Lecture 5
Measures of location and spread

Review

Describing the shape of distribution

- How many peaks? Unimodal, bimodal, multimodal
- Is the distribution symmetric or is it skewed?
- Are there outliers in the data?

Two numbers to describe the center of distribution:

- **Mean** the average value of a distribution – it is NOT resistant to outliers
- **Median** the middle value of a distribution – it is resistant to outliers

The **mode** is a measure of location

- It describes the position of commonly occurring values (i.e the position of peaks in the distribution)

Measures of Location: Quartiles and Percentiles

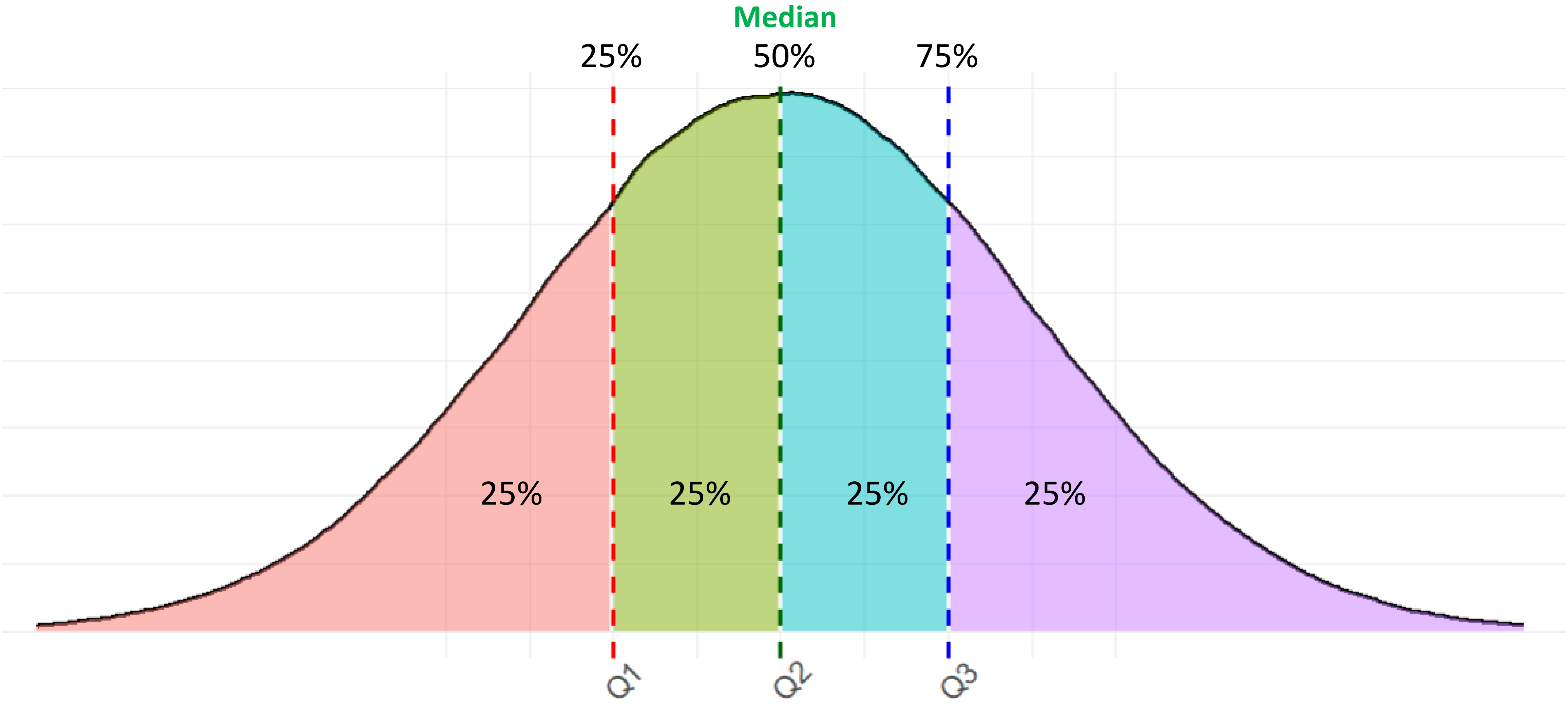
- The $\boldsymbol{p^{th}}$ **percentile** is a value such that p percent of the observations in a sample or population fall at or below that value

- Ex. The 50^{th} percentile is of any dataset is the median

- Three useful percentiles of a distribution are the **quartiles**
  1. The first **quartile Q1** is the 25^{th} percentile of the data.
  2. The second **quartile Q2** is the 50^{th} percentile or median of the data.
  3. The third **quartile Q3** is the 75^{th} percentile of the data.

- The quartiles split a distribution in four equal parts each containing 25% of the observations

Measures of Position: Quartiles and Percentiles



Measures of Position: Quartiles and Percentiles

- How to compute the quartiles:
  1. Arrange the data in increasing order
  2. Find the median and label as **Q2**
  3. Consider the lower half of the observations (excluding the median itself if n is an odd number).
  4. Mark the median for the lower half of the observations and label as **Q1**
  5. Consider the upper half of the observations (again excluding the median itself if n is odd)
  6. Mark the median for the upper half of the observations and label as **Q3**

Ex. Quartiles

- Consider the following data which come from n =20 rolls of a six-sided die

lower half middle upper half

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

Q2 = median $= \frac{2+3}{2} = 2.5$

Q1 = median lower half $= \frac{2+2}{2} = 2$

Q3 = median upper half $= \frac{4+5}{2} = 4.5$

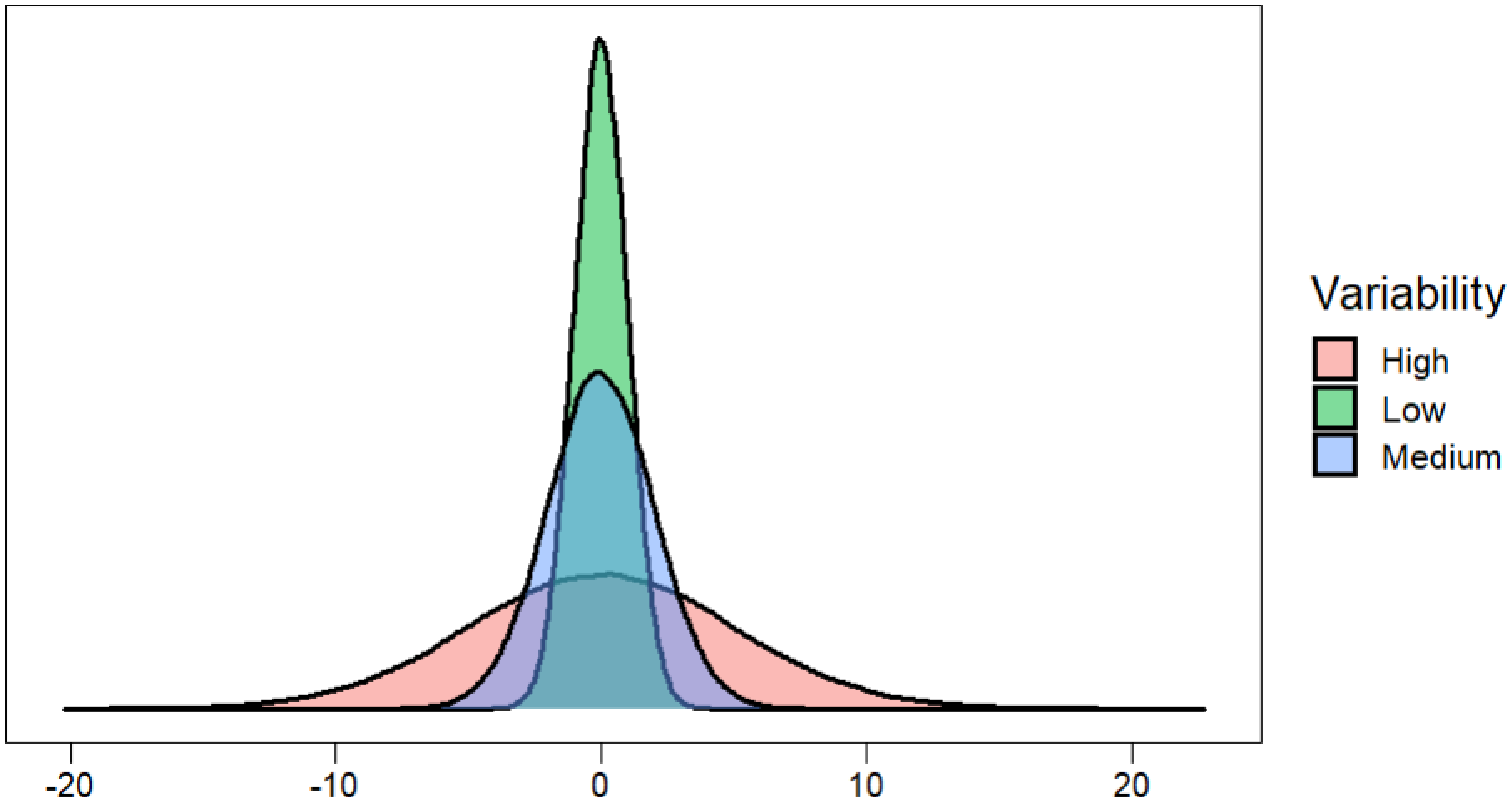
Practice

Consider the following 15 exam scores of students in a statistics course

61,61,65,65,66,68,69,73,74,75,76,78,79,90,94

Compute the 3 quartiles Q1, Q2, and Q3

Variability of A Distribution: Measures of Spread



Measures of Spread: Range

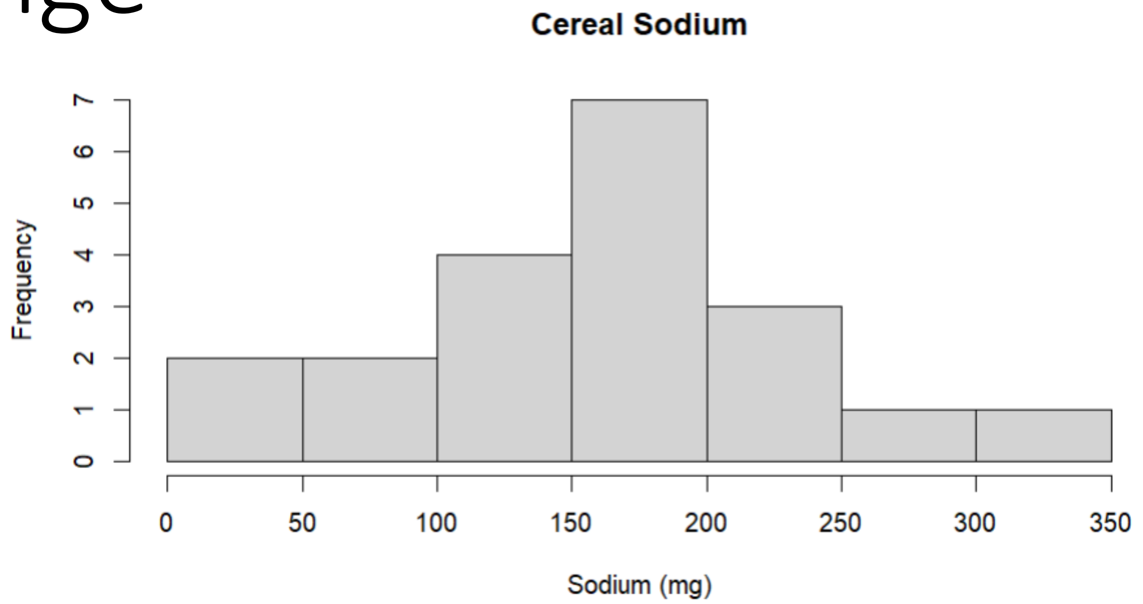
- The **range** is a measure of the distance between the smallest and largest values in the data

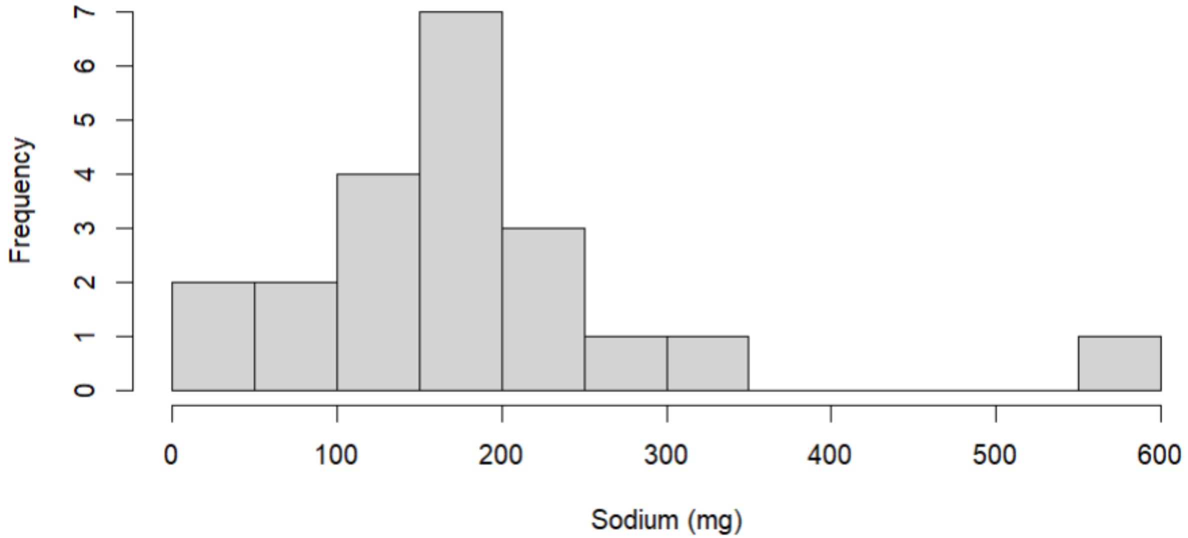
  The range can be computed with only two data points the minimum value and maximum value

- If the range of a set of data is large, then usually this indicates greater dispersion of values

- The range is <u>severely</u> affected by the presence of outliers

- We typically <u>do not</u> use the range to measure variability





Measures of Spread: Interquartile Range

- The **interquartile range (IQR)** measures the spread of the middle 50% of the observations

- It is resistant to outliers

$$\boldsymbol{IQR = Q3 - Q1}$$

- The more variability the larger the value of the **IQR**

- **IQR** is a good choice for distributions that are highly skewed!

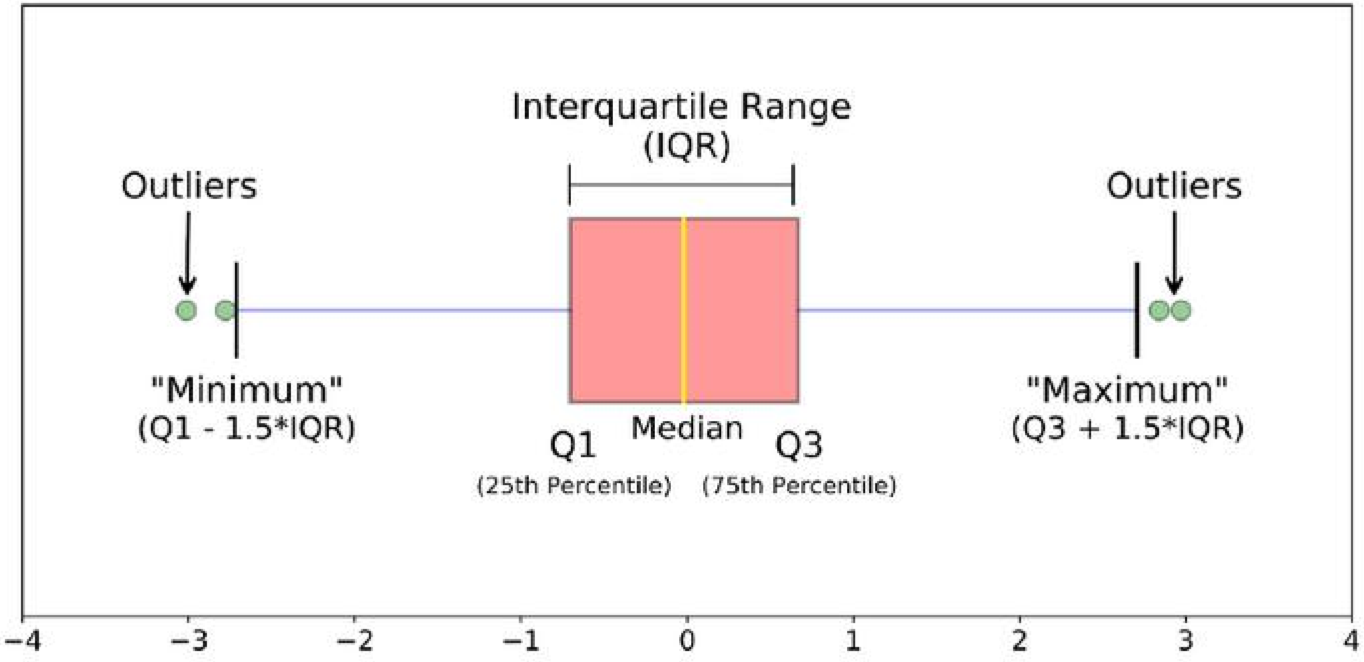
Practice: Finding quartiles and IQR

Exam Scores 61,61,65,65,66,68,69,73,74,75,76,78,79,90,94

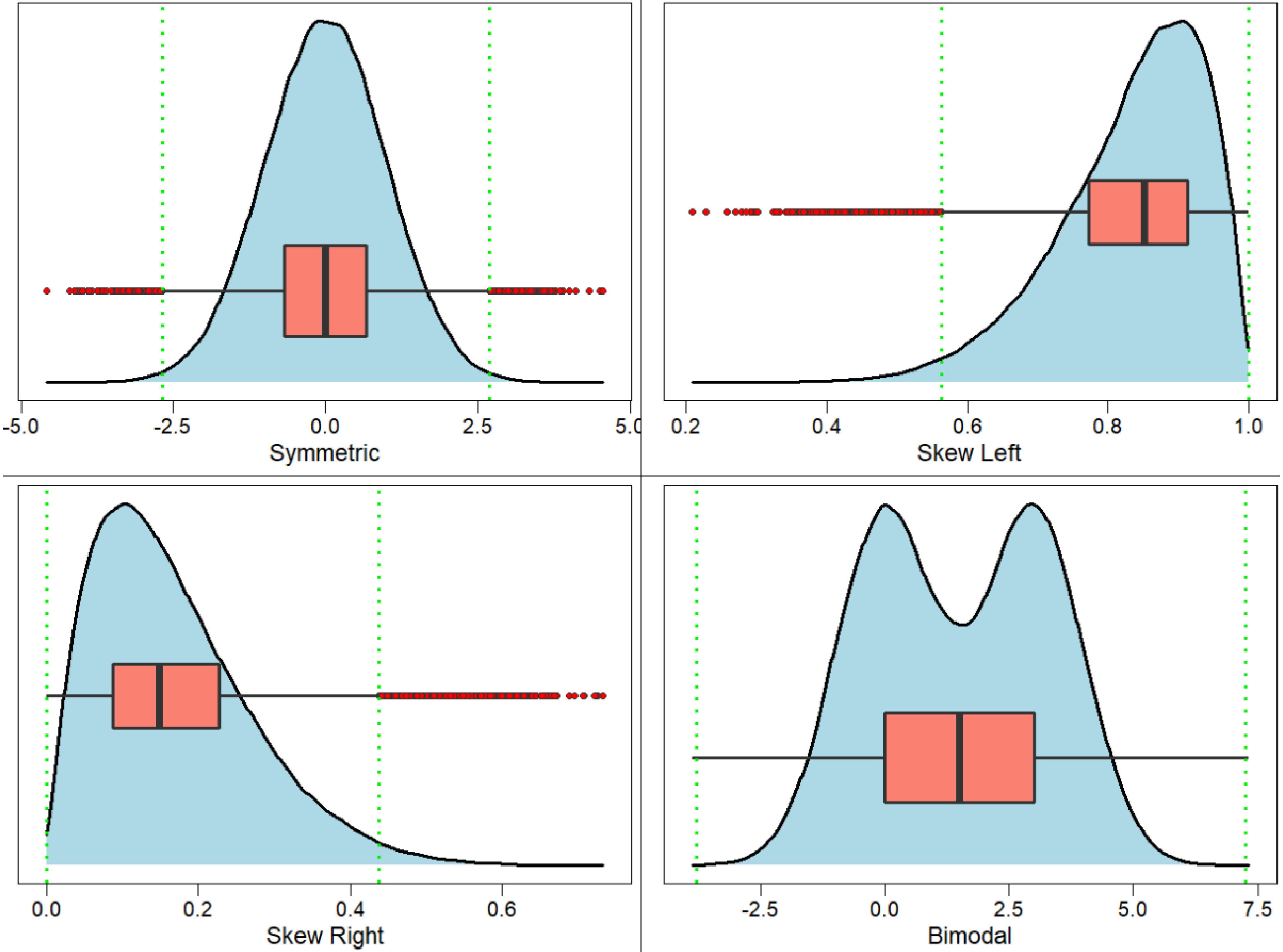
Compute the IQR

The Boxplot (Box and Whisker Plot): A five number summary

- Pros:
  - good for describing shape and location
  - Can be used to identify outliers
  - Length of whiskers indicates skew
  - Good for comparing two distributions or across categories
- Cons:
  - does not show certain features like mounds, or gaps as well as a histogram



Different parts of a boxplot



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

Ex. Construct a Boxplot

- Consider the following data which come from 20 rolls of a six-sided die

lower half | middle | upper half

- Data = $\boxed{1, 1, 1, 2, \mathbf{2}, \mathbf{2}, 2, 2, 2, \mathbf{2},}\boxed{\mathbf{3}, 3, 3, 3, \mathbf{4}, \mathbf{5}, 5, 6, 6, 6}$

Q2 = median $= \frac{2+3}{2} = 2.5$

Q1 = median lower half $= \frac{2+2}{2} = 2$

Q3 = median upper half $= \frac{4+5}{2} = 4.5$

Practice

Construct the boxplot for student exam scores

Exam Scores 61,61,65,65,66,68,69,73,74,75,76,78,79,90,94

Ex2. Construct a Boxplot

Consider the following 12 observations of a quantitative variable X

X = { -5.7, -2.6, -1.5, -1.3, -0.4, 0.2, 1.5, 2.2, 2.3, 2.6, 2.9, 10.4 }

Compute the 5 number summary and draw a boxplot

Compare the IQR with the range

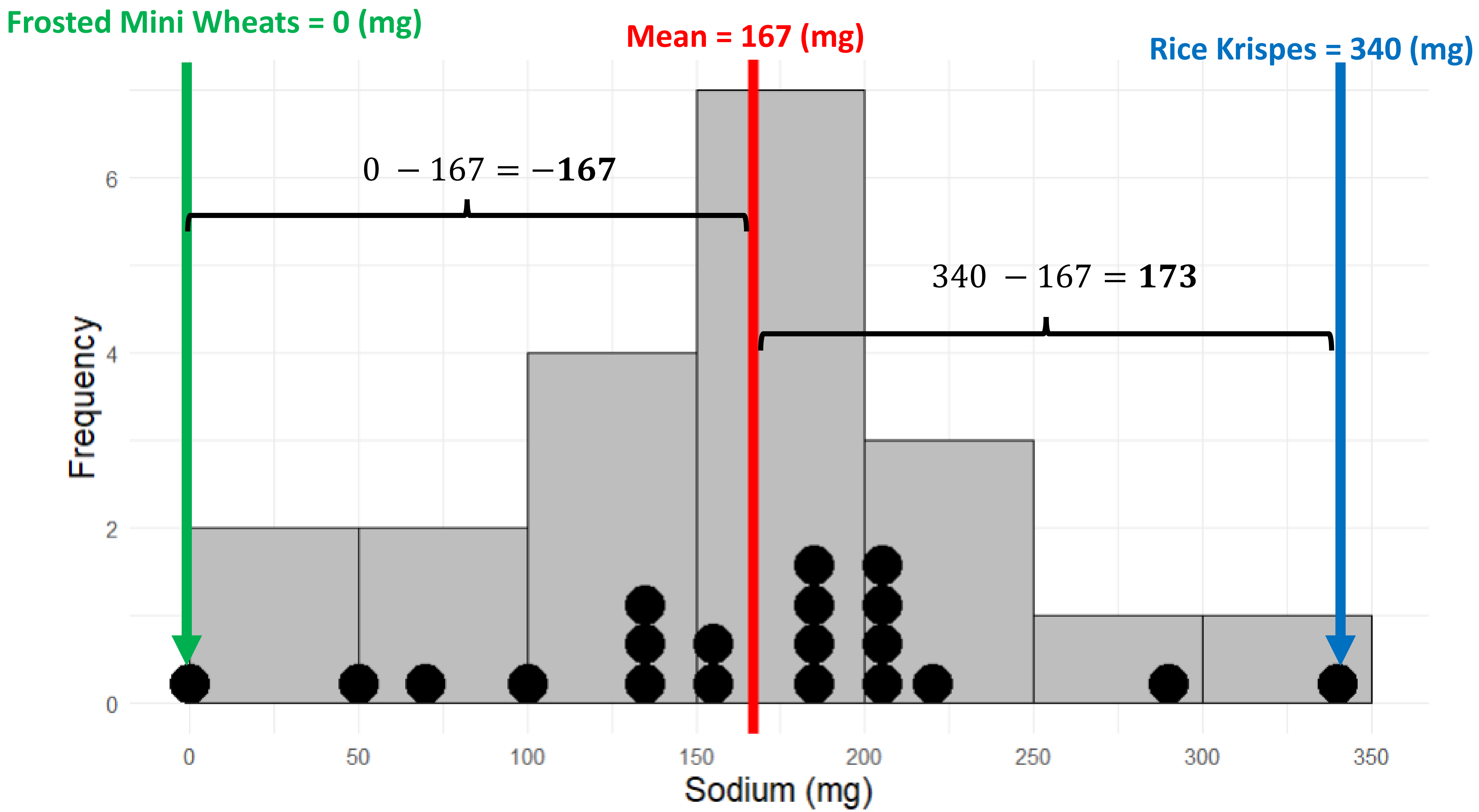
Measures of Spread: Deviation

- 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 average 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 <u>has</u> 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