Measures of Central Tendency

- **What is a measure of central tendency?**
- **Measures of Central Tendency**
  - **Mode**
  - **Median**
  - **Mean**
- **Shape of the Distribution**
- **Considerations for Choosing an Appropriate Measure of Central Tendency**

What is a measure of Central Tendency?

- **Numbers that describe what is average or typical of the distribution**
- **You can think of this value as where the middle of a distribution lies.**

The Mode

- **The category or score with the largest frequency (or percentage) in the distribution.**

- **The mode can be calculated for variables with levels of measurement that are: *nominal, ordinal, or interval-ratio.***

The Mode: An Example

- **Example:** ***Number of Votes for Candidates for Mayor*. The mode, in this case, gives you the "central" response of the voters: the most popular candidate.**

Candidate A – 11,769 votes

Candidate B – 39,443 votes

Candidate C – 78,331 votes

The Mode:

"Candidate C"

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it.

- The median is the middle score, or average of middle scores in a distribution.

Median Exercise #1 (*N is odd*)

Calculate the median for this hypothetical distribution:

Job Satisfaction	Frequency
Very High	2
High	3
Moderate	5
Low	7
Very Low	4
TOTAL	21

Median Exercise #2 (*N is even*)

Calculate the median for this hypothetical distribution:

Satisfaction with Health	Frequency
Very High	5
High	7
Moderate	6
Low	7
Very Low	3
TOTAL	28

Finding the Median in Grouped Data

$$Median = L + \frac{N(.5) - Cf}{f} \times w$$

Percentiles

- **A score below which a specific percentage of the distribution falls.**
- **Finding percentiles in grouped data:**

$$25\% = L + \frac{N(.25) - Cf}{f} \times w$$

The Mean

- **The arithmetic average obtained by adding up all the scores and dividing by the total number of scores.**

Formula for the Mean

$$\bar{Y} = \frac{\sum Y}{N}$$

"Y bar" equals the sum of all the scores, Y, divided by the number of scores, N.

Calculating the mean with grouped scores

$$\bar{Y} = \frac{\sum fY}{N}$$

where: fY **= a score multiplied by its frequency**

Mean: Grouped Scores

Table 4.5 Hours Spent Watching Television

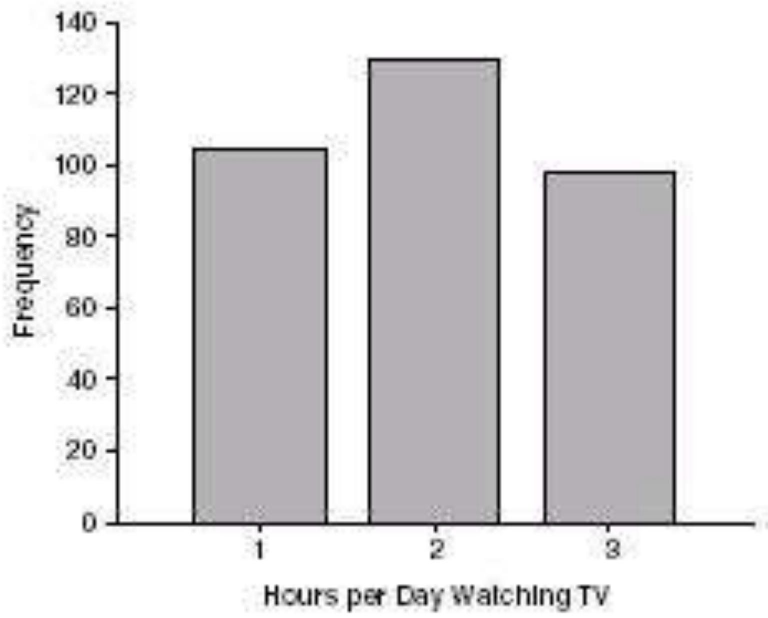
Hours Spent Watching TV	*Frequency (f)*	*fY*	*Percentage*	*C%*
1	104	104	31.3	31.3
2	130	260	39.2	70.5
3	98	294	29.5	100.0
Total	332	658	100.0	

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98$$

Median = 2.0
Mode = 2.0

Mean: Grouped Scores

Figure 4.7 Hours Spent Watching Television



Grouped Data: the Mean & Median

Calculate the median and mean for the grouped frequency below.

Number of People Age 18 or older living in a U.S. Household in 1996 (GSS 1996)

Number of People	Frequency
1	190
2	316
3	54
4	17
5	2
6	2
TOTAL	581

Shape of the Distribution

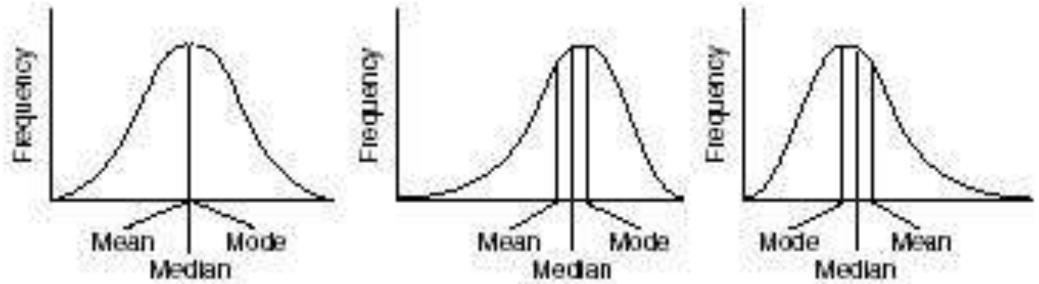
- **Symmetrical** (mean is about equal to median)
- **Skewed**
  - **Negatively** (example: years of education)

    mean < median
  - **Positively** (example: income)

    mean > median
- **Bimodal** (two distinct modes)
- **Multi-modal** (more than 2 distinct modes)

Distribution Shape

Figure 4.6 Types of Frequency Distributions



a. Symmetrical distribution

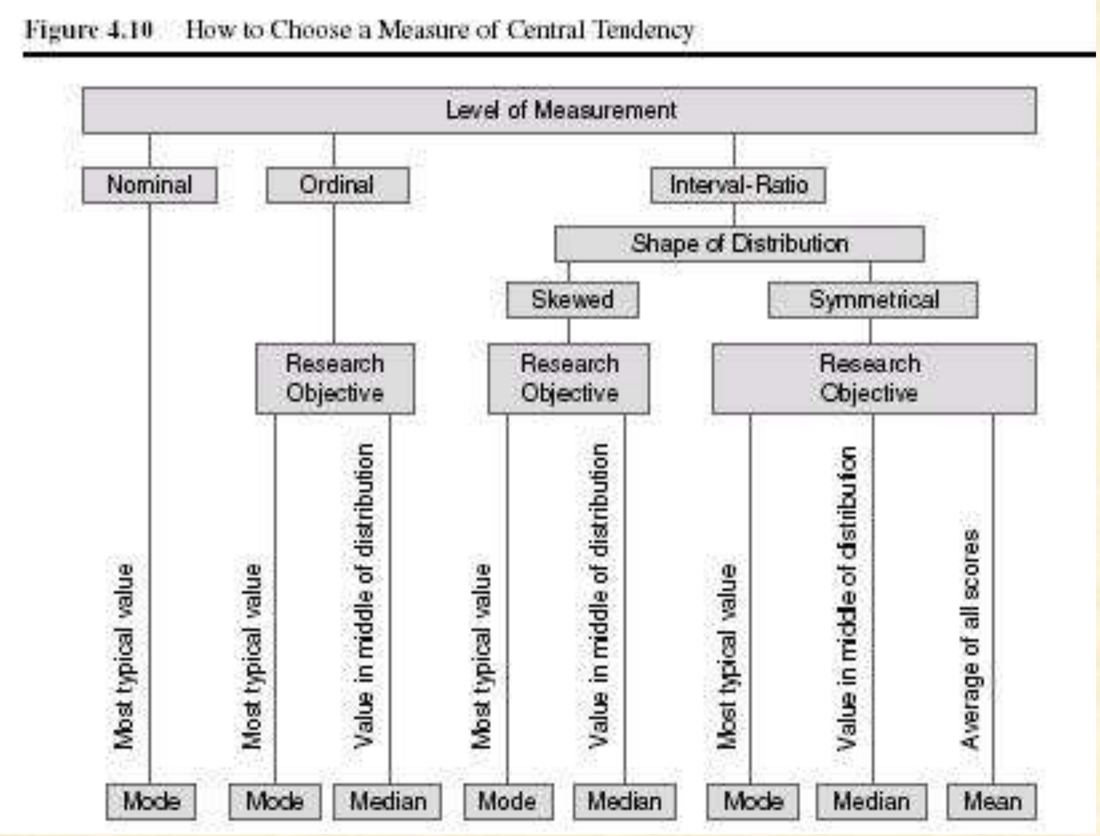
b. Negatively skewed distribution

c. Positively skewed distribution

Considerations for Choosing a Measure of Central Tendency

- **For a *nominal* variable, the mode is the only measure that can be used.**
- **For *ordinal* variables, the mode and the median may be used. The median provides more information (taking into account the ranking of categories.)**
- **For *interval-ratio* variables, the mode, median, and mean may all be calculated. The mean provides the most information about the distribution, but the median is preferred if the distribution is skewed.**

Central Tendency



Measures of Central Tendency

Introduction

- How well did my students do on the last test?
- What is the average price of gasoline in the Phoenix metropolitan area?
- What is the mean number of home runs hit in the National League?
- These questions are asking for a statistic that describes a large set of data.
- In this section we will study the *mean, median,* and *mode.*
- These three statistics describe an average or center of a distribution of numbers.

Sigma notation Σ

- The sigma notation is a shorthand notation used to sum up a large number of terms.
- $\Sigma x = x_1 + x_2 + x_3 + \ldots + x_n$
- One uses this notation because it is more convenient to write the sum in this fashion.

Definition of the mean

- Given a sample of n data points, x_1, x_2, x_3, … x_n, the formula for the mean or average is given below.

$$\bar{x} = \frac{\sum x}{n} = \frac{the \quad sum \quad of \quad the \quad data \quad pts}{the \quad number \quad data \quad pts}$$

Find the mean

- My 5 test scores for Calculus I are 95, 83, 92, 81, 75. What is the mean?
- ANSWER: sum up all the tests and divide by the total number of tests.
- Test mean = (95+83+92+81+75)/5 = 85.2

Example with a range of data

- When you are given a range of data, you need to find midpoints.
- To find a midpoint, sum the two endpoints on the range and divide by 2.
- Example $14 \le x < 18$. The midpoint $(14+18)/2=16$.
- The total number of students is 5,542,000.

Age of males	Number of students
$14 \le x < 18$	94,000
$18 \le x < 20$	1,551,000
$20 \le x < 22$	1,420,000
$22 \le x < 25$	1,091,000
$25 \le x < 30$	865,000
$30 \le x < 35$	521,000
Total	5,542,000

Continuing the previous example

- What we need to do is find the midpoints of the ranges and then multiply then by the frequency. So that we can compute the mean.
- The midpoints are 16, 19, 21, 23.5, 27.5, 32.5.
- The mean is [16(94,000)+19(1,551,000)+21(1,420,000)+ 23.5(1,091,000)+27.5(865,000)+32.5(521,000)] /5,542,000.=22.94

The median

- The **median** is the middle value of a distribution of data.
- How do you find the median?
- First, if possible or feasible, arrange the data from smallest value to largest value.
- The location of the median can be calculated using this formula: (n+1)/2.
- If (n+1)/2 is a whole number then that value gives the location. Just report the value of that location as the median.
- If (n+1)/2 is not a whole number then the first whole number less than the location value and the first whole number greater than the location value will be used to calculate the median. Take the data located at those 2 values and calculate the average, this is the median.

Find the median.

- Here are a bunch of 10 point quizzes from MAT117:
- 9, 6, 7, 10, 9, 4, 9, 2, 9, 10, 7, 7, 5, 6, 7
- As you can see there are 15 data points.
- Now arrange the data points in order from smallest to largest.
- 2, 4, 5, 6, 6, 7, 7, 7, 7, 9, 9, 9, 9, 10, 10
- Calculate the location of the median: (15+1)/2=8. The eighth piece of data is the median. Thus the median is 7.
- By the way what is the mean???? It's 7.13…

The mode

- The **mode** is the most frequent number in a collection of data.
- Example A: 3, 10, 8, 8, 7, 8, 10, 3, 3, 3
- The mode of the above example is 3, because 3 has a frequency of 4.
- Example B: 2, 5, 1, 5, 1, 2
- This example has no mode because 1, 2, and 5 have a frequency of 2.
- Example C: 5, 7, 9, 1, 7, 5, 0, 4
- This example has two modes 5 and 7. This is said to be bimodal.

Section 4.2 #13

- Find the mean, median, and mode of the following data:
- Mean = [3(10)+10(9)+9(8)+8(7)+10(6)+ 2(5)]/42 = 7.57
- Median: find the location (42+1)/2=21.5 Use the 21st and 22nd values in the data set.
- The 21st and 22nd values are 8 and 8. Thus the median is (8+8)/2=8.
- The modes are 6 and 9 since they have frequency 10.

Score	**Number of students**
10	3
9	10
8	9
7	8
6	10
5	2

Skewness, Moments & Kurtosis

SKEWNESS

A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other-to left or right.

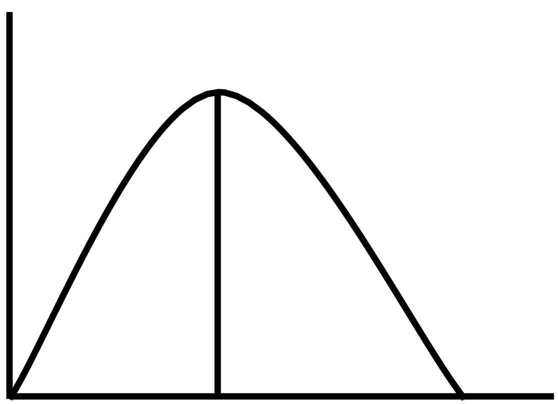
Measures of skewness tell us the direction and the extent of Skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness

SYMMETRICAL DISTRIBUTION

A frequency distribution is said to be symmetrical if the frequencies are equally distributed on both the sides of central value. A symmetrical distribution may be either bell – shaped or U shaped.

1- Bell – shaped or unimodel Symmetrical Distribution

A symmetrical distribution is bell – shaped if the frequencies are first steadily rise and then steadily fall. There is only one mode and the values of mean, median and mode are equal.



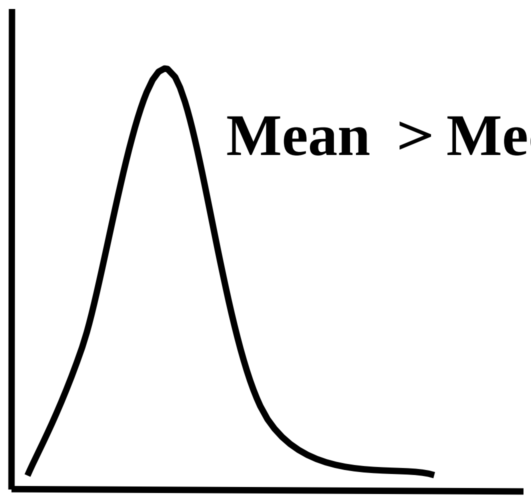
Mean = Median = Mode

SKEWED DISTRIBUTION

A frequency distribution is said to be skewed if the frequencies are not equally distributed on both the sides of the central value. A skewed distribution may be

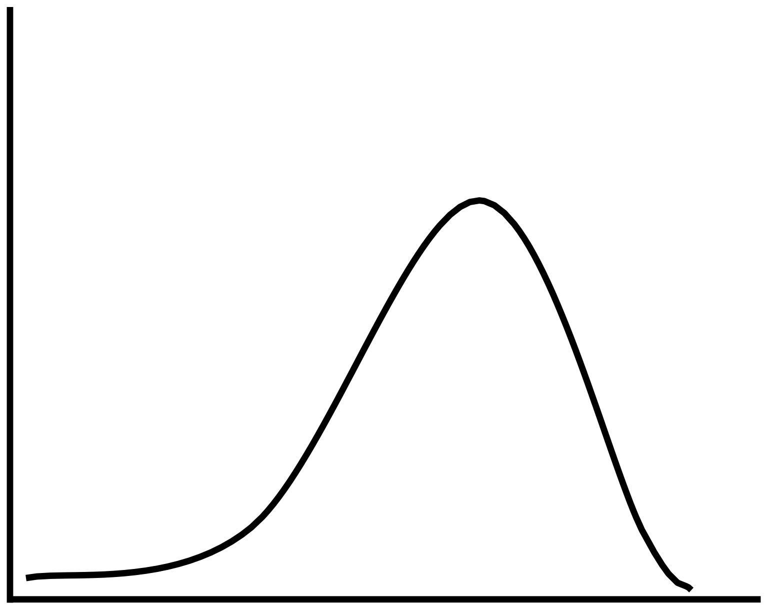
- Positively Skewed
- Negatively Skewed
- ‘L’ shaped positively skewed
- ‘J’ shaped negatively skewed

Positively skewed



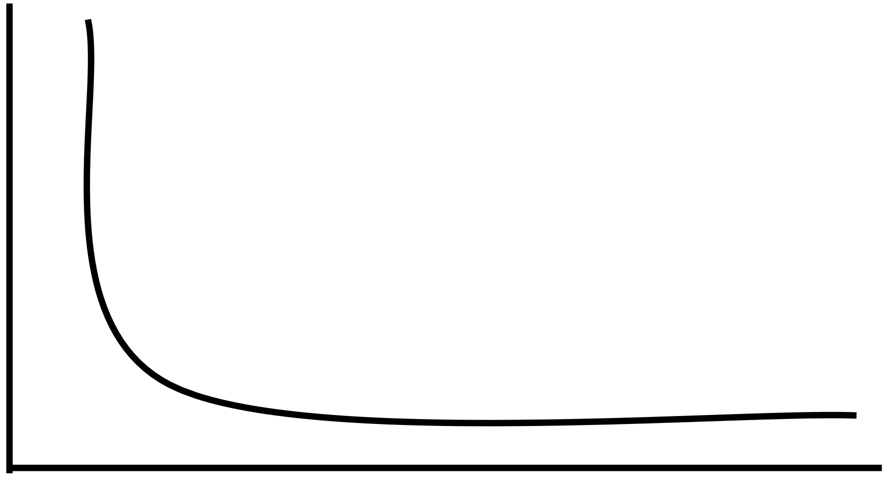
Mean > Median > Mode

Negatively skewed



Mean < Median < Mode

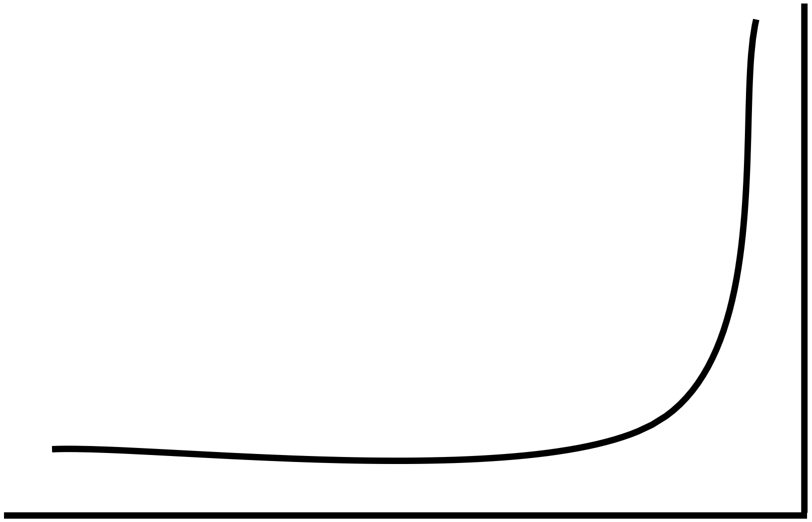
‘L’ Shaped Positively skewed



Mean < Mode

Mean < Median

‘J’ Shaped Negatively Skewed



Mean > Mode

Mean > Median

TESTS OF SKEWNESS

In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

- The values of mean, median and mode do not coincide.
- When the data are plotted on a graph they do not give the normal bell shaped form i.e. when cut along a vertical line through the centre the two halves are not equal.
- The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
- Quartiles are not equidistant from the median.
- Frequencies are not equally distributed at points of equal deviation from the mode.

MEASURES OF SKEWNESS

There are four measures of skewness. The measures of skewness are:

- Karl Pearson's Coefficient of skewness
- Bowley’s Coefficient of skewness
- Kelly’s Coefficient of skewness

Karl Pearson's Coefficient of skewness:-

The formula for measuring skewness as given by Karl Pearson is as follows

$$SK_P = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

Where,

SK_P = Karl Pearson's Coefficient of skewness,

σ = standard deviation

In case the mode is indeterminate, the coefficient of skewness is:

$$SK_P = \frac{\text{Mean} - (3\text{ Median} - 2\text{ Mean})}{\sigma}$$

Now this formula is equal to

$$SK_P = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

The value of coefficient of skewness is zero, when the distribution is symmetrical.

Normally, this coefficient of skewness lies between +1. If the mean is greater than the mode, then the coefficient of skewness will be positive, otherwise negative.

Bowley's Coefficient of skewness

Bowley developed a measure of skewness, which is based on quartile values. The formula for measuring skewness is:

$$SK_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

Where,

SK_B = Bowley's Coefficient of skewness,

Q_1 = Quartile first

Q_2 = Quartile second

Q_3 = Quartile Third

The above formula can be converted to

$$SK_B = \frac{Q_3 + Q_1 - 2Median}{(Q_3 - Q_1)}$$

The value of coefficient of skewness is zero, if it is a symmetrical distribution.

if the value is greater than zero, it is positively skewed and if the value is less than zero, it is negatively skewed distribution.

Kelly's Coefficient of skewness:-

Kelly developed another measure of skewness, which is based on percentiles and deciles. The formula for measuring skewness is as follows:

Based on Percentile

$$SK_k = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

Where,

SK_K = Kelly's Coefficient of skewness,

P_{90} = Percentile Ninety.

P_{50} = Percentile Fifty.

P_{10} = Percentile Ten.

Based on Deciles

$$SK_k = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1}$$

Where,

SK_K = Kelly's Coefficient of skewness,

D_9 = Deciles Nine.

D_5 = Deciles Five.

D_1 = Deciles one.

MOMENTS

In mechanics, the term **moment** is used to denote the rotating effect of a force. In Statistics, it is used to indicate peculiarities of a frequency distribution. The utility of moments lies in the sense that they indicate different aspects of a given distribution. Thus, by using moments, we can measure the central tendency of a series, dispersion or variability, skewness and the peakedness of the curve. The moments about the actual arithmetic mean are denoted by μ. The first four moments about mean or central moments are as follows:

In case of ungrouped data

First moment $\mu_1 = \frac{1}{N} \Sigma (X_1 - \overline{X})$

Second moment $\mu_2 = \frac{1}{N} \Sigma (X_1 - \overline{X})^2$

Third moment $\mu_3 = \frac{1}{N} \Sigma (X_1 - \overline{X})^3$

Fourth moment $\mu_4 = \frac{1}{N} \Sigma (X_1 - \overline{X})^4$

In case of grouped data

First moment $\mu_1 = \frac{1}{N} \Sigma f (X_1 - \bar{X})$

Second moment $\mu_2 = \frac{1}{N} \Sigma f (X_1 - \bar{X})^2$

Third moment $\mu_3 = \frac{1}{N} \Sigma f (X_1 - \bar{X})^3$

Fourth moment $\mu_4 = \frac{1}{N} \Sigma f (X_1 - \bar{X})^4$

Two important constants calculated from μ_2, μ_3 and μ_4 are:-

- β_1 (read as beta one)
- β_2 (read as beta two)

β_1 (Beta one):- Beta one is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

β_1 is used to measures of skewness. In symmetrical distribution β_1 shall be zero. However, the coefficient β_1 as a measure of skewness has serious limitations. β_1 as a measure of skewness cannot tell us about the direction of skewness that is whether it is positive or negative.

This is for the simple reason that μ_3 being the sum of the cubes of the deviation from the mean may be positive or negative but μ_3^2 is always positive. Also μ2 being the variance is always positive. Hence $\beta_1 = \mu_3^2 / \mu_2^2$ is always positive.

This drawback is removed if we calculate karlpearson's Υ_1 (Pronounced as Gamma one). Υ_1 is defined as the square root of β_1 that is

$$\Upsilon_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^2}} = \frac{\mu_3}{\mu_2^{3/2}}$$

The sign of skewness would depend upon the value of μ_3. If μ_3 is positive we will have positive skewness and if μ_3 is negative we will have negative skewness.

β_2 (Beta two):-

Beta two measures Kurtosis and is defined as:

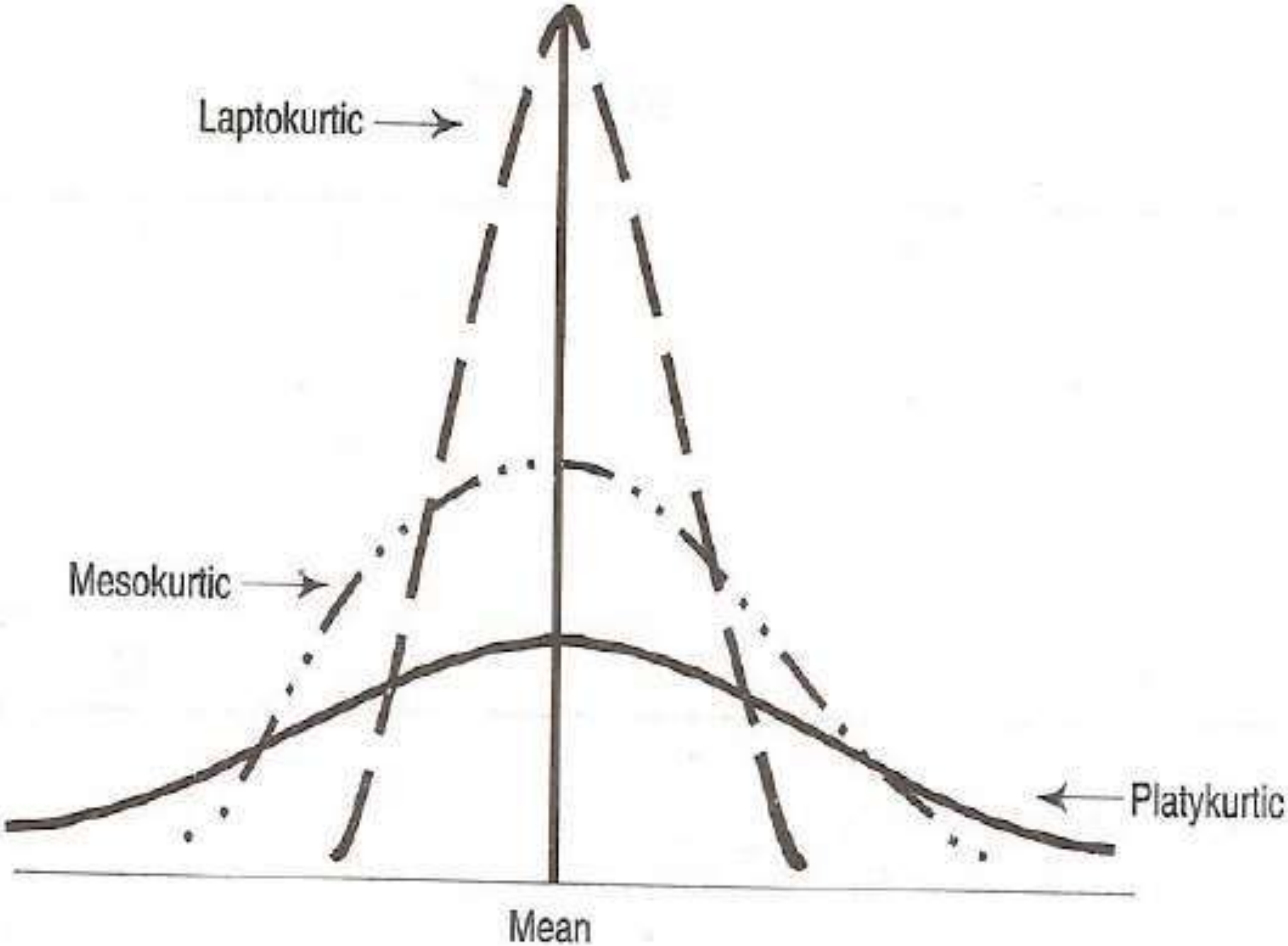
$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

KURTOSIS

Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess.

While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution. Karl Pearson classified curves into three types on the basis of the shape of their peaks.

These are Mesokurtic, leptokurtic and platykurtic. These three types of curves are shown in figure below:



Laptokurtic
Mesokurtic
Platykurtic
Mean

Measures of Kurtosis

Kurtosis is measured by β_2, or its derivative Υ_2

Beta two measures Kurtosis and is defined as:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

And

$$\Upsilon_2 = \beta_2 - 3$$

- In case of a normal distribution, that is, mesokurtic curve, the value of β_2 = 3.

- If β_2 turn out to be greater than 3, the curve is called a leptokurtic curve and is more peaked than the normal curve.

- When β_2 is less than 3, the curve is called a platykurtic curve and is less peaked than the normal curve.

- The measure of kurtosis is very helpful in the selection of an appropriate average. For example, for normal distribution, mean is most appropriate; for a leptokurtic distribution, median is most appropriate; and for platykurtic distribution, the quartile range is most appropriate