Lecture 6

Variance and standard deviation
Cumulative distributions
The normal distribution

Review

- Percentiles
- Range and IQR
- The five number summary and boxplot

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

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**

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

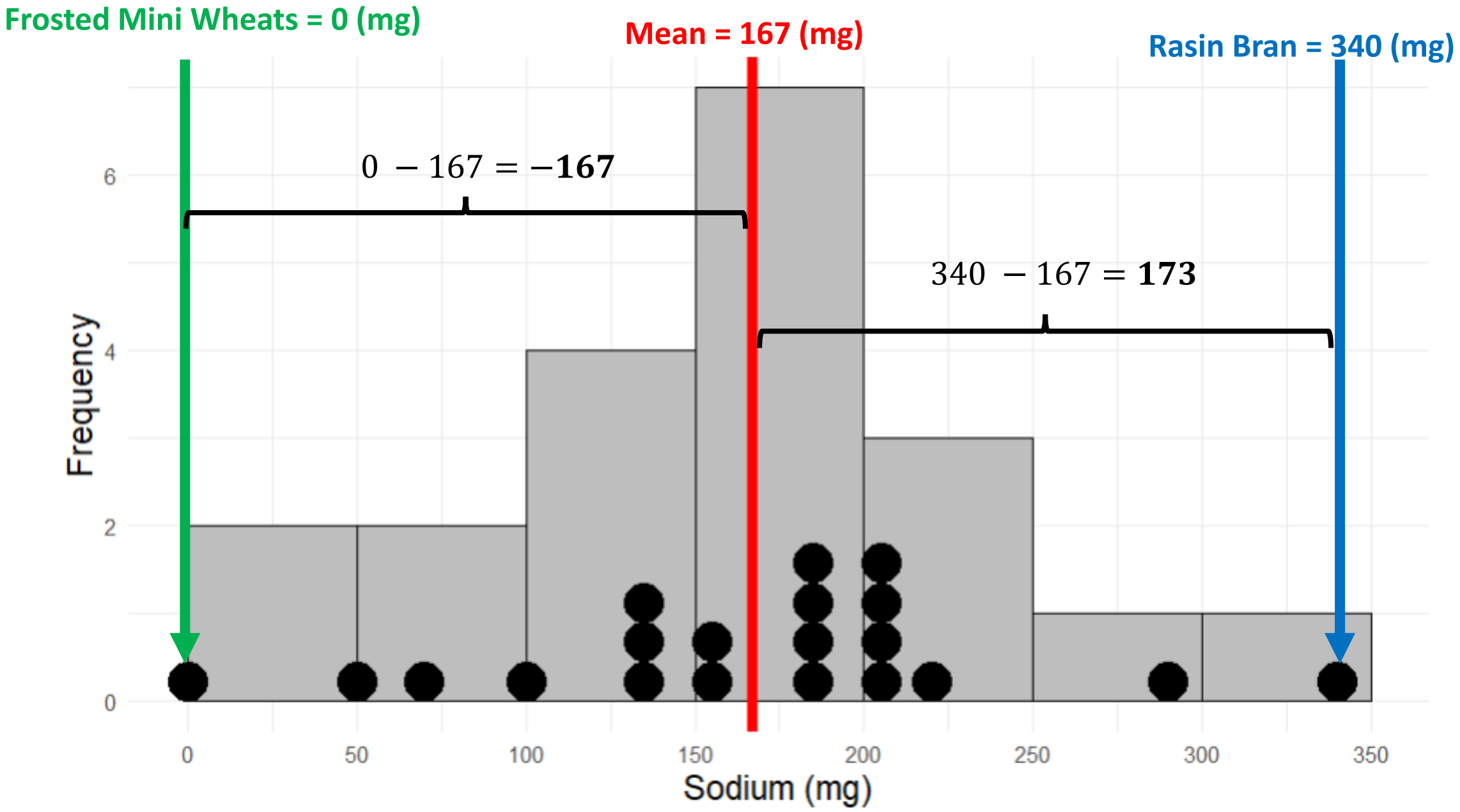
- The standard deviation represents (roughly) the average distance of an observation from the mean

- 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

Try it out: Computing s and s^2

- Data = 61,62,62,68,75
- Mean = 3.8


Frosted Mini Wheats = 0 (mg)
Mean = 167 (mg)
Rasin Bran = 340 (mg)
$0 - 167 = -167$
$340 - 167 = 173$
Frequency
0
2
4
6
0
50
100
150
200
250
300
350
Sodium (mg)

Why divide by $n-1$?

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

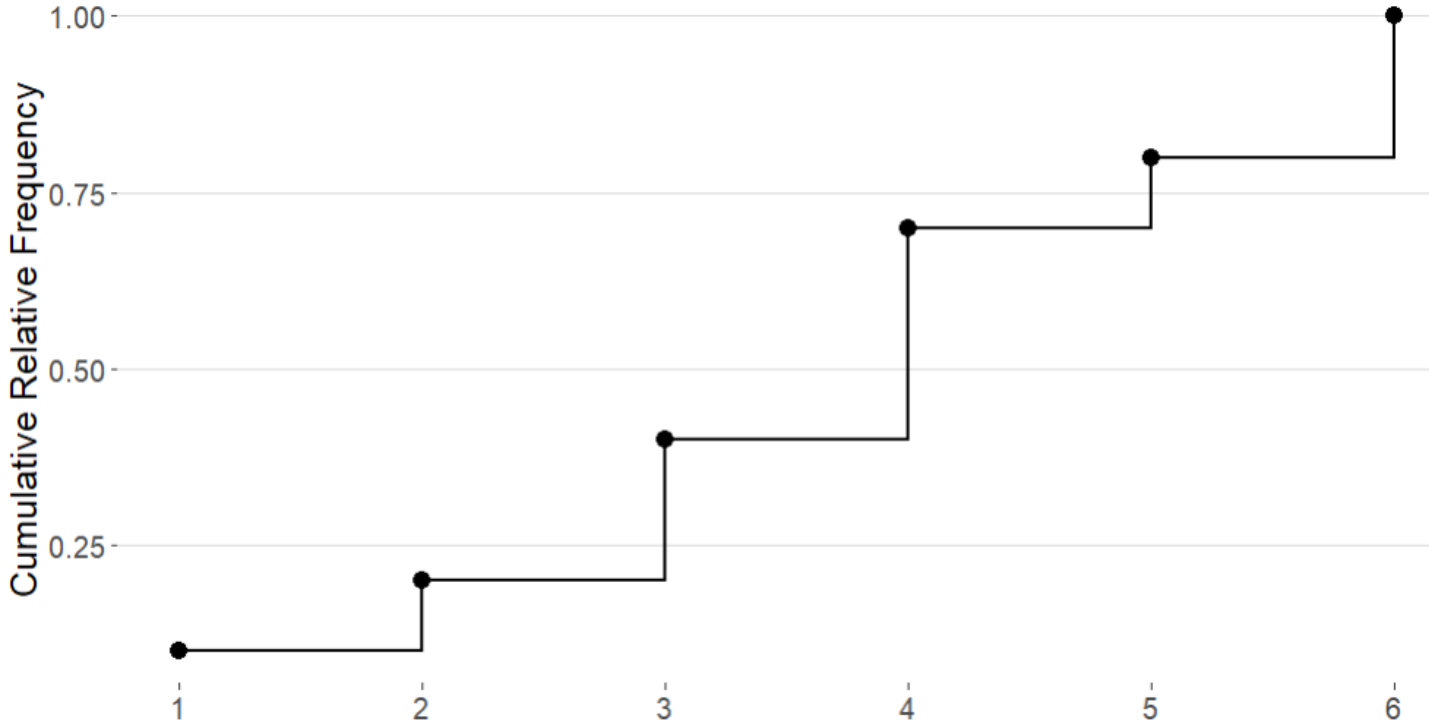
- Since the sum of the deviations must add to zero, then if we know the first $n-1$ deviations we can always figure out the last one

- 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.

Cumulative Distributions

- A **cumulative distribution** shows the relationship between the value of a variable and the **cumulative relative frequency**
- We represent the cumulative distribution using a step function
- Data = 1,2,3,3,4,4,4,5,6,6

x	$F(x)$	$RF(x)$	$CRF(x)$
1	1	0.1	0.1
2	1	0.1	0.2
3	2	0.2	0.4
4	3	0.3	0.7
5	1	0.1	0.8
6	2	0.2	1.0



Finding Percentiles from Cumulative Distributions

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

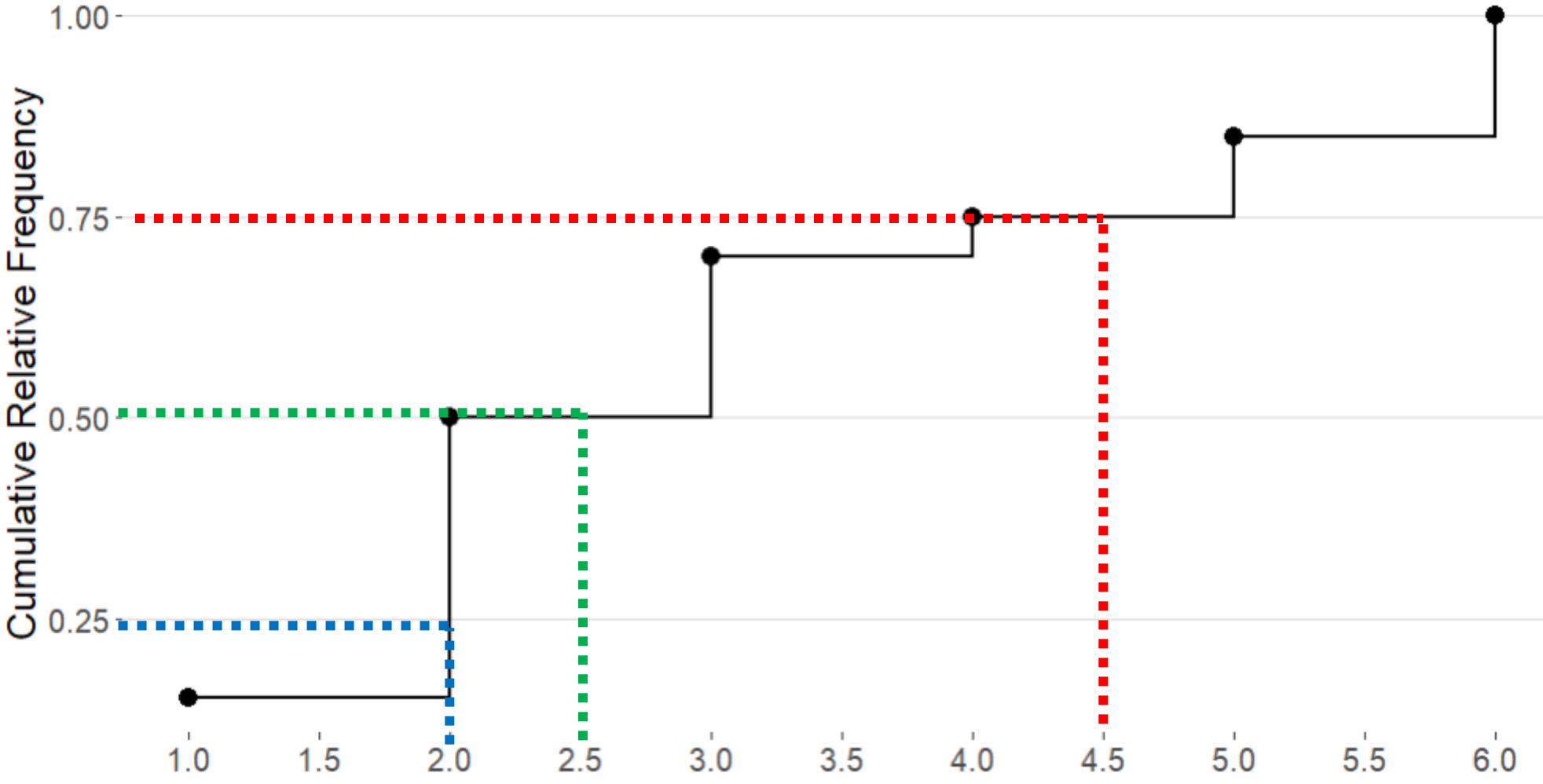
$Q2 = 2.5$

$Q1 = 2$

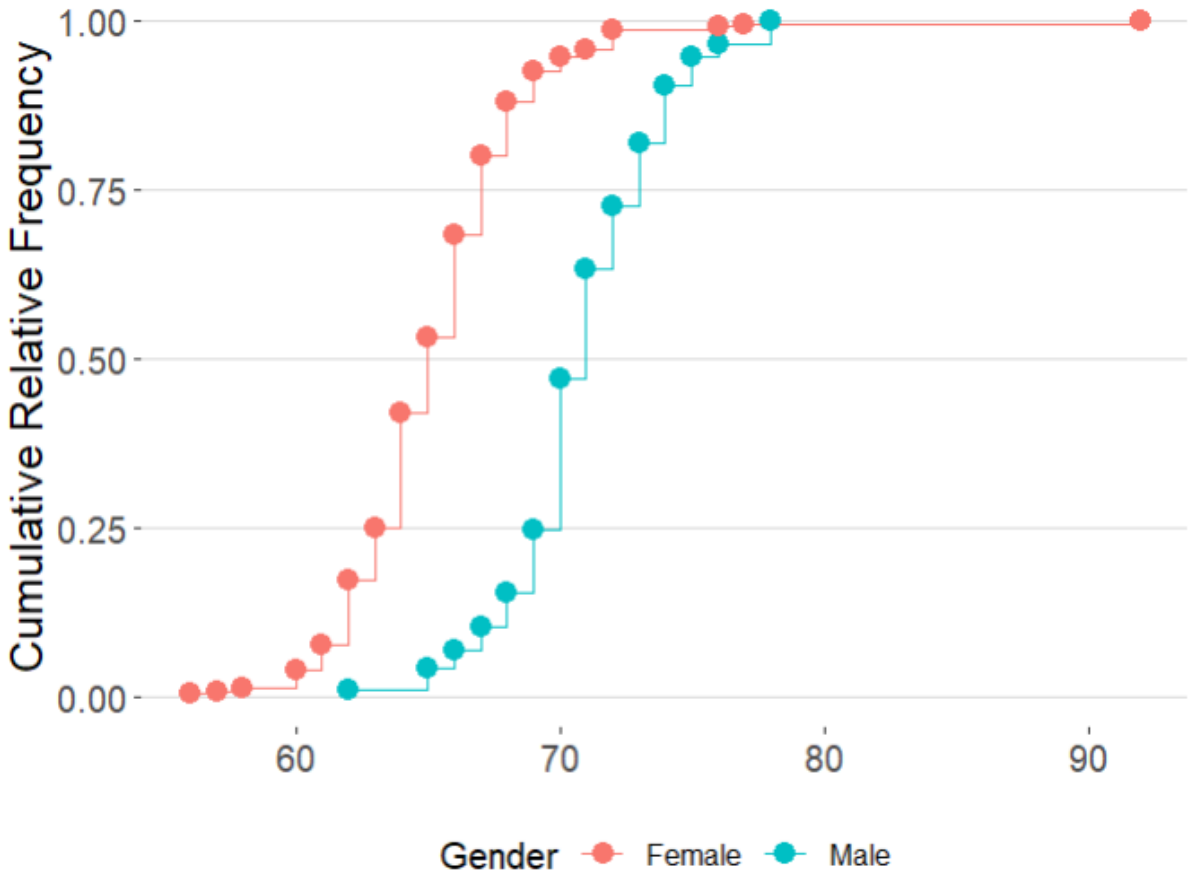
$Q3 = 4.5$

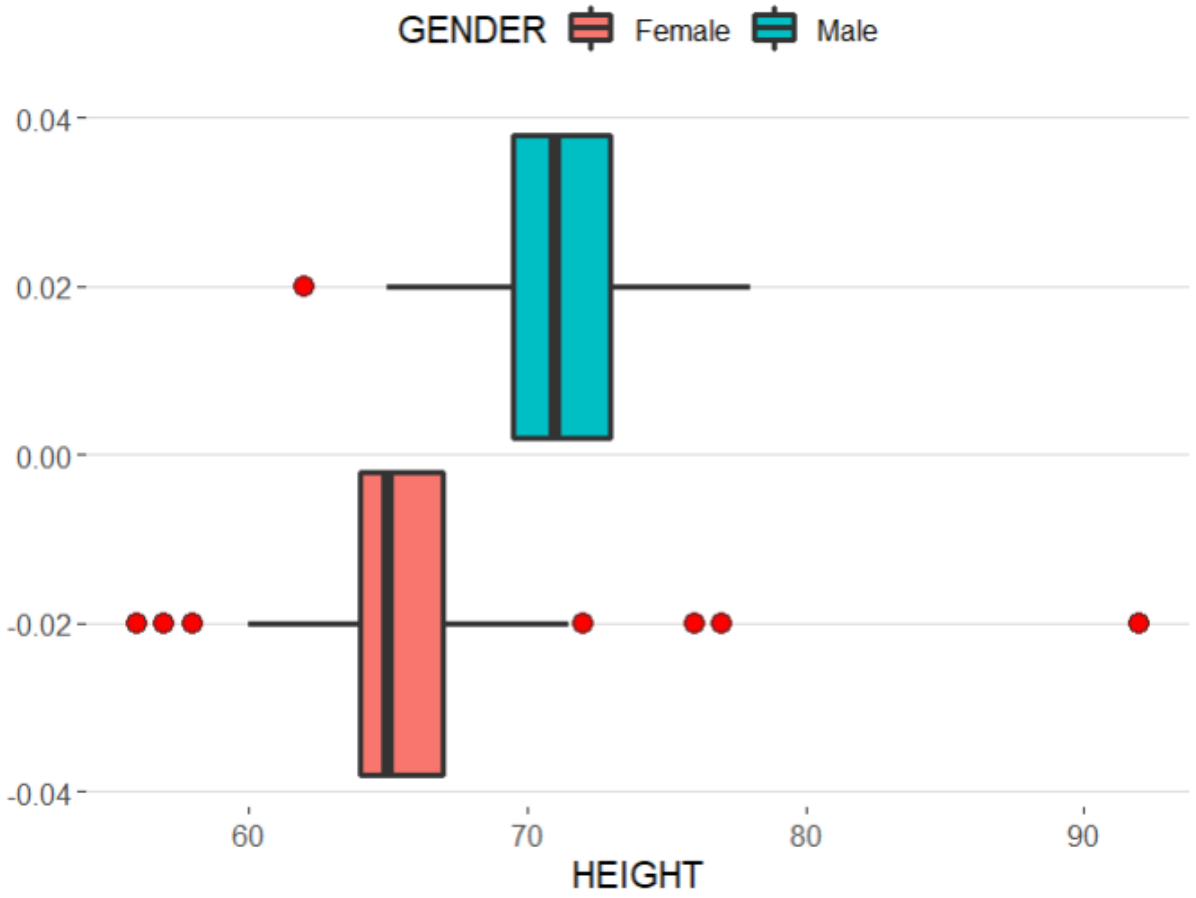
What is the IQR?

- IQR = 4.5 – 2 = 2.5



College Student Heights





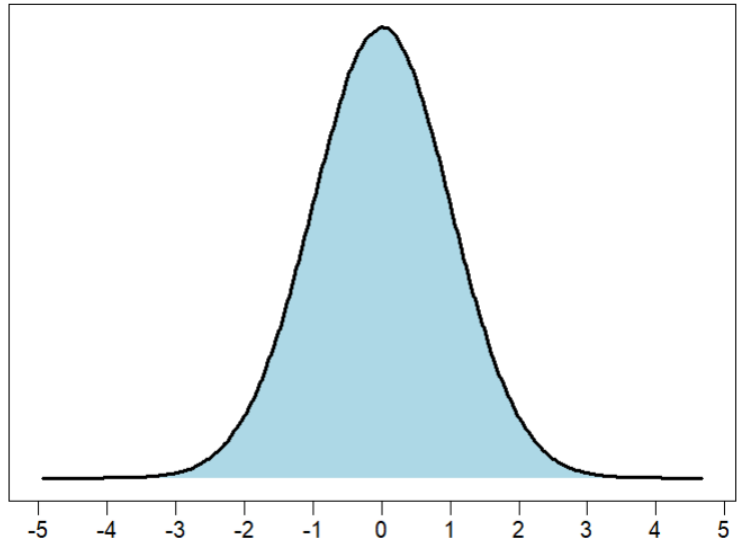
The Normal Distribution

- A family of smooth, bell-shaped (symmetric) distributions that arise often in statistics

- Shape is determined by two parameters: the mean and the standard deviation

  The mean is located where the (relative) frequency is at its peak.

  The standard deviation is the distance from the mean to the value of the variable where the (relative) frequency is a little less than 3/4 of the way (actually about 68%) to its maximum.



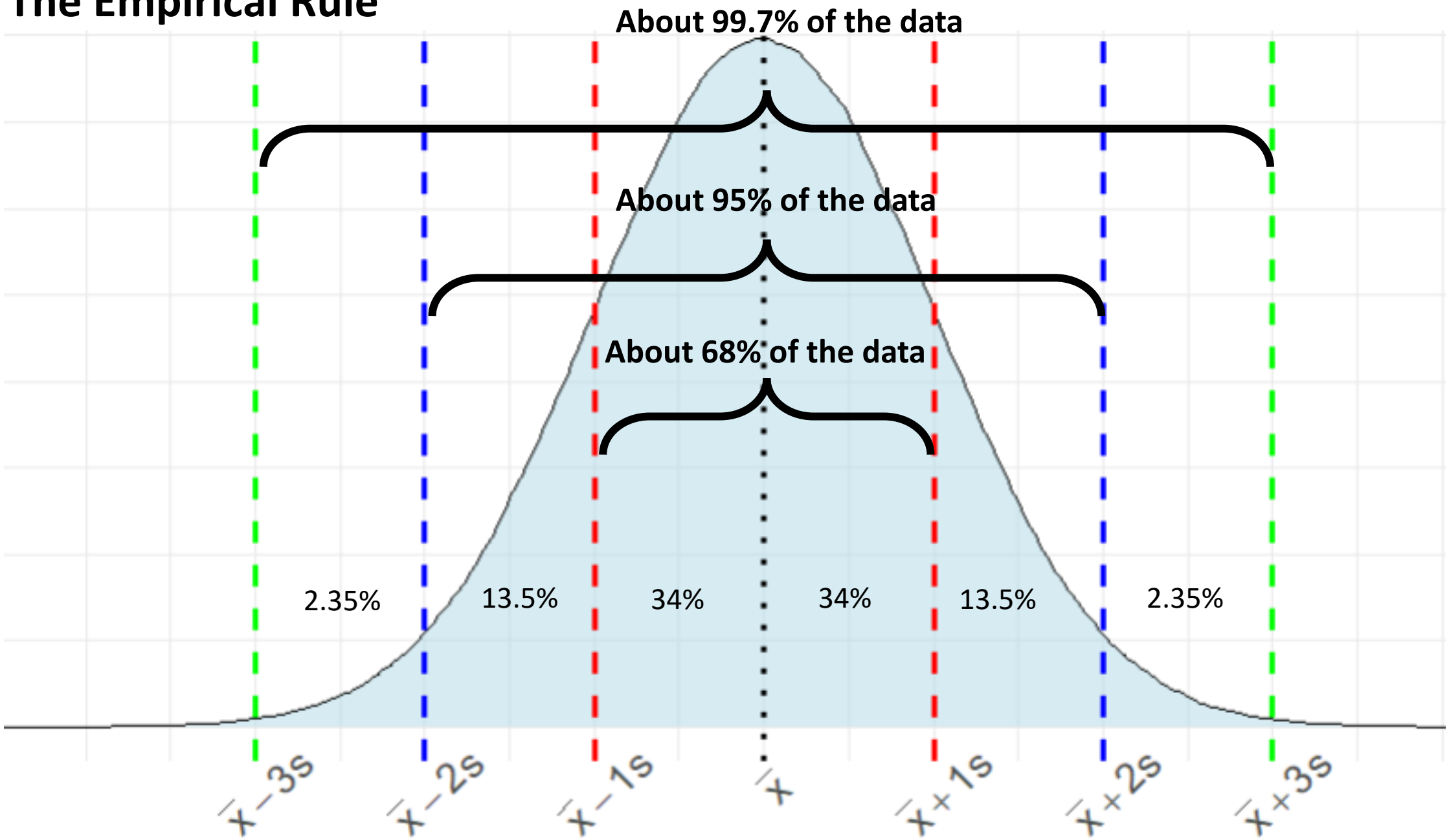
- We denote the normal distribution for a population as

$$x \sim N(\mu, \sigma)$$

- And for a sample as

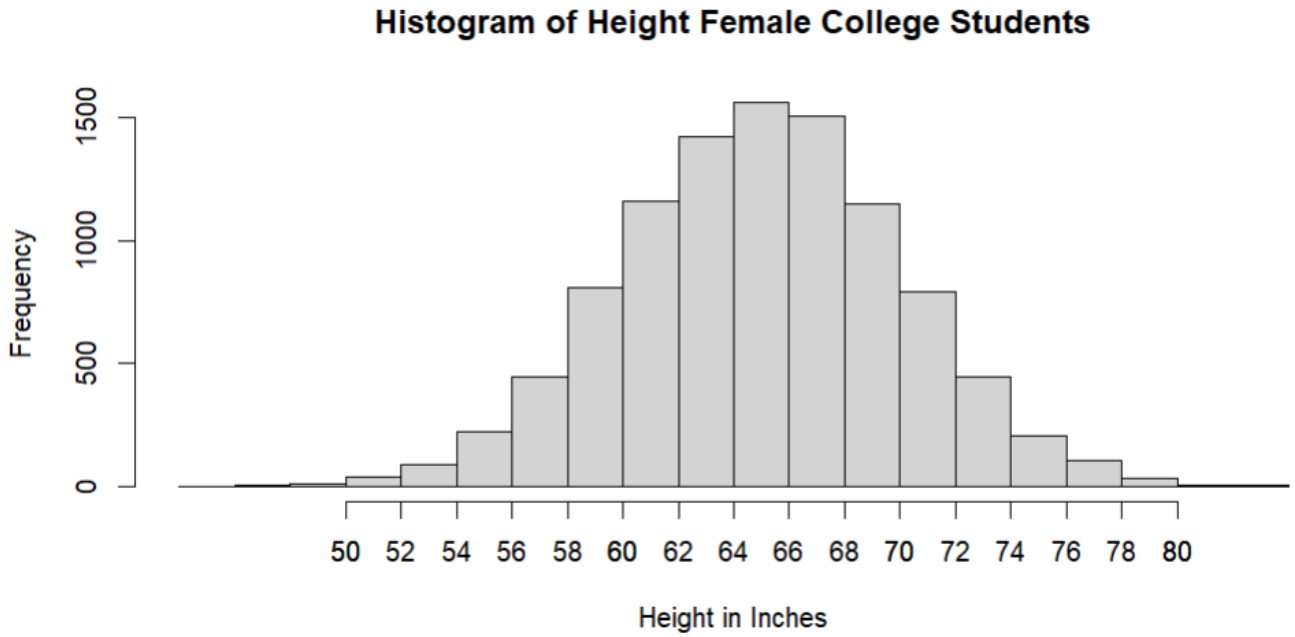
$$x \sim N(\bar{x}, s)$$

The Empirical Rule



Practice

- Suppose the distribution to the left represents the heights of a sample of female college students in the U.S. this distribution has mean and standard deviation
- $\bar{x} \approx 65$ inches
- $s \approx 5$ inches

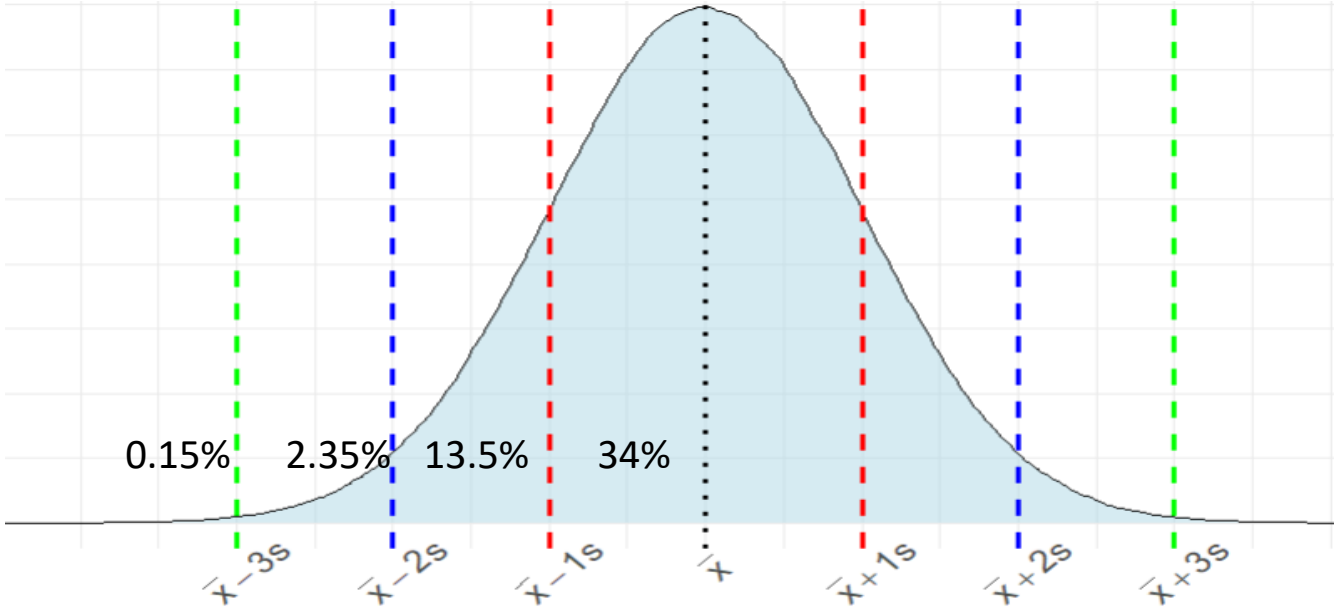


What percentage of students in the sample are shorter than mean height?

50%

What percentage of students in the sample are more than 2 standard deviations above the average height?

About 2.5%



Identifying Outliers: Normal Distributions

- The empirical rule: It is fairly unlikely to observe a value that is more than 2 standard deviations from the mean

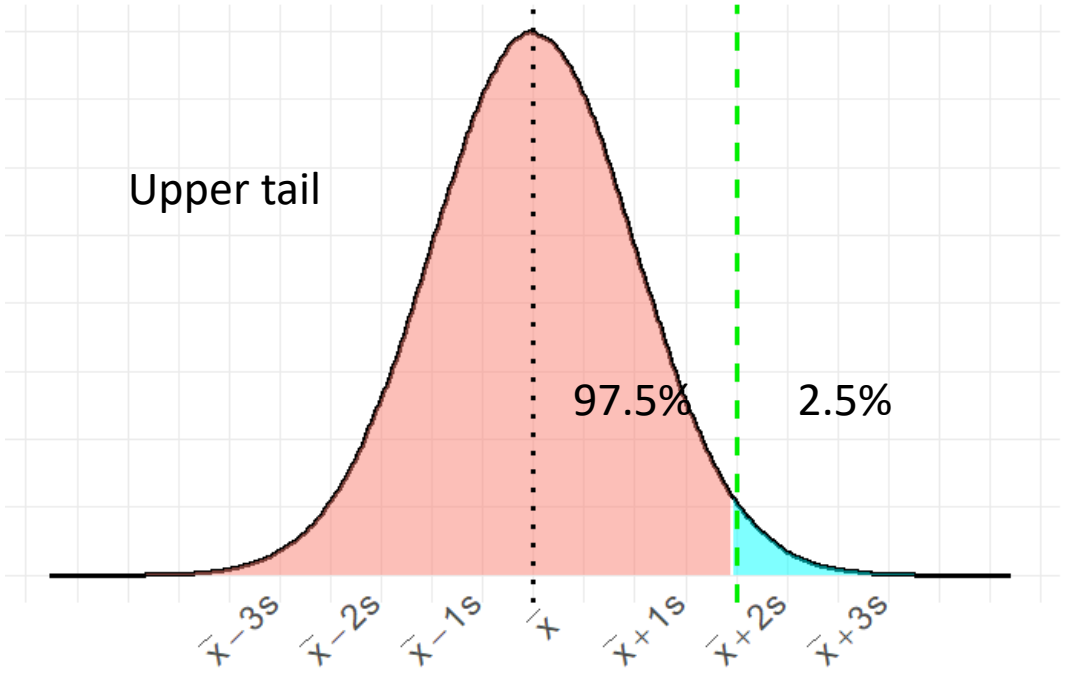
- Therefore, when data are approximately Normally distributed, we can regard all values $\geq 2s$ distance from the mean as outliers

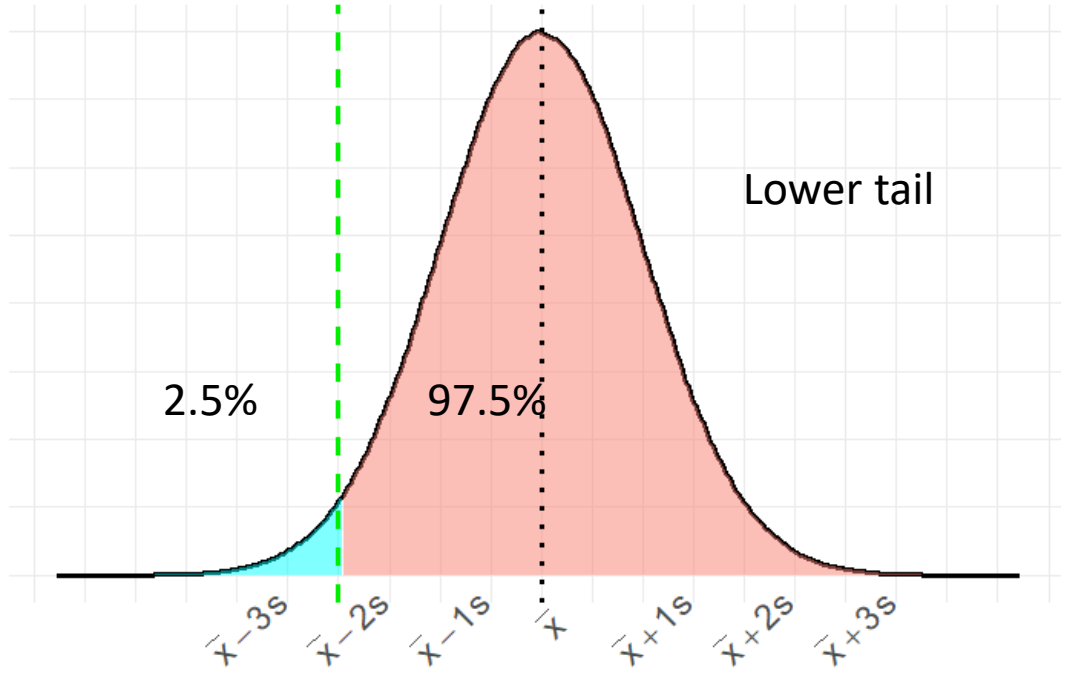
- $\boldsymbol{Z}$ **−score**: The number of standard deviations a value falls from mean

$$z_i = \frac{observation\ - mean}{standard\ deviation}$$

$$= \frac{x_i - \bar{x}}{s} \sim N(0,1) \text{ if } x_i \sim N(\mu, \sigma)$$


Outliers
Outliers
95%
2.5%
2.5%
x̄−3s
x̄−2s
x̄−1s
x̄
x̄+1s
x̄+2s
x̄+3s
x


Upper tail
97.5%
2.5%
x̄−3s
x̄−2s
x̄−1s
x̄
x̄+1s
x̄+2s
x̄+3s


Lower tail
2.5%
97.5%
x̄−3s
x̄−2s
x̄−1s
x̄
x̄+1s
x̄+2s
x̄+3s

Try it out: Female College Student Heights

Height	F(x)	RF(x)	CRF(x)
56	1	0.0038	0.0038
57	1	0.0038	0.0076
58	1	0.0038	0.0115
60	7	0.0267	0.0382
61	10	0.0382	0.0763
62	25	0.0954	0.1718
63	20	0.0763	0.2481
64	45	0.1718	0.4198
65	29	0.1107	0.5305
66	40	0.1527	0.6832
67	31	0.1183	0.8015
68	21	0.0802	0.8817
69	12	0.0458	0.9275
70	5	0.0191	0.9466
71	3	0.0115	0.9580
72	8	0.0305	0.9885
76	1	0.0038	0.9924
77	1	0.0038	0.9962
92	1	0.0038	1.0000

Compute the z-score for a female with a height of 70 inches

Compute the z-score for a female with a height of 92 inches

$\bar{x} = 65.4$

$s = 3.38$

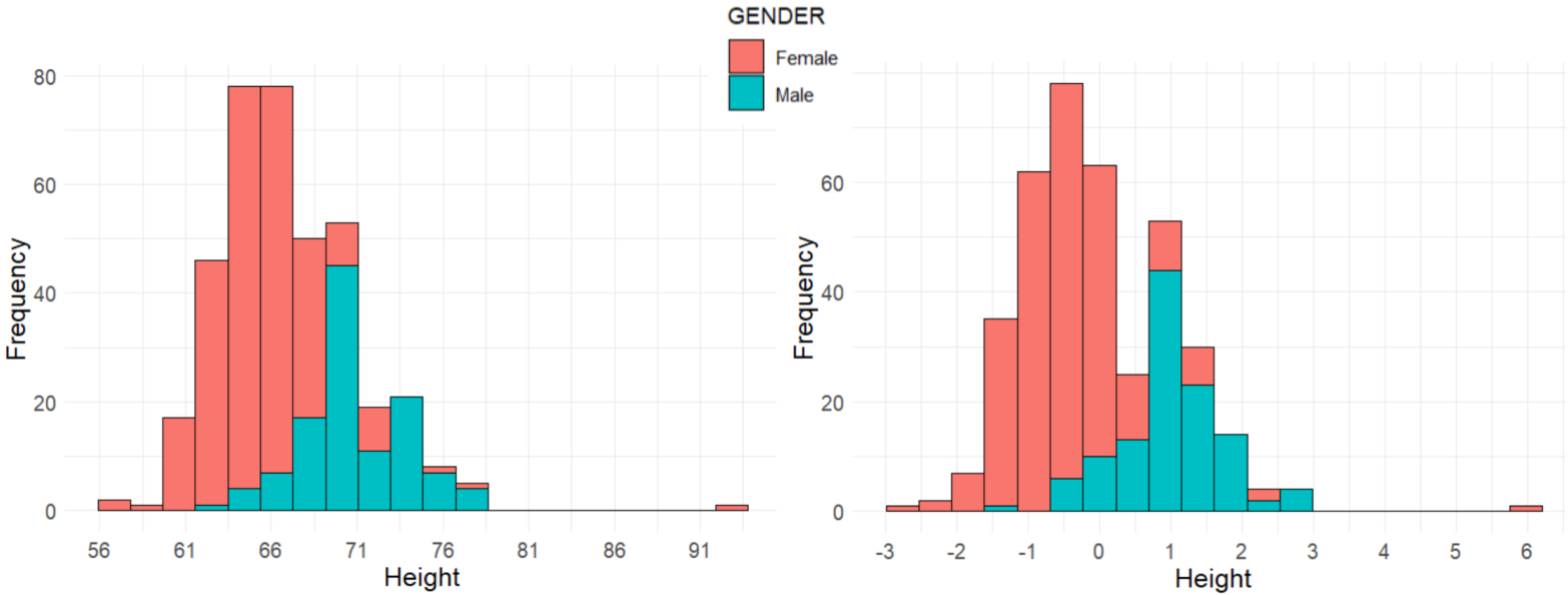
A Note About Transformations of Variables…

- We often need to change the units of measurement of a variable such as from Fahrenheit to Celsius, Feet to meters, dollars to euros etc.

- Linear transformations: adding, subtracting, multiplying, dividing
  - Linear transformations take the form $y = ax + b$ (scaling + shift)
  - a is a scaling constant, b is a shifting constant, x is the original variable and y the transformed variable
  - The z —score is a linear transformation
  - Linear transformations preserve the shape of variables distribution

- Nonlinear transformations: squaring, taking roots, logarithm, exponentiation, etc
  - **Do not** preserved the shape of the variables distribution

College Student Heights



More properties of Linear Transformations

- For a linear transformation of x to y: $y = ax + b$
- $\bar{y} = a\bar{x} + b$
- $median(y) = a \cdot median(x) + b$
- $s_y = |a| \cdot s_x$ (the standard deviation is not affected by shift b)
- $IQR_y = |a| \cdot IQR_x$ (the IQR is not affected by shift b)