



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

Drawing Conclusions from Measurement

Recap

- Supervised learning:
 - Data annotation
 - Classification models
- Running example: do Republicans talk more about taxes than Democrats?
 - We can estimate the difference, but how can we tell if the estimated difference is meaningful?

Overview

- Hypothesis Testing:
 - Background
 - Examples
- Additional Metrics:
 - Effect size
 - Confidence intervals
- Starting next topic: Causal Inference

Hypothesis Testing

Formalizing Hypotheses

- Formalize a research question into testable hypothesis

Question	Hypothesis
What strategies does the Russian government use to manipulate public opinion?	State-affiliated and independent news outlets in Russia covered the Russia-Ukrainian War at different rates

Formalizing Hypotheses

- Null hypothesis: a claim, assumed to be true, that we'd like to test (because we think it is wrong)

Question	Hypothesis	Null Hypothesis
What strategies does the Russian government use to manipulate public opinion?	State-affiliated and independent news outlets in Russia covered the Russia-Ukrainian War at different rates	Level of coverage is the same

Key Idea

- If the null hypothesis were true, how likely is it that you'd see this data?

Example:

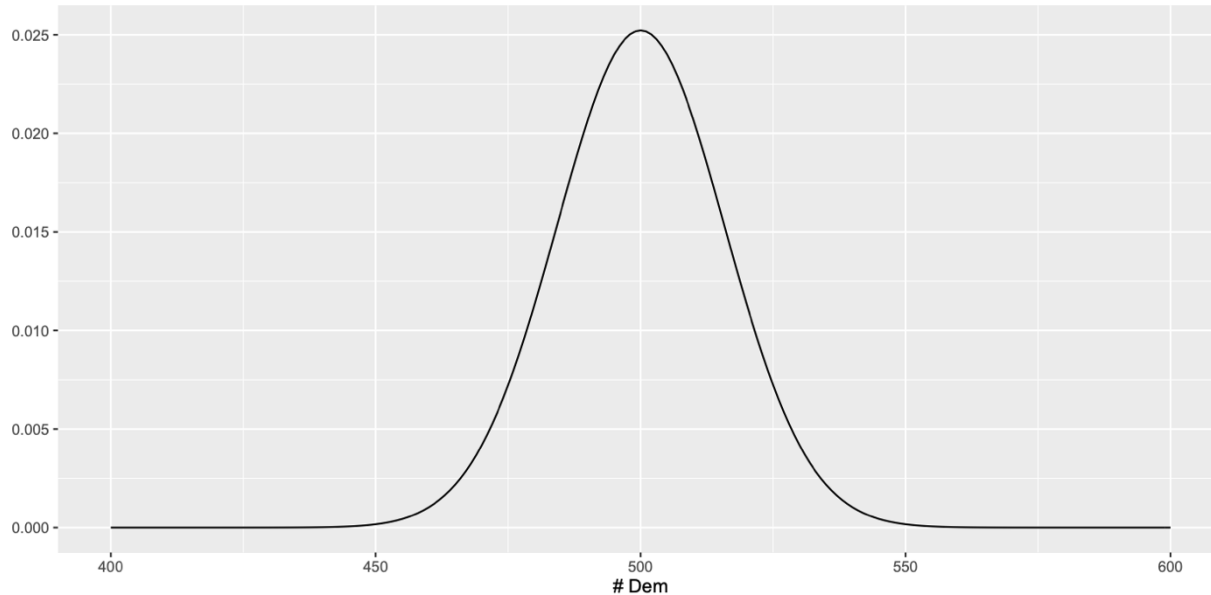
- Hypothesis: Baltimore residents tend to be politically liberal
- H_0 : Among all N registered {Democrat, Republican} primary voters, there are an equal number of Democrats and Republicans in Baltimore

$$\frac{\#dem}{N} = \frac{\#rep}{N} = 0.5$$

Hypothesis Testing

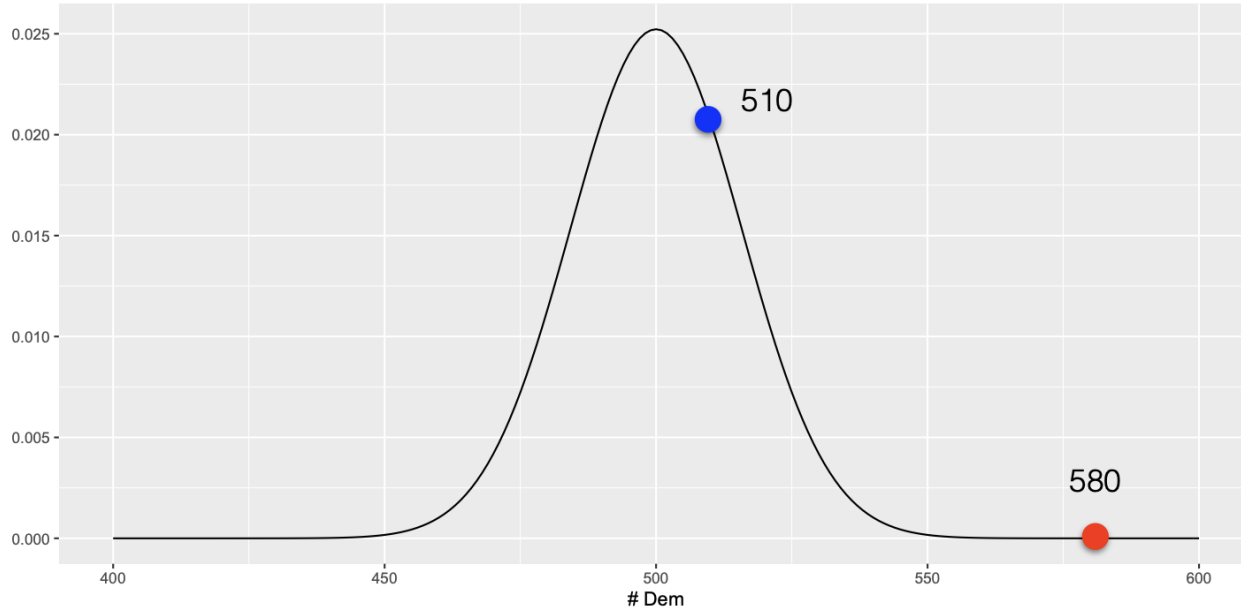
- If we had access to the party registrations (and knew the population), we would have our answer.
- Hypothesis testing measures our confidence in what we can say about a null *from a sample*

Hypothesis Testing: Example



Binomial probability distribution for number of democrats in $n=1000$ with $p = 0.5$

Hypothesis Testing: Example



At what point is a sample statistic unusual enough to reject the null hypothesis?

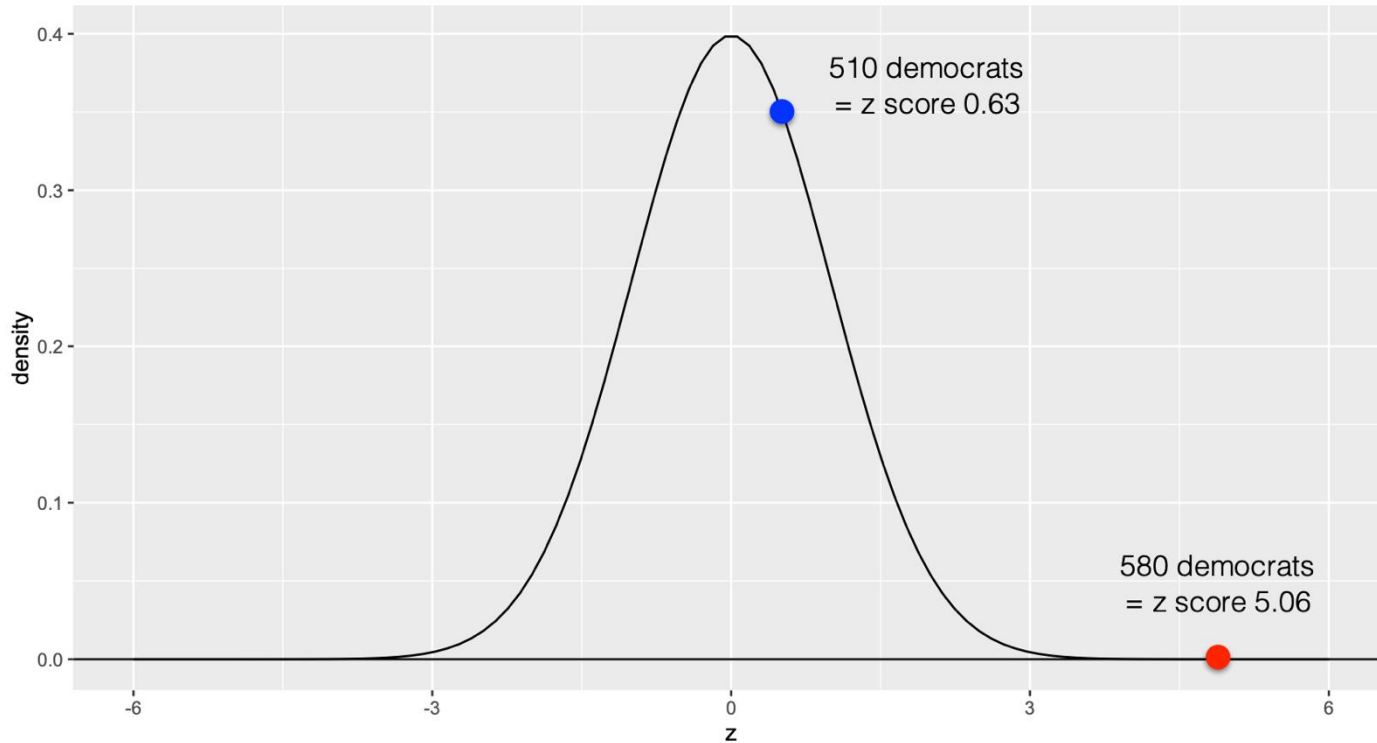
Hypothesis Testing: Example

- The form we assume for the null hypothesis lets us quantify that level of surprise.
- We can do this for many parametric forms that allows us to measure $P(X \leq x)$ for some sample of size n ; for large n , we can often make a normal approximation.

Compute a test statistic: Z-score

- Number of standard deviations away a data point or observed value is from the mean
- For Normal distributions, transform into standard normal (mean = 0, standard deviation = 1)
 - $Z = \frac{X - \mu}{\sigma / \sqrt{n}}$
- For Binomial distributions, normal approximation (for large n)
 - $Z = \frac{Y - np}{\sqrt{np(1-p)}}$
 - Y = 580 (democrats in sample)
 - p = 0.5 (proportion we are testing)
 - n = 1000 (sample size)

Hypothesis Testing: Example

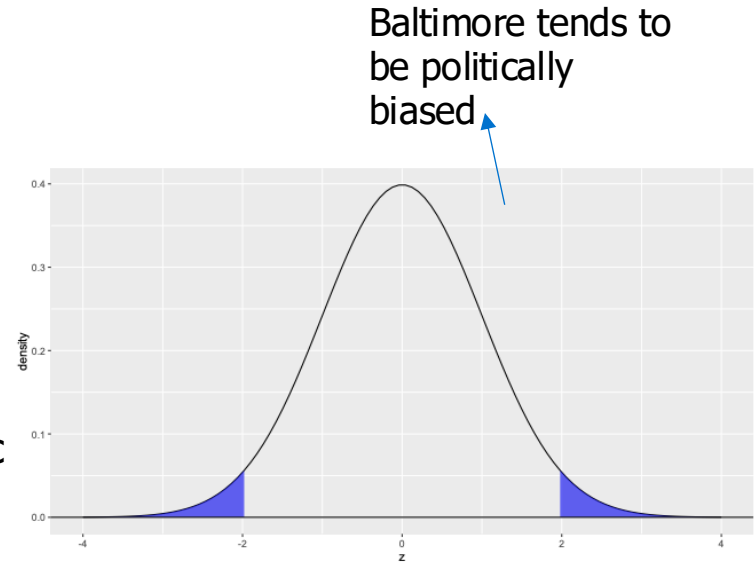


Significance Testing

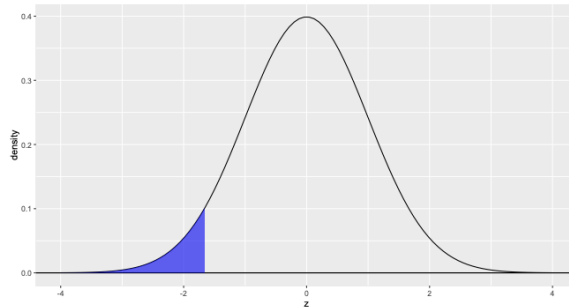
- Decide on the level of significance $\alpha \in \{0.05, 0.01\}$
 - 0.05: we reject the null when the probability of getting a sample mean is less than 5% if the null were true
- Testing is evaluating whether the sample statistic falls in the rejection region defined by α

Tails

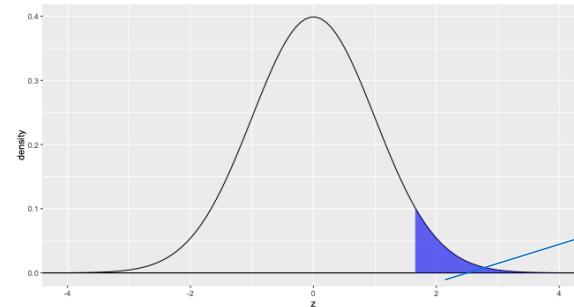
- Two-tailed tests measured whether the observed statistic is different (in either direction)
- One-tailed tests measure difference in a specific direction
- All differ in where the rejection region is located; $\alpha = 0.05$ for all.



two-tailed test



lower-tailed test



upper-tailed test

Baltimore tends to be liberal

P-value

- A p-value is the probability of observing a statistic at least as extreme as the one we did if the null hypothesis were true.
- Two-tailed test
 - $\text{p-value}(z) = 2 * P(Z \leq -|z|)$
- Lower-tailed test
 - $\text{p-value}(z) = P(Z \leq z)$
- Upper-tailed test
 - $\text{p-value}(z) = 1 - P(Z \leq z)$

Choosing a Parametric hypothesis test

	Predictor variable	Outcome variable	Research question example
Paired t-test	<ul style="list-style-type: none"> • Categorical • 1 predictor 	<ul style="list-style-type: none"> • Quantitative • groups come from the same population 	What is the effect of two different test prep programs on the average exam scores for students from the same class?
Independent t-test	<ul style="list-style-type: none"> • Categorical • 1 predictor 	<ul style="list-style-type: none"> • Quantitative • groups come from different populations 	What is the difference in average exam scores for students from two different schools?
ANOVA	<ul style="list-style-type: none"> • Categorical • 1 or more predictor 	<ul style="list-style-type: none"> • Quantitative • 1 outcome 	What is the difference in average pain levels among post-surgical patients given three different painkillers?
MANOVA	<ul style="list-style-type: none"> • Categorical • 1 or more predictor 	<ul style="list-style-type: none"> • Quantitative • 2 or more outcome 	What is the effect of flower species on petal length, petal width, and stem length?

Assumes data is normally distributed

Z-test if sample size is large or variance is known

Errors

		Test Results	
		Keep null	Reject null
Truth	Keep null		Type I Error
	Reject null	Type II Error	

- For any significance level α and n hypothesis tests, we can expect $\alpha \times n$ type I errors.
- $\alpha=0.01, n=1000 = 10$ "significant" results simply by chance

Multiple Hypothesis Correction

- Bonferroni correction:
 - For family-wise significance level α_0 with n hypothesis tests:
$$\alpha \leftarrow \frac{\alpha_0}{n}$$
 - [Very strict; controls the probability of at least one type I error.]
 - Less strict alternative: Holm method
- Alternative: Benjamini-Hochberg (False discovery rate)
 - Less strict, controls the expected proportion of false discoveries

Example statistical tests in the wild

- Research question:
 - Do Wikipedia articles perpetuate gender and racial bias and stereotypes?
- Hypothesis:
 - Wikipedia articles about women have longer “personal life” sections
 - Wikipedia articles about men have longer “career” sections
 - [Null: section sizes are the same]

Why do we care about Wikipedia?

- Extremely widely read platform--bias and stereotypes on Wikipedia can influence readers
- Wikipedia is widely used as training data in NLP models
 - We know that models are liable to absorbing and amplifying data biases
- In some aspects, Wikipedia can be seen as a reflection of society – we can learn about bias and stereotypes in society by studying Wikipedia
- Bonus: Wikipedia has an active editor community, and studies that have revealed biases on the platform have motivated the editor community to correct them [Langrock and Sandra González-Bailón 2020; Reagle and Rhue 2011]

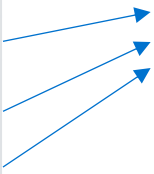
Example statistical tests in the wild

- Method:
 - Pair each article about a women with an article about a man with similar attributes (e.g. profession, nationality)
 - For each article, compute the % of words in each subsection
 - Controls for variable article length
 - Compare proportions of subsections
- Significant test:
 - Paired t-test
 - Multiple hypothesis correction

Results for “personal”, “life”, “early” and “career” sections

Section	Average % women	Average % men	p-value
career	17.47%	15.55%	4.20E-41
early life and education	1.43%	1.09%	1.08E-31
personal life	2.47%	2.12%	1.35E-25

Women have longer “career” and “life” sections?



Example statistical tests in the wild

- Research question:
 - Do Wikipedia articles perpetuate gender and racial bias and stereotypes?

Example statistical tests in the wild

- Research question:
 - Do Wikipedia articles perpetuate gender and racial bias and stereotypes?
- Hypothesis:
 - Wikipedia articles about women have longer “personal life” sections
 - Wikipedia articles about men have longer “career” sections
 - [Null: section sizes are the same]
- Conclusion:
 - We do find significance differences in article sections
 - Men tend to have fine—grained career descriptions, while women have more generic career descriptions and more words discussing “personal life”

Example statistical tests in the wild

- Research question:
 - Do Wikipedia articles reflect language and nationality bias?
 - “Local heroes” effect
- Hypothesis:
 - Articles about Hispanic/LatinX Americans are available in fewer languages than articles about non-Hispanic/LatinX people
 - Articles about Hispanic/LatinX Americans are available in Spanish more often
- Which statistical test to use?
 - t-test/z-test are for continuous data where we can assume a normal distribution
 - Test for *paired categorical* data: McNemar’s Test

Example statistical tests in the wild

	Hispanic/LatinX	Non-Hispanic/LatinX	P-value
# of Languages	7.54	7.62	0.668
% available in Spanish	34.09%	27.54%	2.15E-12

Paired t-test, not significant

McNemar's test, significant

Non-parametric hypothesis tests

	Predictor variable	Outcome variable	Use in place of...
Spearman's r	<ul style="list-style-type: none"> Quantitative 	<ul style="list-style-type: none"> Quantitative 	Pearson's r
Chi square test of independence	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Categorical 	Pearson's r
Sign test	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Quantitative 	One-sample t -test
Kruskal-Wallis H	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 	ANOVA
ANOSIM	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 2 or more outcome variables 	MANOVA
Wilcoxon Rank-Sum test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from different populations 	Independent t -test
Wilcoxon Signed-rank test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from the same population 	Paired t -test

Considerations for choosing a hypothesis test

- If data is continuous, categorical, or binary
- Sample size
- If data can be assumed to be normally distributed (can depend on sample size)
- Number of variables to be analyzed
- If data is paired



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

Effect size and confidence intervals

Other useful ways of conveying results

- Effect size
- Confidence intervals
 - Bootstrap

Effect Size

- Hypothesis testing tells us if difference exists – not how big the differences is
- Common measurement: Cohen's d:
 - difference in means divided by (pooled) standard deviation of data
 - $\frac{avg(x_1) - avg(x_2)}{s}$

d	size
0.01	Very small
0.2	small
0.5	medium
0.8	large
1.2	Very large
2	huge

Confidence Intervals

- Interpretation:
 - The probability that a parameter will fall within a particular range X% of the time
 - 95% confidence interval: If we repeat this experiment infinite times, 95% of the time, the outcome will be in range we computed
 - Handwave: we're 95% certain this range contains the true value [it could be the 5% of experiments where we got the wrong range]
 - The 95% confidence interval for an effect will exclude the null value if and only if the test of significance yields a p value of less than 0.05

Confidence Intervals: Computation

$$CI = \bar{X} \pm z \frac{\sigma}{\sqrt{n}}$$

\bar{X} : sample mean

z : controls size of confidence interval

σ : population standard deviation

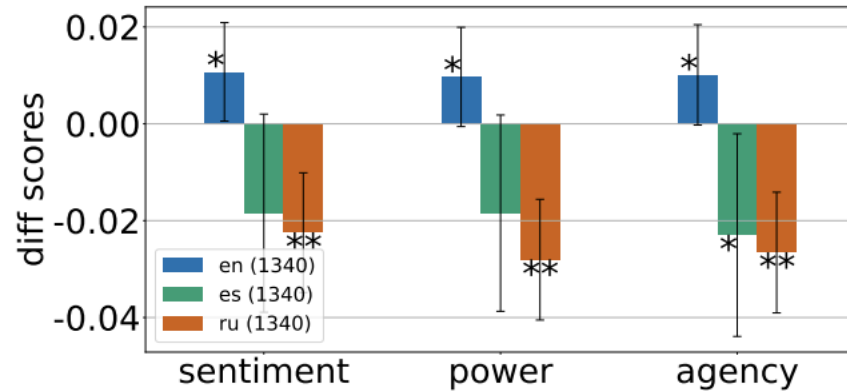
n : sample size

Requires that sample is random, variable of interest has a normal distribution, and that we know population standard deviation

Bootstrap

- Resampling process for estimating summary statistics
- Choose number of bootstrap samples and sample size
 - For each bootstrap sample:
 - Draw a sample with replacement with the chosen size
 - Calculate the statistic on the sample
- Calculate the mean of the sample statistics. Can calculate confidence intervals from ranked statistics
- Can be used to evaluate predictive power of ML models: bootstrap sampling over the *training* set and retrain model over each sample

Example: Connotations about LGBT people in Wikipedia articles

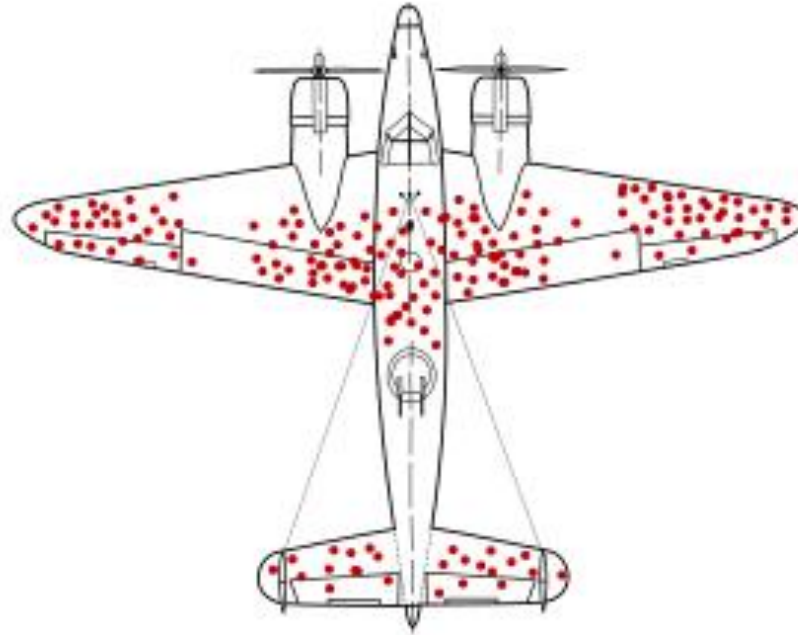


- English articles: LGBT people are portrayed with more **positive** connotations than non-LGBT people
- Spanish articles: results are inconclusive with wide confidence intervals
- Russian articles: LGBT people are portrayed with more **negative** connotations than non-LGBT people
- [Significance measured using paired t-test]

Hypothesis testing can't guarantee your results are valid!

- Examples:
 - Selection/sampling/survivorship bias
 - Simpson's paradox

Survivorship Bias



Example from Wikipedia

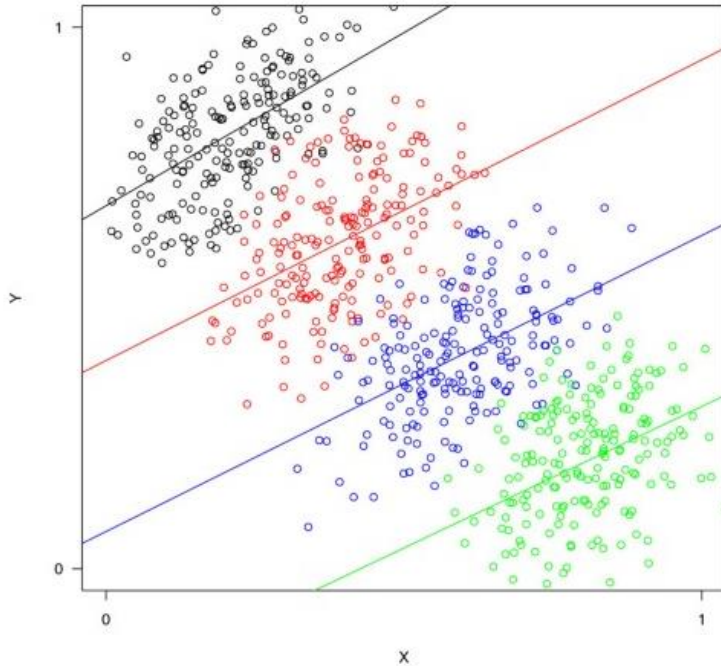
- Numerous studies have shown that Wikipedia articles about women tend to be longer than articles about men [Graells-Garrido et al. 2015; Reagle & Rhue 2011; Wagner et al. 2015; Young et al. 2016]
- Naïve conclusion:
 - Wikipedia editors are biased towards women and spend more time writing detailed accounts of their lives
- Hypothesis [Wagner et al. 2015]
 - There is a higher bar for women to have Wikipedia articles – only the most famous women with lots of accomplishments have Wikipedia articles
 - “Glass ceiling effect”

Simpson's Paradox

	Total Average Length
Articles about men	587.85 (n = 342,677)
Articles about women	634.89 (n = 108,915)

Articles about women are longer (significantly)

Simpson's Paradox



- Data looks negatively correlated overall
- Subsetting data shows positive correlations

Simpson's Paradox

	Total Average Length	Articles about non-athletes	Articles about athletes
Articles about men	587.85 (n = 342,677)	900 (n = 142,677)	365.17 (n = 200,000)
Articles about women	634.89 (n = 108,915)	600 (n = 50,000)	195.74 (n = 58,915)

- It's true that Wikipedia articles about men are dominated by athlete "stubs"
- It's true that when you control for things like "is the article a stub" and "is the person an athlete", you no longer see articles about women are longer
- [I made up the actual numbers]

Takeaways

- Basic idea behind hypothesis testing
- Ability to interpret use of hypothesis tests by others
- Ability to select a hypothesis test based on data characteristics
- Selection bias, Simpson's paradox

- Where to learn more?
 - Statistics classes

- Next class: causal inference
 - How can we make sure we are measuring the right thing?

Quiz

1. Which of the following is not a consideration for choosing a hypothesis test?

- A. Sample size
- B. Whether the data is paired
- C. The variance of the data (assuming variance is known)
- D. Whether the data is categorical or continuous

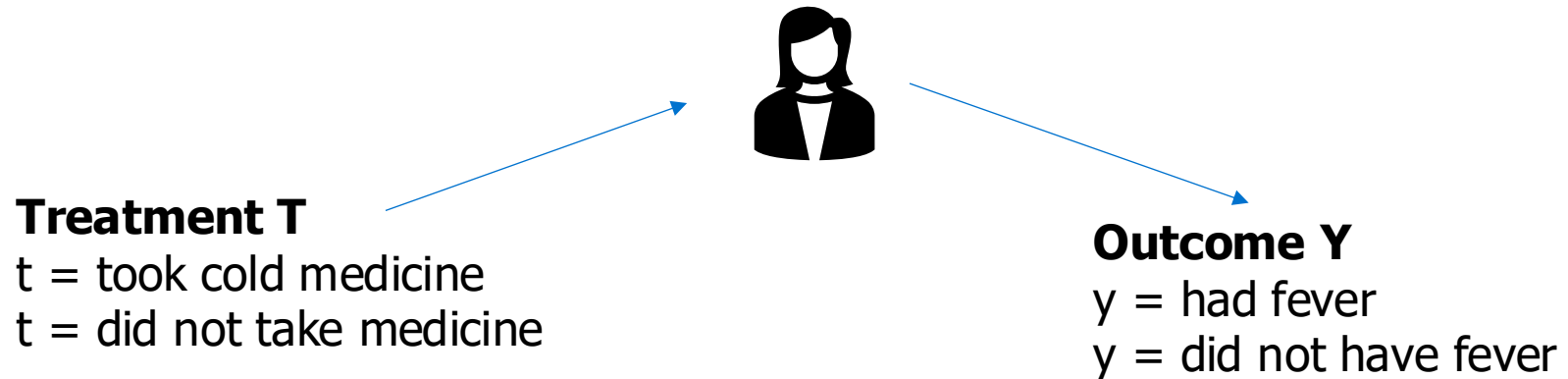
2. In our running example about emotions expressed in tweets about the Black Lives Matter movement, where we see positive emotions increase over time, what would be a hypothetical example of Simpson's paradox?

- A. People who were experiencing sadness chose not to post on social media
- B. People shifted toward using Instagram instead of Twitter over time, and tended to use Instagram for less positive messaging
- C. Model accuracy was higher for tweets earlier in the summer because of their use of popular hashtags

Causal Inference: Definitions

What is causal inference?

- Process of (1) establishing, and (2) quantifying causal relationships empirically (using statistics + data)
- Classic setup: is cold medicine effective?



Some notation: Potential Outcomes

- T: Treatment
- Y: Outcome

- $Y(t)$: The *potential outcome* you would observe under treatment t
 - Outcome that you can *potentially* observe, but that you may not actually observe

- Estimands: causal effects that we want to measure from the data

Causal Estimand: Individual Treatment Effect (ITE)

- For each individual i ,
 - $ITE = Y_i(t = 1) - Y_i(t = 0)$
 - [Outcome had person i taken cold medicine – outcome if they did not]
- ITE is often what we actually care about: should you take cold medicine?
- But we can't measure it!
 - Either you take the medicine or you don't: we can't observe both $Y_i(t = 1)$ and $Y_i(t = 0)$
 - **Fundamental problem of causal inference**
 - Once an outcome is observed, the unobserved outcome is called the *counterfactual*

Causal Estimand: Average Treatment Effect (ATE)

i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

- $ATE = E[Y_i(T = 1) - Y_i(T = 0)] = E[Y(1)] - E[Y(0)]$
- Does $ATE = E[Y | T = 1] - E[Y | T = 0]$?
 - Can we just average over the data in the table, ignoring the missing values?

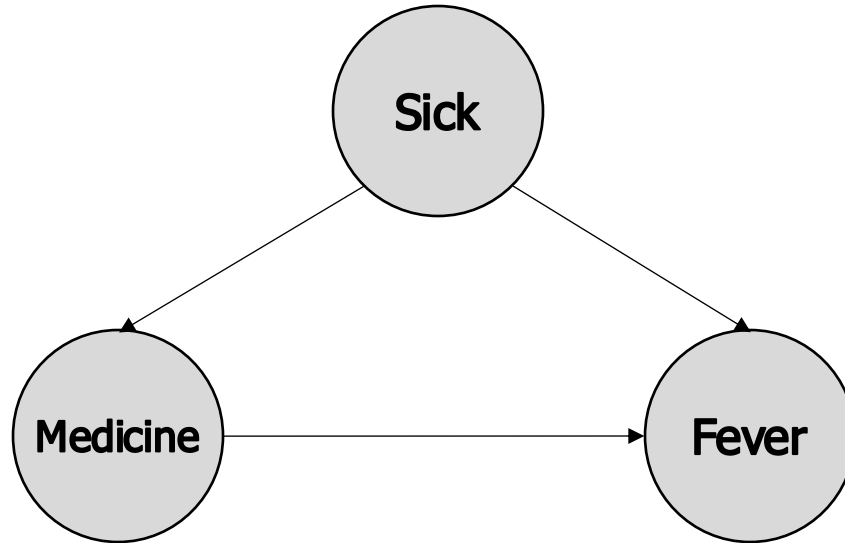
Does $ATE = E[Y | T = 1] - E[Y | T = 0]$?

- Let's pretend we surveyed a bunch of people. We asked them if they took medicine on Sunday and if they had a fever on Monday

		Has Fever (Y)		Sum
		Yes	No	
Took Medicine (T)	Yes	61	43	104
	No	12	80	92
Sum		73	113	186

- $E[Y = fever | T = medicine] - E[Y = fever | T = no medicine]$
- $\frac{61}{104} - \frac{12}{92} = 0.4561$
- Value is positive \rightarrow taking medicine causes fevers?!

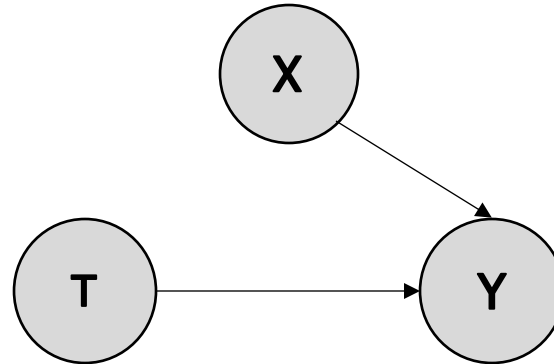
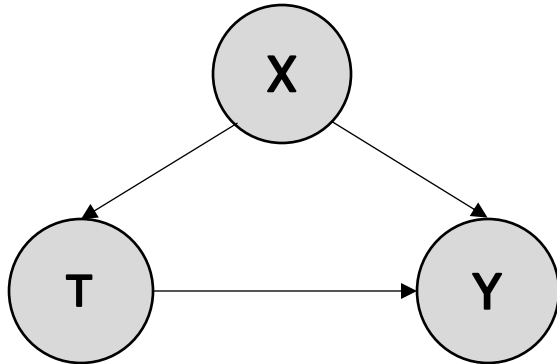
Causal graph and Confounder



- People only took medicine if they were already feeling sick
- **Confounder:** Variable that affects both probability of receiving treatment and outcome

How can we measure ATE without this problem?

- Randomized control trial (RCT)
- More realistic scenario:
 - We'll probably study effects of medicine on someone who is sick
 - If we survey people, there still might be differences: lower income person may not be able to afford medicine and may also have worse nutrition that leads to more severe illness: income is a confounder (X)
- Instead of surveying people, we take a group of people and randomly assign them to "treatment" or "control" group



Randomized Control Trials

- How can we conduct randomized control trials for:
 - Effects of smoking on lung cancer
 - We randomly assign people to smoke or not smoke?
 - Effect of gender on hiring decisions: do men get more higher paying job offers?
 - We randomly assign people genders??
- Many of the questions we may want to study are not possible to investigate through randomized control trials
- We have to rely on ***observational*** data

References and Acknowledgements

- Anjalie Field, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. "Controlled Analyses of Social Biases in Wikipedia Bios" WebConf (2022)
- Chan Young Park*, Xinru Yan*, Anjalie Field*, and Yulia Tsvetkov. "Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia" ICWSM (2021)
- Hypothesis testing slides were inspired by slides from Diyi Yang and David Bamman