STATISTICS

Measures of Central Tendency

One of the main objectives of a statistical analysis is to obtain a value that describes the characteristics of the entire data. Such a value is called a central tendency. Measures of central tendency also facilitate the comparison between data by reducing the mass of data to one single value. Measures of Central Tendency are –

Mean

The most widely used measure of central tendency is the mean. Mean is computed by adding all the values of the variable and dividing that by the total number of values, i.e. the sum of all observations divided by the total number of observations.

For eg:- Find the mean of the even numbers between 1 to 11.

The even number between 1 to 11 are 2, 4, 6, 8 and 10.

The sum of these numbers is 30 and the total number of values is 5.

There the mean is,

Mean = 30/5

Mean = 6

Median

The median refers to the middle value in a distribution. In the case of median, one half of the values of the distribution are equal to or less than the median and the other half are equal to or greater than the median. It splits the observation into two halves. The score in the middle when the observations are ordered from the smallest to the largest. If the total number of observations n is an odd number, then the number on the position is the median. If n is an even number, then the average of the two numbers on the $n/2$ and $n/2 + 1$ positions is the median.

For eg:- Find the median of 5, 6, 11, 10, 4, 9, 7

4, 5, 6, 7, 9, 10, 11; Thus, Median = 7

Mode

The modal value or the mode is that value in a series of observations which occurs with the greatest frequency. It is the value that occurs most often in the data. If two numbers tie then the observation will have two modes and is called Bimodal

For eg:- Find the mode of 2, 6, 3, 9, 5, 6, 2, 6

2, 2, 3, 5, 6, 6, 6, 9; Thus, Mode = 6

Relationship between Mean, Median and Mode

Mode = 3 Median - 2 Mean **or**

Mean - Mode = 3 (Mean - Median)

Scales of Measurement:-

1. Nominal Scale – That can simply be broken down into categories
2. Ordinal Scale – That can be categorized and can be placed in order or ranking
3. Interval Scale – That can be ranked but has no absolute zero point
4. Ratio Scale – That allows to compare and has meaningful zero values

For Nominal scale, the mode is the only measure that can be used. For Ordinal Scale, the mode and the median may be used. For Interval – Ratio Scale, the mean, median and mode all can be used.

Partition Values

If the samples are arranged in ascending or descending order, then the measures of central tendency divides the observations in two equal parts. Similarly, there are other measures which divide a series into equal parts which are quartiles, deciles and percentiles.

Quartiles

Quartiles divides a series into 4 equal parts i.e. Q_1, Q_2 and Q_3. Q_1 is known as first or lower Quartile covering 25% observations. Q_2 is known as second Quartile is the same as Median of the series. Q_3 is known as third or upper Quartile covering 75% observations.

$$Q_1 = l + \frac{\left(\frac{n}{4} - cf\right)}{f} \times i$$

$$Q_3 = l + \frac{\left(\frac{3n}{4} - cf\right)}{f} \times i$$

Where,

l = lower limit of median class; i = class interval

cf = total of all frequencies before median class

f = frequency of median class; n = total number of observations

Deciles Similar to Quartiles, deciles divides a series into 10 equal parts i.e. D_1, D_2, D_3,.....D_{10}.

$$D_1 = l + \frac{\left(\frac{n}{10} - cf\right)}{f} \times i$$

$$D_2 = l + \frac{\left(\frac{2n}{10} - cf\right)}{f} \times i$$

$$D_3 = l + \frac{\left(\frac{3n}{10} - cf\right)}{f} \times i$$

And so on…..

Where,

l = lower limit of median class; i = class interval

cf = total of all frequencies before median class

f = frequency of median class; n = total number of observations

Percentiles

Percentiles divide a series into 100 equal parts i.e., P_1, P_2, P_3,…..P_{99}, P_{100} etc.

$$P_1 = l + \frac{\left(\frac{n}{100} - cf\right)}{f} \times i$$

$$P_2 = l + \frac{\left(\frac{2n}{100} - cf\right)}{f} \times i$$

$$P_{99} = l + \frac{\left(\frac{99\,n}{10} - cf\right)}{f} \times i$$

$$P_{100} = l + \frac{\left(\frac{100n}{100} - cf\right)}{f} \times i$$

Where,

l = lower limit of median class; i = class interval

cf = total of all frequencies before median class

f = frequency of median class; n = total number of observations

Measures of Dispersion

The measures of central tendency give us one single figure that represents the entire data. Ut the average alone cannot sufficiently describe the set of observations, unless all the observations are the same. It is necessary to see the how the data varies from the central value and how it is

scattered around it. It is necessary to describe the dispersion of the observations. Following are the important measures of dispersion –

Range

Range is the simplest method of studying dispersion. It is the difference between the value of the smallest item and the value of the largest item included in the distribution.

Range = Largest value- Smallest value

Range = $\frac{L-S}{L+S}$

Where,

L is the largest value

S is the smallest value

Iner-Quartile Range

Inter-Quartile Range is the difference between the third Quartile and the first Quartile. It is also known as the range of middle 50% values.

Inter-Quartile range = $Q_3 – Q_1$

Percentile Range

It the difference between the 90th and 10th percentile. It is also known as the range of middle 80% values.

Percentile range = $P_{90} – P_{10}$

Quartile Deviation

Quartile deviation gives the average amount by which two quartiles differ from the median. It is the average difference between the third Quartile and the first Quartile. It is an Absolute measure of dispersion.

Quartile Deviation = $\frac{Q_{3-}Q_1}{2}$

Coefficient of Quartile Deviation = $\frac{Q_{3-}Q_1}{Q_{3+}Q_1}$

Mean Deviation

Mean deviation, also known as the average deviation, is the average difference between the items in a distribution and the median or mean of that series. It is the mean of the deviations of the values from a fixed point.

Mean Absolute Deviation = $\frac{\sum_{i=1}^{n} | x\ i-\bar{x}}{n}$

Where,

n = Number of observations

$\bar{X}$= Mean

Standard Deviation

The standard deviation measures the absolute dispersion; the greater the standard deviation, the greater will be the magnitude of the deviation of values from their mean. It is defined as the square root of the mean of the squared deviations of individual values around their mean. If the values of the observations are same, then standard deviation is zero and it is least affected by fluctuations.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{i=N}|X_i-\bar{X}|^2}{N}}$$

Where,
σ = Standard Deviation
S2 = Variance
$\sum_{i=1}^{i=N}|X_i - \bar{X}|^2 =$ sum of the square of deviations from the mean
N = total number of observations

Measures of dispersion assist in the description of the width of the distribution, but they don't give any information about the shape of the distribution. There are further statistics that give information about the shape of the distribution. They are:

- 1st moment Mean (describes central value)

1st-moment = $\frac{\sum_{i=1}^{i=N}|X_i - \bar{X}|^1}{n}$, is equal to zero

- 2nd moment Variance (describes dispersion)

$\frac{\sum_{i=1}^{i=N}|X_i - \bar{X}|^2}{n}$, gives information on the spread or scale of the distribution of numbers

- 3rd moment Skewness (describes asymmetry)

$\frac{\sum_{i=1}^{i=N}|X_i - \bar{X}|^3}{n}$, gives information on the Skewness of the distribution

- 4th moment Kurtosis (describes peakedness)

$\frac{\sum_{i=1}^{i=N}|X_i - \bar{X}|^4}{n}$, gives information on the Kurtosis of the distribution

Skewness

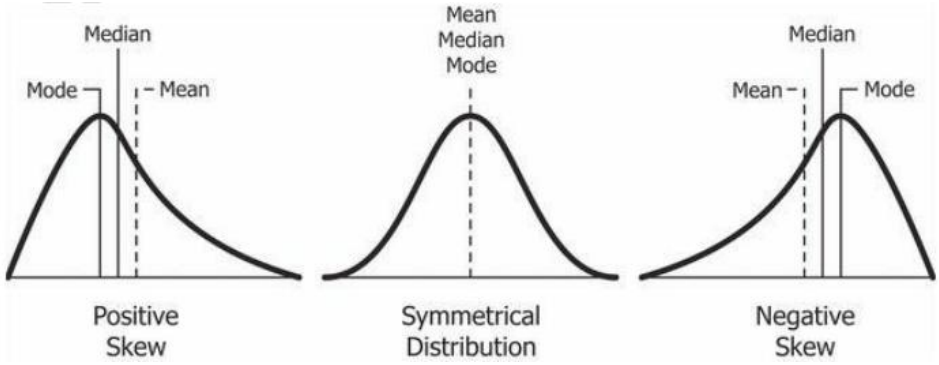
While measures of dispersion show us the variation of a set of values around a central value, skewness shows us the asymmetry of the data set. It tells us about the direction of variation of data. It is a measure of symmetry i.e. the values to the right and left of the central value.

A distribution can be positively skewed, negatively skewed, or there can be 0 skewness.

- **Symmetrical Distribution** - When the values of mean, median and mode are equal, there is no skewness. Such a distribution is called a symmetrical distribution. The spread of the frequencies is the same on both sides of the centre point of the curve.

- **Positively Skewed Distribution** - In a positively skewed distribution, the value of the mean is maximum and that of the mode is the least, while the median lies in between, i.e. Mean > Median > Mode.

- **Negatively Skewed Distribution** - In a negatively skewed distribution, the value of the mode is maximum and that of the mean is the least, while the median lies in between, i.e. Mode > Median > Mean.



Karl Pearson's Coefficient of Skewness = $\frac{Mean-Mode}{SD}$

If the value of Mode is not defined, the formula can be written as,

Pearson's Coefficient of Skewness = $\frac{3(Median-Mean)}{SD}$

It ranges between -3 to 3.

Bowley's Coefficient of Skewness = $\frac{Q_3-2M_d+Q_1}{Q_3-Q_1}$

Where,

M_d = Median

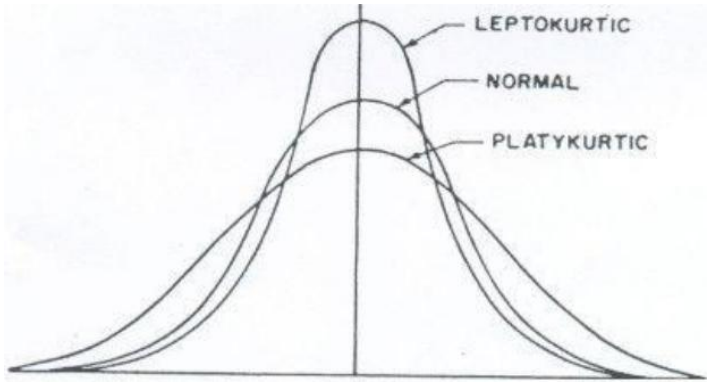
Kelly's Coefficient of Skewness = $\frac{P_{90}-2P_{50}+P_{10}}{P_{90}-P_{10}}$

Where,

P90 = 90th Percentile; P50 = 50th Percentile; P10 = 10th Percentile

Kurtosis

Kurtosis is the degree of peakedness of a distribution. It measures the relative peakedness or flatness of a distribution compared to the normal distribution. A tall thin distribution is Leptokurtic, a flat distribution is said to be Platykurtic and a normal distribution is called Mesokurtic.



- When Kurtosis > 0, the peak of a curve becomes relatively high and that curve is called Leptokurtic. The positive Kurtosis indicates a flat distribution with long tails
- When Kurtosis < 0, the curve is flat-topped, then it is called Platykurtic. The negative Kurtosis indicates a peaked distribution with short tails
- A normal curve is neither very peaked nor very flat-topped, so it is taken as a basis for comparison. The normal curve is called Mesokurtic. For a normal distribution, kurtosis is equal to 3.

The measure of Kurtosis, known as Percentile coefficient of kurtosis is:

Kurtosis = $\frac{QD}{Q_{90+}Q_{10}}$

Where,

Q.D is semi-interquartile range, Q.D = $\frac{Q_{3-}Q_1}{2}$

P_{90} = 90th Percentile;

P_{10} = 10th Percentile