Clinical Trials 4H - lecture notes

Rachel Oughton

2024-02-19

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.*

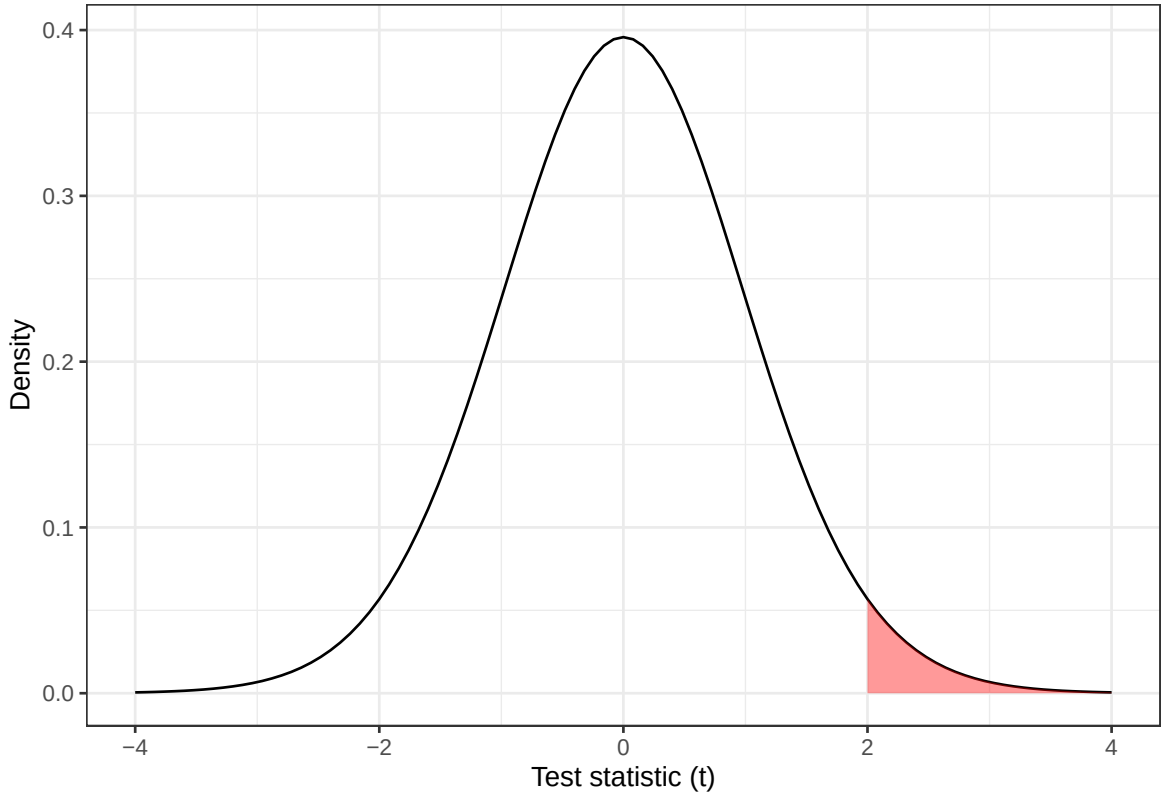


Figure 2.1: The distribution t_{31}, with the area corresponding to $t > 2$ shaded.

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.

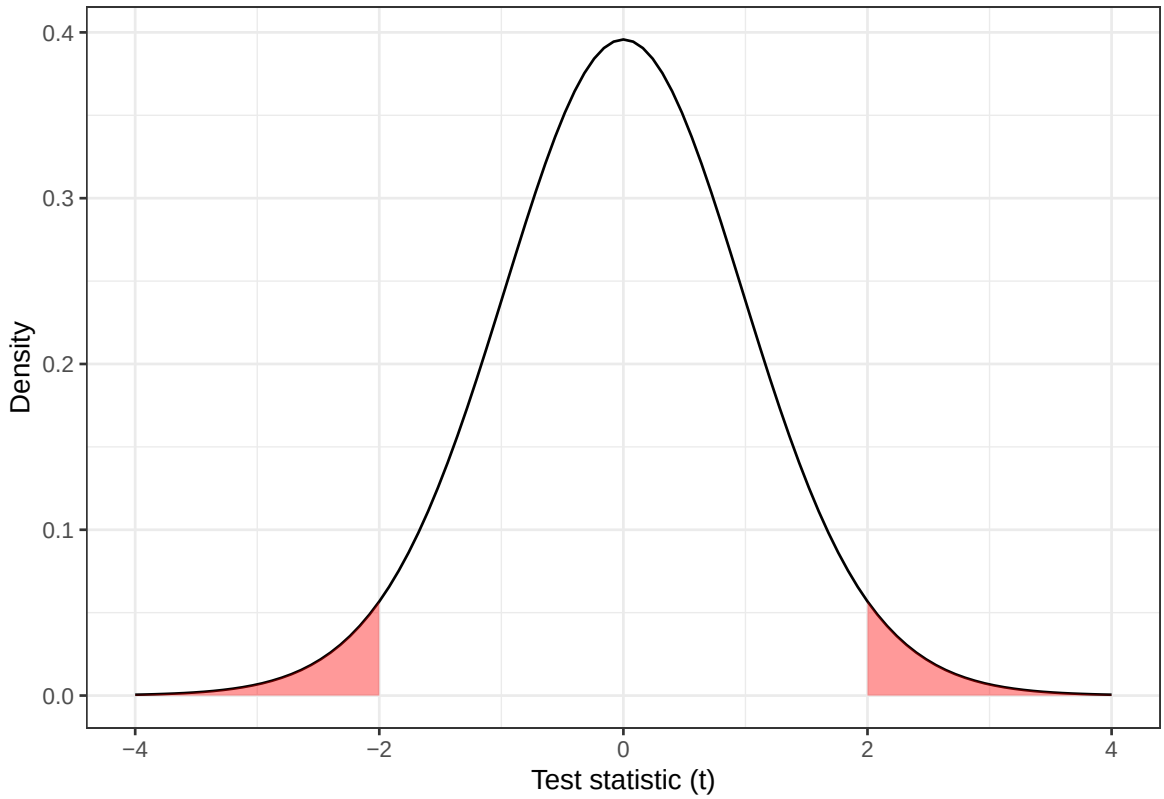


Figure 2.2: The distribution t_{31}, with the area corresponding to $|t| > 2$ shaded.

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

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

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).

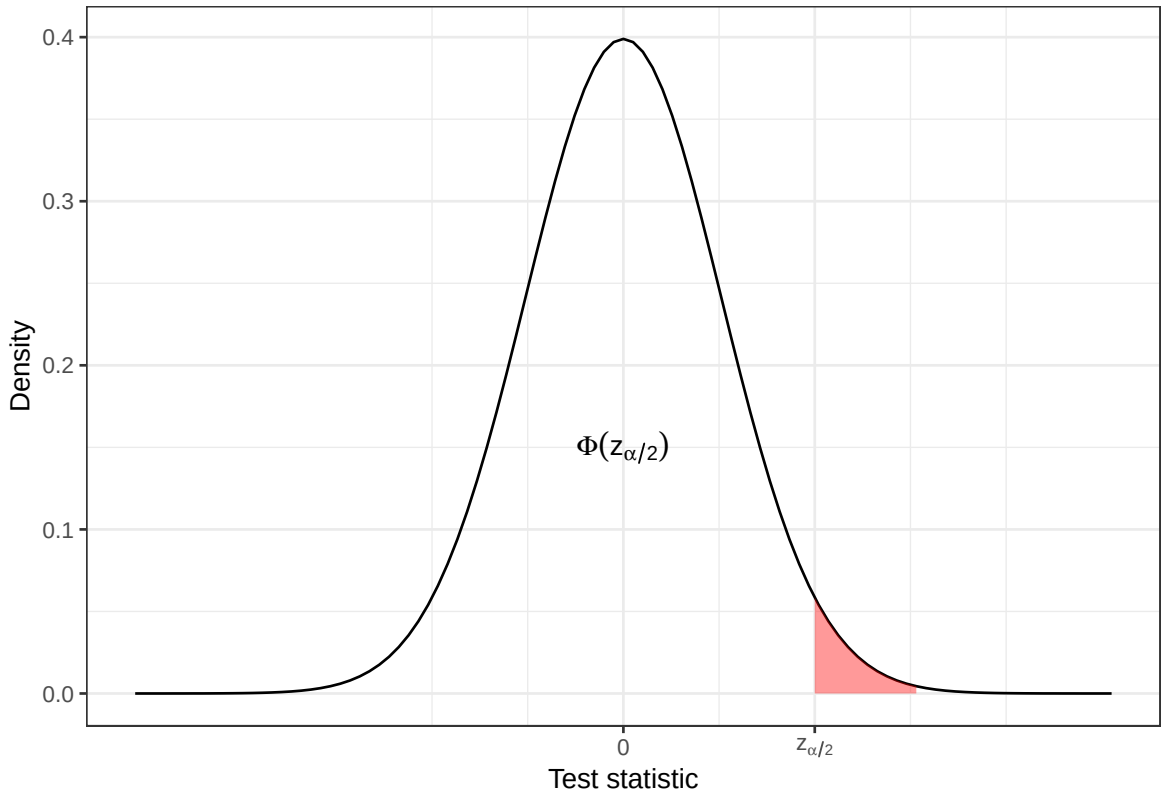
So, if we have run a trial, and have obtained n values of X from group C and m values of X from group T, we can compute D. If D lies outside the interval $\left[-z_{\alpha/2}, z_{\alpha/2}\right]$ then we reject H_0.

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.

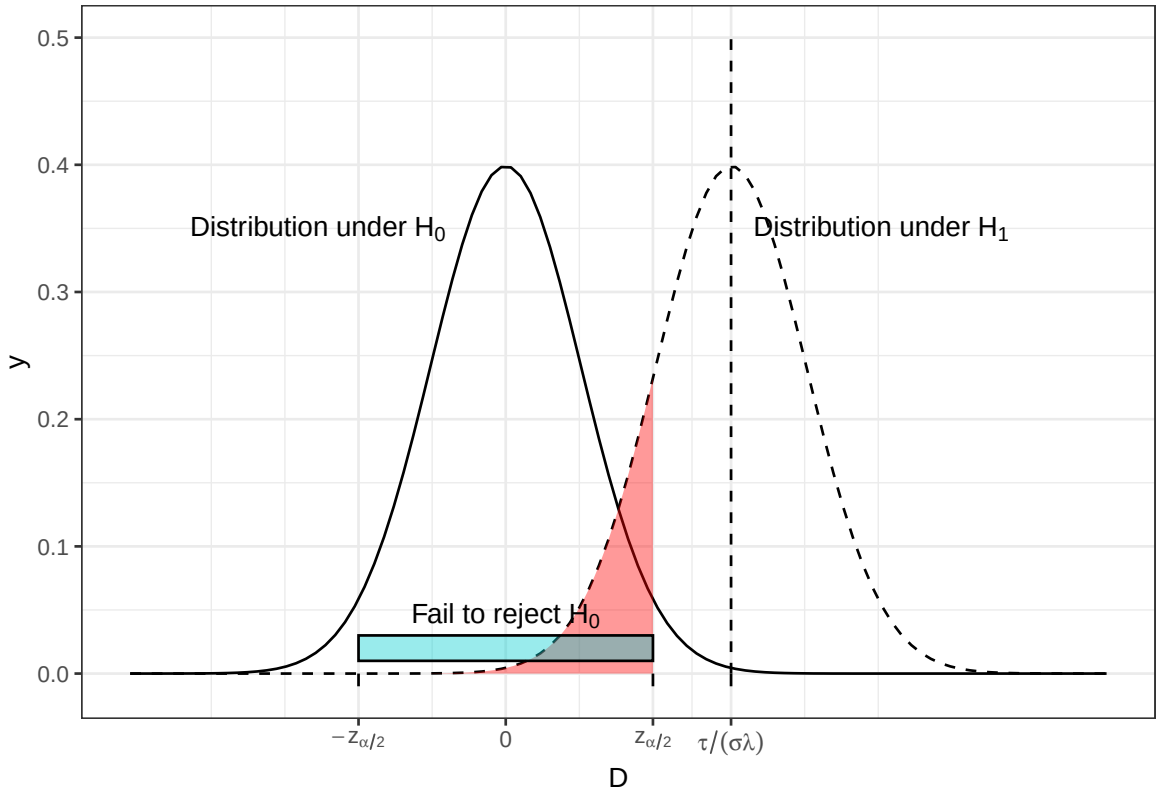


Figure 2.3: The distribution of D under both H_0 and H_1 for some arbitrary values of treatment effect, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

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

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

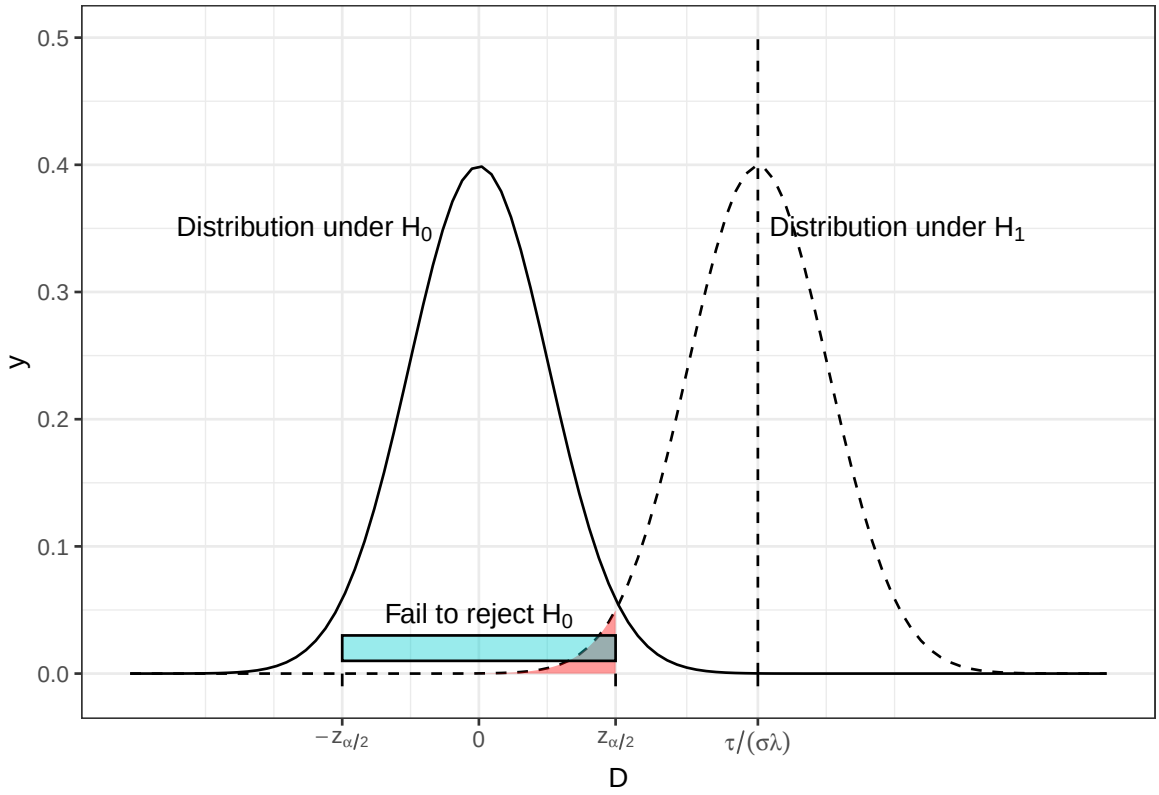


Figure 2.4: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

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.

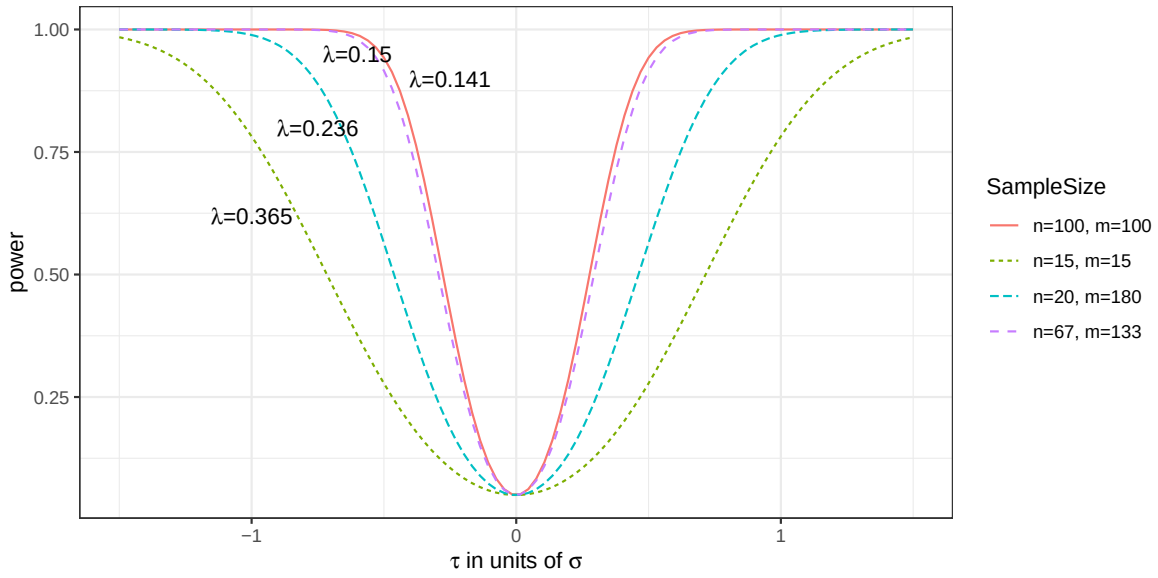


Figure 2.5: Power curves for various values of n and m, with effect size in units of standard deviation, given a type I error rate of 0.05.

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

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

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

Chapter 3

(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.

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

3.2 (Lecture 5) Allocation methods

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.

3.2.1 Simple random allocation

Simplest method: '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 ideal:

- generates the random sample we need
- All participants are allocated independently - not predictable
- No 'master' randomisation

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.

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

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

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

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

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

The probability that $N_C = N_T = n$ is

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

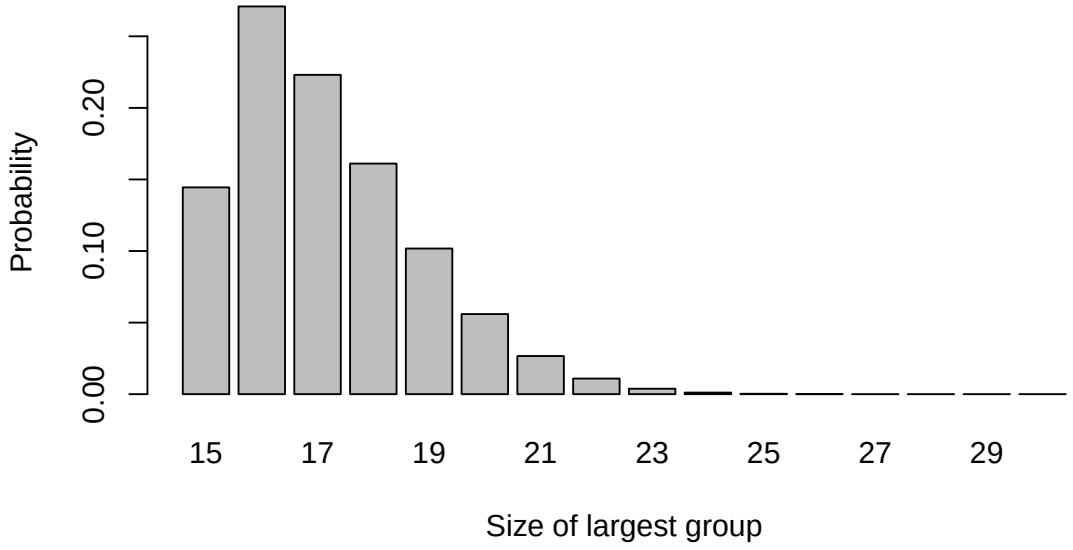


Figure 3.1: The probability distribution of largest group size for n=15.

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

For larger trials, this imbalance will be less pronounced, for example Figure 3.2 shows the same for $n = 300$.

In this case the $P(N_{max} \geq 400) \approx 10^{-16}$, so the chance of highly imbalanced groups (2:1 or worse) 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

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

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

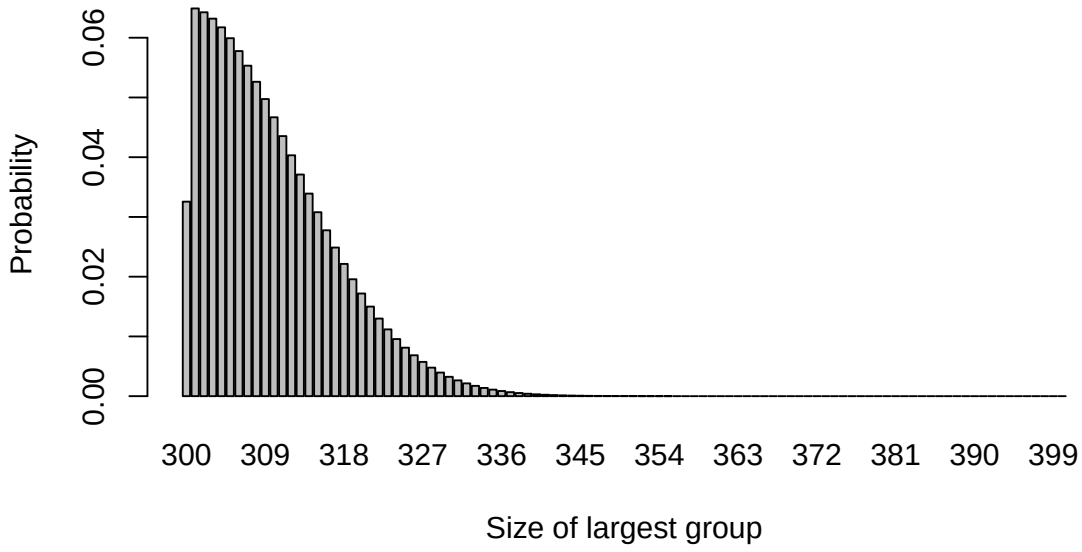


Figure 3.2: The probability distribution of largest group size for n=300.

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

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

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

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

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

If we have two groups, A and B, then there are six *blocks* of length 4 containing two As and two Bs

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

Randomly generate a sequence from $\{1, 2, 3, 4, 5, 6\}$, with equal probability. This sequence will correspond to a sequence in A and B with four times the length. For example, suppose the sequence begins $2, 1, 3, 6, \ldots$. Replacing each number by its block, we have $ABAB\ AABB\ ABBA\ BBAA\ \ldots$.

Advantages:

- Each patient is equally likely to receive A and B
- Difference between N_A and N_B never more than m (two in this case)
- Easy to implement

Disadvantages:

- Some allocations can be predicted (last in every block, sometimes more) *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

Remove predictability (almost) by randomly varying the length of the block. Eg. there are

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

possible 6-blocks. Instead of always using 4-blocks, we can do the following.

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 ($X = 4$) or 1 to 20 ($X = 6$).
3. The Y^{th} block is chosen and participants assigned accordingly.
4. If more participants are needed, go back to step 1.

This method

- Ensures patients equally likely to be in A or B
- group sizes never differ by more than 3
- reduces selection bias (only predictably if difference is 3)

3.2.3 Biased coin designs and urn schemes

What if we want to retain the pure stochasticity of SRS, but with more balance?

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 have groups T and C, and have so far allocated n participants. We write $N_T(n)$ for the number of participants allocated to treatment T, and $N_C(n)$ for the number of participants allocated to treatment C. We denote the *imbalance* by

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

We use $D(n)$ 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 $P \in \left(\frac{1}{2}, 1\right)$.

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

$$P\Big(|D(n+1)| = 1 \,\Big|\, |D(n)| = 0\Big) = 1$$
$$P\Big(|D(n+1)| = j+1 \,\Big|\, |D(n)| = j\Big) = 1 - P$$
$$P\Big(|D(n+1)| = j-1 \,\Big|\, |D(n)| = j\Big) = P$$

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.

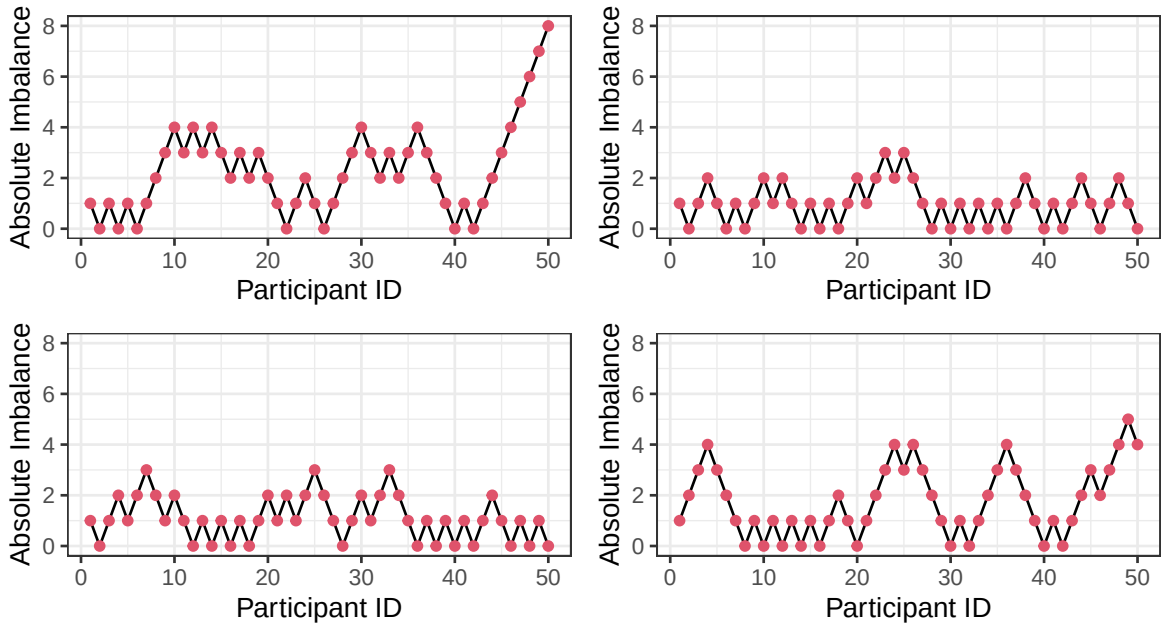


Figure 3.3: Absolute imbalance for a biased-coin scheme with $P = 0.667$.

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

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

In summary

- P closer to 0.5 $/rightarrow$ less predictable, more imbalance
- P closer to 1 $/rightarrow$ more predictable, less imbalance

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.

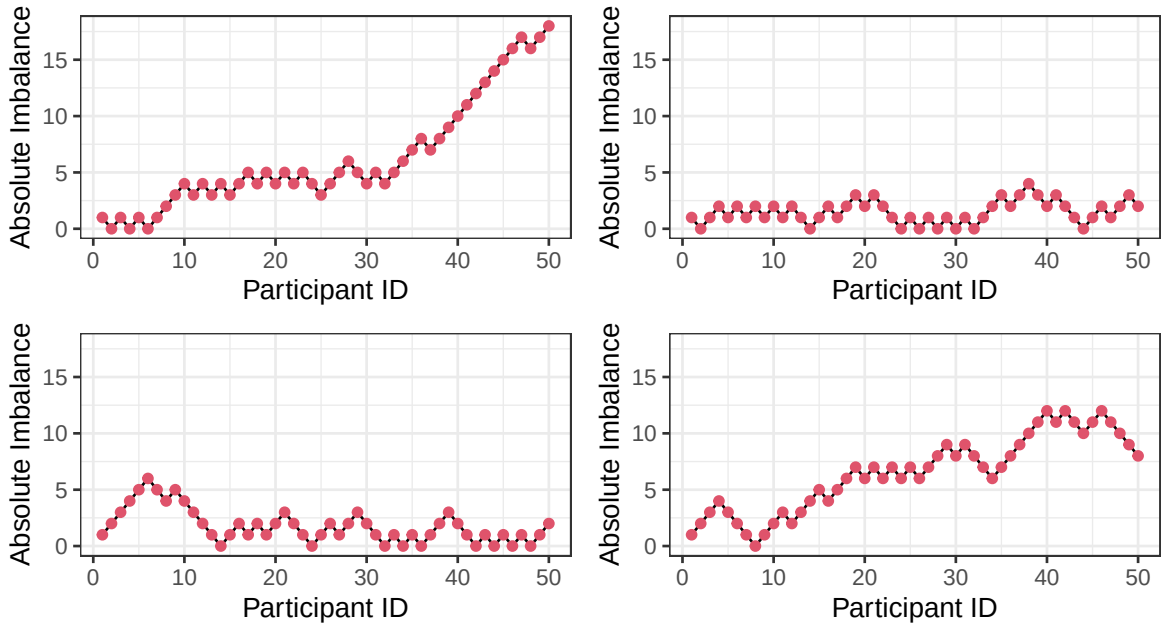


Figure 3.4: Absolute imbalance for a biased-coin scheme with $P = 0.55$.

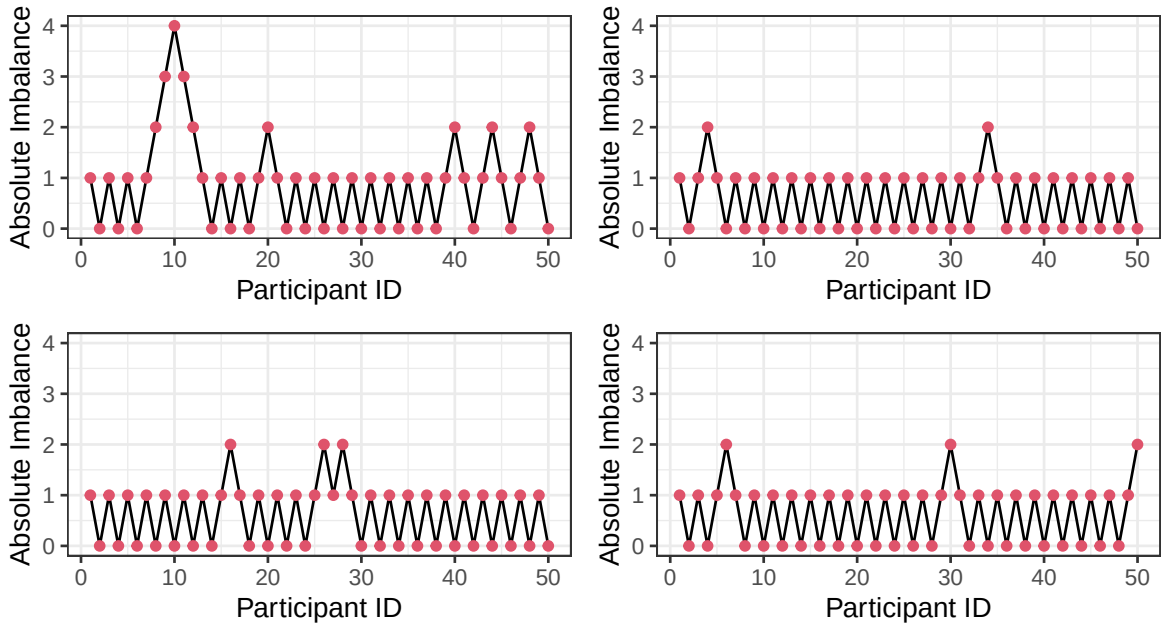


Figure 3.5: Absolute imbalance for a biased-coin scheme with $P = 0.9$.

3.2.3.2 Urn models

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 $UD(1,1)$; one ball for each treatment to start with, one ball added after each allocation.

More generally we assume a $UD(r,s)$ scheme:

- r balls for each treatment to begin with
- s balls 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 after allocation n there is imbalance 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

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

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

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

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

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

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

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

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

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

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

Particular advantages of the urn design are:

- Easy to implement for $K > 2$ groups
- Responds to higher imbalance

We've alluded to the idea of performing these algorithms within specific groups, to achieve balance among participants with particular characteristics. Next lecture we'll think about allocation methods that put more focus on balance in relation to factors.

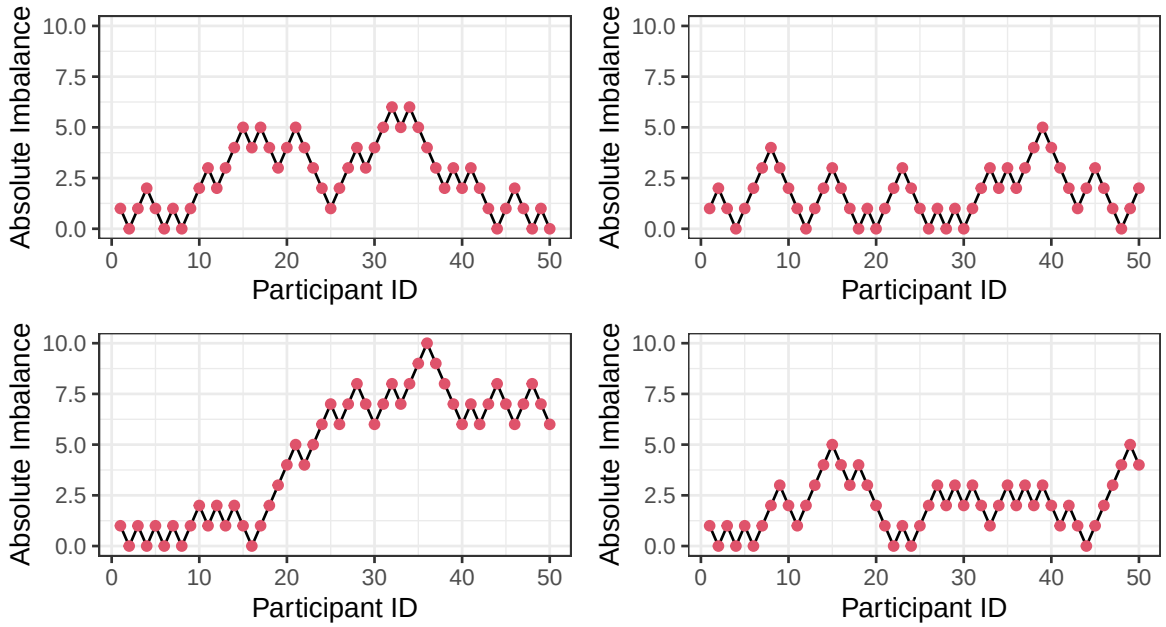


Figure 3.6: Four realisations of absolute imbalance for r=1, s=1, N=50.

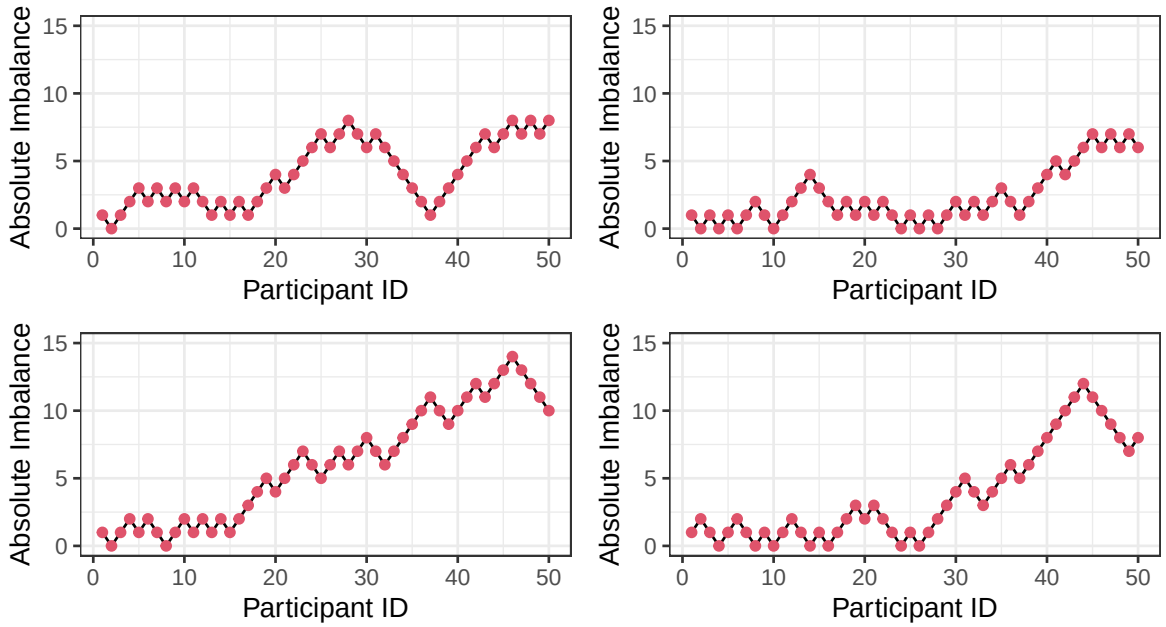


Figure 3.7: Four realisations of absolute imbalance for r=1, s=8, N=50.

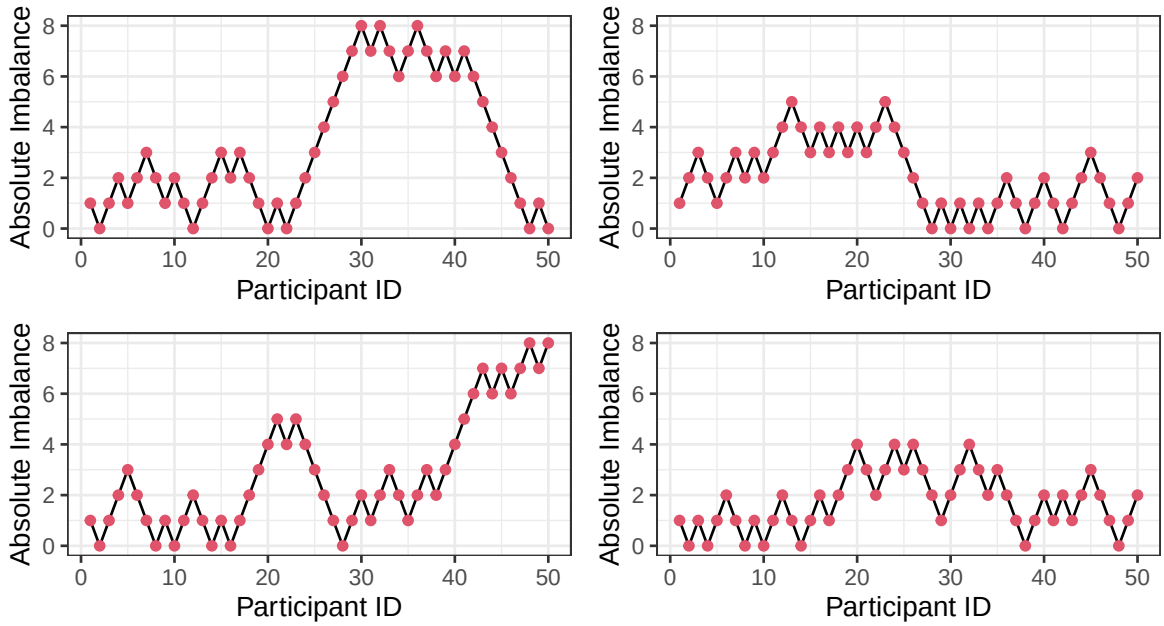


Figure 3.8: Four realisations of absolute imbalance for r=8, s=1, N=50.

3.3 (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.

Baseline measurements can be used in allocation, eg.

- baseline value of X, age, sex, condition specifics,...
- Must **not** be affected by the trial

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.

Divide each prognostic factor into ≥ 2 levels, consider allocation separately within each combination of factor levels. Each such group is a 'stratum'.

Within each stratum, apply RBP, biased coin, urn model or your allocation method of choice (one that aims for balance).

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).*

Difficulty:

- Number of strata quickly becomes too large: eg. four factors with three levels each $\rightarrow 3^4 = 81$ strata.
- Unwieldy practically
- Very likely to have v small numbers in some strata

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 (T *and* C*). It was developed for use with strata, as an alternative to randomly permuted blocks. Although the method was developed in the seventies, it has only gained popularity relatively recently, mainly as computers have become widely available.*

Aims to achieve balance between groups T and C, taking prognostic factors into account. Has become more popular recently, because of computing availability.

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 between the two groups. The minimization method has evolved since its conception, and exists in several forms. Two areas in which methods vary are

- Whether continuous variables have to be binned
- Whether there is any randomness

It is generally agreed that if the risk of selection bias cannot be avoided, there should be an element of randomness. It is also usually accepted that if a variable is included in the minimization, it should also be included in the statistical analysis.

3.5.1 Minimization algorithm

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

Suppose we require balance over several factors, 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. We will use this example to explain minimisation.

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 example

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

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

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
$$S = \left(n^A_{w+++} - n^B_{w+++}\right) + \left(n^A_{+x++} - n^B_{+x++}\right) + \left(n^A_{++y+} - n^B_{++y+}\right) + \left(n^A_{+++z} - n^B_{+++z}\right).$$
5. If $S < 0$ (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 $S > 0$, they are allocated to arm B with probability P. If the $S = 0$, 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

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

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

3.6.1 Communication!

3.6.2 Interactions

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	

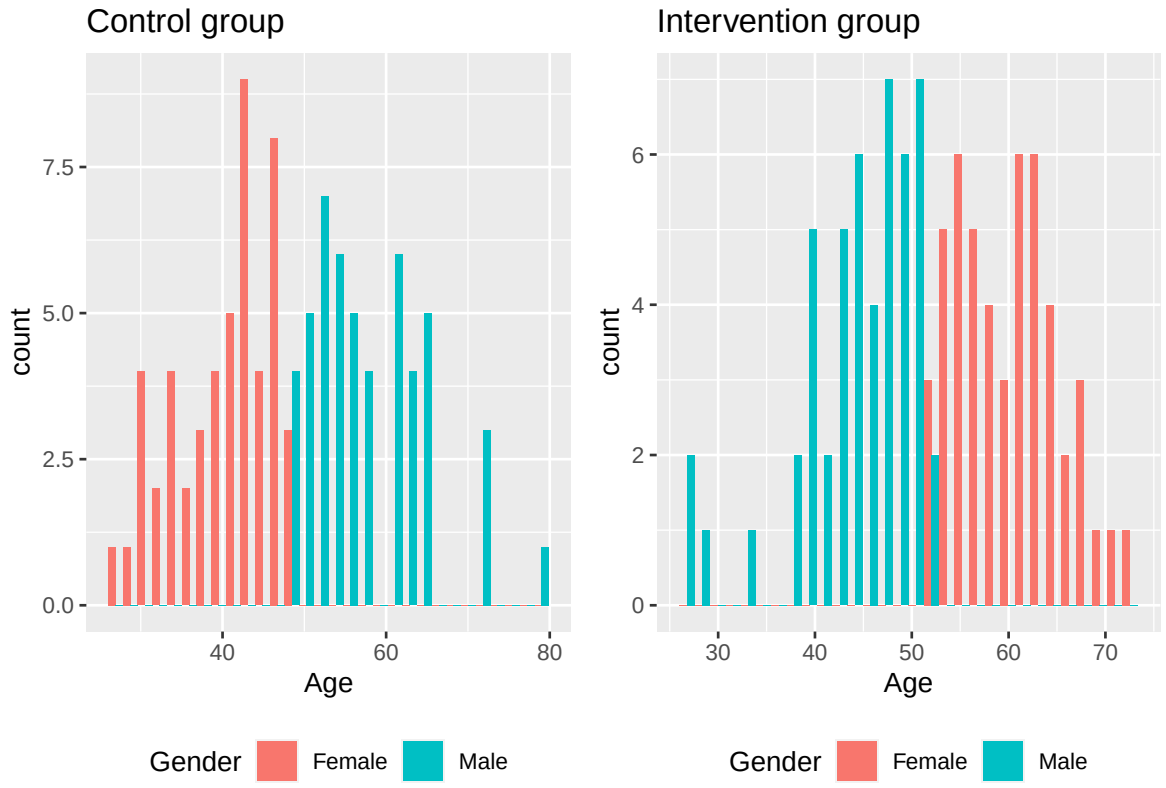
Summary statistics presented as percent of patients or mean ± SD.
[a]Standardized difference (licorice − sugar-water) defined as the difference in means or proportions divided by the pooled standard deviation; >0.2 in absolute value indicates imbalance.
[b]Surgery size: small (thoracoscopy); medium (thoracotomy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

Chapter 4

(Lecture 7) 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.

In this chapter we'll keep our focus on the scenario where the trial outcome is measured on a continuous scale, but in later weeks we'll go on to look at other types of data.

We now have trial data, and want to estimate the treatment effect τ.

Example 4.1.

- From Hommel et al. (1986).
- 16 diabetes patients
- Group T receive Captopril, a drug that may reduce blood pressure.
- Placebo given to group C
- Primary outcome X is systolic blood pressure (mmHg)

This is important, since for those with diabetes, high blood pressure can exacerbate kidney disease (specifically diabetic nephropathy, a complication of diabetes). To participate in the trial, people had to be insulin-dependent and already affected by diabetic nephropathy. In the trial, systolic blood pressure was measured before participants were allocated to each trial arm, and then measured again after one week on treatment. A placebo was given to the control group, so that all participants were blinded.

The baseline and outcome blood pressure measurements (in mmHg) are shown in Table 4.1. We see that nine participants were assigned to the treatment arm (Captopril) and the remaining seven to the placebo group. Hommel et al. (1986) say that the patients were 'randomly allocated' to their group.

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

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

4.1 Confidence intervals and P-values

Because the randomization process should produce groups that are comparable, we can in principle compare X between the groups.

Table 4.1: Data for the Captopril trial from @hommel1986effect.

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

Table 4.2: Summary statistics for each group.

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

Example 4.2. Summary statistics of the outcome for each group are shown below.

We have

$$\bar{x}_T = 135.33$$
$$\bar{x}_C = 141.86.$$

Difference in average of X between the two groups $141.86 - 135.33 = 6.53$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:

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

In our case

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

We can do an independent two-sample t-test,

$$\begin{aligned} t &= \frac{\bar{X_C} - \bar{X_T}}{s_p \sqrt{\frac{1}{n_C} + \frac{1}{n_T}}} \\ &= \frac{6.53}{7.82\sqrt{\frac{1}{7} + \frac{1}{9}}} \\ &= 1.65. \end{aligned}$$

Note that here the placebo group is group C, and the Captopril group is group T.

Under H_0, this value should be t_{14} ($n_i - 1$ for each group).

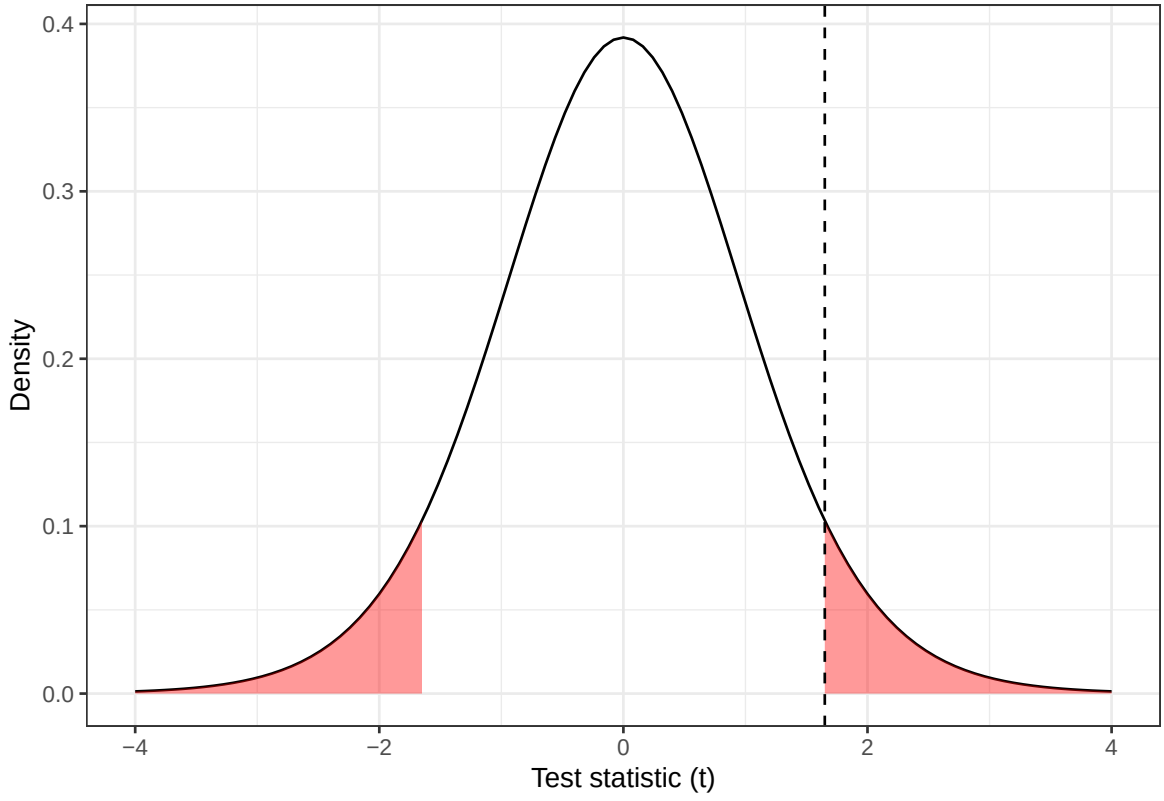


Figure 4.1: The distribution t_{14}, with $t = 1.65$ shown by the dashed line and the 'more extreme' areas shaded.

The dashed line 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
```

So $p = 0.121$ - not significant even at $\alpha = 0.1$.

4.1.1 What do we do with this outcome?

Worst case scenario:

- $\hat{\tau} = 6.53$ mmHg is large enough to be compelling
- Dataset is too small for it to be statistically significant
- Can't confidently conclude that Captopril has any effect on blood pressure (reject H_0).
- Can't say that there is no effect.

This is exactly the sort of scenario we hoped to avoid when planning our study.

Consider the range of treatment effects that are compatible with our trial data. That is, we find the set

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

Suppose the true treatment effect is τ^*, and we test $H_0 : \tau = \tau^*$.

- For all τ^* inside this range, our data are not sufficiently unlikely to reject the H_0 at the 0.05 level.
- For all values of τ^* outside this range, our data are sufficiently unlikely to reject that hypothesis.

Rearrange to give a 95% confidence interval for τ,

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

Example 4.3. Continuing our example, we have

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

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

Rearranging as above, this works out to be the interval

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

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

Some things to note

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

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

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.

You can't calculate the power of a trial after it's happened, but if we fail 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.

4.2 (Lecture 8) 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.

In Table 4.1 we also have *baseline* measurements: measurements of systolic blood pressure for each patient from before the intervention period. We'll denote these B_T and B_C.

Baseline measurements are useful primarily for two reasons:

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

We will demonstrate these by returning to our Captopril example.

Example 4.4. Balance:

- Placebo group: $\bar{b}_C = 146.6$ mmHg and $SD(b_C) = 12.3$ mmHg
- Captopril group: $\bar{b}_T = 148$ mmHg and $SD(b_T) = 11.4$ mmHg

Not identical, but sufficiently similar not to suspect systematic imbalance. In a study this small there is likely to be some difference.

Note, it's sensible to compare things informally like this, but sometimes people suggest using a t-test or similar to test whether the two groups are similar enough. This is a flawed exercise: a formal

Table 4.3: 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 4.4: 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

hypothesis test is testing whether what we see is likely to have arisen by random change. We know that this arose by random chance, because we randomly allocated the patients! A formal test only makes sense if you suspect that something has systematically disrupted the randomisation.

Analysis:

- Interested in whether Captopril has reduced BP for each individual - makes sense to compare change in BP ($X - B$), rather than final BP measurement.

We can see individual data in Table 4.3 and summary statistics in Table 4.4.

Now we can perform our test as before,

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

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

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

Taking into account the baseline values in this way has

- slightly reduced the P-value (though still $p > 0.05$)
- shifted the confidence interval slightly lower

4.2.1 Why did the CI and p-value change?

Notation:

- Baseline: B_C, B_T
- Outcome : X_C, X_T

All participants have been randomised from the same population, so

$$\mathrm{E}(B_C) = \mathrm{E}(B_T) = \mu_B.$$

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

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

Sometimes we'll have $\mu_B = \mu$, but this won't always be the case.

Usually we will assume that

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

and this is generally fairly reasonable in practice.

For our two analyses so far we have

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

that is, both are unbiased estimators of τ.

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

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

where ρ is the true correlation between X and B, and is assumed to be the same in either group.

Similarly,

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

Using this to work out the variance of the estimator $\hat{\tau}$ we find that for comparing means we have

$$\mathrm{var}(\tau) = \mathrm{var}(\bar{x}_T - \bar{x}_C) = \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right).$$

whereas for comparing differences from baseline

Table 4.5: Summary statistics for the dodgy analysis

	t_stat	df	p_value
Captopril	4.23	8	0.003
Placebo	1.58	6	0.170

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

Therefore,

- If $\frac{1}{2} < \rho \leq 1$ there will be a smaller variance when comparing differences
- If $0 \leq \rho < \frac{1}{2}$there will be a smaller variance 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.

4.2.2 A dodgy way to use baseline variables (skip this in lectures!)

Sometimes the analysis performed on a dataset is rather spurious, but it isn't always immediately obvious why. We'll look at one example now, because it is done sometimes.

Look at each group separately and determine whether there has been a significant change in the outcome variable (assumes $\mu_B = \mu$).

For Captopril data, we could perform a paired t-test on the difference between baseline B and outcome X for each patient, for each group.

If we do this, we find the summary statistics in Table 4.5.

We find

- Strong evidence for a change in blood pressure for the Captopril patients (group T)
- No such evidence for the placebo patients.

Can we therefore conclude that Captopril is significantly better than the placebo? No! The analysis is flawed:

- The p-value of 0.17 in the control group doesn't show that H_0 is true *(no treatment effect for the control group) is true, just that we can't reject the null hypothesis. It is quite possible that there is a difference in the control group, and that numerically it could even be comparable to that in the treatment group, so although we can say that there is a significant reduction in blood pressure for the captopril group, we can't conclude that Captopril is better than the placebo.*
- Having set up the experiment as a randomised controlled trial, with a view to comparing the two groups, it seems strange to then deal with them separately.

4.3 Analysis of covariance (ANCOVA)

Previously: based our analysis on the baseline values being statistically identical draws from the underlying distribution, and therefore having the same expectation and variance.

However, in the event there will be some imbalance.

We can see this in our Captopril example. Eg. we saw that difference in baseline means is $\bar{b}_C - \bar{b}_T = 1.4$ mmHg.

- Not clinicaly significant
- Not large enough to make us doubt our randomisation
- A difference nonetheless

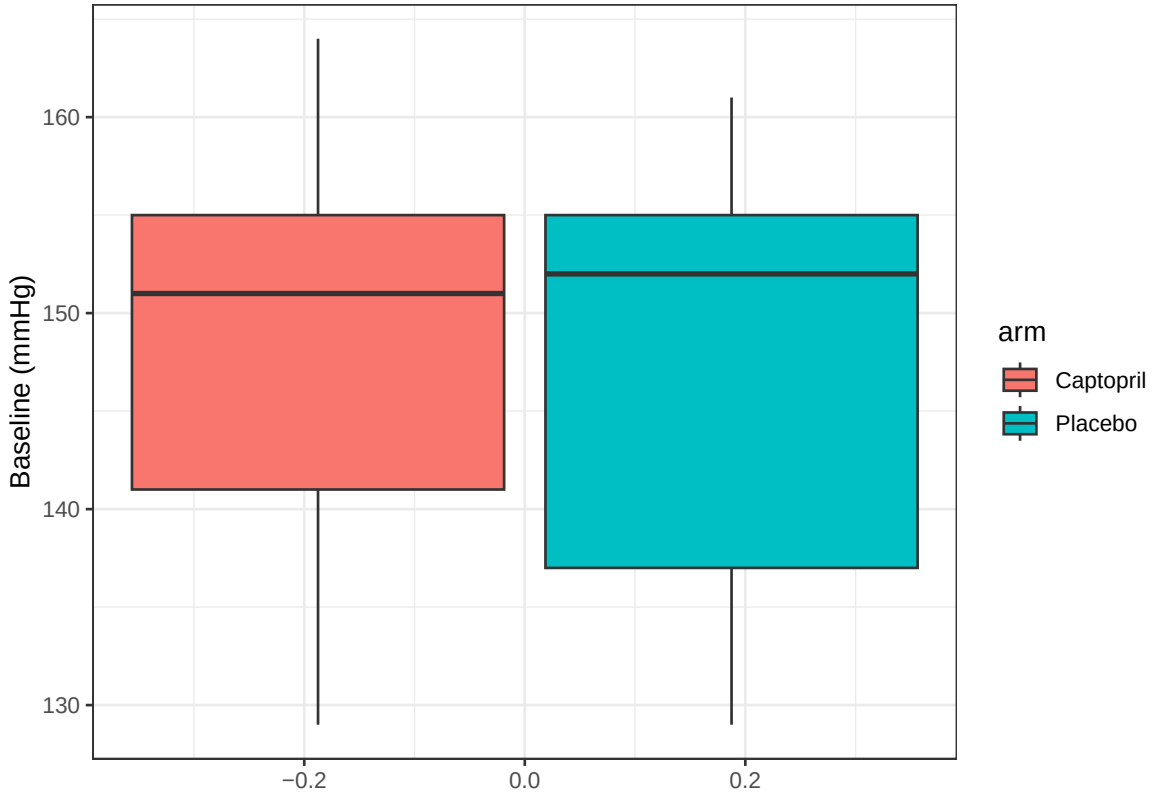


Figure 4.2: Baseline measurements from the Captopril trial.

Basic principle of ANCOVA: if there is some correlation between baseline and outcome, then baselines differing leads to outcomes differing, even if there is no treatment effect (ie. if $\tau = 0$).

Indeed, how do we decide how much of the difference in outcome is down to the treatment itself, and how much is simply the difference arising from different samples?

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

4.3.1 The theory

Usual notation and assumptions:

- Outcome X, baseline B
- $E(X_T) = \mu + \tau$
- $E(X_C) = \mu$
- $var(X_T) = var(X_C) = \sigma^2$
- Correlation between B and X is ρ **within each group**
- We want to learn about τ

Baseline B (same measurement as X) is measured before start of trial. Assume $E(B_T) = E(B_C) = \mu_B$ and $var(B_C) = var(B_T) = \sigma^2$. Also assume both groups have size N, therefore $2N$ patients in all.

We observe baseline measurements $b_1, b_2, \ldots, b_{2N}$.

Given these values, we have

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

From this, we find that

$$\mathrm{E}\left(\bar{X}_T - \bar{X}_C \mid \bar{b}_T, \bar{b}_C\right) = \tau + \rho\left(\bar{b}_T - \bar{b}_C\right). \tag{4.1}$$

If there is a difference in the baseline mean between 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
- 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 4.2, 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

$$\hat{\tau} = \left(\bar{X}_T - \bar{X}_C\right) - \rho\left(\bar{b}_T - \bar{b}_C\right) \tag{4.2}$$

then, following from Equation (4.1) we have

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

4.3.1.1 What's the variance of this estimator?

To work out the variance of $\hat{\tau}$ in Equation (4.2) we need to think about bivariate normal variables.

Let's suppose that random variables X and Y are jointly normally distributed with correlation ρ

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right). \tag{4.3}$$

From Equation (4.3), we know that $\mathrm{E}(Y) = \mu_Y$.

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

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1.5 \\ 1.5 & 3 \end{pmatrix} \right). \tag{4.4}$$

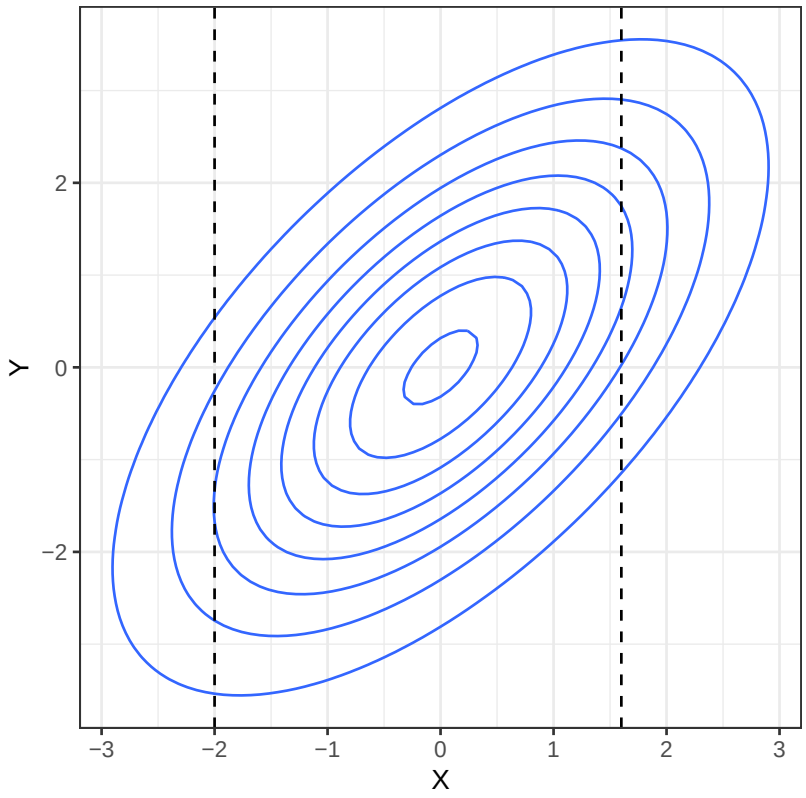


Figure 4.3: A bivariate normal density.

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

$$\mathrm{E}(Y \mid X = x) \neq \mathrm{E}(Y).$$

If we have another RV, W, $W \perp Y$ (and note that if two normally distributed variables are uncorrelated, they are also independent), *then observing $W = w$ doens't give us any information about the distribution of Y, so we have*

$$\mathrm{E}(Y \mid W = w) = \mathrm{E}(Y).$$

We can combine this information to work out $\mathrm{E}(Y \mid X = x)$.

First, $cov(X,\ Y - kX)$, for some constant k. We can find this by

$$\begin{aligned}\operatorname{cov}(X,\, Y - kX) &= \mathrm{E}\left[(X - \mu_X)(Y - kX - \mu_Y + k\mu_X)\right] \\ &\quad \text{(using that } \operatorname{cov}(\mathrm{X}, \mathrm{Y}) = \mathrm{E}\left[(\mathrm{X}\text{- }\mathrm{E}(\mathrm{X}))(\mathrm{Y}\text{- }\mathrm{E}(\mathrm{Y}))\right] \\ &= \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y) - k(X - \mu_X)^2\right] \\ &= \rho\sigma_X\sigma_Y - k\sigma_X^2.\end{aligned}$$

If we set

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

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

$$\mathrm{E}(Y - \beta X \mid X = x) = \mathrm{E}(Y - \beta X) = \mu_Y - \beta\mu_X.$$

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

$$\begin{aligned}\mathrm{E}(Y \mid X = x) &= \mathrm{E}(\beta X \mid X = x) + \mu_Y - \beta\mu_X \\ &= \mu_Y + \beta(x - \mu_X).\end{aligned}$$

We can use the same idea to find $\operatorname{var}(Y \mid X = x)$.

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

$$\operatorname{var}(Y \mid X = x) = \mathrm{E}\left(Y^2 \mid X = x\right) - [\mathrm{E}(Y \mid X = x)]^2. \tag{4.5}$$

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

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

$$\begin{aligned}\operatorname{var}(X) &= \mathrm{E}\left(X^2\right) - [\mathrm{E}(X)]^2 \\ \text{and therefore} \\ \mathrm{E}\left(X^2\right) &= \sigma_X^2 + \mu_X^2,\end{aligned}$$

we find that

$$\mathrm{E}\left[(Y - \beta X)^2 \mid X = x\right] = \mathrm{E}\left[(Y - \beta X)^2\right] = S^2 + (\mu_Y - \beta\mu_X)^2, \tag{4.6}$$

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

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

$$\mathrm{E}\left[Y^2 \mid X = x\right] - 2\beta\,\mathrm{E}\left(Y \mid X = x\right) + \beta^2 x^2 = \mathrm{E}\left[Y^2 \mid X = x\right] - 2\beta x\left(\mu_Y - \beta\mu_X\right) - \beta^2 x^2.$$

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

$$\mathrm{E}\left[Y^2 \mid X = x\right] = S^2 + \left(\mu_Y - \beta\mu_X\right)^2 + 2\beta x\left(\mu_Y - \beta\mu_X\right) + \beta^2 x^2.$$

Now we can expand out

$$\mathrm{E}\left(Y \mid X = x\right) = \mu_Y + \beta\left(x - \mu_X\right) = \left(\mu_Y - \beta\mu_X\right) + \beta x$$

to find

$$\left[\mathrm{E}\left(Y \mid X = x\right)\right]^2 = \left(\mu_Y - \beta\mu_X\right)^2 + 2\beta x\left(\mu_Y - \beta\mu_X\right) + \beta^2 x^2.$$

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

$$\begin{aligned}\mathrm{var}\left(Y \mid X = x\right) &= \mathrm{E}\left[Y^2 \mid X = x\right] - \left[\mathrm{E}\left(Y \mid X = x\right)\right]^2 \\ &= S^2 \\ &= \sigma_Y^2\left(1 - \rho^2\right).\end{aligned}$$

One thing to notice is that this conditional variance of doesn't depend on the observed value of $X = x$. It can also never exceed σ_Y^2, and is only equal to σ_Y^2 if X and Y are uncorrelated (ie. if $\rho = 0$).

Back to our estimator!

Recall that in ANCOVA our estimator of the treatment effect τ is

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

and that we have

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

Therefore, using the result we just found,

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

Notice that unlike our first estimator that used baseline values, in Section 4.2, the variance of the ANCOVA estimate can never exceed $\frac{2\sigma^2}{N}$; if the baseline and outcome are uncorrelated, ANCOVA will perform as well as a t-test.

4.3.2 (Lecture 9) The practice

In the previous section we established an unbiased estimate of the treatment effect that takes into account the baseline measurements.

We can't use this estimator as a model, because:

- $\hat{\tau}$ relies on the ρ, 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:

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

Where

- ϵ_i are independent errors with distribution $N\left(0,\, \sigma^2\right)$
- 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

$$G_i = \begin{cases} 1 & \text{if participant } i \text{ is in Group } T \\ 0 & \text{if participant } i \text{ is in Group } C \end{cases}$$

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 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

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

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.

The estimates $\hat{\mu}$, $\hat{\tau}$ and $\hat{\gamma}$ are then used to estimate σ^2, using

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

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
- 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 is known as **ANCOVA** (short for the **An**alysis of **Cova**riance)

Can be implemented in R and many other statistical software packages. Notice that it is really just a linear model (the like of which you have seen many times) with at least one factor variable, and with a particular focus (application-wise) on the coefficient of the treatment group variable.

Example 4.5. 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.

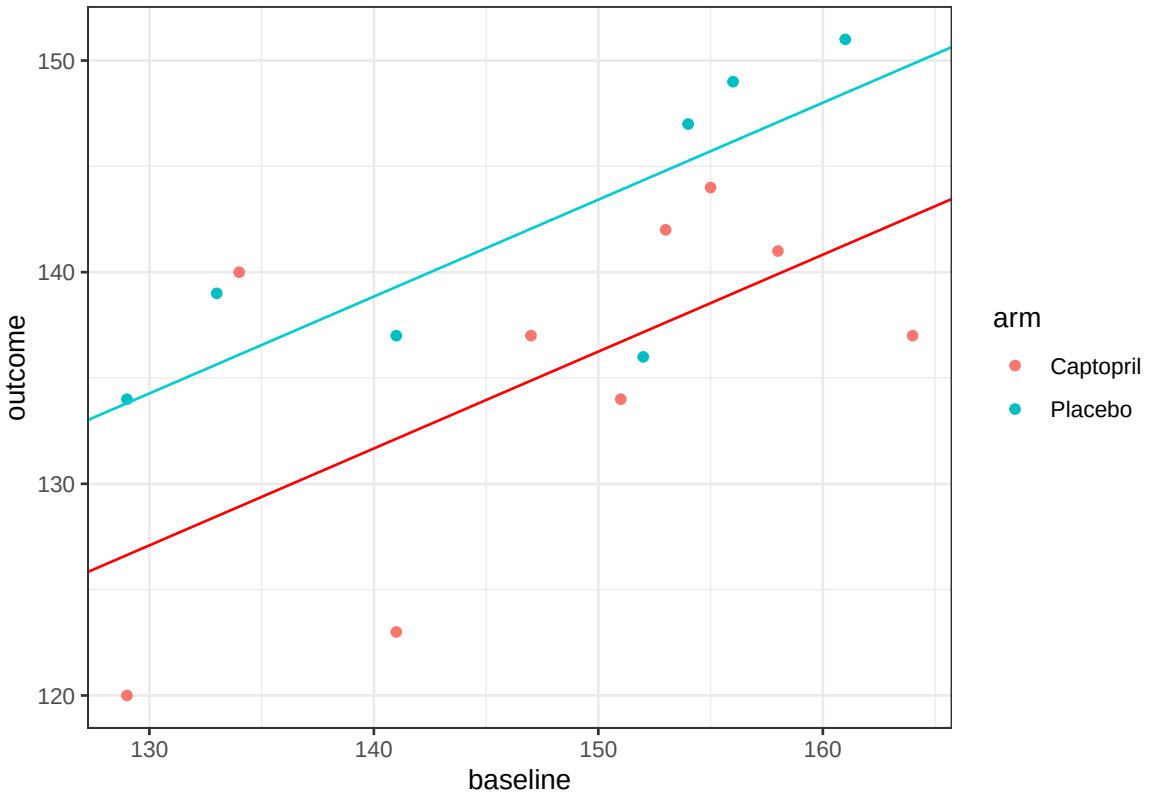
```
lm_capt = lm(outcome ~ baseline + arm, data = df_hommel)
summary(lm_capt)
```

```
##
## 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 we are estimating $p = 3$ parameters.

Notice that unlike our previous confidence intervals, this doesn't contain zero, and so our analysis has enabled us to conclude that there is a significant reduction in blood pressure with Captopril. Also $p < 0.05$. However, you can tell from the width of the interval (and the fact that p is still quite close to 0.05) that there is still a lot of uncertainty about τ.

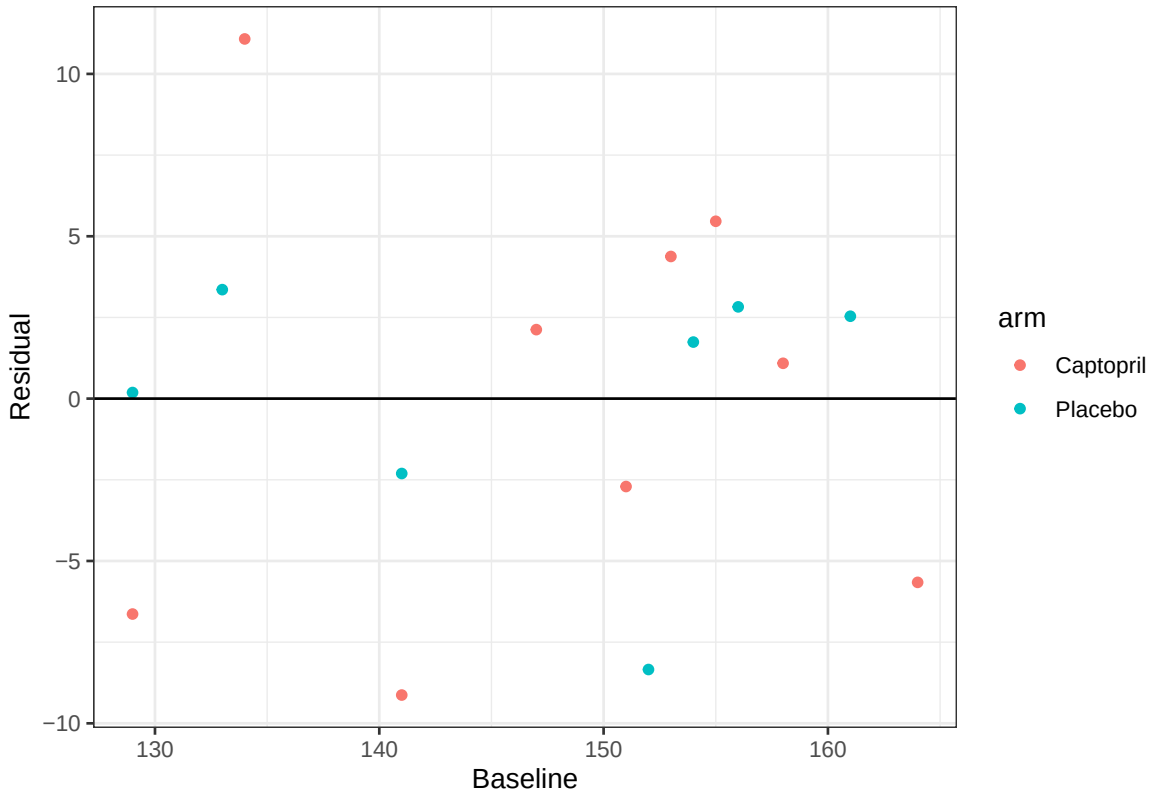
The 'Residual standard error' term near the bottom of the linear model summary is the estimate of $\hat{\sigma}$, so here we have $\hat{\sigma}^2 = 5.869^2 = 34.44$.

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()
```



Residuals look OK:

- No clear patterns
- Distribution appears to be similar for each treatment group

Though, with such a small sample it's difficult really to assess the fit of the model.

4.4 Some follow-up questions. . . .

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

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

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

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

that is, neither of these quantities are unbiased estimators of τ (except in very specific circumstances).

Is this a contradiction? No!

The first two estimators were blind to the values of B_T and B_C, and used their a priori statistical properties. ANCOVA uses the observed values, which will have exactly those properties.

Because of the randomisation procedure, a priori they can be treated the same. However, once we have observed values for the baseline, b_T and b_C, they are very unlikely to be exactly the same.

They are also (along with all other baseline measurements, often things like age, sex, height etc.) definitely not affected by the trial, since they are taken before any placebo or treatment has been administered, and often even before allocation. However, conditioning on their observed values can reduce the variance of our estimate of τ, as we have seen.

In this sense, the observed baseline means $\bar{b}_T$ and $\bar{b}_C$ are known as **ancillary statistics**:

- they contain no direct information about the parameter we are interested in (in this case τ) *our inferences can be improved by conditioning on their observed values.

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

In the analysis above, we 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 4.6. 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
```

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.

4.4.3 Can we include any other baseline covariates?

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

- Must be baseline measurements
- Any covariate used in allocation should be included in analysis

Not to be confused with 'the baseline', which would generally mean the same measurement as the primary outcome, but before treatment). This is because they cannot, at that point, have been affected by the treatment, or have had an influence on the post-trial outcome measurement.

Example 4.7. 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.

- Two treatment levels: `yes` or `no`
- Three exercise levels: `low`, `moderate` and `high`
- 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.

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

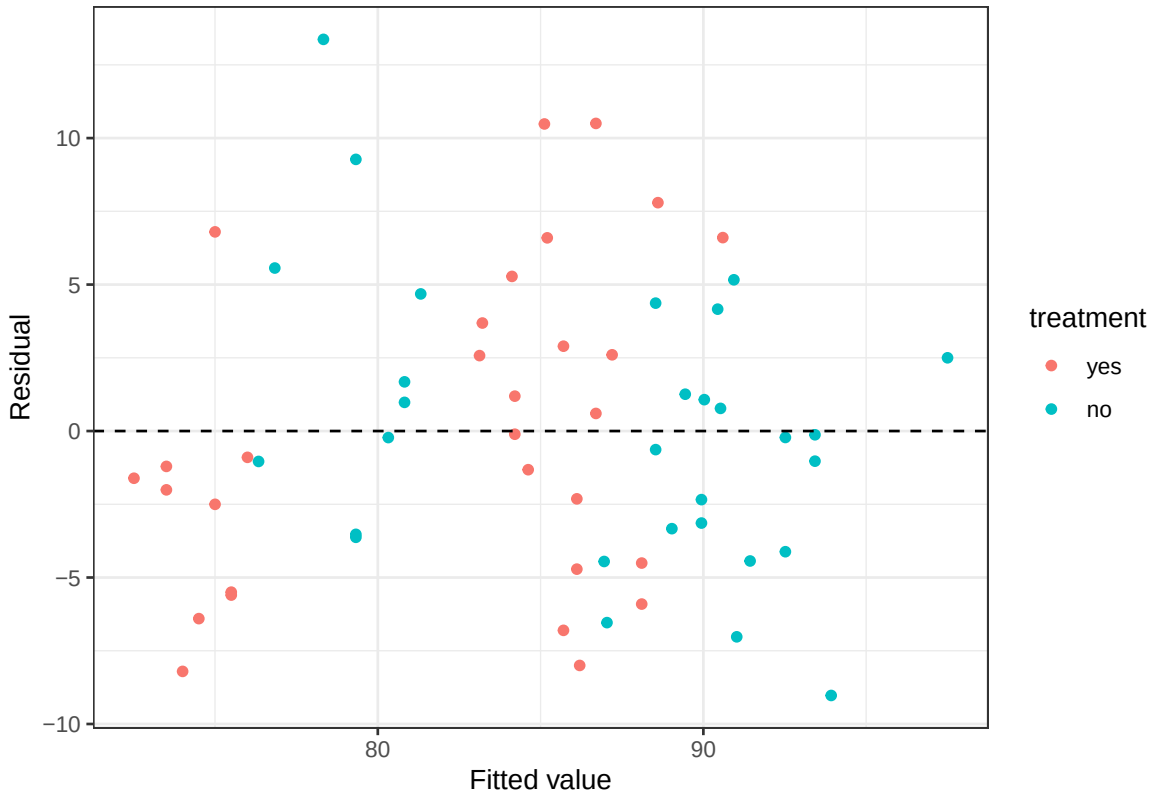
Fitting a linear model, we see that treatment, high levels of exercise and age have an effect on stress.

```
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



And these look reasonably OK. We could also test for interactions, firstly across all factors:

```
##
## Call:
## lm(formula = score ~ (treatment + exercise + age):(treatment +
##     exercise + age), data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5637 -3.3982  0.4173  2.3827 10.3907
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    61.25416   19.86949   3.083  0.00333 **
## treatmentno                     3.89781   24.16324   0.161  0.87250
## exercisemoderate              -14.60897   24.31690  -0.601  0.55070
## exercisehigh                  -12.03441   29.53812  -0.407  0.68544
## age                             0.43121    0.32097   1.343  0.18518
## treatmentno:exercisemoderate   -0.20723    3.35949  -0.062  0.95106
## treatmentno:exercisehigh        8.12783    3.72077   2.184  0.03365 *
## treatmentno:age                -0.03769    0.38851  -0.097  0.92311
## exercisemoderate:age            0.24286    0.40215   0.604  0.54864
## exercisehigh:age               -0.03524    0.50722  -0.069  0.94488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.106 on 50 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6043
## F-statistic: 11.01 on 9 and 50 DF,  p-value: 3.181e-09
```

and then restricted to the interactions that seem important:

```
##
## Call:
## lm(formula = score ~ treatment + exercise + age + treatment:exercise,
##     data = stress)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3250 -3.0192  0.2745  2.4650 10.6667
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   56.79090   10.41383   5.453 1.32e-06 ***
## treatmentno                    1.52858    2.23026   0.685   0.4961
## exercisemoderate               0.01746    2.25662   0.008   0.9939
## exercisehigh                 -13.70331    2.36314  -5.799 3.78e-07 ***
## age                            0.50355    0.16684   3.018   0.0039 **
## treatmentno:exercisemoderate   0.15503    3.16129   0.049   0.9611
## treatmentno:exercisehigh       8.21822    3.15375   2.606   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.985 on 53 degrees of freedom
## Multiple R-squared:  0.6613, Adjusted R-squared:  0.623
## F-statistic: 17.25 on 6 and 53 DF,  p-value: 6.167e-11
```

Notice that now, the effect of the treatment on its own is not significant. Also notice that for both the linear exercise terms and the interactions between the exercise and treatment, the effects of moderate and low exercise are very similar. Combining the coefficients, someone who does a high level of exercise:

- is likely to reduce their stress score by 13.7 if they receive the treatment
- is likely to reduce their stress score by $13.7 - 8.2 = 5.5$ if they don't receive the treatment

Returning to our initial look at the dataset, the fact that age is a factor, and high levels of exercise are clearly very important should worry us slightly, since there are very few older people doing high levels of exercise. This may mean our model is inaccurate.

An important caution!

As you'll have seen if you read Kendall (2003) (for formative assignment 1), we should have everything in place, including a statistical analysis plan, **before** the trial. We should already know which covariates we plan to include in our model, and how. 'Trawling' for the best possible model by trying lots of different things (and inevitably settling on the one that leads to the most significant conclusion) is poor practice, and can increase the type I error rate (α).

I realise that is sort of what we've done in this Section on Analysis, but that was to demonstrate and compare the different methods. Proceeding in the way we have, trying lots of different models, when analysing and writing up a trial would be very poor practice!

There's another excellent episode of the JAMA Evidence podcast, with a focus on adjusting for covariates, that talks about this issue (you can find it here and linked from Ultra).

That draws to a close our work with continuous outcome variables. In the next lecture, we'll start thinking about binary outcome variables.

Part I

Part II: Binary outcome variable

Chapter 5

(Lecture 10) Sample size for a binary variable

So far almost everything we've covered has related to continuous outcome variables, which we assumed to be normally distributed. This allowed us to use familiar techniques such as the t-test, and to take baseline information into account in an accessible way (the linear model / ANCOVA). However, very often clinical trials do not have a continuous, normally distributed output, and in the next two sections we will look at two other common possibilities: binary data (this section) and survival data (next section).

Binary outcome:

- 'the patient was alive 2 years after the procedure' or not
- 'the patient was clear of eczema within a month' or not.

Often coded as 'success' or 'failure', or 1 or 0.

For a trial whose primary outcome variables are binary, the sample size calculations we derived in Chapter 2 will not work, so in this section we'll work through a similar method developed for binary variables.

Suppose we conduct a trial with a binary primary outcome variable and two groups, T and C, containing n_T and n_C participants respectively. The number of successes in each group, R_T and R_C, will be Binomially distributed,

$$R_T \sim Bi\,(n_T,\, \pi_t)$$
$$R_C \sim Bi\,(n_C,\, \pi_C).$$

Now we have

$$H_0 : \pi_T = \pi_C$$
$$H_1 : \pi_T \neq \pi_C.$$

We will need enough participants to test this hypothesis with sufficient power.

With the trial data we find estimates

$$p_T = \frac{R_T}{n_T}$$
$$p_C = \frac{R_C}{n_C}.$$

Recall that

$$\text{var}\left(p_X\right) = \pi_X\left(1 - \pi_X\right),$$

where X is T or C. Notice, **the variance depends on the mean → no free parameter**.

This means there is no free parameter equivalent to σ in the binary situation, and the number of participants required will depend on the approximate value of π_T and π_C. This makes the derivation of a sample size formula somewhat more complicated, and so we first of all make a transformation to remove the dependence of mean and variance. To do this we use an approximation technique called the delta method.

5.1 The Delta Method

Suppose a random variable X has mean μ and variance $\sigma^2 = \sigma^2(\mu)$, ie. its variance depends on its mean.

If we have a 'well-behaved' (infinitely differentiable etc.) function $f(X)$, what are its mean and variance?

To find this exactly requires us to evaluate a sum or integral, and this may be analytically intractable, so we use instead a crude approximation.

First, we expand $f(X)$ in a first-order Taylor series about μ, which gives us

$$f(X) \approx f(\mu) + (X - \mu) f'(\mu) \tag{5.1}$$

and therefore

$$\left(f(X) - f(\mu)\right)^2 \approx (X - \mu)^2 \left[f'(\mu)\right]^2. \tag{5.2}$$

Taking expectations of Equation (5.1) we find $E\left(f(X)\right) \approx f(\mu)$.

Use this in the LHS of Equation (5.2) so that when we take expectations of Equation (5.2) we find

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

where both sides come from

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

This series of approximations, which generally works well, is the Delta method.

We use the Delta method to find a transformation $f(X)$ for which (at least approximately) the variance is unrelated to the mean. To do this, we solve the differential equation

$$\text{var}\left[f\left(X\right)\right] = \sigma^2\left(\mu\right)\left[f'\left(\mu\right)\right]^2 = \text{constant}.$$

In the case of proportions for a binary variable, this becomes

$$\frac{\pi\left(1-\pi\right)}{n}\left[f'\left(\pi\right)\right]^2 = K$$

for some constant K. We can rearrange this to

$$f'\left(\pi\right) = \sqrt{\frac{Kn}{\pi\left(1-\pi\right)}} \propto \sqrt{\frac{1}{\pi\left(1-\pi\right)}}.$$

So we need

$$\int^{\pi} \sqrt{\frac{1}{u\left(1-u\right)}} du,$$

where the notation indicates that we want the anti-derivative, evaluated at π. By substituting $u = w^2$ we find

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

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

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

`Finally, we can substitute this into Equation \\eqref{eq:delta3}, with $f\left(X\right) = \arcsin\lef`

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

and we have achieved our aim of finding a transformation of X whose variance is not related to the mean. This is sometimes called the *angular transformation*.

5.2 A sample size formula

Our estimate p_X (the proportion of successes in group X) is approximately normally distributed (by CLT).

Not true for $n \leq 30$ ish (very small for a clinical trial) or for π close to 0 or 1, say $\pi < 0.15$ or $\pi > 0.85$ (this is more likely to be an issue for some trials).

The linear approximation in Equation (5.1) shows us that if p_X is normally distributed then $f(p_X) = \arcsin\left(\sqrt{p_X}\right)$ will be [approximately] normally distributed too.

In fact, $\arcsin\left(\sqrt{p_X}\right)$ is approximately normally distributed with mean $\arcsin\left(\sqrt{\pi_X}\right)$ and variance $1/(4\pi_X)$.

Using this information, we can test $H_0 : \pi_T = \pi_C$ at the $100\alpha\%$ confidence level by using the variable

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

which is analogous to the variable D constructed in Section 2.3; the difference in $f(p_T)$ and $f(p_C)$ divided by the standard error of the difference.

Using the same logic as in Sections 2.4 and 2.5, the starting place for a sample size formula to achieve significance level α and power β is

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

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

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

Because $\arcsin\left(\sqrt{\pi_T}\right) - \arcsin\left(\sqrt{\pi_C}\right)$ is not a function of $\pi_T - \pi_C$, we cannot express this in terms of the difference itself, but instead need to specify the expected probabilities of success in each group.

In practice, it is likely that (π_C) is well understood, and (π_T) can be specified by using the nearest clinically important value of π_T.

Example 5.1. (From Smith et al. 1994)

Two approaches to managing malignent low bile duct obstruction:

- surgical biliary bypass
- endoscopic insertion of a stent.

Primary outcome: 'Did the patient die within 30d of the procedure?'

Trial designed to have $\alpha = 0.05$, $\beta = 0.95$, which gives $z_{\frac{\alpha}{2}} = 1.96$, $z_\beta = 1.65$.

The trial wanted to be able to determine a change in 30 day mortality rate from 0.2 to at most 0.05. Equation (5.4)) gives

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

and so each group in our trial should contain 115 patients.

If instead our aim had been to detect a change from around 0.5 to 0.35 (the same in terms of $\pi_T - \pi_C$) then

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

that is, 281 patients per trial arm.

Later in the course we'll look at another way of estimating sample size.

Chapter 6

(Lecture 11) Analysis for binary outcomes

For a group of n participants, we will have allocated n_C to the control group (group C), and n_T to the treatment group (group T).

Natural to model number of 'successes' R_C by

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

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

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

and the focus of our analysis is on comparing π_C and π_T.

To do this we need

- point estimates of π_C and π_T
- interval estimates for some measure of the discrepancy between them
- ways to test $H_0: \ \pi_C = \pi_T$.

6.1 Point estimates and Hypothesis tests

First of all, we can tabulate the results of a trial with a binary outcome like this:

	Successes	Failures	Total
Treatment	r_T	$n_T - r_T$	n_T
Control	r_C	$n_C - r_C$	n_C
Total	r	$n - r$	n

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

We can estimate π_C and π_T by the sample proportions

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

We know that

$$\mathrm{E}\left(p_C\right) = \pi_C$$

and

$$\mathrm{Var}\left(p_C\right) = \frac{\pi_C\left(1-\pi_C\right)}{n_C},$$

and similarly for $\mathrm{E}\left(p_T\right)$ and $\mathrm{Var}\left(p_T\right)$.

If Y_{iC} is the outcome of the i-th patient in group C, with

- $Y_{iC} = 1$ if the participant's outcome is 'success'
- $Y_{iC} = 0$ otherwise.

Then we have

$$r_C = \sum_{i=1}^{n_C} y_{iC},$$

and similarly for group T.

Since p_C and p_T are therefore sample means, the Central Limit Theorem $\implies$ p_C and p_P can be approximated by normal distributions:

$$p_C \sim N\left(\pi_C,\, \frac{\pi_C\left(1-\pi_c\right)}{n_C}\right)$$
$$p_T \sim N\left(\pi_T,\, \frac{\pi_T\left(1-\pi_T\right)}{n_T}\right).$$

This means we can test the null hypothesis that $\pi_C = \pi_T$ by referring our observed value of $p_T - p_C$ to a normal distribution with mean 0 and variance

$$\frac{\pi_T\left(1-\pi_T\right)}{n_T} + \frac{\pi_C\left(1-\pi_c\right)}{n_C},$$

which we can approximate by substituting in p_C and p_T.

Under H_0, $\pi_C = \pi_T = \pi$, so it would be more appropriate to use this as the common variance.

The variance of $p_T - p_C$ becomes

$$\pi\left(1-\pi\right)\left(\frac{1}{n_C} + \frac{1}{n_T}\right),$$

and in calculations we replace π with $p = r/n$.

Putting all this together, our test statistic is

$$Z = \frac{p_T - p_C}{\sqrt{p\left(1-p\right)\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}}.$$

Example 6.1. From al (1948).

- 109 patients with tuberculosis
- Assigned to either receive Streptomycin (group T), or placebo (group C)
- Primary outcome variable is whether or not the patient was improved after the treatment period.

The data include several other covariates, including gender, baseline condition (good, fair or poor) and whether the patient had developed resistance to streptomycin after 6 months.

```
##               improved
## arm            FALSE TRUE
##   Streptomycin    17   38
##   Control         35   17
```

We therefore have

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

and can calculate our Z statistic to be

$$\begin{aligned} Z &= \frac{0.691 - 0.327}{\sqrt{0.514\,(1-0.514)\left(\frac{1}{52}+\frac{1}{55}\right)}} \\ &= 3.765. \end{aligned}$$

Finally, we can find the p-value of this test statistic (making sure to have two tails!)

```
2*(1-pnorm(3.765, mean=0, sd=1))
```

```
## [1] 0.0001665491
```

So we can reject the hypothesis that streptomycin has no effect on tuberculosis at the $\alpha = 0.05$ level (and indeed many lower levels).

6.1.1 An alternative approach: chi-squared

Another way to approach this would be to conduct a **chi-squared** test.

- Calculate the **expected** values (E_i) for each box of the summary table
- Compare them to the **observed** values (O_i) by finding the summary statistic

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i}.$$

Under $H_0: \pi_C = \pi_T$, the test statistic $X^2 \sim \chi^2_1$.

We see that the larger the differences between the observed and expected values, relative to the expected values, the larger the test statistic, and therefore the less probably under the χ^2_1 distribution.

Example 6.2. Continuing our streptomycin example, we can calculate a table of expected values by observing that proportion $p = 0.514$ of the total number of patients were improved. There are 52 in the control group, therefore we expect $0.514 \times 52 = 26.73$ improved patients in the control group, and by the same logic $0.514 \times 55 = 28.27$ in the treatment group. Our expected table is therefore

```
##               improved
## arm             FALSE   TRUE
##   Streptomycin 26.730 28.270
##   Control      25.272 26.728
```

We can therefore calculate the χ^2 statistic by looping through the elements of the tables:

```
sum_chi_sq = 0 # set a running total going
# in the following, tab_obs is the table of observed values and
# tab_exp is the table of expected values
for (i in 1:2){
  for (j in 1:2){
    tmp = ((tab_obs[i,j] - tab_exp[i,j])^2)/tab_exp[i,j]
    sum_chi_sq = sum_chi_sq + tmp
  }
}
sum_chi_sq
```

```
## [1] 14.17595
```

```
1-pchisq(sum_chi_sq, df=1)
```

```
## [1] 0.0001664847
```

and again we have a very significant result.

In fact, these two tests are almost equivalent, and we have that $\sqrt{X^2} = Z$:

```
sqrt(sum_chi_sq)
```

```
## [1] 3.765097
```

6.1.2 (Lecture 11) Likelihood: A more rigorous way

Our method above was quite informal, and also made heavy use of the central limit theorem. We can use maximum likelihood to derive a more formally justified test for binary outcomes. This also lays a good foundation for more complex situations.

Using same notation y_{iC} to be outcome variable (0 or 1, in this case) of the i-th participant in the control group (and so on).

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

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

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

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

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

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

Wwe can find the complete likelihood function

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

The log-likelihood is therefore

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

If we differentiate with respect to π_C, we find

$$\frac{\mathrm{d}l\left(\pi_C,\pi_T \mid n_C,n_T,r_C,r_T\right)}{\mathrm{d}\pi_C} = \frac{r_C}{\pi_C} - \frac{n_C-r_C}{1-\pi_C}.$$

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

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

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

where we use p, r, n to denote the pooled values ($n = n_C + n_T$ etc.).

Each term in the final line corresponds to a subgroup of the participants, as labelled, and if we rearrange them slightly we see that

This can be re-written as

$$\lambda_{LR} = 2\sum_{i\in G} o_i \log\left(\frac{o_i}{e_i}\right),$$

where G is the set of subgroups (group C success etc.).

Under the null hypothesis that $\pi_C = \pi_T = \pi$, and for sufficiently large n_C, n_T, λ_{LR} has a χ^2 distribution with one degree of freedom.

Example 6.3. Continuing with the streptomycin example, we can calculate this new test statistic in R by looping through the subgroups.

```
sum_LR = 0 # set a running total going
# in the following, tab_obs is the table of observed values and
# tab_exp is the table of expected values
for (i in 1:2){
  for (j in 1:2){
    tmp = tab_obs[i,j] * log(tab_obs[i,j]/tab_exp[i,j])
    sum_LR = sum_LR + tmp
  }
}
teststat_LR = 2*sum_LR
teststat_LR
```

```
## [1] 14.5028
```

```
1-pchisq(teststat_LR, df=1)
```

```
## [1] 0.0001399516
```

Not surprisingly, this value is quite close to the one we obtained earlier!

Having thought about tests for one proportion, we now move on to thinking how we might compare proportions.

Brief overview of the different measure's we'll be looking at:

Absolute risk difference (ARD): $\pi_T - \pi_C$

Number needed to treat (NNT): $\frac{1}{\text{ARD}} = \frac{1}{\pi_T - \pi_C}$

Risk ratio (RR): $\frac{\pi_T}{\pi_C}$

Odds Ratio (OR): $\frac{\pi_T/(1-\pi_T)}{\pi_C/(1-\pi_C)}$

6.2 (Lecture 12) Measures of difference for binary data

Important note: treating $\pi_T > \pi_C$ as good

Our question in the last lecture was 'is what we've observed statistically significant?' For streptomycin example the answer was a resounding 'Yes!'.

However, for questions like

- 'How big is the difference between the effects of each treatment?'
- 'What is the treatment effect?'

things are a bit less clear.

For continuous X, it made sense to think about the treatment effect as $\mu_T - \mu_C$.

In the binary case there are several ways we can think of the difference between two proportions π_C and π_T.

Each requires a different approach, so we will work our way through them in the next couple of lectures.

6.2.1 Absolute risk difference and Number Needed to Treat

The **absolute risk difference** is

$$\text{ARD} = \pi_T - \pi_C,$$

and is sometimes used. Loses a lot of information that we'd like to keep in: a change from $\pi_C = 0.03$ to $\pi_T = 0.01$ is very different from $\pi_C = 0.57$ to $\pi_T = 0.55$.

Eg., suppose a treatment reduces the incidence of some terrible symptom from $\pi_C = 0.03$ to $\pi_T = 0.01$. The absolute risk difference is 0.02 here. For some other treatment that results in a reduction from $\pi_C = 0.57$ to $\pi_T = 0.55$ we have the same absolute risk difference, even though it feels (and is!) a much less significant reduction.

BUT these numbers are (usually) about people. If the outcome is 'cured' or 'not cured', then for N patients, $N \times$ ARD is the number of extra patients you would expect to cure if you used treatment T instead of treatment C *(which may be nothing or may some usual course of treatment).*

Linked to this is the **number needed to treat** (NNT), which is defined as

$$\text{NNT} = \frac{1}{\pi_T - \pi_C} = \frac{1}{\text{ARD}}.$$

The NNT is the number of patients you'd need to treat (with treatment T rather than C) before you would bring benefit to one extra patient.

The website TheNNT collects together results from many clinical trials and uses the NNT as a summary. Some of the results are quite surprising, compared to how effective we think medicines are!

- Popular as a clinical benchmark
- Provides useful intuition in terms of the number of people it will help

Eg. $\pi_T = 0.25$, $\pi_C = 0.2$, then ARD $= 0.05$ and NNT $= 20$.

After treating 20 patients with treatment C we expect to cure (say) 4, whereas treating 20 patients with treatment T it is expected that we will cure 5.

For very small proportions, the NNT can be large even for what appears to be an important difference. For example, if $\pi_C = 0.005$ and $\pi_T = 0.015$ then ARD $= 0.01$ and NNT $= 100$.

It might be decided that the necessary changes and costs are not worth it for such a small difference.

That said, the NNT is not the easiest statistic to work with, as we shall see!

6.2.1.1 Confidence intervals for ARD and NNT

Let's make a confidence interval for the treatment difference $\tau_{ARD} = \pi_T - \pi_C$.

Using the same normal approximation as before, we can estimate τ_{ARD} by $p_T - p_C$, and $\text{var}\,(p_T - p_C)$ by

$$\frac{p_T\,(1-p_T)}{n_T} + \frac{p_C\,(1-p_C)}{n_C}.$$

Our $100\,(1-\alpha)\%$ confidence interval is therefore given by

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

Example 6.4. Back to our streptomycin example, we can now construct a $100\,(1-\alpha)\%$ confidence interval for the ARD.

Our estimated treatment effect is (to 3 decimal places)

$$\hat{\tau} = p_T - p_C = \frac{38}{55} - \frac{17}{52} = 0.364.$$

Our estimate of the standard error of $\hat{\tau}$ is

$$\frac{p_T\left(1-p_T\right)}{n_T}+\frac{p_C\left(1-p_C\right)}{n_C}=\frac{\frac{38}{55}\times\frac{17}{55}}{55}+\frac{\frac{17}{52}\times\frac{35}{52}}{52}$$
$$=0.0811$$

and therefore a 95% confidence interval for τ_{ARD} is

$$\left(0.364-z_{0.975}\sqrt{0.0811},\ 0.364+z_{0.975}\sqrt{0.0811}\right)=\left(0.187,\ 0.541\right).$$

As we should expect from the very low p-value we saw, the 95% confidence interval does not contain zero.

Our expected value of τ_{NNT} is

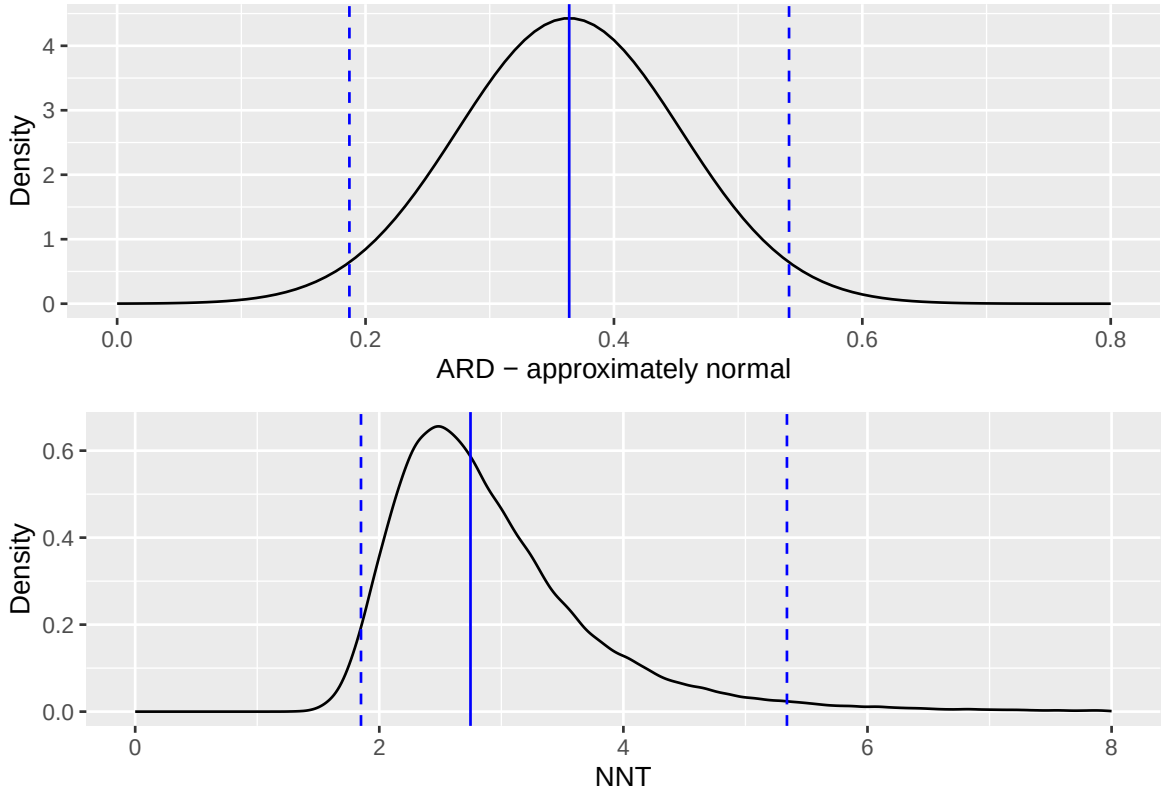
$$\text{NNT}=\frac{1}{\tau_{ARD}}=\frac{1}{0.364}=2.75.$$

That is, we would expect to treat three patients before one is improved (in terms of their tuberculosis symptoms).

We can use the limits of the 95% CI for τ_{ARD} to form a 95% CI for NNT: *simply by taking the reciprocals of the limits to get*

$$\left(\frac{1}{0.541},\ \frac{1}{0.178}\right)=\left(1.85,\ 5.34\right).$$

Because the NNT is the reciprocal of something approximately normally distributed, it has a distribution with a long tail, and we see that the confidence interval is therefore skewed.



6.2.1.2 What if the difference is not significant?

In the above section you might have already wondered what happens if the confidence interval for the absolute risk difference (ARD) contains zero.

To illustrate this, we will make up some data for a small trial:

	Successes	Failures	Total
Treatment	9	5	14
Control	4	8	12
Total	13	13	26

The ARD is now

$$\frac{9}{14} - \frac{4}{12} = \frac{3}{13} \approx 0.310$$

and our 95% confidence interval for τ_{ARD} is $(-0.0567,\ 0.676)$.

Our CI is very wide (this is not a very good trial!), and now contains zero.

It looks very likely that the treatment is effective (the interval only just contains zero) but how many patients might we need to treat before we expect to see an extra success?

The expected value of NNT is

$$\frac{1}{0.310} = 3.23,$$

which is fine.

However, our CI contains the possibility that $\tau_{ARD} = 0$, in which case the NNT is in some sense infinite:

no matter how many patients we treat, we don't expect to see any extra improvements.

Therefore, it feels appropriate that our CI for τ_{NNT} should contain infinity.

When thinking about a confidence interval for the NNT, we need to think about signs, and what negative and positive values mean.

- *If both the lower and upper limits of the confidence interval for ARD are positive, there is no issue - the treatment is effective, and our NNT confidence interval is another entirely positive interval.*
- *If the confidence interval for ARD is entirely negative, we have an entirely negative interval for NNT. A negative value of NNT can be thought of as the 'number needed to treat to harm one extra person'.*

The tricky situation is when the CI for τ_{ARD} is $(-L, U)$ with $L, U > 0$, ie. an interval containing zero.

As we approach zero from U, the upper limit of the CI for $\pi_T - \pi_C$, the number of patients we need to treat increases, since the treatment effect is getting smaller, until at $\pi_T - \pi_C = 0$ the NNT is infinite. Therefore, the part of the CI for NNT corresponding to the positive part of the CI for ARD is

$$\left(\frac{1}{U},\ \infty\right)$$

As we approach zero from the left in the interval (ie. from $-L$), we need to treat more and more patients to **harm** one more. In this region the NNT is negative, since if we deny some patients the treatment we will benefit a few.

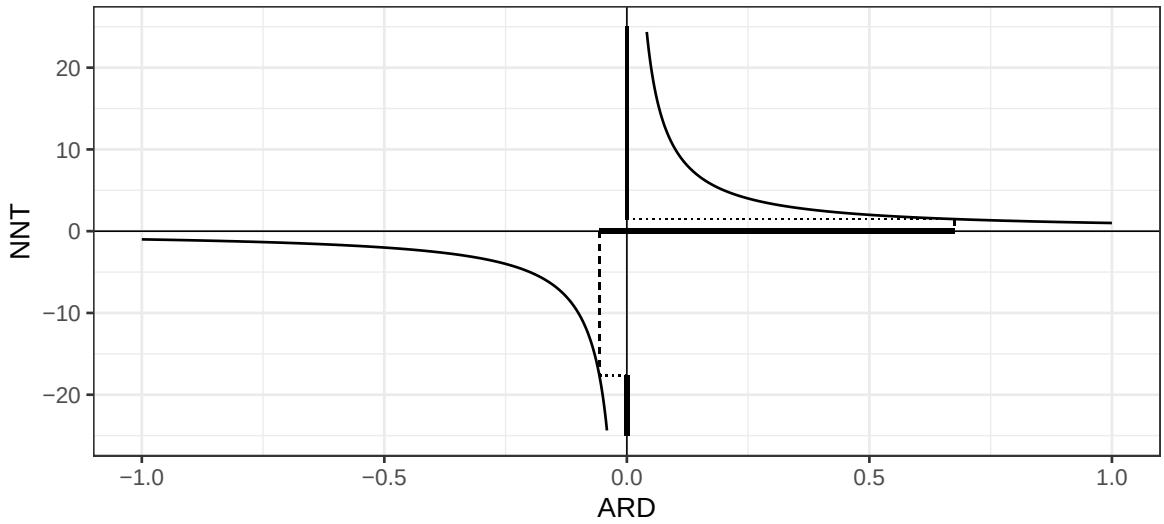
Therefore the CI for τ_{NNT} corresponding to the negative part of the CI for τ_{ARD} is

$$\left(-\infty,\ -\frac{1}{L}\right),$$

and altogether the confidence interval for the number needed to treat (NNT) is the union of these two intervals,

$$\left(-\infty,\ -\frac{1}{L}\right) \cup \left(\frac{1}{U},\ \infty\right).$$

The plot below shows relationship between ARD and NNT, with the intervals for our toy example shown in bold on the respective axis (the NNT interval should continue infinitely in both directions so for obvious reasons this is not all shown!).



Altman (1998) (available here) makes a compelling push for the use of confidence intervals for the number needed to treat. You can decide for yourself whether what you think of it!

Problems with the confidence interval for the ARD

The 'standard' method is not so reliable if the proportion is close to zero or one. The coverage probability of a 95% CI like in Section 6.2.1.1 often turns out to be more like 90% or even 85%.

Also, the limits of the 'standard' CI aren't forced to be in $[-1, 1]$.

Newcombe Method

Step 1: find an interval estimate for a single proportion π. As before, this can be written

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

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

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

'standard' method:

Substitute p (the estimated value of π from our sample) into the right hand side of Equation (6.1) for π, to get

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

which we solve to get the limits

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

Newcombe's method

Keep π in the right hand side and solve the quadratic in Equation (6.1) in terms of π.

The benefit of this new method will be most obvious for a probability that is close to 0 or 1.

Eg. Suppose we have 1 success out of 50 patients, so $p = 0.02,\ n = 50$.

The limits of a standard 95% confidence interval will be

$$\left(0.02 - z_{0.975}\sqrt{\frac{0.02 \times 0.98}{50}},\ 0.02 + z_{0.975}\sqrt{\frac{0.02 \times 0.98}{50}}\right) = (-0.0188,\ 0.0588),$$

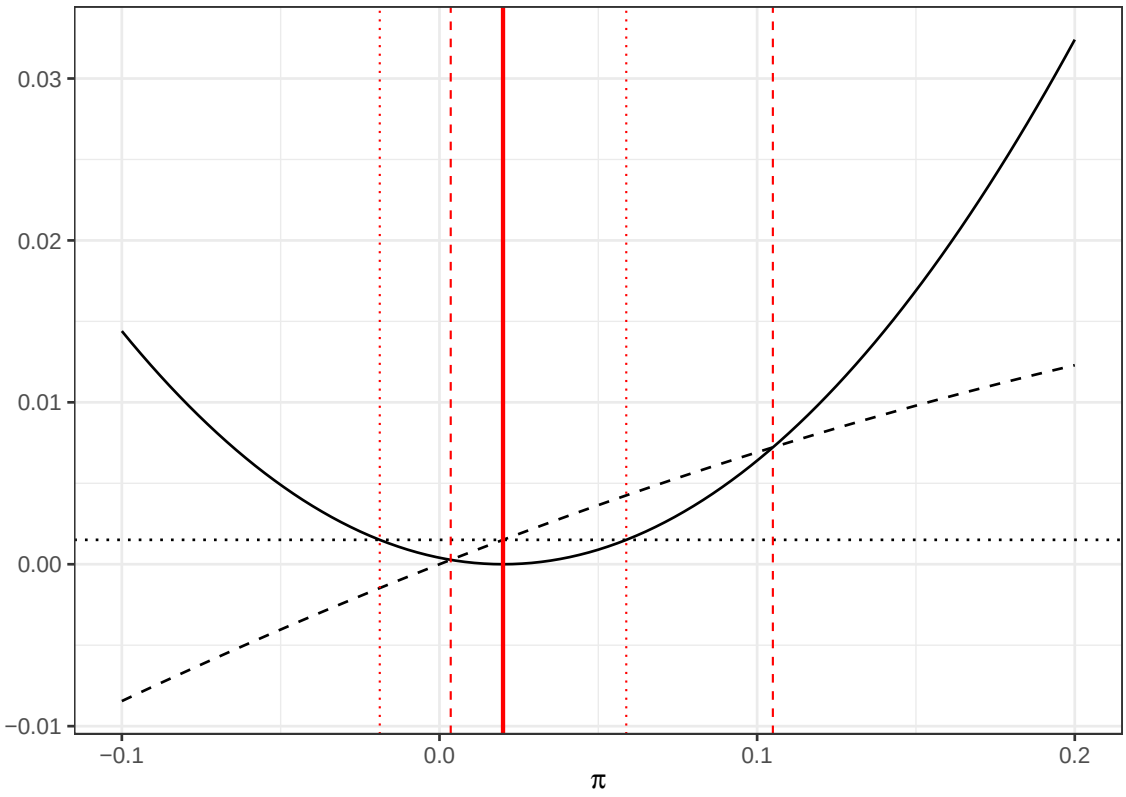
whereas the limits to the Newcombe 95% CI will be the roots of

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

which work out to be

```
## [1] 0.003539259 0.104954436
```

Visually, we can represent this as below by plotting the LHS (solid) and RHS (dashed for new method, dotted for standard method). The thick solid red line shows p_T, the estimated proportion, the thinner dashed red lines show the Newcombe 95% CI and the dotted red lines show the standard 95% CI. Notice that the limits of each confidence interval are formed by the points at which the solid line (LHS) crosses the dashed / dotted lines (RHS).



Example 6.5. Returning to our streptomycin example, our estimate of the probability of success for the treatment group is $p_T = \frac{38}{55}$, $n_T = 55$, and therefore our equation becomes

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

Solving this equation in the usual way (using the quadratic formula) we find the limits

```
## [1] 0.5597141 0.7971771
```

By contrast, in our standard method we have

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

which is

```
## [1] 0.5687797 0.8130385
```

We can see this graphically

Notice that the interval with the new method is now asymmetrical, which is more realistic.

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

```
## [1] 0.2152207 0.4624381
```

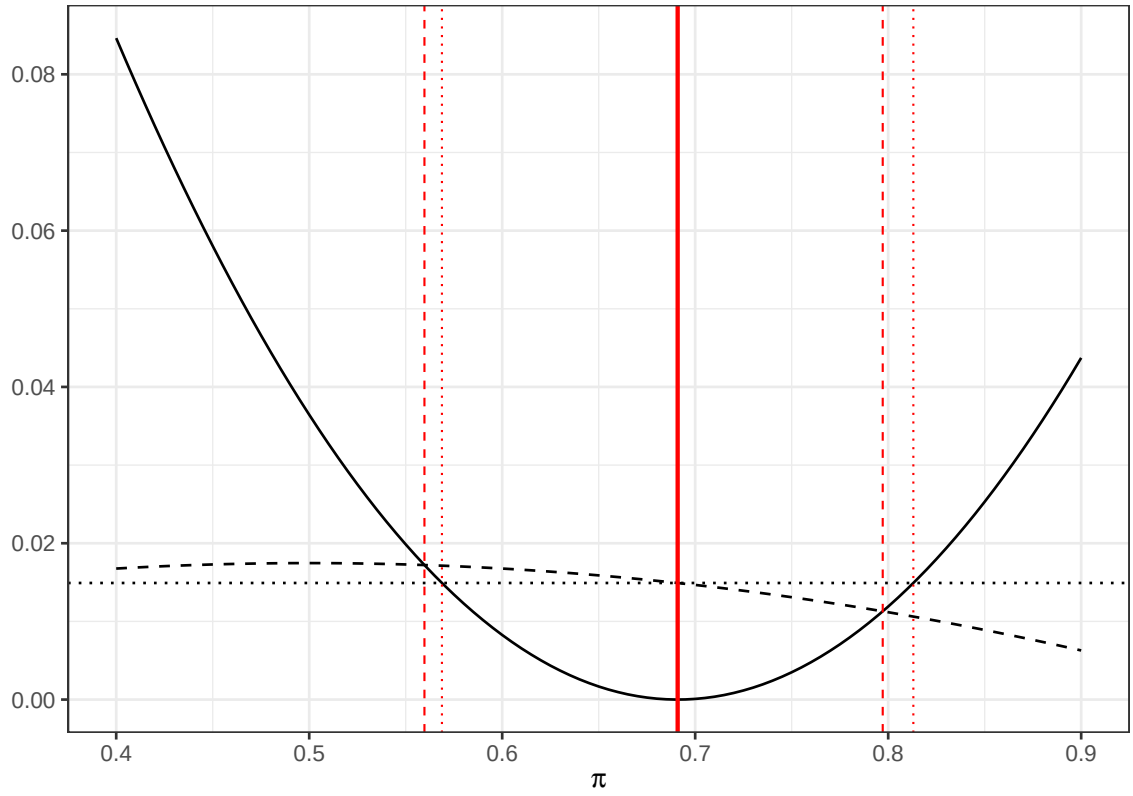


Figure 6.1: As before, dashed for Newcombe, dotted for standard

compared to the standard confidence interval

```
## [1] 0.1994256 0.4544205
```

Again, we can see this graphically.

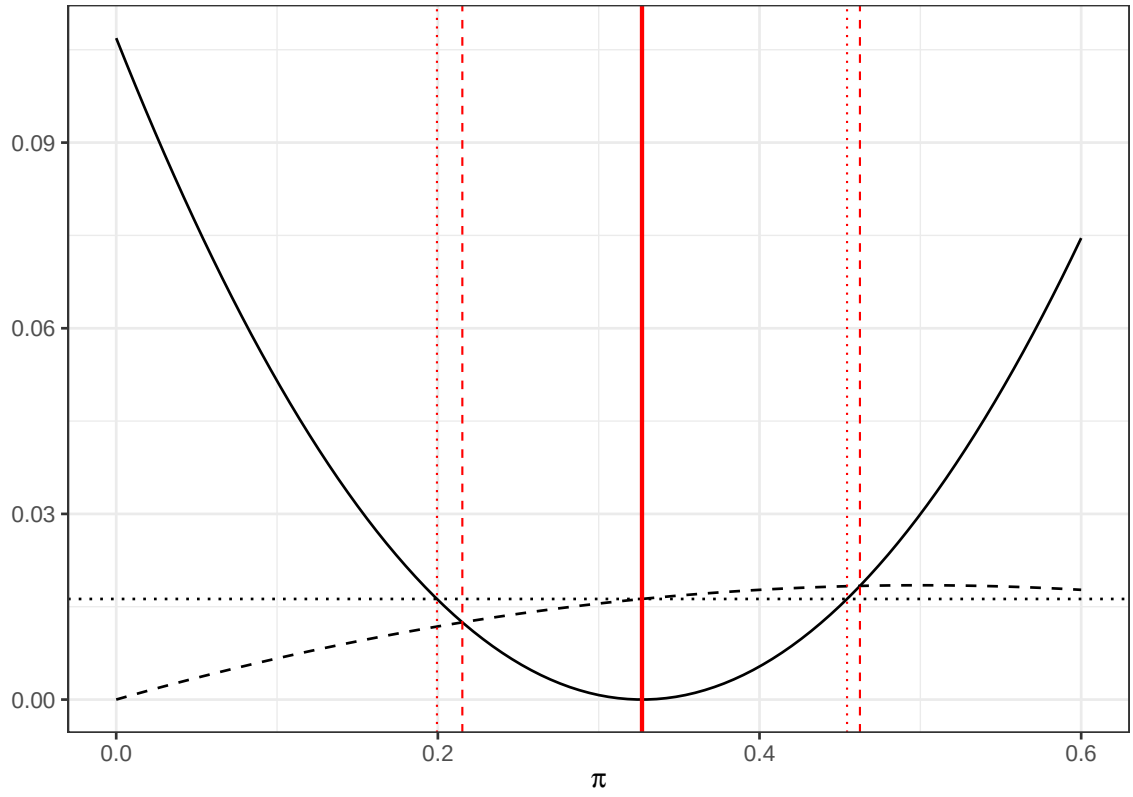


Figure 6.2: As before, dashed for Newcombe, dotted for standard

6.2.1.3 Extending this to $\pi_T - \pi_C$

What the Newcombe interval has given us is a superior method for creating confidence intervals for proportions. But, what we would like is a method for calculating a confidence interval for the difference in two proportions. You'll be relieved to hear that there is such a method, and we'll give a sketch here of how it works.

The limits of the 'standard method' confidence interval at significance level α are given by

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

We can rewrite this as

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

where ω_T and ω_C are the widths of the separate single-sample 'standard' confidence intervals for p_T and p_C.

Newcombe's method: proceed in the same way, but instead use the widths of the Newcombe confidence intervals for the individual probabilities p_T and p_C.

A bit more complicated, since for a Newcombe $(p_X - l_X) \neq (u_X - p_X)$ (for $X = T$ or C)

So, we have

$$\left(p_T - p_C - \sqrt{(p_T - l_T)^2 + (u_C - p_C)^2},\ p_T - p_C + \sqrt{(u_T - p_T)^2 + (p_C - l_C)^2}\right).$$

These differences must be calculated using the individual sample confidence interval method.

Example 6.6. Applying this Newcombe method to our Streptomycin example, recall that we have

$$\begin{aligned}
p_T &= \frac{38}{55} \\
p_T - l_T &= \frac{38}{55} - 0.5597 = 0.1312 \\
u_T - p_T &= 0.7972 - \frac{38}{55} = 0.1064 \\
p_C &= \frac{17}{52} \\
p_C - l_C &= \frac{17}{52} - 0.2152 = 0.1117 \\
u_C - p_C &= 0.4624 - \frac{17}{52} = 0.1355.
\end{aligned}$$

Our 95% confidence interval is therefore

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

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

6.2.2 (Lecture 13) Risk Ratio (RR) and Odds ratio (OR)

Measure so far, esp ARD, quite analagous to the continuous normally distributed case.

However, there are yet more commonly used measures of difference for proportions, which need to be dealt with differently, but also afford more opportunities for modelling.

The **risk ratio** is defined as

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

The **odds ratio** is defined as

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

Note that for both RR and OR:

- the null value is one, not zero.
- Always positive (assuming $\pi_C, \pi_T \neq 0, 1$)

We think about things multiplicatively, so for example if $RR = 3$ we can say that the event is "3 times more likely" in group T than in group C.

Odds

Reminder: The odds of some event A are

$$\frac{p(A)}{1 - p(A)}$$

So, if (for some event A), $p(A) = 0.2$, the odds of A are

$$\frac{p(A)}{p(A')} = \frac{0.2}{0.8} = \frac{1}{4},$$

which we say as "1 to 4" or 1:4. For every one time A occurs, we expect it not to occur four times.

*The **odds ratio** compares the odds of the outcome of interest in the Treament group with the odds of that event in the Control group. It tells us how the odds of the event are affected by the treatment (vs control).*

For probbabilities near zero, RR and OR are quite similar.

Example 6.7. For our Streptomycin example, we estimated the ARD by

$$\hat{\tau}_{ARD} = p_T - p_C = \frac{38}{55} - \frac{17}{52} = 0.364,$$

or could have alternatively had

$$\hat{\tau}_{ARD} = p_C - p_T = \frac{17}{52} - \frac{38}{55} = -0.364.$$

For the risk ratio, we have

$$\hat{\tau}_{RR} = \frac{p_T}{p_C} = \frac{38/55}{17/52} = 2.113,$$

or could alternatively have

$$\hat{\tau}_{RR} = \frac{p_C}{p_T} = \frac{17/52}{38/55} = 0.473 = \frac{1}{2.113}.$$

We could say that a patient is "more than twice as likely to be cured with streptomycin than by the control".

For the odds ratio, we have

$$\hat{\tau}_{OR} = \frac{p_T/(1-p_T)}{p_C/(1-p_C)} = \frac{(38/55)/(17/55)}{(17/52)/(35/52)} = 4.602,$$

and therefore the odds of recovery are around 4.6 greater for Streptomycin than for the control. Similarly, we could reframe this as

$$\hat{\tau}_{OR} = \frac{p_C/(1-p_C)}{p_T/(1-p_T)} = \frac{(17/52)/(35/52)}{(38/55)/(17/55)} = 0.217 = \frac{1}{4.602}.$$

6.2.2.1 Confidence intervals for RR and OR

Symmetry works differently on the RR and OR scale from on the ARD scale.

There is an equivalence between an interval (l, u) (with $l, u > 1$) and $\left(\frac{1}{u}, \frac{1}{l}\right)$, since these intervals would equate to comparing the same two treatments in different directions

Similarly, on this scale the interval

$$\left(\frac{1}{k}, k\right) \text{ for some } k > 1$$

can be thought of as symmetric, in that one treatment may be up to k times more effective than the other, in either direction.

Therefore, to build a confidence interval for OR or RR, we will not be following the usual formula

$$\text{point estimate } \pm z \times SE.$$

You may have already been thinking that a log transformation would be useful here, and you'd be correct! The sort-of symmetric intervals we've been discussing here actually are symmetric (about zero) on the log scale.

With a log transformation, the above interval becomes $(-\log k,\ \log k)$.

Firstly we'll consider the risk ratio. Let's define

$$\phi = \log\left(\frac{\pi_T}{\pi_C}\right).$$

The natural way to estimate this is with the sample proportions

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

These estimated proportions should be approximately normal and independent of one another, and so $\log\left(\frac{p_T}{p_C}\right)$ is approximately normal with mean ϕ (the true value) and variance

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

We can now apply the Delta method.

Reminder: If RV X has mean μ and variance $\sigma^2(\mu)$ then

$$var\left[f(X)\right] \approx \sigma^2(\mu)\left[f'(\mu)\right]^2.$$

We find

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

Since we estimate π_T by r_T/n_T this can be estimated by

$$\frac{1}{r_T} - \frac{1}{n_T}.$$

Notice that we are relying on the derivative of $\log(x)$ being x^{-1}, so we must always use natural logarithms.

This leads us to the result that, approximately

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

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

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

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

Example 6.8. Returning once again to our streptomycin example, recall that we have

$$\begin{aligned} r_T &= 38 \\ n_T &= 55 \\ r_C &= 17 \\ n_C &= 52 \end{aligned}$$

and so the limits of the confidence interval (with $\alpha = 0.05$) on the log scale are

$$\log\left(\frac{38/55}{17/52}\right) \pm 1.96\sqrt{\frac{1}{38} - \frac{1}{55} + \frac{1}{17} - \frac{1}{52}} = \log(2.11) \pm 1.96 \times 0.218$$

which gives us (0.320, 1.176) on the log scale, and a 95% CI for the risk ratio of (1.377, 3.243).

So, we've seen that we can find confidence intervals for each of our four measures of difference. But we probably want to also be able to incorporate baseline measurements, as we did for continuous outcome variables.

6.3 Accounting for baseline observations: logistic regression

We saw with the continuous outcomes that it is often advantageous to include baseline measurements of the outcome (if they are known) in our analysis, and this is the same for binary outcomes.

In this section we use the term 'baseline observations' to mean any measurement that was known before the trial started.

Unlike with continuous measurements, with a binary outcome, there is not usually a pre-trial value of the primary outcome.

A binary outcome is often already relative to pre-trial (for example 'Have the patient's symptoms improved?') or refers to an event that definitely wouldn't have happened pre-trial (for example 'Did the patient die within the next 6 months?' or 'Was the patient cured?').

However, as we saw with ANCOVA, we can include other sorts of covariates in a linear model, so this is fine.

The general form of model that we would like for patient i is

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

where

- G_i is an indicator function taking values 1 if patient i was in group T and 0 if they were in group C,
- $\text{baseline}_1, \ldots, \text{baseline}_p$ are p baseline measurements

Problems with binary variables.

The outcome for patient i will be either 0 or 1, but the terms in the model above do not guarantee this at all. Adding a normally distributed error term doesn't really make sense in this context, so we will remove it.

We can also make the LHS continuous by modelling mean outcome rather than a single outcome.

This makes sense, since if several patients were identical to patient i (in the sense of having the same baseline covariate values and being allocated to the same treatment), we probably wouldn't expect them all to have exactly the same outcome.

In which case our model becomes

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

However now, our LHS is in $[0, 1]$ but the RHS could take any real value.

To address this we use the **logit** transformation, which takes the mean outcome from $[0, 1]$ to $\mathbb{R}$.

The **logit** function is the log of the odds,

$$\text{logit}\,(\pi) = \log \frac{\pi}{1-\pi}.$$

As π tends to zero, $\text{logit}\,(\pi)$ tends to $-\infty$, and as π tends to one, $\text{logit}\,(\pi)$ tends to ∞.

The derivative of the logit function is

$$\frac{d\,\text{logit}\,(\pi)}{d\pi} = \frac{1}{\pi\,(1-\pi)}$$

which is always positive for $\pi \in [0,1]$. This means that we can use it to transform our mean outcome (which we will now call π, since the mean outcome is the estimate of the probability of success) in the model

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

and any value in $\mathbb{R}$ is allowed on both sides.

This model is known as **logistic regression**, and belongs to a class of models called **Generalized Linear Models**.

If you did Advanced Statistical Modelling III you'll have seen these before. If you haven't seen them, and want to know more, this article gives a nice introduction (and some useful R tips!).

6.3.1 What does this model tell us?

We now have an equation for a model that makes sense, but what is it actually modelling? And what does it tell us about the effect of the treatment?

Consider the difference between two patients who are the same in every respect except one is assigned to group C (so $G = 0$) and the other to group T (so $G = 1$).

The model gives:

$$\text{logit}\,(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\text{Odds of success group T}\right) = \mu + \tau + \beta_1 x_1 + \ldots + \beta_p x_p \qquad \text{(group T)}$$

$$\text{logit}\,(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\text{Odds of success group C}\right) = \mu + \beta_1 x_1 + \ldots + \beta_p x_p \qquad \text{(group C)}$$

Subtracting one from the other, we find

$$\begin{aligned}\log(\text{Odds of success for group T}) - \log(\text{Odds of success for group C}) & \\ = \log\left(\frac{\text{Odds of success for group T}}{\text{Odds of success for group C}}\right) &= \log\left(OR\right) \\ = \tau. &\end{aligned}$$

That is, τ is the log of the OR, or e^{τ} is the OR adjusted for variables $x_1, \ldots, x_p$.

While the baseline covariates $x_1, \ldots, x_p$ affect the probability of 'success', τ is a measure of the effect of the treatment compared to control given some set of baseline covariate values.

A logistic regression model can also be used to predict the odds of 'success' for a patient with particular characteristics.

Next Lecture:

- Fitting a logistic regression model
- Diagnostics for logistic regression

6.3.2 Fitting a logistic regression model

Logistic regression models are generally fitted using *maximum likelihood*. In the notation of Equation (6.4), the parameters we need to fit are the coefficients μ, τ and $\beta_1, \ldots, \beta_p$. To ease notation, we will collect these into a vector $\boldsymbol{\beta}$, with $\beta_0 = \mu$, $\beta_1 = \tau$ and $\beta_2, \ldots, \beta_{p+1}$ the original $\beta_1, \ldots, \beta_p$. Sorry this is confusing - we won't really use the vector $\boldsymbol{\beta}$ after this, or think about the parameters individually (apart from τ).

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

$$x_i^T \boldsymbol{\beta} = \sum_{j=0}^{q} x_{ij}\beta_j,$$

where

- $x_{i0} = 1$ (so that β_0 is the intercept μ)
- $x_{i1} = \begin{cases} 0 \text{ if participant } i \text{ is in group } C \\ 1 \text{ if participant } i \text{ is in group } T \end{cases}$
- $x_{i2}, \ldots, x_{iq}$ are the baseline covariates.

If π_i is the probability that the outcome for participant i is 1, where $i = 1, \ldots, n$, then the logistic model specifies these n parameters through the $q + 1$ parameters β_j, via the n expressions

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

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

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

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

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

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

This is generally done some via some numerical method, and we won't go into that here. The method used by R will generate the MLE $\hat{\beta}_j$ for each β_j, and also an estimate of the standard error of each $\hat{\beta}_j$. In particular there will be an estimate of the standard error of $\hat{\beta}_1$, better known as $\hat{\tau}$, the estimate of the treatment effect. This is important, because it means we can test the hypothesis that $\tau = 0$, and can form a confidence interval for the adjusted log odds ratio.

Example 6.9. This study is detailed in Elmunzer et al. (2012). ERCP, or endoscopic retrograde cholangio-pancreatogram, is a procedure performed by threading an endoscope through the mouth to the opening in the duodenum where bile and pancreatic digestive juices are released into the intestine. ERCP is helpful for treating blockages of flow of bile (gallstones, cancer), or diagnosing cancers of the pancreas, but has a high rate of complications (15-25%). The occurrence of post-ERCP pancreatitis is a common and feared complication, as pancreatitis can result in multisystem organ failure and death, and can occur in ~ 16% of ERCP procedures. This study tests whether the use of anti-inflammatory NSAID therapies at the time of ERCP reduce the rate of this complication. The study had 602 participants.

The dataset contains 33 variables, but we will focus on a small number:

- X: (primary outcome) - incidence of post-ercp pancreatitis 0 (no), 1 (yes).
- Treatment arm `rx`: 0 (placebo), 1 (treatment)
- Site: 1, 2, 3, 4
- Risk: Risk score (1 to 5). Should be factor but treated as continuous.
- Age: from 19 to 90, mean 45.27, SD 13.30.

The correlation between `risk` and `age` is -0.216, suggesting no problems of collinearity between those two variables.

Note: an obvious one to include would be `gender`, but I tried it and it is not at all significant, so I have pre-whittled it down for [even more] simplicity.

```
data("indo_rct")
summary(indo_rct[ ,c(1,2,3,4,6,32)])
```

```
##        id           site          age             risk        outcome                  rx
##  Min.   :1001   1_UM  :164   Min.   :19.00   Min.   :1.000   0_no :523   0_placebo     :307
##  1st Qu.:1152   2_IU  :413   1st Qu.:35.00   1st Qu.:1.500   1_yes: 79   1_indomethacin:295
##  Median :2138   3_UK  : 22   Median :45.00   Median :2.500
##  Mean   :1939   4_Case:  3   Mean   :45.27   Mean   :2.381
##  3rd Qu.:2289                3rd Qu.:54.00   3rd Qu.:3.000
##  Max.   :4003                Max.   :90.00   Max.   :5.500
```

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

## Checking for collinearity with factor variables

# No consistent patterns between age and site or risk and site
indo_rct%>%
  group_by(site) %>%
  summarise(
    meanage=mean(age), sdage=sd(age),
    meanrisk = mean(risk), sdrisk=sd(risk)
    )
```

```
## # A tibble: 4 x 5
```

```
##   site   meanage sdage meanrisk sdrisk
##   <fct>    <dbl> <dbl>    <dbl>  <dbl>
## 1 1_UM      47.2  14.2     2.06  0.888
## 2 2_IU      44.4  12.9     2.52  0.846
## 3 3_UK      45.9  11.6     2.23  0.896
## 4 4_Case    47.3  22.9     1.67  0.289
```

```
## We will try models with age and age^2

glm_indo_agelin = glm(outcome ~ age + site + risk + rx, data=indo_rct,
                      family = binomial(link = "logit"))
glm_indo_agesq = glm(outcome ~ I(age^2) + site + risk + rx, data=indo_rct,
                     family = binomial(link = "logit"))

summary(glm_indo_agelin)
```

```
##
## Call:
## glm(formula = outcome ~ age + site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.786293   0.641354  -2.785  0.00535 **
## age               -0.008458   0.009921  -0.853  0.39390
## site2_IU          -1.229290   0.269258  -4.565 4.98e-06 ***
## site3_UK          -1.127935   0.775917  -1.454  0.14603
## site4_Case       -13.864394 827.921132  -0.017  0.98664
## risk               0.561880   0.142342   3.947 7.90e-05 ***
## rx1_indomethacin  -0.763269   0.261538  -2.918  0.00352 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.07  on 595  degrees of freedom
## AIC: 441.07
##
## Number of Fisher Scoring iterations: 14
```

```
summary(glm_indo_agesq)
```

```
##
## Call:
## glm(formula = outcome ~ I(age^2) + site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.954e+00  4.930e-01  -3.963 7.39e-05 ***
```

```
## I(age^2)         -9.388e-05  1.081e-04  -0.869  0.38498
## site2_IU         -1.231e+00  2.693e-01  -4.571 4.87e-06 ***
## site3_UK         -1.135e+00  7.759e-01  -1.463  0.14355
## site4_Case       -1.385e+01  8.275e+02  -0.017  0.98664
## risk              5.593e-01  1.427e-01   3.919 8.88e-05 ***
## rx1_indomethacin -7.617e-01  2.614e-01  -2.914  0.00357 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.03  on 595  degrees of freedom
## AIC: 441.03
##
## Number of Fisher Scoring iterations: 14
```

Since neither `age` nor `age^2` appear influential, we'll remove it and keep the other covariates.

```
glm_indo = glm(outcome ~ site + risk + rx, data=indo_rct, family = binomial(link = "logit"))
summary(glm_indo)
```

```
##
## Call:
## glm(formula = outcome ~ site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.2307     0.3814  -5.848 4.97e-09 ***
## site2_IU          -1.2204     0.2689  -4.539 5.66e-06 ***
## site3_UK          -1.1289     0.7755  -1.456  0.14546
## site4_Case       -13.8400   833.2426  -0.017  0.98675
## risk               0.5846     0.1395   4.191 2.78e-05 ***
## rx1_indomethacin  -0.7523     0.2610  -2.883  0.00395 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.81  on 596  degrees of freedom
## AIC: 439.81
##
## Number of Fisher Scoring iterations: 14
```

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

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

This model supports the hypothesis that the treatment difference isn't zero. We do see however from the Null deviance and the Residual deviance that the model isn't explaining a huge proportion of the variation.

We can also use the model to estimate the odds of 'success' (the outcome 1) for different groups of patients, by fixing the values of the covariates. The linear expression $x^T\hat{\boldsymbol{\beta}}$ for given values of x gives us as estimate of

$$\log\left(\frac{p(X=1)}{1-p(X=1)}\right),$$

where X here is the primary outcome. The exponent of this therefore gives the odds, and this can be rearranged to find the probability,

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

where logit_i is the fitted value of the linear model (on the logit scale) given all the baseline characteristics of some patient i. This will be the probability, according to the model, that a patient with this particular combination of baseline characteristics will have outcome 1.

Example 6.10. Continuing with Example 6.9, we can find estimates of the log odds (and therefore the odds) of post-ECRP pancreatitis for various categories of patient.

For this we will make heavy use of the summary table

```
summary(glm_indo)
```

```
##
## Call:
## glm(formula = outcome ~ site + risk + rx, family = binomial(link = "logit"),
##     data = indo_rct)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.2307     0.3814  -5.848 4.97e-09 ***
## site2_IU          -1.2204     0.2689  -4.539 5.66e-06 ***
## site3_UK          -1.1289     0.7755  -1.456  0.14546
## site4_Case       -13.8400   833.2426  -0.017  0.98675
## risk               0.5846     0.1395   4.191 2.78e-05 ***
## rx1_indomethacin  -0.7523     0.2610  -2.883  0.00395 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 468.01  on 601  degrees of freedom
## Residual deviance: 427.81  on 596  degrees of freedom
## AIC: 439.81
##
## Number of Fisher Scoring iterations: 14
```

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

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

which translates to a probability of post-ECRP pancreatitis of

$$\frac{0.6207}{1 + 0.6207} = 0.383.$$

By contrast, a patient in group T, from site 2, at risk level 1, would have odds

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

which is equivalent to a probability of post-ECRP pancreatitis of

$$\frac{0.0268}{1 + 0.0268} = 0.0261.$$

Being more methodical we can collect these into a table. Since the site 3 and 4 coefficents are not significant (mainly due to a lack of data), we will treat them as zero and lump them in with the site 1 participants

```
##      site risk Odds_groupC Prob_groupC Odds_groupT Prob_groupT
## 1       2    1       0.057       0.054       0.027       0.026
## 2       2    2       0.102       0.093       0.048       0.046
## 3       2    3       0.183       0.155       0.086       0.079
## 4       2    4       0.329       0.248       0.155       0.134
## 5       2    5       0.590       0.371       0.278       0.218
## 6   Not 2    1       0.193       0.162       0.091       0.083
## 7   Not 2    2       0.346       0.257       0.163       0.140
## 8   Not 2    3       0.621       0.383       0.293       0.227
## 9   Not 2    4       1.114       0.527       0.525       0.344
## 10  Not 2    5       1.998       0.666       0.942       0.485
```

Some cautions

As with any linear model, we need to ensure that it is appropriate for our dataset. Two key things we need to check for are:

- **Collinearity**: we should make sure that none of the independent variables are highly correlated. This is not uncommon in clinical datasets, since measurements are sometimes strongly related. Sometimes therefore, this can mean choosing only one out of a collection of two or more strongly related variables.
- **linear effect across the range of the dataset**: a linear model is based on the assumption that the effect of the independent variables is the same across the whole range of the data. This is not always the case. For example, the rate of deterioration with age can be more at older ages. This can be dealt with either by binning age into categories, or by using a transformation, eg. age^2. Note that this would still be a linear model, because it is linear in the coefficients.

6.4 Diagnostics for logistic regression

There are many diagnostic techniques for binomial data (see eg. Collett (2003)) but we will only touch on a small number. Unlike with a linear regression model, we don't have residuals to analyse, because our model output is fundamentally different from our data: our model outputs are probabilities, but our data is all either 0 or 1. Just because a particular patient had an outcome of `1`, we can't conclude that their probability should have been high. If the 'true' probability of $X = 1$ for some group of similar (in the baseline covariates sense) patients is 0.9, this means we should expect 1 in 10 of these patients to have $X = 0$.

This makes diagnostics somewhat trickier.

Diagnostics for logistic regression fall into two categories: **discrimination** and **calibration**. We will look at each of these in turn, though by no means exhaustively.

6.4.1 Discrimination

Here we are thinking of the logistic regression model as a classifier: for each participant the model outputs some value, on the $\text{logit}(p)$ scale. If that value is below some threshold, we classify that participant as 0 If the value is above the threshold, we classify them as 1. Here, we are slightly abandoning the notion that the model is predicting probabilities, and instead testing whether the model can successfully order the patients correctly. Can we set some threshold on the model output that (almost) separates the cohort into its ones and zeros?

A classic way to asses this is by using Receiver Operating Characteric (ROC) analysis. ROC analysis was developed during the second world war, as radar operators analysed their classification accuracy in distinguishing signal (eg. an enemy plane) from noise. It is still widely used in the field of statistical classification, including in medical diagnostics. ROC analysis can be applied to any binary classifier, not just logistic regression.

6.4.1.1 ROC analysis

To understand ROC analysis, we need to revisit two concepts relating to tests or classifiers that you might not have seen since Stats I, and we will introduce (or remind ourselves of) some notation to do this:

- $\hat{p}_i \in (0, 1)$ is the fitted value of the logistic regression model for patient i
- $X_i = 0$ or 1 is the true outcome for patient i
- $t \in (0, 1)$ is the threshold value.

If $\hat{p}_i < t$ we classify patient i as 0, if $\hat{p} \geq t$ we classify them as 1. The language of ROC analysis is so entrenched in diagnostic/screening tests that I have kept it here for consistency. A 'positive' result for us is $X = 1$, and a 'negative' result is $X = 0$.

Definition 6.1. The **sensitivity** of a test (or classifier) is the probability that it will output positive (or 1) if the true value is positive (or 1):

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

Definition 6.2. The **specificity** of a test (or classifier) is the probability that it will output negative (or 0) if the true value is negative (or 0):

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

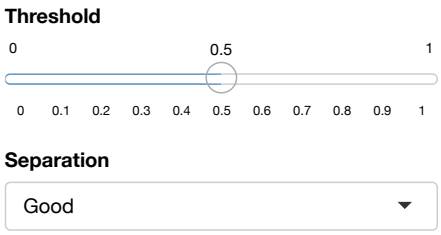
We estimate these by the proportions within the dataset.

These are very commonly used for thinking about diagnostic tests and screening tests, and in these contexts a 'success' or 'positive' is almost always the presence of some condition or disease. In our context, we need to be mindful that a 1 could be good or bad, depending on the trial.

The core part of a ROC analysis is to plot **sensitivity** against **1-specificity** for every possible value of the threshold. In a logistic regression context, the lowest the threshold can be is zero. If we set the $t = 0$, the model will predict everyone to have an outcome of 1. The sensitivity will be 1 and the specificity will be 0. At the other extreme, if we set $t = 0$, we will classify everyone as a 0, and have sensitivity 0 and specificity 1. If we vary the threshold from 0 to 1 the number of people classified in each group will change, and so will the sensitivity and specificity. This forms a **ROC curve**.

The dashboard below shows the distributions of fitted values for patients with $X = 0$ and $X = 1$, with options for good, moderate and poor separation, and the corresponding ROC curve. You can move the threshold to see the sensitivity and specificity at that value. Also note the AUC (area under the curve) which is an overall summary of the model's predictive efficacy. If AUC=1, the model is perfect. If AUC is 0.5, the model is no better than random guessing. Generally it is thought that AUC around 0.8 is quite good, and AUC around 0.9 is excellent.

If you're viewing this in PDF you'll just have a static image, but you can find the dashboard at https://racheloughton.shinyapps.io/ROCplots/.



Note that I've used beta distributions for some hypothetical distributions of fitted values for the different groups, but this is just for convenience: ROC analysis makes no distributional assumptions.

Example 6.11. Let's look at the model we fitted in Example 6.9. To draw the ROC curve of this

data, we will use the R package `pROC`.

```
fit_indo = fitted(glm_indo)    # Fitted values from glm_indo
out_indo = indo_rct$outcome    # outcome values (0 or 1)
roc_indo_df = data.frame(fit = fit_indo, out = out_indo)
```

The main function in the package `pROC` is `roc`, which creates a `roc` object. and `ggroc` that sort and plot the data for us:

```
roc_indo = roc(data=roc_indo_df, response = out, predictor=fit)
```

With that object we can do various things, such as plot the ROC curve:

```
ggroc(roc_indo, legacy.axes=T) + geom_abline(slope=1, intercept=0, type=2)
```

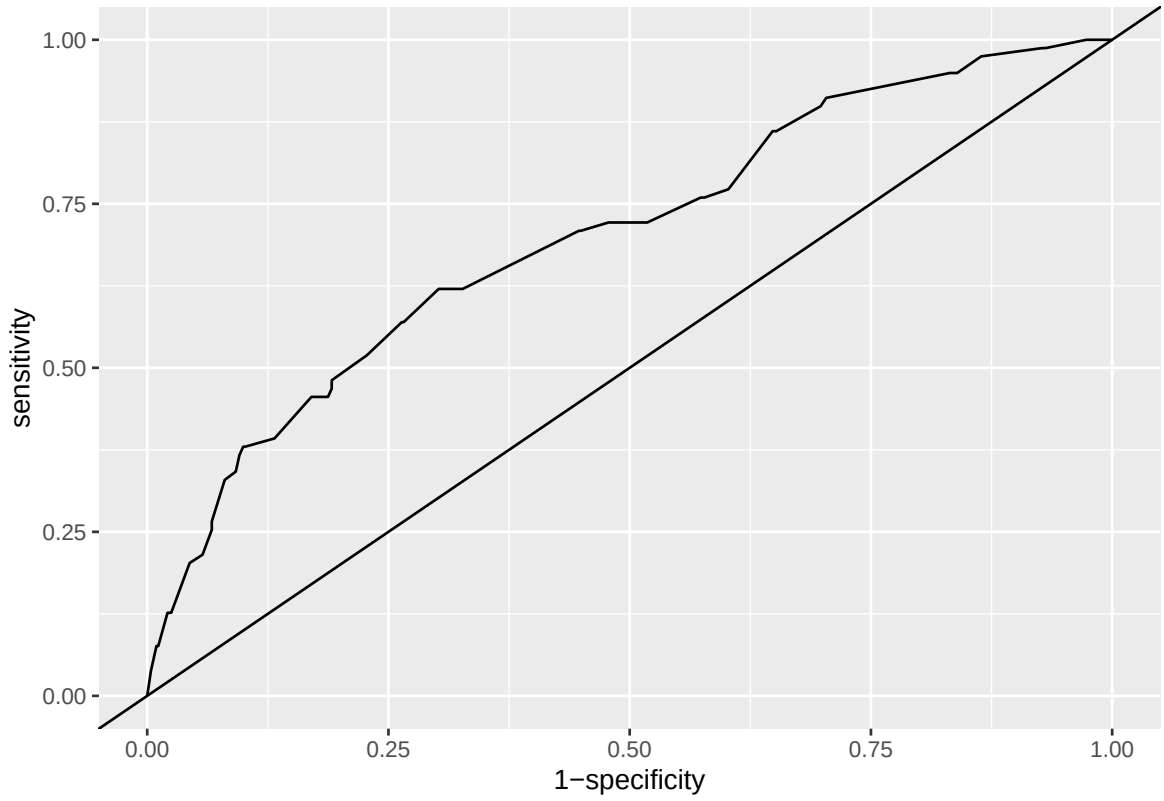


Figure 6.3: ROC curve for our logistic regression model of the indo RCT data (solid line). The dotted line shows the ROC curve we'd expect with random guessing.

and find the area under the curve for the model

```
auc(roc_indo)
```

```
## Area under the curve: 0.7
```

So we see that our model is better than random guessing, but really not all that good! In particular, wherever we put a threshold (if we use the model that way), many people will be mis-classified. It's also worth noting that here we're performing the diagnostics on the data we used to fit the model: if we were to use the model on a new set of patients, the fit would likely be slightly worse.

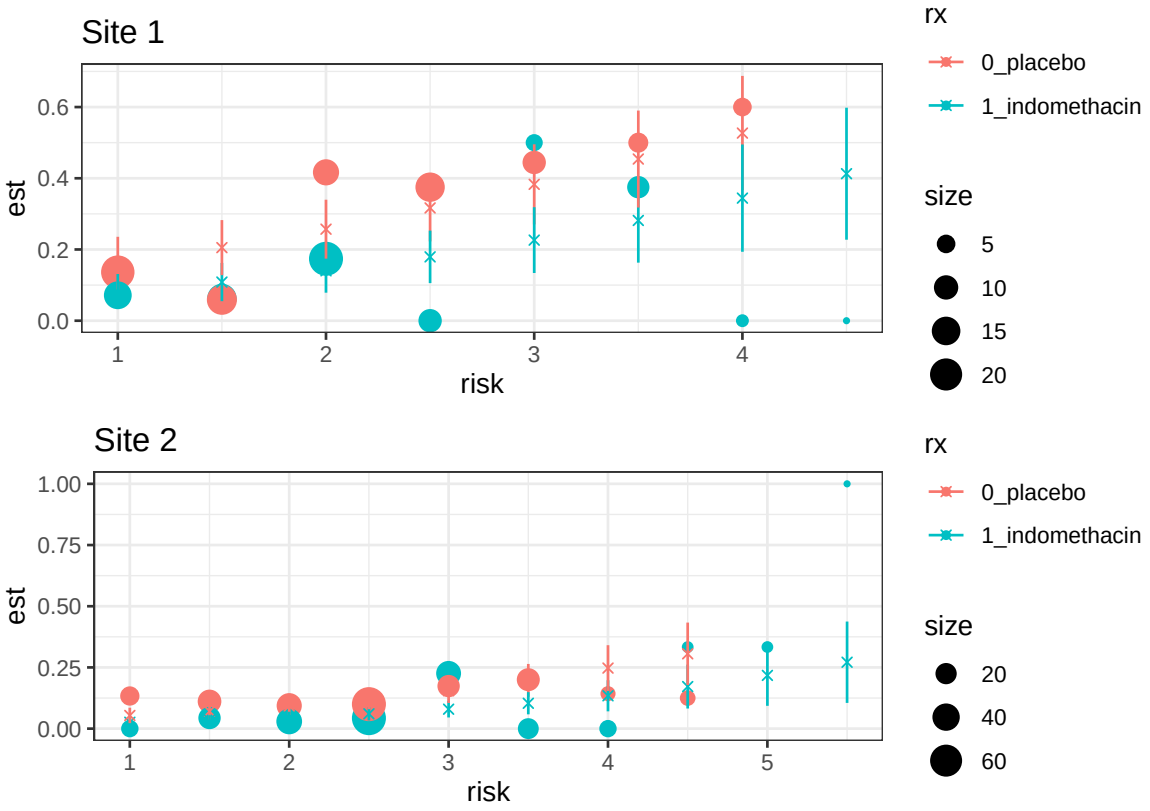
6.4.2 Calibration

Now we are thinking of the model as actually predicting probabilities, and therefore we want to determine whether these probabilities are, in some sense, 'correct' or 'accurate'. One intuitive way to do this is to work through different 'types' of patient (by which we mean different combinations of baseline covariate values) and see whether the proportions of ones in the data broadly match the probability given by the model.

If the explanatory variables are factors, and we have repeated observations for the different combinations of factor levels, then for each combination we can estimate the probability of success (or whatever our outcome variable is) using the data, and compare this to the fitted model value.

Example 6.12. This example uses the model fitted in Example 6.9.

The trial has 602 participants and there are many fewer than 602 combinations of the above factor variables, so for many such combinations we will have estimates. Since we are in three dimensions, plotting the data is moderately problematic. We will have a plot for each site (or for the two main ones), use risk score for the x axis and colour points by treatment group. The circles show the proportions of ones in the data, and are sized by the number of observations used to calculate that estimate, and the crosses and lines show the mean and 95% CI of the fitted value.



These plots are not the easiest to interpret, but there seems to be no evidence of systematic trends away from the model.

We will look some more at this in the upcoming practical class, as well as some further principles of model validation.

For now, we're done with Binary data, and in our next few lectures we'll think about survival, or time-to-event data.

References

This sections lists the references used in the course - it will be updated as the notes are updated. Some of the more accessible (dare I say 'interesting') resources are linked from the notes. If you want to read any of these articles, the easiest way is to copy the title into Google scholar.

al, Geoffrey Marshall et. 1948. "STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS a MEDICAL RESEARCH COUNCIL INVESTIGATION." *British Medical Journal*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2091872/pdf/brmedj03701-0007.pdf.

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

———. 1998. "Confidence Intervals for the Number Needed to Treat." *Bmj* 317 (7168): 1309–12.

Collett, David. 2003. *Modelling Binary Data*. 2nd ed. Texts in Statistical Science. Chapman & Hall.

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.

Elmunzer, B Joseph, James M Scheiman, Glen A Lehman, Amitabh Chak, Patrick Mosler, Peter DR Higgins, Rodney A Hayward, et al. 2012. "A Randomized Trial of Rectal Indomethacin to Prevent Post-ERCP Pancreatitis." *New England Journal of Medicine* 366 (15): 1414–22.

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.

Hommel, EHEBMJ, Hans-Henrik Parving, Elisabeth Mathiesen, Berit Edsberg, M Damkjaer Nielsen, and Jørn Giese. 1986. "Effect of Captopril on Kidney Function in Insulin-Dependent Diabetic Patients with Nephropathy." *Br Med J (Clin Res Ed)* 293 (6545): 467–70.

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.

Kendall, John. 2003. "Designing a Research Project: Randomised Controlled Trials and Their Principles." *Emergency Medicine Journal: EMJ* 20 (2): 164.

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.

Smith, AC, JF Dowsett, RCG Russell, ARW Hatfield, and PB Cotton. 1994. "Randomised Trial of

Endoscopic Steriting Versus Surgical Bypass in Malignant Low Bileduct Obstruction." *The Lancet* 344 (8938): 1655–60.

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.