Clinical Trials 4H

Rachel Oughton

2023-12-11

Contents

Course Information

Welcome to Clinical Trials 4H!

This is a 10 credit module, open to fourth year students. An overview of what we'll cover is

An introduction to clinical trials - part 1 Designing and planning a randomised clinical trial - parts 2 and 3

Lectures

Computer classes

Assessment

This module is assessed through two equally weighted pieces of coursework. I'll give you more details nearer the time, but one will be assigned around half way through the term, the second towards the end of term.

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)

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.

In a clinical trial, we will have two groups [is this true!?]:

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.

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.

The goal of the RCT is to estimate the **treatment effect**, with some specified level of confidence. 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.

1.1 A brief history of RCTs

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.

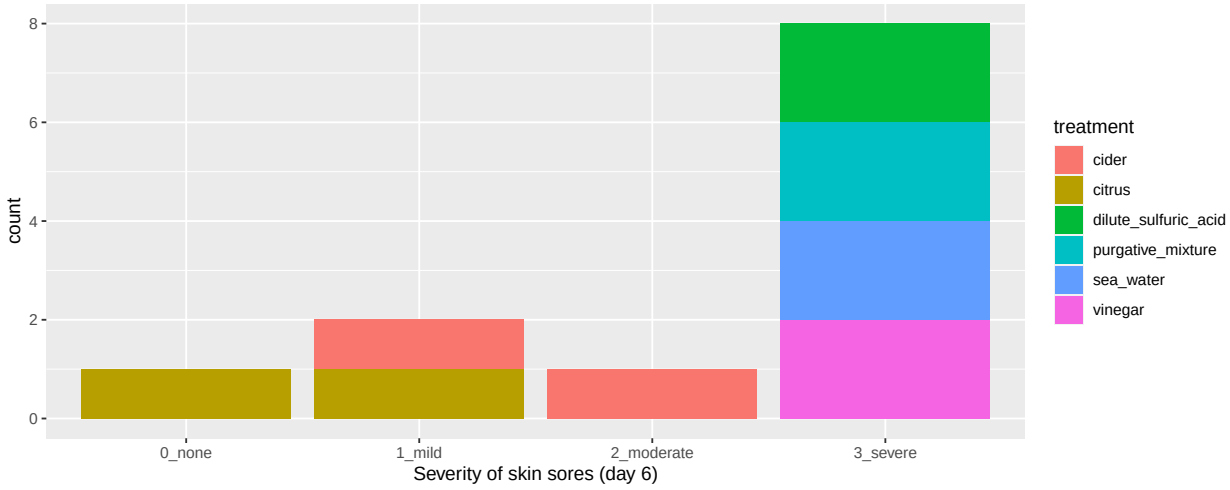
[some examples from agriculture and from further back? Check Spiegelhalter]

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), and in 1757 Lind decided to test six of them, 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,"



1.2 The structure of a clinical trial

Research questions, relating real world problem / question to trial, causal analysis. Stuff from Hulley et al. (2013). Causal inference, relating study to the real world etc. Brief overview of less statistical issues.

1.2.1 The primary outcome

In an RCT, 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.

[examples of primary outcome variables]

Chapter 2

Designing and planning a randomised controlled trial (RCT)

In the first half 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 we will go on to think about binary variables and survival times.

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.

2.1 Sample size for a normally distributed primary outcome variable

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.

2.1.1 The treatment effect

In Section 1.2.1 we discussed the need to settle on a **primary outcome variable**. A further reason this is important is that we base our sample size calculations on the primary outcome variable.

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

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

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

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

Definition 2.1. 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, it is a measure of the probability of obtaining whatever result (eg. treatment effect) we have have found simply by random chance, when in fact there is no treatment effect. Generally, a P-value of 0.05 is accepted as sufficient evidence to reject the null hypothesis, although in clinical settings it can often be smaller (eg. 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. This is 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 $p = 0.000049$.

2.1.2.1 Insignificant results

If our P-value is relatively large, say 0.3 or 0.5, then our result is not at all unlikely under the null hypothesis, and provides no evidence to reject H_0. However, it is not inconsistent with the existence of a treatment effect, so we don't say there is evidence to accept H_0. One can imagine that if the true treatment effect τ were tiny, many trials would fail to find evidence to reject H_0. However, if our sample size were sufficiently large, we should be able to detect it. Conversely, if τ is very large, even a relatively small sample size is likely to provide enough evidence to reject H_0.

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

[link to examples of primary outcome variables in section 1.2.1]

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

2.1.2.2 One-sided or two-sided?

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

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

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

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 $1 - F(2, df = 31) = 0.0272$ (where $F(\cdot)$ is the

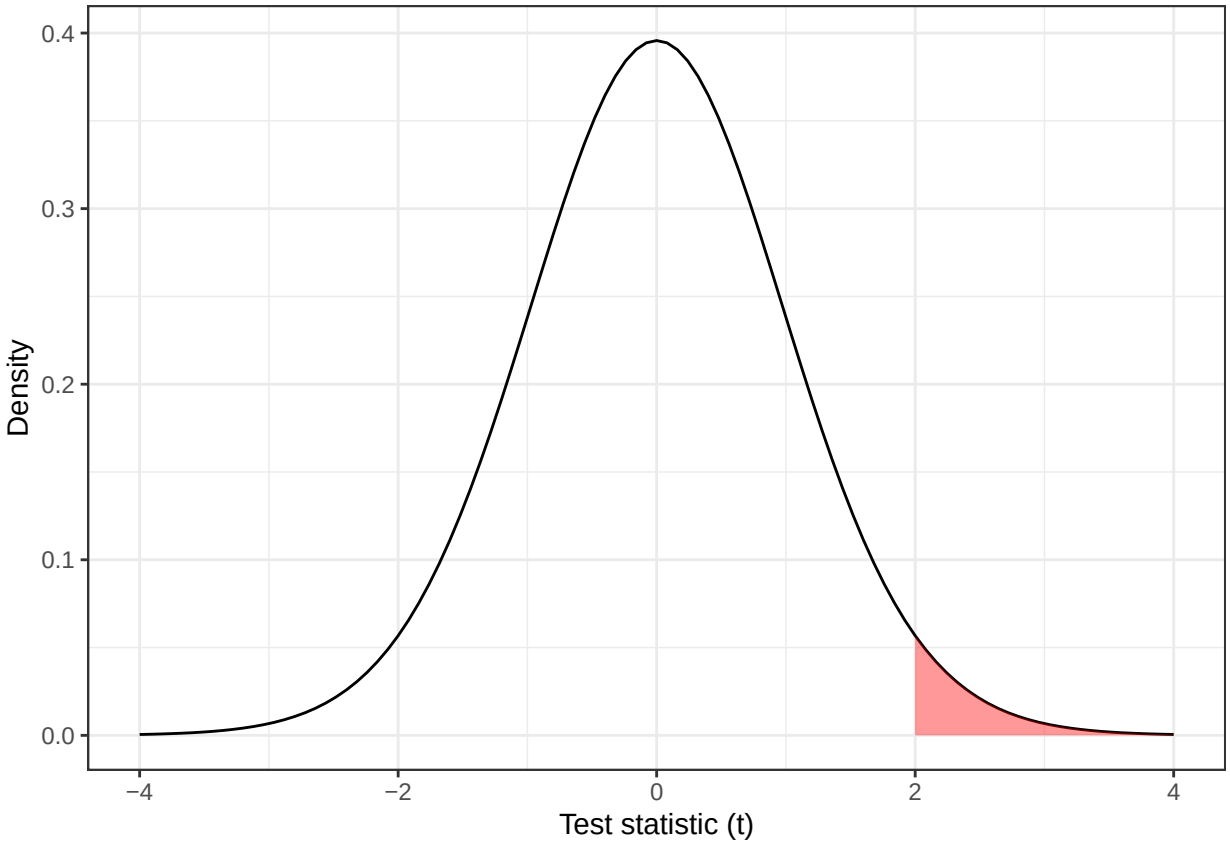


Figure 2.1: The distribution t_{31}, with the area corresponding to $t > 2$ shaded.

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 value of t with $t < -2$ are considered 'equivalent' to those with $t > 2$, in the sense of how unlikely they are under H_0.

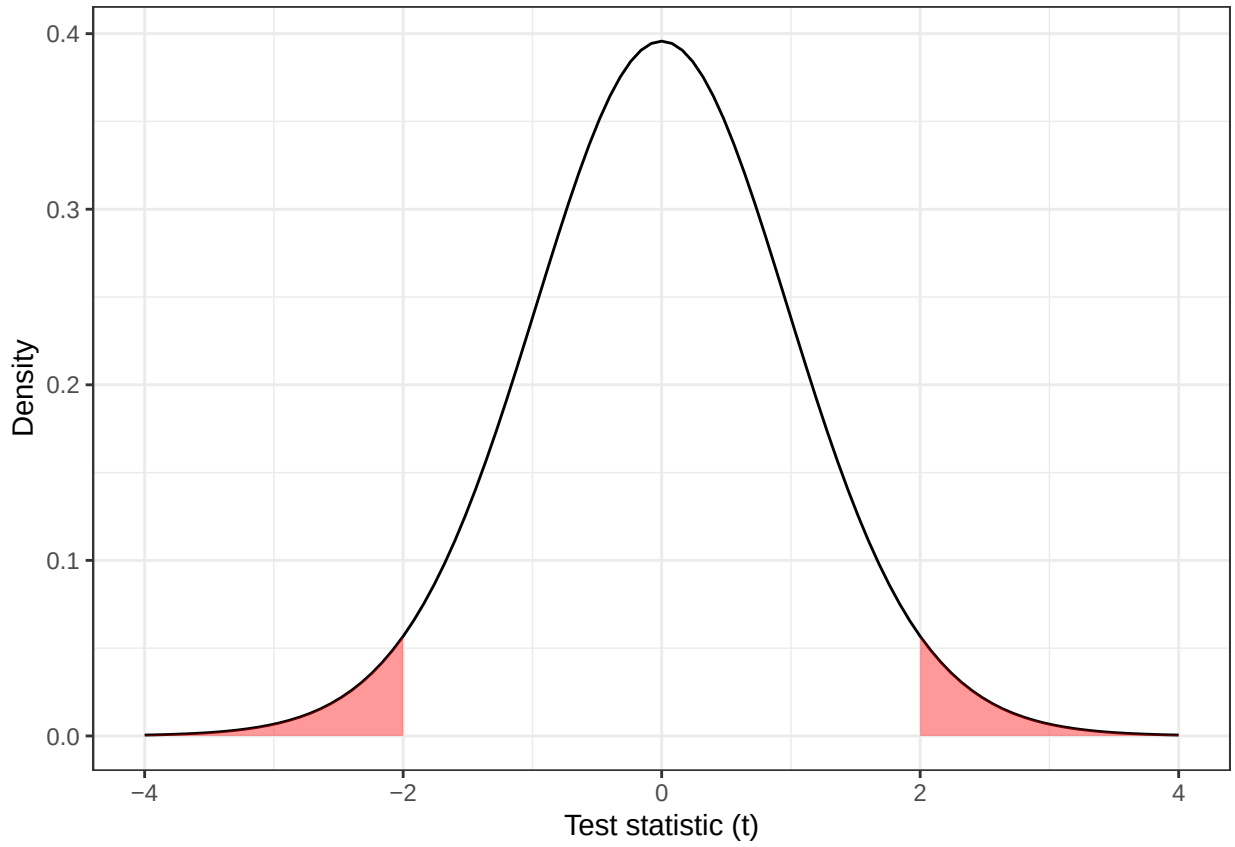


Figure 2.2: The distribution t_{31}, with the area corresponding to $|t| > 2$ shaded.

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-sided tests.

2.1.3 Constructing a measure of effect size

To understand the theory behind calculating an appropriate sample size, we will initially consider just the scenario in which our primary outcome variable X is normally distributed.

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

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

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 $\bar{x}_A$ and $\bar{x}_B$ from each group, and a pooled estimate of the standard deviation

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

where s_A and s_B are the sample standard deviations for groups A and B respectively, for example

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

Using these values we can compute

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

as a standardised measure of the effect τ.

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

Proof. Under H_0,

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

and therefore

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

and

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

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

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

and so we have

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

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

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

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

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

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

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

We can therefore use D to test our hypotheses; if D is such that

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

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

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 $100\alpha\%$ level if $|D| > z_{\alpha/2}$, where z are standard normal values.

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

This is equivalent to $\bar{x}_B - \bar{x}_A$ falling outside the interval

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

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

2.1.4 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 a high probability of rejecting H_0.

Definition 2.2. The **power function** of a test is the probability that we reject H_0, given that H_0 is false. The notation we will use is

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

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

Under H_1, we have (approximately)

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

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

Figure 2.3 shows the distribution of D under H_0 and H_1 for some arbitrary (non-zero) effect size τ. The turquoise bar shows the acceptance region of H_0, ie. the range of observed values of D for which we will fail to reject H_0. We see that this contains 95% of the area of the H_0 distribution (we are assuming $\alpha = 0.05$ 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. One minus this area (ie. the area under H_1 that leads us to accept H_1) is the power.

We can see that if the distributions have better separation, as in Figure 2.4, the power becomes greater.

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

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

The first term in Equation (2.1) is the area in the direction of τ. In Figures 2.3 and 2.4 this is the region to the right of the interval for which we fail to reject H_0, ie. where

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

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

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

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

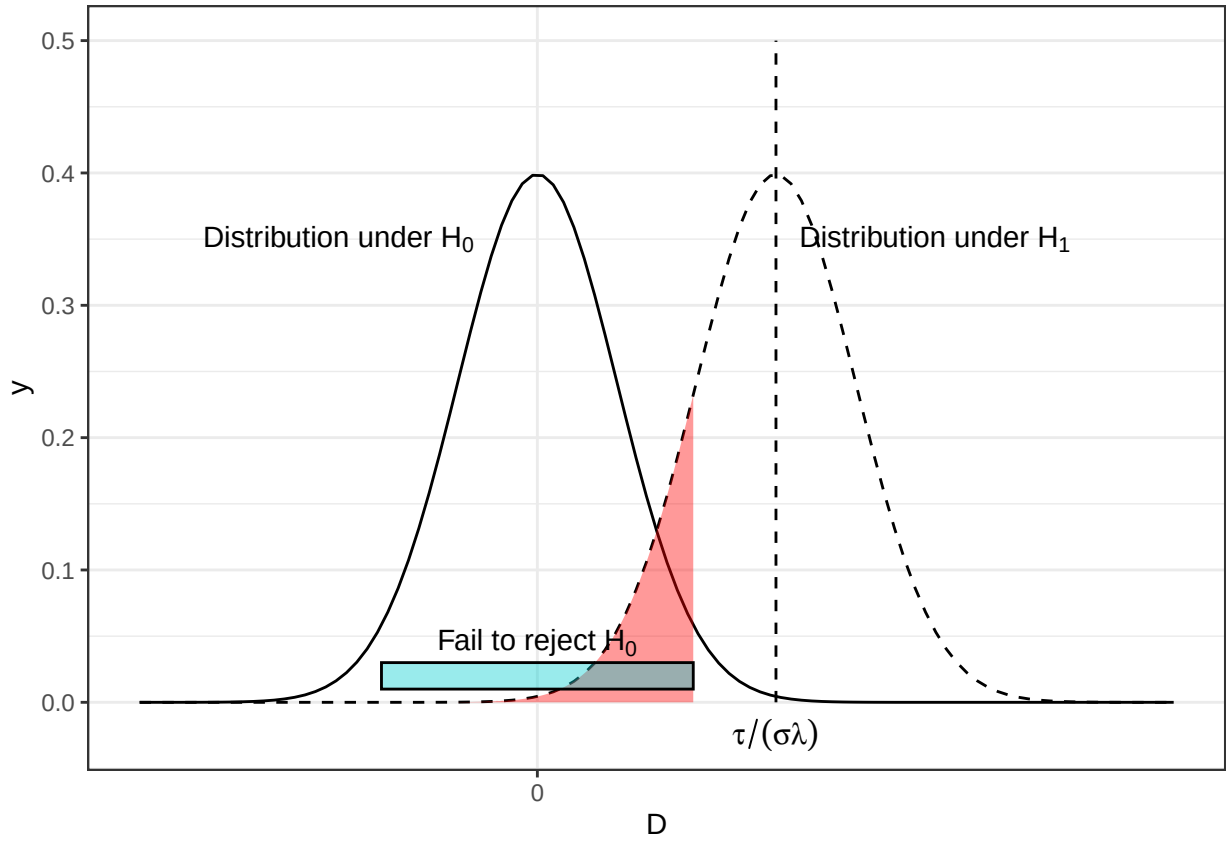


Figure 2.3: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the acceptance region of H_0 shown by the turquoise bar.

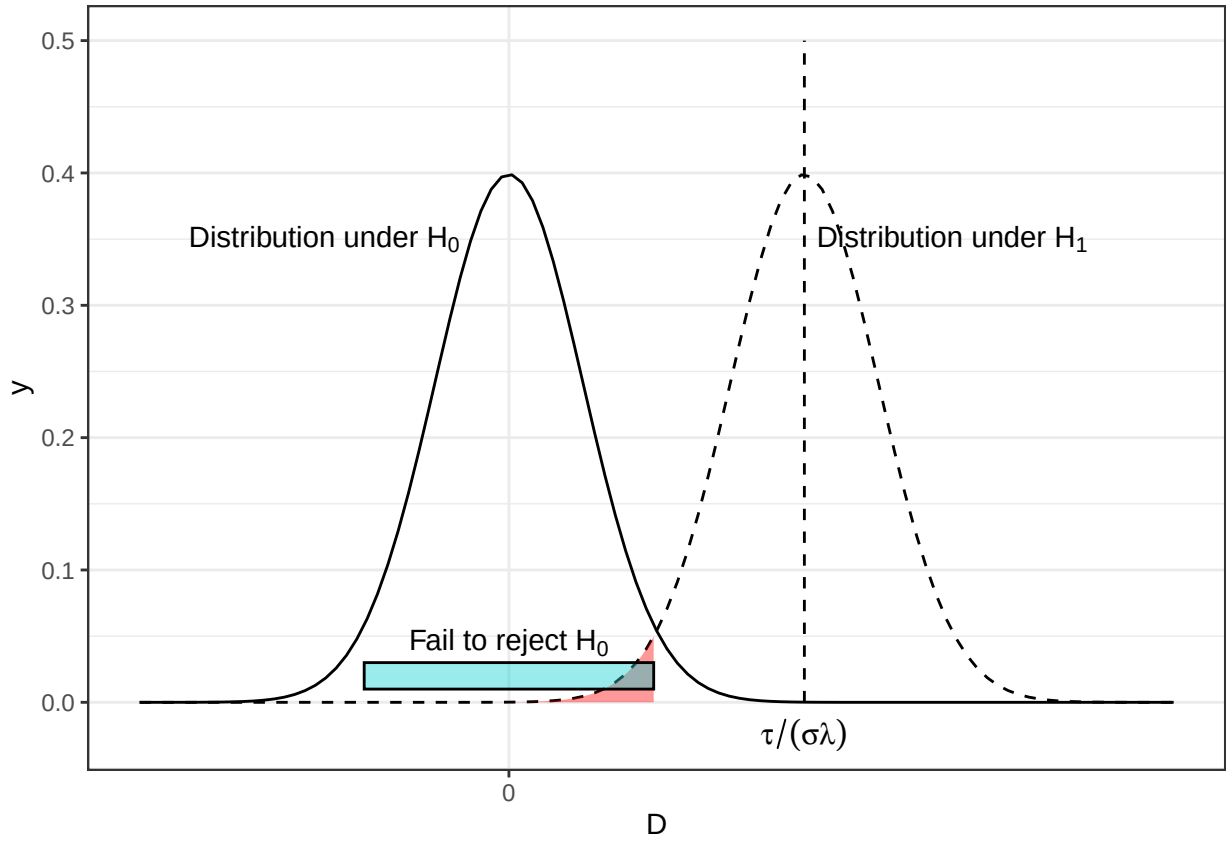


Figure 2.4: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the acceptance region of H_0 shown by the turquoise bar.

Figure 2.5 shows the power function $\Psi(\tau)$ for τ in units of σ (or you could think of this as for $\sigma = 1$), for three different pairs of values of n and m. We see that in general the power is higher for larger sample sizes, and that of the two designs where $n + m = 200$, the balanced one with $n = m = 100$ achieves the greatest power.

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

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

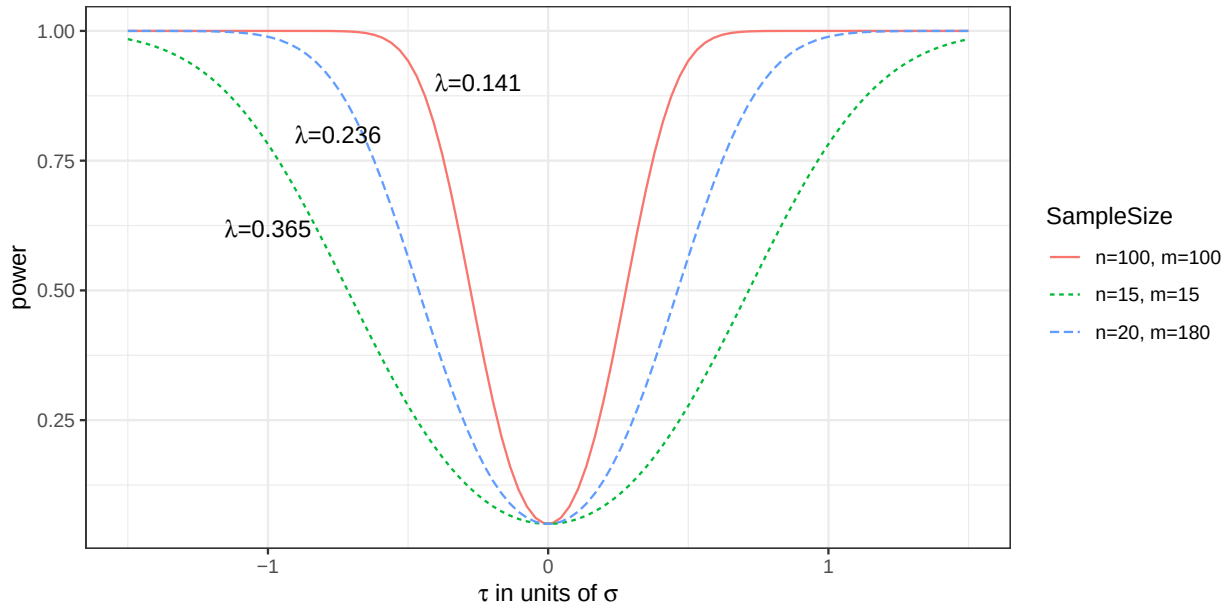


Figure 2.5: Power curves for various values of n and m, with effect size in units of standard deviation, given a type I error rate of 0.05.

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

2.1.5 A sample size formula

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

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

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

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

Definition 2.3. The **minimum detectable effect size** τ_M for a particular trial is the smallest value of effect size that is able to be detected with power β (for some specified value of β).

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

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

Therefore, Equation (2.1) becomes

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

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

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

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

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

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 $\beta = 0.8$, 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

$$\begin{aligned} N &= \frac{2 \times 6^2 (0.842 + 1.96)^2}{3^2} \\ &= 111.6, \end{aligned}$$

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

2.2 Sample size by simulation

A method for sample size calculation that has become increasingly popular in recent years is to use simulation. In simple terms, we write code that runs the trial many, many times in order to determine how many participants we need to achieve the power required.

This approach has the following advantages over the formula-based methods presented in Section 2.1:

1. **Transparency**: If the data generating mechanism is made clear, then the assumptions behind the trial are also clear, and the simulation can be replicated by anyone. Reproducibility is a big issue in clinical trials.
2. **Flexibility**: Whereas the methods above are limited to very specific circumstances, one can simulate arbitrarily complex or unusual trials.
3. **Practice**: This process requires us to perform our planned analysis at the planning stage, thus raising any potential issues early enough to adapt the plan.

Arguably the first and third advantages could be true of a well-planned trial that used conventional sample size formulae, but the second is an advantage unique to simulation.

2.2.1 The simulation method

The first step in the simulation method is to generate trial data. To do this, we use the underlying probabilistic model we have used (or would use) to develop a sample size formula.

Chapter 3

Bias

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

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

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

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

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

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

In this case, $E(s^2) \neq \sigma^2$, and s^2 is an biased estimator. However, we know that

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

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

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

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

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

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

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.

Many of the ideas in this section are closely linked to the topic of **allocation**, which we will discuss in detail in Section 4).

3.1 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.

3.1.1 Selection bias

Selection bias occurs when certain patients or subjects are systematically more (or less) likely be entered into the trial because of the treatment they will receive. This shouldn't be able to happen, because it is usually 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 mean that the estimated treatment effect is worryingly far from the true population treatment effect (where the population is the group of all eligible patients), but this is not technically selection bias.

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

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

$$E(X \mid 1) = \mu_1$$
$$E(X \mid 2) = \mu_2$$

and we assume $\mu_1 > \mu_2$.

For the overall trial group, for untreated patients we have

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

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

$$E(X \mid N, 1) = \mu_1 + \tau_N$$
$$E(X \mid N, 2) = \mu_2 + \tau_N$$
$$E(X \mid S, 1) = \mu_1 + \tau_S$$
$$E(X \mid S, 2) = \mu_2 + \tau_S.$$

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

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

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

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

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

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

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

We can split these according to disease severity, so that

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

and similarly for group S.

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

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

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

Following the same process for group S, we arrive at

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

.

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

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

and

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

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

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

where the third term represents the bias.

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

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

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

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

3.1.2 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.

3.1.3 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. 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.

Chapter 4

Allocation

Historically (and probably still, to an extent), clinical trials have not necessarily used random allocation to assign participants to groups. Altman and Bland (1999b) gives an overview of why this has led to bias, and gives some examples. Altman and Bland (1999a) and Treasure and MacRae (1998)

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

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

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

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

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

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

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

4.1 Allocation methods

4.1.1 Simple random allocation

Perhaps intuitively the most simple method would be a 'toin coss', where each participant has a probability 0.5 of being placed in each group. As participants arrive, assignment A or B 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, A of size N_A and B of size N_B, with $N_A + N_B = 2n$. Patients are allocated independently with equal probability, which means

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

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

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

The probability that $N_A = N_B = n$ is

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

These probabilities are shown in Figure 4.1. We can see that this method leads to very unequal groups relatively easily; with $n = 15$, $P\left(N_{max} \geq 20\right) = 0.099$, so there is around a one in ten chance that one group will be double or more the size of the other.

As we have seen when thinking about sample sizes in Section 2.1.4, this will reduce the power Ψ of the trial, since it depends on $\lambda\left(N_A, N_B\right) = \sqrt{\frac{1}{N_A} + \frac{1}{N_B}}$.

For larger trials, this imbalance will be less pronounced, for example Figure 4.2 shows the same for $n = 200$.

In this case the $P\left(N_{max} \geq 220\right) = 0.051$, 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.

4.1.2 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 m, 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 containing two As and two Bs

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

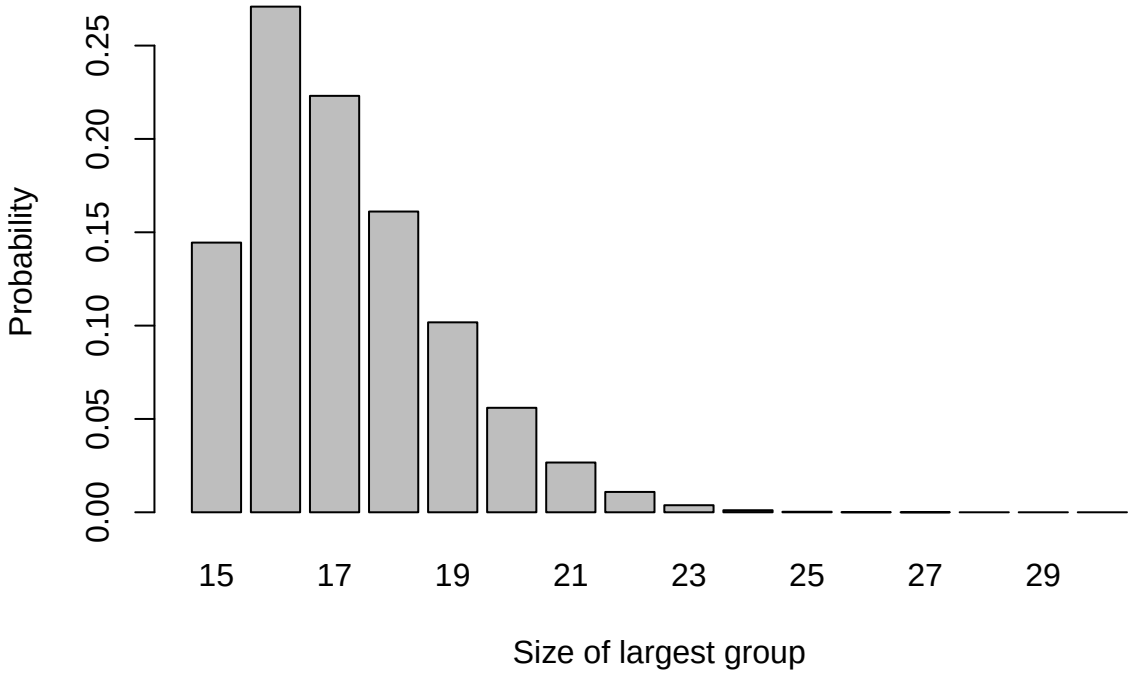


Figure 4.1: The probability distribution of largest group size for n=15.

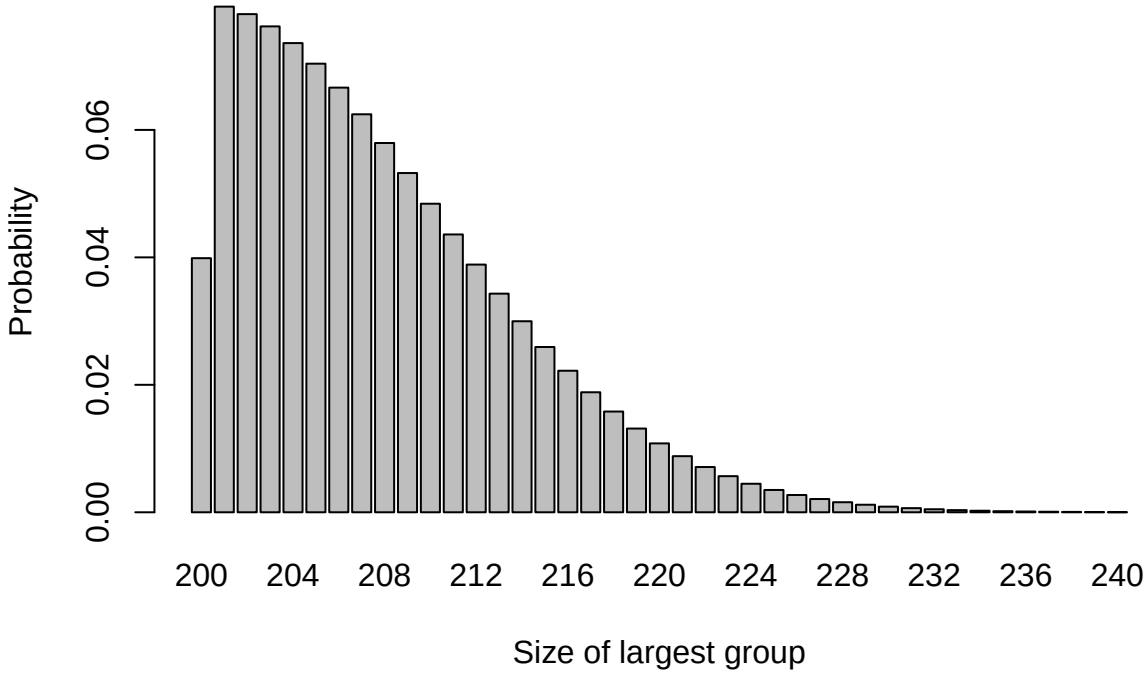


Figure 4.2: The probability distribution of largest group size for n=200.

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

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

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

4.1.2.1 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.

4.1.3 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.

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.

4.1.3.1 Biased coin designs

Suppose we are using a biased coin design for a trial to compare two treatments, A and B. At the point where some number n (not the total trial cohort) have been allocated, we can use the notation $N_A\,(n)$ for the number of participants allocated to treatment A, and $N_B\,(B)$ for the number of participants allocated to treatment B. Using these, we can denote the *imbalance* in treatment numbers as

$$D(n) = N_A(n) - N_B(n) = 2N_A(n) - n.$$

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

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

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

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

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

or

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

Because of the way we have set up the scheme,

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

and

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

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

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

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

Figure 4.3 shows four realisations of this random walk with $P = 0.667$. We see that sometimes the imbalance gets quite high, but in general it isn't too far from 0.

Figure 4.4 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 4.5, 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.

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.

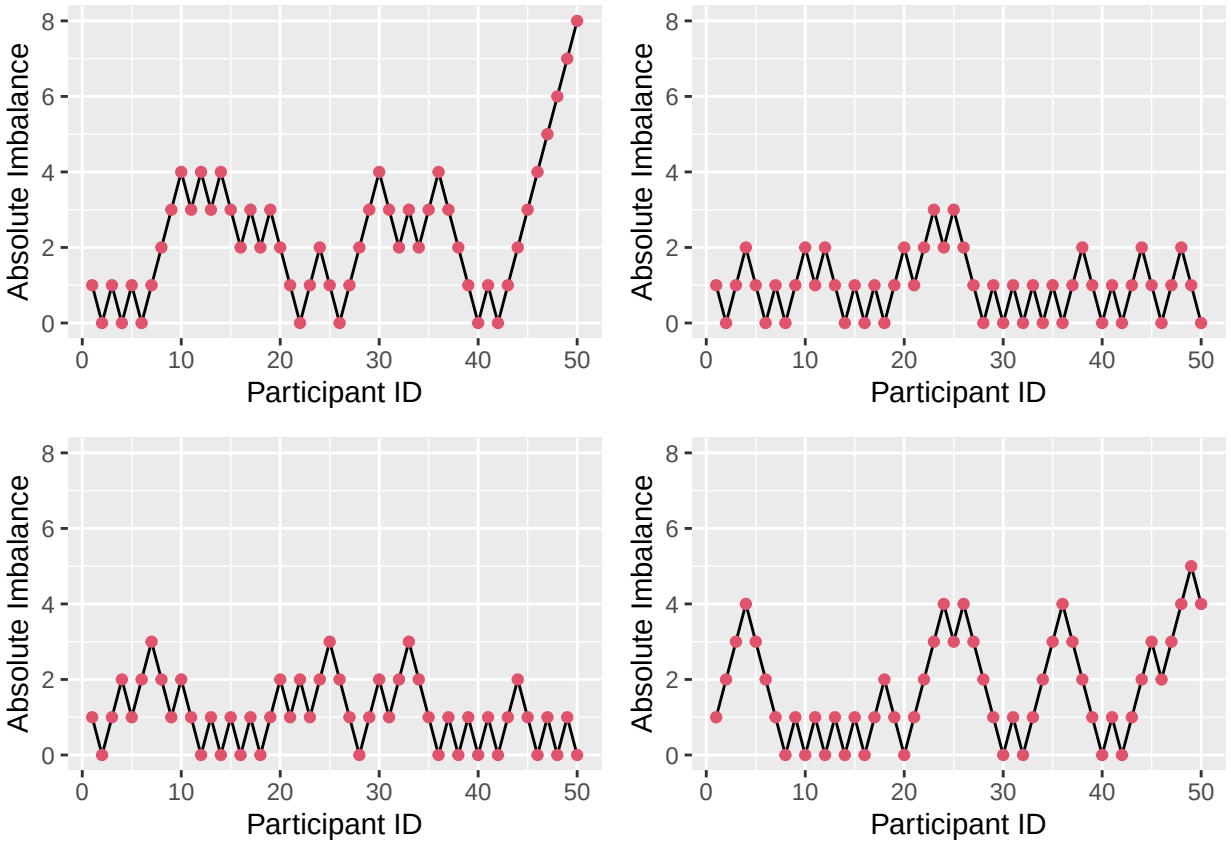


Figure 4.3: Absolute imbalance for a biased-coin scheme with $P = 0.667$.

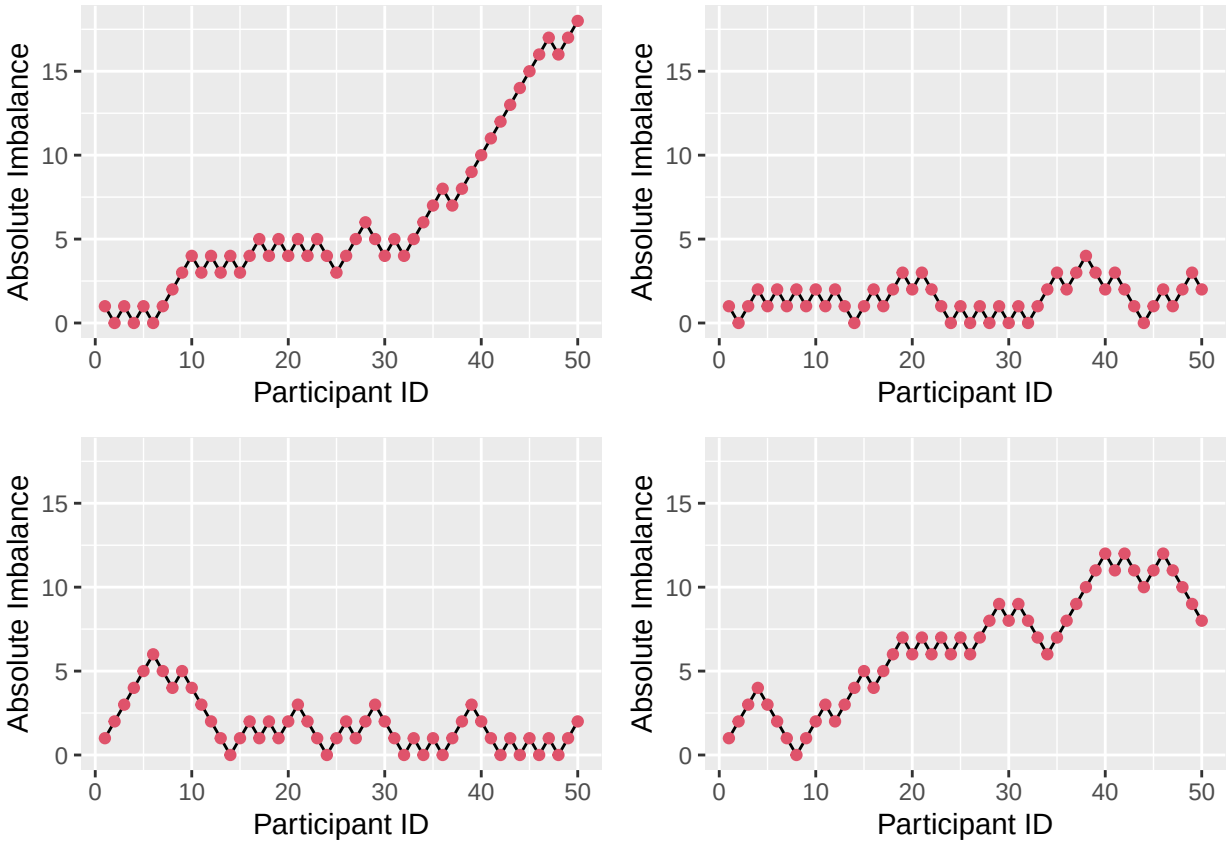


Figure 4.4: Absolute imbalance for a biased-coin scheme with $P = 0.55$.

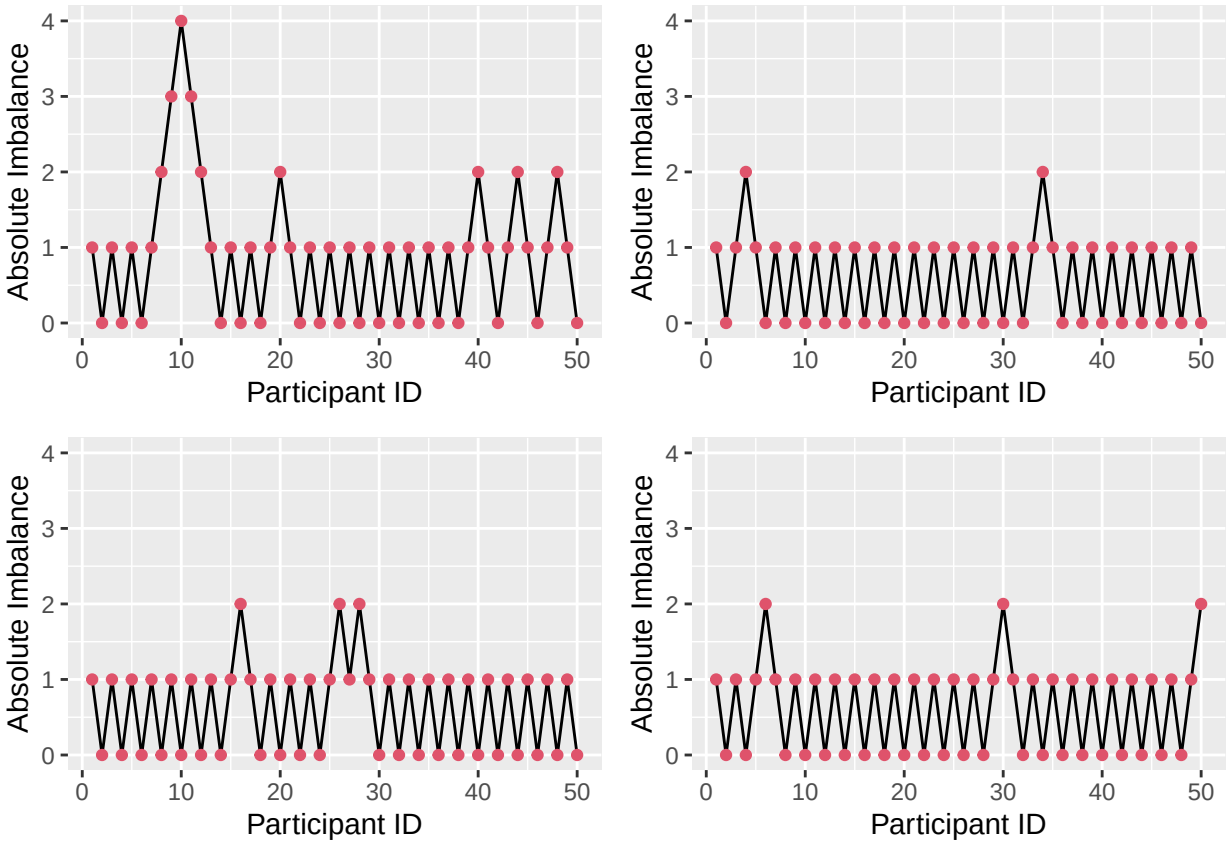


Figure 4.5: Absolute imbalance for a biased-coin scheme with $P = 0.9$.

4.1.3.2 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). 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, $N_A(n)$ participants have been allocated to treatment A, and $N_B(n) = n - N_A(n)$ to treatment B. The imbalance is therefore

$$D(n) = N_A(n) - N_B(n) = 2N_A(n) - n.$$

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, $r + N_B(n)\,s$ will be labelled by treatment A and $r + N_A(n)\,s$ by treatment B.

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 A, then the probability that it becomes less imbalanced at the next allocation is the probability of the next allocation being to treatment B, which is

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

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

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

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

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

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

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

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

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

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

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

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 4.7. By contrast, if r is large and s is small, as in Figure 4.8, the probabilities stay closer to $\left(\frac{1}{2}, \frac{1}{2}\right)$ and so more imbalance occurs early on.

4.2 Problems with allocation

THIS SECTION PROBABLY NEEDS TO GO / BE CHANGED, AND I NEED TO MAKE SURE THEY CAN TAKE INTO ACCOUNT THE AGE VARIABLE IN THE ASSIGNMENT

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 4.9 shows an example.

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.

Variable	Control (n=100)	Intervention (n=100)
Gender (female, %)	50	50
Age, y (mean +/- SD)	50 +/- 10	50 +/- 10

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

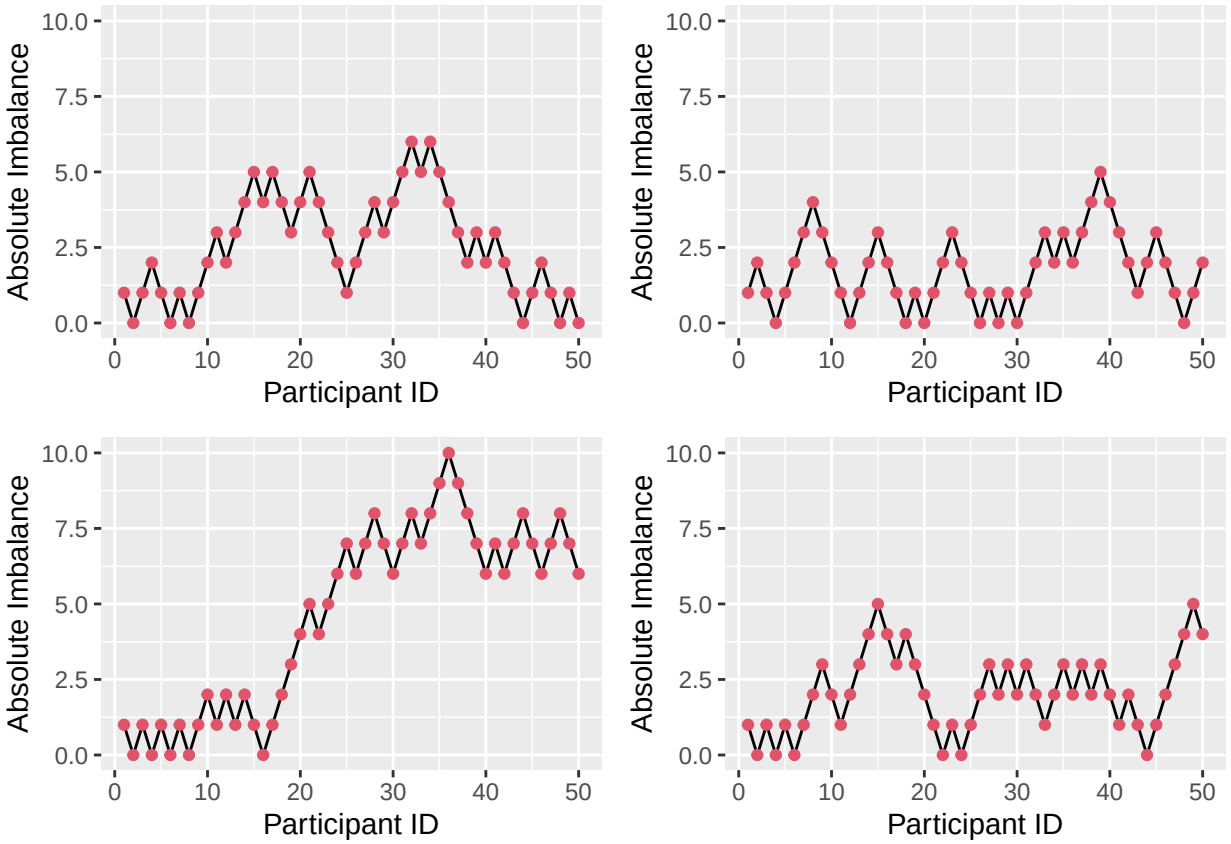


Figure 4.6: Four realisations of absolute imbalance for r=1, s=1, N=50.

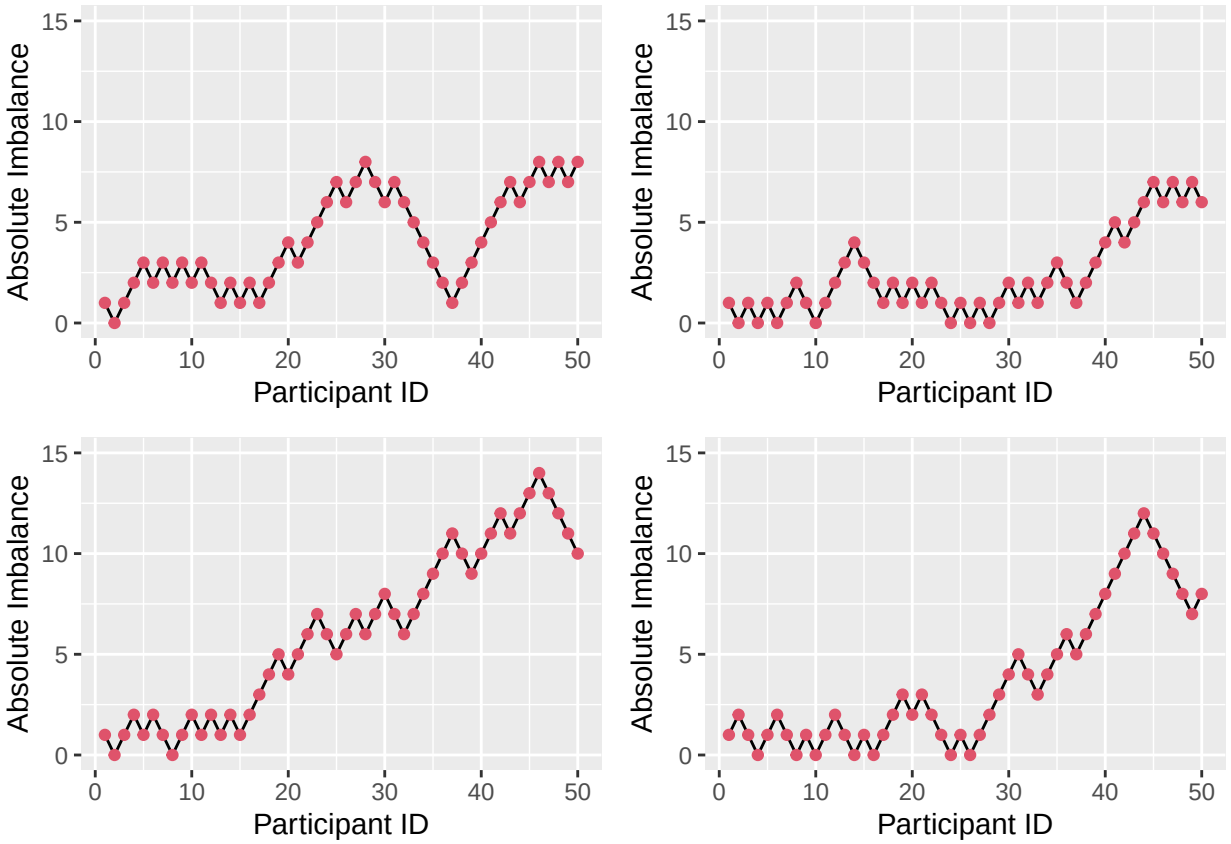


Figure 4.7: Four realisations of absolute imbalance for r=1, s=8, N=50.

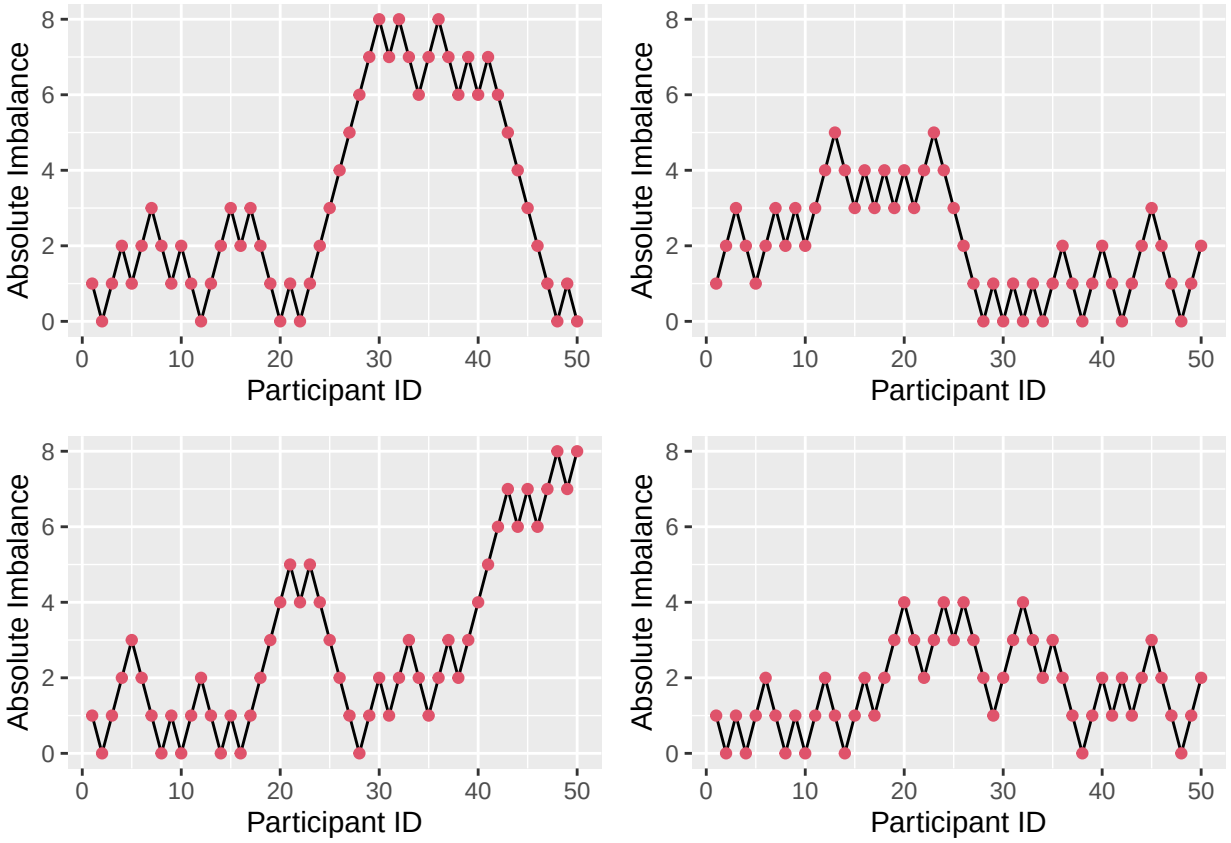


Figure 4.8: Four realisations of absolute imbalance for r=8, s=1, N=50.

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

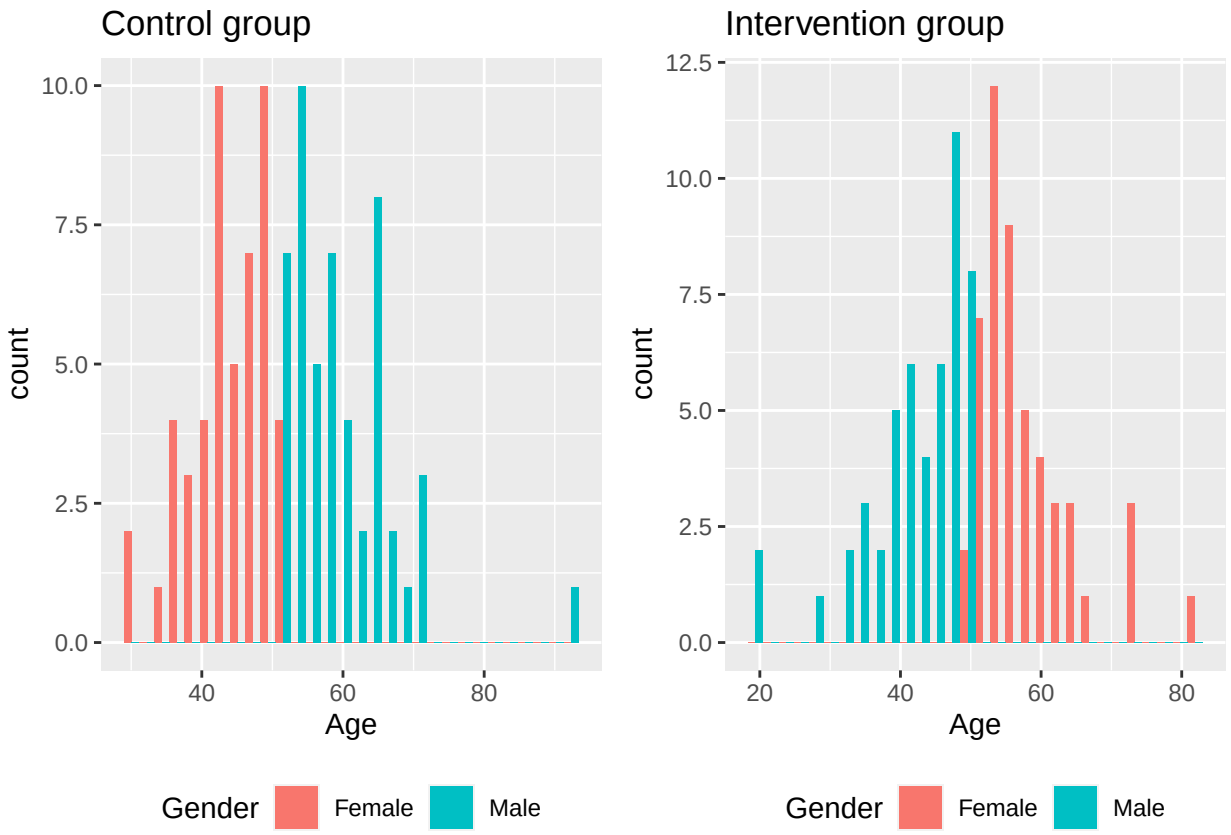
Variable	Licorice (*N* = 118)	Sugar-water (*N* = 117)	Standardized difference[a]
Age, y	57 ± 15	58 ± 16	−0.09
Gender (female), %	42	38	0.08
Body mass index, kg/m²	26 ± 4	26 ± 4	−0.01
Smoking, %			−0.01
Current	38	38	
Past	31	31	
Never	31	31	
Pain (yes), %	0	2	−0.19
ASA physical status, %			−0.07
I	19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

Summary statistics presented as percent of patients or mean ± SD.

[a]Standardized difference (licorice – sugar-water) defined as the difference in means or proportions divided by the pooled standard deviation; >0.2 in absolute value indicates imbalance.

[b]Surgery size: small (thoracoscopy); medium (thoracotomy <3 h), large (thoracotomy >3 h or blood loss >1000 mL).

Figure 4.9: Summary statistics for an RCT comparing a licorice gargle (the intervention) to a sugar-water gargle (the standard). From @ruetzler2013randomized



	Male	Female
Control	59.4 (7.37)	44.02 (5.49)
Intervention	42.22 (7.2)	57.84 (6.7)

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.

4.3 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. This whole procedure is analogous to performing a factorial experiment, without being able to control the factor levels of the experimental units.

Example 4.1. Suppose we are planning a trial involving people over the age of 50, and we anticipate that age and sex might both play an important role in how participants respond to the treatment.

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

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

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

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

4.4 Minimization

Altman and Bland (1999a) tees up minimization. Treasure and MacRae (1998) a good source.

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 us 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. As an example, in a study comparing aspirin to a placebo preceding coronary artery surgery, Kallis et al. (1994) chose age (≤ 50 or 50), sex (M or F), operating surgeon (3 possibilities) and number of coronary arteries affected (1 or 2).

When a patient enters the trial, these factors are listed. The patient is then allocated in such a way as to minimise any difference in these factors. 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.

4.4.1 Minimization algorithm

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In order to implement minimisation, we follow these steps:

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

4. We form the sum

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

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

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

Example 4.2. From Altman (1990) (citing Fentiman et al. (1983)). In this trial, 46 patients with breast cancer were allocated to receive either Mustine (arm A) or Talc (arm B) as treatment for pleural effusions (fluid between the walls of the lung). They used four prognostic factors: age (≤ 50 or > 50), stage of disease (I or II, III or IV), time in months between diagnosis of breast cancer and diagnosis of pleural effusions (≤ 30 or > 30) and menopausal status (Pre or post).

Let's suppose that 15 patients have already been allocated. The totals of patients in each treatment arm in terms of each level of each prognostic factor are shown in Table @ref(tab:minim_eg).

Example after 15 allocations

Mustine (A)

Talc (B)

Age

1. 50 or younger

3

4

2. >50

4

4

Stage

1. I or II

1

2

2. III or IV

6

6

Time interval

1. 30 months or less

4

2

2. >30 months

4

5

Menopausal Status

1. Pre

4

3

2. Post

5

3

Allocations of first 15 patients, divided by diagnostic factor

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

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

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

4.5 Some simulated examples

To conclude this section we will demonstrate methods using simulated datasets with the same underlying probability distributions. We consider four factors:

- Age: 44 and under, 45 - 54, 55-64, 65-74 and 75 and over (5 levels)
- Sex: M or F
- Smoking: current, past or never
- Hypertension: Yes or no

These will be generated using the following distributions:

- Age: $N\left(60, 10^2\right)$ (then binned as above)
- Sex: $p\left(M\right) = p\left(F\right) = 0.5$
  - Smoking: $p\left(\text{Current}\right) = 0.4$, $p\left(\text{Past}\right) = p\left(\text{Never}\right) = 0.3$
  - Hypertension:
    $$p(Y) = \begin{cases} 0.25 & \text{Age under 45} \\ 0.55 & \text{Age 45 to 74} \\ 0.75 & \text{Age 75 and over} \end{cases}$$

  To demonstrate our allocation methods, we simulate one dataset from the distribution described, containing 100 participants. This dataset is shown in the Table below, and in Figures 4.10 and 4.11.

ID	Sex	Age	Smoking	Hypertension
1	Male	65 to 74	Never	Yes
2	Male	75 and over	Current	No
3	Male	75 and over	Past	Yes
4	Female	55 to 64	Current	Yes
5	Male	55 to 64	Never	Yes
6	Male	75 and over	Never	No
7	Male	45 to 54	Never	Yes
8	Male	55 to 64	Past	No
9	Female	75 and over	Current	No
10	Female	55 to 64	Past	Yes
11	Male	45 to 54	Never	No
12	Female	44 and under	Never	No
13	Female	55 to 64	Current	Yes
14	Female	45 to 54	Current	Yes
15	Male	65 to 74	Never	Yes
16	Male	65 to 74	Current	No
17	Female	55 to 64	Current	No
18	Female	65 to 74	Current	Yes
19	Male	55 to 64	Past	Yes
20	Male	55 to 64	Current	Yes
21	Female	65 to 74	Past	No
22	Female	45 to 54	Past	Yes
23	Male	55 to 64	Past	Yes
24	Female	55 to 64	Current	No
25	Female	65 to 74	Current	No
26	Male	45 to 54	Never	Yes
27	Male	55 to 64	Current	Yes
28	Female	45 to 54	Never	No
29	Female	55 to 64	Past	Yes
30	Female	45 to 54	Never	No
31	Female	55 to 64	Current	Yes
32	Male	44 and under	Never	Yes
33	Female	65 to 74	Past	No
34	Female	55 to 64	Never	No
35	Female	65 to 74	Current	Yes
36	Male	65 to 74	Past	Yes
37	Male	55 to 64	Current	No
38	Female	45 to 54	Current	Yes
39	Male	45 to 54	Current	Yes
40	Female	45 to 54	Current	Yes
41	Female	55 to 64	Current	No
42	Female	75 and over	Past	Yes
43	Male	45 to 54	Past	Yes
44	Female	45 to 54	Never	Yes
45	Male	75 and over	Never	Yes
46	Male	45 to 54	Never	Yes
47	Female	55 to 64	Past	Yes
48	Male	65 to 74	Current	Yes
49	Male	65 to 74	Never	No
50	Male	65 to 74	Never	Yes
51	Male	55 to 64	Past	No
52	Male	65 to 74	Never	No
53	Male	55 to 64	Never	No
54	Female	55 to 64	Past	Yes
55	Female	44 and under	Never	No
56	Male	75 and over	Past	No
57	Female	65 to 74	Current	Yes

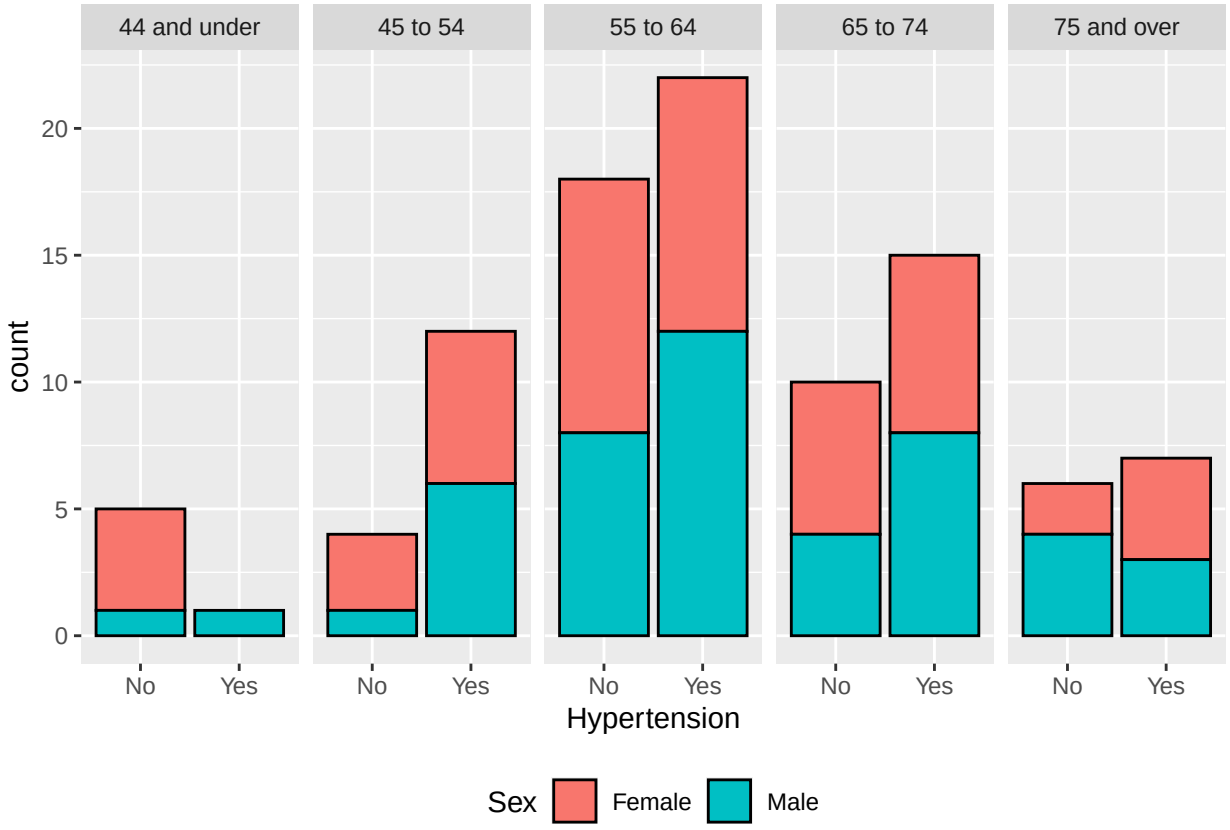


Figure 4.10: A simulated dataset of 100 participants, using the distributions described above, split by hypertension, age and sex.

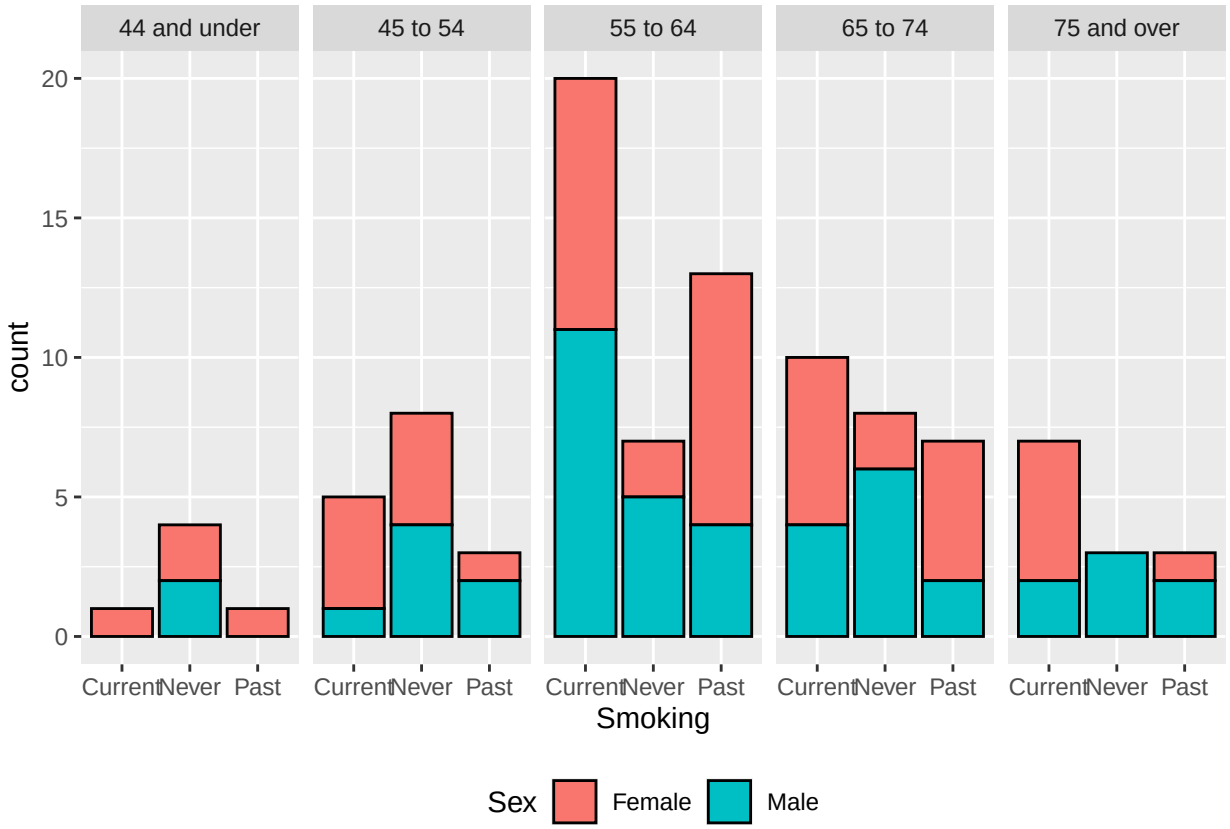


Figure 4.11: A simulated dataset of 100 participants, using the distributions described above, split by smoking history, age and sex.

4.5.1 Simple random allocation

In simple random allocation, each participant is allocated to one of the two trial arms with equal probability. In terms of our allocation,

Chapter 5

The intervention

Having settled on a sample size and an allocation strategy, the trial can now be run.

Chapter 6

Analyzing RCT data

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

We'll first focus on the scenario where the trial outcome is measured on a continuous scale, and then we'll go on to look at other types of data.

Example 6.1. To illustrate the theory and methods, we'll use an example dataset from Hommel et al. (1986) (this example is also used by Matthews (2006)). The data involves a trial of 16 diabetes patients, and focusses on a drug (Captopril) that may reduce blood pressure. This is important, since for those with diabetes, high blood pressure can exacerbate kidney disease (specifically diabetic nephropathy, a complication of diabetes). To participate in the trial, people had to be insulin-dependent and already affected by diabetic nephropathy. In the trial, systolic blood pressure was measured before participants were allocated to each trial arm, and then measured again after one week on treatment. A placebo was given to the control group, so that all participants were blinded.

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

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

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

6.1 Confidence intervals and P-values

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

Example 6.2. Summary statistics of the outcome for each group are shown below.

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

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

Table 6.1: Data for the Captopril trial from @hommel1986effect.

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

Table 6.2: Summary statistics for each group.

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

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

In our case

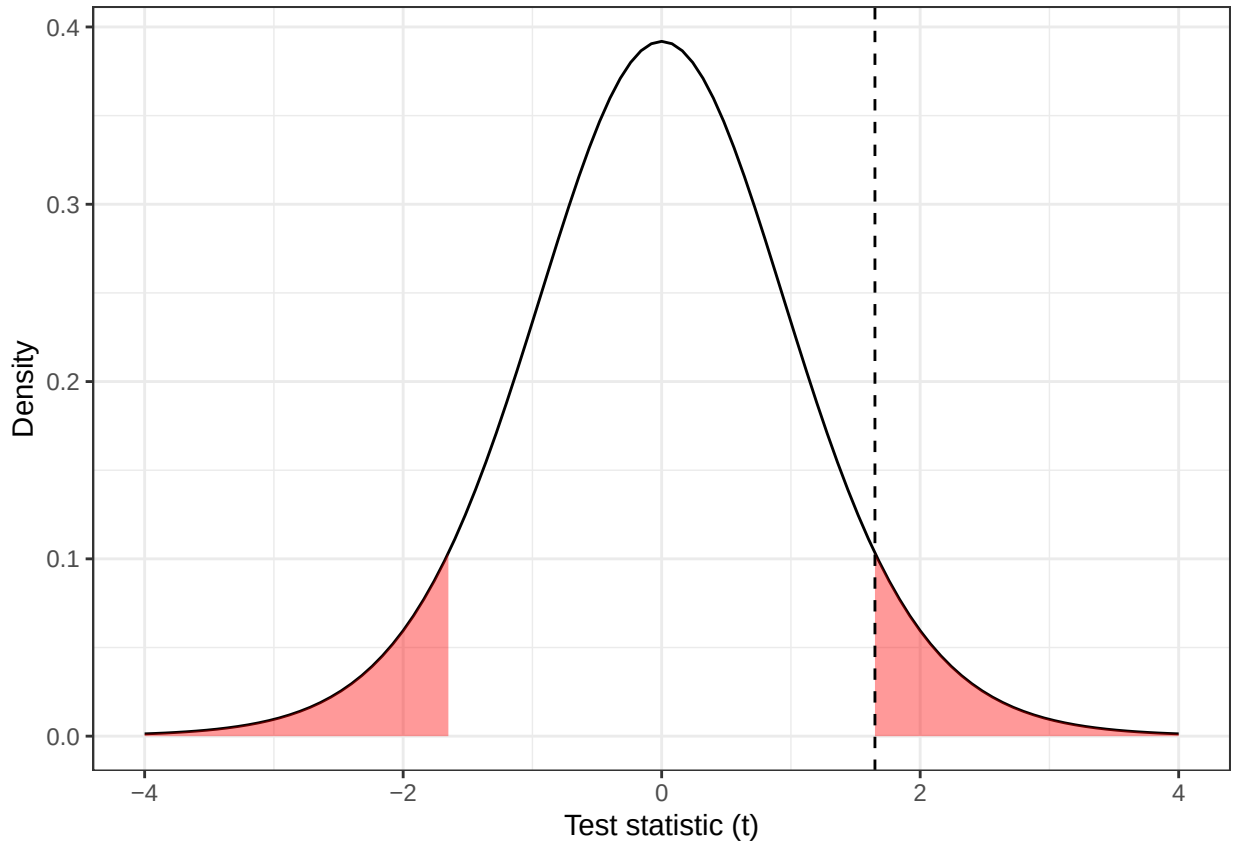
$$\begin{aligned} s_p &= \sqrt{\frac{8 \times 8.43^2 + 6 \times 6.94^2}{8 + 6}} \\ &= 7.82 \text{ mmHg}. \end{aligned}$$

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

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{6.53}{7.82 \sqrt{\frac{1}{9} + \frac{1}{7}}} \\ &= 1.65. \end{aligned}$$

Note that here we are taking the placebo group to be group 1, and the Captopril group to be group 2.

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



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

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

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

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

6.1.0.1 What do we do with this outcome?

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

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

$$\left\{ \tau \mid \frac{|\bar{x}_1 - \bar{x}_2 - \tau|}{s\sqrt{n_1^{-1} + n_2^{-1}}} \leq t_{n_1+n_2-2;\,0.975} \right\},$$

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

$$\left\{ \tau \mid \bar{x}_1 - \bar{x}_2 - t_{n_1+n_2-2;\,0.975}\, s\sqrt{n_1^{-1} + n_2^{-1}} \leq \tau \leq \bar{x}_1 - \bar{x}_2 + t_{n_1+n_2-2;\,0.975}\, s\sqrt{n_1^{-1} + n_2^{-1}} \right\}$$

Example 6.3. Continuing our example, we have

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

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

Rearranging as above, this works out to be the interval

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

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

Some things to note

- We can compute this confidence interval whether or not we failed to reject the null hypothesis that $\tau = 0$, and for significance levels other than 0.05.

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

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

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

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

6.1.1 Bonferroni correction

(and possibly other types of adjustment? - where to put this?!)

6.2 Using baseline values

In our example above, our primary outcome variable X was the systolic blood pressure of each participant at the end of the intervention period. However, we see in Table 6.1 that we also have *baseline* measurements: measurements of systolic blood pressure for each patient from before the intervention period. Baseline measurements are useful primarily for two reasons:

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

We will demonstrate these by returning to our Captopril example.

Example 6.4. Firstly, we use the baseline systolic blood pressure to assess balance. The placebo group has a mean of 146.6 mmHg and an SD of 12.3 mmHg, whereas the Captopril group has mean 148.0 mmHg, SD 11.4 mmHg. While these aren't identical, they are sufficiently similar to think they wouldn't affect our analysis. In a study this small there is likely to be some difference.

Secondly, since we are interested in whether the use of Captopril has reduced blood pressure for each individual, and these individuals had different baseline values, it makes sense to compare not just the outcome but the difference from baseline to outcome for each individual. We can see individual data in Table 6.3 and summary statistics in Table 6.4.

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

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

Table 6.3: Data for the Captopril trial from @hommel1986effect, with differences shown.

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

Table 6.4: Summary statistics for each group.

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

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

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

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

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

Let's label the baseline measurement for each group B_1 and B_2, and the outcome measurements X_1, X_2, where we will take group 1 to be the placebo group and group 2 to be the treatment group. Because all participants have been randomised from the same population, we have

$$\mathrm{E}\,(B_1) = \mathrm{E}\,(B_2) = \mu_B.$$

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

$$\begin{aligned} \mathrm{E}\,(X_1) &= \mu \\ \mathrm{E}\,(X_2) &= \mu + \tau. \end{aligned}$$

Usually we will assume that

$$\mathrm{Var}\,(X_1) = \mathrm{Var}\,(X_2) = \mathrm{Var}\,(B_1) = \mathrm{Var}\,(B_2) = \sigma^2,$$

and this is generally fairly reasonable in practice.

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

$$\begin{aligned} \mathrm{E}(X_2) - \mathrm{E}(X_1) &= (\mu + \tau) - \mu = \tau \\ \mathrm{E}(X_2 - B_2) - \mathrm{E}(X_1 - B_1) &= (\mu - \mu_B + \tau) - (\mu - \mu_B) = \tau, \end{aligned}$$

that is, both are unbiased estimators of τ.

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

$$\begin{aligned} \mathrm{Var}(X_2 - B_2) &= \mathrm{Var}(X_2) + \mathrm{Var}(B_2) - 2\,\mathrm{cov}(X_2, B_2) \\ &= \sigma^2 + \sigma^2 - 2\rho\sigma^2 \\ &= 2\sigma^2(1 - \rho), \end{aligned}$$

where ρ is the true correlation between X and B, and is assumed to be the same in either group. Therefore, if $\frac{1}{2} < \rho \leq 1$ there will be a smaller variance when comparing differences. However, if $0 \leq \rho < \frac{1}{2}$, the variance will be smaller when comparing outcome variables.

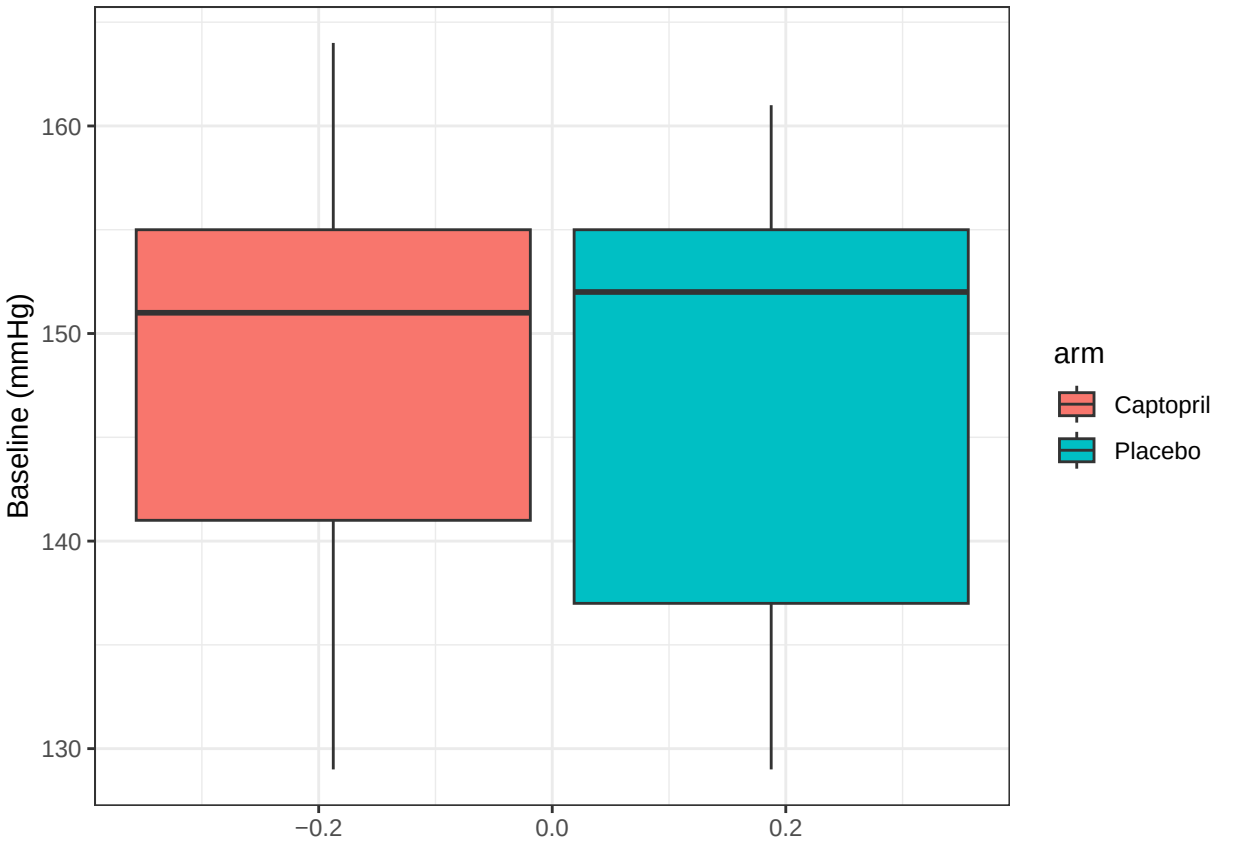
Intuitively, this seems reasonable: if the correlation between baseline and outcome measurements is very strong, then we can remove some of the variability between participants by taking into account their baseline measurement. However, if the correlation is weak, then by including the baseline in the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

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

6.3 Analysis of covariance (ANCOVA)

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

However, although this is theoretically true, in real life trials there will be some imbalance in the baseline measurements for the different treatment arms. We can see this in our Captopril example, in Figure **??**.2.1



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

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

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

6.3.1 The theory

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

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

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

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

From this, we find that

$$\mathrm{E}\left(\bar{X}_T - \bar{X}_C \mid \bar{b}_T, \bar{b}_C\right) = \tau + \rho\left(\bar{b}_T - \bar{b}_C\right). \tag{6.1}$$

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

Comparing the difference between outcome and baseline, as we did in 6.2, does not solve this problem, since we have

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

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

Notice, however, that if we use as our estimator

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

then, following from Equation (6.1) we have

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

6.3.2 The practice

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

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

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

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

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

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

where G_i is 1 if participant i is in group T and 0 if they're in group C. This is a factor variable, which you may remember from Stats Modelling II (if you took it). If $G_i = 1$ (ie. participant i is in group T) then τ is added. If $G_i = 0$ (ie. participant i is in group C) then it isn't.

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

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

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

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

The general form for this is

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

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

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

The technique described above is a well-established statistical method known as **ANCOVA** (short for the **An**alysis of **Cova**riance), which can be implemented in R and many other statistical software packages.

Example 6.5. Let's now implement ANCOVA on our Captopril data in R. We do this by first fitting a linear model using 'lm', with baseline measurement and arm as predictor variables and outcome as the predictand.

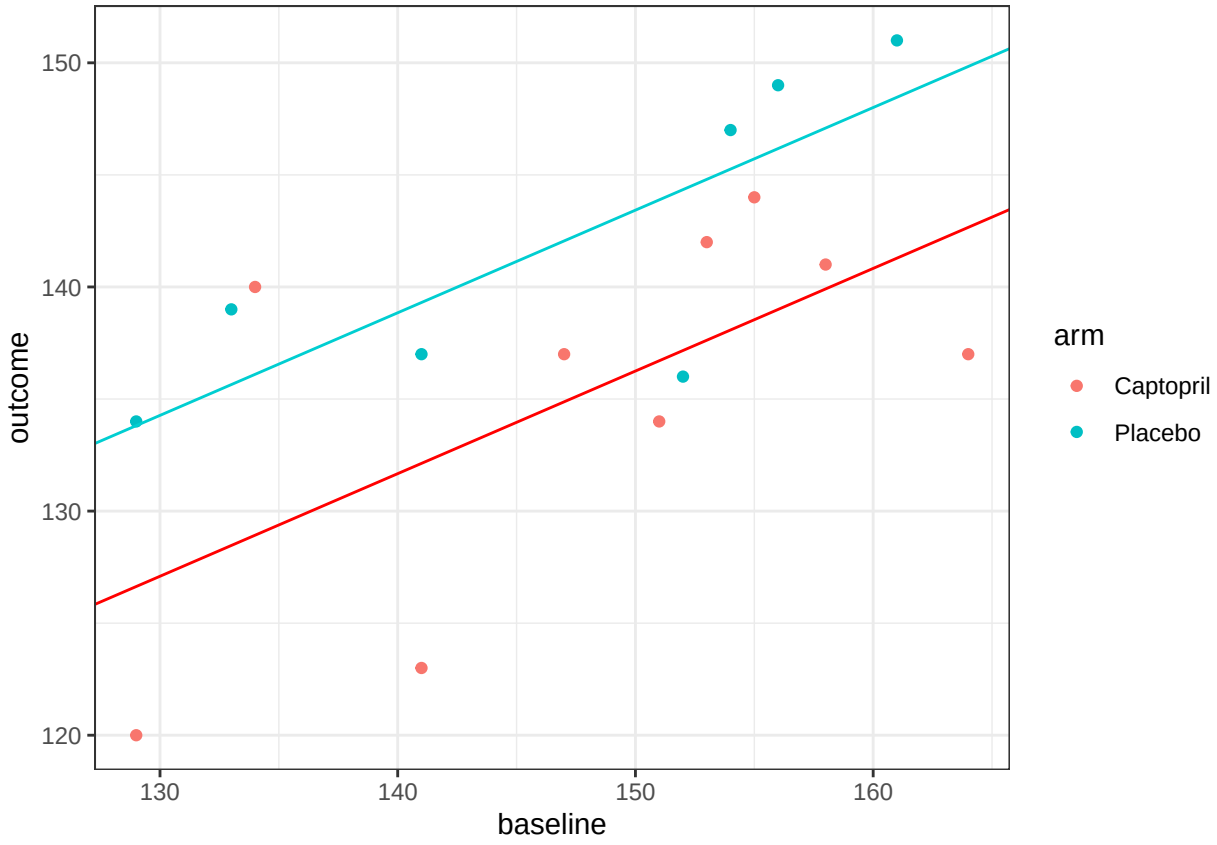
```
# Gives same answers as above, so use this!
# Find out what the numbesrs mean
# Background on ANOVA and factor models (SMII)
# Where to include random effects models? Is it in Matthews?
lm_capt = lm(outcome ~ baseline + arm, data = df_hommel)
summary(lm_capt)
```

```
##
## Call:
## lm(formula = outcome ~ baseline + arm, data = df_hommel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.129 -3.445  1.415  2.959 11.076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.5731    19.7577   3.420  0.00456 **
## baseline      0.4578     0.1328   3.446  0.00434 **
## armPlacebo    7.1779     2.9636   2.422  0.03079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.869 on 13 degrees of freedom
## Multiple R-squared:  0.5629, Adjusted R-squared:  0.4957
## F-statistic: 8.372 on 2 and 13 DF,  p-value: 0.004608
```

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

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

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



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

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

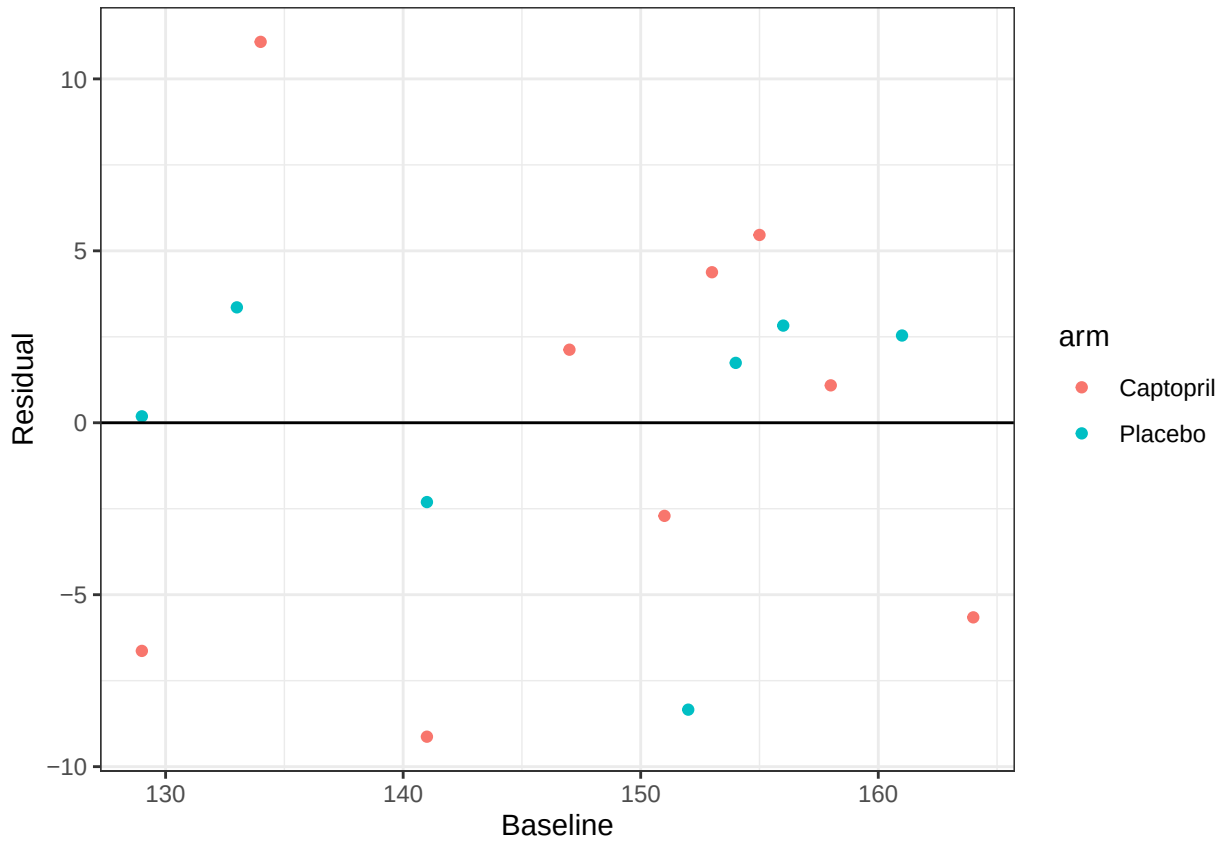
The model has $n - p = 13$ degrees of freedom because there are $n = 16$ data points and are estimating $p = 3$ parameters. Notice that unlike our previous confidence intervals, this doesn't contain zero, and so our analysis has enabled us to conclude that there is a significant reduction in blood pressure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

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

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

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

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



These look pretty good, no clear patterns and a similar distribution for each treatment group.

6.4 Some follow-up questions. . . .

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

6.4.1 How is this related to ANOVA?

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

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

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

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

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

Example 6.6. Continuing once again with the Captopril dataset, we now fit the model

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

```
##
## Call:
## lm(formula = outcome ~ arm + baseline + baseline:arm, data = df_hommel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.094 -3.475  1.412  2.979 11.145
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          66.85150   28.02488   2.385   0.0344 *
## armPlacebo            8.72484   40.93465   0.213   0.8348
## baseline              0.46272    0.18886   2.450   0.0306 *
## armPlacebo:baseline -0.01051    0.27723  -0.038   0.9704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.108 on 12 degrees of freedom
## Multiple R-squared:  0.563,  Adjusted R-squared:  0.4537
## F-statistic: 5.153 on 3 and 12 DF,  p-value: 0.01614
```

```
anova(lm_capt_int) # still don't think I need to include this
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
arm	1	167.5803571	167.5803571	4.4918530	0.0555944
baseline	1	409.1120836	409.1120836	10.9659114	0.0062073
arm:baseline	1	0.0535866	0.0535866	0.0014363	0.9703913
Residuals	12	447.6914727	37.3076227	NA	NA

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

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

6.4.4 Can we include any other covariates?

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

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

and the inclusion of the coefficient γ means that we can include other covariates on different scales too. The key issue is that we can only include as covariates things that were already known before allocation. This is because they cannot, at that point, have been affected by the treatment. Indeed, as a rule, any variable that was used in the randomisation procedure (this particularly applies to minimisation and stratified sampling) should be included in the analysis.

Example 6.7. The data for this example is taken from Kassambara (2019). In this study, 60 patients take part in a trial investigating the effect of a new treatment and exercise on their stress score, after adjusting for age. There are two treatment levels (yes or no) and three exercise levels (low, moderate and high) and 10 participants for each combination of treatment and exercise levels. Because in ANCOVA we fit a coefficient to every covariate, we can include exercise (another factor variable) and age (a continuous variable) in this analysis.

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

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

treatment	exercise	mean	sd
yes	low	61.7	4.691600
yes	moderate	59.6	2.590581
yes	high	57.0	2.211083
no	low	62.1	4.332051
no	moderate	61.4	5.947922
no	high	57.9	3.381321

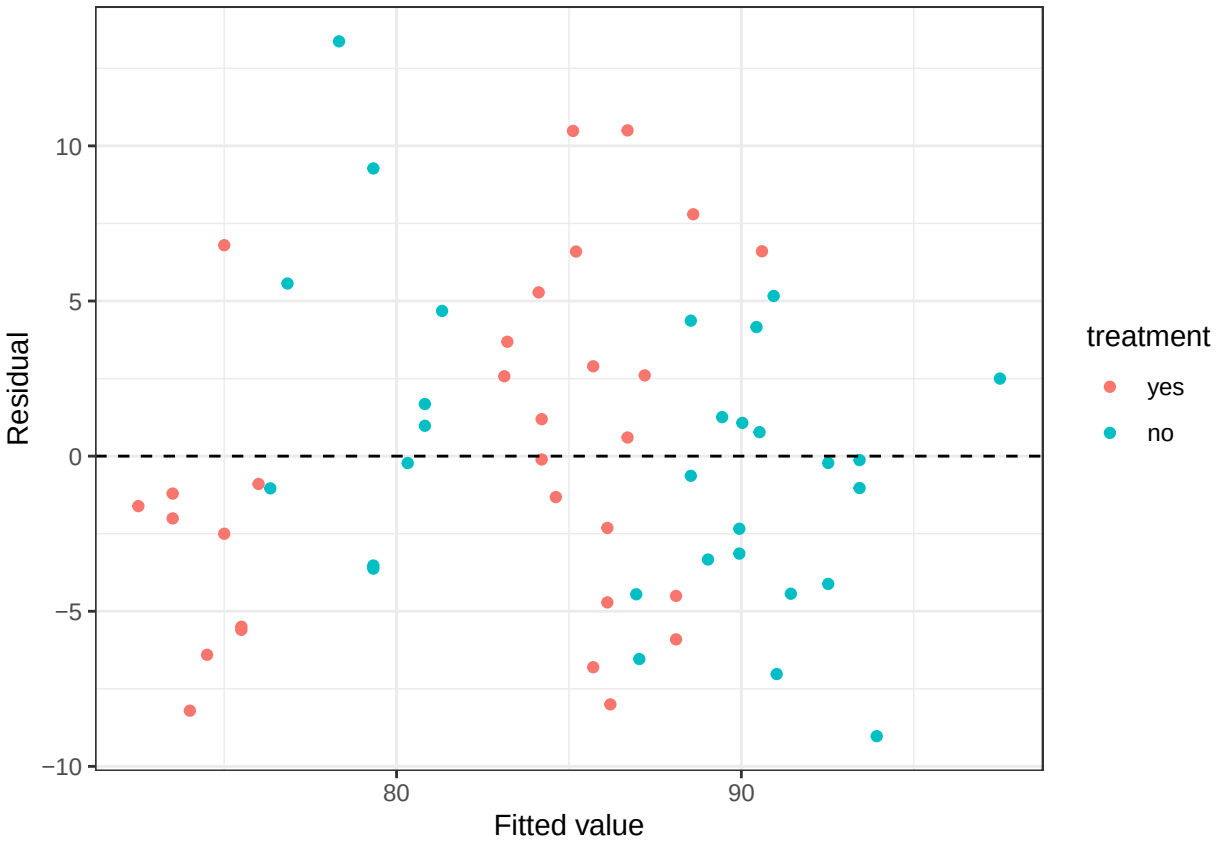
Fitting a linear model, we see that treatment, high levels of exercise and age have an effect on stress.

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

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

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

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



6.5 Some general principles of Analysis

There are some assumptions we're making here, and so we need to be careful when fitting an ANCOVA model.

- We're assuming that the residual variance is the same for both groups
- We're assuming that the coefficient of the baseline is the same for both groups: only the intercept is changing.

We can't check the first until after

Example 6.8. Before fitting our ANCOVA model we should first have checked that there's no significant relationship between the covariate (in this case the baseline) and the treatment group, which we can do using ANOVA:

```
# There's probably a better way to do this - check online articles
model_aov_capt = aov(baseline ~ arm, data = df_hommel)
summary(model_aov_capt)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## arm           1      8    8.04   0.058  0.814
## Residuals    14   1952  139.41
```

If the experiment has been properly designed then there shouldn't be, but this is not always the case, and indeed depending on how we do the randomisation we could be unlucky. If baseline measurements are available before allocation, this can be checked before the trial is run.

Bibliography

Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.

Altman, D. G. and Bland, J. M. (1999a). How to randomise. *Bmj*, 319(7211):703–704.

Altman, D. G. and Bland, J. M. (1999b). Treatment allocation in controlled trials: why randomise? *Bmj*, 318(7192):1209–1209.

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.

Hommel, E., Parving, H.-H., Mathiesen, E., Edsberg, B., Nielsen, M. D., and Giese, J. (1986). Effect of captopril on kidney function in insulin-dependent diabetic patients with nephropathy. *Br Med J (Clin Res Ed)*, 293(6545):467–470.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., and Newman, T. B. (2013). *Designing clinical research, fourth edition*. Lippincott Williams & Wilkins.

Kallis, P., Tooze, J., Talbot, S., Cowans, D., Bevan, D., and Treasure, T. (1994). Pre-operative aspirin decreases platelet aggregation and increases post-operative blood loss–a prospective, randomised, placebo controlled, double-blind clinical trial in 100 patients with chronic stable angina. *European journal of cardio-thoracic surgery: official journal of the European Association for Cardio-thoracic Surgery*, 8(8):404–409.

Kassambara, A. (2019). *datarium: Data Bank for Statistical Analysis and Visualization*. R package version 0.1.0.

Matthews, J. N. (2006). *Introduction to randomized controlled clinical trials*. CRC Press.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115.

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.

Zhong, B. (2009). How to calculate sample size in randomized controlled trial? *Journal of thoracic disease*, 1(1):51.