Gov 50: 6. Descriptive Statistics

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda

2. Measurement

3. Descriptive Statistics

4. Wrap-up

1/ Today's agenda

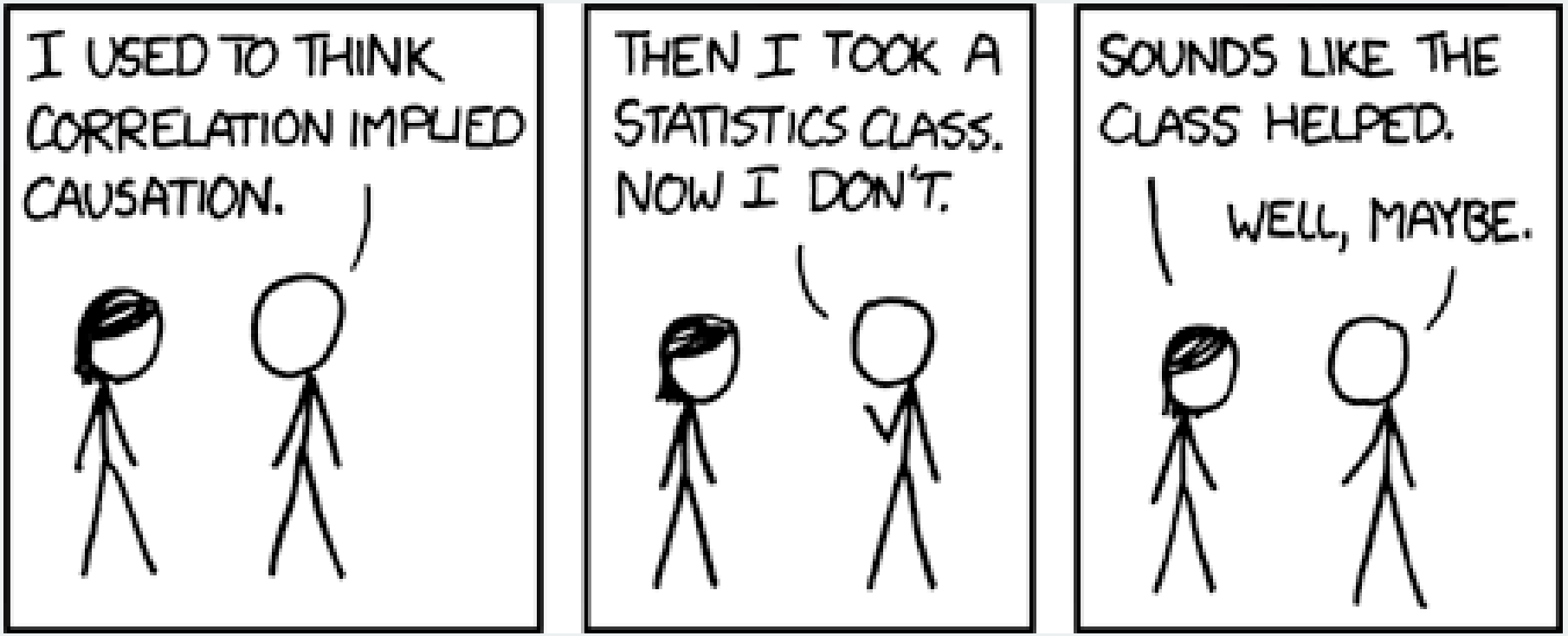
Logistics

- Homework 1
  - Due tonight by midnight.
  - Submit your Rmd and pdf files.
  - Partial credit, so attempt all parts!
- DataCamp Assignment 3 due next Thursday
- Notetaker

Where are we (going)?

- Last few lectures:
  - What is causality?
  - Using data to estimate causal effects
- Next few lectures:
  - How do we measure concepts?
  - Using data to describe the world
  - Numerical summaries of variables

Causality understanding check



2/ Measurement

Concepts & measurement

- Social science is about developing and testing **causal theories**:
  - Does minimum wage change levels of employment?
  - Does outgroup contact influence views on immigration?
- Theories are made up of **concepts**:
  - Minimum wage, level of employment, outgroup contact, views on immigration.
  - We took these for granted when talking about causality.
- Important to consider how we **measure** these concepts.
  - Some more straightforward: what is your age?
  - Others more complicated: what does it mean to "be liberal"?
  - Have to create an **operational definition** of a concept to make it into a variable in our dataset.

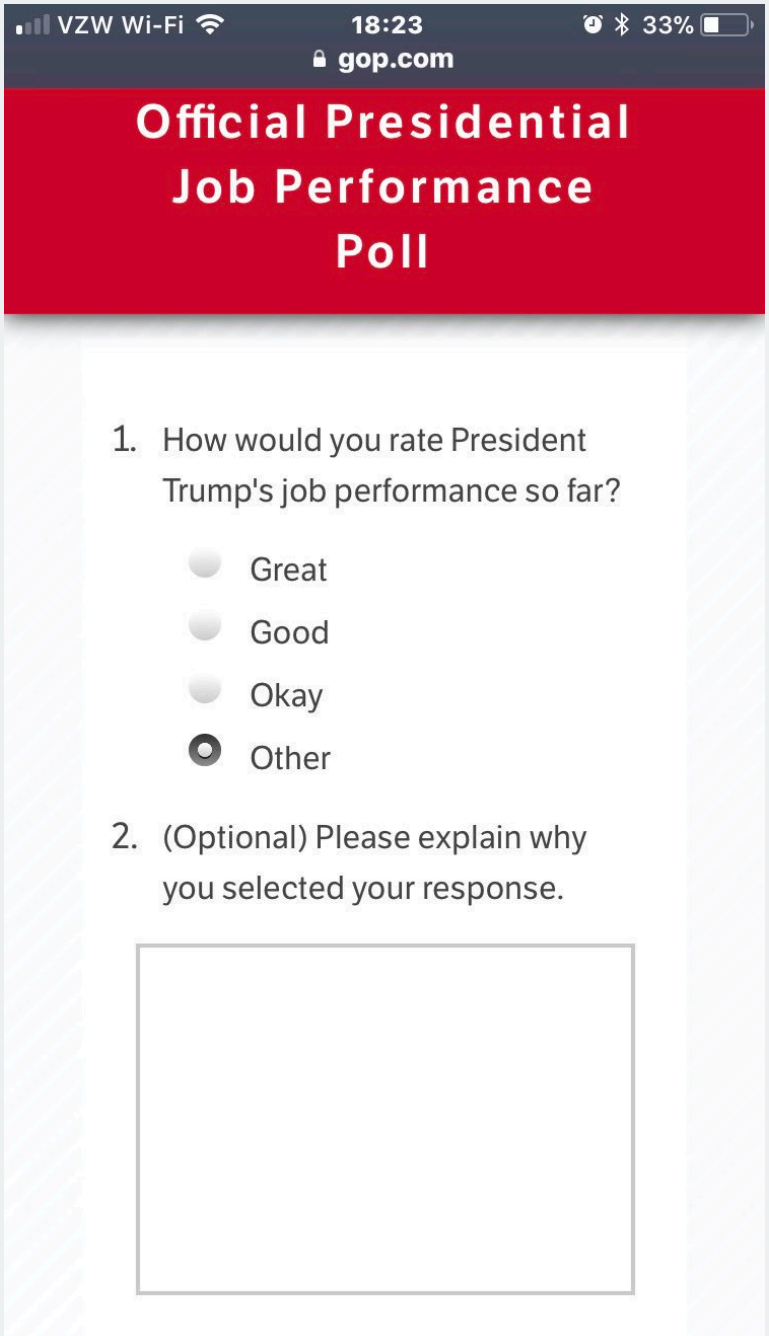
Example

- Concept: presidential approval.
- Conceptual definition:
  - Extent to which US adults support the actions and policies of the current US president.
- Operational definition:
  - “On a scale from 1 to 5, where 1 is least supportive and 5 is more supportive, how much would you say you support the job that Donald Trump is doing as president?”

Measurement error

- **Measurement error**: chance variation in our measurements.
  - individual measurement = exact value + chance error
  - chance errors tend to cancel out when we take averages.
- No matter how careful we are, a measurement could have always come out differently.
  - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012.
  - Data entry errors.
- **Bias**: systematic errors for all units in the same direction.
  - individual measurement = exact value + bias + chance error.
  - "What did you eat yesterday?" $\rightsquigarrow$ underreporting

A biased poll?



3/ Descriptive Statistics

Goal

- A **variable** is a series of measurements about some concept.
- **Descriptive statistics** are numerical summaries of those measurements.
  - If we smart enough, we wouldn't need them: just look at the list of numbers and completely understand.
- 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 the data close 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
- **Median**:

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

- Median more robust to **outliers**:
  - Example 1: data = $\{0, 1, 2, 3, 5\}$. mean = 2.2, median = 2
  - Example 2: data = $\{0, 1, 2, 3, 100\}$. mean = 21.2, median = 2
- What does Mark Zuckerberg do to the mean vs median income?

Minimum wage study

- From QSS: study of how minimum wage increase in New Jersey affected employment, using Pennsylvania as a comparison group.
- Load the data and create subsets:

```
minwage <- read.csv("data/minwage.csv")
minwageNJ <- subset(minwage, subset = (location != "PA"))
minwagePA <- subset(minwage, subset = (location == "PA"))
```

Median wages before and after

```
median(minwageNJ$wageBefore)
```

```
## [1] 4.5
```

```
median(minwageNJ$wageAfter)
```

```
## [1] 5.05
```

```
median(minwagePA$wageBefore)
```

```
## [1] 4.67
```

```
median(minwagePA$wageAfter)
```

```
## [1] 4.5
```

Spread of the data

- Are the data close to the center?
- **Range**: $[\min(X),\ \max(X)]$
- **Quantile** (quartile, quintile, percentile, etc):
  - 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?
- One definition of outliers: over $1.5 \times$ IQR above the upper quartile or below lower quartile.

Quartiles in R

- `summary()` gives quartiles:

```
summary(minwageNJ$wageBefore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.25    4.25    4.50    4.61    4.87    5.75
```

```
summary(minwageNJ$wageAfter)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    5.05    5.05    5.08    5.05    5.75
```

- `IQR()` calculates IQR:

```
IQR(minwageNJ$wageBefore)
```

```
## [1] 0.62
```

```
IQR(minwageNJ$wageAfter)
```

```
## [1] 0
```

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 $n - 1$
  5. Take the square root.
- Sometimes n instead of $n - 1$
- **Variance** = standard deviation^2
- Why not just take the average deviations from mean without squaring?

SDs in R

- Minimum wage data:

```
sd(minwageNJ$wageBefore)
```

```
## [1] 0.343
```

```
sd(minwageNJ$wageAfter)
```

```
## [1] 0.106
```

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, different 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 Hi Rise Bakery, 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 \$1.86 in tips. How big is this?

$$\text{z-score} = \frac{186 - 156}{20} = \frac{30}{20} = 1.5$$

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

$$\text{z-score} = \frac{56 - 156}{20} = \frac{-100}{20} = -5$$

z-scores in R

- Calculate the z-score:

```
wage.mean <- mean(minwageNJ$wageAfter)
wage.sd <- sd(minwageNJ$wageAfter)
minwageNJ$wageAfter.z <- (minwageNJ$wageAfter - wage.mean)/wage.sd
```

- Compare original to z-scores:

```
## original
summary(minwageNJ$wageAfter)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    5.05    5.05    5.08    5.05    5.75
```

```
## z-scores
summary(minwageNJ$wageAfter.z)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.77   -0.30   -0.30    0.00   -0.30    6.33
```

4/ Wrap-up

For next time

- What did we cover:
  - Measurement is about turning concepts into variables.
  - How can we summarize a single variable: center and spread.
- Next time:
  - Read Section 3.3 of QSS.
  - Visualizing a single variable.