# YData: An Introduction to Data Science

## Lecture 20: Causality

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

# Reminders

- Assignment 06 due on Thursday
- Midterm next Friday
  - Practice midterm posted; second posted on Gradescope
  - Review session on Monday (no new topics next week)
  - Exam available Friday during normal class time (no class)
- Assignment 07 posted Friday; due April 1
- Break day next Wednesday (no class or OH)

# A/B Testing

- **Null**: The two samples are drawn randomly from the same underlying distribution.

- If the null is true, all rearrangements of the variable values among the two samples are equally likely. So:
  - compute the observed test statistic
  - then shuffle the values and recompute the statistic; **repeat**; compare with the observed statistic

# Deflategate

# 2015 AFC Championship Game

# Deflategate

**Wikipedia**:
The 2015 AFC Championship Game football tampering scandal, commonly referred to as Deflategate, or Ballghazi

...

# Null hypothesis

**The 4 Colts footballs are like a sample drawn at random without replacement from all 15 balls.**

- To test this hypothesis, repeat this process:
  - Randomly permute all 15 balls
  - Label 11 of them "Patriots" and the remaining 4 "Colts"
  - Compare the averages of the two groups

(DEMO)

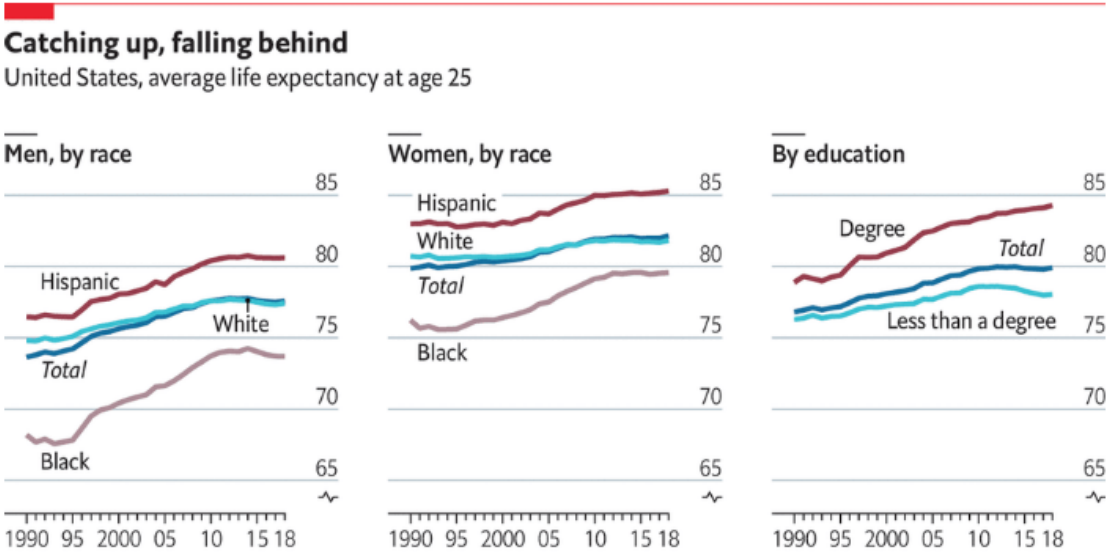
# Causality

# Recent study

Daily chart

## Educated Americans live longer, as others die younger

A college degree is becoming the main factor accounting for the difference in expected lifespans




Source: "Life expectancy in adulthood is falling for those without a BA degree, but as educational gaps have widened, racial gaps have narrowed" by Anne Case and Angus Deaton, PNAS, 2021.

The Economist

# Recent study

- Is there a *causal* connection between education and mortality?
- What are some of the difficulties in answering this question?
- What are some potential confounding variables?

# Randomized Controlled Experiment

- Sample A: **control group**
- Sample B: **treatment group**

- **The treatment and control groups are selected at random; this allows causal conclusions.**

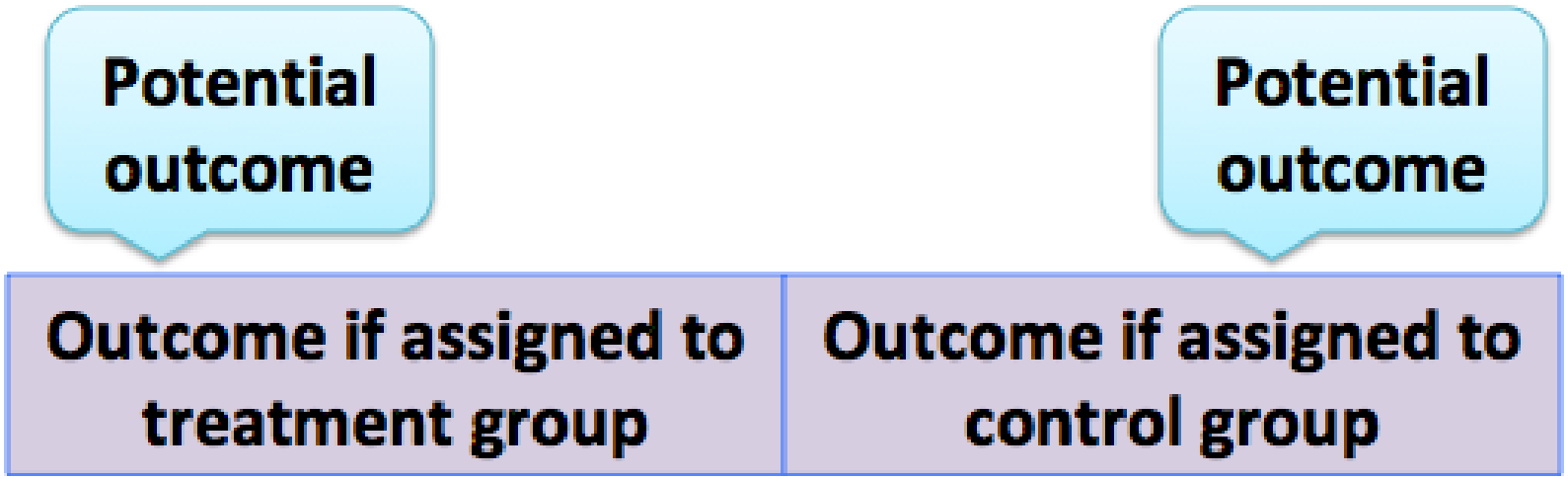
- Any difference in outcomes between the two groups could be due to
  - chance
  - the treatment

# Case study

- RCT to study Botulinum Toxin A (BTA) as a treatment to relieve chronic back pain
- 15 patients in the treatment group (received BTA)
- 16 in the control group (normal saline)
- Trials were run double-blind (neither doctors nor patients knew which group they were in)
- Only 2 patients in the control group had relief from pain (outcome=1); 9 patients in the treatment group had relief.
- Can this difference be just due to chance?

# Potential Outcomes

- In the population there is one imaginary ticket for each of the 31 participants in the experiment.
- Each participant's ticket looks like this:



# The Data

16 randomly picked tickets show:

| | Outcome if assigned to control group |
|---|---|

The remaining 15 tickets show:

| Outcome if assigned to treatment group | |
|---|---|

# The Hypotheses

- **Null**:
  - The distribution of all 31 potential control scores is the same as the distribution of all 31 potential treatment scores.

- **Alternative**:
  - The distribution of all 31 potential control scores is different from the distribution of all 31 potential treatment scores.

We only observe a sample from each set of potential outcomes

(DEMO)

# AstraZeneca Vaccine

Science & technology

Vaccination vacillation

## EU countries pause AstraZeneca's covid-19 jab over safety fears

An abundance of caution could well backfire

# AstraZeneca Vaccine (from The Economist)

- On March 15th France, Germany and Italy announced they were halting use of the AstraZeneca vaccine.
- Why? A Norwegian medical regulator reported four cases of blood clotting in adults given the vaccine. Similar–and similarly small—reports have come from Denmark, Italy and Austria.
- The World Health Organization (WHO) and European Medicines Agency (EMA) said they have no reason to believe the vaccine is unsafe.

Any confusion lies in the difficulty of disentangling causation from correlation. The EMA reckons that as of March 10th there had been 30 "thromboembolic events" among the 5m people in the EU who have received AZ's vaccine. By itself, that is no more remarkable than the fact that some of them will have suddenly had relief from chronic back pain or seen their cancer go into remission. The question is whether the rates are higher than would otherwise be expected. With clots, even as evidence from specific cases needs investigating, the mass of overall data shows that the vaccines are safe.

The Economist

# Discussion Question: A/B Testing of Covid Data

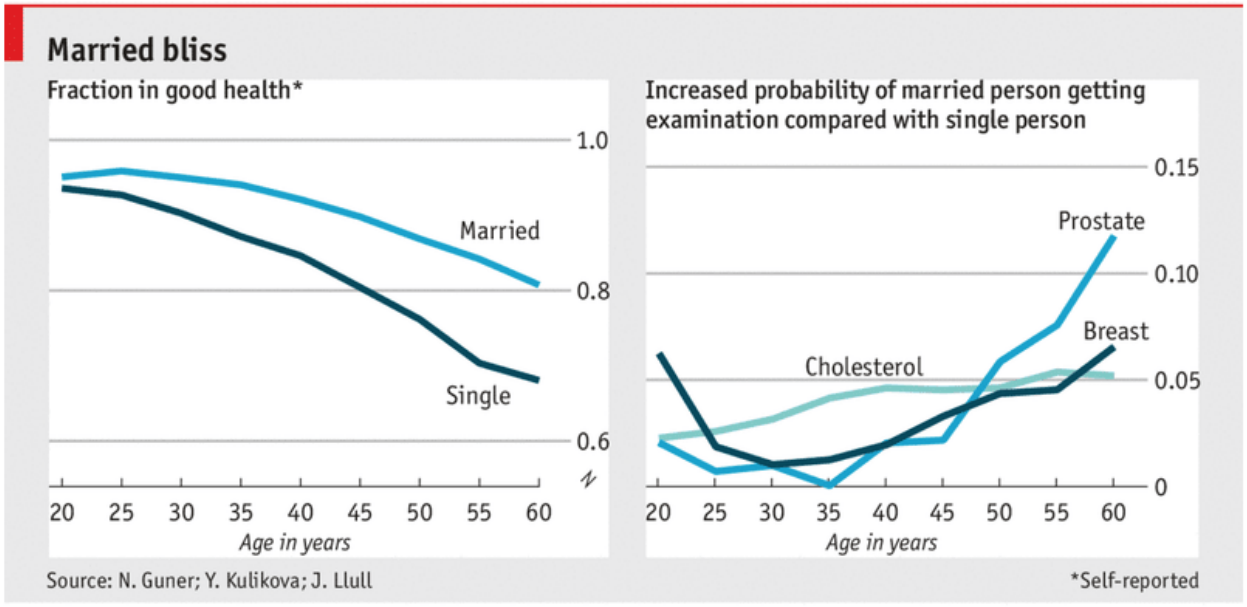
Suppose that clinical trials of the AstraZeneca vaccine resulted in data of this form:

| Treatment | Symptoms | Thrombosis |
|---|---|---|
| Placebo | False | False |
| Vaccine | False | False |
| Placebo | True | False |
| Vaccine | False | False |
| Vaccine | True | False |
| Placebo | False | False |
| Vaccine | False | False |
| Vaccine | False | False |

... (9992 rows omitted)

How would we perform an A/B test to decide whether or not there is excess risk of blood clotting, compared with random chance?

# Many other examples...




The link between marriage and better health is well established. Less clear is whether marriage causes good health or vice versa; healthy people may simply be more likely to marry in the first place. A group of researchers at the Universitat Autònoma de Barcelona—Nezih Guner, Yuliya Kulikova and Joan Llull—looked at data on Americans between the ages of 20 and 64 in order to try to work out which way the causation runs. Does marriage make people healthier?

The Economist

Have a nice weekend!