# Descriptive Statistics

PLSC 309
25 January 2019

# What is description?

“If human beings could see in multiple dimensions, we wouldn’t need data analysis.”

-- Pedro Domingos, *University of Washington*

# What is description?

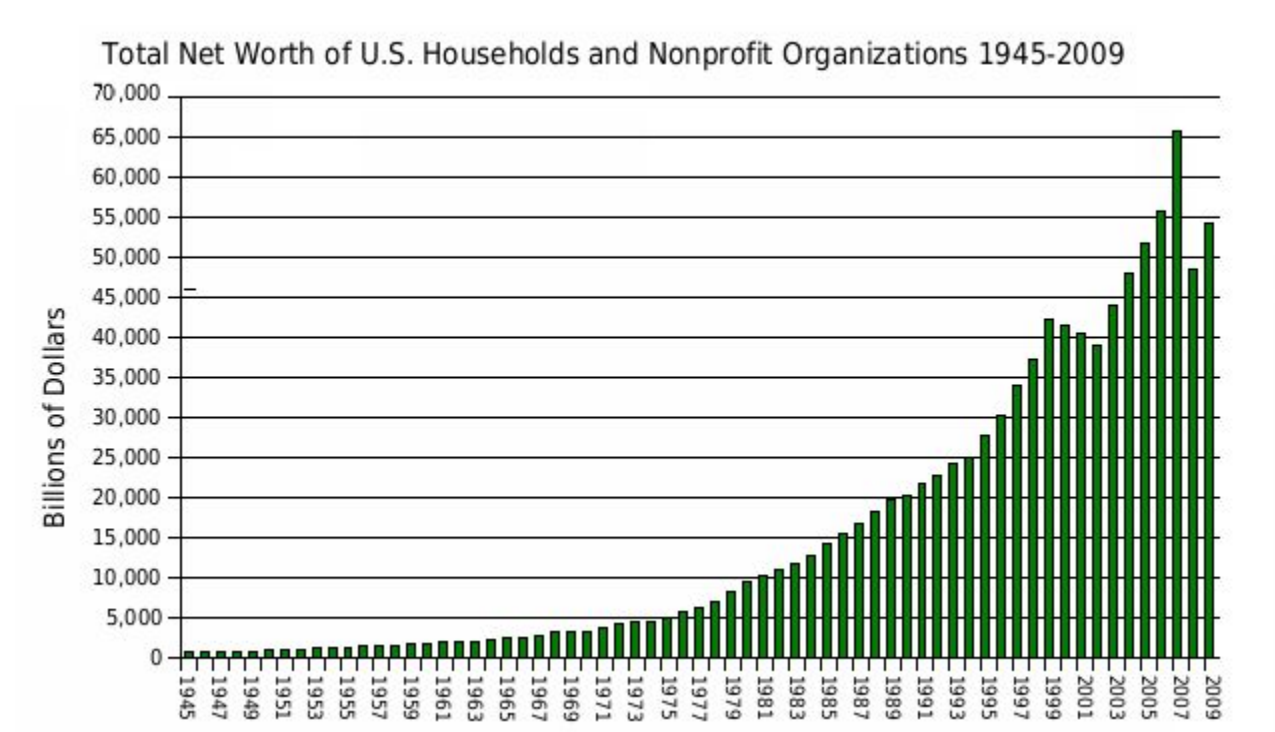
| country | beer_servings | spirit_servings | wine_servings |
|---|---|---|---|
| Afghanistan | 0 | 0 | 0 |
| Albania | 89 | 132 | 54 |
| Algeria | 25 | 0 | 14 |
| Andorra | 245 | 138 | 312 |
| Angola | 217 | 57 | 45 |
| Antigua & Barbuda | 102 | 128 | 45 |
| Argentina | 193 | 25 | 221 |
| Armenia | 21 | 179 | 11 |
| Australia | 261 | 72 | 212 |
| Austria | 279 | 75 | 191 |
| Azerbaijan | 21 | 46 | 5 |
| Bahamas | 122 | 176 | 51 |
| Bahrain | 42 | 63 | 7 |
| Bangladesh | 0 | 0 | 0 |
| Barbados | 143 | 173 | 36 |
| Belarus | 142 | 373 | 42 |
| Belgium | 295 | 84 | 212 |
| Belize | 263 | 114 | 8 |
| Benin | 34 | 4 | 13 |

| | |
|---|---|
| Mean | ? |
| Median | ? |
| Range | ? |

# Two ways to summarize data

- *Centrality*
  - What is the middle point of the data?
  - Describes the average response
- *Spread*
  - What are the range of values?
  - How common are observations further away from the center?
  - Agnostic to direction

# Two ways to summarize data



# Centrality: Mean

1. Mean (geometric average)
2. Add up all values of a variable, divide by number of observations
3. Just use software!

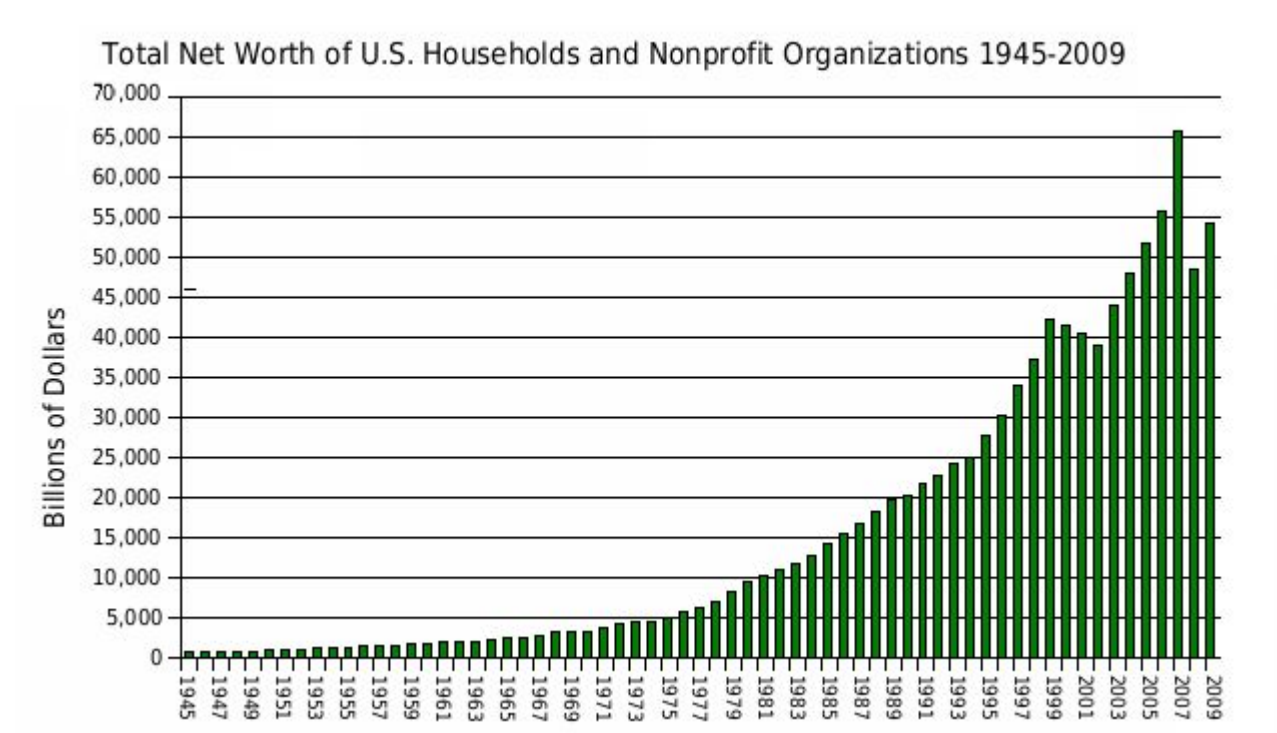
# Centrality: Median

1. Middle point of data
2. Sort all values of a variable, find the value that's at the number of observations / 2
3. Just use software!

# Centrality: Mode

1. The most frequently occuring value
2. Use only for categorical variables or discrete variables with a small amount of values

# Mean vs. Median




| | |
|---|---|
| Mean | 301,000 |
| Median | 45,000 |

# Mean vs. Median

- For the median, *all observations have the same weight*
- For the mean, *higher value observations have a higher weight*

# Spread: Range

- The lowest and highest values
- Not very informative...

# Spread: Variance

# Spread: Variance

$$\sigma^2 = \sum (X_i - \bar{X})^2 / N$$

$\sigma^2$ = *variance*

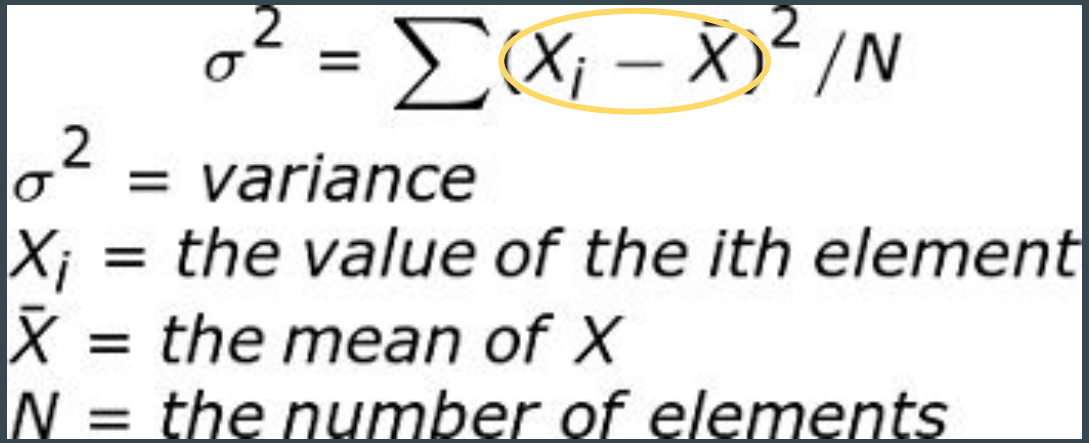
$X_i$ = *the value of the ith element*

$\bar{X}$ = *the mean of X*

$N$ = *the number of elements*

# Spread: Variance

The difference between a single observation and the mean



# Spread: Variance

Squared to equalize positive and negative distances

$$\sigma^2 = \sum (X_i - \bar{X})^2 / N$$

$\sigma^2$ = variance

$X_i$ = the value of the ith element

$\bar{X}$ = the mean of $X$

$N$ = the number of elements

# Spread: variance

All the differences summed together and divided by total number of observations

$$\sigma^2 = \sum (X_i - \bar{X})^2 / N$$

$\sigma^2$ = variance

$X_i$ = the value of the ith element

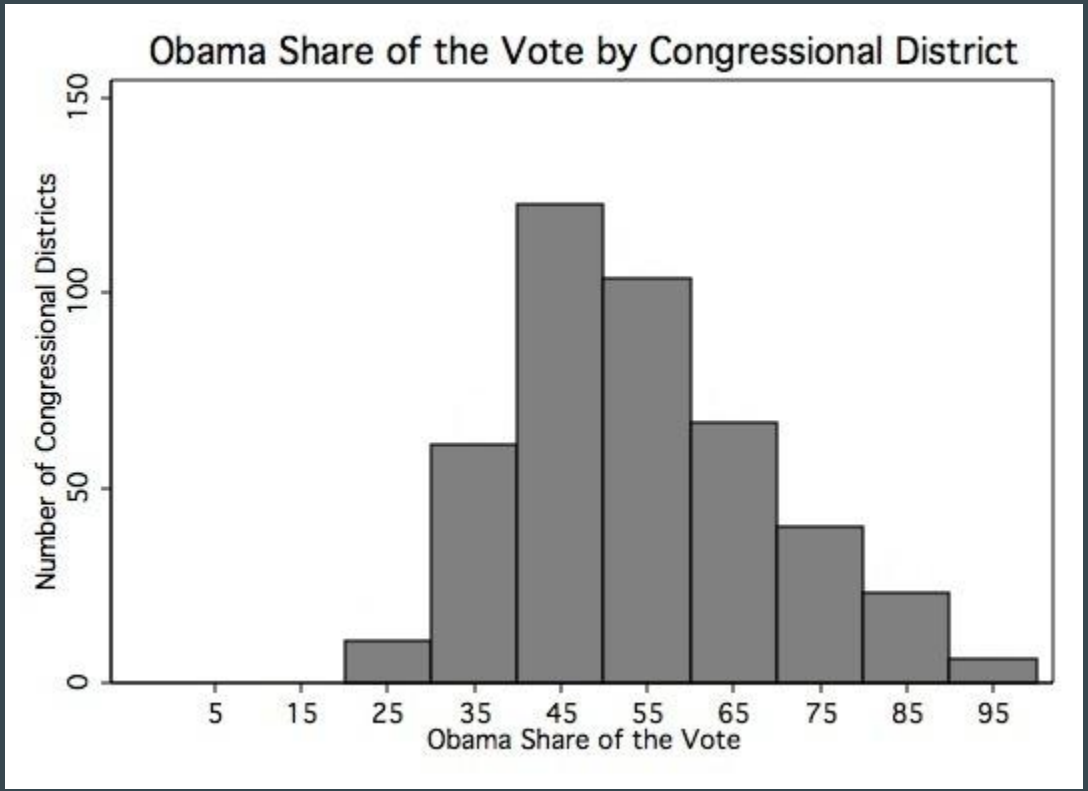
$\bar{X}$ = the mean of $X$

$N$ = the number of elements

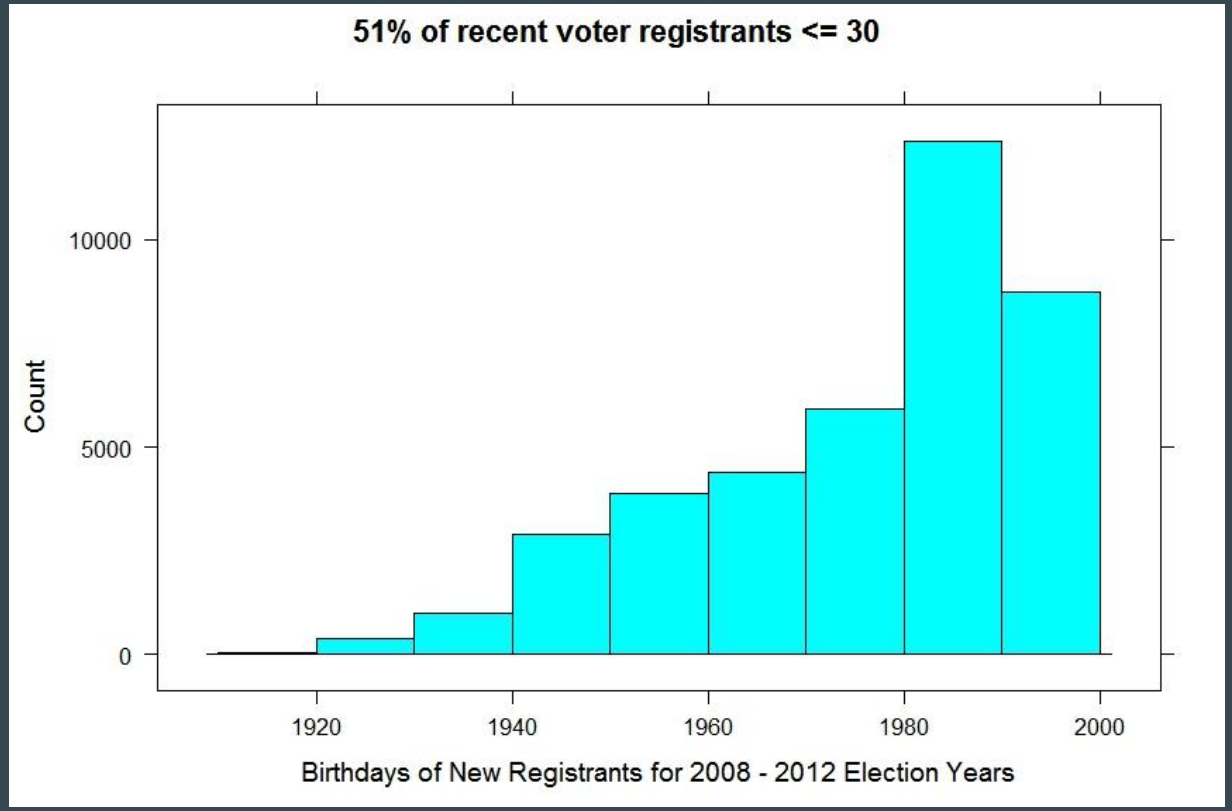
# Data visualization: histogram

- Y-axis represents number of observations
- X-axis represents values of variable

# Data visualization: histogram



# Data visualization: histogram



# Review

- Descriptive statistics are a *lower dimensional* representation of data
- Centrality measures a “typical” or “most likely” value
  - Mean
  - Median
  - Mode
- Spread measures the average distance of observations from the center
  - Range
  - Variance
- We will come back to variance and histograms next week!