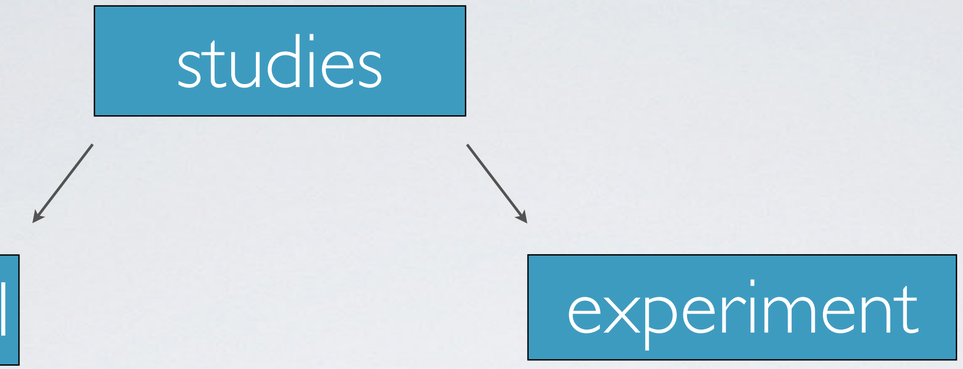
Part I - Introduction to Data

Dr. Mine Çetinkaya-Rundel
Duke University

Duke
UNIVERSITY

observational studies & experiments

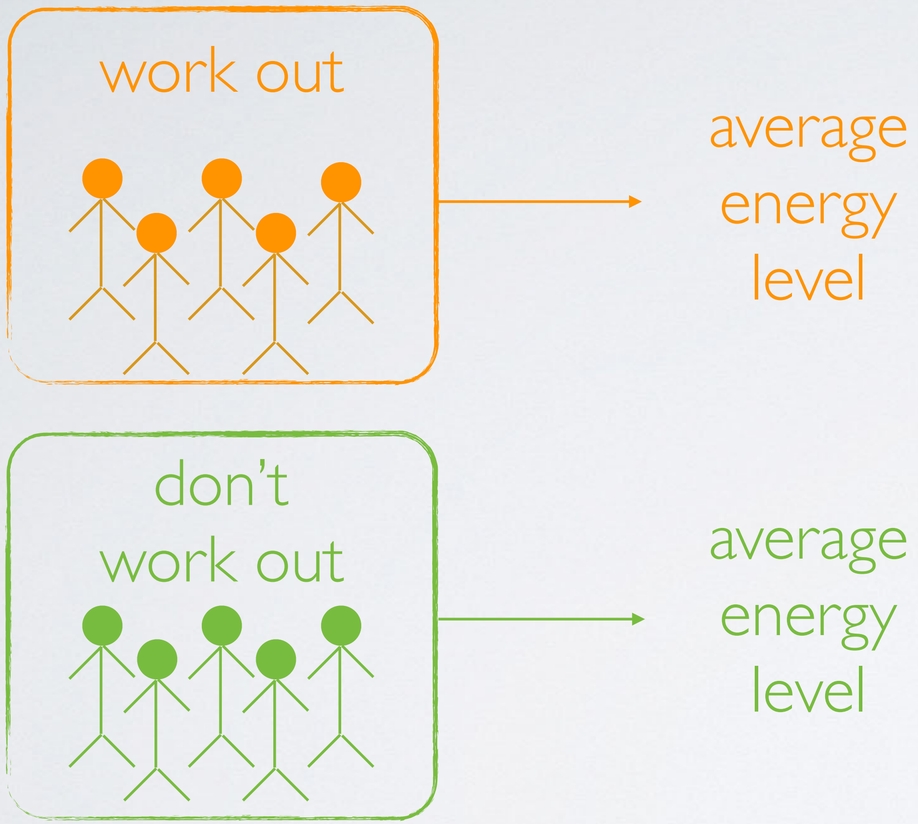
- define observational studies and experiments
- correlation vs. causation



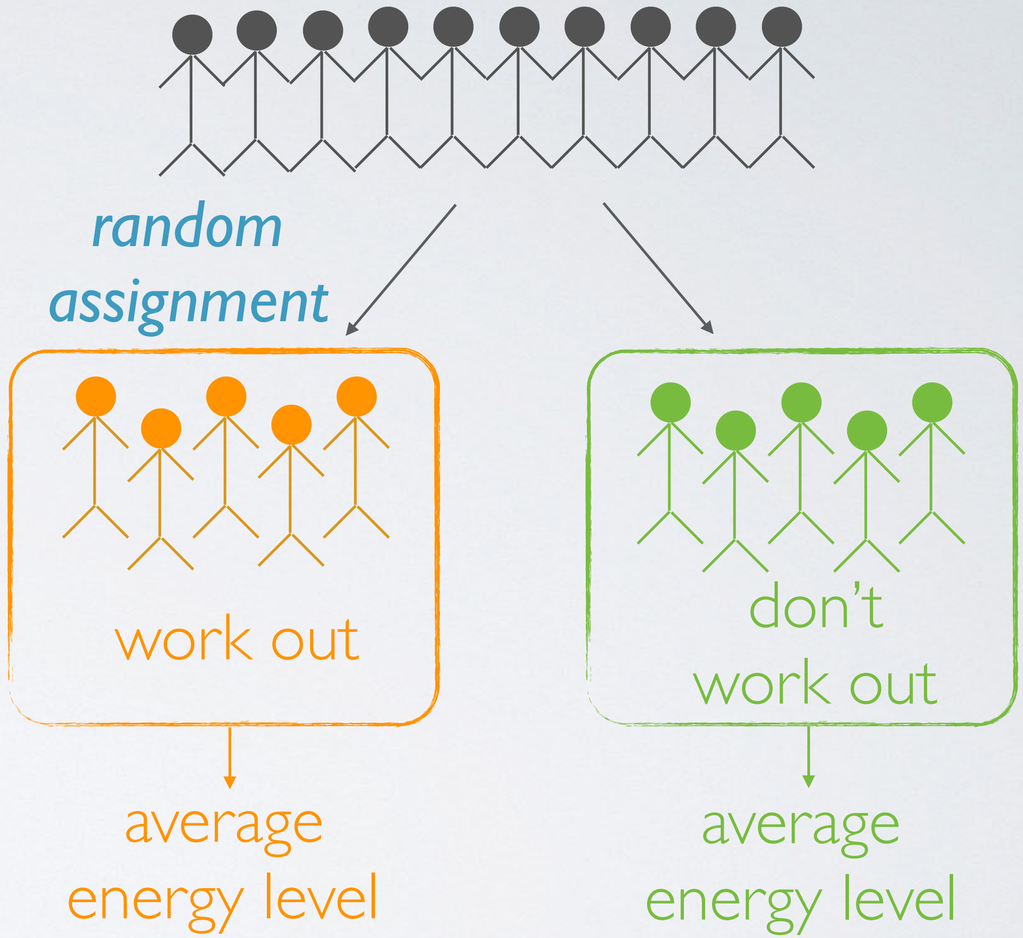
- collect data in a way that does not directly interfere with how the data arise ("observe")
- only establish an association
- retrospective: uses past data
- prospective: data are collected throughout the study

- randomly assign subjects to treatments
- establish causal connections

observational study



experiment



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.

[…]

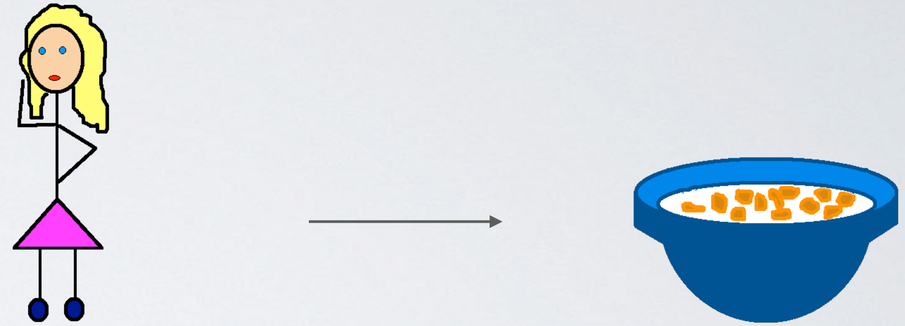
As part of the survey, the girls were asked once a year what they had eaten during the previous three days.

[…]

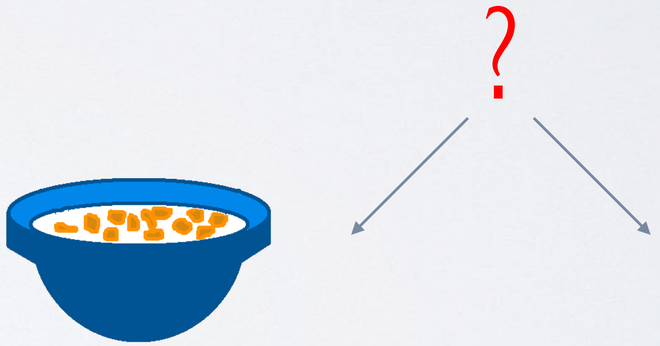
3 possible explanations

1. eating breakfast causes girls to be slimmer

2. being slim causes girls to eat breakfast

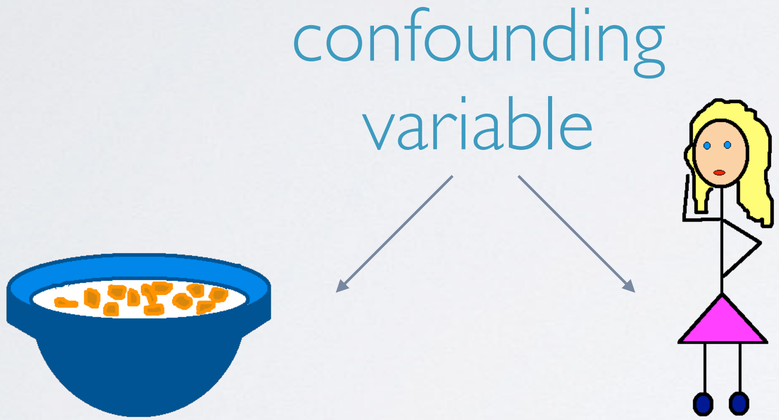


3. a third variable is responsible for both

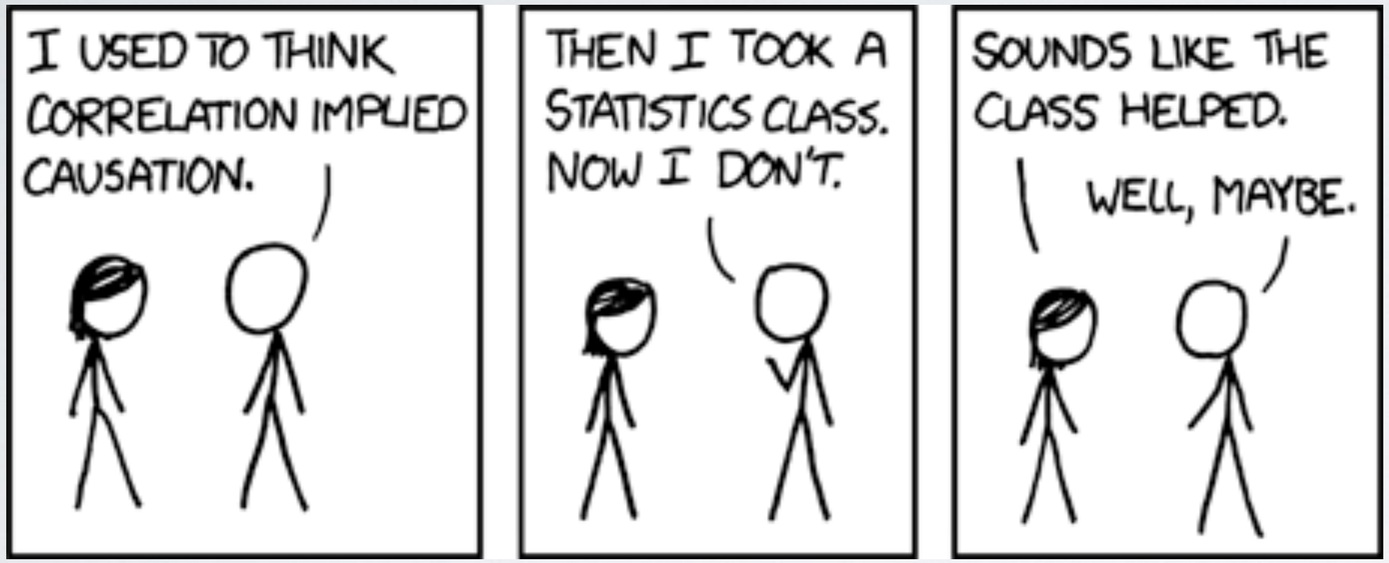


confounding variables

extraneous variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them



correlation does not imply causation



sampling & sources of bias

- census vs. sample
- sources of bias
- sampling methods

census

Wouldn't it be better to just include everyone and "sample" the entire population, i.e. conduct a 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.

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

npr

from KJZZ

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

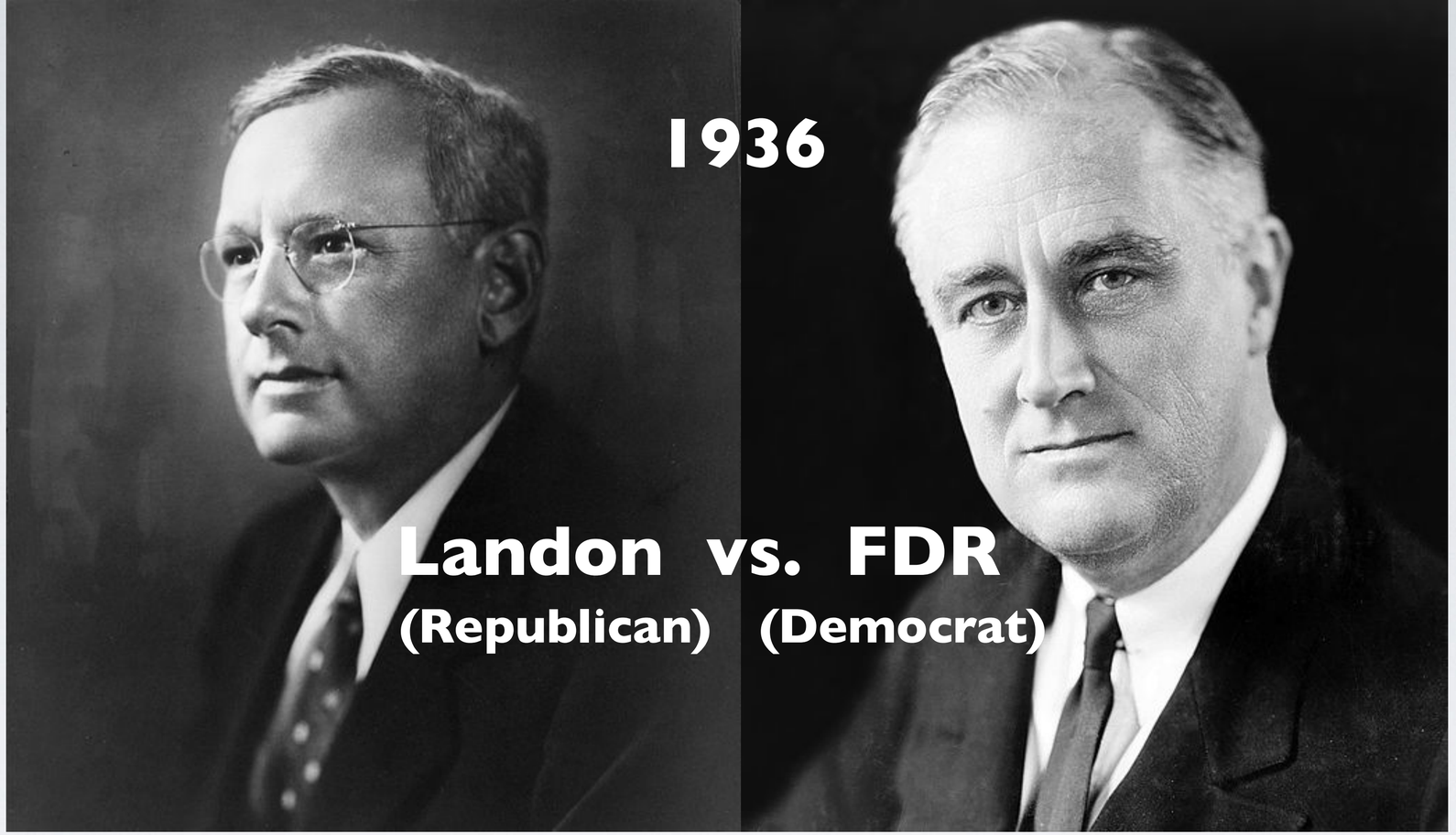
Listen to the NPR story at http://www.npr.org/templates/story/story.php?storyId=125380052

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

a few 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



The Literary Digest

Election results

Lose with 43% of the votes

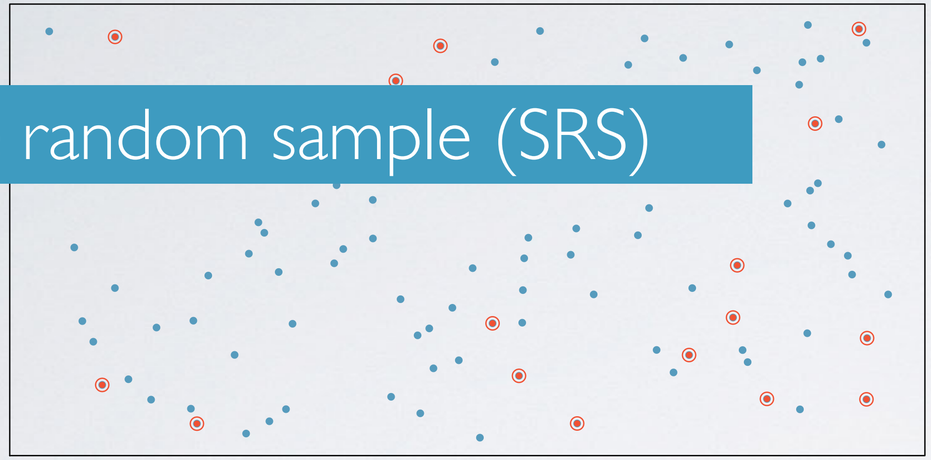
Win with 62% of the votes

Image sources: http://en.wikipedia.org/wiki/File:LandonPortr.jpg, http://en.wikipedia.org/wiki/File:FDR_in_1933.jpg, and http://en.wikipedia.org/wiki/File:LiteraryDigest-19210219.jpg

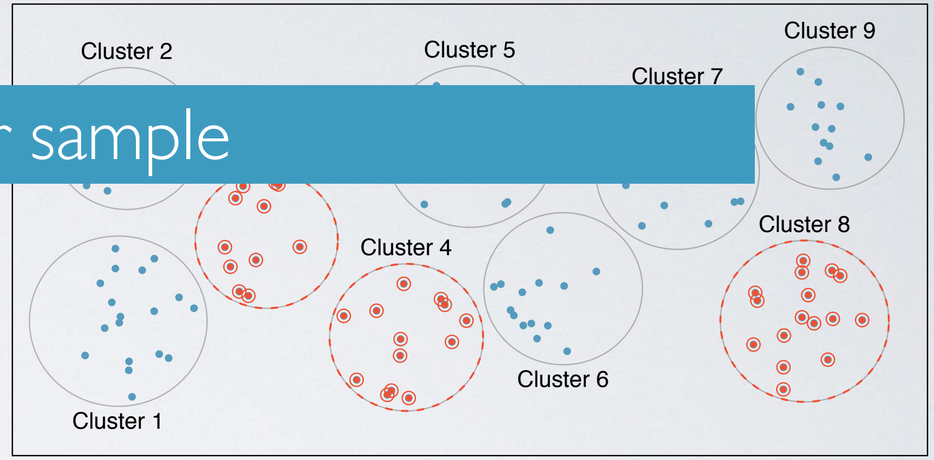
Image credit: Wonderlane CC BY 2.0 http://www.flickr.com/photos/wonderlane/6231888661

sampling methods

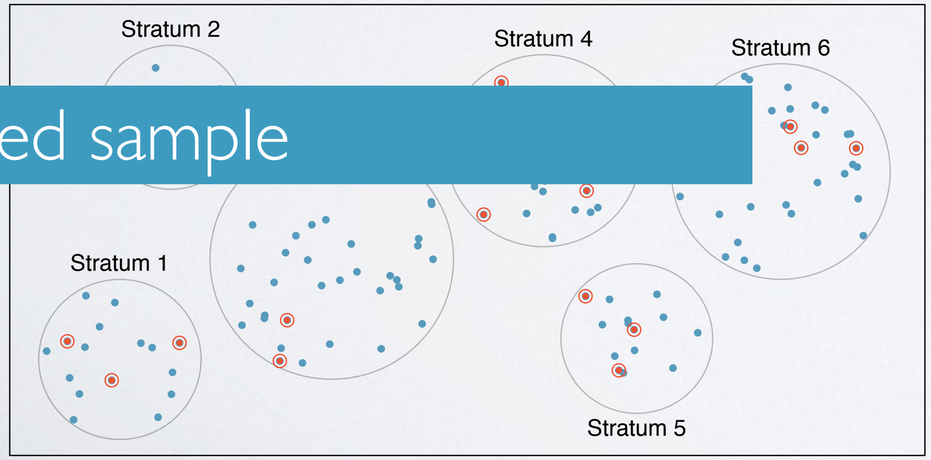
simple random sample (SRS)



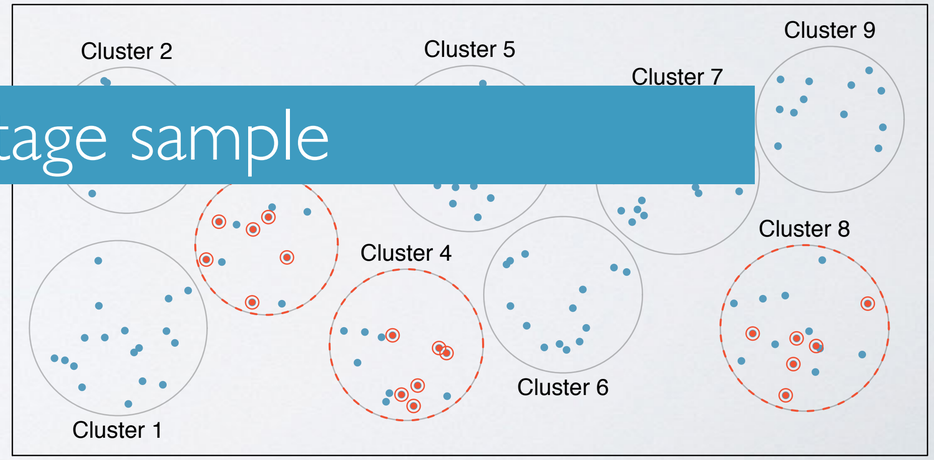
cluster sample



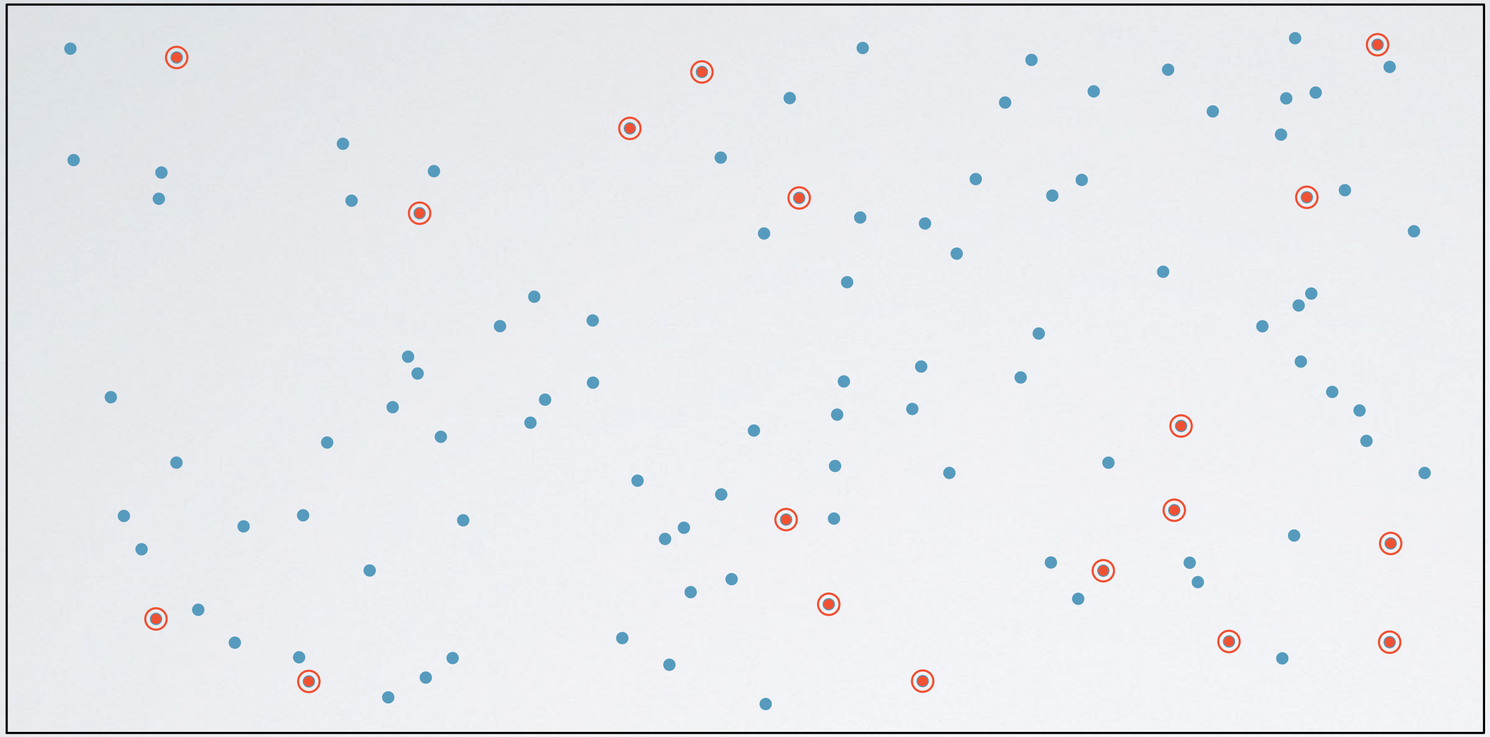
stratified sample



multistage sample

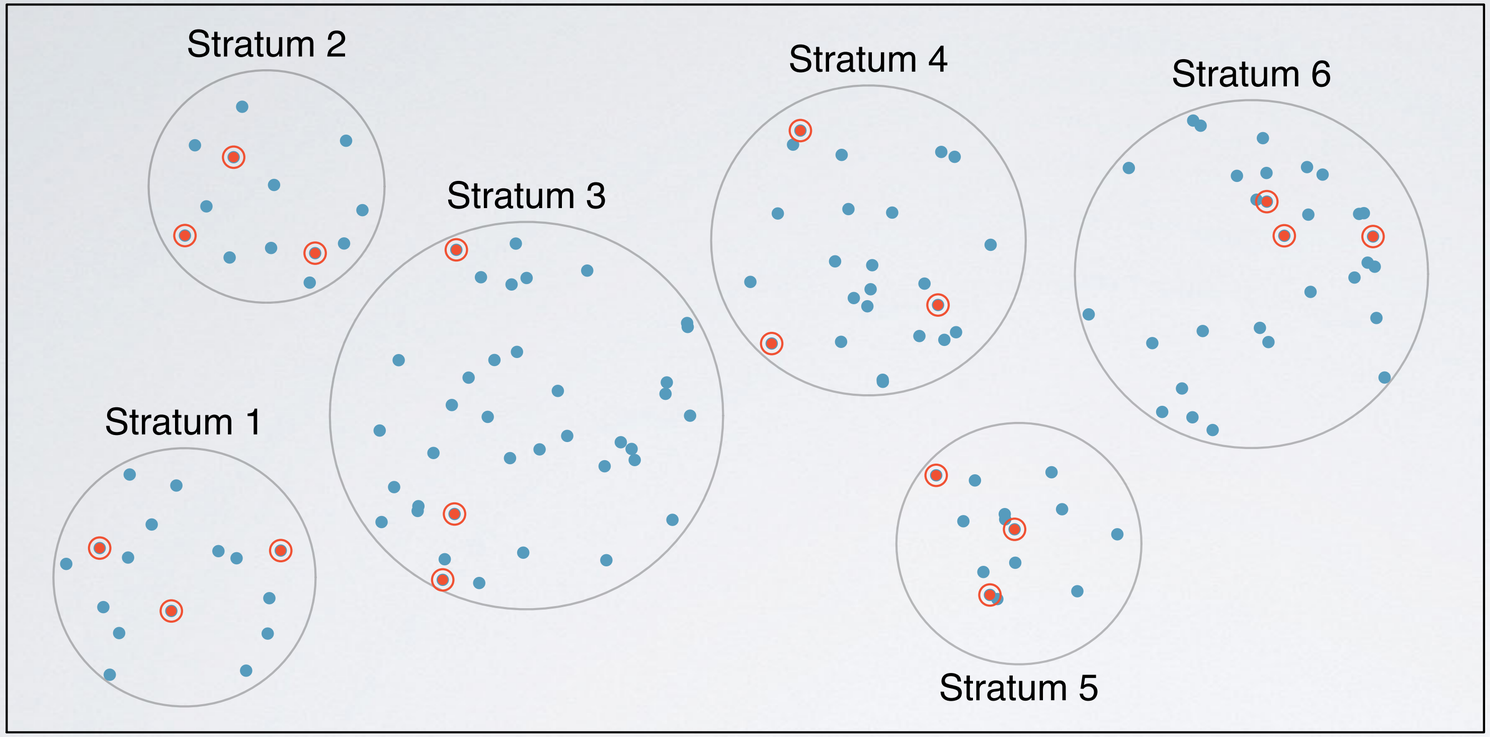


simple random sample (SRS)



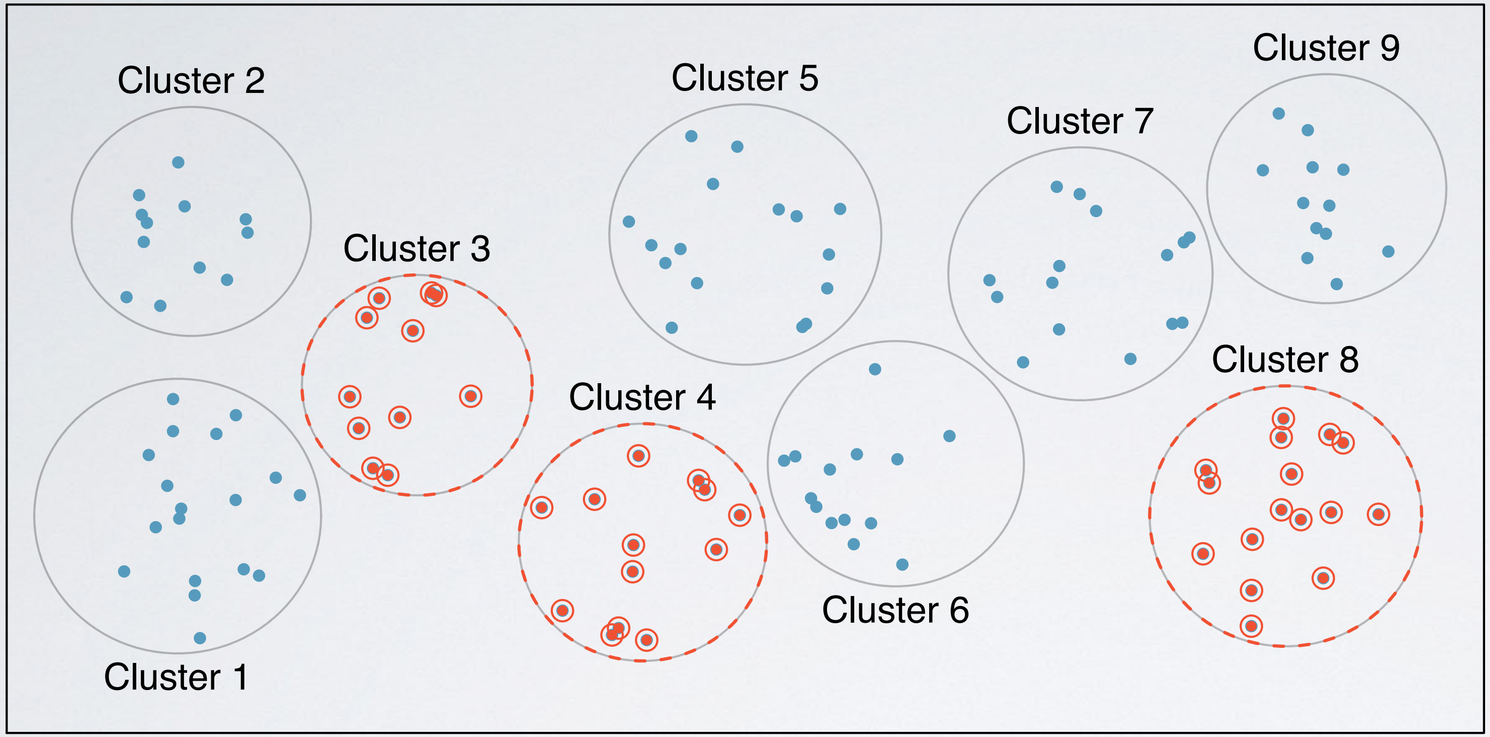
each case is equally likely to be selected

stratified sample



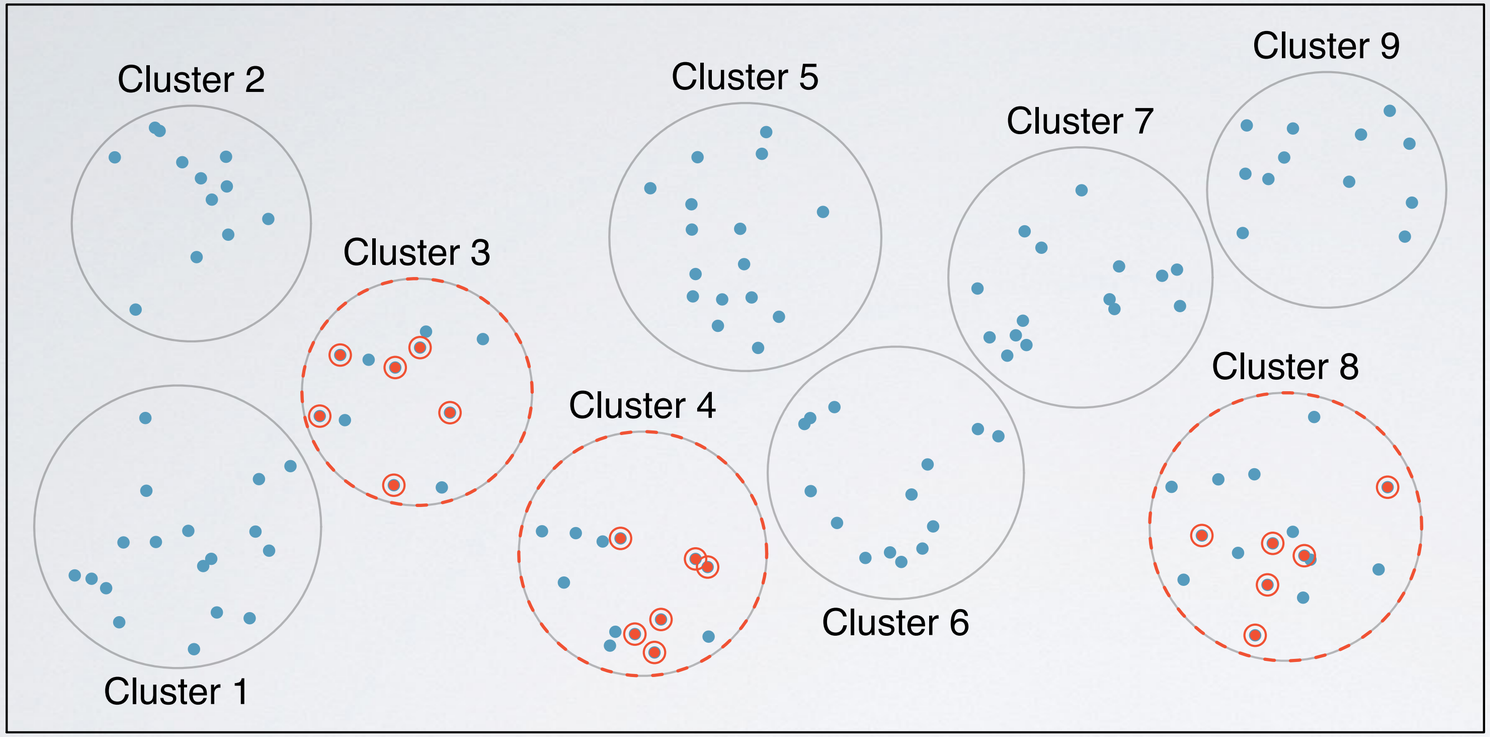
divide the population into homogenous strata,
then randomly sample from within each stratum

cluster sample



divide the population clusters,
randomly sample a few clusters,
then sample all observations within these clusters

multistage sample



divide the population clusters,
randomly sample a few clusters,
then randomly sample within these clusters

experimental design

- principles of experimental design
- experimental design terminology

principles of experimental design

(1) control

compare treatment of interest to a control group

(2) randomize

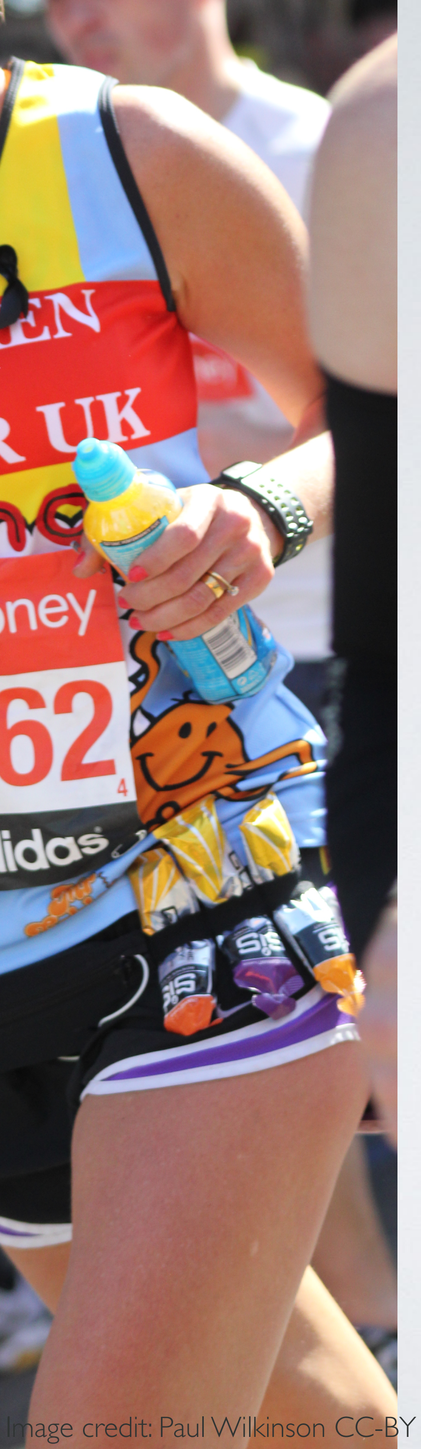
randomly assign subjects to treatments

(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

experimental terminology

placebo

fake treatment, often used as the control group for medical studies

placebo effect

showing change despite being on the placebo

blinding

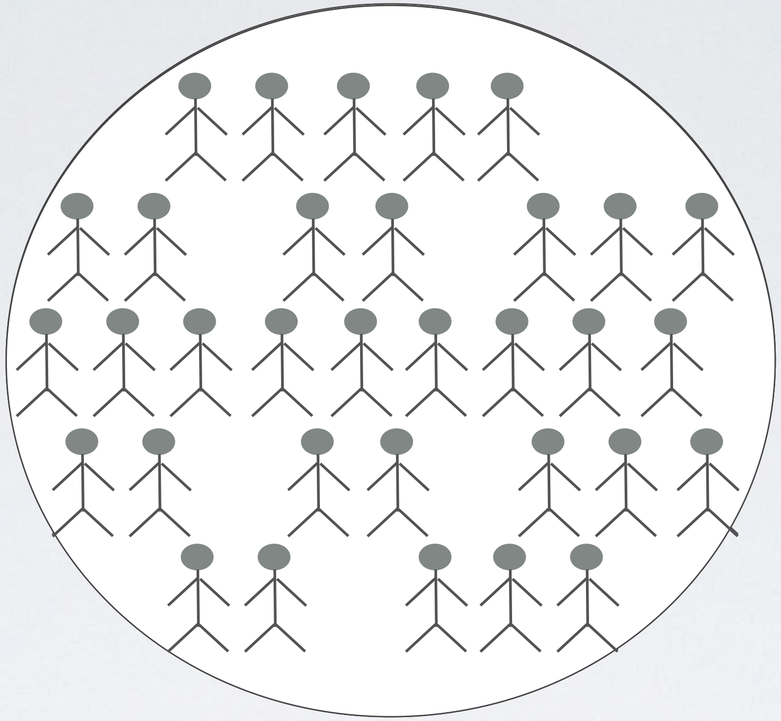
experimental units don't know which group they're in

double-blind

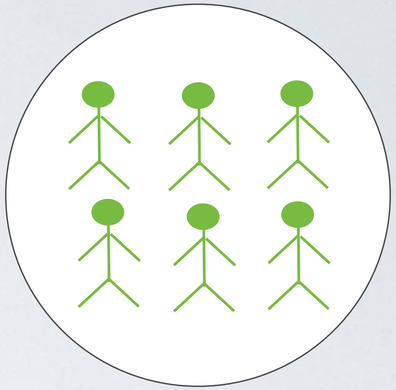
both the experimental units and the researchers don't know the group assignment

random sampling vs. assignment

random sampling



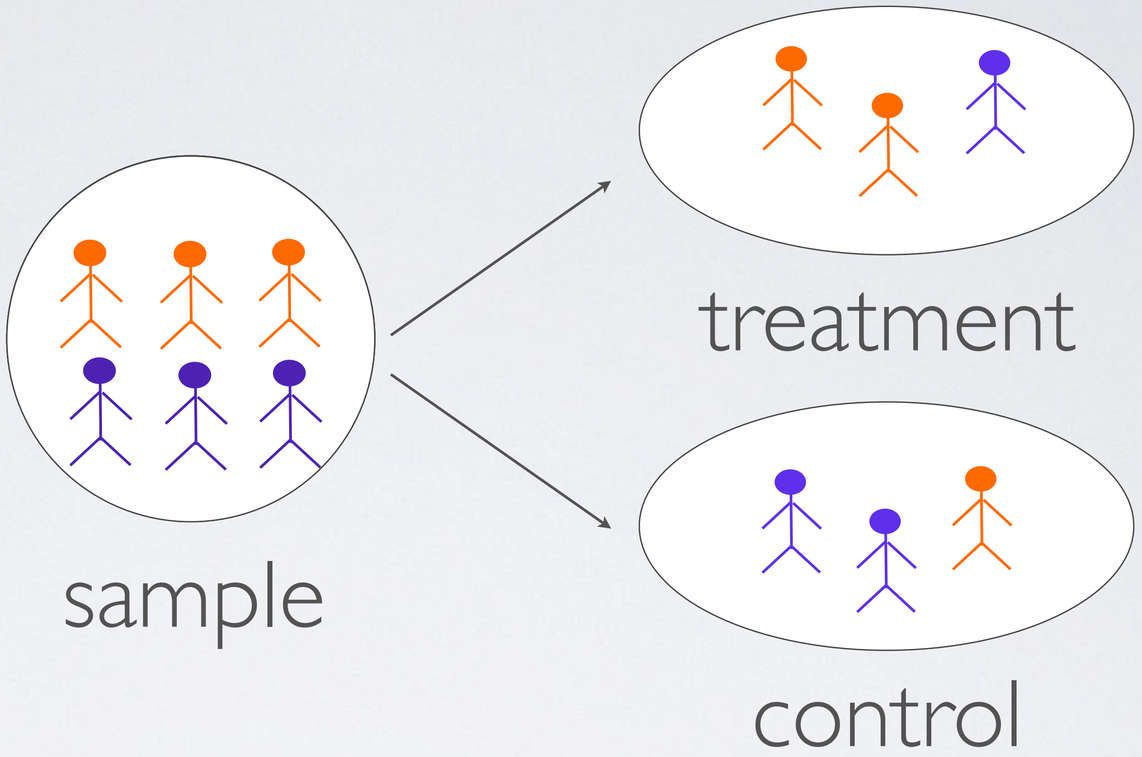
population



sample

generalizability

random assignment



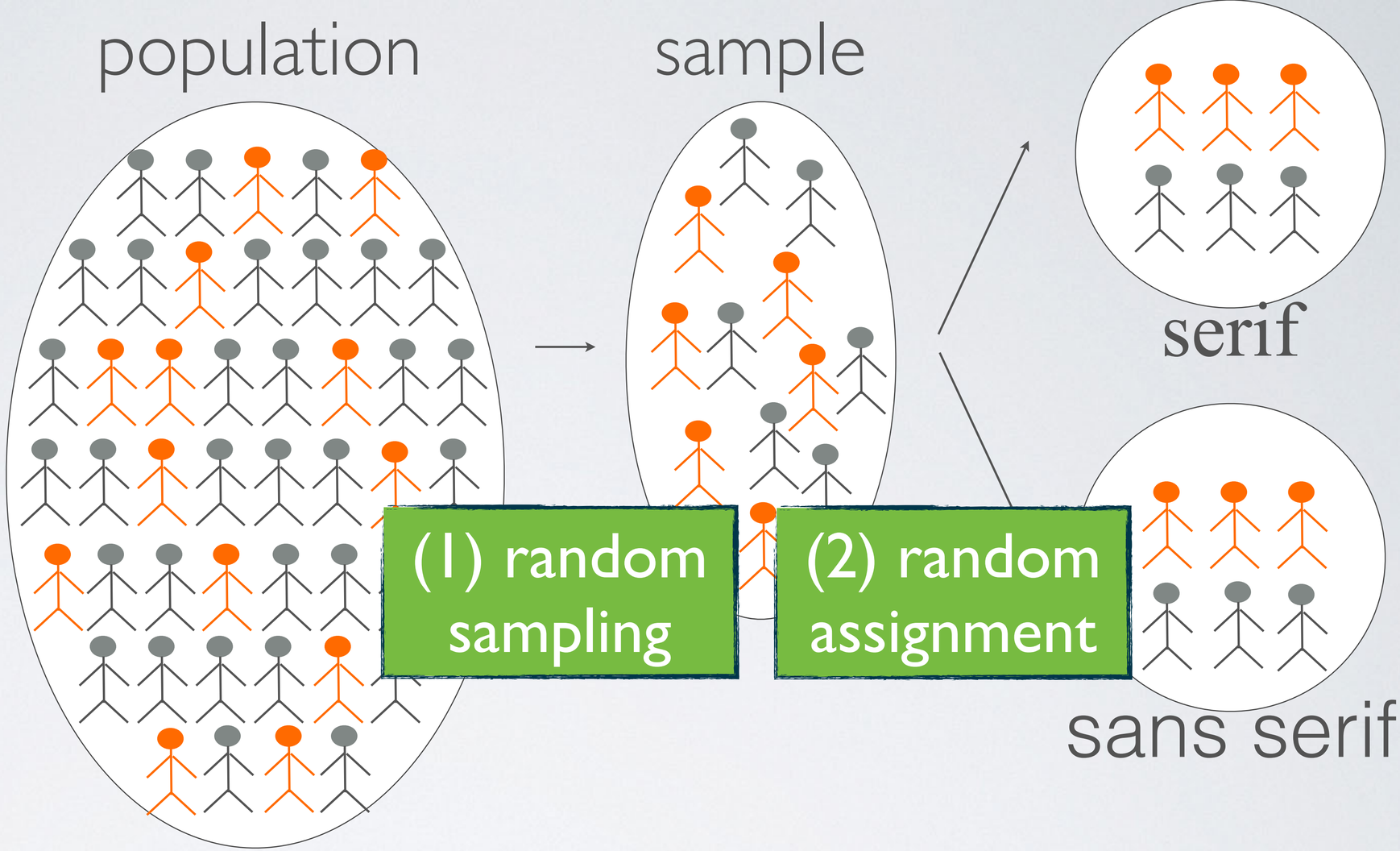
causality

example

Serif

Sans Serif

confounding variable



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
	C		

most experiments

DO YOU SUFFER FROM ASTHMA AND ALLERGIES?

Volunteers, who are non-smokers, diagnosed

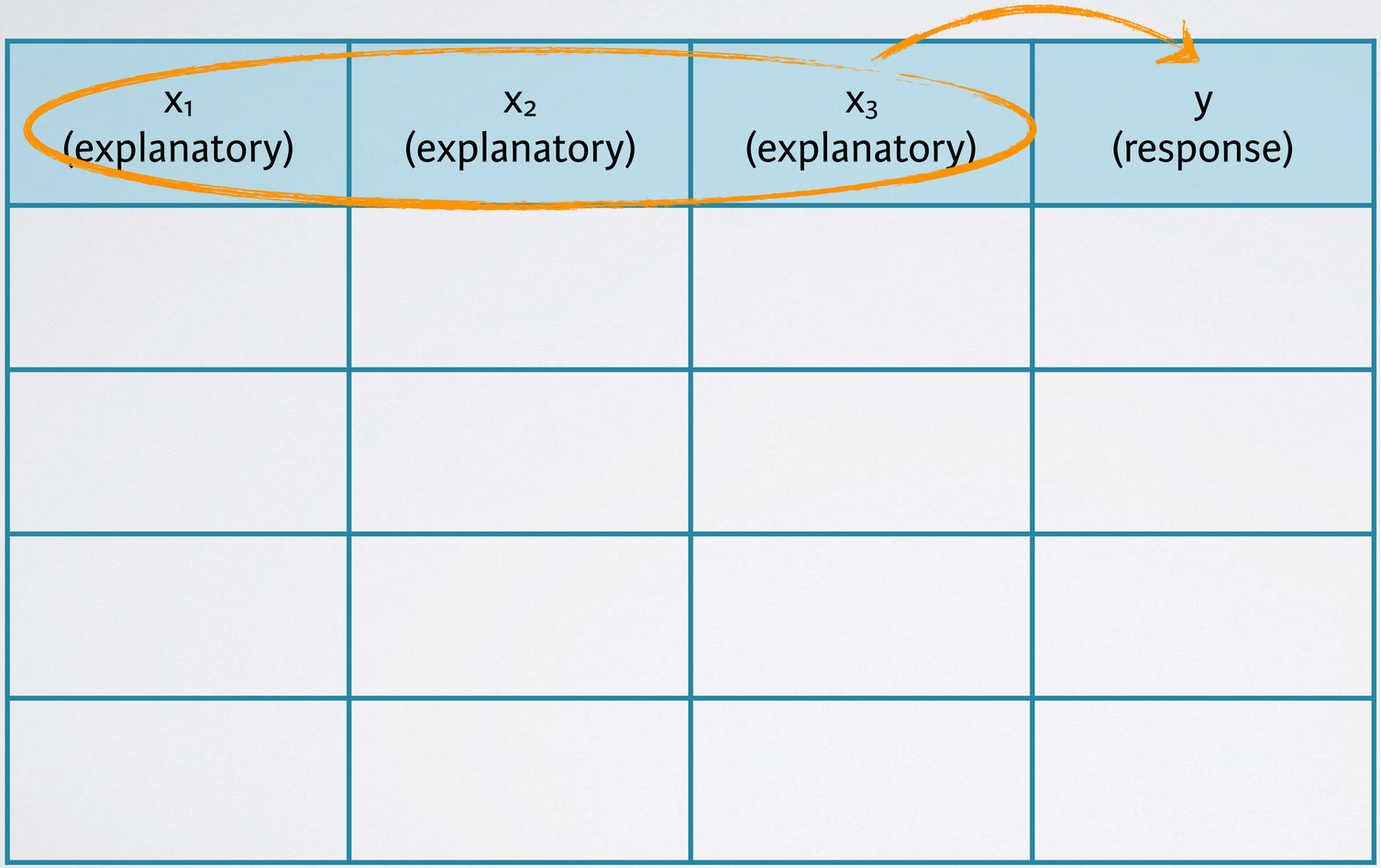
exercise

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

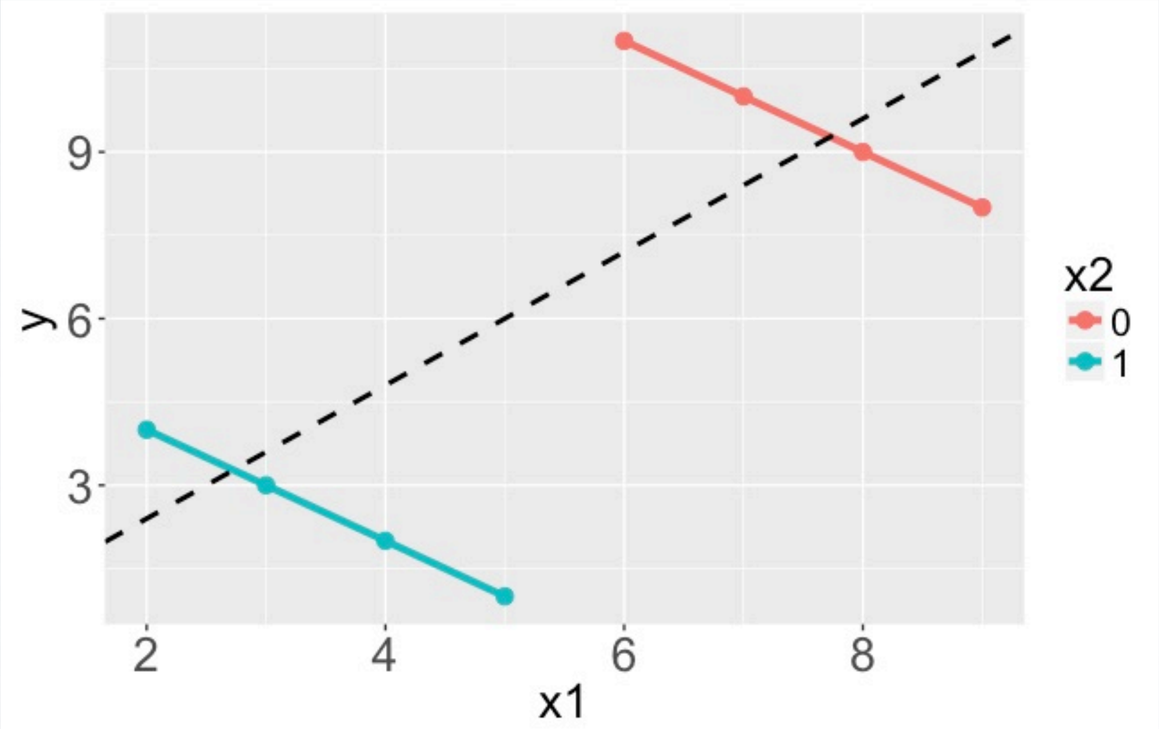
multivariate relationships



multivariate relationships

calories (explanatory)	age (explanatory)	fitness (explanatory)	heart health (response)

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

exercise

berkeley admission data

	Admitted	Rejected
Male	1198	1493
Female	557	1278