# Business Statistics

**by**

Prof. Kewal Dhiraj Malde

**for the course**

# F.Y.B.Com.

# (Entrepreneurship)

**Affiliated to**

## University of Mumbai

**(Academic Year 2021-2022)**

# Contents


| | | |
|---|---|---|
| **1.** | **Contents** | **ii** |
| **1.** | **Measure of Central Tendency & Dispersion** | **1** |
| 1.1 | Unit Structure | 1 |
| 1.2 | Frequency distribution | 1 |
| 1.2.1 | Raw data | 1 |
| 1.2.2 | Variable | 2 |
| 1.3 | Classification of Data | 2 |
| 1.3.1 | Objectives of Classification | 3 |
| 1.3.2 | Basis of Classification | 3 |
| 1.4 | Ungrouped Frequency Distribution | 5 |
| 1.5 | Cumulative Frequency Distribution | 6 |
| 1.6 | Relative Frequency Distribution | 6 |
| 1.7 | Ogive Definition: | 7 |
| 1.7.1 | Ogive Graph | 8 |
| 1.8 | Central Tendency: | 12 |
| 1.8.1 | Arithmetic Mean (A.M) | 12 |
| 1.8.2 | Median | 13 |
| 1.8.3 | Mode | 15 |
| 1.9 | Measures of Dispersion | 16 |
| 1.9.1 | Range | 16 |
| 1.9.2 | Standard Deviation | 17 |
| 1.9.3 | Coefficient of Variance | 18 |
| 1.9.4 | Combined Mean: | 19 |

| | | |
|---|---|---|
| 1.9.5 | Combined Standard Deviation | 20 |
| 1.10 | Skewness: | 21 |
| 1.10.1 | Definition of Skewness: | 21 |
| 1.10.2 | Types of skewness | 22 |
| 1.10.3 | Pearson’s first coefficient of skewness | 23 |
| 1.10.4 | Pearson’s second coefficient of Skewness | 24 |
| 1.10.5 | Galton skewness (also known as Bowley's skewness) | 24 |
| 1.11 | Kurtosis | 25 |
| 1.11.1 | Excess Kurtosis | 26 |
| 1.11.2 | Types of excess kurtosis | 26 |
| **2.** | **Correlation and Regression** | **28** |
| 2.1 | Unit Structure | 28 |
| 2.2 | Introduction | 28 |
| 2.3 | Types of Correlation | 29 |
| 2.3.1 | Positive correlation | 29 |
| 2.3.2 | Negative correlation | 29 |
| 2.4 | Measurement of Correlation: | 30 |
| 2.4.1 | Scatter Diagram: | 30 |
| 2.4.2 | Karl Pearson’s co-efficient of correlation | 32 |
| 2.4.3 | Spearman’s rank correlation coefficient | 35 |
| 2.5 | Regression Analysis | 39 |
| **3.** | **Index Number and Time Series** | **43** |
| 3.1 | Unit Structure | 43 |
| 3.2 | Index number: | 43 |
| 3.3 | Importance of Index Number | 44 |

| | | |
|---|---|---|
| 3.4 | Characteristics of Index Numbers | 44 |
| 3.5 | Types of Index Numbers | 44 |
| 3.5.1 | Value Index | 44 |
| 3.5.2 | Quantity Index | 45 |
| 3.5.3 | Price Index | 45 |
| 3.6 | Uses of Index Number in Statistics | 45 |
| 3.7 | Advantages of Index Number | 46 |
| 3.8 | Limitations of Index Number | 46 |
| 3.9 | Price Index Numbers | 47 |
| 3.9.1 | Simple (Unweighted) Price Index Number By Aggregative Method: | 47 |
| 3.9.2 | Simple (Unweighted) Price Index Number by Average of Price Relatives Method: | 49 |
| 3.9.3 | Weighted Index Numbers by Aggregative Method: | 50 |
| 3.9.4 | Aggregate Expenditure Method: | 55 |
| 3.9.5 | Family Budget Method: | 55 |
| 3.9.6 | Method of Moving Averages: | 57 |
| 3.9.7 | Least Squares Method: | 62 |
| **4.** | **Probability** | **67** |
| 4.1 | Introduction | 67 |
| 4.2 | Concept of random experiment/trial and possible outcomes | 67 |
| 4.2.1 | Random Experiments | 67 |
| 4.2.2 | Mutually exclusive events | 71 |
| 4.2.3 | Mutually exhaustive events | 71 |
| 4.2.4 | Collectively Exhaustive Events | 71 |
| 4.2.5 | Complementary Events | 72 |

4.2.6 Classical definition of Probability 73

4.2.7 Addition theorem (without proof) 73

4.2.8 Conditional probability 75

4.2.9 Multiplication theorem 75

# Unit 1

# Measure of Central Tendency & Dispersion

## 1.1 Unit Structure

- Frequency distribution: Raw data, attributes and variables, Classification of data, frequency distribution, cumulative frequency distribution, Histogram & Ogive curves.
- Concept of central tendency, Desirable Properties for good measures of central tendency.
- Measures of central tendency: Arithmetic mean, median and mode for grouped and ungrouped data, Combined mean for two groups.
- Appropriate choice of measures.
- Measures of dispersion: Range, Standard deviation (S.D.) for grouped and ungrouped data, combined S.D., Variance.
- Measures of relative dispersion: coefficient of range, coefficient of variation
- Skewness and Kurtosis.

## 1.2 Frequency distribution

### 1.2.1 Raw data

Raw data is the unorganized data when we're done with the collection stage. This is because it is similar to a lump of clay with no identity and also of no practical use. It is important to realize that organized data facilitates comparison and meaningful conclusions. Further, to organize the data we need to look for similarities or group the data. In this way, we effectively convert heterogeneous data into homogeneous

data. To do so, an investigator has to classify the data in the form of a series. Series refer to those data which are in some order and sequence. Thus, if we arrange the data in the example mentioned in the introduction according to the classes in your school, we will eventually classify the data in form of a statistical series. Note that we can also arrange them according to their heights. Hence, this basis of the arrangement of raw data can vary from purpose to purpose.

### 1.2.2 Variable

A variable is simply something that can vary with time and we can measure this variation. In other words, a variable is a characteristic or a phenomenon which is capable of being measured and changes its value over time. A variable is classified into two:

#### 1.2.2.1 Discrete

Value of a discrete variable changes only in complete numbers or increases in jumps. Thus the phenomenon or characteristic, a discrete variable represents should be such that its value cannot be infractions but only in whole numbers. For example, the number of children in a family can be 2, 3, 4 etc but not 2.5, 3.5 etc.

#### 1.2.2.2 Continuous

A continuous variable assumes fractional values or its value does not increase in jumps. For example, the heights of students, the weights of students and so on.

## 1.3 Classification of Data

The main objective of the organization of data is to arrange the data in such a form that it becomes fairly easy to compare and analyze. Generally, we can do this by distributing data into various classes on the basis of some attribute or characteristic. This distribution of data into classes is the classification of data. Further, each division of data is a class. All in all, through the process of

classification we can group and divide data into classes according to a general attribute, which facilitates comparison and analysis.

## 1.3.1 Objectives of Classification

1) Simplification and Briefness: Classification presents data in a brief manner. Hence, it becomes fairly easy to analyze the data.

2) Utility: As classification highlights the similarity in the data, it brings out its utility.

3) Distinctiveness: With the help of grouping data into different classes, classification also brings out the distinctiveness in data.

4) Comparability: As already mentioned, it facilitates comparison of data.

5) Scientific Arrangement: Classification arranges data on scientific lines. Thus it also increases the reliability of data.

6) Attractive and Effective: Lastly, through the process of classification, data becomes effective and attractive.

## 1.3.2 Basis of Classification

Definitely, we can classify a given data according to various characteristics, depending on the purpose of our study. Evidently, there is the various basis of classification.

### 1.3.2.1 Geographical classification

When we classify data according to different locations, it is termed as a geographical classification of data. For example, a classification of the data about the number of children aged between 3-8 according to the various cities in India.

### 1.3.2.2 Chronological Classification

In chronological classification, we classify data according to time i.e., it follows a chronological sequence. For example, the classification of the data about the number of deaths in India according to the years.

#### 1.3.2.3 Qualitative Classification

Here, we classify data according to the qualities or attributes of data. One key point to remember is that an attribute is qualitative in nature i.e. we cannot measure an attribute in quantitative terms like 5, 1, 2 etc. This qualification is further of two types:

##### *1.3.2.3.1 Simple*

In the simple qualitative classification of data, we qualify data exactly into two groups. One group has data items that exhibit the quality, the other group doesn't. Evidently, it is also known as classification according to a dichotomy. Example of classes can be educated-uneducated, male-female and so on.

##### *1.3.2.3.2 Manifold*

Here we classify data according to more than one characteristic of an attribute. This means one we classify data into two groups according to an attribute; the two groups are further divided into two according to another attribute. As a result, there can be many levels of classification couples with more than just two classes. For example, the classification of data about students in a class, according to their gender, followed by classification according to whether they are fat or not.

#### 1.3.2.4 Quantitative or Numerical Classification

Unlike qualitative classification, quantitative classification allows numerical division of data into classes. Here, each class represents a range of numerical values for the phenomenon under consideration. Accordingly, we frame each class with a lower and higher value and according to the range of data.

Again, the phenomenon should be such that it can be expressed in numerical terms. As it is classified into classes with a different range of values, this classification is effectively the representation of the change of the value of a

phenomenon over time or across different regions. Which means its value varies. Accordingly, quantitative classification is also known as classification by variables.

The different types of frequency distributions are ungrouped frequency distributions, grouped frequency distributions, cumulative frequency distributions, and relative frequency distributions.

#### 1.3.2.5 Grouped Frequency Distribution

Sometimes to make deriving insights from an observation easily, we group them into class intervals.

Calculate the maximum and minimum value of the data set.

Divide this range by the number of groups you intend to have in your analysis.

Segregate the data within this small sub-group basis the class width.

Calculate the frequency of data within each group.

## 1.4 Ungrouped Frequency Distribution

The ungrouped cumulative distribution is similar to grouped frequency distribution except for the fact that class intervals are not created, and values are ordered from minimum to maximum.

List the unique values as the first column.

Calculate the repeated instances of each unique value and record it.

## 1.5 Cumulative Frequency Distribution

When you add or subtract the frequencies of all the previous class intervals to determine the frequency of a particular class interval, it results in a cumulative frequency distribution. Also, another major difference is that class intervals do not

denote a range but instead represent a logical conclusion like greater than a threshold value or less than a threshold value.

Calculate frequencies for every category.

Arrange in ascending or descending order according to categories/class intervals based on whether one wants to prepare an increasing/decreasing cumulative frequency distribution.

Total all the preceding frequencies. E.g., the second category's frequency is calculated by the sum of the first and second category's individual frequencies. Third is calculated by the sum of the first, second, third category's individual frequencies.

## 1.6 Relative Frequency Distribution

A relative frequency distribution is extensively used in our day-to-day statistical applications, which refers to the proportion of total observations associated with each category. It is calculated for individual class intervals by dividing them by the total observed frequencies. Relative frequencies can be written as a percentage, fraction, or decimal points. Cumulative relative frequency is the total of all preceding relative frequencies. To find the cumulative relative frequency, total all the previous relative frequencies till the current category.

**Solved Examples**

1) A research was done in 20 homes in Chennai Avadi. People were asked how many bikes did they own?

   The results were:1, 4, 3, 0, 5, 1, 2, 2, 1, 5, 2, 3, 2, 2, 0, 1, 2, 0, 3, 2.

   Present this data in Frequency Distribution Table. Also, find the maximum number of homes owning the same number of bikes.

**Solution:**

Divide the number of bikes in every home into different intervals. Every house can own either 0,1,2,3, etc. bikes. All these numbers form the rows. Now calculate the number of homes having {0,1,2,3, etc.} bikes. This is called the frequency. When you plot this in the form of a table:

| Number of Bikes | Frequency |
|---|---|
| 0 | 3 |
| 1 | 4 |
| 2 | 6 |
| 3 | 3 |
| 4 | 2 |
| 5 | 2 |

It can be seen from the table that 6 homes have 2 bikes and a lesser number of people own other numbers of bikes. Hence the answer is 6 homes.

## 1.7 Ogive Definition:

The Ogive is defined as the frequency distribution graph of a series. The Ogive is a graph of a cumulative distribution, which explains data values on the horizontal plane axis and either the cumulative relative frequencies, the cumulative frequencies or cumulative per cent frequencies on the vertical axis.

Cumulative frequency is defined as the sum of all the previous frequencies up to the current point. To find the popularity of the given data or the likelihood of the data that fall within the certain frequency range, Ogive curve helps in finding those details accurately.

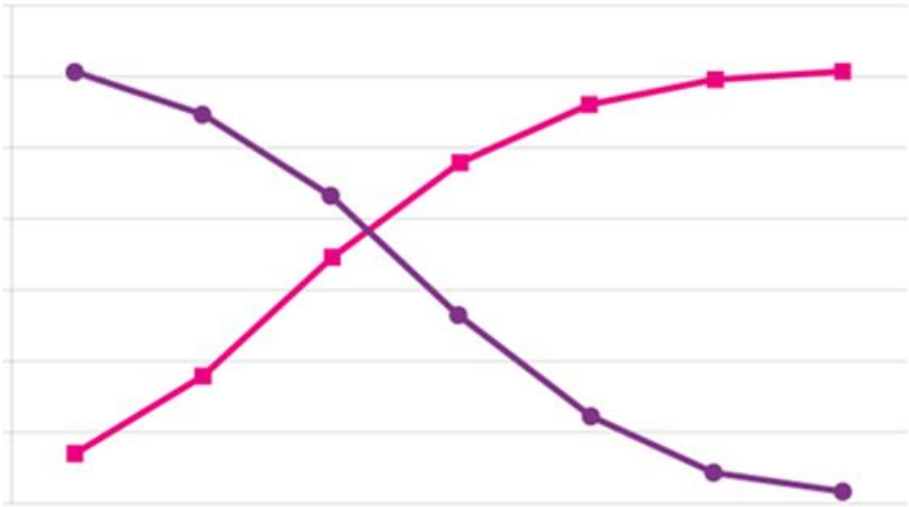
Create the Ogive by plotting the point corresponding to the cumulative frequency of each class interval. Most of the Statisticians use Ogive curve, to illustrate the data in the pictorial representation. It helps in estimating the number of observations which are less than or equal to the particular value.

### 1.7.1 Ogive Graph

The graphs of the frequency distribution are frequency graphs that are used to exhibit the characteristics of discrete and continuous data. Such figures are more appealing to the eye than the tabulated data. It helps us to facilitate the comparative study of two or more frequency distributions. We can relate the shape and pattern of the two frequency distributions.

The two methods of Ogives are:

- Less than Ogive
- Greater than or more than Ogive



The graph given above represents less than and the greater than Ogive curve. The rising curve (Brown Curve) represents the less than Ogive, and the falling curve (Green Curve) represents the greater than Ogive.

#### 1.7.1.1 Less than Ogive

The frequencies of all preceding classes are added to the frequency of a class. This series is called the less than cumulative series. It is constructed by adding the first-class frequency to the second-class frequency and then to the third class frequency and so on. The downward cumulation results in the less than cumulative series.

#### 1.7.1.2 Greater than or More than Ogive

The frequencies of the succeeding classes are added to the frequency of a class. This series is called the more than or greater than cumulative series. It is constructed by subtracting the first class, second class frequency from the total, third class frequency from that and so on. The upward cumulation result is greater than or more than the cumulative series.

#### 1.7.1.3 Ogive Chart

An Ogive Chart is a curve of the cumulative frequency distribution or cumulative relative frequency distribution. For drawing such a curve, the frequencies must be expressed as a percentage of the total frequency. Then, such percentages are cumulated and plotted, as in the case of an Ogive.

Below are the steps to construct the less than and greater than Ogive.

***How to Draw Less Than Ogive Curve?***

Step 1. Draw and mark the horizontal and vertical axes.

Step 2. Take the cumulative frequencies along the y-axis (vertical axis) and the upper-class limits on the x-axis (horizontal axis).

Step 3. Against each upper-class limit, plot the cumulative frequencies.

Step 4. Connect the points with a continuous curve.

***How to Draw Greater than or More than Ogive Curve?***

Step 1. Draw and mark the horizontal and vertical axes.

Step 5. Take the cumulative frequencies along the y-axis (vertical axis) and the lower-class limits on the x-axis (horizontal axis).

Step 6. Against each lower-class limit, plot the cumulative frequencies.

Step 7. Connect the points with a continuous curve.

#### 1.7.1.4 Uses of Ogive Curve

Ogive Graph or the cumulative frequency graphs are used to find the median of the given set of data. If both, less than and greater than, cumulative frequency curve is drawn on the same graph, we can easily find the median value. The point in which, both the curve intersects, corresponding to the x-axis, gives the median value. Apart from finding the medians, Ogives are used in computing the percentiles of the data set values.

**Ogive Example**

1) Construct the more than cumulative frequency table and draw the Ogive for the below-given data.

| Marks | 1-10 | 11-20 | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | 71-80 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 3 | 8 | 12 | 14 | 10 | 6 | 5 | 2 |

**Solution**:

“More than” Cumulative Frequency Table:

| Marks | Frequency | More than Cumulative Frequency |
|---|---|---|
| More than 1 | 3 | 60 |
| More than 11 | 8 | 57 |
| More than 21 | 12 | 49 |
| More than 31 | 14 | 37 |
| More than 41 | 10 | 23 |
| More than 51 | 6 | 13 |
| More than 61 | 5 | 7 |
| More than 71 | 2 | 2 |

Plotting an Ogive:

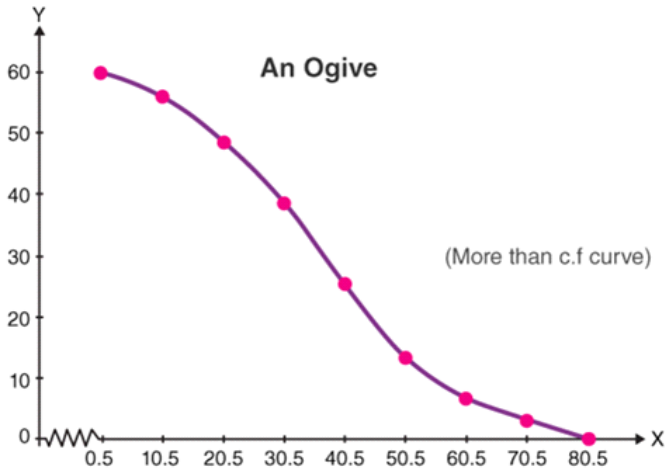
Plot the points with coordinates such as (70.5, 2), (60.5, 7), (50.5, 13), (40.5, 23), (30.5, 37), (20.5, 49), (10.5, 57), (0.5, 60).

An Ogive is connected to a point on the x-axis, that represents the actual upper limit of the last class, i.e.,( 80.5, 0)

Take x-axis, 1 cm = 10 marks

Y-axis, 1 cm = 10 c.f

More than the Ogive Curve:




## 1.8 Central Tendency:

Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution.” It aims to provide an accurate description of the entire data. It is the single value that is most typical / representative of the collected data. The term “number crunching” is used to illustrate this aspect of data description. The mean, median and mode are the three commonly used measures of central tendency.

**Properties of a good measure of central tendency:**

1) It is rigidly defined

7) It is based on all values of the data

8) It should not be affected by the extreme values of the data

9) It should have the sampling stability

10) It should be capable of further statistical analysis.

## 1.8.1 Arithmetic Mean (A.M)

Arithmetic Mean is defined as the sum of all the observations divided by the total number of observations in the data and is denoted by x, which is read as 'x-bar'

#### 1.8.1.1 Raw Data

#### 1.8.1.2 Ungrouped Frequency Data

#### 1.8.1.3 Grouped Frequency Data

## 1.8.2 Median

The median by definition refers to the middle value in a distribution. Median is the value of the variable which divides the distribution into two equal parts. The 50% observations lie below the value of the median and 50% observations lie above it. Median is called a positional average. Median is denoted by M.

#### 1.8.2.1 For Raw data

Median is defined as the value of the middle item of a series when the observations have been arranged in ascending or descending order of magnitude.

Example:

#### 1.8.2.2 For Ungrouped Frequency Distribution

#### 1.8.2.3 For Grouped Data

## 1.8.3 Mode

#### 1.8.3.1 For Raw Data

Mode is the value which occurs most frequently, in a set of observations. It is a value which is repeated maximum number of times and is denoted by Z.

Example 19: Find mode for the following data.

64, 38, 35, 68, 35, 94, 42, 35, 52, 35

Solution:

As the number 35 is repeated maximum number of times that is 4 times.

Mode=35 units.

For ungrouped frequency distribution:

Mode is the value of the variable corresponding to the highest frequency.

Example 20: Calculate the mode for the following data.

| Size of Shoe: | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| No. of Pairs: | 38 | 43 | 48 | 56 | 25 | 22 |

Solution: Here the highest frequency is 56 against the size 8.

Modal size = 8.

#### 1.8.3.2 For Grouped data:

In a Continuous distribution first the modal class is determined. The class interval corresponding to the highest frequency is called modal class.

## 1.9 Measures of Dispersion

### 1.9.1 Range

Range is the simplest measure of dispersion.

When the data are arranged in an array the difference between the largest and the smallest values in the group is called the Range.

Symbolically: Absolute Range = L - S, [where L is the largest value and S is the smallest value]

Amongst all the methods of studying dispersion range is the simplest to calculate and to understand but it is not used generally because of the following reasons:

Since it is based on the smallest and the largest values of the distribution, it is unduly influenced by two unusual values at either end. On this account, range is usually not used to describe a sample having one or a few unusual values at one or the other end It is not affected by the values of various items comprised in the distribution. Thus, it cannot give any information about the general characters of the distribution within the two extreme observations.

For example, let us consider the following three series:

Series A: 6 46 46 46 46 46 46 46

Series B: 6 6 6 6 46 46 46 46

Series C: 6 10 15 25 30 32 40 46

It can be noted that in all three series the range is the same, i.e. 40, however the distributions are not alike: the averages in each case is also quite different. It is because range is not sensitive to the values of individual items included in the distribution. It thus cannot be depended upon to give any guidance for determining the dispersion of the values within a distribution.

## 1.9.2 Standard Deviation

As we have seen range is unstable, quartile deviation excludes half the data arbitrarily and mean deviation neglects algebraic signs of the deviations, a measure of dispersion that does not suffer from any of these defects and is at the same time useful in statistic work is standard deviation. In 1893 Karl Pearson first introduced the concept. It is considered as one of the best measures of dispersion as it satisfies the requisites of a good measure of dispersion. The standard deviation measures the absolute dispersion or variability of a distribution. The

greater the amount of variability or dispersion greater is the value of standard deviation. In common language a small value of standard deviation means greater uniformity of the data and homogeneity of the distribution. It is due to this reason that standard deviation is considered as a good indicator of the representativeness of the mean.

It is represented by σ (read as _sigma‘).

σ2 i.e., the square of the standard deviation is called variance. Here, each deviation is squared.

The measure is calculated as the average of deviations from arithmetic mean. To avoid positive and negative signs, the deviations are squared. Further, squaring gives added weight to extreme measures, which is a desirable feature for some types of data. It is a square root of arithmetic mean of the squared deviations of individual items from their arithmetic mean.

The mean of squared deviation, i.e., the square of standard deviation is known as variance. Standard deviation is one of the most important measures of variation used in Statistics. Let us see how to compute the measure in different situation.

## 1.9.3 Coefficient of Variance

## 1.9.4 Combined Mean:

A combined mean is a mean of two or more separate groups, and is found by:

Calculating the mean of each group,

Combining the results.

Formula:

A combined mean is simply a weighted mean, where the weights are the size of each group.

### 1.9.5 Combined Standard Deviation

## 1.10 Skewness:

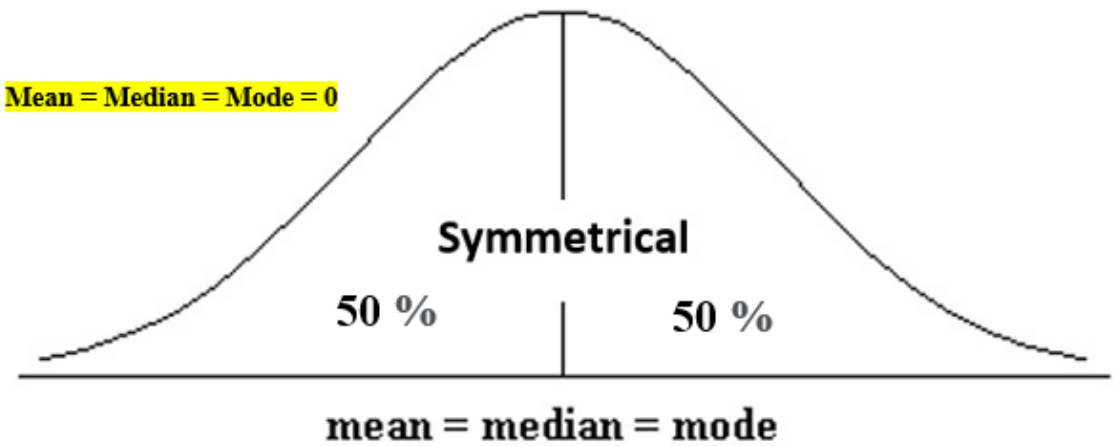
Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

### 1.10.1 Definition of Skewness:

If the values of a specific independent variable (feature) are skewed, depending on the model, Skewness may violate model assumptions or may reduce the interpretation of feature importance.

In statistics, Skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.

The normal distribution helps to know Skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero Skewness as all measures of a central tendency lies in the middle.



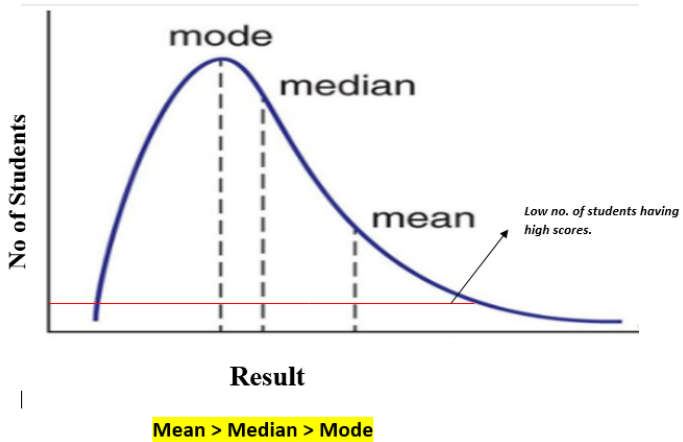

When data is symmetrically distributed, the left-hand side, and right-hand side, contains the same number of observations. (If the dataset has 90 values, then the

left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time Skewness comes into the picture.

## 1.10.2 Types of skewness

### 1.10.2.1 Positive skewed or right-skewed

In statistics, a positively skewed distribution is a sort of distribution where, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other, with positively skewed data, the measures are dispersing, which means Positively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.




Mean > Median > Mode

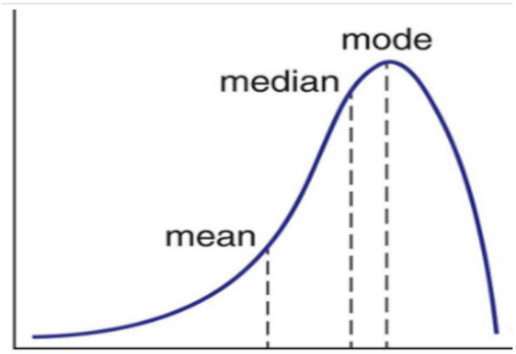
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the highest value

The extreme positive Skewness is not desirable for distribution, as a high level of Skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

### 1.10.2.2 Negative skewed or left-skewed

A negatively skewed distribution is the straight reverse of a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.




Mode > Median > mean

Median is the middle value, and mode is the highest value, and due to unbalanced distribution median will be higher than the mean.

## 1.10.3 Pearson's first coefficient of skewness

Subtract a mode from a mean, then divides the difference by standard deviation.

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the

standard deviation, it truly scales the value down to a limited range of -1 to +1. That accurately the range of the correlation values.

Pearson's first coefficient of Skewness is helping if the data present high mode. But, if the data have low mode or various modes, Pearson's first coefficient is not preferred, and Pearson's second coefficient may be superior, as it does not rely on the mode.

### 1.10.4 Pearson's second coefficient of Skewness

Multiply the difference by 3, and divide the product by standard deviation.

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$

If the Skewness is between -0.5 & 0.5, the data are nearly symmetrical.

If the Skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.

If the Skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

### 1.10.5 Galton skewness (also known as Bowley's skewness)

It is defined as

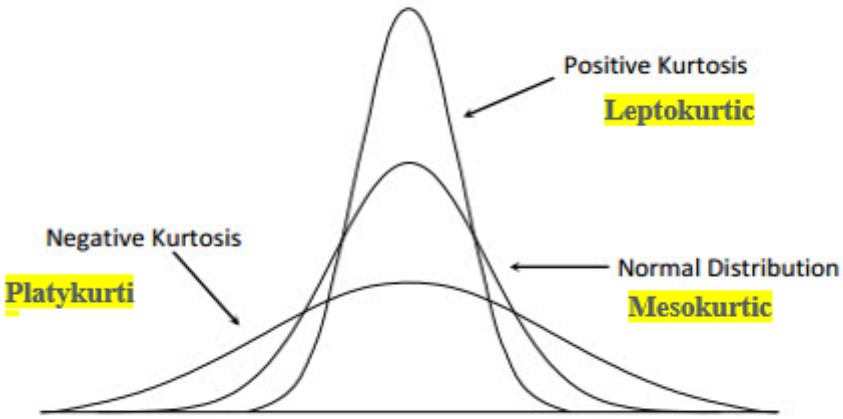
where Q1 is the lower quartile, Q3 is the upper quartile, and Q2 is the median.

## 1.11 Kurtosis

Kurtosis refers to the degree of presence of outliers in the distribution.

Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

The histogram is an effective graphical technique for showing both the Skewness and kurtosis of data set.




In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

### 1.11.1 Excess Kurtosis

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero

(Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

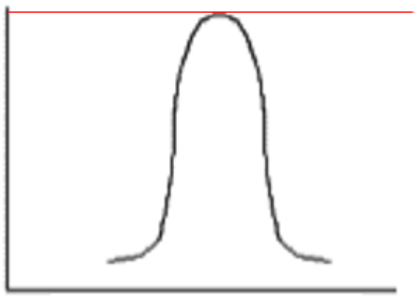
Excess kurtosis = Kurt – 3

## 1.11.2 Types of excess kurtosis

- Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
- Mesokurtic (kurtosis same as the normal distribution).
- Platykurtic or short-tailed distribution (kurtosis less than normal distribution).

### 1.11.2.1 Leptokurtic (kurtosis > 3)

Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.



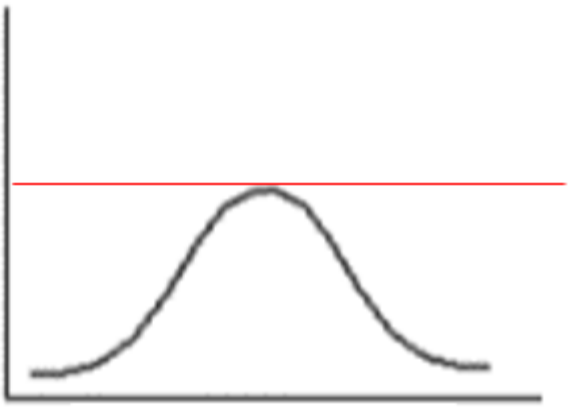
Leptokurtic

### 1.11.2.2 Platykurtic (kurtosis < 3)

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

### 1.11.2.3 Mesokurtic (kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near to 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.



**Mesokurtic**

**Mesokurtic = 3 – 3 = 0**

# Unit 2

# Correlation and Regression

## 2.1 Unit Structure

- Concept and types of correlation
- Scatter diagram
- Interpretation with respect to magnitude and direction of relationship.
- Karl Pearson's coefficient of correlation for ungrouped data.
- Spearman's rank correlation coefficient.
- Concept of regression
- Lines of regression for ungrouped data.
- Prediction using lines of regression.
- Regression coefficients and their properties.

## 2.2 Introduction

In the statistical analysis we come across the study of two or more relevant characteristics together in terms of their interrelations or interdependence. e.g. Interrelationship among production, sales and profits of a company. Inter relationship among rainfall, fertilizers, yield and profits to the farmers.

Relationship between price and demand of a commodity When we collect the information (data) on two of such characteristics it is called bivariate data. It is

generally denoted by (X,Y) where X and Y are the variables representing the values on the characteristics.

**Following are some examples of bivariate data:**

- Income and Expenditure of workers.
- Marks of students in the two subjects of Maths and Accounts.
- Height of Husband and Wife in a couple.
- Sales and profits of a company.

Between these variables we can note that there exists some sort of interrelationship or cause and effect relationship. i.e. change in the value of one variable brings out the change in the value of other variable also. Such relationship is called as correlation. Therefore, correlation analysis gives the idea about the nature and extent of relationship between two variables in the bivariate data.

## 2.3 Types of Correlation

There are two types of correlation: Positive correlation and Negative correlation.

### 2.3.1 Positive correlation

When the relationship between the variables *X* and Y is such that increase or decrease in X brings out the increase or decrease in *Y* also, i.e. there is direct relation between X and Y, the correlation is said to be positive. In particular when the change in X equals to change in *Y* the correlation is perfect and positive. e.g. Sales and Profits have positive correlation.

### 2.3.2 Negative correlation

When the relationship between the variables X and Y is such that increase or decrease in X brings out the decrease or increase in Y, i.e. there is an inverse relation between X and Y, the correlation is said to be negative. In particular when

the change in X equals to change in Y but in opposite direction the correlation is perfect and negative. e.g. Price and Demand have negative correlation.

## 2.4 Measurement of Correlation:

The extent of correlation can be measured by any of the following methods:

- Scatter diagrams
- Karl Pearson‘s co-efficient of correlation
- Spearman‘s Rank correlation

### 2.4.1 Scatter Diagram:

The Scatter diagram is a chart prepared by plotting the values of X and Y as the points (X, Y) on the graph. The pattern of the points is used to explain the nature of correlation as follows. The following figures and the explanations would make it clearer

#### 2.4.1.1 Perfect Positive Correlation

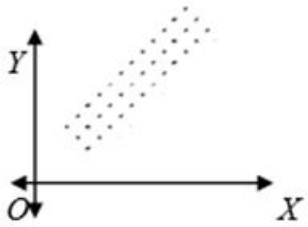
If the graph of the values of the variables is a straight line with positive slope, we say there is a perfect positive correlation between X and Y.

Here r = 1.


#### 2.4.1.2 Imperfect Positive Correlation

If the graph of the values of X and Y show a band of points from lower left corner to upper right corner , we say that there is an imperfect positive correlation.

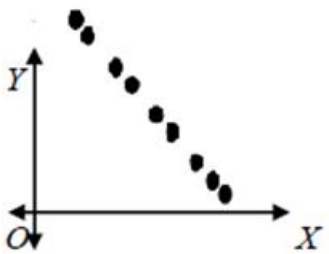
Here 0 < r < 1.




### 2.4.1.3 Perfect Negative Correlation

If the graph of the values of the variables is a straight line with negative slope we say there is a perfect negative correlation between X and Y.

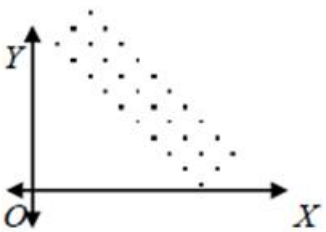
Here r = –1.




### 2.4.1.4 Imperfect Negative Correlation

If the graph of the values of X and Y show a band of points from upper left corner to the lower right corner, then we say that there is an imperfect negative correlation.

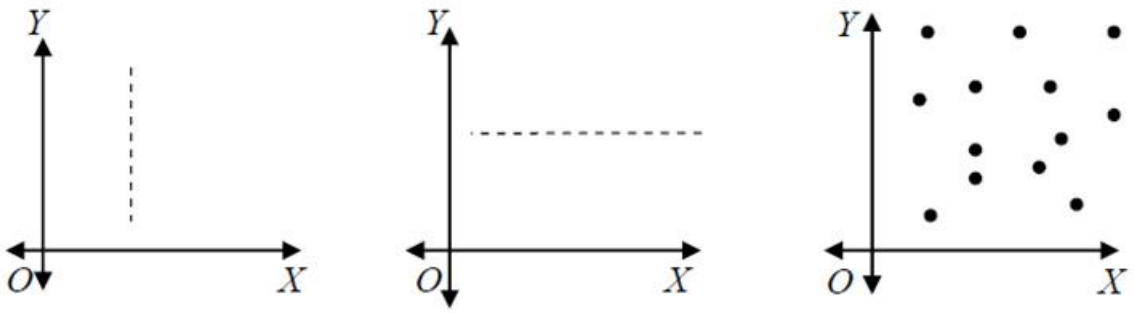
Here –1 < r < 0.



#### 2.4.1.5 Zero Correlation

If the graph of the values of X and Y do not show any of the above trend then we say that there is a zero correlation between X and Y. The graph of such type can be a straight line perpendicular to the axis, or may be completely scattered.

Here r = 0.




## 2.4.2 Karl Pearson's co-efficient of correlation

This co-efficient provides the numerical measure of the correlation between the variables X and Y. It is suggested by Prof. Karl Pearson and calculated by the formula

$$= \frac{Cov(x, y)}{\sigma_x . \sigma_y}$$

Where, Cov(x,y) : Covariance between $x$ & $y$

$\sigma_x$.: Standard deviation of $x$ & $\sigma_y$: Standard deviation of $y$

Also, $Cov(x,y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}$

$$S.D.(x) = \sigma x = \sqrt{\frac{1}{n}\Sigma(x-\bar{X})^2} = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{X}^2} \quad \text{and}$$

$$S.D.(y) = \sigma y = \sqrt{\frac{1}{n}\Sigma(y-\bar{Y})^2} \quad \sqrt{\frac{1}{n}\Sigma y^2 - \bar{Y}^2}$$

Remark : We can also calculate this co-efficient by using the formula given by

$$r = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{Y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y-\bar{Y})^2}} = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{Y}}{\sqrt{\left(\frac{\Sigma x^2}{n} - \bar{X}^2\right)\left(\frac{\Sigma y^2}{n} - \bar{Y}^2\right)}}$$

The Pearson's Correlation co-efficient is also called as the 'product moment correlation co-efficient'

#### 2.4.2.1 Properties of correlation co-efficient 'r'

- The value of 'r' can be positive (+) or negative (-)
- The value of 'r' always lies between –1 & +1, i.e. –1< r<+1.
- Significance of 'r' equals to –1, +1 & 0.
- When 'r'= +1; the correlation is perfect and positive.
- When 'r'= -1; the correlation is perfect and negative.
- When there is no correlation 'r'= 0.

Example: Let us calculate co-efficient of correlation between Marks of students in the Subjects of Maths & Accounts in a certain test conducted.

Table of calculation:

| Marks InMaths X | Marks In Accounts Y | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 28 | 30 | 840 | 784 | 900 |
| 25 | 40 | 1000 | 625 | 1600 |
| 32 | 50 | 1600 | 1024 | 2500 |
| 16 | 18 | 288 | 256 | 324 |
| 20 | 25 | 500 | 400 | 625 |
| 15 | 12 | 180 | 225 | 144 |
| 19 | 11 | 209 | 361 | 121 |
| 17 | 21 | 357 | 289 | 441 |
| 40 | 45 | 1800 | 1600 | 2025 |
| 30 | 35 | 1050 | 900 | 1225 |
| $\Sigma x = 242$ | $\Sigma y = 287$ | $\Sigma xy = 7824$ | $\Sigma x^2 = 6464$ | $\Sigma y^2$ 9905 |

n=10

Now Pearson's co-efficient of correlation is given by the fomula,

$$r = \frac{Cov(x, y)}{\sigma_x . \sigma_y}$$

Where,

$$\bar{x} = \frac{\Sigma x}{n} = \frac{242}{10} = 24.2 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{287}{10} = 28.7$$

$$\text{Cov}(x,y) = \frac{\Sigma xy}{n} - \bar{X}\bar{Y} \qquad \sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{X}^2} \qquad \sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{Y}^2}$$

$$= \frac{7824}{10} - 24.2 \times 28.7 \qquad \sigma_x = \sqrt{\frac{6464}{10} - 24.2^2} \qquad \sigma_y = \sqrt{\frac{9905}{10} - 28.7^2}$$

$$= 782.4 - 694.54 \qquad = \sqrt{60.76} \qquad = \sqrt{166.81}$$

$\text{Cov}(x,y) = 87.86, \quad \sigma_x = 7.79$ and $\sigma_y = 12.91$

$\therefore$ **Cov(x,y) = 87.86** $\sigma_x$ **= 7.79** **and** $\sigma_y$ **= 12.91**

Substituting the values in the formula of **r** we get

$$r = \frac{87.86}{7.79x12.91} = \mathbf{0.87}$$

$\therefore$ **r = 0.87**

### 2.4.3 Spearman's rank correlation coefficient

In many practical situations, we do not have the scores on the characteristics, but the ranks (preference order) decided by two or more observers. Suppose, a singing competition of 10 participants is judged by two judges A and B who rank or assign scores to the participants on the basis of their performance. Then it is quite possible that the ranks or scores assigned may not be equal for all the participants. Now the difference in the ranks or scores assigned indicates that there is a difference of opinion between the judges on deciding the ranks. The rank correlation studies the association in this ranking of the observations by two or more observers. The measure of the extent of association in rank allocation by the two judges is calculated by the co-efficient of Rank correlation 'R'. This co-efficient was developed by the British psychologist Edward Spearman in 1904.

Mathematically, Spearman's rank correlation co-efficient is defined as,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Where d= rank difference and n= no of pairs.

Remarks: We can note that, the value of 'R' always lies between –1 and +1

The positive value of 'R' indicates the positive correlation (association) in the rank allocation. Whereas, the negative value of 'R' indicates the negative correlation (association) in the rank allocation.

Example:

When ranks are given:-

Data given below read the ranks assigned by two judges to 8 participants. Calculate the co-efficient of Rank correlation.

| Participant No. | Ranks by Judge | | Rank diff Square $d^2$ |
|---|---|---|---|
| | A | B | |
| 1 | 5 | 4 | $(5-4)^2 = 1$ |
| 2 | 6 | 8 | 4 |
| 3 | 7 | 1 | 36 |
| 4 | 1 | 7 | 36 |
| 5 | 8 | 5 | 9 |
| 6 | 2 | 6 | 16 |
| 7 | 3 | 2 | 1 |
| 8 | 4 | 3 | 1 |
| N = 8 | | Total | $104 = \Sigma d^2$ |

Spearman's rank correlation co-efficient is given by

$$R=1-\frac{6\Sigma d^2}{n(n^2-1)}$$

Substituting the values from the table we get,

$$R= 1-\frac{6x\ 104}{8(8^2-1)} = -0.23$$

The value of correlation co-efficient is - 0.23. This indicates that there is negative association in rank allocation by the two judges A and B

When scores are given:-

| Student No | Marks by Examiner | | Ranks | | Rank difference square |
|---|---|---|---|---|---|
| | A | B | RA | RB | $D^2$ |
| 1 | 85 | 80 | 2 | 2 | 0 |
| 2 | 56 | 60 | 8 | 7 | 1 |
| 3 | 45 | 50 | 10 | 10 | 0 |
| 4 | 65 | 62 | 6 | 6 | 0 |
| 5 | 96 | 90 | 1 | 1 | 0 |
| 6 | 52 | 55 | 9 | 8 | 1 |
| 7 | 80 | 75 | 3 | 4 | 1 |
| 8 | 75 | 68 | 5 | 5 | 0 |
| 9 | 78 | 77 | 4 | 3 | 1 |
| 10 | 60 | 53 | 7 | 9 | 1 |
| N = 10 | | | | Total | 5 |

The data given below are the marks given by two Examiners to a set of 10 students in a aptitude test. Calculate the Spearman's Rank correlation co-efficient, 'R'

Spearman's rank correlation co-efficient is given by

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Substituting the values from the table we get,

$$R = 1 - \frac{6x5}{10(10^2-1)}$$

R = 1-0.03

R = 0.97

The value of correlation co-efficient is + 0.97. This indicates that there is positive association in assessment of two examiners, A and B

### 2.4.3.1 Case of repeated values:-

It is quite possible that the two participants may be assigned the same score by the judges. In such cases Rank allocation and calculation of rank correlation can be explained as follows.

Example: The data given below scores assigned by two judges for 10 participants in the singing competition. Calculate the Spearman's Rank correlation co-efficient

| Participant No | Score assigned By Judges | | Ranks | | Rank difference square |
|---|---|---|---|---|---|
| | A | B | RA | RB | D2 |
| 1 | 28 | 35 | 9 (8.5) | 6 | (8.5-6)2 =6.25 |
| 2 | 40 | 26 | 3 | 10(9.5) | 42.25 |
| 3 | 35 | 42 | 5 (4.5) | 3 | 2.25 |
| 4 | 25 | 26 | 10 | 9 (9.5) | 0.25 |
| 5 | 28 | 33 | 8 (8.5) | 7 | 2.25 |
| 6 | 35 | 45 | 4 (4.5) | 2 | 6.25 |
| 7 | 50 | 32 | 1 | 8 | 49 |
| 8 | 48 | 51 | 2 | 1 | 1 |
| 9 | 32 | 39 | 6 | 4 | 4 |
| 10 | 30 | 36 | 7 | 5 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| N = 10 | | | | Total | $\Sigma$ d2 =117.5 |

**Explanation:-** In the column of A and B there is repetition of scores so while assigning the ranks we first assign the ranks by treating them as different values and then for repeated scores we assign the average rank.

In col A the score 35 appears 2 times at number 4 and 5 in the order of ranking so we calculate the average rank as (4+5)/2 = 4.5.

Hence the ranks assigned are 4.5 each. The other repeated scores can be ranked in the same manner.

**Note:** In this example we can note that the ranks are in fraction e.g. 4.5, which is logically incorrect or meaningless. Therefore in the calculation of 'R' we add a correction factor(C.f.). $d^2$ calculated as follows

Table of correction factor (C.F.)

| Value Repeated | Frequency M | $m(m^2-1)$ |
|---|---|---|
| 35 | 2 | $2x(2^2-1)=6$ |
| 28 | 2 | 6 |
| 26 | 2 | 6 |
| | Total | $m(m^2-1)=18$ |

$$\text{Now } C.F. = \frac{\Sigma(m^3 - m)}{12} = 18/12 = 1.5$$

$d^2$ = 117.5+1.5= 119

We use this value in the calculation of 'R'

Now the Spearman's rank correlation co-efficient is given by

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$\text{Substituting the values we get, } R = 1 - \frac{6x\ 119}{10(10^2-1)} = 1 - 0.72 = 0.28$$

## 2.5 Regression Analysis

As the correlation analysis studies the nature and extent of interrelationship between the two variables X and Y, regression analysis helps us to estimate or approximate the value of one variable when we know the value of other variable. Therefore we can define the

'Regression' is the estimation (prediction) of one variable from the other variable when they are correlated to each other. e.g. We can estimate the Demand of the commodity if we know it's Price.

Why are there two regressions?

When the variables X and Y are correlated there are two possibilities,

Variable X depends on variable Y. in this case we can find the value of x if know the value of y. This is called regression of x on y.

Variable Y depends on variable X. we can find the value of y if know the value of X. This is called regression of y on x.

Hence there are two regressions,

- Regression of X on Y;
- Regression of X on Y.

Formulas on Regression equation,

| **Regression of** X **on** Y | **Regression of** X **on** Y |
|---|---|
| i. Assumption: X depends on Y The regression equation is $(x - \bar{x}) = b_{xy}(y - \bar{y})$<br>ii. $b_{xy}$=Regression co-efficient of<br>$\text{X on Y} = \frac{Cov(x, y)}{V(y)}$ | i. Y depends on X<br>The regression equation is $(y - \bar{y}) = b_{yx}(x - \bar{x})$<br>ii. $b_{yx}$= Regression co-efficient of<br>$\text{Y on X} = \frac{Cov(x, y)}{V(x)}$ |

Where,

$$\text{Cov}(x,y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}$$

$$V(x) = \frac{1}{n}\Sigma(x-\bar{x})^2 \quad \text{and} \quad V(y) = \frac{1}{n}\Sigma(y-\bar{y})^2$$

$$V(x) = \frac{1}{n}\Sigma x^2 - \bar{x}^2 \quad \text{and} \quad V(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2$$

Use: To find X  Use: To find γ

**Example 2**:

Obtain the two regression equations and hence find the value of y when x=10

Data:-

| X | Y | XxY | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 12 | 25 | 300 | 144 | 625 |
| 20 | 18 | 360 | 400 | 324 |
| 8 | 17 | 136 | 64 | 289 |
| 14 | 13 | 182 | 196 | 169 |
| 16 | 15 | 240 | 256 | 225 |
| $\Sigma x = 70$ | $\Sigma y = 88$ | $\Sigma xy = 1218$ | $\Sigma x^2 = 1060$ | $\Sigma y^2 = 1632$ |

And n= 5

Now the two regression equations are,

$(x-\bar{x}) = b_{xy}(y-\bar{y})$ -------x on y (i)

$(y-\bar{y}) = b_{yx}(x-\bar{x})$ ------- y on x (ii)

Where,

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{70}{5} = 14 \quad \text{and} \quad \bar{y} = \frac{1}{n}\Sigma y = \frac{88}{5} = 17.6$$

Also,

$$\text{Cov}(x,y) = \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y} = \frac{1218}{5} - 14 \times 17.6 = 243.6 - 246.4$$

$\therefore$ **Cov(x,y)** = -2.8

$$V(x) = \frac{1}{n}\Sigma x^2 - \bar{x}^2 = \frac{1060}{5} - 14^2 = 212 - 196$$

$V(x) = 16$

$$V(y) = \frac{1}{n}\Sigma y^2 - \bar{y}^2 = \frac{1632}{5} - 17.6^2 = 326.4 - 309.76$$

$V(y) = 16.64$

Now we find,

Regression co-efficient of X on Y

$$b_{xy} = \frac{Cov(x,y)}{V(y)} = \frac{2.8}{16.64}$$

$\therefore$ $b_{xy} = -0.168$

Regression co-efficient of X on Y

$$b_{yx} = \frac{Cov(x,y)}{V(x)} = -\frac{2.8}{16.64}$$

$b_{yx} = 0.175$

Now substituting the values of $\bar{x}$, $\bar{y}$, $b_{xy}$ and $b_{yx}$ in the regression equations we get,

$(x-14) = -0.168(y-17.6)$ -------x on y (i)

$(y-17.6) = -0.175(x-14)$ ------- y on x (ii)

as the two regression equations.

Now to estimate y when x =10, we use the regression equation of y on x

$\therefore$ $(y-17.6) = -0.175(10-14)$

$\therefore$ **y =17.6 +0.7 = 24.3**

We can also obtain the regression coefficients $b_{xy}$ and $b_{yx}$ from standard deviations, $\sigma_x$., $\sigma_y$ and correlation coefficient 'r' using the formulas

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Also consider,

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y}\, r\frac{\sigma_y}{\sigma_x} = r^2 \qquad \text{i.e. } r = \sqrt{b_{xy} \times b_{yx}}$$

Hence the correlation coefficient 'r' is the geometric mean of the regression coefficients, $b_{xy}$ and $b_{yx}$

**Example**

You are given the information about advertising expenditure and sales:

| Exp. on Advertisiment (Rs. In Lakh) | Sales (Rs. In Lakh) | |
|---|---|---|
| Mean | 10 | 90 |
| S.D. | 3 | 12 |

Coefficient of correlation between sales and expenditure on Advertisement is 0.8.Obtain the two regression equations.

Find the likely sales when advertisement budget is Rs. 15 Lakh.

**Solution**: We define the variables,

X: Expenditure on advertisement

Y: Sales achieved.

Therefore we have,

$\bar{x} = 10$, $\bar{y} = 90$, $6x = 3$, $6y = 12$ and $r = 0.8$

Now, using the above results we can write the two regression equations as

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \quad \text{-------x on y (i)}$$

$$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \text{ ------- y on x (ii)}$$

Substituting the values in the equations we get,

$$(x-10) = 0.8\frac{3}{12}(y-90)$$

i.e $x - 10 = 0.2\,(y-90)$ -------x on y (i)

$$\text{also} \quad (y-90) = 0.8\frac{12}{3}(x-10)$$

$$\text{i.e.} \quad y-90 = 3.2\,(x-10) \qquad \text{-------y on x (ii)}$$

Now when expenditure on advertisement (x) is 15, we can find the sales from eqn (ii) as,

$$y-90 = 3.2\,(15-10)$$

$$\therefore \quad y = 90 + 16 = 106$$

Thus the likely sales are Rs.106 Lakh.

# Unit 3

# Index Number and Time Series

## 3.1 Unit Structure

- Concept of index number, price index number, price relatives.
- Problems in construction of index number.
- Construction of price index number: Weighted index Number, Laspeyre's, Paasche's and Fishers method.
- Cost of living/ consumer price index number: Definition and problems in construction, method of construction: family budget and aggregate expenditure.
- Uses of index numbers, commonly used index numbers.
- Types of business models
- Moving Averages
- Least square method of fitting model

## 3.2 Index number:

Index number in statistics is the measurement of change in a variable or variables across a determined period. It will show general relative change and not a directly measurable figure. An index number is expressed in percentage form.

## 3.3 Importance of Index Number

Index numbers occupy an important place due to its efficacy in measuring the extent of economic changes across a stipulated period. It helps to study such changes' effects due to factors that cannot be directly measured.

## 3.4 Characteristics of Index Numbers

**The main features of index numbers are –**

- It is a special category of average for measuring relative changes in such instances where absolute measurement cannot be undertaken.
- Index number only shows the tentative changes in factors that may not be directly measured. It gives a general idea of the relative changes.
- The method of index number measure alters from one variable to another related variable.
- It helps in the comparison of the levels of a phenomenon concerning a specific date and to that of a previous date.
- It is representative of a special case of averages especially for a weighted average.
- Index numbers have universal utility. The index that is used to ascertain the changes in price can also be used for industrial and agricultural production.

## 3.5 Types of Index Numbers

### 3.5.1 Value Index

A value index number is formed from the ratio of the aggregate value for a particular period with that of the aggregate value that is found in the base period. The value index is utilized in for inventories, sales and foreign trade, among others.

### 3.5.2 Quantity Index

A quantity index number is used to measure changes in the volume or quantity of goods that are produced, consumed and sold within a stipulated period. It shows the relative change across a period for particular quantities of goods. Index of Industrial Production (IIP) is an example of Quantity Index.

### 3.5.3 Price Index

A price index number is used to measure how price alters across a period. It will indicate the relative value and not the absolute value. The Consumer Price Index (CPI) and Wholesale Price Index (WPI) are major examples of a price index.

## 3.6 Uses of Index Number in Statistics

- It helps in measuring changes in the standard of living as well as the price level.
- Wage rate regulation is consistent with the changes in the price level. With the determination of price levels, wage rates may be revised.
- Government policies are framed following the index number of prices. This price stability inherent to fiscal and economic policies is based on index numbers.
- It gives a pointer for international comparison concerning different economic variables—for instance, living standards between two countries.

## 3.7 Advantages of Index Number

- It adjusts primary data at varying costs, which is useful for deflating. It facilitates the transformation from nominal wage to real wage.

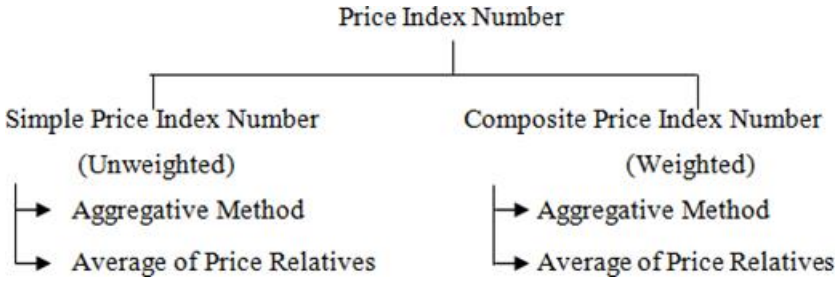
- Index numbers find extensive usage in economics and help in the framing of appropriate policies. Such findings help with the establishment of researches as well.
- It helps in case of trends such as drawing outcomes for irregular forces and cyclical forces.
- Index number can be leveraged in case of future development of activities in the economic sphere. This time series analysis is utilised for the determination trends and cyclical developments.
- The number is useful in measuring the changes that take place in the standard of living in different countries over an established period.

## 3.8 Limitations of Index Number

- There are chances for errors given that index numbers come as a result of samples. These samples are put together after deliberation, which creates chances for errors. It can also be found in weights or base periods etc.
- It is always calculated based on items. Items that are so selected may not exactly be in trend, which in turn creates an inaccurate analysis.
- Multiple methods can be used to formulate index numbers. Due to this multiplicity of methods, outcomes may bring forward a different set of values which may further lead to confusion.
- The index numbers show the approximate indications of the relative changes that occur. Moreover, the changes in variables that are compared over a prolonged time may fall short on reliability.
- The selection of representative commodities may be skewed. It is since these commodities are based on samples.

## 3.9 Price Index Numbers

The price index numbers are classified as shown in the following diagram:




Notations:

- $P_0$: Price in Base Year
- $Q_0$: Quantity in Base Year
- $P_1$: Price in Current Year
- $Q_1$: Quantity in Current Year

The suffix '0' stands for the base year and the suffix '1' stands for the current year.

### 3.9.1 Simple (Unweighted) Price Index Number By Aggregative Method:

In this method we define the price index number as the ratio of sum of prices in current year to sum of prices in base year and express it in percentage. i.e. multiply the quotient by 100.

Symbolically,

**Steps for computation:**

- The total of all base year prices is calculated and denoted by
- The total of all current year prices is calculated and denoted by
- Using the above formula, simple price index number is computed.

**Example 1**

For the following data, construct the price index number by simple aggregative method:

| Commodity | Unit | Price in | |
|---|---|---|---|
| | | 1985 | 1986 |
| A | Kg | 10 | 12 |
| B | Kg | 4 | 7 |
| C | Litre | 6 | 7 |
| D | Litre | 8 | 10 |

Solution: Following the steps for computing the index number, we find the totals of the 3rd and 4th columns as shown below:

| **Commodity** | **Unit** | **Price in** | |
|---|---|---|---|
| | | **1985($P_0$)** | **1986($P_1$)** |
| **A** | Kg | 10 | 12 |
| **B** | Kg | 4 | 7 |
| **C** | Litre | 6 | 7 |
| **D** | Litre | 8 | 10 |

### 3.9.2 Simple (Unweighted) Price Index Number by Average of Price Relatives Method:

In this method the price index is calculated for every commodity and its arithmetic mean is taken, i.e. the sum of all price relative is divided by the total number of commodities.

Symbolically, if there are n commodities in to consideration, then the simple price index number of the group is calculated by the formula:

$$I = \frac{1}{n}\Sigma\left(\frac{P_1}{P_0} \text{ x } 100\right)$$

**Example 2**

Construct the simple price index number for the following data using average of price relative method:

| Commodity | Unit | Price in | |
|---|---|---|---|
| | | 1997 | 1998 |
| Rice | Kg | 10 | 13 |
| Wheat | Kg | 6 | 8 |
| Milk | Litre | 8 | 10 |
| Oil | Litre | 15 | 18 |

Solution: In this method we have to find price relatives for every commodity and then total these price relatives. Following the steps for computing as mentioned above, we introduce first, the column of price relatives. The table of computation is as follows:

| Commodity | Unit | Price in | | |
|---|---|---|---|---|
| | | 1997(P0) | 1998(P1) | |
| Rice | Kg | 10 | 13 | 130 |
| Wheat | Kg | 6 | 8 | 133.33 |
| Milk | Litre | 8 | 10 | 125 |
| Oil | Litre | 15 | 18 | 120 |
| | | | Total | 508.33 |

Now, n = 4 and the total of price relatives is 508.33

$$\therefore I = \frac{1}{n}\Sigma\left(\frac{P_1}{P_0} \text{ x } 100\right) = \frac{508.33}{4} = 127.08$$

The prices in 1998 have increased by 27 % as compared with in 1997.

**Remark:**

- The simple aggregative method is calculated without taking into consideration the units of individual items in the group. This may give a misleading index number.
- This problem is overcome in the average of price relatives method, as the individual price relatives are computed first and then their average is taken.
- Both the methods are unreliable as they give equal weightage to all items in consideration which is not true practically.

### 3.9.3 Weighted Index Numbers by Aggregative Method:

In this method weights assigned to various items are considered in the calculations. The products of the prices with the corresponding weights are computed; their totals are divided and expressed in percentages.

Symbolically, if W denotes the weights assigned and P0, P1 have their usual meaning, then the weighted index number using aggregative method is given by the formula:

$$I = \frac{\Sigma P_1 W}{\Sigma P_0 W} \text{ x } 100$$

**Example 3**

From the following data, construct the weighted price index number:

| Commodity | A | B | C | D |
|---|---|---|---|---|
| Price in 1982 | 6 | 10 | 4 | 18 |
| Price in 1983 | 9 | 18 | 6 | 26 |
| Weight | 35 | 30 | 20 | 15 |

Solution: Following the steps mentioned above, the table of computations is as follows:

| Commodity | Weight (W) | Price in 1982 (P0) | P0W | Price in 1983 (P1) | P1W |
|---|---|---|---|---|---|
| A | 35 | 6 | 210 | 9 | 315 |
| B | 30 | 10 | 300 | 18 | 540 |
| C | 20 | 4 | 80 | 6 | 120 |
| D | 15 | 18 | 270 | 26 | 390 |
| Total | - | - | P0W = 860 | - | P1W = 1365 |

$$\text{Weighted Index Number } I = \frac{\Sigma P_1 W}{\Sigma P_0 W} \text{x} 100 = \frac{1365}{860} \text{x } 100 = 158.72$$

There are different formulae based on what to be taken as the weight while calculating the weighted index numbers. Based on the choice of the weight we are going to study here three types of weighted index numbers:

- Laspeyre‘s Index Number,
- Paasche‘s Index Number and

- Fisher‘s Index Number.

### 3.9.3.1 Laspeyre’s Method

Laspeyre was of the view that base year quantities must be chosen as weights. Therefore the formula is:

Here, ∑P1Q0= Summation of prices of current year multiplied by quantities of the base year taken as weights and ∑P0Q0= Summation of, prices of base year multiplied by quantities of the base year taken as weights.

### 3.9.3.2 Paasche’s Method

Unlike the above mentioned, Paasche believed that the quantities of the current year must be taken as weights. Hence the formula:

Here, ∑P1Q1= Summation of, prices of current year multiplied by quantities of the current year taken as weights and ∑P0Q1= Summation of, prices of base year multiplied with quantities of the current year taken as weights.

### 3.9.3.3 Fisher’s Method

Fisher combined the best of both above-mentioned formulas which resulted in an ideal method. This method uses both current and base year quantities as weights as follows:

NOTE: Index number of base year is generally assumed to be 100 if not given.

**Fisher’s Method is an Ideal Measure:**

As noted Fisher’s method uses views of both Laspeyres and Paasche. Hence it takes into account the prices and quantities of both years. Moreover, it is based on

the concept of the geometric mean, which is considered as the best mean method. However, the most important evidence for the above affirmation is that it satisfies both time reversal and factor reversal tests. Time reversal test checks that when we reverse the current year to base year and vice-versa, the product of indexes should be equal to unity. This confirms the working of a formula in both directions. Also, factor reversal test implies that interchanging the piece and quantities do not give varying results. This proves the consistency of the formula.

**Common Problems with Construction of Index Numbers**

Due to the availability of a wide range of index numbers we have to select an index number that matches the objective we want to fulfill. For example, to study the impact of a change in the government's budget on people, one should refer to the price index number. It must be noted that the selected base year should be a normal one. In other words, there should be no reforms in that year which can influence the economy in a drastic manner. If such is chosen as the base year there will be a big variation in the index numbers, which would not reflect the accurate changes over the years. Also, it is not possible to include all the goods and services along with their prices in our calculations. This means we need to select various goods and services that can effectively represent all of them. In a word, a sample size has to be selected. Larger the sample size more is the accuracy. And we need to select the method of calculation that suits best with the objective in hand.

**Solved Example**

1) Construct index numbers of prices of items in the year 2012 from the following data by: Laspeyres method, Paasche's method, Fisher's method

| Items | Price (2004) | Quantity(2004) | Price(2012) | Quantity (2012) |
|---|---|---|---|---|
| A | 10 | 10 | 5 | 25 |
| B | 35 | 4 | 35 | 10 |
| C | 30 | 3 | 15 | 15 |

| | | | | |
|---|---|---|---|---|
| D | 10 | 25 | 20 | 20 |
| E | 40 | 3 | 40 | 5 |

Solution:

| Items | P0 | Q0 | P1 | Q1 | P0Q0 | P0Q1 | P1Q0 | P1Q1 |
|---|---|---|---|---|---|---|---|---|
| A | 10 | 10 | 5 | 25 | 100 | 250 | 50 | 125 |
| B | 35 | 4 | 35 | 10 | 140 | 350 | 140 | 350 |
| C | 30 | 3 | 15 | 15 | 90 | 450 | 45 | 225 |
| D | 10 | 25 | 4 | 20 | 250 | 200 | 100 | 80 |
| E | 40 | 3 | 40 | 5 | 120 | 200 | 120 | 200 |
| | | | | | ∑=700 | ∑=1450 | ∑=455 | ∑=980 |

- Laspeyre's method= (455/700) × 100 = 65
- Paasche's method= (980/1450) × 100= 67.58
- Fisher's method= √0.43927 × 100 = 66.27

## 3.9.4 Aggregate Expenditure Method:

Under this method, we take the quantities of consumption of various commodities by a particular section of the people in the base year as weights. We then calculate the total expenditure of each commodity for each year.

For this, we need to multiply the price of the current year with the quantity or weight of the base year and add these products. Similarly, we have to calculate the total expenditure for the base year of each commodity.

Thus, in order to calculate the index numbers, we have to divide the total expenditure of the current year by the total expenditure of the base year and multiply the resulting figure by 100.

This method is somewhat like the Laspeyres' Method.

Here,

p1 = prices of the current year

p0 = prices of the base year

q0 = quantity consumed in base year.

### 3.9.5 Family Budget Method:

Under this method, we study the family budgets of a large number of people and estimate the aggregate expenditure of the average family for various items. These values are used as weights. We then convert the current year's prices into price relatives on the basis of the base year's prices. We then multiply these price relatives by the respective values of the commodities of the base year. Now, we need to divide the total of these products by the sum of the weights. This method is similar to the weighted average of price relative method. Its formula is:

Consumer Price Index = ∑PW∑W

Where,P = \(\frac{p1}{p0}\) x 100

V = Value weights or p0q0

Uses of Consumer Price Index Number

We use it to develop economic policy and also to evaluate the real earnings.

It is also helpful in measuring the purchasing power of the consumer. The formula for measuring the purchasing power is:

Purchasing Power = 1ConsumerPriceIndex x 100

It is also used in the process of deflating. The formula to express the process of deflating is:

Real Wage = MoneyValueConsumerPriceIndex x 100

It is also useful in the negotiation of wages and wage contracts. It is also used in the calculation of Dearness Allowance.

**Solved Example on Aggregate Expenditure**

From the following information calculate the changes in the cost of living of people of Indore in 2018 in comparison with 2017.

| | Food | Clothing | Rent | Education | Others |
|---|---|---|---|---|---|
| Expenses | 40% | 20% | 10% | 15% | 15% |
| Price in 2017 | 200 | 120 | 70 | 100 | 50 |
| Price in 2018 | 220 | 150 | 80 | 120 | 70 |

Answer:

Calculation of Cost of Living

| Items | Expenses % (W) | Price in 2017 (p0) | Price in 2018 (p1) | P | PW |
|---|---|---|---|---|---|
| Food | 40 | 200 | 220 | 220/200 x 100 | 220/200 x 100 x 40 = 4400 |
| Clothing | 20 | 120 | 150 | 150/120 x 100 | 150/120 x 100 x 20 = 2500 |
| Rent | 10 | 70 | 80 | 80/70 x 100 | 80/70 x 100 x 10 = 1143 |
| Education | 15 | 100 | 120 | 120/100 x 100 | 120/100 x 100 x 15 = 1800 |
| Others | 15 | 50 | 70 | 70/50 x 100 | 70/50 x 100 x 15 = 2100 |
| | ∑W = 100 | | | | ∑PW = 11943 |

Consumer Price Index = ∑PW/∑W = 11943/100 = 119.43

We can hence conclude that the cost of living has increased by 19.43% in 2018 as compared to 2017.

## 3.9.6 Method of Moving Averages:

This is a simple method in which we take the arithmetic average of the given times series over a certain period of time. These average move over period and are hence called as moving averages. The time interval for the average is taken as 3 years, 4 years or 5 years and so on. The average are thus called as 3 yearly, 4 yearly and 5 yearly moving average. The moving averages are useful in smoothing the fluctuations caused to the variable. Obviously larger the time interval of the average more is the smoothing. We shall study the odd yearly (3 and 5) moving average first and then the 4 yearly moving average.

### 3.9.6.1 Odd Yearly Moving Average

In this method the total of the value in the time series is taken for the given time interval and is written in front of the middle value. The average so taken is also written in front of this middle value. This average is the trend for that middle year. The process is continued by replacing the first value with the next value in the time series and so on till the trend for the last middle value is calculated. Let us understand this with example:

Example 1:

Find 3 years moving averages and draw these on a graph paper. Also represent the original time series on the graph.

| Year | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|
| Production (in thousand unit) | 12 | 15 | 20 | 18 | 25 | 32 | 30 | 40 | 44 |

Solution:

| Year | Production (in thousand unit) | 3 Years Total | 3yearly Moving Average |
|---|---|---|---|
| 1999 | 12 | | |
| 2000 | 15 | 12 + 15 + 20 = 47 | 47 / 3 = 15.6 |
| 2001 | 20 | 15 + 20 + 18 = 53 | 53 / 3 = 17.6 |
| 2002 | 18 | 20 + 18 + 25 = 63 | 63 / 3 = 21.0 |
| 2003 | 25 | 18 + 25 + 32 = 75 | 75 / 3 = 25.0 |
| 2004 | 32 | 25 + 32 + 30 = 87 | 87 / 3 = 29.0 |
| 2005 | 30 | 32 + 30 + 40 = 102 | 102 / 3 = 34.0 |
| 2006 | 40 | 30 + 40 + 44 = 114 | 114 / 3 = 38.0 |
| 2007 | 44 | | |

We calculate arithmetic mean of first three observations viz. 12, 15 and 20, then we delete 12 and consider the next one so that now, average of 15, 20 and 18 is calculated and so on. These averages are placed against the middle year of each group, viz. the year 2000, 2001 and so on. Note moving averages are not obtained for the year 1999 and2007.

Example 2:

Determine the trend of the following time series using 5 yearly moving averages.

| Year | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Exports in'000 Rs | 78 | 84 | 80 | 83 | 86 | 89 | 88 | 90 | 94 | 93 | 96 |

Solution: The time series is divided into overlapping groups of five years, their 5 yearly total and average are calculated as shown in the following table.

| Year | Export (Y) | 5 – yearly total (T) | 5 – yearly moving average: (T/5) |
|---|---|---|---|
| 1981 | 78 | | |
| 1982 | 84 | | |
| 1983 | 80 | | |
| 1984 | 83 | 78+84+80+83+86 = 411 | 411 / 5 = 82.2 |
| 1985 | 86 | 84+80+83+86+89 = 422 | 422 / 5 = 84.4 |
| 1986 | 89 | 80+83+86+89+88 = 426 | 85.2 |
| 1987 | 88 | 83+86+89+88+90 = 436 | 87.2 |
| 1988 | 90 | 86+89+88+90+94 = 447 | 89.4 |
| 1989 | 94 | 89+88+90+94+93 = 454 | 90.8 |
| 1990 | 93 | 88+90+94+93+96 = 461 | 92.2 |
| 1991 | 96 | | |

**Observations:**

In case of the 5 – yearly moving average, the total and average for the first two and the last two in the time series is not calculated. Thus, the moving average of the first two and the last two years in the series cannot be computed.

To find the 3 – yearly total (or 5 – yearly total) for a particular years, you can subtract the first value from the previous year‘s total, and add the next value so as to save your time!

### 3.9.6.2 Even yearly moving averages:

In case of even yearly moving average the method is slightly different as here we cannot find the middle year of the four years in consideration. Here we find the total for the first four years and place it between the second and the third year value of the variable. These totals are again sunned into group of two, called as centered total and is placed between the two totals. The 4 – yearly moving average is found by dividing these centered totals by 8. Let us understand this method with an example

Example 4:

Find the moving average of length 4 for the following data. Represent the given data and the moving average on a graph paper.

| Year | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sales (in thousand unit) | 60 | 69 | 81 | 86 | 78 | 93 | 102 | 107 | 100 | 109 |

Solution: We prepare the following table.

| Year | Sale (in thousand unit) | 4 Yearly Totals | Centred Total | Moving Avg. Central = Total / 8 |
|---|---|---|---|---|
| 1998 | 60 | | | |
| 1999 | 69 | | | |
| | | 60 + 69 + 81 + 86 = 296 | | |
| 2000 | 81 | | 296 + 314 = 610 | 76.25 |
| | | 69 + 81 + 86 + 78 = 314 | | |
| 2001 | 86 | | 314 + 338 = 652 | 81.5 |
| | | 81 + 86 + 78 + 93 = 338 | | |
| 2002 | 78 | | 338 + 359 = 697 | 87.125 |
| | | 86 + 78 + 93 + 102 = 359 | | |
| 2003 | 93 | | 359 + 380 = 739 | 92.375 |
| | | 78 + 93 + 102 + 107 = 380 | | |
| 2004 | 102 | | 380 + 402 = 782 | 97.75 |
| | | 93 + 102 + 107 + 100 = 402 | | |
| 2005 | 107 | | 402 + 418 = | 102.5 |

| | | | | |
|---|---|---|---|---|
| | | | 820 | |
| | | 102 + 107 + 100 + 109 = 418 | | |
| 2006 | 100 | | | |
| 2007 | 109 | | | |

Note that 4 yearly total are written between the years 1999-2000, 2000- 01, 2001-02 etc. and the central total are written against the years 2000, 2001, 2002 etc. so also the moving average are considered w.r.t. years; 2000, 2001 and so on. The moving averages are obtained by dividing the certain total by 8.

The graph of the given set of values and the moving averages against time representing the trend component are shown below. Note that the moving averages are not obtained for the years 1998, 1999, 2006 and 2007. (i.e., first and last two extreme years).

When the values in the time series are plotted, a rough idea about the type of trend whether linear or curvilinear can be obtained. Then, accordingly a linear or second degree equation can be fitted to the values. In this chapter, we will discuss linear trend only.

### 3.9.7 Least Squares Method:

Let $y = a + bx$ be the equation of the straight line trend where a, b are constant to be determined by solving the following normal equations,

$\sum y = na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

where y represents the given time series.

We define x from years such that $\sum x = 0$. So substituting $\sum x = 0$ in the normal equation and simplifying, we get

Using the given set of values of the time series, a, b can be calculated and the straight line trend can be determined as y = a + bx. This gives the minimum sum of squares line deviations between the original data and the estimated trend values. The method provides estimates of trend values for all the years. The method has mathematical basis and so element of personal bias is not introduced in the calculation. As it is based on all the values, if any values are added, all the calculations are to be done again.

### 3.9.7.1 Odd number of years in the time series

When the number of years in the given time series is add, for the middle year we assume the value of x = 0. For the years above the middle year the value given to x are …, -2, -1 while those after the middle year are values 1,2, … and so on.

### 3.9.7.2 Even number of years in the time series

When the number of years in the time series is even, then for the upper half the value of x are assumed as…., -5, -3, -1. For the lower half years, the values of x are assumed as 1, 3, 5, …. And so on.

Example 5:

Fit a straight line trend for the following data giving the annual profits (in lakhs of Rs.) of a company. Estimate the profit for the year 1999.

| Years | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|---|---|
| Profit | 30 | 34 | 38 | 36 | 39 | 40 | 44 |

Solution: Let y = a + bx be the straight line trend.

The number of years is seven, which is sold. Thus, the value of x is taken as 0 for the middle years 1995, for upper three years as -3, -2, -1 and for lower three years as 1, 2, 3.

The table of computation is as shown below:

| Years | Profit (y) | x | xy | $x^2$ | Trend Value: Yt |
|---|---|---|---|---|---|

| | | | | | = a + bx |
|---|---|---|---|---|---|
| 1992 | 30 | -3 | -90 | 9 | 31.41 |
| 1993 | 34 | -2 | -68 | 4 | 33.37 |
| 1994 | 38 | -1 | -38 | 1 | 35.33 |
| 1995 | 36 | 0 | 0 | 0 | 37.29 |
| 1996 | 39 | 1 | 39 | 1 | 39.25 |
| 1997 | 40 | 2 | 80 | 4 | 41.21 |
| 1998 | 44 | 3 | 132 | 9 | 43.17 |
| Total | $\sum y = 261$ | $\sum x = 0$ | $\sum xy = 55$ | $\sum x^2 = 28$ | |

From the table: n = 7, $\sum xy = 55$, $\sum x2 = 28$, $\sum y = 261$

b=1.96 and a=37.29

Thus, the straight line trend is y = 37.29 + 1.96x.

The trend values in the table for the respective years are calculated by substituting the corresponding value of x in the above trend line equation.

For the trend value for 1992: x = -3:

y1992 = 37.29 + 1.96 (-3) = 37.29- 5.88 = 31.41

Similarly, all the remaining trend values are calculated.

(A short-cut method in case of odd number of years to find the remaining trend values once we calculate the first one, is to add the value of b to the first trend value to get the second trend value, then to the second trend value to get the third one and so on. This is because the difference in the values of x is 1.)

To estimate the profit for the years 1999 in the trend line equation, we substitute the prospective value of x, if the table was extended to 1999. i.e. we put x = 4, the next value after x = 3 for the year 1998.

$y_{1999}$ = 37.29 + 1.96 (4) = 45.13

Therefore the estimated profit for the year 1999 is Rs. 45.13 lakhs.

Example 6:

Fit a straight line trend to the following data. Draw the graph of the actual time series and the trend line. Estimate the sales for the year 2007.

| Year | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|---|---|---|---|
| Sales in'000 Rs | 120 | 124 | 126 | 130 | 128 | 132 | 138 | 137 |

Solution: let y = a + bx be the straight line trend.

The number of years in the given time series is eight, which is an even number. The upper four years are assigned the values of x as 1, 2, 3, and 7. Note that ere the difference between the values of x is 2, but the sum is zero.

| Years | Profit (y) | X | Xy | $X_2$ | Trend Value: $Y_t$ = a + bx |
|---|---|---|---|---|---|
| 1998 | 120 | -7 | -840 | 49 | 120.84 |
| 1999 | 124 | -5 | -620 | 25 | 123.28 |
| 2000 | 126 | -3 | -378 | 9 | 125.72 |
| 2001 | 130 | -1 | -130 | 1 | 128.16 |
| 2002 | 128 | 1 | 128 | 1 | 130.06 |
| 2003 | 132 | 3 | 396 | 9 | 133.04 |
| 2004 | 138 | 5 | 390 | 25 | 135.48 |
| 2005 | 137 | 7 | 359 | 49 | 137.92 |
| Total | ∑y = 1035 | ∑x = 0 | ∑xy = 205 | $\sum x^2$ = 168 | |

From the table: n = 8, ∑xy = 205, $\sum x^2$ = 168, ∑y = 1035

b=1.22 and a=129.38

Thus, the straight line trend is y = 129.38 + 1.22x.

The trend values in the table for the respective years are calculated by substituting the corresponding value of x in the above trend line equation.

For the trend value for 1998: x = -7:

y1998 = 129.38 + 1.22 (-7) = 129.38- 8.54 = 120.84

Similarly, all the remaining trend values are calculated.

(A short-cut method in case of even number of years to find the remaining trend values once we calculate the first one, is to add twice the value of b to the first trend value to get the second trend value, then to the second trend value to get the third one and so on. This is because the difference in the values of x is 2. In this example we add 2 x 1.22 = 2.44)

Estimation:

To estimate the profit for the years 2007 in the trend line equation, we substitute the prospective value of x, if the table was extended to 2007. i.e. we put x =11, the next value after x = 9 for the year 2006 and x = 7 for 2005.

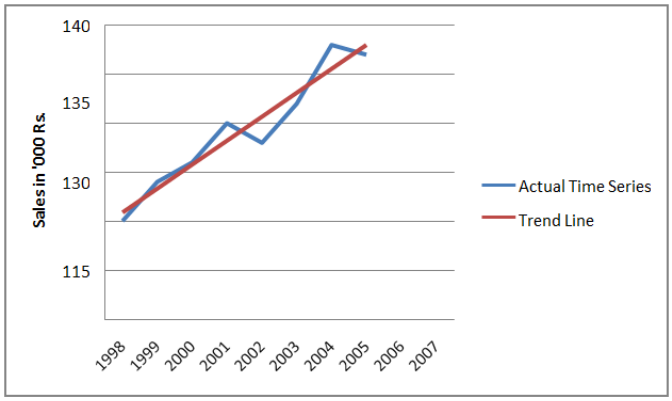
y2007 = 129.38 + 1.22 (11) = 142.8

Therefore the estimated profit for the year 2007 is Rs. 1,42,800.

Now we draw the graph of actual time series by plotting the sales against the corresponding year, the period is taken on the X-axis and the sales on the Y-axis. The points are joined by straight lines. To draw the trend line it is enough to plot any two point (usually we take the first and the last trend value) and join it by straight line.

To estimate the trend value for the year 2007, we draw a line parallel to Y-axis from the period 2007 till it meet the trend line at a point say A. From this point we draw a line parallel to the X-axis till it meet the Y-axis at point say B. This point is our

estimate value of sales for the year 2007. The graph and its estimate value (graphically) is shown below:




From the graph, the estimated value of the sales for the year 2007 is 142 i.e. Rs 1,42,000 (approximately).

# Unit 4

# Probability

## 4.1 Introduction

- Concept of random experiment/trial and possible outcomes.
- Sample Space: Discrete and Continuous.
- Events and different types (mutually exclusive, exhaustive and complimentary).
- Algebra of Events.
- Classical definition of Probability, Addition theorem (without proof),
- Conditional probability.
- Multiplication theorem and Independence of Events: $P( A \cap B ) = P(A) P(B)$.

## 4.2 Concept of random experiment/trial and possible outcomes

### 4.2.1 Random Experiments

Before rolling a die you do not know the result. This is an example of a random experiment. In particular, a random experiment is a process by which we observe something uncertain. After the experiment, the result of the random experiment is known. An outcome is a result of a random experiment. The set of all possible outcomes is called the sample space. Thus in the context of a random experiment,

the sample space is our universal set. Here are some examples of random experiments and their sample spaces:

- Experiment 1: Tossing a coin
    - Possible outcomes are head or tail.
    - Sample space, S = {head, tail}
- Experiment 2: Tossing a die
    - Possible outcomes are the numbers 1, 2, 3, 4, 5, and 6
    - Sample space, S = {1, 2, 3, 4, 5, 6}

When we repeat a random experiment several times, we call each one of them a trial. Thus, a trial is a particular performance of a random experiment. In the example of tossing a coin, each trial will result in either heads or tails. Note that the sample space is defined based on how you define your random experiment. For example,

## Sample Space

A **sample space** is the set of all possible outcomes in an experiment.

Example:
Two coins are tossed. Represent the sample space for this experiment by making a list, a table, and a tree diagram.
(H – Head, T – Tail)

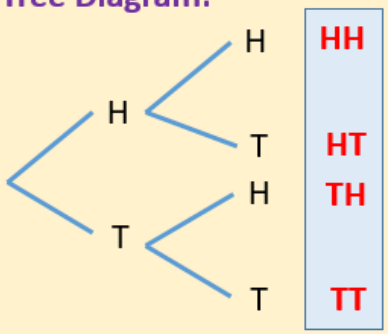
**List:**

HH HT TH TT

**Table:**

| | H | T |
|---|---|---|
| H | HH | HT |
| T | TH | TT |

**Tree Diagram:**




The sample space is {HH, HT, TH, TT}

Our goal is to assign probability to certain events. For example, suppose that we would like to know the probability that the outcome of rolling a fair die is an even number. In this case, our event is the set E = {2, 4, 6}. If the result of our random experiment belongs to the set E, we say that the event E has occurred. Thus an event is a collection of possible outcomes. In other words, an event is a subset of the sample space to which we assign a probability. Although we have not yet discussed how to find the probability of an event, you might be able to guess that the probability of {2, 4, 6} is 50 percent which is the same as 1/2 in the probability theory convention.

#### 4.2.1.1 Union and Intersection:

If A and B are events, then A∪B and A∩B are also events.

By remembering the definition of union and intersection:

We observe that A∪B occurs if A or B occur.

Similarly, A∩B occurs if both A and B occur.

Similarly, if A1, A2, ⋯, An are events,

then the event A1∪A2∪A3⋯∪An occurs if at least one of A1, A2, ⋯, An occurs.

The event A1∩A2∩A3⋯∩An occurs if all of A1, A2, ⋯, An occur.

It can be helpful to remember that the key words "or" and "at least" correspond to unions and the key words "and" and "all of" correspond to intersections.

#### 4.2.1.2 Discrete Sample Space

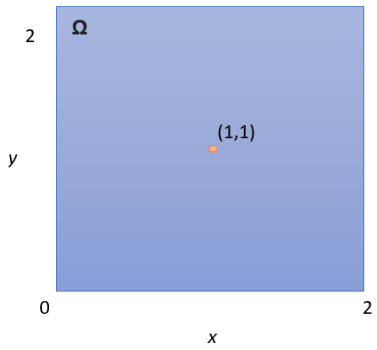
Sample spaces introduced in early probability classes are typically discrete. That is, they are made up of a finite (fixed) amount of numbers. For example, if you roll a die, the sample space (S) is [1, 2, 3, 4, 5, 6]. Rolling a twenty-sided die or choosing a card from a deck all produce discrete sample spaces.

#### 4.2.1.3 Continuous sample space

A continuous sample space is based on the same principles, but it has an infinite number of items in the space. In other words, you can't write out the space in the same way that you would write out the sample space for a die roll. With a continuous sample space, you would still be writing numbers long after the sun has imploded into a black hole.

**Continuous Sample Space Example**

The following image shows a continuous sample space — area 2 units high and 2 units wide. Imagine that the space represents a game where a dart is dropped into the space and lands on a random spot. What are the odds of the dart landing on the spot pictured?




It may surprise you to learn that the odds of the ball landing on either of those two spots are zero. In fact, the odds of the dart falling in any spot at all equals zero. Let's say the dart lands at (1, 1). It could also land a hair to the right, at (1.001, 1). Or, an even tinier fraction to the right, at (1.0000000001, 1). If you try and write down all of the possible places the dart could land, you won't be able to, because there are an infinite number of spaces the ball could land on. If you have a problem with this logic, you aren't alone. It's an example of how we think the world works, and how it actually works. A famous example of how infinity messes with logic is Zeno's paradox.

A more modern approach is to use the problem in terms of limits. In simple terms, a limit is the number very close to the one you're measuring. In the case of all these zeros (1.001, 1.0000001, 1. 00000000000001 etc.), you could say they all are very close to 1, giving you far fewer points to deal with. For the above problem, you could come up with a very reasonable probability if you could all the possible whole numbered coordinates: (0, 1), (0, 2), (0, 3) and so on.

### 4.2.2 Mutually exclusive events

Two or more events are said to be mutually exclusive if they don't have any element in common. i.e. if, the occurrence of one of the events prevents the occurrence of the others then those events are said to be mutually exclusive. In other words, mutually exclusive events are those events that do not occur at the same time. For example, when a coin is tossed then the result will be either head or tail, but we cannot get both the results. Such events are also called disjoint events

since they do not happen simultaneously. If A and B are mutually exclusive events then its probability is given by,

Probability of Disjoint (or) Mutually Exclusive Event = P ( A ∩ B) = 0

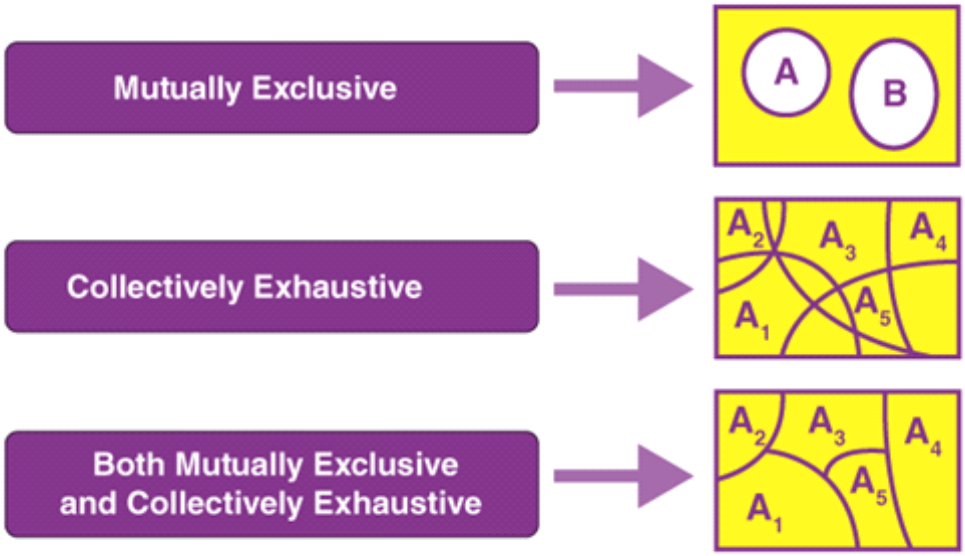
### 4.2.3 Mutually exhaustive events

Two events are said to be mutually exhaustive if there is a certainty of occurring at least one of those two events. i.e. one of those events will definitely happen. In other words, mutually exhaustive events are those events that combine to form entire sample space. For example, when a coin is tossed then the result will be either head or tail, so combination of both will give us entire sample space. If A and B are mutually exhaustive events then its probability is given by,

Probability of Mutually Exhaustive Event = P ( A ∪ B) = 1.

### 4.2.4 Collectively Exhaustive Events

In probability theory, a set of events can be either jointly or collectively exhaustive if at least one of the events must occur for sure. We can verify that because the outcomes comprise the entire range of possible outcomes, i.e. sample space for an experiment. For example, when throwing an unbiased six-sided die, the outcomes 1, 2, 3, 4, 5, and 6 are collectively exhaustive. Similarly, when a coin is tossed, the outcome can either be heads or tails. Therefore, considering each can occur during an experiment, they are both described as exhaustive events. However, the union of all those events comprises the sample space of that experiment known as the exhaustive events.




### 4.2.5 Complementary Events

Two events are said to be complementary when one event occurs if and only if the other does not. The probabilities of two complimentary events add up to 1.

For example, rolling a 5 or greater and rolling a 4 or less on a die are complementary events, because a roll is 5 or greater if and only if it is not 4 or less. The probability of rolling a 5 or greater is $\frac{2}{6} = \frac{1}{3}$, and the probability of rolling a 4 or less is $\frac{4}{6} = \frac{2}{3}$. Thus, the total of their probabilities is $\frac{1}{3} + \frac{2}{3} = \frac{3}{3} = 1$.

### 4.2.6 Classical definition of Probability

Probability is a statistical concept that measures the likelihood of something happening. Classical probability is the statistical concept that measures the likelihood of something happening, but in a classic sense, it also means that every statistical experiment will contain elements that are equally likely to happen.

The typical example of classical probability would be a fair dice roll because it is equally probable that you will land on any of the 6 numbers on the die: 1, 2, 3, 4, 5, or 6.

Another example of classical probability would be a coin toss. There is an equal probability that your toss will yield a heads or tails result.

## 4.2.7 Addition theorem (without proof)

If A and B are any two events then,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If A, B and C are any three events then,

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

**Example**

If P(A) = 0.37 , P(B) = 0.42 , P (A ∩ B) = 0.09 then find P (A U B) .

Solution

P(A) = 0.37 , P(B) = 0.42 , P (A ∩ B) = 0.09

P (A U B) = P (A) + P (B ) − P (A ∩ B)

P (A U B) = 0.37 + 0.42 − 0.09 = 0. 7

**Example**

A card is drawn from a pack of 52 cards. Find the probability of getting a king or a heart or a red card.

**Solution**

Total number of cards = 52; n(S) = 52

Let A be the event of getting a king card. n(A) =4

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52}$$

Let B be the event of getting a heart card. n(B) =13

$$P(B) = \frac{n(B)}{n(S)} = \frac{13}{52}$$

Let C be the event of getting a red card. n(C) =26

$$P(C) = \frac{n(C)}{n(S)} = \frac{26}{52}$$

P (A ∩ B) = P (getting heart king) = 1/52

P (B ∩ C) = P (getting red and heart)) = 13/52

P (A ∩ C) = P (getting red king) = 2/52

P (A ∩ B ∩ C) = P (getting heart, king which is red) = 1/52

Therefore, required probability is

P(A ∪ B ∪ C) = P(A) + P(B) + P(C) − P(A∩B) − P(B∩C) − P(A∩C) + P(A∩B∩C)

= 4/52 + 13/52 + 26/52 – 1/52 – 13/52 – 2/52 + 1/52

= 28/52

= 7/13

## 4.2.8 Conditional probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional

probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding or conditional event.

**For example:**

Event A is that an individual applying for college will be accepted. There is an 80% chance that this individual will be accepted to college.

Event B is that this individual will be given dormitory housing. Dormitory housing will only be provided for 60% of all of the accepted students.

P (Accepted and dormitory housing) = P (Dormitory Housing | Accepted) P (Accepted) = (0.60)*(0.80) = 0.48.

A conditional probability would look at these two events in relationship with one another, such as the probability that you are both accepted to college, and you are provided with dormitory housing.

## 4.2.9 Multiplication theorem

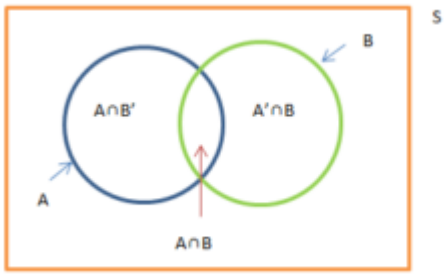
**Theorem 1**

For two events A and B,

$P(A \cap B) = P(A)\, P(B \mid A), P(A) > 0.$

or, $P(A \cap B) = P(B)\, (A \mid B), P(B) > 0.$

Here, P(B | A) represents the conditional probability of occurrence of B when the event A had already occurred. P(A | B) represents the conditional probability of occurrence of A when the event B had already happened.




**Proof:**

From the concept of conditional probability, we have

$P(A \mid B) = P(A \cap B) / P(B)$.

Re-writing the above, we have, $P(A \cap B) = P(B)\, P(A \mid B)$.

Similary, $P(B \mid A) = P(A \cap B) / P(A)$.

$\Rightarrow P(A \cap B) = P(A)\, P(B \mid A)$.

The mathematical theorem on probability shows that the probability of the simultaneous occurrence of two events A and B is equal to the product of the probability of one of these events and the conditional probability of the other, given that the first one has occurred.

**Theorem 2**

For two events A and B such that $P(B) > 0$, $P(A \mid B) \leq P(A)$.

**Proof:**

It is obvious that the number of common outcomes in A and B is either less or equal to the number of outcomes in any of the event.

$n(A \cap B) \leq n(A)$ … (i),

and, $n(B) \leq n(S)$ … (ii)

Dividing (i) and (ii), we get,

$n(A \cap B) / n(B) \leq n(A) / n(S)$

$\Rightarrow P(A \mid B) \leq P(A)$.

#### 4.2.9.1 Multiplication Theorem for Independent Events

The multiplication theorem on probability for dependent events can be extended for the independent events.

From the theorem, we have, $P(A \cap B) = P(A)\, P(B \mid A)$.

If the events A and B are independent, then, $P(B \mid A) = P(B)$.

The above theorem reduces to

$P(A \cap B) = P(A)\, P(B)$.

This shows that the probability that both of these occur simultaneously is the product of their respective probabilities.

#### 4.2.9.2 Extension of Multiplication Theorem of Probability to n Events

For n events A1, A2, … , An, we have

$P(A1 \cap A2 \cap \ldots \cap An) = P(A1)\, P(A2 \mid A1)\, P(A3 \mid A1 \cap A2) \ldots \times P(An \mid A1 \cap A2 \cap \ldots \cap An\text{-}1)$

#### 4.2.9.3 Extension of Multiplication Theorem of Probability to n Independent Events

For n independent events, the multiplication theorem reduces to

$P(A1 \cap A2 \cap \ldots \cap An) = P(A1)\, P(A2) \ldots P(An)$.