

# Descriptive Statistics

---

Aleksandr Fisher

1. Overview of Last Week
2. Measurement
3. Descriptive Statistics
4. Readings
5. Wrap up

# 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.
  - “Did you vote?” ~ overreporting

# Definitions

- 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:** the “middle” of a sorted list of numbers.
- 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

## 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  $1.5 \times \text{IQR}$  above the upper quartile or below lower quartile.



# Standard deviation

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

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

- 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

- **Variance** = standard deviation (squared)

$$\frac{\sum((x - \bar{x})^2)}{n - 1}$$

# How large is large?

- Need a way to put any variable on common units.

- **z-score:**

$$\frac{x - \bar{x}}{\sigma}$$

- 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.

# Studying Feelings Toward Democracy

- 2018 Pew Study
- 27 countries, ~ 30,000 respondents
- **Question:** How satisfied are you with the way democracy is working in our country – very satisfied, somewhat satisfied, not too satisfied, or not at all satisfied?
  1. very satisfied
  2. somewhat satisfied
  3. not too satisfied
  4. not at all satisfied
  5. don't know
  6. refused

- Load the data:

```
library(haven) # package to read the data
library(dplyr) # package for data manipulation
pew2018 <- read_sav("Pew 2018.sav",
                  user_na=TRUE) %>%
                  as_factor()
pew2018 = pew2018 %>% select(COUNTRY, satisfied_democracy, age, sex)
head(pew2018) # top 6 observations
```

```
## # A tibble: 6 x 4
##   COUNTRY      satisfied_democracy age  sex
##   <fct>      <fct>                <fct> <fct>
## 1 United States Not at all satisfied 60   Male
## 2 United States Not too satisfied   69   Male
## 3 United States Not too satisfied   71   Male
## 4 United States Somewhat satisfied  82   Female
## 5 United States Somewhat satisfied  46   Male
## 6 United States Not at all satisfied 57   Female
```

## Glimpsing at data

- `dim()`: Retrieve the dimension
- `names()`: Get the names
- `str()`: Display compactly the internal structure
- `glimpse()`: is the dplyr-version of `str()` showing values of each variable the whole screen width, but does not display the number of levels and names of factor variables. But this feature of `str()` cannot be displayed completely with either many or long levels names.
- `View()`: With RStudio you can see and inspect the data set comfortably. The `View()` function invokes a spreadsheet-style data viewer.

# Glimpsing at data

```
dim(pew2018) # Dimensions
```

```
## [1] 30109      4
```

```
names(pew2018) # Column names
```

```
## [1] "COUNTRY"          "satisfied_democracy" "age"  
## [4] "sex"
```

```
glimpse(pew2018) # Structure of data
```

```
## Observations: 30,109  
## Variables: 4  
## $ COUNTRY          <fct> United States, United States, United State...  
## $ satisfied_democracy <fct> Not at all satisfied, Not too satisfied, N...  
## $ age              <fct> 60, 69, 71, 82, 46, 57, 64, 81, 84, Refuse...  
## $ sex              <fct> Male, Male, Male, Female, Male, Female, Fe...
```

# Contingency table

- The `_table()` function shows us how many respondents are in each category of a categorical variable:

```
table(pew2018$satisfied_democracy)
```

```
##
##          Very satisfied      Somewhat satisfied      Not too satisfied
##                3403                10402                8801
## Not at all satisfied Don't know (DO NOT READ)  Refused (DO NOT READ)
##                6785                662                56
```

- We can use `prop.table()` to show what proportions of the data each response represents:

```
prop.table(table(pew2018$satisfied_democracy))
```

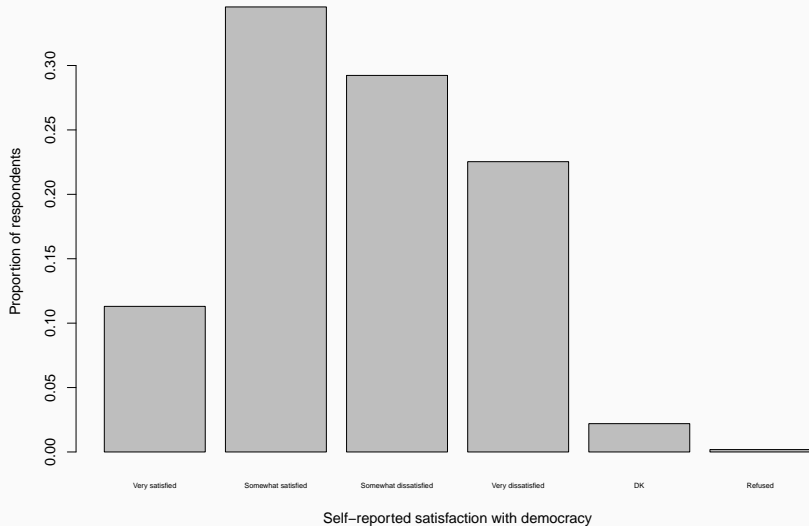
```
##
##          Very satisfied      Somewhat satisfied      Not too satisfied
##          0.113022684          0.345478096          0.292304627
## Not at all satisfied Don't know (DO NOT READ)  Refused (DO NOT READ)
##          0.225347903          0.021986781          0.001859909
```



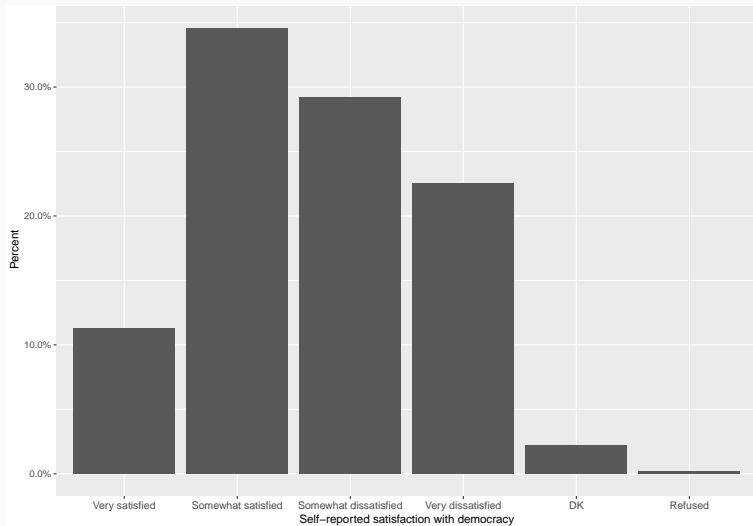
- The `barplot()` function can help us visualize a contingency table:

```
barplot(prop.table(table(pew2018$satisfied_democracy)),  
xlab = "Self-reported satisfaction with democracy", ylab = "Proportion of respondents")
```

- Arguments:
  - First is the height each bar should take (we're using proportions in this case)
  - names are the labels for the each category
  - `xlab`, `ylab` are axis labels



## Bar plot (Using ggplot)



# Histogram