

SSNAP - Statistical Analysis

Part I - Introduction to Data

observational studies &
experiments

observational study

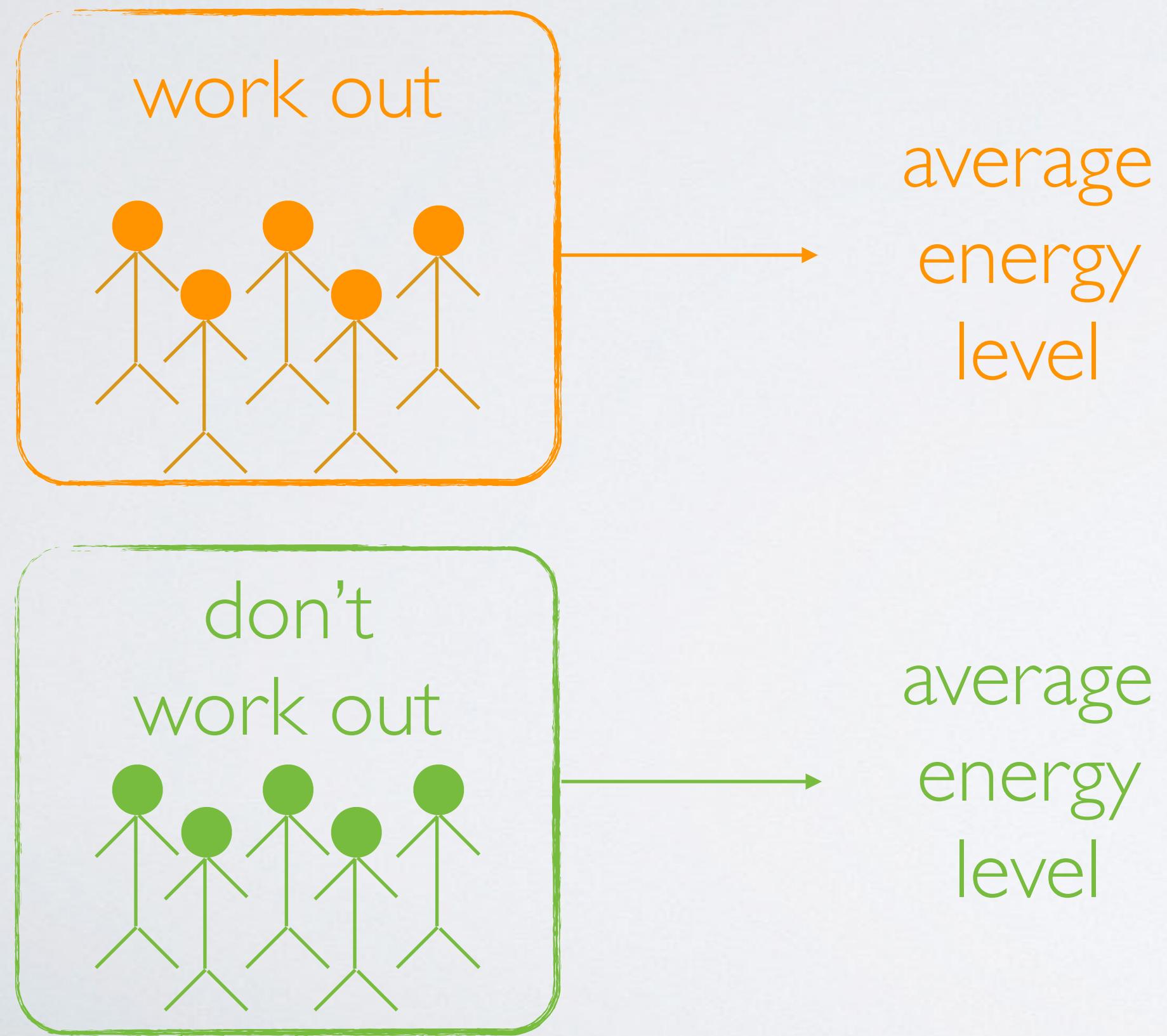
- ▶ collect data in a way that does not directly interfere with how the data arise (“observe”)
- ▶ only establish an association
- ▶ **retrospective:** uses past data
- ▶ **prospective:** data are collected throughout the study

experiment

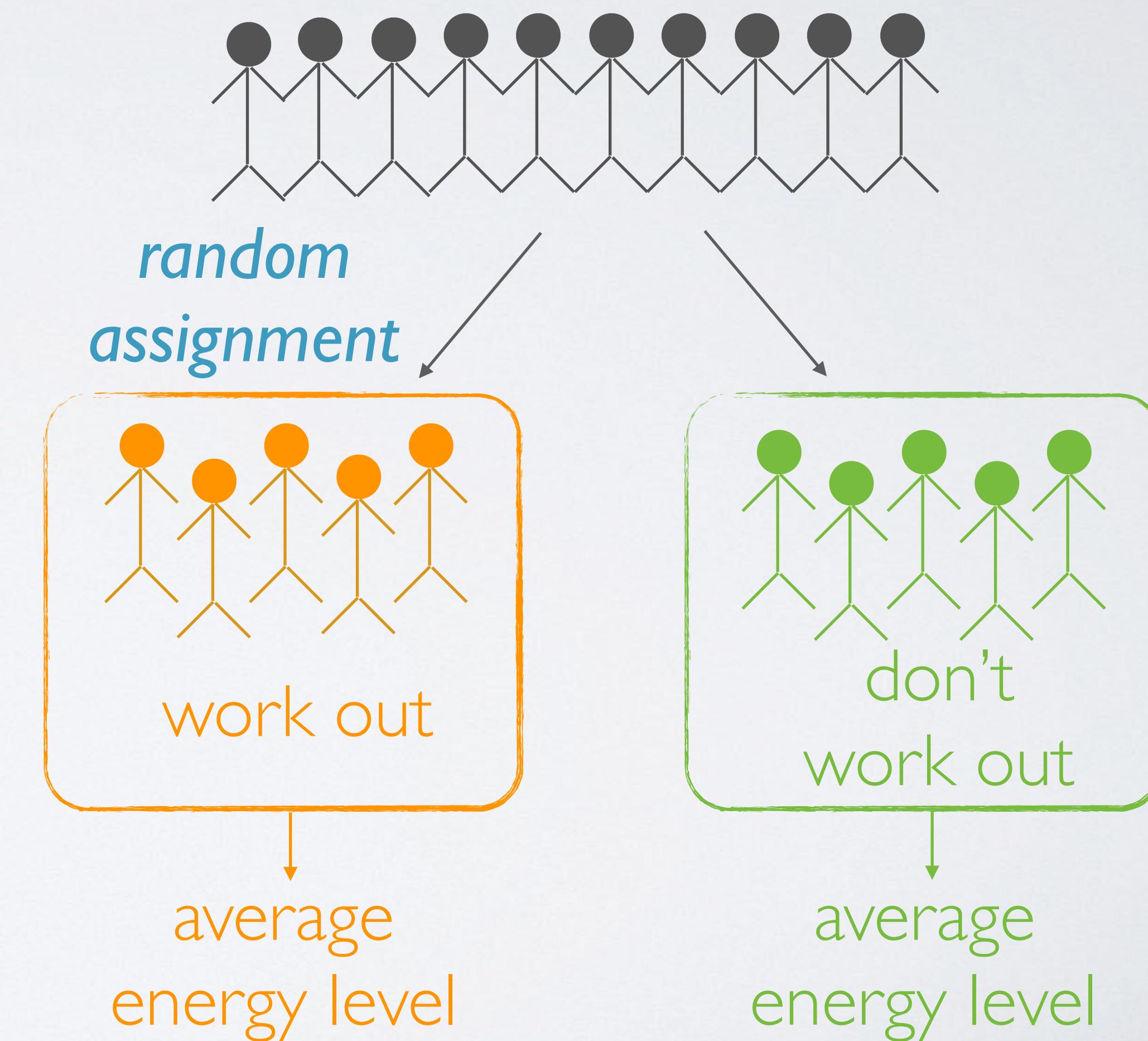
- ▶ randomly assign subjects to treatments
- ▶ establish causal connections

example: working out

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.

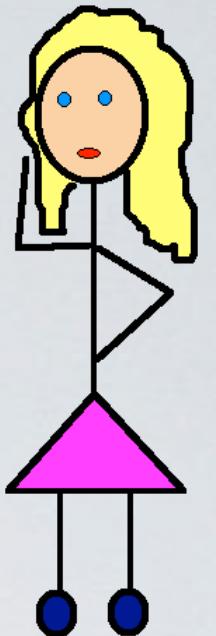
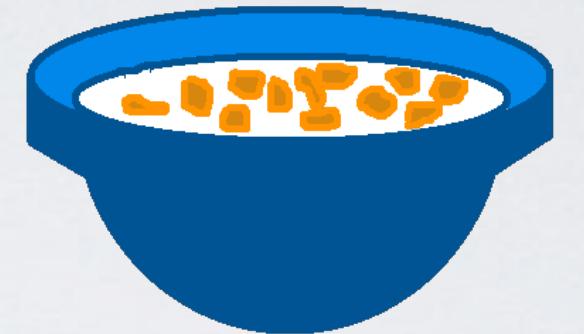
[...]

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

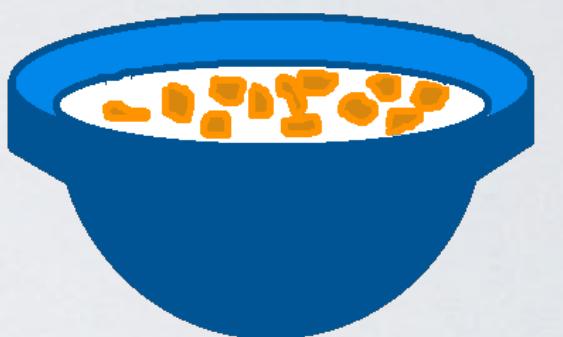
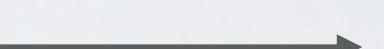
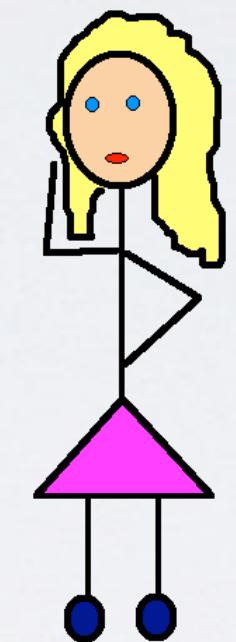
[...]

example: eating breakfast

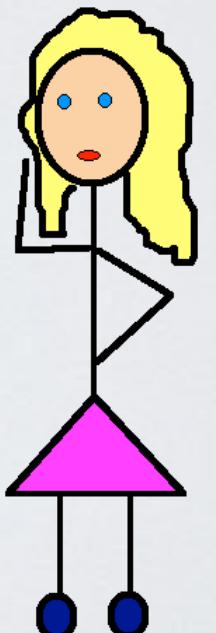
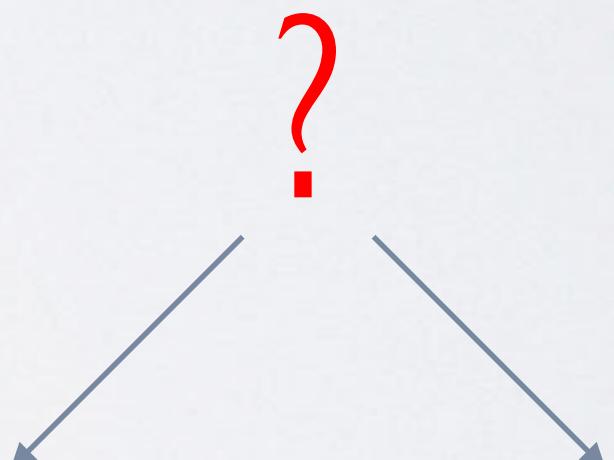
1. eating breakfast causes girls to be slimmer



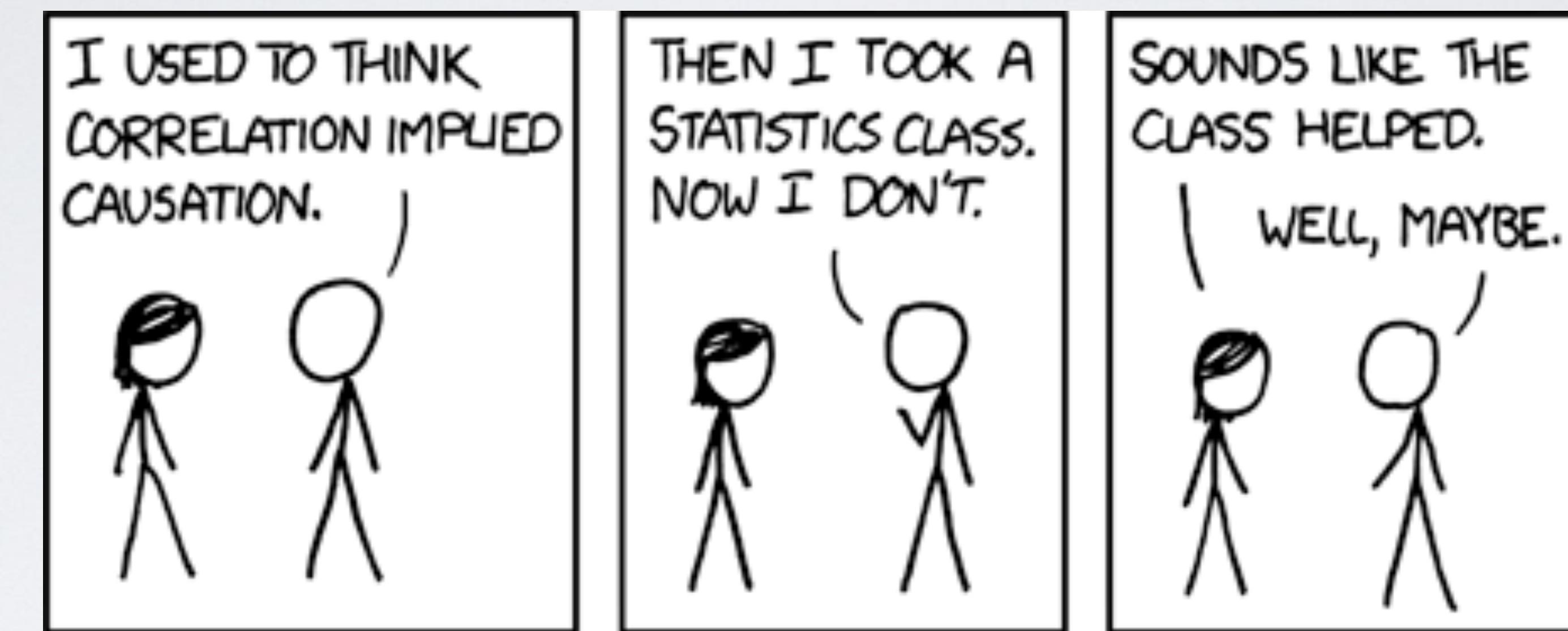
2. being slim causes girls to eat breakfast



3. a third variable is responsible for both



correlation does not imply causation



sampling & experimental design

- ▶ 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



inference

**representative
sample**

**exploratory
analysis**

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



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

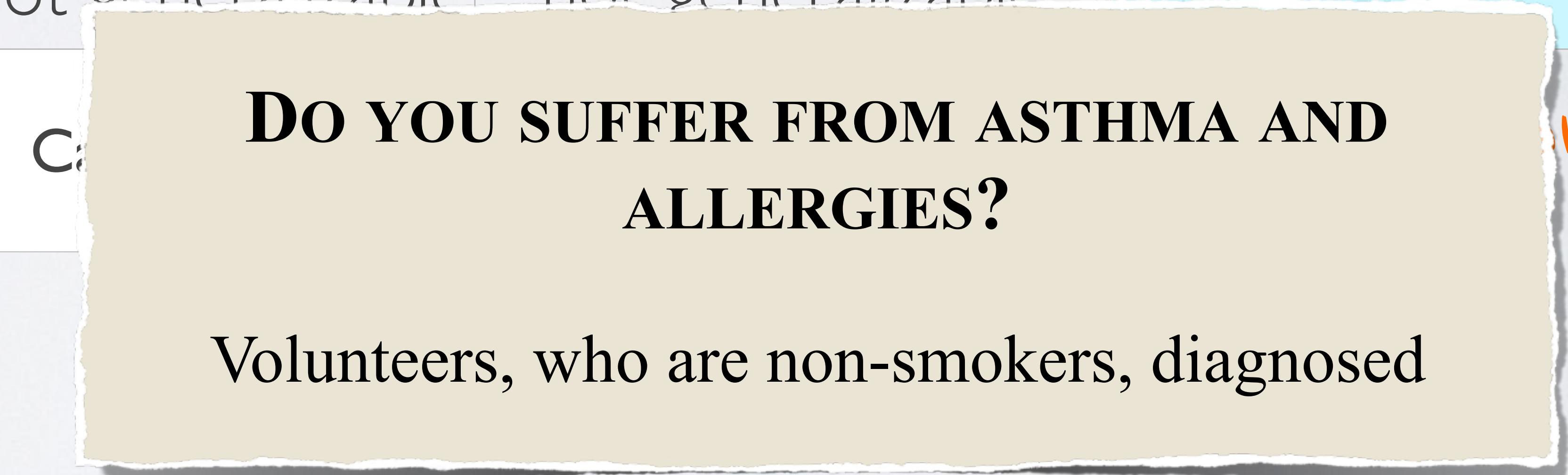
- ▶ Design an experiment investigating whether energy gels help you run faster:
 - ▶ Treatment: energy gel
 - ▶ Control: no energy gel
- ▶ Energy gels might affect pro and amateur athletes differently
- ▶ Block for pro status:
 - ▶ Divide the sample to pro and amateur
 - ▶ Randomly assign pro and amateur 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

	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
most experiments			



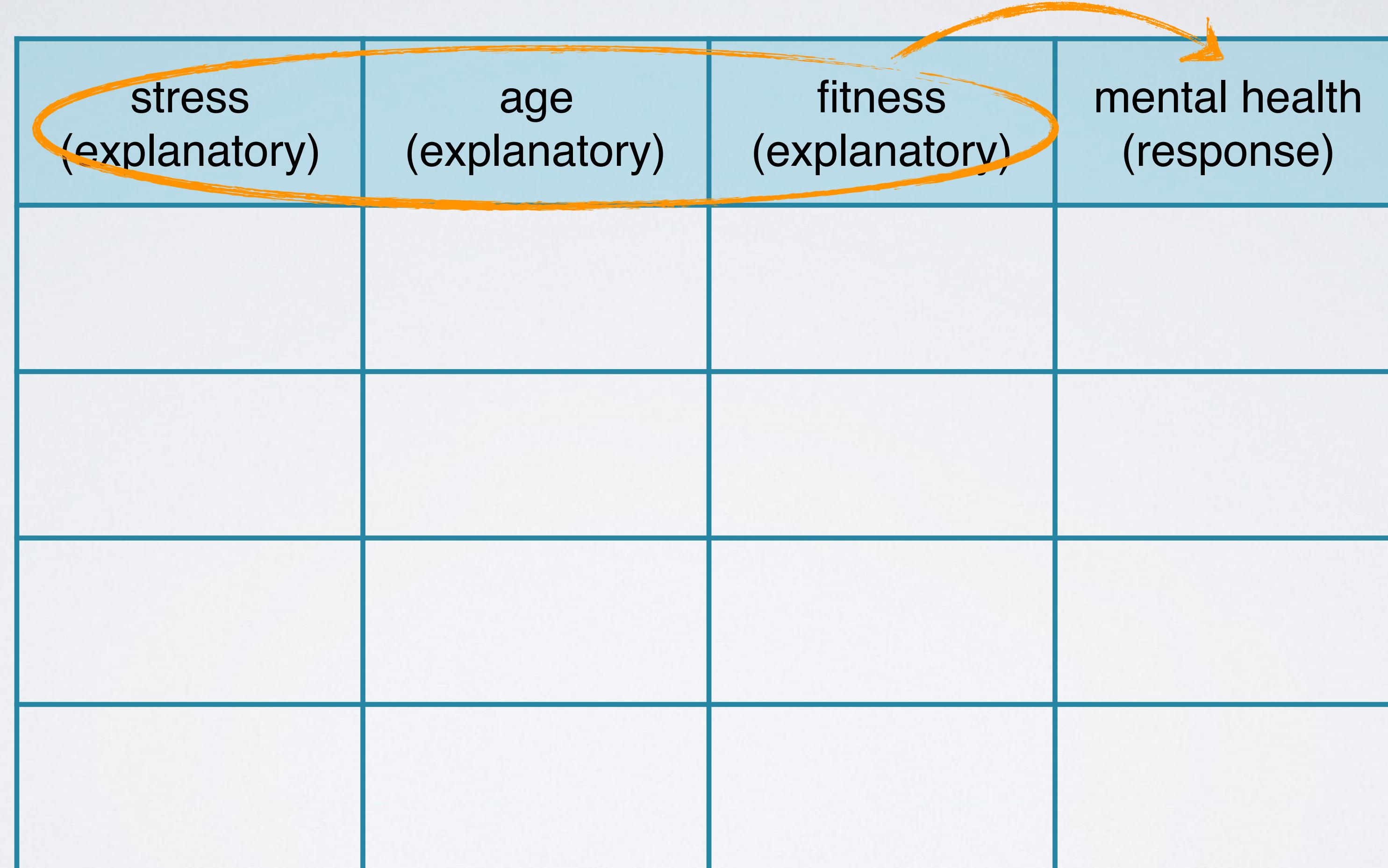
simpson's paradox

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

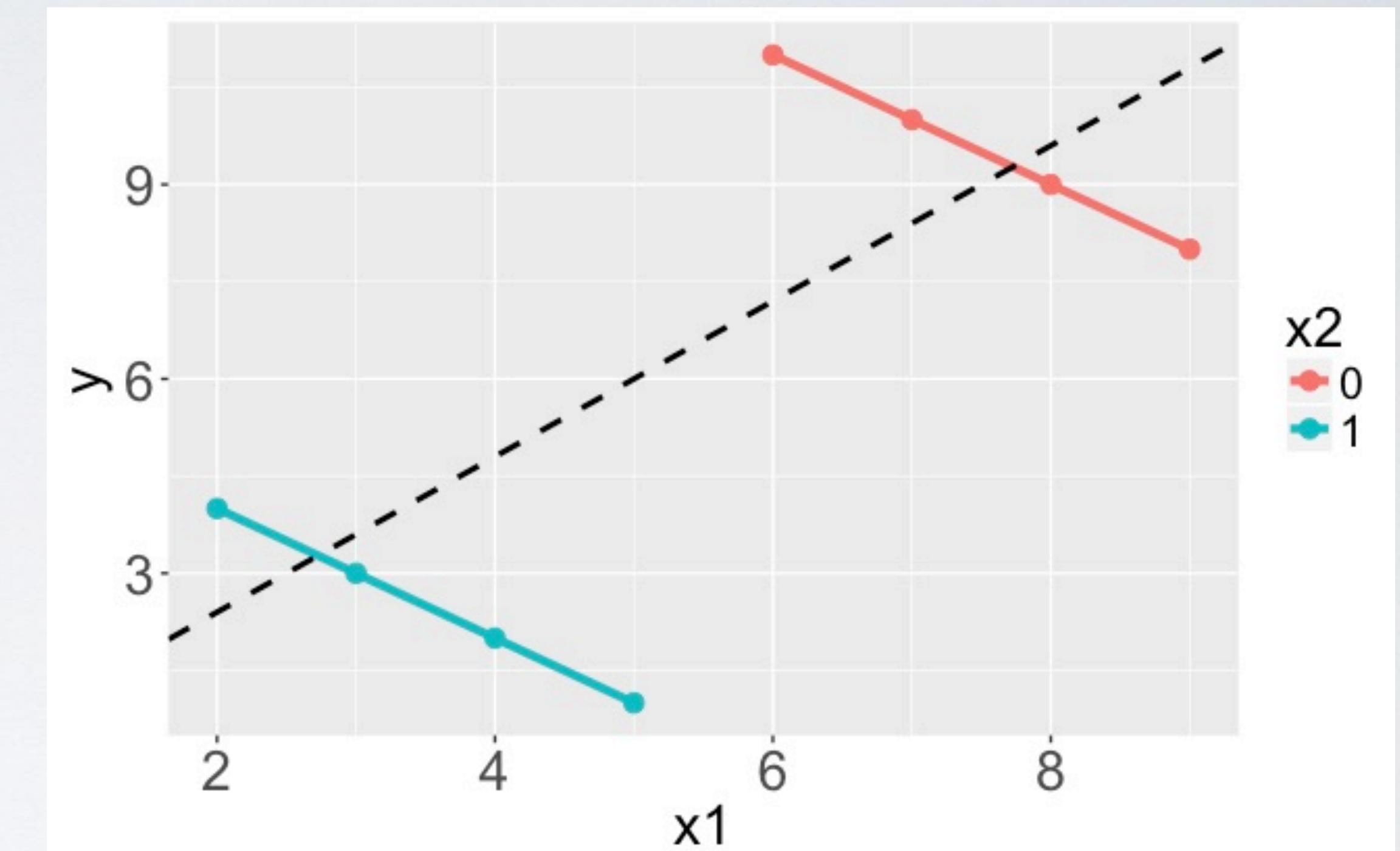
fitness (explanatory)	mental health (response)

multivariate relationships



simpson's paradox

- ▶ 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



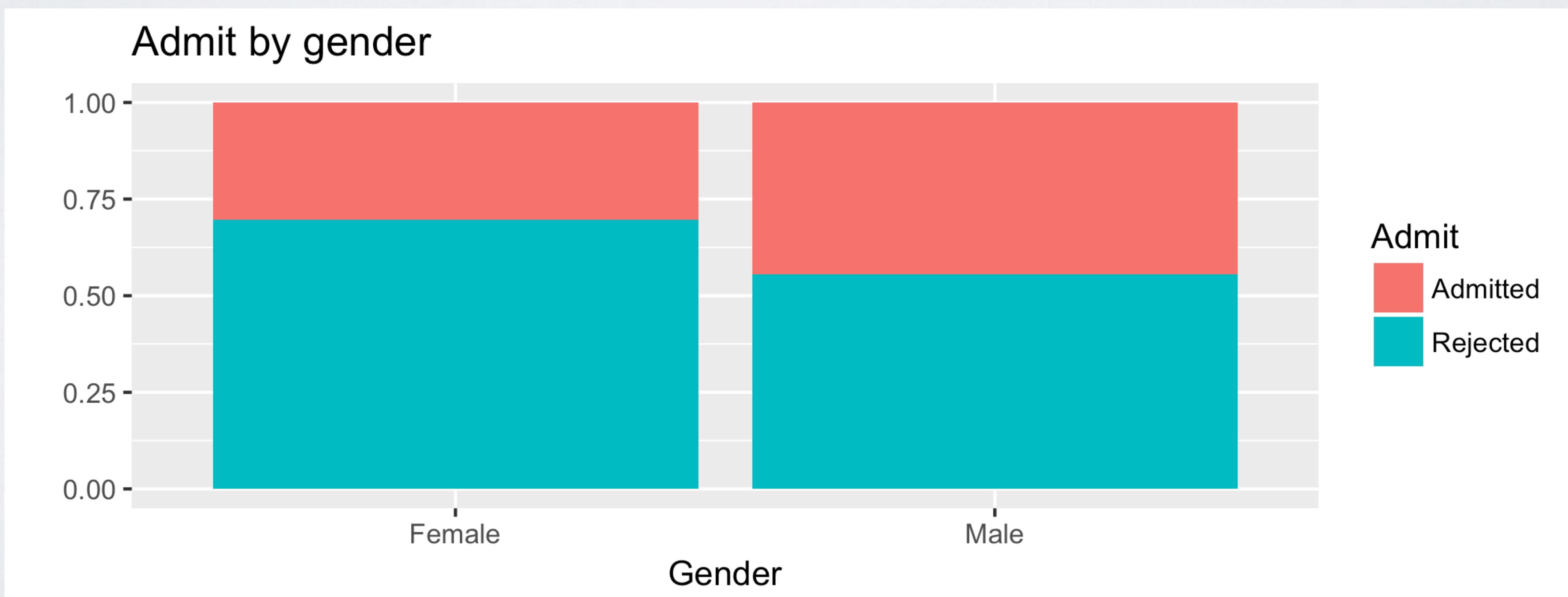
example: Berkeley admission

- ▶ 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 in graduate admissions
- ▶ The data come from six departments. For confidentiality we'll call them A-F.
- ▶ 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?

Admit by gender by department

