Directorate of Distance and Continuing Education

Manonmaniam Sundaranar University

Tirunelveli-627 012, Tamil Nadu.

B.A. ECONOMICS

(First Year)

STATISTICS FOR ECONOMICS – II

JMEC22

Compiled by

Dr. G. Monikanda Prasad

Assistant Professor of Economics
Manonmaniam Sundaranar University
Tirunelveli - 627 012.

STATISTICS FOR ECONOMICS - II

UNIT	Contents
I	**Index Numbers** Index Numbers – Methods – Unweighted and Weighted Index Numbers – Aggregate and Relative Index Numbers – Chain and Fixed based Index Numbers – Test of Adequacy of Index Numbers – Wholesale Price Index – Consumer Price Index – Cost of Living Index.
II	**Time Series Analysis** Definition – components and Measurement – Graphic Method – Methods of Semi Average, Moving Average and Method of Least Squares – Use of Time Series Analysis.
III	**Theory of Probability** Key Concepts of Probability – Importance – Theorems of Probability: Addition, Multiplication and Bayes' Theorem – Discrete and Continuous Random Variables – Theoretical Distributions – Binomial, Poisson and Normal – Properties – Uses and Application**.**
IV	**Sampling** Sampling – Census and Sample Method – Theoretical Basis of Sampling – Methods of sampling – Random and Non _ Random Sampling – Size of Sample – Merits and Limitations of Sampling and Non – Sampling Errors.
V	**Testing of Hypothesis** Hypothesis Testing – Meaning, Types, Source and Functions of Hypothesis – Test: Null and Alternative Hypothesis – Type – I and Type – II Errors-'t' Test – Paired 't' – test –Chi-Square test, 'F' test – Analysis of Variance – One way and Two – way ANOVA.
	Text books
1	S.P.Gupta, (2017) "Statistical Methods", Sultan Chand & Sons.
2	Anderson, Sweeney and Williams (2012), "Statistics for Business and Economics
3	Pillai R.S.N. &Bagavathi V (2012), " Statistics: Theory and Practice" S. Chand & Company Ltd. New Delhi.
4	Dr.T.K.V.Iyengar, Dr.B.Krishna G and hi S.Ranganantham, Dr. M.V.S.S.N Prasad, Probability and Statistics, S. Chand and Co, 2020.
5	Prof S.G. Vekatachalapathy and Dr.H. Premraj (2018) Statistical Methods Margham Publications.

UNIT – I
INDEX NUMBER

Introduction

Historically, the first index was constructed in 1764 to compare the Italian price index in 1750 with the price level in 1500. Though originally developed for measuring the effect of change in prices. Index numbers have today become one of the most widely used statistical devices and there is hardly any field where they are not used. Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular period compared to the previous period as disclosed by index numbers. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary tendencies. In fact, they are described as 'barometers if economic activity'. i.e., if one wants to get an idea as to what is happening to an economy, he should look to important indices like the index number of industrial production, agricultural production, business activity, etc.

1. "Index numbers are devices for measuring differences in the magnitude of a group of related variables – Croxton & Cowden

Use of index numbers

Index numbers are indispensable tools of economic and business analysis. Their significance can be best appreciated by the following points:

(i) They help in framing suitable policies. Many of the economic and business policies are guided by index numbers. For example, while deciding the increase in dearness allowance of the employees, the employers have to depend primarily upon the cost of living index. If wages and salaries are not adjusted in accordance with the cost of living, very often it leads to strikes and lock-outs which in turn cause considerable waste of resources. The index numbers provide some guideposts that one can use in making decisions.

(ii) They reveal trends and tendencies. Since index numbers are most widely used for measuring changes over a period of time, the time series so formed enable us to study the general trends of the phenomenon under study. For example, by examining index number of imports for India for the last 8-10 years we can say that our imports are showing an upward tendency, i.e., they are rising year after year. Similarly, by examining the index numbers of industrial production, business activity, etc., for the last few years we can conclude about the trend of production and business activity. By examining the trend of the phenomenon under study we can draw very important conclusions as to how much change is taking place due to the effect of seasonality, cyclical forces, irregular forces, etc. thus index numbers are highly useful in studying the general business conditions.

(iii) They are important in forecasting future economic activity. Index numbers are useful not only in studying the past and present workings of our economy, but they are also important in forecasting future economic activity. Index numbers then are often used in time series analysis, the historical study of long-term trend, seasonal variations and business cycle development, so that business leader may keep pace with changing economic and business conditions and have better information available for decision-making purpose.

(iv) Index numbers are very useful in deflating. Index numbers are highly useful in deflating, i.e., they are used to adjust the original data for price changes, or to adjust wages for cost of living changes and thus transform nominal wages for cost of living changes and thus transform nominal wages into real wages. Moreover, nominal income can be transformed into real income and nominal sales into real sales through appropriate index numbers.

Classification of Index Numbers

Index numbers may be classified in terms of what they measure, in economics and business the classifications are (1) price; (2) quantity; (3) value; and (4) special purpose.

Only price and quantity index numbers are discussed in detail. The others will be mentioned, but without detail, of how to construct them since both value and special purpose index numbers do not offer new problems in construction. Since the method of construction of various types of index number can be understood if the technique of constructing price index number is clear, we shall devote major attention to them.

Problems in the construction of index numbers

Before constructing index numbers a careful thought must be given to the following problems:

1. The Purpose of the Index: At the very outset the purpose of constructing the index must be very clearly decided-what the index is to measure and why? There is no all-purpose index. Every index is of limited and particular use. Thus, a price index that is intended to measure consumers price must not include wholesale price. And if such an index is intended to measure the cost of living of poor families, great care should be taken not to include goods ordinarily used by middle class and upper-income groups. Failure to decide clearly the purpose of the index would lead to confusion and wastage of time with no fruitful results. Other problems such as the base year, the number of commodities to be included, the price of the commodities, etc., are decided in the light of the purpose for which the index is being constructed.

2. Selection of a Base Period: whenever index numbers are constructed a reference is made to some base period. The base period of an index number (also called the reference period) is the period against which comparisons are made.

(i) The base period should be a normal one. The period that is selected as base should be normal, i.e., it should be free from abnormalities like wars, earthquakes, famines, booms, depression, etc. however, at times it is really difficult to select a year which is normal in all respects-a year which is normal in one respect may be abnormal in another. To solve this problem an average of a number of years, 3 or 4, may be taken as the base. The process of averaging will reduce the effect of extremes. Thus the average of the period from 2000 to 2002 may be considered normal whereas no individual year in that span can be considered normal.

(ii) The base period should not be too distant in the past. It is desirable to have an index based on a fairly recent period, since comparisons with a familiar set of circumstances are more helpful than comparisons with vaguely remembered conditions. For example, for deciding increase in dearness allowance at present there is no advantage in taking 1970 or 1980 as the base: the comparison should be with the preceding year or the year after which dearness allowance has not been revised.

(iii)Fixed base or chain base. While selecting the base a decision has to be made as to whether the base shall remain fixed or not, i.e., whether we have a fixed base or chain base index. In the fixed base method, the year or the period of years to which all other prices are related is constant for all times. On the other hand, in the chain base method the prices of a year are linked with those of the preceding year and not with the fixed year. Naturally the chain base method gives a better picture than what is obtained by the fixed base method. However, much would depend upon the purpose of constructing the index.

3. **Selection of Number of Items:** The items included in an index should be determined by the purpose for which the index is constructed. Every item cannot be included while constructing an index number and hence one has to select a sample. For example, while constructing a price index it is impossible to include each and every commodity. Hence it is necessary to decide what commodities to include. The commodities should be selected in such a manner that they are representative of the tastes. Habits and customs of the people for whom the index is meant. Thus in a consumer price index for working class, items like scooters, motor cars, refrigerator, cosmetics, etc., find to place. A decision must also be made on the number of commodities to be included and their qualities. Here we should note that the larger the number of commodities included, the more representative shall be the index but at the same time the greater shall be the cost and the time taken. The purpose of the index shall help in deciding the number of commodities. Thus, in a general price index a larger number of commodities shall have to be included as compared to a specific purpose index as the index number of the prices of food grains or industrial raw materials.

4. **Price quotations:** After the commodities have been selected, the next problem is to obtain price quotations for these commodities. It is a well known fact that price of many commodities vary from place to place and even from shop to shop in the same market. It is impracticable to obtain price quotations from all the places where a commodity is dealt in. A selection must be made of representative places and persons. These places should be those which are well known for trading for that particular commodity. After the places from where the price quotations are to be obtained is decided, the next thing is to appoint some person or institution who can supply price quotations as and when required. Great care must be exercised to see that the price reporting agency is unbiased. In order to check the inaccuracy of price quotations supplied by an agency quotations are obtained from more than one agency. If there is some reliable journal or magazine supplying price quotations then it may be utilized.

5. **Choice of an Average:** since index numbers are specialized averages a decision has to be made as to which particular average should be used for constructing the index. Median, mode and mean are almost never used in the construction of index numbers. Basically a choice has to be made between arithmetic mean and geometric mean. Theoretically speaking, geometric mean is the best average in the construction of index numbers because of the following reasons : (I) in the construction of index numbers we are concerned with rations of change; (ii) geometric mean is less susceptible to major variations as a result of violent fluctuations in the values of the individual items; and (iii)Index numbers calculated by using this average are reversible and therefore, base shifting is easily possible.The geometric mean index always satisfies the time reversal test.

6. **Selection of Appropriate Weights:** The problem of selecting suitable weights is quite important and at the same time quite difficult to decide. The term ‘weight’ refers to the relative importance of the different items in the construction of the index. All item are not of equal importance and hence it is necessary to device some suitable method whereby the varying importance of the different items is taken into account. This is done by allocating weights. Thus, we have broadly two types of indices unweighted indices and weighted indices. In the former case, no specific weights are assigned whereas in the latter case specific weights are unweighted in the strict sense of the term as weights implicitly enter in unweighted indices because we are giving equal importance to all the items and hence weights are unity. It is, therefore, necessary to adopt some suitable method of weighting so that arbitrary and haphazard weights may not affect the results.

7. **Selection of an Appropriate Formula:** A large number of formulae have been devised for constructing the index. The problem very often is that of selecting the most appropriate formula. The choice of the formula would depend not only on the purpose of the index but also on the data available. Prof. Irving Fisher has suggested that an appropriate index is that which satisfies time reversal test and factor reversal test. Theoretically, Fisher's method is considered as "ideal" for constructing index number. However, from a practical point of view there are certain limitations of this index which shall be discussed later. As such, no one particular formula can be regarded as the best under all circumstances. On the basis of this knowledge of the characteristics of the different formulae, a discriminating investigator will choose technical methods adapted to his data and appropriate to his purpose.

Methods of Constructing Index Numbers

A large number of formulae have been devised for constructing index numbers. Broadly speaking they can be grouped under two heads:

(a) Unweighted indices; and

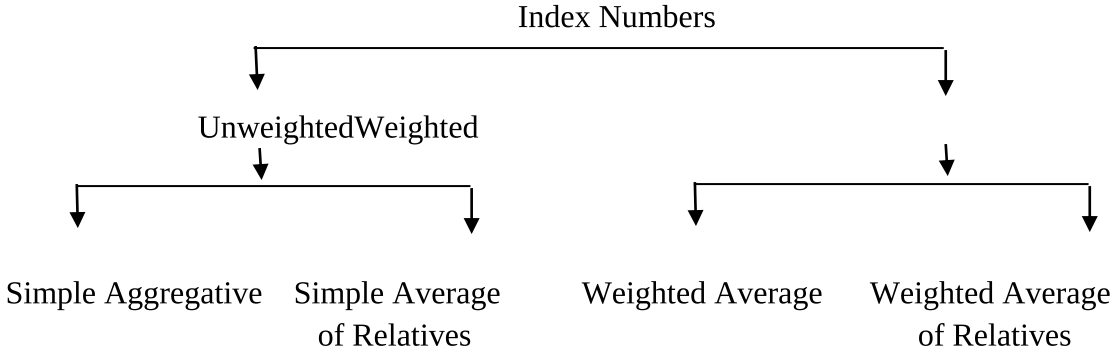
(b) Weighted indices.

In the unweighted indices weight are not expressly assigned whereas in the weighted indices weights are assigned to the various items. Each of these types may be further divided under two heads:

(i) Simple Aggregative, and

(ii) Simple Average of Relatives.

The following chart illustrates the various methods:

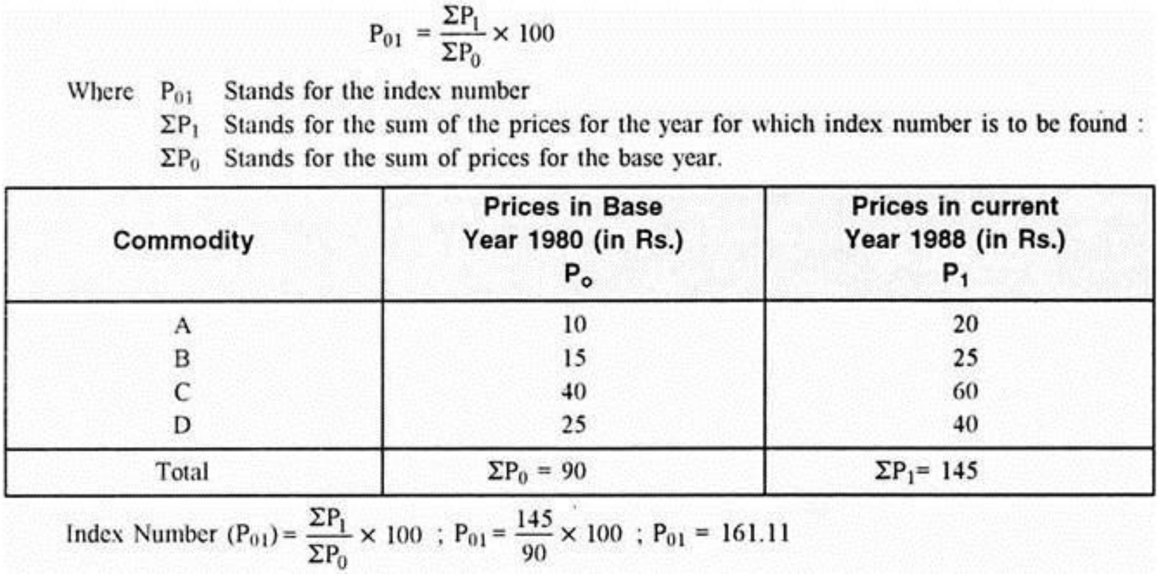


Unweighted Index Numbers

1. **Simple Aggregative Method:** this is the simplest method of constructing index numbers. When this method is used to construct a price index, the total of current year prices for the various commodities in question is divided by the total of base year prices and the quotient is multiplied by 100. Symbolically :

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

$$\sum P_1 = total\,of\ current\ prices$$
$$\sum P_0 = total\,of\ base\ prices$$



$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where P_{01} Stands for the index number
ΣP_1 Stands for the sum of the prices for the year for which index number is to be found :
ΣP_0 Stands for the sum of prices for the base year.

Commodity	Prices in Base Year 1980 (in Rs.) P_o	Prices in current Year 1988 (in Rs.) P_1
A	10	20
B	15	25
C	40	60
D	25	40
Total	$\Sigma P_0 = 90$	$\Sigma P_1 = 145$

Index Number $(P_{01}) = \frac{\Sigma P_1}{\Sigma P_0} \times 100$; $P_{01} = \frac{145}{90} \times 100$; $P_{01} = 161.11$

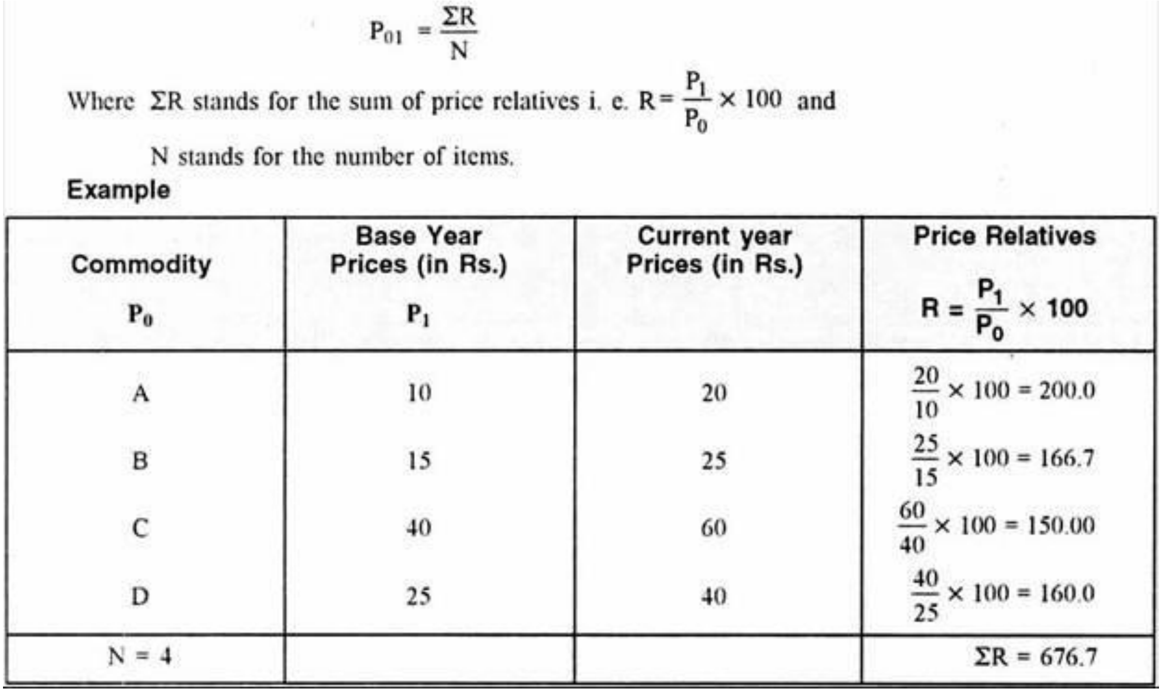
2. Simple Average of Price Relative Method

In this method, average of price relative of commodity, first of all price relatives are obtained for the various items included in the index and then average of these relatives is obtained using any one of the measures of central value.

Steps involved

1. Find price relative for each commodity for the current year using the formula R = (P1 / P0) × 100.
2. Add all price relatives of all the commodities.
3. Divide sum obtained in step 2 by the number of commodities (N).
4. Overall formula for the method is.

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0}\times 100\right)}{N}$$



$$P_{01} = \frac{\Sigma R}{N}$$

Where ΣR stands for the sum of price relatives i. e. $R = \frac{P_1}{P_0} \times 100$ and

N stands for the number of items.

Example

Commodity P_0	Base Year Prices (in Rs.) P_1	Current year Prices (in Rs.)	Price Relatives $R = \frac{P_1}{P_0} \times 100$
A	10	20	$\frac{20}{10} \times 100 = 200.0$
B	15	25	$\frac{25}{15} \times 100 = 166.7$
C	40	60	$\frac{60}{40} \times 100 = 150.00$
D	25	40	$\frac{40}{25} \times 100 = 160.0$
N = 4			ΣR = 676.7

Weighted Index Numbers

It is quite important to meet the needs of any sort of simple or unweighted methods. So, in such a case, we weigh the value of any commodity using any factor that deems fit. This factor is usually the quantity we sell it for during the base year. The categories of these indices are:

Weighted Aggregative Index and Weighted Average of Relatives

Let’s have a close look at the following two indices.

Weighted Aggregative Index Method

We generally use this method to weigh out the price of any commodity. The weighing is done using a very approximate factor. These factors are likely to vary and can be anything. It can be a quantity or it can be the volume that it is selling off for during the base year.

The year not necessarily needs to be the base year but can also be an average of other years or any year in general. Well, the choice of it will totally depend on the importance of the specific year. So, besides the quantity, it is on us about terming the importance of a specific year.

Weighted Aggregative Index generally comes off in the form of percentages. As a result, there are different formulas that we use for the same. Some of them are:

1. Laspeyres Index

Under this type of index, the quantities in the base year are the values of*weights.*

Formula –$(\sum P_nQ_o/\sum P_oQ_o)*100$

2. Passche’s Index

Under this type of Index, the quantities in the current year are the values of *weights.*

Formula – $(\sum P_nQ_n/\sum P_oQ_n)*100$

3. Some of the methods that depend on a typical time period:

Index $(\sum P_nQ_t/\sum P_oQ_n) * 100$, here, the subscript “*t”* symbolizes the typical period of time in years. The quantities of these years are the values of weight.

Note: Using the following formulas, the indices are subject to return the values in the form of percentages.

Marshall-Edgeworth Index

Under this type of index, we take both i.e. the current year as well as the base year into consideration for specifying the methods.

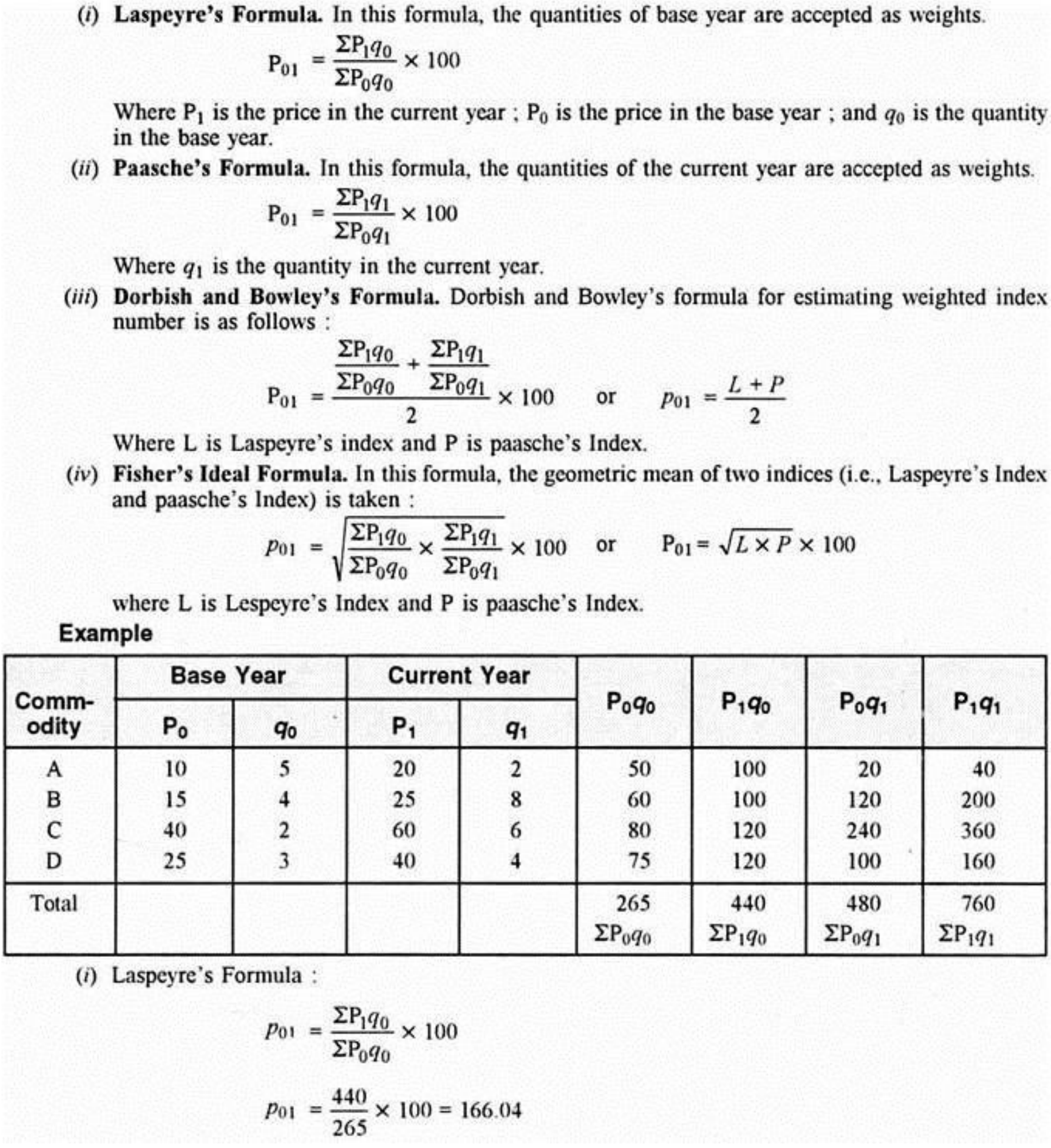
Marshall-Edgeworth Index – $[\sum P_n(Q_o+ Q_n)/\sum P_o(Q_o+ Q_n)] * 100$

4. Fisher's Ideal Price Index

The geometric mean of Laspeyres' and Paasche's is the Fisher's Ideal Price Index.

Formula –$\sqrt{[(\sum P_nQ_o/\sum P_oQ_o)*(\sum P_nQ_n/\sum P_oQ_n)]}* 100$

Some of them are explained below:



(*i*) **Laspeyre's Formula.** In this formula, the quantities of base year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

Where P_1 is the price in the current year ; P_0 is the price in the base year ; and q_0 is the quantity in the base year.

(*ii*) **Paasche's Formula.** In this formula, the quantities of the current year are accepted as weights.

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

Where q_1 is the quantity in the current year.

(*iii*) **Dorbish and Bowley's Formula.** Dorbish and Bowley's formula for estimating weighted index number is as follows :

$$P_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 \quad \text{or} \quad p_{01} = \frac{L + P}{2}$$

Where L is Laspeyre's index and P is paasche's Index.

(*iv*) **Fisher's Ideal Formula.** In this formula, the geometric mean of two indices (i.e., Laspeyre's Index and paasche's Index) is taken :

$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100 \quad \text{or} \quad P_{01} = \sqrt{L \times P} \times 100$$

where L is Lespeyre's Index and P is paasche's Index.

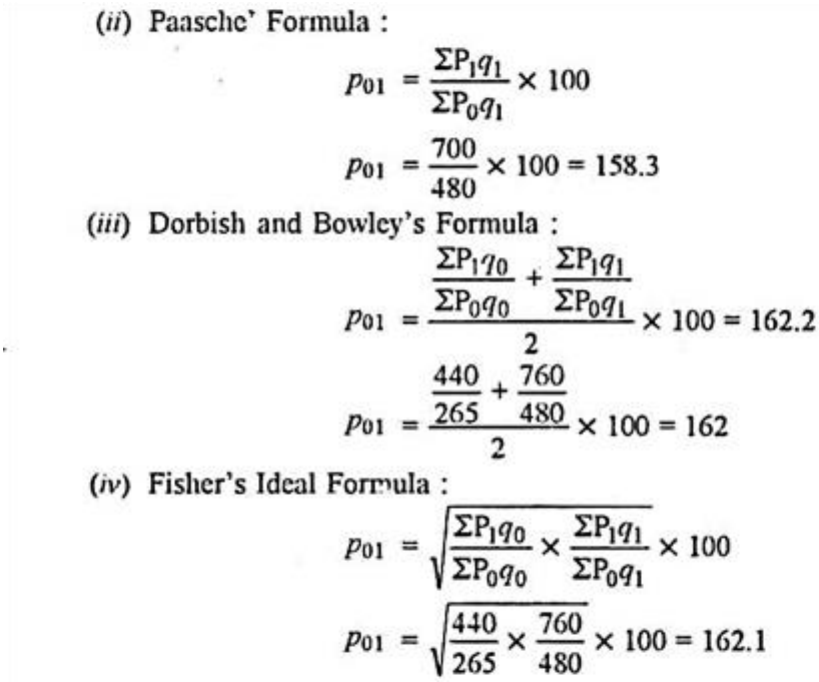
Example

Comm-odity	Base Year		Current Year		P_0q_0	P_1q_0	P_0q_1	P_1q_1
	P_0	q_0	P_1	q_1				
A	10	5	20	2	50	100	20	40
B	15	4	25	8	60	100	120	200
C	40	2	60	6	80	120	240	360
D	25	3	40	4	75	120	100	160
Total					265 $\Sigma P_0 q_0$	440 $\Sigma P_1 q_0$	480 $\Sigma P_0 q_1$	760 $\Sigma P_1 q_1$

(*i*) Laspeyre's Formula :

$$p_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$$

$$p_{01} = \frac{440}{265} \times 100 = 166.04$$



(*ii*) Paaschc' Formula :

$$p_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100$$

$$p_{01} = \frac{700}{480} \times 100 = 158.3$$

(*iii*) Dorbish and Bowley's Formula :

$$p_{01} = \frac{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} + \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}}{2} \times 100 = 162.2$$

$$p_{01} = \frac{\frac{440}{265} + \frac{760}{480}}{2} \times 100 = 162$$

(*iv*) Fisher's Ideal Formula :

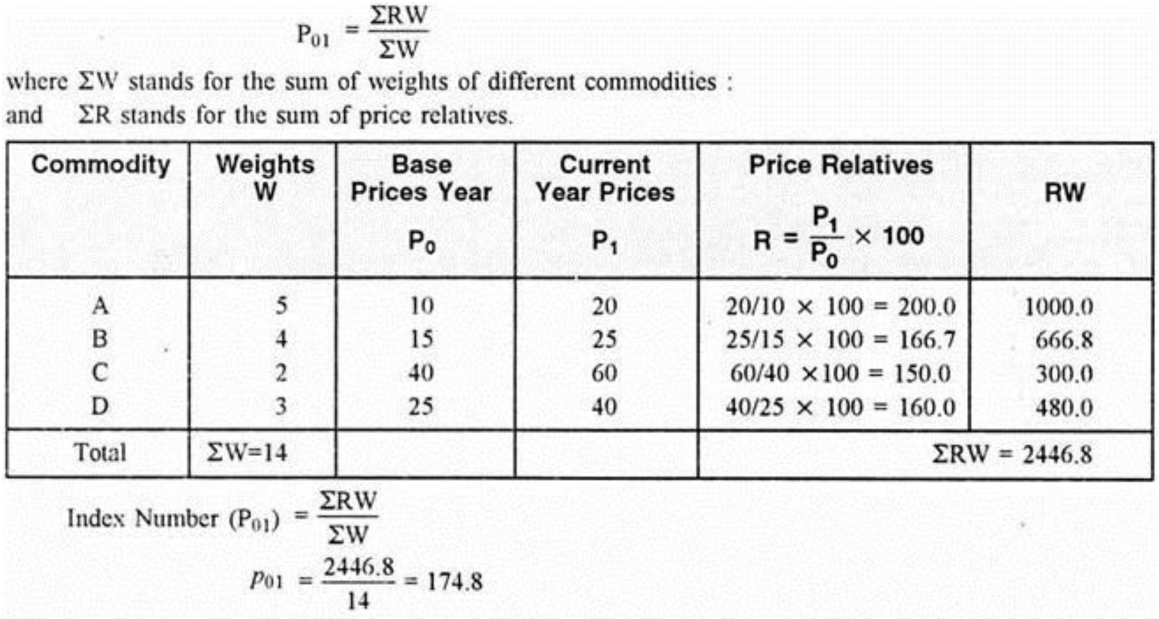
$$p_{01} = \sqrt{\frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1}} \times 100$$

$$p_{01} = \sqrt{\frac{440}{265} \times \frac{760}{480}} \times 100 = 162.1$$

Weighted Average of Relatives

We use the weighted average of relatives to avoid the disadvantage that comes along with the simple average method. Furthermore, the preference is weighted geometric mean but weighted arithmetic mean is used otherwise. Therefore, the representation of the weighted AM using the values of base year weights is:

The price index number is found out with the help of the following formula:



$$P_{01} = \frac{\Sigma RW}{\Sigma W}$$

where ΣW stands for the sum of weights of different commodities :
and ΣR stands for the sum of price relatives.

Commodity	Weights W	Base Prices Year P_0	Current Year Prices P_1	Price Relatives $R = \frac{P_1}{P_0} \times 100$	RW
A	5	10	20	20/10 × 100 = 200.0	1000.0
B	4	15	25	25/15 × 100 = 166.7	666.8
C	2	40	60	60/40 ×100 = 150.0	300.0
D	3	25	40	40/25 × 100 = 160.0	480.0
Total	ΣW=14				ΣRW = 2446.8

$$\text{Index Number } (P_{01}) = \frac{\Sigma RW}{\Sigma W}$$

$$p_{01} = \frac{2446.8}{14} = 174.8$$

There are a number of theoretical as well as practical difficulties in the construction of price index numbers. Moreover, the index number technique itself has many limitations.

(A) Conceptual Difficulties:

The following are the conceptual difficulties during the construction of price index numbers:

1. Vague Concept of Value of Money:

The concept of money is vague, abstract and cannot be clearly defined. The value of money is a relative concept which changes from person to person depending upon the type of goods on which the money is spent.

2. Inaccurate Measurement:

Price index numbers do not measure the changes in the value of money accurately and reliably. A rise or fall in the general level of prices as indicated by the price index numbers does not mean that the price of every commodity has risen or fallen to the same extent.

3. Reflect General Changes:

Price index numbers are averages and measure general changes in the value of money on the average. Therefore, they are not of much significance for the particular individuals who may be affected by the changes in the actual prices quite differently from that indicated by the index numbers.

Limitations of Wholesale Price Index:

The wholesale price index numbers, which are generally used to measure changes in the value of money, suffer from certain limitations:

(a) They do not reflect the changes in the cost of living because retail prices are generally higher than the wholesale prices.

(b) They ignore some of the important items concerning the urban population, such as, expenditure on education, transport, house rent, etc.

(c) They do not take into consideration the changes in the consumers' preferences.

(B) Practical Difficulties:

The practical difficulties in the way of constructing price index numbers, and therefore, in measuring changes in the value of money are as follows:

1. Selection of Base Year:

While preparing the index number, first difficulty arises regarding the selection of base year. The base year should be a normal year. But, it is very difficult to find out a fully normal year free from any unusual happening. There is every possibility that the selected base year may be an abnormal year, or a distant year, or may be selected by an immature or biased person.

2. Selection of Items:

The selection of the representative commodities is the second difficulty in the construction of index numbers:

(a) With the passage of time the quality of the product may change ; if the quality of a product changes in the year of enquiry from what it was in the base year, the product becomes irrelevant,

(b) The relative importance of certain commodities may change due to a change in the consumption pattern of the people in the course of time; for example, Vanaspati Ghee was not an important item of consumption in India in the pre-war period, but today it has become an item of necessity. Under such conditions, it is not easy to select the appropriate commodities.

3. Collection of Prices:

It is also difficult to obtain correct, adequate and representative data regarding prices. It is not an easy job to select representative places from which the information about prices to be collected and to select the experienced and unbiased individuals or institutions who will supply price quotations. Moreover, there is the problem of deciding which prices (wholesale or retail) are to be taken into consideration. It is comparatively easy to get information about wholesale prices which vary considerably.

4. Assigning Weights:

Another important difficulty that arises in preparing the index numbers is that of assigning proper weights to different items in order to arrive at correct and unbiased conclusions. As there are no hard and fast rules to weights for the commodities according to their relative importance, there is very likelihood that the weights are decided arbitrarily on the basis of personal judgement and involve biasness.

5. Selection of Averages:

Another major problem is that which average should be employed to find out the price relatives. There are many types of averages such as arithmetic average, geometric average, mean, median, mode, etc. The use of different averages gives different results. Therefore, it is essential to select the method with great care. Dr. Marshall has advocated the use of chain index number to solve the problem of averaging and weighing.

6. Problem of Dynamic Changes:

In the dynamic world, the consumption pattern of the individuals and the number and varieties of goods undergo continuous changes.

They create difficulties for preparing index numbers and making temporal comparisons:

(a) Since, in the course of time, old commodities may disappear and many new ones come into existence, the long-run comparison may become difficult,

(b) The quantity and quality of commodities may also change over the period of time, thus making the choice of commodities for constructing index numbers difficult,

(c) A number of factors, like income, education, fashion, etc., bring changes in the consumption pattern of the people which render the index numbers uncomparable.

Types of Index Numbers:

Index numbers are of different types.

Important types of index numbers are discussed below:

1. Wholesale Price Index Numbers:

Wholesale price index numbers are constructed on the basis of the wholesale prices of certain important commodities. The commodities included in preparing these index numbers are mainly raw-materials and semi-finished goods. Only the most important and most price-sensitive and semi- finished goods which are bought and sold in the wholesale market are selected and weights are assigned in accordance with their relative importance.

The wholesale price index numbers are generally used to measure changes in the value of money. The main problem with these index numbers is that they include only the wholesale prices of raw materials and semi-finished goods and do not take into consideration the retail prices of goods and services generally consumed by the common man. Hence, the wholesale price index numbers do not reflect true and accurate changes in the value of money.

2. Retail Price Index Numbers:

These index numbers are prepared to measure the changes.in the value of money on the basis of the retail prices of final consumption goods. The main difficulty with this index number is that the retail price for the same goods and for continuous periods is not available. The retail prices represent larger and more frequent fluctuations as compared to the wholesale prices.

3. Cost-of-Living Index Numbers:

These index numbers are constructed with reference to the important goods and services which are consumed by common people. Since the number of these goods and services is very large, only representative items which form the consumption pattern of the people are included. These index numbers are used to measure changes in the cost of living of the general public.

4. Working Class Cost-of-Living Index Numbers:

The working class cost-of-living index numbers aim at measuring changes in the cost of living of workers. These index numbers are consumed on the basis of only those goods and services which are generally consumed by the working class. The prices of these goods and index numbers are of great importance to the workers because their wages are adjusted according to these indices.

5. Wage Index Numbers:

The purpose of these index numbers is to measure time to time changes in money wages. These index numbers, when compared with the working class cost-of-living index numbers, provide information regarding the changes in the real wages of the workers.

6. Industrial Index Numbers:

Industrial index numbers are constructed with an objective of measuring changes in the industrial production. The production data of various industries are included in preparing these index numbers.

Importance of Index Numbers:

Index numbers are used to measure all types of quantitative changes in different fields.

Various advantages of index numbers are given below:

1. General Importance:

In general, index numbers are very useful in a number of ways:

(a) They measure changes in one variable or in a group of variables.

(b) They are useful in making comparisons with respect to different places or different periods of time,

(c) They are helpful in simplifying the complex facts.

(d) They are helpful in forecasting about the future,

(e) They are very useful in academic as well as practical research.

2. Measurement of Value of Money:

Index numbers are used to measure changes in the value of money or the price level from time to time. Changes in the price level generally influence production and employment of the country as well as various sections of the society. The price index numbers also forewarn about the future inflationary tendencies and in this way, enable the government to take appropriate anti- inflationary measures.

3. Changes in Cost of Living:

Index numbers highlight changes in the cost of living in the country. They indicate whether the cost of living of the people is rising or falling. On the basis of this information, the wages of the workers can be adjusted accordingly to save the wage earners from the hardships of inflation.

4. Changes in Production:

Index numbers are also useful in providing information regarding production trends in different sectors of the economy. They help in assessing the actual condition of different industries, i.e., whether production in a particular industry is increasing or decreasing or is constant.

5. Importance in Trade:

Importance in trade with the help of index numbers, knowledge about the trade conditions and trade trends can be obtained. The import and export indices show whether foreign trade of the country is increasing or decreasing and whether the balance of trade is favourable or unfavourable.

6. Formation of Economic Policy:

Index numbers prove very useful to the government in formulating as well as evaluating economic policies. Index numbers measure changes in the economic conditions and, with this information, help the planners to formulate appropriate economic policies. Further, whether particular economic policy is good or bad is also judged by index numbers.

7. Useful in All Fields:

Index numbers are useful in almost all the fields. They are specially important in economic field.

Limitations of Index Numbers:

Index number technique itself has certain limitations which have greatly reduced its usefulness:

(i) Because of the various practical difficulties involved in their computation, the index numbers are never cent per cent correct.

(ii) There are no all-purpose index numbers. The index numbers prepared for one purpose cannot be used for another purpose. For example, the cost-of-living index numbers of factory workers cannot be used to measure changes in the value of money of the middle income group.

(iii) Index numbers cannot be reliably used to make international comparisons. Different countries include different items with different qualities and use different base years in constructing index numbers.

(iv) Index numbers measure only average change and indicate only broad trends. They do not provide accurate information.

(v) While preparing index numbers, quality of items is not considered. It may be possible that a general rise in the index is due to an improvement in the quality of a product and not because of a rise in its price.

Tests of adequacy of Index Number

An index number, which measures the change in the level of a phenomenon from one period to another,can satisfy the certain tests.

There are three methods to test the adequacy of Index numbers

1. Unit Test
2. Time Reversal test
3. Factor reversal test
4. Circular test

Unit Test

In this test, the expressed units of prices and quantities should be independent

All methods, except simple aggregative method satisfied the test

All other indices satisfy the test except unweighted aggregative index number.

Time Reversal test

1. This method works for both backward and forward
2. The test is based on the analogy that the principle, which holds good for a single
3. commodity, should also be true for the index number as a whole.
4. Ratio between one point of time and the other, no matter which of the two is taken as the base.
5. In other words, when the data for any two years are treated by the same method, but
6. with the base is reversed, the two index numbers should be reciprocals of eachother.

Symbolically the test is $P_{01} \times P_{10}=1$

Time reversal test Symbolically the test is $P_{01}XP_{10}=1$ P_{01} is the index for the time ‘1’ on time ‘0’ P_{10} is the index for the time ‘0’ on time ‘0’.

Under this method if product is not the unity, the method suffers from the time basis. This method is satisfied for the following methods;

1. Simple aggregative method
2. Fisher's method
3. Marshall aggregative method
4. Kelly’s method

Factor Reversal Test

According to this method, the price index and quantity index should be equal to value index.

Each method in index number interchanging the prices and quantities without giving inconsistant results which means two results multipies together which gives the true value.

This test describes change in price multipied by change in quantity can equal to total change in value.

The test is symbolically represented as $P_{01}XQ_{01}=V_{01}$

Circular Test

1. Mostly two types of base periods are used to construct the index numbers, namely,
   a) Fixed base, b) Chain base.
2. But in most cases fixed base method used.
3. It cannot take into account any changes in price or quantity in any other year
4. It fails to include new commodities gaining importance at a later date or exclude commodities losing significance in course of time. These problems can be overcome by chain index numbers.

The formulae satisfying the requirements of circular test are:

1) Simple aggregative index

2) Simple geometric mean of relatives

3) Weighted aggregative index (such as Laspeyres' index with constant weights)

4) Weighted geometric mean of relatives with constant weights.

Wholesale Price Index

Wholesale Price Index number (WPI)

The wholesale price index number represents the price of a basket of wholesale items. WPI index is concerned with the cost of goods sold between businesses. It does not focus on consumer-purchased commodities. WPI’s major goal is to track pricing changes in manufacturing, construction, and industry, representing demand and supply. WPI aids in measuring an economy’s microeconomic conditions and macroeconomics.

Importance of WPI index

The inflation rate, which is derived using the Wholesale Price Index (WPI), is an important metric for tracking price changes.

WPI is widely used by the government, banks, industry, and business circles since it captures price variations comprehensively.

WPI movements are frequently linked to significant monetary and fiscal policy shifts.

Similarly, the movement of the WPI is a key factor in the formation of trade, fiscal, and other economic policies by the Indian government.

Escalation clauses in the supply of raw materials, machinery, and construction work are also based on the WPI indexes.

The WPI is used to deflate various nominal macroeconomic indicators, including GDP (GDP).

Let us dive a bit into the details of the wholesale price index and consumer price index.

Wholesale Price Index (WPI) in India

Inflation rates are generally calculated using the WPI and CPI (Consumer Price Index). In India, inflation rates are calculated using the Wholesale Price Index (WPI), which the Ministry of Commerce and Industry publishes. The Consumer Price Index (CPI) is a weighted average of prices for a basket of consumer goods and services consumed by households, such as transportation, food, and medical care. In September 1974, India had its highest inflation rate of 34.68 per cent. In May 1976, the lowest rate was -11.31 per cent.

Consumer Price Index Number

The consumer price index is often used to look at the average weight of a price for a collection of products and services, including food, transportation, and medical care. It calculates the average change in pricing for the basket of items and services that the client is expected to spend. Pricing changes associated with the cost of living are approached using changes within customer price index numbers. The customer index number is a useful metric for identifying periods of deflation and inflation. It's also known as an economic metric.

Consumer Price Index Number for Industrial Workers

The consumer price index for industrial employees is calculated to track price changes for a specific category of goods and services over time. This is absorbed by a certain demographic, in this case, workers in the manufacturing industry. The consumer index is compiled for all industrial workers in 70 cities around the country with significant industrial importance. These 70 centres were distributed to all states in their industrial growth.

Each month, the indices of all 70 centres are collected and released based on weights obtained from working-class families. It is also based on the expenditure survey taken between 1981 and 1982 and current prices of individual commodities collected from 226 markets served by 70 centres. The All-India Index, likewise a weighted average, is calculated using the 70 centres.

In addition, the Labour Bureau collects and delivers indices from the other six centres to suit the needs of specific index consumers. Aside from serial data, the magazine also includes information like inflation rates, all-India products, and connecting factors for new and old series, among other things.

Applications of the Consumer Price Index Number

The consumer price index number represents the change in consumer prices. As a result, it assists the government in formulating a number of policies relating to taxation, price, exports, and imports of all commodities.

It is used to give staff allowances and other incentives.

It is also used to determine the purchasing power of a unit of currency.

It also compares changes in different groups' probability of survival.

Using the index number, data on wages, living costs, and national income are likewise deflated.

It also serves as an economic indicator, similar to financial instruments and commodity prices.

Acts as a policymaker at the global, state, and national levels.

Conclusion

In the above notes, we have read about wholesale and consumer prices. A wholesale price index (WPI) measures and monitors changes in the price of items before they reach the retail level. This is a term used to describe goods sold in large quantities and transferred between entities or businesses (instead of between consumers). The WPI, which is usually represented as a ratio or percentage, displays the average price change of the products covered; it is frequently used to measure a country's level of inflation. Don't skip these notes if you are preparing for an upcoming examination.

Consumer Price Indexes

Consumer price indexes (CPIs) are index numbers that measure changes in the prices of goods and services purchased or otherwise acquired by households, which households use directly, or indirectly, to satisfy their own needs and wants.

CPI stands for consumer price index and measures the ongoing change in the costs of goods and services. This can include almost any good or service, like transportation, medical care, food or other merchandise items. Many use it to predict and determine the cost of living and economic growth in certain areas. This is one of the more popular ways for professionals within the economic or financial industries to locate inflation or deflation periods within the economy.

Formula for CPI

When calculating the consumer price index, the final consumer price index result represents the average change in prices that consumers will spend on a basket of goods and services over time. This is how economic and financial professionals identify and determine inflation. They can then use CPI to determine the economy's aggregate price levels to measure the purchasing price of an entire country or a specific area. The consumer price index formula is:

Cost of products or services in a current period/cost of products or services in a previous time period × 100 = consumer price index

Merits and Demerits of Consumer Price Index Number

Merits:

Due to the Consumer Price Index functions, it has numerous advantages. Some advantages of CPI are listed below:

1. Firstly, it helps the government analyze the risks of elevating prices for development. As a result, it allows the government to increase prices without affecting the cost of living among the masses
2. The state government uses the Consumer Price Index to decide wage contracts and additional benefits like dearness allowances for the workers
3. Lastly, the CPI also calculates the national income and acts as the income deflator

Demerits:

Although CPI has advantages, it also has some disadvantages or limitations. Listed below are some disadvantages of CPI:

1. The CPI is it fails to represent the diversity amongst the masses in various sectors
2. The retail price is different for all markets. Therefore, finding one retail price representative of collective retail prices isn’t easy
3. Lastly, consumers’ consumption patterns and ratios may change from time to time

Cost of Living Index

A cost-of-living index is a theoretical price index that measures relative cost of living over time or regions. It is an index that measures differences in the price of goods and services, and allows for substitutions with other items as prices vary.

Methods Used to Construct Consumer Price Index Numbers

Following two methods are used to construct consumer price index numbers – i) Family Budget method or weighted average of relatives ii) Aggregative Expenditure

The following points highlight the three Measurements for Cost of Living. The Measurements are: 1. Consumer Price Index (CPI) 2. Producer Price Index (PPI).

Measurement # 1. Consumer Price Index (CPI):

The consumer price index (CPI) is the most widely used measure of the level of prices.

It is constructed by collecting the prices of thousands of goods and services.

We know that the GDP expresses the quantity of diverse goods and services into a single number which is used as a measure of the value of society's output.

In a like manner the CPI expresses the prices of numerous goods and services into a single index for measuring the general price level.

The CPI is a weighted average of all prices. It is the price of a basket of goods and services relative to the price of the same basket in some base year.

Example:

Let us suppose our representative consumer (Mr. X) buys 5 bananas and 2 cakes every month. Then the basket of goods consists of 5 bananas and 2 cakes and, in this case,

$$\text{CPI} = \frac{(5\text{ x current price of bananas}) + (2\text{ x current price of cakes}}{(5\text{ x price of bananas in year 2000}) + (2\text{ x price of cakes in year 2000})}$$

In this CPI, 2000 is taken as the base year. This index indicates the cost of buying 5 bananas and 2 cakes today compared to how much it had cost to buy the same basket in 2000.

UNIT –II

TIME SERIES ANALYSIS

The analysis of time series is of great significance not only to the economist and businessman but also to the scientist, astronomist, geologist, sociologist, biologist, research worker, etc., for reasons given below:

1. It helps in understanding past behavior. By observing data over a period of time one can easily understand what changes have taken place in the past. Such analysis will be extremely helpful in predicting the future behavior.
2. It helps in planning future observations. Plans for the future cannot be made without forecasting events and relationship they will have. Statistical techniques have been evolved which enable time series to be analyzed in such a way that the influence which have determined the form of that series may be ascertained.
3. It helps in evaluating current accomplishments. The actual performance can be compared with the expected performance and the cause of variation analyzed.
4. It facilitates comparison. Different time series are often compared and important conclusion drawn there from.

Components of Time Series

It is customary to classify the fluctuations of a time series into four basic types of variations, which superimposed and acting all in concert account for changes in the series over a period of time. Those four types of patterns, movements, or, as they are often called, components or elements of a time series, are:

(1) Secular Trend

(2) Seasonal Variations

(3) Cyclical Variations

(4) Irregular Variations

1. **Secular Trend :**

   Secular trend movements are attributable to factors such as population change, technological progress and large-scale shifts in consumer tastes.

   - Linear or Straight Line Trends
   - Non-linear Trends

2. **Seasonal variations**

   Seasonal variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself.

(i) Climate and weather conditions. The most important factor causing seasonal variations is the climate. Changes in the climate and weather conditions such as rainfall, humidity, heat, etc., act on different products and industries differently. For example, during winter there is greater demand for woolen clothes, hot drinks, whereas in summer cotton clothes, cold drinks have a greater sale. Agriculture is influenced very much by the climate. The effect of the climate is that there are generally two seasons in agriculture the growing season and harvesting season which directly affect the income of the farmer which, in turn, affects the entire business activity.

(ii) Customs, traditions and habits. Though nature is primarily responsible for seasonal variations in time series, customs, traditions and habits also have their impact. For example on certain occasions like Deepawali, Dussehra, Christmas, etc.,

3. **Cyclical Variations**

The term cycle refers to the recurrent variations in time series that usually last longer than a year and are regular neither in amplitude nor in length. Cyclical fluctuations are long term movements that represent consistently recurring rises and declines in activity. A business cycle consists of the recurrence of the up and down movements of business activity from some sort of statistical trend or normal. There are four well-defined periods or phases in the business cycle namely : i) prosperity, ii) decline, iii) depression and iv) improvement.

Each phase changes gradually into the phase which follows it in the order given.

4. Irregular Variations

Irregular variations also called erratic accidental random refer to such variations in business activity which does not repeat in a definite pattern. Irregular movements on the other hand, are considered to be largely random, being the result of change factors which like those determining the fall of a coin, are wholly unpredictable.

Irregular variations are caused by such isolated special occurrences as floods, earthquakes, strikes and wars. Sudden changes in demand or very rapid technological progress may also be included in this category. By their very nature these movements are very irregular and unpredictable. Quantitatively it is almost impossible to separate out the irregular movements and the cyclical movements.

There are two reasons for recognizing irregular movements:

(i) To suggest that on occasions it may be possible to explain certain movements in the data due to specific causes and to simplify further analysis.

(ii) To emphasize the fact that predictions of economic conditions are always subject to degree of error owing to the unpredictable erratic influences which may enter.

Preliminary Adjustment before Analysing Time Series

Before beginning the actual work of analysing a time series it is necessary to make certain adjustments in the raw data. The adjustments may be needed for:

- Calendar Variations.
- Population Changes.

- Price Changes.
- Comparability.

(i) **Calendar Variations**.

A vast proportion of the important time series is available in a monthly form and it is necessary to recognize that the month is a variable time unit. The actual length of the shortest month is about 10 per cent less than that of the longest , and if we take into account holidays and weekends, the variation may be even greater. The adjustment for calendar variations is made by dividing each monthly totally by the number of days in the month thus arriving at daily average for each month.

(ii) Populations changes.

Certain types of data call for adjustment for population changes. Changes in the size of population can easily distort comparisons of income, production and consumption figures. For example national income may be increasing year after year, but per capita income may be declining because of greater pressure of population.

(iii) Price changes:

An adjustment for price changes is necessary whenever we have a value series and are interested in quantity changes alone. Because of rising prices the total sale proceeds may go up even when there is a tall in the number of units sold. For example, if in 2001, 1,000 units of a commodity that is priced Rs. 10 are sold, the total sale proceeds would be 1,000*10=Rs. 10,000.,

(iv) Comparability.

For any meaningful analysis of time series, it is necessary to see that the data are strictly comparable throughout the time period under investigation. Quite often it is difficult or even impossible to get strictly comparable data. For example, if we are observing a phenomenon over the last 25 years the comparability may be observes by differences in definition, differences in geographical coverage, differences in the method adopted, changes in the method of reporting, etc.

Measurement of Trend

Given any long-term series, we wish to determine and present the direction which it takes-is it growing or declining? There are two important reasons for trend measurement:

(i) To find out trend characteristics in and of themselves. In studying trend in and of itself. We ascertain the growth factor. For example, we can compare the growth in the texitile industry with the growth in the economy as a whole or with the growth in other industries, or we can compare the growth in one firm of the textile industry with the growth in the industry as a whole. Moreover, we can compare through trend characteristics the growth of the textile industry in India with that of other countries. The growth factor also helps us in predicting the future behavior of the data. If a trend can be determined, the rate of change can be ascertained and tentative estimates concerning future made accordingly.

(ii) To enable us to eliminate trend in order to study other elements. The elimination of trend leaves us with seasonal, cyclical and irregular factors. We can then, in two or more series, compare or use the impact of these three relatively short-term elements divorced from the long-term factor.

The various methods that can be used for determining trend are:

- Freehand or graphic method.
- Semi-average method.
- Moving average method.
- Method of least squares.

Each of these methods is discussed below.

Freehand Graphic Method

This is the simplest method of studying trend. The procedure of obtaining a straight line trend by this method is given below:

1. Plot the time series on a graph.

2. Examine carefully the direction of the trend based on the plotted information.

3. Draw a straight line which will beast fit to the data according to personal judgment. The line now shows the direction of the trend.

Merits and Limitations

1. This is the simplest method of measuring trend.

2. This method is very flexible in that it can be used regardless of whether the trend is a straight line or a curve.

3. The trend line drawn by a statistician experienced in computing trend and having knowledge of the economic history of the concern or the industry under analysis may be a better expression of the secular movement than a trend fitted by the use of a rigid mathematical formula which while providing a good fit to the points, may have no other logical justification. In fact a specialist of long experience who is familiar with the institutional setting, history and behavior of the series may well be able manually to fit a trend superior of one derived by mathematical means. Although the freehand method is not recommended for beginners, it has considerable merit in the hands of experienced statisticians and is widely used in applied situations.

Limitations

- This method is highly subjective because the trend line depends on the personal judgment of the investigator and, therefore, different persons may draw different trend lines from the same set of data. Moreover, the work cannot be left to clerks and it must be handled by skilled and experienced people who are well conversant with the history of the particular concern.

- Since freehand curve fitting is subjective it cannot have much value if it is used as a basis for predictions.
- This method appears simple and direct. However, it is very time-consuming to construct a freehand trend if a careful and conscientious job is done.

It is only after long experience in trend fitting that a statistician should attempt to fit a trend line by inspection.

Method of Semi-Averages

When this method is used, the given data is dividing into two parts. Preferably with the same number of years. For example, if we are given data from 1986 to 2003, over period of 18 years, the two equal parts will be each nine years, from 1986 to 1994 and from 1995 to 2003. In case of odd number of years like 9, 13, 17, etc., two equal parts can be made simply by omitting the middle year. For example, if data are given for 19 years from 1995 to 2003 the two equal parts would be from 1985 to 1993 and from 1995 to 2003 the middle year 1994 will be omitted.

After the data have been divided into two parts, an average of each part is obtained. We thus get two points. Each point is plotted at the mid-point of the class interval covered by the respective part and then the two points are joined by a straight line which gives us the required trend line. The line can be extended downwards or upwards to get intermediate values or to predict future values.

The following example shall illustrate this method:

Fit a trend line to the following data by the method of semi-average:

Year	**Sales of Firm A (thousand units)**	**Year**	**Sales of Firm A (thousand units)**
1997	102	2001	108
1998	105	2002	116
1999	114	2003	112
2000	110		

Solution : Since seven years are given, the middle year shall be left out and an average of the first three years and the last three years shall be obtained. The average of the first three years is $\frac{102+105+114}{3} = \frac{321}{3} = 107$

And the average of the last three years is

$$Y_c = a + bX \ \frac{108+116+112}{3} = \frac{336}{3} = 112$$

Thus we get two points 107 and 112 which shall be plotted corresponding to their respective middle years.

Method of Moving Averages

When a trend is to determined by the method of moving averages, the average value for a number of years is secured, and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the averages. The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

While applying this method, it is necessary to select a period for moving average such as 3-yearly moving average, 5-yearly moving average, 8-yearly moving average, etc. The period of moving average is to be decided in the light of the length of the cycle. Since the moving average method is most commonly applied to data which are characterized by cyclical movements it is necessary to select a period for moving average which considers with the length of the cycle, otherwise the cycle will not be entirely removed. The danger is more severe, the shorter the time period represented by the average. When the period of moving average and the period of the cycle do not coincide, the moving average will display a cycle which has the same period as the cycle in the data, but which has less amplitude than the cycle in the data. Often we find that the cycles in the data are not of uniform length. In such a case we should take a moving average period equal to or somewhat greater than the average period of the cycle in the data. Ordinarily, the necessary period will range between three and ten years for general business series but even longer periods are required for certain types of data.

The 3-yearly moving average shall be computed as following:

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3}, ...$$

And for 5-yearly average:

$$\frac{a+b+c+d+e}{3}, \frac{b+c+d+e+f}{3}, \frac{c+d+e+f+g}{3}, ...$$

Calculate the 3-yearly moving averages of the production figures given below and draw the trend:

Year	**Production (in m. tones)**	**Year**	**Production (in m. tones)**
1989	15	1997	63
1990	21	1998	70
1991	30	1999	74
1992	36	2000	82
1993	42	2001	90
1994	46	2002	95
1995	50	2003	102

Solution

Year	**Production (in million tonnes)**	**3-yearly total**	**3-yearly moving average**
1989	15	_	_
1990	21	66	22.00
1991	30	87	29.00
1992	36	108	36.00
1993	42	124	41.33
1994	46	138	46.00
1995	50	152	50.67
1996	56	169	56.33
1997	63	189	63.00
1998	70	207	69.00
1999	74	226	75.33
2000	82	246	82.00
2001	90	267	89.00
2002	95	287	95.67
2003	102	_	_

The Method of Least Squares

This method is most widely used in practice. It is a mathematical method and with its help a trend line is fitted to the data in such a manner that the following two conditions are satisfied:

(1) $\sum(y - y_c) = 0,$

The sum of deviations of the actual values of Y and the computed values of Y is zero.

(2) $\sum(y - y_c)^2$ is least,

The sum of the squares of the deviations of the actual and computed values is least from this line and hence the name method of least squares. The line obtained by this method is known 'as the line of beast fit'.

The method of least squares may be used either of fit a straight line trend or a parabolic trend.

The straight line trend is represented by the equation

$Y_c = a + bX$

Where Y_c is used to designate the trend values to distinguish them from the actual Y values, as is the Y intercept or the computed trend figure of the Y variable when X=0, b represents the slope of the trend line or amount of change in Y variable that is associated with a change of one unit in X variable. The X variable in time series analysis represents time.

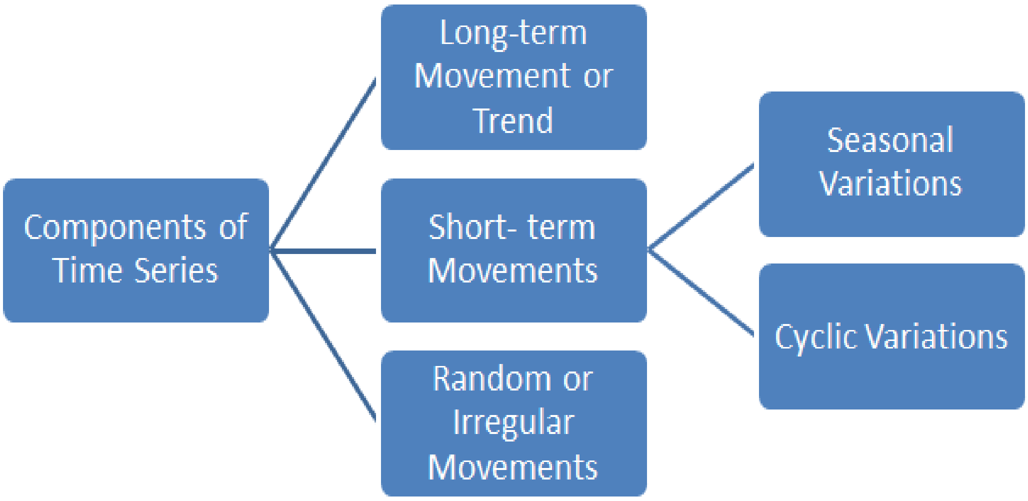
Uses of Times Series

Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves ...

1. The most important use of studying time series is that it helps us to predict the future behaviour of the variable based on past experience
2. It is helpful for business planning as it helps in comparing the actual current performance with the expected one
3. From time series, we get to study the past behaviour of the phenomenon or the variable under consideration
4. We can compare the changes in the values of different variables at different times or places, etc.

Components for Time Series Analysis

The various reasons or the forces which affect the values of an observation in a time series are the components of a time series. The four categories of the components of time series are Trend, Seasonal Variations, Cyclic Variations and Random or Irregular movements. Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.


Components of Time Series
Long-term Movement or Trend
Short- term Movements
Random or Irregular Movements
Seasonal Variations
Cyclic Variations

Trend

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable. The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Linear and Non-Linear Trend

If we plot the time series values on a graph in accordance with time t. The pattern of the data clustering shows the type of trend. If the set of data cluster more or less round a straight line, then the trend is linear otherwise it is non-linear (Curvilinear).

Periodic Fluctuations

There are some components in a time series which tend to repeat themselves over a certain period of time. They act in a regular spasmodic manner.

Seasonal Variations

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.

The effect of man-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Cyclic Variations

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the ‘Business Cycle’.

It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Random or Irregular Movements

There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

UNIT III

THEORY OF PROBABILITY

The word probability or a chance is very commonly used in day-to-day conversation and generally people have a vague about its meaning. The theory of probability has its origin in the games of chance related to gambling such as throwing a die, tossing a coin, drawing cards from a pack of cards etc.

Theory of probability was developed in the middle of the 17th century. The names which are associated with probability are Hugenes, Pascal, Format, Berouli, Laplace, Bayes. Today, the theory of probability has been developed to a great extent and extensively used in various subjects. The theory of probability was developed from gambling as it is a game of chance.

Key Concepts of Probability

1. **Random experiment**: An experiment or trial outcome is uncertain is called random experiment. Though an experiment is repeated under the same conditions, individual outcome is not predictable. Such an experiment is called random experiment.
2. **Event:** Any possible outcome of a random experiment is called an event. Performing a random experiment is trail and the occurrence or non-occurrence of something is an event. The occurrence of an event which is inevitable, when a random experiment is performed, is called sure event or certain event. On the other hand is the occurrence is impossible, it is called impossible event.
3. **Simple and compound events**: Here the classification is made on the basic of the number of events in question. If only one event is take place at a time, it is called simple event.
4. **Mutually exclusive event**: If the occurrence of the event excludes the occurrence of the alternative, the events are called mutually exclusive. Simply, in mutually

exclusive events simultaneous occurrence of events is not possible. Such events are called alternative events or incompatible events.

5. **Equally likely events**: When the change of occurrence of each events is the same, the events are called equally likely events. They are also called equiprobable events. In such events, one events does not occur more often the other.
6. **Independent** events: In independent events, the occurrence of one event doesn't affect and is not affected by the other. For instance, when we toss a coin twice, the result of the second toss will not in any way be affected by the result of the first toss. In other words, the result of the first toss does not affect the result of the second event.
7. **Dependent events:** Two events are said to be dependent, if the occurrence or non-occurrence of one event affects occurrence of the other. In other words, the occurrence of an event affects the result of the subsequent trail, then such events are called department events.

Importance of Probability

- The whole sampling theory, particularly the principle of law of statistical regularity and the law of inertia of large numbers, was developed on the basis of probability theory.
- It is very useful to solve problems related to betting gambling.
- The decision theory is also constructed on the probability theory.
- Different tests used for testing of hypothesis are derived from probability.

Theorem of probability

Theorems explain functional relationship existing between variables or attributes. The functional relationships or laws are formed to tackle complex situations. The law of probability has also helped to tackle complex situations that arise in the field of probability. There are two important theorems,

1. Addition theorem

2. Multiplication theorem

1. Addition theorem: The addition rule is the simplest and frequently used rule to determine probability. If two events are mutually exclusive, then the probability of happening either ‘A’ or ‘B’ is the sum of their separate probabilities. It is also called theorem of total probability.

$$P(A or\, B) = P(A) + P(B)$$

Proof of the Theorem:

If an event A can happen in $a_1$ways andB can happen in a_2 ways, then the number of ways in which either event can happen is a_1+ a_2.

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n}$$

$$\frac{a_1}{n} = P(A)$$

$$\frac{a_2}{n} = P(B)$$

$$P(A or\, B) = P(A) + P(B)$$

Hence, the theorem can be extended to three or more mutually exclusive events.

$$P(A or\, B) = P(A) + P(B) + \mathrm{P(C)}$$

2. Multiplication theorem: The multiplication law states that “the probability of happening of given 2 events or in different words the probability of the intersection of 2 given events is equivalent to the product achieved by finding out the product of the probability of happening of both the events.”

Proof of the Theorem:

If an event A can happen in $n_1$ways of which a_1 are successful and the event B can happen in n_2 ways, of which $a_2$are successful, then the number of ways in which either event can happen is n_1+ n_2.

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} + \frac{a_2}{n_2}$$

$$\frac{a_1}{n_1} = P(A)$$

$$\frac{a_2}{n_2} = P(B)$$

$$P(A\, and\; B) = P(A) \times P(B)$$

Hence, the theorem can be extended to three or more mutually exclusive events.

$P(A, B\, and\; C) = P(A) \times P(B) \times \mathrm{P(C)}$

Example:1

One card is drawn from a standard pack of 52. What is the probability that it is either a king or a queen?

Solution

There are 4 kings and 4 queens in a pack of 52 cards.

The probability that the card drawn is a king = $\frac{4}{52}$ and the probability that the card drawn is a queen = $\frac{4}{52}$

Since the events are mutually exclusive, the probability that the card drawn is either a king or a queen = $\frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$

Example:2

A man wants to marry a girl having white complexion-the probability of getting such a girl is one in 20, handsome dowry-the probability of getting this is one in fifty, westernized manners and etiquettes-the probability here is one in hundred. Find out the probability of his getting married to such a girl when the possession of these three attributes is independent.

Solution:

Probability of a girl with the complexion = $\frac{1}{20} = 0.05$

Probability of a girl with handsome dowry = $\frac{1}{50} = 0.02$

Probability of a girl with westernized manners = $\frac{1}{100} = 0.01$

Since the events are independent, the probability of simultaneous occurrence of all these qualities

$$= \frac{1}{20} \times \frac{1}{50} \times \frac{1}{100} = 0.05 \times 0.02 \times 0.01$$

$$= 0.00001$$

Conditional Probability

In probability theory, conditional probability is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) is already known to have occurred. This particular method relies on event B occurring with some sort of relationship with another event A. In this event, the event B can be analyzed by a conditional probability with respect to A. If the event of interest is *A* and the event *B* is known or assumed to have occurred, "the conditional probability of *A* given *B*", or "the probability of *A* under the condition *B*", is usually written as P(*A*/*B*) or occasionally $P_B(A)$. This can also be understood as the fraction of probability B that intersects with A, or the ratio of the probabilities of both events happening to the "given" one happening (how many times A occurs rather than not assuming B has occurred):

If two events A and B are dependent, then the conditional probability of B given A is

$$P(B/A) = \frac{P(AB)}{P(A)}$$

Example:3

A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.

Solution:

Probability of drawing a black ball in the first attempt is $P(A) = \frac{3}{5+3} = \frac{3}{8}$

Probability of drawing a second black ball given that the first ball drawn is black $P(\mathrm{B/\,A}) = \frac{2}{5+2} = \frac{2}{7}$

The probability that both balls drawn are black is given by

$$P(\mathrm{AB}) = P(A) \times P(B/A) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$

Discrete and Continuous

Discrete data is information that has noticeable gaps between values. Continuous data is information that occurs in a continuous series. Discrete data is made up of discrete or distinct values. Directly in opposition, continuous data includes any value that falls inside a range

Random variables

Key Takeaways: A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. A random variable can be either discrete (having specific values) or continuous.

A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. Random variables are often designated by letters and can be classified as discrete, which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.

Types of Random Variables

Likelihood that any of the possible values would occur. Let's say that the random variable, Z, is the number on the top face of a die when it is rolled once. The possible values for Z will thus be 1, 2, 3, 4, 5, and 6. The probability of each of these values is 1/6 A random variable has a probability distribution that represents the as they are all equally likely to be the value of Z.

For instance, the probability of getting a 3, or P (Z=3), when a die is thrown is 1/6, and so is the probability of having a 4 or a 2 or any other number on all six faces of a die. Note that the sum of all probabilities is 1.

A random variable can be either discrete or continuous.

Discrete Random Variables

Discrete random variables take on a countable number of distinct values. Consider an experiment where a coin is tossed three times. If X represents the number of times that the coin comes up heads, then X is a discrete random variable that can only have the values 0, 1, 2, or 3 (from no heads in three successive coin tosses to all heads). No other value is possible for X.

Continuous Random Variables

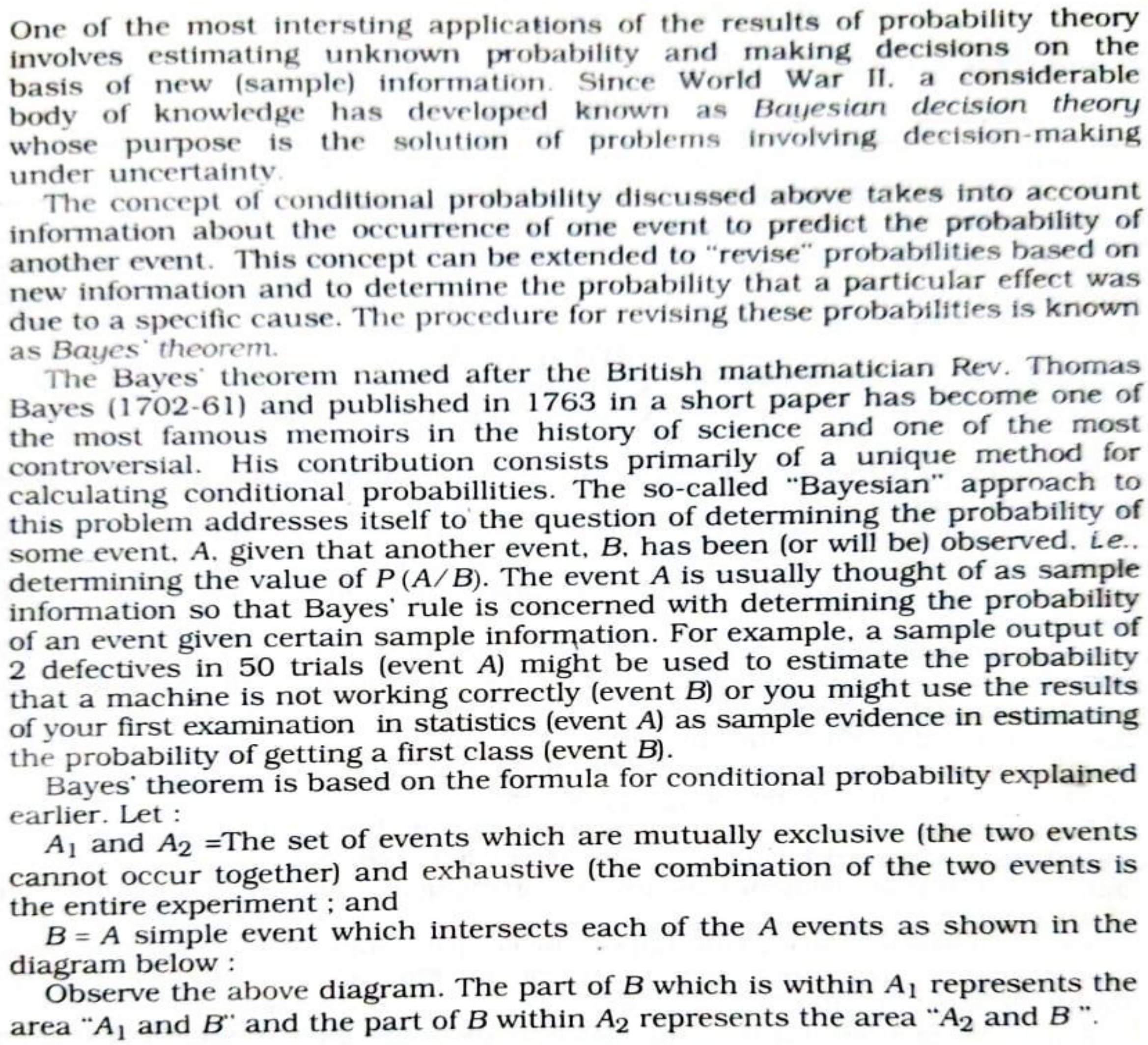
Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values. An example of a continuous random variable would be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.

Bayes Theorem

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

Bayes formula used

The Bayes theorem is a mathematical formula for calculating conditional probability in probability and statistics. In other words, it's used to figure out how likely an event is based on its proximity to another.



One of the most intersting applications of the results of probability theory involves estimating unknown probability and making decisions on the basis of new (sample) information. Since World War II. a considerable body of knowledge has developed known as *Bayesian decision theory* whose purpose is the solution of problems involving decision-making under uncertainty.

The concept of conditional probability discussed above takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to "revise" probabilities based on new information and to determine the probability that a particular effect was due to a specific cause. The procedure for revising these probabilities is known as *Bayes' theorem.*

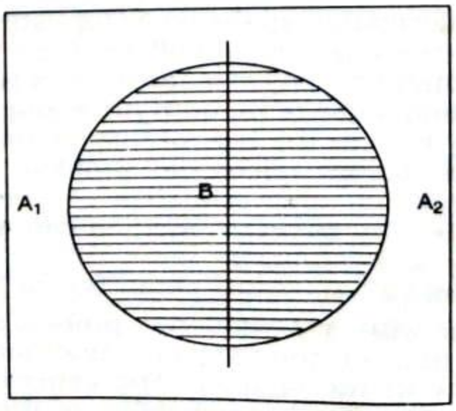
The Bayes' theorem named after the British mathematician Rev. Thomas Bayes (1702-61) and published in 1763 in a short paper has become one of the most famous memoirs in the history of science and one of the most controversial. His contribution consists primarily of a unique method for calculating conditional probabillities. The so-called "Bayesian" approach to this problem addresses itself to the question of determining the probability of some event, A, given that another event, B, has been (or will be) observed, *i.e.*, determining the value of $P(A/B)$. The event A is usually thought of as sample information so that Bayes' rule is concerned with determining the probability of an event given certain sample information. For example, a sample output of 2 defectives in 50 trials (event A) might be used to estimate the probability that a machine is not working correctly (event B) or you might use the results of your first examination in statistics (event A) as sample evidence in estimating the probability of getting a first class (event B).

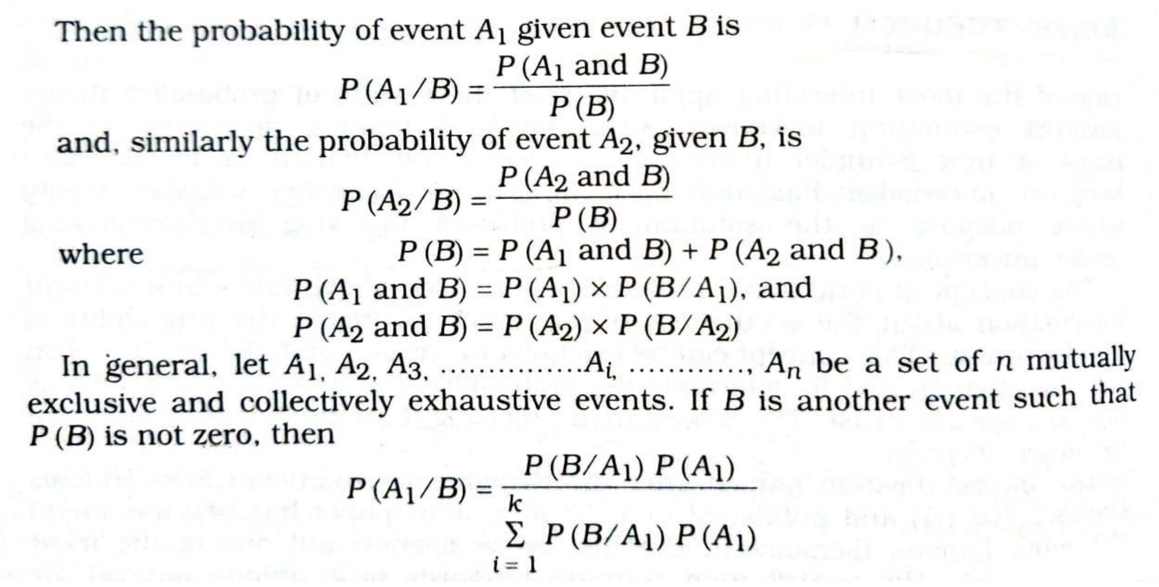
Bayes' theorem is based on the formula for conditional probability explained earlier. Let :

A_1 and A_2 =The set of events which are mutually exclusive (the two events cannot occur together) and exhaustive (the combination of the two events is the entire experiment ; and

$B = A$ simple event which intersects each of the A events as shown in the diagram below :

Observe the above diagram. The part of B which is within A_1 represents the area "A_1 and B" and the part of B within A_2 represents the area "A_2 and B ".





Then the probability of event A_1 given event B is

$$P(A_1/B) = \frac{P(A_1 \text{ and } B)}{P(B)}$$

and, similarly the probability of event A_2, given B, is

$$P(A_2/B) = \frac{P(A_2 \text{ and } B)}{P(B)}$$

where $P(B) = P(A_1 \text{ and } B) + P(A_2 \text{ and } B)$,

$$P(A_1 \text{ and } B) = P(A_1) \times P(B/A_1), \text{ and}$$

$$P(A_2 \text{ and } B) = P(A_2) \times P(B/A_2)$$

In general, let $A_1, A_2, A_3, \ldots\ldots\ldots\ldots A_i, \ldots\ldots\ldots\ldots, A_n$ be a set of n mutually exclusive and collectively exhaustive events. If B is another event such that $P(B)$ is not zero, then

$$P(A_1/B) = \frac{P(B/A_1)\, P(A_1)}{\sum_{i=1}^{k} P(B/A_1)\, P(A_1)}$$

Example of Bayesian theory

For example, if a disease is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have the disease, compared to the assessment of the probability of disease made without knowledge of the person's age.

Theoretical Distributions

In other words, theoretical distribution is a statistical distribution received by a set of logical and mathematical reasoning from given principles or assumptions. Theoretical distribution is the opposite of distribution derived by real-world data derived by empirical research.

Binomial Distribution

The binomial distribution also known as ‘Bernoulli Distribution’ is associated with the name of a Swiss mathematician James Bernoulli also known as Jasques or Jakob (1654-1705). Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure.

This distribution has been used to describe a wide variety of processes in business and the social sciences as well as other areas. The type of process which give rise to this distribution is usually referred to as Bernoulli trail or as a Bernoulli process.

Properties of the Binomial Distribution

1. The shape and location of binomial distribution changes as P changes for a given n or as n changes for a given p. As p increases for a fixed n, the binomial distribution shifts to the right.

2. The mode of the binomial distribution is equal to the value of x which has the largest probability.

Importance of the Binomial distribution

The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life events. For example, a quality control inspector wants to know the probability of defective light bulbs in a random sample of 10 bulbs if 10 per cent of the bulbs are defective. He can quickly obtain the answer from tables of the binomial probability distribution. The binomial distribution can be used when:

1. The outcome or results of each trial in the process are characterized as one of two types of possible outcomes. In other words, they are attributes.

2. The possibility of outcome of any trial does not change and is independent of the results of previous trials.

The Binomial Distribution

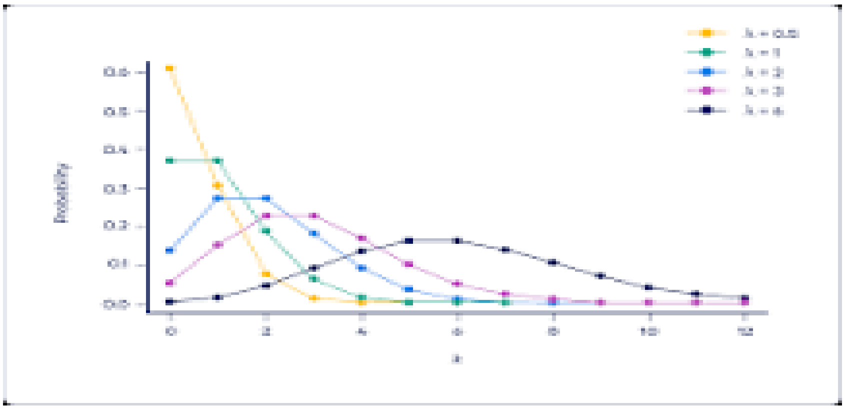
$$\begin{aligned}
&P(r) = {}^{n}C_{r}q^{n-r}p^{r} \\
&Where \\
&P = \Pr obability\, of\ sucess\, in\, a\, \text{single trail} \\
&q = 1 - p \\
&n = Number\, of\ trails \\
&r = Number\, of\ successes\, in\, n\, trails.
\end{aligned}$$

Poisson Distribution:

Poisson distribution is a discrete probability distribution and is very widely used in statistical work. It was developed by a French mathematician, Simeon Denis Poisson (1781-1840). Poisson distribution may be expected in cases where the chance of any individual event being a success is small.

Unlike a normal distribution, which is always symmetric, the basic shape of a Poisson distribution changes. For example, a Poisson distribution with a low mean is highly skewed, with 0 as the mode. All the data are “pushed” up against 0, with a tail extending to the right.

Poisson distribution

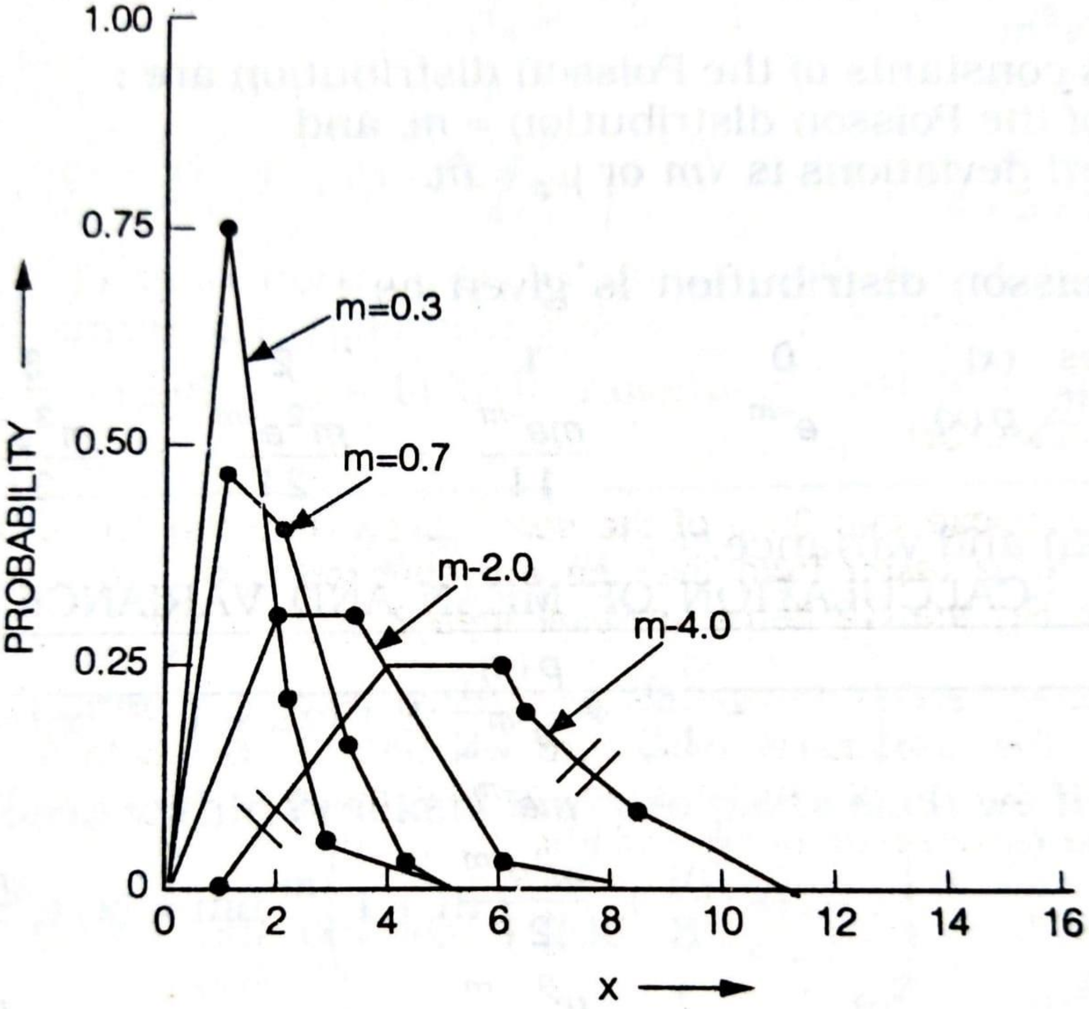


A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space. The Poisson distribution has only one parameter, λ (lambda), which is the mean number of events.

$$P(r) = \frac{e^{-m} m^{r}}{r!}$$

$$\begin{aligned}
&\textit{where}\\
&r = 0,1,2,3,4....\\
&e = 2.7183(\textit{the base of natural Logarithms})\\
&m = \textit{the mean of the Poisson distribution}
\end{aligned}$$

The Poisson distribution is a discrete distribution with a single parameter m. As m increases, the distribution shifts to the right. This is explained from the below diagram.



All Poisson probability distributions are skewed to the right. This is the reason why the Poisson probability distribution has been called the probability distribution of area events.

Role of Poisson Distribution:

1. It is used in quality control statistics to count the number of defects of an item.
2. In biology to count the number of bacteria.
3. In physics to count the number of particles emitted from a radioactive substance.
4. In insurance problems to count the number of causalities.

5. In waiting-time problems to count the number incoming telephone calls or incoming customers.
6. Number of traffic arrivals such as trucks at terminals, aeroplanes at airports, slips at docks, and so forth.
7. In determining the number of deaths in a district in a given period, say, a year, by a rare disease,
8. The number of typographical errors per page in typed material, the number of deaths as a result of road accidents, etc.,
9. In problems dealing with the inspection of manufactured products with the probability that any one piece is defective is very small and the lost are very large
10. To model the distribution of the number of persons joining a queue to receive a service or purchase of a product.

Normal Distribution:

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The binominal and the Poisson distribution described above are the most useful theoretical distributions for discrete variables, i.e., they relate to the occurrence of distinct events. In order to have mathematical distribution suitable for dealing with quantities whose magnitude is continuously variable, a continuous distribution in needed. The normal distribution, also called the normal probability distribution, happens to be most useful theoretical distribution for continuous variables. Many statistical data concerning business and economic problems are displayed in the form of normal distribution. In fact normal distribution is the cornerstone of modern statistics.

Importance of the Normal Distribution

The normal distribution has been long occupied a central place in the theory of statistics. Its importance will be clear from the following points.

1. The normal distribution has the remarkable property stated in the socalled central limit theorem. According to this theorem as the sample size n increases the distribution of mean, $\overline{X}$ of a random sample taken from practically any population approaches a normal distribution.
2. As n becomes large the normal distribution serves as a good approximation of many discrete distributions whenever the exact discrete probability is laborious to obtain or impossible to calculate accurately.
3. In theoretical statistic many problems can be solved only under the assumption of a normal population
4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. The normal distribution is used extensively in statistical quality control in industry in setting up of control limits.

Properties of Normal Distribution:

The following are the important properties.

1. The normal curve is bell-shaped curve and symmetrical in its appearance. If the curves were folded along its vertical axis, the two halves would coincide.
2. The height of the normal curve is at its maximum at the mean. Hence, the mean and mode of the normal distribution coincide. Thus mean, median and mode are equal.
3. There is one maximum point of the normal curve which occurs at the mean. The height of the curve declines as we go in either direction from the mean.

4. The curve approaches nearer and nearer to the base but it never touches it, i.e., the curve is asymptotic to the base on either side. Hence its range is unlimited or infinite in both directions.

5. Since there is only one maximum point, the normal curve is unimodal, i.e., it has only one mode.

6. The points of inflexion, I, e., the points where the change in curvature occurs are $\bar{X} \pm \sigma$.

7. As distinguished from Binomial and Poisson distribution where the variable is discrete, the variable distributed according to the normal curve is a continuous one.

8. The first and third quartiles are equidistant from the median.

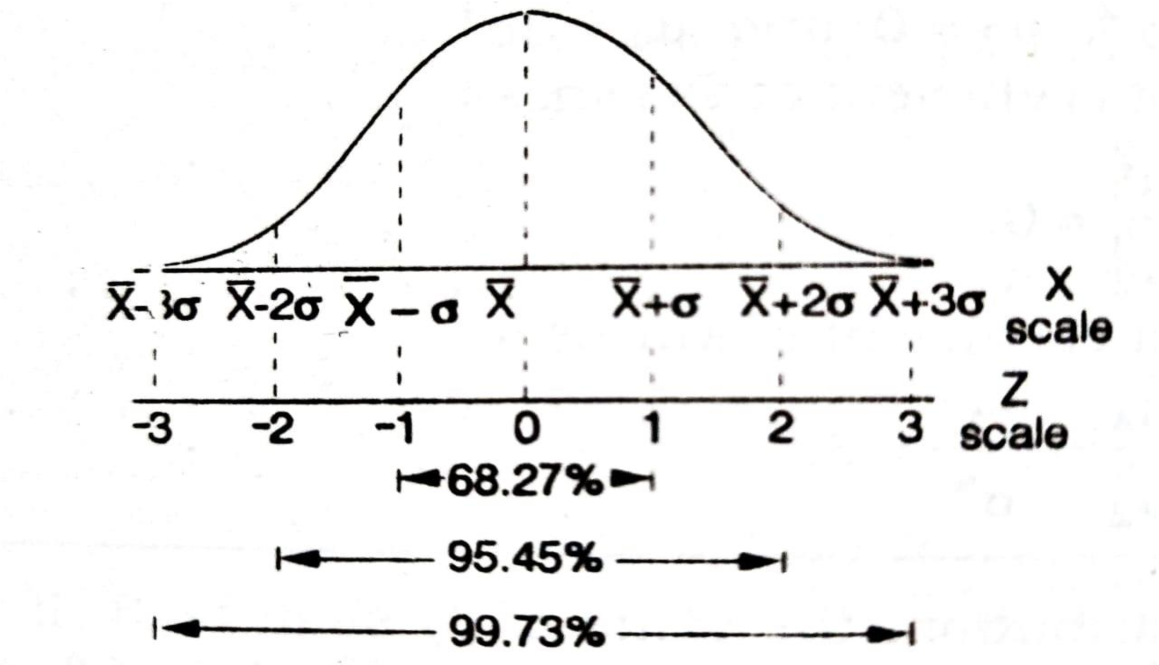
9. The mean deviation is 4th or more precisely 0.7979 of the standard deviation.

10. The area under the normal curve distributed as follows:

$a) Mean \pm 1\sigma \operatorname{cov} ers\, 68.27\%\, area; 34.135\%\, area\, will\, lie\, on\, either\, side\, of\, the\, mean$

$b) Mean \pm 2\sigma \operatorname{cov} ers\, 95.45\%\, area$

$c) Mean \pm 3\sigma \operatorname{cov} ers\, 99.73\%\, area$



Significance of the Normal Distribution

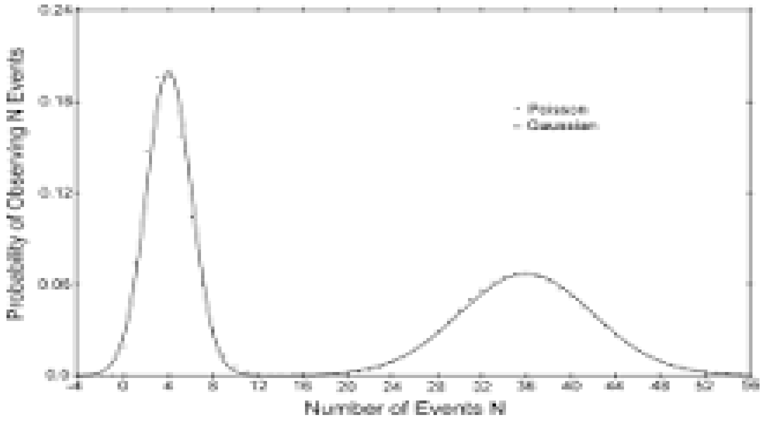
Normal distribution is mostly used for the following purposes.

1. To approximate of fit a distribution of measurement under certain conditions.
2. To approximate the binomial distribution and other discrete of continuous probability distribution under suitable condition.
3. To approximate the distribution of means and certain other quantities calculated from samples, especially large samples.

Difference between Poisson process and normal distribution

A Poisson distribution model helps find the probability of a given number of events in a time period, or the probability of waiting time until the next event in a Poisson continuous process (where certain events occur randomly and independently but at a rate).

Difference between Poisson and Gaussian curve



The Poisson function is defined only for a discrete number of events, and there is zero probability for observing less than zero events. The Gaussian function is continuous and thus takes on all values, including values less than zero as shown for the μ = 4 case.

UNIT IV

SAMPLING

Methods of Sampling

There are two type of method for finding sampling namely probability of sampling method and non – probability of sampling method let us we take probability of sampling method.

Meaning of Sampling:

Sampling is the process of selecting a proportion of the whole population or universe and drawing conclusions about the characteristics of the entire population or universe. For example, if we want to know the quality of sugar in a bag, only a spoonful of sugar may be taken and thereby get an idea about the quality of the sugar in the entire bag. Similarly, if a doctor wants to diagonise a patient 's disease, a few drops of blood can be sucked for the test and thereby gets an idea about the patient's disease. Accordingly, he may prescribe for the disease.

METHODS OF SAMPLING

Sampling methods are of two broad categories-random or probability sampling and non-random or non-probability sampling.

RANDOM SAMPLING METHODS

The random or probability sampling includes:

- Simple Random Sampling
- Systematic Random Sampling
- Stratified Random Sampling
- Cluster Sampling
- Multi-Stage Sampling.

NON-RANDOM SAMPLING METHODS

The non-random or non-probability sampling methods includes

- Convenience Sampling
- Judgement Sampling
- Quota Sampling

Judgement Sampling:

Judgement sampling is where the auditor using his own experience and knowledge of the client's business and circumstances selects the sample to be tested without use of any mathematical or statistical tools. Statistical sampling is the drawing of inferences about a large volume of data by an examination of a sampling using statistical methods in its selection.

Merits

a)The apporach is well understood and has been refined refined by experience over many years;

b)The auditor is given an opportunity to bring his judgement and expertise into play. After all auditing is an exercise in professional judgement;

c) No special knowledge of statistics is required;

d) No time is wasted playing mathematics;

Demerits

a) It is unscientific;
b) It is wasteful and usually too large samples are selected;
c) You cannot extrapolate the results to the population as a whole as the sampling are not representative;
d) Personal bias in selecting the sample is unavoidable;
e) There is no logic to the selection of the sample or its size;

f) The sample selection is so erratic that it cannot be said to have applied to all item in the year;

g) The conclusions reached are usually vague.

Judgement sampling is still the preferred methods by the majority of auditors and this defended on the grounds that the auditor several pieces of evidence and is investigating several things at the same time that the whole process is too complex to be reduced to sample formulas.

Convenience Sampling:

Convenience sampling is a non- random sampling method in which the investigation will decide the selection of sampling units based on their convenience. The sampling units for this type of sampling are selected from a telephone directory, newspaper subscribers list, departmental stores, etc.

For example, if we want to know the particulars of oil merchants, we can take the selection of oil merchants from telephone directory.

Hence the results obtained by following convenience sampling methods can hardly be representative of the population they are generally based and unsatisfactory, However, convenience sampling is often used for making pilot studies, Questions may be tested and preliminary information may be obtained by the chunk before the final sampling design is decided upon.

Merits:

a) Simplicity of sampling and the ease of research

b) Helpful for pilot studies and for hypothesis generation

c) Data collection can be facilitated short duration of time

d) Cheapest to implement that alternative sampling methods

Demerits:

a) Highly vulnerable to selection bias and influences beyond the control of the researcher
b) High level of sampling error
c) Studies that use convenience sampling have little credibility due to reasons above.

Quota Sampling:

Quota sampling is a type of judgment sampling and is perhaps the most commonly used sampling technique in non-probability category. In a quota sample, quota are set up according to some specified characteristics such as so many in each of several income groups, so many in each age, so many with certain political or religious affiliations, and so on. Each interviewer is then told to interview a certain number of persons which constitute hid quota. Within the quota, the selection of sample items depends on personal judgment, For example, in a radio listening survey, the interviewers may be told to interview 500 people living in a certain area and that out of every 100 persons interviewed 60 are to be housewives, 25 farmers and 15 children under the age of 15 Within these quotas the interviewer is free select the people to be interviewed. The cost per person interviewed may be relatively small for a quota sample but there are numerous opportunities for bias which may invalidate the result. For example, interviewers may miss farmers working in the fields or talk with those housewives who are at home. If a person refuses to respond, the interviewer simply selects someone else. Because of the risk of personal prejudice and bias entering the process of selection, the quota sampling is not widely used in a practical work.

Quota sampling and stratified random sampling are similar in as much as in both methods the universe is divided into parts and the total sample is allocated among the parts. However, the two procedures diverge radically. In stratified random sampling the sample with each stratum is chosen at random. In quota sampling, the sampling within each cell is not done at random, the field representatives are given wide latitude in the selection of respondents to meet their quotas.

Merits:

a) Relatively easy to administer

b) Can be performed quickly

c) Cost-effective

d) Accounts for population proportions

e) A useful methods when probability sampling techniques are not possible

Demerits:

a) Sample selection is not random

b) There is a potential for selection bias, which can result in a sample that is unrepresentative of the population

PROBABILITY SAMPLING:

- Each simple unit has an equal chance of being selected
- Sampling unit has an varying probability of being selected
- Probability of selection of a unit is proportional to the sample size

Simple random sampling:

Simple random sampling is the technique in which sample is so drawn that each and every unit in the population has an equal and independent change of being included in the sample

Selection of a Simple Random Sampling:

Which is based on the population

- Lottery method
- Use of table of random Number

Lottery Method:

The simplest method of drawing a random sample is the lottery system this consists in identifying each and every number or unit of the population with a distinct number which is recorded on a slip or a card. These slips should be as homogeneous as possible in shape, size, colour to avoid the human bias these sampling units corresponding to the numbers on the selected slip will constitute a random sample

Use of Table of Random Numbers :

The lottery method described above is quite time consuming and cumbersome to use if the population to be sampled is sufficiently large moreover in this method it is not humanly possible to make all the slip or card exactly alike and as such some bias is likely to be introduction statisticians have avoided this difficult by considering random sampling number series most of these series are the result of actual sampling operations recorded for future use

Restricted Sampling:

Restricted sampling included stratified, systematic and multistage sampling

Stratified Sampling :

- ➢ When the population is heterogeneous with respect to the variable or characteristic under study then the technique of stratified random sampling is used to obtain more efficient result stratification means division into layers or groups

Merits

- ❖ The division of the population into relatively homogenous sub group bring administration convenience unlike random sample the stratified sample are expected to be localised geographically this ultimately result in reduction in cost and saving in term of known precision for each of each of the stratum

- Sometime it is desired to achieve different of accuracy for different segment of the population stratified random sampling is the only sampling plan which enables us to obtain the result of known precision for each of the stratum.

Demerits :

a) Disproportional stratified sampling requires the assignment of weights to different strata and if the weight assigned are faulty, the resulting sample will not be and might based result.

b) Effective stratification of the universe into homogeneous strata. Appropriate size of the samples to be drawn the error due to wrong stratification cannot be compensated by taking large samples.

Systematic Sampling :

Systematic sampling is slight variation of the simple random sampling in which only

- The first sample unit is selected at random and remaining units are automatically selected in a definite sequence at equal spacing from one another.

Merits:

- Systematic sampling is very easy to operate and checking can also be done quickly accordingly, it results in considerable saving in tome and labour relative to simple random sampling or stratified random sampling .

Demerits:

Systematic sampling works well only if the complete and up to data frame is available and if they are randomly arranged. However, these requirement are randomly arranged. However, these requirements are not generally fulfilled.

Multistage Sampling:

Multistage sampling consists in sampling first stage units by some suitable method of sampling from among the selected first stage units, a sub-sample of secondary stage units is drawn by some suitable method of sampling which may be same as or different from the method used in selecting first stage units further stages may be added to arrive at a sample the desired sampling units.

Merits:

Multistage sampling is more flexible as compared to other methods of sampling it is simple to carry out and results in administrative convenience by permitting the field work to be concentrated and yet converting large area.

Demerits:

Errors are likely to be larger in this method than in any other method. The variability of the estimates under this method may be greater than that estimates based on simple random sampling. This variability depends on the composition of the primary units. In general, a general a multistage.

Sampling is usually less efficient than a suitable single stage sampling of the same size.

Sampling Theory

The best way to represent a population is to enumerate its members before selecting a random sample from that population. When properly implemented, this guarantees that the sample will formally represent the population within known limits of sampling error.

Probability Sampling Types

Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

Importance in sampling theory

The idea behind importance sampling is that certain values of the input random variables in a simulation have more impact on the parameter being estimated than others. If these "important" values are emphasized by sampling more frequently, then the estimator variance can be reduced.

Sampling Distribution

A sampling distribution isa probability distribution of a statistic that is obtained through repeated sampling of a specific population. It describes a range of possible outcomes for a statistic, such as the mean or mode of some variable, of a population.

Types of Sampling Distributions

Here is a brief description of the types of sampling distributions:

- **Sampling Distribution of the Mean:** This method shows a normal distribution where the middle is the mean of the sampling distribution. As such, it represents the mean of the overall population. In order to get to this point, the researcher must figure out the mean of each sample group and map out the individual data.
- **Sampling Distribution of Proportion:** This method involves choosing a sample set from the overall population to get the proportion of the sample. The mean of the proportions ends up becoming the proportions of the larger group.
- **T-Distribution:** This type of sampling distribution is common in cases of small sample sizes. It may also be used when there is very little information about the entire population. T-distributions are used to make estimates about the mean and other statistical points.

Parameter and Statistic

A parameter is a number describing a whole population (e.g., population mean), while a statistic is a number describing a sample (e.g., sample mean).

The difference between parameter and statistic

The key difference between parameters and statistics is that parameters describe populations, while statistics describe samples. You can easily remember this distinction using the alliterations for population, parameter, and sample statistic.

Points	Statistic	Parameter
1	Derived from sample data	Derived from population data
2	Used to estimate population characteristics	Represents population characteristics
3	Subject to sampling variability	Fixed value
4	Provides information about a sample	Provides information about a population
5	Varied values across different samples	Consistent value for the entire population
6	Estimation based on inference techniques	Known or can be determined with complete data
7	Used to draw conclusions about a population	Describes a population
8	Often denoted using Greek letters	Often denoted using English letters
9	Can change with different samples	Remains constant for a specific population
10	Used in hypothesis testing and confidence intervals	Used in defining populations and subgroups

UNIT V
TESTING OF HYPOTHESIS

Testing of Hypothesis

Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

There are three types of hypothesis tests: right-tailed, left-tailed, and two-tailed. When the null and alternative hypotheses are stated, it is observed that the null hypothesis is a neutral statement against which the alternative hypothesis is tested.

Procedure of Testing Hypothesis

The procedure of testing hypothesis is as follows:

1. Set up a hypothesis

The null hypothesis can be thought of as the opposite of the "guess" the researchers made: in this example, the biologist thinks the plant height will be different for the fertilizers. So the null would be that there will be no difference among the groups of plants. Specifically, in more statistical language the null for an ANOVA is that the means are the same.

2. Set up a suitable significance level

The significance level is typically set equal to such values as 0.10, 0.05, and 0.01. The 5 percent level of significance, that is, $\alpha = 0.05$, has become the most common in practice. Since the significance level is set to equal some small value, there is only a small chance of rejecting H_0 when it is true.

3. Setting a test criterion

This involves selecting an appropriate probability distribution for the particular test, that is, a probability distribution which can properly be applied.

4. Doing Computations

A computation is any type of arithmetic or non-arithmetic calculation that is well-defined. Common examples of computations are mathematical equations.

5. Making Decisions

Finally as a fifth step, we may conclude statistical conclusions and take decisions. A statistical conclusions or statistical decision is a decision either to reject or to accept the null hypothesis.

The steps of testing hypothesis

Table of contents

- Step 1: State your null and alternate hypothesis.
- Step 2: Collect data.
- Step 3: Perform a statistical test.
- Step 4: Decide whether to reject or fail to reject your null hypothesis.
- Step 5: Present your findings.

The advantages of hypothesis

A hypothesis can help you to formulate a specific and testable research problem, and to design an appropriate method to collect and analyze data. A hypothesis can also help you to establish a clear direction and focus for your research, and to communicate your expectations and assumptions to your readers or audience.

Hypothesis test to use

A z-test is used to test a Null Hypothesis if the population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population variance is unknown.

Characteristics of the hypothesis:

- The hypothesis should be clear and precise to consider it to be reliable.

- If the hypothesis is a relational hypothesis, then it should be stating the relationship between variables.
- The hypothesis must be specific and should have scope for conducting more test.

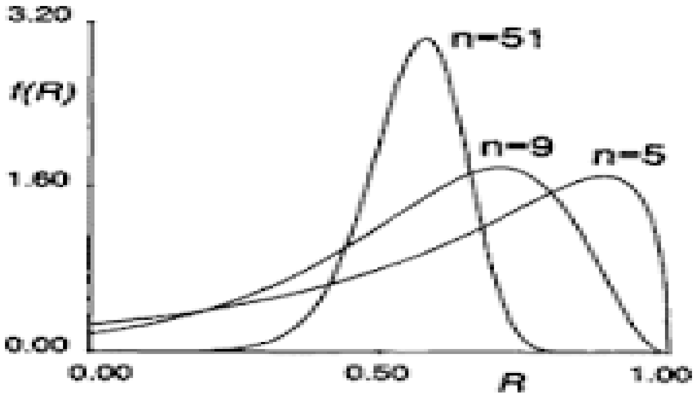
Level of Significance

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true. The level of significance is stated to be the probability of type I error and is preset by the researcher with the outcomes of error.

The 4 levels of significance

The null hypothesis is a hypothesis that states that the data has no result, association between variables, or disparity between variables. There are four levels in statistics that are organized by level of complexity and precision. They are nominal, ordinal, interval, and ratio.

5 significance level



The significance level is typically set equal to such values as 0.10, 0.05, and 0.01. The 5 percent level of significance, that is, α = 0.05, has become the most common in practice. Since the significance level is set to equal some small value, there is only a small chance of rejecting H_0 when it is true.

Two Types of Errors in testing of Hypothesis

1. The hypothesis is true but our test rejects it (Type I error).
2. The hypothesis is false but our test accepts it (Type II error).

3. The hypothesis is true and our test accepts it (Correct Decision).
4. The hypothesis is false and our test rejects it (Correct Decision).

Type I and Type II Errors

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

Table of Type I and Type II Error

The relationship between truth or false of the null hypothesis and outcomes or result of the test is given in the tabular form:

Error Types	**When H_0 is True**	**When H_0 is False**
Don’t Reject	Correct Decision (True negative) Probability = $1 – \alpha$	Type II Error (False negative) Probability = β
Reject	Type II Error (False Positive) Probability = α	Correct Decision (True Positive) Probability = $1 – \beta$

Type I and Type II Errors Example

Check out some real-life examples to understand the type-I and type-II error in the null hypothesis.

Example 1: Let us consider a null hypothesis – A man is not guilty of a crime.

Then in this case:

Type I error (False Positive)	Type II error (False Negative)
He is condemned to crime, though he is not guilty or committed the crime.	He is condemned not guilty when the court actually does commit the crime by letting the guilty one go free.

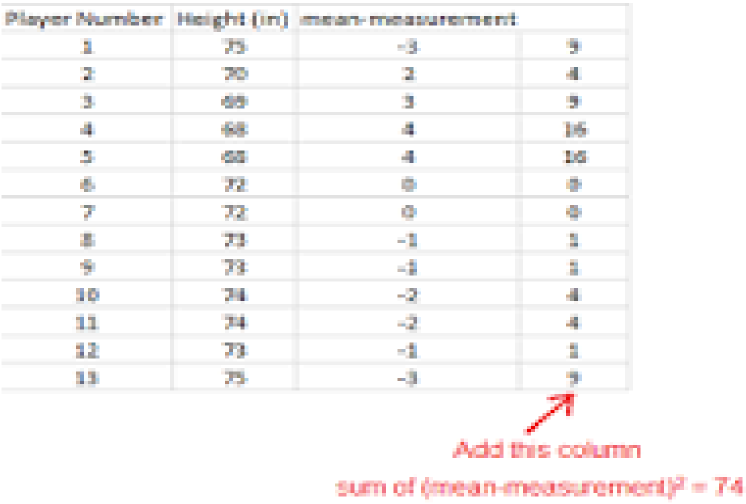
Example 2: Null hypothesis- A patient’s signs after treatment A, are the same from a place box.

Type I error (False Positive)	Type II error (False Negative)
Treatment A is more efficient than the placebo	Treatment A is more powerful than placebo even though it truly is more efficient.

STANDARD ERROR

Standard error is the approximate standard deviation of a statistical sample population. The standard error describes the variation between the calculated mean of the population and one which is considered known, or accepted as accurate.

An example of using standard error



Player Number	Height (in)	mean-measurement	
1	75	-3	9
2	70	2	4
3	69	3	9
4	68	4	16
5	68	4	16
6	72	0	0
7	72	0	0
8	73	-1	1
9	73	-1	1
10	74	-2	4
11	74	-2	4
12	73	-1	1
13	75	-3	9

Add this column

sum of (mean-measurement)^2 = 74

For example, you would construct a 95% confidence interval by adding and subtracting 1.96 times the standard error from the sample mean. Therefore, the 95% confidence interval for high school basketball player height would be 70.65 inches to 73.35 inches.

Standard Error

$$SE = \frac{\sigma}{\sqrt{n}}$$

Standard deviation

Number of samples

Standard error is calculated by dividing the standard deviation of the sample by the square root of the sample size. Calculate the mean of the total population. Calculate each measurement's deviation from the mean.

The symbol for standard error

The standard error of a statistic is usually designated by the Greek letter sigma (σ) with a subscript indicating the statistic. For instance, the standard error of the mean is indicated by the symbol: σ_M.

Properties of Good Estimator

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

An example of estimate

We need to estimate how much paint we'll need for the job. The cost of the project has been estimated at/as about 10 million dollars. He estimates that current oil reserves are 20 percent lower than they were a year ago. Damage from the hurricane is estimated (to be) in the billions of dollars.

Difference between Large and small samples

When the sample size is under 30, statisticians are supposed to use the Student T distribution instead. It has a much greater chance of being wrong. In statistical context, a sample is considered to be large if it is at least 30. On the other hand, a sample is considered small is it is less than 30.

The difference between a small sample size and a large sample size lies in the number of observations or data points included in each sample. Here's a comparison of the characteristics of small and large sample sizes:

Small Sample Size:

1. Limited Representation: A small sample size may not fully represent the population from which it is drawn. It may not capture the full range of variability and characteristics present in the population.

2. Higher Sampling Error: Small samples tend to have higher sampling error or variability. The observed data points may deviate more from the true population values, leading to less precise estimates or conclusions.

3. Reduced Statistical Power: Small samples may have lower statistical power, making it more challenging to detect significant effects or relationships. This can increase the likelihood of Type II errors (failing to detect true effects).

4. Narrow Confidence Intervals: With smaller sample sizes, the confidence intervals around estimates or statistical parameters tend to be wider, reflecting increased uncertainty or imprecision in the results.

5. Limited Generalizability: The findings or conclusions from a small sample may have limited generalizability to a larger population. The relationships or patterns observed in the small sample may not hold true for the broader population.

Large Sample Size:

1. Improved Representation: A large sample size is more likely to capture the characteristics and variability of the population. It provides a better representation of the overall population, leading to more reliable and generalizable results.

2. Lower Sampling Error: Large samples tend to have lower sampling error or variability. The observed data points are closer to the true population values, resulting in more precise estimates and reduced random fluctuations.

3. Higher Statistical Power: Large samples have higher statistical power, enabling a better chance of detecting significant effects or relationships. This reduces the risk of Type II errors.

4. Narrow Confidence Intervals: With larger sample sizes, the confidence intervals around estimates or statistical parameters tend to be narrower, indicating greater precision and reduced uncertainty in the results.

5. Enhanced Generalizability: Findings from a large sample are more likely to be generalizable to the broader population, increasing the confidence in applying the conclusions to a wider context.

In summary, larger sample sizes generally offer more accurate and reliable estimates, higher statistical power, and increased generalizability. However, the appropriate sample size depends on the research question, desired level of accuracy, effect size, variability, and statistical techniques employed. Statisticians employ power analysis and other methods to determine the sample size needed for specific research studies.

Test of Significance for Large Samples

Before moving to large samples, test of significance has to be seen in brief. Test of significance is performed after framing the hypothesis (tentative statements) at say 1%, 5% and 10% level. The level of significance (denoted as α or alpha) represents the probability of error or chances of making wrong decisions.

Statistical test is used for large sample size

If the frequency of success in two treatment groups is to be compared, Fisher's exact test is the correct statistical test, particularly with small samples.

Test for Two Means and Standard Deviations

The 2-Sample Standard Deviation test compares the standard deviations of 2 samples, and the Standard Deviations test compares the standard deviations of more than 2 samples. In this paper, we refer to k-sample designs with k = 2 as 2- sample designs and k-sample designs with k > 2 as multiple-sample designs.

The t-test for means and standard deviation

The t-test is a test used for hypothesis testing in statistics. Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values. T-tests can be dependent or independent.

Used for standard deviation

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance (or standard deviation).

The t-test with means

A t test is used to measure the difference between exactly two means. Its focus is on the same numeric data variable rather than counts or correlations between multiple variables.

Proportion and Confidence of Fit

Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers. The build a confidence interval for population proportion p, we use: ˆp–zα2⋅√ˆp(1−ˆp)n<p<ˆp+zα2⋅√ˆp(1−ˆp)n.

Therefore, the 99% confidence interval is 0.37 to 0.43. That is, we are 99% confident that the true proportion is in the range 0.37 to 0.43.

Small sample Test

t, and F χ -tests are some commonly used small sample tests. Unit which are based on χ2 and F-distributions described in Unit 3 and Unit 4 of this course respectively. This unit is divided into eight sections.

Best for small sample size

If the frequency of success in two treatment groups is to be compared, Fisher's exact test is the correct statistical test, particularly with small samples.

Small sample size sampling

The size of the sample is small when compared to the size of the population. When the target population is less than approximately 5000, or if the sample size is a significant proportion of the population size, such as 20% or more, then the standard sampling and statistical analysis techniques need to be changed.

t-test:

A t-test is a statistical test that compares the means of two samples. It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

The three types of t-tests

There are three forms of Student's t-test about which physicians, particularly physician-scientists, need to be aware: (1) one-sample t-test; (2) two-sample t-test; and (3) two-sample paired t-test.

(1) one-sample t-test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

$$t = \frac{(\overline{X_1} - \mu)\sqrt{n}}{S}$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$$

(2) Two-sample t-test

The two-sample t-test (also known as the independent samples t-test) is a method used to test whether the unknown population means of two groups are equal or not.

$$t = \frac{\overline{X_1} - \overline{X_2}}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$S = \sqrt{\frac{\sum (X_1 - \overline{X}_1)^2 + \sum (X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}}$$

Example:1

Two types of drugs were used on 5 and 7 patients for reducing their weight.

Drug A was imported and drug B indigenous. The decrease in the weight after using the drugs for six months was as follows.

Drug A	10	12	13	11	14		
Drug B	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs? If no, which drug should you buy. (forv=10, $t_{0.05}$=2.223)

Hypothesis: There is no significant difference in the efficacy of the two drugs. Apply t-test.

X_1	**(X_1- $\overline{X}$)**	**(X_1- $\overline{X_1}$)^2**	**X_2**	**(X_2- $\overline{X}_2$)**	**(X_1- $\overline{X}_2$)^2**
10	-2	4	8	-3	9
12	0	0	9	-2	4
13	1	1	12	1	1
11	-1	1	14	3	9
14	2	4	15	4	16
			10	-1	1
			9	-2	4
$\sum X_1 = 60$		$\sum (X_1 - \overline{X_1})^2 = 10$	$\sum X_2 = 77$		$\sum (X_2 - \overline{X_2})^2 = 44$

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{60}{5} = 12$$

$$\overline{X_2} = \frac{\sum X_2}{n_2} = \frac{77}{7} = 11$$

$$S = \sqrt{\frac{\sum (X_1 - \overline{X}_1)^2 + \sum (X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{10+44}{5+7-2}} = \sqrt{\frac{54}{10}} = 2.324$$

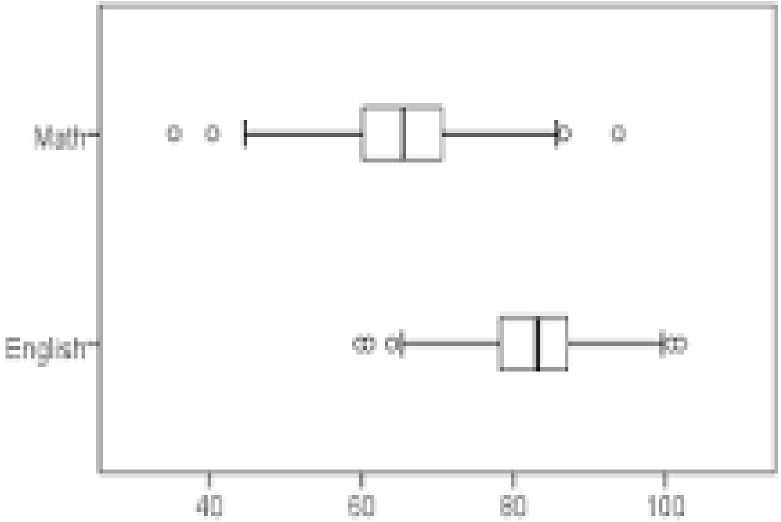
$$t = \frac{\overline{X_1} - \overline{X_2}}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$t = \frac{12-11}{2.324} \times \sqrt{\frac{5\times 7}{5+7}}$$

$$t = \frac{1.708}{2.324} = 0.735$$

The calculated value of t is less than the table value, the hypothesis is accepted. Hence, the hypothesis is accepted. We should buy indigenous drug.

(3) two-sample paired t-test.



The Paired Samples t Test compares the means of two measurements taken from the same individual, object, or related units. These "paired" measurements can represent things like: A measurement taken at two different times (e.g., pre-test and post-test score with an intervention administered between the two time points).

Chi-square test:

The χ^2 test (chi-square) is one of the simplest and most widely used non-parametric test in statistical work. The symbol χ^2 is the Greek letter Chi. The χ^2 test was first used by Karl Pearson in the year 1900.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

For a Chi-square test, a p-value that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. You can conclude that a relationship exists between the categorical variables.

It is defined as $\chi^2 = \frac{\sum (O-E)^2}{E}$

Where O refers the observed frequencies and E refers to the expected frequencies. To calculate the expected frequencies

$$E = \frac{RT \times CT}{N}$$

E = Expected frequency

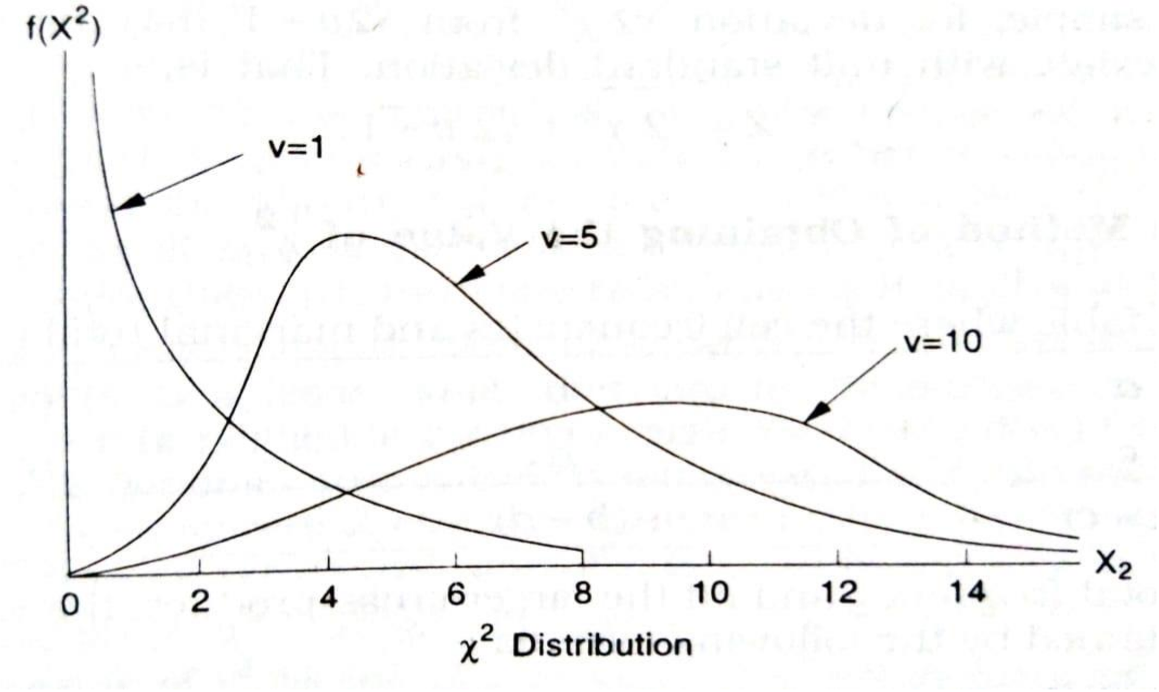
RT = The row for the row containing the cell

CT = the column total for the column containing the cell

N = the total number of observations.

The Chi-Square distribution

The following diagram gives the χ^2 distribution for 1.54 and 10 degree of freedom.



Uses of χ^2 test

1. χ^2 test as a test of independence
2. χ^2 test as a goodness of fit
3. χ^2 test as a test of homogeneity

Example-1

IN an anti-dengue campaign in a certain area, quinine was administered to 812 persons out of total population3248. The number of fever cases as follows:

Treatment	Fever	No Fever	Total
Quinine	**20**	**792**	**812**
No Quinine	**220**	**2216**	**2436**
Total	**240**	**3008**	**3248**

Check the usefulness of quinine in checking dengue.

Hypothesis: There is not effective in checking dengue. Apply χ^2 test.

Solution:

Expectation of (AB) $= \frac{A \times B}{N} = \frac{240 \times 812}{3248} = 60$

Likewise find all the expected frequencies.

60	**752**	**812**
180	**2258**	**2436**
240	**3008**	**3248**

O	E	$(O-E)^2$	$(O-E)^2 / E$
20	**60**	**1600**	**26.667**
220	**180**	**1600**	**8.889**
792	**752**	**1600**	**2.128**
2216	**2256**	**1600**	**0.709**
$\sum (O-E)^2 / E$			**38.393**

$$\chi^2 = \sum (O-E)^2 / E$$
$$= 38.393$$

The calculated value of χ^2 is greater than the table value. The hypothesis is rejected. Hence the quinine is useful in checking dengue.

Test and Goodness of Fit

In the previous chapter various tests of significance such as t, F and Z were discussed. These tests were based on the assumption that the samples were drawn from normally distributed populations, or more accurately that the sample means were normally distributed. Since the testing procedure request assumption about the type of the population or parameters.

Though non-parametric theory developed as early as the middle of the nineteenth

Century, it was only after 1945 that non-parametric test came to be used widely. Originated in sociological and psychological research, non-parametric tests today are very popular in behavioural sciences. The following three reasons account for the increasing use of non-parametric tests in business research:

(1) These statistical tests are distribution-free (can be used with any shape of population distribution)

(2) They are usually computationally easier to handle and understand than parametric tests;

(3) They can be used with types of measurements that prohibit the use of parametric tests.

The increasing popularity of non-parametric tests should not lead the reader to form an impression that they are usually superior to the parametric tests. In fact, in a situation where parametric and non-parametric tests both apply, the former are more desirable than the latter.

Introduction of F Test Formula

The F Test Formula is a Statistical Formula used to test the significance of differences between two groups of Data. It is often used in research studies to determine whether the difference in the means of two populations is statistically significant. It is based on the F Statistic, which is a measure of how much variation exists in one group of Data compared to another. Students who are studying for their Statistics course will need to be familiar with this Formula. Our article will provide a detailed explanation of how to use the F Test Formula. It will also provide examples of how to use it in practice. The use of the F Test Formula is a critical step in any research study, and it is important to understand how to use it correctly. You will be able to find the F Test Formula in most Statistics textbooks.

Definition of F-Test Statistic Formula

It is a known fact that Statistics is a branch of Mathematics that deals with the collection, classification and representation of data. The tests that use F - distribution are represented by a single word in Statistics called the F Test. F Test is usually used as a generalized Statement for comparing two variances. F Test Statistic Formula is used in various other tests such as regression analysis, the chow test and Scheffe test. F Tests can be conducted by using several technological aids. However, the manual calculation is a little complex and time-consuming.

F-Test is a test Statistic that has an F distribution under the null hypothesis. It is used in comparing the Statistical model with respect to the available Data set. The name for the test is given in honour of Sir. Ronald A Fisher by George W Snedecor. To perform an F Test using technology, the following aspects are to be taken care of.

- State the null hypothesis along with the alternative hypothesis.
- Compute the value of ‘F’ with the help of the standard Formula.
- Determine the value of the F Statistic.
- The ratio of the variance of the group of means to the mean of the within-group variances.
- As the last step, support or reject the Null hypothesis.

Assumptions in F test

- The populations are characterized as having a normal distribution.
- The populations are independent from one another.
- When calculating the F-statistic, the larger variance is used as the numerator, and the smaller variance is used in the denominator.

Assumptions of the F-test

The F-test is based on two assumptions: (1) the samples are normally distributed, and (2) the samples are independent of each other. If these assumptions are fulfilled and H_0 is true, the statistic F follows an F-distribution.

F-Test Equation to Compare Two Variances:

In Statistics, the F-test Formula is used to compare two variances, say σ1 and σ2, by dividing them. As the variances are always positive, the result will also always be positive. Hence, the F Test equation used to compare two variances is given as:

$$F=\frac{S_1^{\,2}}{S_2^{\,2}}\ ,\qquad S_1^{\,2}=\frac{(X_1-\overline{X_1})^2}{n_1-1}\ ;\qquad\qquad S_2^{\,2}=\frac{(X_2-\overline{X_2})^2}{n_2-1}$$

It should note that $S_1^{\,2}$ is always the larger estimate of variance.

$$F=\frac{L\arg er\,estimate\,of\ \operatorname{var}iance}{Smaller\,estimate\,of\ \operatorname{var}iance}$$

F Test Formula helps us to compare the variances of two different sets of values. To use F distribution under the null hypothesis, it is important to determine the mean of the two given observations at first and then calculate the variance.

In the above formula, σ2 is the variance x is the values given in a set of data x is the mean of the given data set n is the total number of values in the data set, While running an F Test, it is very important to note that the population variances are equal. In more simple words, it is always assumed that the variances are equal to unity or 1. Therefore, the variances are always equal in the case of the null hypothesis.

The conditions to use F-test

In order to use an ANOVA F-Test, each group must be normally distributed, the groups must have the "same" variance, and the samples must be randomly selected in an independent manner.

Analysis of Variance: Assumptions

There are three primary assumptions in ANOVA: The responses for each factor level have a normal population distribution. These distributions have the same variance. The data are independent.

1. Normality
2. Homogeneity
3. Independence of error

One-way and two-way Classification

The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two. One-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.

An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

Basically, you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

What Does "One-Way" or "Two-Way Mean?

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test.

- One-way has one independent variable (with 2 levels). For example: brand of cereal,
- Two-way has two independent variables (it can have multiple levels). For example: brand of cereal, calories.

Types of Tests.

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

- One-way ANOVA between groups: used when you want to test two groups to see if there’s a difference between them.
- Two way ANOVA without replication: used when you have one group and you’re double-testing that same group. For example, you’re testing one set of individuals before and after they take a medication to see if it works or not.
- Two way ANOVA with replication: Two groups, and the members of those groups are doing more than one thing. For example, two groups of patients from different hospitals trying two different therapies.

One Way ANOVA

A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution. The null hypothesis for the test is that the two means are equal. Therefore, a significant result means that the two means are unequal.

Examples of when to use a one way ANOVA

Situation 1: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.
Situation 2: Similar to situation 1, but in this case the individuals are split into groups based on an attribute they possess. For example, you might be studying leg strength of people according to weight. You could split participants into weight categories (obese, overweight and normal) and measure their leg strength on a weight machine.

Limitations of the One Way ANOVA

A one way ANOVA will tell you that at least two groups were different from each other. But it won’t tell you which groups were different. If your test returns a significant f-statistic, you may need to run an ad hoc test (like the Least Significant Difference test) to tell you exactly which groups had a difference in means.

Two Way ANOVA

A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable. With a Two Way ANOVA, there are two independents. Use a two way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two way ANOVA is appropriate.

For example, you might want to find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called factors in Two Way ANOVA.

The factors can be split into levels**.** In the above example, income level could be split into three levels: low, middle and high income. Gender could be split into three levels: male, female, and transgender. Treatment groups are all possible combinations of the factors. In this example there would be 3 x 3 = 9 treatment groups.

Assumptions for Two Way ANOVA

1. The population must be close to a normal distribution.
2. Samples must be independent.
3. Population variances must be equal (i.e. homoscedastic).
4. Groups must have equal sample sizes.

An example of a two way classification

For example, one way classifications might be: gender, political party, religion, or race. Two way classifications might be by gender and political party, gender and race, or religion and race. Each classification variable is a called a factor and so there are two factors, each having several levels within that factor.

Example-1

Apply F-test. Two random samples were drawn from the two populations and their values are:

A	66	67	75	76	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance. ($F_{0.05}$= 3.36)

Hypothesis: The two populations have the same variance.

Calculation of F-test:

A	$x_1 = (X_1 - \overline{X_1})$	X_1^2	B	$(X_2 - \overline{X_2})$	X_2^2
66	-14	196	64	-19	361
67	-13	169	66	-17	289
75	-5	25	74	-9	81
76	-4	16	78	-5	25
82	2	4	82	-1	1
84	4	16	85	2	4
88	8	64	87	4	16
90	10	100	92	9	81
92	12	144	93	10	100
			95	12	144
			97	14	196
$\sum X_1 = 720$	$\sum x_1 = 0$	$\sum x_1^2 = 734$	$\sum X_2 = 913$	$\sum x_2 = 0$	$\sum x_2^2 = 1298$

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{720}{9} = 80 \qquad \overline{X_2} = \frac{\sum X_2}{n_2} = \frac{913}{11} = 83$$

$$F = \frac{S_1^2}{S_2^2} \qquad S_1^2 = \frac{(X_1 - \overline{X_1})^2}{n_1 - 1}, \qquad S_2^2 = \frac{(X_2 - \overline{X_2})^2}{n_2 - 1}$$

$$S_1^2 = \frac{734}{9-1} = 91.75 \qquad S_2^2 = \frac{1298}{11-1} = 129.8 \qquad F = \frac{S_1^2}{S_2^2} = \frac{129.8}{91.75}$$

$$F = 1.415$$

The calculated value of F is greater than the table value. The hypothesis is accepted. Hence, that the two populations have the same variance.

***