Statistical biases

Introduction to data analysis: Lecture 9

Ori Plonsky

Spring 2023

"I can't understand why the whole audience hated my Simpson's Paradox joke. I tried it on the men and the women in the crowd separately and each group loved it!"

Case study: Berkeley admissions

- In the 70's, a study was conducted to evaluate whether there was gender-based discrimination in admissions to the university.
- The data is from six departments, called A-F.
- We have information on whether the applicant was male or female and whether they were admitted or rejected.
- We will examine whether there is discrimination against females in admission.

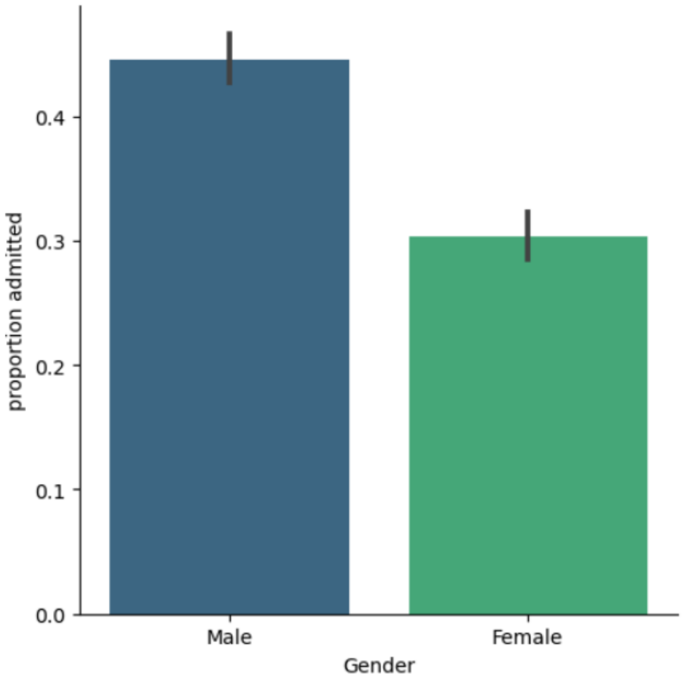
Is there gender discrimination?

- What is our null hypothesis?
- What is the alternative hypothesis?
- How can we test our hypothesis?

[notebook]

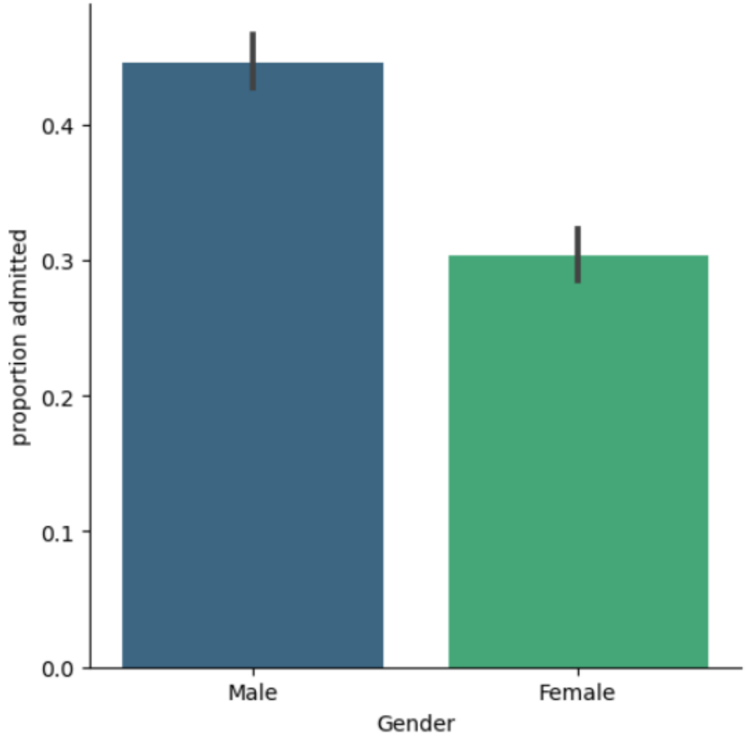
What happened in our analysis?

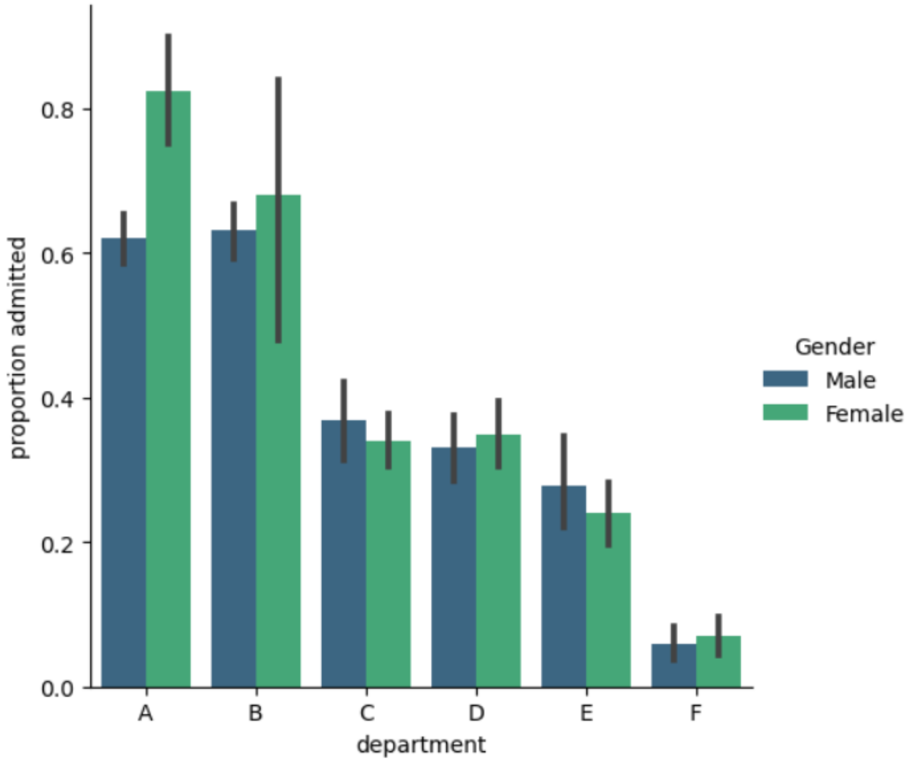
- Overall, it seemed like admission rates are lower for women
- The 95% confidence interval for the difference in proportion of males accepted and proportion of female accepted is [0.11, 0.17]
- We reject the null hypothesis and argue that the proportions are not equal



What happened in our analysis?

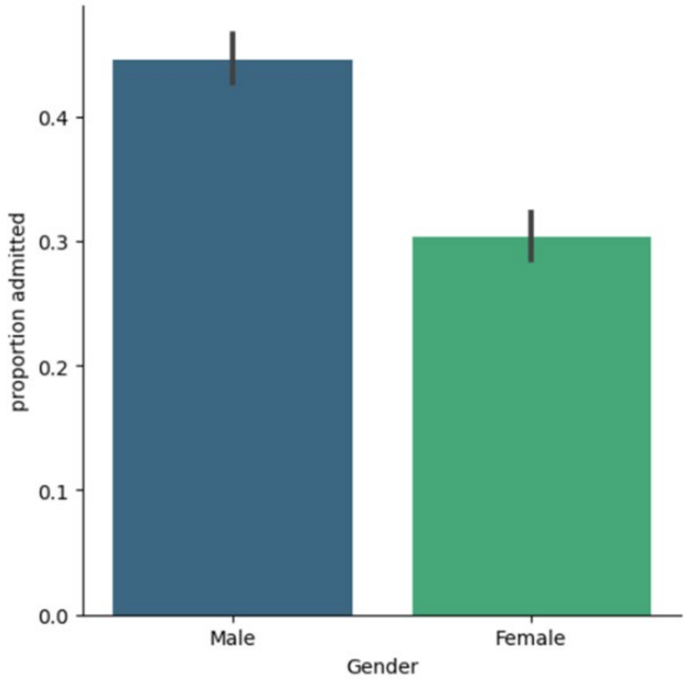
- Overall, it seemed like admission rates are lower for females
- But for each department separately admission rates are the same or even higher for females???



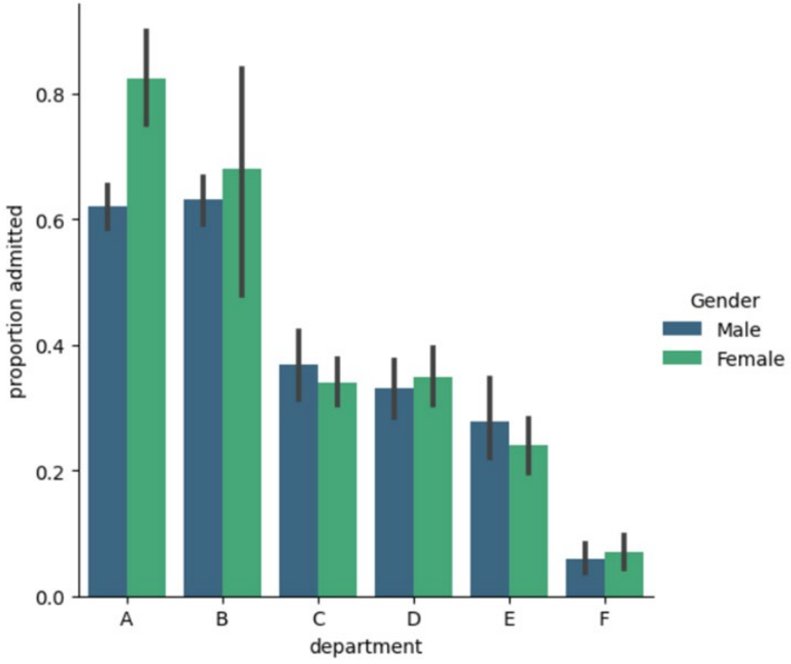


What happened in our analysis?

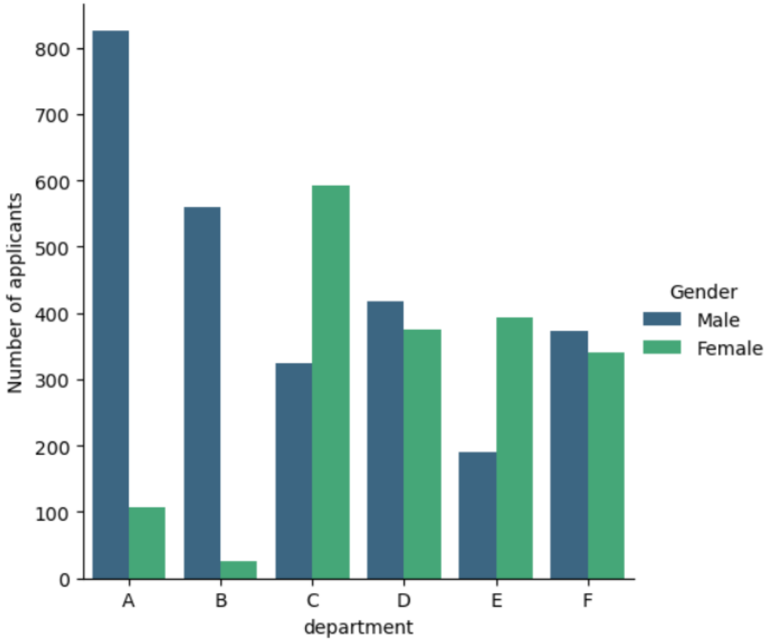
- Overall, it seemed like admission rates are lower for females
- But for each department separately admission rates are the same or even higher for females???



Average admission rate by gender



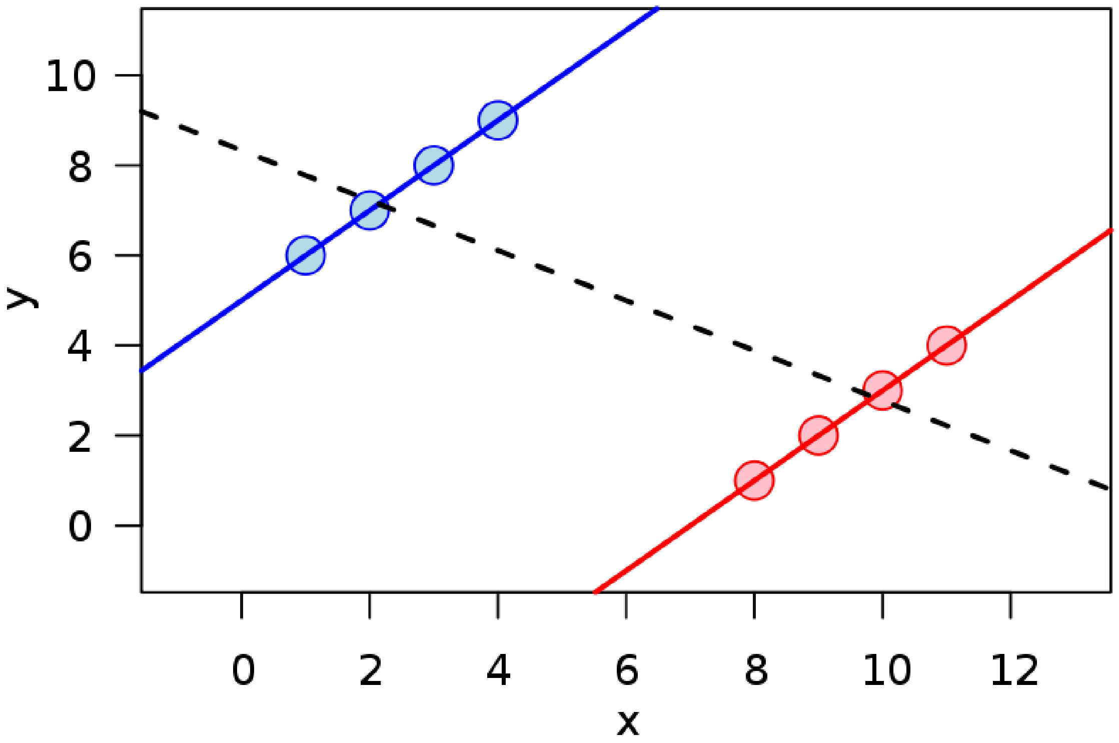
Average admission rate by gender and department



Number of applicants by gender and department

Simpson’s Paradox

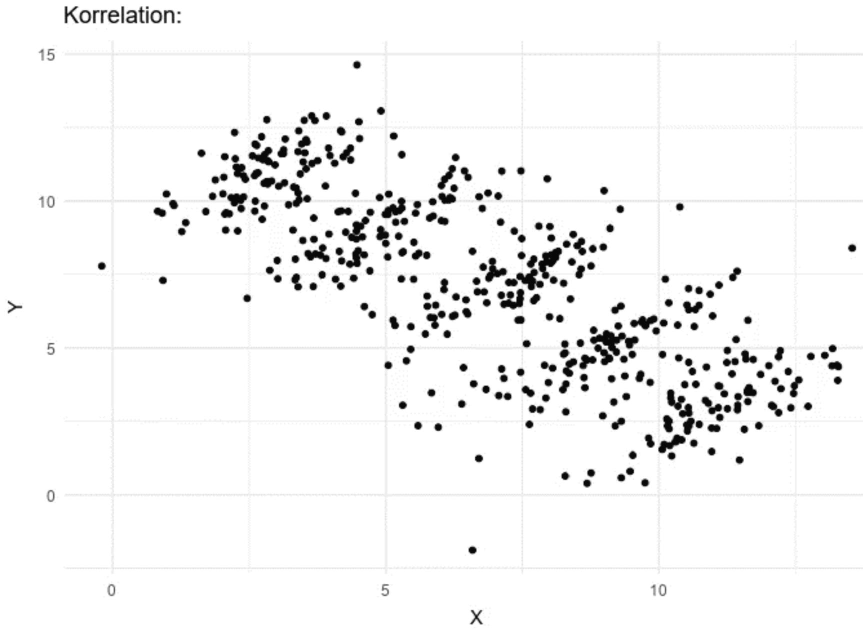
- Not considering an important variable when studying a relationship can result in what we call a Simpson's paradox



Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

Simpson’s Paradox

- Not considering an important variable when studying a relationship can result in what we call a Simpson's paradox



Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

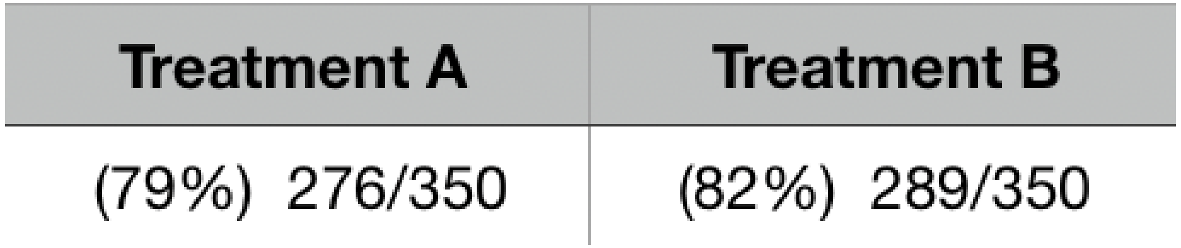
Another example: kidney stones

- There are two treatments for kidney stones: A and B
- 350 patients got Treatment A and 350 patients got Treatment B
- Number of patients who got better:

Source: https://towardsdatascience.com/solving-simpsons-paradox-e85433c68d03

Another example: kidney stones

- There are two treatments for kidney stones: A and B
- 350 patients got Treatment A and 350 patients got Treatment B
- Number of patients who got better:
    - Overall, looks like B was somewhat more effective



Treatment A	Treatment B
(79%) 276/350	(82%) 289/350

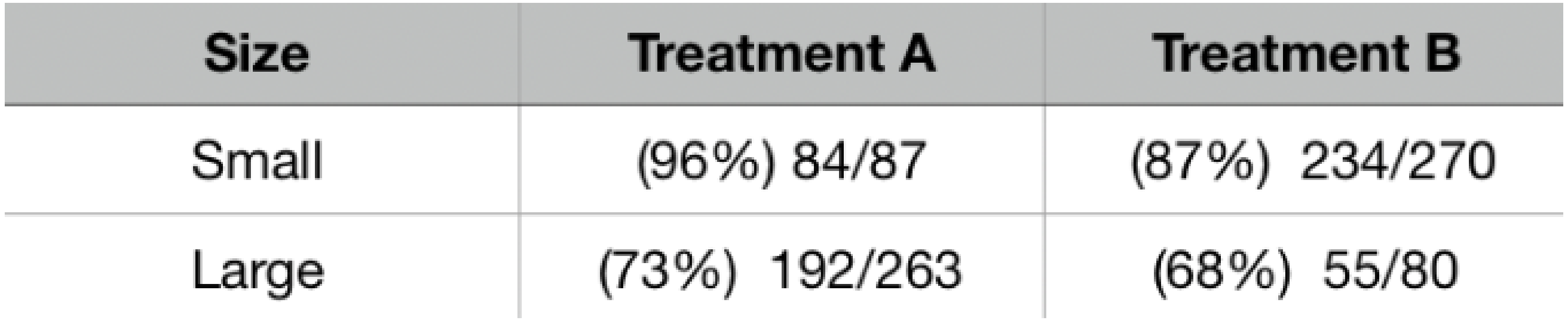
Source: https://towardsdatascience.com/solving-simpsons-paradox-e85433c68d03

Another example: kidney stones

- What happens if we look by the type of kidney stones?
  - Specifically, kidney stones can be large or small

Another example: kidney stones

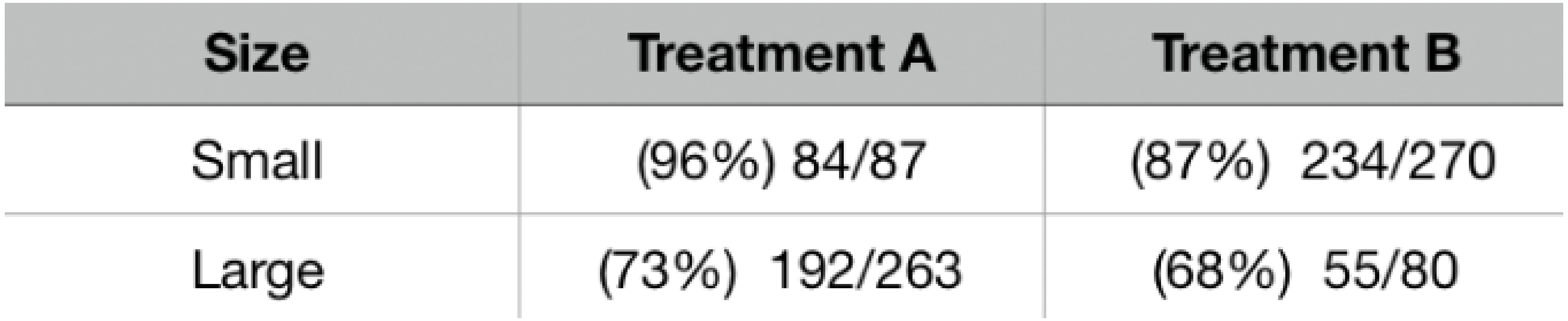
- What happens if we look by the type of kidney stones?
  - Specifically, kidney stones can be large or small
- Treatment A is open surgery; Treatment B is minimally invasive



Size	Treatment A	Treatment B
Small	(96%) 84/87	(87%) 234/270
Large	(73%) 192/263	(68%) 55/80

Another example: kidney stones

- What happens if we look by the type of kidney stones?
  - Specifically, kidney stones can be large or small
- Treatment A is open surgery; Treatment B is minimally invasive
- Patients were not assigned randomly to these treatments!



Size	Treatment A	Treatment B
Small	(96%) 84/87	(87%) 234/270
Large	(73%) 192/263	(68%) 55/80

Case study: Purdue class size

- Students complain that the classes are too big
- Dean: mean class size is ~35 students.
- Student Body: No, it is ~90
- What’s going on?

[notebook]

Case study: Purdue class size

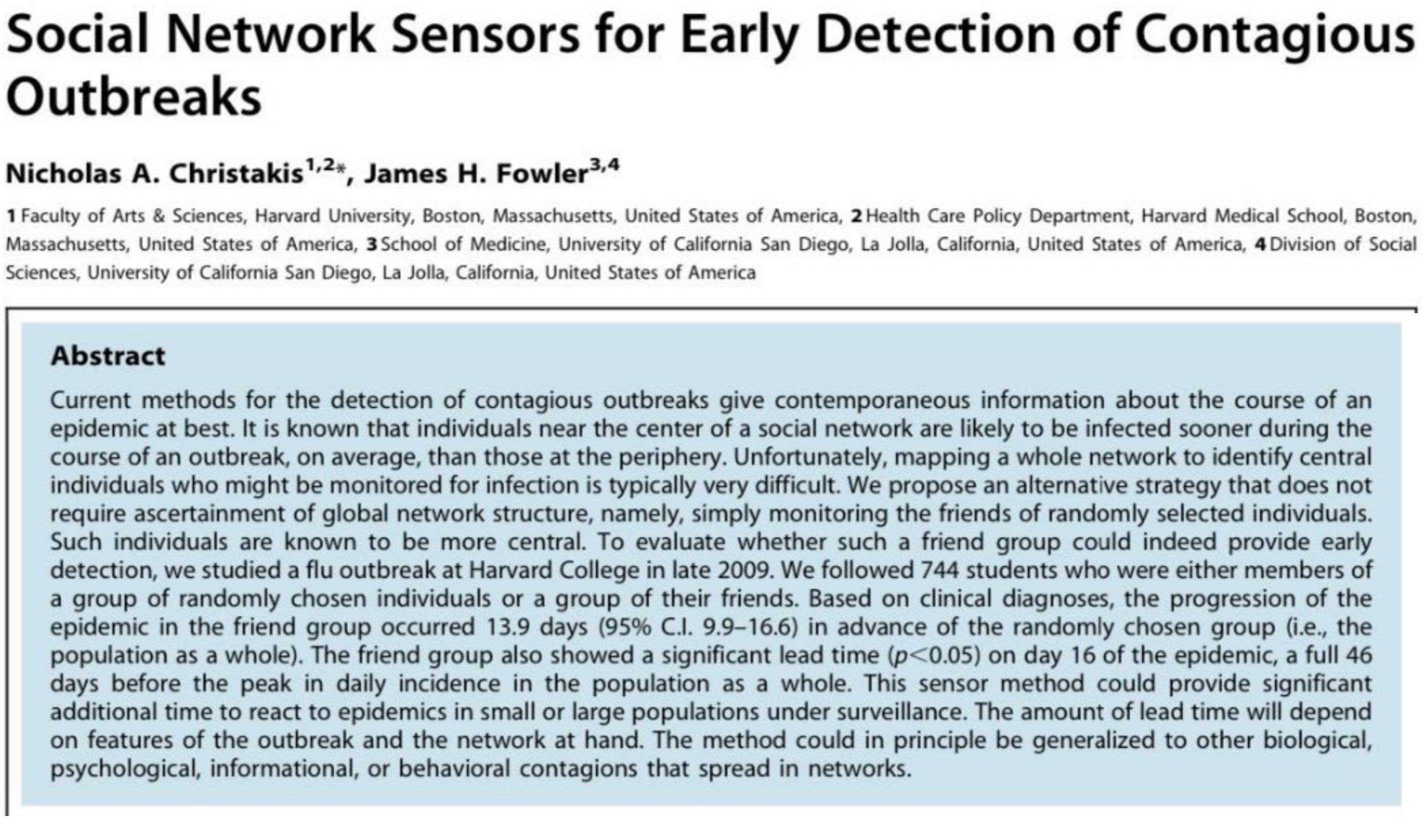
- Students complain that the classes are too big
- Dean: mean class size is ~35 students.
- Student Body: No, it is ~90
- What's going on?

[notebook]

- When sampling a student completely at random, that student is more likely to come from a large class than from a small class!
    - If there are 10 students in a class, you have 10 chances to sample that class; if there are 100 students, you have 100 chances. In general, if the class size is x, it will be overrepresented in the sample by a factor of x.
- Need to sample classes!

Inspection paradox/bias

- When the probability of observing a quantity is related to the quantity being observed, sampling “at random” may result in biased distributions
  - Waiting times for busses/trains
    - More likely to arrive to station when span between trains is longer
    - (notebook)
  - Number of friends in social media
    - A user’s friends are more likely to have many friends than a random user



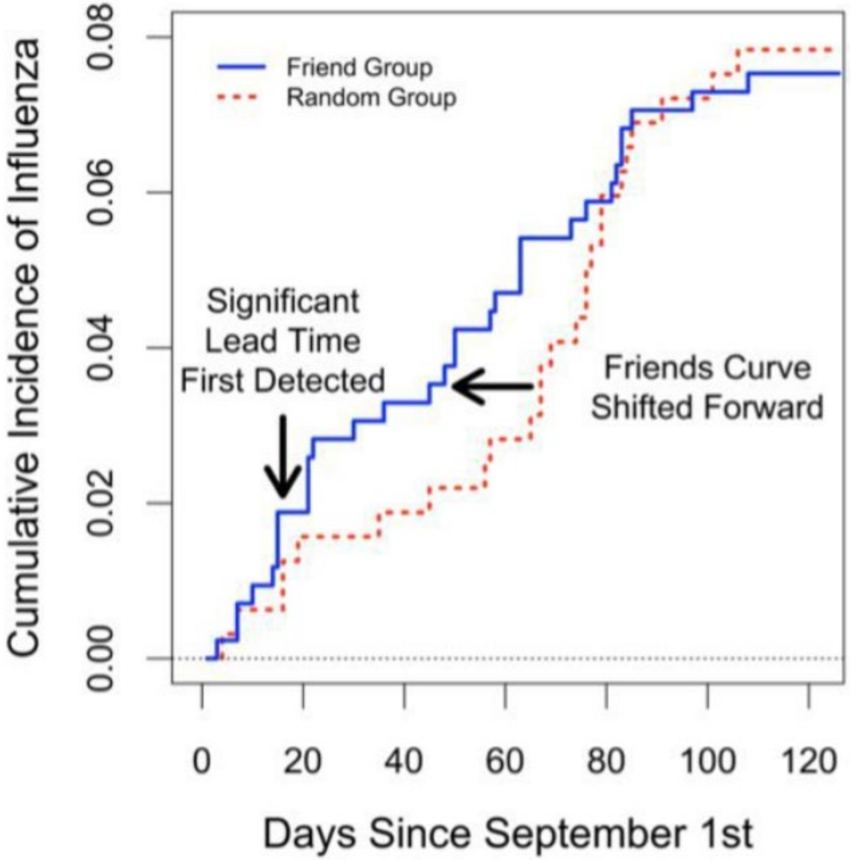
Social Network Sensors for Early Detection of Contagious Outbreaks

Nicholas A. Christakis[1,2]*, James H. Fowler[3,4]

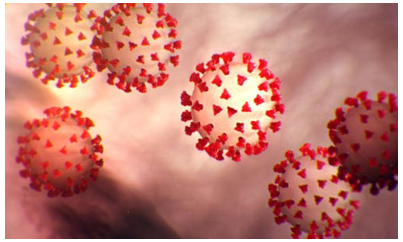
1 Faculty of Arts & Sciences, Harvard University, Boston, Massachusetts, United States of America, 2 Health Care Policy Department, Harvard Medical School, Boston, Massachusetts, United States of America, 3 School of Medicine, University of California San Diego, La Jolla, California, United States of America, 4 Division of Social Sciences, University of California San Diego, La Jolla, California, United States of America

Abstract

Current methods for the detection of contagious outbreaks give contemporaneous information about the course of an epidemic at best. It is known that individuals near the center of a social network are likely to be infected sooner during the course of an outbreak, on average, than those at the periphery. Unfortunately, mapping a whole network to identify central individuals who might be monitored for infection is typically very difficult. We propose an alternative strategy that does not require ascertainment of global network structure, namely, simply monitoring the friends of randomly selected individuals. Such individuals are known to be more central. To evaluate whether such a friend group could indeed provide early detection, we studied a flu outbreak at Harvard College in late 2009. We followed 744 students who were either members of a group of randomly chosen individuals or a group of their friends. Based on clinical diagnoses, the progression of the epidemic in the friend group occurred 13.9 days (95% C.I. 9.9–16.6) in advance of the randomly chosen group (i.e., the population as a whole). The friend group also showed a significant lead time ($p<0.05$) on day 16 of the epidemic, a full 46 days before the peak in daily incidence in the population as a whole. This sensor method could provide significant additional time to react to epidemics in small or large populations under surveillance. The amount of lead time will depend on features of the outbreak and the network at hand. The method could in principle be generalized to other biological, psychological, informational, or behavioral contagions that spread in networks.

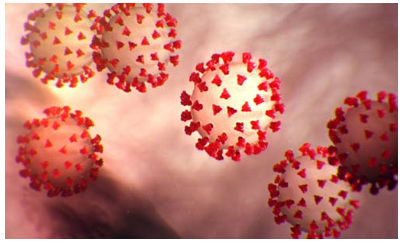


Case study: disease screening



2% of the population has a certain disease

Case study: disease screening

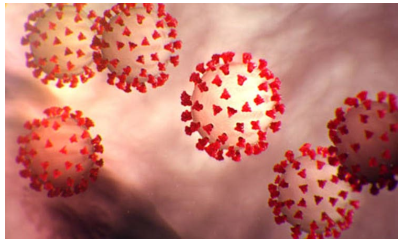


2% of the population has a certain disease

A test screens for the disease

- If a tested person is sick, there's 100% chance the test comes out positive ("sick")
  - No False Negatives (test has 100% sensitivity: later in course)
- If a tested person isn't sick, there's 5% chance the test comes out positive ("sick")
  - 5% False Positives (test has 95% specificity: later in course)

Case study: disease screening

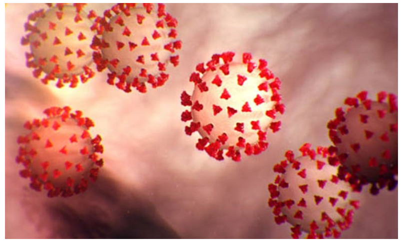


2% of the population has a certain disease

A test screens for the disease

- If a tested person is sick, there's 100% chance the test comes out positive ("sick")
  - No False Negatives (test has 100% sensitivity: later in course)
- If a tested person isn't sick, there's 5% chance the test comes out positive ("sick")
  - 5% False Positives (test has 95% specificity: later in course)
- If a test comes out positive, what are the chances that the person is sick?
  a) >99%
  b) ~95%
  c) ~80%
  d) ~50%
  e) ~30%

Case study: disease screening



2% of the population has a certain disease

A test screens for the disease

- If a tested person is sick, there’s 100% chance the test comes out positive (“sick”)
  - No False Negatives (test has 100% sensitivity: later in course)
- If a tested person isn’t sick, there’s 5% chance the test comes out positive (“sick”)
  - 5% False Positives (test has 95% specificity: later in course)
- If a test comes out positive, what are the chances that the person is sick?

  a) >99%
  b) ~95%
  c) ~80%
  d) ~50%
  [✓ e) ~30%]

Base rate neglect

- Say 1000 people are tested
- 2% of the population are sick ➔ 20 are truly sick
- Of 20 that are sick, all come out positive
- Of 980 that are not sick, 49 come out positive
    - 5% of 980
- In total, 69 people come out positive, but only 20 are really sick
- Probability of being sick given test is positive: 20/69 ≈ 30%

Base rate neglect

- Say 1000 people are tested
- 2% of the population are sick ➔ 20 are truly sick
- Of 20 that are sick, all come out positive
- Of 980 that are not sick, 49 come out positive
    - 5% of 980
- In total, 69 people come out positive, but only 20 are really sick
- Probability of being sick given test is positive: 20/69 ≈ 30%

	Person is sick	Person is not sick
Test comes out positive (“sick”)	20	49
Test comes out negative (“not sick”)	0	931

What if the base rate is higher?

50% of the population has a certain disease

A test screens for the disease

- If a tested person is sick, there’s 100% chance the test comes out positive (“sick”)
  - No False Negatives (100% sensitivity, later in course)
- If a tested person isn’t sick, there’s 5% chance the test comes out positive (“sick”)
  - 5% False Positives (95% specificity, later in course)
- If a test comes out positive, what are the chances that the person is sick?

  a) >99%
  b) ~95%
  c) ~80%
  d) ~50%
  e) ~30%

What if the base rate is higher?

- Say 1000 people are tested
- 50% of the population are sick ➔ 500 are truly sick
- Of 500 that are sick, all come out positive
- Of 500 that are not sick, 25 come out positive
  - 5% of 500
- In total 525 people come out positive, and 500 of them are really sick
- Probability of being sick given test is positive: 500/525 ≈ 95%

What if the base rate is higher?

- Say 1000 people are tested
- 50% of the population are sick ➔ 500 are truly sick
- Of 500 that are sick, all come out positive
- Of 500 that are not sick, 25 come out positive
  - 5% of 500
- In total 525 people come out positive, and 500 of them are really sick
- Probability of being sick given test is positive: 500/525 ≈ 95%

	Person is sick	Person is not sick
Test comes out positive (“sick”)	500	25
Test comes out negative (“not sick”)	0	475

The relation to p-values

- p-value threshold = probability of rejecting the null given that it is true

The relation to p-values

- p-value threshold = probability of rejecting the null given that it is true

	Null is false (Person is sick)	Null is true (Person is not sick)
Test rejects the null (Test comes out positive)	✓	X
Test doesn’t reject the null (Test comes out negative)	X	✓

The relation to p-values

- p-value threshold = probability of rejecting the null given that it is true

	Null is false (Person is sick)	Null is true (Person is not sick)
Test rejects the null (Test comes out positive)	✓	X
Test doesn’t reject the null (Test comes out negative)	X	✓

The relation to p-values

- p-value threshold = probability of rejecting the null given that it is true

	Null is false (Person is sick)	Null is true (Person is not sick)
Test rejects the null (Test comes out positive)	✓	X
Test doesn’t reject the null (Test comes out negative)	X	✓

- It does **not answer the question** “how likely the null hypothesis is to be false”
  - Or even how likely it is given the test rejects it!

The relation to p-values

- p-value threshold = probability of rejecting the null given that it is true

	Null is false (Person is sick)	Null is true (Person is not sick)
Test rejects the null (Test comes out positive)	✓	X
Test doesn't reject the null (Test comes out negative)	X	✓

- It does **not answer the question** "how likely the null hypothesis is to be false"
  - Or even how likely it is given the test rejects it!

The relation to p-values

- Say p-value = 0.05, but in only ~2% of the “worlds” the null is false

	Null is false (Person is sick)	Null is true (Person is not sick)
Test rejects the null (Test comes out positive)	20	50
Test doesn’t reject the null (Test comes out negative)	0	950

The relation to p-values

- Say p-value = 0.05, and in ~50% of the “worlds” the null is false

	Null is false (Person is sick)	Null is true (Person is not sick)
Test rejects the null (Test comes out positive)	1000	50
Test doesn’t reject the null (Test comes out negative)	0	950