

CMPT 733 – Big Data Programming II

Hypothesis Testing

Instructor Steven Bergner

Course website <https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages/>

Slides by: Jiannan Wang

Why Hypothesis Testing?

We want to make a claim from our data

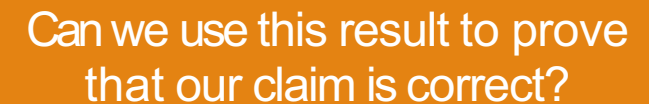
But, data is just a sample

How to prove our claim in this situation?

Using Hypothesis Testing

Example

- Claim: A data scientist earns more money than a data engineer
- Data: A sample of 50 data scientists and 50 data engineers
- Result: 100k vs. 70k



Can we use this result to prove that our claim is correct?

Hypothesis Testing

Equivalent Terms

- Hypothesis == Claim
- Hypothesis Testing == Claim Proving

Key Idea

- Prove by contradiction

Analogy

- How to prove: There exists no smallest positive rational number.
- Hint: a rational number is any number that can be expressed as the fraction a/b of two integers

Alternative and Null Hypotheses

Alternative Hypothesis (H_a)

- This is the claim that you want to prove it's correct

Null Hypothesis (H_0)

- The opposite side of H_a

Possible Outcomes

- Reject H_0 (a contradiction is found) → Accept H_a
- Fail to reject H_0 (no contradiction is found)

Example

Alternative Hypothesis (H_a)

- A data scientist earns **more** money than a data engineer

NULL Hypothesis (H_0)

- A data scientist earns **less (or equal)** money than a data engineer

If H_0 is true, what's the probability of seeing:

- ~~Data Scientist (100 K) vs. Data Engineer (70 K)~~
- $\text{Salary}(\text{Data Scientist}) - \text{Salary}(\text{Data Engineer}) \geq 30 \text{ K}$

This is called P-value

Make a decision based on p-value

We hope that

- p-value is as low as possible so that we can reject H_0 (i.e., accept H_a)

Level of Significance (e.g., $\alpha = 0.01$)

- How low do we want p-value to be?

Level of Confidence (e.g., $c = 1 - \alpha = 99\%$)

- How confident are we in our decision?

p-Hacking (Cheating on a p-Value)

Common Mistakes

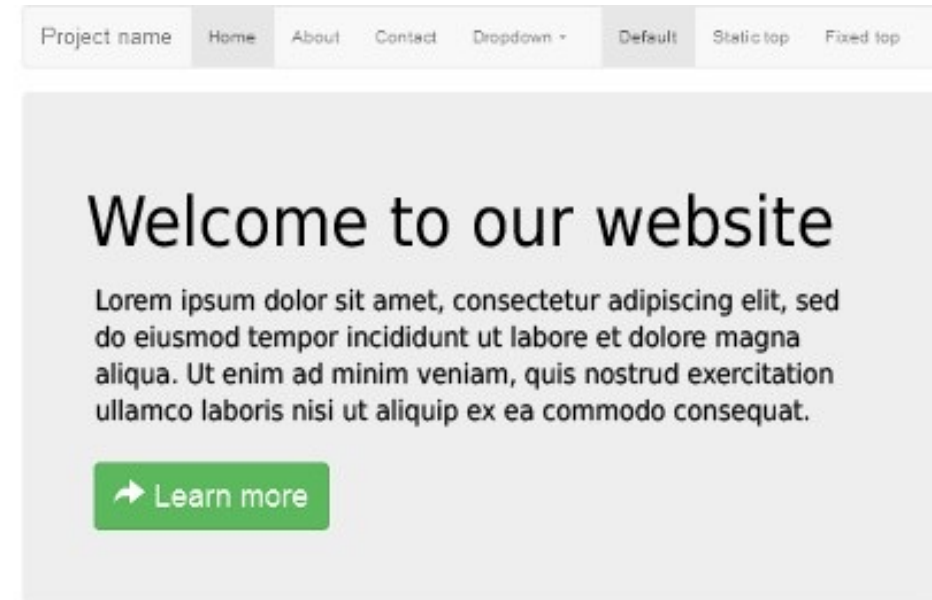
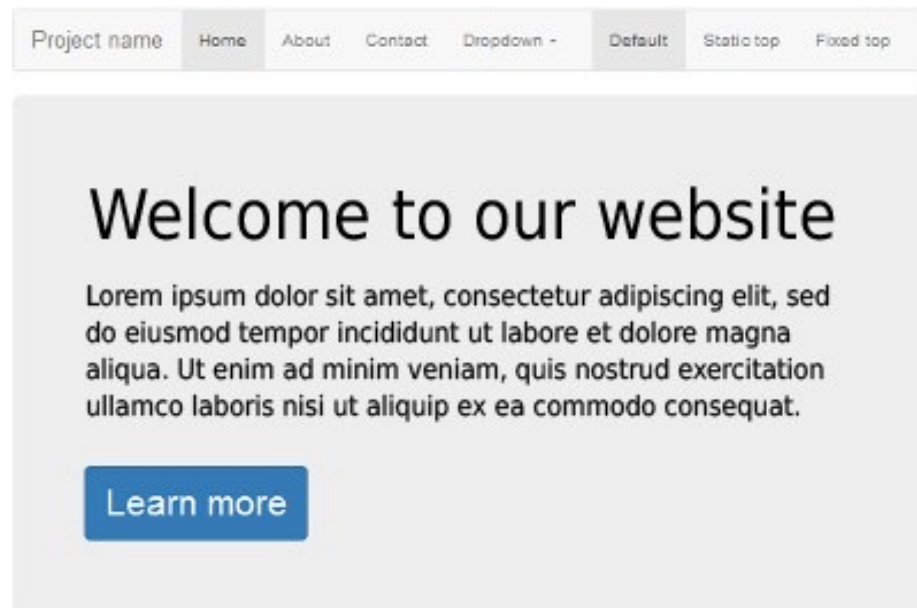
1. Collect data until the hypothesis testing is passed
2. Keep doing analysis on the same data until you find something significant

Solution

- You should know what you're looking for (H_0 and H_a) before you start
- Decrease the level of significance (e.g., $\alpha/2$ for two hypothesis tests on the same data)

A/B Testing

What UI is better?

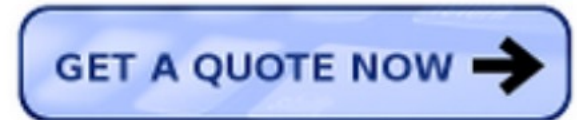


Surprising A/B Tests

A. Get \$10 off the first purchase. Book online now!

B. Get an additional \$10 off. Book online now.

Control Button



Experiment Button

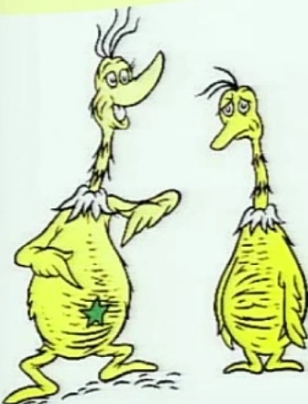


<https://www.wordstream.com/blog/ws/2012/09/25/a-b-testing>

Permutation Test

<https://youtu.be/lq9DzN6mvYA?t=8m9s>

**Sneetches:
Stars and
Intelligence**



Test Scores

★		×	
84	72	81	69
57	46	74	61
63	76	56	87
99	91	69	65
		66	44
		62	69

★ mean: 73.5
× mean: 66.9
difference: 6.6

PYCON 2016
ROSE CITY
PORTLAND, OREGON
MAY 28TH - JUNE 5TH

2023-03-16

8:51 / 40:44

Steven Bergner - CMPT 733

CC HD

Conclusion

- Hypothesis Testing
 - Null Hypothesis (H_0) and Alternative Hypothesis (H_a)
 - P-value and P-hacking
 - A/B Testing

CMPT 733 – Big Data Programming II

Causal Inference

Instructor

Steven Bergner

Course website

<https://coursys.sfu.ca/2024sp-cmpt-733-g1/pages/>

Slides by

Yi Xie, Nathan Yan, Jiannan Wang

Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

Questions Data Scientists Can Answer

Is This A or B?	Classification
How much or How Many?	Regression
Is This Weird?	Anomaly Detection
How Is This Organized?	Clustering
What if?	Causal Inference

From Prediction to Causation

Predicting user activity for Xbox

- Y = logins in next month
- X = logins in past month, number of friends,...

What if we increase the number of friends?

- Would it increase user activity?

Maybe or May be not (!)

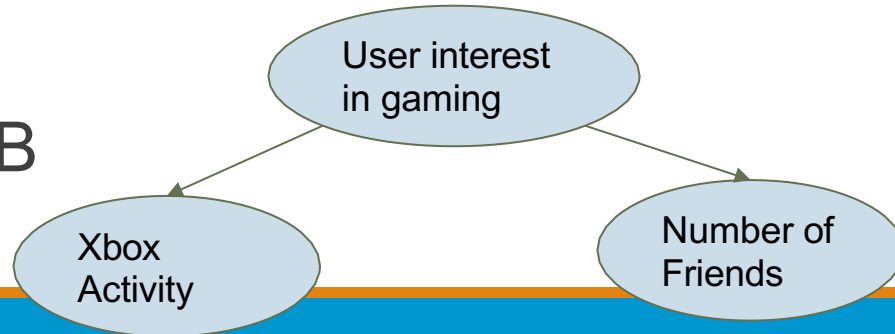
A causes B



B causes A



C causes A and B



A/B Testing Helps!

Treatment Group

- A random sample of users
- Launch a campaign to increase friends
- Average activity next month

Control Group

- A random sample of users
- **Not** Launch a campaign to increase friends
- Average activity next month



**Hypothesis
Testing**

A/B Testing Does **Not** Work

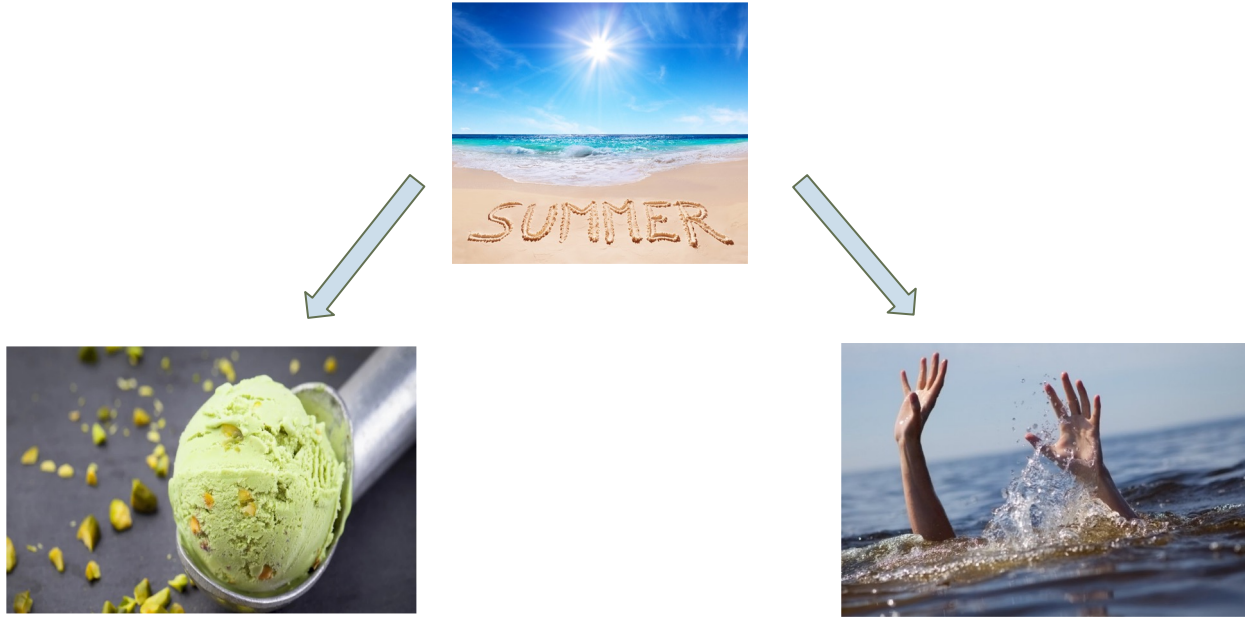
- **It is infeasible to do A/B testing**
 - What if you went to UBC rather than SFU, would it help you land a better job?
- **It is unethical to do A/B testing**
 - What if subscription price is set to \$69 rather than \$99, would it increase revenue?

Example 1: Causality \neq Correlation

When a team was investigating the city death rate data, they found that reported cases of teenage drowning death increase while the sales of ice cream also increase.



Example 1: Causality \neq Correlation



Example 2: Causality \neq Correlation

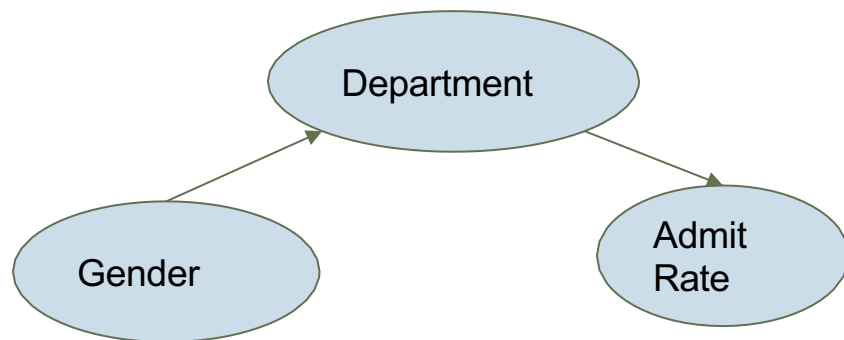
Is Berkeley gender biased?

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Simpson's paradox

Example 2: Causality \neq Correlation

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%



Example 3: Causality \neq Correlation



- In city areas with nearby trees and natural landscapes, there is less domestic violence.
- Apartment complexes with many trees had 52% fewer crimes.
- On tree-lined streets, people drive more slowly, reducing accident risk.

Outline

Why Should Data Scientists Care?

Basic Concepts

- Outcome / Treatment Variables
- Intervention and Do Operation
- Counterfactual
- Causal Graph

Causal Inference

Outcome / Treatment Variables

What if participating **Study Program**, would it improve **Grade**?

Treatment



Outcome



Student	Gender	Class	Study Program	Grade
Jacky	male	1	0	78
Terry	male	1	1	82
Mary	female	1	0	86
Sarah	female	2	1	83

Intervention and Do Operation

- ❖ Do operator will signal the experimental intervention (invented by Judea Pearl)

$$P(\text{grade}|\text{do}(\text{enrolled in study program}))$$

represents the distribution of grad if the person is enrolled in study program

Intervention and Do Operation

- ❖ $P(A \mid B = b)$: probability of A being true given that B is observed as $B = b$
- ❖ $P(A \mid \text{do}(B) = b)$: probability of A being true given an intervention that sets B to b

Counterfactual

❖ What would have happened if I had changed “Treatment Variable”

Student	Gender	Class	Study Program	Actual Grade
Jacky	male	1	0	78
Terry	male	1	1	82
Mary	female	1	0	86
Sarah	female	2	1	83

Counterfactual

❖ What would have happened if I had changed “Treatment Variable”

Student	Gender	Class	Study Program	Actual Grade	Counterfactual Grade (You cannot get it in reality)
Jacky	male	1	0	78	82
Terry	male	1	1	82	82
Mary	female	1	0	86	90
Sarah	female	2	1	83	85

Causality

Causality Definition

- The difference between actual outcome and counterfactual outcome

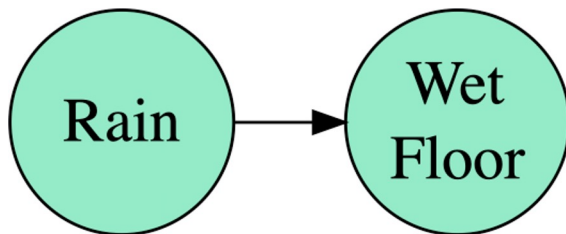
The Fundamental Problem of Causal Inference

- We cannot observe the counterfactual outcome

Causal Graph

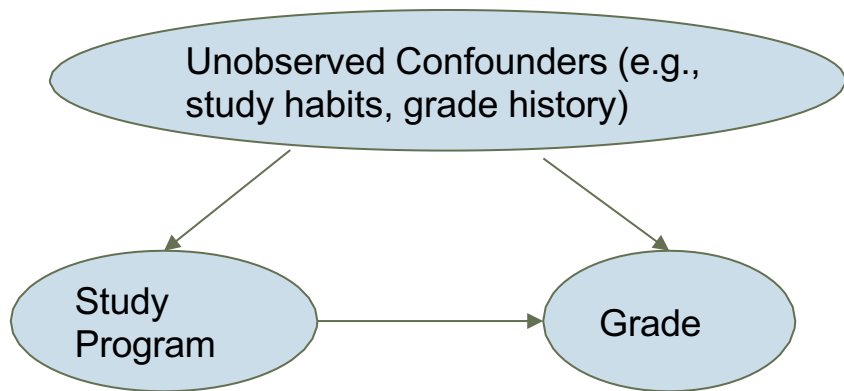
Causal graph is a **directed** graph

- Nodes: variables
- Directed Edges: X affects Y



Causal Graph

- ❖ Confounding variables: common cause of treatment and outcome



Why Causal Graph?

- Helpful to identify which variables to control for
- Make assumptions explicit

Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice

Statistical vs. Causal Inferences

Statistical inference

- Data is just a sample
- Your goal is to infer a population
- Think about how to go “backwards” from sample to population

Causal inference

- Derive a treatment group from Data
- Derive a control group (i.e., without treatment) from Data
- Think about how to infer the actual effect of treatment from the derived treatment and control groups

Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice

Individual Treatment Effect

What is the grade difference between enrolling and not enrolling in the study program?

Student	Gender	Class	Enroll in Study Program	Not Enroll in Study Program	
Jacky	male	1	82	78	4
Terry	male	1	82	82	0
Mary	female	1	90	86	4
Sarah	female	2	83	85	-2

Average Treatment Effect (ATE)

The average of all values for individual treatment effects

$$ATE = (4 + 0 + 4 + -2) / 4 = 1.5$$

Student	Gender	Class	Enroll in Study Program	Not Enroll in Study Program	
Jacky	male	1	82	78	4
Terry	male	1	82	82	0
Mary	female	1	90	86	4
Sarah	female	2	83	85	-2

ATE Estimation

ATE estimation methods

- ❖ Matching based:
 - Perfect matching
 - Nearest neighbor matching
 - Propensity score matching
- ❖ ML based:
 - Regression method
 - Representation learning

Perfect Matching

student	gender	class	Study program	grade
Terry	male	1	1	82
Sarah	female	2	1	83
Jacky	male	1	0	78
Mary	female	1	0	86

Find the “perfect matching in counterfactual world”

Perfect Matching

student	gender	class	Study program	grade
Terry	male	1	1	82
Sarah	female	2	1	83
Jacky	male	1	0	78
Mary	female	1	0	86

4

Find the “perfect matching in counterfactual world”

Perfect Matching

Student	Gender	Class	Study Program	Grade
Terry	male	1	1	82
Sarah	female	2	1	83
Jacky	male	1	0	78
Mary	female	1	0	86

By perfect matching, we can't find ATE over the population, because for tuple Sarah, we can't find the perfect match in control group

Nearest Neighbor Matching

Student	Gender	Class	Study Program	Grade
Terry	male	1	1	82
Sarah	female	2	1	83
Jacky	male	1	0	78
Mary	female	1	0	86

-3

Find the “nearest matching in counterfactual world”

Nearest Neighbor Matching

Student	Gender	Class	Study Program	Grade	
Terry	male	1	1	82	4
Sarah	female	2	1	83	-3
Jacky	male	1	0	78	
Mary	female	1	0	86	

$$ATE = \frac{1}{2} [3 + 4] = 3.5$$

Propensity Score Matching

Steps

- ❖ Using logistic regression to infer $e(x) = Pr[T=1|X=x]$
- ❖ Match users with treatment 0 with users with treatment 1 based on propensity score, using matching methods

Propensity Score Matching

Student	Gender	Class	Study Program	Grade	PSE
Terry	male	1	1	82	0.7
Sarah	female	2	1	83	0.6
Jacky	male	1	0	78	0.7
Mary	female	1	0	86	0.55

Based on PSM derived by logistic regression, the match we'll find for Terry is Jacky, and the match we'll find for Sarah is Mary

ATE = 3.5, computation is similar with nearest neighbor matching

Regression Method

Intuition: the distribution of Y given X is different when treatment is different

Train two separate regression models under Treatment = 0 or Treatment = 1, infer $p(Y | T = 0, X)$ and $p(Y | T = 1, X)$

Regression Method

Student	Gender	Class	Study Program	Grade
Terry	male	1	1	82
Sarah	female	2	1	83
Jacky	male	1	0	78
Mary	female	1	0	86

Treat one regression model 1 on [Terry, Sarah] where Study Program = 1

treat another regression model 2 on [Jacky, Mary] where Study Program = 0

Using model 1 to compute counterfactual outcome of [Jacky, Mary], same for model 2

Other ML-based Methods

- ❖ Representation learning
- ❖ Intuition: transform dataset into a space where treatment assignment is more evenly distributed
- ❖ For other more techniques, please check latest publications

Outline

Why Should Data Scientists Care?

Basic Concepts

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice

Causal Inference Case Study

What if you went to UBC rather than SFU, would it help you land a better job?

1. What data to collect?
2. How to use data to answer this question?

Allocating Policy for Homelessness

- ❖ Researchers allocate different interventions (like emergency shelter, rapid rehousing) for homelessness based on causal inference
- ❖ Report published in AAI 2019

Amanda Kube, Sanmay Das, Patrick J. Fowler: [Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services](#). AAI 2019: 622-629

Social Media

- ❖ For Twitter, the impact of race, gender and closeness on persuasion is studied using causal inference
- ❖ Result published in NeurIPS 2019

https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/a/238/files/2019/12/ld_104_final.pdf

DoWhy Python Library

- ❖ DoWhy is a Python library for causal inference that supports explicit modeling and testing of causal assumptions, developed by Microsoft.
- ❖ DoWhy is based on a unified language for causal inference, combining causal graphical models and potential outcomes frameworks.

Summary

Why Should Data Scientists Care?

- “What if” Question ?
- Why not A/B testing?
- Causality \neq Correlation

Basic Concepts

- Outcome / Treatment Variables
- Intervention and Do Operation
- Counterfactual
- Causal Graph

Causal Inference

- Statistical Inference vs. Causal Inference
- Average Treatment Effect (ATE) Estimation
- Causal Inference in Practice