Clinical Trials 4H - lecture notes

Rachel Oughton

2024-01-16

Contents

Chapter 1

Introduction

This is just here to preserve the numbering!

Chapter 2

(Lecture 2) Sample size

For most of this course, our trial will have two arms and our unit of randomization will be individual participants. In this section we'll focus on continuous primary outcome variables.

Will go on to think about binary variables and time-to-event data.

The topics we'll cover fall into two categories:

- Before the trial - design and planning
- After the trial - analysis and communication

but there is some interaction between these phases.

The first big question asked of a trial statistician is usually **how many participants does the trial need in order to be viable?**

Can also be asked about the design itself - lots of different sorts of trials. But not always!

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

1. Not enough particpants may mean not enough evidence to come to a conclusion. This is both scientifically disappointing and 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. Too many patients *(ie. we would be sufficiently likely to reach a conclusion with many fewer)* means subjecting more patients than necessary to an inferior treatment. *Possibly also taken up more time and resources than was necessary.*

This has been quite woolly so far, but now we'll start to think more carefully.

2.1 The treatment effect

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

We base our sample size calculations on the primary outcome variable.

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

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

Before we can construct a method to calculate sample size, we need to think about what we'll do with the trial data once we have it, so we now have a brief-ish segue into hypothesis tests.

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

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

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

The p-value is the probability of obtaining whatever result (eg. treatment effect) we have have found simply by random chance, when in fact H_0 is true and there is no treatment effect (ie. $\tau = 0$). Generally, a P-value of $\alpha = 0.05$ is accepted as sufficient evidence to reject the null hypothesis, although in clinical settings it can often be smaller (eg. $\alpha = 0.01$). It is conventional to present the P-value by simply saying whether it is smaller than some threshold (often 0.05), rather than giving the exact value.

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

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

2.2.1 Insignificant results

If our P-value is 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.

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 our results are consistent with $H_0 : \tau = 0$, and also with some small treatment effect.

Key issue: what size of treatment effect do we care about?

Our sample size should be big enough to be sufficiently likely to detect a clinically meaningful treatment effect.

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

2.2.2 One-sided or two-sided?

The trial clinicians will have strong beliefs about the direction of the treatment effect. Assuming that a larger value of the primary outcome variable X is good, they will expect $\tau > 0$ (or be prepared to accept $\tau = 0$, no effect).

Therefore should we perform a one-sided test, with

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

ANNOTATE PLOT: Suppose our test statistic $\sim t_{31}$ *and we find* $t - 2$*, as shown in plot. Then* $p = 1 - F_t(2, df = 31) = 0.0272$ *(where* $F_t(\cdot)$ *is the cumulative distribution function of the* t *distribution) , and the result would be considered significant at the 0.05 level.*

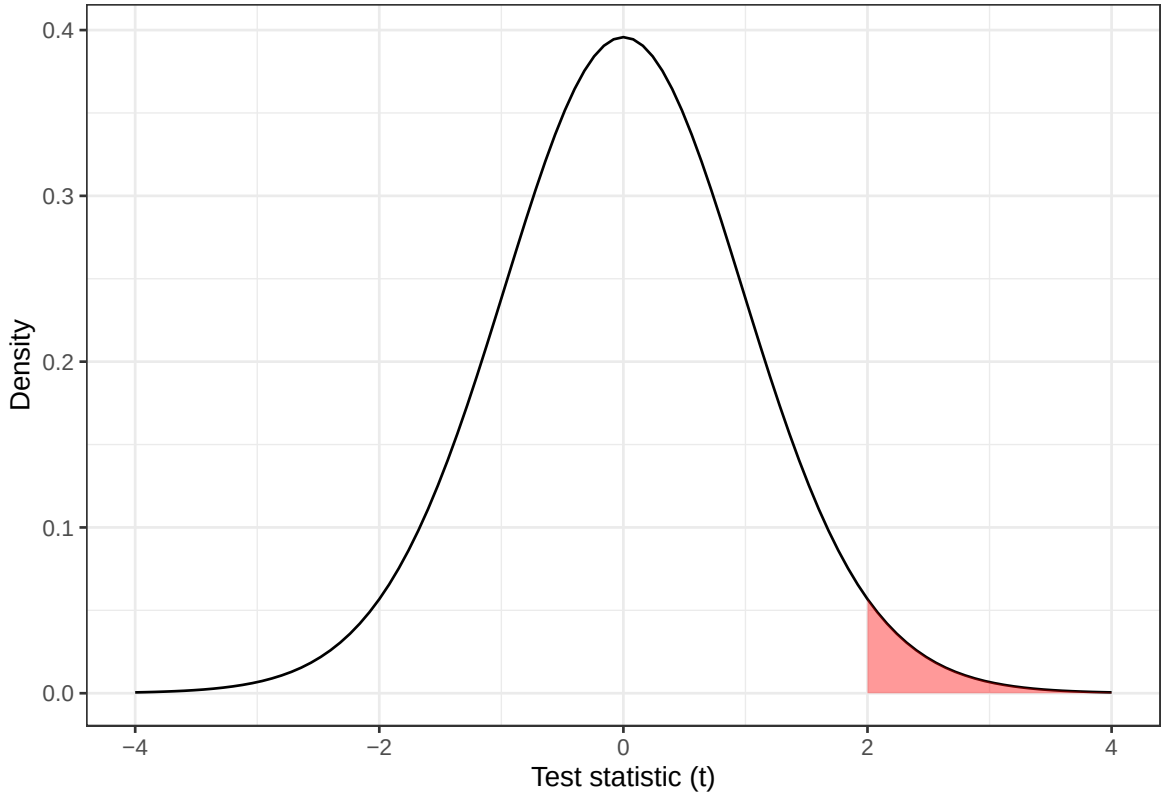


Figure 2.1: The distribution t_{31}, with the area corresponding to $t > 2$ shaded.

If $t \gg 0$, we obtain a small P-value, and reject H_0. Conclusion: the intervention is effective (in a good way). But what if we obtain $t \ll 0$? 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 way to conclude that an intervention has a significantly negative effect.*

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

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

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

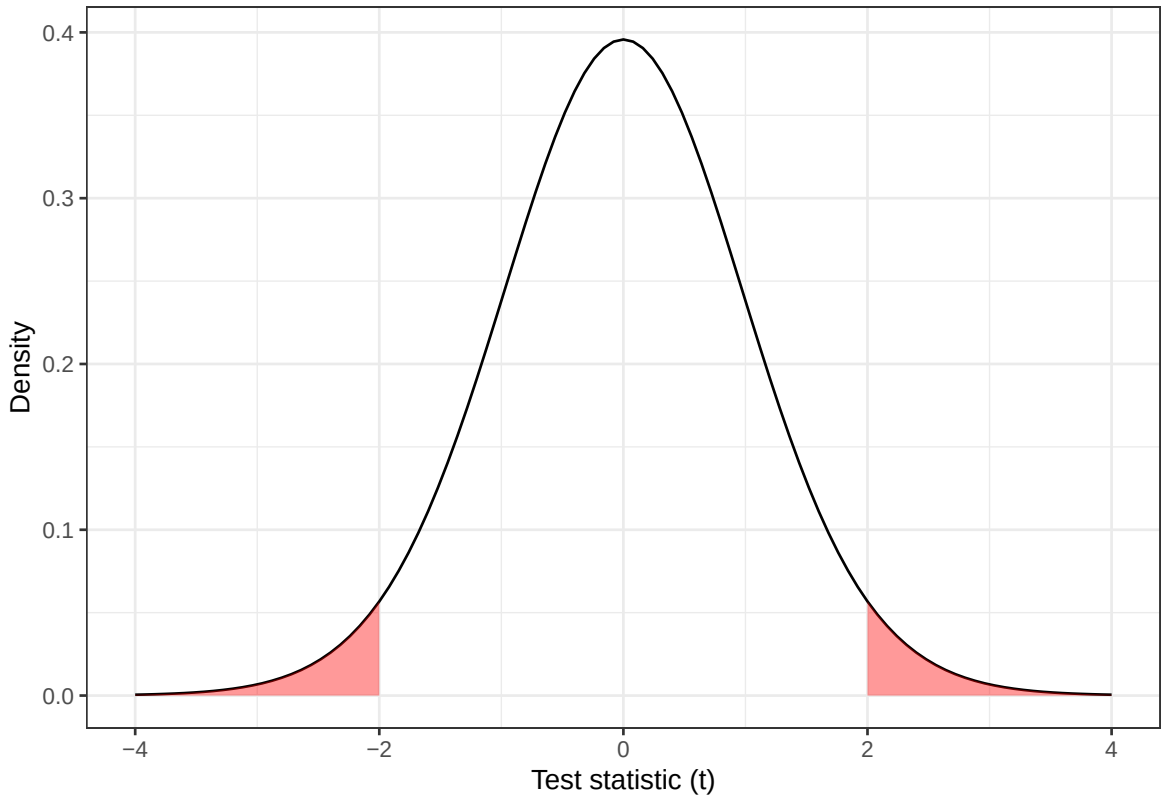


Figure 2.2: The distribution t_{31}, with the area corresponding to $|t| > 2$ shaded.

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

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

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

2.3 Constructing a measure of effect size

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

Talk about blinding - should really have A and B, and statistician not know which is T and C. This is for simplicity and clarity.

Suppose we have n patients in group C, and m in group T, and

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

The primary outcome variable X is normally distributed with mean μ in group C (the control group) and mean $\mu + \tau$ in group T (the intervention group), and common standard deviation σ. We will use X for the primary outcome variable

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

Using the trial data we find sample means $\bar{x}_C$ and $\bar{x}_T$ from each group, and a pooled estimate of the standard deviation

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

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

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

Using these values we can compute

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

as a standardised measure of the effect τ.

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

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

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

and therefore

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

and

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

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

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

and so we have

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

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

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

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

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

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

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

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

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

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

Generally we approximate this with a normal distribution (since n and m are usually sufficiently large).

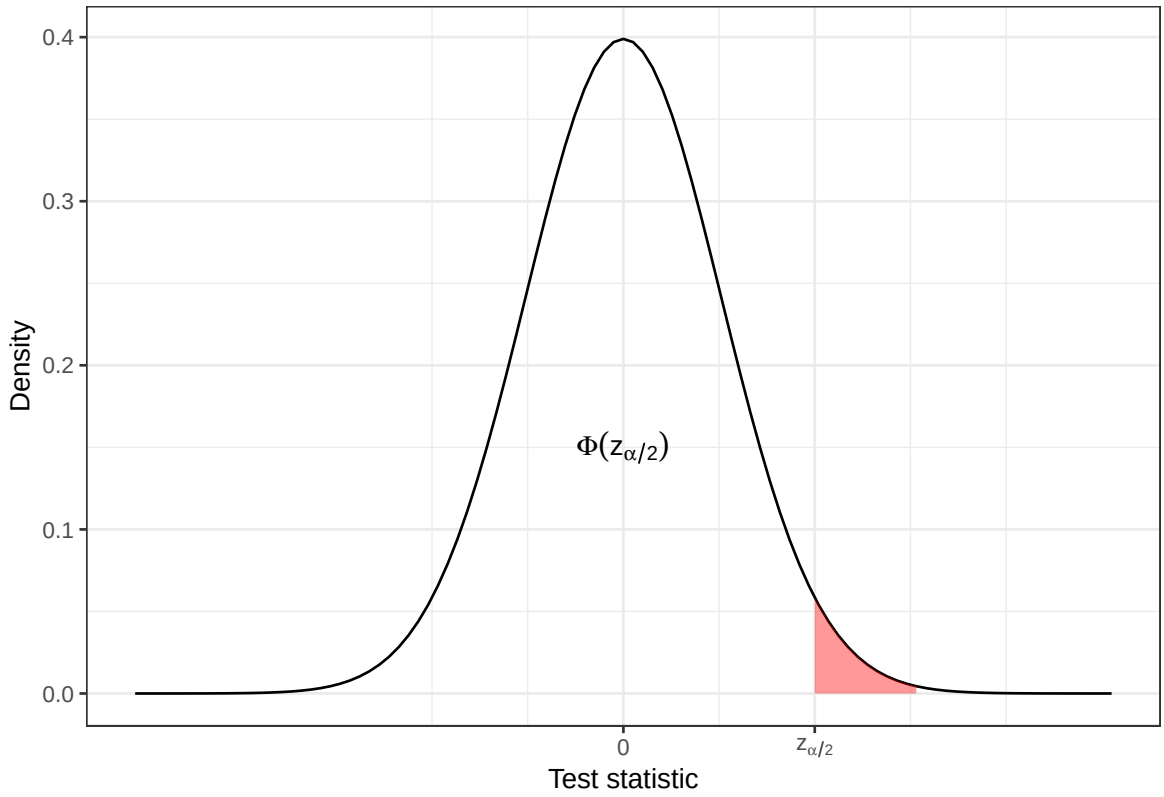
So, if we have run a trial, and have obtained n values of X from group C and m values of X from group T, we can compute D. If D lies outside the interval $\left[-z_{\alpha/2}, z_{\alpha/2}\right]$ then we reject H_0.

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

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

Brief aside on notation

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



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

We have constructed our whole argument under the assumption that H_0 *is true, and that the probability of such a value is therefore* 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?*

Our argument is based on H_0 being true - but what if it isn't?

Lecture 3

Recap:

- We constructed a measure $D = \frac{\bar{x}_T - \bar{x}_C}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$ that we can use to test $H_0 : \tau = 0$, since under H_0, $D \sim t_{n+m-2}$ (and approximately $D \sim N(0,1)$).

2.4 Power: If H_0 is false

So far, if H_0 is true, we have a small probability of rejecting H_0 (type I error rate).

Flip side: if H_0 is false, and $\tau \neq 0$, we want a high probability of rejecting H_0.

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

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

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

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

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

Under H_1, we have (approximately)

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

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

$$D = \frac{\bar{x}_T - \bar{x}_C}{s\lambda(n,m)}.$$

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 have set $\alpha = 0.05$ here), so under H_0, we have a 0.95 probability of observing a value of D that is consistent with H_0.

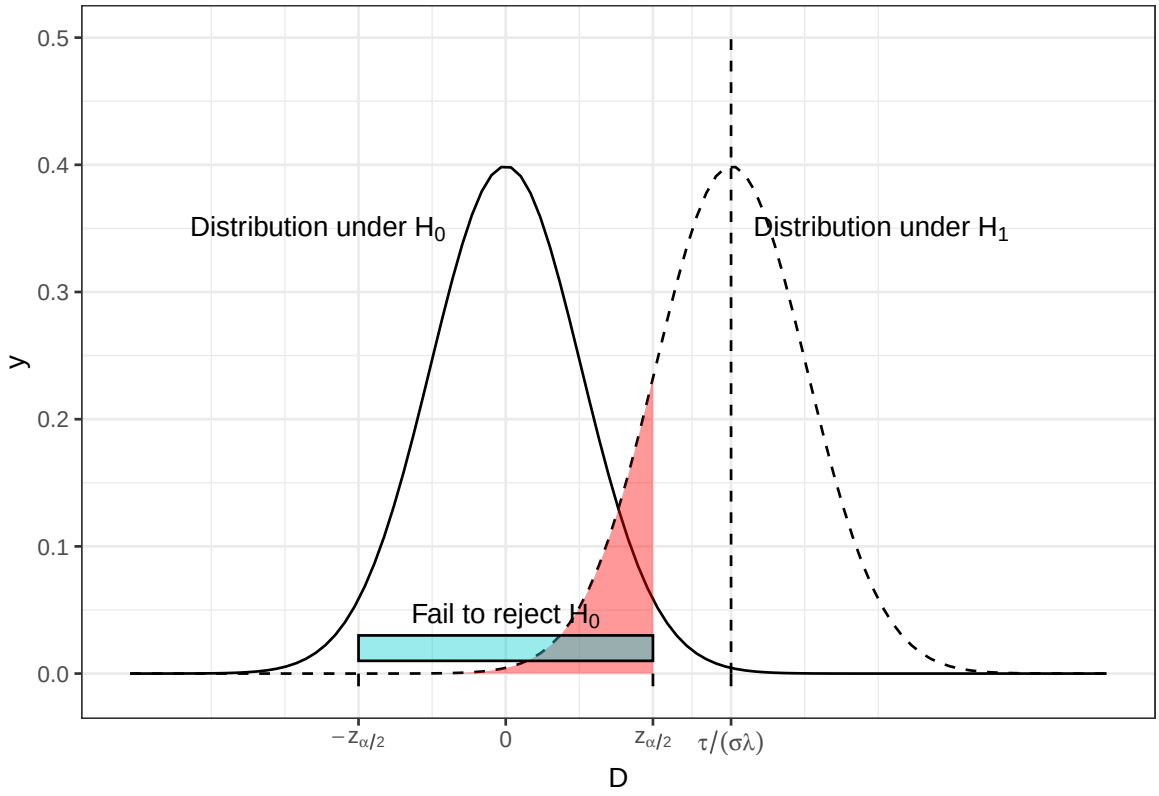


Figure 2.3: The distribution of D under both H_0 and H_1 for some arbitrary values of treatment effect, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

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

We can see that if the distributions have better separation, as in Figure 2.3, the power becomes greater. This can be as a result of a larger τ, a smaller σ or a smaller λ (therefore larger m and/or n).

[FIGURE!!]

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

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

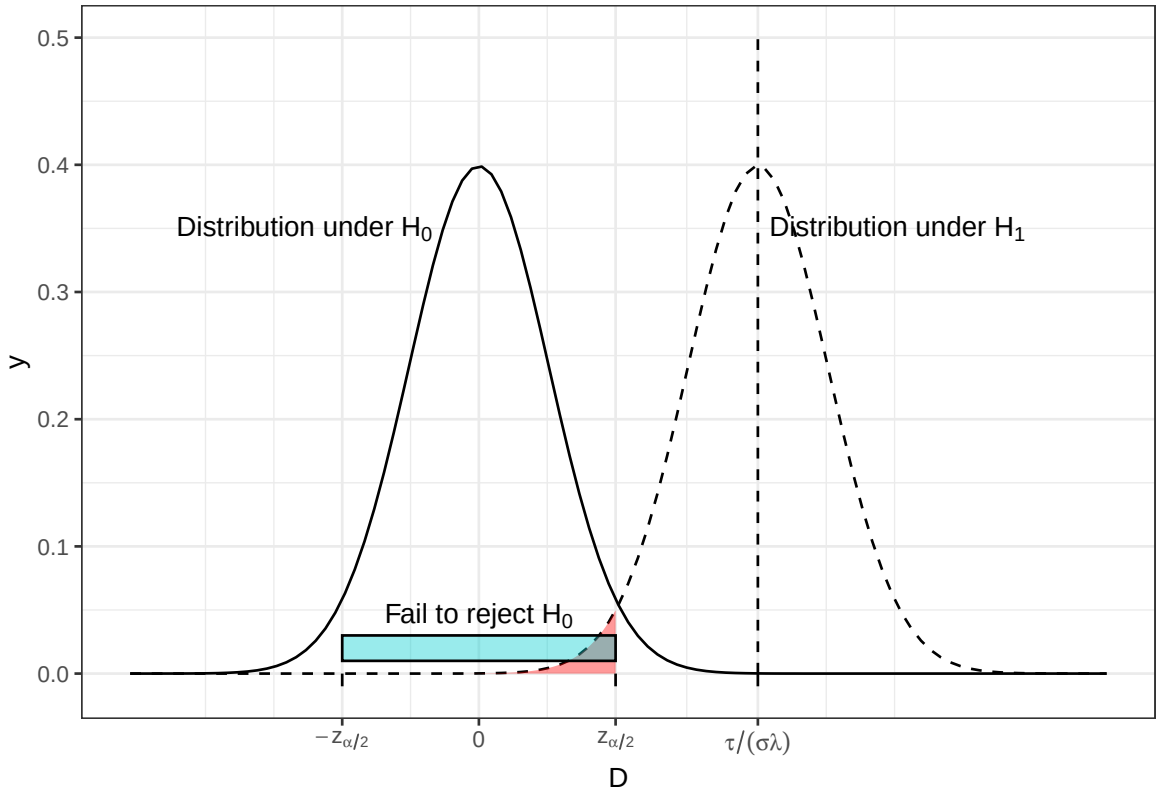


Figure 2.4: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

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

- Larger sample size → greater power
- Equal groups → greater 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.

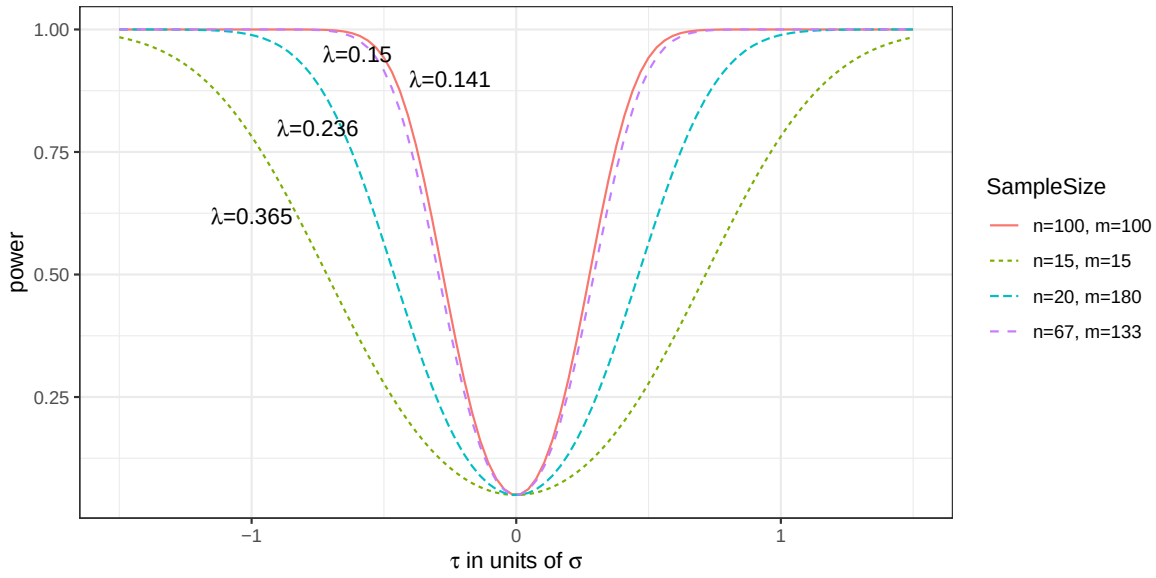


Figure 2.5: Power curves for various values of n and m, with effect size in units of standard deviation, given a type I error rate of 0.05.

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.

2.5 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.

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

Values for α and β are often specified by those planning the trial as around $\alpha \in [0.01, 0.05]$, $1-\beta \in [0.8, 0.9]$.

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

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

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

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

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

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

Therefore, Equation (2.1) becomes

$$\Psi(\tau) = 1-\beta = \left[1-\Phi\left(z_{\frac{\alpha}{2}} - \frac{\tau_M}{\sigma\lambda}\right)\right]. \tag{2.3}$$

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

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

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

$$\begin{aligned} z_\beta &= \frac{\tau_M}{\sigma\lambda} - z_{\frac{\alpha}{2}} \\ z_\beta + z_{\frac{\alpha}{2}} &= \frac{\tau_M}{\sigma\lambda}. \end{aligned} \tag{2.4}$$

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

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

and Equation (2.4) rearranges to

$$N = \frac{2\sigma^2 \left(z_\beta + z_{\frac{\alpha}{2}}\right)^2}{\tau_M^2}. \tag{2.5}$$

Example 2.1. (from Zhong 2009) A trial is being planned to test whether there is a difference in the efficacy of ACEII antagonist (a new drug) and ACE inhibitor (the standard drug) for the treatment of primary hypertension (high blood pressure). The primary outcome variable is change in sitting diastolic blood pressure (SDBP, mmHg) compared to a baseline measurement taken at the start of the trial. The trial should have a significance level of $\alpha = 0.05$ and a power of $1 - \beta = 0.8$, with the same number of participants in each group. The minimum clinically important difference is $\tau_M = 3$ mmHg and the pooled standard deviation is $s = 8$ mmHg. Therefore, using equation (2.5) the sample size should be at least

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

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

Chapter 3

(Lecture 4) Allocation

[Finished with sample size for now - check out JAMAevidence: JAMA Guide to Statistics and Methods Interviews about the statistical and methodological foundations of clinical research. Esp sample size, linked from Ultra]

Once we've decided how many participants we need in our trial, and they've been recruited, we next need to determine which participants should be assigned to which trial arm. This is process is known as **allocation** (or sometimes as **randomization**).

Before we think about methods for allocation, we are going to spend some time talking about bias.

3.1 Bias

In statistics, *bias* is a systematic tendency for the results of our analysis to be different from the true value, eg. when 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 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.

Recall the standard deviation. If we have some data $x_1, \ldots, x_n$ that are IID $N\left(\mu, \sigma^2\right)$, we can find the sample variance

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

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

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

We can apply this correction to produce an unbiased estimate of σ^2.

Now, suppose our sample $x_1, \ldots, x_n$ were drawn from $N\left(\mu, \sigma^2\right)$, 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 * 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 is much closer to what we mean when we talk about *bias* in a clinical trial setting.

Suppose we are testing some new treatment T against the standard C. We measure some outcome X for each patient, and our hypothesis is that X behaves differently for those in the treatment group than for those in the control group. It is common practice to express this additively,

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

where τ is our treatment effect, which we can estimate using the difference in the groups' means, $\bar{X}_T - \bar{X}_C$. Our null hypothesis is that $\tau = 0$, and our alternative hypothesis is that $\tau \neq 0$, and therefore an estimate of τ from our data is very important!

Clinical trials are all about estimating the treatment effect τ, so it is important that there is no bias in our estimates of $\bar{X}_C$ and $\bar{X}_T$.

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

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

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

3.1.1 Sources of bias

Bias is very serious - where does it come from? Most sources of bias creep in during the selection or allocation.

Selection bias

Certain patients are systematically more (or less) likely be entered into the trial because of the treatment they will receive.

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

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

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

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

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

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

and we assume $\mu_1 > \mu_2$.

For untreated patients we have

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

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

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

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

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

Similarly, $E(X \mid S) = \mu + \tau_S$ and $\tau = \tau_N - \tau_S$ 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.

Suppose a reduced proportion $q = 1 - p$ of those with severity 2 assigned to surgery enter the trial (event A):

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

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

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

We can split these according to disease severity, so that

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

and similarly for group S.

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

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

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

Following the same process for group S, we arrive at

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

which we will call b.

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

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

and

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

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

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

where the third term represents the bias.

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

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

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

If N is the better treatment, then $\tau_N - \tau_S > 0$ and the bias will cause the trial to underplay the treatment effect. Conversely, if S is better, then $\tau_N - \tau_S < 0$ and the trial will exaggerate the treatment effect.

This is because more severely ill patients have been assigned to N than to S, which reduces the average survival time for those in group N.

Allocation bias

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

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

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

Assessment bias

Measurements are made on participants throughout and during the trial.

Often objective: eg. weight, or concentration of blood sugar. Some measurments are subject to the individual practitioner assessing the patient. Eg, 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. Blood pressure used to rely on pracitioner's hearing and judgement.

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

Slight aside: publication bias

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

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

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

3.1.2 Implications for allocation

Clinical trials haven't always used random allocation to assign participants to groups. Some popular alternatives:

- 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.

Two important aspects to the allocation being *random*:

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. Eg. strict alternation ($ABABAB\ldots$), using 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.

[ASSIGNMENT! Explain details a bit, deadline is Monday 29th Jan.]

3.2 (Lecture 5) Allocation methods

3.2.1 Simple random allocation

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

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

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

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

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

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

The probability that $N_C = N_T = n$ is

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

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

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

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

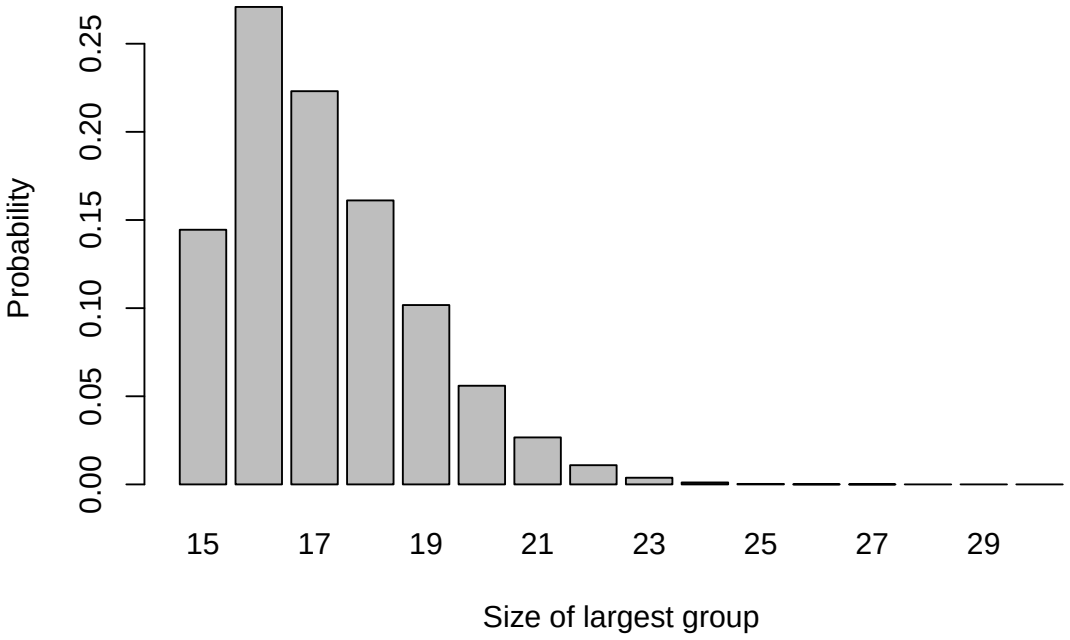


Figure 3.1: The probability distribution of largest group size for n=15.

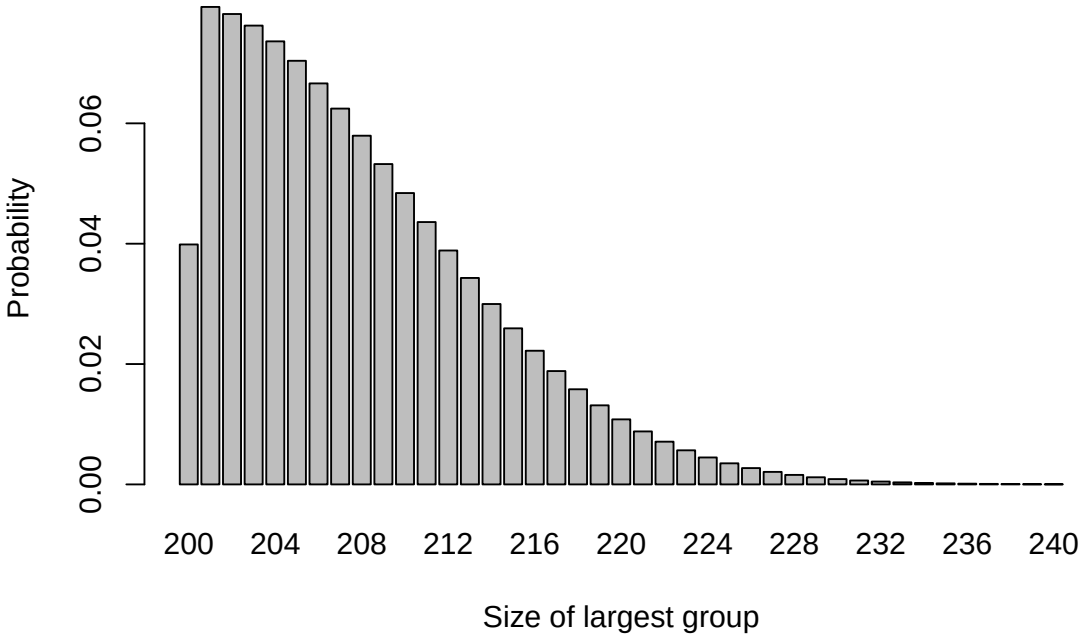


Figure 3.2: The probability distribution of largest group size for n=200.

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

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

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

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

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

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

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

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

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

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

3.2.2 Random permuted blocks

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

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

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

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

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

We can also randomly generate a sequence of numbers from $\{1, 2, 3, 4, 5, 6\}$, where each number has equal probability. This sequence will correspond to a sequence in A and B with four times the length.

In this method, each patient is equally likely to receive A and B, but there will never be a difference of more than two between the size of the two groups.

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

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

3.2.2.1 RPBs with random block length

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

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

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

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

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

3.2.3 Biased coin designs and urn schemes

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

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

3.2.3.1 Biased coin designs

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

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

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

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

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

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

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

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

or

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

Because of the way we have set up the scheme,

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

and

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

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

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

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

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

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

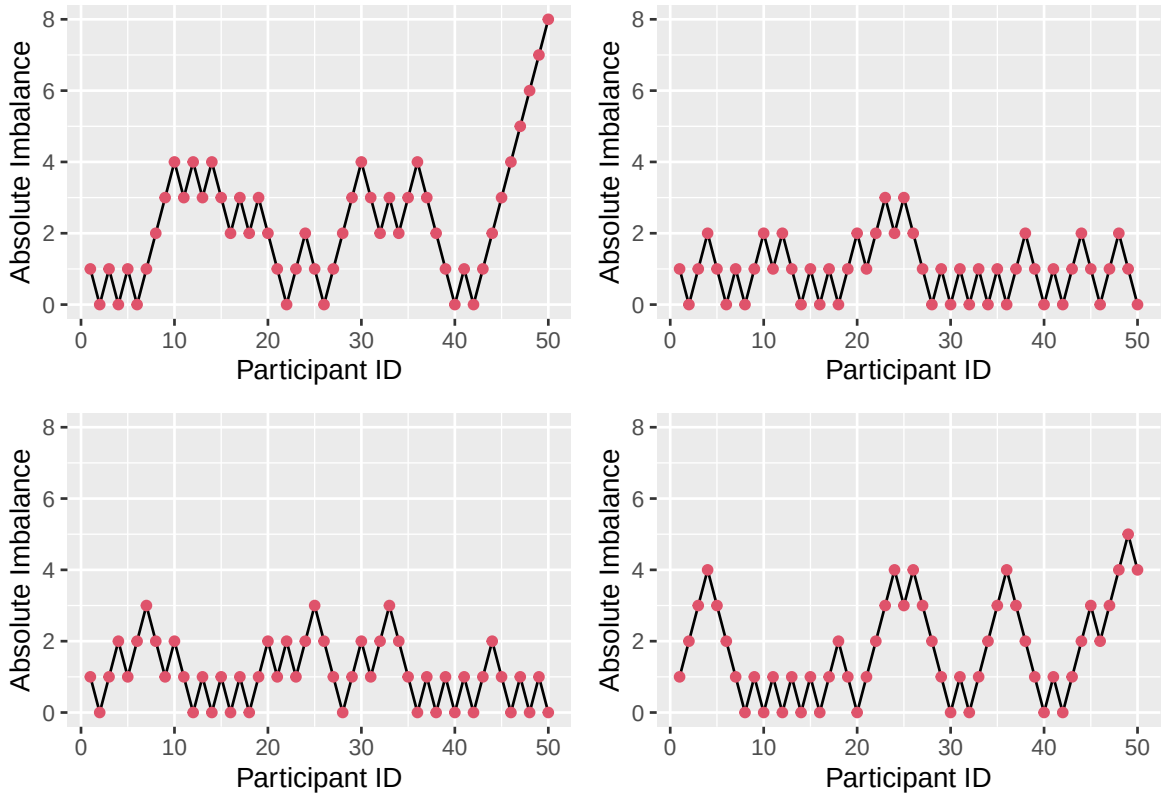


Figure 3.3: Absolute imbalance for a biased-coin scheme with $P = 0.667$.

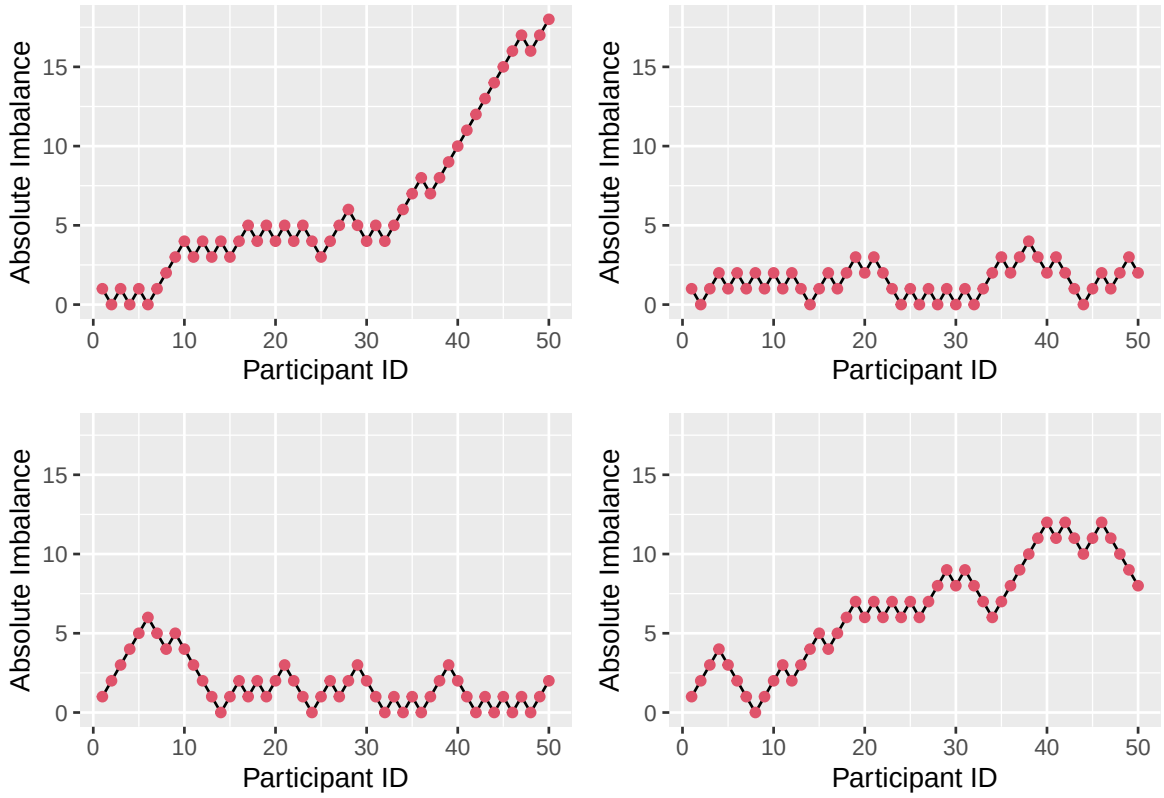


Figure 3.4: Absolute imbalance for a biased-coin scheme with $P = 0.55$.

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

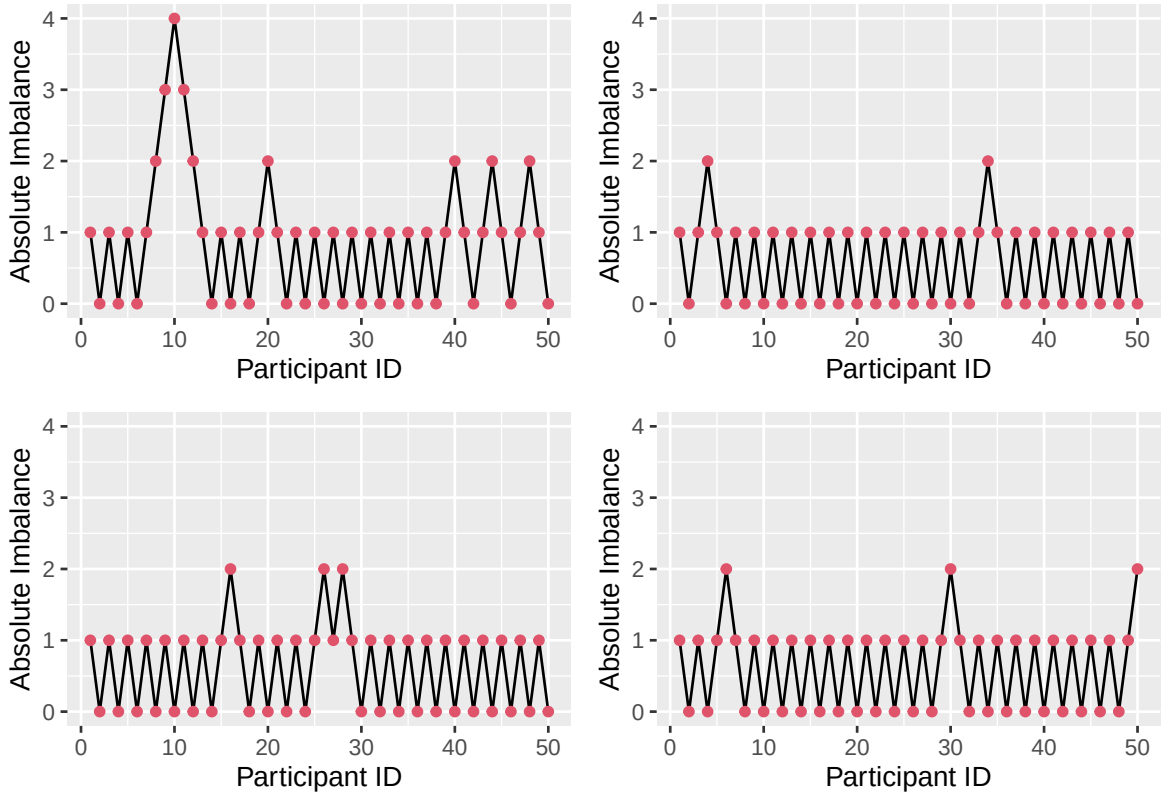


Figure 3.5: Absolute imbalance for a biased-coin scheme with $P = 0.9$.

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

3.2.3.2 Urn models

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

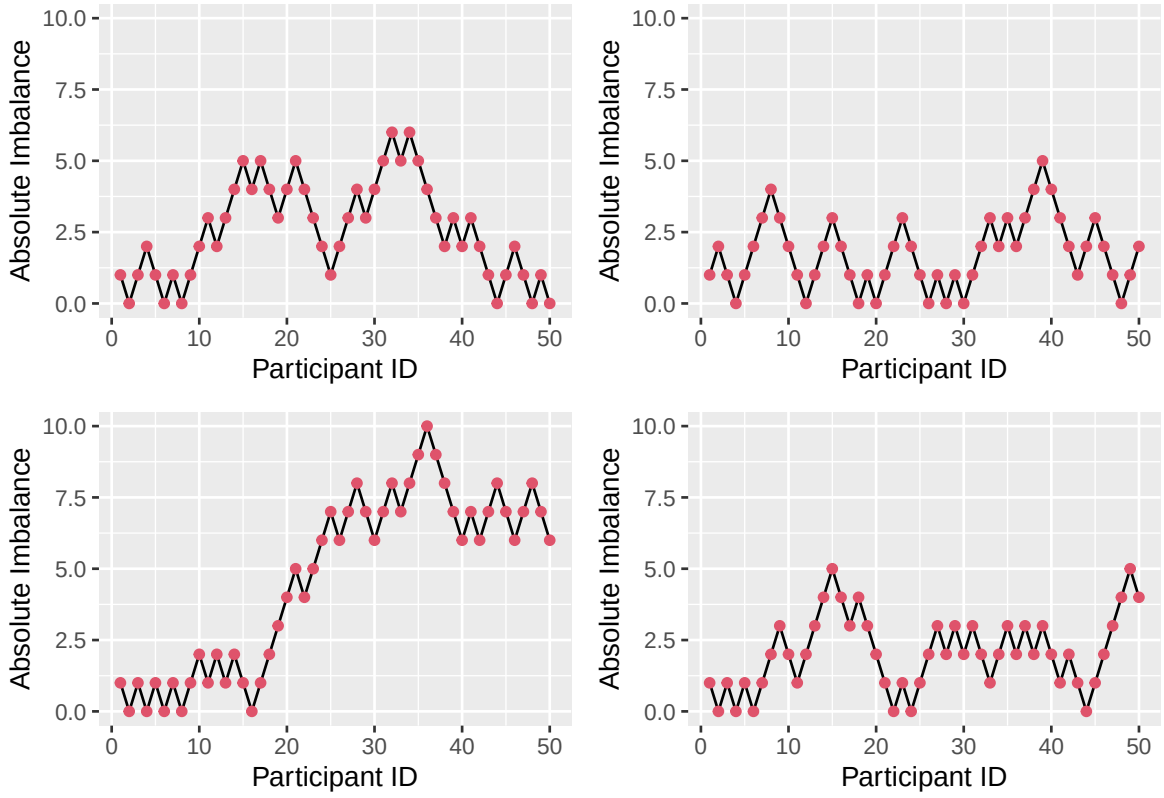


Figure 3.6: Four realisations of absolute imbalance for r=1, s=1, N=50.

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

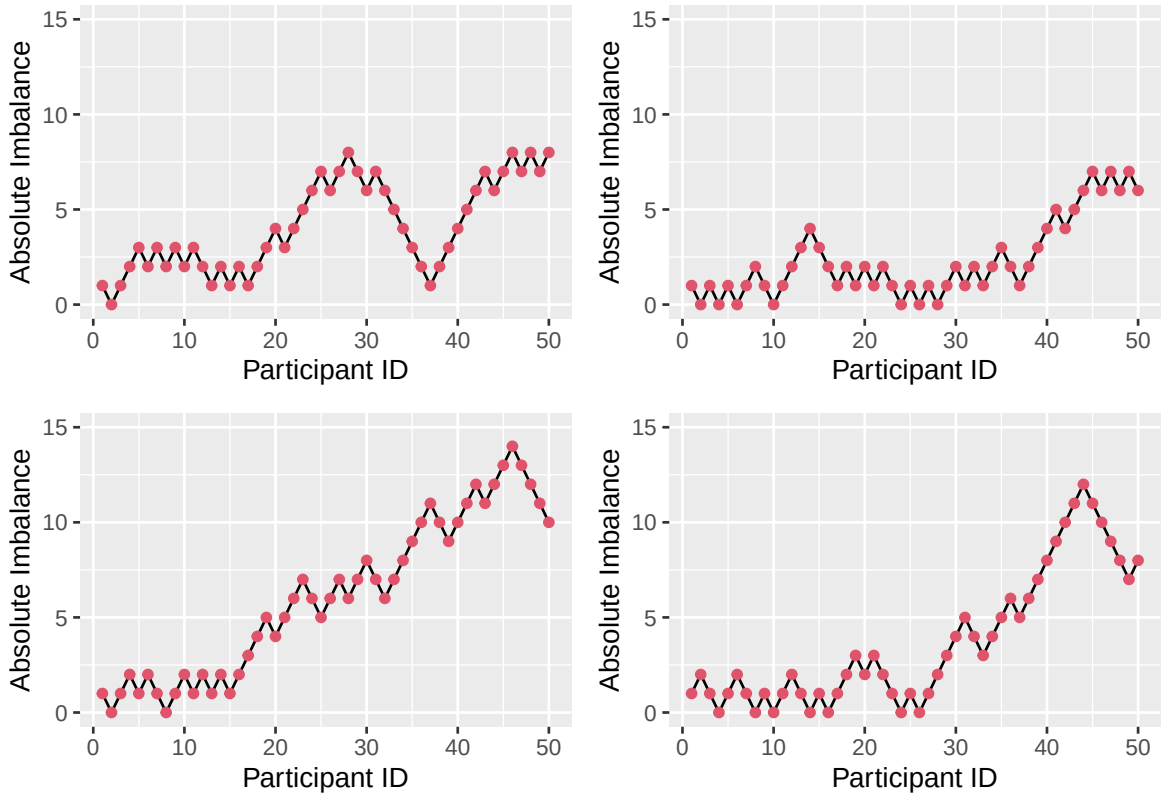


Figure 3.7: Four realisations of absolute imbalance for r=1, s=8, N=50.

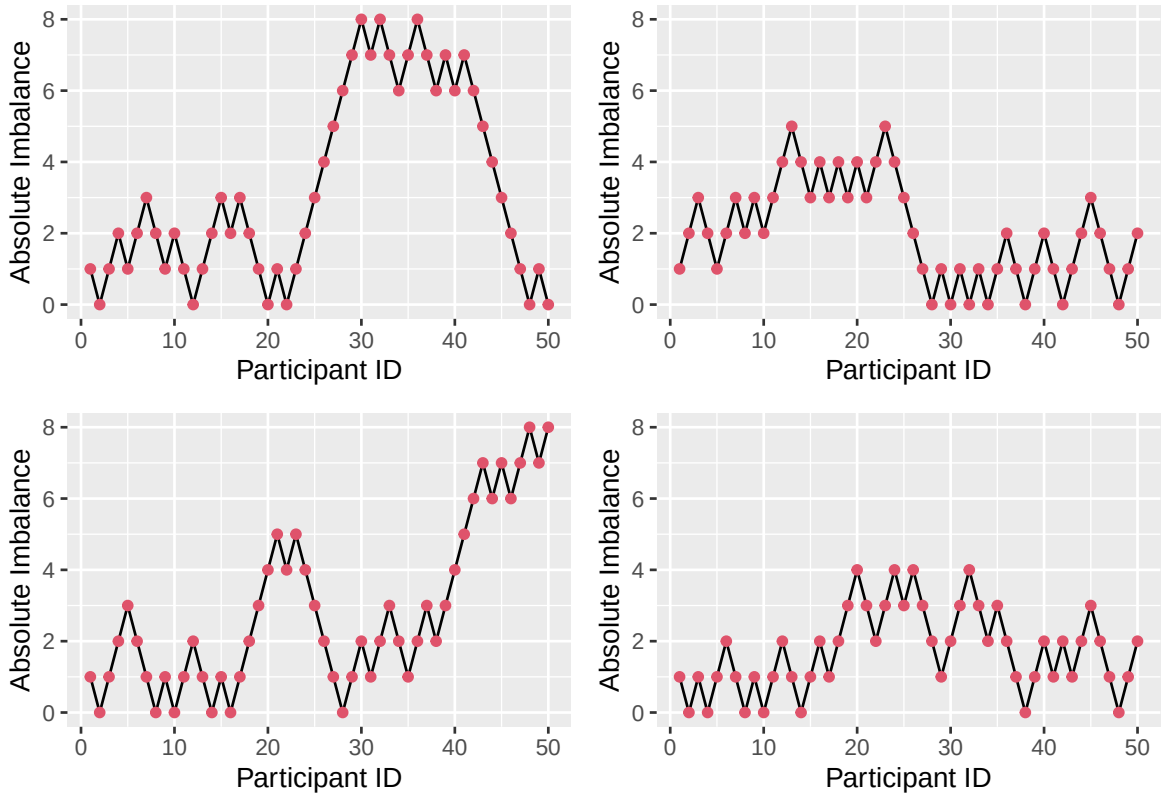


Figure 3.8: Four realisations of absolute imbalance for r=8, s=1, N=50.

3.3 (Lecture 6) Incorporating baseline measurements

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

Such baseline measurements can be used in allocation.

3.4 Stratified sampling

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

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

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

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

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

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

3.5 Minimization

Minimization was first proposed by Taves (1974), then shortly after by Pocock and Simon (1975) and Freedman and White (1976). The aim of minimization is to minimize the difference between the two groups. It was developed for 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.

3.5.1 Minimization algorithm

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

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

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

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

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

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

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

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

The first point is a pragmatic one - the method usually guaranteed to achieve good balance is likely to fail, at least for some strata. The second is more theoretical. In general, we require that groups be balanced according to each individual prognostic factor, but not to interactions. For example, it is

often believed that younger patients would have generally better outcomes, but that other factors do not systematically affect this difference.

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

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

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

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

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

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

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

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

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

3.6 Problems with allocation

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

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

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

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

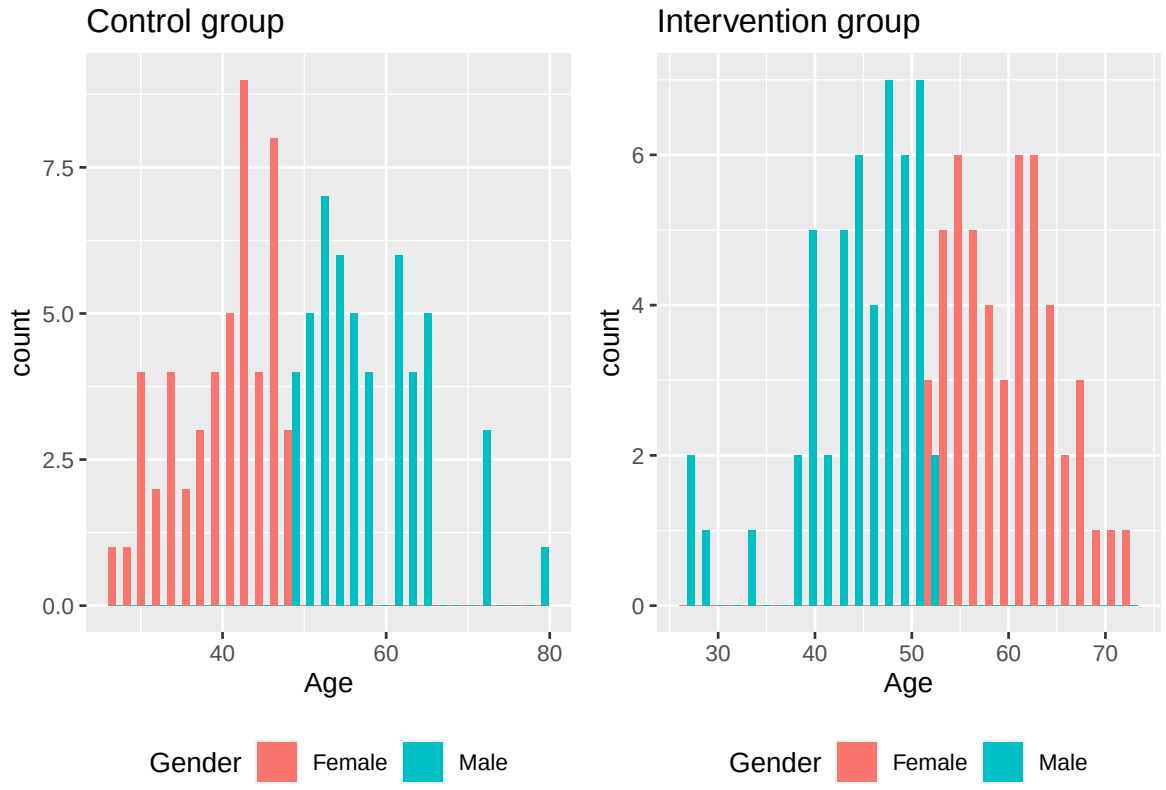
Figure 3.9: Summary statistics for an RCT comparing a licorice gargle (the intervention) to a sugar-water gargle (the standard). From @ruetzler2013randomized

The problem here is that only the marginal distributions are compared for similarity. Consider the following (somewhat extreme and minimalistic) scenario. A study aims to investigate the effect of some

treatment, and to balance for gender and age in their allocation, resulting in the following summary table.

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

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



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

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

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

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

Altman, Douglas G. 1990. *Practical Statistics for Medical Research*. CRC press.

Cottingham, Marci D, and Jill A Fisher. 2022. "Gendered Logics of Biomedical Research: Women in US Phase i Clinical Trials." *Social Problems* 69 (2): 492–509.

Fentiman, Ian S, Robert D Rubens, and John L Hayward. 1983. "Control of Pleural Effusions in Patients with Breast Cancer a Randomized Trial." *Cancer* 52 (4): 737–39.

Freedman, LS, and Susan J White. 1976. "On the Use of Pocock and Simon's Method for Balancing Treatment Numbers over Prognostic Factors in the Controlled Clinical Trial." *Biometrics*, 691–94.

Health, National Institute of. 2023. "History of Women's Participation in Clinical Research." Office of Research on Women's Health. https://orwh. od. nih. gov/toolkit .... https://orwh.od.nih.gov/toolkit/recruitment/history.

Kallis, P, JA Tooze, S Talbot, D Cowans, DH Bevan, and T Treasure. 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): 404–9.

Kar, Sumit, Ajay Krishnan, Preetha K, and Atul Mohankar. 2012. "A Review of Antihistamines Used During Pregnancy." *Journal of Pharmacology and Pharmacotherapeutics* 3 (2): 105–8.

Pocock, Stuart J, and Richard Simon. 1975. "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial." *Biometrics*, 103–15.

Ruetzler, Kurt, Michael Fleck, Sabine Nabecker, Kristina Pinter, Gordian Landskron, Andrea Lassnigg, Jing You, and Daniel I Sessler. 2013. "A Randomized, Double-Blind Comparison of Licorice Versus Sugar-Water Gargle for Prevention of Postoperative Sore Throat and Postextubation Coughing." *Anesthesia & Analgesia* 117 (3): 614–21.

Taves, Donald R. 1974. "Minimization: A New Method of Assigning Patients to Treatment and Control Groups." *Clinical Pharmacology & Therapeutics* 15 (5): 443–53.

Treasure, Tom, and Kenneth D MacRae. 1998. "Minimisation: The Platinum Standard for Trials?: Randomisation Doesn't Guarantee Similarity of Groups; Minimisation Does." *Bmj*. British Medical Journal Publishing Group.

Vitale, Cristiana, Massimo Fini, Ilaria Spoletini, Mitja Lainscak, Petar Seferovic, and Giuseppe MC Rosano. 2017. "Under-Representation of Elderly and Women in Clinical Trials." *International Journal of Cardiology* 232: 216–21.

Zhong, Baoliang. 2009. "How to Calculate Sample Size in Randomized Controlled Trial?" *Journal of Thoracic Disease* 1 (1): 51.