# Measures of Central Tendency

Mean, Median, and Mode

# Population study

- The population under study is the 65 students in the 2017 class.
- We asked how many languages they speak (besides English).
- Here were their responses:

0 1 0 0 1 1 0 0 1 1 2 1 4 1 2 1 2 2 1 2 0 1 1 2 1 0 2 0 1 0 0 1

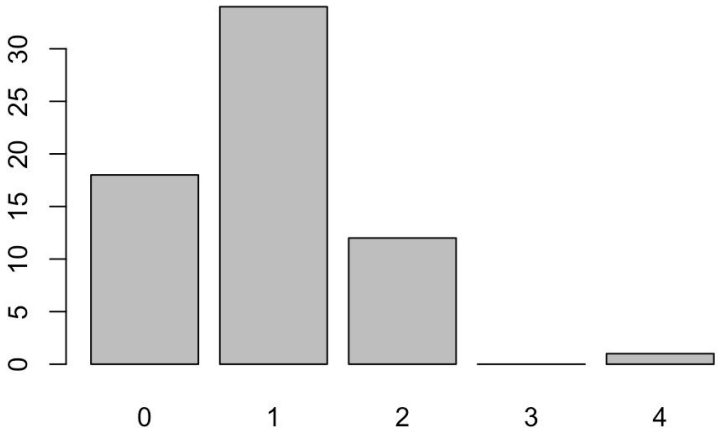
1 2 0 1 1 1 1 0 2 1 1 1 0 1 1 2 0 1 2 0 1 0 1 1 1 1 1 1 1 0 2 0 1

- Each response is a measurement, or an outcome
- In this lecture, we will discuss ways to summarize data from a sample, using these responses as an example

# Frequency distribution

- A count of the # of times each outcome occurs is called a frequency distribution.
- This information can be conveyed in a table or a plot.
- A plot of a frequency distribution of numerical data is called a histogram.

| | frequency |
|---|---|
| 0 | 18 |
| 1 | 34 |
| 2 | 12 |
| 3 | 0 |
| 4 | 1 |



# The mean of a frequency distribution

- The 2017 dataset

0 1 0 0 1 1 0 0 1 1 2 1 4 1 2 1 2 2 1 2 0 1 1 2 1 0 2 0 1 0 0 1

1 2 0 1 1 1 1 0 2 1 1 1 0 1 1 2 0 1 2 0 1 0 1 1 1 1 1 1 1 0 2 0 1

- The mean (sum/sample size) of these numbers is 0.95
- Note that no student reports knowing 0.95 languages
- Still, this number is one way of summarizing the frequency distribution:

"On average, the students in this class know 0.95 languages (beyond English)."

# Another way to calculate the mean

- Use the frequency distribution
- Calculate the sum of the data points by first multiplying the outcome and frequency columns together, and then adding up the results

  sum = (0 x 18) + (1 x 34) + (2 x 12) + (4 x 1) = 0 + 34 + 24 + 4 = 62

- The mean is then the sum divided by the sample size (65)

  mean = sum/sample size = 62/65 ∽ 0.95

| | frequency | partial_sums |
|---|---|---|
| 0 | 18 | 0 |
| 1 | 34 | 34 |
| 2 | 12 | 24 |
| 3 | 0 | 0 |
| 4 | 1 | 4 |

# Yet another way to calculate the mean

- We could also calculate the mean by multiplying the outcome column by relative frequencies (frequency/sample size), and adding up the results

  mean = (0 x 0.28) + (1 x 0.52) + (2 x 0.18) + (3 x 0) + (4 x 0.02) ∽ 0.95

- The mean of a sample is the weighted average of the distinct outcomes
- The weights are the relative frequencies with which those outcomes occur

Relative Frequencies:

18/65 ∽ 0.28

34/65 ∽ 0.52

...

| | frequency | relative_frequency | partial_weighted_sum |
|---|---|---|---|
| 0 | 18 | 0.28 | 0.00 |
| 1 | 34 | 0.52 | 0.52 |
| 2 | 12 | 0.18 | 0.36 |
| 3 | 0 | 0.00 | 0.00 |
| 4 | 1 | 0.02 | 0.08 |

# But the mean is not perfect

(No descriptive statistic ever is)

Example:

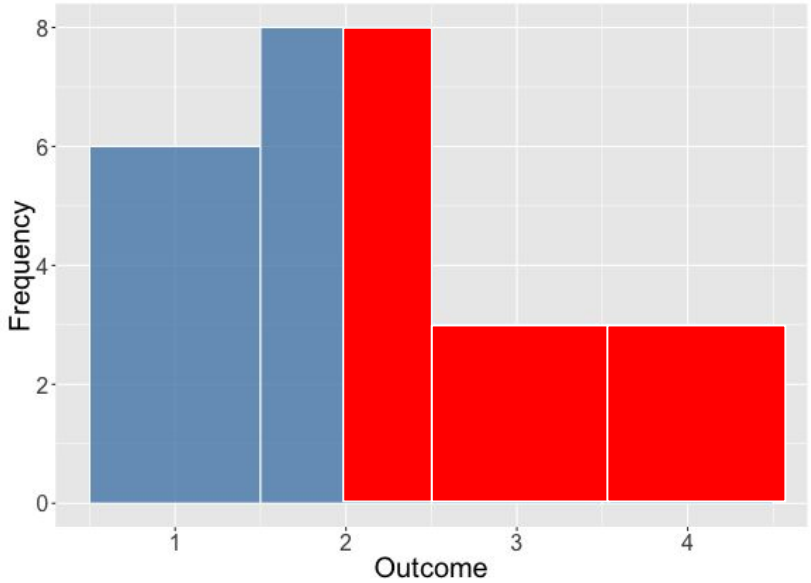
- A ten person first-year seminar has 9 first-year students who are 18 years old, and one who is 45
- The mean student age is 20.7
- Yet almost no one in the class is even 20!
- The mean is not an ideal way to summarize the data in this case

# The median of a frequency distribution

- The median is the “middle” of a frequency distribution of ordinal data
- Assume a sample: 12, 5, 3, 4, 5
- Sorting these data gives: 3, 4, 5, 5, 12
- The median value of this sample is 5
- If the sample size is even: e.g., 2, 3, 4, 5, 5, 12
- The median is the mean of the middle two numbers
- (4 + 5)/2 = 4.5
- Q: What is the median age in the aforementioned first-year seminar?

# The median is the “middle” point of a distribution

- Consider another sample: 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4
- The median is the outcome that divides the frequency distribution of outcomes in half
- 50% of the outcomes are at or below the median, and 50% are at or above it
- The area in a histogram to the right of the median equals the area to its left



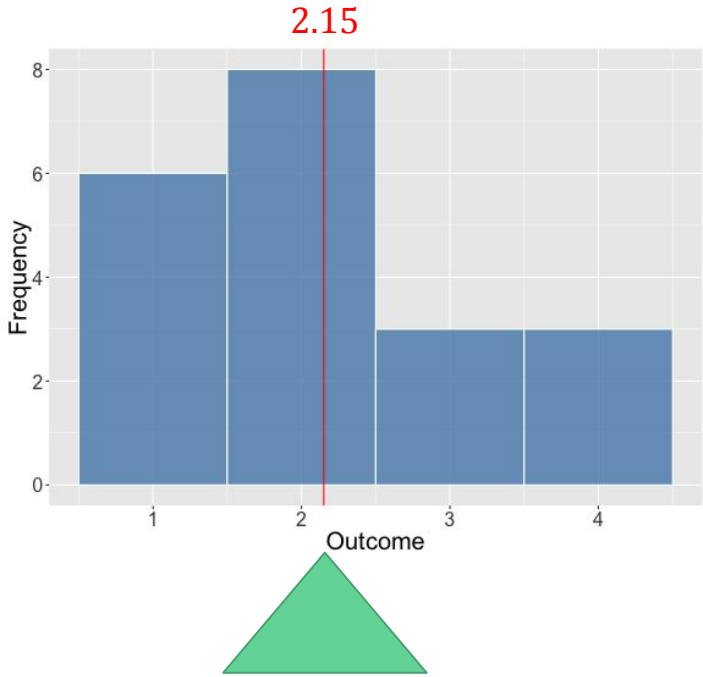

Blue Area = Red Area

$6 \times 1 + 8 \times 0.5 = 8 \times 0.5 + 3 \times 1 + 3 \times 1$

$10 = 10$

# The mean is the “balance” point of a distribution

- You can think of the mean as the a distribution’s center of gravity
- The sum total of the distances of all the points less than the mean equals the sum total of the distances of all the points greater




(2.15-1) x 6 + (2.15-2) x 8 = (3-2.15) x 3 + (4-2.15) x 3

(1.15) x 6 + (.15) x 8 = (.85) x 3 + (1.85) x 3

8.1 = 8.1

# iClicker Q: The mean vs. the median

If a student scores above the median on an exam (in the top half of the class), did the student score above the mean?

A: Yes, the student scored above the mean

B: No, the student scored below the mean

C: Not sure; there isn't enough information to answer this question

D: Not sure, because I don't understand the mean and median well enough yet to even hazard a guess (but I'm interested in learning this stuff!)

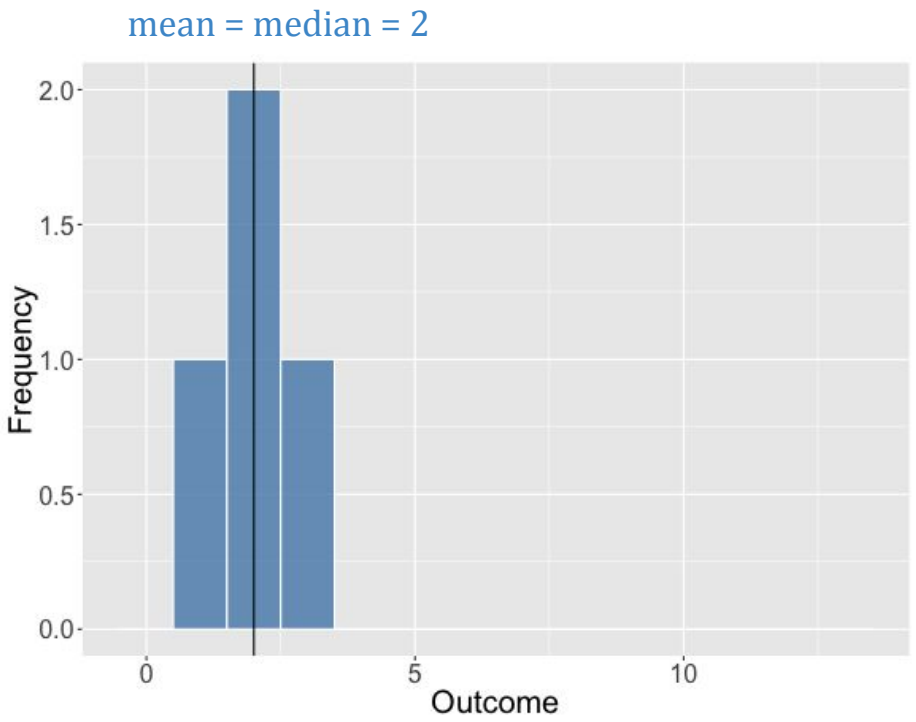
# C (or D): Not Enough Information

- The student is not necessarily above (or below) the mean
- The mean is the “balance” point of a distribution
- The median is the “middle” point of a distribution
- The order of these two points can vary with the distribution

# Symmetric (i.e., non-skewed) distributions

- In a symmetric distribution, the left mirrors the right
- So the mean (the balance point) equals the median (the middle)
- E.g., the mean and the median of the frequency distribution for the sample 1, 2, 2, 3 are both 2

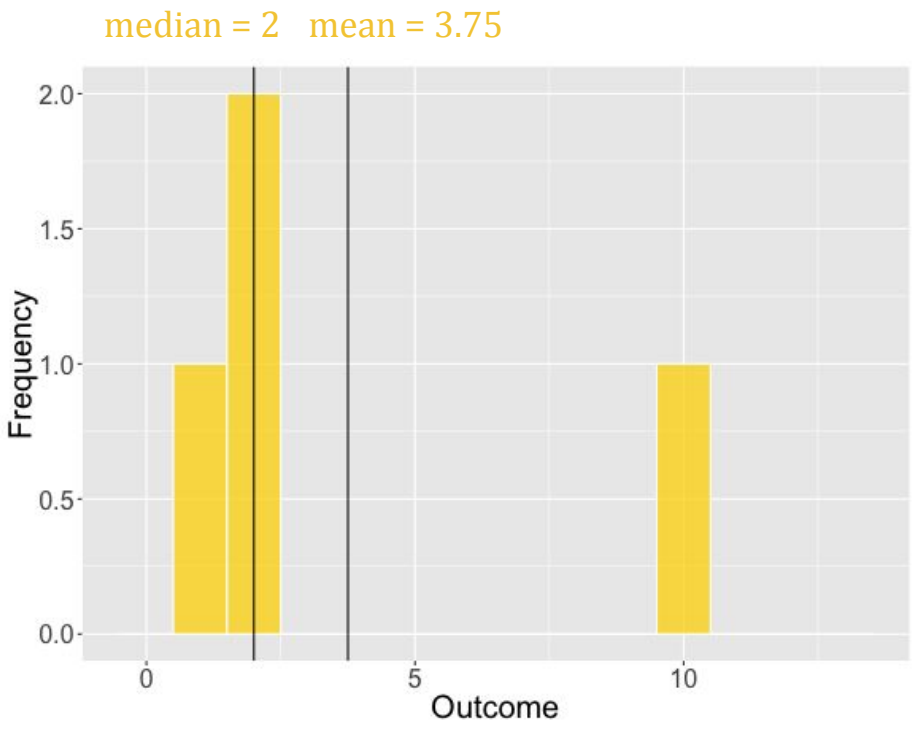
| | outcome | frequency |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 2 |
| 3 | 3 | 1 |



# Asymmetric (i.e., skewed) distributions

- In an asymmetric distribution, the left does not mirror the right
- So the mean and the median are not necessarily equal
- E.g., the mean and the median of the frequency distribution for the sample 1, 2, 2, 10 are not equal

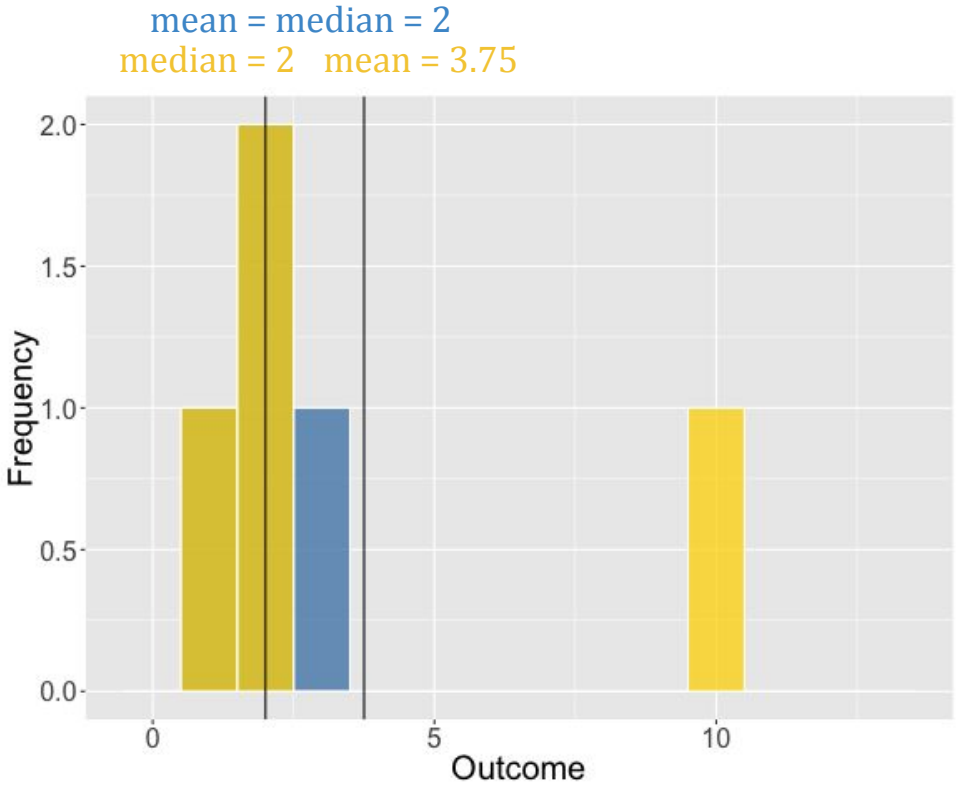
| | outcome | frequency |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 2 |
| 3 | 10 | 1 |



# Mean vs. Median

- The gold sample has an outlier: 10
- So its histogram is skewed right
- Likewise, its balance point (the mean) falls to the right of the middle point (the median)
- Ultimately, only 25% of the outcomes are above the mean

| | outcome | blue_frequency | gold_frequency |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 |
| 3 | 3 | 1 | 0 |
| 4 | 10 | 0 | 1 |



# iClicker Q: The mean vs. the median

If a student scores above the median on an exam (in the top half of the class), did the student score above the mean?

A: Yes, the student scored above the mean

B: No, the student scored below the mean

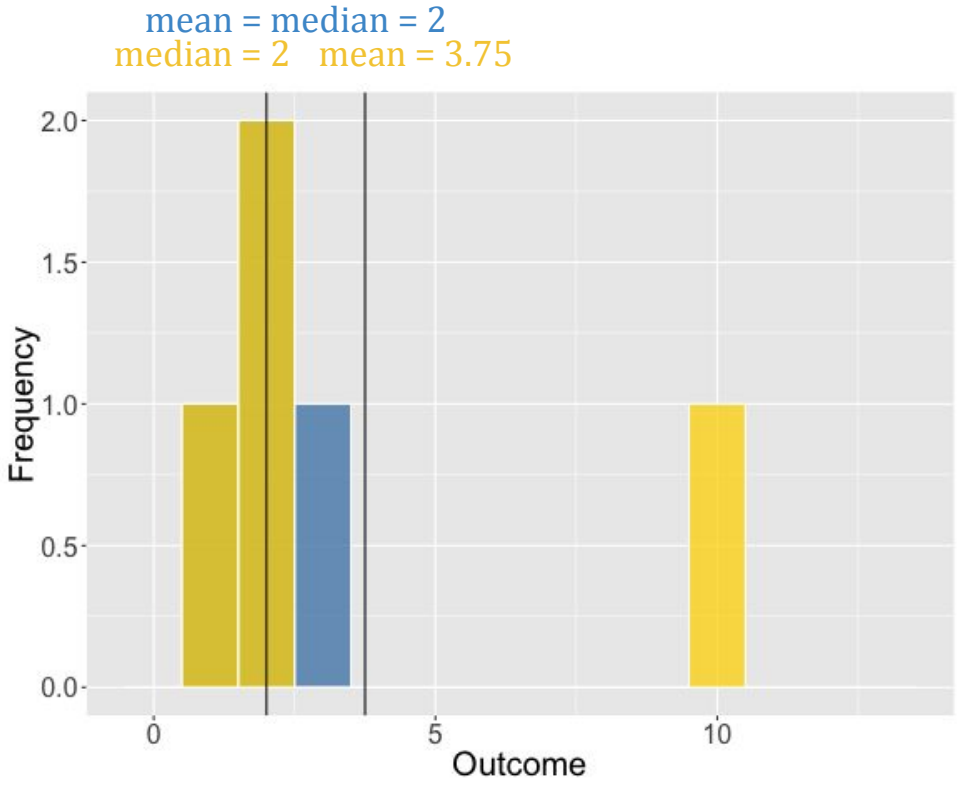
C: Not sure; there isn't enough information to answer this question

D: Not sure, because I don't understand the mean and median well enough yet to even hazard a guess (but I'm interested in learning this stuff!)

# C: Not Enough Information

- A student who scores above the median might have scored 10, but they also might have scored 3
- It's impossible to tell if they scored above the mean or not!

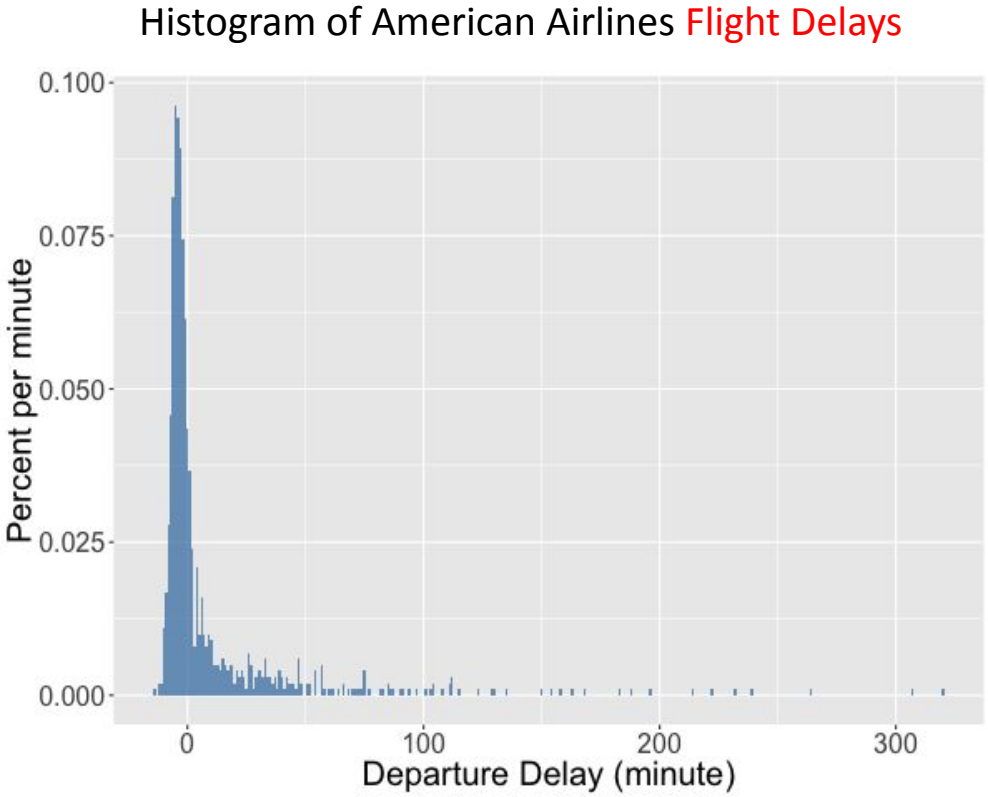
| | outcome | blue_frequency | gold_frequency |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 |
| 3 | 3 | 1 | 0 |
| 4 | 10 | 0 | 1 |



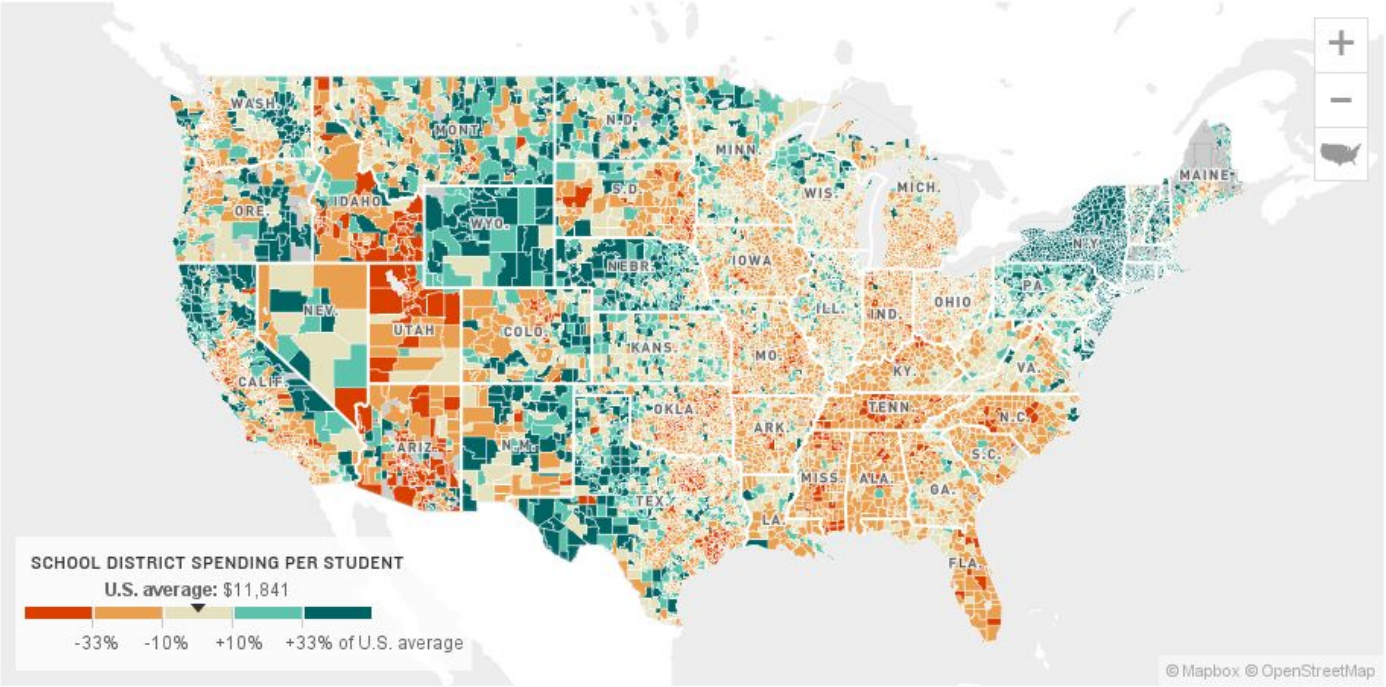
# A real-world example of a skewed distribution

- Delay times are in minutes
- A negative observation means the flight left earlier than scheduled
- The histogram is right skewed
- It has a long right-hand tail
- The mean is being pulled away from the median in the direction of the tail, so we expect the mean to be greater than the median:

mean = 9.135913  median = -2



Another example of a skewed real-world distribution: Public school funding



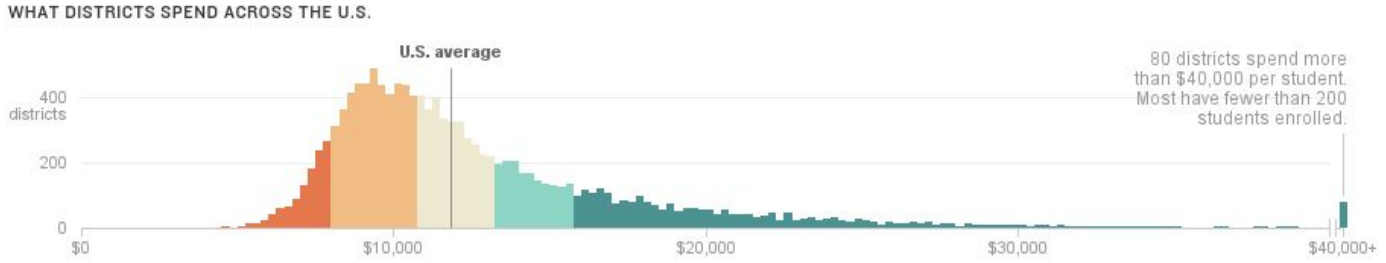




Image source

# Mean vs. median, rule-of-thumb

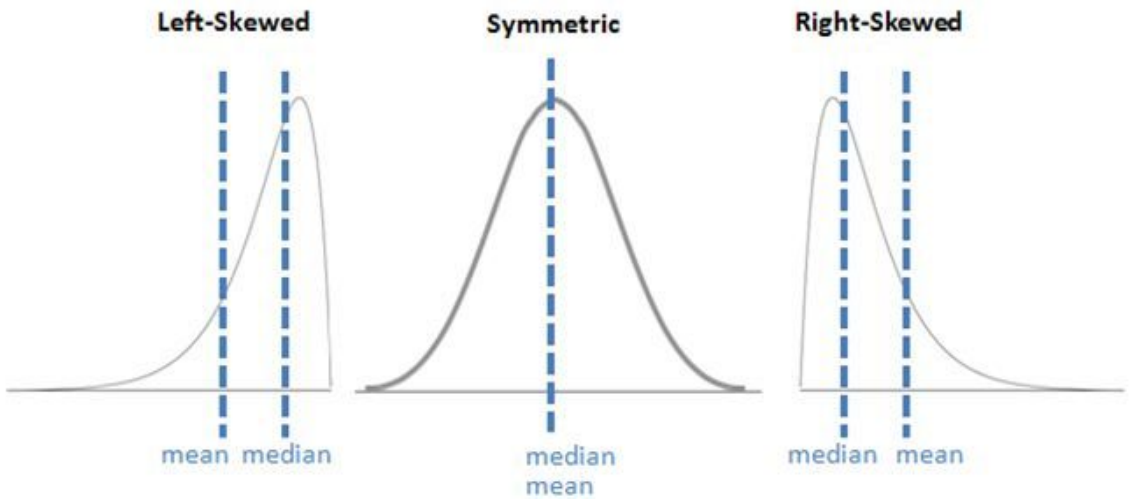



Image Source

# Bush Tax Cuts (2001/3)

- GOP reported that 92 million Americans would get a tax cut, averaging $1,083
- But actually, the median was less than $100
- Incredibly rich outliers received much larger tax cuts and skewed the mean

# Bernie Sanders' Talking Points (2016)

- Average contribution is $27
- Distribution should have been right-skewed (because it is lower-bounded by 0)
- So isn't the median smaller still?
- I'm guessing it wasn't

# Rule of Thumb

- For skewed distributions, the mean lies in the direction of the skew (towards the longer tail) relative to the median
- But this rule of thumb is not always true!

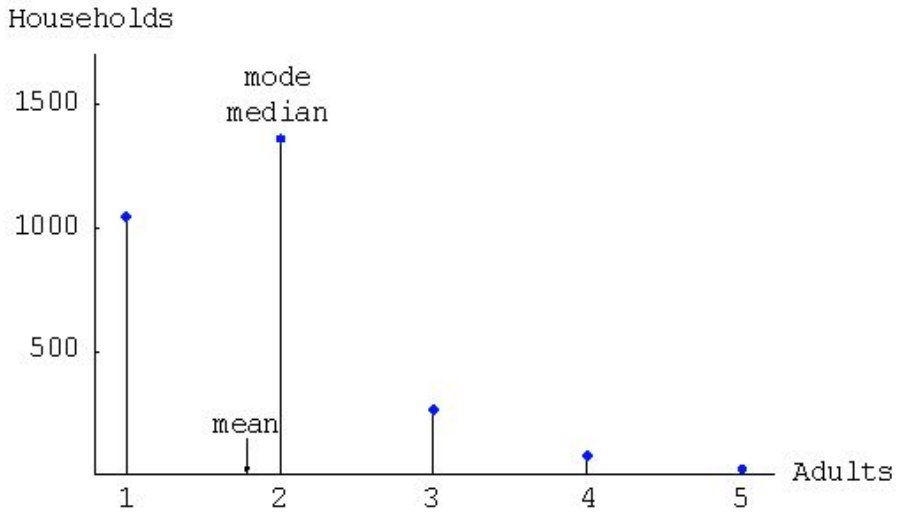



Image Source

# Mode

- The mode is the most popular outcome; the one that occurs most often.
- The mode is useful for summarizing categorical data. In the histogram of letters in English text, the mode is e.
- In plurality voting, the winning candidate is the mode; the one with the most votes.

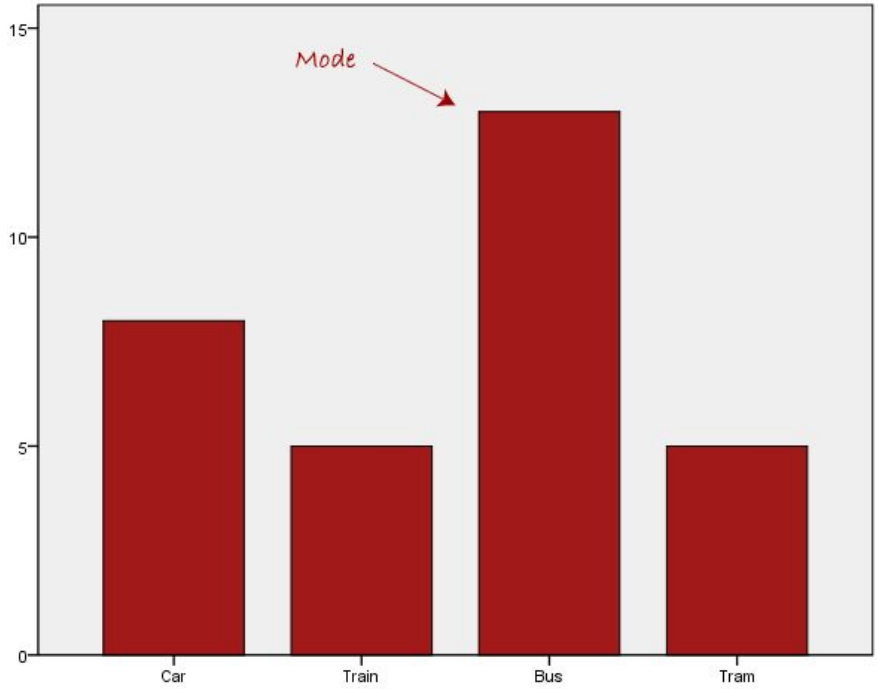



Image Source

# Mode

- Continuous data are often discretized before computing the mode.
- The mode is intended as a measure of central tendency.
- But the mode need not be unique; some histograms are bi- or multi-modal.
- And the mode may not be representative of the bulk of the data.

Image Source

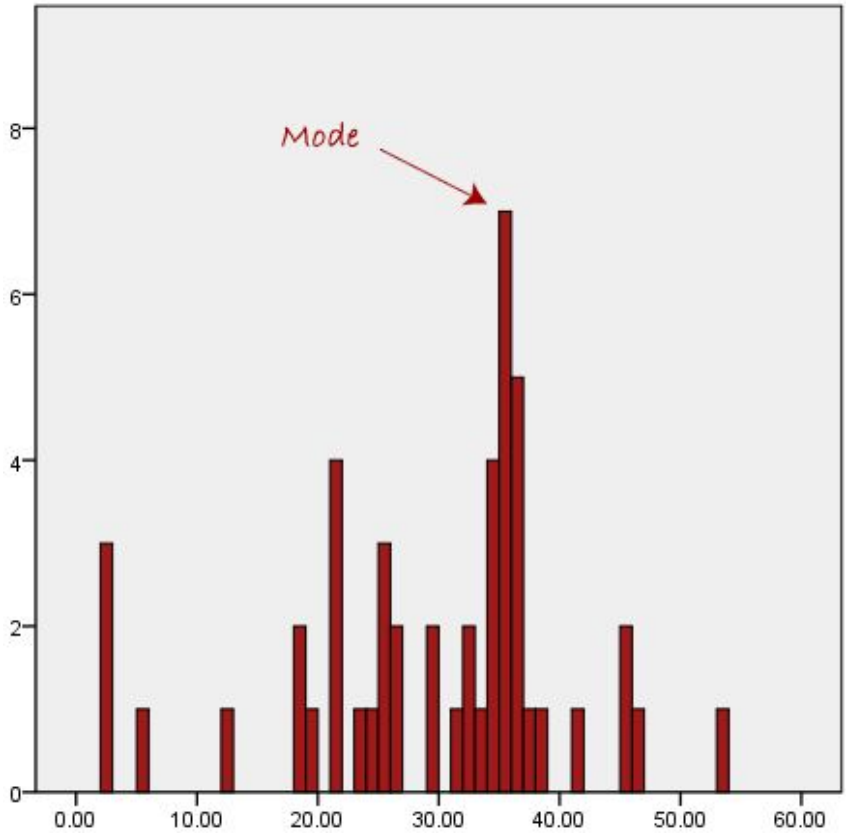
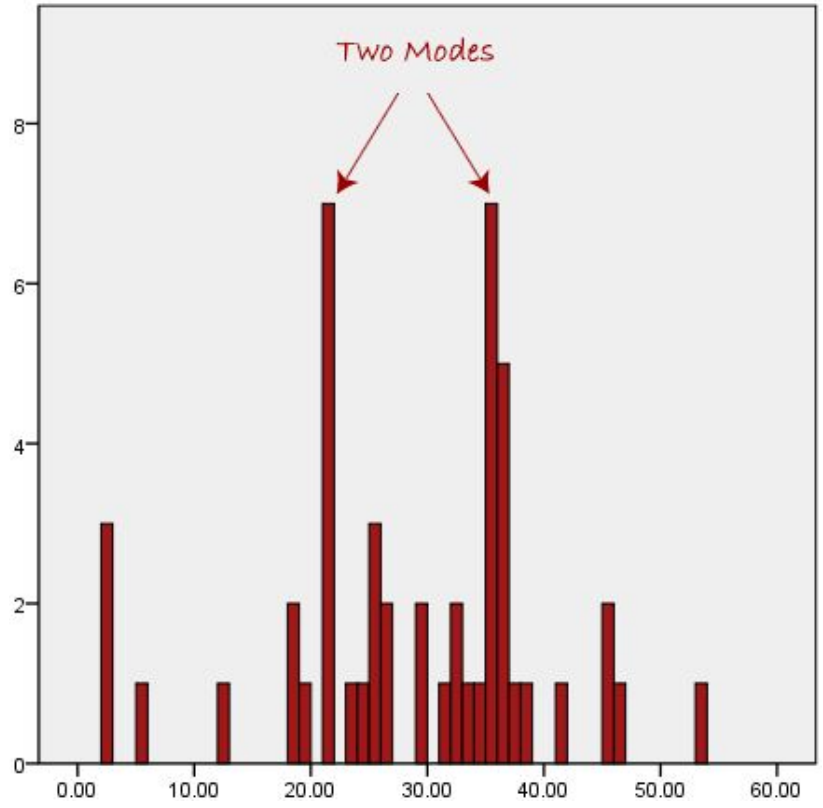
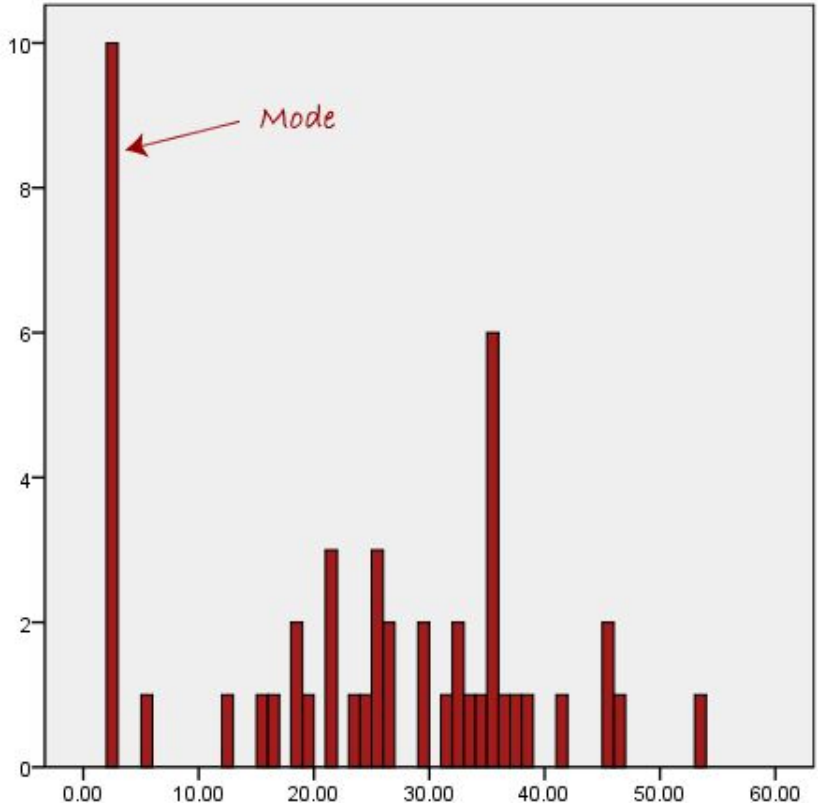


Image Source

# Mode



# Measures of Central Tendency

- Mean: numerical, preferably symmetric data without outliers
- Median: ordered, possibly skewed data, robust to outliers
- Mode: any data, but the only choice for categorical data
  - e.g., hair color, food items, movies, etc.