# Sullivan Section 3.1

Gene Quinn

# Measures of Central Tendency

By "measures of central tendency" we mean:

measures that numerically describe the **average** or **typical** data value.

# Measures of Central Tendency

By "measures of central tendency" we mean:

measures that numerically describe the **average** or **typical** data value.

We will consider three measures of central tendency:

- the (arithmetic) mean
- the median
- the mode

# Measures of Central Tendency

By "measures of central tendency" we mean:

measures that numerically describe the **average** or **typical** data value.

We will consider three measures of central tendency:

- the (arithmetic) mean
- the median
- the mode

You may find any or all of these measures being reported as "an average".

# Parameters versus Statistics

We will need to distinguish two kinds of descriptive measures:

Descriptive measures of the first kind are measures of the whole **population**.

# Parameters versus Statistics

We will need to distinguish two kinds of descriptive measures:

Descriptive measures of the first kind are measures of the whole **population**.

These descriptive measures are known as **parameters**

# Parameters versus Statistics

We will need to distinguish two kinds of descriptive measures:

Descriptive measures of the first kind are measures of the whole **population**.

These descriptive measures are known as **parameters**

Descriptive measures of the second kind are measures of a **sample** taken from the population.

# Parameters versus Statistics

We will need to distinguish two kinds of descriptive measures:

Descriptive measures of the first kind are measures of the whole **population**.

These descriptive measures are known as **parameters**

Descriptive measures of the second kind are measures of a **sample** taken from the population.

These descriptive measures are known as **statistics**

# Parameters versus Statistics

**Example:**

We compute the average SAT verbal score for all persons who took the test May 16th.

This would be considered a **parameter** because it measures the whole population.

# Parameters versus Statistics

**Example:**

We compute the average SAT verbal score for all persons who took the test May 16th.

This would be considered a **parameter** because it measures the whole population.

We compute average SAT verbal score for 1000 students randomly selected from those who took the test on May 16th.

This would be considered a **statistic**.

# Parameters versus Statistics

**Example:**

The average height of all persons living in the United States.

This would be considered a **parameter** because it measures the whole population.

# Parameters versus Statistics

**Example:**

The average height of all persons living in the United States.

This would be considered a **parameter** because it measures the whole population.

The average height of all persons in the locker room at halftime of a professional basketball game could be thought of as a *cluster sample* of the population of the United States.

# Parameters versus Statistics

**Example:**

The average height of all persons living in the United States.

This would be considered a **parameter** because it measures the whole population.

The average height of all persons in the locker room at halftime of a professional basketball game could be thought of as a *cluster sample* of the population of the United States.

This measure would be considered a **statistic**

# Parameters versus Statistics

We can obtain the value of a **statistic** by taking a sample.

# Parameters versus Statistics

We can obtain the value of a **statistic** by taking a sample.

The only way to obtain the value of a **parameter** is to take a census of the entire population.

# The Arithmetic Mean

The first measure of central tendency we will consider is the

**arithmetic mean**

# The Arithmetic Mean

The first measure of central tendency we will consider is the

**arithmetic mean**

The **arithmetic mean** of a variable is computed in two steps.

First, add up all of the values of the variable in the data.

# The Arithmetic Mean

The first measure of central tendency we will consider is the

**arithmetic mean**

The **arithmetic mean** of a variable is computed in two steps.

First, add up all of the values of the variable in the data.

Next, divide the sum obtained in step one by the number of observations.

# The Arithmetic Mean

**Example:**

The **arithmetic mean** of the ten numbers

$$1, 2, 3, 2, 5, 1, 6, 2, 3, 3, 4, 8$$

is computed by adding the ten numbers, and dividing the sum by 10:

$$\frac{1+2+3+2+5+1+6+2+3+3+4+8}{10}$$

# The Arithmetic Mean

**Example:**

The **arithmetic mean** of the ten numbers

$$1, 2, 3, 2, 5, 1, 6, 2, 3, 3, 4, 8$$

is computed by adding the ten numbers, and dividing the sum by 10:

$$\frac{1+2+3+2+5+1+6+2+3+3+4+8}{10}$$

After adding the 10 numbers in the numerator, the fraction is

$$\frac{40}{10} \quad = \quad 4$$

# Notation: Population Arithmetic Mean

If our arithmetic mean is taken over the entire population, the following notation is used:

The number of observations is the population size and is denoted by $N$.

# Notation: Population Arithmetic Mean

If our arithmetic mean is taken over the entire population, the following notation is used:

The number of observations is the population size and is denoted by $N$.

The arithmetic mean is denoted by the Greek letter $\mu$ (pronounced "mew").

# Notation: Population Arithmetic Mean

If our arithmetic mean is taken over the entire population, the following notation is used:

The number of observations is the population size and is denoted by $N$.

The arithmetic mean is denoted by the Greek letter $\mu$ (pronounced "mew").

The formula for $\mu$ is

$$\mu = \frac{x_1 + x_2 + \cdots x_N}{N}$$

# Notation: Population Arithmetic Mean

The sum in the denominator is usually abbreviated using the upper case Greek letter sigma:

$$\mu = \frac{\sum x_i}{N}$$

$\mu$ is a **parameter**

# Notation: Sample Arithmetic Mean

If our arithmetic mean is taken over a *sample* of the population, the following notation is used:

The number of observations is the *sample* size and is denoted by $n$.

# Notation: Sample Arithmetic Mean

If our arithmetic mean is taken over a *sample* of the population, the following notation is used:

The number of observations is the *sample* size and is denoted by $n$.

The arithmetic mean is denoted by the symbol $\overline{x}$ (pronounced "x-bar").

# Notation: Sample Arithmetic Mean

If our arithmetic mean is taken over a *sample* of the population, the following notation is used:

The number of observations is the *sample* size and is denoted by $n$.

The arithmetic mean is denoted by the symbol $\overline{x}$ (pronounced "x-bar").

The formula for $\overline{x}$ is

$$\overline{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

# Notation: Sample Arithmetic Mean

The sum in the denominator is usually abbreviated using the upper case Greek letter sigma:

$$\overline{x} = \frac{\sum x_i}{n}$$

$\overline{x}$ is a **statistic**

# The Median

The second measure of central tendency we will consider is the

**median**

The **median** of a variable is the value that lies in the middle of the list when the data is arranged in ascending order.

# The Median

The second measure of central tendency we will consider is the

**median**

The **median** of a variable is the value that lies in the middle of the list when the data is arranged in ascending order.

In other words, if the data is arranged in an ordered list, half of the observations fall below the median, and half fall above.

# The Median

The second measure of central tendency we will consider is the

**median**

The **median** of a variable is the value that lies in the middle of the list when the data is arranged in ascending order.

In other words, if the data is arranged in an ordered list, half of the observations fall below the median, and half fall above.

The median is usually denoted by the symbol $M$.

# Computing the Median - n odd

The median $M$ is computed in one of two ways, depending on whether $n$ is even or odd.

# Computing the Median - n odd

The median $M$ is computed in one of two ways, depending on whether $n$ is even or odd.

When $n$ is **odd**, we use the following procedure:

Arrange the data values in ascending order.

The median $M$ is the value that lies exactly in the middle of the list

# Computing the Median - n odd

The median $M$ is computed in one of two ways, depending on whether $n$ is even or odd.

When $n$ is **odd**, we use the following procedure:

Arrange the data values in ascending order.

The median $M$ is the value that lies exactly in the middle of the list

If the ascending list is numbered from $1$ to $n$, the median $M$ is the observation that is in position

$$\frac{n+1}{2}$$

# Computing the Median - n even

When $n$ is **even**, we use the following procedure:

Arrange the data values in ascending order.

# Computing the Median - n even

When $n$ is **even**, we use the following procedure:

Arrange the data values in ascending order.

The median $M$ is the **arithmetic mean** of the two middle observations in the data set.

# Computing the Median - n even

When $n$ is **even**, we use the following procedure:

Arrange the data values in ascending order.

The median $M$ is the **arithmetic mean** of the two middle observations in the data set.

If the ascending list is numbered from $1$ to $n$, the median $M$ is the arithmetic mean of the data values in positions

$$\frac{n}{2} \quad \text{and} \quad \frac{n}{2}+1$$

in the list.

# Computing the Median

**Example:**

Find the median of the numbers

$$2, 5, 3, 1, 7, 9, 5, 2, 1, 1, 6$$

# Computing the Median

First arrange the list in ascending order:

$$1, 1, 1, 2, 2, 3, 5, 5, 6, 7, 9$$

# Computing the Median

First arrange the list in ascending order:

$$1, 1, 1, 2, 2, 3, 5, 5, 6, 7, 9$$

Since $n = 11$ is odd, the median $M$ is the number in position

$$\frac{n+1}{2} = \frac{12}{2} = 6$$

in the list: $M = 3$.

# Computing the Median

**Example:**

Find the median of the numbers

$$2, 5, 3, 1, 7, 9, 5, 2, 1, 1, 6, 10$$

# Computing the Median

First arrange the list in ascending order:

$$1, 1, 1, 2, 2, 3, 5, 5, 6, 7, 9, 10$$

# Computing the Median

First arrange the list in ascending order:

$$1, 1, 1, 2, 2, 3, 5, 5, 6, 7, 9, 10$$

Since $n = 12$ is even, the median $M$ is the arithmetic mean of the values in positions

$$\frac{n}{2} = 6 \quad \text{and} \quad \frac{n}{2} + 1 = 7$$

in the list:

$$M = \frac{3 + 5}{2} = 4$$

# The Mode

The third measure of central tendency we will consider is the

**mode**

The **mode** of a variable is the value of that variable that occurs most frequently in the data.

# The Mode

The third measure of central tendency we will consider is the

**mode**

The **mode** of a variable is the value of that variable that occurs most frequently in the data.

In other words, compute the number of times each value of the variable appears in the data.

The value with the highest count is the **mode**.

# The Mode

**Example:**

Find the mode of the following data:

$$1, 4, 3, 2, 7, 5, 4, 6, 5, 2, 2, 6, 2, 1, 7, 9$$

# The Mode

**Example:**

Find the mode of the following data:

$$1, 4, 3, 2, 7, 5, 4, 6, 5, 2, 2, 6, 2, 1, 7, 9$$

The number of times each value appears is:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 9 |
|---|---|---|---|---|---|---|---|
| 2 | 4 | 1 | 2 | 2 | 2 | 2 | 1 |

The mode is the value with the highest count: $2$

# The Mode

Unlike the mean and the median, the mode of a variable may not exist.

# The Mode

Unlike the mean and the median, the mode of a variable may not exist.

If every value in the data set appears exactly once, the data is considered not to have a mode.

# The Mode

Unlike the mean and the median, the mode of a variable may not exist.

If every value in the data set appears exactly once, the data is considered not to have a mode.

The following data has no mode:

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

# The Mode

On the other hand, a set of data may have more than one mode if several values are tied for the highest frequency.

# The Mode

On the other hand, a set of data may have more than one mode if several values are tied for the highest frequency.

The following data has three modes, $2$, $4$, and $6$:

$$1, 2, 2, 3, 4, 4, 5, 6, 6, 7$$

# The Mode

We note in passing that qualitative data can have a mode.

# The Mean, Median, and Skewness

Together the mean and median determine the skewness of the data set:

If the median is **below** the mean, the data is said to be **skewed right**

# The Mean, Median, and Skewness

Together the mean and median determine the skewness of the data set:

If the median is **below** the mean, the data is said to be **skewed right**

If the median is **above** the mean, the data is said to be **skewed left**

# The Mean, Median, and Skewness

Together the mean and median determine the skewness of the data set:

If the median is **below** the mean, the data is said to be **skewed right**

If the median is **above** the mean, the data is said to be **skewed left**

If the median is **equal to** the mean, the data is said to be **symmetric**

# Mean vs Median

Often real world data contains values called **outliers** that differ markedly from the rest of the sample or population.

# Mean vs Median

Often real world data contains values called **outliers** that differ markedly from the rest of the sample or population.

Suppose a street has of four modest houses built in the 1940's, and one enormous house built in 1996.

Assume for tax purposes the houses are valued at:

- 1. 125,000
- 2. 175,000
- 3. 145,000
- 4. 160,000
- 5. 3,120,000

# Mean vs Median

- 1. 125,000
- 2. 175,000
- 3. 145,000
- 4. 160,000
- 5. 3,120,000

The mean of these values is 745,000 - not very representative.

# Mean vs Median

- 1. 125,000
- 2. 175,000
- 3. 145,000
- 4. 160,000
- 5. 3,120,000

The mean of these values is 745,000 - not very representative.

The median is the 160,000, which is more representative of the typical house on the street.

# Mean vs Median

- 1. 125,000
- 2. 175,000
- 3. 145,000
- 4. 160,000
- 5. 3,120,000

The mean of these values is 745,000 - not very representative.

The median is the 160,000, which is more representative of the typical house on the street.

The mean is sensitive to outliers, while the median essentially ignores them.