Lecture 18: Decision Errors and Power of a Test

- Decision errors
- Power calculation
- Statistical significance v.s. practical significance

Introduction

– Today we will build on what we learned about hypothesis testing in the last lecture. We will learn about where α, the error rate, really comes from.

– This would lead us to type I and type II errors in hypothesis testing, and power calculations, determining a proper sample size.

– Lastly, we will discuss statistical significance with practical significance.

Duality between HT and CI

Hypothesis testing is much like setting a confidence interval. A two-sided test of $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$ for a given α is often equivalent to whether or not a two-sided $\big(1 - \alpha\big)100\%$ confidence interval contains θ_0 (and similarly for one-sided tests and one-sided intervals).

Decision Errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Comparison

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	*True Negatives*	*False Positive*
	H_A true	*False Negative*	*True Positives*

One never "accepts" or "proves" the null or alternative hypotheses; one simply rejects or fails to reject the null hypothesis.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty
- Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"

– William Blackstone

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α – increasing α increases the Type 1 error rate.

Choosing a significance level

- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.
- We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

Notation

Let α be the probability of type I error and β the probability of type II error, then, any two of the following three quantities determines the third:

- n, the sample size in the test;
- α, the probability of Type I error; and
- β, which is the probability of Type II error.

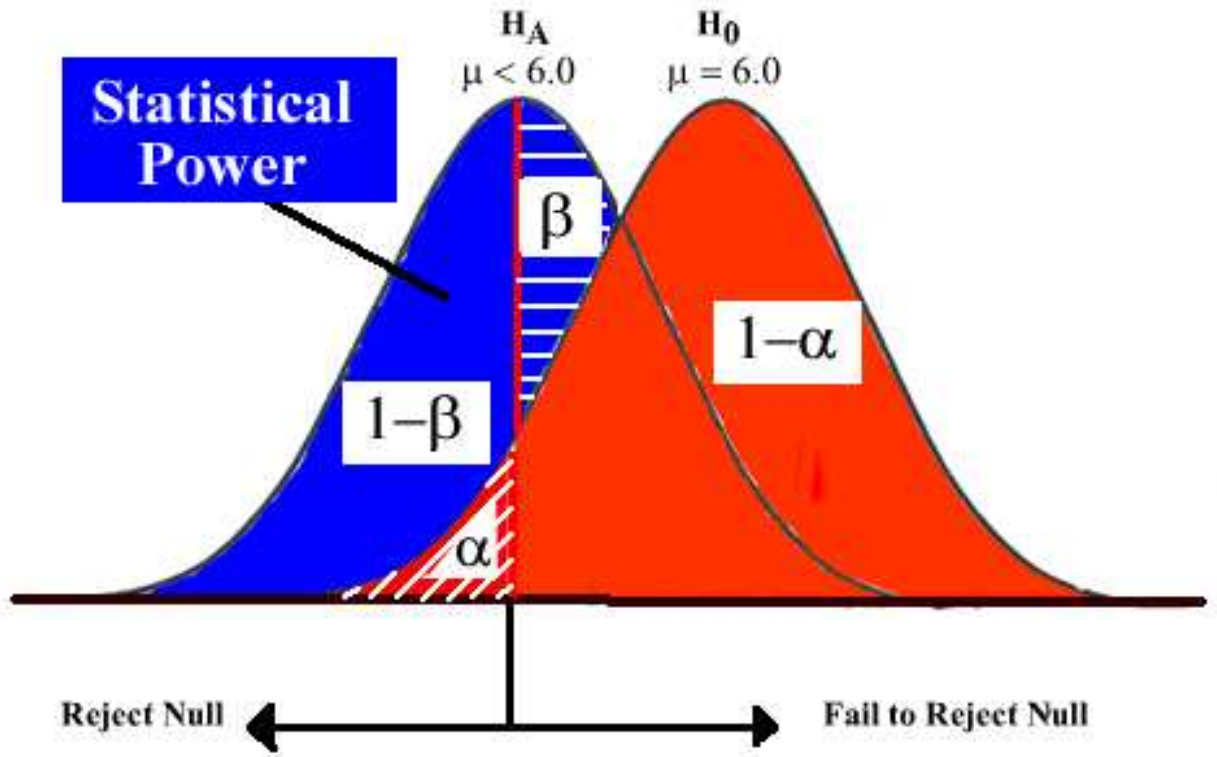
Typically, circumstances force you to pick α and n.

The **power** of a test is the $1-\beta$, which is the probability that your test correctly rejects the null hypothesis when the null hypothesis is false (the second quadrant of the previous picture).

In practice, one picks α at the outset, and then obtains the largest sample size n that one can afford, in order to maximize the power of the test.

Illustration

This figure assumes a one-sided test of $H_0: \mu \geq 6$ versus $H_A: \mu < 6$ with σ known and some level α (say 0.05).



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true		

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1-\beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

Type 2 error rate

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

- The answer is not obvious.
- If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- Clearly, β depends on the *effect size* δ

Example - Blood Pressure (BP), hypotheses

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication, and half of the subjects are given the new drug (treatment) and the other half continue to take their current medication through generic-looking pills to ensure blinding (control). What are the hypotheses for a two-sided hypothesis test in this context?

$$H_0 : \mu_{treatment} - \mu_{control} = 0$$
$$H_A : \mu_{treatment} - \mu_{control} \neq 0$$

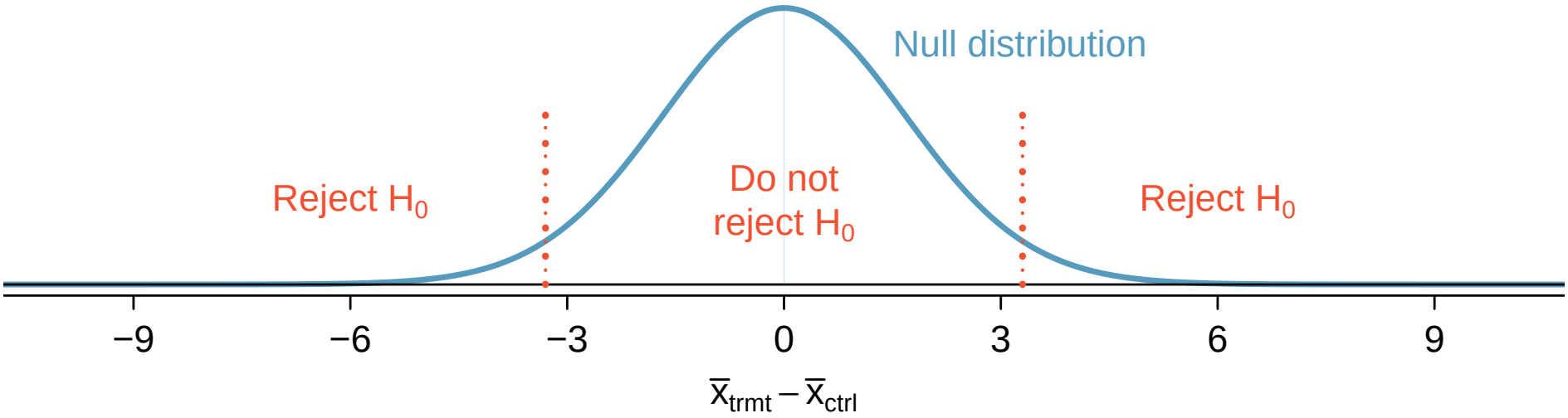
Example - BP, standard error

Suppose researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric. If we had 100 patients per group, what would be the approximate standard error for difference in sample means of the treatment and control groups?

$$SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

Example - BP, minimum effect size required to reject H_0

For what values of the difference between the observed averages of blood pressure in treatment and control groups (effect size) would we reject the null hypothesis at the 5% significance level?



The difference should be at least

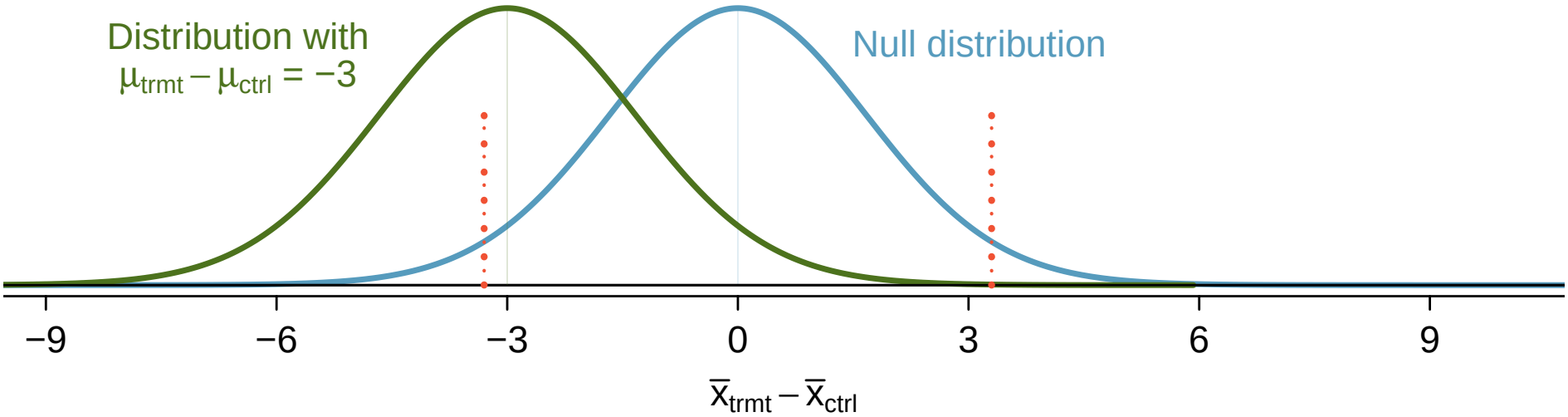
$$1.96 * 1.70 = 3.332$$

or at most

$$-1.96 * 1.70 = -3.332.$$

Example - BP, power

Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. What is the power of the test that can detect this effect?



$$Z = \frac{-3.332 - (-3)}{1.70} = -0.20$$

$$P(Z < -0.20) = 0.4207$$

Steps for Calculating Power

1. Pick a meaningful effect size δ and a significance level α
2. Find the rejection region for the point estimate where you would reject H_0 at the chosen α level
3. Calculate the probability of observing a value from preceding step if the sample was drawn from a population where $\mu = \mu_0 + \delta$

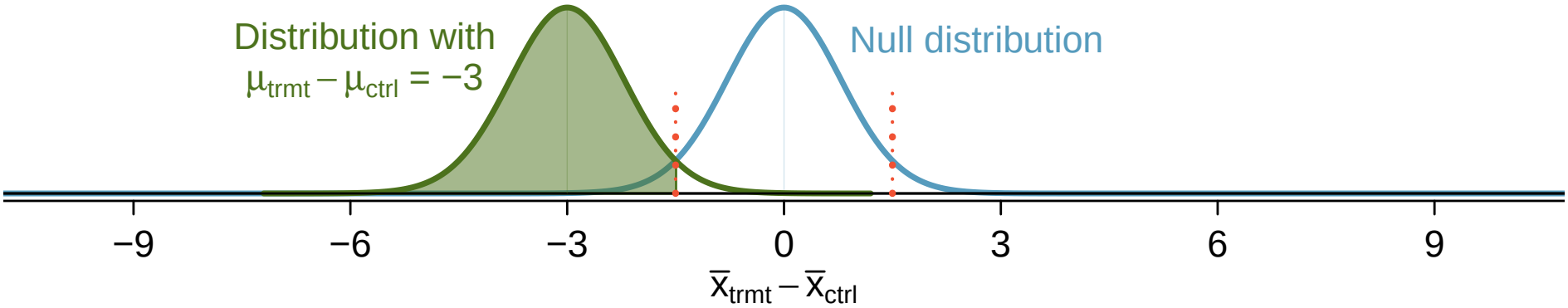
Determining a proper sample size

If we have a sample size of 100 in each group, we can only detect an effect size of 3 mmHg with a probability of about 0.42. Suppose the researchers moved forward and only used 100 patients per group, and the data did not support the alternative hypothesis, i.e. the researchers did not reject H_0. This is a very bad situation to be in for a few reasons:

- In the back of the researchers' minds, they'd all be wondering, maybe there is a real and meaningful difference, but we weren't able to detect it with such a small sample.
- The company probably invested hundreds of millions of dollars in developing the new drug, so now they are left with great uncertainty about its potential since the experiment didn't have a great shot at detecting effects that could still be important.
- Patients were subjected to the drug, and we can't even say with much certainty that the drug doesn't help (or harm) patients.
- Another clinical trial may need to be run to get a more conclusive answer as to whether the drug does hold any practical value, and conducting a second clinical trial may take years and many millions of dollars.

Example - BP, required sample size for 80% power

What sample size will lead to a power of 80% for this test?



$> \text{qnorm}(p = 0.8)$
$[1] \quad 0.8416212$

$$SE = \frac{3}{0.84 + 1.96} = \frac{3}{2.8} = 1.07142$$

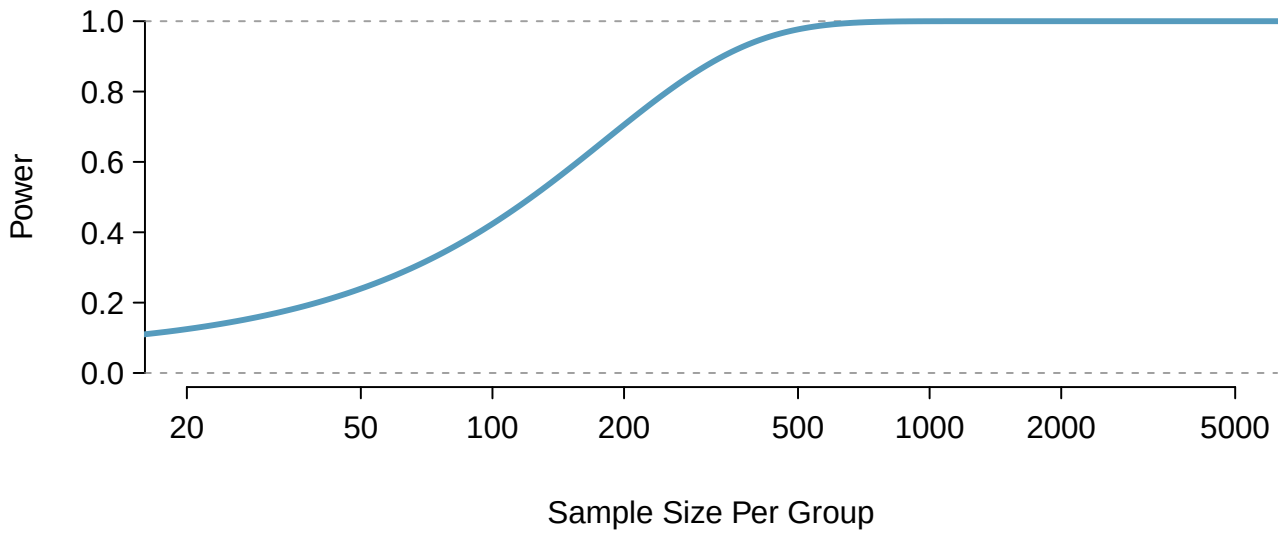
$$1.07142 = \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = 250.88 \rightarrow n \geq 251$$

Power v.s. Sample Size

- Calculate required sample size for a desired level of power
- Calculate power for a range of sample sizes, then choose the sample size that yields the target power (usually 80% or 90%)

From 20 patients to 5,000 patients when $\alpha = 0.05$ and the true difference is -3.



Having more than 250 to 350 observations doesn't provide much additional value in detecting an effect when $\alpha = 0.05$.

Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.

You must pick your null and alternative hypotheses before seeing the data. Also, you must pick two of α, β and n before looking at the data. Doing otherwise is cheating.

Example

In many cases one can calculate the power of a test. This is important when deciding how large a sample you need—if your test is underpowered, you can improve it by investing in a larger sample size.

Example 1: You have a sample of size 100 from a normal population with known standard deviation 4. You want to test $H_o : \mu \geq 6$ versus $H_A : \mu < 6$ with a Type I error rate of 0.05.

Suppose the population actually has a true mean of 5. What will be the power of your test?

Example (Cont'd)

$$
\begin{aligned}
\textbf{power} = 1-\beta &= 1-\mathbb{P}[\textbf{fail to reject null when null is false}] \\
&= 1-\mathbb{P}[ts > -1.645] = \mathbb{P}[\frac{\bar{x}-6}{4/\sqrt{100}} < -1.645] \\
&= \mathbb{P}[\frac{\bar{x}-6}{0.4} < -1.645] \\
&= \mathbb{P}[\frac{\bar{x}-5+5-6}{0.4} < -1.645] \\
&= \mathbb{P}[\frac{\bar{x}-5}{0.4} + \frac{5-6}{0.4} < -1.645] \\
&= \mathbb{P}[\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < -1.645 - \frac{5-6}{0.4}] \\
&= \mathbb{P}[Z < 0.855] \quad (\textbf{CLT}) \\
&= 0.8023
\end{aligned}
$$

So the test has about an 80% chance of correctly rejecting the null hypothesis.

Example (Cont'd)

Often one picks α and β, and then those determine n. For example, to obtain NIH funding to run a clinical trial, you might decide to use $\alpha = 0.01$ and you want power $1 - \beta = 0.9$ for **detecting an increase in average lifespan of 1 year.**

You know that the average U.S. life expectancy is 77.6 years, with a standard deviation of about 14.5 years.

You want to show that your drug extends lives. The hypotheses are:

$$H_o : \mu_D \leq 77.6 \qquad \textbf{vs.} \qquad H_A : \mu_D > 77.6.$$

The test statistic is

$$ts = \frac{\overline{x} - 77.6}{14.5/\sqrt{n}}.$$

and the critical value is $z_{0.99} = 2.33$.

Example (Cont'd)

$$
\begin{aligned}
\textbf{power} = 0.9 &= 1 - \mathbb{P}[\textbf{fail to reject null when null is false}] \\
&= 1 - \mathbb{P}[ts < 2.33] = \mathbb{P}[\frac{\bar{x} - 77.6}{14.5/\sqrt{n}} > 2.33] \\
&= \mathbb{P}[\frac{\bar{x} - 78.6 + 78.6 - 77.6}{14.5/\sqrt{n}} > 2.33] \\
&= \mathbb{P}[\frac{\bar{x} - 78.6}{14.5/\sqrt{n}} + \frac{1}{14.5/\sqrt{n}} > 2.33] \\
&= \mathbb{P}[Z > 2.33 - \frac{1}{14.5/\sqrt{n}}].
\end{aligned}
$$

From the z-table, $0.9 = \mathbb{P}[Z > -1.28]$, so

$$-1.28 = 2.33 - \frac{1}{14.5/\sqrt{n}}.$$

Solving shows that the least **integer** that achieves this power is $n = 2740$.

With large samples, one can get a statistically significant result that is of no practical importance.

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu > 49.5$.

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.0062$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

As n increases - $SE \downarrow$, $Z \uparrow$, p-value $\downarrow$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10,000$?

(a) $n = 100$

(b) $n = 10,000$

Suppose $\bar{x} = 50$, $s = 2$, $H_0 : \mu = 49.5$, and $H_A : \mu > 49.5$.

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad \text{p-value} = 0.0062$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad \text{p-value} \approx 0$$

As n increases - $SE \downarrow$, $Z \uparrow$, p-value $\downarrow$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$			
$n = 5000$			

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$		
$n = 5000$			

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$		
$n = 5000$	$p - value = 0.04$		

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	
$n = 5000$	$p - value = 0.04$		

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

Test the hypothesis $H_0 : \mu = 10$ vs. $H_A : \mu > 10$ for the following 6 samples. Assume $\sigma = 2$.

$\overline{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.

Statistical vs. practical significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (*effect size*), even when the difference is not practically significant.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).
- The role of a statistician is not just in the analysis of data, but also in planning and design of a study.

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." – R.A. Fisher

Recap

Today we learned about decision errors, how to calculate the power of a test and determine a proper sample size.

Suggested reading:

- OpenIntro3: Sec. 4.3, 5.4