DS100: Fall 2018

Lecture 20 (Josh Hug): Hypothesis Testing

- Case Study: United States vs. Kristen Gilbert
- Case Study: Gender Bias in Student Evaluations
- A Quick Note on p-values

Case Study:
United States vs. Kristen Gilbert

From “Statistics in the Courtroom”: http://www.stat.ucla.edu/~nchristo/statistics100B/article.pdf

Defendant Background

Background information on the nurse in our story:

- First nursing job at the VA Clinic in Northampton, WA in 1989.
- Established a reputation of being good in a crisis. When a patient went into cardiac arrest, she would:
  - Sound the code blue alarm.
  - Stay calm.
  - Administer a shot of epinephrine to restart the heart.

Initial suspicions

By the mid 1990s, coworkers had become suspicious of this nurse.

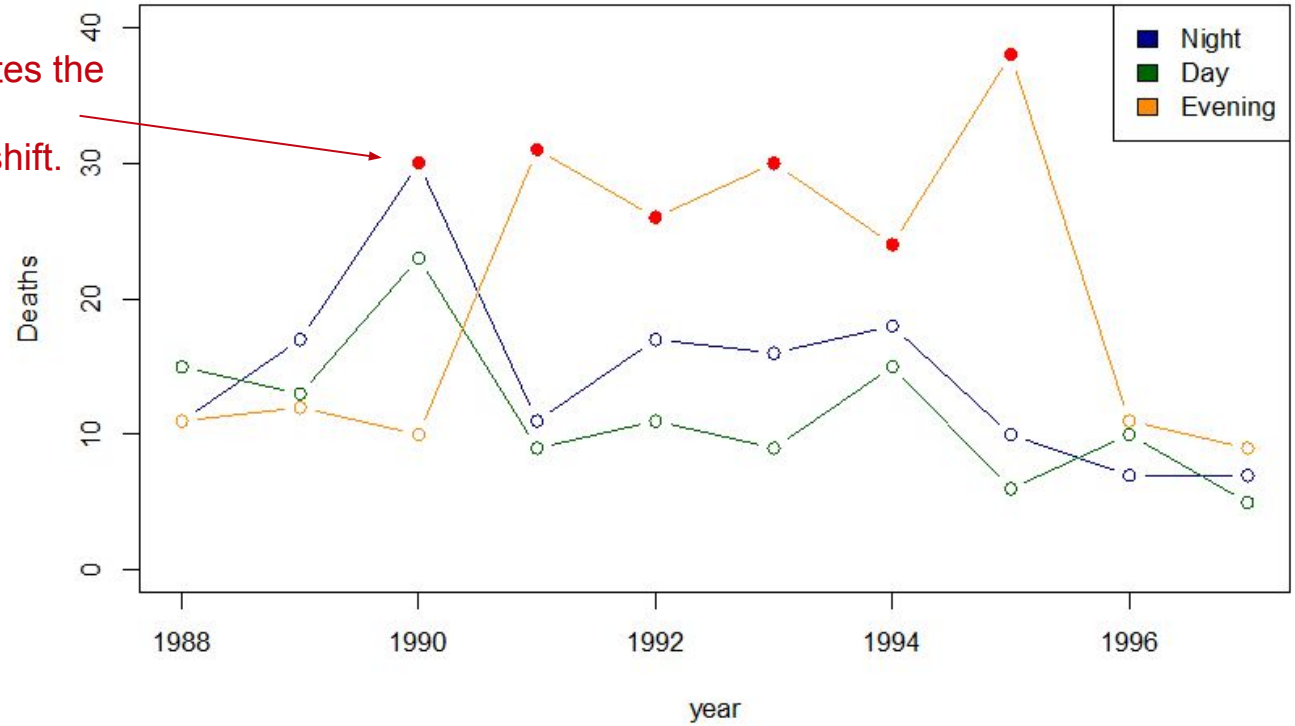
- Too many code blue calls and too many crises when this nurse was on shift.
- Initival VA report found that the number of deaths was consistent with patterns at other VA hospitals.
- Despite this report, suspicions among other staff remained.
- A later report found statistical anomalies, which we'll now review.

Deaths (Coarse Grained Data)

Below are the number of deaths at each of the three shifts at the hospital.

- Example: ~30 total patients died at this hospital during night shifts in 1990.

A red dot indicates the presence of this nurse during a shift.

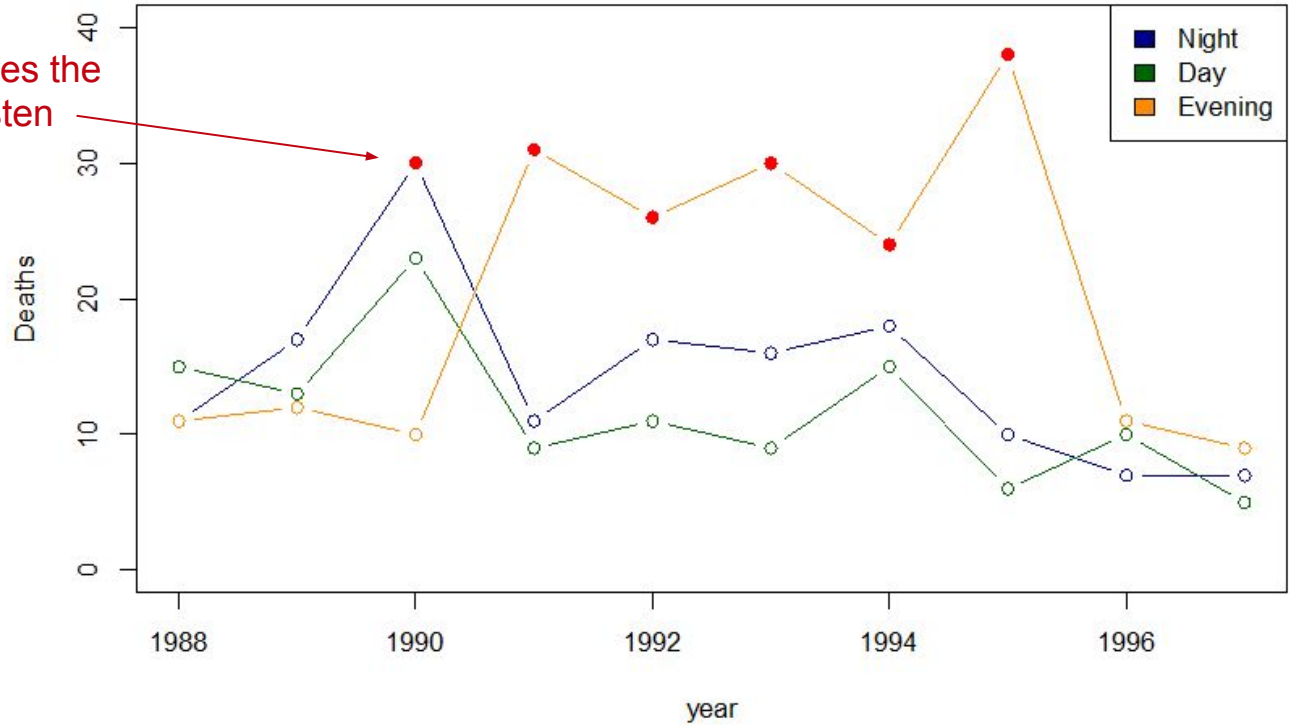


Two Possible Hypotheses

Null hypothesis: The greater number of deaths during nurse's shift was random.

Alternate hypothesis: The greater number of deaths was not random.

A red dot indicates the presence of Kristen during a shift.



Resolving Our Hypotheses: Finer Grained Analysis

Consider the 18 months leading up to February 1996 before the nurse left for medical leave.

- 547 days in this period with 3 shifts each, for a total of 1641 shifts. For each shift, we know:
  - Whether this nurse was working: 0 = no, 1 = yes.
  - Whether a death occurred: 0 = no, 1 = yes.

By looking at joint distribution of these two indicators, we can try to resolve which hypothesis is correct.

- Null hypotheses: Deaths are independent of whether nurse was on shift.
- Alternate hypothesis: Deaths are not independent of whether nurse was on shift.

Finer Grained Analysis

Some key observations:

- Out of 1641 shifts, this nurse worked on 257.
- Out of 1641 shifts, 74 had at least one death.
  - Of these 74, this nurse was working 40.

	DEATH ON SHIFT?		
GILBERT PRESENT?	**Yes**	**No**	**Total**
Yes	**40**	**217**	**257**
No	**34**	**1350**	**1384**
Total	**74**	**1567**	**1641**

Statistical question: If this nurse's presence was independent of deaths, how likely is it that she'd observe deaths on 40 or more shifts?

- Let's call this number IND40.

Finer Grained Analysis

Statistical question: If this nurse's presence was independent of deaths, how likely is it that she'd observe deaths on 40 or more shifts?

- Let's perform a statistical test by doing a computational experiment.

	DEATH ON SHIFT?		
GILBERT PRESENT?	**Yes**	**No**	**Total**
Yes	**40**	**217**	**257**
No	**34**	**1350**	**1384**
Total	**74**	**1567**	**1641**

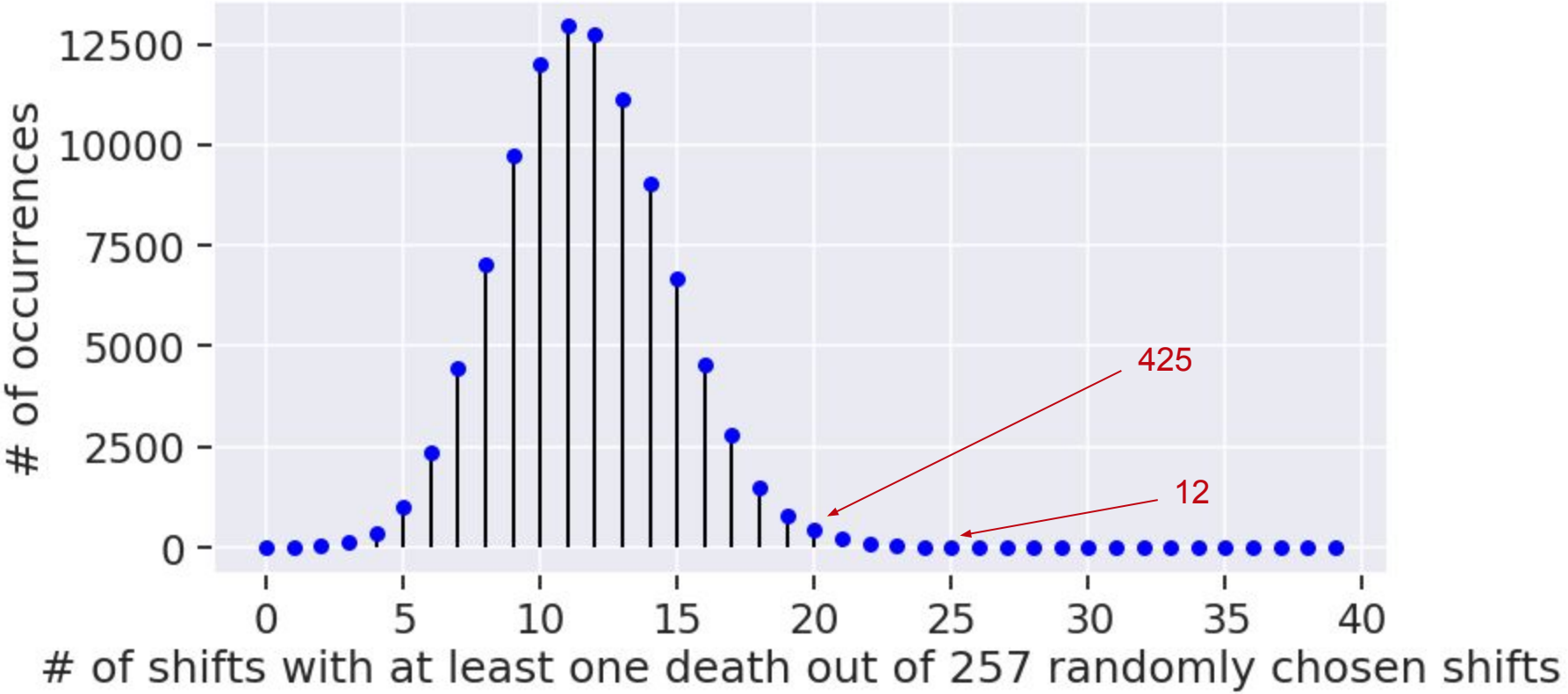
Computational Experiment

Goal: What is the chance that we'd see 40 or more deaths just by pure random chance?

See notebook.

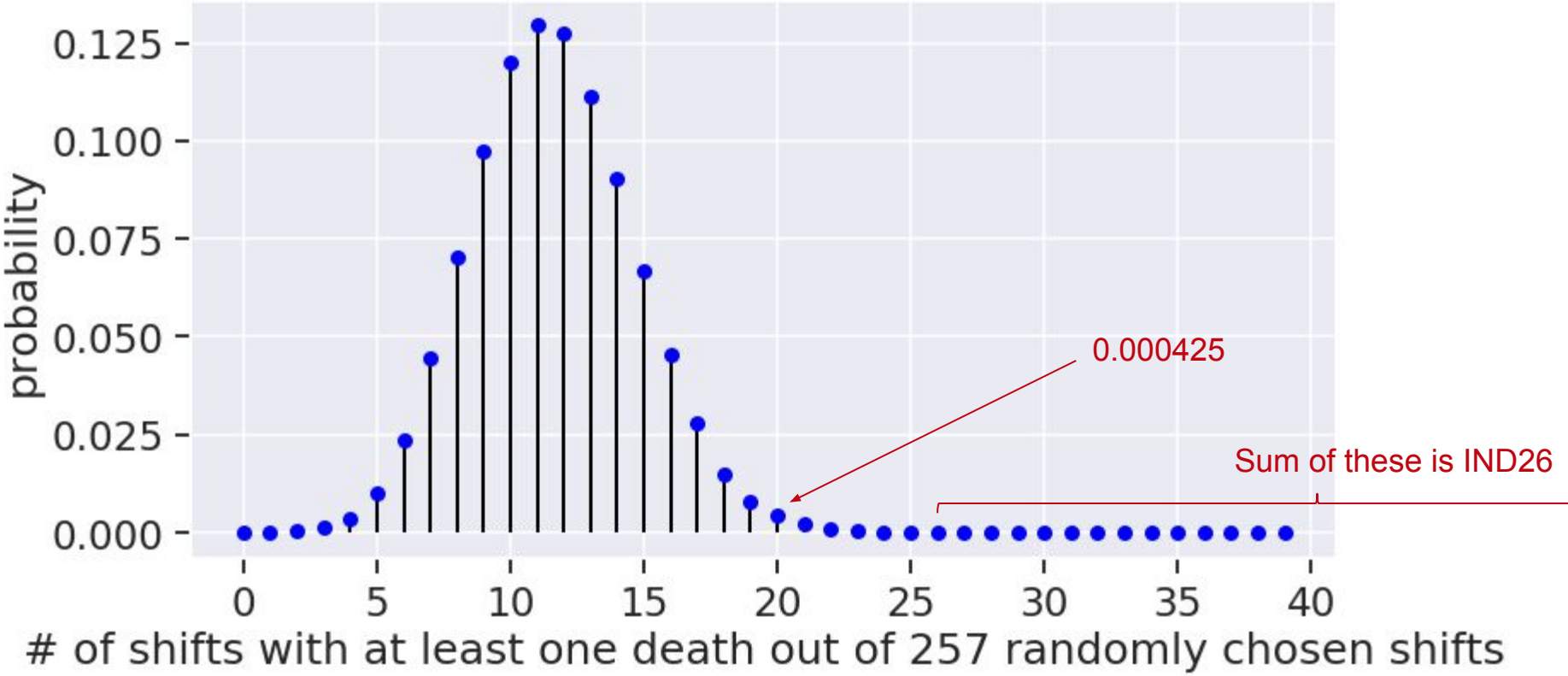
Simulation

Out of 100,000 trials, never see more than 25 shifts with deaths.



Notebook Results

Computationally, it seems IND40 and even IND26 is zero! This is unreasonable.



The Hypergeometric Distribution

This distribution is known as the hypergeometric distribution. Given:

- a population of N=1647 possible shifts,
- where K=74 involved deaths, and
- a sample of n=257 worked shifts,
- where k=40 worked shifts involved deaths.

Then the hypergeometric pmf p(N, K, n, k) = $\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Chance of randomly seeing 40 or more deaths is IND40 = $p_{40}(1674, 74, 257, 40)$ + $p_{41}(1674, 74, 257, 40)$ + … = 9.1121622071405691e-16

- IND 40 is roughly 1 in a trillion!

Data 8 Review [Link]

Reminder: All statistical tests choose between two views of the world:

- **Null hypothesis**: The data was generated by random chance under some set of assumptions, which we call our **null model.**
- **Alternate hypothesis**: Some reason other than chance made the data differ from the predictions of the model.

To resolve which hypothesis is correct we will compute a **test statistic.**

- Our test statistic is the number of deaths observed.

The **p-value** is the chance, under the null model, that the test statistic is equal to or worse than our observation.

- In other words, p-value is the chance that we observe 40 deaths or more.

Question 1: True or False yellkey.com/rule

Our computed p-value was 9.1121622071405691e-16.

True or false: The nurse was not assigned to her shifts at random so our computation of the p-value is flawed.

Question 1: True or False

Our computed p-value was 9.1121622071405691e-16.

True or false: The nurse was not assigned to her shifts at random so our computation of the p-value is flawed.

False: The p-value has nothing to do with the actual process that generated the data. It is the probability that the null (a.k.a. random) model could generate the data (or worse).

Question 2-6: True or False. yellkey.com/practice

Questions to ponder:

- Q2: The chance that the nurse is innocent is approximately 1 in a quadrillion.
- Q3: The chance that the nurse is guilty is approximately 999,999,999,999,999 out of a quadrillion.
- Q4: Our p-value is so small that we're sure the nurse's shifts are not independent of whether or not there was a death. However, that doesn't prove she's guilty. There could be other explanations.
- Q5: The fact that our p-value is so small is a good reason alone to bring this nurse to trial.
- Q6: The fact that our p-value is so small is a good reasons alone to convict this nurse of murder.

Question 2-6: True or False

Questions to ponder:

- Q2: The chance that the nurse is innocent is approximately 1 in a trillion.
  - False!
- Q3: The chance that the nurse is guilty is approximately 999,999,999,999 out of a trillion.
  - False!

This is the “prosecutor’s fallacy”. It is a **very common mistake** in statistical reasoning.

- Yes, the small p-value means that **random variation is not the cause** of the excess deaths on the nurse’s shift. **Test statistic is inconsistent with null hypothesis.**
- However, **we have said nothing about other possible causes** (the nurse was framed, the nurse was host to some sort of pathogen, etc.).

Question 2-6: True or False

Questions to ponder:

- Q4: Our p-value is so small that we're sure the nurse's shifts are not independent of whether or not there was a death. However, that doesn't prove she's guilty. There could be other explanations.
  - True (see previous slide).
- Q5: The fact that our p-value is so small is a good reason alone to bring this nurse to trial.
  - Arguably true: It does mean further investigation is needed!
- Q6: The fact that our p-value is so small is a good reasons alone to convict this nurse of murder.
  - False! As mentioned above, we have ONLY shown that the excess deaths are not due to random chance.

What Happened in Real Life

A “grand jury” (decides on whether a trial should occur) was presented with the statistical evidence by expert witnesses hired by the prosecutors.

- Grand jury decision: Send the nurse was sent to trial.

During trial, defense attorneys had the prosecutors’ expert statistical witness explain why the statistical evidence was insufficient (much like we have today).

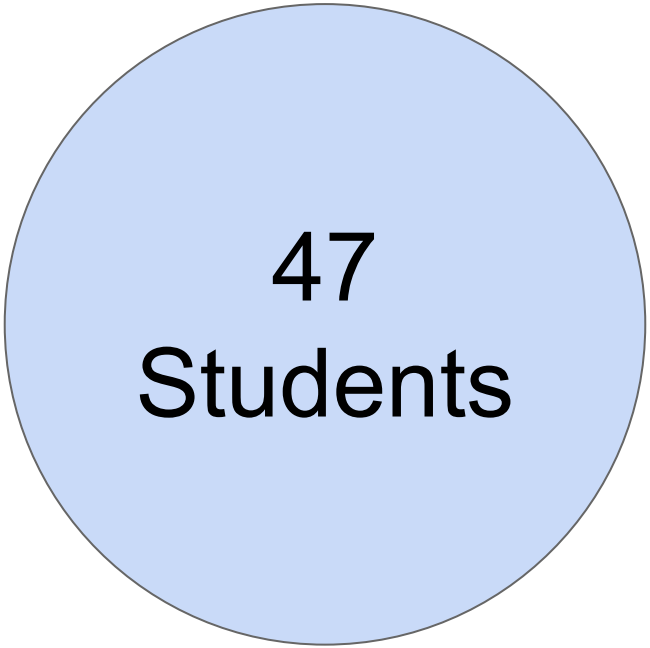
- Statistical evidence was therefore not considered in determining guilt.
- Jury decision: Nurse was guilty of 3 murders and 2 attempted murders based on specific non-statistical evidence, sentenced to life imprisonment by the jury.

Case Study: Gender Bias in Student Evaluations

From Boring, Ottoboni, and Stark (2016) “Student evaluations of teaching (mostly) do not measure teaching effectiveness” [Link] and Macnell, Driscoll, Hunt (2015): “What’s in a Name: Exposing Gender Bias in Student Ratings of Teaching” [Link]

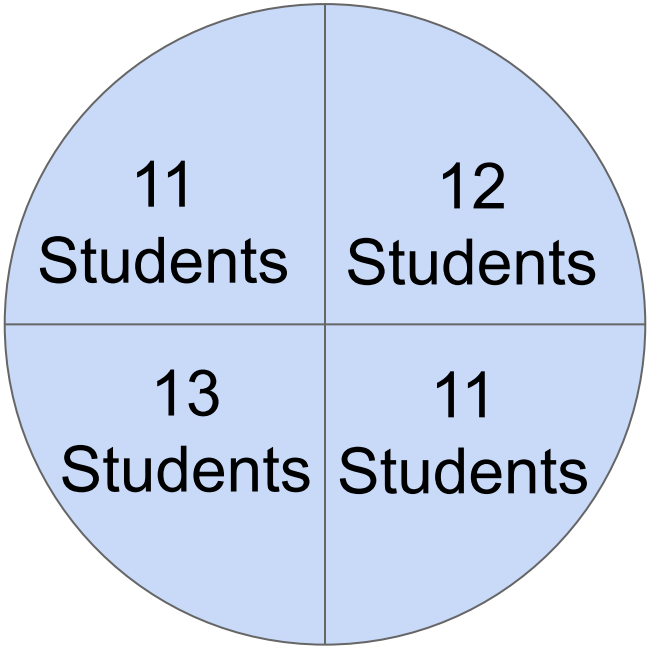
Background

47 students were enrolled in a purely online class.



Background

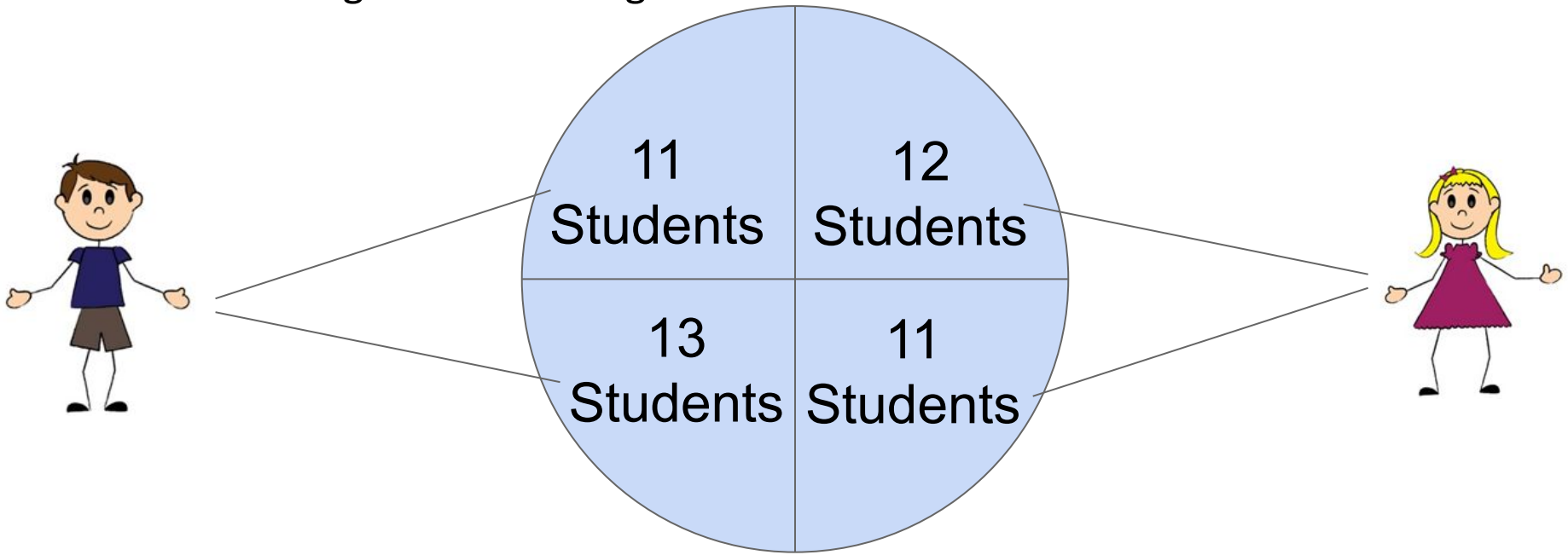
Students were randomly assigned to one of 4 sections.



Background

Students were randomly assigned to one of 4 sections.

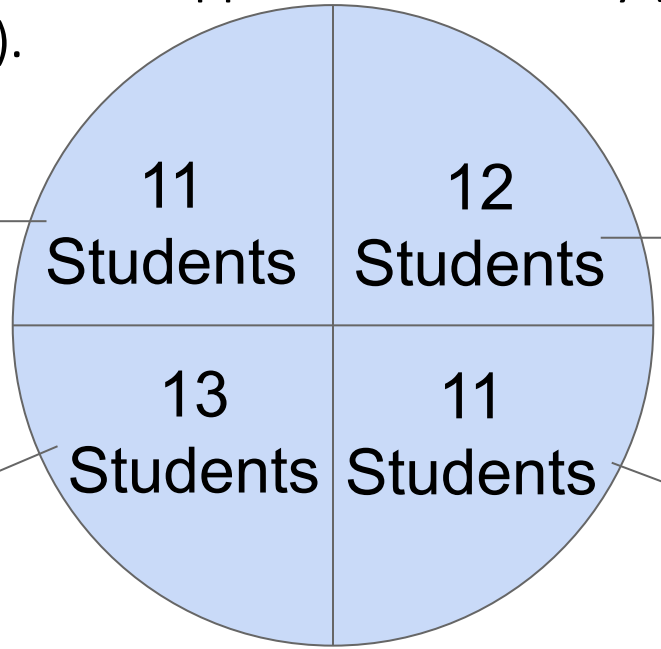
- Two teaching assistants taught two sections each.

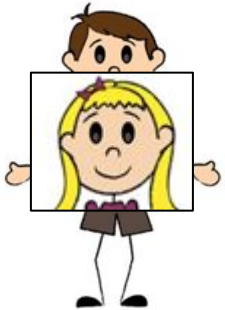


Background

Students were randomly assigned to one of 4 sections taught by 2 TAs.

- For one section, TA used the opposite TA's identity (all interactions were text based, no video!).





Student Evaluations of Teaching (SETs)

Students provided evaluations of their TAs on many dimensions.

- Such evaluations are sometimes called “Student Evaluations of Teaching” or SETs.
- We’ll focus on homework return promptness.
  - If you return the homework promptly (somewhat unlike DS100), then you will get a high score in promptness.

All HW grading was done together, and all HWs returned at the exact same time.

Observed Data

	group	tagender	taidgender	prompt
0	3	0	1	4.0
1	3	0	1	5.0
2	3	0	1	5.0
3	3	0	1	5.0
4	3	0	1	3.0
5	3	0	1	4.0
6	3	0	1	4.0
7	3	0	1	5.0
8	3	0	1	4.0
9	3	0	1	4.0
10	3	0	1	5.0

`group`: Which section the student belonged to.

`tagender`: The true gender of the TA.

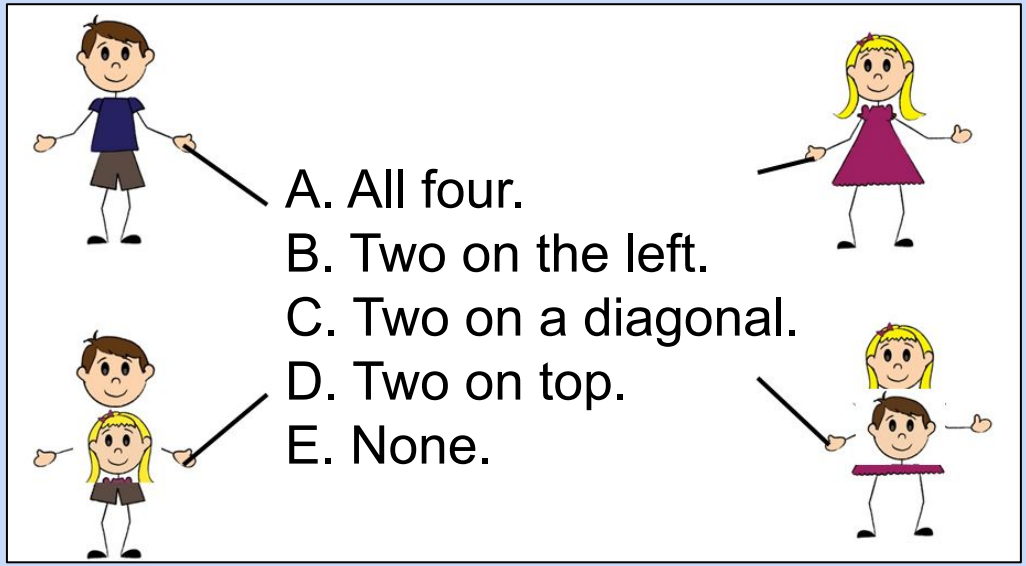
`taidgender`: The gender presented to students.

`prompt`: Rating on promptness of HW on a scale from 1 to 5.

Student Evaluations yellkey.com/drug

All HW grading was done together, and all HWs returned at the exact same time.

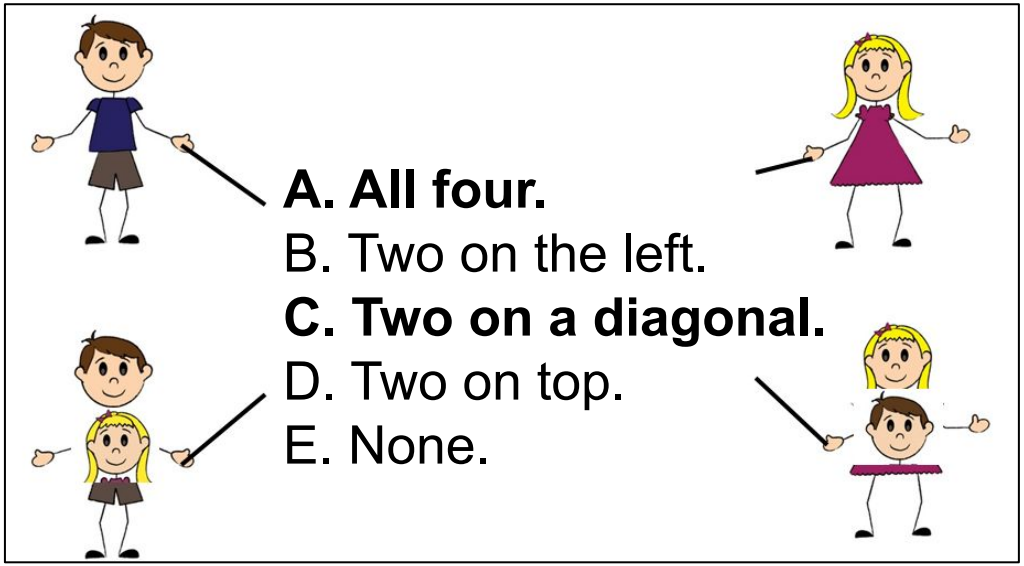
Which evaluations do you expect to be the same?



Student Evaluations

All HW grading was done together, and all HWs returned at the exact same time.

Which evaluations do you expect to be the same?



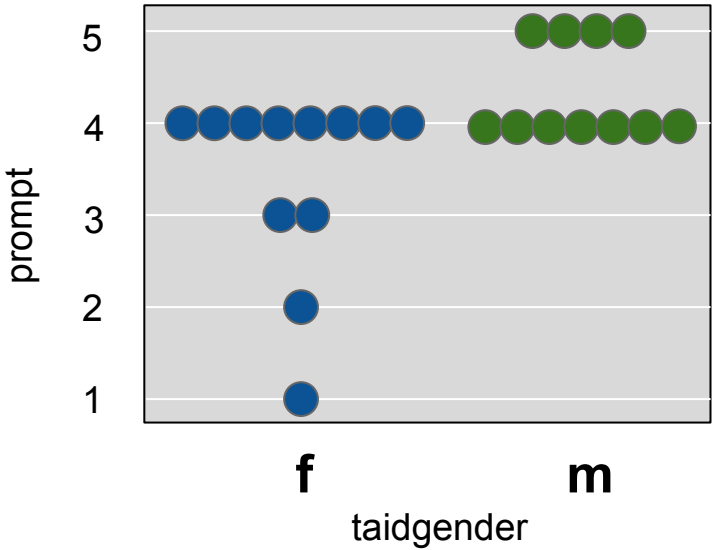
Student Evaluations

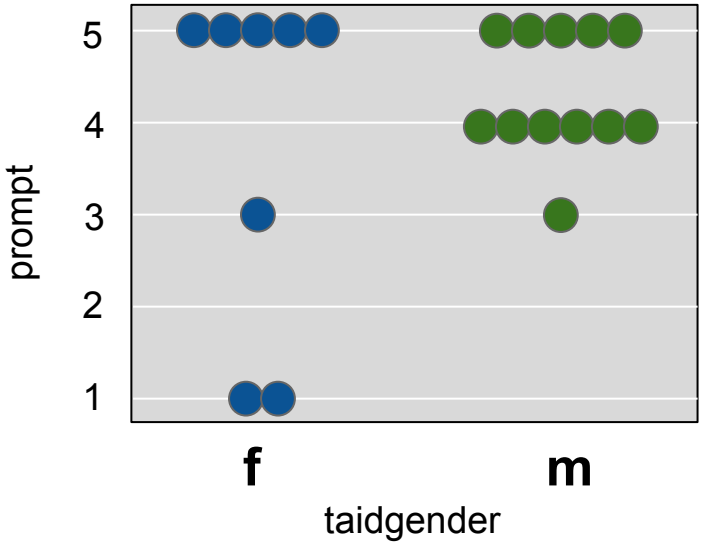
All HW grading was done together, and all HWs returned at the exact same time.

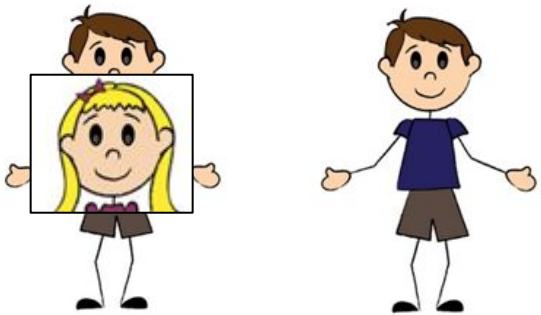
- You'd expect in a world without gender bias, that students would give the same average promptness rating in all four sections.
- You'd expect in a world with gender bias that the diagonals would be equal.

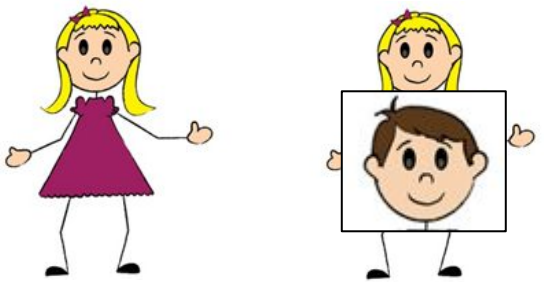
This top answer (all equal) is not what happened.

Observed Data Visualized







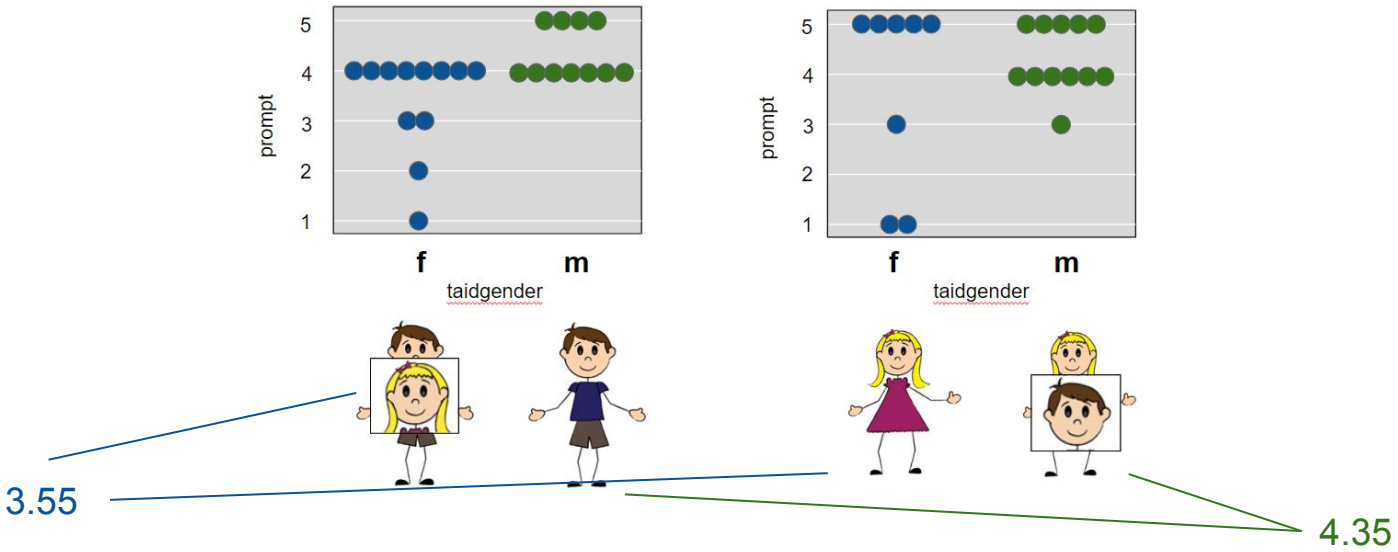


Student Evaluations

Average for instructor identified as male: (9*5 + 13*4 + 1*3)/23 = 4.35

Average for instructor identified as female: (5*5 + 8*4 + 3*3 + 1*2 + 3*1)/20 = 3.55

Could this have simply happened by chance?



Student Evaluations

Average for instructor identified as male: (9*5 + 13*4 + 1*3)/23 = 4.35

Average for instructor identified as female: (5*5 + 8*4 + 3*3 + 1*2 + 3*1)/20 = 3.55

Could this have simply happened by chance?

- Null hypothesis: The gap exists simply due to random chance.
- Alternate hypothesis: The gap exists due to gender perception bias.
  - Our alternate hypothesis here is stronger, because we conducted a randomized controlled trial where the ONLY difference is the perceived gender.

Can compute a p-value that the null hypothesis is true… but how?

Student Evaluations

Average for instructor identified as male: (9*5 + 13*4 + 1*3)/23 = 4.35

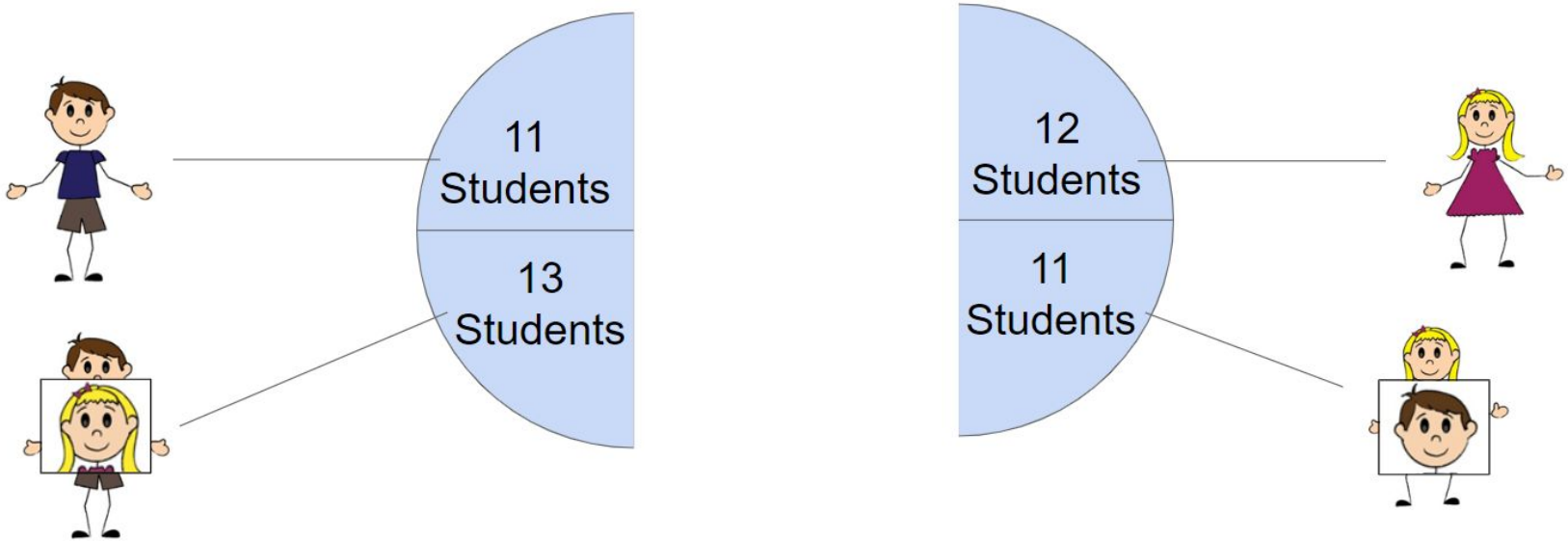
Average for instructor identified as female: (5*5 + 8*4 + 3*3 + 1*2 + 3*1)/20 = 3.55

Could this have simply happened by chance?

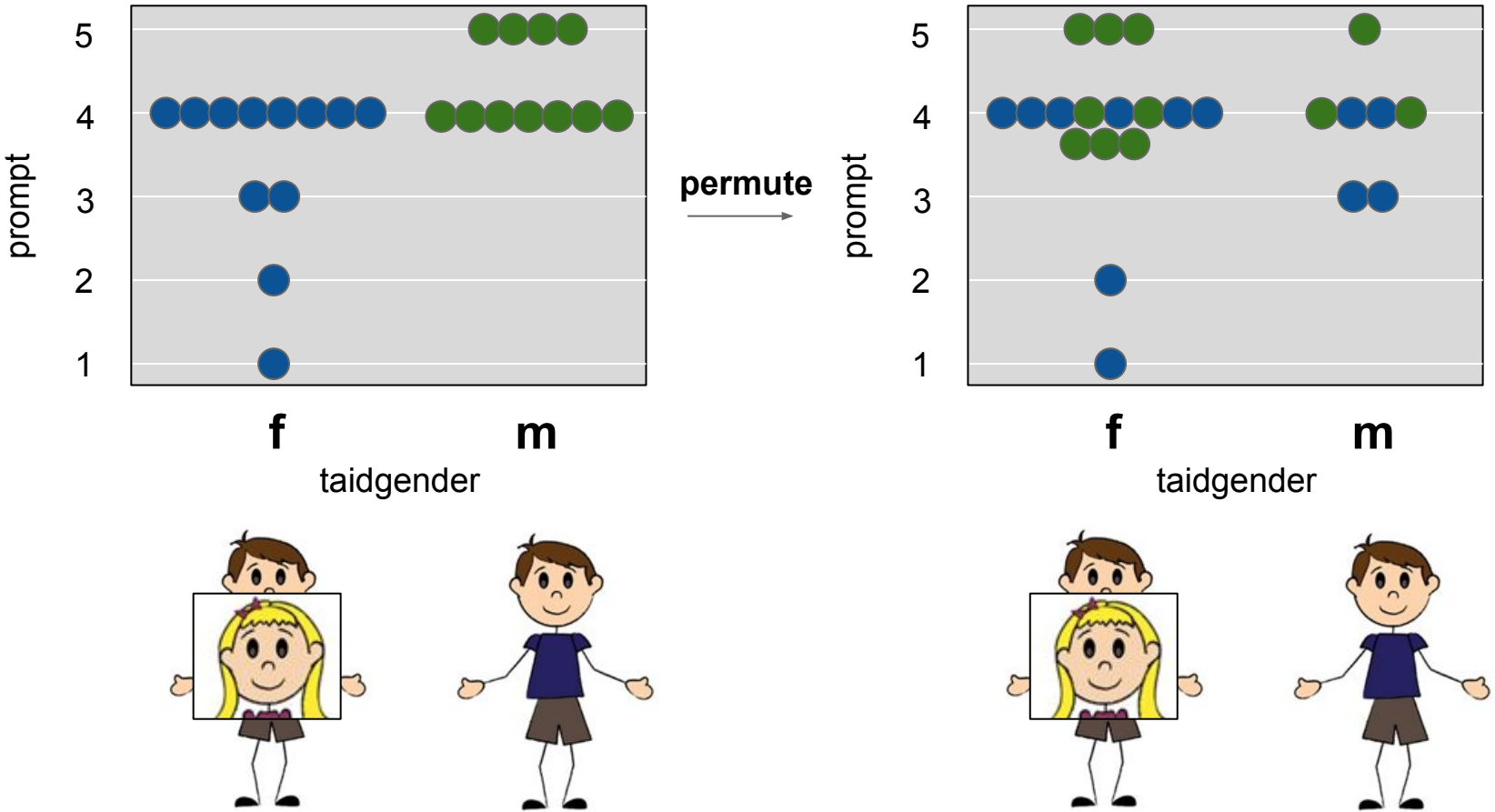
- p-value calculation in paper by Macnell, Driscoll, Hunt involves assumptions about how students make ratings, e.g. that student evaluation ratings are drawn from a gaussian.
  - Such assumptions are hard to defend, so p-values may not be meaningful. Thus, we will not discuss the details.
- p-value calculation in paper by Boring, Ottoboni, Stark makes weaker assumptions by using a procedure known as “permutation testing”.

Permutation Testing

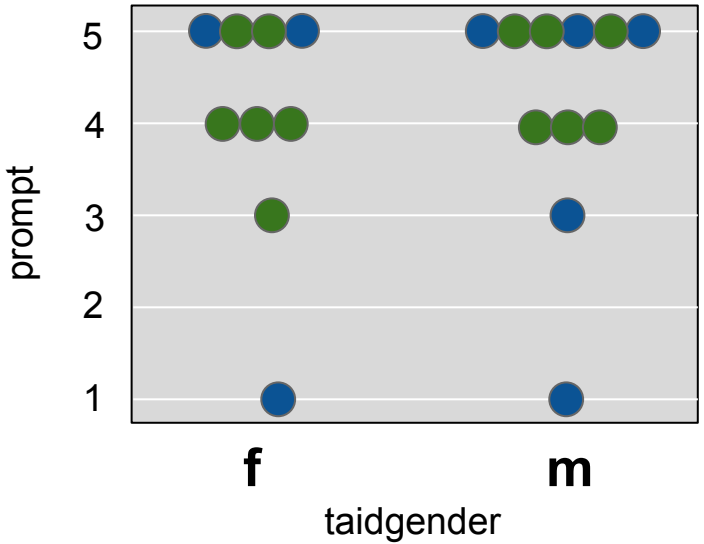
Basic Idea: If random chance caused the difference, then if we randomly shuffle all the genderid=male TA's students, AND also randomly shuffle all the genderid=female TA's students, then the gap of 0.8 should not be unlikely.



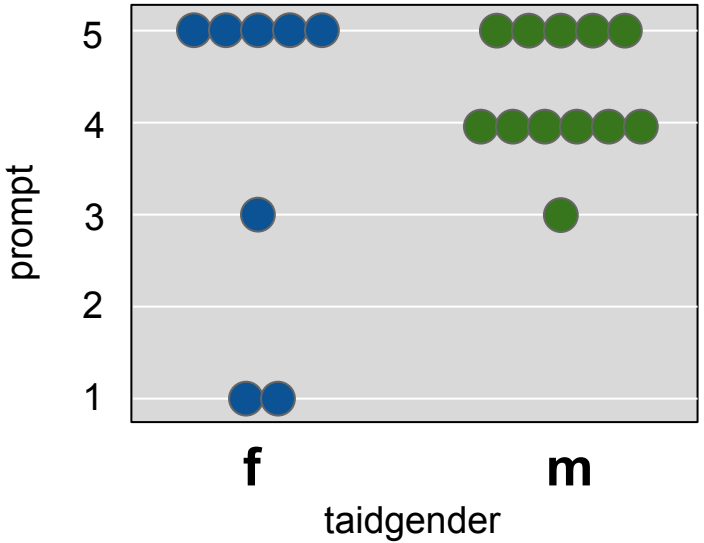
Example Permutation of the Left (tagender=male) TA Side

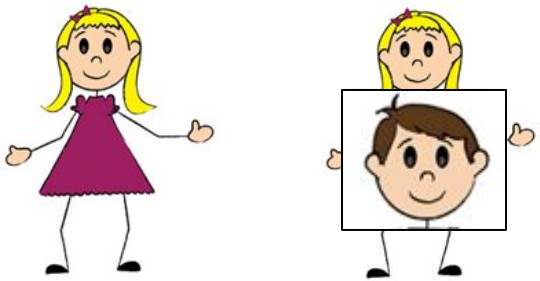


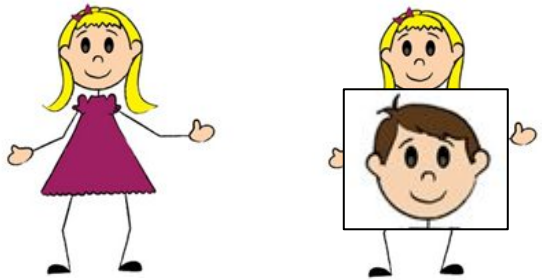
Example Permutation of the Right (tagender=female) TA Side



permute



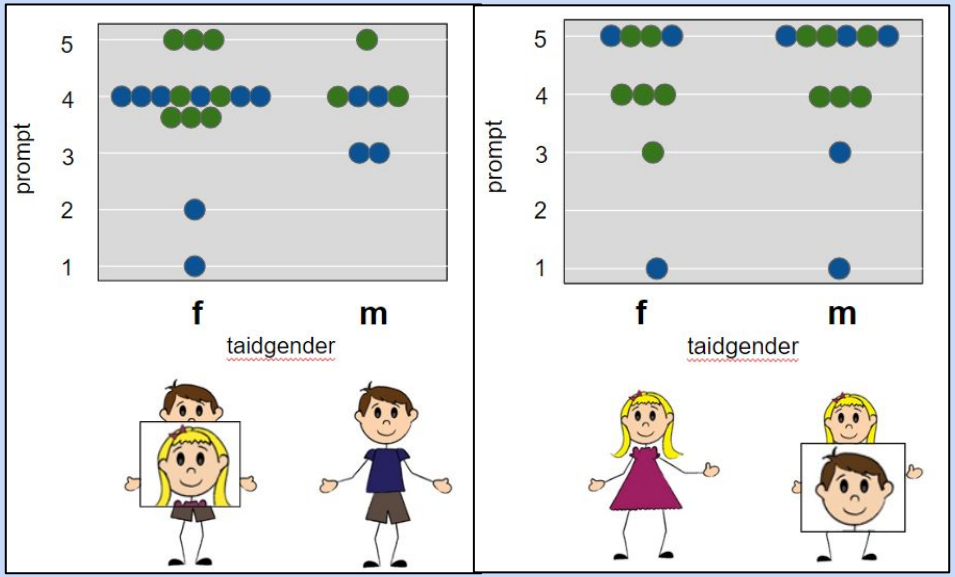




Gap Computation for Permutation Example yellkey.com/long

Average for instructor identified as male: ??? [write an expression]

Average for instructor identified as female:

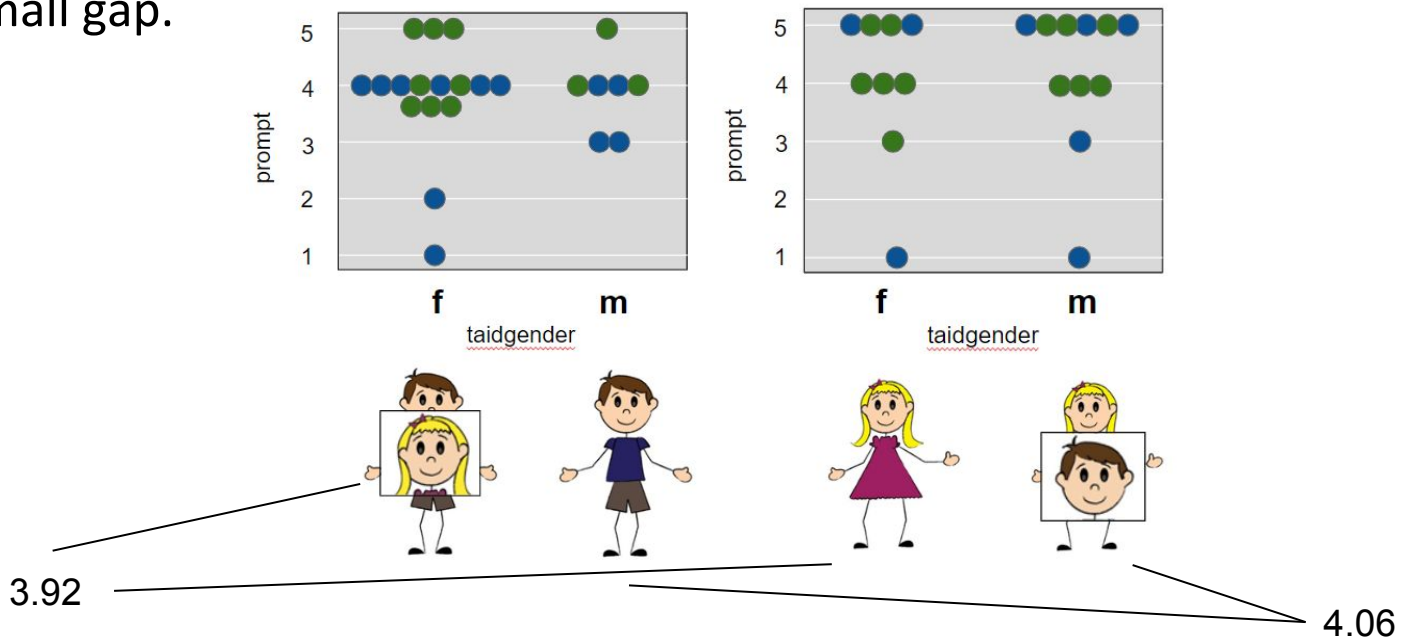


Gap Computation for Permutation Example

Average for instructor identified as male: (7*5 + 7*4 + 3*3 + 1*1)/17 = 4.06

Average for instructor identified as female: (7*5 + 14*4 + 1*3 + 1*2+2*1)/25 = 3.92

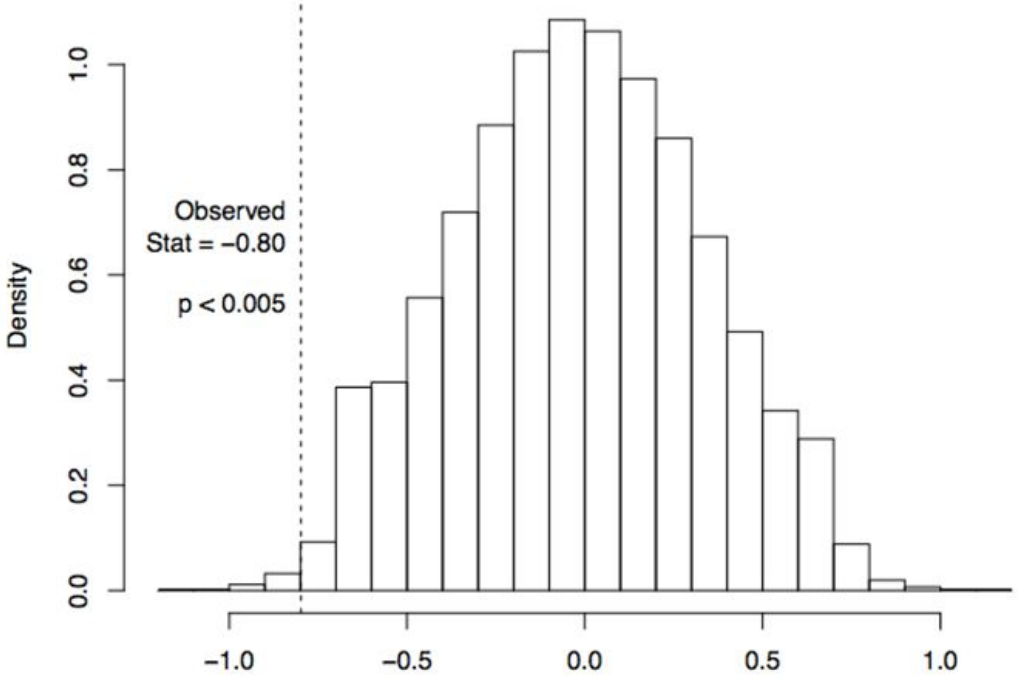
Gap is only -0.14 for this permutation! Random assignment of perceived gender yielded small gap.



Repeated Experiment Result

Result after many simulations:

- Fewer than 5 in 1000 permutations yielded a male female difference of <= -0.8.
- Our p-value is therefore 0.005. Null hypothesis is unlikely!



Correlation and Causation

Correlation and causation:

- Because this was a proper randomized trial (unlike the nurse case), and the ONLY difference was the identified gender of the TA, then we are fairly certain that the perception of gender is causing students to evaluate the TA differently.
- Note: Based on our permutation testing, even if the gender bias hypothesis is false, we'd expect 5 out of 1000 experiment to show results supporting the hypothesis! It is possible that the gender bias hypothesis is false, and our experiment just happened to be unlucky.

Caveat: Exact details of how gender was revealed were unclear to me. If it was based solely on a name, perhaps it was a judgment of the TA's name instead?

Shewchuk vs. Yelick, Fall 1998

Looking at past course survey evaluations for 61B, I noticed a strange anomaly:

- Course was taught by two instructors: Jonathan Shewchuk and Kathy Yelick.
- Same exams, same hws, same projects.

However, Students rated Jonathan's class as a 3.7/5 in difficulty, but Kathy's as a 4.4/5. Anecdotally, this is a huge difference.

- I'd be curious if a similar phenomenon played a role here. No way to really know, though.

Boring, Ottoboni, and Stark (2016)

The Boring, Ottoboni, and Stark paper was a re-analysis of a experimental data collected by Macnell, Driscoll, and Hunt.

- M/D/H paper data was posted publicly for other researchers to evaluate. This is great, because it lets other researchers re-evaluate and expand upon their results.
- Wonderful trend in modern science!

Boring, Ottoboni, and Stark (2016)

From Methods section:

- Previous analyses of these data relied on parametric tests **based on null hypotheses that do not match the experimental design**. For example, the tests assumed that SET of male and female instructors are independent random samples from normally distributed populations with equal variances and possibly different means. **As a result, the p-values reported in those studies are for unrealistic null hypotheses and might be misleading**.
- In contrast, **we use permutation tests** based on the as-if random (French natural experiment) or truly random (US experiment) assignment of students to class sections, **with no counterfactual assumption that the students, SET scores, grades, or any other variables comprise random samples from any populations, much less populations with normal distributions**.

Materials by Boring, Ottoboni, and Stark

Relevant repositories on GitHub.

Code and data for paper:

- https://github.com/kellieotto/SET-and-Gender-Bias

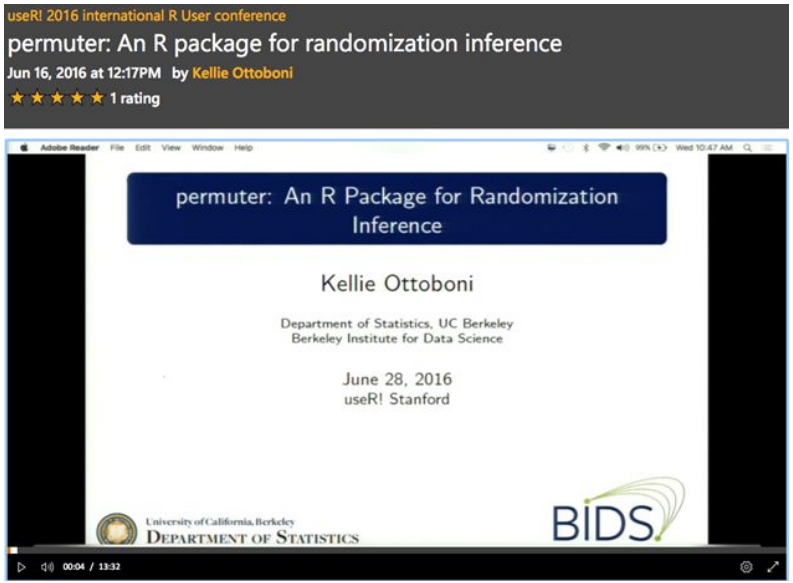
Permute – Python Library:

- https://github.com/statlab/permute

Permuter – R Library:

- https://github.com/statlab/permuter

Presentation video @ useR! 2016 conference



https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016/Wednesday-June-29--1048am-1106am

A Quick Story About p-values

Why Is p < 0.05 A Common Standard?

> “It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not.”
> — Sir Ronald Fisher, Statistical Methods for Research Workers, 1925

> “If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author prefers to set a low standard of significance at the 5 percent point …”.”
> — Sir Ronald Fisher, 1926

p-values Are A Hot Topic

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

- Q: Why do so many colleges and grad schools teach p = 0.05?
  A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use p = 0.05?
  A: Because that's what they were taught in college or grad school.

Interesting links:

- The ASA's Statement on p-Values: Context, Process, and Purpose [Link].
- The Story Behind the ASA Statement on P-Values [Link].

p-values Are A Hot Topic

Redefine statistical significance

- Benjamin et al. (72 authors), Nature Human Behaviour 2, 6–10(2018), doi:10.1038/s41562-017-0189-z
  - “We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.”

The Proposal to Lower P Value Thresholds to .005

- John P. A. Ioannidis, MD, DSc, JAMA. 2018;319(14):1429-1430. doi:10.1001/jama.2018.1536, April 10, 2018

Abandon Statistical Significance

- B. McShane, D. Gal, A. Gelman, C. Robert, J. L. Tackett, https://arxiv.org/abs/1709.07588, April 10, 2018

Deleted Slides

The Lottery

A few days ago, the numbers 5, 28, 62, 65, 70 - 5 were worth a $1.5 billion jackpot.

- Odds of winning were 1 in 302,575,350.

One could argue that the chance of winning by random selection is so low that the winner must have cheated.

- Why is this wrong?