

Task 3 – Data Science

Five Number Summary, Central limit Theorem and Correlation methods

Five Number Summary

- **The 5 number summary consists of the following 5 pieces of information:**

1. **Minimum** - The smallest value.

2. **First Quartile (Q1)** - The lower quartile (the value beneath which 25% of data values lie). 25th Percentile Rank

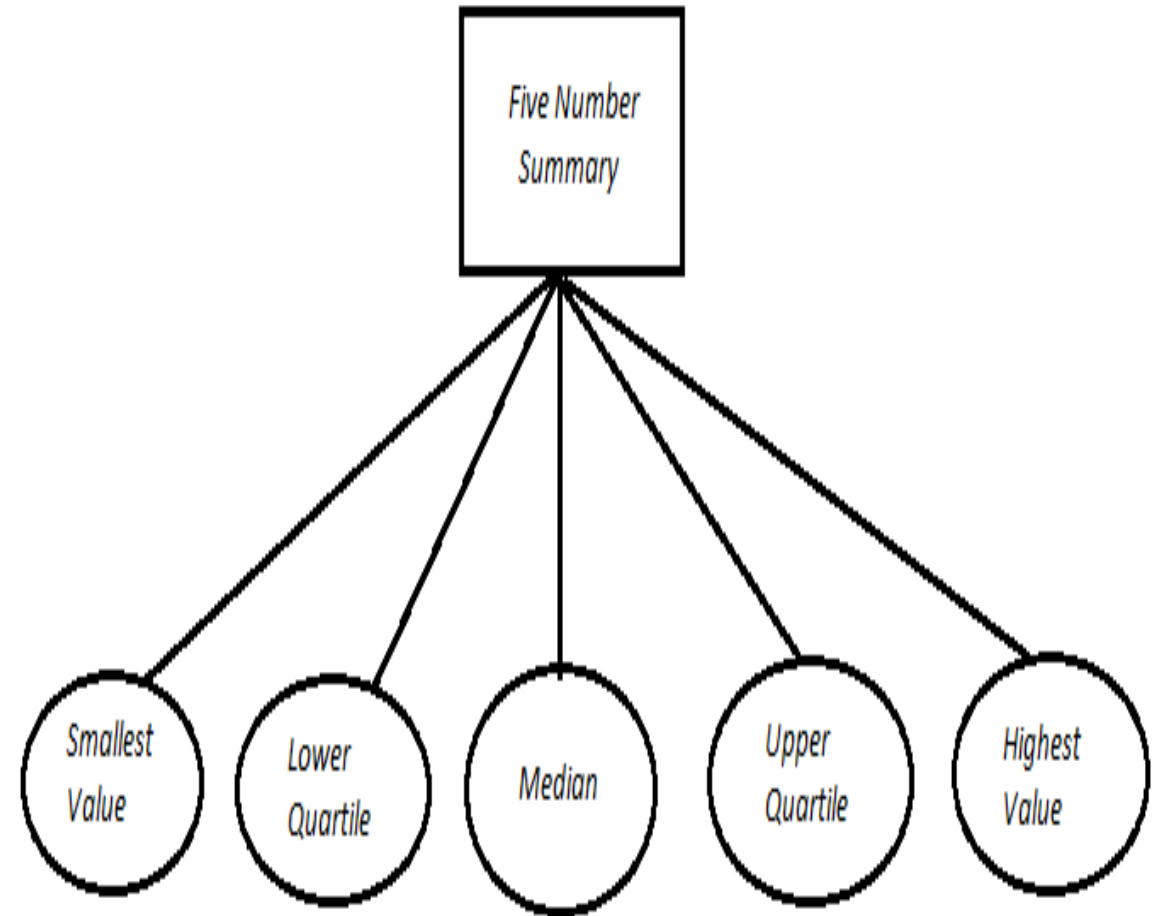
3. **The median** - the value beneath which 50% of the data values lie.

4. **Third Quartile (Q3)** - The lower quartile (the value beneath which 75% of data values lie). 75th Percentile Rank

5. **Maximum** - The highest value

Five Number Summary Definition and its advantage

- The five number summary for a given set of data is a set of 5 representative numerical values obtained from the data which give us an idea of how the data looks like and how it is distributed.
- The main advantage of the 5 number summary is that instead of tediously going through the entire data we can use it to get an overview of the data at just a glance.
- Some outliers in the data may also be included in the five number summary. This is because it contains the highest and lowest values which may turn out to be outliers.



How to calculate five number summary?

- Will find Five number summary for below set of data:
- 1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27
- From above set of data total no.of data sets: $n = 19$
- We have to find Minimum (Lowest value), First Quartile (Q1), Median, Third Quartile (Q3) and Maximum (Highest value)
- By seeing the set of data we know Minimum value is 1
- Now we have to find Lower Fence, Upper Fence and Inter Quartile Range – IQR (Q2)

Required Formulae's....

- $Q1$ (25th Percentile Rank) = $(25/100) * (n + 1) = (25/100) * (19 + 1) = 5$
- Here 5 is index position (index position starts with 1) in Set of data and 3 is the data point in the given set of data. So $Q1 = 3$
- $Q3$ (75th Percentile Rank) = $(75/100) * (n + 1) = (75/100) * (19 + 1) = 15$
- Here 15 is index position (index position starts with 1) in Set of data and 7 is the data point in the given set of data. So $Q3 = 7$
- IQR ($Q2$) = Third Quartile ($Q3$) – First Quartile ($Q1$) = $7 - 3 = 4$
- Lower Fence = $Q1 - 1.5(IQR) = 3 - 1.5(4) = -3$
- Upper Fence = $Q3 + 1.5(IQR) = 7 + 1.5(4) = 13$

Required Formulae's contd....

- Based on Lower Fence and Upper Fence values we came to know that data points in our set of data should be in between -3 and 13
- In our Data set 27 is outlier since 13 is the upper fence and data points falling above 13 are considered as outlier
- Similarly data points which fall below -3 will be considered as outlier
- Maximum value in data set is 9.
- So our Data set after removing outliers
1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9
- Median of above data set is : $5+5/2 = 5$

Calculated Five Numbers

Minimum	First Quartile (Q1)	Median	Third Quartile (Q3)	Maximum
1	3	5	7	9

CORRELATION Definition and Meaning..

Correlation Definition:-

- ✓ Correlation is a statistical measure that expresses the extent to which two variables are linearly related i.e. variables change together at constant rate.
- ✓ Correlation is commonly used for describing simple relationships without making a statement about cause and effect.
- ✓ For example there may exist a relationship between heights and weights of a group of students, the scores of students in two different subjects are expected to have an interdependence or relationship between them

□ Meaning of Correlation:

- ✓ To measure the degree of association or relationship between two variables quantitatively, an index of relationship is used and is termed as co-efficient of correlation.
- ✓ Co-efficient of correlation is a numerical index that tells us to what extent the two variables are related and to what extent the variations in one variable changes with the variations in the other.
- ✓ The co-efficient of correlation is always symbolized either by r or ρ (Rho).
- ✓ The notion ' r ' is known as product moment correlation co-efficient or Karl Pearson's Coefficient of Correlation.
- ✓ The symbol ' ρ ' (Rho) is known as Rank Difference Correlation coefficient or spearman's Rank Correlation Coefficient.

The coefficient of correlation is a number and not a percentage and Need for Correlation...

- The size of ' r ' indicates the amount (or degree or extent) of correlation-ship between two variables.
- If the correlation is positive the value of ' r ' is + ve and if the correlation is negative the value of V is negative
- Thus, the signs of the coefficient indicate the kind of relationship. The value of V varies from +1 to -1.
- Correlation can vary in between perfect positive correlation and perfect negative correlation.
- The top of the scale will indicate perfect positive correlation and it will begin from +1 and then it will pass through zero, indicating entire absence of correlation.
- The bottom of the scale will end at -1 and it will indicate perfect negative correlation.
- Thus numerical measurement of the correlation is provided by the scale which runs from +1 to -1.

Need for Correlation:

- ✓ Correlation gives meaning to a construct
- ✓ Correlational analysis is essential for basic psycho-educational research.
- ✓ Indeed most of the basic and applied psychological research is correlational in nature.

Correlational analysis is required for and Methods of Correlation..

- Finding characteristics of psychological and educational tests (reliability, validity, item analysis, etc.).
- Testing whether certain data is consistent with hypothesis.
- Predicting one variable on the basis of the knowledge of the other(s).
- Building psychological and educational models and theories.
- Grouping variables/measures for parsimonious interpretation of data.
- Carrying multivariate statistical tests (Hotelling's T^2 ; MANOVA, MANCOVA, Discriminant analysis, Factor Analysis).
- Isolating influence of variables.

Methods of Correlation:

In a bivariate distribution, the correlation may be:

1. Positive, Negative and Zero Correlation; and
2. Linear and Curvilinear (Non-linear).

Methods of Correlation Contd..


3. **Methods of Studying Correlation**

- ▶ **Scatter Diagram Method**
- ▶ **Graphic Method**
- ▶ **Karl Pearson's Coefficient of Correlation**
- ▶ **Method of Least Squares**



Types of Correlation Contd..

Types of Correlation Type II

- ▶ **Simple correlation:** Under simple correlation problem there are only two variables are studied.
 - ▶ **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied.
Ex. $Q_d = f (P, P_C, P_S, t, y)$
 - ▶ **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.
 - ▶ **Total correlation:** is based on all the relevant variables, which is normally not feasible.
- 

Positive, Negative or Zero Correlation..

1. Positive, Negative or Zero Correlation:

Positive Correlation:

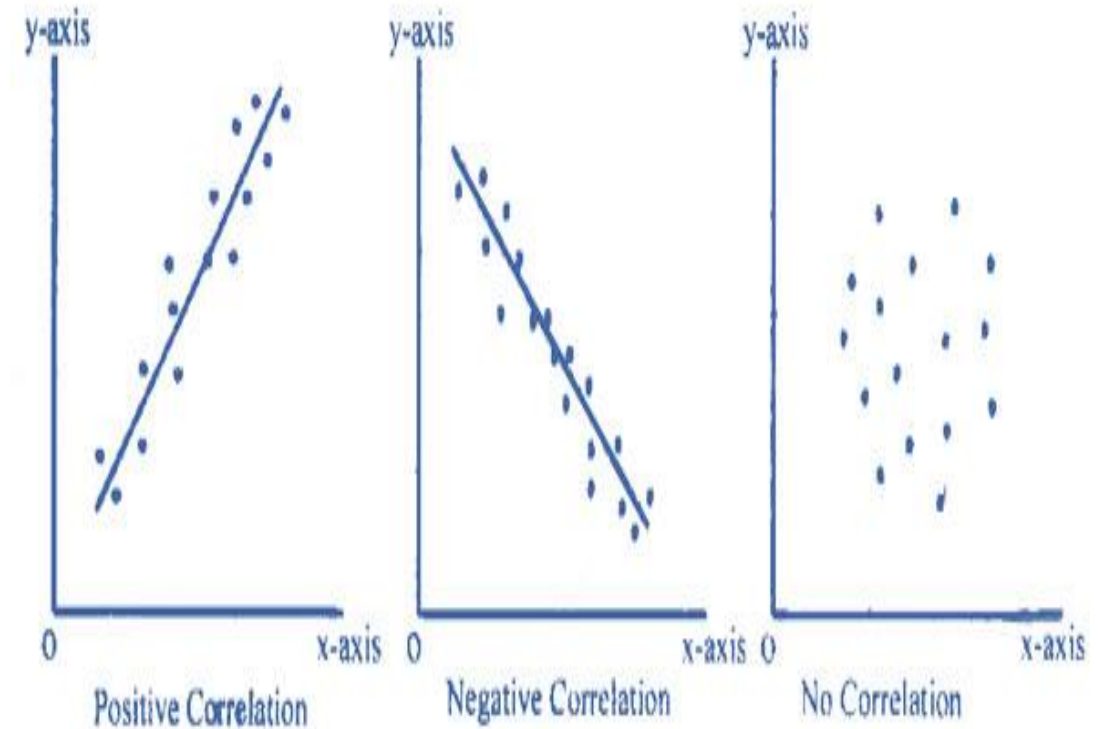
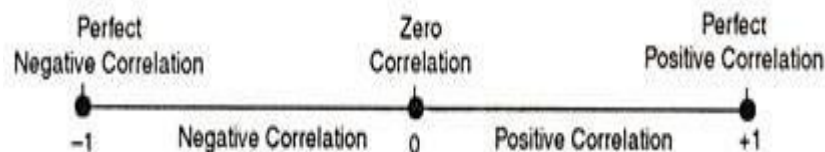
When the increase in one variable (X) is followed by a corresponding increase in the other variable (Y); the correlation is said to be positive correlation. The positive correlations range from 0 to +1; the upper limit i.e. +1 is the perfect positive coefficient of correlation.

Negative Correlation:

The perfect positive correlation specifies that, for every unit increase in one variable, there is proportional increase in the other. For example “Heat” and “Temperature” have a perfect positive correlation.

Zero Correlation:

If, on the other hand, the increase in one variable (X) results in a corresponding decrease in the other variable (Y), the correlation is said to be negative correlation.



Methods of Correlation Contd..

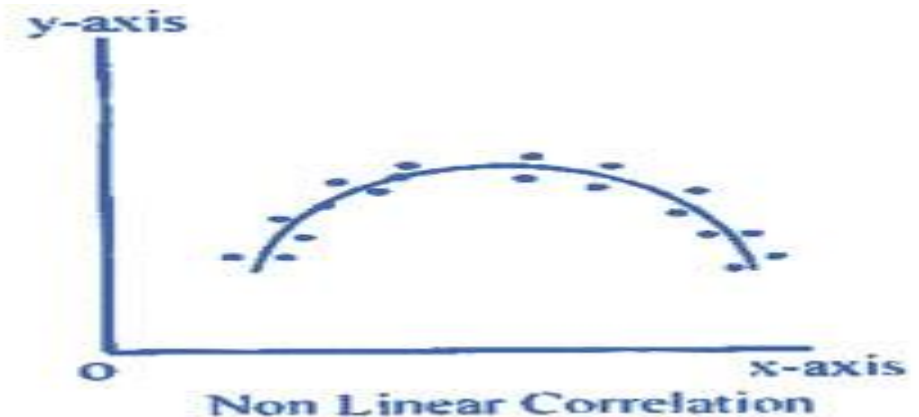
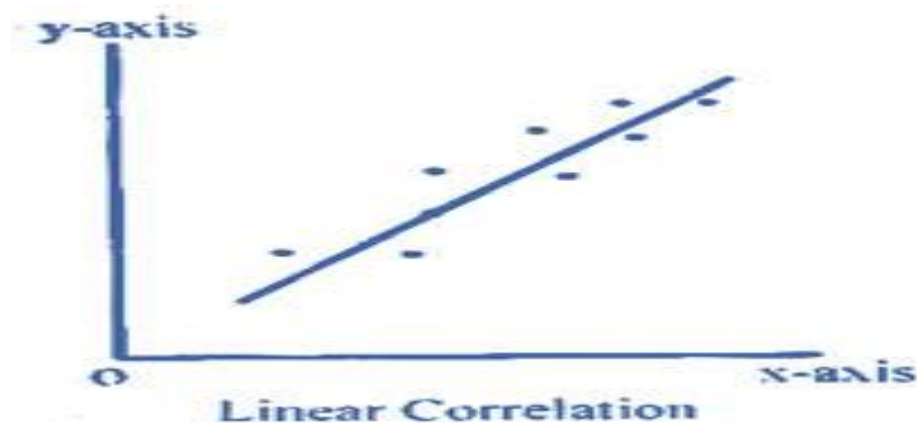
2. Linear or Curvilinear Correlation:

Linear Correlation:-

Linear correlation is the ratio of change between the two variables either in the same direction or opposite direction and the graphical representation of the one variable with respect to other variable is straight line.

Curvilinear or Non-linear Correlation:-

Consider another situation. First, with increase of one variable, the second variable increases proportionately upto some point; after that with an increase in the first variable the second variable starts decreasing. The graphical representation of the two variables will be a curved line. Such a relationship between the two variables is termed as the curvilinear correlation.



Scatter Diagram Method..

- 1. Scatter Diagram Method:
- Scatter diagram or dot diagram is a graphic device for drawing certain conclusions about the correlation between two variables.
- In preparing a scatter diagram, the observed pairs of observations are plotted by dots on a graph paper in a two dimensional space by taking the measurements on variable X along the horizontal axis and that on variable Y along the vertical axis.
- The placement of these dots on the graph reveals the change in the variable as to whether they change in the same or in the opposite directions. It is a very easy, simple but rough method of computing correlation.

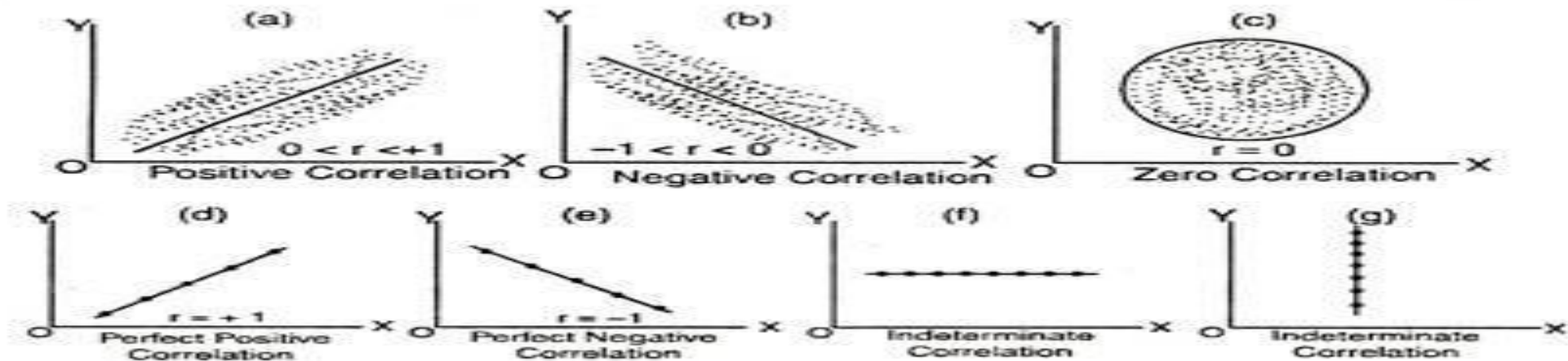


Fig. 5.1 Scatter Diagrams Showing Varying Degree of Relationship between X and Y.

Advantages and Disadvantages of Scatter Diagram Method

- Advantages of Scatter Diagram

- ✓ Simple and Non Mathematical method
- ✓ Non Influenced by the size of extreme item
- ✓ First step in investigating the relationship between two variables

- Disadvantages of Scatter Diagram

- ✓ Can not adopt to an exact degree of correlation.

Pearson Correlation Coefficient (r)..

Coefficient of correlation as ratio (r) :

- The Pearson correlation **measures the strength of the linear relationship between two variables**
- The product-moment coefficient of correlation may be thought of essentially as that ratio which expresses the extent to which changes in one variable are accompanied by—or dependent upon—changes in a second variable.
- It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = Values of the x-variable in a sample

\bar{x} = mean of the values of x-variable

Y_i = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

Advantages and Disadvantages of Pearson Correlation Coefficient

Advantages:-

- It summarizes in one value, the degree of correlation and direction of correlation also.

Disadvantages:-

- Always assume linear relationship
- Interpreting the value of r is difficult
- Value of Correlation coefficient is affected by the extreme values
- Time consuming methods

Spearman Rank Correlation Coefficient formula

- ✓ The Spearman's rank coefficient of correlation is a nonparametric measure of rank correlation (statistical dependence of ranking between two variables).
- ✓ It measures the strength and direction of the association between two ranked variables. But before we talk about the Spearman correlation coefficient, it is important to understand Pearson's correlation first. A Pearson correlation is a statistical measure of the strength of a linear relationship between paired data.
- For the calculation and significance testing of the ranking variable, it requires the following data assumption to hold true:
 - ✓ Interval or ratio level
 - ✓ Linearly related
 - ✓ Bivariant distribut

$$r_R = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Spearman correlation coefficient Formula

Here,

- r_R = Rank Correlation Coefficient
- n = number of data points of the two variables
- d_i = difference in ranks of the "ith" element

Advantages and Disadvantages of Spearman Rank Correlation Coefficient

- Advantages of Scatter Diagram

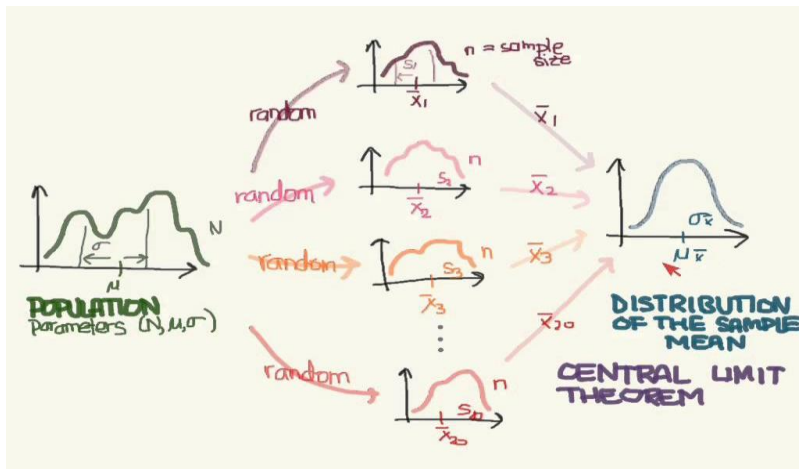
- ✓ This method is simpler to understand and easier to apply compared to Pearson's correlation method
- ✓ This method is useful where we can give ranks and not the actual data. (qualitative term)
- ✓ This method is to use where the initial data is in the form of ranks.

- Disadvantages of Scatter Diagram

- ✓ Cannot be used for finding out correlation in a grouped frequency distribution
- ✓ This method should be applied where n exceeds 30

Central Limit Theorem (CLT)

1. In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.
2. Central Limit Theorem (CLT) is very fundamental and a key concept in probability theory.
3. It says that the statistical and probabilistic methods that work for normal distribution can also be applied to many other problems which deal with different types of distributions.
4. This blog will explain what Central Limit Theorem is, why it is so important and how it solves the problems which does not deal with normal distribution. The below diagram represents the CLT flow



Central Limit Theorem Formula



Sample mean = Population mean = μ

$$\begin{aligned}\text{Sample standard deviation} &= \frac{(\text{Standard deviation})}{\sqrt{n}} \\ &= \frac{\sigma}{\sqrt{n}}\end{aligned}$$