Descriptive Statistics

Aleksandr Fisher

Plan

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?" $\sim$ 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 od 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 $1.5 \times$ 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

- **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.

Shape

- *Skewness*
  - Positive/right skew
  - Symmetric
  - Negative/left skew
- *Kurtosis*: peakedness of a distribution

Skewness

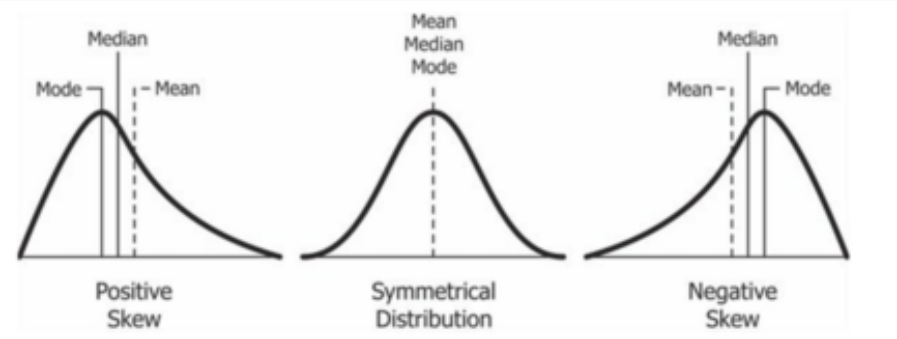


Figure 1: Skewness

Kurtosis

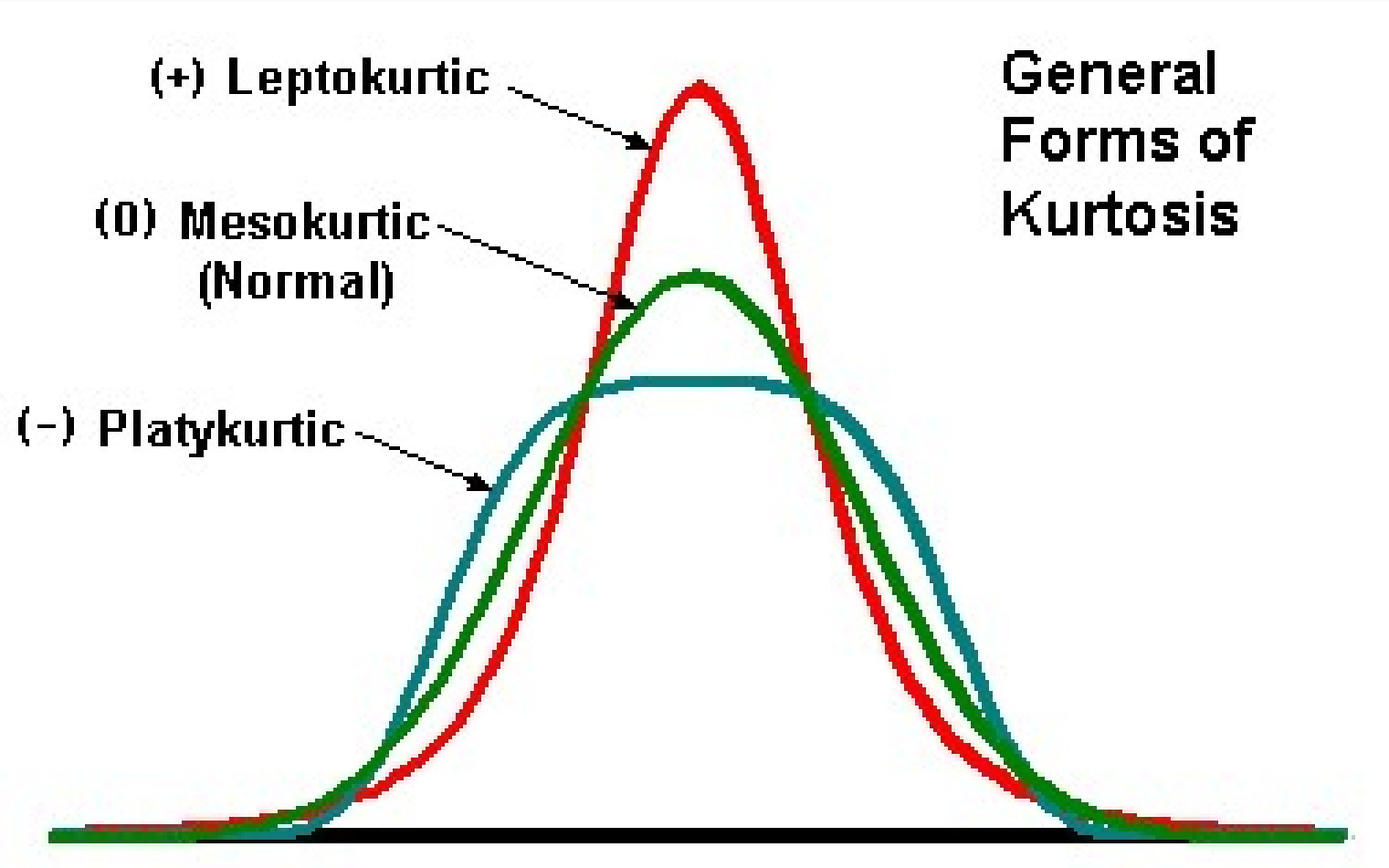


Figure 2: Kurtosis

Relationship

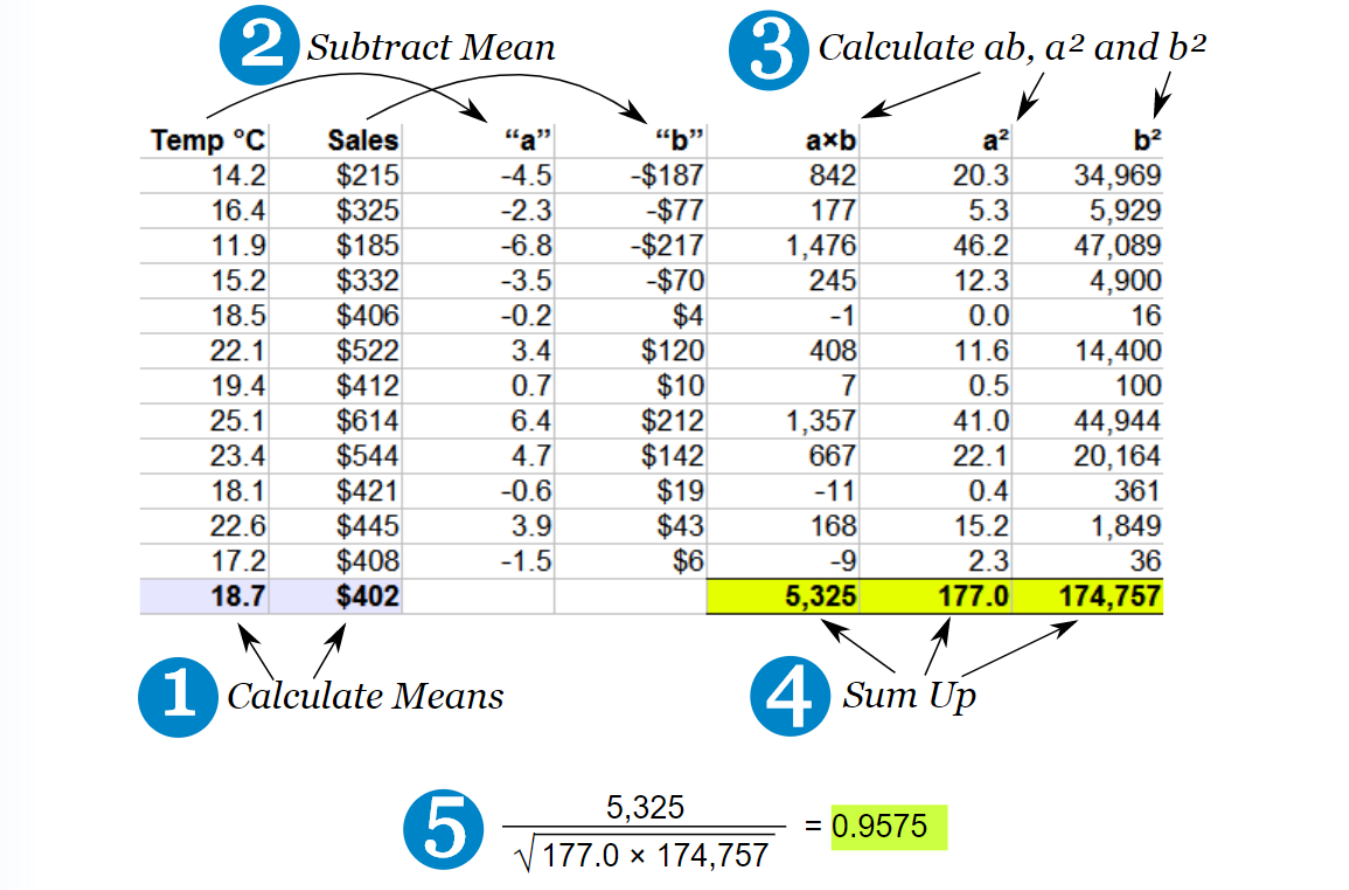
- **Covariance**: provides a measure of the strength of the correlation between two or more sets of random variates.

$$cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

- **Correlation** is the degree to which two or more quantities are linearly associated

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

Correlation



Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

Figure 3: Correlation by hand

Correlation

- Correlation is **Positive** when the values **increase** together
- Correlation is **Negative** when one value **decreases** as the other increases
- Correlation can have a value:
  - 1 is a perfect **positive** correlation
  - 0 is no correlation (the values don't seem linked at all)
  - -1 is a perfect **negative** correlation

Correlation

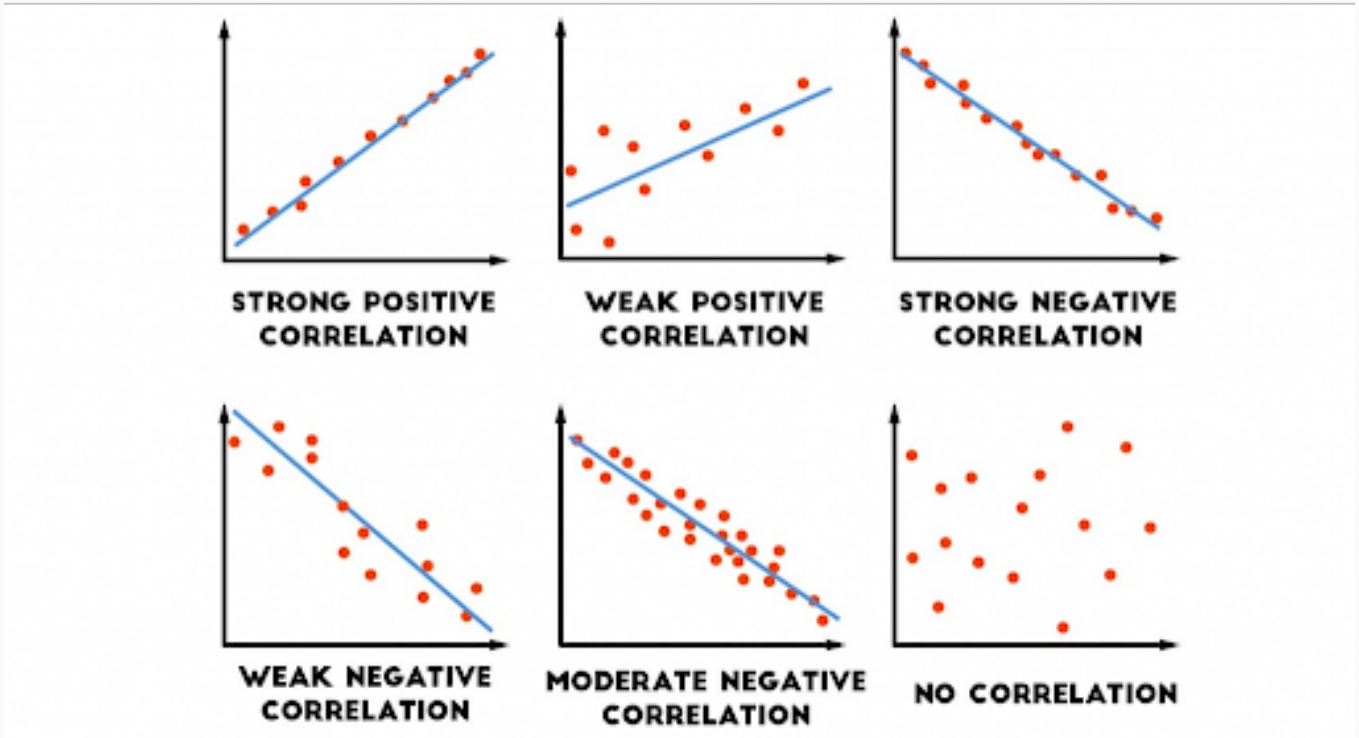
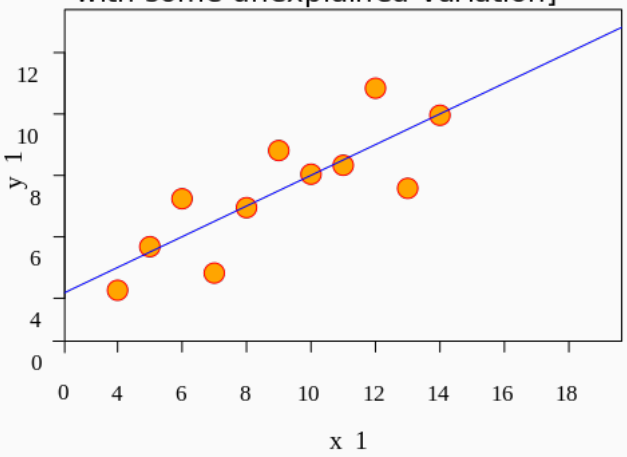


Figure 4: Correlation

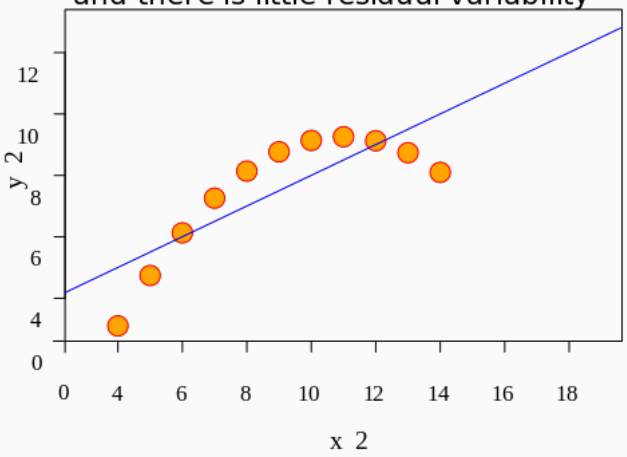
Correlation is not good at curves

Anscombe's Quartet

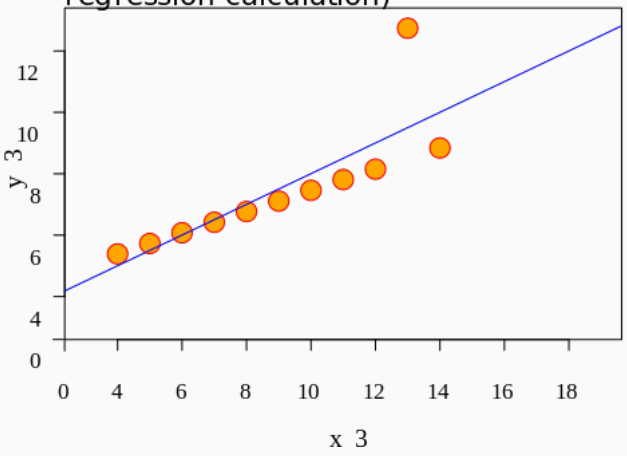
"what most people would see in their mind's eye [for a linear relationship with some unexplained variation]"



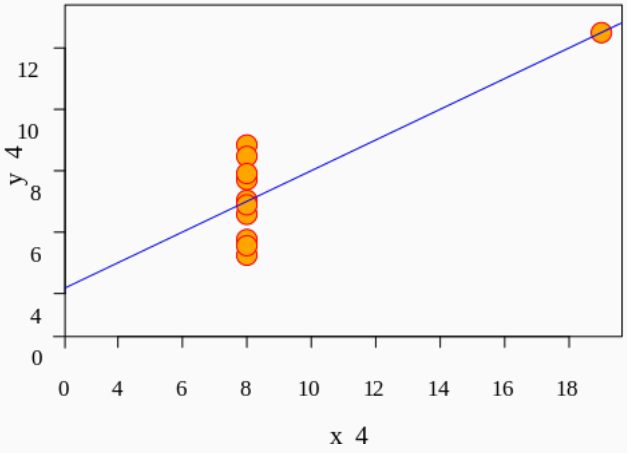
"y has a smooth curved relation with x, possibly quadratic, and there is little residual variability"



"all but one of the observations lie close to a straight line (not the one yielded by the standard regression calculation)"



"all the information about the slope of the regression line resides in one observation"



Correlation is not Causation

- What it really means is that a correlation does not prove one thing causes the other:
  - One thing might cause the other
  - The other might cause the first to happen
  - They may be linked by a different thing
  - Or it could be random chance!

Spurious Correlations

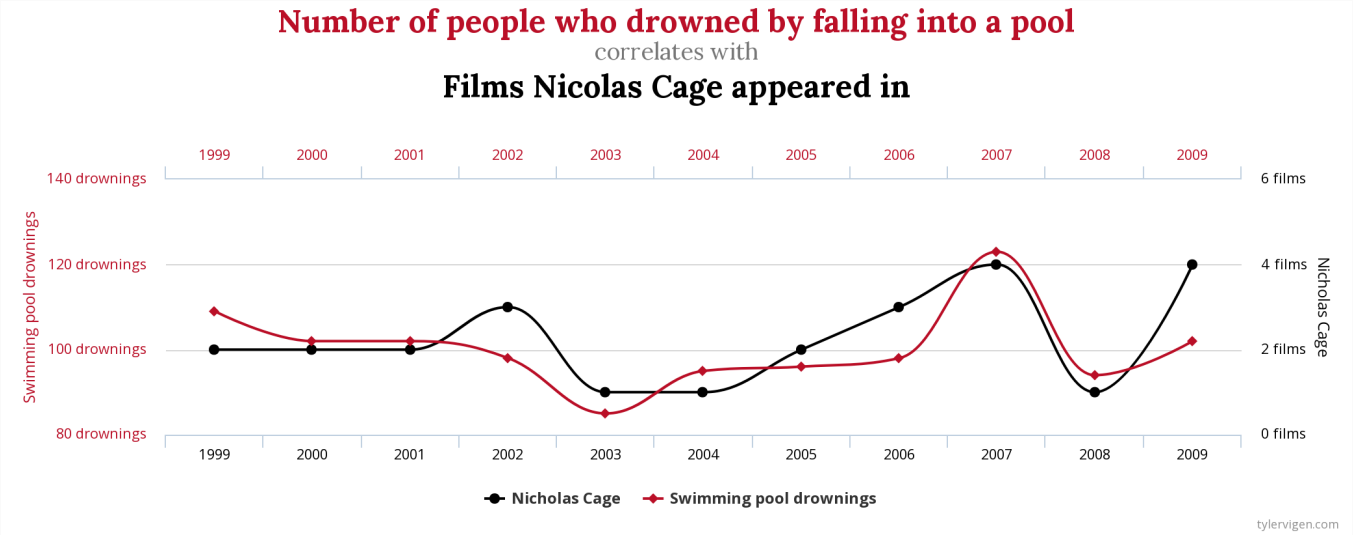


Figure 6: Correlation is NOT Causation

Studying Feelings Toward Democracy

- 2018 Pew Study
- 27 countries, $\sim$ 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

Data

- Load the data:

```
library(haven) # package to read the data
library(dplyr) # package for data manipulation
library(tidyverse) # package for 'tidy' data
pew2018 <- read_sav("Pew 2018.sav",
                    user_na=TRUE) %>%
                    as_factor()
pew2018 = pew2018 %>% select(COUNTRY, satisfied_democracy, age, sex, d_ptyid_us)
pew2018$partyid2 <- fct_collapse(pew2018$d_ptyid_us,
                         DK = c("No preference (DO NOT READ)",
                                "Other party (DO NOT READ)",
                                "Don't know (DO NOT READ)",
                                "Refused (DO NOT READ)"),
                         Rep = "Republican",
                         Ind = "Independent",
                         Dem = "Democrat")
```

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 sceen width, but does not display the number of levels and names of factor variables. But this feature of str() cannot be displayed completly 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     6
```

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

```
## [1] "COUNTRY"             "satisfied_democracy" "age"
## [4] "sex"                 "d_ptyid_us"          "partyid2"
```

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

```
## Rows: 30,109
## Columns: 6
## $ COUNTRY             <fct> United States, United States, United States, Un...
## $ satisfied_democracy <fct> Not at all satisfied, Not too satisfied, Not to...
## $ age                 <fct> 60, 69, 71, 82, 46, 57, 64, 81, 84, Refused (DO...
## $ sex                 <fct> Male, Male, Male, Female, Male, Female, Female,...
## $ d_ptyid_us          <fct> Democrat, Independent, Democrat, Republican, De...
## $ partyid2            <fct> Dem, Ind, Dem, Rep, Dem, Dem, Dem, Dem, DK, DK,...
```

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
```

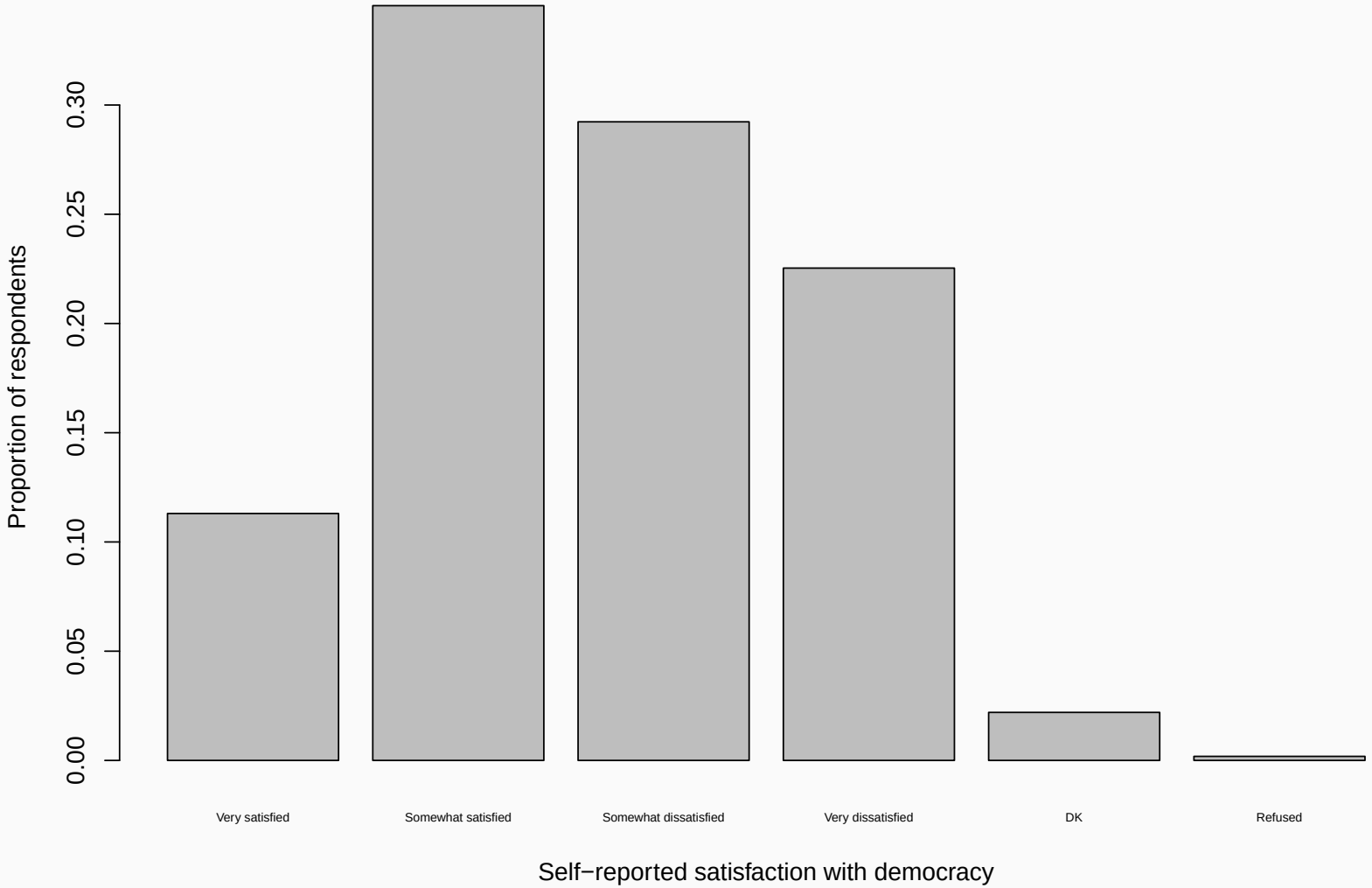
Barplot

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

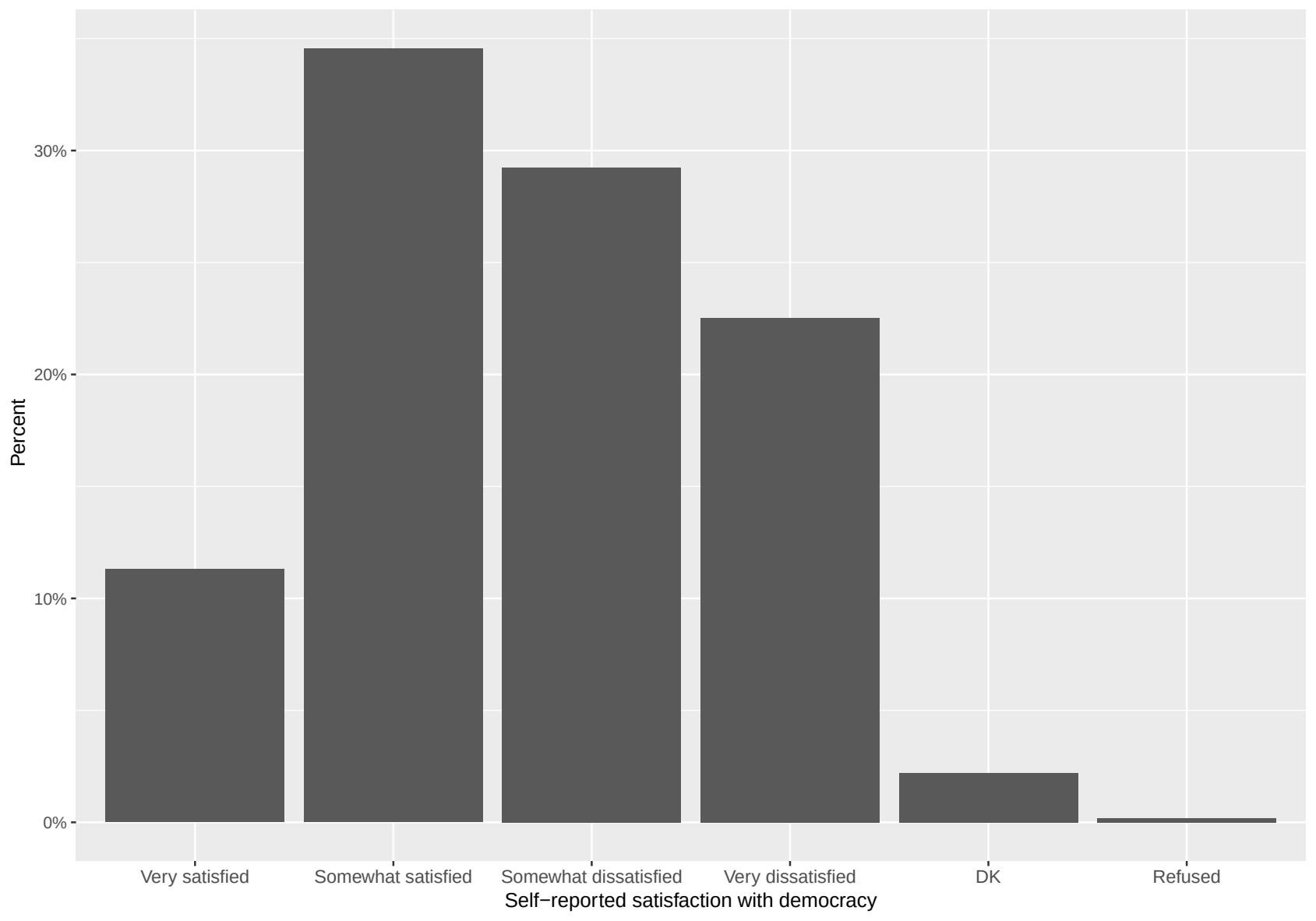
```
barplot(prop.table(table(pew2018$satisfied_democracy)),
xlab = "Self-reported satisfaction with democracy",
ylab = "Proportion of respondents")
```

- 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

Barplot


Proportion of respondents
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Very satisfied
Somewhat satisfied
Somewhat dissatisfied
Very dissatisfied
DK
Refused
Self−reported satisfaction with democracy

Bar plot (Using ggplot)



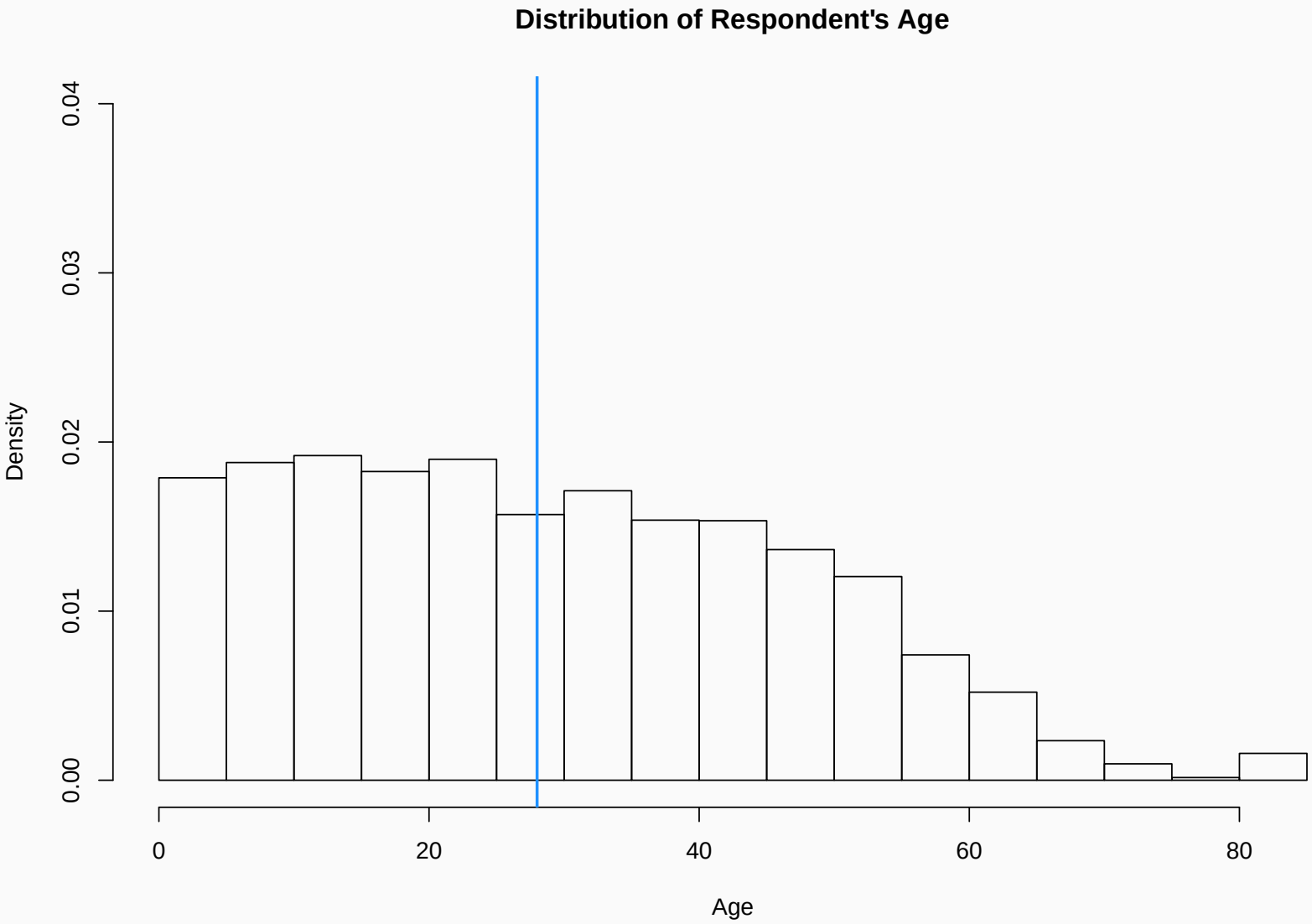
Histogram

- Visualize density of continuous/numeric variable.
- How to create a histogram by hand:
  1. create bins along the variable of interest
  2. count number of observations in each bin
  3. **density** = bin height
- In R, we use `hist()` with `freq = FALSE`:

```
hist(as.numeric(pew2018$age), freq = FALSE, ylim = c(0, 0.04),
xlab = "Age", main = "Distribution of Respondent's Age")
```

- Other arguments:
  - `ylim` sets the range of the y-axis to show (if you don't set it, uses the range of the data).
  - `main` sets the title for the figure.

Histogram


Distribution of Respondent's Age
Density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
Age

What is Density

- The areas of the blocks = proportion of observations in those blocks.
- area of the blocks sum to 1 (100%)
- Can lead to confusion: height of block can go above 1!

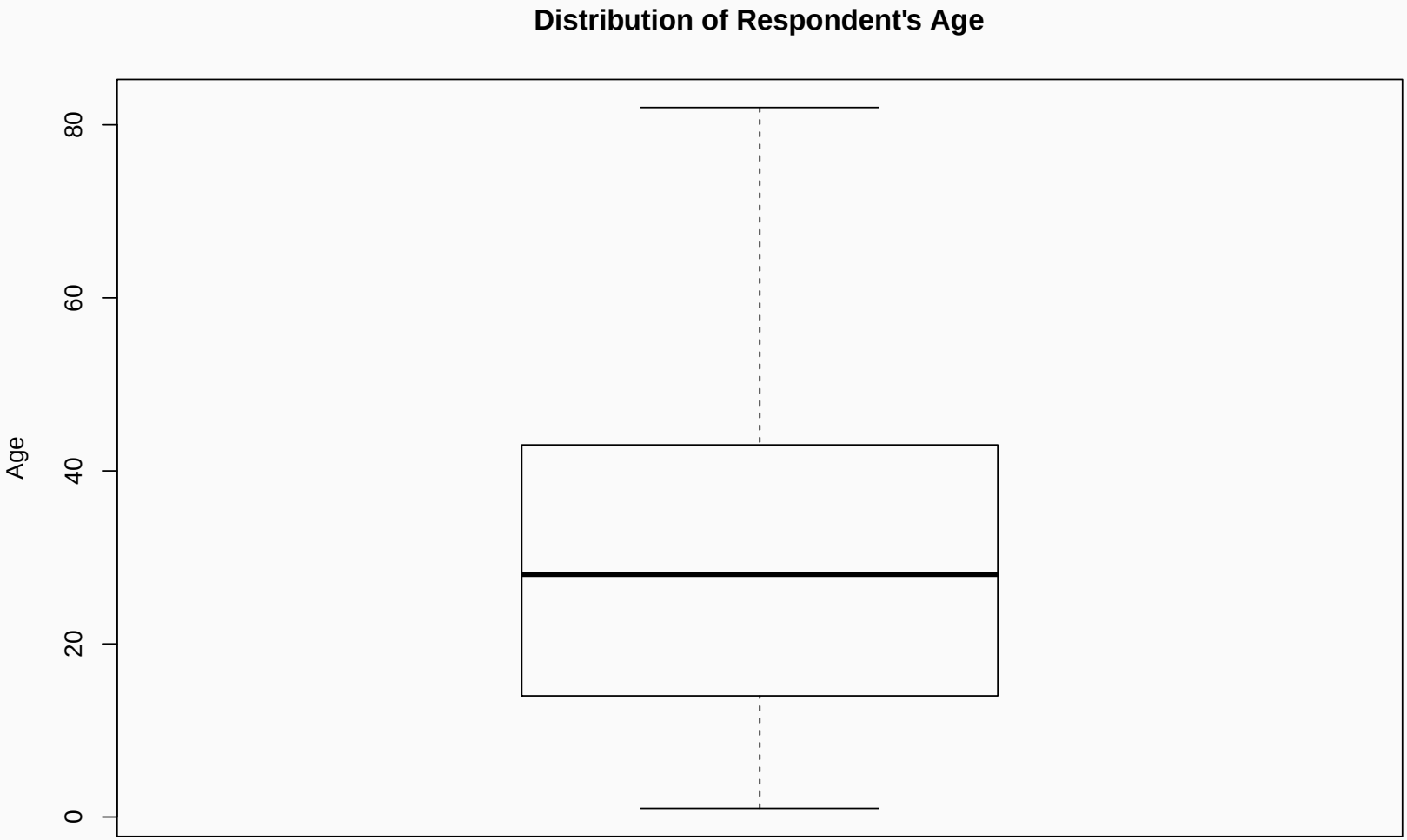
Boxplot

- A boxplot can characterize the distribution of continuous variables
- Use `boxplot()`:

```
boxplot(as.numeric(pew2018$age),
ylab = "Age",
main = "Distribution of Respondent's Age")
```

- "Box" represents range between lower and upper quartile.
- "Whiskers" represents either:
  - 1.5 $\times$ IQR or max/min of the data, whichever is tinyer.
  - Points beyond whiskers are outliers

Boxplot



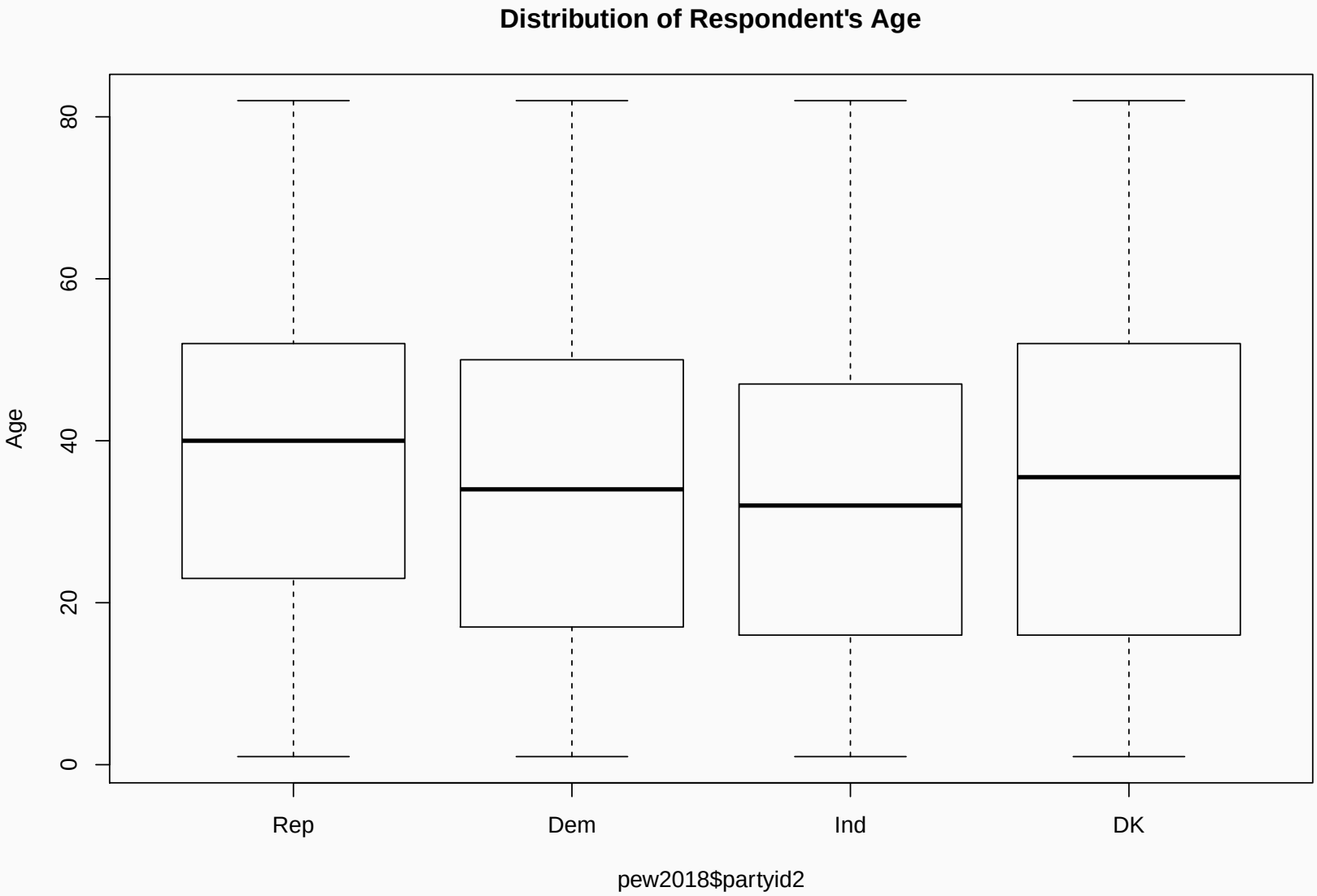
Comparing Distributions with the boxplot

- Useful for comparing a variable across groups:

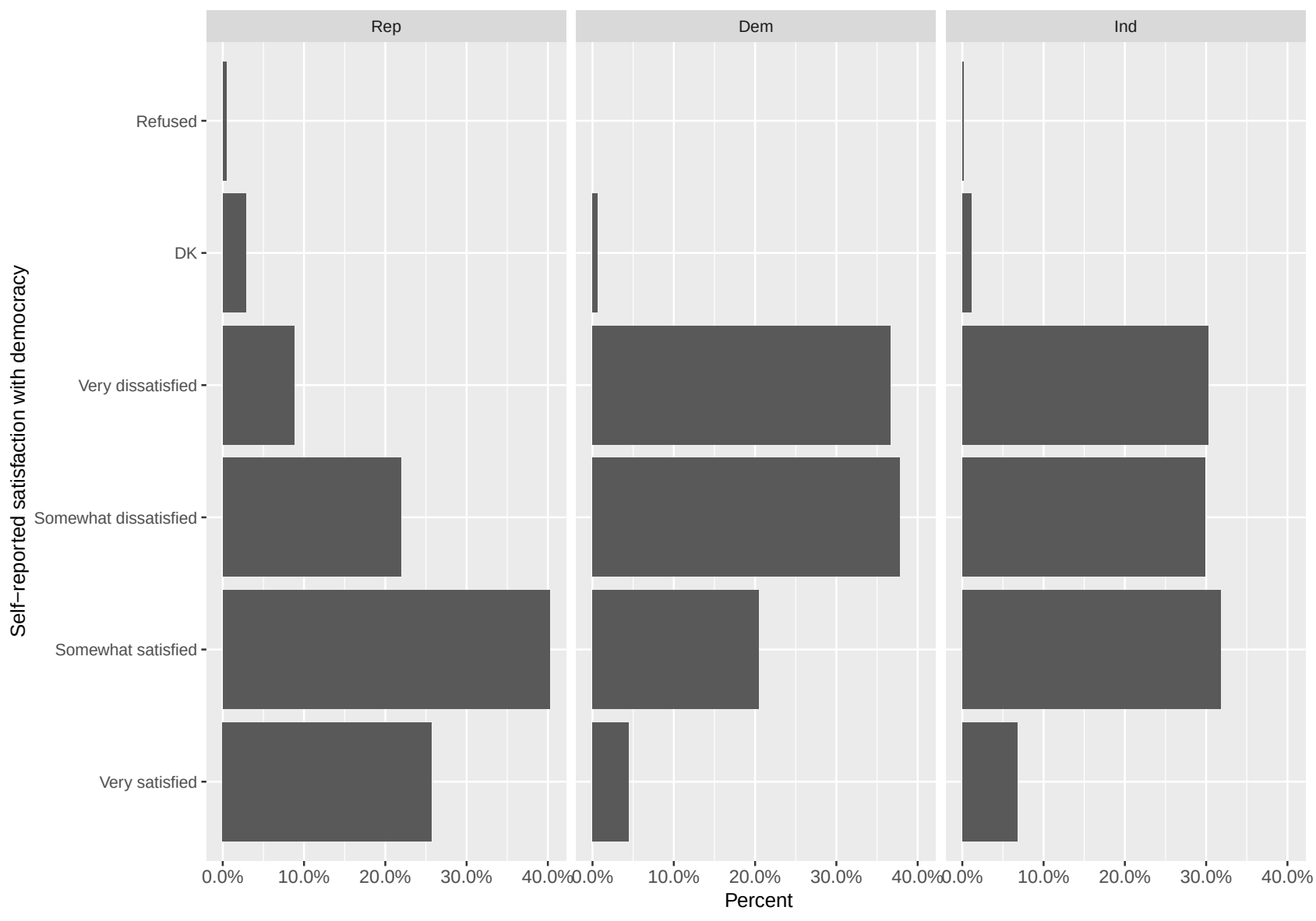
```
boxplot(as.numeric(pew2018$age) ~ pew2018$sex,
ylab = "Age",
main = "Distribution of Respondent's Age")
```

- First argument is called a formula, `y ~ x`:
  - `y` is the continuous variable whose distribution we want to explore.
  - `x` is the grouping variable.
  - When using a formula, we need to add a `data` argument

Comparing Distributions with the boxplot



Satisfaction with democracy by party



Satisfaction with democracy by party

- Why are Republicans more satified with democracy?