Measurement of Central Tendency

We often represent a data set by numerical summary measures, usually called the typical values. A measure of central tendency gives the center of a histogram or a frequency distribution curve. This section discusses three different measures of central tendency: the mean, the median, and the mode; however, a few other measures of central tendency, such as the trimmed mean, the weighted mean, and the geometric mean, are explained in exercises following this section. We will learn how to calculate each of these measures for ungrouped data. Recall that the data that give information on each member of the population or sample individually are called ungrouped data, whereas grouped data are presented in the form of a frequency distribution table.

1- Mean

The mean, also called the arithmetic mean, is the most frequently used measure of central tendency. We can be use the words mean and average synonymously. For ungrouped data, the mean is obtained by dividing the sum of all values by the number of values in the data set:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

The mean calculated for sample data is denoted by (read as $\overline{x}$ bar"), and the mean calculated for population data is denoted by ($\boldsymbol{\mu}$) (Greek letter *mu*). We know from the discussion in previous lectures the number of values in a data set is denoted by *n* for a sample and by *N* for a population. And also, we learned that a variable is denoted by *x*, and the sum of all values of *x* is denoted by $\sum_{i=1}^{n} x_i$. Using these notations, we can write the following formulas for the mean.

Calculating Mean for Ungrouped Data The *mean for ungrouped data* is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data: } \mu = \frac{\Sigma x}{N}$$

$$\text{Mean for sample data: } \overline{x} = \frac{\Sigma x}{n}$$

where Σx is the sum of all values, N is the population size, n is the sample size, μ is the population mean, and $\overline{x}$ is the sample mean.

EXAMPLE [1]

Table (1) lists the total sales (rounded to billions of dollars) of six U.S. companies for 2008.

Company	Total Sales (billions of dollars)
General Motors	149
Wal-Mart Stores	406
General Electric	183
Citigroup	107
Exxon Mobil	426
Verizon Communication	97

Find the 2008 mean sales for these six companies.

Solution :The variable in this example is the 2008 total sales for a company. Let us denote this variable by x. Then, the six values of x are

$$x_1 = 149, \quad x_2 = 406, \quad x_3 = 183, \quad x_4 = 107, \quad x_5 = 426, \quad \text{and} \quad x_6 = 97$$

where $x_1 = 149$ represents the 2008 total sales of General Motors, $x_2 = 406$ represents the 2008 total sales of Wal-Mart Stores, and so on. The sum of the 2008 sales for these six companies is

$$\begin{aligned}\Sigma x &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \\ &= 149 + 406 + 183 + 107 + 426 + 97 = 1368\end{aligned}$$

Note that the given data include only six companies. Hence, they represent a sample. Because the given data set contains six companies, $n = 6$. Substituting the values of $\sum_{i=1}^{n} X_i$ and n in the sample formula, we obtain the mean 2008 sales of the six companies:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1368}{6} = 228 = \$228 \text{ Billion}$$

Thus, the mean 2008 sales of these six companies was 228, or $228 billion.

EXAMPLE [2]

The following are the ages (in years) of all eight employees of a small company:

53 32 61 27 39 44 49 57

Find the mean age of these employees.

Solution : Because the given data set includes *all* eight employees of the company, it represents the population. Hence, $N = 8$. We have

$$\Sigma x = 53 + 32 + 61 + 27 + 39 + 44 + 49 + 57 = 362$$

The population mean is

$$\mu = \frac{\Sigma x}{N} = \frac{362}{8} = \mathbf{45.25\ years}$$

Thus, the mean age of all eight employees of this company is 45.25 years, or 45 years and 3 months.

2 - *Median*

Another important measure of central tendency is the **median**. It is defined as follows.

Definition

Median The *median* is the value of the middle term in a data set that has been ranked in increasing order.

As is obvious from the definition of the median, it divides a ranked data set into two equal parts. The calculation of the median consists of the following two steps:

1. Rank the data set in increasing order.

2. Find the middle term. The value of this term is the median.

Note that if the number of observations in a data set is *odd*, then the median is given by the value of the middle term in the ranked data. However, if the number of observations is *even*, then the median is given by the average of the values of the two middle terms.

EXAMPLE [3]

The following data give the prices (in thousands of dollars) of seven houses selected from all houses sold last month in a city.

312 257 421 289 526 374 497

Find the median.

Solution : First, we rank the given data in increasing order as follows:

257 289 312 374 421 497 526

↑ Median

Thus, the median price of a house is 374, or $374,000.

EXAMPLE [4]

Table below. gives the 2008 profits (rounded to billions of dollars) of 12 companies selected from all over the world.

Company	2008 Profits (billions of dollars)
Merck & Co	8
IBM	12
Unilever	7
Microsoft	17
Petrobras	14
Exxon Mobil	45
Lukoil	10
AT&T	13
Nestlé	17
Vodafone	13
Deutsche Bank	9
China Mobile	11

Find the median for these data.

***Solution*:** First we rank the given profits as follows:

7 8 9 10 11 12 13 13 14 17 17 45

There are 12 values in this data set. Because there is an even number of values in the data set, the median is given by the average of the two middle values. The two middle values are the sixth and seventh in the foregoing list of data, and these two values are 12 and 13. The median, which is given by the average of these two values, is calculated as follows.

7 8 9 10 11 12 13 13 14 17 17 45

↑ Median

$$\text{Median} = \frac{12 + 13}{2} = \frac{25}{2} = 12.5 = \textbf{\$12.5 billion}$$

Thus, the median profit of these 12 companies is $12.5 billion.

The median gives the center of a histogram, with half of the data values to the left of the median and half to the right of the median. The advantage of using the median as a measure of central tendency is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of central tendency for data sets that contain outliers.

3- Mode

Mode is a French word that means *fashion*—an item that is most popular or common. In statistics, the mode represents the most common value in a data set.

Definition

Mode The *mode* is the value that occurs with the highest frequency in a data set.

EXAMPLE [5]

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77 82 74 81 79 84 74 78

Find the mode.

Solution: In this data set, 74 occur twice, and each of the remaining values occurs only once. Because 74 occur with the highest frequency, it is the mode. Therefore,

Mode = 74 miles per hour

A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median. For instance, a data set with each value occurring only once has no mode. A data set with only one value occurring with the highest frequency has only one mode. The data set in this case is called **unimodal**. A data set with two values that occur with the same (highest) frequency has two modes. The distribution, in this case, is said to be **bimodal**. If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be **multimodal**.

EXAMPLE [6]

Last year's incomes of five randomly selected families were $76,150, $95,750, $124,985, $87,490, and $53,740. Find the mode.

Solution: Because each value in this data set occurs only once, this data set contains **no mode**.

EXAMPLE[7]

Refer to the data on 2008 profits of 12 companies given in Table of Example [4]. Find the mode for these data.

Solution: In the data given in Example [4], each of the two values 13 and 17 occurs twice, and each of the remaining values occurs only once. Therefore, that data set has two modes:

$13 billion and $17 billion.

EXAMPLE[8]

The ages of 10 randomly selected students from a class are 21, 19, 27, 22, 29, 19, 25, 21, 22, and 30 years, respectively. Find the mode.

Solution: This data set has three modes: **19**, **21**, and **22**. Each of these three values occurs with a (highest) frequency of 2.

One advantage of the mode is that it can be calculated for both kinds of data—quantitative and qualitative—whereas the mean and median can be calculated for only quantitative data.

_ EXAMPLE [9]

The status of five students who are members of the student senate at a college are senior, sophomore, senior, junior, and senior, respectively. Find the mode.

Solution: Because **senior** occurs more frequently than the other categories, it is the mode for this data set. We cannot calculate the mean and median for this data set.

To sum up, we cannot say for sure which of the three measures of central tendency is a better measure overall. Each of them may be better under different situations. Probably the mean is the most-used measure of central tendency, followed by the median. The mean has the advantage that its calculation includes each value of the data set. The median is a better measure when a data set includes outliers. The mode is simple to locate, but it is not of much use in practical applications.

To find the midpoint of the upper limit of the first class and the lower limit of the second class in Table (5), we divide the sum of these two limits by 2. Thus, this midpoint is

$$\frac{1000 + 1001}{2} = 1000.5$$

The value 1000.5 is called the *upper boundary* of the first class and the *lower boundary* of the second class. By using this technique, we can convert the class limits of Table (5) to **class boundaries**, which are also called *real class limits*. The second column of Table 2.8 lists the boundaries for Table (5).

Definition

Class Boundary The *class boundary* is given by the midpoint of the upper limit of one class and the lower limit of the next class.

The difference between the two boundaries of a class gives the **class width**. The class width is also called the **class size**.

Finding Class Width

Class width = Upper boundary − Lower boundary

Thus, in Table (6), the class widths for the frequency distribution of Table (5) are listed in the third column of Table 2.8. Each class in Table 2.8 (and Table (5) has the same width of 200.

The class midpoint or mark is obtained by dividing the sum of the two limits (or the two boundaries) of a class by 2.

Calculating Class Midpoint or Mark

$$\text{Class midpoint or mark} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Thus, the midpoint of the first class in Table (5) or Table (6) is calculated as follows:

The class midpoints for the frequency distribution of Table (5) are listed in the fourth column of Table (6).

Table (6)

Class Limits	Class Boundaries	Class Width	Class Midpoint
801 to 1000	800.5 to less than 1000.5	200	900.5
1001 to 1200	1000.5 to less than 1200.5	200	1100.5
1201 to 1400	1200.5 to less than 1400.5	200	1300.5
1401 to 1600	1400.5 to less than 1600.5	200	1500.5
1601 to 1800	1600.5 to less than 1800.5	200	1700.5
1801 to 2000	1800.5 to less than 2000.5	200	1900.5

Note that in Table 2.8, when we write classes using class boundaries, we write *to less than* to ensure that each value belongs to one and only one class. As we can see, the upper boundary of the preceding class and the lower boundary of the succeeding class are the same.

Constructing Frequency Distribution Tables

When constructing a frequency distribution table, we need to make the following three major decisions.

- **Number of Classes**

Usually the number of classes for a frequency distribution table varies from 5 to 20, depending mainly on the number of observations in the data set.1 It is preferable to have more classes as the size of a data set increases. The decision about the number of classes is arbitrarily made by the data organizer.

- **Class Width**

Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are the same size, first find the difference between the largest and the smallest values in the data. Then, the approximate width of a class is obtained by dividing this difference by the number of desired classes.

Calculation of Class Width

$$\text{Approximate class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

Usually this approximate class width is rounded to a convenient number, which is then used as the class width. Note that rounding this number may slightly change the number of classes initially intended.

- **Lower Limit of the First Class or the Starting Point**

Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

Example 2–3 illustrates the procedure for constructing a frequency distribution table for quantitative data.

The following data give the total number of iPods® sold by a mail order company on each of 30 days. Construct a frequency distribution table.

8	25	11	15	29	22	10	5	17	21
22	13	26	16	18	12	9	26	20	16
23	14	19	23	20	16	27	16	21	14

Solution In these data, the minimum value is 5, and the maximum value is 29. Suppose we decide to group these data using five classes of equal width. Then,

$$\text{Approximate width of each class} = \frac{29 - 5}{5} = 4.8$$

Now we round this approximate width to a convenient number, say 5. The lower limit of the first class can be taken as 5 or any number less than 5. Suppose we take 5 as the lower limit of the first class. Then our classes will be

5–9, 10–14, 15–19, 20–24, and 25–29

We record these five classes in the first column of Table (7).

One rule to help decide on the number of classes is Sturge's formula:

$$c = 1 + 3.3 \log n$$

Where c is the number of classes and n is the number of observations in the data set. The value of log n can be obtained by using a calculator.

Now we read each value from the given data and mark a tally in the second column of Table (7) next to the corresponding class. The first value in our original data set is 8, which belongs to the 5–9 class. To record it, we mark a tally in the second column next to the 5–9 class.

We continue this process until all the data values have been read and entered in the tally column.

Note that tallies are marked in blocks of five for counting convenience. After the tally column is completed, we count the tally marks for each class and write those numbers in the third column.

This gives the column of frequencies. These frequencies represent the number of days on which iPods indicated in classes are sold. For example, on 8 of 30 days, 15 to 19 iPods were sold.

on iPods Sold

iPods Sold	Tally	f
5–9	|||	3
10–14	𝍸 |	6
15–19	𝍸 |||	8
20–24	𝍸 |||	8
25–29	𝍸	5
		$\Sigma f = 30$