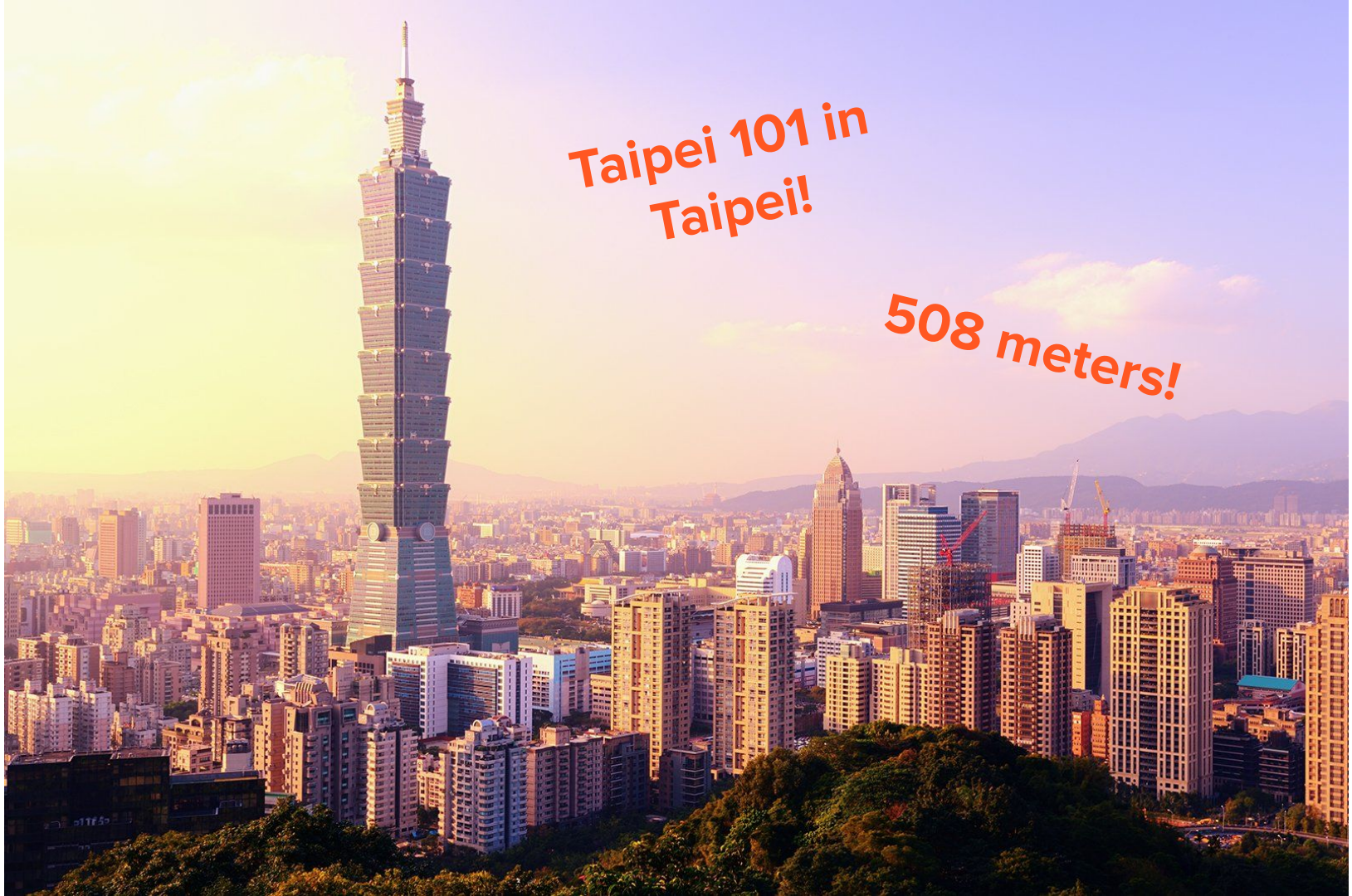
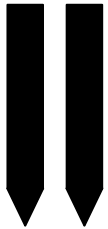
Teaching Demonstration

Eric Yanchenko
Global Connectivity
Akita International University

Taipei 101 in
Taipei!


Taipei 101 in
Taipei!
508 meters!

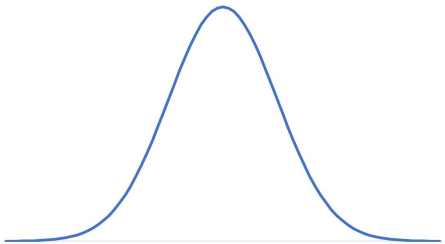
508 m



80 m

Data Science & AI

Statistics



A/B testing

Tech giants like Meta, Google, LINE, etc.

Principled approach to test new features/improvements, typically on websites

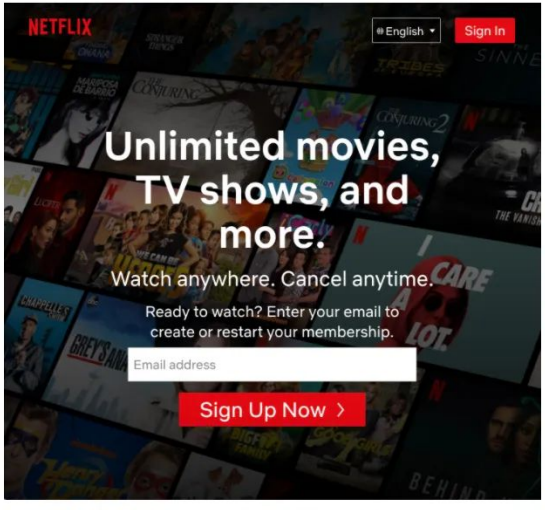
Split users into two (or more) groups

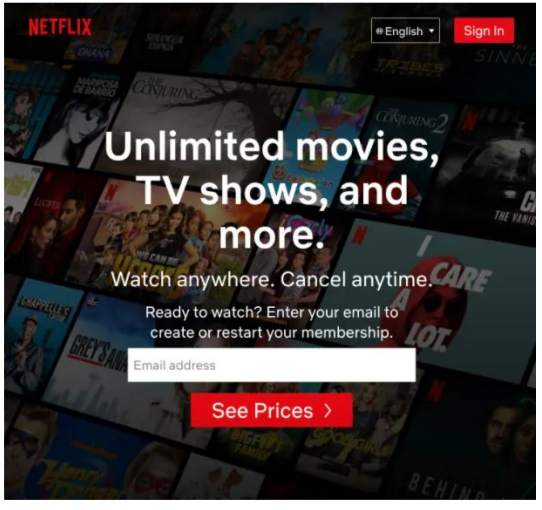
Give each group a different version/feature (treatment vs. control)

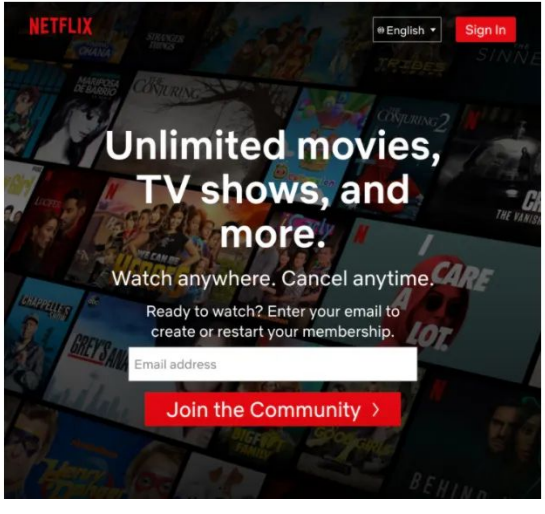
See which leads to a larger metric of interest (CTR, downloads, purchases, etc.)

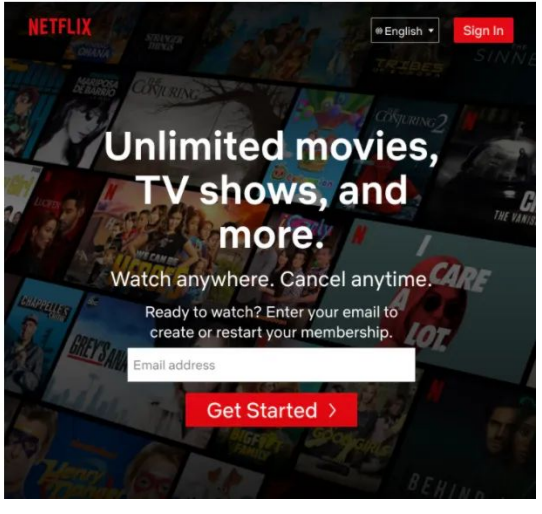
Netflix

In 2021, Netflix varied the Subscribe button to see if it increases subscriptions.



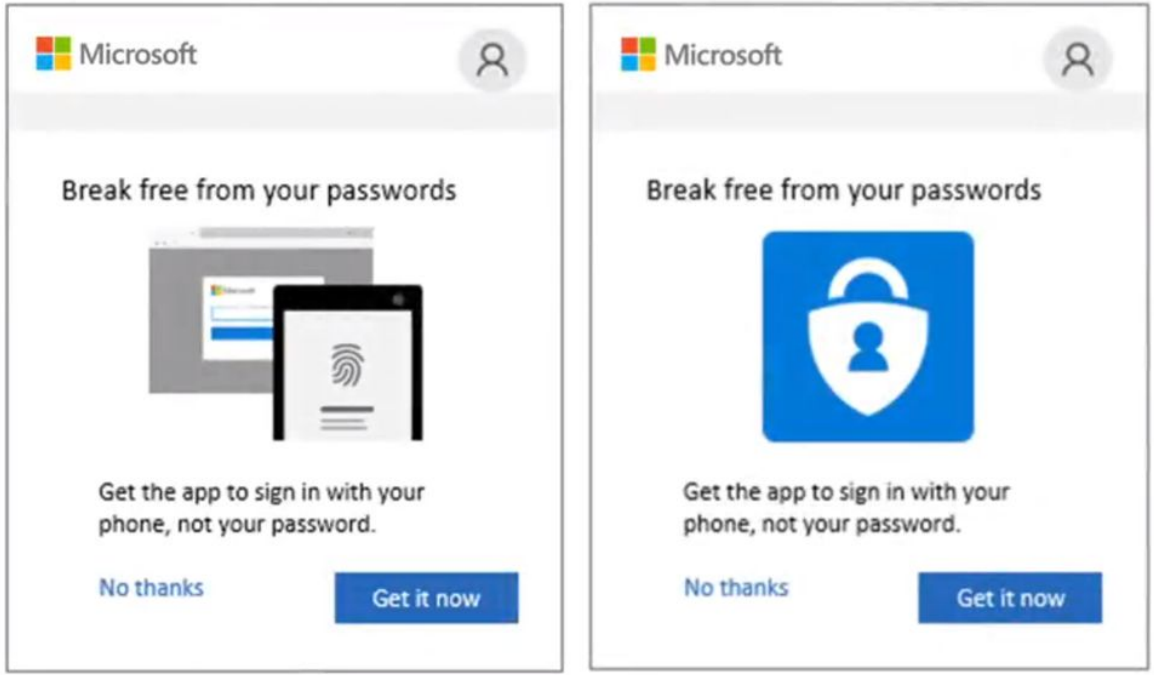






Microsoft

In 2020, Microsoft Azure wanted to increase the downloads of their Authenticator app.



A/B Testing

Statistical hypothesis testing

Motivating example

You are a senior data scientist at Amazon

Determine whether a new cart design leads to an increase in purchases

Set up an A/B testing experiment

Randomly assign 100 users to A (control) and 100 users to B (treatment)

Compare amount spent by users in each group

A

B

Hypothesis testing overview

1. Develop a scientific question of interest
    a. Cart A leads to more spending than cart B, Drug A is more effective than drug B, etc.

Hypothesis testing overview

1. Develop a scientific question of interest
    a. Cart A leads to more spending than cart B, Drug A is more effective than drug B, etc.
2. Formulate this hypothesis statistically
    a. Null (status quo) vs. Alternative

Hypothesis testing overview

1. Develop a scientific question of interest
    a. Cart A leads to more spending than cart B, Drug A is more effective than drug B, etc.
2. Formulate this hypothesis statistically
    a. Null (status quo) vs. Alternative
3. Derive a test statistic and its distribution
    a. E.g., mean of data, largest observation, etc.

Hypothesis testing overview

1. Develop a scientific question of interest
    a. Cart A leads to more spending than cart B, Drug A is more effective than drug B, etc.
2. Formulate this hypothesis statistically
    a. Null (status quo) vs. Alternative
3. Derive a test statistic and its distribution
    a. E.g., mean of data, largest observation, etc.
4. Determine decision criteria
    a. Reject or fail to reject

Shopping cart example notation

μ_A : true mean spending of users with Cart A (parameter)

μ_B : true mean spending of users with Cart B (parameter)

Shopping cart example notation

μ_A : true mean spending of users with Cart A (parameter)

μ_B : true mean spending of users with Cart B (parameter)

X_{iA} : spending of user i with Cart A, $i = 1, \ldots, n_A = 100$ (random variable)

X_{jB} : spending of user j with Cart B, $j = 1, \ldots, n_B = 100$ (random variable)

Shopping cart example notation

μ_A : true mean spending of users with Cart A (parameter)

μ_B : true mean spending of users with Cart B (parameter)

X_{iA} : spending of user i with Cart A, $i = 1, \ldots, n_A = 100$ (random variable)

X_{jB} : spending of user j with Cart B, $j = 1, \ldots, n_B = 100$ (random variable)

$\bar{X}_A$: average spending of users with Cart A (random variable)

$\bar{X}_B$: average spending of users with Cart B (random variable)

Shopping cart example notation

μ_A : true mean spending of users with Cart A (parameter)

μ_B : true mean spending of users with Cart B (parameter)

X_{iA} : spending of user i with Cart A, $i = 1, \ldots, n_A = 100$ (random variable)

X_{jB} : spending of user j with Cart B, $j = 1, \ldots, n_B = 100$ (random variable)

$\bar{X}_A$: average spending of users with Cart A (random variable)

$\bar{X}_B$: average spending of users with Cart B (random variable)

$$X_{1A}, \ldots, X_{100A} \overset{iid}{\sim} \text{Normal}(\mu_A, 10)$$

$$X_{1B}, \ldots, X_{100B} \overset{iid}{\sim} \text{Normal}(\mu_B, 10)$$

Shopping cart example

1. Question: Does shopping cart design B lead to greater sales than cart A?

Shopping cart example

1. Question: Does shopping cart design B lead to greater sales than cart A?
2. Hypothesis: ***How can we mathematically express our question of interest?***
    a. Hint: Describe the null hypothesis (status quo) and alternative hypothesis (what you want to show) using *parameters*.
3. Test statistic: ***How can we construct a statistic to test this hypothesis?***
    a. Hint: Relate the *random variables* to the hypotheses above.
4. Decision criteria: ***How do we decide whether to reject or fail to reject H0?***
    a. Hint: For what values of your test statistic would we reject? Fail to reject? Can be vague (large, small, etc.)

Shopping cart example

1. Question: Does shopping cart design B lead to greater sales than cart A?
2. Hypothesis:

$$H_0 : \mu_A = \mu_B \text{ vs. } H_1 : \mu_A < \mu_B$$

3. Test statistic:

$$Z = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{10^2}{n_A} + \frac{10^2}{n_B}}} = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{2}} \overset{H_0}{\sim} \text{Normal}(0, 1)$$

4. Decision criteria: Reject H0 if Z is “large.”
    a. How large is “large”?

p-values

P-values give a principled way to decide when to reject the null hypothesis

p-values

P-values give a principled way to decide when to reject the null hypothesis

Assume we are testing $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A < \mu_B$

p-values

P-values give a principled way to decide when to reject the null hypothesis

Assume we are testing $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A < \mu_B$

Let Z be our test statistic and assume we observed some z*

p-values

P-values give a principled way to decide when to reject the null hypothesis

Assume we are testing $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A < \mu_B$

Let Z be our test statistic and assume we observed some z*

Assuming the null is true

$$p\text{–value} = P(Z \geq z^* | H_0 \text{ is true})$$

Probability of observing a test statistic as extreme or more extreme than we did, assuming the null is true.

Example

Recall:

$$Z = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{10^2}{n_A} + \frac{10^2}{n_B}}} = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{2}} \overset{H_0}{\sim} \text{Normal}(0, 1)$$

Example

Recall:

$$Z = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{10^2}{n_A} + \frac{10^2}{n_B}}} = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{2}} \overset{H_0}{\sim} \text{Normal}(0, 1)$$

Run A/B testing experiment and observe:

$$\bar{x}_A = 100, \ \bar{x}_B = 102 \implies z^* = \frac{2}{\sqrt{2}}$$

Example

Recall:

$$Z = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{10^2}{n_A} + \frac{10^2}{n_B}}} = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{2}} \overset{H_0}{\sim} \text{Normal}(0, 1)$$

Run A/B testing experiment and observe:

$$\bar{x}_A = 100, \ \bar{x}_B = 102 \implies z^* = \frac{2}{\sqrt{2}}$$

Compute p-value:

$$p - \text{value} = P(Z \geq z^* | H_0 \text{ true}) = P(Z \geq 2/\sqrt{2} | H_0 \text{ true}) = 0.08$$

Example

Recall:

$$Z = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{\frac{10^2}{n_A} + \frac{10^2}{n_B}}} = \frac{\bar{X}_B - \bar{X}_A}{\sqrt{2}} \overset{H_0}{\sim} \text{Normal}(0, 1)$$

Run A/B testing experiment and observe:

$$\bar{x}_A = 100, \ \bar{x}_B = 102 \implies z^* = \frac{2}{\sqrt{2}}$$

Compute p-value:

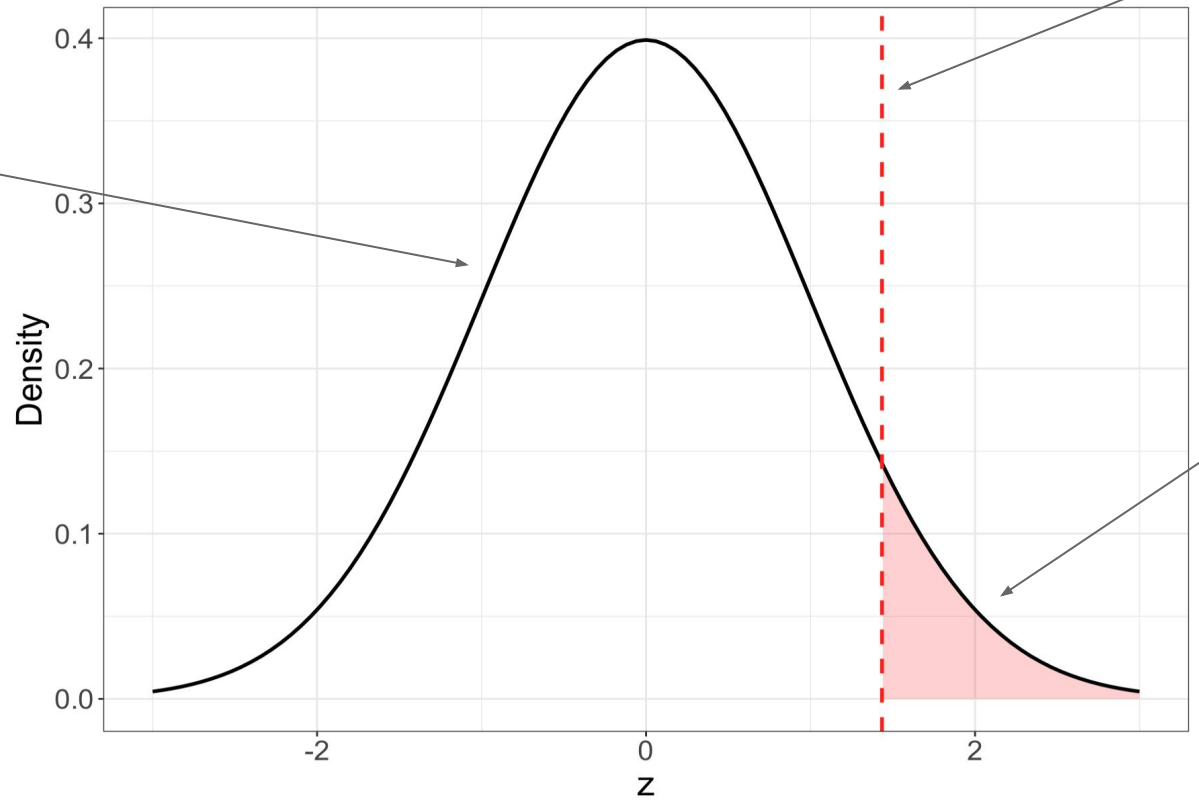
$$p - \text{value} = P(Z \geq z^* | H_0 \text{ true}) = P(Z \geq 2/\sqrt{2} | H_0 \text{ true}) = 0.08$$

Assuming the null hypothesis is true, there is an 8% chance of seeing a result as extreme or more extreme than the one we observed.

Example

Observed value of test statistic: z*

Distribution of Z under null: Normal(0,1)



Probability of more extreme value: p-value

Decisions, decisions

p-value = 0.08

Is this enough evidence to reject the null hypothesis?

Decisions, decisions

p-value = 0.08

Is this enough evidence to reject the null hypothesis?

Conventional threshold: 0.05

True or False, a p-value is:

- Evidence in favor (against) the null?

True or False, a p-value is:

- Evidence in favor (against) the null?

True or False, a p-value is:

- Evidence in favor (against) the null?
- The probability that the null is true?

True or False, a p-value is:

- Evidence in favor (against) the null?
- The probability that the null is true?

True or False, a p-value is:

- Evidence in favor (against) the null?
- The probability that the null is true?
- A statistic, i.e., random so it changes every time you collect data?

True or False, a p-value is:

- Evidence in favor (against) the null?
- The probability that the null is true?
- A statistic, i.e., random so it changes every time you collect data?

True or False, a p-value is:

- Evidence in favor (against) the null?
- The probability that the null is true?
- A statistic, i.e., random so it changes every time you collect data?

Statistical significance IS NOT the same as practical significance

Shiny app

Please go to:

https://eyanchenko.shinyapps.io/teaching/

You will have a chance to simulate data to better understand hypothesis testing

Questions

https://eyanchenko.shinyapps.io/teaching/

- How does increasing n_A, n_B affect the distribution of the sample means?
- How does varying sigma affect the likelihood of rejection?
- How does varying mu_B – mu_A affect the likelihood of rejection?
- If you re-run the experiment with the exact same settings, do you get the same results? How does this depend on the parameter values?
- If mu_B > mu_A, do you always get a small p-value? How does this depend on the other parameter values?
- What's something else interesting that you noticed?
- **For a given setting, what's your decision? Which cart design should we prefer?**

Thank you!

Thank you so much for your time and attention!

I will not be assigning any homework tonight :)

Please refer to my website for these slides as well as a link to the Shiny app.

http://www.ericyanchenko.com/