

Pharmaceutical Statistics

Lecture 3

Descriptive statistics

Indicators of central tendency

Prepared and presented by

Dr. Muna Oqal

Important Characteristics of Data Description

1. **Centre:** A representative or average value that indicates where the middle of the data set is located
2. **Variation:** A measure of the amount that the values vary among themselves
3. **Distribution:** The nature or shape of the distribution of data (such as bell-shaped, uniform, or skewed)
4. **Outliers:** Sample values that lie very far away from the vast majority of other sample values

Measures of Central Tendency

- A measure of central tendency is a measure which indicates where the **value at the center or middle** of a data set is.
- The three most commonly used measures of central tendency are: **Mean, Median, and Mode.**



DESCRIBING DATA

MEAN	Average or arithmetic mean of the data
MEDIAN	The value which comes half way when the data are ranked in order
MODE	Most common value observed

- **In a normal distribution, mean and median are the same**
- **If median and mean are different, indicates that the data are not normally distributed**
- **The mode is of little if any practical use**

The Statistic and The Parameter

- **A Statistic:**
 - It is a descriptive measure computed from the data of a sample (with lower case n = sample size).
- **A Parameter:**
 - It is a a descriptive measure computed from the data of a population (with upper case N =population size).
 - Since it is difficult to measure a parameter from the population, a sample is drawn of size n , whose values are $\chi_1, \chi_2, \dots, \chi_n$. From this data, we measure the statistic.

The Mean or Arithmetic Mean or Average

The number obtained by adding the values and dividing the total by the number of values

- **Notation**

- Σ (sigma) denotes the sum of a set of values
- x is the variable usually used to represent the individual data values
- n represents the number of values in a *sample*
- N represents the number of values in a *population*

- Mean of a set of sample values (read as x-bar)-formula

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

- Mean of all values in a population (read as “mu”)

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

Mean

- For the population mean of N values: $X_1, X_2, X_3, \dots, X_N$, the mean is calculated as:

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

- ❖ which is usually unknown, that's why we use the sample mean to estimate or approximate it.

- For the sample mean of n values (ungrouped or raw data) ($X_1, X_2, X_3, \dots, X_n$), the mean is calculated as:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Mean

Example:

Here is a sample of size 10 of ages, where

$$\chi_1 = 42, \chi_2 = 28, \chi_3 = 28, \chi_4 = 61, \chi_5 = 31, \chi_6 = 23, \chi_7 = 50, \chi_8 = 34, \chi_9 = 32, \\ \chi_{10} = 37.$$

Calculate the arithmetic mean?

$$\bar{x} = (42 + 28 + \dots + 37) / 10 = 36.6$$

Sample mean of grouped or classified data (data that is placed in intervals)

Unlike listed data, the individual values for grouped data are not available and you are not able to calculate their sum. So to calculate the mean of grouped data in frequency table, the following procedure is used:

1. Find the midpoint for each class interval (X_i , $i=1,2,\dots,K$)
2. For each class multiply the frequency with each midpoint ($f_i * X_i$).
3. Find the sum $\sum (f_i * X_i)$
4. Divide the sum in step-3 by the sum of frequencies ($n = \sum f_i$) to get the sample mean as follow:
5. $\text{Mean} = \sum (f_i * X_i) / n$

Worked Example

Calculating the Mean

In Tim's school, there are 25 teachers. Each teacher travels to school every morning in his or her own car. The distribution of the driving times (in minutes) from home to school for the teachers is shown in the table below:

Driving Times (minutes)	Number of Teachers
0 to less than 10	3
10 to less than 20	10
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

The driving times are given for all 25 teachers, so the data is for a population. Calculate the mean of the driving times.

Solution

Worked Example

- **Step 1:** Determine the midpoint for each interval.
- For 0 to less than 10, the midpoint is 5.
- For 10 to less than 20, the midpoint is 15.
- For 20 to less than 30, the midpoint is 25.
- For 30 to less than 40, the midpoint is 35.
- For 40 to less than 50, the midpoint is 45.
- **Step 2:** Multiply each midpoint by the frequency for the class.
- For 0 to less than 10, $(5)(3)=15$
- For 10 to less than 20, $(15)(10)=150$
- For 20 to less than 30, $(25)(6)=150$
- For 30 to less than 40, $(35)(4)=140$
- For 40 to less than 50, $(45)(2)=90$
- **Step 3:** Add the results from Step 2 and divide the sum by 25.
- $15+150+150+140+90=545$
$$\mu=545/25=21.8$$
- Each teacher spends a mean time of 21.8 minutes driving from home to school each morning.

Worked Example

To better represent the problem and its solution, a table can be drawn as follows:

Driving Times (minutes)	Number of Teachers f	Midpoint Of Class x	Product xf
0 to less than 10	3	5	15
10 to less than 20	10	15	150
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90

For the population, $N=25$ and $\sum xf=545$, so using the formula $\mu=\sum xf/N$, the mean would again be $\mu=545/25=21.8$

Properties of the Mean

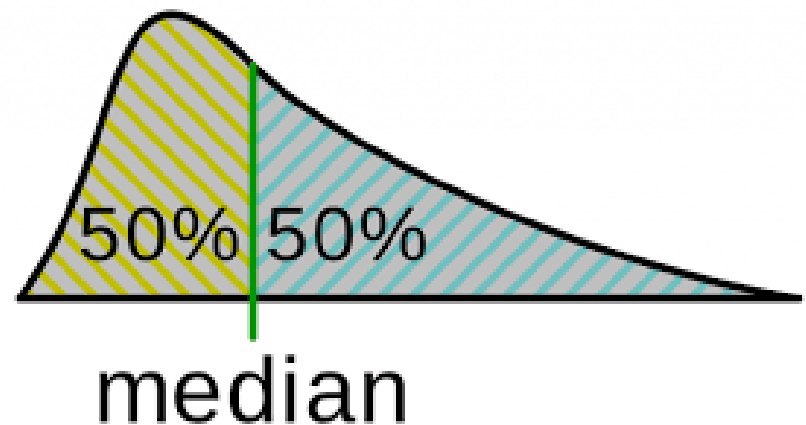
- In general, mean is the most important of all numerical measurements used to describe data
 - Uniqueness. For a given set of data there is one and only one mean.
 - Is used in computing other statistics, such as variance
 - Simplicity. It is easy to understand and to compute.
 - Is affected (not robust) by extremely high or low values (outliers) and may not be the appropriate average to use in those situations since all values enter into the computation.

➤ *Example:* Assume the values are 115, 110, 119, 117, 121 and 126.
The mean = 118.

But assume that the values are 75, 75, 80, 80 and 280. The mean = 118, a value that is not representative of the set of data as a whole.

Median (MD)

- When **ordering** the data, it is the observation that divide the set of observations into **two equal parts** such that half of the data are before it and the other are after it (in other words, the number of values greater than the median is equal to number of values is less than the median).
- The median either will be a specific value in the data set or will fall between two values.



The Median of Ungrouped or (Raw) Data

- If n is **odd**, the median will be the middle of observations. It will be the $(n+1)/2$ th ordered observation.
- ✓ E.g. when $n = 11$, then the median is the 6th observation.
- If n is **even**, there are two middle observations. The median will be the mean of these two middle observations. It will be the $(n+1)/2$ th ordered observation.
- ✓ When $n = 12$, then the median is the 6.5th observation, which is an observation halfway between the 6th and 7th ordered observation.

Example:

- For this random sample, the ordered observations will be as:
23, 28, 28, 31, 32, 34, 37, 42, 50, 61.
Since $n = 10$, then the median is the 5.5th observation,
i.e. $= (32+34)/2 = 33$.

Properties of the Median:

- **Uniqueness:** For a given set of data there is one and only one median.
- **Simplicity.** It is easy to calculate.
- **Robust measure, it is not affected by extreme values (outliers)** as is the mean.

The Median of Grouped Data-Simple Frequency Table

- To calculate the median for a simple frequency table, the following procedure is used:

Step-1: find the cumulative frequency (cf)

Step-2: find the value of $(n/2)$, where n is the total number of observations (the sum of frequencies)

Step-3: the median (MD) is the first value having cumulative frequency (cf) greater than or equal to $(n/2)$

The Median of Grouped Data-Simple Frequency Table

Example

The Gulf Pharmaceutical Industries Company in UAE planning to improve safety plan in its factory. For this, accident data for the last 50 weeks was compiled. These data were listed into the simple frequency table as shown below:

Number of Accidents (x)	2	7	12	17	22
Number of Weeks (f)	5	22	13	8	2

Calculate the median (MD) for the number of accidents per week?

Solution

Step-1: We calculate the value of cumulative frequency (cf) as follows:

No. of Accidents (x)	Frequency, f	cf
2	5	5
7	22	27
12	13	40
17	8	48
22	2	50
Total	50	

Step-2: The value of:

$$n / 2 = 50 / 2 = 25$$

Should be
Equals

Step-3: The median (MD) is the first value having cumulative frequency (cf) greater than or equal to 25, then from the table, the value of the median is MD = 7.



Median (MD) = 7 accidents per week.

The Median of Grouped Data-Frequency Table (Distribution)

- To calculate the median for a grouped frequency table (frequency distribution), the following procedure is used:
- Step-1: construct the cumulative frequency distribution.
- Step-2: decide the class that contain the median (class median).
- Class median: is the first class having cumulative frequency (cf) greater than or equal to $(n/2)$.
- Step-3: find the median (MD) by using the following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

where:

n = the total frequency.

F = the cumulative frequency before class median.

f_m = the frequency of the class median.

i = the class width.

L_m = the lower boundary (limit) of the class median.

The Median of Grouped Data-Frequency Table (Distribution)

- **Example:**

The following frequency table represents the time in minutes of 50 pharmacists to their work selected from the records of a given Pharmaceutical Industries Company in Jordan:

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7



Calculate the median (MD) for the time traveled to the work for pharmacists of this Pharmaceutical Industries Company ?

Solution

1st Step: Construct the cumulative frequency distribution for the data as follows:

Time to Travel to work Class	Frequency (f)	Cumulative Frequency (cf)
1 - 10	8	8
11 - 20	14	22
21 - 30	12	34
31 - 40	9	43
41 - 50	7	50

2nd Step: We find the class median as follows:

$$\frac{n}{2} = \frac{50}{2} = 25 \longrightarrow \text{class median is the 3rd class}$$

**Class
Median**

So, $F = 22$, $f_m = 12$, $L_m = 20.5$ and $i = 10$

3rd Step: The median (MD) is:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Conclusion

Thus, 25 persons take less than 23 minutes to travel to work and another 25 persons take more than 23 minutes to travel to work.

$$= 20.5 + \left(\frac{25 - 22}{12} \right) 10$$

$$= 23$$

The Mode

- Mode is the value which occurs most frequently.
- If all values are different there is no mode.
- Sometimes, there are more than one mode.
- It can be used for quick estimation and for identifying the most common observation.
- It can be unimodal, bimodal, or multimodal
- Example:
For the same random sample, the value 28 is repeated two times, so it is the mode.
- **Properties of the Mode:**
 - Sometimes, it is not unique.
 - Simple
 - It is a robust measure, not affected by extreme values (outliers)
 - It may be used for describing qualitative data.
 - It can be calculated for both quantitative and qualitative data, while mean and median for the quantitative data only.

The Mode of Raw Data and Simple Frequency Table

Example

A survey on the Ministry of Health showed the following distribution for the number of tablets sold in May 2018 for five types of medications used to treat blood pressure:

Medicine Name	Number of Tablets Sold
Almor	632
Lasix	1425
Aldacton	878
Indicardin	95
Diovan	471

Solution

Since the category with the highest frequency is Lasix, then the mode for the number of tablets sold in May 2018 for the five types of medications used to treat blood pressure is the Lasix drug.

The Mode of Grouped Data (Frequency Table)

- For the grouped data in a frequency table, the modal class is the class with the largest (highest) frequency). To find the mode for grouped data, use the following formula:

$$\text{Mode} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

where

- L_{mo} = Lower boundary (limit) of the modal class.
- Δ_1 = The difference of frequency between modal class and class before it.
- Δ_2 = The difference of frequency between modal class and class after it.
- i = class width.

Example

The following frequency table represents the time in minutes of 50 pharmacists to their work selected from the records of a given Pharmaceutical Industries Company in Jordan:

Time to travel to work	Frequency
1 - 10	8
11 - 20	14
21 - 30	12
31 - 40	9
41 - 50	7

Calculate the mode for the time traveled to the work for pharmacists of this Pharmaceutical Industries Company ?

Solution

The modal class is the interval 11 - 20 because it has the largest frequency, then we have:

$$L_{mo} = 10.5, \quad \Delta_1 = (14 - 8) = 6, \quad \Delta_2 = (14 - 12) = 2 \quad \text{and} \quad i = 10$$

The Mode of Grouped Data (Frequency Table)

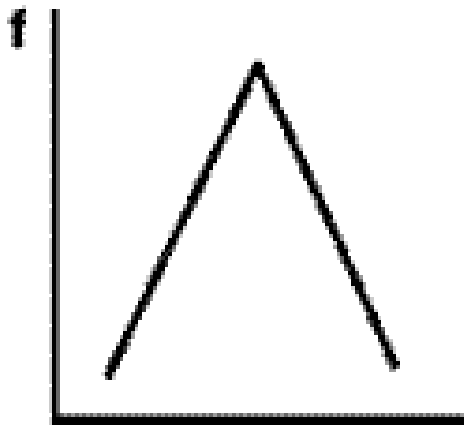
- Use the below formula to calculate the mode:

$$\text{Mode} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

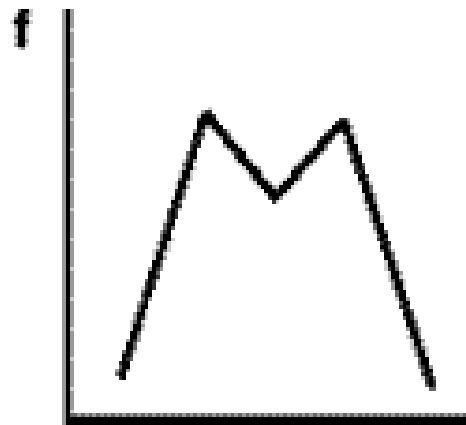
$$\text{Mode} = 10.5 + \left(\frac{6}{6 + 2} \right) 10 = 18 \text{ minutes}$$

Some Definitions

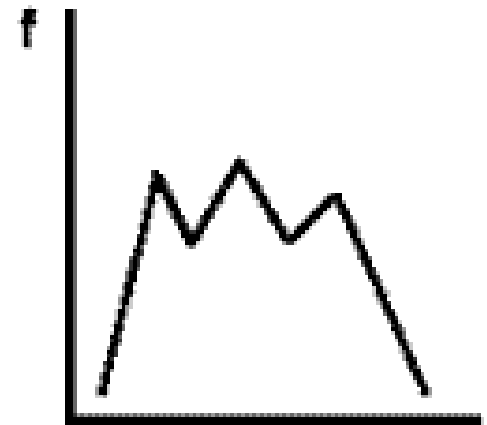
- **Distributions** – (of a variable) tells us what values the variable takes and how often it takes these values.
- Unimodal- having a single peak
- Bimodal-having two peak
- Symmetric-left and right half are mirror images



Unimodal



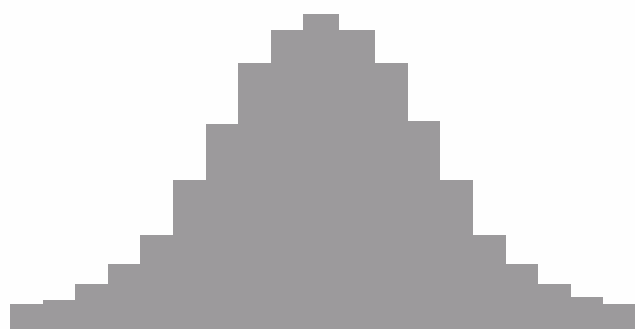
Bimodal



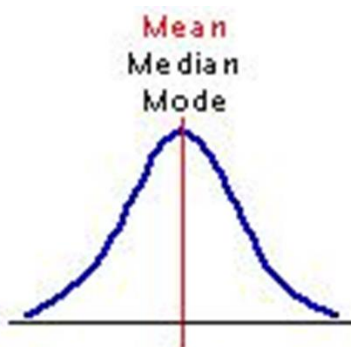
Multimodal

How Mean, Median, and Mode are related?

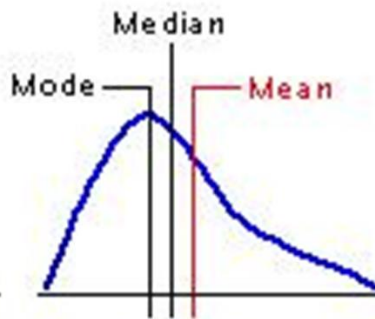
- A comparison of the mean, median, and mode can reveal information about the distribution shape
- A bell-shaped (normal) distribution is **symmetric**
 - Data values are evenly distributed on both sides of the mean
 - Unimodal (one peak)
 - Mean $\approx$ Median $\approx$ Mode



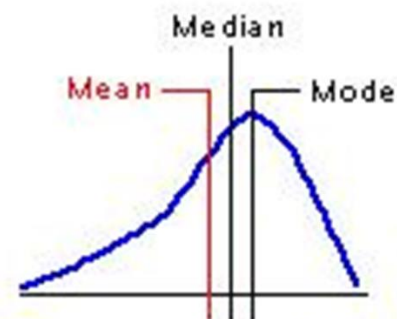
Normal distribution



Symmetrical
Distribution



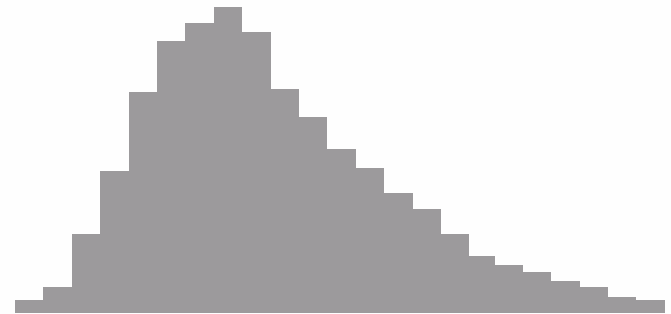
Positive
Skew



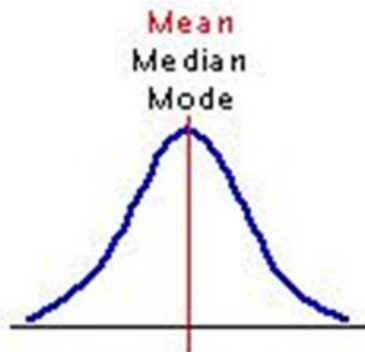
Negative
Skew

How Mean, Median, and Mode are related?

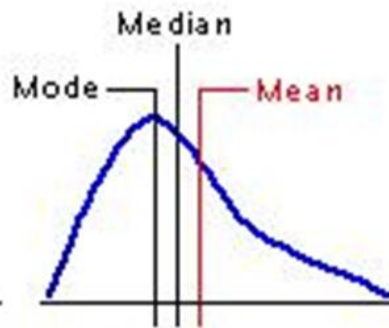
- Right-skewed (or positively) distribution has the majority of data values fall to the left of the mean and cluster at the lower end of the distribution; the “tail” is to the right
- $\text{Mode} < \text{Median} < \text{Mean}$
- Median is the “center” point



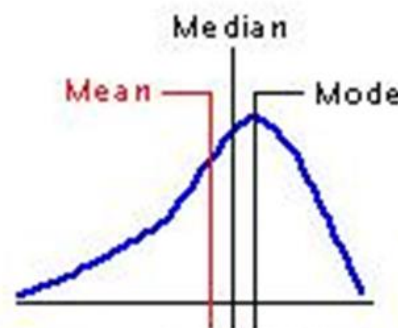
Right-skewed distribution



Symmetrical
Distribution



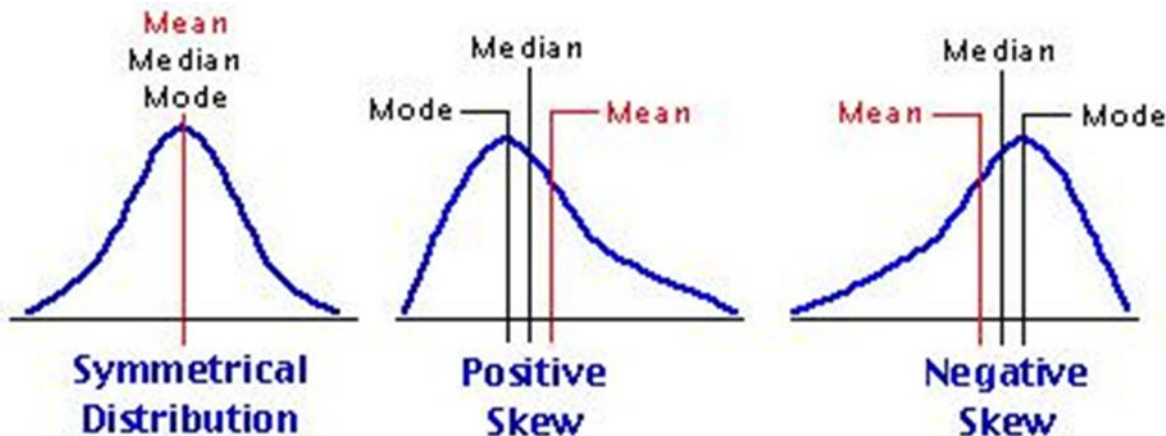
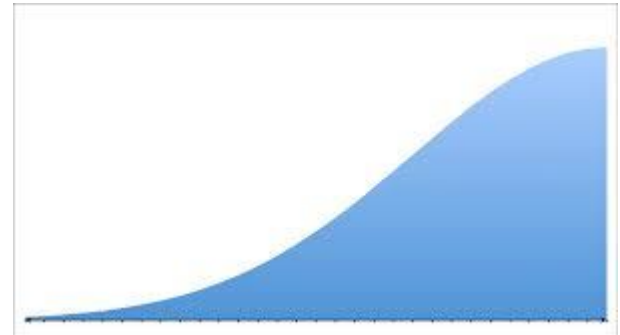
Positive
Skew



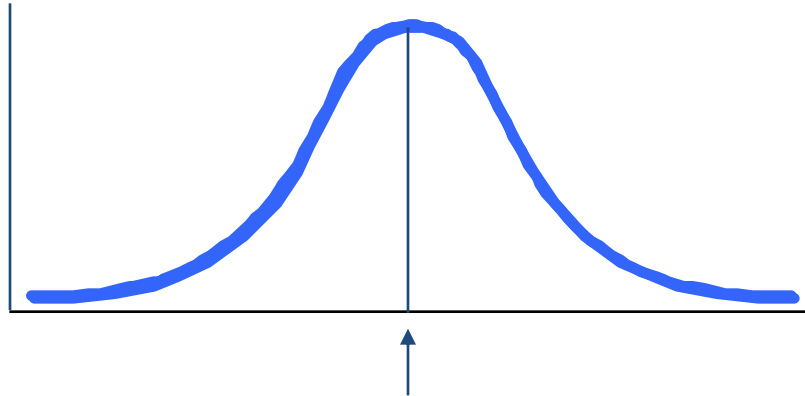
Negative
Skew

How Mean, Median, and Mode are related?

- Left-skewed (or negatively) distribution has the majority of data values to the right of the mean and cluster at the upper end of the distribution, with the tail to the left
- $\text{Mean} < \text{Median} < \text{Mode}$

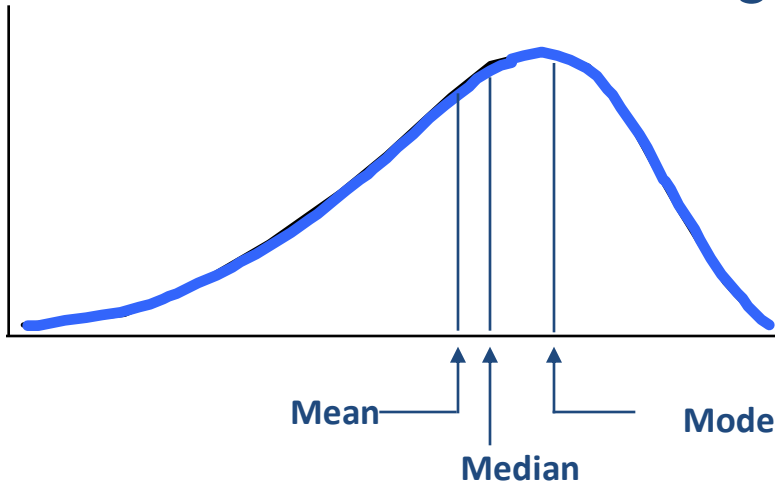


Skewness

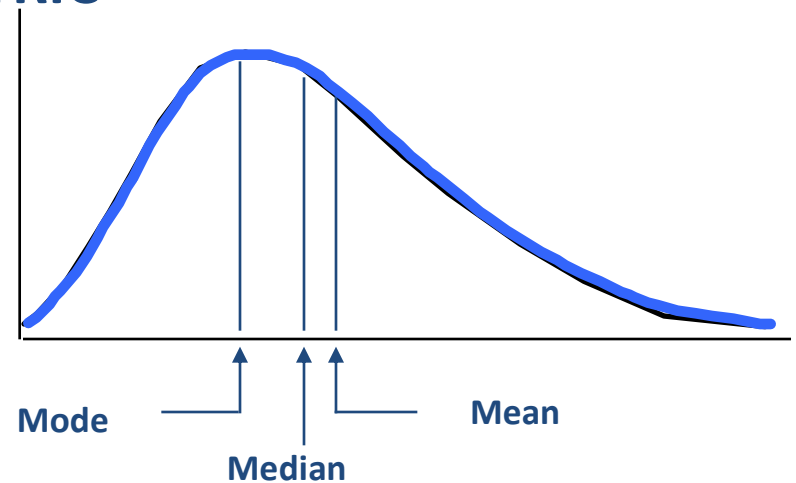


Mode = Mean = Median

SYMMETRIC



SKEWED LEFT
(negatively)



SKEWED RIGHT
(positively)

Which Measure of Central Tendency is best?

- There is no single best answer to that question because there are no objective criteria for determining the most representative measure for all data sets
- Avoid the term “average” , instead use the actual measure of central tendency that is calculated (mean, median, mode)
- Use the advantages and disadvantages stated above to decide which measure of central tendency is best.
- Large sample values tend to inflate the mean. This will happen if the histogram of the data is right-skewed.
- The median is not influenced by large sample values and is a better measure of centrality if the distribution is skewed.
- If $\text{mean} = \text{median} = \text{mode}$ then the data are said to be symmetrical