Power Calculation for 2-Sample Tests

Dr. Ab Mosca (they/them)

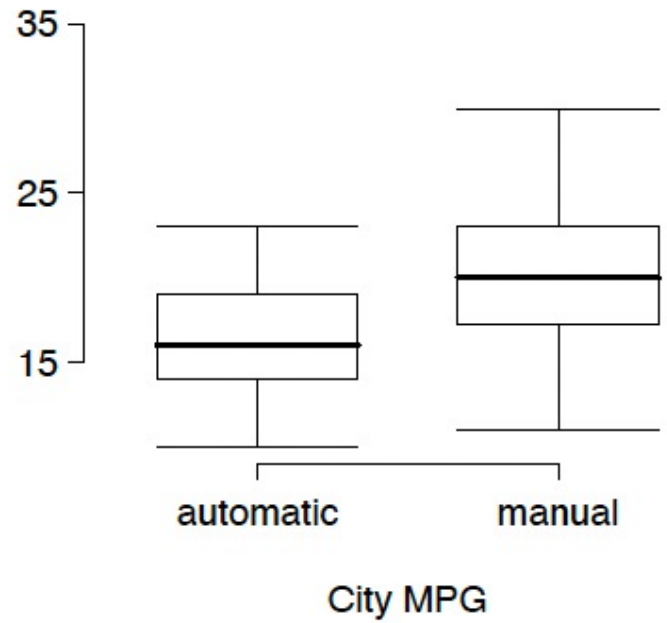
Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

Plan for Today

- Warm-up – Review 2 Sample Hypothesis Tests
- Power Calculations

7.28 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[22]

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



7.28 The hypotheses are as follows:

$H_0 : \mu_{A,c} - \mu_{M,c}$

$H_0 : \mu_{A,c} \neq \mu_{M,c}$

We are told to assume that conditions for inference are satisfied.
Then, the test statistic and the p-value can be calculated as follows:

$$T = \frac{(\bar{x}_{A,c} - \bar{x}_{M,c}) - (\mu_{A,c} - \mu_{M,c})}{\sqrt{\frac{s^2_{A,c}}{n_{A,c}} + \frac{s^2_{M,c}}{n_{M,c}}}} = \frac{(16.12 - 19.85) - 0}{\sqrt{\frac{3.58^2}{26} + \frac{4.51^2}{26}}} = \frac{-3.73}{1.13} = -3.3$$

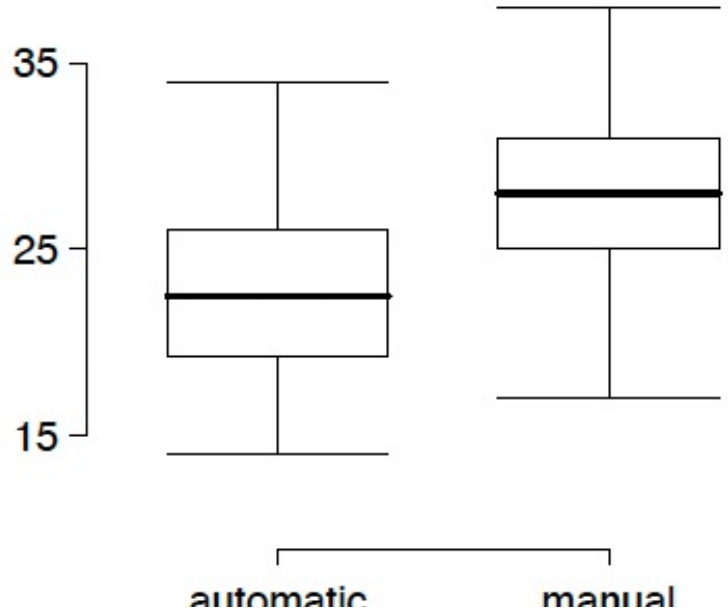
$$df = min(n_{M,c} - 1, n_{A,c} - 1) = min(26 - 1, 26 - 1) = 25$$

$$p - value = P(|T_{25}| > 3.3) < 0.01$$

Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a difference in the average city mileage between cars with automatic and manual transmissions.

7.30 Fuel efficiency of manual and automatic cars, Part II. The table provides summary statistics on highway fuel economy of the same 52 cars from Exercise 7.28. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.[23]

	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26



7.32 True / False: comparing means. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

(a) When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal model for the difference in means since $n_2 \geq 30$.

(b) As the degrees of freedom increases, the t-distribution approaches normality.

(c) We use a pooled standard error for calculating the standard error of the difference between means when sample sizes of groups are equal to each other.

7.30

$$df = min(n_1 - 1, n_2 - 1) = min(26 - 1, 26 - 1) = 25 \ \rightarrow \ t^\star_{25} = 2.49$$

$$\begin{aligned}(\bar{x}_{A,hwy} - \bar{x}_{M,hwy}) \pm t^\star_{df}\sqrt{\frac{s^2_{A,hwy}}{n_{A,hwy}} + \frac{s^2_{M,hwy}}{n_{M,hwy}}} &= (22.92 - 27.88) \pm 2.49 * \sqrt{\frac{5.29^2}{26} + \frac{5.01^2}{26}} \\ &= -4.96 \pm 2.49 \times 1.43 \\ &= -4.96 \pm 3.56 \\ &= (-8.52, -1.4)\end{aligned}$$

We are 98% confident that on the highway cars with manual transmissions get on average 1.4 to 8.52 MPG more than cars with automatic transmissions.

7.32

(a) False, in order to be able to use a Z test both sample sizes need to be above 30.

(b) True.

(c) False, we use the pooled standard deviation when the variability in groups is constant.

Power

- The power of a statistical test is the probably that we detect an effect if there is a real effect

Decision Errors

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

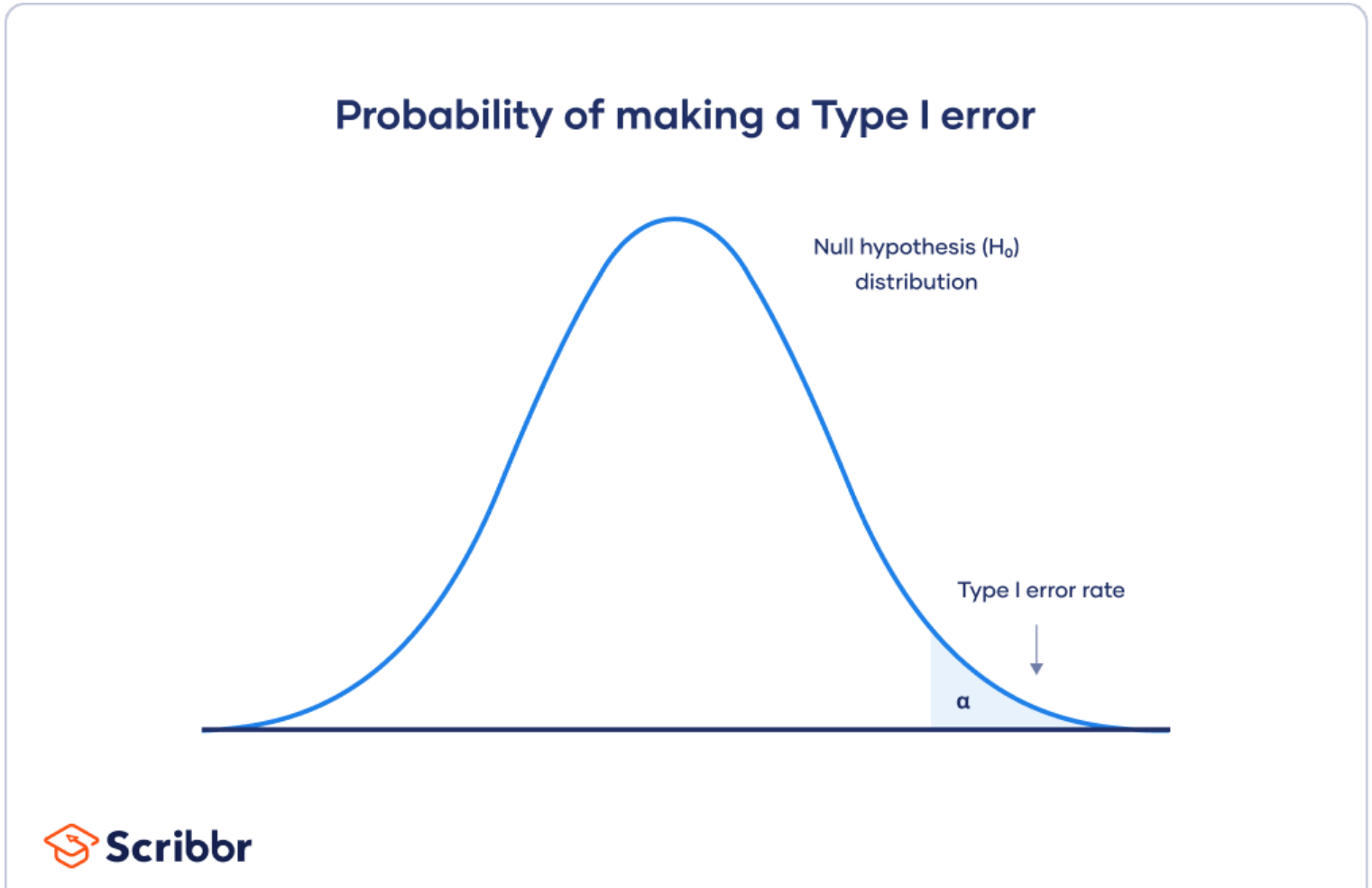
Decision Errors

		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)

Decision Errors

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



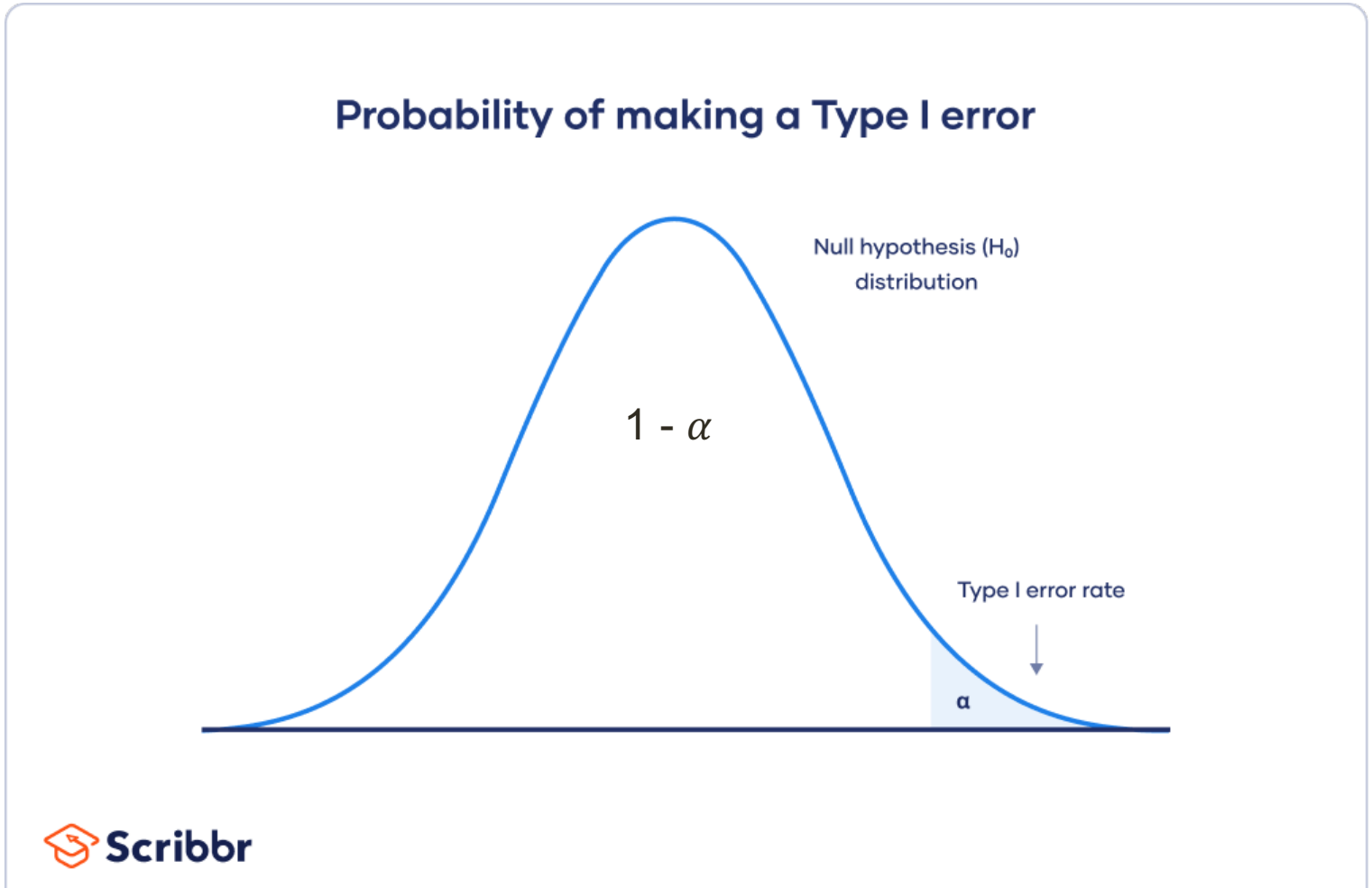
Decision Errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*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)
- The probability of making a correct decision if H_0 is true is 1 - α

Decision Errors

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



Decision Errors

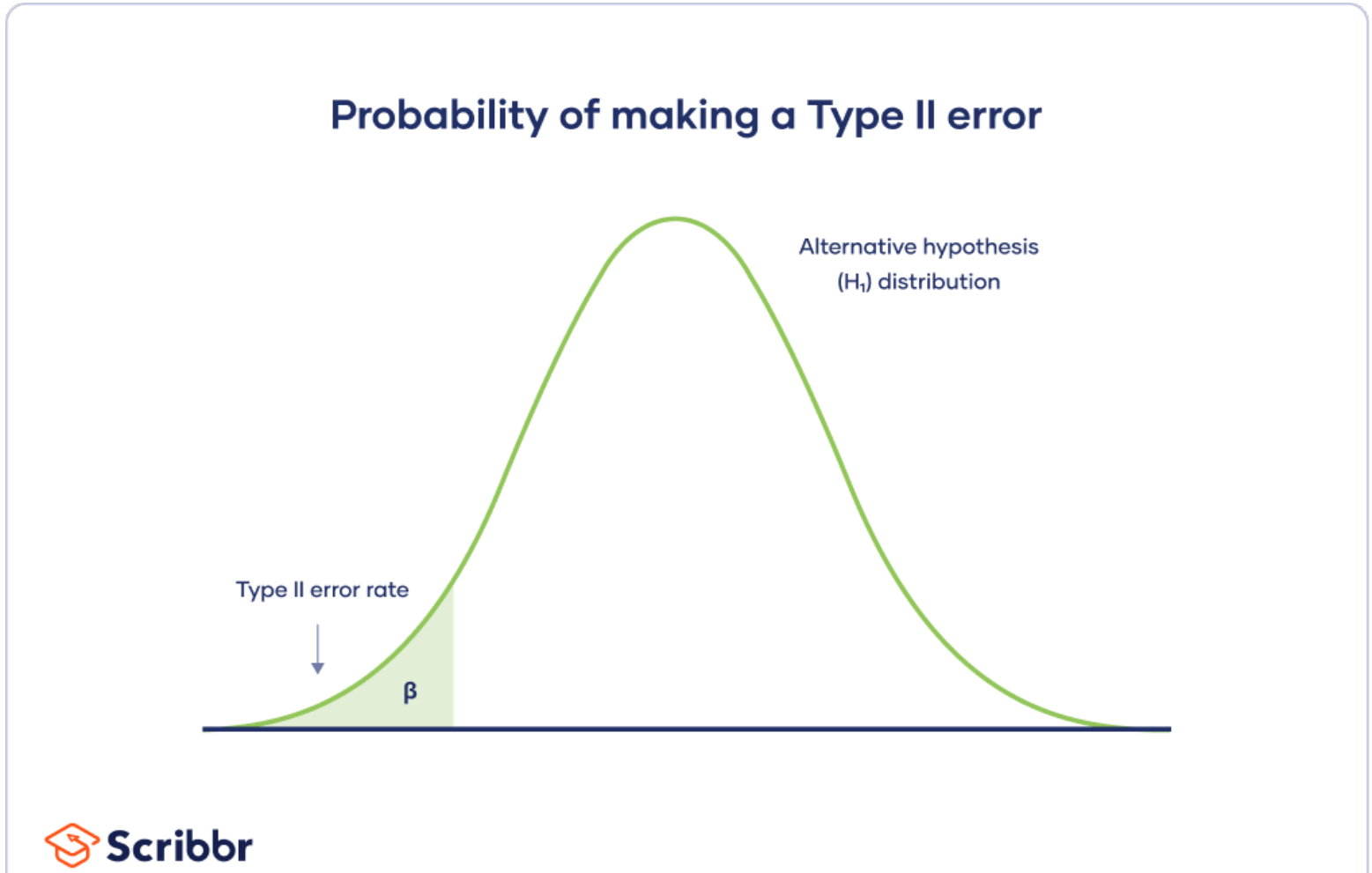
Truth	Decision: fail to reject H_0	Decision: reject H_0
H_0 true	$1 - \alpha$	*Type 1 Error, α*
H_A true	*Type 2 Error, β*	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- The probability of making a correct decision if H_0 is true is 1 - α

- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β

Decision Errors

		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*	



Decision Errors

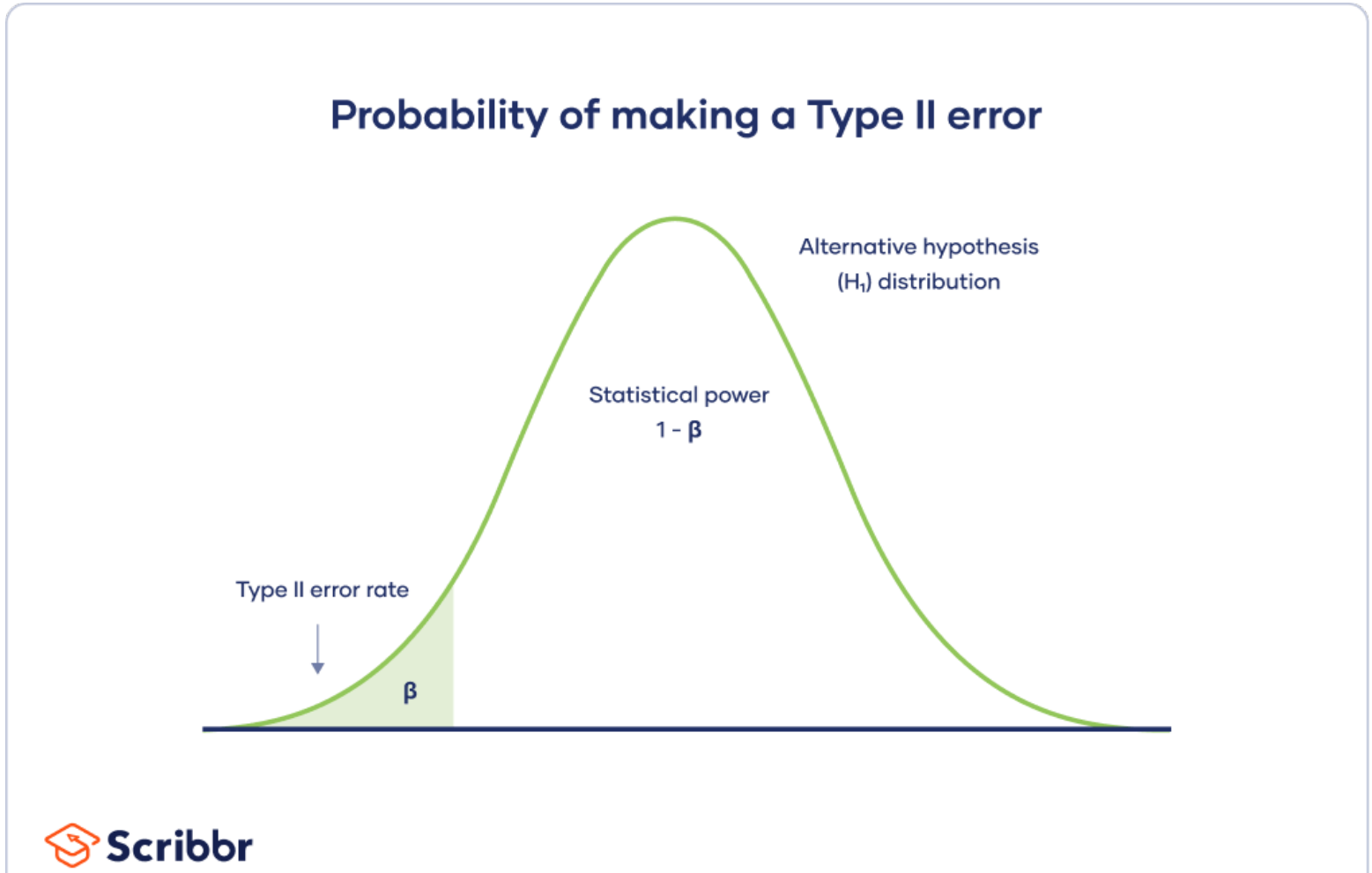
Truth	Decision: fail to reject H_0	Decision: reject H_0
H_0 true	$1 - \alpha$	*Type 1 Error,* α
H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- The probability of making a correct decision if H_0 is true is 1 - α

- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β
- The probability of making a correct decision if H_0 is false is 1 − β; we call this the power of a test

Decision Errors

		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,* $1 - \beta$



Decision Errors

		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, $1 - \beta$

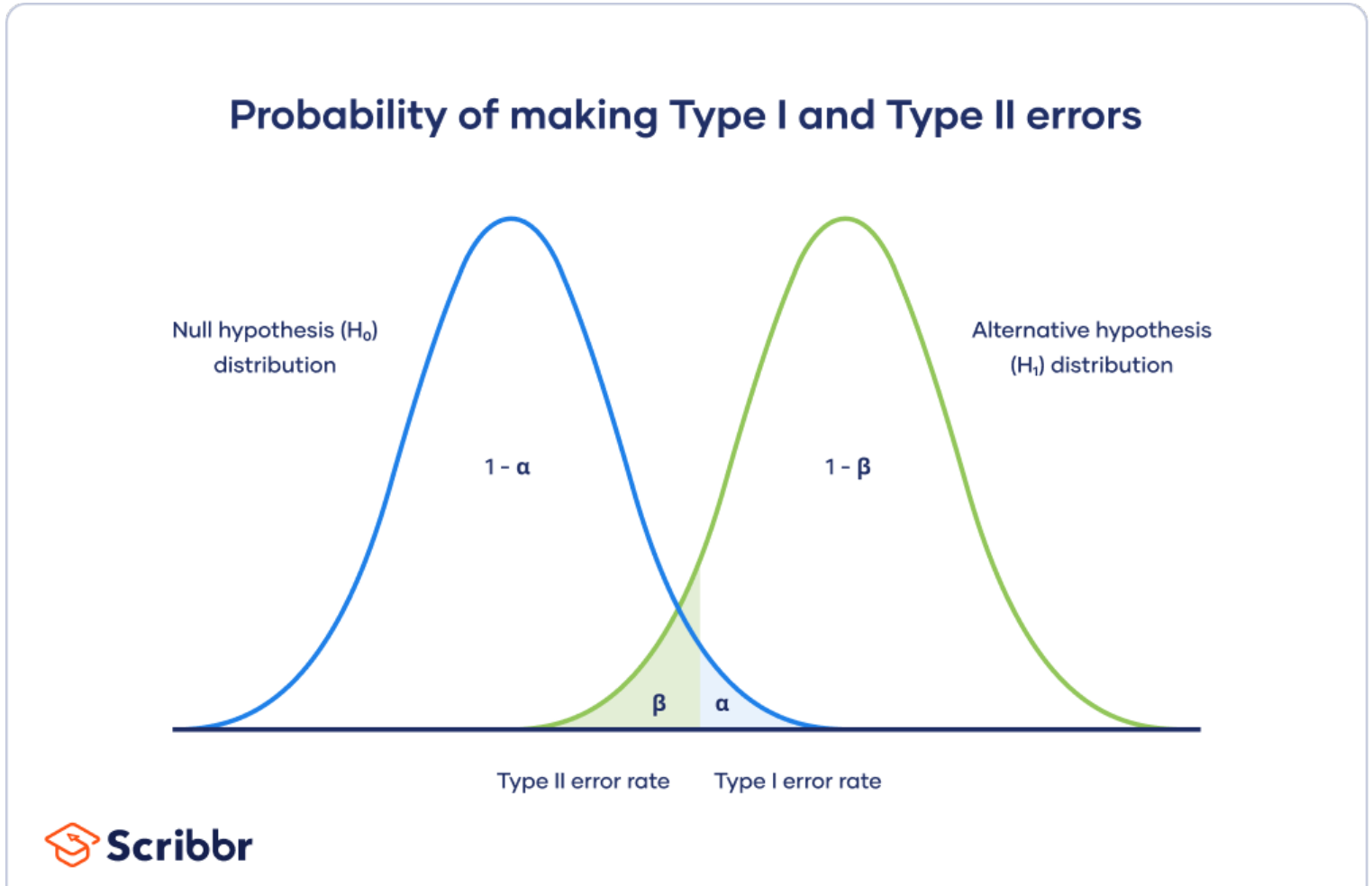
- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- The probability of making a correct decision if H_0 is true is 1 - α

- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β
- The probability of making a correct decision if H_0 is false is 1 − β; we call this the power of a test

- There is a trade off between Type I and Type II errors

Decision Errors

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

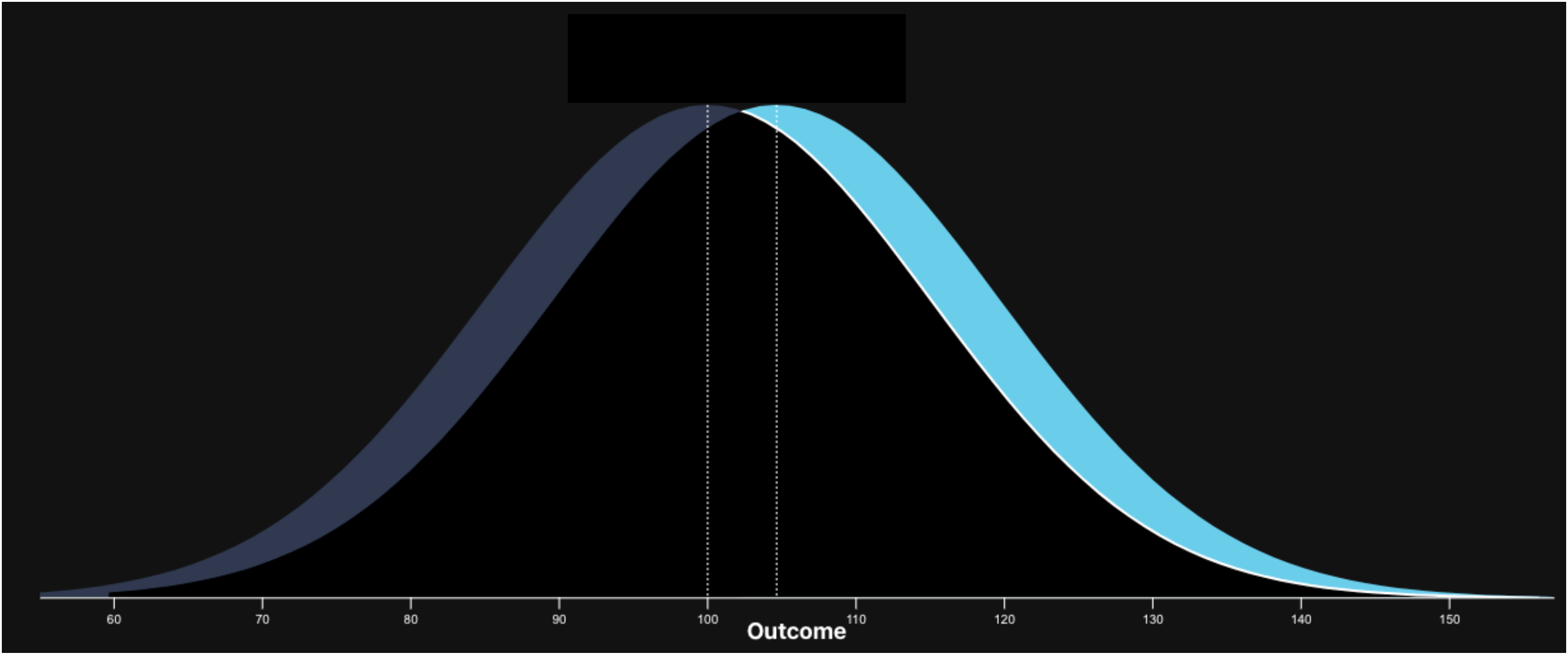


Type 2 error rate

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)

Type 2 error rate

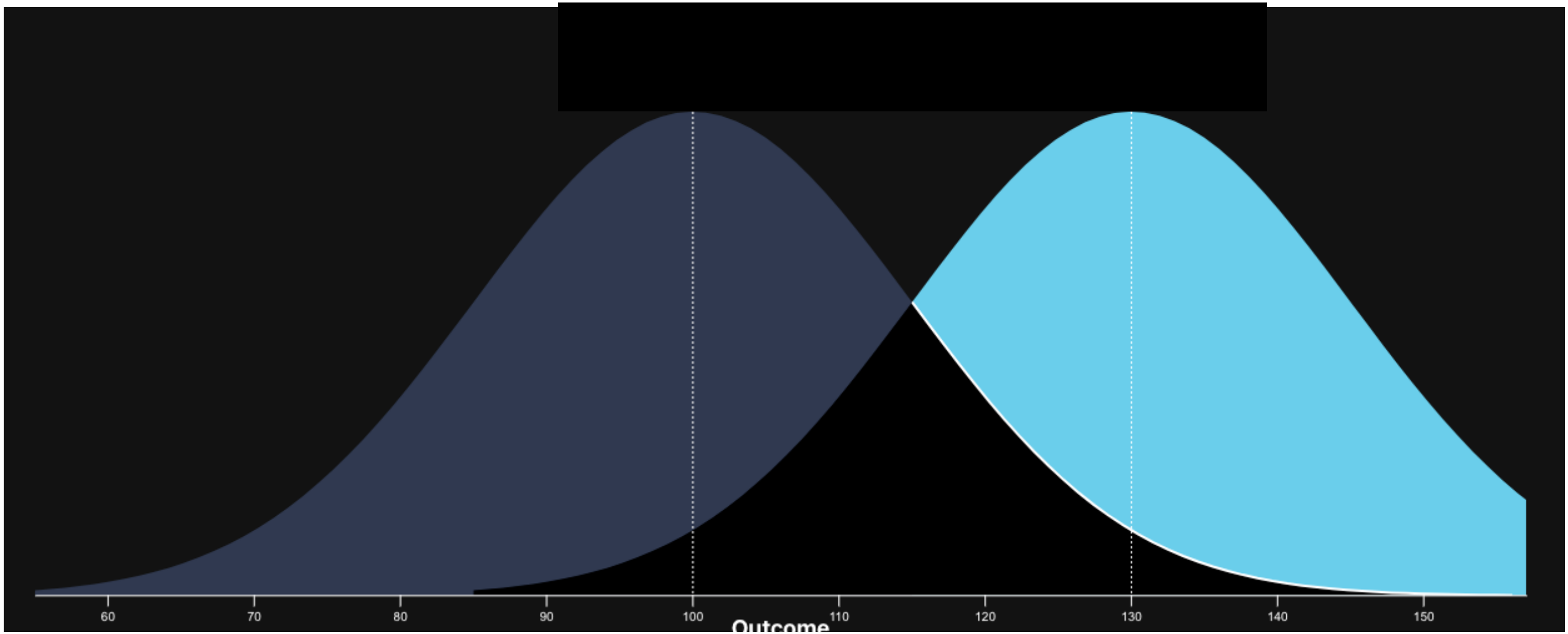


Type 2 error rate

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

Type 2 error rate



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
- The difference between the true population value and the null hypothesis value is called *effect size* (δ)
- β depends on the effect size

Effect Size

- Effect size is the practical significance of the difference between the actual population mean and the null hypothesis mean

Effect Size

- Effect size is the practical significance of the difference between the actual population mean and the null hypothesis mean
- It differs from statistical significance because it is not influenced by sample size (a larger sample increases the likelihood of a statistically significant difference because variance decreases with increased sample size)
- https://medium.com/@dtuk81/sample-sizes-impact-on-effect-size-and-power-fbd5084c7c47

Effect Size

- Effect size is the practical significance of the difference between the actual population mean and the null hypothesis mean
- It differs from statistical significance because it is not influenced by sample size (a larger sample increases the likelihood of a statistically significant difference because variance decreases with increased sample size)
- https://medium.com/@dtuk81/sample-sizes-impact-on-effect-size-and-power-fbd5084c7c47

Example - Blood Pressure (BP)

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 explanatory and response variables for this experiment? What are the levels of the explanatory variable?

Example - Blood Pressure (BP)

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 explanatory and response variables for this experiment? What are the levels of the explanatory variable?

		Response: Blood Pressure
Explanatory: Drug	New Drug	
	Current Medication	

Example - Blood Pressure (BP)

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?

Example - Blood Pressure (BP)

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 in our study, what would be the approximate standard error for difference in sample means of the treatment and control groups?

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

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.

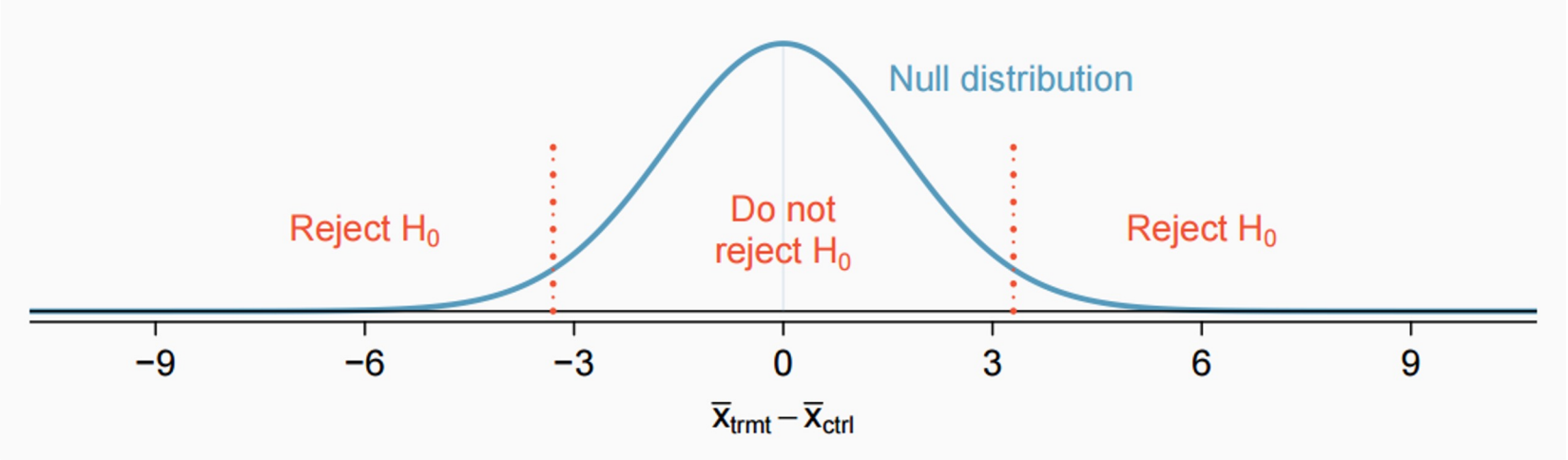
If we had 100 patients per group in our study, what would be the approximate standard error for difference in sample means of the treatment and control groups?

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} \qquad SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

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?

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

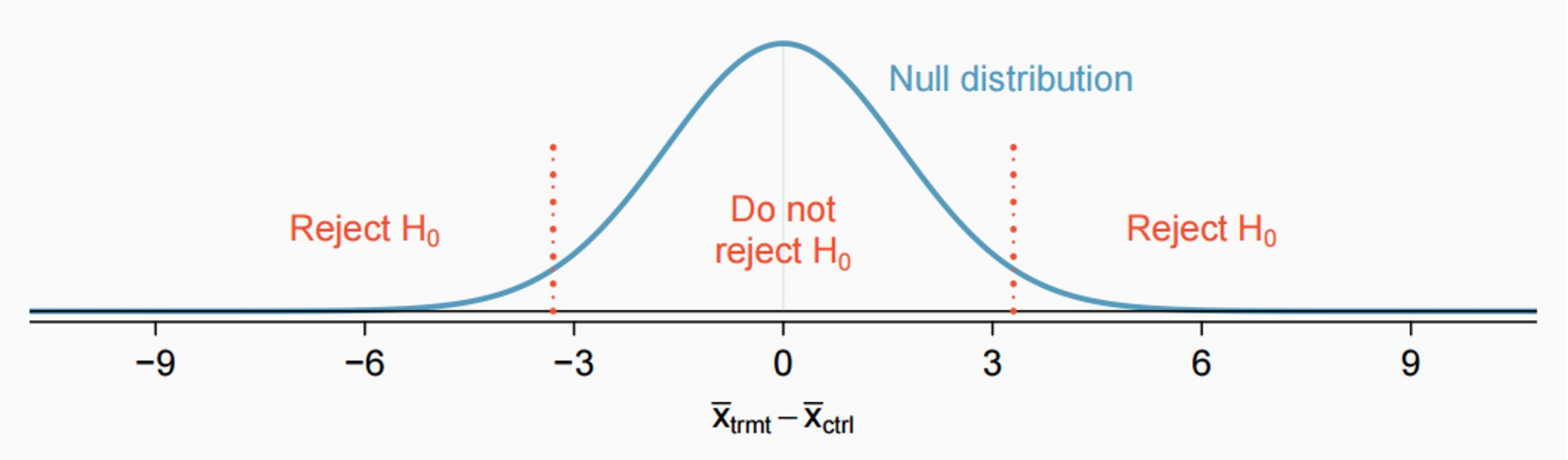


Find the cutoffs for the observed values that would cause us to reject the null hypothesis.

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?

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



The difference should be at least

1.96 * 1.70 = 3.332

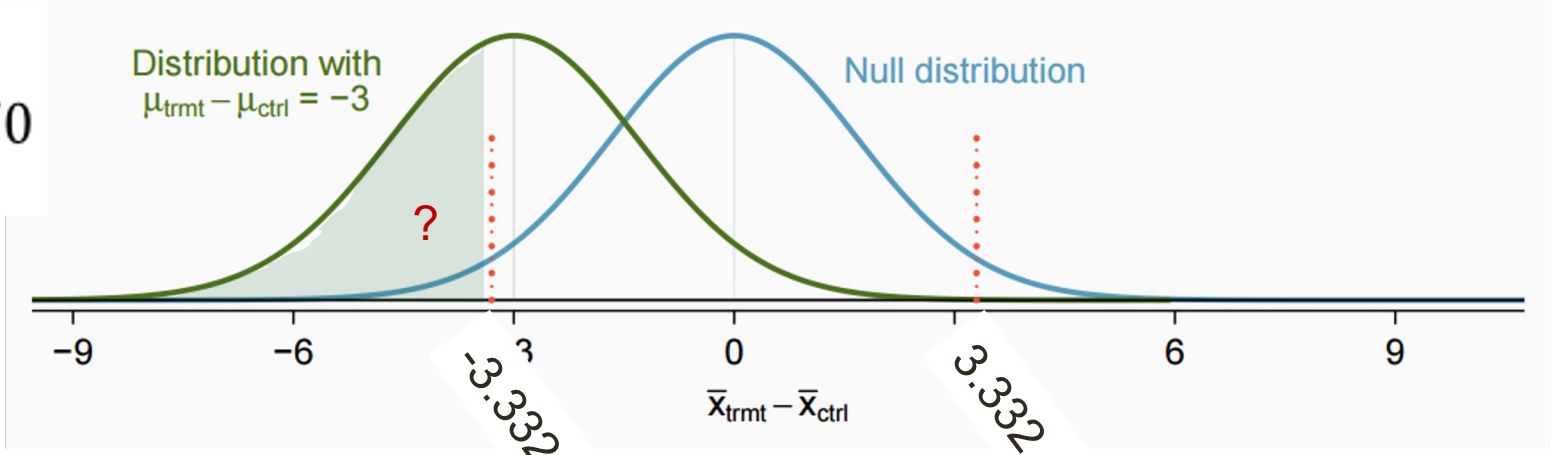
or at most

-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?

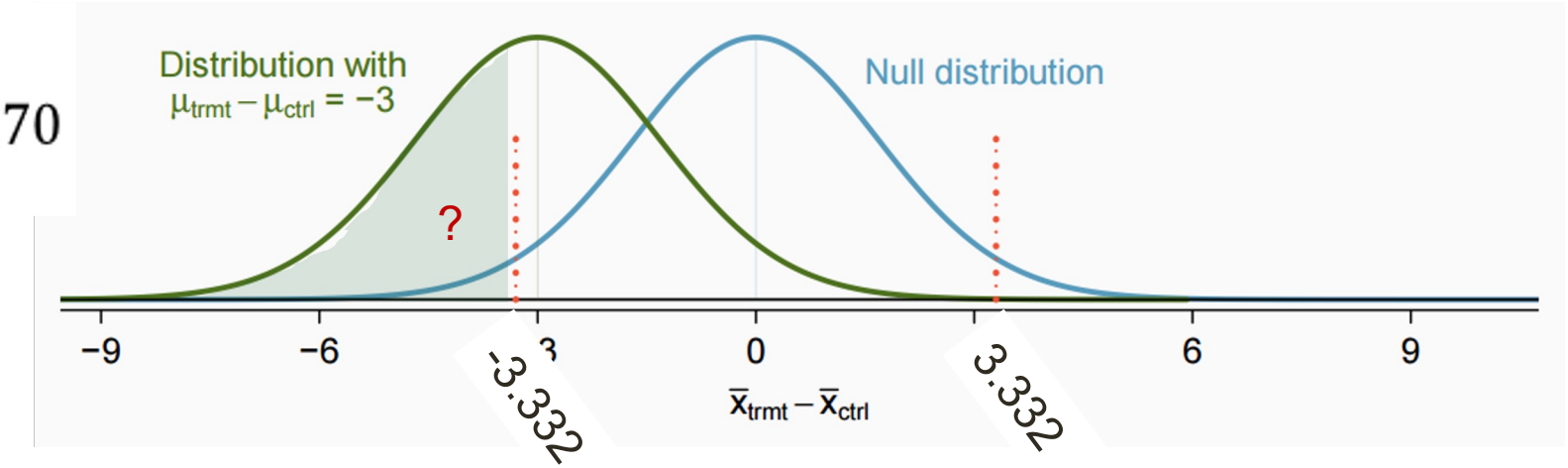
$$SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$



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?

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

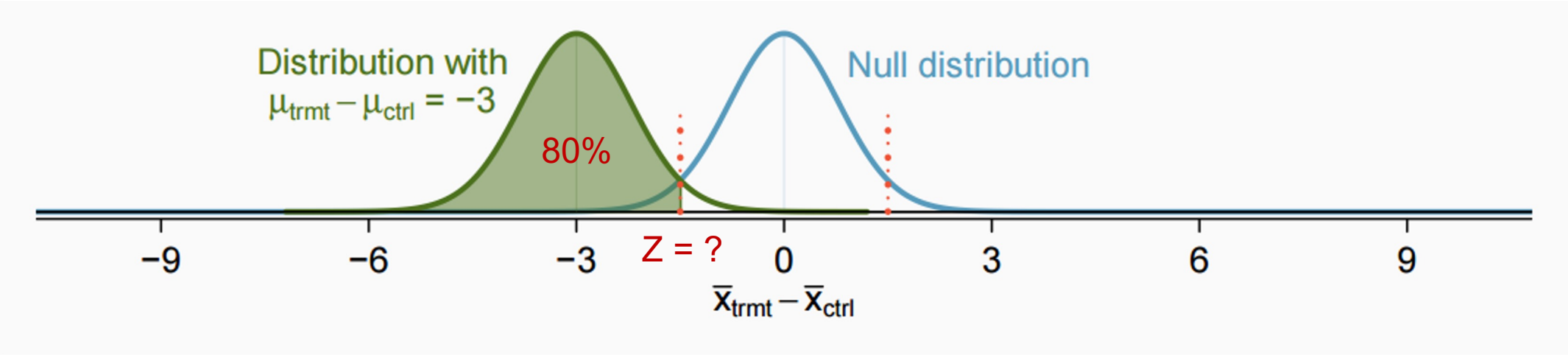


$$Z = \frac{-3.332 - (-3)}{1.70} = -0.20$$

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

Example - BP, required sample size for 80% power

What sample size will lead to a power of 80% for this test?
Use α = 0.05

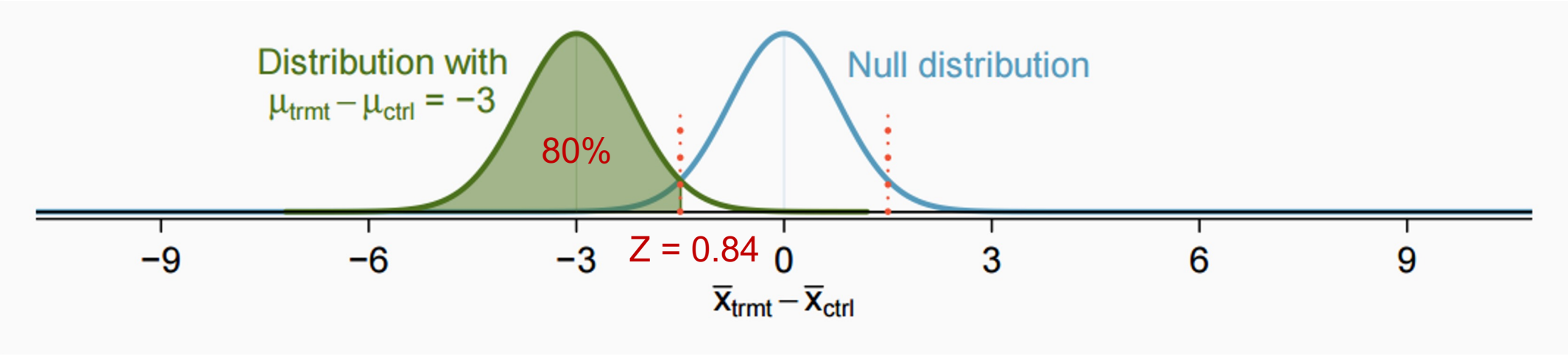


First, find Z

Example - BP, required sample size for 80% power

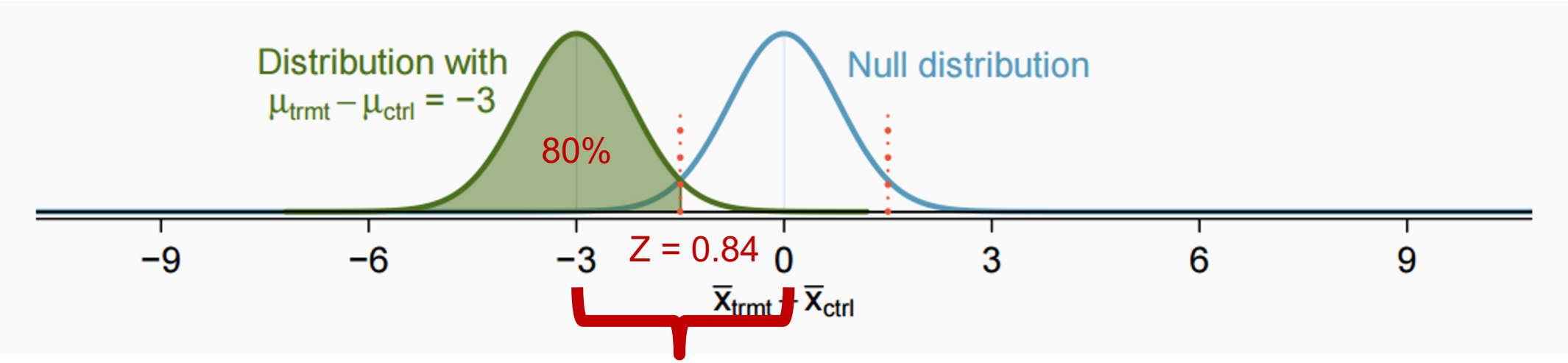
What sample size will lead to a power of 80% for this test?

Use α = 0.05



Example - BP, required sample size for 80% power

What sample size will lead to a power of 80% for this test?
Use α = 0.05

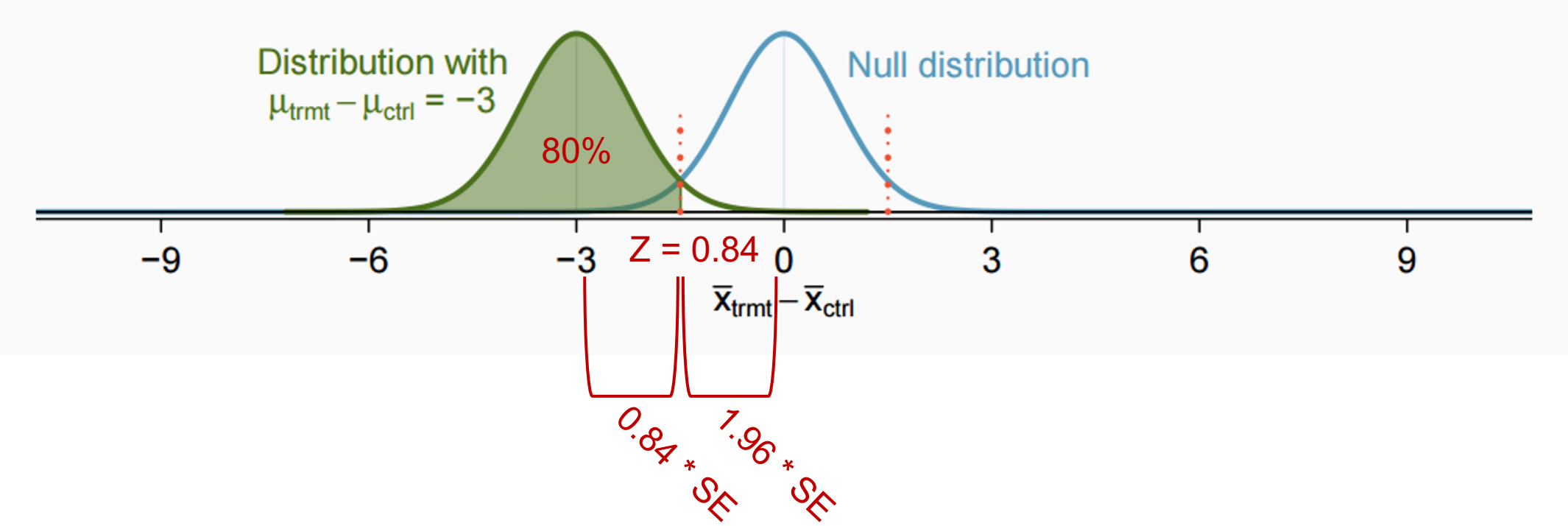


Next, find how many SE's are between the two means.

Example - BP, required sample size for 80% power

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

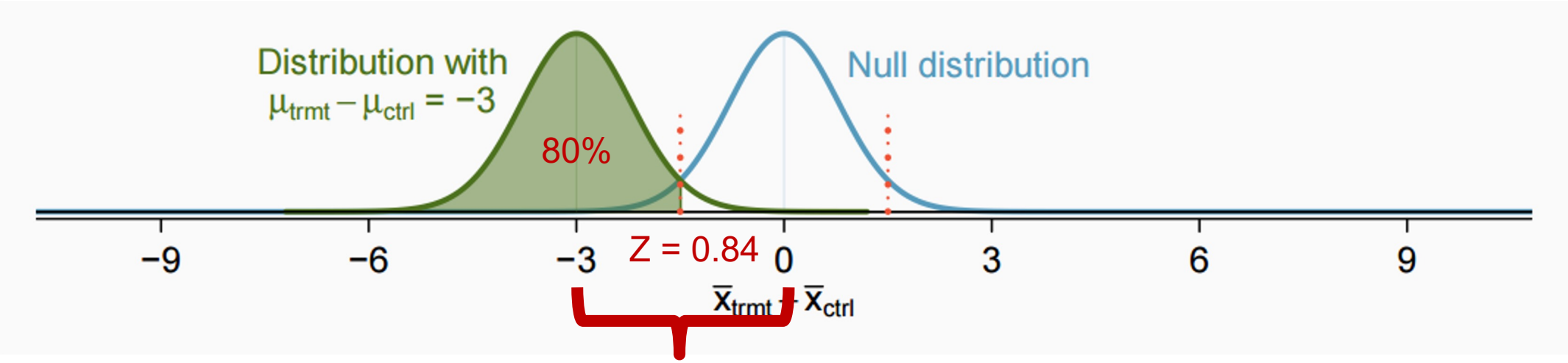
Use α = 0.05



distance = 0.84 * SE + 1.96 SE = 2.8 * SE

Example - BP, required sample size for 80% power

What sample size will lead to a power of 80% for this test?
Use α = 0.05

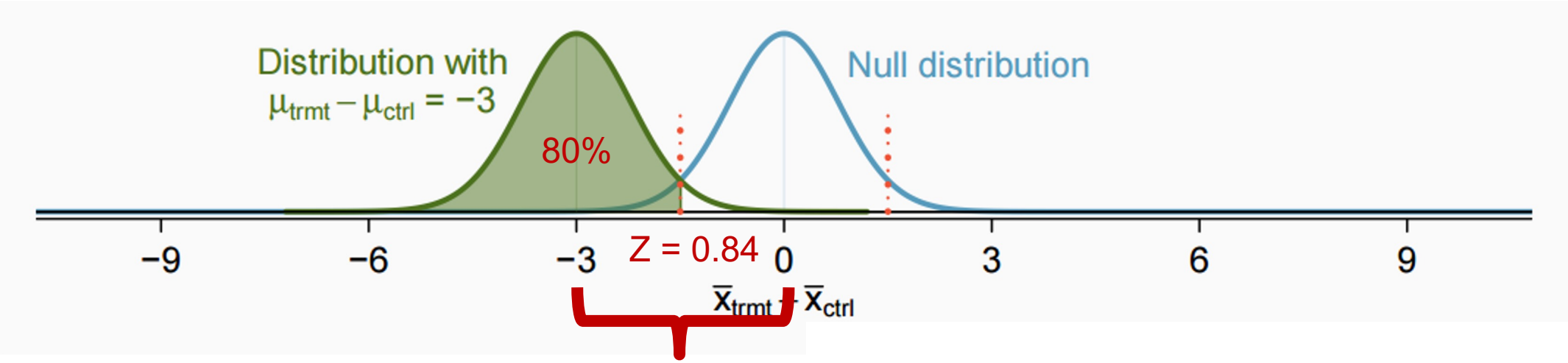


We were given that a difference is practically significant only if it is 3 or greater. Use this, the distance we just found, and the SE equation to find n.

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Example - BP, required sample size for 80% power

What sample size will lead to a power of 80% for this test?
Use α = 0.05



$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

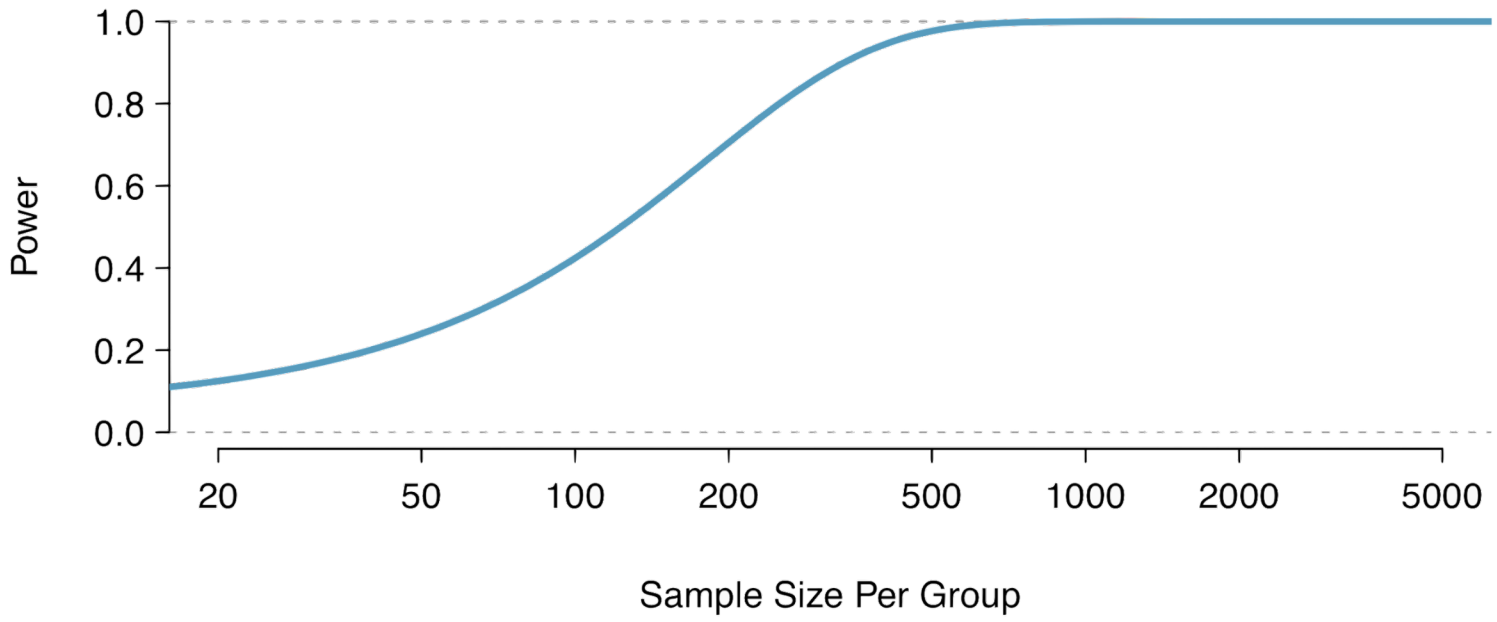
$$3 = 2.8 \times SE$$

$$3 = 2.8 \times \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = \frac{2.8^2}{3^2} \times (12^2 + 12^2) = 250.88$$

Recap

- 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%)



Achieving desired power

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

Achieving desired power

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

1. Increase the sample size

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

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)

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