Lecture 1

Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/08/24

Slides Updated: 2023-08-28

Agenda

1. Meet the instructor
2. Why are you here?
3. Variables 101
4. Data and summarizing data
5. Review syllabus and expectations

Meet the instructor

- Education
  - PhD from NYU Politics in 2019
  - Postdocs at Princeton Niehaus & NYU CSMaP
- Published some things
  - Methods-ey: external validity 1, 2; measurement 3, 4
  - Substantive: economics & populism 1; Covid-19 & U.S. politics 2, 3; IPE 4; academic naval-gazing 5
  - Popular press: 1, 2, Podcasts
- Work
  - World Bank / IFC
  - MarketCast

Why are you here?

- I.e., why are you getting a PhD in political science?
  - You enjoy asking and answering questions about **politics**
  - What questions are you interested in?

How I can help

- Quantitative analysis is one of many tools to answer your questions
  - Based on **numerical** measurements
  - Interested in developing and testing **generalizable** theories
  - Measurements and analyses that are easily **replicable** by others
- Contrast to other two dominant paradigms in political science research
  1. Qualitative analyses
  2. Formal modeling
- Never **ever** fall into the petty trap that quantitative is somehow "better"
  - “A foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines."
- You are here because you do not have a little mind

Quantitative Analysis

- Political scientists work with quantitative data for three reasons:
    1. What can we say about the data **we have**?
    2. What can we say about the data **we don't have**?
    3. What can we say about the data **we'd expect to see**?
- Answering these questions requires three types of statistics
    1. Descriptive
    2. Inferential (from samples to populations)
    3. Prediction (from models to hypotheticals)
- All three of these approaches rely on **a test statistic**
    - A number that **summarizes data**

Background

- Handout work

Variables 101

- We study **units**
  - Phenomena about which we wish to make statements
  - AKA **cases**: people, counties, nations, dyads, etc.
- Units have **attributes**
  - Characteristics of a **unit** that distinguish it from other units
- **Variables** are *logical* groupings of *mutually exclusive* attributes
  - An important part of quantitative research is assigning a **value** to each attribute
  - The variable **GDP per capita** takes on the value **$2,256** for the unit **India**
  - The variable **year** takes on the value **first** for the unit **you**
- In quantitative analysis, we assign **scores** to each **value**

Levels of measurement

- Variables can be measured at four levels
    1. **Nominal**: cannot be ordered in any logical way
    2. **Ordinal**: can be ordered, but no meaning to differences
    3. **Interval**: ordinal variables whose differences can be compared
    4. **Ratio**: zero is meaningful -- *nothing* of the quantity measured
- Mathematical operations can be conducted on these levels
    1. **Nominal**: equality ($=$) only (do these take on the same value or not?)
    2. **Ordinal**: equality, greater than ($>$) or less than ($<$)
    3. **Interval**: addition ($+$), subtraction ($-$), averages ($\frac{1}{n}\sum_i x_i$)
    4. **Ratio**: multiplication ($*$) and division ($/$)

Tricky cases:

- Celsius? Latitude and longitude? Binary variables?

Data structures

- **Data table** (or **data frame**)
  - Rows are **units**
  - Columns are **variables**
  - Cells are **scores**
- **List**: Tree-like structure
  - Units are **outermost node**
  - Attributes are **child nodes**
  - Scores are **children of attributes**
- We will only use **data table** / **data frames** this semester
- But if you want to communicate with CS / engineers / data scientists, **lists** are their world!

Summarizing data: displays

```
##   id state age  GPA
## 1 PE    MI  25 4.04
## 2 VF    MT  24 3.73
## 3 EP    CO  23 3.84
## 4 LD    IA  25 3.90
## 5 OB    IN  29 3.97
## 6 IG    VA  27 3.84
```

- Why not just present this table as is?

```
##    id state age  GPA
## 1  PZ    WI  25 3.90
## 2  VE    DE  24 4.08
## 3  EB    MT  24 3.79
## 4  LO    KY  25 4.04
## 5  OH    WY  24 4.03
## 6  IT    TN  24 3.93
## 7  XP    NY  26 3.90
## 8  FL    MI  23 3.85
## 9  ZC    OH  28 3.86
## 10 DI    SD  24 3.96
## 11 BR    KS  30 4.11
```

Summarizing data: displays

- Fundamental tension in quantitative analysis: **detail versus parsimony**
- Use a frequency table?

```
##   age n
## 1  23 1
## 2  27 1
## 3  26 2
## 4  28 2
## 5  29 3
## 6  30 4
## 7  25 6
## 8  24 7
```

Summarizing data: displays

- What about for GPA?

```
##       GPA n
## 1  3.7609 1
## 2  3.7628 1
## 3  3.7805 1
## 4  3.7861 1
## 5  3.8276 1
## 6  3.8545 1
## 7  3.8597 1
## 8  3.8633 1
## 9  3.8683 1
## 10 3.8808 1
## 11 3.8822 1
## 12 3.8826 1
## 13 3.8830 1
## 14 3.8847 1
## 15 3.8992 1
## 16 3.9007 1
## 17 3.9258 1
## 18 3.9336 1
## 19 3.9550 1
```

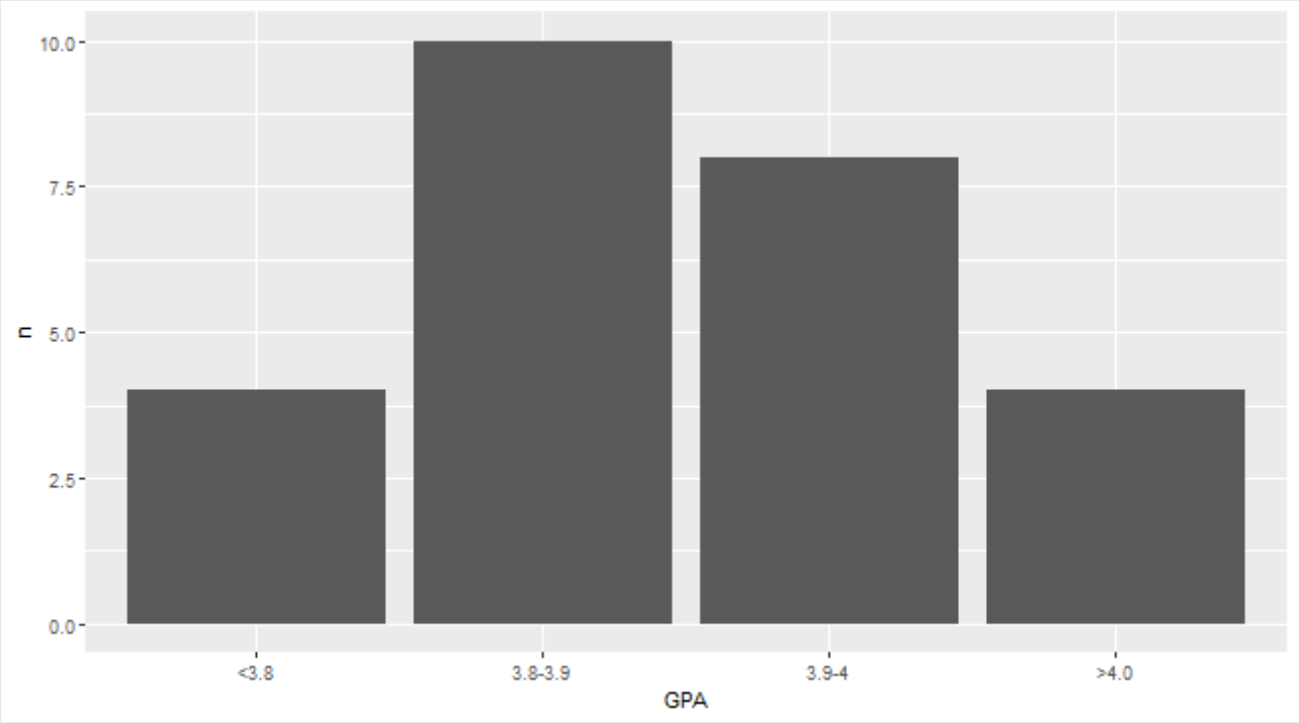
Summarizing data: displays

- **Recode** data into categories, then use a frequency table

```
##       GPA  n
## 1    <3.8  4
## 2 3.8-3.9 10
## 3   3.9-4  8
## 4    >4.0  4
```

Summarizing data: displays

- Also can visualize with a **plot**



Summarizing data: central tendency

- **Central Tendency**: The *typical value*
  - **Mode**: most frequently observed value (which levels of measurement (LOM)?)
  - **Median**: value of smallest observation for which the cumulative percentage is ≥ 50 (which LOM?)
  - **Mean**: average $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ (which LOM?)

Summarizing data: dispersion

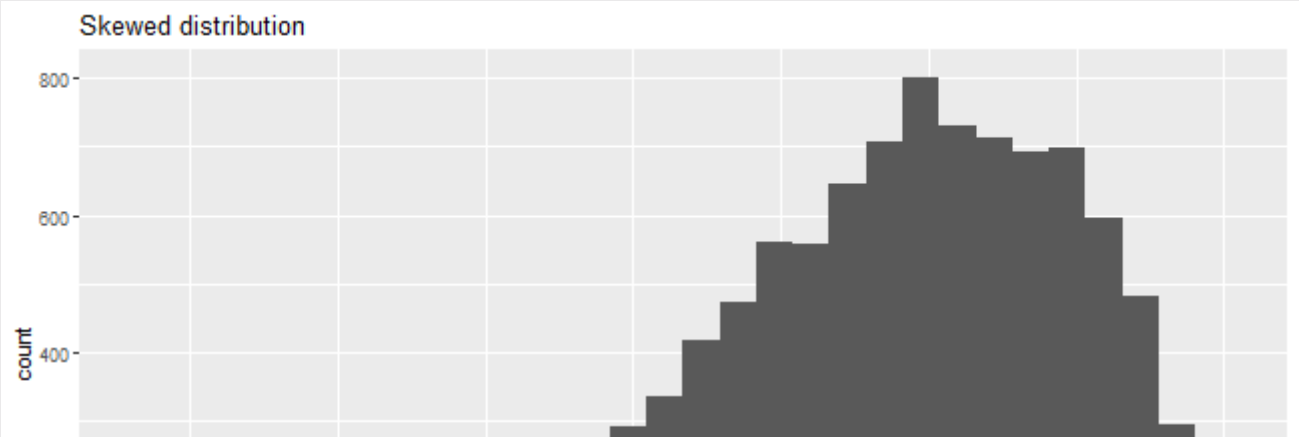
- **Dispersion**: The *spread*
  - **Range**: difference between smallest and largest values (LOM?)
  - **IQR**: difference between 75th%ile and 25%ile (LOM?)
  - **Variance**: $s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$ (LOM?)

Summarizing data: qualitative description

- Frequency distribution may be "symmetric" or "skewed"
    - Median is typically better than mean if data is skewed (why?)

```
data.frame(x = rbeta(10000,5,2)) %>%
  ggplot(aes(x = x)) +
  geom_histogram() +
  labs(title = 'Skewed distribution')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with
## `binwidth`.
```



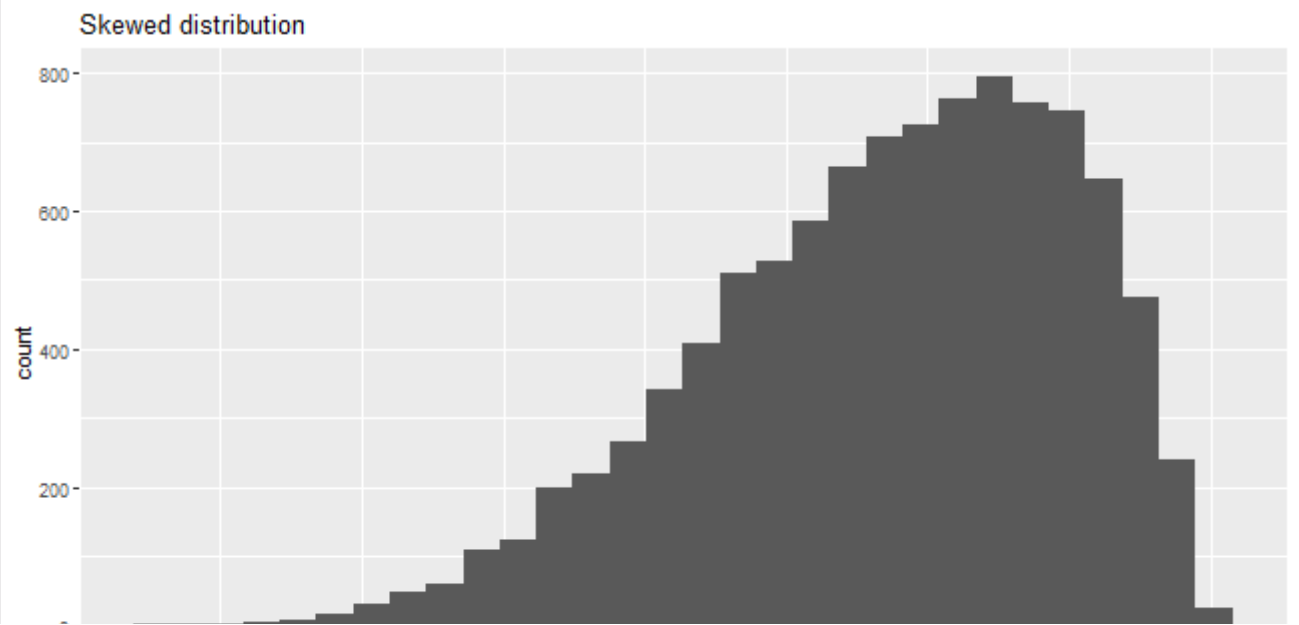
Summarizing data: qualitative description

- May be "unimodal" or "bimodal"
- Qualitative descriptions can be quantified
  - I.e., skew $g_1 = \frac{1}{N*s^3} \sum_{i=1}^{N} (y_i - \bar{y})^3$
  - If $g_1 = 0$, symmetric
  - If $g_1 < 0$, skewed left
  - If $g_1 > 0$, skewed right

Skew

```
data.frame(x = rbeta(10000,5,2)) %>%
  ggplot(aes(x = x)) +
  geom_histogram() +
  labs(title = 'Skewed distribution')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with
## `binwidth`.
```



Logistics

- Syllabus review
- TA office hours and labs