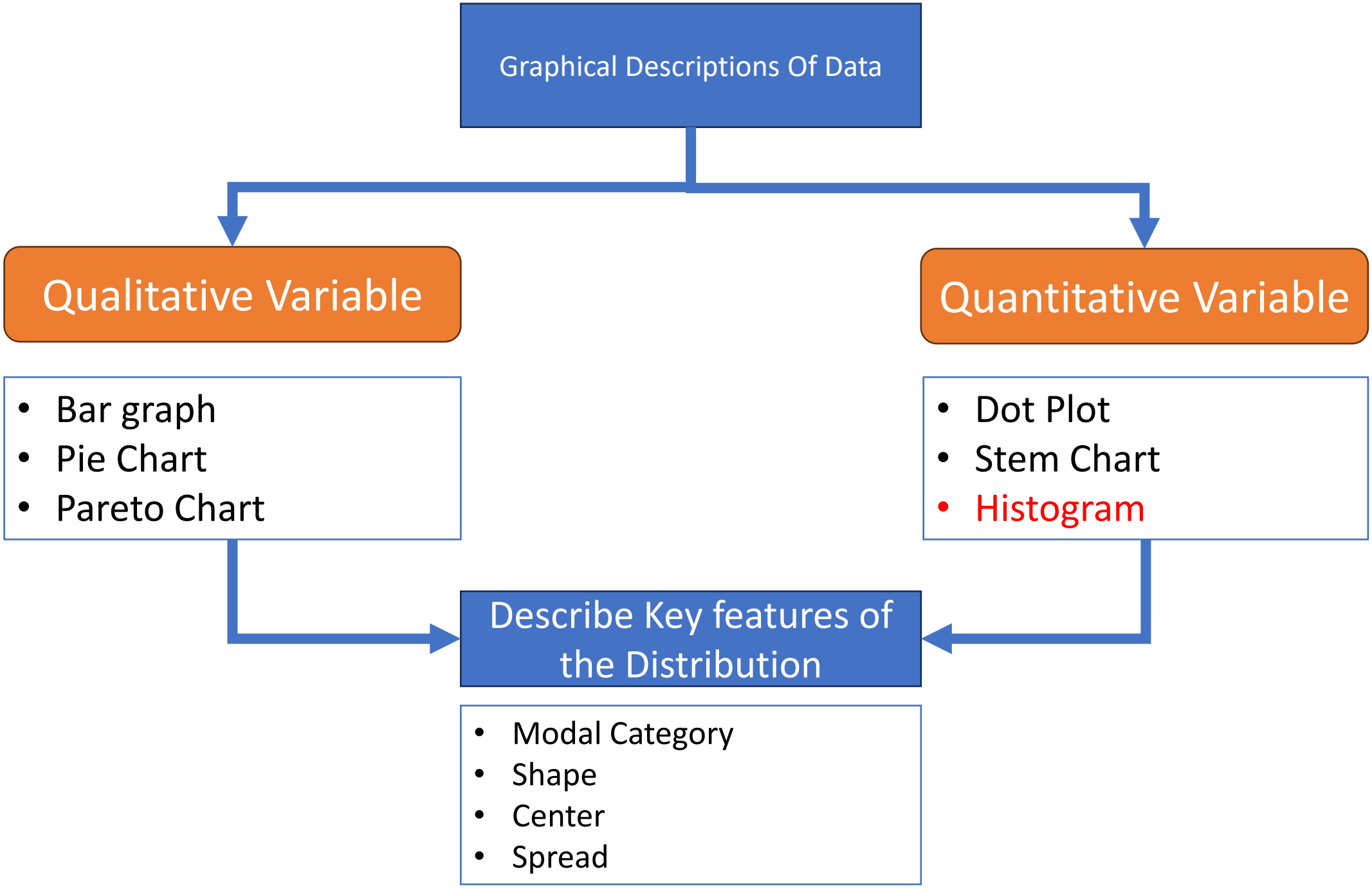
Lecture 3
Describing and Visualizing Distributions Continued

Review

- A natural first step of statistical description is to look graphical summaries of the observations for our variables

- A **distribution** of a variable gives (a) the values that occur and (b) how often each value occurs

- A **frequency table** is a tabular descriptions of the distribution of a variable – it can be applied to either quantitative or qualitative variables


Graphical Descriptions Of Data
Qualitative Variable
Quantitative Variable
• Bar graph
• Pie Chart
• Pareto Chart
• Dot Plot
• Stem Chart
• Histogram
Describe Key features of the Distribution
• Modal Category
• Shape
• Center
• Spread

Visualizing Distributions: Quantitative Variables

Stem and leaf plots and **dot plots** are unwieldy for large n

Histogram – uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable

Steps to construct a histogram

1. **Divide the range of the data into intervals of equal width**
   - If the data are categorical, we simply plot a bar for each distinct value
2. **Compute the frequency of each interval (i.e construct the frequency table)**
3. **Label the x-axis with the values or endpoints of each interval.**
4. **Draw a bar over each value or interval with height equal to its frequency or relative frequency**

How to choose the number of Bins?

- How to choose the best number of bins is not a straightforward question and there is a lot of literature on the subject

- We can construct our histogram using a specific binwidth w or under a set number of bins k
- $w = \frac{\max x - \min x}{k}$ or $k = \frac{\max x - \min x}{w}$

- Square root method: $k = \text{round}(\sqrt{n})$ (A fairly safe and basic rule of thumb)

- Sturges Rule[1]: $k = \text{round}(\log_2 n) + 1$ (not great for $n < 30$)

- Rices Rule[2]: $k = 2\sqrt[3]{n}$

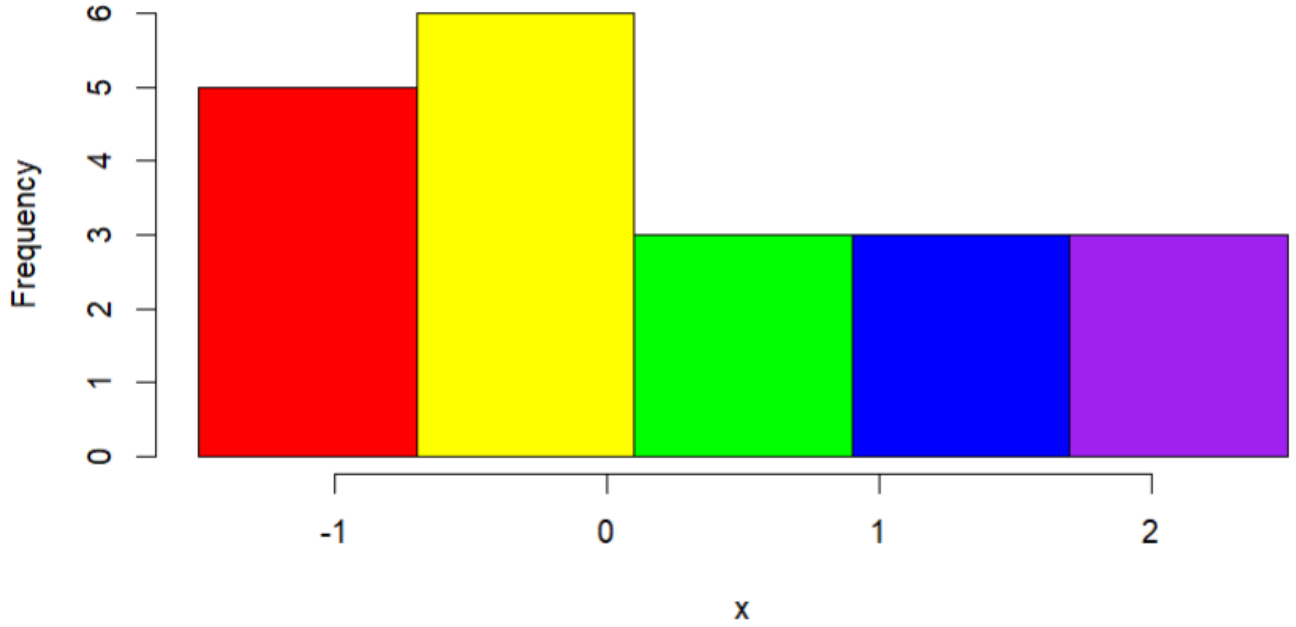
[1] Sturges, Herbert A. "The choice of a class interval." Journal of the american statistical association 21.153 (1926): 65-66.

[2] Lane, David. *Online statistics education: A multimedia course of study*. Association for the Advancement of Computing in Education (AACE), 2003. – Chapter 2 “Graphing Distributions

Try it out: Histogram

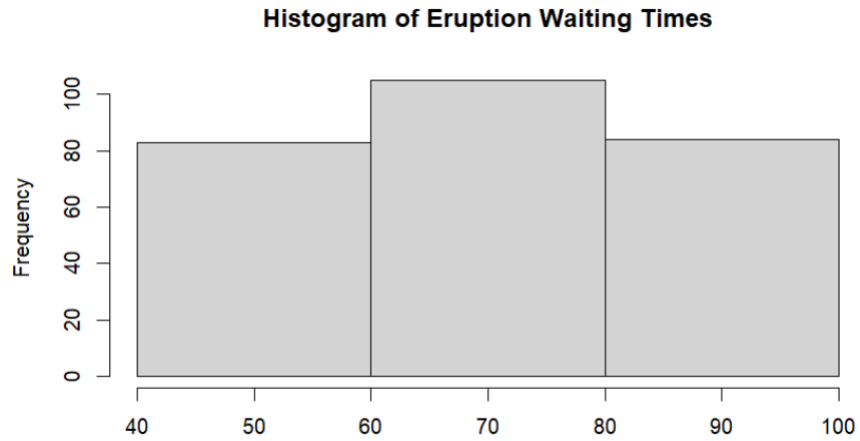
Consider the following $n = 20$ observations of a continuous variable

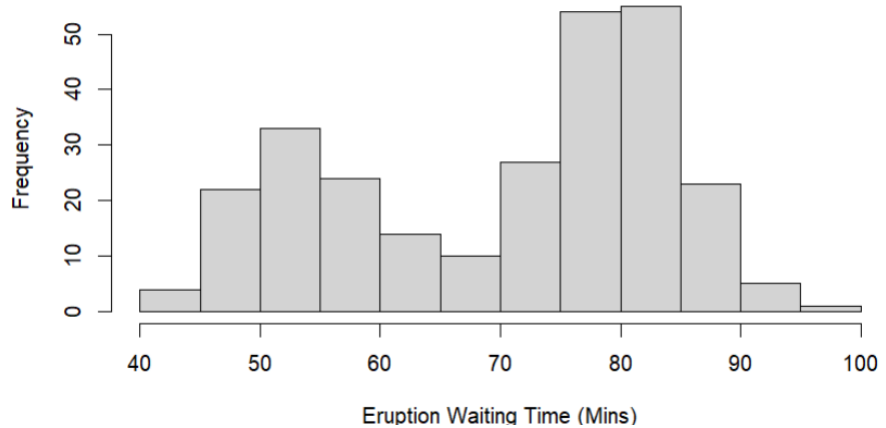
	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
x =	-1.5, -1.2, -1.0, -0.8, -0.7,	-0.6, -0.1, -0.1, 0.1, 0.1, 0.1,	0.6, 0.6, 0.8,	1.1, 1.2, 1.3,	1.8, 1.9, 2.4

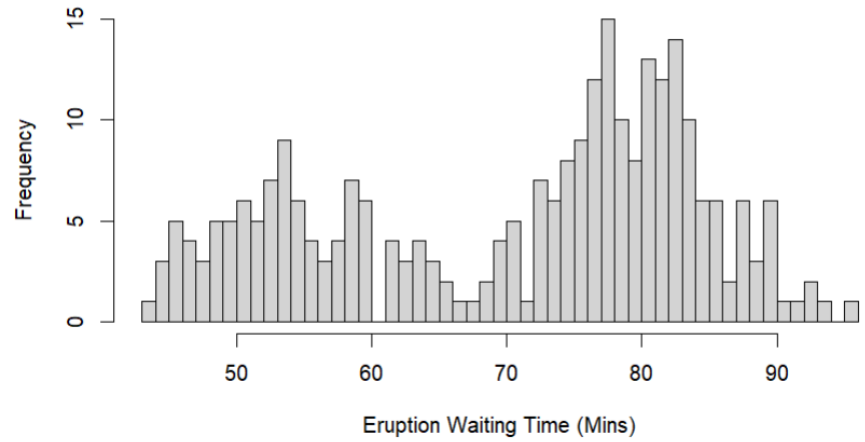


Some tips

- If too few intervals are used, then the graph will be too crude
- If too many intervals are used, graph will contain many short bars and gaps.
  Usually between 5 - 15 intervals are enough.
- Most plotting software will automatically choose the number of bins.
- **ALWAYS** plot the histogram to get an idea about the shape of the distribution of a quantitative variable
- Is the number of observations is small (say n < 50) then it's a good idea to supplement a histogram with a dot plot or stem plot

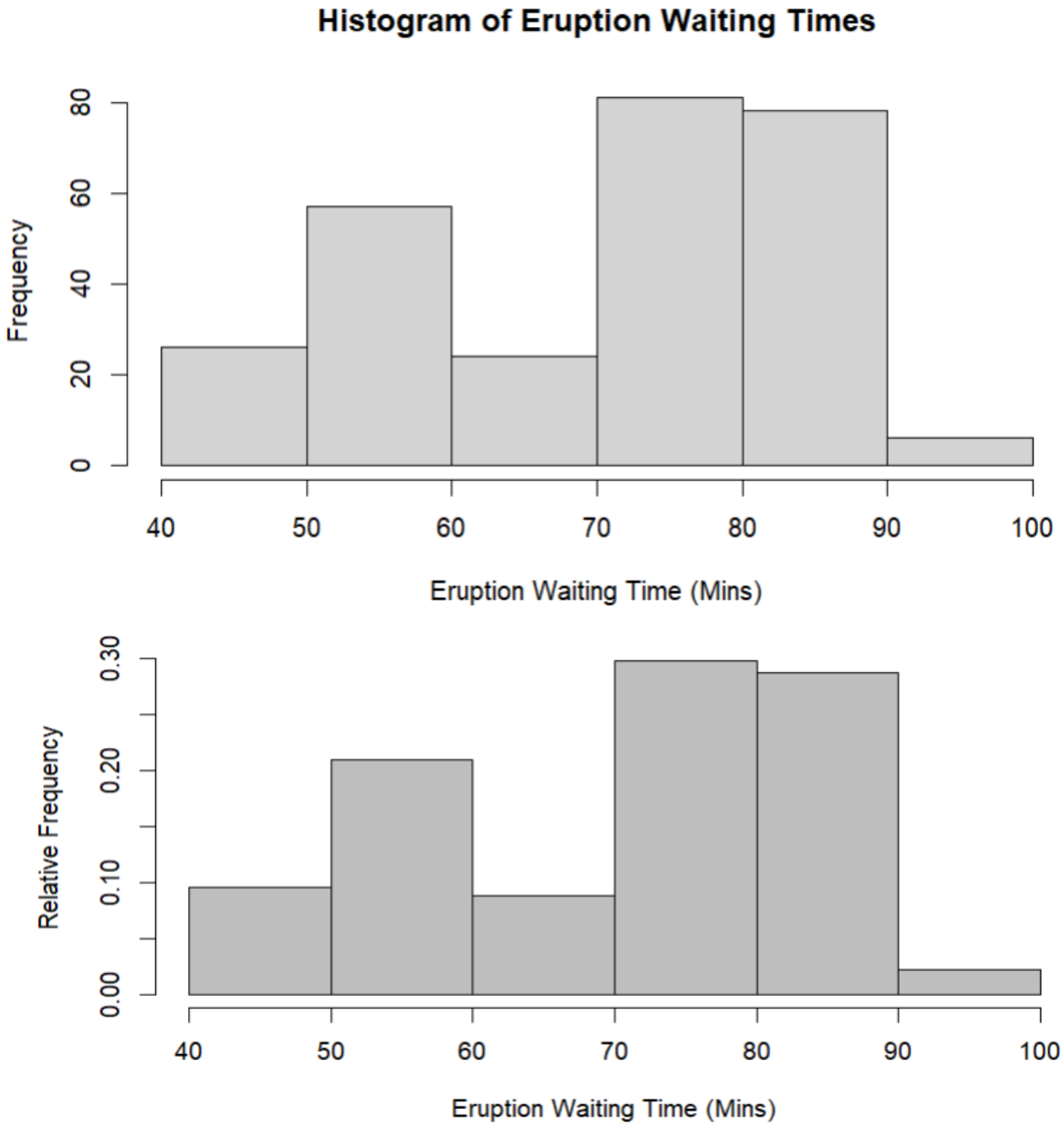




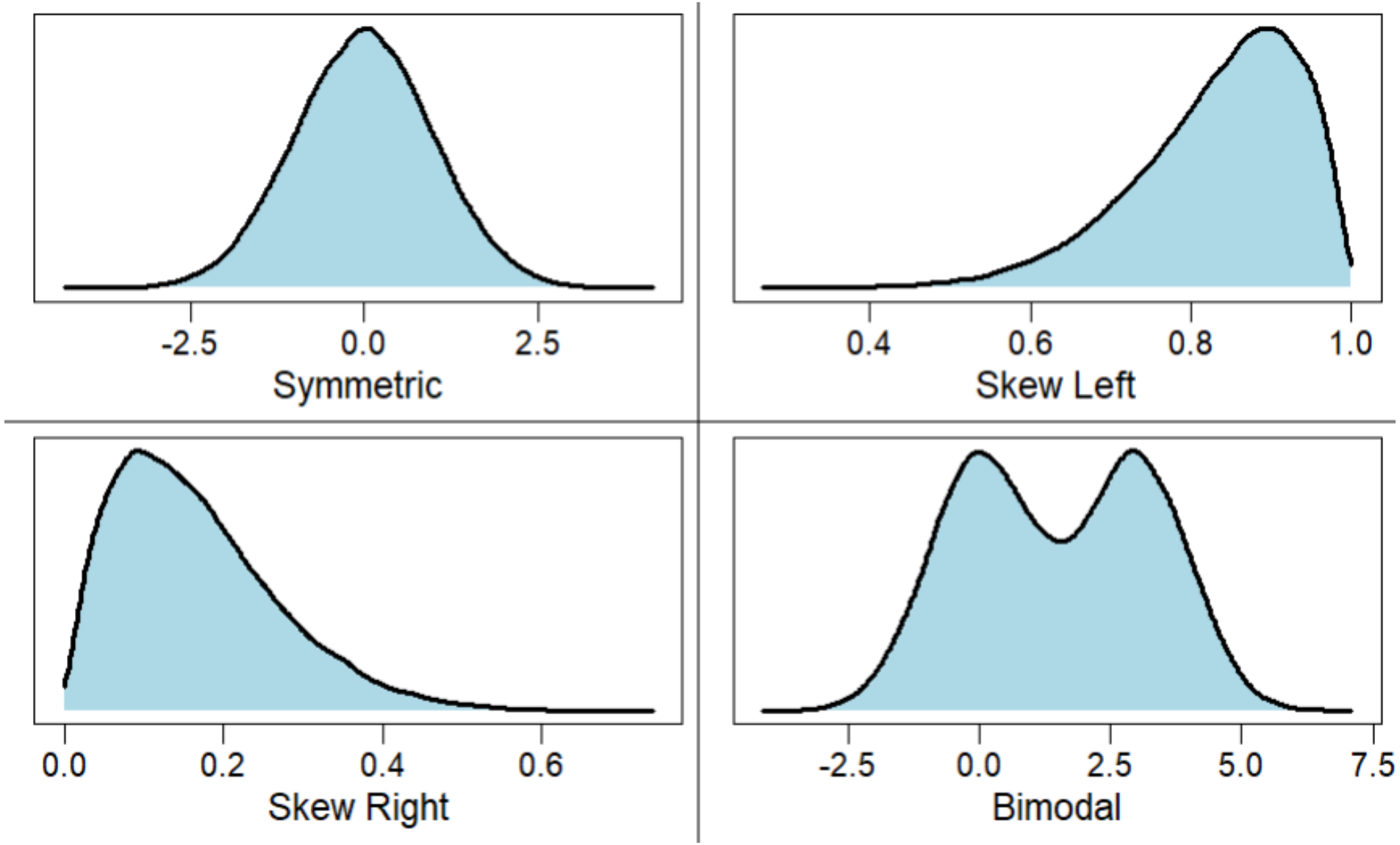


Example: Old Faithful Eruption Times

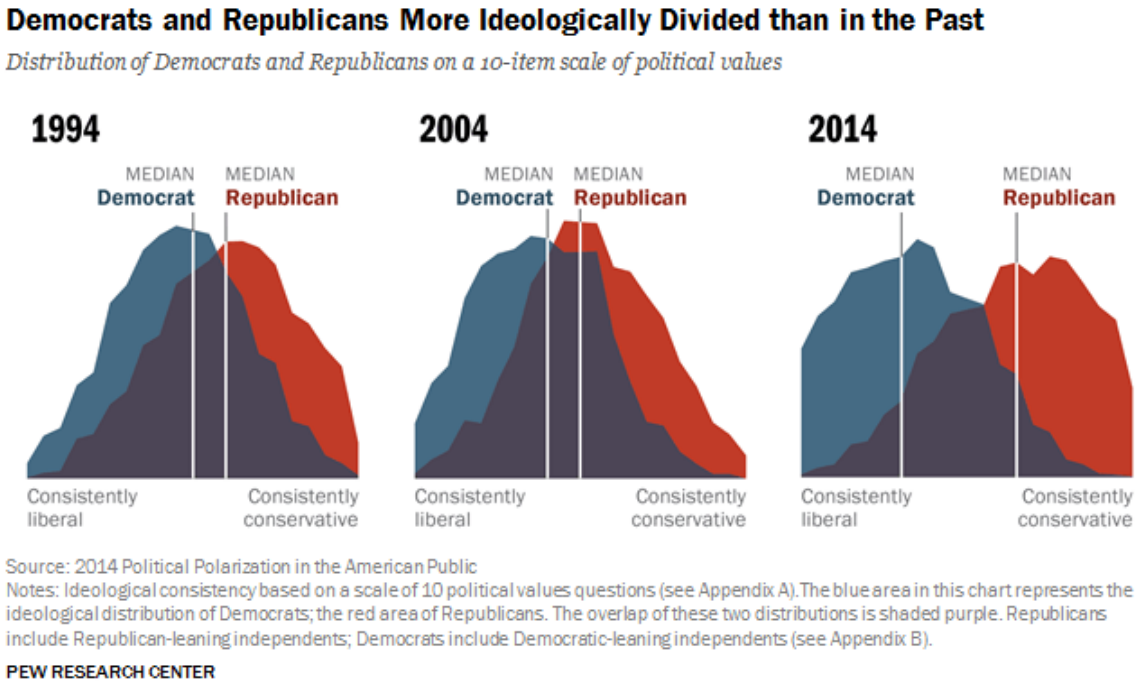
Waiting Time (Min)	Frequency	Relative Frequency	Cumulative Relative Frequency
< 50	21	0.077	0.077
50 - 60	56	0.206	0.283
60 - 70	26	0.096	0.379
70 - 80	77	0.283	0.662
80 - 90	80	0.294	0.956
> 90	12	0.044	1

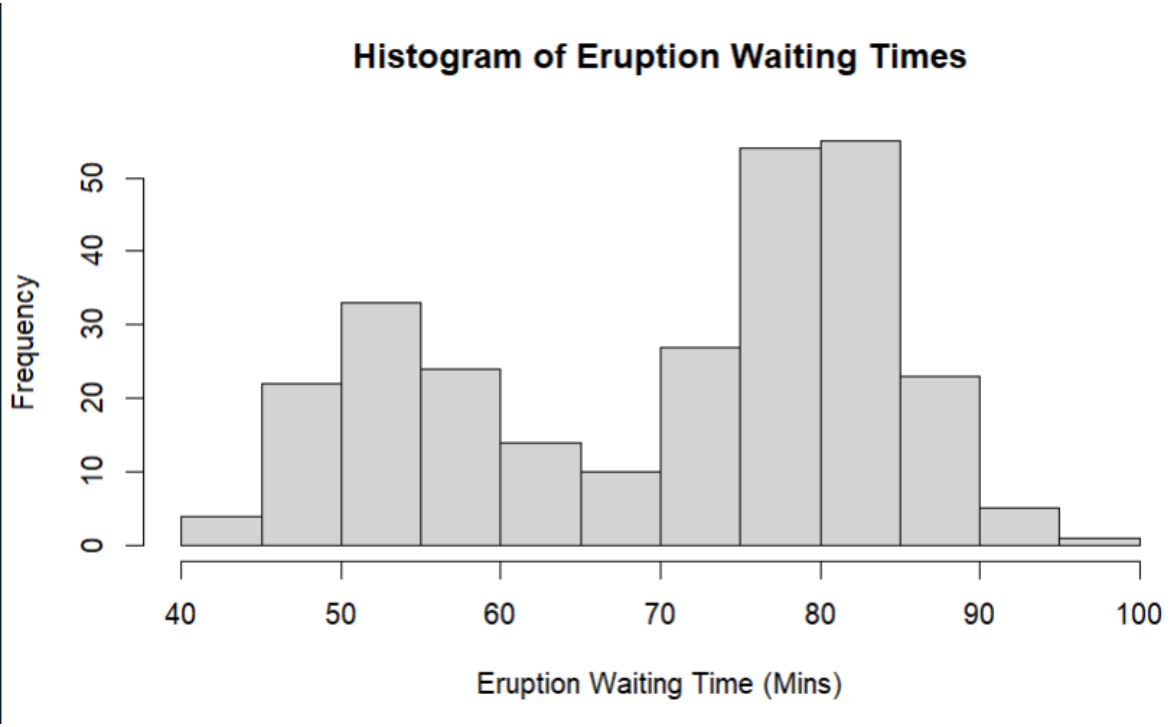


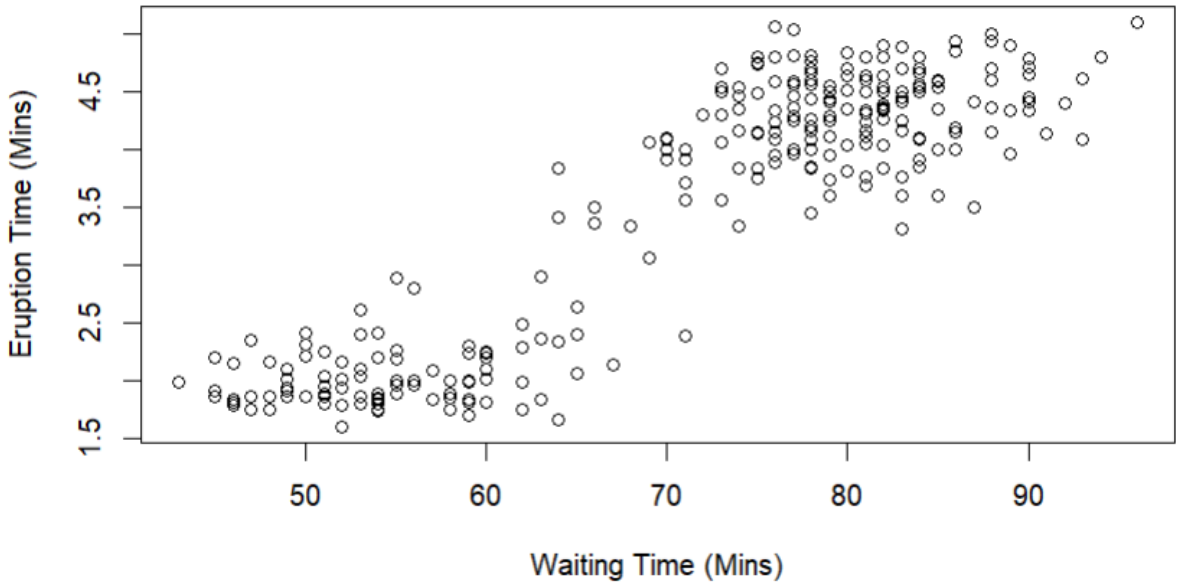
Shape of a distribution



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







Measures of Central Tendency

- The (arithmetic) **mean** is the average of a set of observations
  it measures the center of mass of a distribution (the balancing point)

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

- 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

- 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

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

Ex.)

Data = 1,1,4,5,6

Median = 4

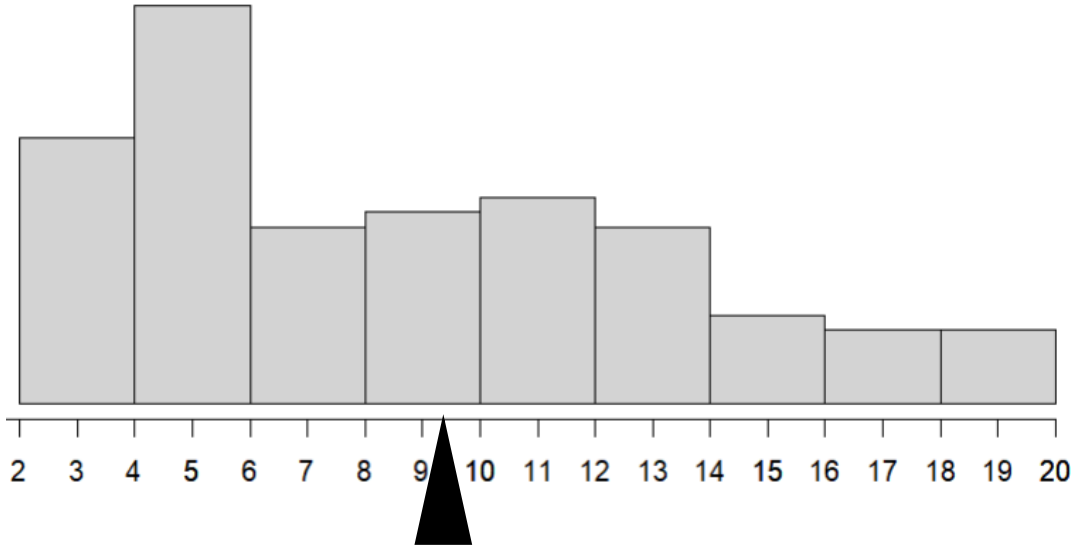
Mode = 1

Data = 1,1,4,5,6,6

Median $= \frac{4+5}{2} = 4.5$

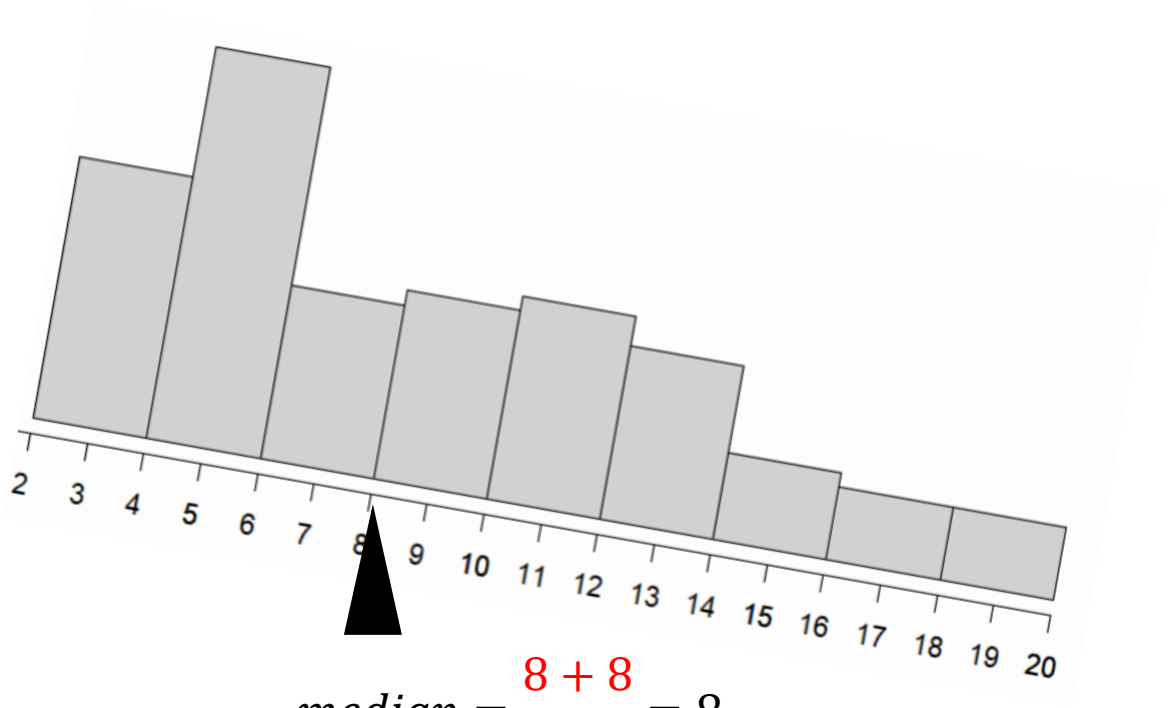
Mode 1, 6

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



$\bar{x} = 9.2$

The mean is the center of gravity



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

The median is the middle value

Measure of Central Tendency

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

Ex.)

Data = 1,1,4,5,6

Mode = 1

Data = 1,1,4,5,6,6

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

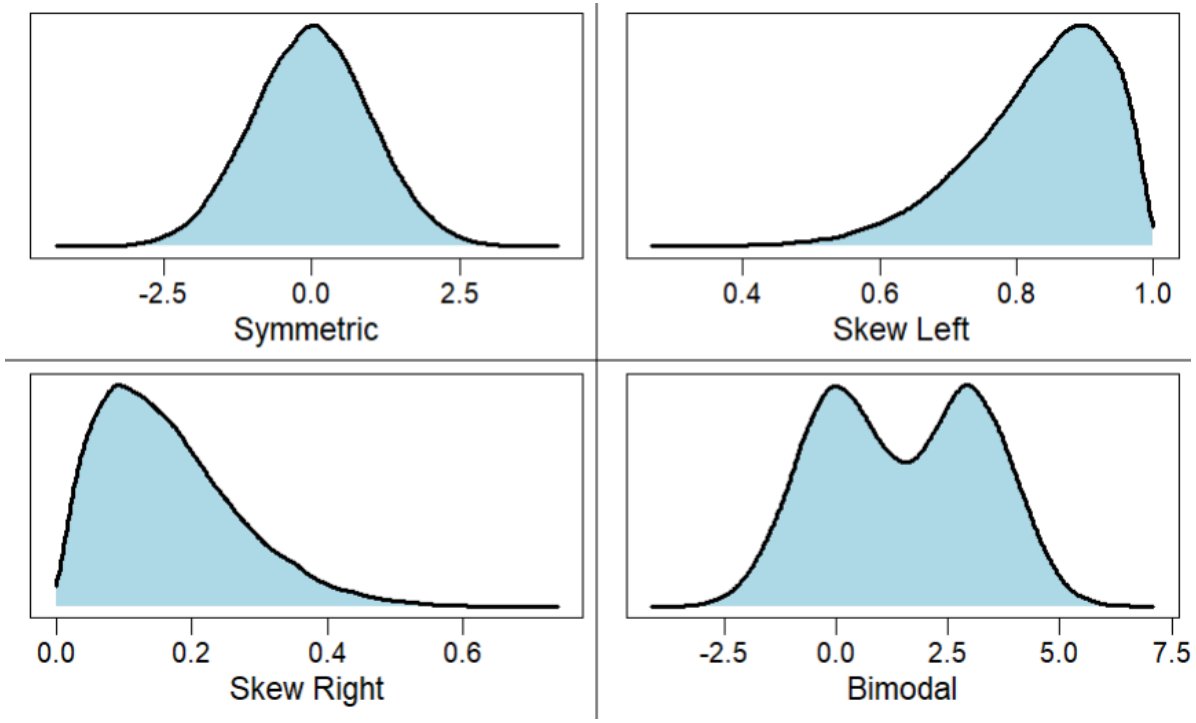
Compute the **mean** using all 3 equations:

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

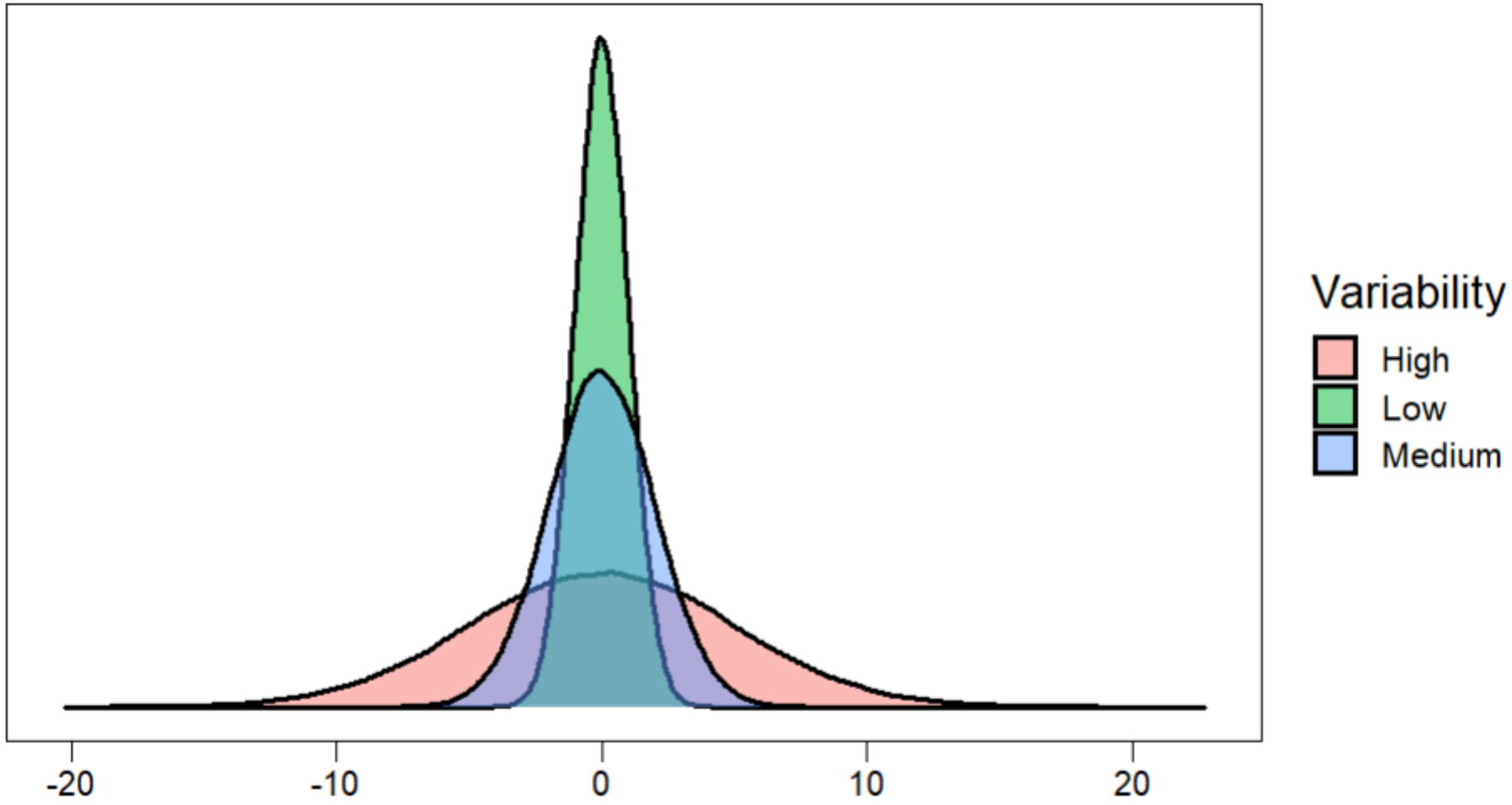
  The **median** is 1.8

- 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

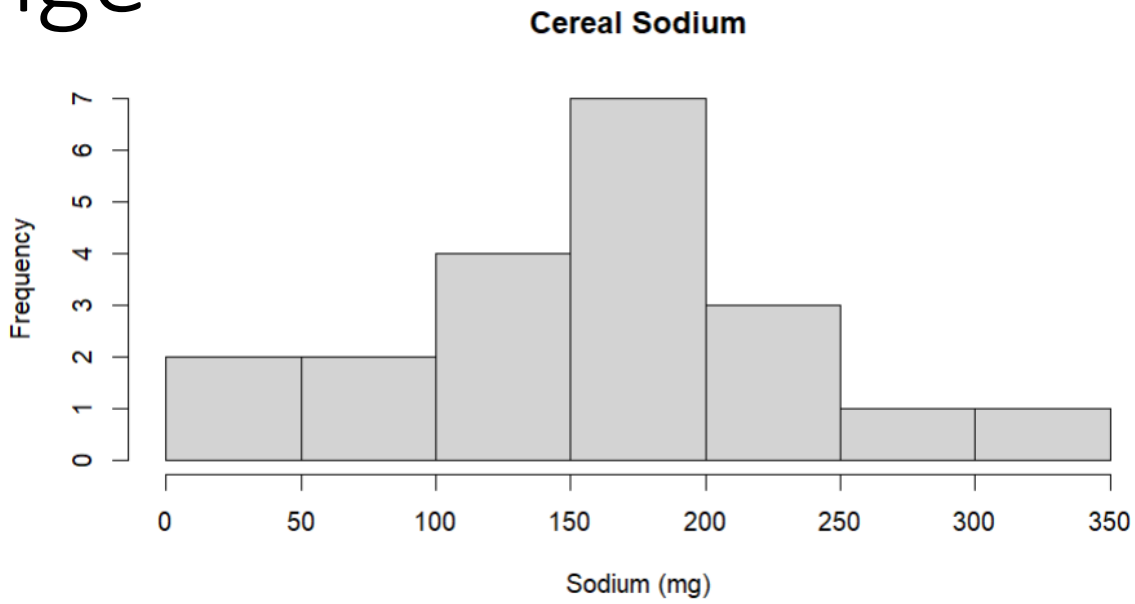
- The **range** is a measure of the distance between the smallest and largest values in the data

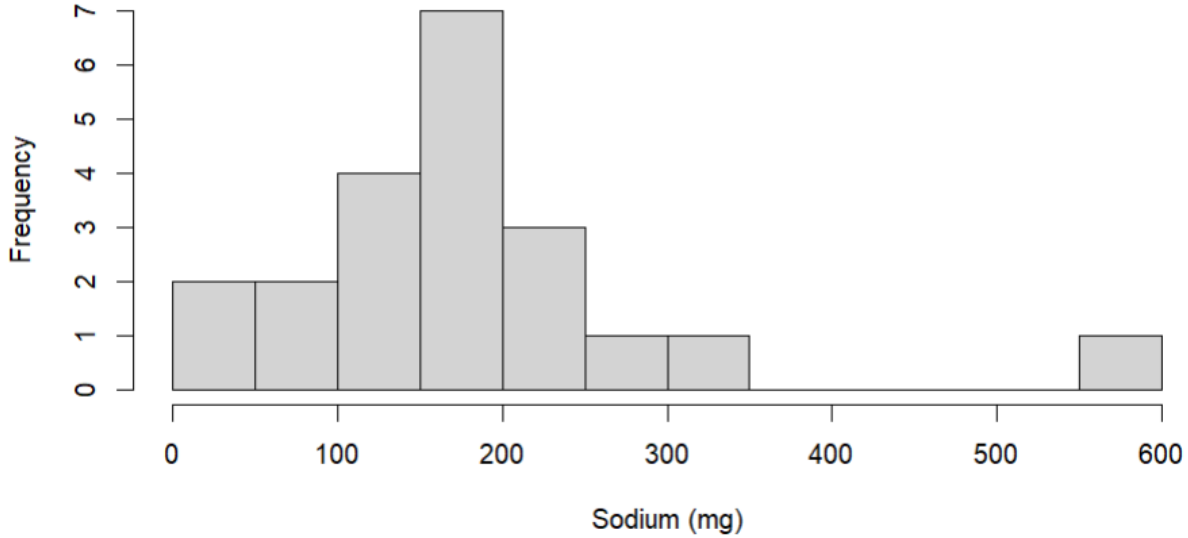
  The range can be computed with only two data points the minimum value and maximum value

- 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





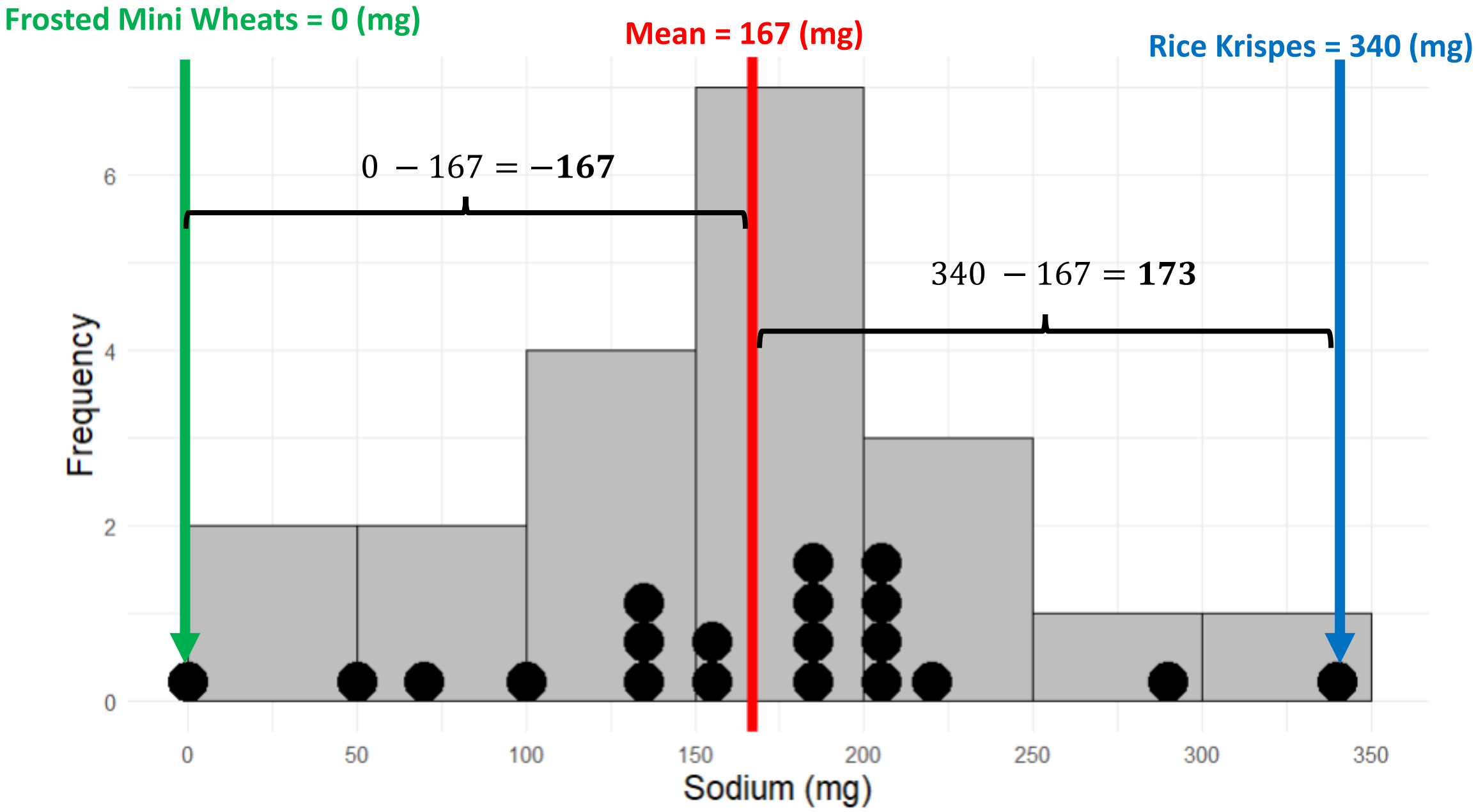
Measures of Spread: Deviation

- 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)
$0 - 167 = -167$
$340 - 167 = 173$
Frequency
0
2
4
6
0
50
100
150
200
250
300
350
Sodium (mg)

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 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