SSNAP - Statistical Analysis

Part I - Introduction to Data

Dr. Mine Çetinkaya-Rundel
Duke University

Duke
UNIVERSITY

slides at **bit.ly/ssnap-2022**

observational studies & experiments

types of studies

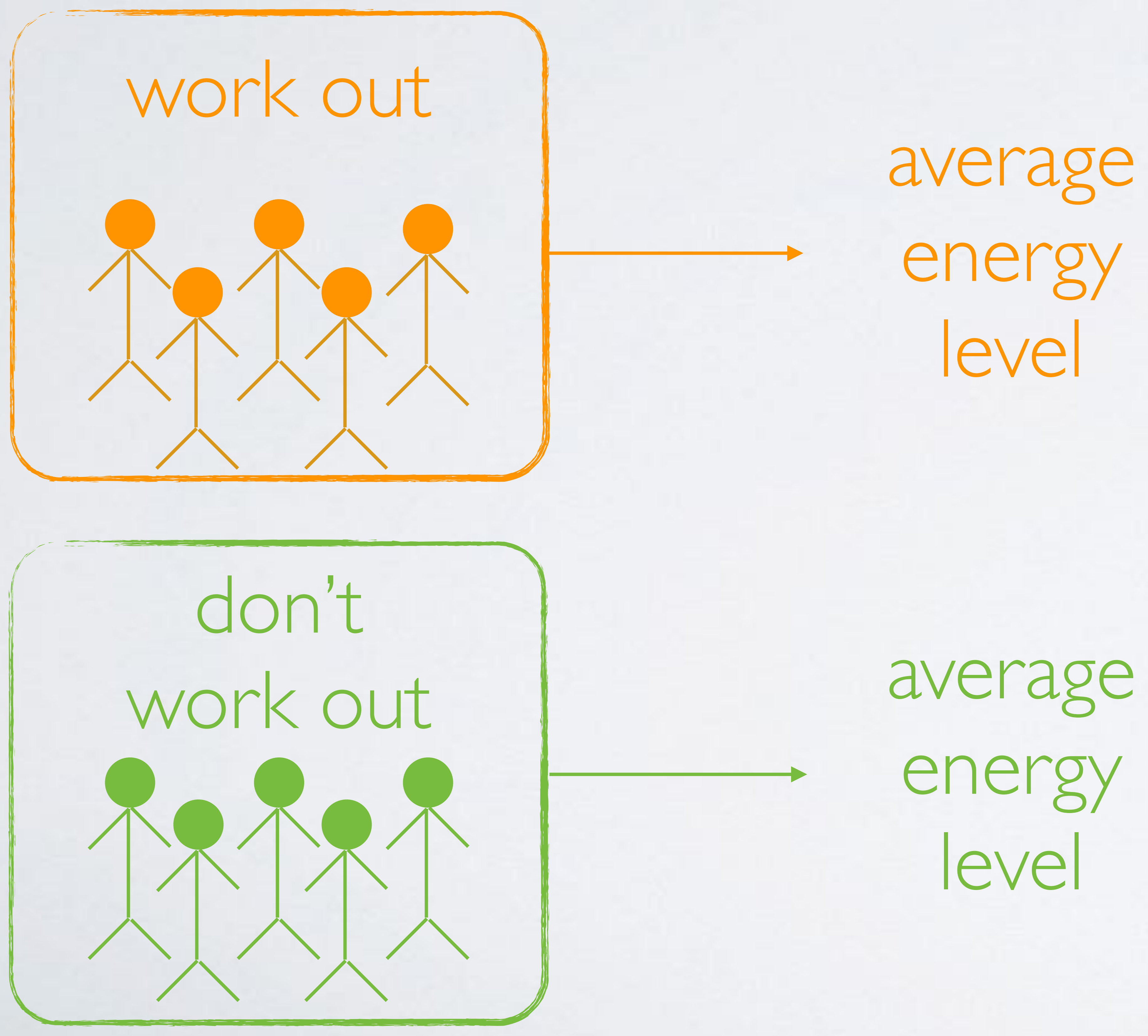
observational study

- collect data in a way that does not directly interfere with how the data arise ("observe")
- only establish an association
- retrospective: uses past data
- prospective: data are collected throughout the study

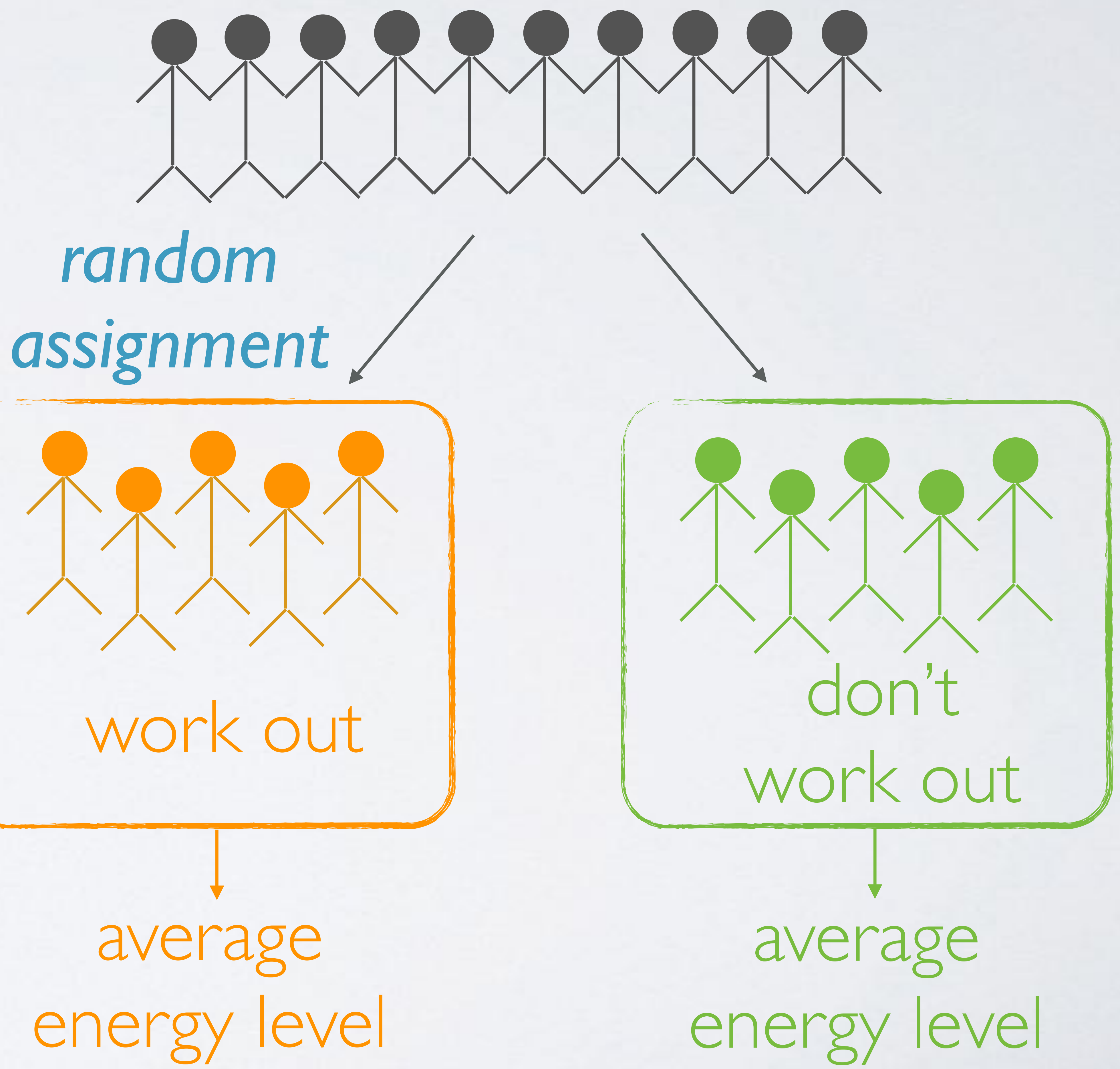
experiment

- randomly assign subjects to treatments
- establish causal connections

observational study



experiment



example: eating breakfast

Study: Breakfast cereal keeps girls slim

USA TODAY

Sept 8, 2005

[…]

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute with funding from the National Institutes of Health (NIH) and cereal-maker General Mills.

[…]

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio, and Maryland who were tracked between the ages of 9 and 19.

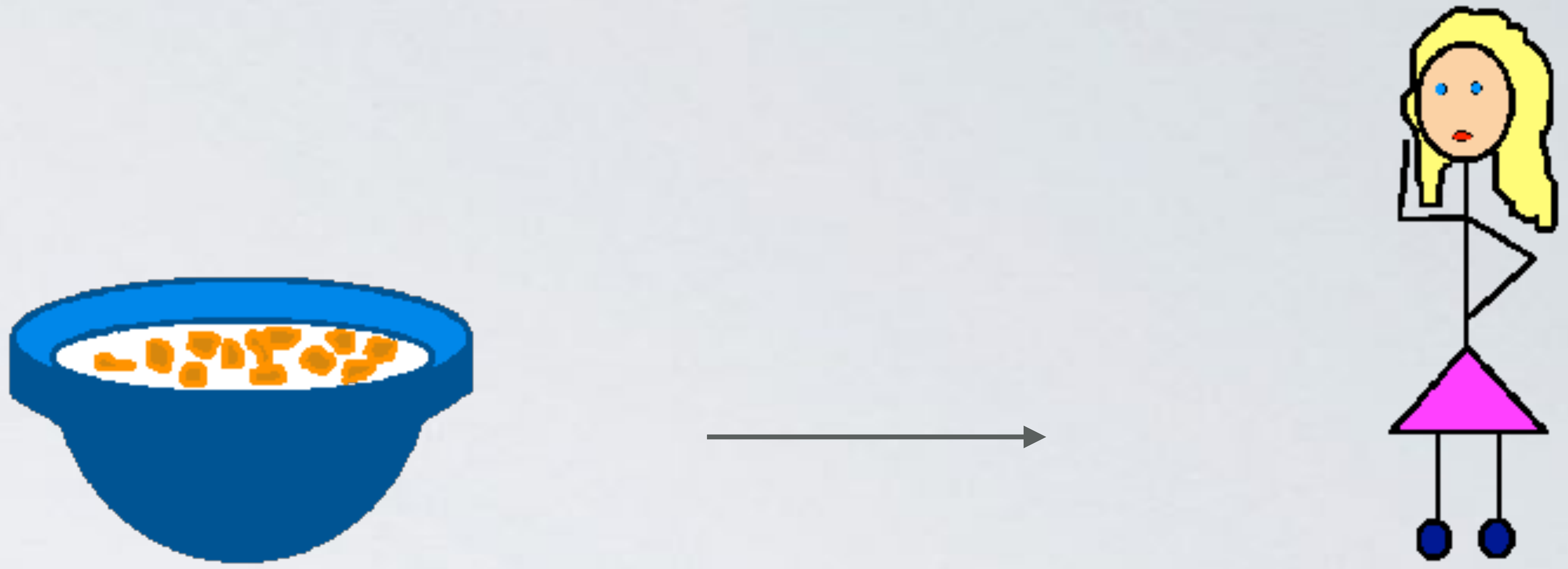
[…]

As part of the survey, the girls were asked once a year what they had eaten during the previous three days.

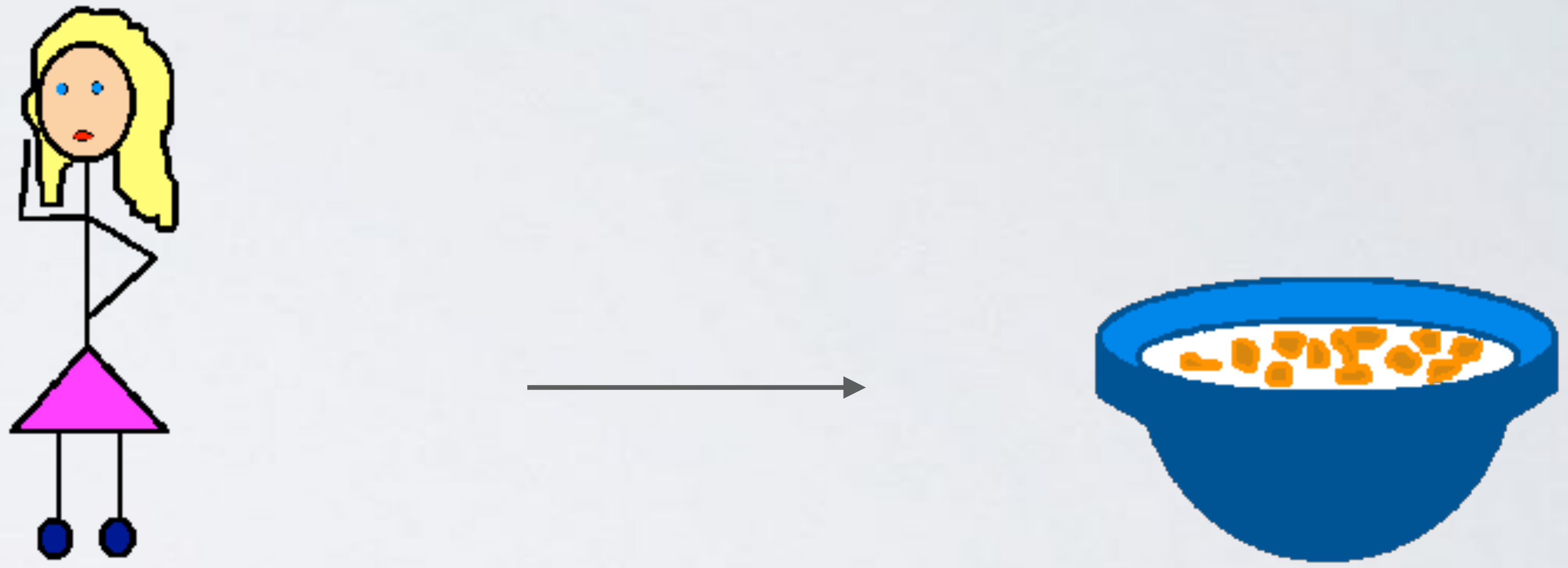
[…]

example: eating breakfast

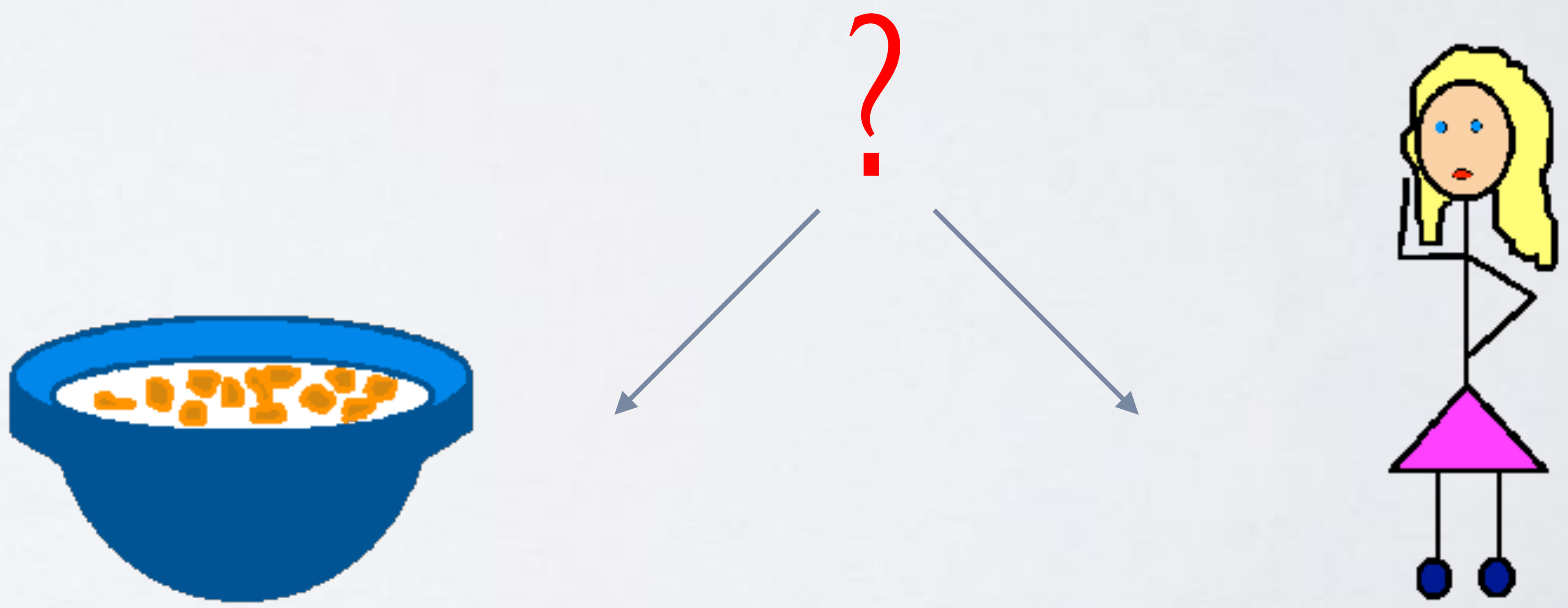
1. eating breakfast causes girls to be slimmer



2. being slim causes girls to eat breakfast



3. a third variable is responsible for both



correlation does not imply causation

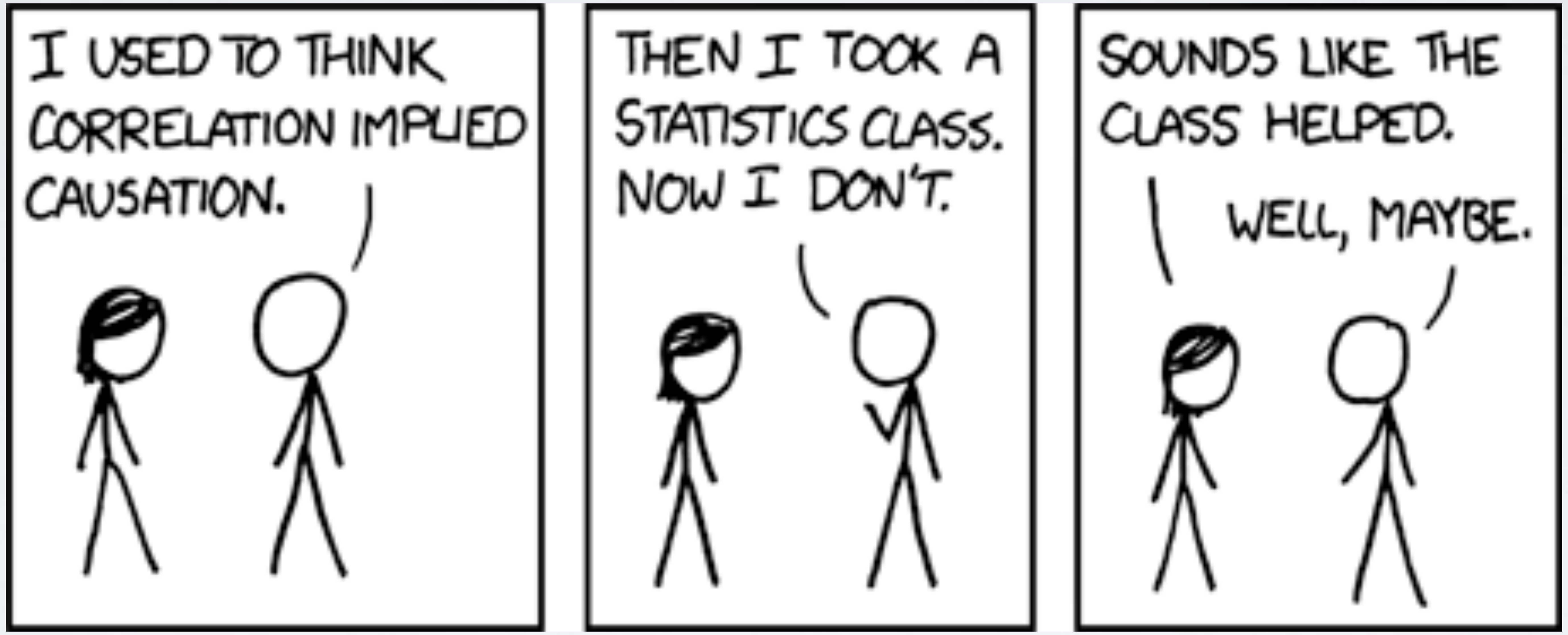
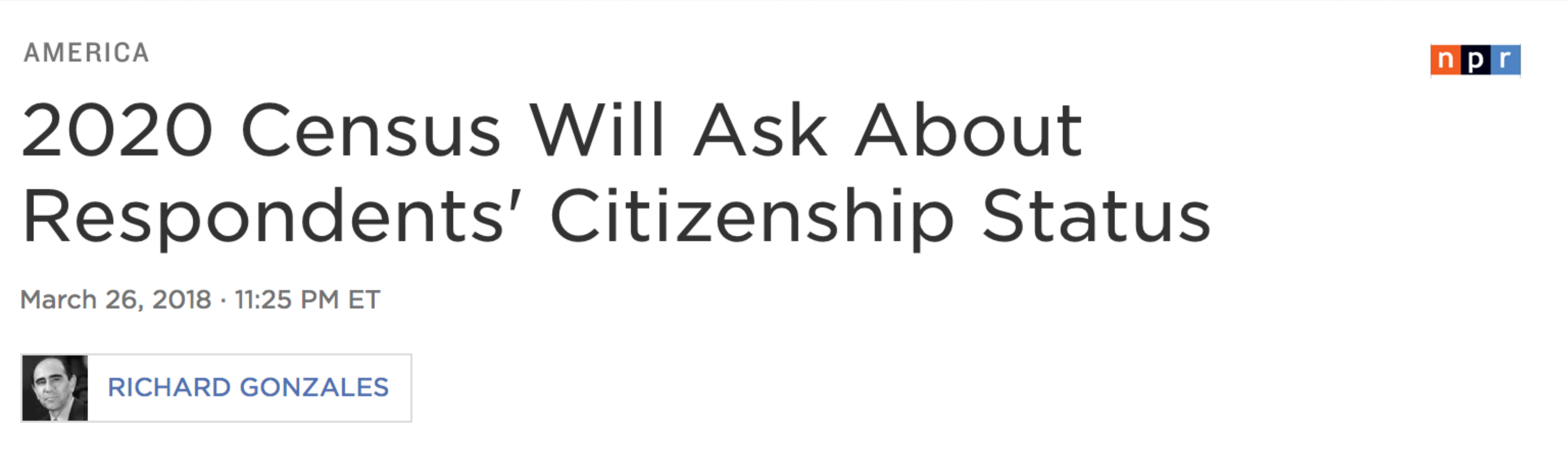


Image credit: Randall Munroe CC BY-NC 2.5 http://xkcd.com/552/

sampling & experimental design

census

- Some individuals are hard to locate or measure, and these people may be different from the rest of the population.
- Populations rarely stand still.



AMERICA

npr

2020 Census Will Ask About Respondents' Citizenship Status

March 26, 2018 · 11:25 PM ET

RICHARD GONZALES

Listen to the NPR story at http://www.npr.org/templates/story/story.php?storyId=125380052

Image credit: Wonderlane CC BY 2.0 http://www.flickr.com/photos/wonderlane/6231888661

sources of sampling bias

- Convenience sample: Individuals who are easily accessible are more likely to be included in the sample
- Non-response: If only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population
- Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue

Poll source: edition.cnn.com, August 29, 2013

principles of experimental design

(1) control

compare treatment of interest to a control group

(2) randomize

randomly assign subjects to treatments

(3) replicate

collect a sufficiently large sample, or replicate the entire study

(4) block

block for variables known or suspected to affect the outcome

more on blocking

- Design an experiment investigating whether energy gels help you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- Energy gels might affect pro and amateur athletes differently
- Block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro and amateur athletes to treatment and control groups
  - Pro and amateur athletes are equally represented in both groups

blocking vs. explanatory variables

- Explanatory variables (factors) - conditions we can impose on experimental units
- Blocking variables - characteristics that the experimental units come with, that we would like to control for
- Blocking is like stratifying:
  - Blocking during random assignment
  - Stratifying during random sampling

ideal experiment

most observational studies

	Random assignment	No random assignment	
Random sampling	causal and generalizable	not causal, but generalizable	Generalizability
No random sampling	causal, but not generalizable	neither causal nor generalizable	No generalizability
	Ca		

most experiments

DO YOU SUFFER FROM ASTHMA AND ALLERGIES?

Volunteers, who are non-smokers, diagnosed

simpson's paradox

explanatory and response

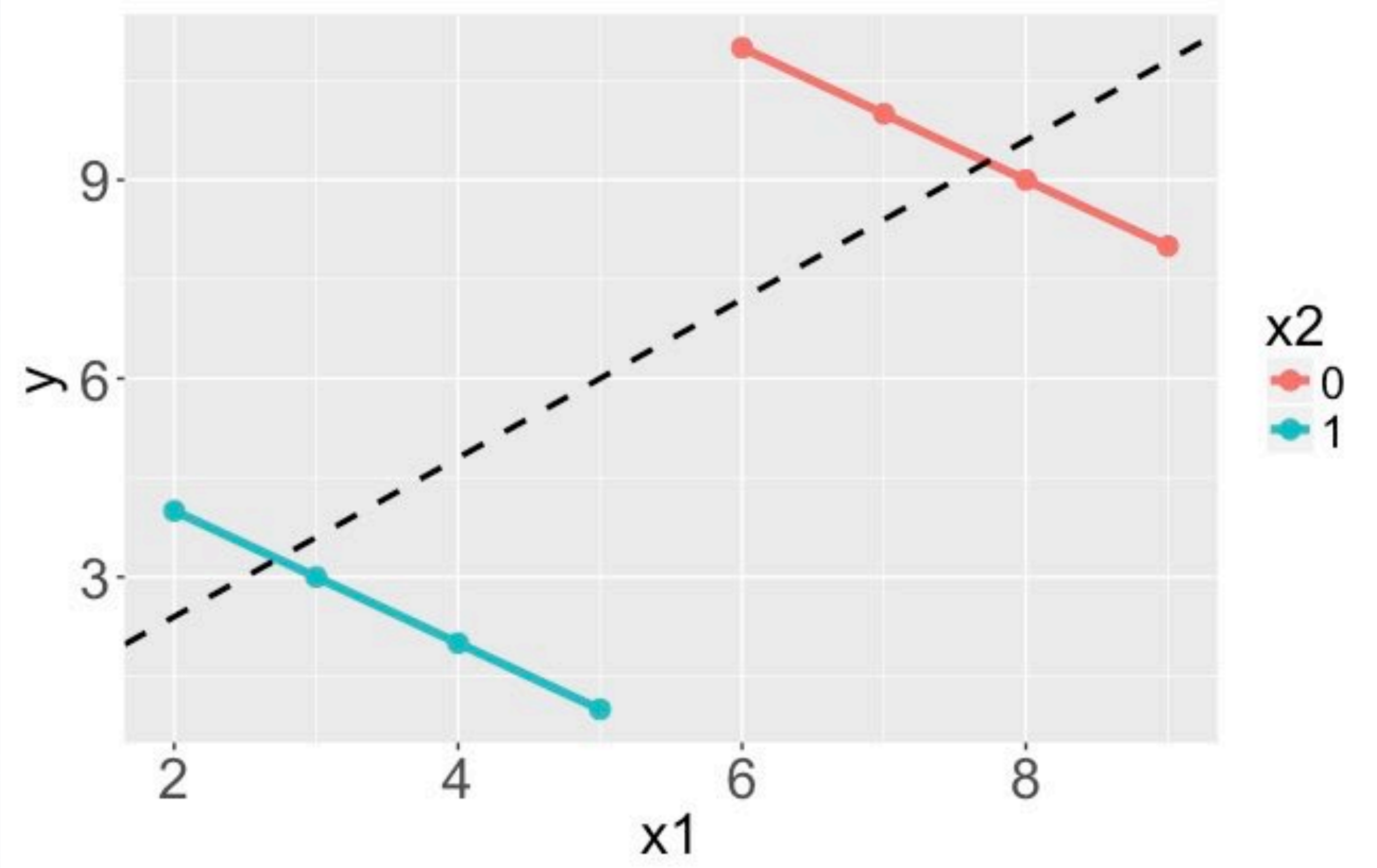
Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified

fitness (explanatory)	mental health (response)

multivariate relationships

stress (explanatory)	age (explanatory)	fitness (explanatory)	mental health (response)

- Not considering an important variable when studying a relationship can result in what we call a Simpson's paradox
- Illustrates the effect the omission of an explanatory variable can have on the measure of association between another explanatory variable and a response variable



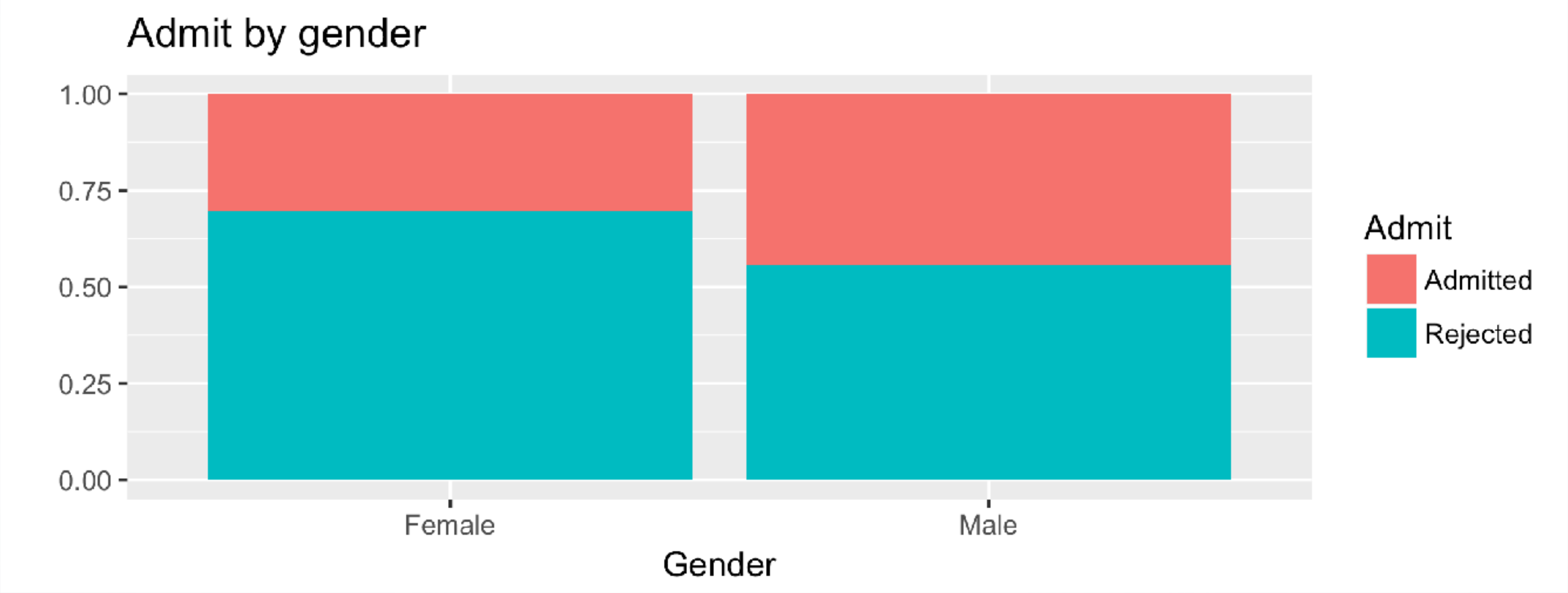
example: Berkeley admission

- Study carried out by the graduate Division of the University of California, Berkeley in the early 70's to evaluate whether there was a gender bias (coded as male and female only) in graduate admissions
- The data come from six departments. For confidentiality, they're labelled A-F in the data.
- We have information on whether the applicant was male or female and whether they were admitted or rejected.

example: Berkeley admission

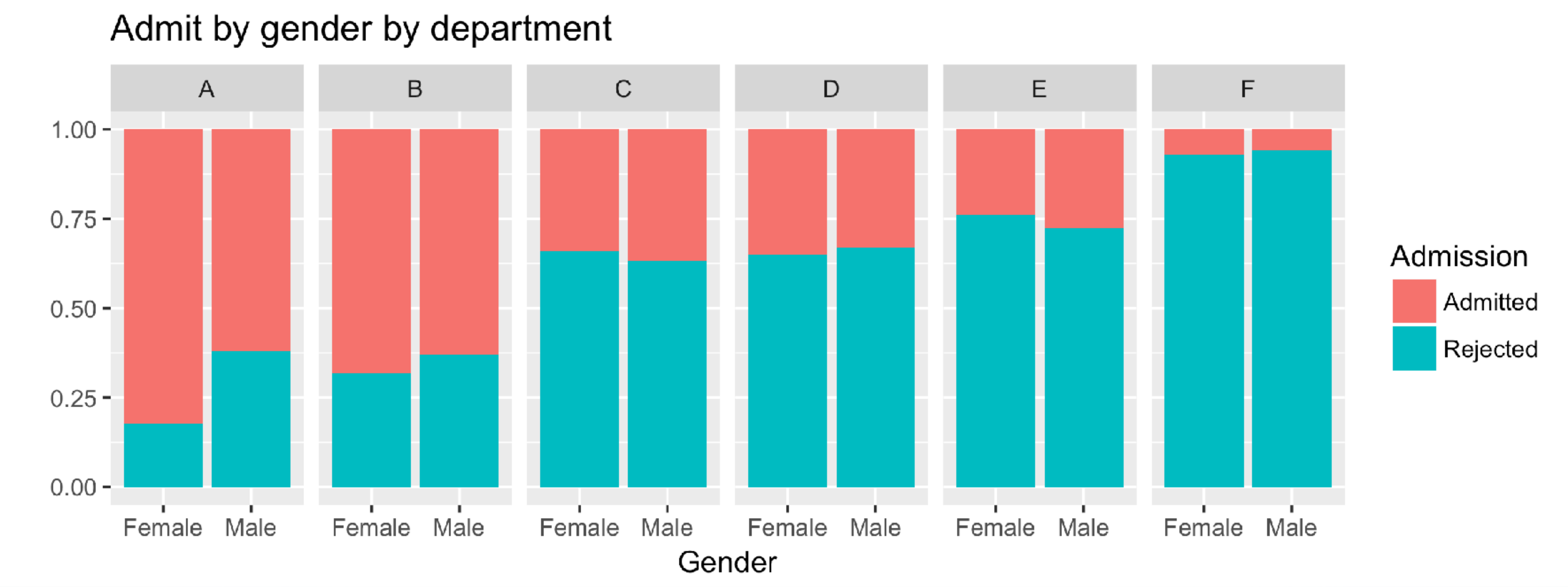
What can you say about the overall gender distribution?

	Admitted	Rejected
Male	1198	1493
Female	557	1278



example: Berkeley admission

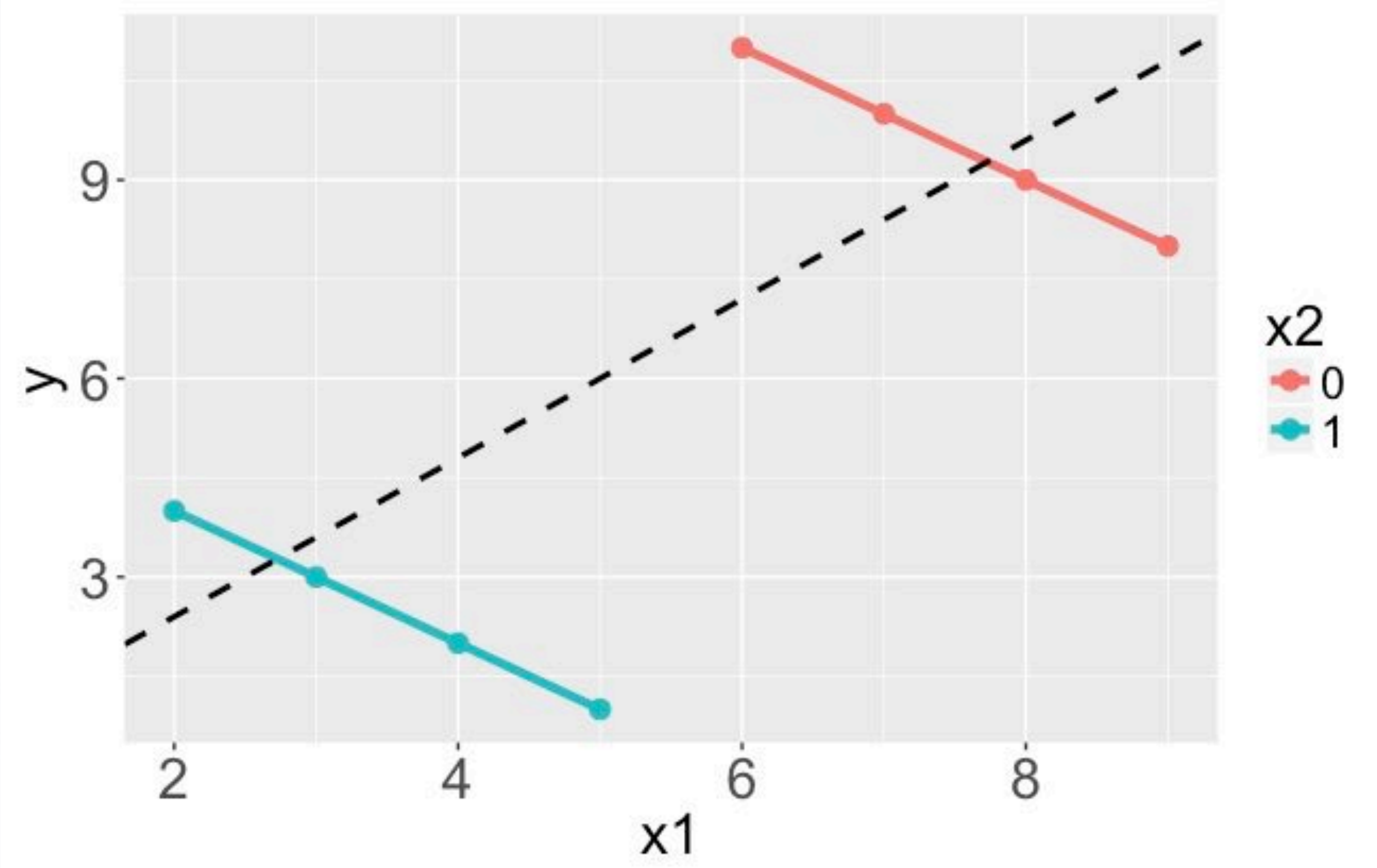
What can you say about the gender distribution by department?



confounding

confounding

- Simpson's paradox is a special (and extreme) case of confounding where the inclusion of a third variable *reverses* the relationship between the other variables
- Confounding can happen if a third variable *changes* the magnitude of the relationship, even if it doesn't reverse it



rewind to Oct 20, 2021…

"nearly 60% of Israeli hospitalized COVID-19 patients are fully vaccinated"

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax	Fully Vax	vs. severe disease
All ages			214	301	**Vax don't work!**

Source: https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

taking into consideration vaccination rate

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**

Efficacy = 1 - V / N

V = rate of infection per 100k for fully vaccinated

N = rate of infection per 100k for unvaccinated

Source: https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

taking into consideration age

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

Source: https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated