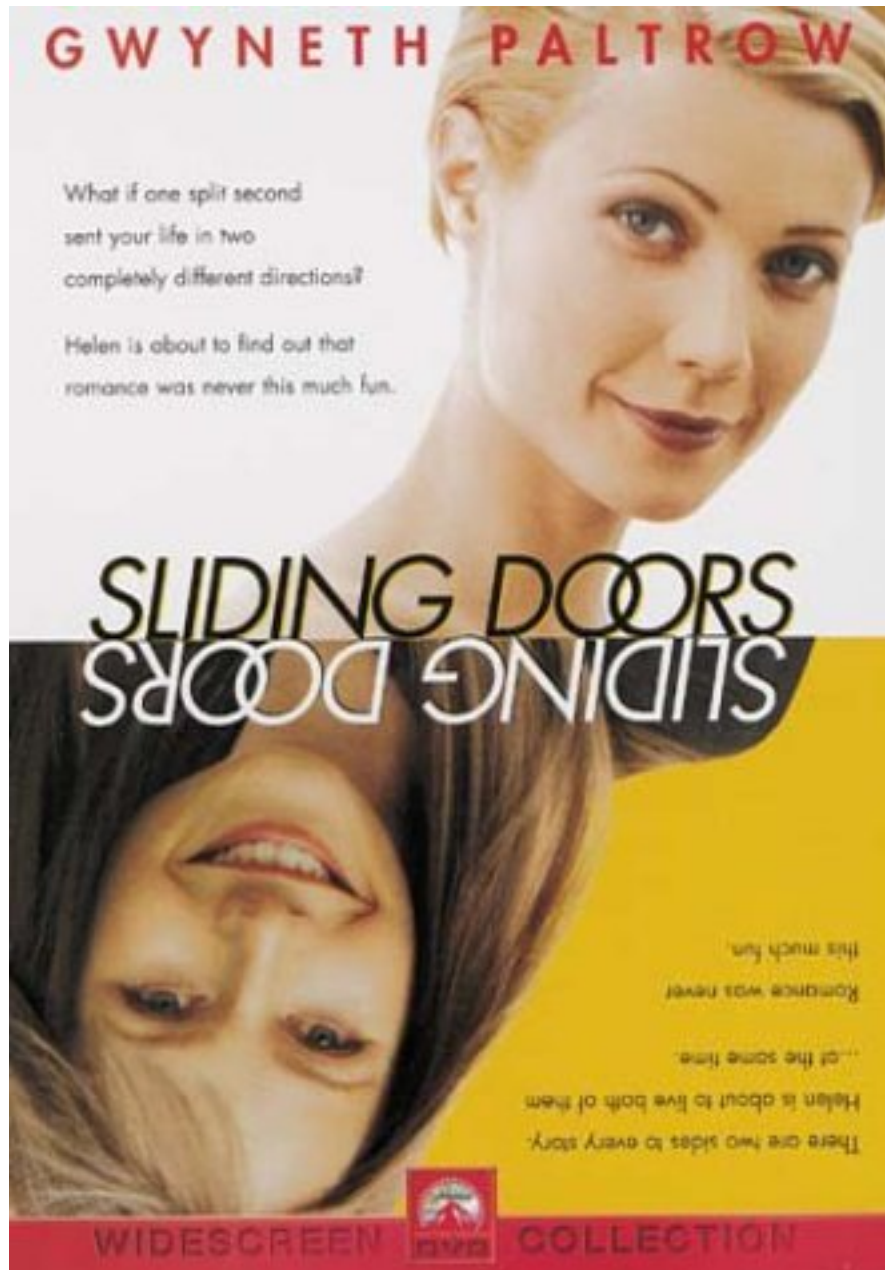


A key thing to notice about the Scurvy example is that Lind went to great lengths to ensure that the treatment was the only thing affecting these 12 sailors: they all started with a similar severity of symptoms, they were kept in the same place and their diet was identical apart from their treatment.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

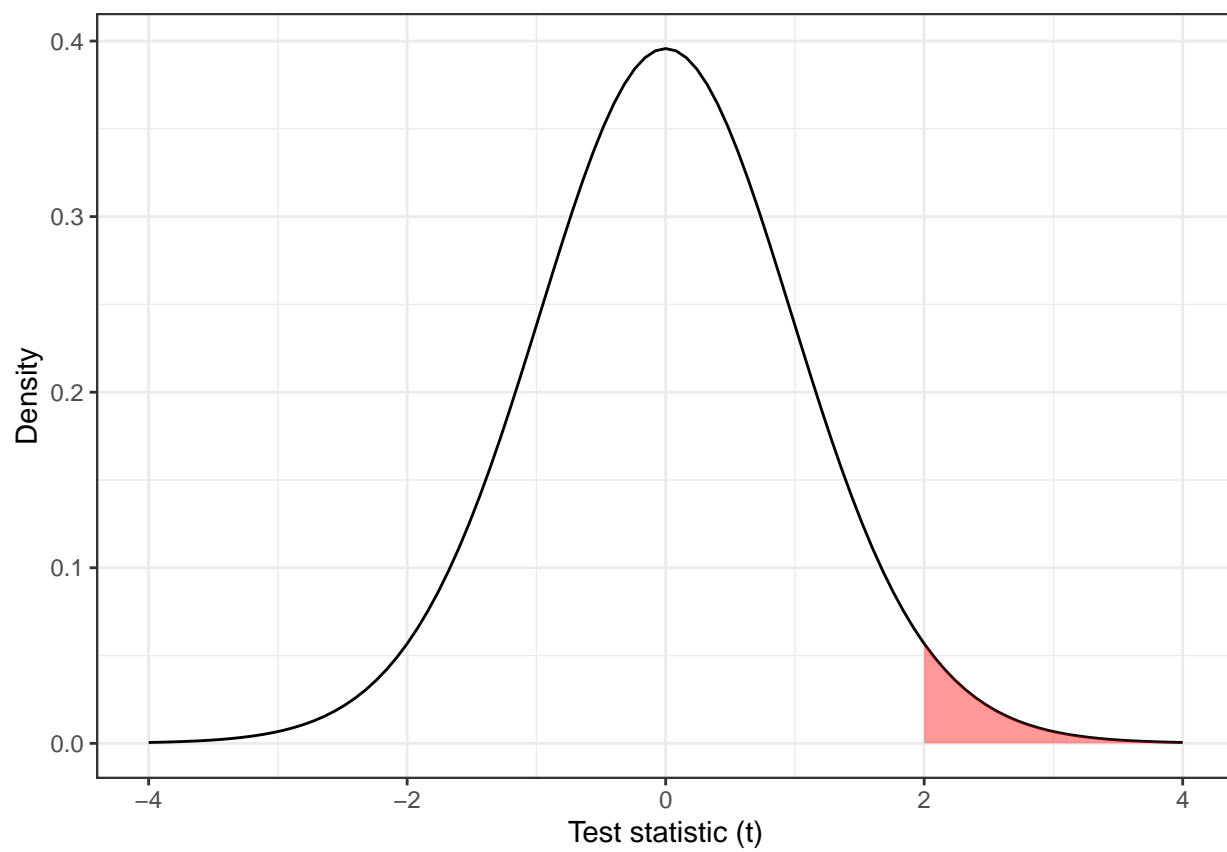


Figure 2.1: The distribution  $t_{31}$ , with the area corresponding to  $t > 2$  shaded.



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

$$H_0 : \tau = 0$$

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

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

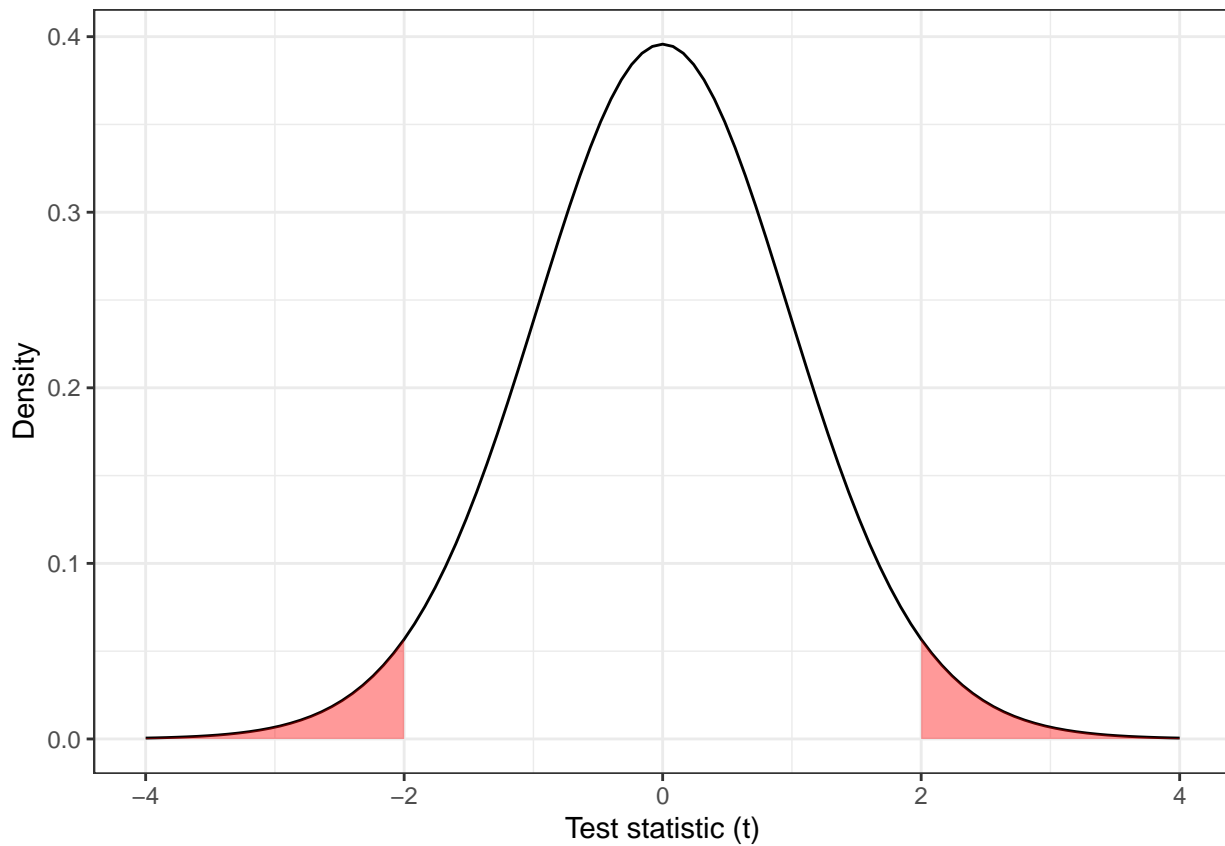


Figure 2.2: The distribution  $t_{31}$ , with the area corresponding to  $|t| > 2$  shaded.

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

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

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



Figure 2.3: A rather ubiquitous two-tailed mermaid

## 2.3 Constructing a measure of effect size

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

Suppose that we have  $n$  patients in group  $C$ , and  $m$  in group  $T$ . The primary outcome variable  $X$  is normally distributed with mean  $\mu$  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.

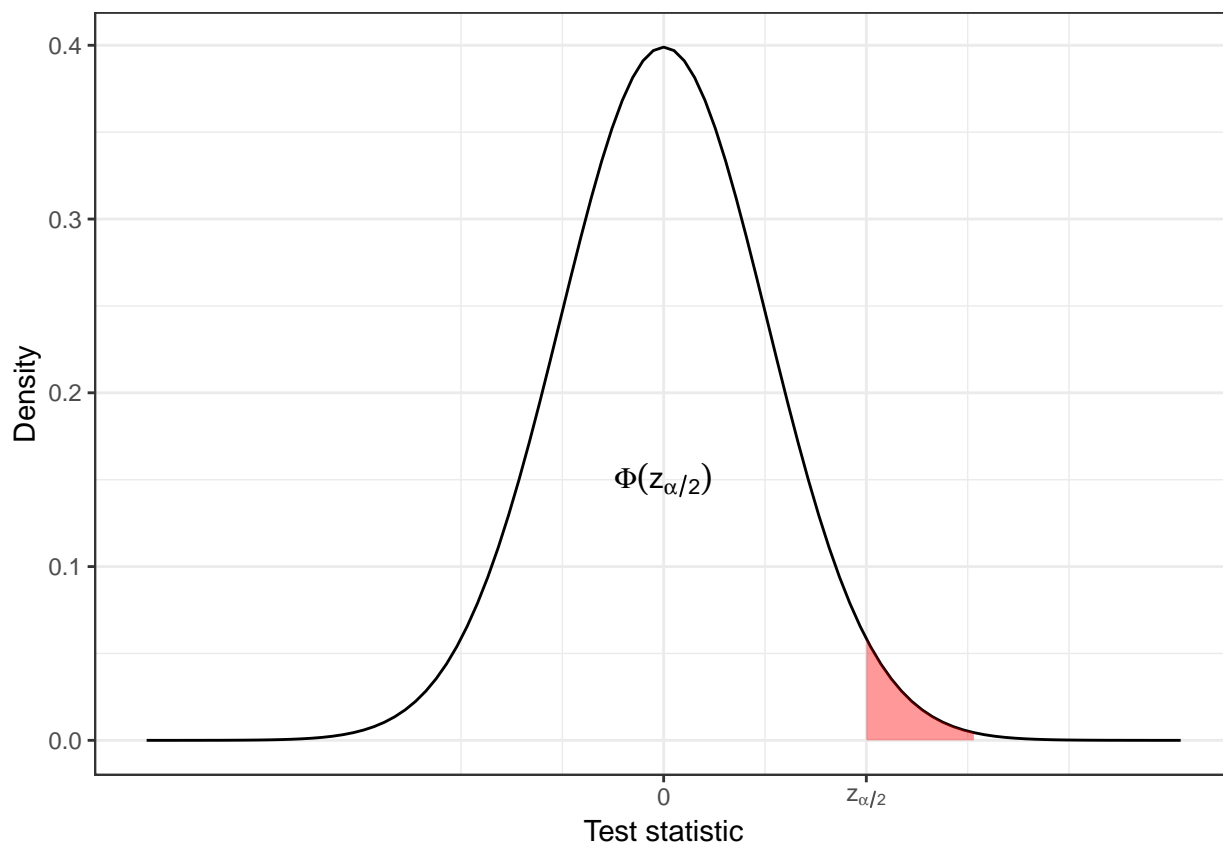
So, if we have run a trial, and have obtained  $n$  values of  $X$  from group  $C$  and  $m$  values of  $X$  from group  $T$ , we can compute  $D$ . If  $D$  lies outside the interval  $[-z_{\alpha/2}, z_{\alpha/2}]$  then we reject  $H_0$ .

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.

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  $\Pr(\text{Accept } H_0 \mid \tau \neq 0)$ , 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  $H_0$ .

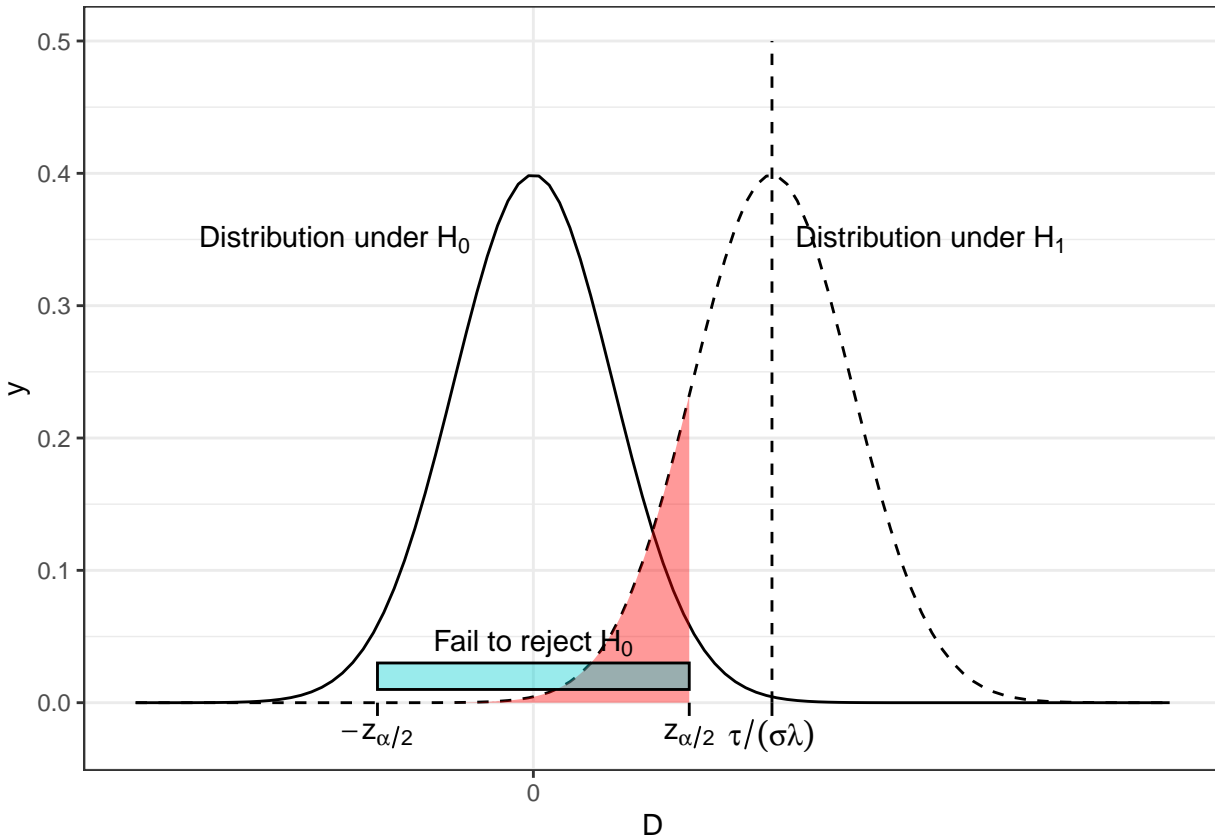


Figure 2.4: The distribution of  $D$  under both  $H_0$  and  $H_1$  for some arbitrary values of effect size, population variance,  $n$  and  $m$ , with the region in which we fail to reject  $H_0$  shown by the turquoise bar and the red shading.

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  $n$ ).

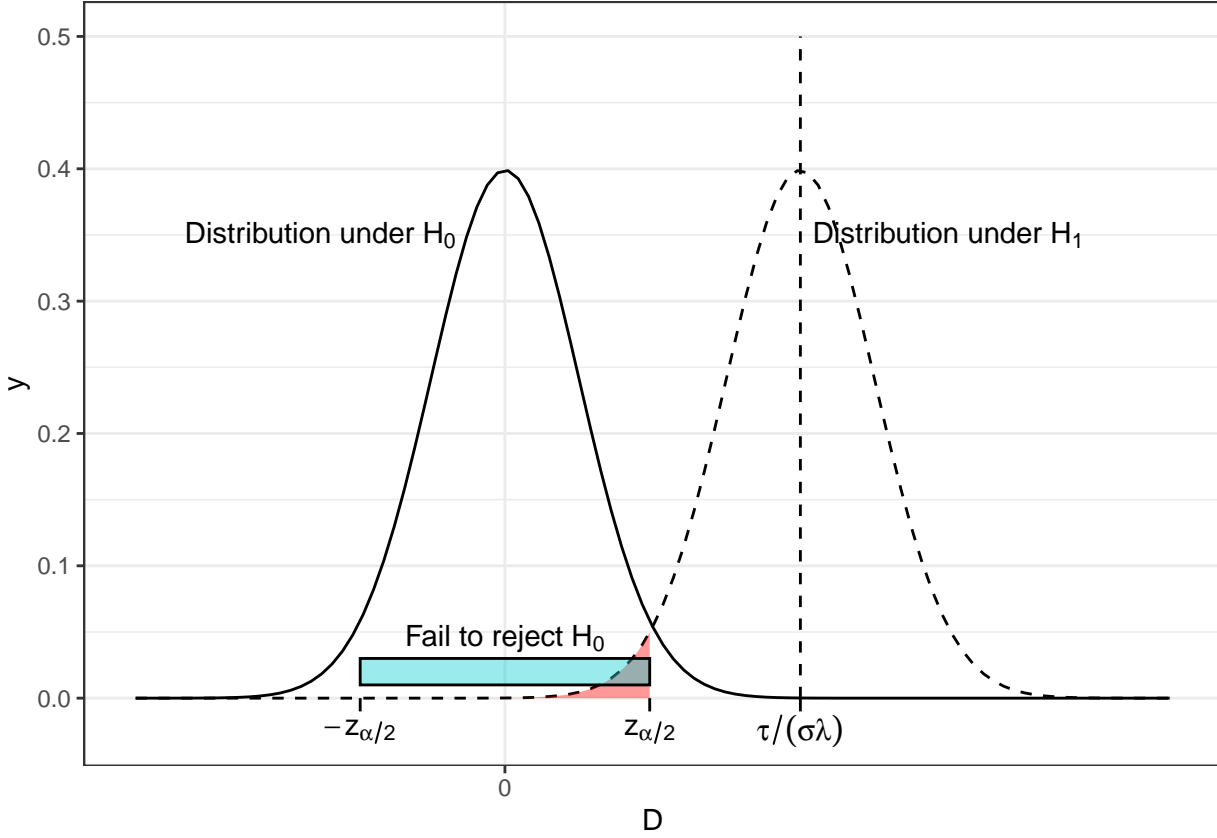


Figure 2.5: The distribution of  $D$  under both  $H_0$  and  $H_1$  for some arbitrary values of effect size, population variance,  $n$  and  $m$ , with the region in which we fail to reject  $H_0$  shown by the turquoise bar and the red shading.

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.

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  $\alpha$ .

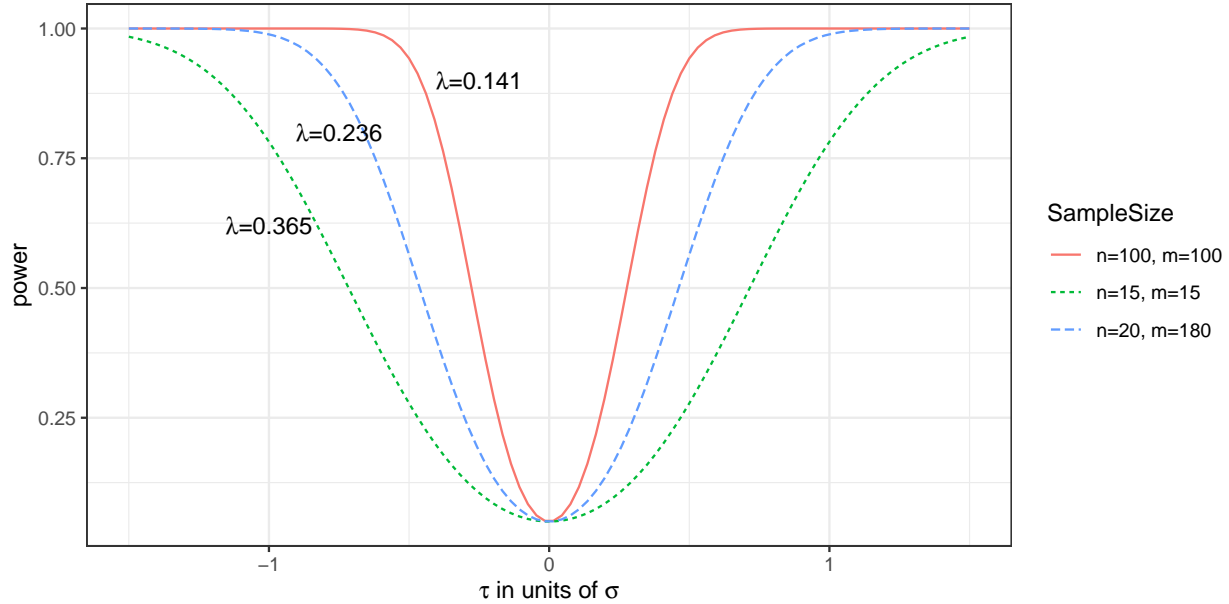


Figure 2.6: Power curves for various values of  $n$  and  $m$ , with effect size in units of standard deviation, given a type I error rate of 0.05.

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

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

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 (??) rearranges to

$$N = \frac{2\sigma^2 (z_\beta + z_{\frac{\alpha}{2}})^2}{\tau_M^2}. \quad (2.3)$$

**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 (2.3) 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.

# Bibliography

- 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.
- Matthews, J. N. (2006). *Introduction to randomized controlled clinical trials*. CRC Press.
- 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.
- Zhong, B. (2009). How to calculate sample size in randomized controlled trial? *Journal of thoracic disease*, 1(1):51.