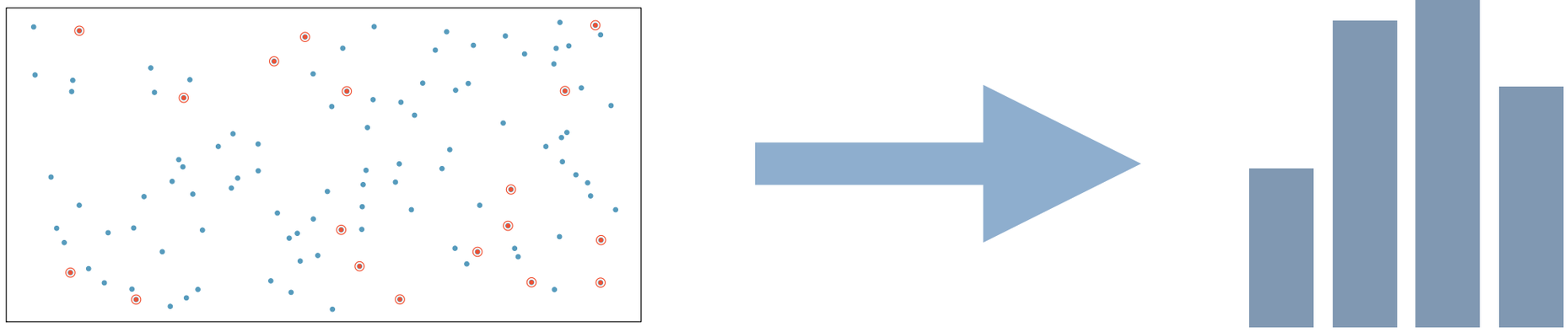


Collecting and summarizing data

From Data to Insight

Dr. Çetinkaya-Rundel
July 8, 2016

~~Data can be misleading.~~

It is possible to summarize and visualize data in a misleading way.

“It is easy to lie with statistics. It is hard to tell the truth without it.”

–Andrejs Dunkels

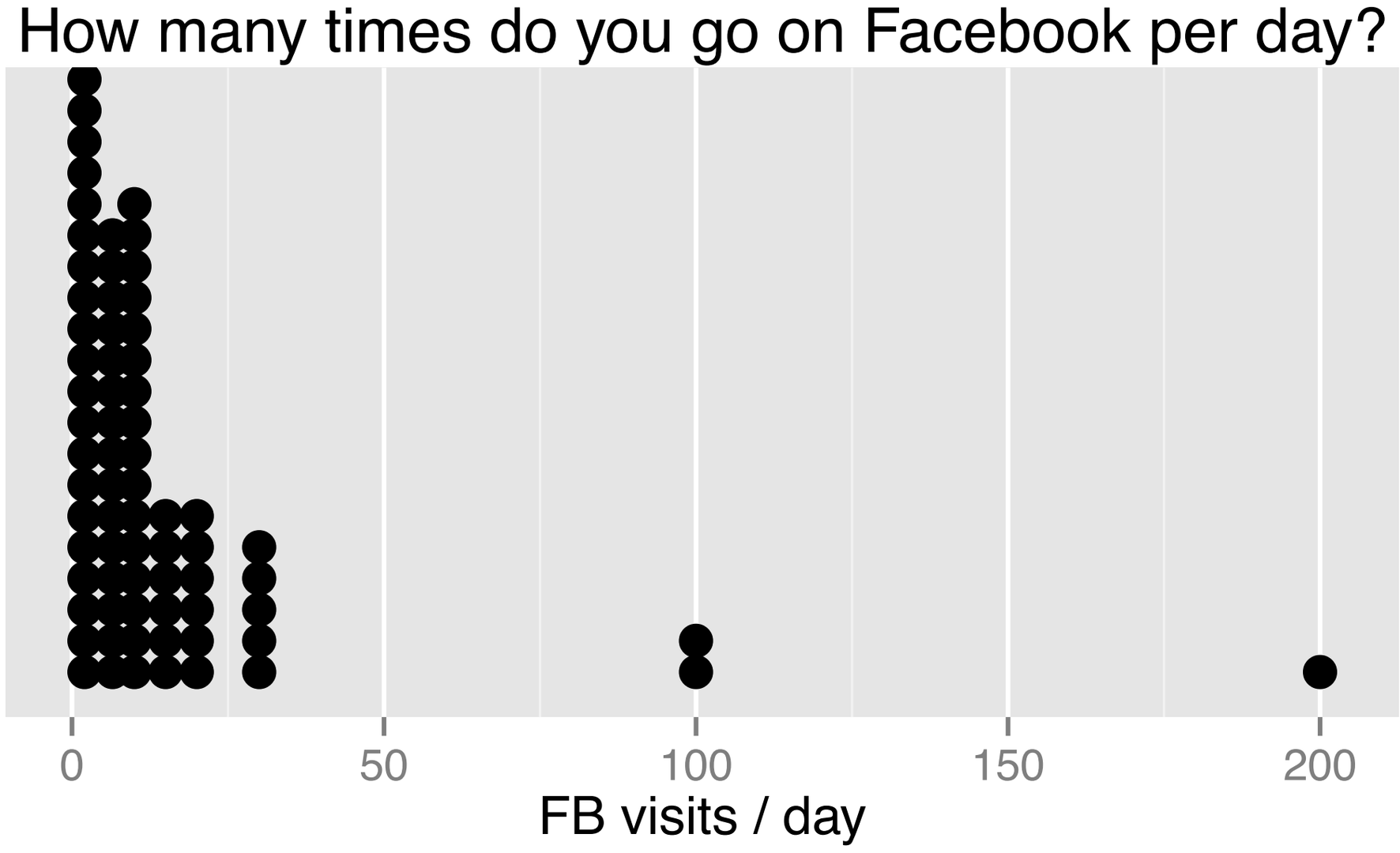
“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”

–H. G. Wells

Always start your exploration with a visualization!

Do you see anything out of the ordinary?

How are people reporting higher vs. lower values of FB visits?



Use the appropriate measure of central tendency

Which of these is most likely to have a roughly symmetric distribution?

(a) salaries of a random sample of people from NC

(b) weights of adult females

(c) scores on an well-designed exam

(d) last digits of phone numbers

How do the mean and median of these two datasets compare?

Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

(a) mean1 = mean2, median1 = median2

(b) mean1 < mean2, median1 = median2

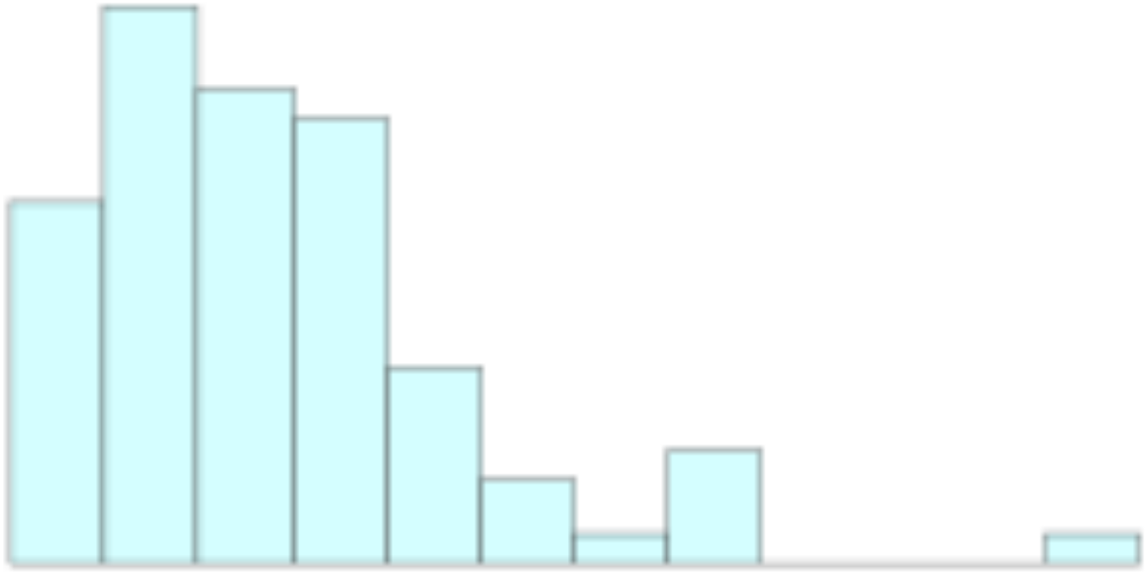
(c) mean1 < mean2, median1 < median2

(d) mean1 > mean2, median1 < median2

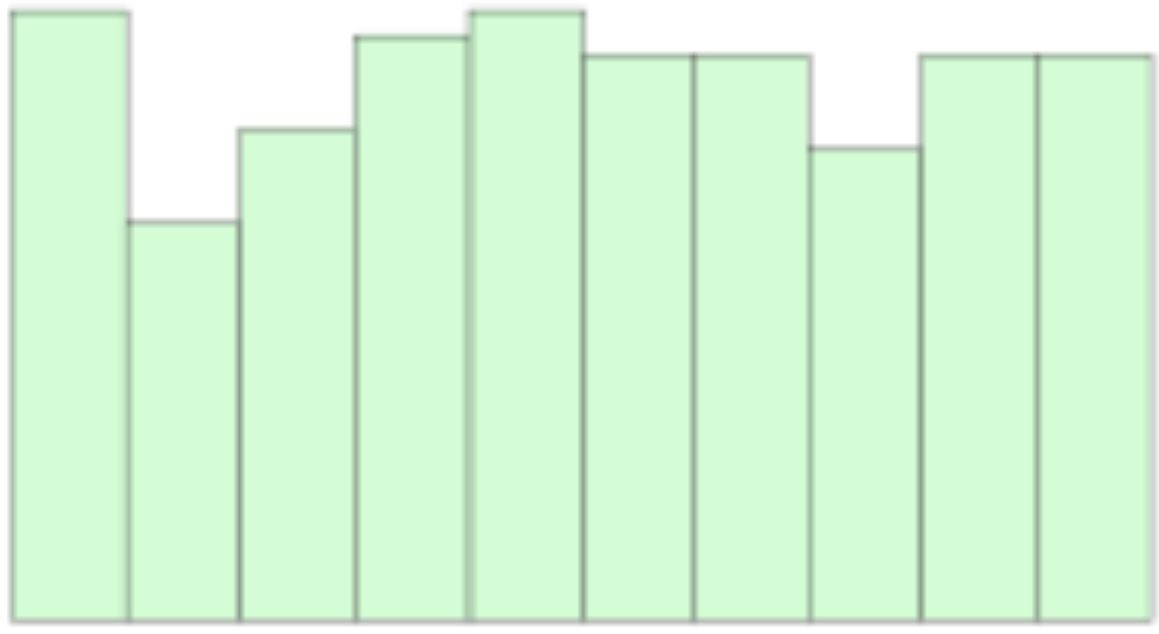
(e) mean1 > mean2, median1 = median2

Which histogram corresponds to the age at which a sample of people applied for marriage licenses and which to the last digit of a sample of social security numbers?

(a)

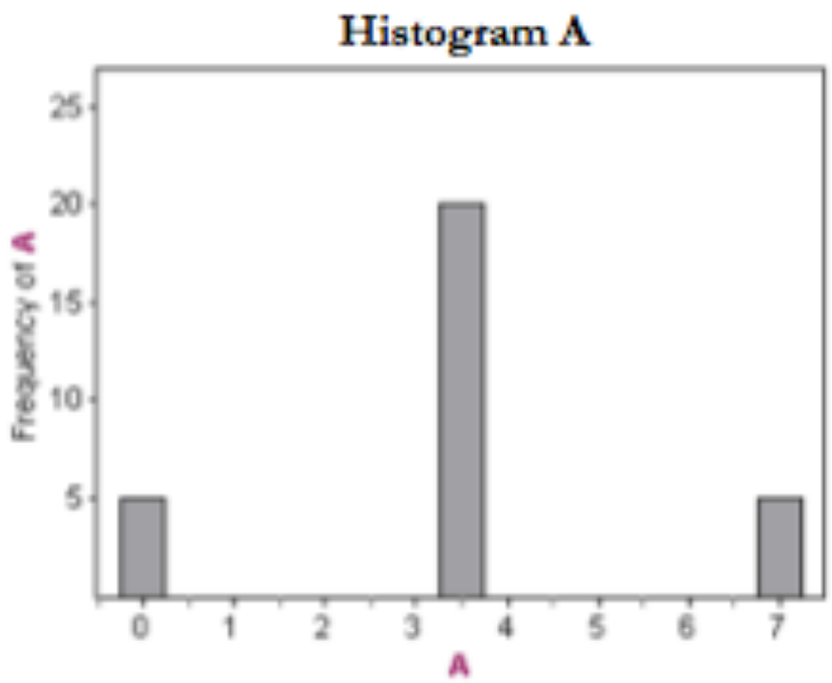


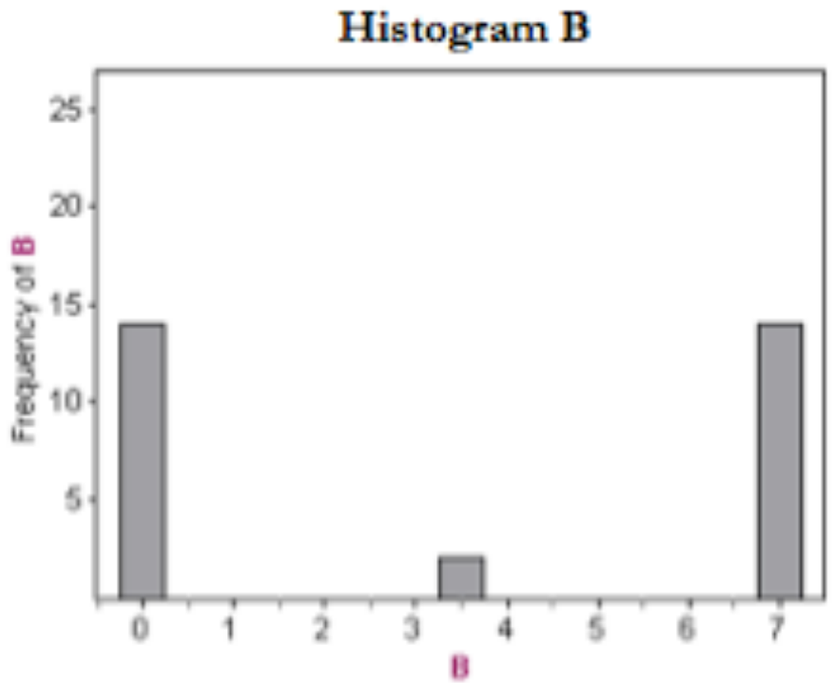
(b)

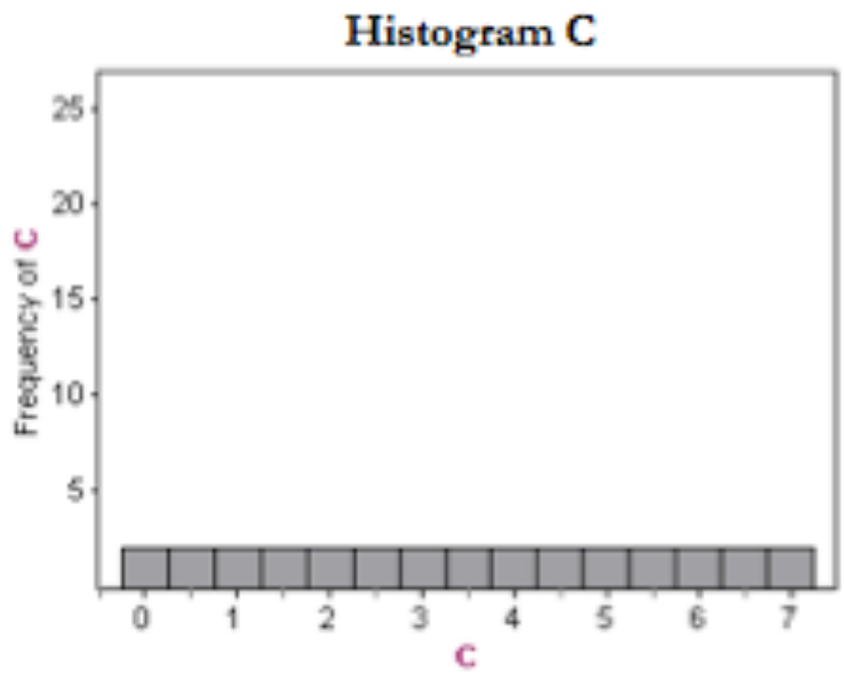


Variability is measured as average deviation from the mean

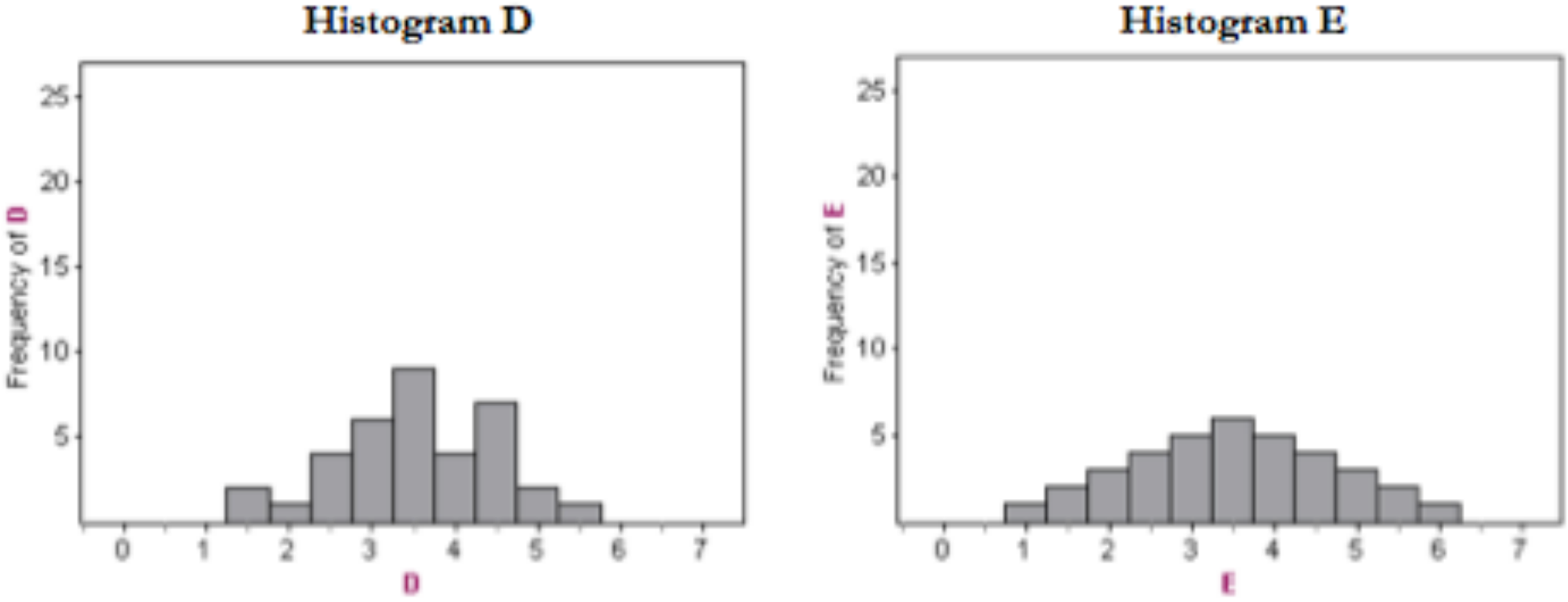
Order histograms from least to most variable.







Which histogram exhibits more variability?

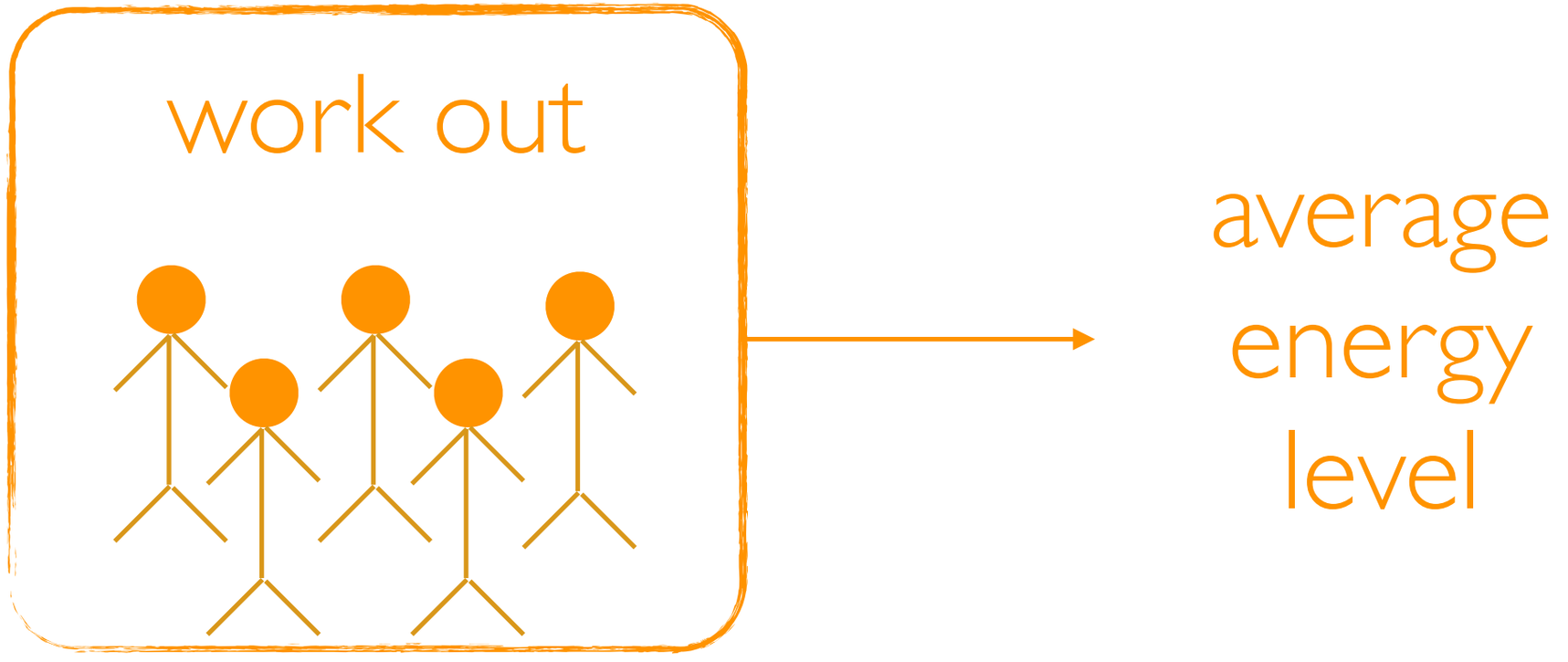


Correlation vs. causation & types of studies

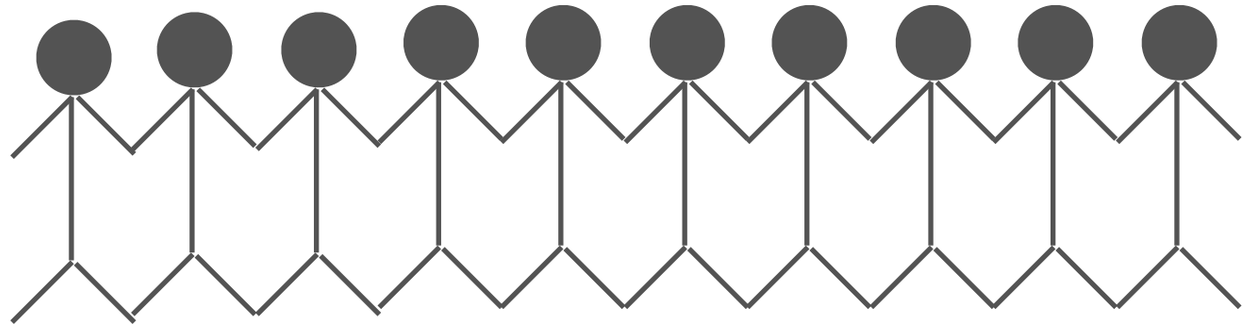
Correlation ≠ causation

- But in certain circumstances it does!
- If the data come from a randomized experiment and a correlation is found, this might also suggest a causation between the variables studied.
  - **Experiment:** Researchers randomly assign subjects to treatments
- If the data come from an observational study and a correlation is found, this does **not** also suggest a causation between the variables studied.
  - **Observational study:** Collect data in a way that does not directly interfere with how the data arise ("observe")

observational study



experiment



random assignment

work out

average energy level

don't work out

average energy level

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.

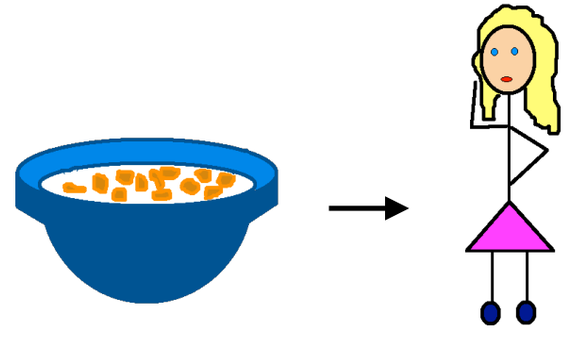
[…]

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

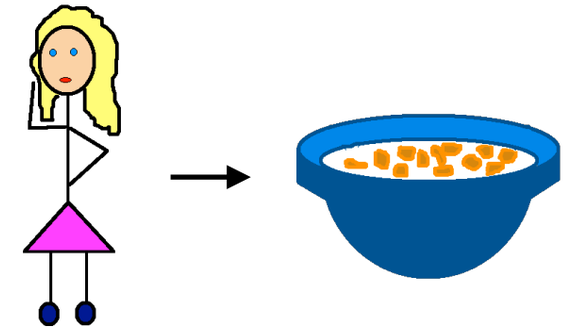
[…]

3 possible explanations

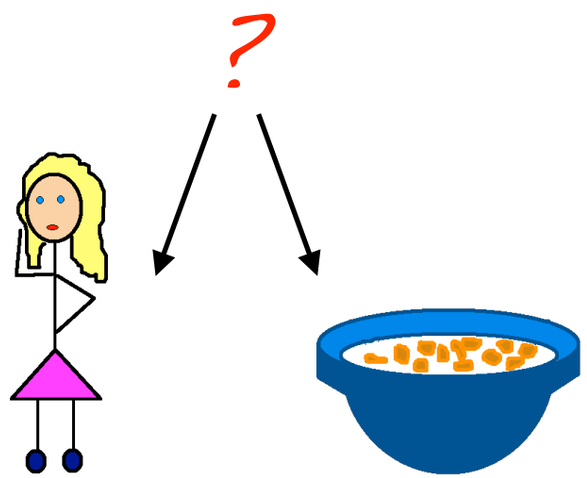
1. eating breakfast causes girls to be slimmer



2. being slim causes to eat breakfast

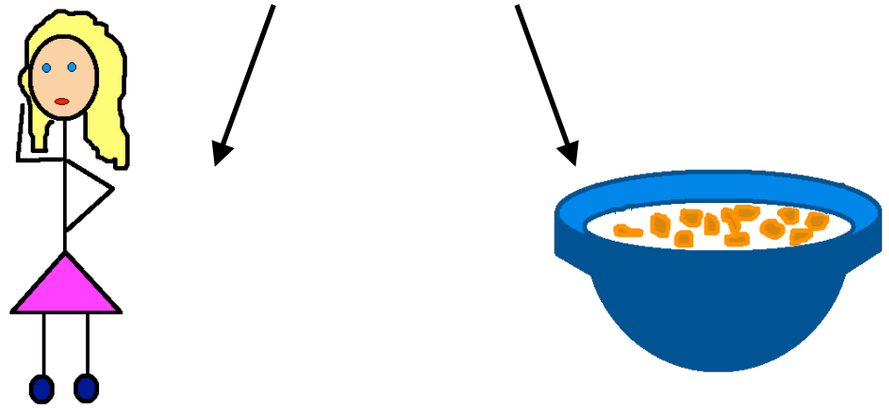


3. a third variable is responsible for both



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



Stress and muscle cramps

- A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

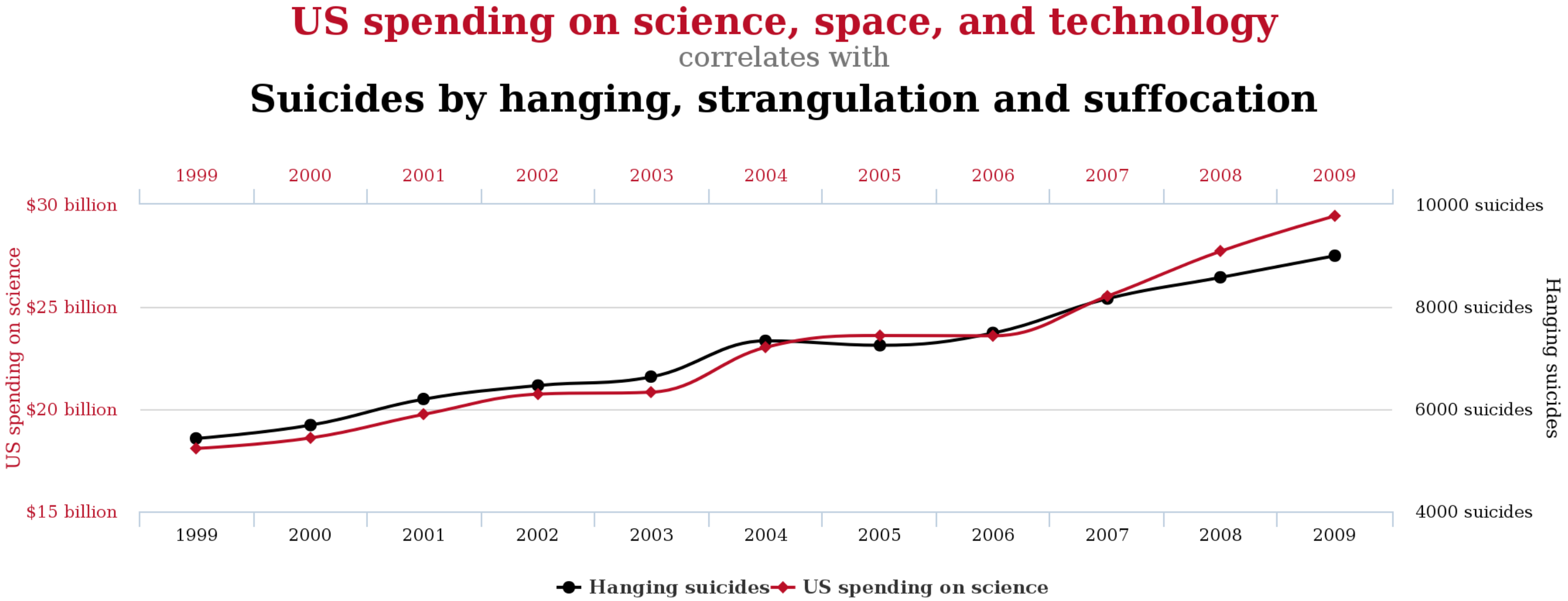
- What is the conclusion of the study?

- Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

Stress and muscle cramps, revisited

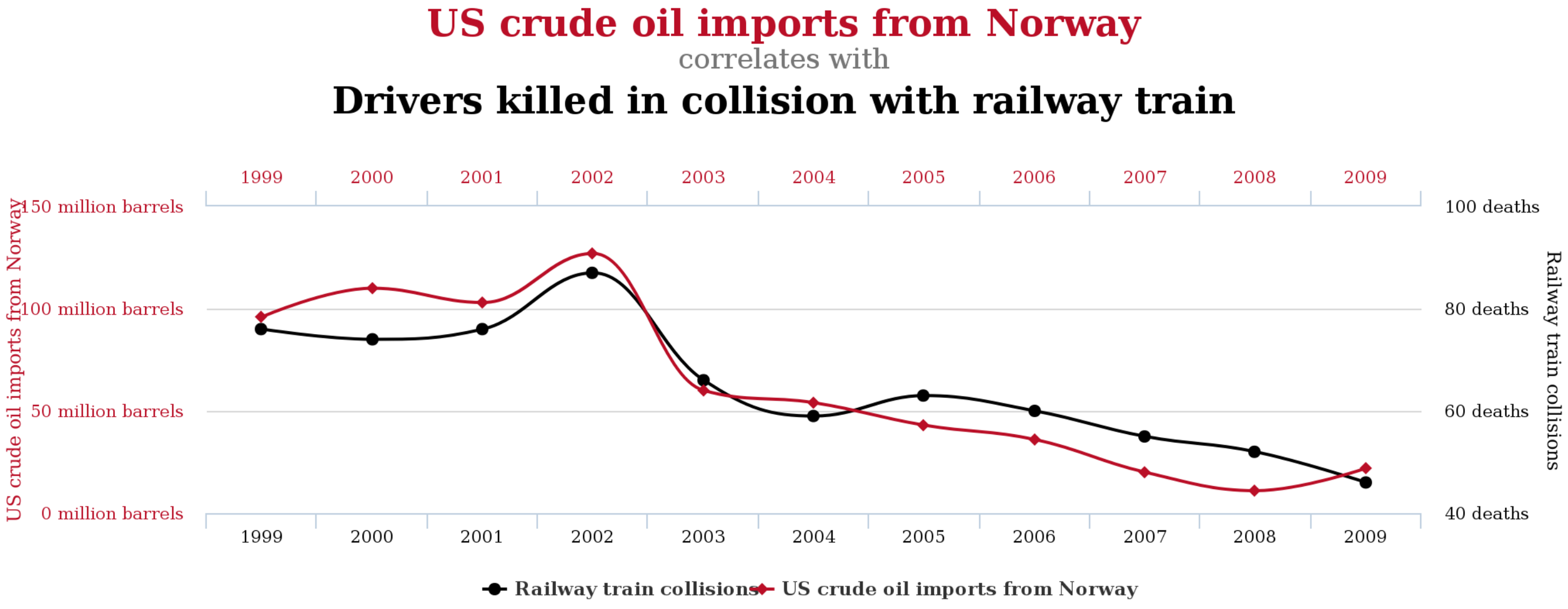
- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people:
  - **Block** for age

Correlation ≠ causation



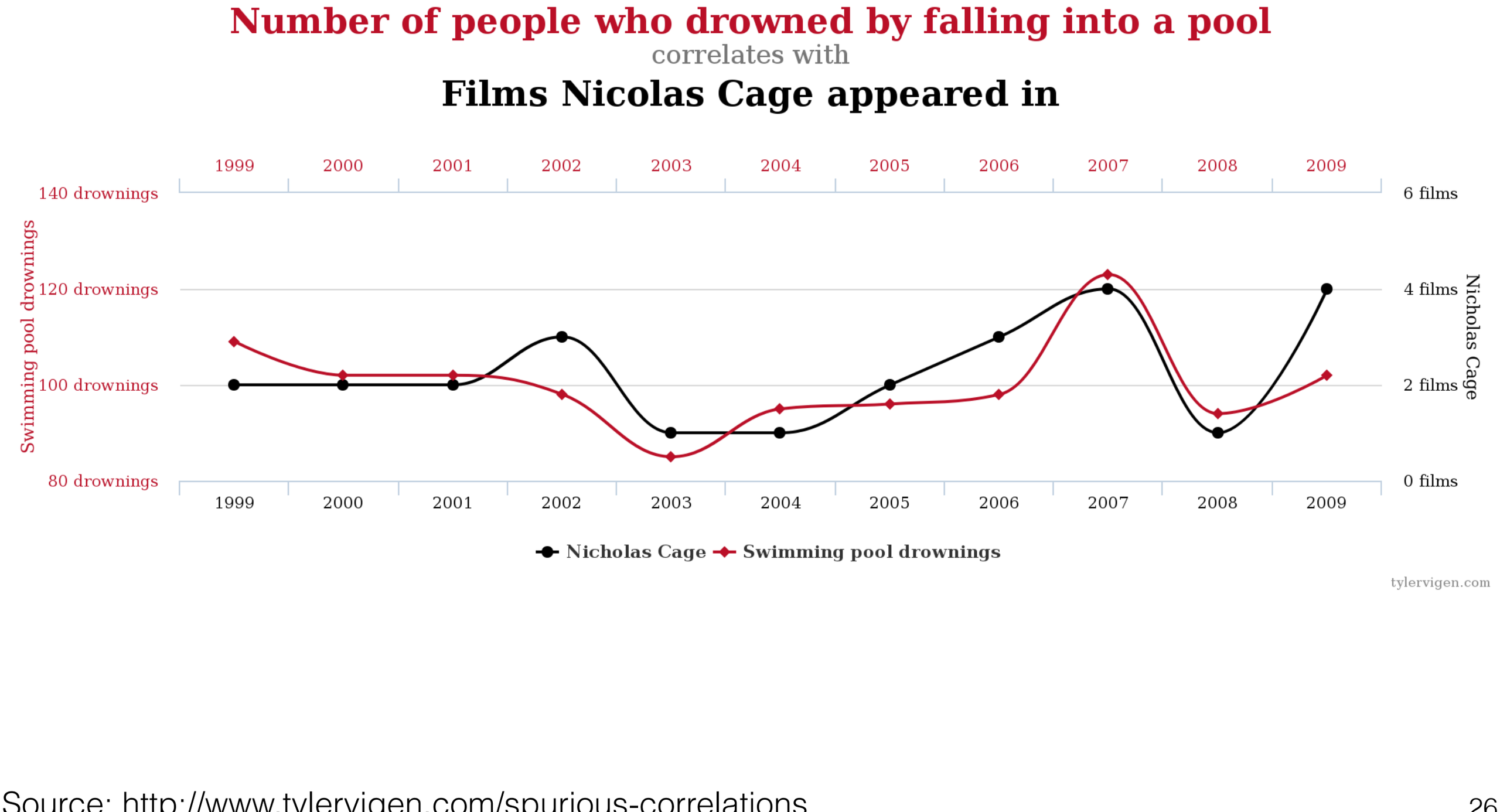
Source: http://www.tylervigen.com/spurious-correlations

Correlation ≠ causation

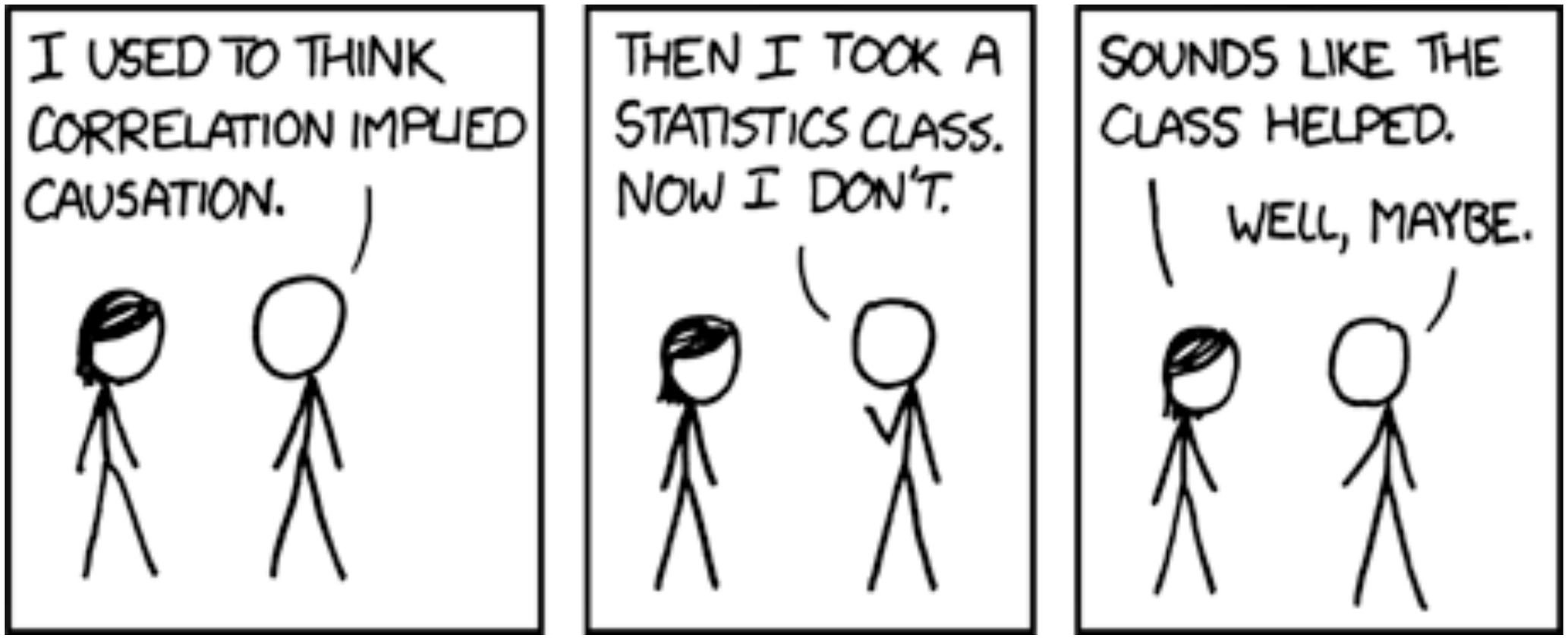


Source: http://www.tylervigen.com/spurious-correlations

Correlation ≠ causation



Source: http://www.tylervigen.com/spurious-correlations



Source: http://xkcd.com/552/

Sampling, and sampling biases

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.

Source: http://www.npr.org/templates/story/story.php?storyId=125380052

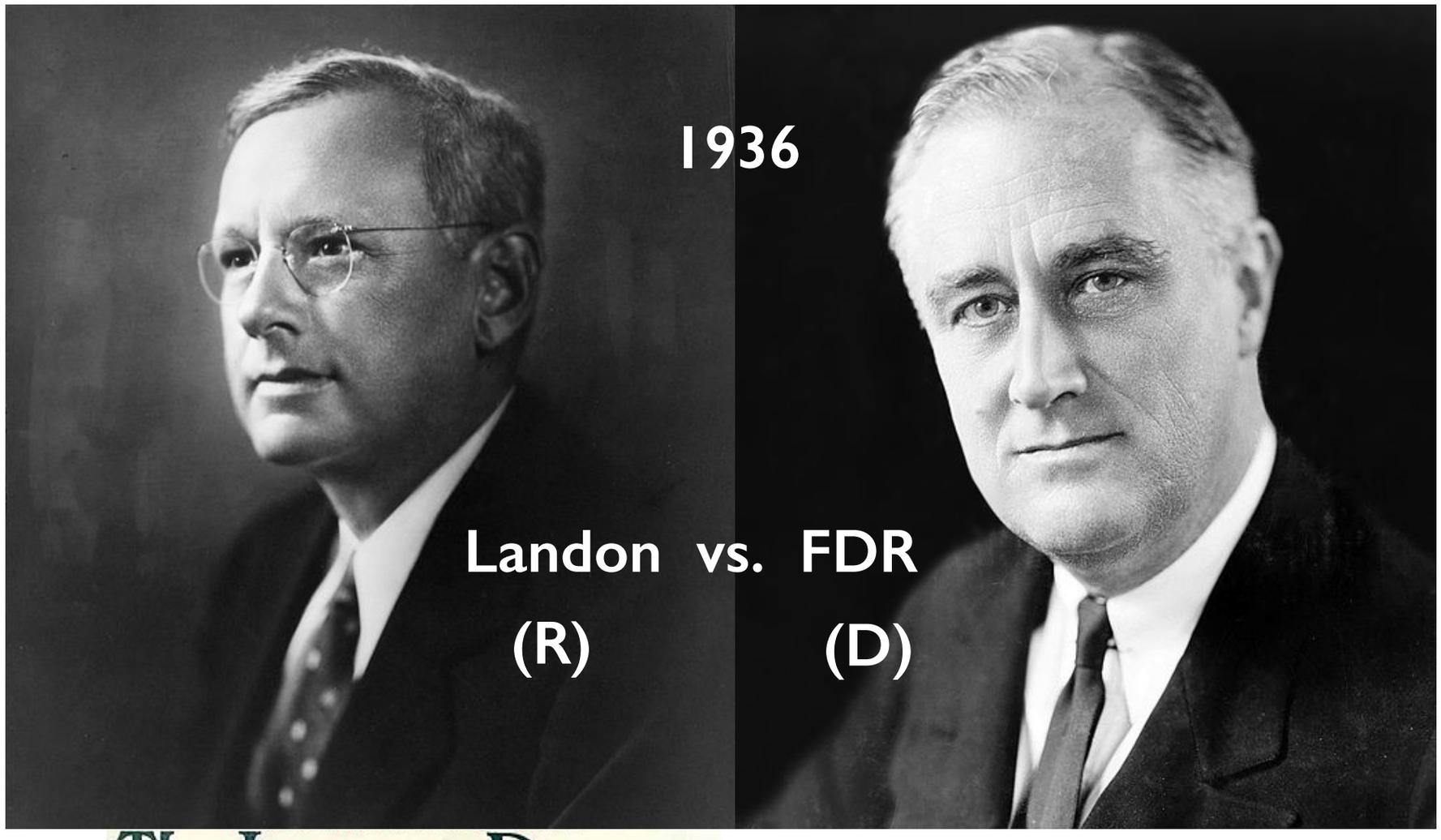
Sampling is natural

- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's **exploratory analysis**.

- If you generalize and conclude that your entire soup needs salt, that's an **inference**.

- For your inference to be valid, the spoonful you tasted (the sample) needs to be **representative** of the entire pot (the population).

Garbage in, garbage out!



The Literary Digest

Election results

Lose with 57% of the votes

Win with 60% of the votes