Clinical Trials 4H - lecture notes

Rachel Oughton

2024-01-08

Contents

Chapter 1

Introduction

This is just here to preserve the numbering!

Lecture 2

Chapter 2

Sample size

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

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

The topics we'll cover fall into two categories:

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

but there is some interaction between these phases.

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

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

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

1. Not enough particpants may mean not enough evidence to come to a conclusion. This is both scientifically disappointing and unethical. *To conduct the trial, some of the patients will have been subject to an inferior treatment (assuming one treatment was actually better), and if there is no conclusion then this was effectively for no purpose.*
2. Too many patients *(ie. we would be sufficiently likely to reach a conclusion with many fewer)* means subjecting more patients than necessary to an inferior treatment. *Possibly also taken up more time and resources than was necessary.*

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

2.1 The treatment effect

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

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

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

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

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

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

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

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

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

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

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

2.2.1 Insignificant results

If our P-value is large, say 0.3 or 0.5, then our result is not at all unlikely under the null hypothesis, and provides no evidence to reject H_0. However, it is not inconsistent with the existence of a treatment effect, so we don't say there is evidence to accept H_0.

If the true treatment effect τ were tiny, many trials would fail to find evidence to reject H_0. However, if our sample size were sufficiently large, we should be able to detect it. Conversely, if τ is very large, even a relatively small sample size is likely to provide enough evidence to reject H_0.

A non-significant P-value means our results are consistent with $H_0 : \tau = 0$, and also with some small treatment effect.

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

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

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

2.2.2 One-sided or two-sided?

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

Therefore should we perform a one-sided test, with

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

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

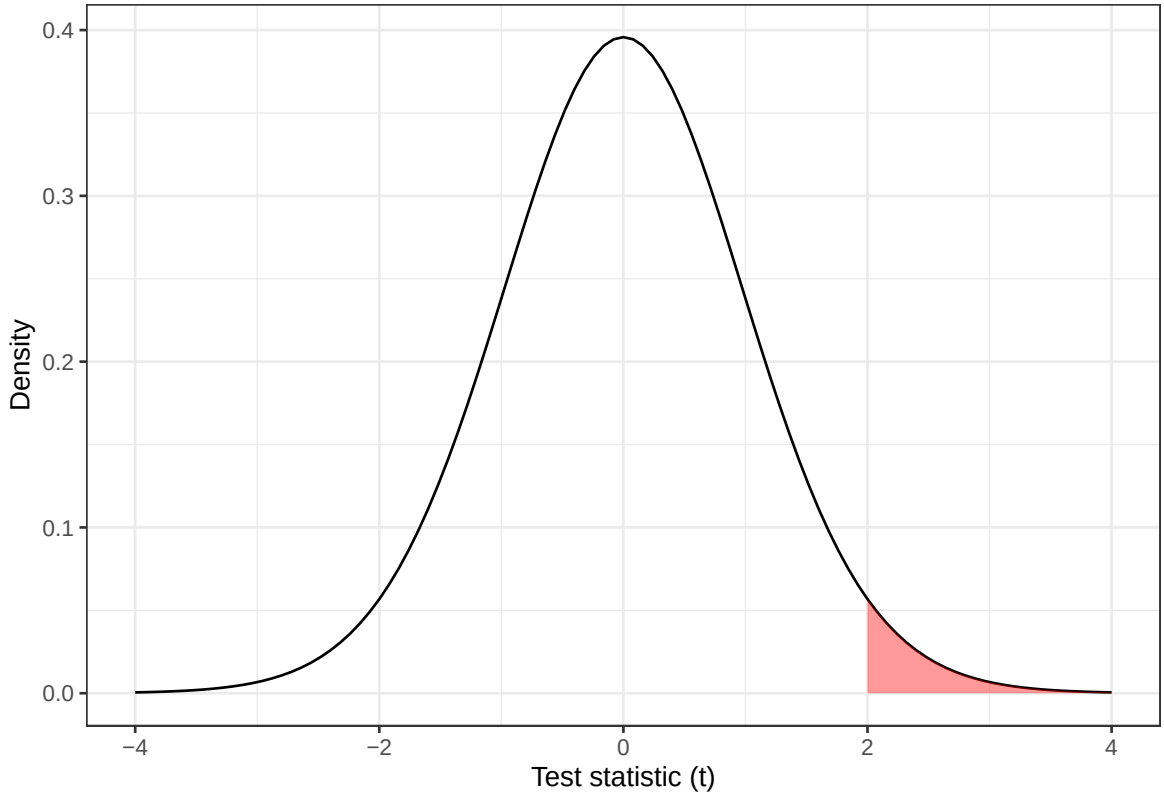


Figure 2.1: The distribution t_{31}, with the area corresponding to $t > 2$ shaded.

If $t \gg 0$, we obtain a small P-value, and reject H_0. Conclusion: the intervention is effective (in a good way). But what if we obtain $t \ll 0$? In this one-sided set-up, there is no value of $t < 0$ that would give a significant result.

Negative values of t *are simply considered consistent with* H_0*, and there is no way to conclude that an intervention has a significantly negative effect.*

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

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

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

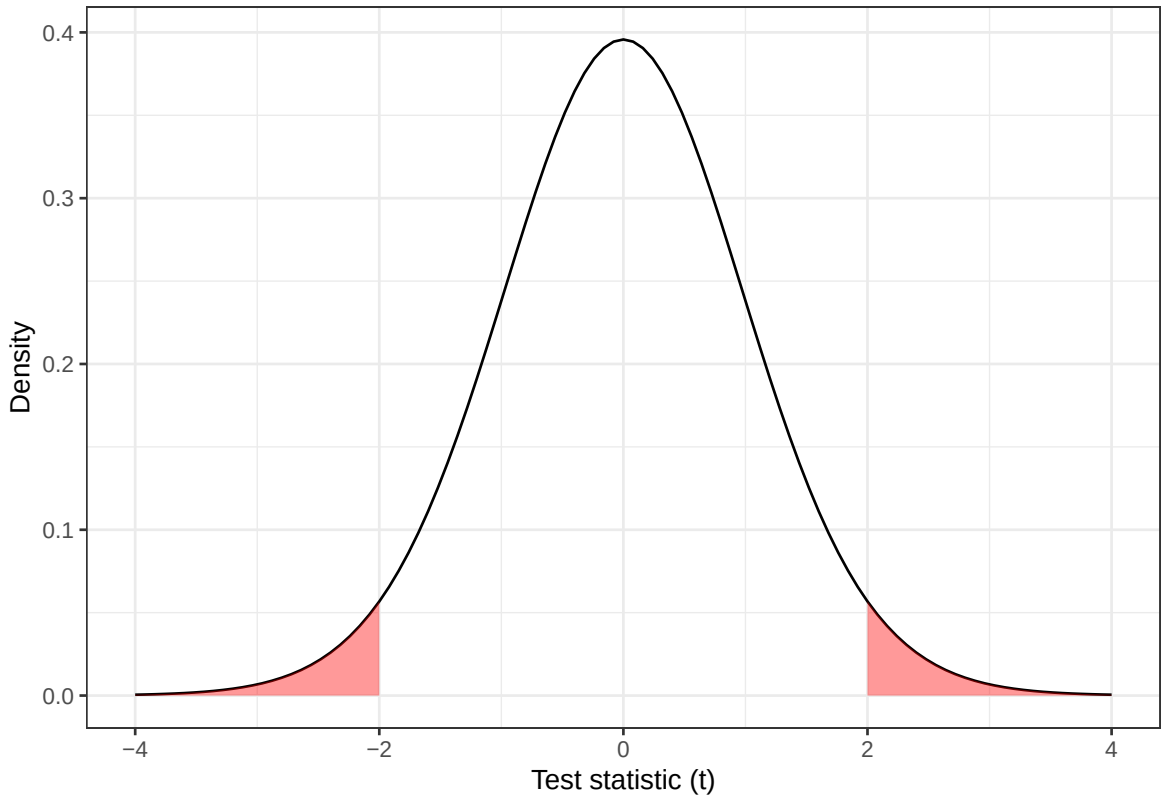


Figure 2.2: The distribution t_{31}, with the area corresponding to $|t| > 2$ shaded.

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

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

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

2.3 Constructing a measure of effect size

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

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

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

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

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

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

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

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

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

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

Using these values we can compute

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

as a standardised measure of the effect τ.

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

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

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

and therefore

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

and

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

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

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

and so we have

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

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

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

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

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

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

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

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

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

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

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

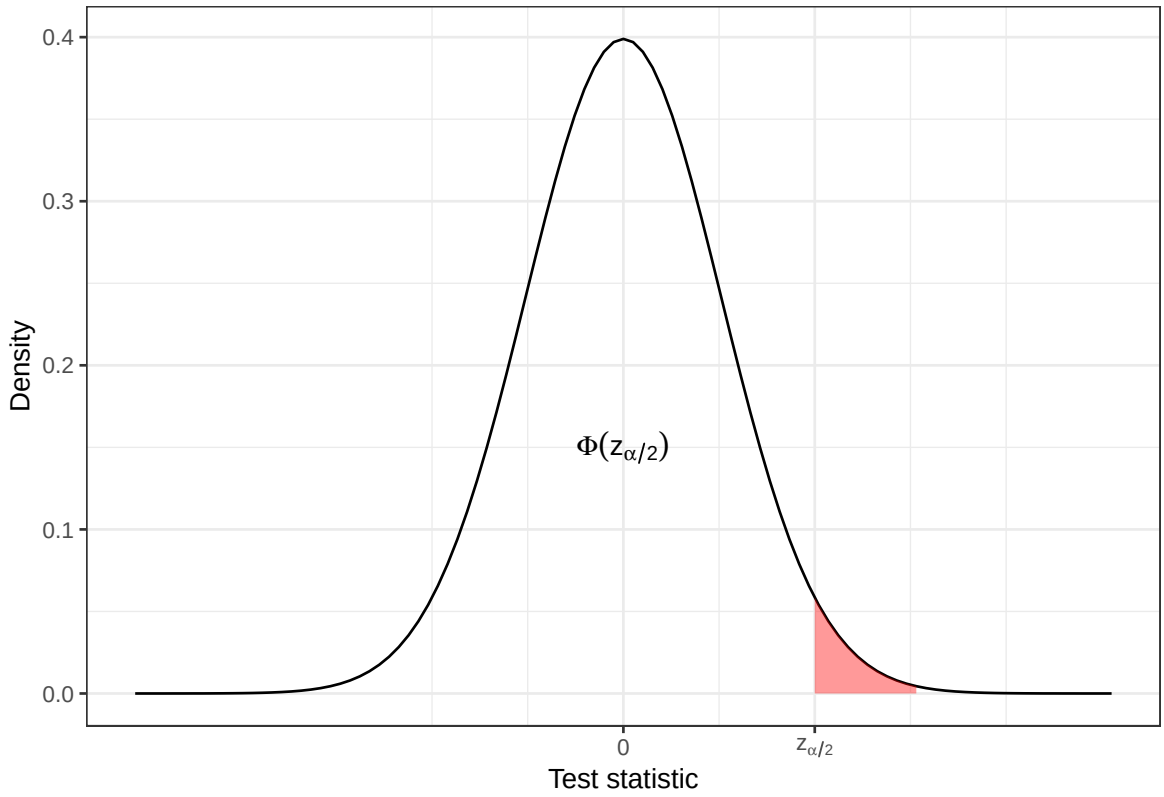
So, if we have run a trial, and have obtained n values of X from group C and m values of X from group T, we can compute D. If D lies outside the interval $\left[-z_{\alpha/2}, z_{\alpha/2}\right]$ then we reject H_0.

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

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

Brief aside on notation

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



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

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

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

Lecture 3

Recap:

- We constructed a measure $D = \frac{\bar{x}_T - \bar{x}_C}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$ that we can use to test $H_0 : \tau = 0$, since [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.

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 $\alpha,\ \beta \ll 0.5$. It is also common to use percentages: a significance level of $\alpha = 0.05$ can also be referred to as "the 95% level", and $\beta = 0.2$ is the same as a "power of 80%". When using percentages, we talk in terms of the amount of time we expect the test to come to the correct conclusion.

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

Under H_1, we have (approximately)

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

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

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

Figure 2.3 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.

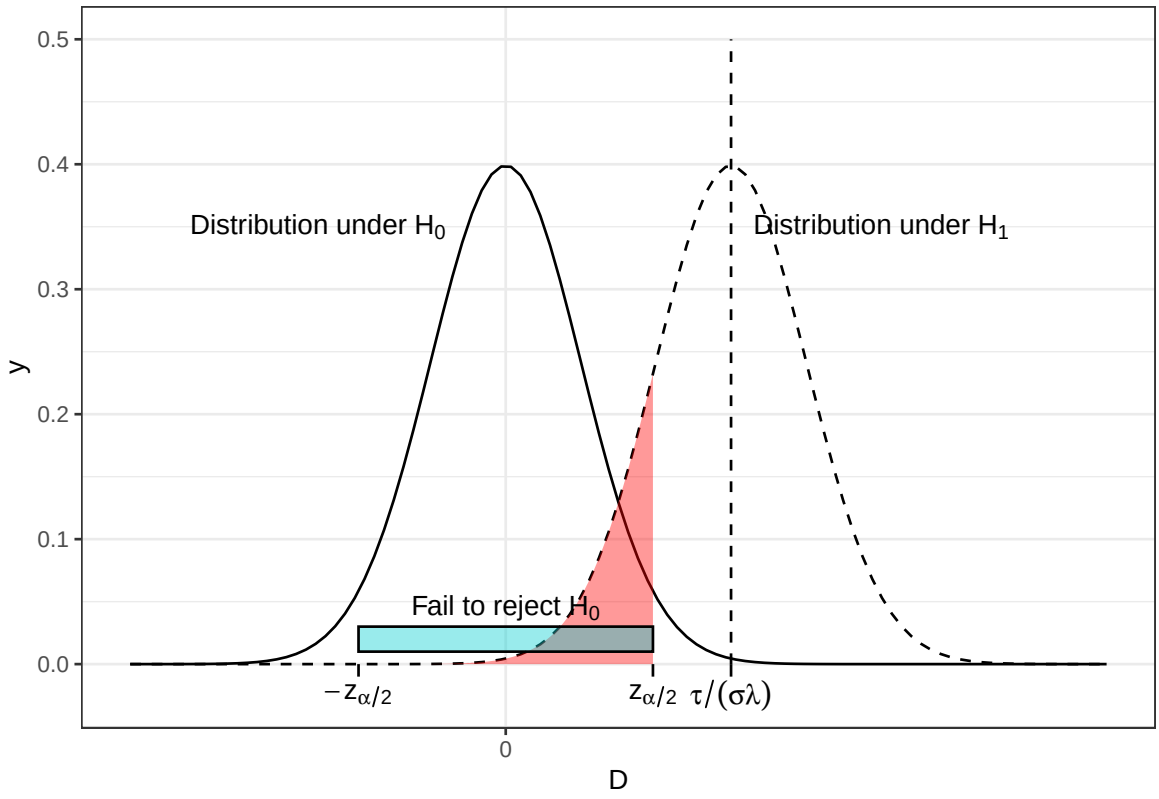


Figure 2.3: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

However, if H_1 is true, and $\tau \neq 0$, there is a non-zero probability of observing a value of D that would lead us to fail to reject H_0. This is shown by the area shaded in red, and it has area 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 2.4, 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).

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

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

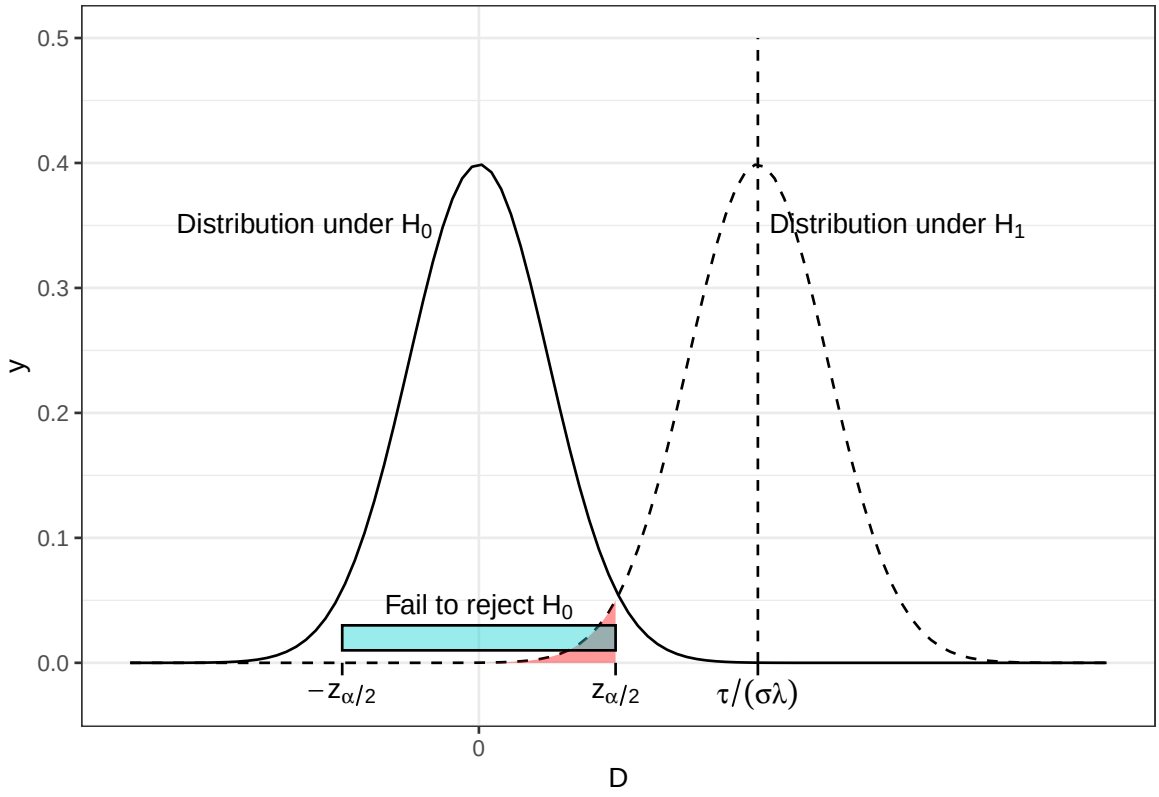


Figure 2.4: The distribution of D under both H_0 and H_1 for some arbitrary values of effect size, population variance, n and m, with the region in which we fail to reject H_0 shown by the turquoise bar and the red shading.

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 2.5 shows the power function $\Psi(\tau)$ for τ in units of σ (or you could think of this as for $\sigma = 1$), for three different pairs of values of n and m (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.

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.

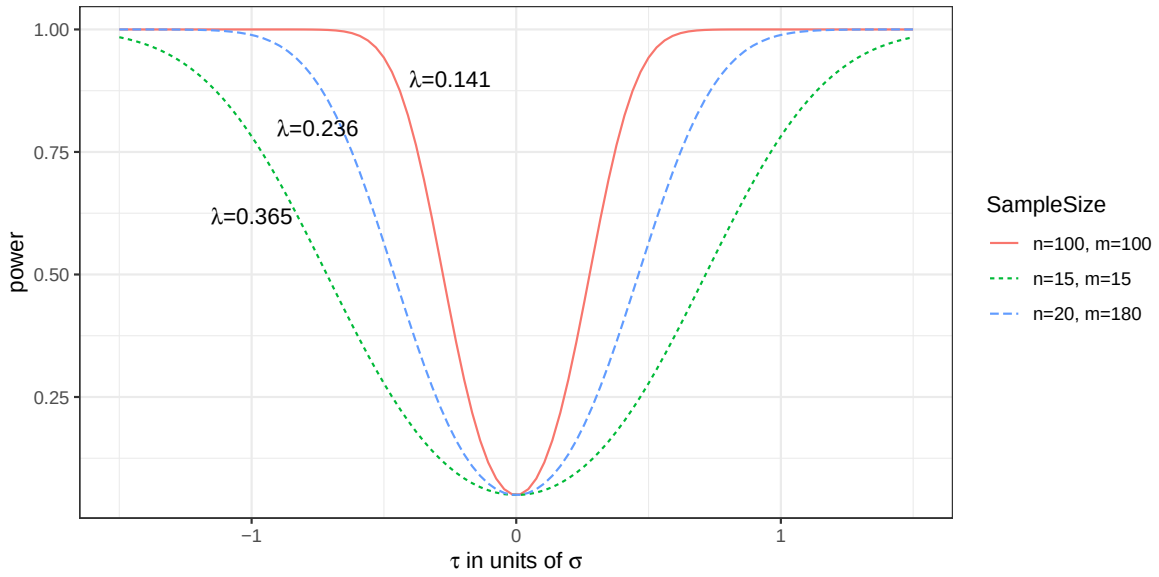


Figure 2.5: Power curves for various values of n and m, with effect size in units of standard deviation, given a type I error rate of 0.05.

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. 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 $\alpha,\ \beta$).

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

Therefore, Equation (2.1) becomes

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

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

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

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

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

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

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

and Equation (**??**) rearranges to

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

Example 2.1. (from Zhong 2009) A trial is being planned to test whether there is a difference in the efficacy of ACEII antagonist (a new drug) and ACE inhibitor (the standard drug) for the treatment of primary hypertension (high blood pressure). The primary outcome variable is change in sitting diastolic blood pressure (SDBP, mmHg) compared to a baseline measurement taken at the start of the trial. The trial should have a significance level of $\alpha = 0.05$ and a power of $1 - \beta = 0.8$, with the same number of participants in each group. The minimum clinically important difference is $\tau_M = 3$ mmHg and the pooled standard deviation is $s = 8$ mmHg. Therefore, using equation (2.3) the sample size should be at least

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

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

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