Observational Studies & Descriptive Statistics

PSC7475: Week 2

Prof. Weldzius

Villanova University

Slides Updated: 2025-01-29

Do newspaper endorsements matter?

- Can newspaper endorsements change voters' minds?
- Why not compare vote choice of readers of different papers?
  - Problem: readers choose papers based on their previous beliefs
  - Liberals $\rightsquigarrow$ New York Times, conservatives $\rightsquigarrow$ Wall Street Journal
- Could do a lab experiment, but there are concerns over **external validity**
- Study for today: British newspapers switching their endorsements.
  - Some newspapers endorsing Tories in 1992 switched to Labour in 1997
  - **Treated group**: readers of Tory $\rightarrow$ Labour papers
  - **Control group**: readers of papers who didn't switch

Observational studies

- Example of an **observational study**:
  - We as researchers observe a naturally assigned treatment
  - Very common: often cna't randomize for ethical/logistical reasons
- **Internal validity**: Are the causal assumptions satisfied? Can we interpret this as a causal effect?
  - RCTs usually have higher internal validity
  - Observational studies less so, because pre-treatment variable may differ between treatment and control groups
- **External validity**: Can the conclusions/estimated effects be generalized beyond this study?
  - RCTs weaker here because often very expensive to conduct on representative samples
  - Observational studies often have larger/more representative samples that improve external validity

Confounding

- **Confounder**: pre-treatment variable affecting treatment and the outcome
  - Leftists (X) more likely to read newspapers switching to Labour (T)
  - Leftists (X) also more likely to vote for Labour (Y)
- **Confounding bias** in the estimated SATE due to these differences
  - $\bar{Y}_{control}$ not a good proxy for $Y_i(0)$ in treated group
  - one type: **selection bias** from self-selection into treatment

Research designs

- How can we find a good comparison group?
- Depends on the data we have available
- Three general types of observational study **research designs**:
  1. **Cross-sectional design**: compare outcomes treated and control units at one point in time
  2. **Before-and-after design**: compare outcomes before and after a unit has been treated, but need over-time data on treated group
  3. **Differences-in-differences design**: use before/after information for the treated and control group; need over-time data on treated and control group

Cross-sectional design

- Compare treatment and control groups after treatment happens
  - Readers of switching papers vs. readers of non-switching papers in 1997
- Treatment and control groups assumed identical on average as in RCT
  - Sometimes called **unconfoundedness** or **as-if randomized**
- Cross-section comparison estimate:

$$\bar{Y}^{after}_{treated} - \bar{Y}^{after}_{control}$$

- Could there be confounders?

Statistical control

- **statistical control**: adjust for confounders using statistical procedures
  - Can help to reduce confounding bias
- One type of statistical control: **subclassification**
  - Compare treated and controls groups within levels of a confounder
  - Remaining effect can't be due to the confounder
- Threat to inference: we can only control for observed variables $\rightsquigarrow$ threat of **unmeasured confounding**

Before-and-after comparison

- Compare readers of party-switching newspapers before and after switch
- Advantage: all person-specific features held fixed
  - comparing within a person over time
- Before-and-after estimate:

$$\bar{Y}_{treated}^{after} - \bar{Y}_{treated}^{before}$$

- Threat to inference: **time-varying confounders**
  - Time trend: Labour just did better overall in 1997 compared to 1992

Differences in differences (Diff-in-Diff)

- Key idea: use the before-and-after difference of **control group** to infer what would have happened to **treatment group** without treatment
- DiD estimate:

$$\left(\bar{Y}^{after}_{treated} - \bar{Y}^{before}_{treated}\right) - \left(\bar{Y}^{after}_{control} - \bar{Y}^{before}_{control}\right)$$

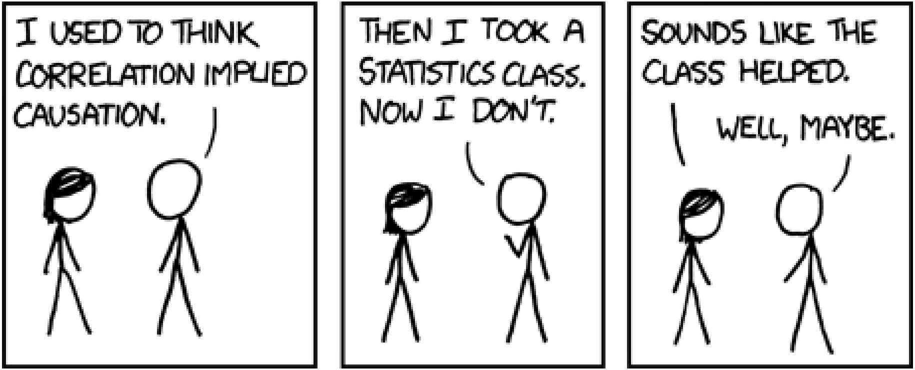
- Change in treated group above and beyond the change in control group
- **Parallel time trend assumption**
  - Changes in vote of readers of non-switching papers roughly the same as changes that readers of switching papers would have been if they read non-switching papers
  - Threat to inference: non-parallel trends

Summarizing approaches:

1. **Cross-sectional comparison** - compare treated units with control units after treatment - Assumption: treated and control units are comparable - Possible confounding
2. **Before-and-after comparison** - Compare the same units before and after treatment - Assumption: no time-varying confounding
3. **Differences-in-differences** - Assumption: parallel trends assumptions - Under this assumption, it accounts for unit-specific and time-varying confounding

- All rely on assumptions that can't be verified to handle confounding
- RCTs handle confounding by design

Causality understanding check



See also: https://www.tylervigen.com/spurious-correlations

Lots of data

- Data from study of the effect of minimum wage

```r
library(tidyverse)
data(minwage, package = "qss")
head(minwage)
```

Lots of data

- Data from study of the effect of minimum wage

```
##         chain location wageBefore wageAfter fullBefore
## 1      wendys       PA       5.00      5.25         20
## 2      wendys       PA       5.50      4.75          6
## 3 burgerking       PA       5.00      4.75         50
## 4 burgerking       PA       5.00      5.00         10
## 5         kfc       PA       5.25      5.00          2
## 6         kfc       PA       5.00      5.00          2
##   fullAfter partBefore partAfter
## 1         0         20        36
## 2        28         26         3
## 3        15         35        18
## 4        26         17         9
## 5         3          8        12
## 6         2         10         9
```

Lots and lots of data

```r
head(minwage$wageAfter, n = 200)
```

```
##   [1] 5.25 4.75 4.75 5.00 5.00 5.00 4.75 5.00 4.50 4.75 4.5
##  [12] 5.00 4.75 4.75 4.75 4.25 5.00 4.90 5.00 4.75 5.00 4.2
##  [23] 4.75 4.25 4.25 4.25 4.25 4.25 4.25 4.38 4.75 4.25 4.5
##  [34] 4.50 4.25 4.25 4.25 4.25 5.05 4.25 4.25 4.25 4.25 4.3
##  [45] 4.50 4.50 5.00 4.75 5.00 4.35 4.25 4.90 4.50 4.50 4.7
##  [56] 6.25 4.35 4.50 4.50 5.00 4.75 4.50 4.75 4.25 4.91 4.4
##  [67] 4.25 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.0
##  [78] 5.05 5.05 5.05 5.50 5.05 5.05 5.05 5.05 5.05 5.05 5.2
##  [89] 5.25 5.05 5.05 5.50 5.05 5.05 5.05 5.05 5.05 5.05 5.0
## [100] 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.2
## [111] 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.67 5.0
## [122] 5.05 5.05 5.05 5.25 5.25 5.05 5.50 5.05 5.05 5.05 5.5
## [133] 5.50 5.05 5.05 5.25 5.05 5.05 5.15 5.05 5.05 5.05 5.0
## [144] 5.00 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.0
```

How to summarize data

- How should we summarize the wages data? Many possibilities!
  - Up to now: focus on **averages** or means of variables
- Two salient features of a variable that we want to know:
  - **Central tendency**: where is the middle/typical/average value
  - **Spread** around the center: are all values to the center or spread out?

Center of the data

- "Center" of the data: typical/average value
- **Mean**: sum of the values divided by the number of observations

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median**:

$$\text{median} = \begin{cases} \text{middle value} & \text{if number of entries is odd} \\ \frac{\text{sum of two middle values}}{2} & \text{if number of entries is even} \end{cases}$$

- In **R**: `mean()` and `median()`

Mean vs median

- Median more robust to **outliers**:
  - Example 1: data $= 0, 1, 2, 3, 5$. Mean? Median?
  - Example 2: data $= 0, 1, 2, 3, 100$. Mean? Median?
- What does Mark Zuckerberg do to the mean vs. median income?

Spread of the data

- Are the values of the variable close to the center?
- **Range**: $[\min(X), \max(X)]$
- **Quantile** (quartile, percentile, etc.): divide data into equal sized groups.
  - 25th percentile: lower quartile (25% of the data below this value)
  - 50th percentile: median (50% of the data below this value)
  - 75th percentile: upper quartile (75% of the data below this value)
- **Interquartile range** (IQR): a measure of variability
  - How spread out is the middle half of the data?
  - Is most of the data really close to the median or are the values spread out?
- **R** function: `range()`, `summary()`, `IQR()`

Standard deviation

- **Standard deviation**: On average, how far away are data points from the mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Steps:
  1. Subtract each data point by the mean
  2. Square each resulting difference
  3. Take the sum of these values
  4. Divide by $n-1$ (or n, doesn't matter much)
  5. Take the square root
- **Variance**: standard deviation^2
- Why not just take the average deviations from mean without squaring?

How large is large?

- Is a wage of 5.30 an hour large?
- Better question: is 5.30 large relative to the distribution of the data?
  - Big in one dataset might be small in another!
  - Different units, difference spreads of the data, etc.
- Need a way to put any variable on **common units**
- **z-score**:

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Interpretation:
  - Positive values above the mean, negative values below the mean
  - Units now on the scale of **standard deviations away from the mean**
  - Intuition: data more than 3 SDs away from mean are rare

z-score example

- Jane works at The Grog where there's a tip jar.
- She's been keeping track of her daily tips:
  - Average tip of \$1.56 with a standard deviation of 20 cents.
- Yesterday, Jane got a \$1.86 tip. How big is this?

- Today she got \$0.56, what about that?