POLS201 Spring 2019

Descriptive Statistics

February 27 and March 1

Before we begin

- Hand in your lab sheets
- FYI: Look for the new improved slide archive
- We lurched into a standard deviation discussion abruptly Monday
- Today: that and more descriptive statisticas
    - Bring your laptop Monday

Takeaways from the Homework

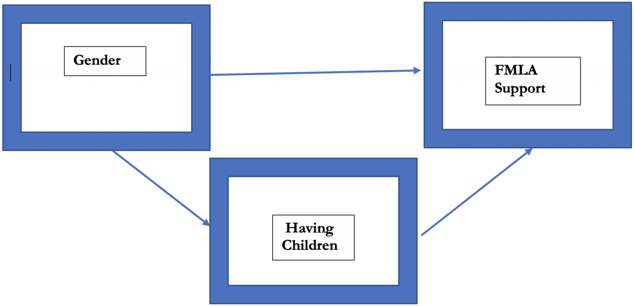
- Let's discuss confounds vs. intervening variables
- Units of analysis
- The diagram for experiments

Confounds are intervening variables are distinct

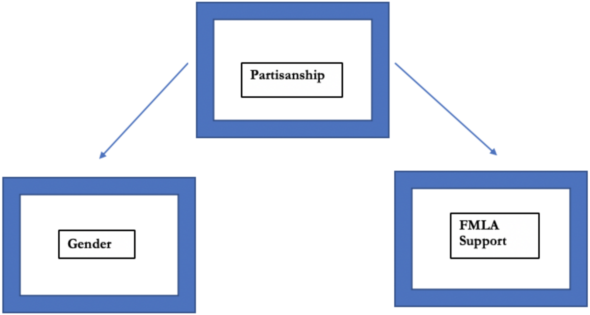
- Consider a causal sequence:
  - *Gender* ——> *Opinion on FMLA*
- A Confound comes from *outside* this link.
  - *Partisanship* correlates with *Gender* and correlates with *FMLA*

An intervening variable sits between your IV and DV

- It follows your IV and precedes your DV
- It may ultimately explain all of the variance, some, or none.



A confounding variable logically precedes or coincides with your IV



Unit of Analysis

- Don't confuse it with your variable or potential variable
- It's just the things you are counting or measuring, usually associated with a level of aggregation.
- In surveys, it's the individual, unless you have aggregated them.

Well designed experiments can be imagined this way:

- Notice there are *four* separate instances of observations.

Random Assignment	Observe Pre-Test	Treat	Observe Post-Test
R	O	~X	O
	O	X	O

Two other small points from Homework 1

- Very few causal stories completely defy the imagintion.
  - From Q4: Does the sequence
    Exposure -> Content -> Opinion
    make sense? It could.
- Even though statistics never prove causation, don't be afraid to imply causation in your hypotheses.
  - Avoid "is correlated with". Don't be passive,

The Cliffhanger from Monday

- Which has greater standard deviation:
  - A survey with 9 Yes and 1 No, or 5 Yes and 5 No
  - Suppose each Yes = 1 and each No = 0
  - Let's compare:
  - A sequence of 1, 0, 0, . . . , 0
  - and a sequence of 1, 0, 1, 0, . . . , 1, 0

In R, we could do this:


```
# create the sequences
first_sequence <- c(1, rep(0, 9))
second_sequence <- rep(c(1, 0), 5)
first_sequence
```

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

```
second_sequence
```

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

```
sd(first_sequence)
```

```
## [1] 0.3162278
```

```
sd(second_sequence)
```

```
## [1] 0.5270463
```

The basic SD formula, in English

- Find the mean
- Subtract each item from the mean, then square each difference.
- Add 'em up
- Divide by n - 1. You now have the variance.
- Get the square root. You have the standard deviation.

The basic SD formula, from last time:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Descriptive Statistics

- Essential Properties of Data that reveal...
  - Central Tendency
  - Dispersion
  - Keep in mind this distinction between the data we observe and the process that generates it
    - We always drawing a single cup from the data stream.
    - The cup reveals something about the stream but never perfectly.

Recall the Levels of Measuremnt

- They are important because they tell us which operations can be performed.
- Assume that you can never take a meaningful average of an ordinal scale.

		Nominal scale	Ordinal scale	Interval scale	Ratio scale	Dummy
Logical/ math operations	× ÷	✗	✗	✗	✓	✓
	+ -	✗	✗	✓	✓	✓
	< >	✗	✓	✓	✓	✓
	= ≠	✓	✓	✓	✓	✓

Recall the Measures of Central Tendency

- Mean
- Median
- Mode

Some important measures of Dispersion

- Variation Ratio
  - Proportion of units that are NOT equal to the mode.
- Range and Interquartile Range
  - Range is the maximum minus the minimum
  - Interquartile range is the 75th - 25th percentiles
- Variance and Standard Deviation

Your goals:

- Learn notation and know how to calculate different measures of central tendency and dispersion.
- Learn which measures are appropriate for different levels of measurement,

Basic Notation

- You will see different mathematical notation for "samples" and "populations"
- We use sample "statistics" to estimate population "parameters"

Basic Notation

- The i subscript is your clue that the variable is going to taken on different values for different observations.
- i stands for an "individual observation" You'll see it in an operation that is repeated.
- Typically, the repeated operation is a summation. The letter sigma (Σ) tells you that the results of some equation are to be summed.
  - ...and a capital Pi (Π) describes a product in the same way.

x_i	x_i
x_1	1
x_2	3
x_3	6
x_4	6

How would you calculate this?

$$\sum_{i=1}^{n} x_i$$

Basic Notation

- Different symbols are used for populations vs. samples

	(Population) Parameter	(Sample) Statistic
Mean	μ	$\bar{x}$
Proportion	P	p
Standard Deviation	σ	s
Variance	σ^2	s^2

Modes and Medians

- Modes are the most frequently occurring value in your variable
- What is the mode? -(1, 2, 2, 3, 3, 3, 4, 5, 6, 7, 8, 9, 9)
- Median represents the middle value in a rank ordering
  - What is the median? -(1, 2, 2, 3, 3, 3, 4, 5, 6, 7, 8, 9, 9)

And yes, R will also tell you these things:


```
values <- c(1, 2, 2, 3, 3, 3, 4, 5, 6, 7, 8, 9, 9)

getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(values)
```

```
## [1] 3
```

```
median(values)
```

```
## [1] 4
```

Let's work through this example

- Recall the Yes/No poll example from today.
- In the data below, what is the mean? The SD?

Name	Vote
Maggie	1
Sam	1
David	0
Maria	1
Maddie	0
Connie	0
Guy	0
Jacob	1
Hillary	1
Jerome	0

Descriptive Statistics

<table>
<tr><th></th><th>Mean</th><th>Median</th><th>Mode</th><th>Range</th><th>SD/Var.</th></tr>
<tr><td>Categorical</td><td></td><td></td><td>X</td><td></td><td></td></tr>
<tr><td>Dummy</td><td>X</td><td colspan="2">X</td><td>X</td><td>X</td></tr>
<tr><td>Ordinal</td><td></td><td>X</td><td>X</td><td>X</td><td></td></tr>
<tr><td>Continuous</td><td>X</td><td>X</td><td>X</td><td>X</td><td>X</td></tr>
</table>

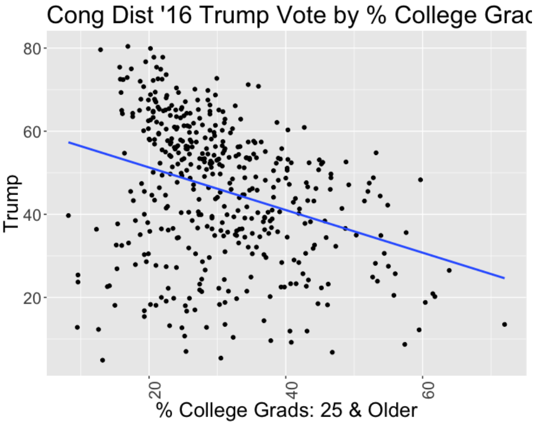
The applicable measure of dispersion for categorical variables is the "variation ratio" (1- % in Mode)

Overview of Linear Correlation

- As we saw earlier, linear correlation provides a numerical measure of direction and strength of association
- Correlations lie between -1 and 1
- Positive values mean a positive association, negative means negative association
- $+1$ and -1 are strongest values; 0 is the weakest

Remember this graph?

- It looks like a mess but actually has a correlation of -.31.



The formula for Pearson's R: You don't need to remember it

- Note it cannot exceed +1 or fall below -1

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y}$$

Anscombe's Quartet

POLS201
Spring 2019

- But know this: Each of these four datasets produces identical correlations of +.82:
- A correlation tells you *something* but you need to see a picture.