Measures Of Central Tendency, Variance

Module 12 Topic 2

You will learn

- Measures of central tendency
  - Mean, mode, median
  - Variability
    - Variance
    - Sd
    - Se
    - Quartile, decile, etc
  - Distribution around a mean
    - Normal curve
    - Skewedness
    - Kurtosis

Measures of Central Tendency

- To measure :
  - (A) Centre of Distribution &
  - (B) How the observations are dispersed around that centre
- Mean, Mode, Median are the measures of the centre of the distribution
- Range, Variance & Standard Deviation measure the dispersion within the distribution

Mode

- It is simply the value of the variable with the greatest frequency of occurrence ; this is the simplest measure and generates the most “popular” value in the distribution
- Modal Class of a continuous variable is the class associated with highest frequency
- Distributions may be unimodal, bimodal or multimodal

Median

- The Sample median is the Centre Point for any distribution of scores ; the median divides the distribution into two equal parts (50th percentile)
- Though it is a better estimate than the mode, yet it is not truly representative of the EXTREME VALUES in the sample distribution

Arithmetic Mean

- The Sample Mean is what is commonly referred to as average. It is the weighted centre point of a distribution, and is computed by summing up all the observations or scores and dividing by the total number of observations

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

The advantage in using the mean over the other measures of central tendency is that it takes into consideration how far the scores or values spread apart and allows for extreme scores

Examples Mean (5)

- Consider the sample consisting of the nine observations 8, 1, 2, 9,3, *2, 8, 1, 2*
- Here n, the sample size, is *9;* X_1, the first observation, is 8; X_2, the second observation, is 1. Similarly, $X_3 = 2$. $X_4 = 9$, *and so on, with* $X_9 = 2$. *Then,*
- Mean$X$ = *(8 + 1 + 2 +9 + 3 + 2 + 8 + 1 + 2)/9*

= 36/9 = 4

Mode is 2

Median the 5th figure is also 2

Why Dispersion?

- An average alone does not tell the full story. It is hardly fully representative of a mass, unless we know the manner in which the individual items scatter around it .... Study of Dispersion is necessary, if we are to gauge how representative the average is

Why Geeta Drowned?

- Geeta wanted to visit her grandma who lived across the river Spiti, and there was no bridge or boat
- Geeta was told that the mean depth of the river was 4 feet, while she was 5’3” tall. Though she did not know how to swim, she knew that the water would just about reach her shoulders, she tried to wade across the river

Spiti is deep

- When the river Spiti was measured from one shore to another, the depth being recorded every yard, the depths were as follows:
  - 1', 3', 4', 5', 6', 7', 7', 6', 4', 2', 2', 1',
- The mean depth was 4', but the water went over Geeta's head as soon as she was over 4 yards from the bank

Dispersion

- In measuring dispersion, it is imperative to know the amount of variation (absolute measure) and the degree of variation (relative measure). In the former case we consider the range, mean deviation, standard deviation etc. In the latter case we consider the coefficient of range, the coefficient mean deviation, the coefficient of variation etc.

Standard Deviation

- Is a measure of how widely values are dispersed from the mean, or how closely they are flocked around the mean

  Formula:

$$SD = \sqrt{\frac{\Sigma d^2}{n-1}}$$

where *d* is difference between individual value and the mean

The depth of Spiti

Mean = 4.0

$\Sigma d^2 = 54$

$\frac{\Sigma d^2}{n} = 4.5$

$SD = \sqrt{\frac{\Sigma d^2}{n-1}} = 2.1213$

Importance of Dispersion

- Let us take the following three sets. Notice the dispersion; are the three samples identical?

Students	Group X	Group Y	Group Z
1	50	45	30
2	50	50	45
3	50	55	75
∴ mean	50	50	50

Calculating Dispersion

Method of limits:

- The range
- Inter-quartile range
- Percentile range

Method of Averages:

- Quartile deviation
- Mean deviation
- Standard Deviation and
- Other measures

Range

- In any statistical series, the difference between the largest and the smallest values is called as the range
- Thus Range (R) = L - S
- **Coefficient of Range:** The relative measure of the range. It is used in the comparative study of the dispersion co-efficient of Range = (L-S/L+S)

Exercises

- Find the range and the co-efficient of the range of the following items :

  110, 117, 129, 197, 190, 100, 100, 178, 255, 790

  Range = 100 to 790

  Coeff of range = 790-100/790+100

  = 690/890

  = 0.775

  If outlier is removed it is 255-100/255+100=0.44

- Just as the Median divides the distribution into TWO (50% of values on either side), the Quartiles (Q1, Q2, Q3, Q4) divide the distribution into FOUR such that 25% of the values lying in each quartile Q2 being the median or the 50th Percentile

P_1: represents the 1st Percentile and

P_{100} : represents the 100th Percentile

P_{50} : coincides with Q2 and the Median

Inter Quartile Range

- If we concentrate on two extreme values (as in the case of range), we don't get any idea about the scatter of the data within the range (i.e. the two extreme values). If we discard these two values the limited range thus available might be more informative. For this reason the concept of interquartile range is developed. It is the range which includes middle 50% of the distribution. Here 1/4 of the lower end and 1/4 of the upper end of the observations are excluded
- Q3 – Q1 : is the Quartile Range but the more in use is (Q3-Q1)/2 called the Semi Quartile Range

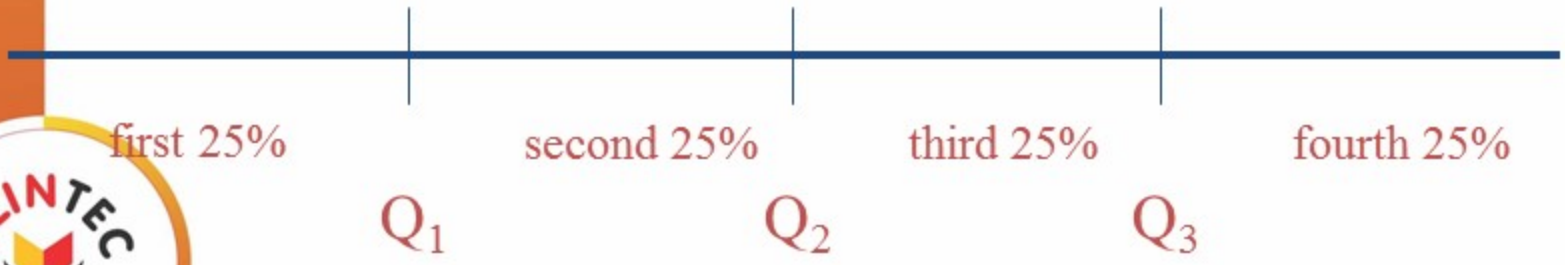
Median and quartiles

Sort the data in increasing order

The median is the middle value (if n is odd) or the average of the two middle values (if n is even), it is a measure of the "center" of the data

Quartiles: dividing the set of ordered values into 4 equal parts

Q_2 = second quartile = median



IQR = Interquartile range = $Q_3 - Q_1$

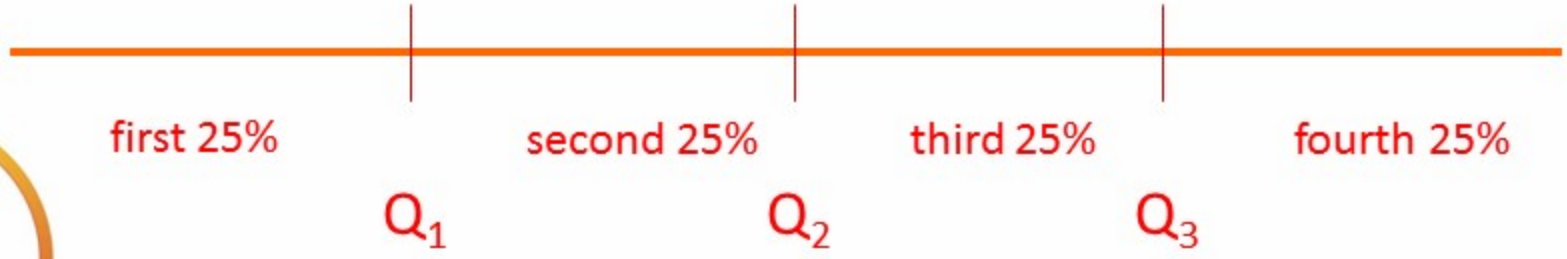
Sort the data in increasing order

Sort the data in increasing order

- The **median** is the middle value (if n is odd) or the average of the two middle values (if n is even), it is a measure of the "center" of the data

Quartiles: dividing the set of ordered values into 4 equal parts

Q_2 = second quartile = median



IQR = Interquartile range = $Q_3 - Q_1$

Percentile Range

- It is also used as one of the measure of dispersion. It is a set of data and is defined as = $P_{90} - P_{10}$ where P90 and P10 are 90^{th} & 10^{th} percentile respectively. Note that P25 is the same as Q1; P50 is the same as Median etc.

Standard Deviation and variance

- It is the most important measure of dispersion and is widely used in many statistical formulae
- Standard deviation is also called Root-Mean Square Deviation
- The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean
- It provides accurate result. Square of standard deviation is called Variance

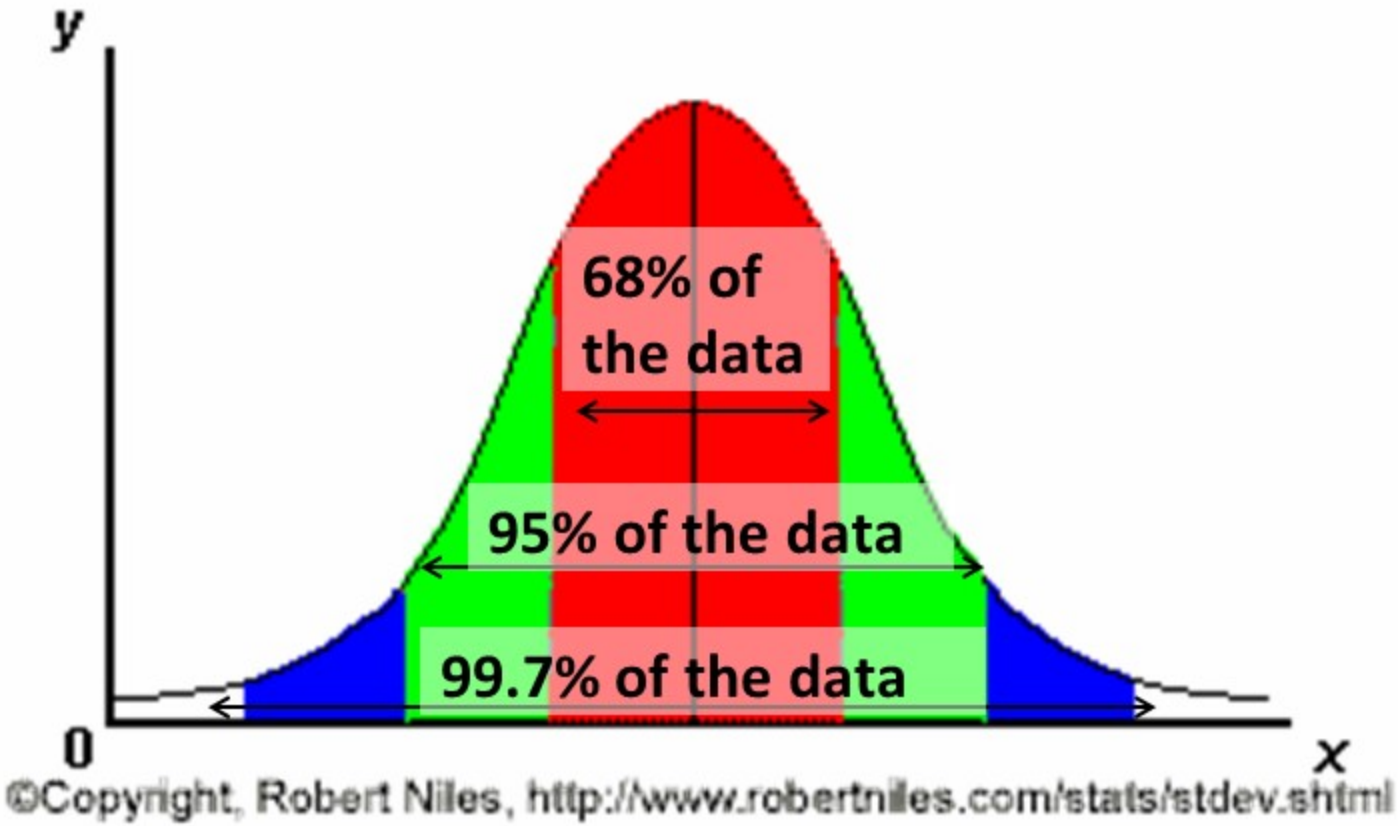
Standard Deviation – meaning? (6)

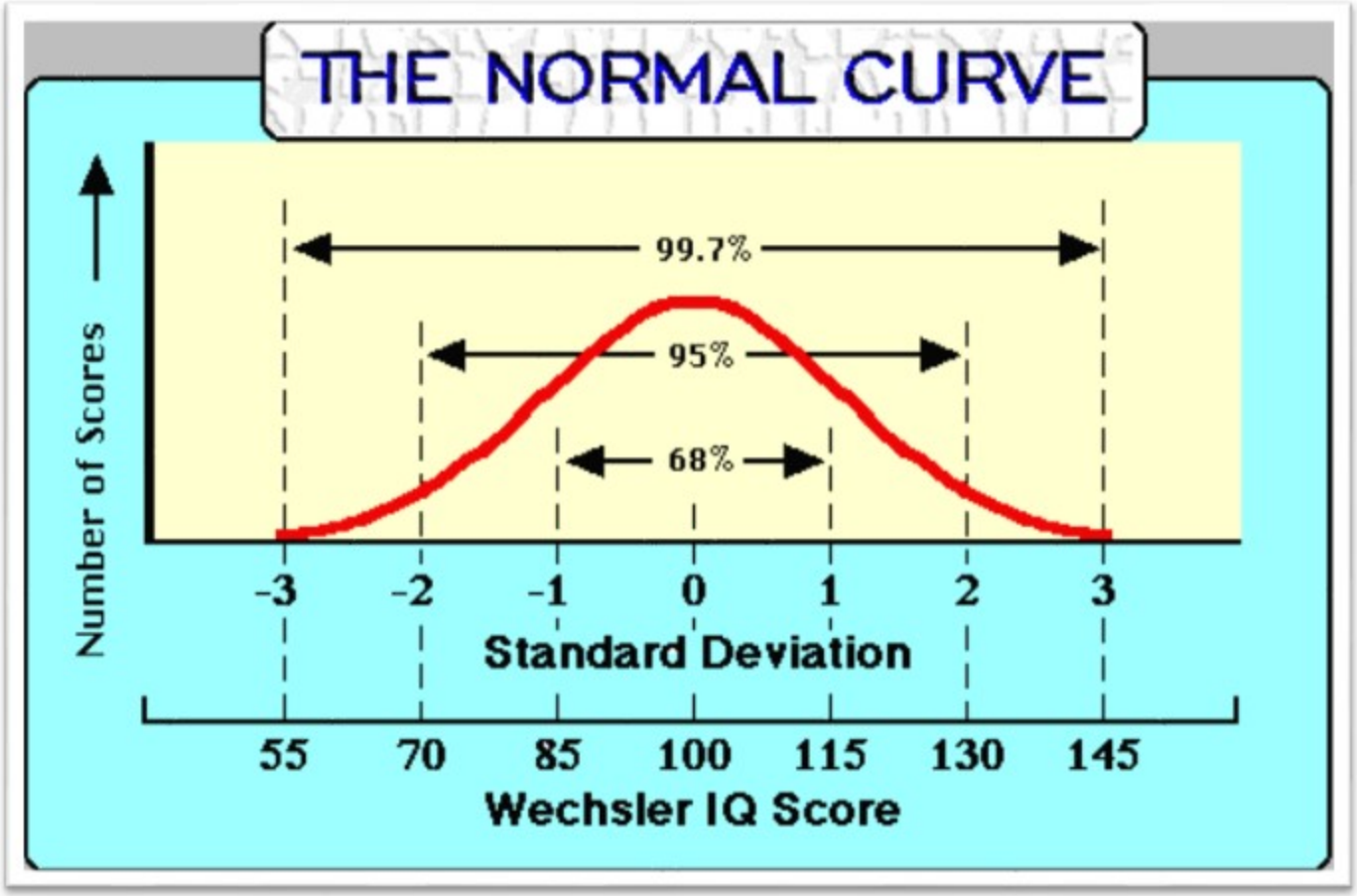
- SD indicates how much a set of values is spread around the average
- A range of one SD above and below the mean includes 68.4% of the data
- 2 SD around mean =95.4% of data and 3 SD around mean =99.7% of data

SD to check normal distribution(6)

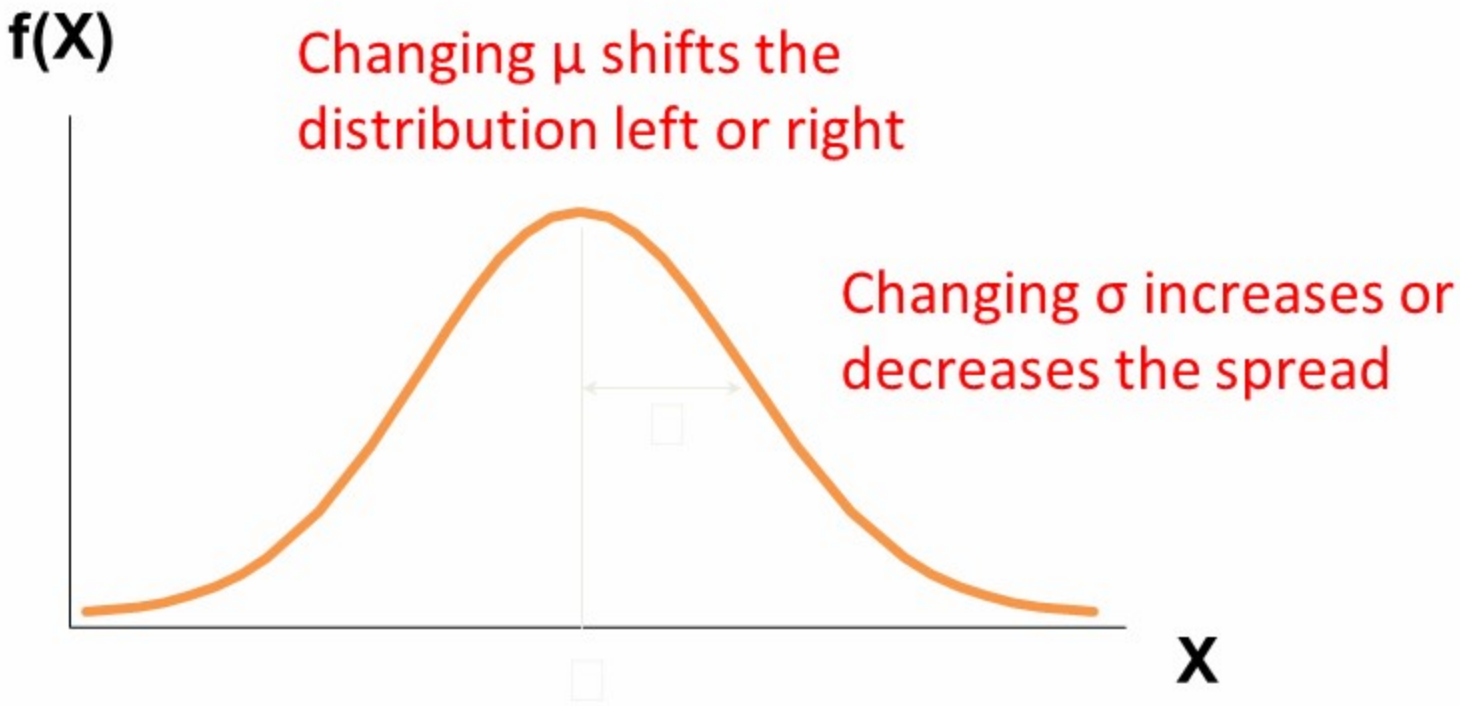
- A simple check for a normal distribution is to see if 2 SDs away from the mean are still within the possible range for the variable
- For example, if we have some length of hospital stay data with a mean stay of 10 days and a SD of 8 days then:
- mean – 2 × SD = 10 – 2 × 8 = 10 – 16 = -6 days
- This is clearly an impossible value for length of stay, so the data cannot be normally distributed

68-95-99.7 Rule




THE NORMAL CURVE
Number of Scores
99.7%
95%
68%
-3
-2
-1
0
1
2
3
Standard Deviation
55
70
85
100
115
130
145
Wechsler IQ Score

The Normal Distribution



Example

- Suppose SAT scores roughly follows a normal distribution in the U.S. population of college-bound students (with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:
  - 68% of students will have scores between 450 and 550
  - 95% will be between 400 and 600
  - 99.7% will be between 350 and 650

Measures of Symmetry

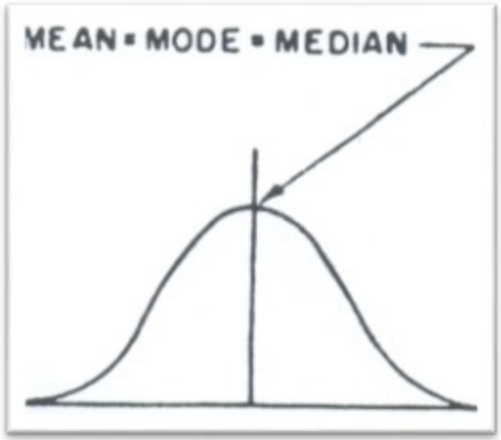
- Skewness
  - Symmetric distribution
  - Positively skewed distribution
  - Negatively skewed distribution

Skewness

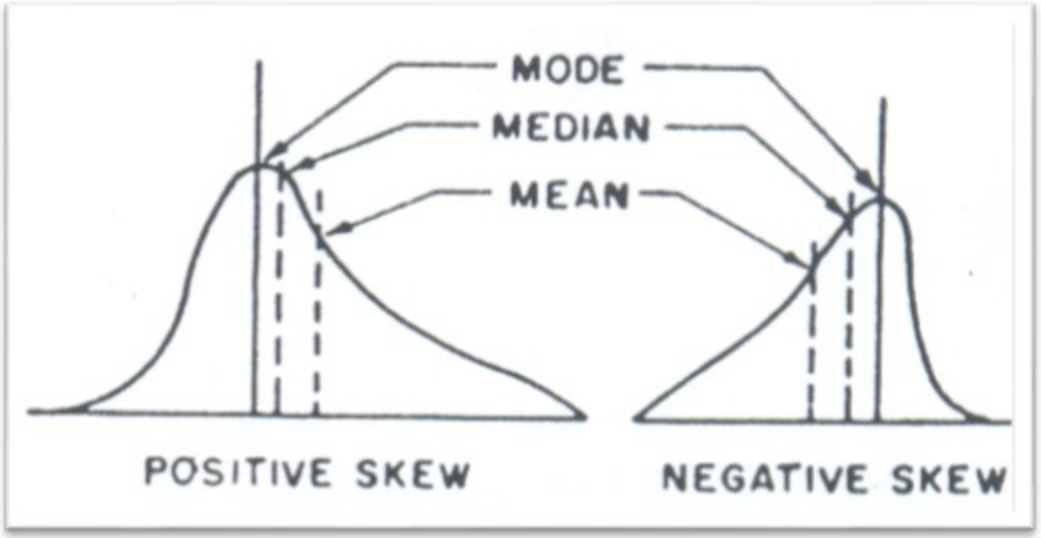
- The term skewness refers to the lack of symmetry. The lack of symmetry in a distribution is always determined with reference to a normal or Gaussian distribution. Note that a normal distribution is always symmetrical
- The skewness may be either positive or negative. When the skewness of a distribution is positive (negative), the distribution is called a positively (negatively) skewed distribution. Absence of skewness makes a distribution symmetrical
- It is important to emphasize that skewness of a distribution cannot be determined simply my inspection

Measures of Skewness

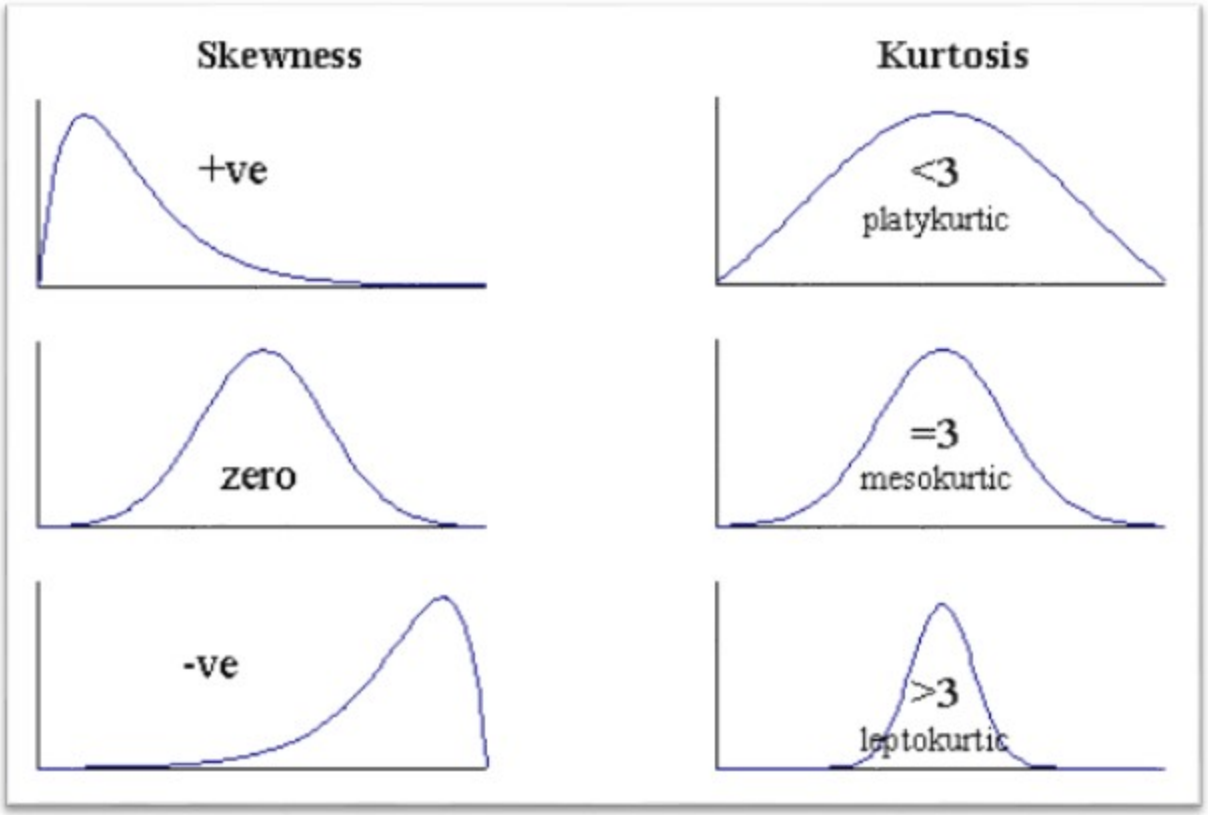
- If Mean > Mode, the skewness is positive.
- If Mean < Mode, the skewness is negative.
- If Mean = Mode, the skewness is zero.



Symmetric



Skewedness and Kurtosis

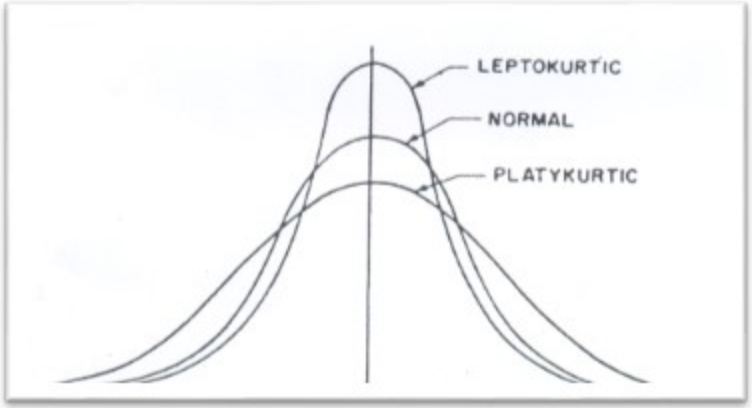


Kurtosis

- A curve having relatively higher peak than the normal curve, is known as **Leptokurtic**
- On the other hand, if the curve is more flat-topped than the normal curve, it is called **Platykurtic**
- A normal curve itself is called **Mesokurtic**, which is neither too peaked nor too flat-topped.

Measures of Peakedness

- Kurtosis:
  - is the degree of peakedness of a distribution, usually taken in relation to a normal distribution
  - Leptokurtic
  - Platykurtic
  - Mesokurtic



References

- Aggarwal YP, Statistical Methods, concepts, application, computation. Sterling publishers (1986)
- Varalakshmi V et al. Statistics – Higher Secondary, First Year. Tamil Nadu Textbook Corporation 2005
- Langley R, Practical Stats. The Chauser press 1968
- Barkan H, Annals of Cardiac Anaesthesia (2015)18:74
- Dunn OJ, Clark VA, Basic stats, a primer for biomed sciences. 4th Ed, 2009, John Weiley and sons inc.
- Harris M, Taylor G. Medical Statistics made easy. 1st Ed, Martin Dunitz, 2003
- Krousel wood M, clinicials guide to stats part 1 The Ochsner Journal (2006) 6:2,68
- KadamP , Sample size calculation. International J of Ayurveda (2010)1:55-57
- Fox N, Hunn A, Mathews N, Sampling and sample size calculation. The NIHR RDS EM/ YH (2009)