Lecture 4
Shape of distribution and Measures of Central Tendency

Review From Friday 1/19

3 features of a distribution that we are interested in:

- Shape
- Center
- Spread or variability

Graphs of data are a good way summarize patterns in data

Graphs for qualitative data are

- Bar graphs, pie charts

Graphs for quantitative data are:

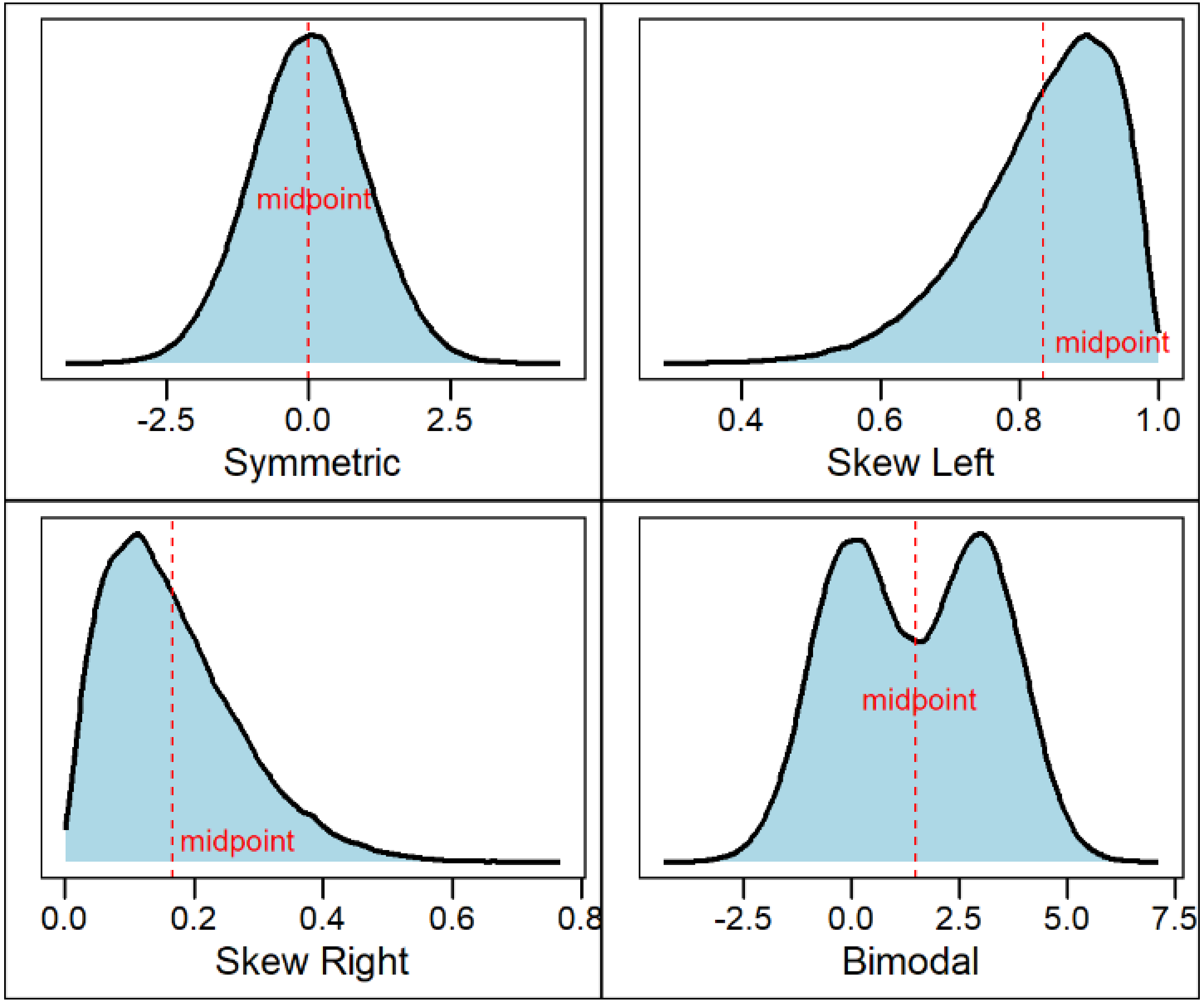
- Stem plot, dot plot, histogram

Practice: Histogram

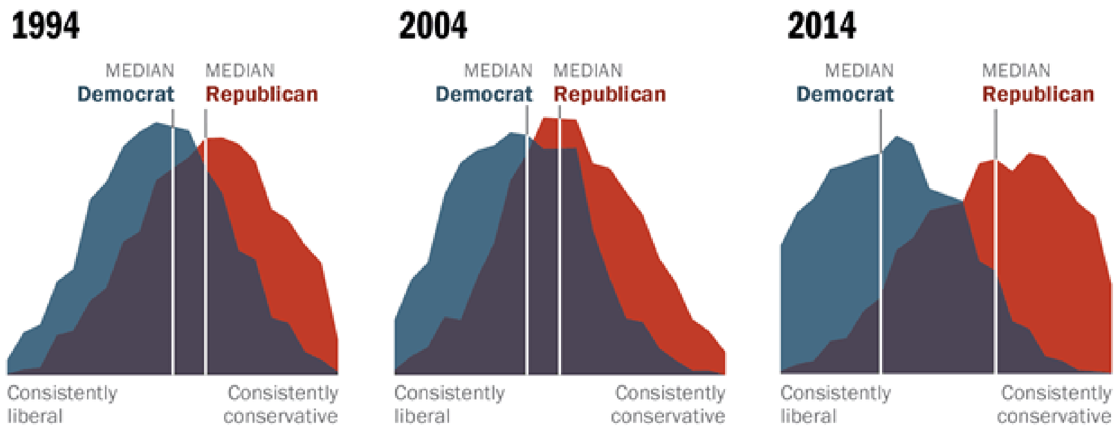
- $X = \{-1.49, -0.65, -0.6, -0.54, -0.45, 0.01, 0.17, 0.27, 0.51, 1.34\}$

Construct a histogram using K = 4 bins/intervals:

Shape of a distribution



- Bimodal distributions can arise when
  - A population is polarized on a controversial issue
  - When observations come from two different sub-populations



Democrats and Republicans More Ideologically Divided than in the Past

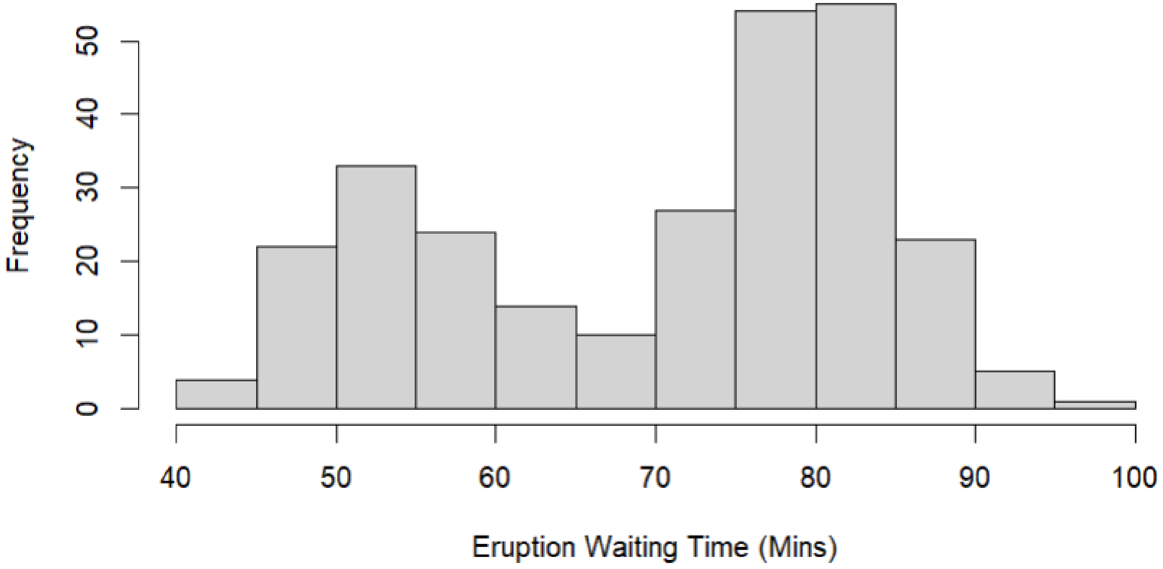
Distribution of Democrats and Republicans on a 10-item scale of political values

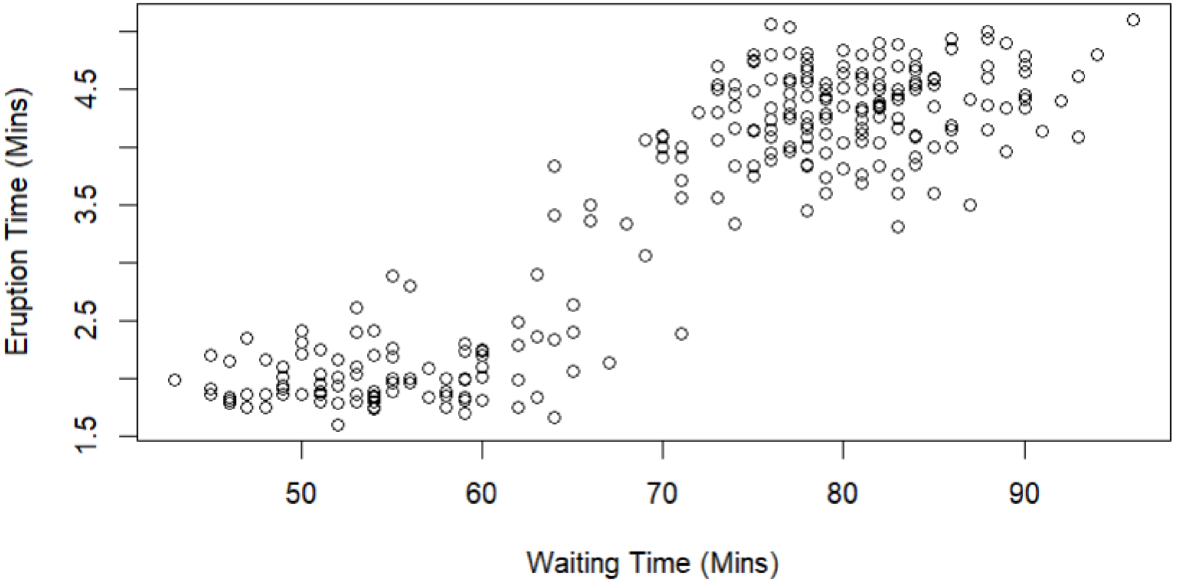
Source: 2014 Political Polarization in the American Public

Notes: Ideological consistency based on a scale of 10 political values questions (see Appendix A). The blue area in this chart represents the ideological distribution of Democrats; the red area of Republicans. The overlap of these two distributions is shaded purple. Republicans include Republican-leaning independents; Democrats include Democratic-leaning independents (see Appendix B).

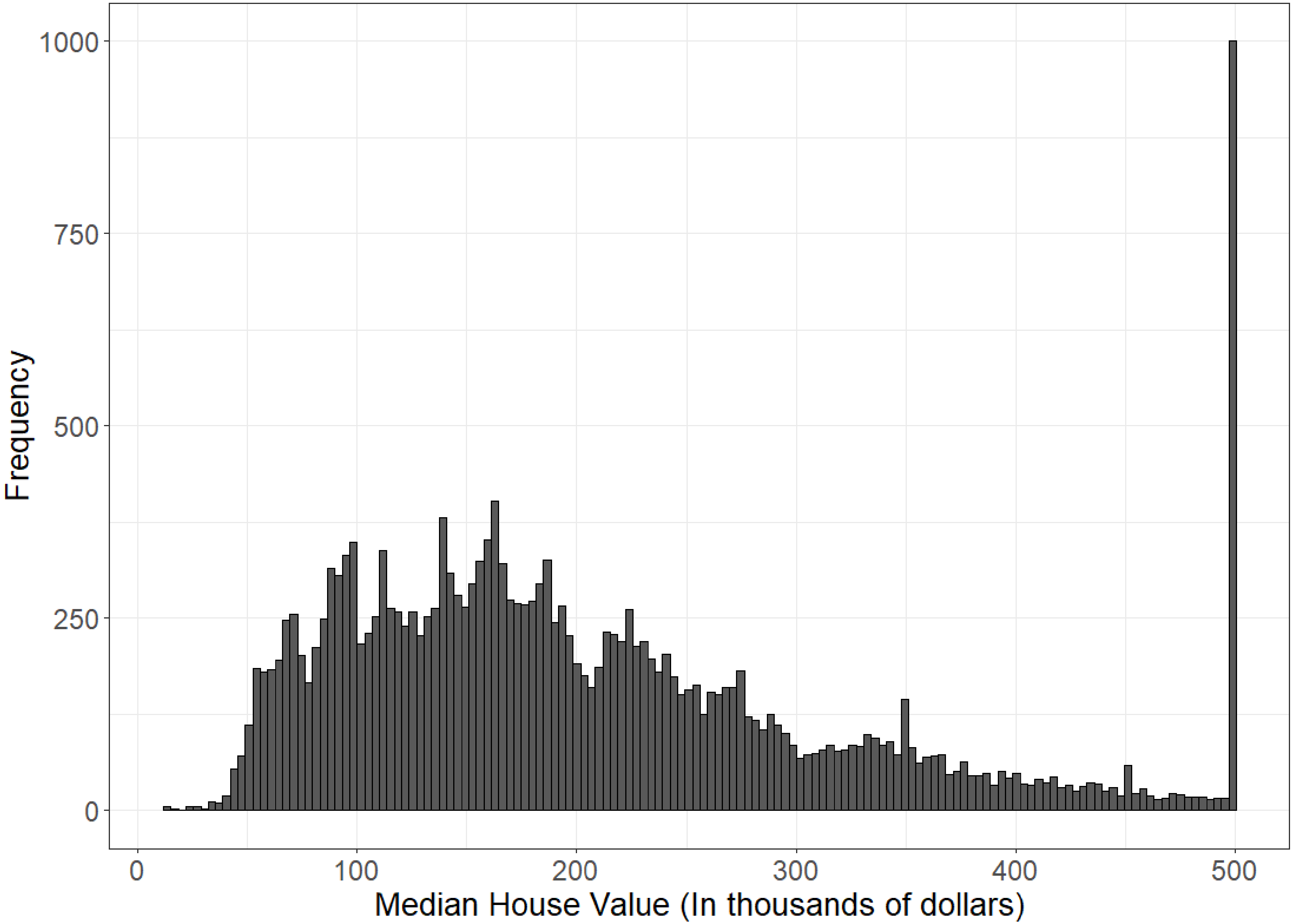
PEW RESEARCH CENTER

Histogram of Eruption Waiting Times





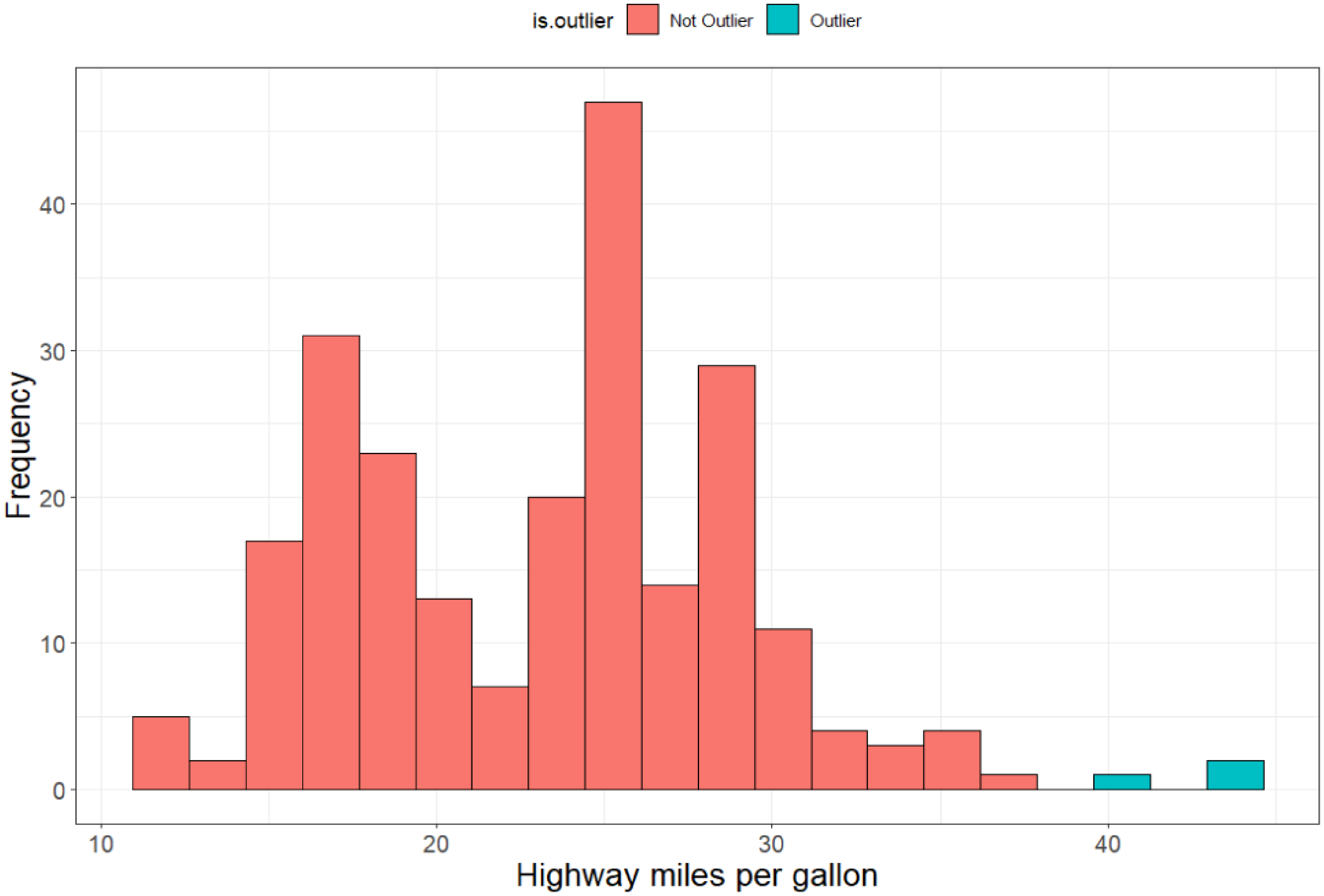
- Consider the following histogram of median housing prices in California from the 1990 national census



- Skewed distributions occur when there is a strict boundary on the possible values of a variable

- **Outliers** are extreme values that fall far away from the midpoint of the data

- Consider the following histogram of the fuel efficiency of cars from 1990 - 2008



Measures of Central Tendency

- The (arithmetic) **mean** is the average value of a set of observations

  it measures the center of mass of a distribution (the balancing point)

$$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

➢the mean is usually not equal to any of the values observed in the sample

➢The mean is highly influenced by **outliers** - observations that take on extreme values relative to the distribution

Practice: Calculate The Mean

- $X = \{1, 3, 5, 5, 6, 7, 7, 8\}$

Measures of Central Tendency

- The **median** is the middle value of a set of observations

How to compute the median:

1. Compute the median by first ordering the observations from smallest value to largest value and choose the number in the middle

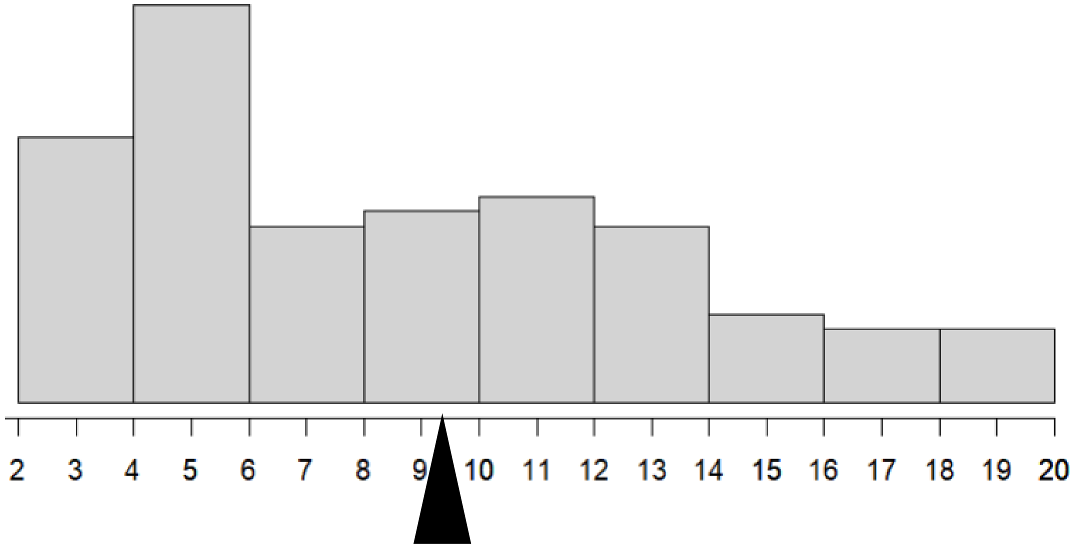
2. If the n is odd the median is the middle number

   - If n is even the median is the sum of the two middle values divided by 2

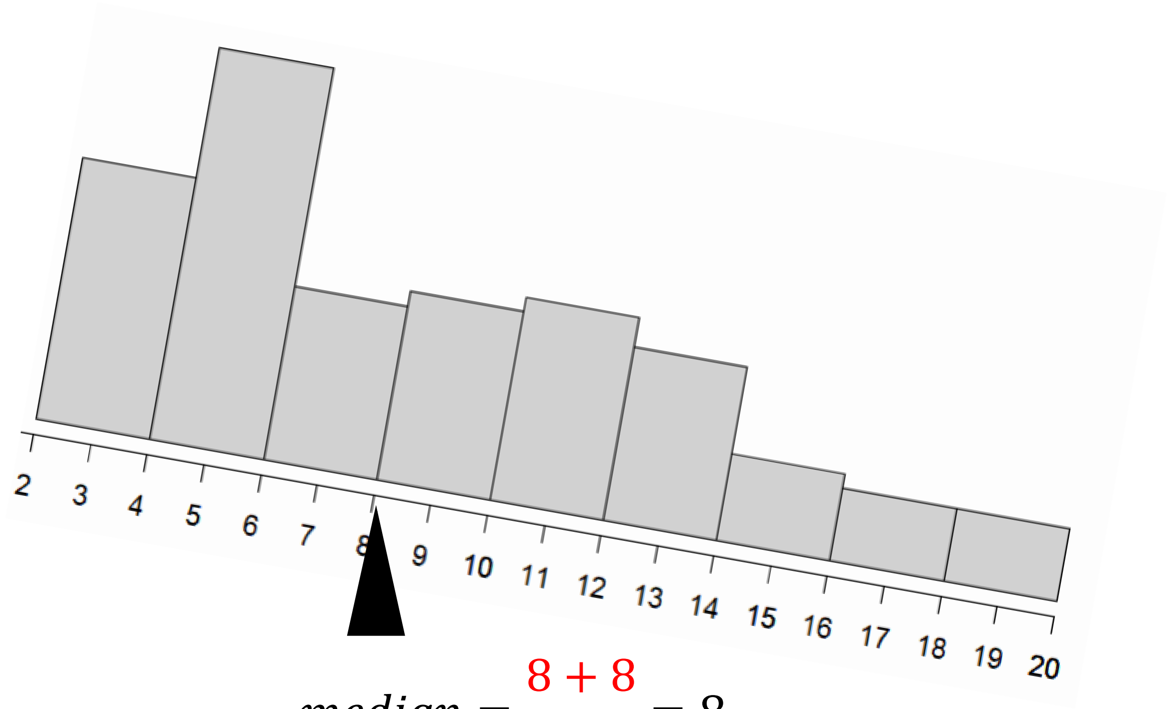
Practice: Calculate the Median

- $X = \{1, 3, 5, 5, 6, 7, 7, 8\}$

Data: 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5
5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7
7 8 8 8 8 8 8 9 9 9 9 9 9 9 10 10 10 10 10 10 10 11 11 11 11 11 12
12 12 12 12 12 12 12 13 13 13 14 14 14 14 14 14 15 15 15 16 16 16 16 17 17 17 18
18 20 20 20 20 20



$\bar{x} = 9.2$



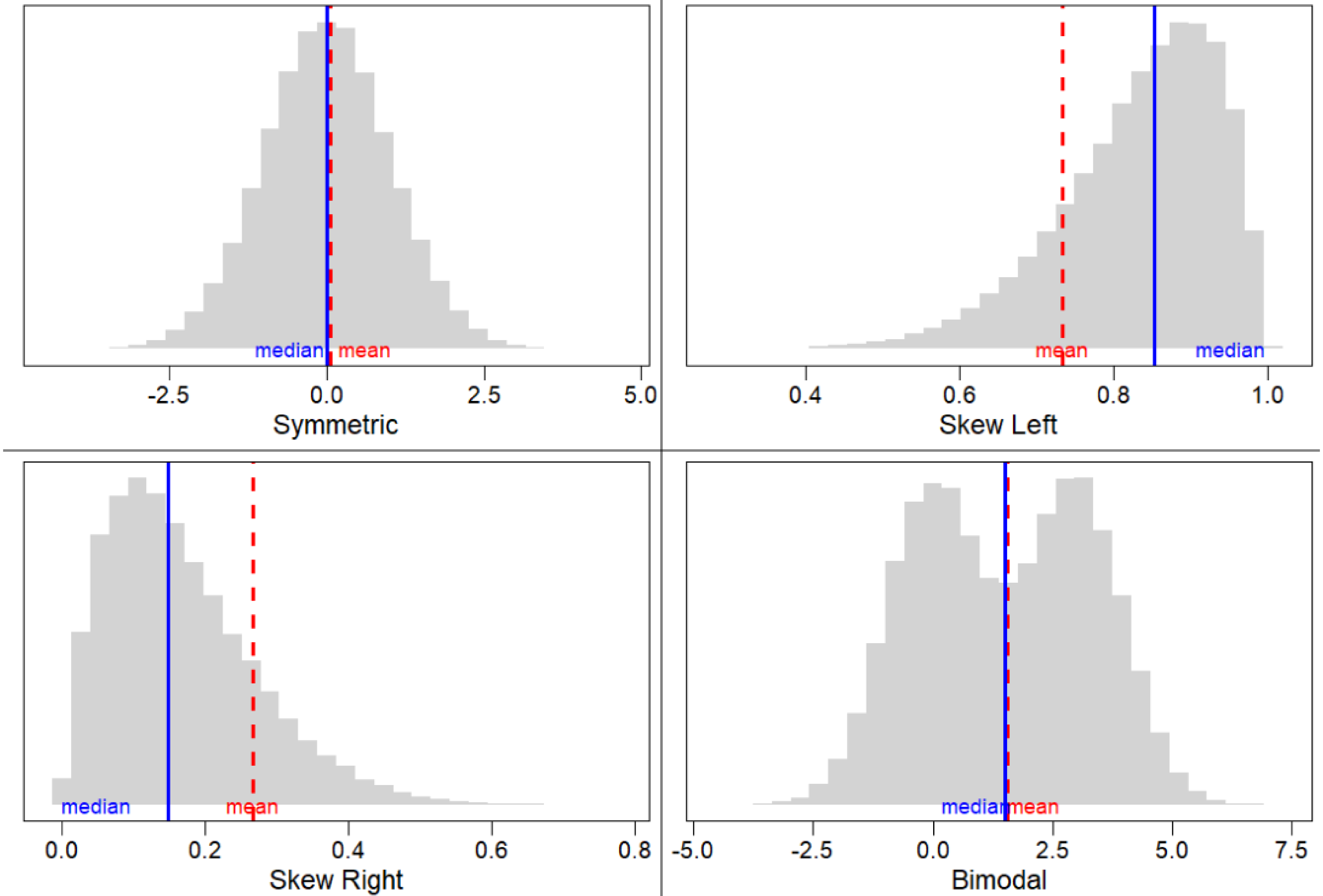
$$median = \frac{8+8}{2} = 8$$

The mean is the center of gravity

The median is the middle value

Mean and median treat outliers differently

- $X = \{1, 3, 5, 5, 6, 7, 7, 8, 32\}$


median
mean
-2.5
0.0
2.5
5.0
Symmetric
mean
median
0.4
0.6
0.8
1.0
Skew Left
median
mean
0.0
0.2
0.4
0.6
0.8
Skew Right
median
mean
-5.0
-2.5
0.0
2.5
5.0
7.5
Bimodal

Alternative formulas for the mean

- We can also express the mean in terms of the frequency F or the relative frequency RF

$$\bar{x} = \frac{1}{n} \sum_{x} x \, F(x) \qquad \text{or} \qquad \bar{x} = \sum_{x} x \, RF(x)$$

Where the sum is over all distinct values of the variable x

Example: Computing the mean from a frequency table

$X = \{1, 3, 5, 5, 6, 7, 7, 8\}$

X	Freq.	Rel. Freq
1	1	0.125
3	1	0.125
5	2	0.250
6	6	0.125
7	2	0.250
8	1	0.125

The mode

- The **mode** is the value with the largest relative frequency (i.e the value that occurs most often)

  - Can be used with categorical data (mean and median cannot)

    - e.g the most frequent category

  - It may not be unique if two or more values have the same frequency

  - <u>**Caution**</u> for quantitative data, the mode <u>may not</u> anywhere near the center of the distribution.

Ex.)

Data = $\boxed{1,1,}4,5,6$

Mode = 1

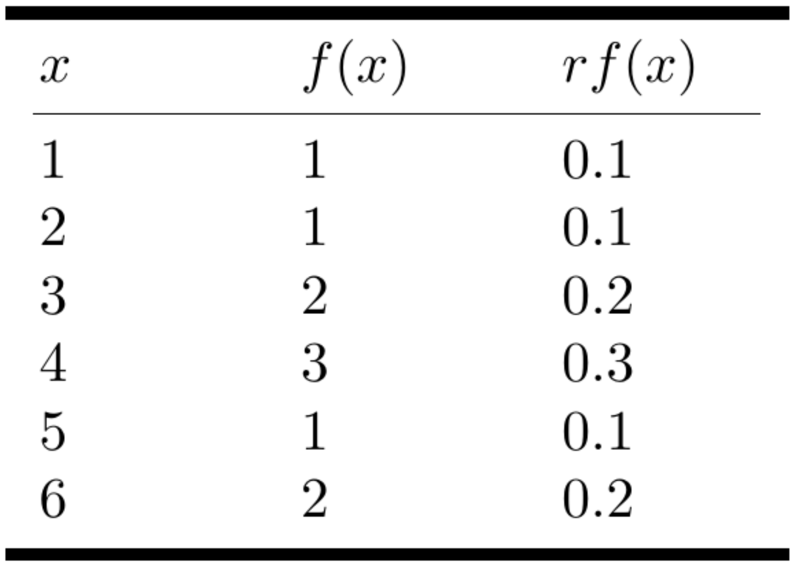
Data = $\boxed{1,1,}4,5,\boxed{6,6}$

Mode   1, 6

Practice:

- Roll a six-sided die $n = 10$ times and record the number rolled each time

- Data = 1,2,3,3,4,4,4,5,6,6



x	$f(x)$	$rf(x)$
1	1	0.1
2	1	0.1
3	2	0.2
4	3	0.3
5	1	0.1
6	2	0.2

Compute the **mean** using all 3 equations:

Compute the **median**

Compute the **mode**

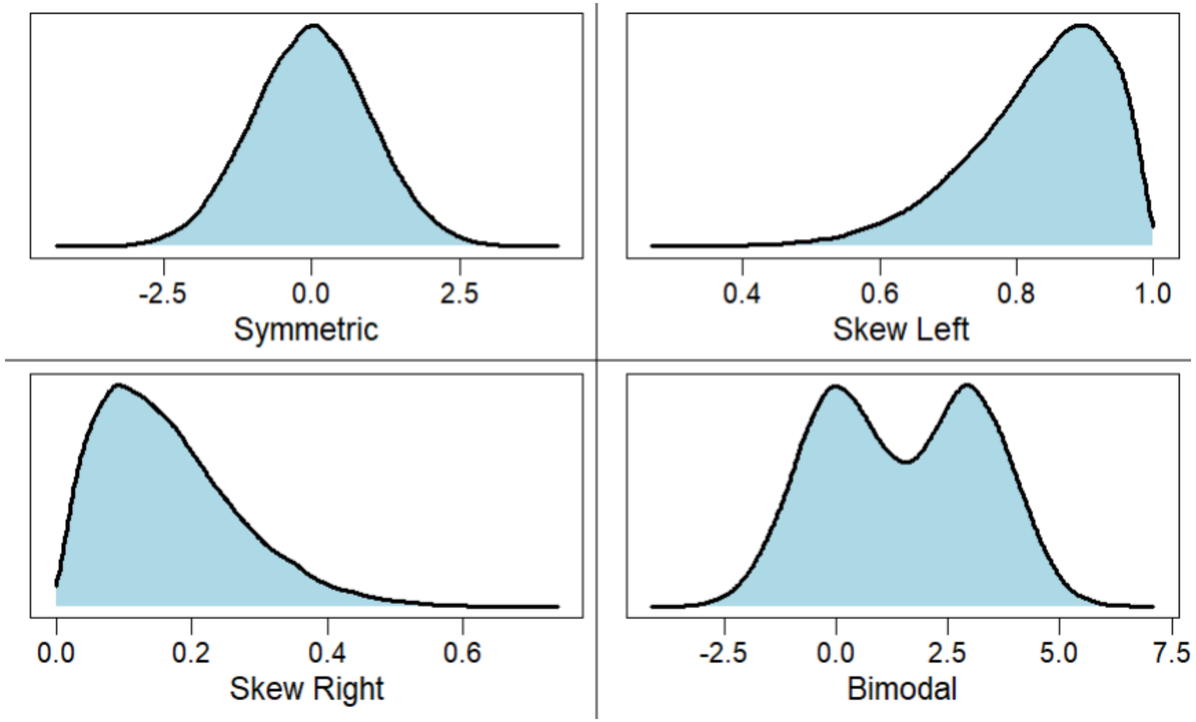
Comparing the Mean, Median, and Mode

- The shape of a distribution influences whether the mean is larger or smaller.

- Skew left = mean < median

- Skew right = mean > median

- When a distribution is symmetric the mean will equal the median



Comparing the Mean, Median, and Mode

- The median is a robust estimate of the mean
- The median is not usually affected by the presence of outliers
- The median is usually preferred for highly skewed distributions

- Ex.) take using the following 9 data points: 0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, 16.9

  The **mean** is about 4.58

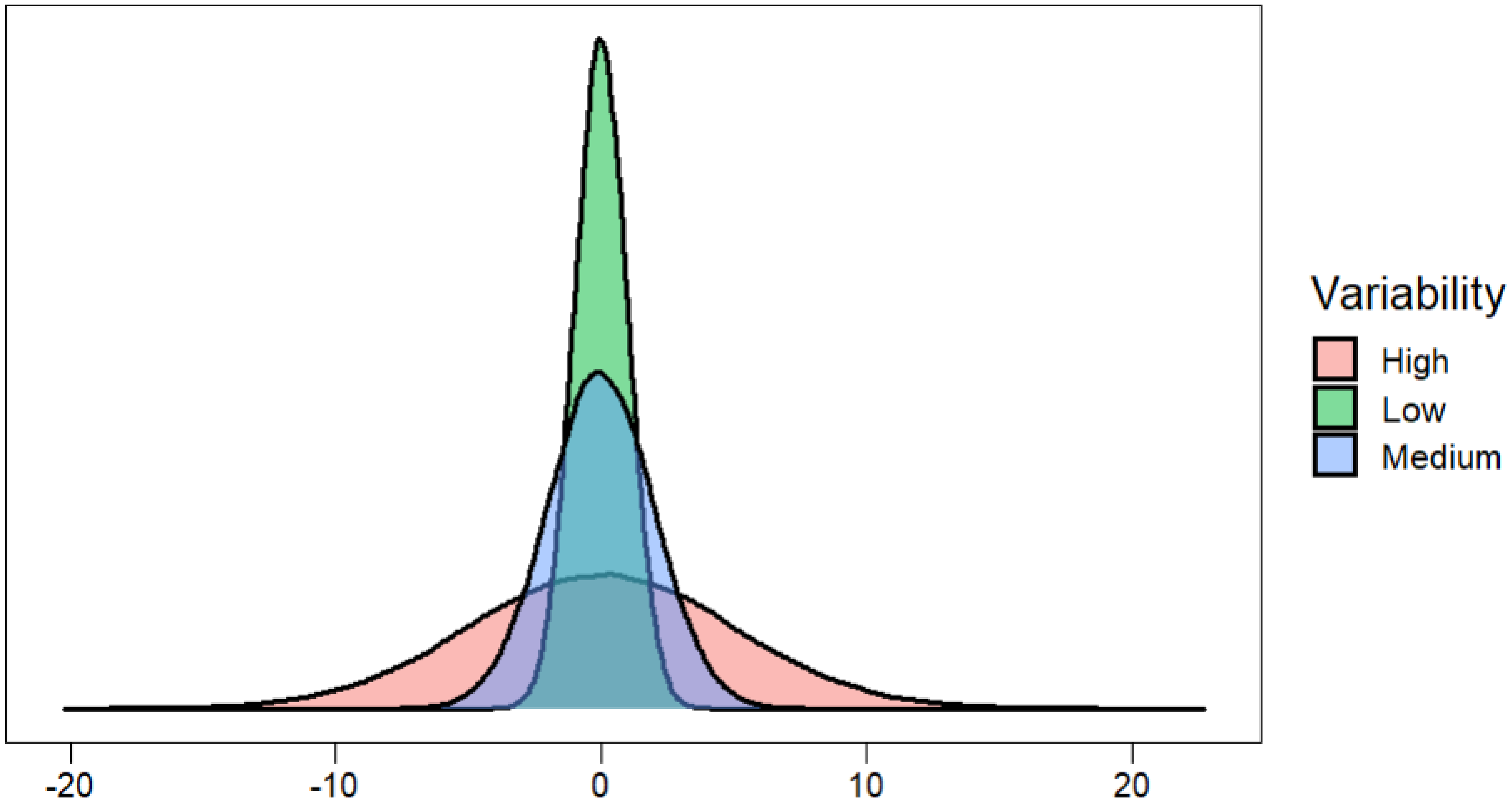
  The **median** is 1.8

- Change one of the data points to be an outlier, for example, we change **16.9** to **90**

  The **mean** becomes 12.7

  While the **median** is still 1.8

Variability of A Distribution: Measures of Spread



Measures of Spread: Range

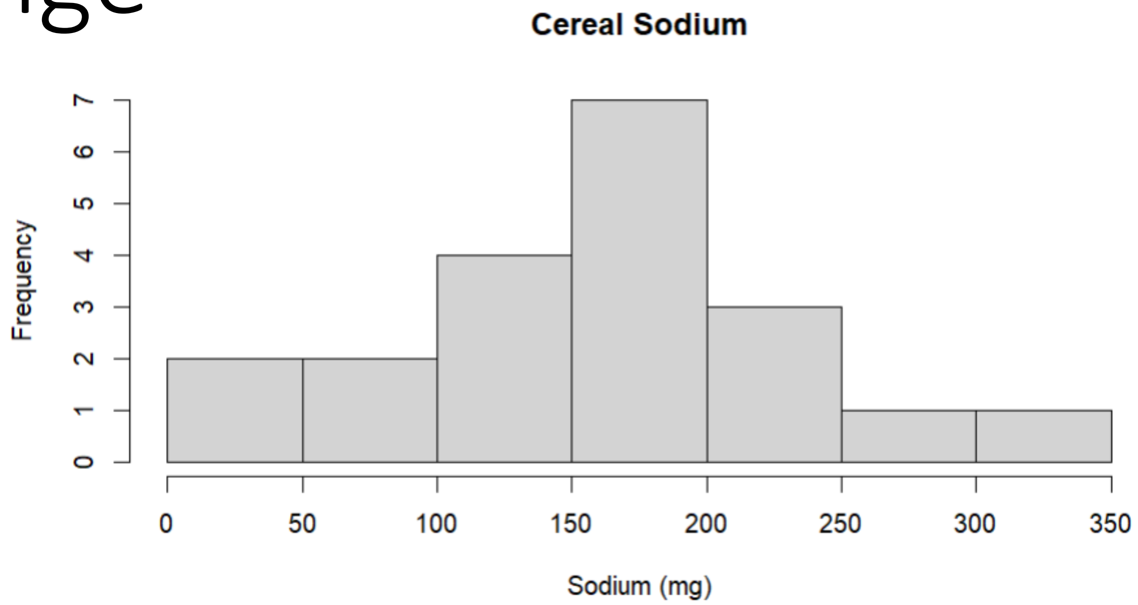
- The **range** is a measure of the distance between the smallest and largest values in the data

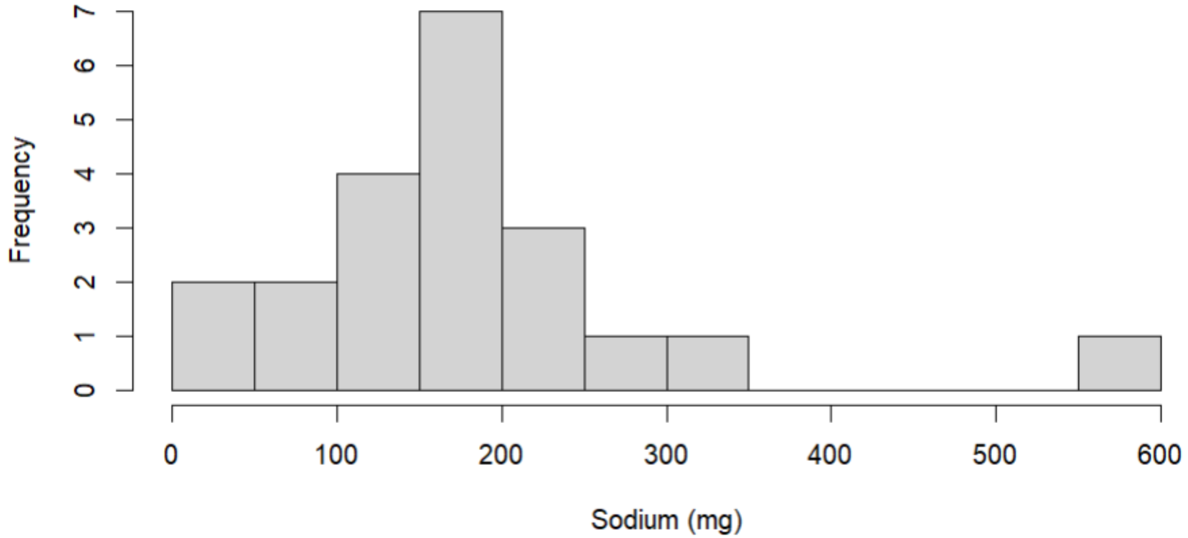
  The range can be computed with only two data points the minimum value and maximum value

- If the range of a set of data is large, then the data vary more

- The range is severely affected by the presence of outliers

- We typically do not use the range to measure variability





Measures of Spread: Deviation

- A better measure of variability that uses *all* the data is based on **deviations**

- **deviations** are the distances of each value from the mean of the data:

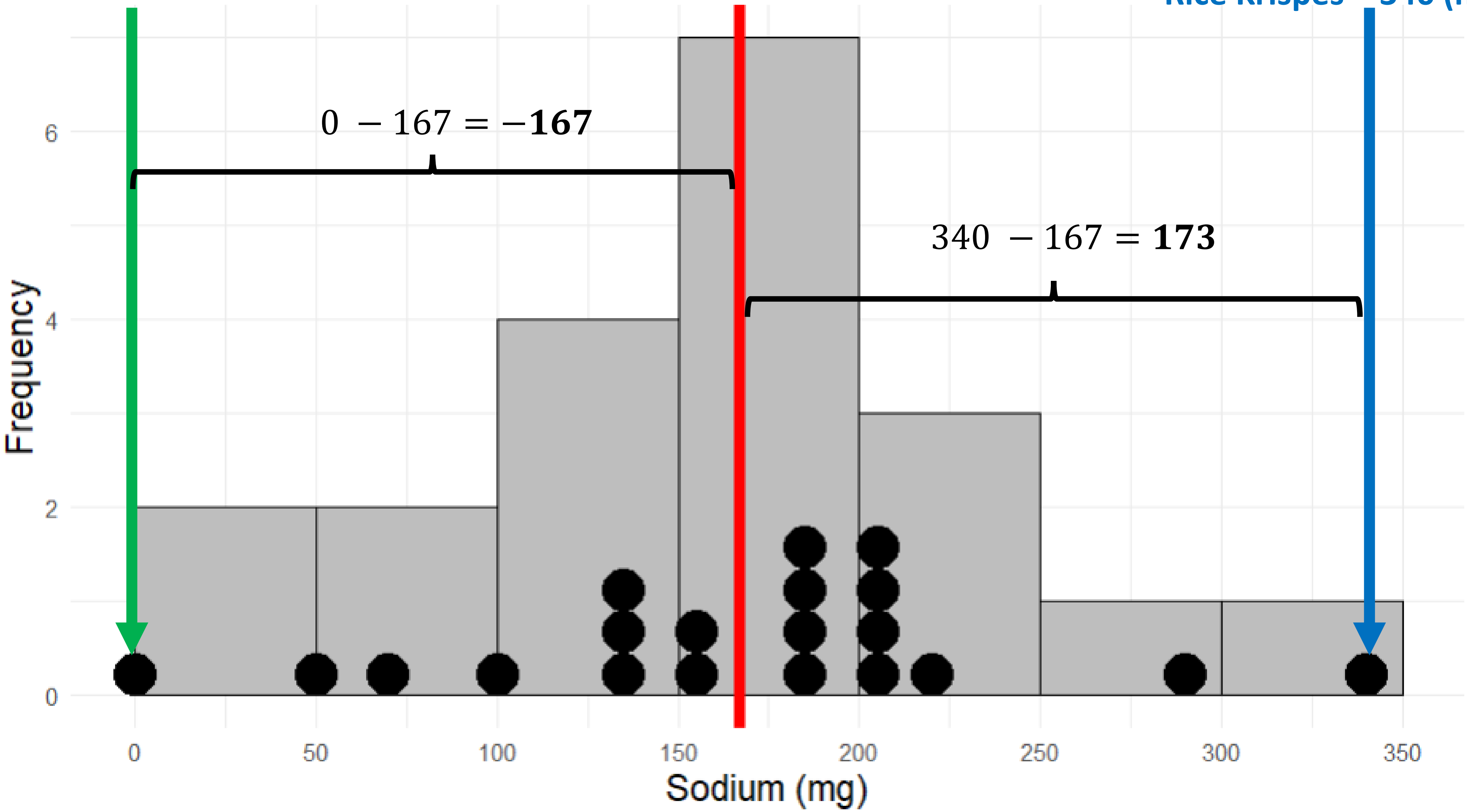
  Deviation of an observation $x_i = (x_i - \bar{x})$

- Every observation will have a deviation from the mean

Frosted Mini Wheats = 0 (mg)

Mean = 167 (mg)

Rice Krispes = 340 (mg)



$0 - 167 = \mathbf{-167}$

$340 - 167 = \mathbf{173}$

Measures of Spread: Variance

- The sum of all deviations is zero. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

- We typically use either the **squared deviations** or their **absolute value**

  Squared deviation of an observation $x_i = (x_i - \bar{x})^2$

- The **Variance** of a distribution is the <u>average</u> squared deviation from the mean

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The sum $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is called the sum of squares

Measures of Spread: Standard Deviation

- Since the variance uses the squared deviation, we usually take its square root called the **standard deviation**

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The standard deviation represents (roughly) the average distance of an observation from the mean

- The greater s is the greater the variability in the data is

- We denote the population parameter for the variance and standard deviation using σ for s and σ^2 for s^2

Why divide by $n - 1$?

- We divide by $n - 1$ because we have only $n - 1$ pieces of independent information for s^2

- Since the sum of the deviations must add to zero, then if we know the first $n - 1$ deviations we can always figure out the last one

- Ex.) suppose we have two data points and the deviation of the first data point is $x - \bar{x} = -5$
  - Then the deviation of the second data point has to be 5 for the sum of deviations to be zero.

Try it out: Computing s and s^2

- Roll a six-sided die $n = 10$ times and record the number rolled each time
- Data = 1,2,3,3,4,4,4,5,6,6
- Mean = 3.8