TW0S-days – Sept 12th/14th 2019 Today's AGENDA

- **September Calendar**
- **Reading & Warm-UP**
- **Begin Chapter 4**
- **TEST Discussion**

Chapter 4: Numerical Methods for Describing Data

Describing Quantitative Data with Numbers

Chapter 4
Exploring Data

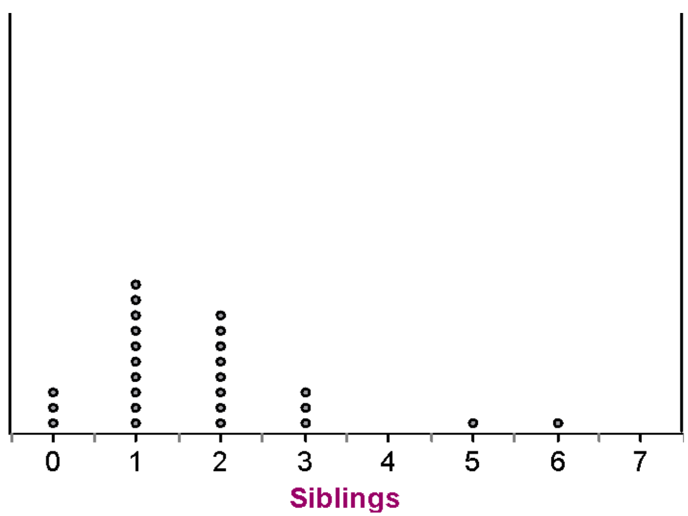
- **4.1** Describing the Center of a Data Set
- **4.2** Describing Variability of a Data Set
- **4.3** Summarizing a Data Set: Boxplots

Warm-Up

1. If percents are referenced by *percentiles*, then quarters must be referenced by ____________

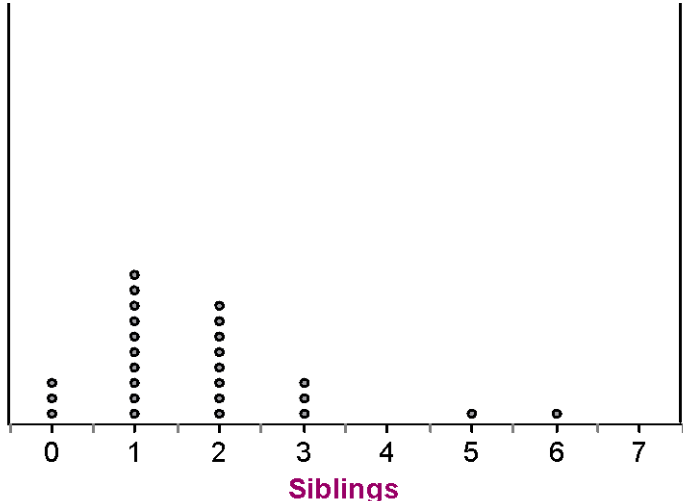
2. What is an outlier?

3. How would you label the shape of this data?



Warm-Up

1. If percents are referenced for *percentiles*, then quarters must be referenced by __*quartiles*__
2. What is an outlier? *Any data that is **unusually** large or **unusually** small compared to the data*
3. How would you label the shape of this data?



Skewed right or positively skewed

Section 4.1
Describing Quantitative Data with Numbers

Learning Objectives: I can…

After this section, you should be able to…

- ✓ MEASURE center with the mean and median
- ✓ MEASURE spread with standard deviation and interquartile range
- ✓ IDENTIFY outliers
- ✓ CONSTRUCT a boxplot using the five-number summary
- ✓ CALCULATE numerical summaries with technology

Measuring Center: The Mean

- The most common measure of center is the ordinary arithmetic average, or **mean**.

Definition:

To find the **mean** $\bar{x}$ (pronounced "x-bar") of a set of observations, add their values and divide by the number of observations. If the *n* observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

In mathematics, the capital Greek letter Σis short for "add them all up." Therefore, the formula for the mean can be written in more compact notation:

$$\bar{x} = \frac{\sum \mathrm{x_i}}{n}$$

Measuring Center: The Median

- Another common measure of center is the **median**. In section 1.2, we learned that the median describes the midpoint of a distribution.

Definition:

The **median M** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1) Arrange all observations from smallest to largest.

2) If the number of observations n is odd, the median M is the center observation in the ordered list.

3) If the number of observations ***n* is even**, the median M is the average of the two center observations in the ordered list.

Measuring Center

- Use the data below to calculate the mean and median of the commuting times (in minutes) of 20 randomly selected New York workers.

Example, page ??

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

$$\bar{x} = \frac{10+30+5+25+\ldots+40+45}{20} = 31.25 \text{ minutes}$$

0	5
1	005555
2	00**05**
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 represents a New York worker who reported a 45-minute travel time to work.

$$M = \frac{20+25}{2} = 22.5 \text{ minutes}$$

Comparing the Mean and the Median

- The mean and median measure center in different ways, and both are useful.
  - *Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.*

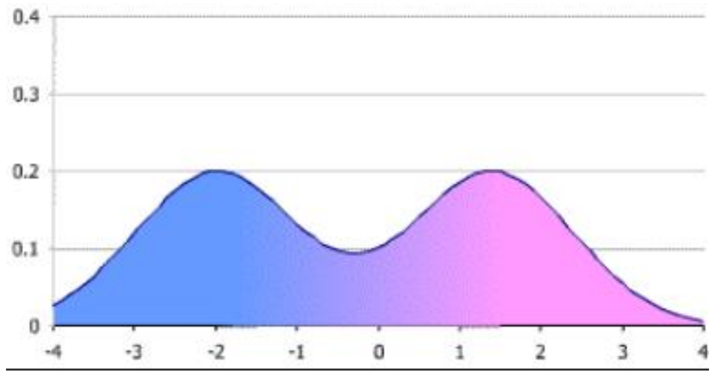
Comparing the Mean and the Median

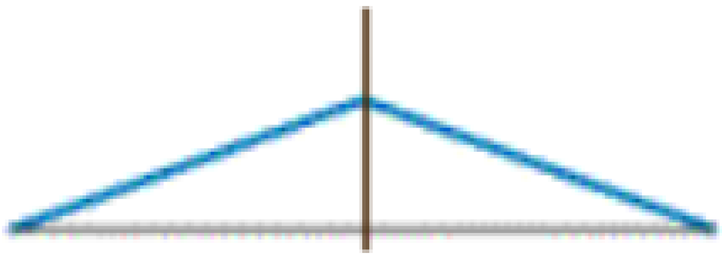
The mean and median of a roughly symmetric distribution are close together.

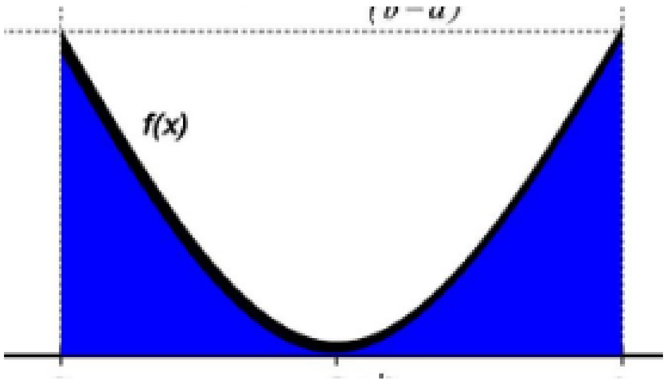
If the distribution is exactly symmetric, the mean and median are exactly the same.

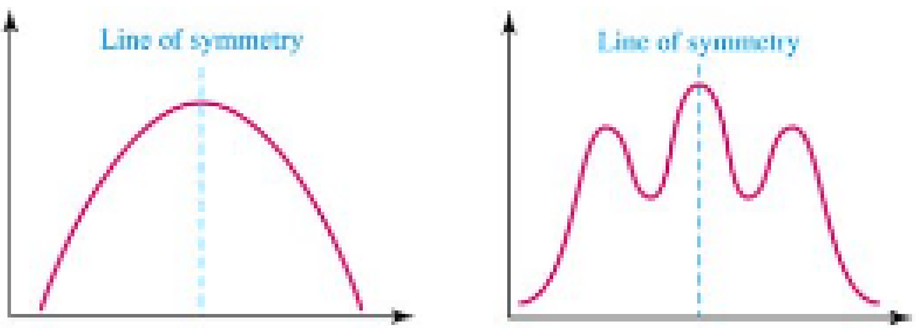
In a skewed distribution, the mean is usually farther out in the long tail than is the median.

Symmetric distribution ≠ Normal distribution



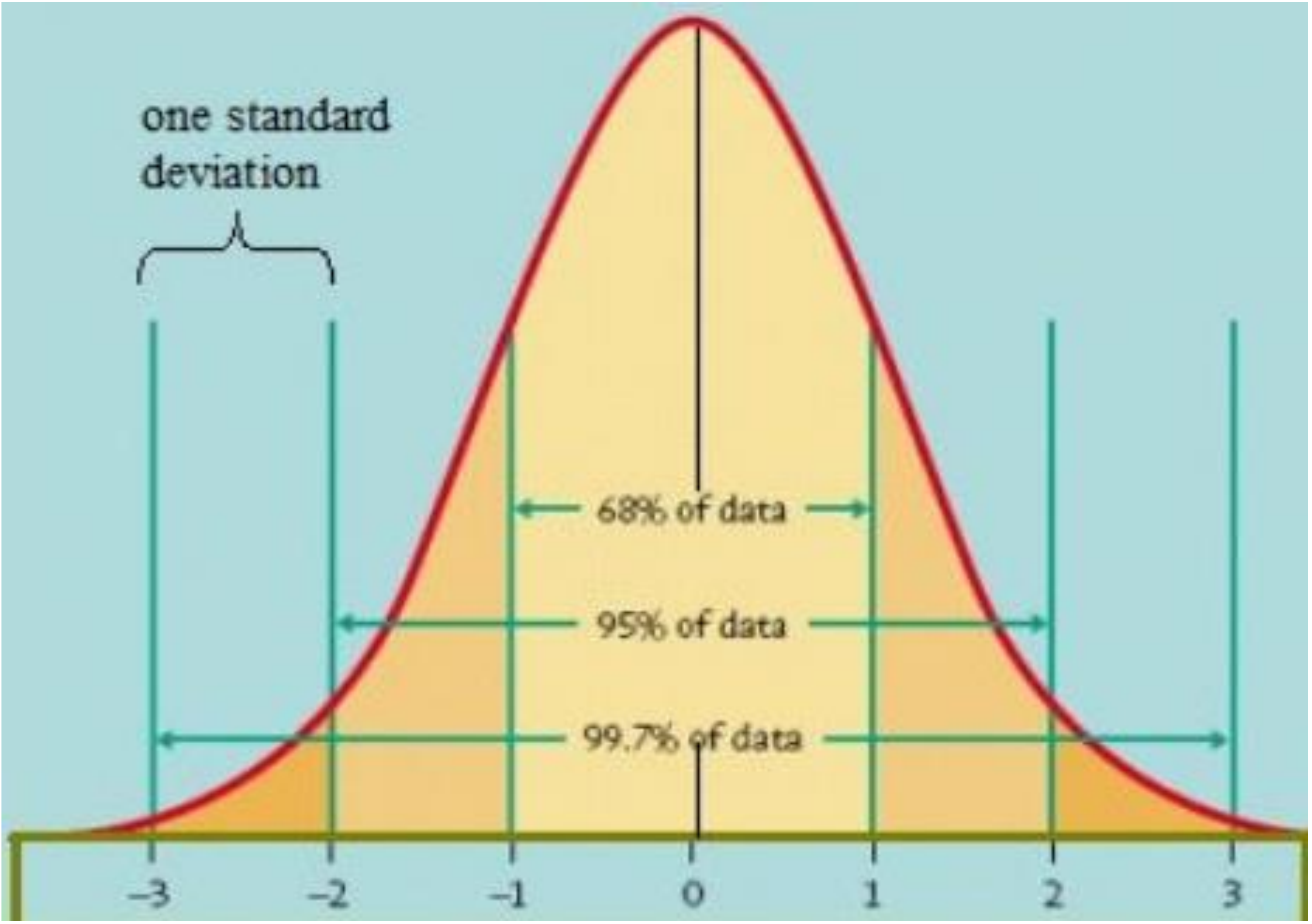






ALL symmetric, but NONE are *Normal distributions*!

Normal distributions are very special symmetric distributions



Sample Proportion of successes

The sample proportion of successes are used when there are only two possible responses, such as male or female, having or not having a driver's license, testing positive or testing negative, etc. Each of these represent a **dichotomy.**

$\hat{p}$ described as "p-hat"

not phat!

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{S}{n}$$

Measuring Spread: The Interquartile Range (*IQR*)

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and a measure of spread.

How to Calculate the Quartiles and the Interquartile Range

To calculate the **quartiles**:

1) Arrange the observations in increasing order and locate the median *M*.

2) The **first quartile** Q_1 is the median of the observations located to the left of the median in the ordered list.

3) The **third quartile** Q_3 is the median of the observations located to the right of the median in the ordered list.

The **interquartile range (*IQR*)** is defined as:

$$IQR = Q_3 - Q_1$$

Find and Interpret the IQR

Example

Travel times to work for 20 randomly selected New Yorkers

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

$Q_1 = 15$ $M = 22.5$ $Q_3 = 42.5$

$$
\begin{aligned}
IQR &= Q_3 - Q_1 \\
&= 42.5 - 15 \\
&= 27.5 \text{ minutes}
\end{aligned}
$$

Interpretation: The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.

Identifying Outliers

- In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

Definition:

The 1.5 x IQR Rule for Outliers

Call an observation an outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile.

Example

In the New York travel time data, we found Q_1=15 minutes, Q_3 = 42.5 minutes, and IQR = 27.5 minutes.

For these data, 1.5 x IQR = 1.5(27.5) = 41.25

Q_1 - 1.5 x IQR = 15 – 41.25 = **-26.25**

Q_3+ 1.5 x IQR = 42.5 + 41.25 = **83.75**

Any travel time shorter than -26.25 minutes or longer than 83.75 minutes is considered an outlier.

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	**5**

The Five-Number Summary

- The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.

- To get a quick summary of both center and spread, combine all five numbers.

Definition:

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

Minimum $\quad Q_1 \quad M \quad Q_3 \quad$ ***Maximum***

Looking Ahead…

In the next part of Chapter 4…

We'll learn how to model distributions of data…

- **Constructing Box Plots**
- **Calculating the IQR & Standard deviation of a distribution**
- **Describing Location in a Distribution**
- **Introduction to Normal Distributions**