Power and AB Testing

Data Science Immersive

Outline

- Type I and Type II error
- Power
- Effect Size
- The 4 elements we can control in an experiment
- Multiple Comparison Problem

Review : Type I and II error

$$H_0 : defendant = innocent$$

$$H_A : defendant \neq innocent$$

		Reality	
		H_0 is true	H_0 is false
Test Result	Fail to Reject H_0	Correct Decision (1-α) **1**	Type II Error (β) **2**
	Reject H_0	Type I Error (α) **3**	Correct Decision (1-β) **4**

True Positive :

- (Correctly Reject the Null Hypothesis)
- Defendant is guilty! **4**

False Positive :

- (Incorrectly Reject the Null Hypothesis)
- Defendant is accused of guilty even though innocent :((((**3**

True Negative :

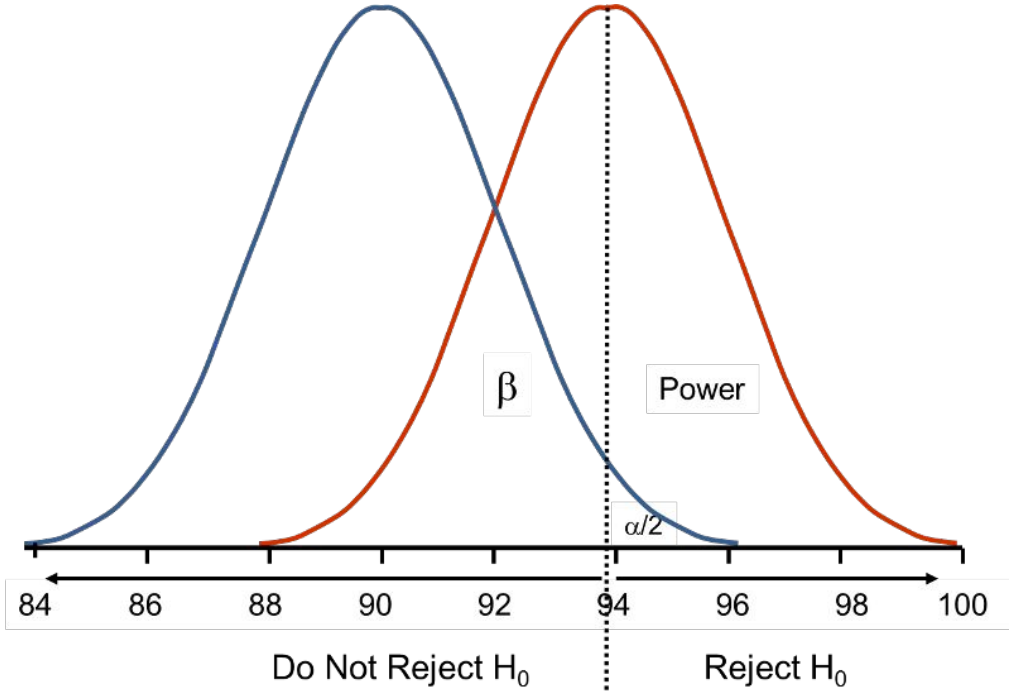
- (Correctly Fail to Reject the Null Hypothesis)
- Defendant is innocent and found innocent **1**

False Negative :

- (Incorrectly Fail to Reject the Null Hypothesis)
- Defendant is guilt but found innocent **2**

Review : Type I and II error

		Reality	
		H_0 is true	H_0 is false
Test Result	Fail to Reject H_0	Correct Decision (1-α)	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Decision (1-β)



Power of a Test

The power of a hypothesis test is the **probability** of making the **correct decision** if the **alternative hypothesis is true**. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_A is the hypothesis that is true.

Probability of having True Positives

For example, if a drug was actually effective, it would be the chance of actually determining that it was effective.

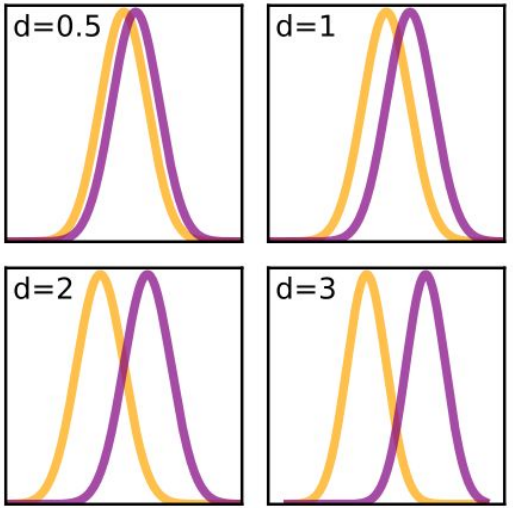
Practically, why is this important?

Review : Effect Size

- A quantitative measurement of the magnitude. Essentially how far away one distribution is from one another
- All else held constant, the higher the effect size, the higher the power.

Simple effect size : $\mu_0 - \mu_1$

Cohen's D: $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}}$



What can we control in an experiment?

When conducting an experiment, there are 4 components we can control:

- **Sample Size (n)**
- **Significance Level (alpha)**
- **Effect Size (Cohen's d)**
- **Power of a Test**

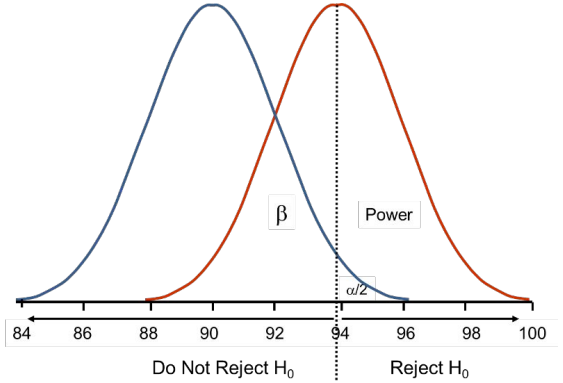
But there's a catch…..

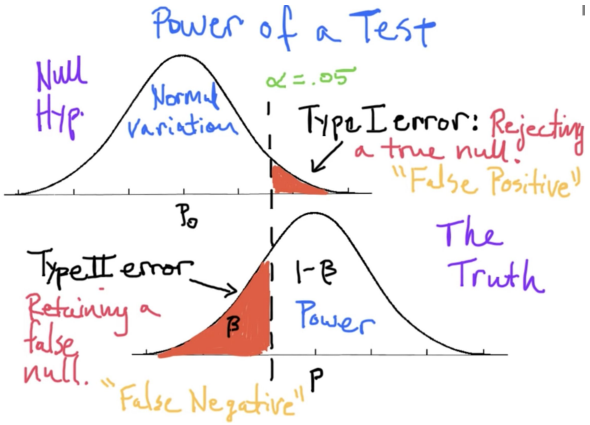
We can only control **3** out of the **4** at any given time

Type I and II error

		Reality	
		H_0 is true	H_0 is false
Test Result	Fail to Reject H_0	Correct Decision (1-α)	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Decision (1-β)

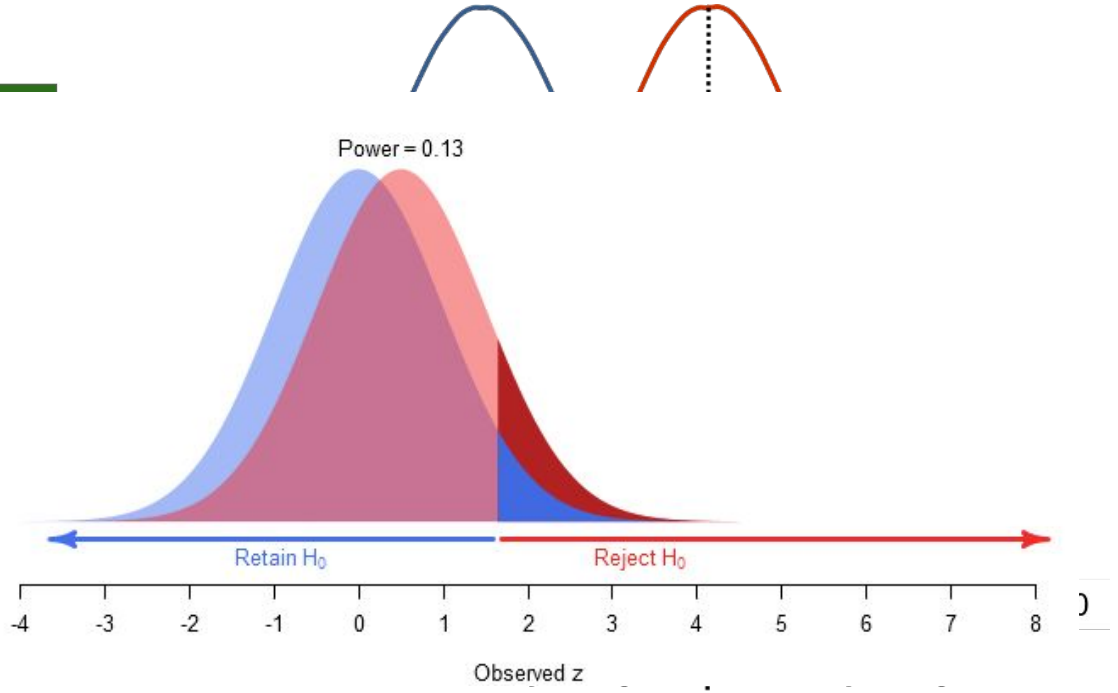
Take the next 4 minutes to determine explore [this interactive distribution] and answer the questions on the next





Type I and II error

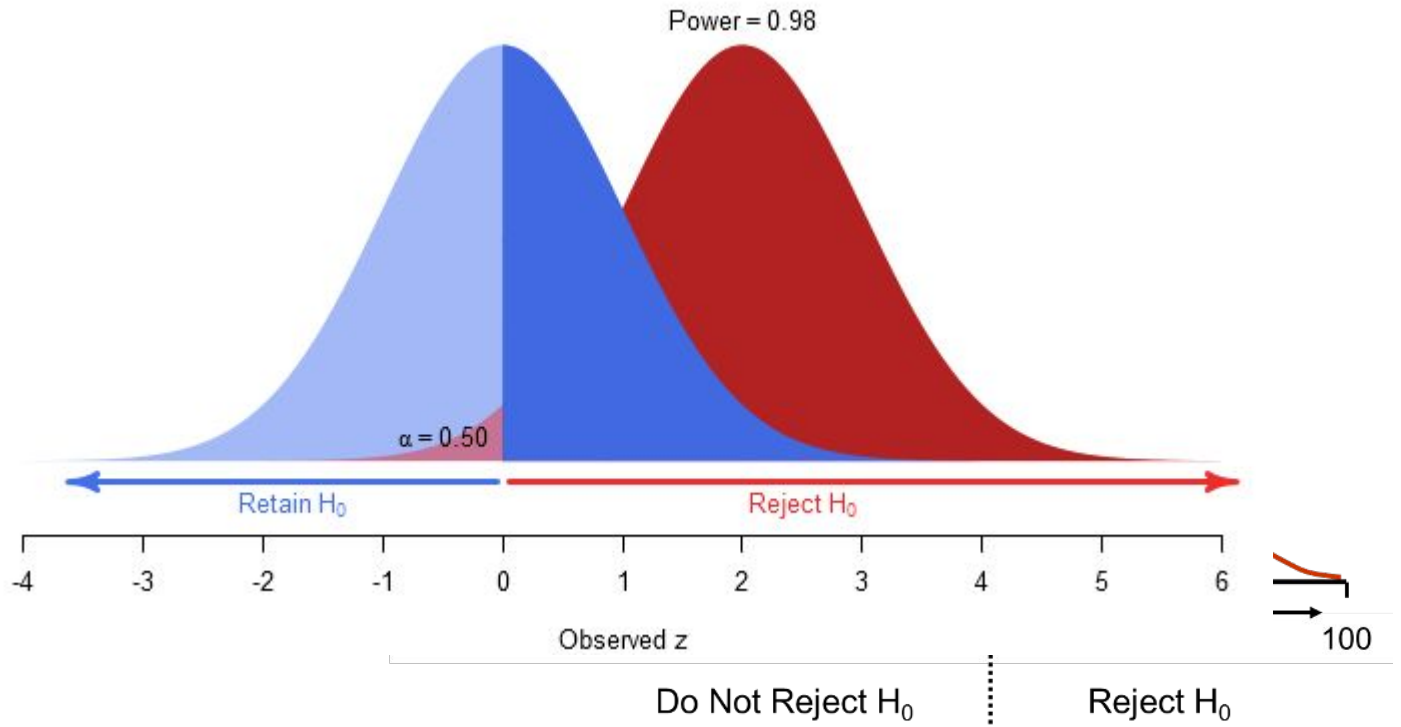
		Reality	
		H_0 is true	H_0 is false
Test Result	Fail to Reject H_0	Correct Decision (1-α)	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Decisi (1-β)



What happens to power when we increase the **effect size**?

Type I and II error

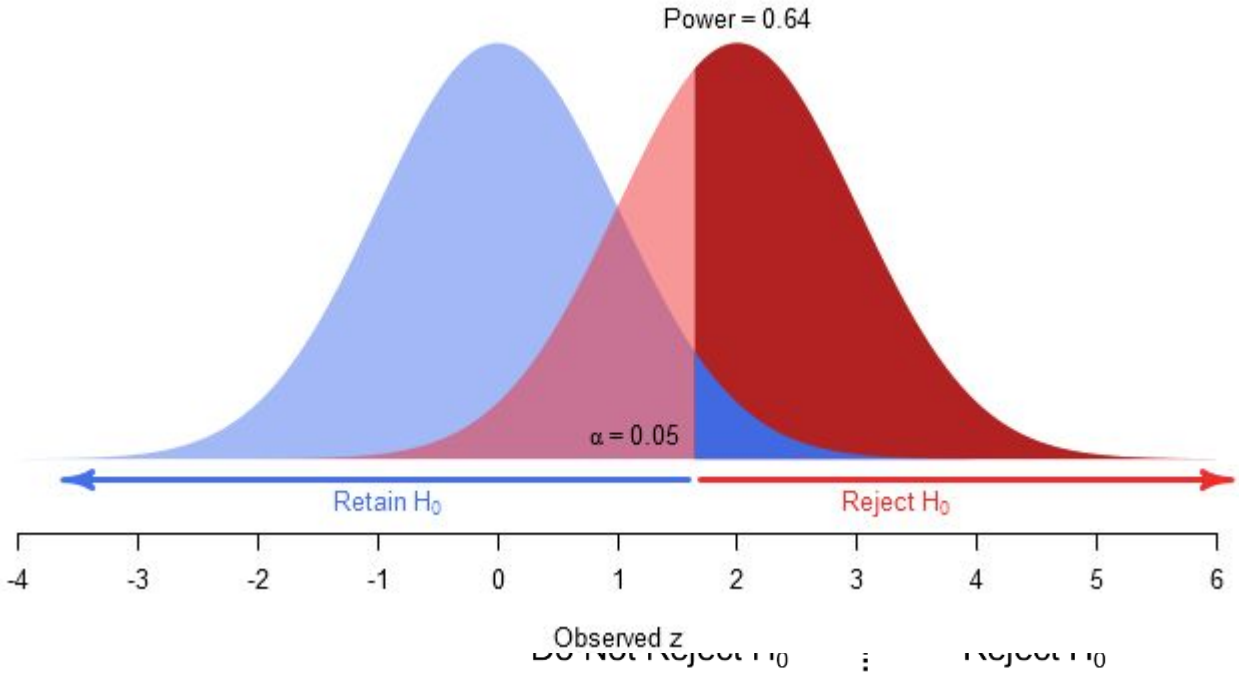
		H_0 is true
Test Result	Fail to Reject H_0	Correct Decis (1-α)
	Reject H_0	Type I Erro (α)



What happens to power when we increase the **alpha**?

Type I and II error

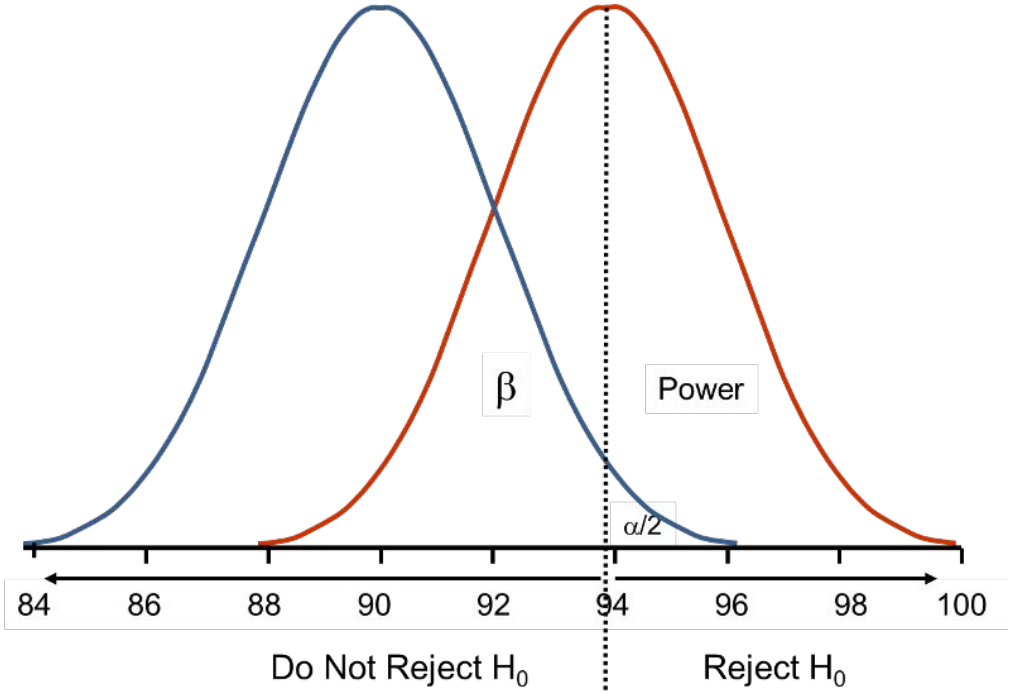
		Reality	
		H_0 is true	
Test Result	Fail to Reject H_0	Correct Decision (1-α)	
	Reject H_0	Type I Error (α)	



What happens to power when we move from a **one-tailed** test to a **two-tailed** test?

Type I and II error

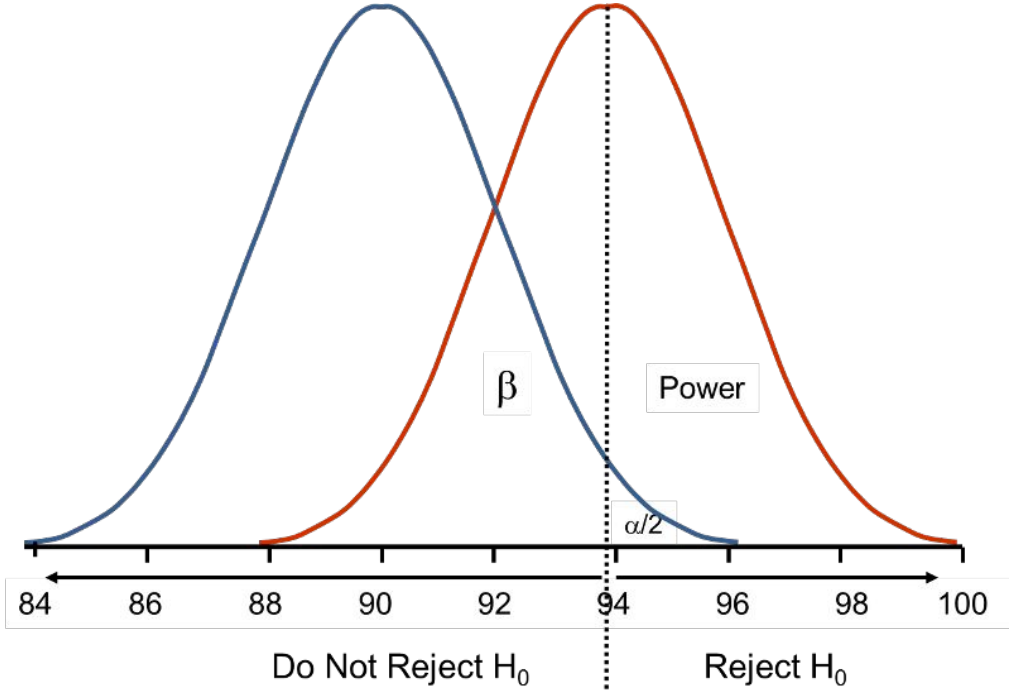
		Reality	
		H_0 is true	H_0 is false
Test Result	Fail to Reject H_0	Correct Decision (1-α)	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Decision (1-β)



What happens to power when we increase the **sample size**?

Type I and II error

		Reality	
		H_0 is true	H_0 is false
Test Result	Fail to Reject H_0	Correct Decision (1-α)	Type II Error (β)
	Reject H_0	Type I Error (α)	Correct Decision (1-β)



What happens to power if we increase the **sample standard deviation**?

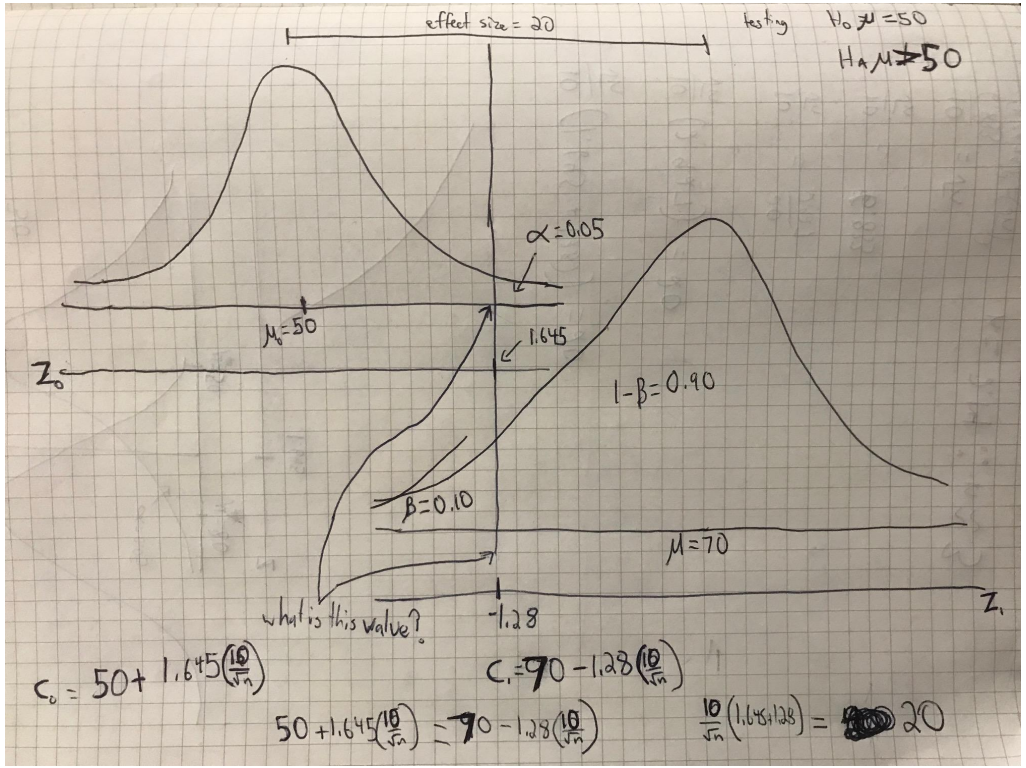
Example

An online storage solution PickUpBox wants to test if it can have an increase in yield for the number of free users that are converting to “premium” membership after seeing an advertisement. Currently, 4% of free members are converting to premium. PickUpBox wants to increase the conversion rate to 8%. If PickUpBox wants to have a 0.01 significance level, with at least a 95% power. How many people will need to view the advertisement before we are able to conclude the results of our hypothesis test?

Determining Sample Size Example

Imagine scientists discover what they believe to be a giant capybara in the Amazonian rainforest. To determine whether or not this is a new species, they want to see if its mass is significantly different than the average mass of capybaras 50 kg and standard deviation of 10 kg. Scientists want to see if this species is different with a simple raw mean difference effect size of 20 kg. What is the minimum sample size required if we want to determine significance at the 0.05 level and with a power of 0.9?

Example



$$\frac{10}{\sqrt{n}}(1.645 + 1.28) = 20$$

$$\frac{10}{\sqrt{n}}(2.927) = 20$$

$$\div$$

$$\frac{10}{\sqrt{n}} = \frac{20}{2.927}$$

$$\frac{10}{\sqrt{n}} = 6.833$$

$$\frac{10}{6.833} = \sqrt{n}$$

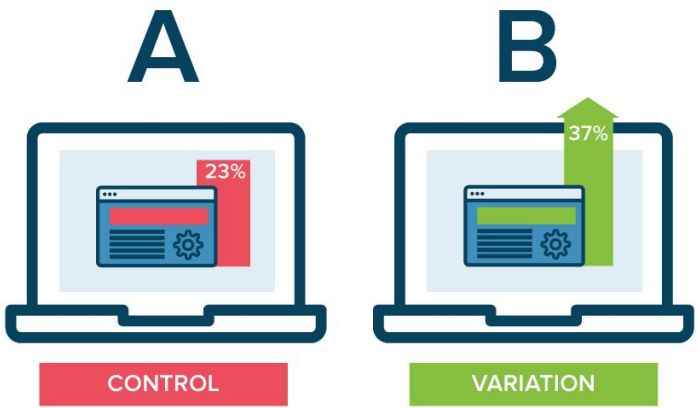
$$(1.4633)^2 = (\sqrt{n})^2$$

$$n = 2.14 \therefore n \approx 3$$

AB Testing

Multiple Comparisons

Repeated hypothesis test to determine which version of something (website, app, advertisement) is most effective.



Suppose you were trying out 20 new different website designs, and you wanted to see the ones that are effective at an α of 0.05. The current clickthrough rate is 0.003 (3 out of every 1000 visitors click on a link)

Multiple Comparison Problem

$$H_0 : \mu_d = 0.003$$

$$H_{A_1} : \mu_d \neq 0.003$$

$$H_{A_2} : \mu_d \neq 0.003$$

.....

$$H_{A_{20}} : \mu_d \neq 0.003$$

If we have an alpha of 0.05, what is the issue with comparing a null hypothesis to 20 separate alternate hypotheses???

Multiple Comparison Problem

Bonferonni Correction:

Divide alpha by the number of comparisons

In this example, our alpha level would become 0.05 / 1000 = 0.00005