Rural Management

Business Analytics II

First Edition

Rural Management - Business Analytics II

MoE | Government of India
Ministry of Education

सत्यमेव जयते

First Edition: 2020

ISBN: 978-93-89431-09-4

Price: ₹ 750/-

5-10-174, Shakkar Bhavan, Ground Floor, Fateh Maidan Road, Hyderabad - 500 004

Telangana State. Tel: 040-23422112, 23212120, Fax: 040-23212114

E-mail : editor@mgncre.org Website : www.mgncre.org

Published by: Mahatma Gandhi National Council of Rural Education (MGNCRE), Hyderabad

About the Book

Business Analytics is a course combining Data analytics and Business Intelligence on a given set of data to develop future business strategies and plans. It is a science of good decision making in the face of uncertainty. It is also called a branch of Applied Statistics to provide knowledge and skills to interpret and use statistical techniques in a variety of business applications. A typical Business Analytics course is intended for business majors and covers mostly the topics of descriptive statistics, probability, estimation, binomial and normal distribution of data, sampling theory, testing of hypotheses, confidence intervals, variance, linear regression, correlation, etc. The objective of this book is to familiarize the students with the basic concepts of statistics; to provide insights on statistical techniques which will be useful in making business decisions; to help students to describe different aspects of data; to familiarize with evidence based interpretation of data.

Modern Rural Development policy claims to be evidence based. While the rationale of rural developments may be a mix of equity, improving economic efficiency and political economy, the details are often dependent on statistical information. Statistics can contribute to rural development policy at all stages of the process. Statistics can help to throw light on the nature of the problems to be tackled. Only when there is required quantitative information available, any action is far more likely. The problems of rural development may be biodiversity, high levels of unemployment, low wages, homelessness, lack of social participation, etc. Statistics can assist in setting up the objectives that have to be achieved to address the problems, such as increasing the percentage of rural population with educational qualification. Statistics for rural development is largely concerned with describing the conditions in areas that can be labeled as rural and comparing them with the developed areas or urban areas and thus it can form an integral part of evaluation of the performance.

The modern technology of Big Data is enormous in every business sector including rural business and its development. The usage so far has been more focused on e-commerce of rural based products. Rural India plays a vital role in economic growth of our country through agriculture, self employment, construction, services, etc. The focus of our government to build Digital India through broadband highways connecting every household, village, panchayat, government department will generate huge amount of data which can be analysed to provide solution to the various problems of Rural India and to create Smarter Villages. The benefits of Big Data are not only transforming the businesses in the urban areas but also extended to the rural parts of our country.Statistics has been given an important place in our courses of studies in various social sciences. The study of statistics has become an integral part of our everyday life.

This text book has though been written with the prime objective of fulfilling the needs of students of B.B.A – Rural management. The main objective of this book is to equip students with requisite quantitative skills that they can employ and build on in flexible ways. Students studying this course are expected to understand the fundamental of probability theory, statistical reasoning, statistical computing, and description-interpretation-exploratory analysis of data by graphical method, etc. This book will also equip the students to be able to read and interpret the statistical data, to develop critical thinking and analytical skills, to make decisions based on the data collected and analysed, and finally to draw inferences from the data.

I thank the contributors: Dr K Duraivelu, Professor of Mechanical Engineering and Dean of Faculty of Engineering & Technology, SRM Institute of Science and Technology, and Dr R Manimaran, Assistant Professor of Mathematics at SRM Institute of Science and Technology, Vadapalani Campus, Chennai to this book for their outstanding insights.

Also, I would like to thank MGNCRE Team members for extending their extreme support in completing this text book.

Dr W G Prasanna Kumar
Chairman MGNCRE

Contents

List of Tables

List of Figures

Chapter 1 Overview of Basics of Statistics

Introduction

The purpose of this chapter is to introduce you the basics of statistics. In the present age of information, data is abundantly available around us. Hence, we should have knowledge of extracting and using the data for the given applications. For example, as a student of BBA, you may be interested to know the companies offering jobs for BBA graduates and the average salary being offered to a fresh BBA graduate. So, you should be able to know where to collect the required statistics and how to interpret from the given statistics. Statistics is a discipline of mathematical science that concerns methods of collecting, organizing, displaying, analyzing, interpreting, and presenting the data in such a way that some meaningful inferences or conclusions may be drawn. The concept of statistics can be applied to scientific, industrial and social problems. Statistics include numerical facts and figures.

Objectives of the Chapter

- To explain evolution of statistics over period of time
- To infer the types, importance, scope, functions and limitations of statistics
- To classify data and explain the rules for classification
- To prepare frequency table for single variable and bi-variables
- To explain the various elements of tabulation of data and the classification of table
- To represent the given data diagrammatically in various formats
- To project the given data in a graphical format
- To apply the procedure to solve various basic statistical problems

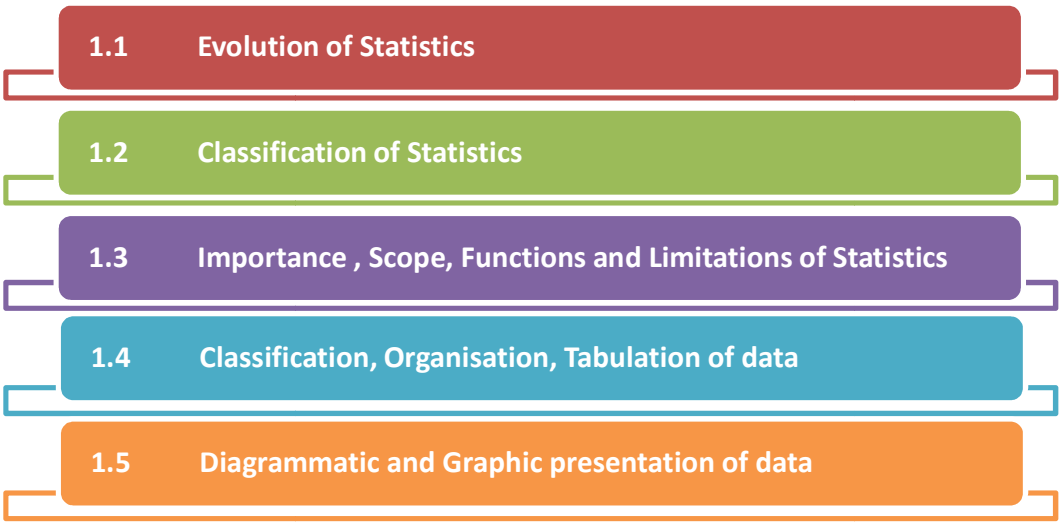


Fig. 1.1 Chapter Flow

1.1 Evolution of Statistics

The terminology 'statistics' is derived from the latin word 'statisticum collegium' and the Italian word 'statista'. During 18th century the terminology 'statistics' was used to mean the systematic collection of demographic and economic data. In the early 19th century, collection of data intensified and the meaning of statistics broadened to include collection, summary and analysis of data. In the current

world driven by technology and information, the computers have expedited more elaborate statistical computation and facilitate the collection and consolidation of data.

In the 19th century, probability theory was increasingly used in statistics. Astronomy used probability models, in particular the method of least squares, to attempt to reduce the errors in estimating the locations of various celestial bodies. The Royal Statistical Society was established in 1834. University College London founded the world's first University statistics department in 1911. Knowing the importance of statistics in almost all the fields, Professor Prasant Chandra Mahalanobis founded Indian Statistical Institute (ISI) on 17th December 1931 at Kolkatta. At present, in the fourth industrial revolution 'Industry 4.0', we talk about the emerging field of 'Machine Learning' which is the application of huge data base (statistics) and the computational algorithms. Also, the modern computer world talks about more on 'Artificial Intelligence' and 'Big data Analytics' which require collection and analysis of data.

Statistics is an indispensable tool of marketing research. It is extensively used in time and motion study, consumer behavior study, performance measurements, investment decisions, inventory management, quality control and management, marketing research, forecasting, distribution channel design, social surveys, etc.

History of Statistics is summarized in fig. 1.2

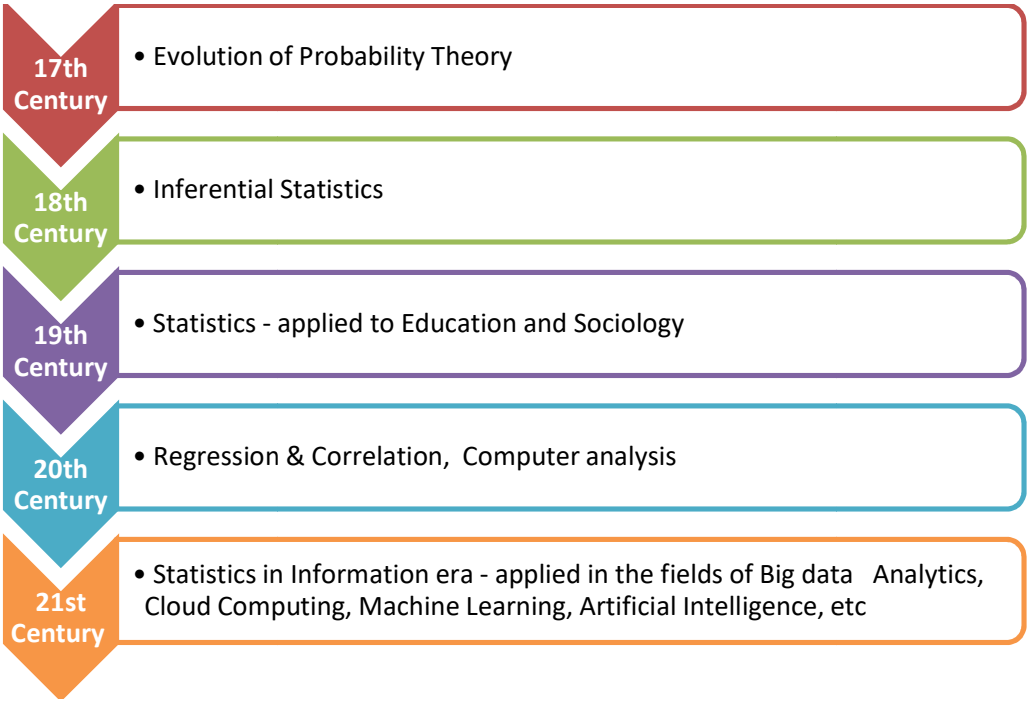


Fig. 1.2 History of Statistics

1.2 Classification of Statistics

Statistical methods are broadly divided into five categories, namely Descriptive Statistics, Inferential Statistics, Analytical Statistics, Inductive Statistics and Applied Statistics. The broad classification in shown in fig.1.3

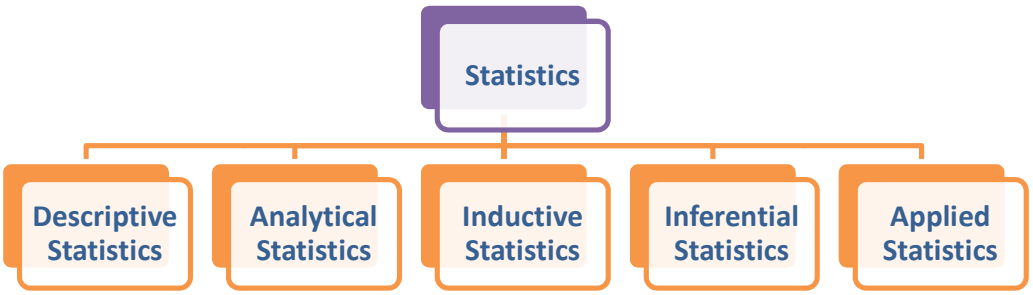


Fig. 1.3 Broad Classification of Statistics

Descriptive and Inferential Statistics

Based on outcome of the statistical procedure, Statistics can be classified into two categories of Descriptive Statistics and Inferential Statistics. Descriptive statistics provides descriptions of the whole population in the form of graphs or statistical table done through numerical calculations. It describes or summarizes the data in a way that it is meaningful and useful. It is the type of statistics that probably comes to every one's mind when the word 'statistics' is heard. All descriptive statistics are either measures of central tendency or measures of variation/dispersion.

Measures of Central tendency

The common measures of central tendency are given in fig.1.4

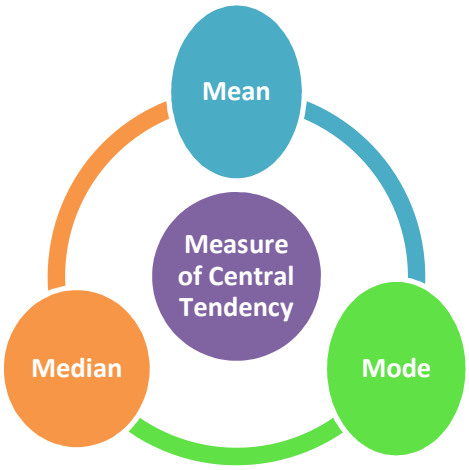


Fig.1.4 Measures of Central Tendency

Illustration No. 1

Assume that the strength of girl students studying across 10 primary schools located in a village is as given below. Let us find out the basic descriptive statistics of the given data.

42, 73, 26, 38, 55, 85, 26, 65, 90.

The mean value of (42+73+26+38+55+85+26+65+90)/10 i.e 50 gives the average strength of girl students studying in all 10 schools. The median value of (found in the middle when arranged in ascending or descending order of the data) 26, 26, 38, 42, **55**, 65, 73, 85, 90 = 55 gives the mid value of strength of girl students studying in all 10 schools.

The mode value (whose frequency of occurrence is more compared with any other data In the set) of 26 (occurring two times here) gives the detail of major strength of girl students studying in many schools.

Measure of Dispersion

The range of 26 to 90 (minimum value to maximum value) gives the detail of variation in the strength of girl students studying in the given 10 schools.

To Do Activity

Find out the various measures of central tendency and dispersion (mean, median, mode and range) of ages of your current class mates

A distribution in which the values of mean, median and mode coincide is known as symmetrical distribution. It is of bell shaped. Normal distribution is one of such symmetrical distributions. Figure 1.5 gives the relationship between these three parameters.

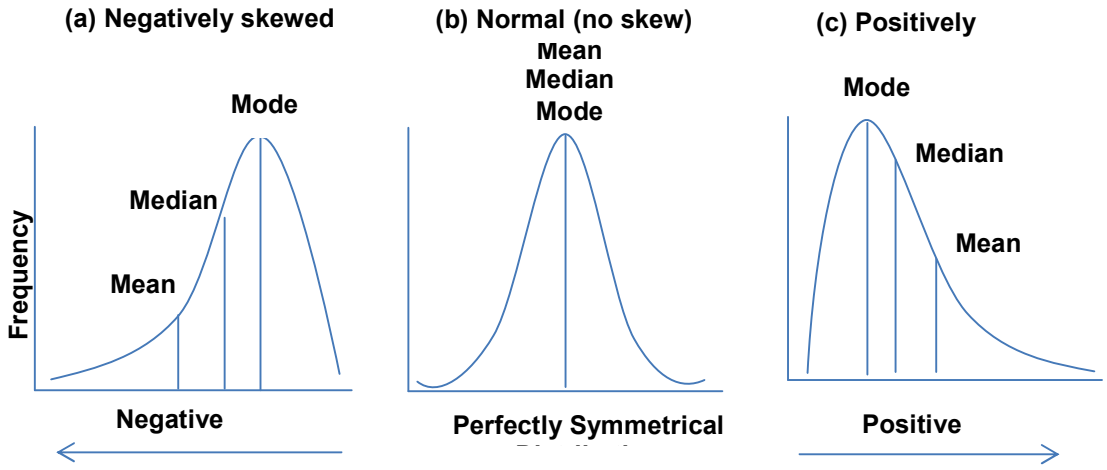


Fig. 1.5 Relationship between mean, median and mode

The relationship between mean, median and mode is given by the relationship

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$
$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

Illustration No.2

In a moderately skewed distribution of data, the value of median is 25 and the mean is 23. Let us find the value of mode.

Mean = 25

Median = 23

Using the relationship, Mean – Mode = 3 (Mean – Median)

we get, 25 – Mode = 3 (25 – 23)

25 – Mode = 6

Mode = 25 – 6 = 19

Inferential statistics provides the inferences and predictions about the whole population based on a sample of data taken from their population. The purpose of the inferential statistics is to draw conclusion from a sample of data. It determines the probability of the characteristics of the sample using probability theory. The common methodologies used in inferential statistics are Hypothesis tests, Analysis of variance, etc.

For example, if it is required to know the performance of all girl students studying in all 10 primary schools, the example given in illustration 1, we cannot do this analysis for all 500 students studying in all schools put together due to time constraints.Hence, we may do this analysis taking 5 students at random from all 10 schools and carryout the analysis on the sample of 50 students and conclude the performance of all 500 students based on the performance measure of these 50 students.

The following table 1.1 gives the differences between descriptive statistics and inferential statistics.

Table 1.1 Differences between Descriptive Statistics and Inferential Statistics

S.No.	Descriptive statistics	Inferential statistics
1	It describes the target population	It makes inferences from the sample and generalizes them to the entire population
2	It organises, analyses and presents the data in a meaningful manner	It compares, tests and predicts the future outcomes
3	Results are shown in the form of graphs, tables and charts	Probability scores are the results
4	It describes the data which is already available	It makes the conclusion about the population which is beyond the data available
5	Tools used : Measures of Central Tendency, Measures of Variation	Tools used : Hypothesis tests, Analysis of Variance, etc

1.3 Importance and Scope of Statistics

Statistics is a tool in the hands of mankind to produce the facts of complex nature in a simple and understandable way. Statistics are more important in our life because we live in the information world today and much of the information is presented using the concept of Statistics. There is no human activity where the application of statistics is not required. Following are a few examples giving the importance of statistics in our day to day life.

i) We watch the weather forecasting in the TV channels. Have you ever thought how does the weather forecaster get the information? There are some computer models which compare the current weather condition with the past weather conditions with the support of software built on statistics and predict the nearby future weather conditions.

ii) News reporter predicts the winner of elections based on statistics of information collected through pre-poll survey.
iii) When statistics is involved in medical field, it would be easy to predict the causes of certain diseases and the impact of certain drugs in healing the diseases.
iv) Statistics are used in checking the quality of products based on statistics obtained from a sample of products where cent percent inspection is not possible or time consuming.
v) Any government is making policies based on statistics of population, national wealth, exports, imports, GDP, etc.
vi) Statistics uses numerical evidence to draw valid conclusions.

Functions of Statistics

Statistics has universal applicability. The functions of statistics are summarized in the following paragraph.

i) Presents facts in definite form: Statistics helps us to present the things in their original form with the help of numbers and figures. For example, if we say simply without any figure that population is increasing at a faster rate; the expression would be vague and indefinite. Instead, if we say that the population in current year is 20% more compared to last year, it would be more definite.
ii) Simplifies complexity: Statistics helps in simplifying complex data into a simple and understandable, by presenting the data in the form of graph or diagram or through some statistical numbers. For example, the marks scored by individual students in a class cannot be remembered ever to judge the quality of the class. But if we know the average marks of the class, we can remember the figure very easily.
iii) Compares the data: The relationship between the two groups can be represented by means of certain mathematical quantity like regression or correlation coefficient. Statistics enables us to compare the past and present results with a view to ascertain the reasons for change taken place and the effect of such changes in the future.
iv) Formulates and tests the hypothesis: Statistical methods help us to formulate and test the hypothesis or new theories. For example, we can test the hypothesis of influence of demonization policy in the sluggish growth of GDP.
v) Enlarges individual experiences and knowledge: When a person goes through various statistical procedures, it widens the knowledge pattern of the person besides widening the thinking and reasoning power. It enables persons to understand and measure the actions of people and make a rationale conclusion.
vi) Forecasts future courses: It helps in forecasting the future tendency of a given phenomenon. Any businessman can exploit the market situation in a successful manner if he has the idea of market trend for which statistical methods are more helpful.
vii) Helps in policy making: On the basis of forecasting, any government makes the policies about various welfare schemes.
viii) Measures uncertainty: Using various statistical techniques like regression, time series, interpolation, etc., we can make correct estimation and reduce the uncertainty of any happenings.

The functions of statistics are summarized in fig. 1.6

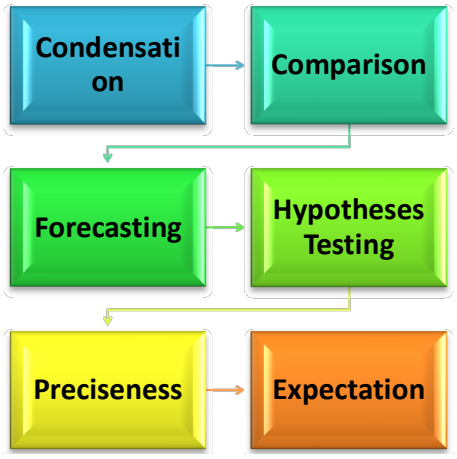


Fig. 1.6 Functions of Statistics

Limitations of Statistics

Though the statistics and its various techniques are widely used in every branch of knowledge, it is not a tool or technique giving solutions to every problem. Following are a few limitations of statistics.

i) Statistics is best applicable to quantitative data only, but qualitative aspects such as poverty, efficiency, intelligence, blindness, etc cannot be studied directly. Only when the qualitative data is converted into quantitative data (numbers), statistical techniques can be applied.

ii) Statistics deals with groups or aggregates only. Hence the scope of statistics lies outside the study of individual facts. For example, per capita income is obtained by dividing the total income of entire population by the total population. This figure does not reveal the income of any individual separately.

iii) Statistics cannot be applied to heterogeneous data. Only uniform and homogeneous data can be compared. Unequal or incomparable data will lead to wrong and misleading results.

iv) Statistics is liable to be misused. It becomes dangerous in the hands of those who do not know its use and deficiencies. It will be of great support only if it is properly used by an expert in the field of Statistics.

v) Errors are possible in statistical decisions. It will not be clear if an error has been committed or not. Hence the results of statistics may mislead to wrong conclusion some times. Hence, enough care is required to be exercised while collection, analysis and interpretation of the data.

The limitations of statistics are summarized in fig. 1.7

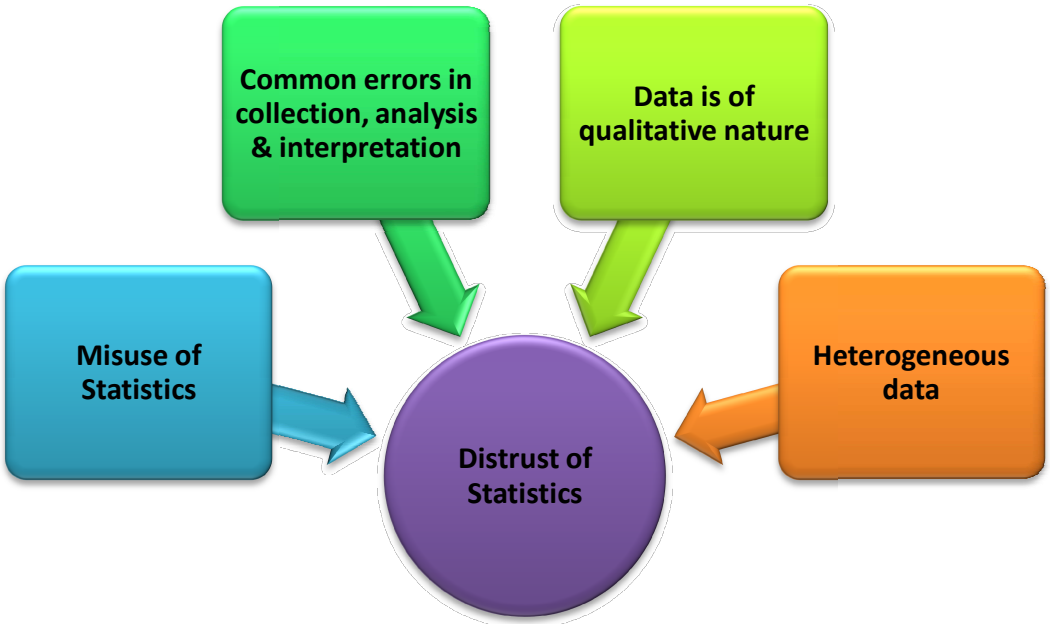


Fig. 1.7 Limitations of Statistics

1.4 Classification of Data

It is the process of arranging the given data/facts into a homogenous classes / groups on the basis of certain common characteristics/properties/similarities/resemblances/attributes. The elements of classification of data is given in fig. 1.8

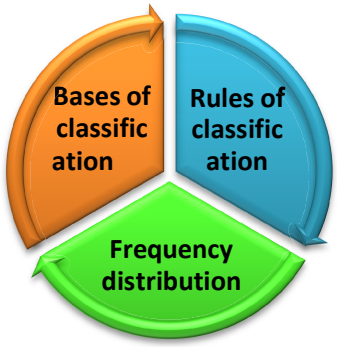


Fig. 1.8 Elements of Classification of data

Objectives of Classification

1. It condenses the mass data into a form suitable for statistical analysis
2. It presents the facts in a simple and understandable form
3. It removes the complexities and highlights the features of the data
4. It brings out the points of similarity and dissimilarity clearly
5. It facilitates comparison and draws inferences from the data
6. It brings out the mutual relationship among elements of a data set
7. It prepares data for tabulation

Rules of Classification

The following rules have to be followed for classification of data.

1. Exactness: The classification should not lead to any ambiguity or confusion
2. Mutually exclusive: The classes should not overlap
3. Flexibility: It should be flexible to adjust to new situations
4. Suitability: It should be suitable for the object of enquiry/interest
5. Stability: Same pattern of classification should be maintained throughout the analysis
6. Homogeneity: The data included in each class should be homogenous
7. Mathematical accuracy: The grand total and subtotal of items included in the classification should match

In general, data are classified on the following four bases.

1. Geographical (Spatial) classification:

In this type of classification, data are classified on the basis of geographical or locational differences between various elements of data such as countries, states, districts, regions, taluks, zones, villages, cities, areas, etc. Table 1.2 is an example of geographical classification.

Table 1.2 Population Distribution - State & Union Territory Wise, in India in 2011

	India	**1,21,01,93,422**
1.	Uttar Pradesh	19,95,81,477
2.	Maharashtra	11,23,72,972
3.	Bihar	10,38,04,637
4.	West Bengal	9,13,47,736
5.	Andhra Pradesh	8,46,65,533
6.	Madhya Pradesh	7,25,97,565
7.	Tamil Nadu	7,21,38,958
8.	Rajasthan	6,86,21,012
9.	Karnataka	6,11,30,704
10.	Gujarat	6,03,83,628
11.	Orissa	4,19,47,358
12.	Kerala	3,33,87,677
13.	Jharkhand	3,29,66,238
14.	Assam	3,11,69,272
15.	Punjab	2,77,04,236
16.	Chhattisgarh	2,55,40,196
17.	Haryana	2,53,53,081
18.	NCT of Delhi*	1,67,53,235
19.	Jammu & Kashmir	1,25,48,926
20.	Uttarakhand	1,01,16,752
21.	Himachal Pradesh	68,56,509
22.	Tripura	36,71,032
23.	Meghalaya	29,64,007
24.	Manipur	27,21,756
25.	Nagaland	19,80,602
26.	Goa	14,57,723
27.	Arunachal Pradesh	13,82,611
28.	Puducherry	12,44,464
29.	Mizoram	10,91,014
30.	Chandigarh	10,54,686
31.	Sikkim	6,07,688
32.	Andaman & Nicobar Island	3,79,944
33.	Dadra & Nagar Haveli*	3,42,853
34.	Daman & Diu	2,42,911
35.	Lakshadweep	64,429

Courtesy: Censusindia.gov.in

To Do Activity
List out any ten districts of our country with their rural population from the latest census records and tabulate the data in a suitable form

2. Chronological Classification

In chronological classification, data are classified on the basis of the time of occurrence, such as years, months, weeks, days, hours, etc. Time series are also called chronological classification, as the data are classified into the period of time. Table 1.3 is an example of chronological classification.

Table 1.3 Population distributions in India – year wise from 1901 to 2011

Census Year	Population
1901	23,83,96,327
1911	25,20,93,390
1921	25,13,21,213
1931	27,89,77,238
1941	31,86,60,580
1951	36,10,88,090
1961	43,92,34,771
1971	54,81,59,652
1981	68,33,29,097
1991	84,64,21,039
2001	1,02,87,37,436
2011	1,21,01,93,422

Courtesy: Censusindia.gov.in

To Do Activity
List out the population growth of any one village of our country for the last 10 years, from the latest census records

3. Qualitative Classification

In qualitative classification, data are classified on the basis of some attributes such as sex, literacy, religion, caste, community, marital status, color, blindness, etc. Table 1.4 gives an example for qualitative classification.

Table 1.4 : Percentage distribution of population by age groups/sex/residence in India in 2011

Residence	Sex	Broad age groups (years)							
		0-4	5-9	10-14	0-14	15-59	60+	15-64	65+
Total	Total	9.7	9.2	10.5	29.5	62.5	8.0	65.2	5.3
	Male	9.9	9.4	10.7	30.0	62.2	7.7	65.0	5.0
	Female	9.5	9.0	10.3	28.8	62.8	8.4	65.5	5.7
Rural	Total	10.3	9.5	11.0	30.9	61.0	8.1	63.7	5.4
	Male	10.5	9.7	11.3	31.5	60.7	7.8	63.4	5.1
	Female	10.1	9.4	10.8	30.3	61.3	8.4	63.9	5.8
Urban	Total	8.2	8.3	9.0	25.5	66.6	7.9	69.4	5.1
	Male	8.3	8.6	9.2	26.1	66.2	7.6	69.1	4.8
	Female	8.0	8.1	8.8	24.9	66.9	8.2	69.7	5.5

Note: Total percentage may not add to 100 on account of rounding in broad age groups

Courtesy: Censusindia.gov.in

To Do Activity
List out the current literacy rate of any ten states our country, from the latest census records

4. **Quantitative Classification**

In this classification, data are classified on the basis of some characteristics which can be measured as income, expenditure, height, weight, price, sales, production, profit, etc. Table 1.5 is an example of Quantitative classification.

Table 1.5 Value of Indian Rupee against US Dollar from 1913-2018

YEAR	1 USD TO INR	YEAR	1 USD TO INR	YEAR	1 USD TO INR	YEAR	1 USD TO INR
1913	0.09	**1982**	9.46	**1964**	4.76	**2001**	47.19
1925	0.1	**1983**	10.1	**1965**	4.76	**2002**	48.61
1947	4.16	**1984**	11.36	**1966**	6.36	**2003**	46.58
1948	3.31	**1985**	12.37	**1967**	7.50	**2004**	45.32
1949	3.67	**1986**	12.61	**1968**	7.50	**2005**	44.1
1950	4.76	**1987**	12.96	**1969**	7.50	**2006**	45.31
1951	4.76	**1988**	13.92	**1970**	7.50	**2007**	41.35
1952	4.76	**1989**	16.23	**1971**	7.49	**2008**	43.51
1953	4.76	**1990**	17.5	**1972**	7.59	**2009**	48.41
1954	4.76	**1991**	22.74	**1973**	7.74	**2010**	45.73
1955	4.76	**1992**	25.92	**1974**	8.10	**2011**	46.67
1956	4.76	**1993**	30.49	**1975**	8.38	**2012**	53.44
1957	4.76	**1994**	31.37	**1976**	8.96	**2013**	56.57
1958	4.76	**1995**	32.43	**1977**	8.74	**2014**	62.33
1959	4.76	**1996**	35.43	**1978**	8.19	**2015**	62.97
1960	4.76	**1997**	36.31	**1979**	8.13	**2016**	66.46
1961	4.76	**1998**	41.26	**1980**	7.86	**2017**	67.79
1962	4.76	**1999**	43.06	**1981**	8.66	**2018**	70.09
1963	4.76	**2000**	44.94	Source: IMF, World Bank			

To Do Activity
Tabulate the height and weight of all your family members in a suitable form

Organization of Data

The collected data are arranged in a proper way. Using array, raw data can be arranged in descending or ascending order of magnitude. But array does not reduce the volume of data. The collected data are first grouped into classes and the number of cases which fall in each class recorded using frequency distribution technique. The data collected can be broadly classified into two categories, i.e., Univariate data and Bivariate data. The differences between these two are illustrated in table 1.6.

Table 1.6 Differences between Univariate Data and Bivariate Data

S.No.	Univariate Data	Bivariate Data
1	It involves a single variable	It involves two variables
2	Purpose of data is to describe the data	Purpose of data is to explain about the data
3	It does not deal with causes or relationships	It deals with causes or relationships between variables
4	It deals with central tendency of data like mean, median and mode	It deals with correlations between two variables
5	It uses Histogram, Bar chart, Pie chart, Line graph, etc.	It uses tables where one variable is contingent on the values of the other variable given in the other table

Single-Variate Frequency Distribution

First let us discuss on single variate frequency distribution of data. Let us take the following example. The marks scored by 30 students in a class in the subject 'Business Statistics' are given below as per the roll number of the students for a maximum score of 20 marks in a midterm test.

Illustration No.3

Let us learn how to organize the given data based on the frequency distribution.

7	10	20	18	10	17	20	15	4	20
10	15	20	15	20	20	15	4	15	17
4	10	17	20	18	4	12	12	9	15

From the above data, we find difficult to understand the significance of marks scored by the 30 students, as it is given in the raw form. We have to form discrete series out of the given data. Let us take the lowest and highest values. In the first column of frequency table, all possible values of the variable (marks) are placed. In the second column, a tally mark 'I' is marked against the variable whenever it occurs. After a particular value occurs fifth time, a cross tally mark ' – ' is marked, cutting the first four tally marks and thus gives a block of five frequency of occurrence. The given set of values (marks) is tabulated as per the above procedure in table 1.7

Table 1.7 Frequency table

Marks	Tally sheet	Frequency umber of students)
4	III	4
7	II	2
9	I	1
10	IIII	3
12	II	2
15	𝍸 I	6
17	III	3
18	II	2
20	𝍸II	7
Total		30

The same data can be expressed using grouped/continuous frequency distribution wherein the class intervals theoretically continue from the beginning of the frequency distribution to the end without break. The class interval is the difference between the lower limit and the upper limit of the class. The formula to find out the class interval of a given problem is given as

$$i = \frac{L - S}{k}$$

where, i = class interval

L = largest value

S = smallest value

k = the preferred number of classes

For the given table, the class interval is calculated below for the preferred number of classes of 5,

$i = \frac{20-4}{5}$ = 3.2 rounded off to next integer value of 4

Thus, the data given in table 1.7 is tabulated using grouped frequency distribution and is given in table 1.8

Table 1.8 Grouped Frequency table

Marks	No. of students
04 - 07	6
08 - 11	4
12 -15	8
16 -19	5
20 - 23	7

The above data is shown diagrammatically in fig. 1.9.

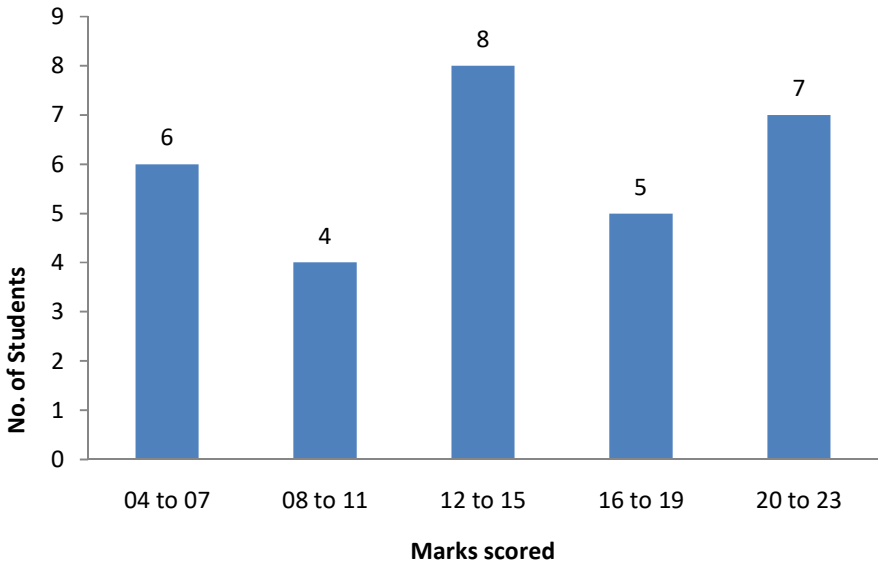


Fig. 1.9 Histogram for single variable data

To Do Activity
Collect the data of 10th standard marks of your relatives and draw a histogram for the class interval of 15

Bivariate Frequency Distribution

The frequency distribution discussed in previous paragraphs involved only one variable and therefore called univariate frequency distribution. In case the data involve two variables such as income and expenditure, ages of husbands and wives, weights and heights of children, scores obtained by two players, supply and demand, etc., then the distribution is called bivariate frequency distribution.

Illustration No. 4

The following data give the weights (in kg) of vegetables produced by two farmers in their respective cultivation lands during a period of 20 days. Let us form a bivariate frequency table. Let us take class intervals of 15-19, 20-24, 25-29, ….for farmer-1 and 20-24, 25-30,…for farmer-2.

Farmer 1 :	17	20	15	18	22	27	19	24	32	35
Farmer 2 :	32	45	47	26	25	23	35	34	30	27
Farmer 1 :	22	25	24	28	27	20	18	17	15	15
Farmer 2 :	25	27	29	30	32	34	30	28	24	20

Steps for Construction of Table

1. Find out the class interval of each of the variables (Farmer 1 & Farmer 2).
2. Write one of the variables (Farmer-1) on the left hand side of the table and the other (Farmer-2)at the top

3. A tally mark has to be put against each of the corresponding row and column. For example, the first value of 17 for Farmer-1 and 32 of Farmer-2 is marked with a tally mark in row 3 and column 1.
4. Repeat the procedure for all values of 20 days
5. Find out the total of the tallies at the bottom and to the right side
6. Totals at the right at the extreme column are for Farmer-1 and those at the bottom row are for Farmer-2.

The completed frequency table for the given problem is given in table 1.9

Table 1.9 Bivariate Frequency Table

Farmer-1 → Farmer -2 ↓	15-19	20-24	25-29	30-34	35-40	Total
20-24	II (2)		I (1)			3
25-29	II (2)	III (3)	I (1)		I (1)	7
30-34	II (2)	II (2)	II (2)	I (1)		7
35-39	I (1)					1
40-44						0
45-49	I (1)	I (1)				2
Total	8	6	4	1	1	20

The above data is shown in radar chart as in fig. 1.10

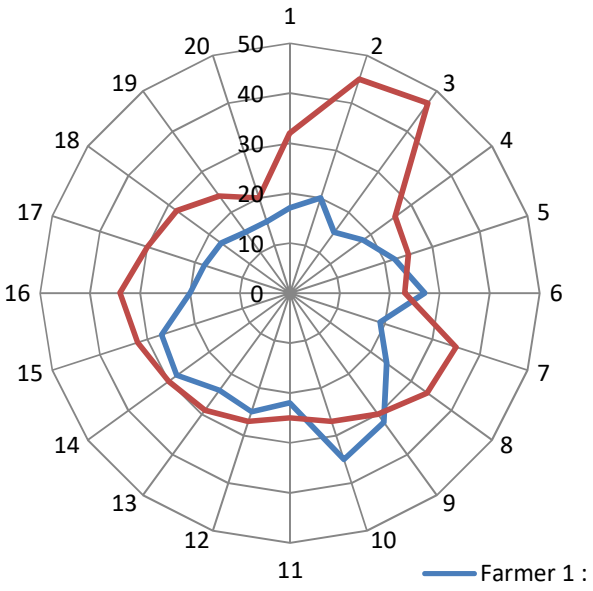


Fig. 1.10 Radar Chart for Bivariate data

To Do Activity
Collect the data of 10th& 12th standard marks of any one of your close friends and compare the marks with your marks in all the subjects using any bivariate chart

Tabulation of Data

It is a process of summarizing and presenting the given data in a systematic form in rows and columns. It facilitates comparison of data and analysis. The following are the parts of tabulation.

i) Table number: A table should always be numbered for easy identification and reference in the future. The table number may be mentioned in the center or side of the table but above the top of the table.
ii) Column number: If the number of columns in a given table is large, then the columns also are numbered so that easy reference to these columns is possible.
iii) Title of the table: Each table should be given a suitable title on the top of the table. It should a) indicate the nature of the data b) explain the locality of data covered c) indicate the time or period of data d) contain the source of data, as a means of verification and as a reference. The source is always mentioned below the table.
iv) Prefatory or Head note: It is a statement, mentioned below the title and enclosed in brackets. For example, the unit of measurement is written as a head note, such as 'in kg', 'in Rs.' etc.
v) Caption: It is a heading for vertical columns. It must be brief and self-explanatory. It must be clear and concise, written in small letters.
vi) Stub: It is a heading for the horizontal rows. Stubs are wider than columns.
vii) Body: It contains the numerical information, the most important part of the table. The arrangement of data is generally from left to right in rows and from top to bottom in columns.
viii) Foot note: Anything written below the table is a footnote. If any explanation or elaboration regarding any item is required, footnote is written.
ix) Source note: It refers to the source from where the data has been taken. It is useful to the reader to check the genuineness of data and to collect additional information. A model table with various table-elements is given in Fig. 1.11

Table number and Title of the table
(Prefatory or Head note if any)

Stub Heading	Caption			Total (for Rows)
	Column head	Column head	Column head	
Stub entries	Body			
Total (for columns)				

Foot note :
Source note :

Fig. 1.11 Elements of a model table

Classification of Table

The classification of table depends on various aspects which are given in the fig. 1.12

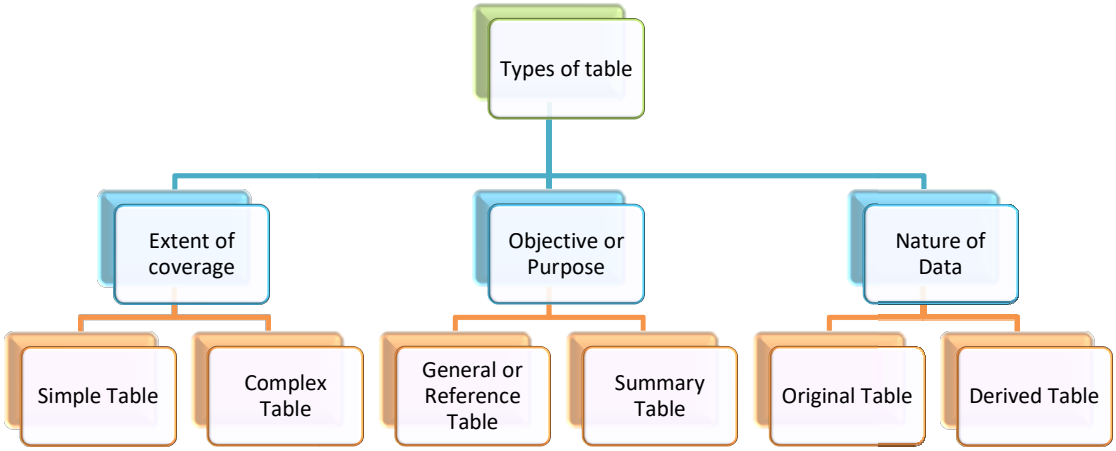


Fig. 1.12 Classification of Table

i) Simple table: Data are presented based on one characteristic.
ii) Complex table : Data are presented based on two or more characteristics
iii) General or Reference table : Data are presented based on general or reference purpose
iv) Summary table : Summary of Data is presented in the table
v) Original table: It is also called classification table, containing the data collected from a primary source.
vi) Derived table: The data presented in the table have been derived from a general table.

1.5 Diagrammatic Representation of Data

The main purpose of Statistics is to simplify complex data and to present the simplified data in a tabulated format. A large amount of data extending over a large number of columns will not create interest to the reader and it is difficult for the reader to understand the significance of the data at a glance. Complicated data presented through a diagram and graph can be easily understood and attractive to the reader. Diagrammatic representation refers to Bars, Circles, Maps, Pictorials, Cartograms, etc.

Advantages of Diagrammatic Representation

i) It is attractive and impressive. It will create interest in the mind of readers.
ii) It saves time and labor. Reader can easily and quickly understand the meaning of the data and draw meaningful inferences from it.
iii) It has universal applicability. Diagrammatic presentation of data is followed universally.
iv) It makes data simple and comparison easy. Diagrams can be remembered easily as it renders comparison in an easy and possible way.
v) It provides more information. Diagram plays a vital role in the modern advertising campaigns and hence the newspapers, articles, journals are filled with more diagrams.

Drawbacks of Diagrammatic Representation

The presentation of a diagram, without care will be misleading. The following are a few drawbacks of it.

i) Drawing a table will be easier than construction of a diagram.
ii) Diagram can be a supplement to a table, but not an alternative to it.
iii) All details cannot be presented diagrammatically.
iv) Diagram can show only approximate values
v) Diagram cannot be analyzed further

Rules for Making Diagram

Construction of a diagram is an art, which can be acquired through practice. The following are a few guide lines to be followed in drawing.

i) Heading : Every diagram should have a title which must be brief, self-explanatory and clear
ii) Size: The size of the diagram should match with the paper size and should be presented in the middle of the paper.
iii) Length and breadth: To make the diagram more attractive, an appropriate proportion between length and breadth of the diagram should be maintained.
iv) Scale: A proper scale should be used to create a visual impact on the reader. At the same time, accuracy should be ensured rather than attractiveness.
v) Neatness: Since impression is needed, drawing should be made with the help of drawing instruments.
vi) Proper selection of diagram: A wrong selection of diagram will lead to wrong and misleading interpretations. An inappropriate diagram may distort the facts and mislead the reader.
vii) Index: When many items are shown in the diagram through different colours, dottings, crossing, etc. an index should be given for identifying and understanding various items in the diagram.
viii) Source: If data has been acquired from some external source, then it has to be indicated at the bottom of the diagram.

Types of Diagram

The following figure 1.13 gives the broad classification of diagrams.

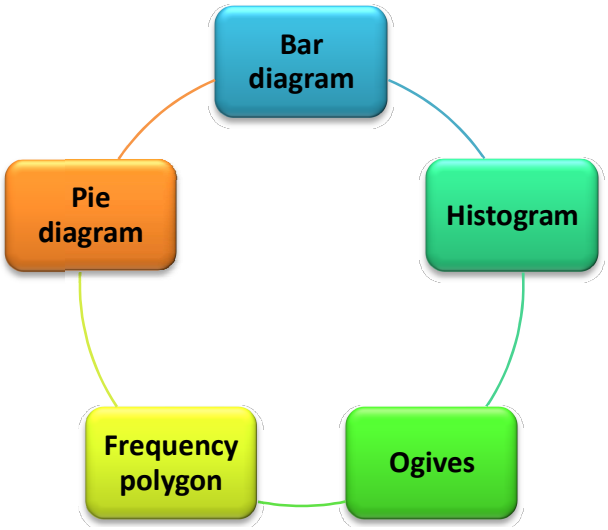


Fig 1.13 Types of Diagram

A few common types of diagrams are briefly explained below.

1) One-dimensional diagram (Line and Bar diagram) :

Line diagram: It is the simplest of all diagrams. On the basis of size of the values of data, heights of lines are drawn. The distance between the lines is kept uniform.

Example of Line diagram is given in fig. 1.14

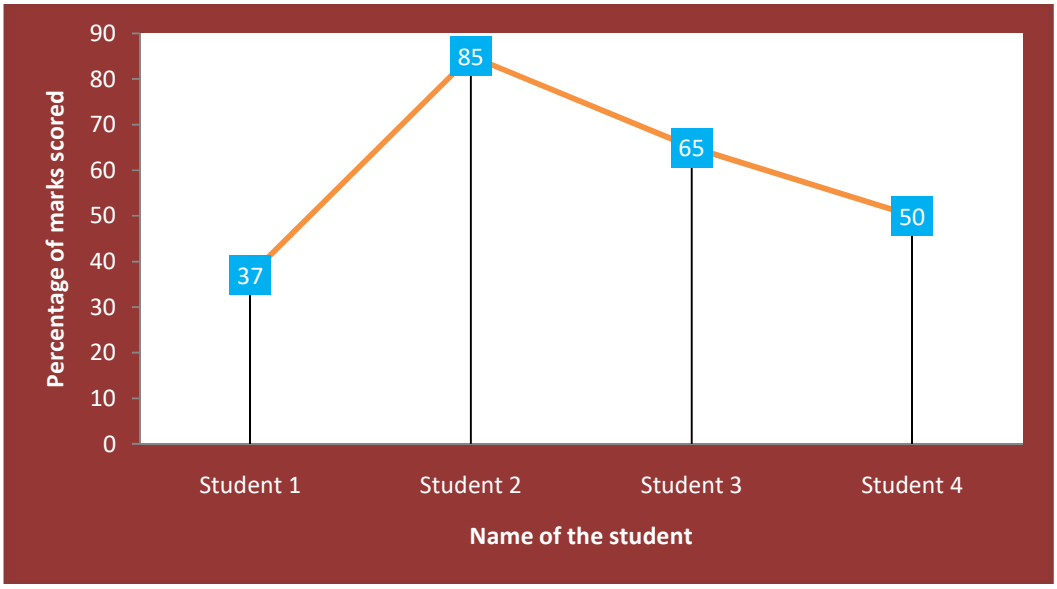


Fig 1.14 Model Line Diagram

Bar diagram: Lines in the line diagram are replaced by bars. Bars can be drawn either on horizontal or vertical base. A model bar diagram is shown in fig. 1.15

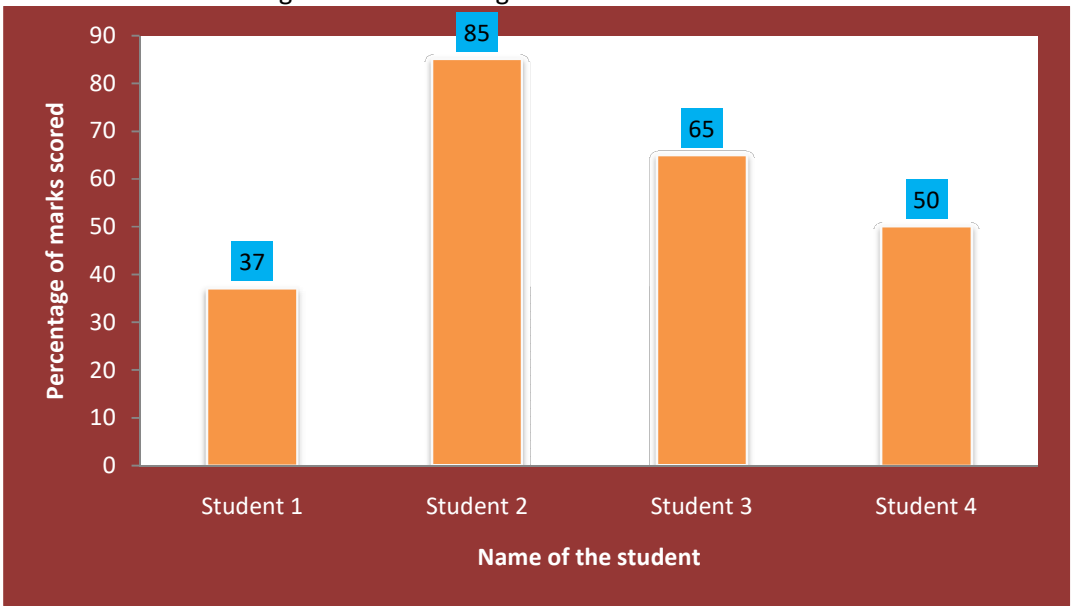


Fig 1.15 Model Bar Diagram

2) Two-dimensional diagram (Rectangle, Square, Circle, etc.)

It is otherwise called as Area or Surface diagram. It is used when two or more magnitudes with different components have to be compared. A model Circle (Pie) diagram showing the marks scored by 4 students in a class is given in fig. 1.16

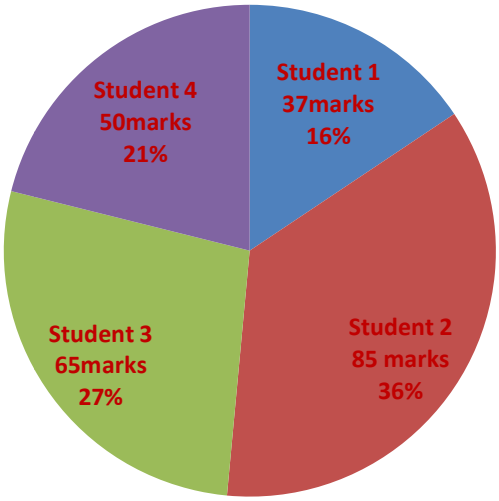


Fig 1.16 Model Pie Chart

3) Three dimensional diagram (Cube, Sphere, Cylinder, etc.)

When the quantities to be represented are diverse in nature, three dimensional diagrams are used. A model three-dimensional diagram is given in fig. 1.17

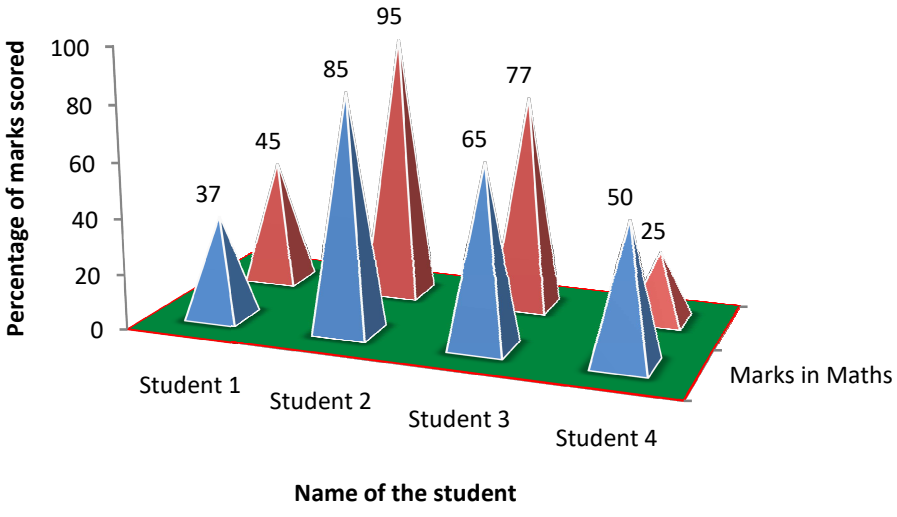


Fig 1.17 Model 3-D Diagram

4) Pictogram and Cartogram

In this type, statistical data is presented in the form of pictures (pictogram) or through maps (cartogram). For the purpose of propaganda, the pictorial presentations of facts are quite popular especially in exhibitions and in primary schools. A sample pictogram depicting the population of 5 cities is given fig. 1.18.

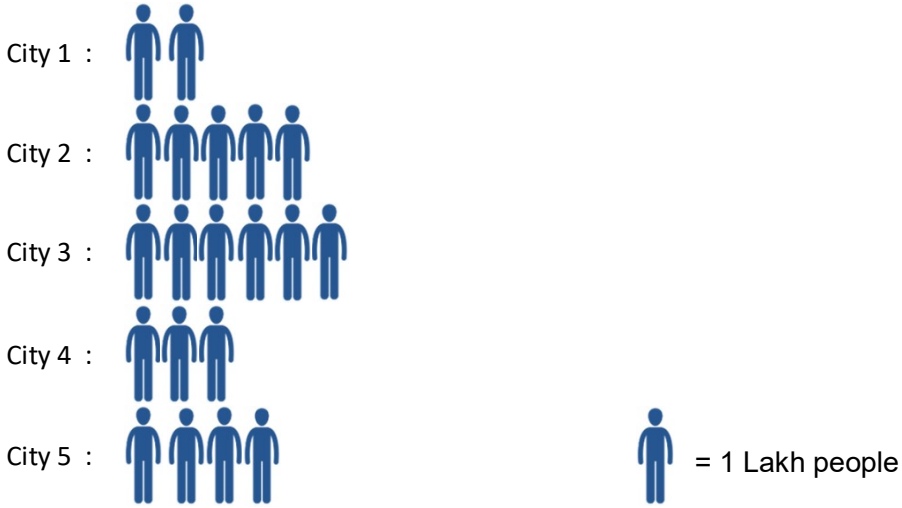
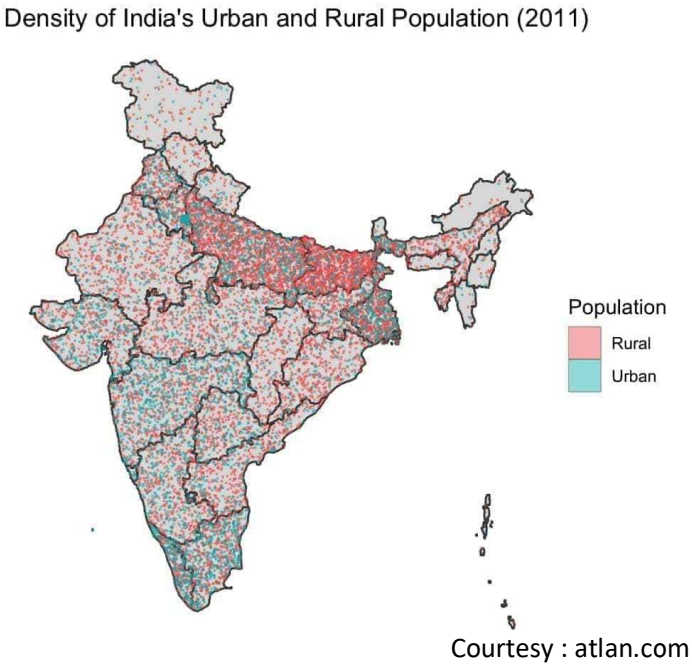


Fig. 1.18 Model Pictogram

A model Cartogram is given in fig. 1.19 in which the population density of rural and urban areas in India is shown with different colour dots in the following map.



Courtesy : atlan.com

Fig. 1.19 Model Cartogram

Graphic Presentation of Data

Graphic presentation of statistical data gives a pictorial effect. It is a visual form of presentation of data. Graphs are generally drawn on a graph paper containing thin horizontal and vertical lines. A graph is divided into 4 quadrants but normally the first quadrant of the graph is used.

The graphic presentation of the same data given for three-dimensional diagram is given fig. 1.20.

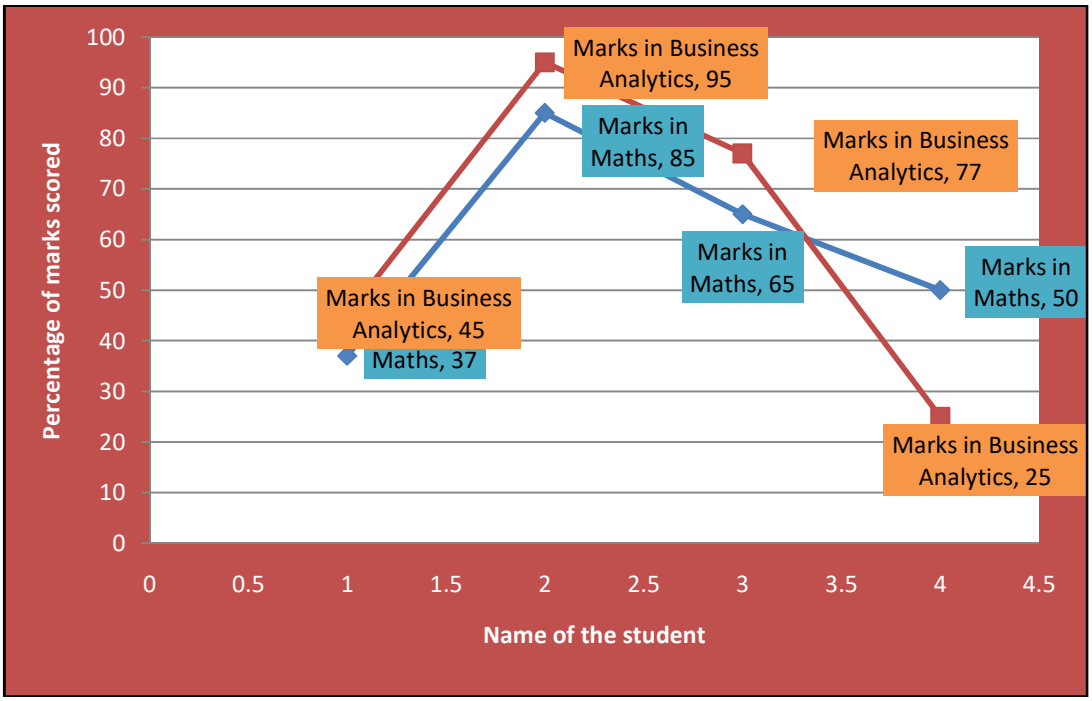


Fig. 1.20 Model graphic presentation of data

The following table 1.10 gives a few differences between a diagram and a graph.

Table 1.10 Differences between a diagram and a graph

S.No.	Diagram	Graph
1	It can be drawn on an ordinary paper	It can be drawn on a graph sheet
2	It is not used in interpolation and extrapolation techniques	It is used in interpolation and extrapolation techniques
3	Median and mode cannot be estimated	Median and mode can be estimated
4	It is used for comparisons only	It gives mathematical relationship between two variables
5	Data are represented by bars, rectangles, pictures, etc.	Data are represented by points of lines of different kinds

Summary

In this chapter, we have learnt the meaning of statistics and its development over the period of time.

The scope of statistics in various fields also is discussed in detail in this chapter besides its functions and limitations. The differences between descriptive and inference statistics have been discussed with suitable examples. A few problems have been presented to describe the descriptive statistics. The latter part of the chapter discussed about the tabulation of data and the various types of tables. The rules for classifying the data are narrated in detail. The procedure to prepare frequency table for single and bi-variables has been explained. The last part of the chapter has thrown light on representation of data in graphical and diagrammatic formats.

Model Questions

Problems

1. In a primary school of a village, there are 10 classes offering various standards. The following is the strength of girl students studying in different classes. Find the average strength of the girl students, the maximum strength of girls studying in any class andmore the number of similar strength of girl students studying in different classes.

   25 32 24 25 41 36 24 31 45 27

   Ans : Average = 31, Maximum strength = 45, Mode = 24 and 25

2. The average strength of students in a school is found to be 42 out of 13 sections. The median value is found to be 50. Calculate the maximum strength of students studying in many classes.

   Ans : 66

3. Vegetables cultivated and harvested during every month of 30 days period by a farmer in his farm are given below in kg.

   a) Arrange the harvests in ascending order
   b) Arrange the harvests in descending order
   c) Convert the harvests into a continuous series of a class-interval of 10 kg

55	47	48	32	55	65	38	80	53	20
48	34	51	25	35	41	48	53	42	33
44	60	71	41	47	31	55	42	35	25

4. Under the scheme of 'Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA)', the following information was obtained. There were both male and female workers who got benefitted from the scheme. Both skilled and unskilled workers were involved in the work. Present the data in a suitable tabular form.

   Village A : Female workers were 40% ; skilled workers were 30% and male unskilled workers were 20%

   Village B : Female workers were 45% ; skilled workers were 25% and male unskilled workers were 35%

5. Present the following data of production of rice in bags of 50 kg each, from a 50 pieces of cultivation land of 1 acre size in a village. Take class size as 0-9.

   35 28 32 37 24 23 19 32 34 28 23 24 32
   19 29 36 34 28 36 35 27 30 28 29 36
   35 37 30 34 28 26 17 36 12 27 35 38 32

35 30 29 19 36 16 35 35 27 30 25 22

6. The marks obtained by 25 students in Economics and Business Analytics are given below in percentage. The first value in the bracket indicates the marks obtained in Economics whereas the second value in the same bracket indicates the marks obtained in Business Analytics by the same student. Prepare a two-way frequency table taking the width of each class interval as 5 marks, the first being less than 5.

   (30, 45) (20, 32) (40, 45) (45, 65) (10, 35) (32, 30) (60, 85) (10, 25) (90, 94) (60, 75) (70, 65) (5, 15) (40, 25) (4, 10) (45, 45) (10, 55) (20, 35) (10, 35) (30, 50) (34, 67) (23, 72) (13, 54) (3, 8) (70, 95) (5, 9)

7. Represent the following data by means of percentage bar diagram. Values are given in Rupees in Hundreds for one month period.

Items of expenses	Family – A (in '00 Rs.)	Family – B (in '00 Rs.)
Food	100	70
House Rent	250	120
Children Education	50	20
Clothing	27	12
Travel expenses	30	15
Miscellaneous	50	22

8. Represent the following data by a pie diagram.

Items of expenses	Monthly Expenditure (in '00 Rs.)
Food	90
House Rent	120
Children Education	60
Clothing	25
Travel expenses	45
Miscellaneous	50
Total	390

9. The following figures relate to the cost of house construction in rural place in Tamil Nadu. Represent the data by a suitable diagram.

Items of expenses	Expenditure (in percentage)
Cement	18
Steel	15
Bricks	14
Sand	8
Timber	13
Labour	22
Miscellaneous	10

10. The following table gives the egg production during a year in a poultry farm. Represent the data graphically.

No. of eggs	0-20	21-30	31-40	41-50	51-60
No. of hens	2	4	5	12	25

No. of eggs	61-70	71-80	81-90	91-100	Above 100
No. of hens	48	39	16	12	4

Theoretical Questions

1. What is statistics?
2. 'Statistics are numerical statements of facts in any department of enquiry, placed in relation to each other'. Explain
3. No isolated facts constitute statistics – Discuss
4. Statistics is said to be both a science and an art – Why?
5. Explain briefly why the current age of information depends on statistics completely without which the modern technology of Artificial Intelligence, Machine Learning, Cloud Computing will not exist.
6. Briefly explain the development of Statistics over period of time.
7. Classify Statistics with suitable examples
8. Write down the differences between descriptive and inferential statistics
9. Discuss the importance and scope of statistics in present age of information
10. What are the characteristics that statistics should possess?
11. Write down the functions of Statistics
12. What are the difficulties being faced while using statistics in various fields?
13. Give briefly the objectives of classification
14. Write down the rules to be followed in classifying the data
15. Explain briefly the classification of data
16. Write down the differences between univariate and bivariate data with suitable examples
17. With a model table, explain the various elements of a table to be constructed for entering the data collected
18. What are the requisites of a good table?
19. State the rules that serve as a guide in tabulating statistical data.
20. What are the types of table generally followed?
21. Compare Diagrammatic representation with Graphic representation of data with examples
22. Explain briefly the types of diagrams used for data representation.
23. What are the rules to be followed in making a diagram?
24. Write down the differences between diagram and graph
25. What is meant by Cartogram? Give an example.

Multiple Choice Questions:

1. Both descriptive and inferential statistics are used for the purpose to change the data into information which in turn is converted into _____ that leads to better decision making.
   a) Process

b) Knowledge
c) Forecast
d) Parameter

2. Inferential statistics involves all the following except
   a) estimating a parameter
   b) testing a hypothesis
   c) estimating a statistic.
   d) analyzing relationships
3. A frequency distribution formed involving two variables at a time is called
   a) Univariate frequency distribution
   b) Trivariate frequency distribution
   c) Bivariate frequency distribution
   d) Bimodal distribution
4. In a statistical table, row captions are termed as
   a) Box head
   b) Body
   c) Stub
   d) Title
5. The graph of the cumulative frequency distribution is called
   a) Ogive
   b) Histogram
   c) Frequency Polygon
   d) Pictogram
6. Histogram and frequency polygon are two graphical representations of
   a) Class boundaries
   b) Frequency Distribution
   c) Class Intervals
   d) Class Marks
7. Inferential statistics are useful for
   a) Interviews
   b) Observing natural behavior
   c) Construct validity
   d) Determining the probability of something
8. There are basically two types of statistics namely descriptive and inferential. Which of the following sentences are true about descriptive statistics?
   a) Descriptive statistics enable you to make decisions about your data
   b) Descriptive statistics enable you to draw inferences about your data
   c) Descriptive statistics describe the data.
   d) Descriptive statistics from a population is used to estimate the sample characteristics

9. A random sample of 250 cellphone users was asked which network they subscribe to. What type of data has been collected and which graphical technique would be the most appropriate to highlight the various market shares, amongst those listed below?
   a) Quantitative data to be represented in a pie chart
   b) Qualitative data to be represented in a pie chart
   c) Qualitative data to be represented in a histogram
   d) Quantitative data to be represented in a bar chart
10. A certain company employs a large number of senior managers earning very high salaries and a few others who earn comparatively small salaries. What is a histogram of salaries for this company likely to look like?
   a) positively skewed
   b) negatively skewed
   c) symmetrical
   d) bimodal
11. Which of the following would be most helpful in the construction of a pie chart?
   a) ogive
   b) frequency distribution
   c) cumulative percentages
   d) relative frequencies
12. What does the highest bar in a histogram represent?
   a) The class with the highest frequency
   b) The class with the lowest frequency
   c) The class with the highest cumulative frequency
   d) The class with the lowest relative frequency
13. The width of a class interval in a bar chart will be approximately equal to the range of the data divided by the ____
   a) average of the data set
   b) lowest value in the data set
   c) highest value in the data set
   d) number of class intervals
14. Inferential statistics are useful for
   a) Determining the probability of something
   b) Observing natural behavior
   c) Interviews
   d) construct validity
15. Which of the following is considered to be inferential statistics?
   a) Computing a mean
   b) Setting up a frequency table
   c) Testing the significance of a correlation coefficient
   d) Calculating gamma

16. The arrangement of data in rows and columns is called
    a) Classification
    b) Tabulation
    c) Frequency distribution
    d) Cumulative frequency distribution
17. Histogram can be drawn only for
    a) Discrete frequency distribution
    b) Relative frequency distribution
    c) Cumulative frequency distribution
    d) Continuous frequency distribution
18. Cumulative frequency polygon can be used to compute
    a) Mean
    b) Median
    c) Mode
    d) Geometric mean
19. Number of employees according to human resource manager is an example of
    a) Flowchart variable
    b) Discrete variable
    c) Continuous variable
    d) Measuring variable
20. In classification, the data are arranged according to
    a) Percentages
    b) Similarities
    c) Differences
    d) Ratios

Answers for MCQ

1	2	3	4	5	6	7	8	9	10
b	c	c	c	a	b	d	c	b	b
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
d	a	d	a	c	b	d	b	b	b

References

1. Sharma, J.K. (2014). *Business Statistics – Problems and Solutions*. New Delhi : Vikas Publishing House Pvt Ltd.
2. Pillai, R.S.N. &Bagavathi, V. (1999). *Statistics*. New Delhi :S.Chand& Company Ltd.
3. Gupta, S.P. (2010). *Statistical Methods.* New Delhi :S.Chand& Company Ltd.
4. Beri, G.C. (2011). *Business Statistics.* New Delhi : Tata McGraw Hill Educations Pvt Ltd.
5. Foster, D. & Stine, E.R. (2010). *Statistics for Business : Decision Making and Analysis.* New Delhi : Pearson Publishers.
6. Gupta, S.C. & Kapoor, V.K. (2006). Fundamentals of Mathematical *Statistics.* New Delhi :S.Chand& Company Ltd.
7. Srivastava, S.C & Srivastava, S. (2003). Fundamentals of Statistics. New Delhi : Anmol Publications Pvt. Ltd.

Chapter 2 Measure of Central Tendency, Dispersion, Skewness and Kurtosis

Introduction

This chapter is to teach you the concepts and characteristics of measures of central tendency and dispersion. The concept of central tendency plays an important role in the study of statistics. In many frequency distributions of data, you can notice that the tabulated values have a tendency towards a group of data around a typical central value. Similarly, you can observe that the tabulated values are dispersed around a central value. Sometimes a group of data will have a departure from symmetrical distribution around the central value. Studying about the central tendency and dispersion of data around a central value is of much importance to take remedial action to bring the deviated values into a common group in order to achieve the collective desired property.

Objectives of the Chapter

- To understand the concept of central tendency
- To describe all the measures of central tendency, i.e. arithmetic mean, weighted mean, median and mode
- To explain merits and demerits of central tendency
- To explain measures of dispersion such as mean deviation and standard deviation
- To explain partition values or positional measures such quartiles, deciles and percentiles
- To discuss co-efficient of variation
- To explain skewness and kurtosis
- To understand measures of moments and their applications

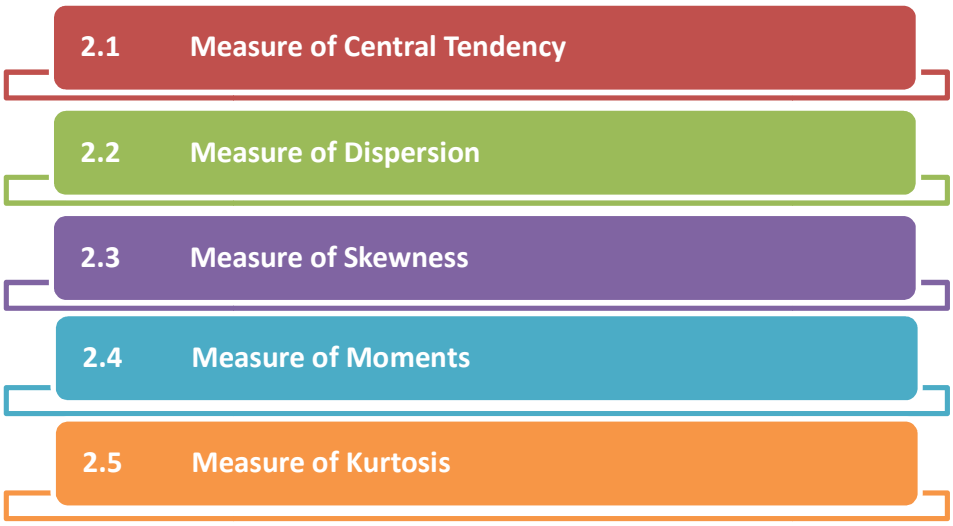


Fig 2.1 Chapter Flow

2.1 Measure of Central Tendency

Measure of central tendency is a typical value of the entire group of data. The numerical value of an observation, also called central value, around which most other numerical values in a set of data show a tendency to cluster or group. This property is called central tendency. Central tendency describes the characteristics of the entire mars of data. This property reduces the complexity of data and makes them to compare.

Characteristics of Central Tendency

A good measure of central tendency should possess the following characteristics.

- Easy to understand
- Simple to calculate
- Not unduly influenced by any single data or a group of data
- Based on all observations
- Rigidly defined so that there is no confusion with regard to its meaning
- Its definition should be in the form of a mathematical expression
- Should be capable of further algebraic treatment
- Should have sampling stability

Functions of Central Tendency

The following are the functions of central tendency.

1. To facilitate quick and easy understanding of complex data. The purpose of mean is to represent a group of individual values in a simple and understandable manner so that we can get a quick understanding of entire group of data. For example, it is very difficult to remember the income of every individual in a village. But we can remember the per capita income of the village and hence quickly understand the economy of the village.
2. To facilitate comparison. A quick comparison between two different groups of values is easy with average values of the groups and conclusions can be drawn easily.
3. To understand the nature of entire group from a sample. Similarly, the average of samples gives a better idea about the average of entire population.
4. To facilitate in decision making. In the process of experimentation and research, averages are more important in setting the standards and estimation.
5. To establish mathematical relationship among different groups of data.

The functions of central tendency is summarized in fig. 2.2

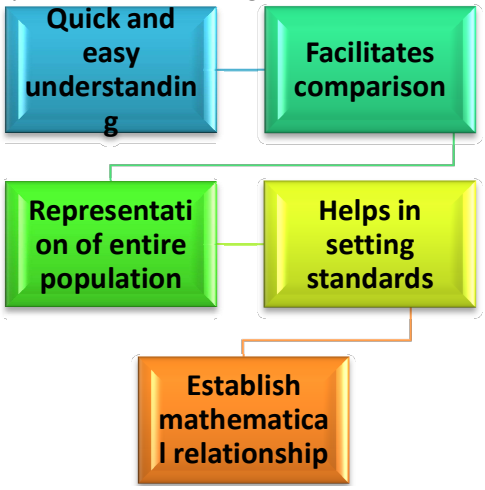


Fig. 2.2 Functions of Central Tendency

The following are the important elements of central tendency.

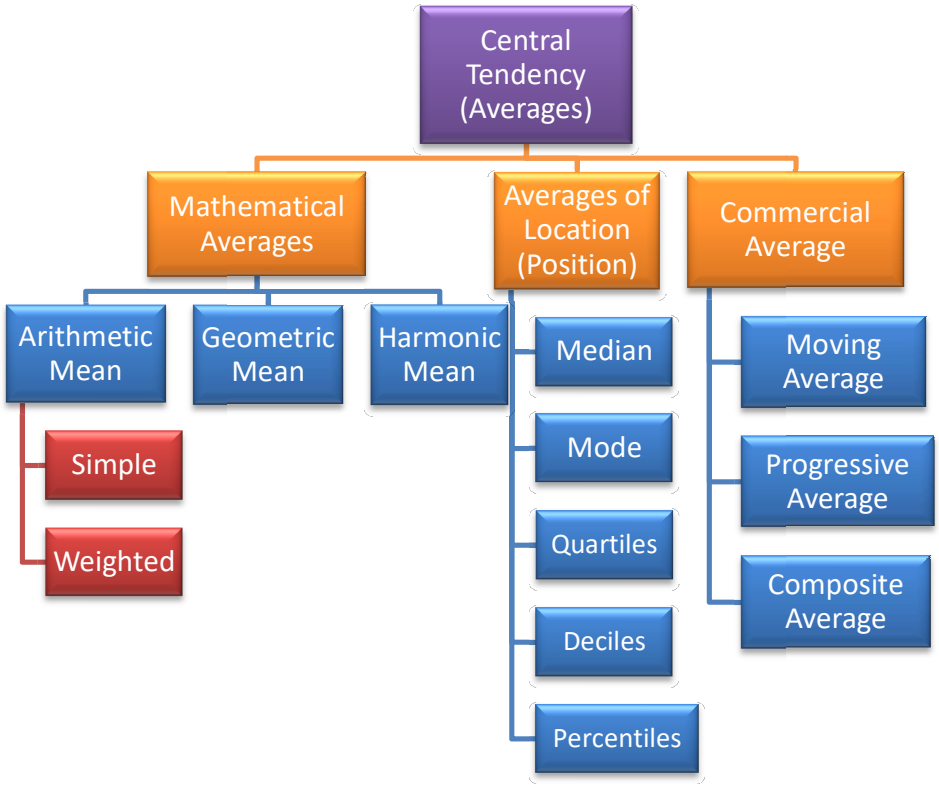


Fig. 2.3 Elements of Central Tendency

Arithmetic Mean

It is also called as mean. It is the quotient obtained by dividing the sum of the various items by their number. It is of two types, namely

i) Simple arithmetic mean

ii) Weighted arithmetic mean

Simple Arithmetic Mean

Let us see how to find out simple arithmetic mean using the following two methods.

a) **Direct method:** The arithmetic mean is simply calculated just by summing up all the given values and dividing the sum by the number of values.

Illustration No.1

Let us find out the simple arithmetic mean of a blacksmith's monthly income in a village for a year.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Income in Rs.	4000	4200	3700	3500	4500	5000	3800	3900	4100	4200	4250	4100

Solution

Month (i)	Income (X_i)
Jan	4000
Feb	4200
Mar	3700
Apr	3500
May	4500
Jun	4950
Jul	3800
Aug	3900
Sep	4100
Oct	4200
Nov	4250
Dec	4100
N = 12	$\Sigma X_i = 49200$

Mean, $\bar{X} = \frac{\Sigma X_i}{N} = \frac{49200}{12}$

= 4100

So, the average monthly income of the blacksmith is Rs.4,100

To Do Activity
Calculate the average monthly salary of your parent/guardian for the last financial year

b) **Short cut method :** This method reduces the amount of calculation. In this method, any one value is considered as a working mean or arbitrary mean or assumed mean (A). Then the difference of each value from the assumed mean is calculated as a deviation (d = X - A). Add all the deviations (Σd). Then apply the following formula to find out the arithmetic mean.

$$Mean, \bar{X} = A \pm \frac{\Sigma X_i}{N}$$

Illustration No.2 :

Let us find out the simple arithmetic mean for the same problem given in illustration no.1
Let us assume, A = 4000

Month (i)	Income (X_i)	d = X_i - A
Jan	4000	0
Feb	4200	200
Mar	3700	-300
Apr	3500	-500
May	4500	500
Jun	4950	950
Jul	3800	-200
Aug	3900	-100
Sep	4100	100
Oct	4200	200
Nov	4250	250
Dec	4100	100
N = 12		Σ d = 1200

$$Mean, \bar{X} = A \pm \frac{\Sigma X_i}{N}$$

$$\bar{X} = 4000 + \frac{1200}{12}$$

= 4100

We get the same mean value of Rs. 4,100 using short cut method also.

Weighted Arithmetic Mean

If the relative importance of the items in the distribution is not the same, then it is required to apply weightage to every item while computing the mean value which is called weighted arithmetic mean.

Weighted Arithmetic Mean ,$\bar{X} = \frac{\Sigma W_i X_i}{\Sigma W_i}$

Where W_i = Weightage of individual item

X_i = Value of individual item

Let us calculate the weighted arithmetic mean for the problem given in illustration no.3.

Illustration No.3

Find out the first semester grade point average (SGPA) of a BBA student who has got marks in his first semester as follows.

Name of the Subject	English-1	Regional Language	Foundations of Management	Ecology and Environment	Business Analytics	Management Decision Tools	Field work
Weightage (Credit)	3	2	3	3	3	3	2
Grade point scored	7	7	8	7	10	9	8

Solution

Subject (i)	Weightage (credit), W_i	Grade point scored, X_i	Weighted Score, $W X_i$
English-1	3	6	18
Regional Language	2	7	14
Rural Society and Polity	3	7	21
Foundations of Management and Entrepreneurship	3	8	24
Ecology and Environment	3	7	21
Business Analytics	3	10	30
Management Decision Tools	3	9	27
Field work	2	8	16
	$\Sigma W_i = 22$		$\Sigma W_i X_i = 171$

$$\bar{X} = \frac{\Sigma W_i X_i}{\Sigma W_i} = \frac{171}{22} = 7.77$$

Hence the SGPA of the student in the first semester is 7.77

To Do Activity
Calculate your first semester GPA

Similarly, the arithmetic mean can be found out for discrete data with different frequency also as described in illustration no.4

Illustration No.4

The daily wage (in Rs.) of employees working on a daily basis in a village are given by

Daily earnings (in Rs.)	200	220	250	270	300	320	350	370	400
No. of employees	5	8	10	12	18	24	25	20	15

Let us calculate the average daily wage of all employees. Let us use short cut method as the values are in three digits. Let us assume assumed mean wage as A = 300

No. of employees, f_i	Daily wage, X_i	$d_i = X_i - A$	Weighted Score, $f_i d_i$
5	200	-100	-500
8	220	-80	-640
10	250	-50	-500
12	270	-30	-360
18	300	0	0
24	320	20	480
25	350	30	750
20	370	40	800
15	400	100	1500
$\Sigma f_i = 137$			$\Sigma f_i d_i = 1530$

$$Average\ wage, \bar{X} = A \pm \frac{\Sigma f_i d_i}{\Sigma f_i}$$

$$\bar{X} = 300 + \frac{1530}{137} = 311.17$$

Hence the average daily wage of all employees is Rs.311.17

To Do Activity
Calculate the average daily wages of a set of 50 people in a village known to you

Now, let us learn to compute arithmetic mean for continuous series of data through the illustration no.5

Illustration No.5

The following data gives the information on the strength of people found in different age groups in a village. Let us calculate the average age of the village.

Age group (in years)	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of persons	135	190	250	230	225	210	150	54	12	4

Solution Let us find out the mid value of age in each group as m_i. Though mid age value is in two digits, let us use short cut method as mid value is in fraction and also the frequency value is in three digits for easy computation. Let us take assumed age, A = 50.5

Age group (X_i)	Mid age (m_i)	No. of persons (f_i)	$d_i = m_i - A$	$f_i d_i$
00-09	04.5	135	-46	-6210
10-19	10.5	190	-40	-7600
20-29	20.5	250	-30	-7500
30-39	30.5	230	-20	-4600
40-49	40.5	225	-10	-2250
50-59	50.5	210	0	0
60-69	60.5	150	10	1500
70-79	70.5	54	20	1080
80-89	80.5	12	30	360
90-99	90.5	4	40	160
		$\Sigma f_i = 1460$		$\Sigma f_i d_i = -25{,}060$

$$Average\ age, \bar{X} = A \pm \frac{\Sigma f_i d_i}{\Sigma f_i}$$

= 50.5 - (25,060 / 1460)

= 33.33

So, the average age of the people in the village is 33.33 years.

To Do Activity
Calculate the average age of your family taking at least 20 members of your relatives

Correcting Incorrect Mean

Sometimes the wrong items are included instead of correct items due to mistake or oversight. To correct it, the wrong items have to be deducted from ΣX and correct items have to be added to ΣX and then the new ΣX has to be divided by N.

Illustration No. 6

The average mark of a student for 10 subjects was found to be 40. Later it was found that mark 55 was misread and wrongly entered as 65. Find the correct average mark.

Solution

Given values are : N = 10, $\bar{X}$ (Incorrect) = 40, incorrect item = 65, correct item = 55

$From\ \bar{X} = \frac{\Sigma X_i}{N}$, we can compute $\Sigma\bar{X}$(Incorrect) as

$\Sigma\bar{X}$(Incorrect) = $\bar{X}$ (Incorrect) x N

= 40 x 10 = 400

Corrected $\Sigma\bar{X}$ = $\Sigma\bar{X}$(Incorrect) – incorrect item + correct item

= 400 – 65 + 55

= 390

Corrected $\bar{X}$ $= \frac{\Sigma X_i}{N}$

= 390/10 = 39 marks

Combined Arithmetic Mean

If we know the means and the number of items in two or more related groups, the combined mean can be computed with following formula

If we have the values of $\bar{X}_1$&$\bar{X}_2$ and N_1& N_2, then combined $\bar{X}$ is given by

$\bar{X}$ (combined) = $(N_1\bar{X}_1 + N_2\bar{X}_2) / (N_1 + N_2)$

Illustration No. 7

The first semester & second semester GPA of a student is 8 and 9 for the total credits of 22 and 20 respectively. Find out the cumulative GPA of the student after second semester.

Solution$\bar{X}_1$ = 8, N_1 = 22, $\bar{X}_2$ = 9, N_1 = 20, What is combined $\bar{X}$?

$\bar{X}$ (combined) $= (N_1\bar{X}_1 + N_2\bar{X}_2) / (N_1 + N_2)$

$= (22 \times 8 + 20 \times 9)/ (22+20)$

= 8.476

So, the CGPA of the student after second semester will be 8.476

To Do Activity
Calculate the cumulative GPA using above method for any of your senior batch student after second semester results.

Finding out Missing Items

Let us see the following illustration to learn to find out the missing items in a set of data if we know the cumulative frequency and total number of items.

Illustration No. 8

Calculate the missing frequencies in the following table with the mean value of 3

Number of accidents	No. of days happened
0	8
1	10
2	?
3	?
4	15
5	20
	93

Solution

Number of accidents (X_i)	No. of days happened (f_i)	$f_i X_i$
0	8	0
1	10	10
2	?	$2f_2$
3	?	$3f_3$
4	15	60
5	20	100
	$\Sigma f_i = 53 + f_2 + f_3$	$170 + 2f_2 + 3f_3$

we know $93 = 53 + f_2 + f_3$

$f_2 + f_3 = 40$ .....(1)

$$\text{and } \bar{X} = \frac{\Sigma f_i X_i}{\Sigma f_i}$$

$3 = (170 + 2f_2 + 3f_3)/93$

$170 + 2f_2 + 3f_3 = 93 \times 3$

$2f_2 + 3f_3 = 109$ .....(2)

Multiplying equation (1) by 2 and subtracting it from equation (2), we get

$2f_2 + 3f_3 = 109$

$2f_2 + 2f_3 = 80$

$f_3 = 29$ .....(3)

Substituting f_3 value in equation (1), we get

$f_2 + 29 = 40$

$f_2 = 11$

Hence the missing frequencies are 11 and 29

Geometric mean

It is defined as the N^{th} root of the product of N items.

It is used

- in construction of index numbers
- in economic and social sciences, in which more weightage is given to smaller items and small weightage is given to large items
- in averaging ratios, percentages and rate of increase between two periods

Geometric mean, GM = $\sqrt[N]{X_1 X_2 X_3 \ldots\ldots X_N}$

For easy calculation purpose, it can be expressed as

GM = Antilog of $\frac{\sum_1^N \log X}{N}$

For example, GM of the series of 2, 4 and 6 is $\sqrt[3]{2 \; x \; 4 \; x \; 6}$ = 3.634

Illustration No.9

Find out the geometric mean of the yield of rice from all the farms of a village whose yields are given below.

Yield of Rice (in 50 kg bags)	No. of farms
18	23
12	9
27	1
9	5
15	19
24	4
21	7

Solution

Yield of Rice (in 50 kg bags), X_i	log X_i	No. of farms, f_i	$f_i \log X_i$
18	1.2553	23	28.8719
12	1.0792	9	9.7128
27	1.4314	1	1.4314
9	0.9542	5	4.7710
15	1.1761	19	22.3459
24	1.3802	4	5.5208
21	1.3222	7	9.2554
		$\Sigma f_i = 68$	$\Sigma f_i \log X_i = 81.9092$

GM = Antilog of $\frac{\sum f_i \log X_i}{\sum f_i}$

= Antilog of 81.9092/68

= Antilog of 1.2045

= 16.016 ≈ 16 bags

Harmonic mean

It is the reciprocal of the arithmetic average of the reciprocal of values of various items in the variable. It is appropriate for situations when the average of rates such flow rate, speed rate, density (mass rate), resistance rate, etc are desired.

Harmonic mean, HM = $\frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}.....\frac{1}{X_N}}$

Illustration No. 10

An investor purchases Rs. 10,000 worth of shares of a company for three months continuously. During 3 months, he purchases the shares at a price of Rs.100, Rs.120 and Rs.180 respectively. After 3 months, what is the average price paid by the investor for the shares?

Solution

In this problem, the worth of shares is redundant to find out the average price of the shares. Harmonic mean is the appropriate to find out the average price of the shares purchased by the investor.

Hence the Harmonic Mean, HM $= \frac{3}{\frac{1}{100}+\frac{1}{120}+\frac{1}{180}}$

$= \frac{3}{0.01+0.00833+0.00556}$

= Rs. 125.58

Median

It is the value of item which will divide the series into equal parts. It is also called as positional average.

Illustration No.11

Find out the median of the following items.

13 19 16 12 10

Solution

The given items are arranged in ascending or descending order first as done below

10 12 13 16 19

Median = Value of $\frac{N+1}{2}^{th}$ item

= Value of $\frac{5+1}{2}^{th}$ item

= Value of 3^{rd} item

= 13

Illustration No.12

Find out the median of the following items.

13 19 16 12 10 25

Solution

The given items are arranged in ascending order first as done below

10 12 13 16 19 25

Median = Value of $\frac{N+1}{2}^{th}$ item

= Value of $\frac{6+1}{2}^{th}$ item

= Value of 3.5^{th} item

= Value of $\frac{3rd\ item + 4th\ item}{2}$

$= \frac{13+16}{2}$

= 14.5

Illustration No.13

Find out the median of the following items.

No. of bikes crossing a traffic signal	Frequency for every signal in a day
12	10
16	13
20	18
22	15
25	13

Solution

No. of bikes crossing a traffic signal, X_i	Frequency for every signal in a day, f_i	Cumulative frequency, cf
12	10	10
16	13	23
20	18	41
22	15	56
25	13	69

In this problem, the total occurrence, N = 69

Hence, Median = Value of $\frac{N+1}{2}^{th}$ item

= Value of $\frac{69+1}{2}^{th}$ item

= Value of 35^{th} item

= 20

i.e. the 35^{th} item (occurrence) is found after 23^{rd} item and before 41th item, whose value is 20.

To Do Activity
Calculate the mean traffic intensity (number of buses crossing a particular signal) in your area in a day. Take the data for a week.

Illustration No.14

Find out the median of the following items.

No. of bikes crossing a traffic signal	**Frequency for every signal in a day**
12	10
16	13
20	18
22	15
25	12

Solution

No. of bikes crossing a traffic signal, X_i	**Frequency for every signal in a day, f_i**	**Cumulative frequency, cf**
12	10	10
16	13	23
20	18	41
22	15	56
25	12	68

In this problem, the total occurrence, N = 68

Hence, Median = Value of $\frac{N+1}{2}^{th}$ item

= Value of $\frac{68+1}{2}^{th}$ item

= Value of 34.5^{th} item

= Value of $\frac{34th\ item + 35th\ item}{2}$

= (20 + 20) /2 = 20

i.e. both the 34^{th} item and 35^{th} item are found after 23^{rd} item and before 41^{th} item, whose value is 20.

Illustration No.15

A survey was carried out to find out the age (in years) of 180 cows in a village. The result of the survey is as follows:

Age of cow (in years)	Less than 5	Less than 10	Less than 15	Less than 20	Less than 25
No. of cows	28	64	119	161	180

What is the median age of the cows?

Solution

The table is redrawn with frequency for respective class as shown below.

Age of cow (in years)	No. of cows (f)	Cumulative frequency (cf)
0-4	28	28
5-9	36	64
10-14	55	119
15-19	42	161
20-24	19	180
	N = 180	

Median is given by $\frac{N}{2}^{th}$ item, which is 180/2 i.e. 90^{th} item.

The 90^{th} item falls in the class interval of 10-14. To find out the exact value between 10 & 14, the following formula is used in such cases.

Median = $L + \frac{\left(\frac{N}{2}\right) - cf}{f}\ C$;

where L is the low value of the identified class,

f is the frequency of the identified class,

C is the class interval;

cf is the cumulative frequency of previous class interval

In this problem, the identified class is 10-14 for the $\frac{N}{2}^{th}$ item (for continuous series). The corresponding values are listed below.

C= 5 ; L = 10; cf = 64; f = 55

Hence, Median = $L + \frac{\left(\frac{N}{2}\right) - cf}{f}\ C$

= $10 + \frac{\left(\frac{180}{2}\right) - 64}{55}\ x\ 5$

= 12.36

Hence, the median age of 180 cows is 12.36 years

To Do Activity
Calculate the mean age of buffaloes in your village, taking a minimum count of 50 buffaloes.

Mode

It is the value of the variable which occurs most frequently in a distribution. It is the item around which there is a maximum concentration.

When all the items have same frequency, then there is no mode in the distribution. When there is only one mode in the series, it is called unimodal. If there are two modes, it is called bimodal. If there are three modes, it is called trimodal. For more than 3 modes, it is called multi-modal.

The relationship between mean, median and mode is given by

Mean – Mode = 3 (Mean – Median)
Mode = 3 Median – 2 Mean

Illustration No. 16

Find out the mode for the following cases.

a) 35, 65, 32, 28, 56, 66 Ans : No mode

b) 35, 65, 32, 28, 56, 66, 32, 28, 55, 72 Ans : Bimodal. Mode-1 : 28 Mode-2 : 32

c)

Size	5	8	10	12	15	18	20
Frequency	10	14	12	10	8	6	2

Ans : Mode : 8 (due to more frequency of 14)

Illustration No. 17

Find out the mode for the problem discussed in illustration no.15

Solution

Age of cow (in years)	No. of cows (f)
0-4	28
5-9	36
10-14	55
15-19	42
20-24	19
	N = 180

From the above table, we understand that the class interval 10-14 has got more frequency compared to other class intervals and hence the mode value is found between 10 and 14. To find out the mode in such cases, the following formula is used.

Mode, $M_o = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} C$

Where L = Lower value of the modal class interval identified

f_1 = frequency of the modal class interval identified

f_0 = frequency of the class interval preceding modal class interval

f_2 = frequency of the class interval succeeding modal class interval

C = value of class interval

In this problem, the modal class is identified as 10-14

So, L = 10, $f_1 = 55$, $f_0 = 36$, $f_2 = 42$, C= 5

Substituting the values, Mode, $M_o = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} C$

$= 10 + \frac{55 - 36}{2\,(55) - 36 - 4} x\ 5$

$= 12.97$

Quartiles

The values of the observations in a distribution, when arranged in an ordered sequence, can be divided into four equal parts or quarters, using three quartiles namely Q_1, Q_2 & Q_3. Q_1 divides the distribution in such a way that 25% of observations have a value less than Q_1. Similarly Q_2 divides for 50% and Q_3 divides for 75%. Q_2 is otherwise called as median.

The formula to find out the required quartile is given by

$$Q_i = L + \frac{i\left(\frac{N}{4}\right) - cf}{f} C;\ i = 1, 2, 3$$

Deciles

The values of the observations in a distribution, when arranged in an ordered sequence, can be divided into 10 equal parts, using nine deciles namely D_1, D_2..... D_9.

The formula to find out the required decile is given by

$$D_i = L + \frac{i\left(\frac{N}{10}\right) - cf}{f} C;\ i = 1, 2, \ldots. 9$$

Percentiles

The values of the observations in a distribution, when arranged in an ordered sequence, can be divided into 100 equal parts, using ninty nine percentiles namely P_1, P_2..... P_{99}.

The formula to find out the required percentile is given by

$$P_i = L + \frac{i\left(\frac{N}{100}\right) - cf}{f} C;\ i = 1, 2, \ldots. 99$$

To Do Activity

Calculate your position in terms of percentile in your class based on first semester total marks scored by all your class mates

Illustration No. 18

Find out the Q_3, D_6& P_{70} for the problem discussed in illustration no.15

Solution

The table is repeated here

Age of cow (in years)	No. of cows (f)	Cumulative frequency (cf)
0-4	28	28
5-9	36	64
10-14	55	119
15-19	42	161
20-24	19	180
	N = 180	

For Q_3 :

Q_3 = Value of $(3N/4)^{th}$ observation
= Value of $\{3(180)/4\}^{th}$ observation
= Value of 135^{th} observation
is found in the class interval of 15-19

$$Q_3 = L + \frac{3\left(\frac{N}{4}\right) - cf}{f} C$$
$$= 15 + \frac{3\left(\frac{180}{4}\right) - 119}{42} x\ 5$$
= 16.905

For D_5 :

D_6 = Value of $(6N/10)^{th}$ observation
= Value of $(6x180/10)^{th}$ observation
= Value of 108^{th} observation
is found in the class interval of 10-14

$$D_6 = L + \frac{6\left(\frac{N}{10}\right) - cf}{f} C$$
$$= 10 + \frac{6\left(\frac{180}{10}\right) - 64}{55} x\ 5$$
= 14

For P_{70} :

P_{70} = Value of $(70N/100)^{th}$ observation
= Value of $(70x180/100)^{th}$ observation
= Value of 126^{th} observation
is found in the class interval of 15-19

$$P_{70} = L + \frac{70\left(\frac{N}{100}\right) - cf}{f} C$$
$$= 15 + \frac{70\left(\frac{180}{100}\right) - 119}{42} x\ 5$$
= 15.833

2.2 Measure of Dispersion

Central tendency gives the general level of magnitude of the distribution, but it fails to show anything further about the distribution. The following three kinds of statistical measures, describe the constitution or shape or scatter of values of the items of a statistical distribution.

i) Measures of variation or dispersion – explain how the items are dispersed away from average
ii) Measures of skewness – explain whether a distribution is symmetrical or asymmetrical
iii) Measures of Kurtosis – describe the peakedness of the distribution or relative influence of central deviation

Objectives of Measure of Dispersion

- To test the reliability of an average
- To test as a basis for control of variability
- To compare two or more series with regard to their variability
- To facilitate as a basis for further statistical analysis

Properties of a Good Measure of Dispersion

An ideal measure of dispersion should

- Be simple to understand and easy to calculate
- Be rigidly defined
- Be based on each individual item of the distribution
- Be capable of further mathematical treatment
- Have sampling stability
- Not be affected by extreme observations

Methods of Measuring Dispersion

The following are the methods of measuring dispersion

a) Range
b) Quartile Deviation
c) Mean Deviation
d) Standard Deviation

Range

It is the difference between the largest and smallest value of data in the distribution. The value of range will become unfit for comparison if the distribution of data is in different units. Coefficient of range serves the purpose.

Range = Largest value – Smallest Value……(Absolute measure)

$R = L - S$

Co-efficient of Range = $\frac{L-S}{L+S}$……(Relative measure)

Illustration No. 19

Find out the range and co-efficient of range for the following data.

Age	5-10	11-15	16-19	20-25	26-45	46-65
No. of persons	23	45	54	72	65	34

Solution

In the given problem, the variable is 'Age' for which we need to find out the range and co-efficient of range. Frequency of data does not affect the value of range.

Range = Largest value – Smallest value

= 65 -5 = 60

Co-efficient of Range $= \frac{65-5}{65+5}$

= 0.857

Quartile Deviation

The inter-quartile range is a measure of dispersion or spread of values in the distribution of data between the third quartile, Q_3 and first quartile, Q_1

Inter-quartile Range = $Q_3 - Q_1$……(Absolute measure)

Half the distance between Q_3 and Q_1 is called semi-quartile range or Quartile deviation

Quartile Deviation, QD = $\frac{Q_3 - Q_1}{2}$……(Absolute measure)

Co-efficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$……(Relative measure)

QD is an improved measure over Range, as it is not calculated from extreme items, but on quartiles.

Illustration No. 20

Find out the Quartile Deviation and Co-efficient of Quartile Deviation for the problem discussed in illustration no. 15

Solution

The cf table is reproduced here

Age of cow (in years)	No. of cows (f)	Cumulative frequency (cf)
0-4	28	28
5-9	36	64
10-14	55	119
15-19	42	161
20-24	19	180
	N = 180	

For Q_3 :

Q_3 = Value of $(3N/4)^{th}$ observation

= Value of $\{3(180)/4\}^{th}$ observation

= Value of 135^{th} observation

is found in the class interval of 15-19

$$Q_3 = L + \frac{3\left(\frac{N}{4}\right) - cf}{f} C$$

$$= 15 + \frac{3\left(\frac{180}{4}\right) - 119}{42} x\ 5 = 16.905$$

For Q_1 :

Q_1 = Value of $(N/4)^{th}$ observation
= Value of $180/4^{th}$ observation
= Value of 45^{th} observation
is found in the class interval of 5-9

$$Q_1 = L + \frac{\left(\frac{N}{4}\right) - cf}{f} C$$

$$= 5 + \frac{\left(\frac{180}{4}\right) - 28}{36} x\ 5$$

= 7.361

Quartile Deviation, QD = $\frac{Q_3 - Q_1}{2}$

$$= \frac{16.905 - 7.3611}{2}$$

= 4.772

Co-efficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

$$= \frac{16.905 \quad .361}{16.905 + 7.361}$$

= 0.393

Mean Deviation

The range and quartile deviations are not based on all data of observations and hence they do not show any scatter of all data of observations from the mean. But, mean deviation is a measure of dispersion based on all data of observations. It is the arithmetic mean (i.e ignoring sign value) of the deviations of a series computed from any measure of central tendency, i.e mean, median or mode. In general, it is computed from mean.

Mean Deviation, MD = $\frac{\sum |D|}{N}$...(Absolute measure)

where, D = deviation of observation from the mean of its series

N = number of observations

Co-efficient of Mean Deviation = $\frac{Mean\ Deviation}{Mean}$ ....(Relative measure)

Illustration No. 21

Let us find out the co-efficient of mean deviation, of a blacksmith's monthly income in a village for a year (illustration no.1 problem)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Income in Rs.	4000	4200	3700	3500	4500	5000	3800	3900	4100	4200	4250	4100

Solution

Mean value, $\bar{X}$ is taken from illustration no.1 as 4100 (Refer illustration no.1)

Let us add one more column after finding out mean value in the solution table as given below.

Month (i)	Income (X_i)	$\|D\| = \|X_i - \bar{X}\|$
Jan	4000	100
Feb	4200	100
Mar	3700	400
Apr	3500	600
May	4500	400
Jun	4950	850
Jul	3800	300
Aug	3900	200
Sep	4100	0
Oct	4200	100
Nov	4250	150
Dec	4100	0
N = 12	ΣX_i = 49200	$\Sigma\|D\| = 3200$

Mean Deviation, MD $= \frac{\Sigma|D|}{N}$

= 3200/12

= Rs. 266.66

Co-efficient of Mean Deviation $= \frac{Mean\ Deviation}{Mean}$

= 266.66 / 4100

= 0.065

Illustration No. 22

Let us find out mean and co-efficient of mean deviation of the following problem. A survey was carried out to find out the age (in years) of 180 cows in a village. The result of the survey is as follows:

Age of cow (in years)	Less than 5	Less than 10	Less than 15	Less than 20	Less than 25
No. of cows	28	64	119	161	180

Solution

The table is redrawn taking class interval of 5 and mid value of the class is taken as variable, m

Age of cow (in years)	Mid value of the class, m	No. of cows (f)	d_i = m - A where A=12	f d	$\|D\| = \|X_i - \bar{X}\|$	f $\|D\|$
0-4	2	28	-10	-280	9.67	270.76
5-9	7	36	-5	-180	4.67	168.12
10-14	12	55	0	0	0.33	18.15
15-19	17	42	5	210	5.33	223.86
20-24	22	19	10	190	10.33	196.27
		N = 180		Σfd = -60		Σ f $\|D\|$ =877.16

Let us use shortcut method to find out mean as the mid values are high for calculation. Let assume A = 12. Corresponding deviations are found as d_i

Mean, $\bar{X}$= A + $\frac{\Sigma fd}{N}$

= 12 – (60/180)

= 11.67 years

Now the deviation of mid value (variable)from mean is calculated as D

Mean Deviation, MD = $\frac{\sum f|D|}{N}$

= 877.16/180

= 4.873 years

Co-efficient of Mean Deviation = $\frac{Mean\ Deviation}{Mean}$

= 4.873 / 11.67

= 0.418

Standard Deviation

The drawback of ignoring the algebraic sign in mean deviation, is overcome by taking the square of deviation, thereby making all the deviations as positive. Standard deviation is defined as a positive square root of the arithmetic mean of squares of the deviations of the given observation from the arithmetic mean. It is generally denoted as 'σ'

Standard deviation, σ = $\sqrt{\frac{\Sigma(X_i-\bar{X})^2}{N}}$ or $\sqrt{\frac{\Sigma(D)^2}{N}}$ or $\sqrt{\frac{\Sigma fD^2}{\Sigma f}}$

Illustration No.23

Let us find out the standard deviation for the previous problem (illustration no.21)

After finding out mean, $\bar{X}$, the columns for D, D^2 and fD^2 are drawn as shown below.

Age of cow (in years)	Mid value of the class, m	No. of cows (f)	d_i = m - A where A=12	f d	D = X -$\bar{X}$ Where 'm' is taken as 'X'	D^2	f D^2
0-4	2	28	-10	-280	-9.67	93.51	2618.28
5-9	7	36	-5	-180	-4.67	21.81	785.16
10-14	12	55	0	0	0.33	0	0
15-19	17	42	5	210	5.33	28.41	1193.22
20-24	22	19	10	190	10.33	106.71	2027.49
		N = 180		Σfd = -60			ΣfD^2= 6624.15

Mean, $\bar{X}$= A + $\frac{\Sigma fd}{N}$

= 12 – (60/180)
= 11.67 years

Standard deviation, $\sigma = \sqrt{\frac{\Sigma f D^2}{\Sigma f}}$

$= \sqrt{\frac{6624.15}{180}}$

$= \sqrt{36.8}$

= 6.066 years

Illustration No.24

Let us find out the standard deviation for the illustration problem no.1

Solution

Mean value, $\bar{X}$ is taken from illustration no.1 as 4100 (Refer illustration no.1)

Let us add two more columns for D and D^2 in the solution table as given below.

Month (i)	Income (X_i)	$D = X_i - \bar{X}$	D^2
Jan	4000	-100	10000
Feb	4200	100	10000
Mar	3700	-400	160000
Apr	3500	-600	360000
May	4500	400	160000
Jun	4950	850	722500
Jul	3800	-300	90000
Aug	3900	-200	40000
Sep	4100	0	0
Oct	4200	100	10000
Nov	4250	150	22500
Dec	4100	0	0
N = 12	ΣX_i = 49200		ΣfD^2= 1425000

Standard deviation, $\sigma = \sqrt{\frac{\Sigma(D)^2}{N}}$

$= \sqrt{\frac{1425000}{12}}$

= Rs. 344.6

2.3 Measure of Skewness

Measures of Central Tendency and Measure of dispersion discussed in previous units do not indicate whether the distribution is symmetric or not. There are some frequency distributions which differ widely in nature and composition. Skewness is a measure of degree and direction of departure from symmetry.

The following table 2.1 gives the differences between Dispersion and Skewness.

Table 2.1 Differences between Dispersion and Skewness

S.No.	Dispersion	Skewness
1	Gives the spread of individual values about mean	Gives the departure from symmetry
2	Judges the truthfulness of the central tendency	Judges the differences between the central tendencies
3	It is an averages of deviation – averages of second order	It is not an average, but measured by the average, median and mode
4	Shows the degree of variability	Shows if the concentration is in the side of higher or lower values

Skewness can be positive or negative or zero.
When mean = median = mode, there is no skewness
When mean > median > mode, skewness will be positive
When mean < median < mode, skewness will be negative

The above cases are shown in fig. 2.4

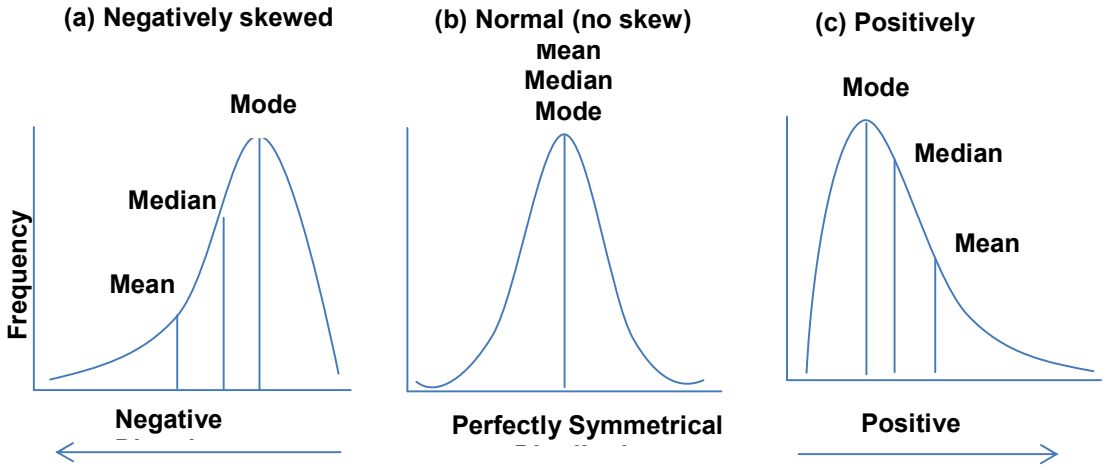


Fig 2.4 Three different situations of Skewness

Measures of Skewness can be computed by

a) Karl-Pearson's method
b) Bowley's method
c) Kelly's method

Out of the above three methods, Karl-Pearson's method is widely used to compute Skewness and hence this method only is discussed in this chapter.

Karl-Pearson's co-efficient or Pearsonian co-efficient of skewness is given by

$Sk_p = \frac{\bar{X} - Mode}{\sigma}$

$= \frac{3(\bar{X} - Median)}{\sigma}$; when mode is ill defined or multi-modal distribution

Illustration No. 25

Let us find out the Pearsonian co-efficient of skewness for the problem discussed in illustration no.23

The mean, standard deviations are directly taken from the problem no. 23 as

$\bar{X}$ = 4100 and σ = 344.6

In this problem, there are two modes and hence we will go for the formula with median instead of mode.

$Sk_p = \frac{3(\bar{X} - Median)}{\sigma}$

We need to find out median now to use the above formula.

Let us reproduce the given table with the values in ascending order

Si.No.	1	2	3	4	5	6	7	8	9	10	11	12
Income(X)	3500	3700	3800	3900	4000	4100	4100	4200	4200	4250	4500	4950

Median = Value of $\frac{N+1^{th}}{2}$ item

= Value of $\frac{12+1^{th}}{2}$ item

= Value of 6.5^{th} item

= Value of $\frac{6th\ item + 7th\ item}{2}$

$= \frac{4100+4100}{2}$

= 4100

Hence, $Sk_p = \frac{3(\bar{X} - Median)}{\sigma}$

$= \frac{3(4100 - 4100)}{344.6}$

= 0

The zero value of skewness indicates that the given frequency distribution of values is perfectly symmetrical.

Illustration No. 26 :

Let us find out the co-efficient of skewness for the problem discussed in Illustration No. 22. Also we will find out the median value from the skewness index.

Let us take the required mean and standard deviation values directly from the problem no.22 as

$\bar{X}$ = 11.67 and σ = 6.066

Let us find out the mode now. We will reproduce the table with class and frequency.

Age of cow (in years)	Mid value of the class, m	No. of cows (f)
0-4	2	28
5-9	7	36
10-14	12	55
15-19	17	42
20-24	22	19
		N = 180

The mode lies in class 10-14 as it has got more frequency of 55 compared to other classes. The mode, M_o is given by

$$M_o = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} C$$

$$= 10 + \frac{55-3}{2(55) - 36 - 42} x\ 5$$

$$= 12.97$$

Hence, $Sk_p = \frac{\bar{X} - Mode}{\sigma}$

$$= \frac{11.67 - 12.97}{6.066} = -0.214$$

The negative value indicates that distribution is negatively skewed (i.e. towards left in the normal curve)
To find median,

$$Sk_p = \frac{3(\bar{X} - Median)}{\sigma}$$

-0.214 = 3 (11.67-Median) / 6.066
-1.298 = 35.01 – 3 Median
3 Median = 36.308
Median = 12.1 (closer to value of 12.3 as determined in illustration no. 15)

2.4 Measure of Moments

Moment describes various characteristics of frequency distribution. It is defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution. It is denoted by the letter 'μ'. The measures of tendency, dispersion and skewness discussed in previous units to describe a frequency distribution, may be classified into following two groups,

i) Percentile system
ii) Moment system

Percentile system includes measures like mean, quartile, decile, percentile and so on as the value of these measures represent a given proportion of the observation.

Moment system includes measures like mean, mean deviation, standard deviation and so on as the value of these measures represent the deviation of individual observations from a given point.

The first four moments about arithmetic mean are given below.

Type of moment	for individual series	for discrete series
First moment about the mean, $\mu_1 = 0$	$\frac{\sum(X-\bar{X})}{N} = \frac{\sum d}{N} = 0$	$\frac{\sum f(X-\bar{X})}{\Sigma f} = \frac{\sum fd}{\Sigma f}$
Second moment about the mean, μ_2 =Variance	$\frac{\sum(X-\bar{X})^2}{N} = \frac{\sum d^2}{N}$	$\frac{\sum f(X-\bar{X})^2}{\Sigma f} = \frac{\sum fd^2}{\Sigma f}$
Third moment about the mean, μ_3 μ_3 = 0 for symmetrical distribution	$\frac{\sum(X-\bar{X})^3}{N} = \frac{\sum d^3}{N}$	$\frac{\sum f(X-\bar{X})^3}{\Sigma f} = \frac{\sum fd^3}{\Sigma f}$
Fourth moment about the mean, μ_4 = Kurtosis	$\frac{\sum(X-\bar{X})^4}{N} = \frac{\sum d^4}{N}$	$\frac{\sum f(X-\bar{X})^4}{\Sigma f} = \frac{\sum fd^4}{\Sigma f}$

It will be difficult to compute the values, if the mean value is fractional. In such cases, moments are calculated about an assumed mean first and then actual moment is calculated from it using the following formulae. Moment about working origin is denoted by μ'

The first four moments about assumed mean (working origin) are given below.

Type of moment	for individual series (ungrouped data)	for discrete series (grouped data)
First moment about the working origin, μ_1' = Mean about origin	$\frac{\sum(X-A)}{N} = \frac{\sum d}{N}$	$\frac{\sum f(X-A)}{\Sigma f} = \frac{\sum fd}{\Sigma f}$
Second moment about the working origin, μ_2'	$\frac{\sum(X-A)^2}{N} = \frac{\sum d^2}{N}$	$\frac{\sum f(X-A)^2}{\Sigma f} = \frac{\sum fd^2}{\Sigma f}$
Third moment about the working origin, μ_3'	$\frac{\sum(X-A)^3}{N} = \frac{\sum d^3}{N}$	$\frac{\sum f(X-A)^3}{\Sigma f} = \frac{\sum fd^3}{\Sigma f}$
Fourth moment about the working origin, μ_4'	$\frac{\sum(X-A)^4}{N} = \frac{\sum d^4}{N}$	$\frac{\sum f(X-A)^4}{\Sigma f} = \frac{\sum fd^4}{\Sigma f}$

The moments from-the mean are calculated from the moments from-the working origin as given below.

$\mu_1 = \mu_1' - \mu_1' = 0$

$\mu_2 = \mu_2' - (\mu_1')^2$

$\mu_3 = \mu_3' - 3\,\mu_1'\mu_2' + 2\,(\mu_1')^3$

$\mu_4 = \mu_4' - 4\,\mu_1'\,\mu_3' + 6\,(\mu_1')^2\,\mu_2' - 3\,(\mu_1')^4$

Other expressions from moments are :

1) Karl Pearson's coefficient of skewness is given by

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3}$$

2) Standard deviation, $\sigma = \sqrt{\mu_2}$

Illustration No. 27

Let us find out the nature of the distribution which has the first four moments about the origin as 1, 4, 10 and 46 respectively.

Solution

Given data are A = 0, $\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$, $\mu_4' = 16$

Let us find out the moments about mean as follows

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2 = 4 - 1^2 = 3$

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = 10 - 3(1)(4) + 2(1)^3 = 0$

$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4 = 16 - 4(1)(10) + 6(1)^2(4) - 3(1)^4 = -3$

Parameters of frequency distribution :

1) Mean about working origin, $\mu_1' = 1$

   Hence, mean about origin = A ± μ_1' = 1 (here working origin = origin as A = 0)

2) Karl Pearson's coefficient of skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0^2/3^3 = 0$

3) Since $\beta_1 = 0$, the given distribution is symmetrical and hence Mode = Median = Mean = 1

4) Standard deviation, $\sigma = \sqrt{\mu_2} = \sqrt{3} = 1.732$

Illustration No. 28

Let us find out the parameters of distribution of data given in illustration no. 21 using moments method The table is reproduced here from problem no. 21

Age of cow (in years)	Mid value of the class, m	No. of cows (f)	d_i = m - A where A=12	f d	fd^2	fd^3	fd^4
0-4	2	28	-10	-280	2800	- 28000	280000
5-9	7	36	-5	-180	900	- 4500	22500
10-14	12	55	0	0	0	0	0
15-19	17	42	5	210	1050	5250	26250
20-24	22	19	10	190	1900	19000	190000
		Σf= 180		Σfd = -60	Σfd^2= 6650	Σfd^3= -8250	Σfd^4= 518750

Let us find out the moments about assumed mean or working origin as follows

$\mu_1' = \frac{\Sigma fd}{\Sigma f}$ = -60/180 = -0.333,

$\mu_2' = \frac{\Sigma fd^2}{\Sigma f}$ = 6650/180 = 36.944

$\mu_3' = \frac{\Sigma fd^3}{\Sigma f}$ = -8250/180 = -45.833

$\mu_4' = \frac{\Sigma fd^4}{\Sigma f}$ = 518750/180 = 2881.944

Let us find out the moments about mean as follows

$\mu_1 = 0$ (by default)

$\mu_2 = \mu_2' - (\mu_1')^2 = 36.944 - (-0.333)^2 = 36.833$

$\mu_3 = \mu_3' - 3\,\mu_1'\mu_2' + 2\,(\mu_1')^3 = -45.833 - 3(-0.333)(36.944) + 2(-0.333)^3 = -9$

$\mu_4 = \mu_4' - 4\,\mu_1'\,\mu_3' + 6\,(\mu_1')^2\,\mu_2' - 3\,(\mu_1')^4$

$= 2881.944 - 4(-0.333)(-45.833) + 6(-0.333)^2\,(36.944) - 3(-0.333)^4 = 2845.438$

Parameters of frequency distribution :

1) Mean = $A \pm \mu_1' = 12 - 0.333 = 11.667$ (same as in problem no.21)
2) Karl Pearson's coefficient of skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = (-9)^2/36.833^3 = 0.0016$
3) Since $\beta_1 \approx 0$, the given distribution is almost symmetrical and hence Mode = Median = Mean = 11.667
4) Standard deviation, $\sigma = \sqrt{\mu_2} = \sqrt{36.833} = 6.069$ (same as in problem no.22)

Illustration No. 29

The first three moments of a variable measured from the value '2' are 1, 16 and -40 respectively. Find out the mean, variance and fourth moment from mean.

Solution

The first three moments about A = 2 are

$\mu_1' = 1$; $\mu_2' = 16$; $\mu_3' = -40$

Mean = $A + \mu_1'$

$= 2 + 1 = 3$

Variance, $\mu_2 = \mu_2' - (\mu_1')^2$

$= 16 - 1^2 = 15$

Third moment, $\mu_3 = \mu_3' - 3\,\mu_1'\mu_2' + 2\,(\mu_1')^3$

$= -40 - 3(1)(16) + 2(1)^3$

$= -86$

2.5 Measure of Kurtosis

The expression 'Kurtosis' is used to describe the peakedness of a frequency distribution curve of observations. It measures the extent to which a distribution is more peaked or more flat topped than the normal curve. Fig 2.5 shows the three forms of kurtosis.

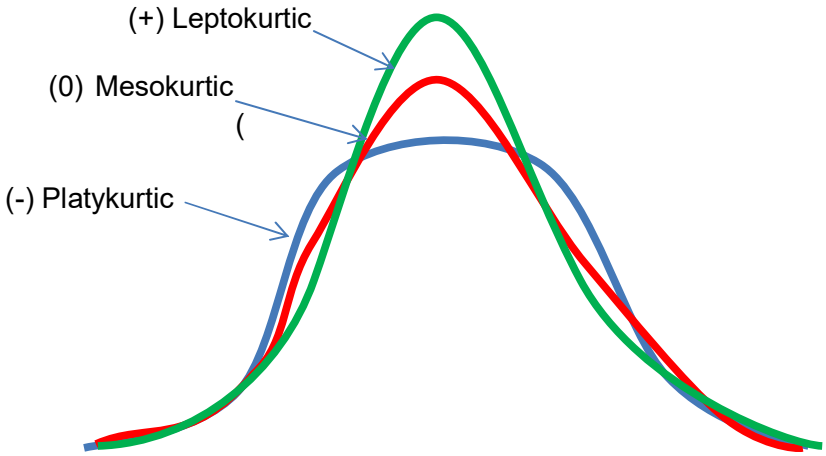


Fig 2.5 Forms of Kurtosis

Measures of kurtosis of a frequency distribution are based on the fourth moment about the mean of distribution. It is generally denoted as β_2

$\beta_2 = \frac{\mu_4}{\mu_2^2}$ where μ_4= 4th moment, μ_2= 2nd moment

If β_2 = 3, then the distribution is said to be normal and the curve also is normal (mesokurtic)

If β_2> 3, then the distribution is said to be more peaked and the curve is leptokurtic

If β_2< 3, then the distribution is said to be flat topped and the curve is platykurtic

Illustration No. 30 :

Let us check the type of kurtosis for the distribution discussed in previous problem.

Let us reproduce the values of moments about mean as follows

$\mu_1 = 0$; $\mu_2 = 36.833$; $\mu_3 = -9$; $\mu_4 = 2845.438$

Measures of kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2}$

$= \frac{2845.438}{36.833^2}$

$= 2.097$

Since β_2< 3, the given distribution is said to be flat topped and the curve is platykurtic.

Summary

In this chapter, we have learnt the various characteristics of distribution of data. Measures of central tendency like mean, median, mode, quaterile, decile and percentile give the general level of magnitude of distribution, but fail to show anything further about the distribution. Measures of variation or dispersion explain how the items are dispersed away from mean. Measures of skewness explain whether a distribution is symmetrical or asymmetrical. Measures of Kurtosis describe the peakedness of the distribution or relative influence of central deviation. Measure of first four moments of distribution of data supports measure of central tendency, measure of skewness and measure of kurtosis. Various illustrations were discussed to learn the procedure to find out the required parameter of the distribution of given data.

Model Questions

Problems

11. Find out the mean, median, mode, standard deviation of the following distribution of data. Also check if the curve representing the given data is of symmetrical; platykurtic using moments method.

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	20	15	10	5

Ans :mean, median and mode = 4, standard deviation = 2, $\beta_2 = 2.35 < 3$

12. Find out the Pearsonian co-efficient of skewness for the following data of distribution. Also find out the mean, mode and standard deviation.

Cultivation of rice (in 25 kg bags)	400-500	500-600	600-700	700-800	800-900
No. of farms (in acres)	8	16	20	17	3

Ans :mean = 635.94, mode = 657.14, standard deviation = 108.7, Co-eff = -0.195

13. Find out the first four moments about the mean and hence the skewness and kurtosis for the following data.

Yield income of a farmer in thousands Rs.	0-10	10-20	20-30	30-40
No. of farmers	1	3	4	2

Ans :First four moments about mean = 0,81,-144,14817, skewness = 0.039, kurtosis = 2.26

14. 50 students in a class have scored the following marks in the subject 'Business Analytics'. If 60% of the students have passed the exam, find out the minimum marks obtained by a pass student. Hint : Find out the fourth decile, D_4

Marks more than	0	10	20	30	40	50
No. of students	50	46	40	20	10	3

Ans : 25

15. Find out the missing frequency of the following data of distribution. Arithmetic mean is found to be 28.

No. of persons in a self help group	0-10	10-20	20-30	30-40	40-50	50-60
No. of groups	12	18	27	?	17	6

Ans : 20

16. Check if Geometric and Harmonic mean for the following are same.
0.5 0.4 5 125 130 75 10 45 500 150

Ans : GM = 22.98 HM =2.06

17. Find out the lower quartile, 7^{th} decile and 85^{th} percentile of the following distribution of data

Marks scored in Business Statistics	Below 10	10-20	20-30	30-40	40-50	50-60	60-70	Above 70
No. of students	8	12	20	32	30	28	12	4

Ans : Lower quartile = 28.25, 7^{th} Decile = 50.07, 85^{th} percentile = 57.9

18. Following is the data of income of two villages. Find out which village has got more variation in income.

	Village A	Village B
No. of people in the village	600	500
Average income per person in Rs.	175	186
Variance in income in Rs.	100	81

Ans :C.V (A) = 5.71% , CV(B) = 43.84%

19. The first two moments of a distribution about a value 5 of the variable are 2 and 20. Find out the mean and variance.
Ans : 7 & 16
20. For a moderately skewed data, the arithmetic mean is 200, the coefficient of variation is 8 and pearsonian coefficient of skewness is 0.3. Find the mode and median.

Ans : mode = 195.2, median = 198.4

Theoretical Questions

1. What is meant by central tendency?
2. What are the characteristics a good measure of central tendency should possess?
3. Explain briefly the functions of central tendency
4. What is the difference between arithmetic mean and weighted mean?
5. List out the differences between Geometric mean and Harmonic mean.
6. List out any one application of various forms of average.
7. Write down the relationship between mean, median and mode.
8. Define the following
   a) Mean b) Median c) Mode d) Quartile e) Decile f) Percentile
9. List out the objectives of measure of dispersion
10. What are the properties a good measure of dispersion should possess?
11. Define 'Range' and write down its applications in statistics.
12. Write down the difference between quartile deviation, mean deviation and standard deviation.
13. Differentiate dispersion from skewness
14. Show by means of a diagram how mean, mode and median change the form of skewness.
15. Define measures of moments. How are they useful in deriving various elements of distribution?
16. What are the forms of kurtosis?

Multiple Choice Questions

1. If the value of mode is 14 and value of mean is 5 then value of median is
   a) 12
   b) 8
   c) 18
   d) 14

2. Numerical value which shows tendency around central value of cluster is classified as
   a) Cluster tendency
   b) Central tendency
   c) Group tendency
   d) Numerical tendency
3. Concept used in calculation of index numbers and where smaller observations must be taken into consideration is called
   a) Geometric mean
   b) Deviation square mean
   c) Paired mean
   d) Harmonic mean
4. If the arithmetic mean is 20 and the harmonic mean is 30, then the geometric mean is
   a) 14.94
   b) 34.94
   c) 44.94
   d) 24.94

5. Properties such as variation, central tendency and shape of distribution of frequency are the bases to extract information from
   a) Descriptive measures
   b) Extended measures
   c) Skewed measures
   d) Ordinal measures
6. Distribution in which values of median, mean and mode are not equal is considered as
   a) experimental distribution
   b) symmetrical distribution
   c) asymmetrical distribution
   d) exploratory distribution

7. Types of descriptive measures include
   a) measures of skewness
   b) measures of dispersion
   c) measures of central tendency
   d) all of above
8. The number of observations is 30 and the value of arithmetic mean is 15, then sum of all values is
   a) 15
   b) 200
   c) 45
   d) 450

9. The mean and median of 100 items are 50 and 52 respectively. The value of largest item is 100. It was later found that it is 110 not 100. The true mean and median are
   a) 50.1, 51.5
   b) 50, 52
   c) 50.1, 52
   d) 50, 51.5

10. Coefficients of variation of two distributions are 50 and 60 and their arithmetic means are 30 and 25 respectively. Difference of their standard deviation is
   a) 0
   b) 1
   c) 1.5
   d) 2.5

11. If quartile deviation of a sample is 20, then the most likely value of standard deviation is
   a) 18
   b) 12
   c) 30
   d) 13

12. Measurement techniques used to measure extent of skewness in data set values are called
   a) measure of distribution width
   b) measure of median tail
   c) measure of tail distribution
   d) measure of skewness

13. If first and third quartiles are as 32 and 35 respectively with median of 20 then distribution is skewed to
   a) closed end tail
   b) upper tail
   c) lower tail
   d) open end tail

14. Moment about mean which is indication if distribution is symmetrical or asymmetrical is considered as
   a) first moment
   b) second moment
   c) third moment
   d) fourth moment

15. Kurtosis defines peakness of curve in region which is
   a) around the mode
   b) around the mean
   c) around the median
   d) around the variance

16. The sum of squares of the deviations is minimum, when deviations are taken from
   a) Median

b) Mean
c) Mode
d) Range

17. Half of the difference between upper and lower quartiles is called
    a) Mean deviation
    b) Interquartile range
    c) Quartile deviation
    d) Standard deviation
18. Measure of variation which is useful for highly skewed distribution is
    a) Inter quartile range
    b) Quartile deviation
    c) Inter quartile deviation
    d) Quartile range
19. Value of first quartile is 23 and inter quartile range is 20, then the value of third quartile is
    a) 63
    b) 37
    c) 53
    d) 43
20. The degree of peakness or flatness of a uni- model distribution is called
    a) Skewness
    b) Kurtosis
    c) Dispersion
    d) Symmetry

Answers for MCQ

1	2	3	4	5	6	7	8	9	10
b	b	a	d	a	c	d	d	c	a
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
c	d	c	c	a	b	c	b	d	b

References

1. Sharma, J.K. (2014). *Business Statistics – Problems and Solutions*. New Delhi : Vikas Publishing House Pvt Ltd.
2. Pillai, R.S.N. &Bagavathi, V. (1999). *Statistics*. New Delhi :S.Chand& Company Ltd.
3. Gupta, S.P. (2010). *Statistical Methods.* New Delhi :S.Chand& Company Ltd.
4. Beri, G.C. (2011). *Business Statistics.* New Delhi : Tata McGraw Hill Educations Pvt Ltd.
5. Foster, D. & Stine, E.R. (2010). *Statistics for Business : Decision Making and Analysis.* New Delhi : Pearson Publishers.
6. Gupta, S.C. & Kapoor, V.K. (2006). Fundamentals of Mathematical *Statistics.* New Delhi :S.Chand& Company Ltd.
7. Srivastava, S.C & Srivastava, S. (2003). Fundamentals of Statistics. New Delhi : Anmol Publications Pvt. Ltd.

Chapter 3 Probability, Distributions and Estimation

Introduction

The word probability refers to the study of randomness and uncertainty. Consider the following example:

I. A plant may or may not be infected by species during the rainy seasons.
II. The price of the gold is increasing under economic conditions in a country.
III. The birth of an individual depends on a biological chance resulting in a male or a female.

Objectives of the Chapter

- To be aware of random experiments, trials, outcomes, events and sample space.
- To provide insights on mutually exclusive and mutually exhaustive events
- To provide insights about the theorems on probability and applies in problems.
- To provide insights about independent events and multiplication theorem on probability.
- To be aware of the concept of conditional probability and independent events.
- To be aware of the application of Bayes' theorem.
- To be aware of the concept of Binomial, Poisson, and Normal distribution.
- To be aware of properties of normal probability curve.
- To be aware of sampling distribution and estimation.

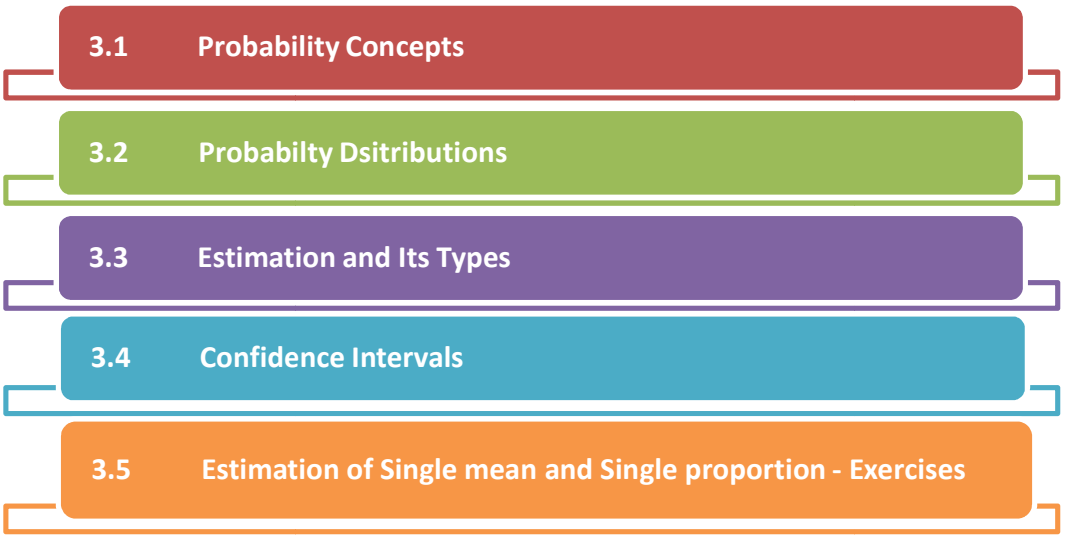


Fig 3.1 Chapter Flow

In the following example given in fig. 3.2, these is an amount of uncertainty prevails. The word probability involves synonyms such as 'chance possible' 'probably' 'likely' 'odds' 'uncertainty' 'incidence' 'risk' 'expectancy' etc. Our entire world is filled with uncertainty. It is very much essential to determine a quantitative value to the chance of probability.

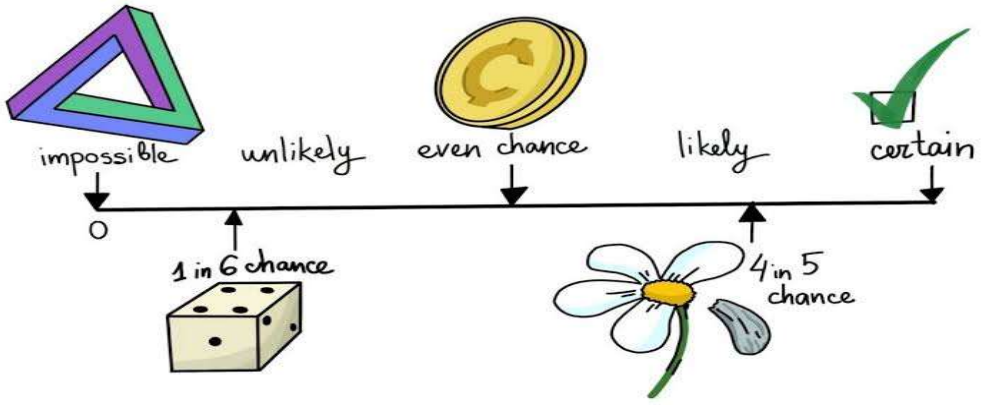


Fig 3.2 Events

3.1 Probability Concepts

Experiment

In statistics the word experiment means, it is a process for which its result is well defined a laboratory experiment.

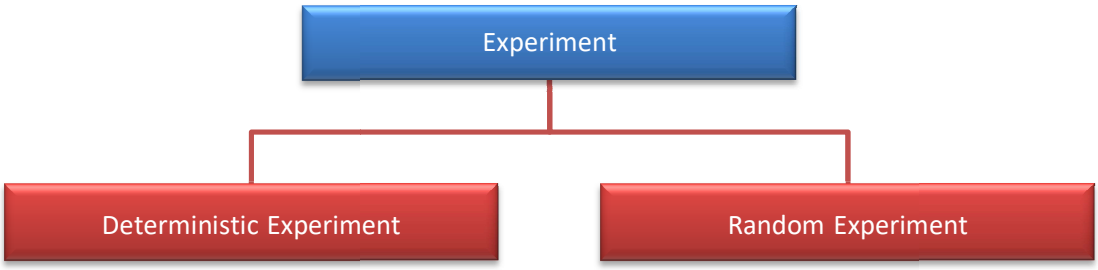


Fig 3.3 Types of experiment

Following are the two types of experiment:

1. Deterministic experiment
2. Random experiment

Deterministic Experiment

The outcome of an experiment is predicted with certain under real conditions.

Examples

The experiments that we conduct to verify the laws of science or established laws of other areas are the best examples,

I. An experiment conducted to verify the Newton's Laws of motion.
II. An experiment conducted to verify the economic law of demand.

Random Experiment

In an experiment if,

1. All the possible outcomes of the experiment are predictable in advance
2. Outcome of any trial of the experiment is not known, in advance and

3. Can be repeated under identical conditions.

then, the experiment is called a random experiment.

Examples

1. Flipping a coin
2. Tossing a dice
3. Taking a card from a pack of playing cards.
4. Obtaining blood samples from a group of persons.

Outcome : The result of random experiment will be called an outcome.

Trial : Any particular performance of a random experiment is called a trial.

Event : The possible outcome of an experiment is called event.

Types of Events

The various types of events are shown in fig. 3.4 & 3.5

1. Sure event.
2. Impossible event.
3. Complementary event.
4. Mutually exclusive event.
5. Mutually exhaustive event.
6. Independent and Dependent events.
7. Equally likely event.

Sure Event : The sample space is called sure event.

Impossible Event : The null set φ is called an impossible event.

Complementary Event : The event $\bar{A}$ is called the complement event of A.

Example

Let S = {1,2,3,4,5,6}

A = {2,4,6}

$\bar{A}$ = {1,3,5}

Mutually Exclusive Event : Two events A and B are said to be mutually exclusive or disjoint events if $A \cap B = \phi$.

Mutually Exhaustive Event : Two events A and B are said to be mutually exhaustive if $A \cup B$ = Sample space **S**.

Illustration:

Let S = {HH, HT, TH, TT}, when two coins are Tossed.

Let A = {HH, TT}, B = {HT, TH}

$A \cap B = \phi$ and $A \cup B = S$

↓ ↓

Mutually Exclusive **Mutually Exhaustive**

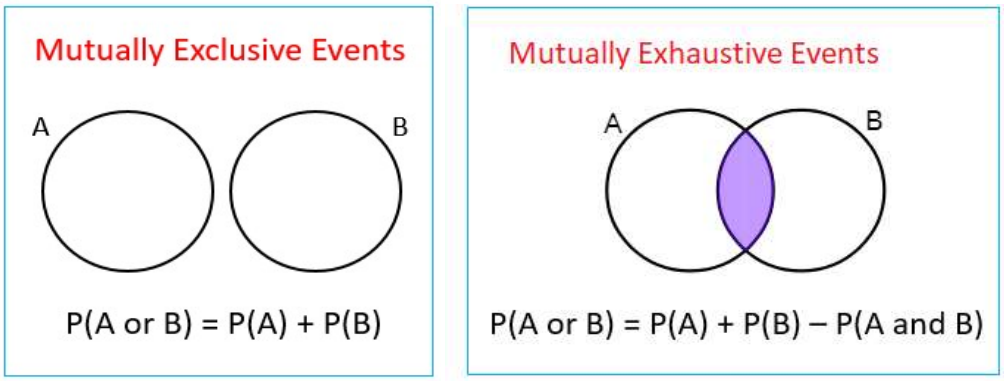


Fig 3.4 Types of event

Independent and Dependent Events

Two or more number of events are said to be independent, if the happening of another event does not affect the happening of the other event.

i.e. Two events A and B are said to be independent if $P(A \cap B) = P(A)\,P(B)$

Two or more number of events are said to dependent, if the happening of one event affect the happening of the other event

i.e. $P(A \cap B) \neq P(A)\,P(B)$

Equally Likely Event

Two events A and B are said to be equally likely if one of them cannot be expected in preference to other.

Notations

Let E_1 and E_2 be two events

1. $E_1 \cup E_2$ stands for the occurrence of E_1 or E_2 or both
2. $E_1 \cap E_2$ →simultaneous occurrence of E_1 and E_2
3. $\overline{A}$ →non-occurrence of A
4. $A \cap \overline{B}$ → occurrence of only A
5. $\overline{A} \cap B$ → occurrence of only B.

Sample Space

The set of all possible outcomes of random experiment is called the sample space and it is generally denoted by the letter **'s'.**

Example-1

1. It a coin is tossed then the sample space s= {H, T}
2. A die is rolled, then the sample s= {1,2,3,4,5,6}
3. Consider an experiment in which each of the automobiles taking a particular freeway exit turns left (L) or right (R) at the end of the exit ramp.

Sample space S={LLL, RLL, LRL, LLR, LRR, RLR, RRL, RRR}

Example-2

Let us take an experiment tossing a coin until a head appear. In this experiment one cannot say in

advance how many tosses will be required; and the sample space s= {H, TH,TTH,TTTH,TTTTH,.....} In an infinite set

Example-3

The life Length (t: in Hours) of a Tube Light Is S={t:0<t<2000}

From the examples (2) & (3), one needs to define the difference between two types of infinite sets.

i) Countable infinite

ii) Uncountable infinite

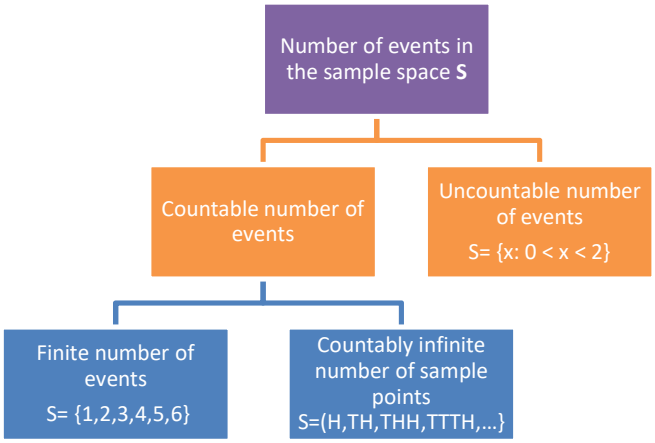


Fig 3.5 Types of event

Probability

Let S be the sample space associated with a random experiment and let E be an event.

Let n(S) and n(E) be the number of elements in S and E respectively. Then the probability of the event E is defined as

$$P(E) = \frac{n(E)}{n(S)} = \frac{Number\ of\ favourable\ cases\ to\ A}{Total\ number\ of\ possible\ cases}$$

Every probabilistic model involves an underlying process shown in fig. 3.6 and fig. 3.7

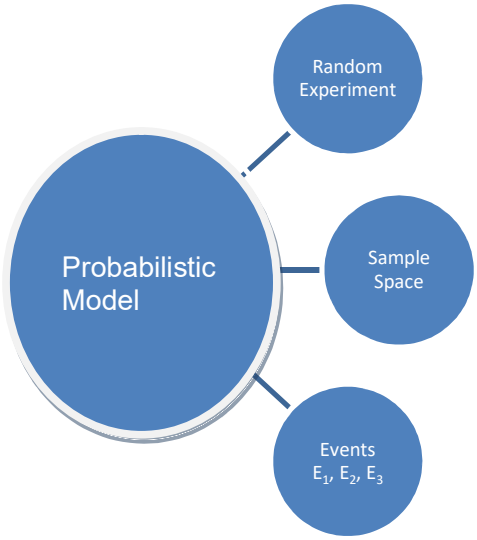


Fig 3.6 Probabilistic model

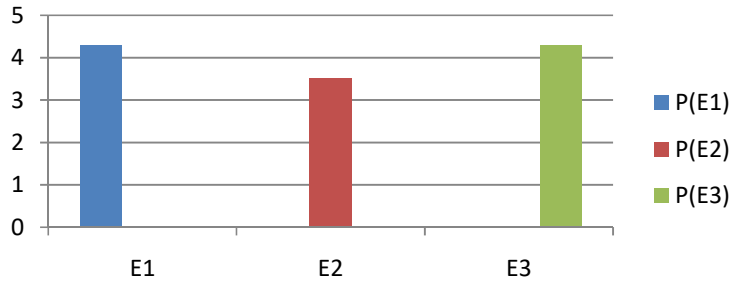


Fig 3.7 Graphical representation

Arithmetic approach to probability let sbe the sample space associated with a random experiment and E ∈ S be any event.

Axiom

For any event E,P(E) ≥0.

P(S) = 1.

(a) If E_1, E_2.....E_n is a collection of mutually exclusive events then

P ($E_1 U E_2 U$.....$U E_n$) = P(E_1) + P(E_2) +.....+ P(E_n) = $\sum_{i=1}^{n} p(Ei)$

(b) If E_1, E_2 ..... is an infinite collection of mutually exclusive events then

P (E_1 U E_2 U....) = $\sum_{i=1}^{n} p(Ei)$

Properties of Probability

Property 1: The probability of impossible event is zero. i.e. P (ϕ) = 0.

Let $\phi \in S$ be any event.

S U ϕ = S and S n = Φ

S and ϕ are mutually exclusive and exhaustive events.

P (SU ϕ) = P(S)

P(S) + P (ϕ) = P(S)

p (ϕ) =0

Property II: If S is the sample space and E ∈ S be any event then P ($\bar{E}$) = 1-P(E)

i.e. P (Not happening) = 1-P(happening)

Let E and $\bar{E} \in S$

E ∪ $\bar{E}$ = S and E ∩ $\bar{E}$ = ϕ

E and ϕ are mutually exclusive and exhaustive events.

P (EU$\bar{E}$) = P(S)

P(E) + P($\bar{E}$) = 1 , Since

P($\bar{E}$) =1- P(E)	p(S)=1

Property III: If A and B mutually exclusive events then P (A∩B) = 0

A and B are mutually exclusive,
A ∩ B = ϕ
p (A∩B) = P(ϕ)
P (A ∩ B) = 0

Property IV: Addition theorem (shown in fig. 3.8)
If A and B are any two events then P(AUB) = P(A) + P(B) - P(A∩ B)
Let A, B Ɛ S be any two events.

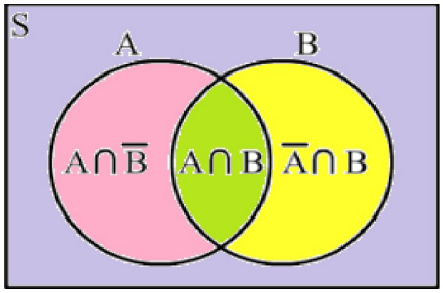


Fig 3.8 Addition theorem

According to the above figure 3.8,
AUB = AU ($\bar{A}$∩B)
Since A and $\bar{A}$∩B are mutually exclusive,
P (AUB)= P(A) + P($\bar{A}$∩B) ---------(1)
Now, B= (A∩B) U ($\bar{A}$∩B)
Since A and B and $\overline{A}$∩B are mutually exclusive P(B)= P(A∩B) + P($\bar{A}$∩B)
P(B) - P(A∩ B) = P(A`∩ B) ----------(2)
Using 2 in 1 we get
P (AUB) = P(A) + P(B) - P(A∩ B)

Alternate proof:
In set theory,
n(AUB) = n(A) + n(B) - n(A∩B)
Divide by n(s) we get , $\frac{n(AUB)}{n(S)} = \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(AnB)}{n(s)}$
P(AUB) = P(A) + P(B) - P(A∩B)
Note : If A,B,C are any three events then
P(AUBUC) = P(A) + P(B) + P(C) - P(A∩B) - P(B∩C) - P(A∩C) + P(A∩B∩C)

Illustration No. 1 :
In a certain village 60% of all households subscribe to the Jio setup box given by Reliance in a nearby showroom. 80% subscribe to the package contains serials channels or sports and 50% of all households subscribe both the channels. If a hose hold is selected at random what is the probability that the household subscribes to

i. at least one of the packages and
ii. exactly one of the two packages

Solution

Let A→ subscribe the package contains Serial or sports

B→ subscribes both the packages.

By given information,

P(A) = 60% = 0.6

P(B) = 80% = 0.8 and P(A∩B) = 50% = 0.5

(i) P(a house hold subscribes at least one of the package)= P(AUB)

= P(A) + P(B) - P(A∩B)

= 0.6 + 0.8 - 0.5 =0.9

(ii) P(exactly one of the two package) = P(An$\bar{B}$) + P($\bar{A}$∩B)

= [P(A)-P(A∩B]+[P(B)-P(A∩B]

= (0.6-0.5)+(0.8-0.5) =0.1+0.3 =0.4

Illustration No. 2

A box contains 3 Apples and 4 Oranges, find the probability of selecting 2 fruits of same variety.

Solution

Number of fruits = 3 + 4 = 7 = n

A: selecting 2 Apples from 3Apples

B:selecting 2 Oranges from 4 Oranges

P(A) = $\frac{3}{7} \times \frac{2}{6} = \frac{1}{7}\left(\frac{3C_2}{7C_2}\right)$

P(B) = $\frac{4}{7} \times \frac{3}{6} = \frac{2}{7}\left(\frac{4C_2}{7C_2}\right)$

P (AUB) = P(A)+P(B) = $\frac{1}{7} + \frac{2}{7} = \frac{3}{7}$

Illustration No. 3

Let S be the sample space associated with an experiment throwing a die, if the die is constructed so that any one of the three even outcomes is twice as likely to occur as any of the three odd outcomes.

Find the probability that

(i) the outcome is even

(ii) the outcome ≤ 3

Solution

Let S = {1,2,3,4,5,6.}

Given, P(1) = P(3) = P(5) = K

and P(2) = P(4) = P(6) = 2K

We have $\sum p(x) = 1$

P(1) + P(2) + P(3) + P(4) +P(5) + P(6) =1

K + 2K + K + 2K + K + 2K = 1

9K = 1 ; K = 1/9

P(1) = P(3) =P(5) = $\frac{1}{9}$and P(2) = P(4) = P(6) = $\frac{2}{9}$

Let A= {outcome is even}

B = {outcome ≤ 3}

(1) P(the outcome is even)= P(2) + P(4) + P(6)

P (A) = $\frac{2}{9} + \frac{2}{9} + \frac{2}{9} = \frac{6}{9} = \frac{2}{3}$

(2) P(B) = P(1) + P(2) + P(3) = $\frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}$

Illustration No. 4

100 carrots were found to be stale in a basket of 800 carrots. Find the probability that a carrot selected from the basket is not stale.

Solution

Total number of carrots =800

Non stale carrots = 800 – 100 = 700

P(non-stale carrots) = $\frac{700}{800} = \frac{7}{8}$

Conditional Probability

Let us consider a situation, where a fair die is being rolled and you were asked to give the probability to get a five. We know that here are six equally liked outcomes, so your answer will be 1/6. But, assume that, if before you answer, you get extra information that the number rolled was an odd number. Since we know that there are only three odd numbers which are possible, out of which one of the number is five, you would certainly revise your calculation for the likelihood that a five was rolled, from 1/6 to 1/3. This revised probability that an event A has occurred, considering the additional information, that another event B has definitely occurred on this trial of the experiment, is called conditional probability of A given B and is denoted by P(A/B).

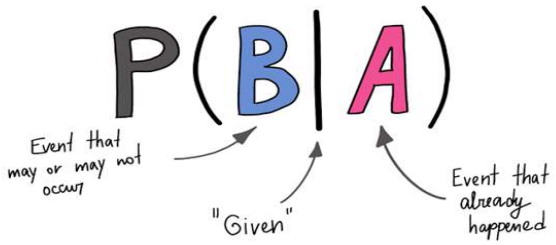


Fig 3.9 Conditional probability

For any two events A and B with P(B) > 0 the conditional probability of A given that B has already happened is defined by

$$P(A/B) = \frac{P(AnB)}{P(B)}, P(B) \neq 0$$

$$\text{Similarly, } P(B/A) = \frac{P(AnB)}{P(A)}, P(A) \neq 0$$

Multiplication theorem

The probability of simultaneous happening of two events A and B is given by

(i) $P(A \cap B) = P(A)\,P(B/A)$

(ii) $P(A \cap B) = P(B)\,P(A/B)$

Illustration No. 5

Suppose that of all individuals buying daily newspaper 60% include a Daily Thanthi in their purchase. 40% include "The Hindu" and 30% include both types of papers.

Calculate,

i. The probability that an individual purchase Daily Thanthi given that an individual purchase the Hindu.

ii. The probability that an individual purchase the Hindu given that an individual purchase Daily Thanthi.

Solution

Let A= {an individual buys Daily Thanthi}

B= {an individual buys The Hindu}

Then P (A) = 0.60, P (B) = 0.40 and P (A∩B) = 0.30.

I. P(A/B) = $\frac{P(AnB)}{P(B)}$

$= \frac{0.30}{0.40} = 0.75$

II. P(B/A) = $\frac{P(AnB)}{P(A)}$

$= \frac{0.30}{0.60} = 0.50$

Illustration No. 6

An unbiased die is thrown. If A is the event "The number appearing is odd" and B be the event "the number appearing is a prime number "then verify A&B are independent.

Solution

When a die is thrown, S = {1,2,3,4,5,6}

A = {1,3,5}, B = {2,3}

A∩B = {3}

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

$$P(B) = \frac{2}{6} = \frac{1}{3}$$

$$P(A)\,P(B) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} = P(A \cap B)$$

∴ A &B are independent

Illustration No. 7

Let $P(A) = \frac{3}{7}$ and $P(B) = \frac{2}{7}$ find P(A ∩ B) if A and B are independent.

Solution

If A and B independent $P(A \cap B) = P(A)P(B)$

$$= \frac{3}{7} \times \frac{2}{7} = \frac{6}{49}$$

Illustration No. 8

A number is selected randomly from the digits 21 to29. Consider the events, A = {21,24,26,28,29}, B={22,24,28,29}, C={23,25,28,29} Find,

i. P(A/B)
ii. P(A/C)
iii. P(B/A)

Solution

$A \cap B = \{24,28,29\}$

$A \cap C = \{28,29\}$

$B \cap C = \{28,29\}$

$P(A) = \frac{5}{9}, P(B) = \frac{4}{9} = P(C)$

$P(A \cap B) = \frac{3}{9}, \quad P(A \cap C) = \frac{2}{9} = P(B \cap C)$

$(i) \quad P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{3}{9}}{\frac{4}{9}} = \frac{3}{4}$

$(ii) \quad P(A/C) = \frac{P(A \cap C)}{P(C)} = \frac{\frac{2}{9}}{\frac{4}{9}} = \frac{2}{4} = \frac{1}{2}$

$(iii) \quad P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{3}{9}}{\frac{5}{9}} = \frac{3}{5}$

To Do Activity
Three seeds are planted. Each has probability of 0.3 of growing. Draw the distribution of the number of seeds that grow.

Law of Total Probability

Let $E_1, E_2, \ldots\ldots E_n$ be mutually exclusive and exhaustive events. Then for any other event A,

$P(A) = P(A/E_1)\,P(E_1) + P(A/E_2)\,P(E_2) + \ldots\ldots\ldots + P(A/E_n)P(E_n)$

$= \sum_{i=1}^{n} P(A/Ei)P(Ei)$

Baye's Theorem

Let $E_1, E_2, \ldots\ldots\ldots E_n$ be a collection of n mutually exclusive And exhaustive events with $P(E_i) > 0$, For i = 1,2…….n, then for any other events A for which P(A) > 0 ,

$$P(Ei/A) = \frac{P(Ei \cap A)}{P(A)}$$

$$= \frac{P(A/Ei)P(Ei)}{\sum_{i=1}^{n} P\left(E_i/_A\right) P(Ei)}$$

Illustration No. 9

The medical survey says 1 in 10,000 buffalo is affected Rota virus for which a diagnostics test has been developed. The test is such that, when a buffalo actually has the virus, a positive result will occur 98% of the time whereas a buffalo without the disease will show a positive test result only 2 % of the time. If a randomly selected buffalo is tested and the result is positive, then what is the probability that the buffalo has the virus?

Solution

Let E_1 = {buffalo has the virus}

E_2 = {buffalo does not have the virus}

A = {positive test result}

P (E_1) = 0.0001; P (E_2) = 0.9999

P (A/E_1) = 0.98; P (A/E_2) = 0.02

P (A) =P (E_1) P (A/E_1) +P (E_2) P (A/E_2)

$= 0.0001 \times 0.98 + 0.9999(0.02)$

= 0.020096

$$P\ (E_1/A) = \frac{0.0001 \times 0.98}{0.020096} = 0.0049$$

3.2 Probability Distributions

There are two types of theoretical distribution which are shown in fig. 3.10

i. Discrete distributions
ii. Continuous distribution

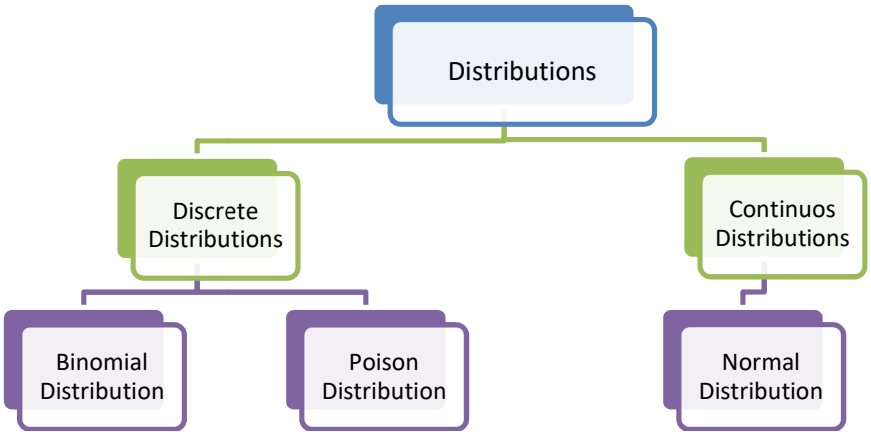


Fig 3.10 Types of Theoretical Distribution

Binomial Distribution

Binomial distribution was discovered by James Bernoulli (1654-1705) in the year published in 1713 eight years his death. A random experiment whose outcomes are of two types namely success S and failure F, with probabilities p and q is called a Bernoulli trial.

An experiment which contains repeated number of Bernoulli trials is called a binomial experiment and the distribution based on binomial experiment is called the binomial distribution.A discrete random variable X is said to follow binomial distribution if it has the probability mass function.

$$P(x=n)=P(n)=\begin{cases} {}^{n}C_{x}p^{x}q^{n-x}, x = 0,1,2,3,4 \ldots . n \\ 0, \quad otherwise \end{cases}$$

where, q= 1-p and n and p are called the parameters.

Constants of Binomial Distribution

Mean = np

Variance = npq

Standard Deviation $=\sqrt{npq}$

The binomial distribution can be used under the following conditions.

1) n, the number of trials is finite.
2) the trials are independent of each other.
3) p, probability of success is constant for each trial.
4) In each trial there are only two possible outcomes success or failure.

Note: In a binomial distribution Mean> Variance

Illustration No. 10

Verify the following statement: The mean of a binomial distribution 16 and its standard deviation is 5

Solution

Given mean =16

Standard deviation $\sigma = 5$

Variance σ^2= 25

So, mean < variance

$\therefore$ *Given statement is wrong.*

Illustration No. 11 :

9 coins are flipped simultaneously. Find the probability of getting 7 heads.

Solution

Here n=9, p =P (getting a head) =1/2

q = 1-p $= 1 - \frac{1}{2} = \frac{1}{2}$

$P(X=n) = {}^{n}c_{n}p^{x}q^{n-x}$, n=0,1,2,.....n

$= {}^{9}c_{x}(1/2)^{x}(1/2)^{9-x}$, x= 0,1,2.....9

$=9c_x (1/2)^9$

$=\frac{1}{512} 9c_n$, x=0,1,2,……9

P (getting at least 7 heads) = P (x≥7)

= P(x=7) + P(x=8) + P (x=9)

$=\frac{1}{512} 9c_7 + \frac{1}{512} 9c_8 + \frac{1}{512} 9c_9$

$=\frac{1}{512}[9c_7 + 9c_8 + 9c_9]$

we have $nc_r = nc_{n-r}$

$9c_7 = 9c_{9-7} = 9c_2$

$9c_8 = 9c_1$

$9c_9 = 9c_0$

$=\frac{1}{512}[9c_2 + 9c_1 + 9c_0] = \frac{1}{512}[36 + 0 + 1] = \frac{46}{512} = \frac{23}{256}$

Illustration No. 12

A pair of dice is thrown 5 times. If getting a doublet is considered as success, find the probability of getting 3 successes.

Solution

Here n=5

A—› doublet={(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)}

$p = \frac{6}{36} = \frac{1}{6}$

$q = 1 - P = 1 - \frac{1}{6} = \frac{5}{6}$

$P(X = n) = nC_X\, p^N q^{n-X}$, X=0,1,2,….5

$P(X = n) = 5C_X \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{5-X}$, $X = 0,1,2,\ldots 5$

P (3 success) = P (x=3)

$= 5C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2$

$= 10 \times \frac{1}{216} \times \frac{25}{36}$

$= \frac{125}{3,888}$

Illustration No. 13

Among 20 vegetables in a basket of 100 are rotten. If 10 vegetables are selected at random, find the probability that

i. 10 are rotten

ii. at most 3 are rotten

Solution

Here n=10

Let X be a random variable represent rotten vegetables

Given p = $\frac{20}{100}$= 1/5

$q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}$

P(x=n) = $nc_n p^x q^{n-x}$, n=0.1,2,.....n

=10 Cx $\left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{10-x}$,x=0,1,2,....10

(i) P (10 are rotten) = P (X =10)

$=10c_{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^{10-1}$

$=\left(\frac{1}{5}\right)^{10}$

(ii) P (at most 3 are rotten) =$P(x \leq 3)$

=P(x=0) +P (x=1) + P(x=2) + P (x=3)

$= 10c_0\left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} + 10c_1\left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + 10c_2\left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8 + 10c_3\left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$

= 0.879

To Do Activity

You will need to pair up. One person will have the set of 10 pop quizzes with answers as True or False. The other partner will guess the correct answers. For each test, keep track of correct responses. After your pop quizzes are over, now reverse roles. When completed with both partners completing the pop quizzes, tally your results with the other groups. Calculate the frequency of the event, getting exactly 4 questions correct, P(X = 4)

Poisson Distribution

In 1837 French mathematician Simeon Dennis Poisson derived the distribution as a limiting case of Binomial distribution. It is called after his name as Poisson distribution.

Poisson distribution is a limiting from of binomial distribution under the following conditions:

i. n, number of trials is indefinitely large (i.e.) $n \to \infty$
ii. p, probability of success is very small (i.e.) p—›0
iii. $np = \lambda$ is finite

Definition

A discrete random variable x is said to follow Poisson distribution if it has the probability mass function

$$P(X = n) = \begin{cases} \frac{e^{-\lambda}\lambda^n}{n!}, n = 1,2,3,4 \ldots . \\ 0, \text{otherwise} \end{cases}$$

Here λ is called the parameter.

Constants of Poisson Distribution

1. Mean = λ

2. Variance = λ
3. standard deviation $= \sqrt{\lambda}$

Note: In a Poisson distribution, n Mean = Variance

Illustration No. 14

Suppose the number x of tornados observed in a particular region during a 1 year period has a Poisson distributions with λ=8

i. Compute $P(X \leq 3)$
ii. Compute $P(2 \leq X \leq 5)$
iii. compute $P(3 \leq X)$
iv. How many tornados can be expected to be observed during the 1-year period? What is the variance of the number of observed tornados? (e^{-8}=0.0003)

Solution

Given, λ = 8

P(X = n) = $\frac{e^{-\lambda}\lambda^n}{n!}$, $n = 0,1,2,3,4 \ldots\ldots$

= $\frac{e^{-8}8^n}{n!}$, $n = 0,1,2,3,4 \ldots\ldots$

i. P(x 3) = P (x=0) + P (x=1) + P(x=2) + P(x=3)

$= \frac{e^{-8}8^0}{0!} + \frac{e^{-8}8^1}{1!} + \frac{e^{-8}8^2}{2!} + \frac{e^{-8}8^3}{3!}$

$= e^{-8}[1+8+32+\left(\frac{256}{3}\right)]$

= 0.0379

ii. P(2≤×≤5) = P(x=2) + P(x=3) + P(x=4) + P(x=5)

$= \frac{e^{-8}8^2}{2!} + \frac{e^{-8}8^3}{3!} + \frac{e^{-8}8^4}{4!} + \frac{e^{-8}8^5}{5!}$

= 0.1683

iii. P(3 x) = P(x≥3)

= 1- P(x<3)

=1- [P(x=0)+P(x=1)+P(x=2)]

$=1-\left\{\frac{e^{-8}8^0}{0!} + \frac{e^{-8}8^1}{1!} + \frac{e^{-8}8^2}{2!}\right\}$= 0.9877

iv. Number of expected torn = λ = 8

variance = λ= 8

Illustration No. 15

A land lord has two tractors. The demand for a tractor on each day is distributed as a Poisson variation with mean 1.5. Calculate the proportion of days on which

i. neither tractor is used

ii. the land lord refuses some demand.

Solution

Let x be a random variable represent the demand of taxi

Given, λ= 1.5

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0,1,2,3,4 \ldots\ldots$

$= \frac{e^{-1.5}1.5^x}{x!},\ x = 0,1,2,3,4 \ldots$

(i) P (neither car is used)=P(x=0)

$= \frac{e^{-1.5}1.5^0}{0!}$

= 0.2231

(ii) P (some demand is refused)

=P(x>2) =1- P(x 2)

= 1 - [P(x=0) + P(x=1) + P(x=2)]

$=1-\left\{\frac{e^{-1.5}1.5^0}{0!} + \frac{e^{-1.5}1.5^1}{1!} + \frac{e^{-1.5}1.5^2}{2!}\right\}$

=1- 0.1804

= 0.8196

Illustration No. 16

From the information from medical board 1 in 200 people affected by mouth cancer because of tobacco. In a sample of 100 individuals what is the approximate distribution of the number who affected by mouth cancer? Use this distribution to calculate the approximate probability that

i. between 5 and 8 (include)affected by mouth cancer

ii. at least 2 affected by mouth cancer.

Solution

Let, x be a random variable represent the number of individuals carry the gene

Given, $p = \frac{1}{200}$, n=number of individuals =1000

λ= np = 1000× $\frac{1}{200}$

λ= 5

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2,3,4 \ldots\ldots$

$= \frac{e^{-5}5^x}{x!},\ x = 0,1,2,3,4 \ldots\ldots$

i. P($5 \leq x \leq 8$) = P(x=5) + P(x = 6) + P(x =7) + P (x = 8)

$= \frac{e^{-5}5^5}{5!} + \frac{e^{-5}5^6}{6!} + \frac{e^{-5}5^7}{7!} + \frac{e^{-5}5^8}{8!}$

= 0.4886

ii. P (at least 3 carry the gene)

= P (x≥2) = 1-P (x < 2)

= 1- {P (x=0) + P(x=1) }

= 1-$\left(\frac{e^{-5}5^0}{0!}+\frac{e^{-5}5^1}{1!}\right)$

= 0.9598

To Do Activity

You notice that a news reporter says "uh" on average two times per broad cast. What is the probability that the news reporter says "uh" more than two times per broad cast.

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. In graph form, normal distribution will appear as a bell curve. (Fig 3.11). Normal distribution is one of the most widely used continuous distributions. It was introduced by Demoivre in 1733 in the development of probability. The normal distribution is one of the most common types of distribution utilized in technical stock market analysis and in other types of statistical analysis.

The normal distribution is a limiting form of binomial distribution and also it is called Gaussian distribution. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is a perfectly symmetrical distribution. Symmetrical distribution is one where a dividing the area bounded by a normal curve in to two equal parts.

Definition

A continuous random variable X is said to follow normal distribution with parameter μ and σ^2 if its probability density function is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},-\infty<x<\infty,\mu\,\sigma>0.$$

and, it is denoted by $X\sim N(\mu,\sigma^2)$

Note

Normal distribution is a limiting case of binomial distribution under the following conditions

i. n, number of trials is indefinitely large (i.e.) n —›

ii. neither p nor q is small.

Characteristics of Normal Distribution

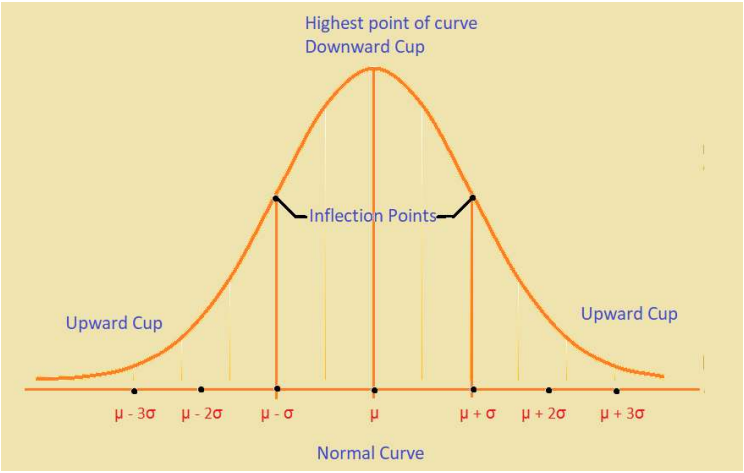


Fig 3.11 Normal curve

1. The curve is bell shaped and symmetrical aloud the lime x=μ.
2. Mean = Median = Mode.
3. X axis is an asymptote to the normal curve.
4. The points of inflection of the curve at $X = \mu \pm \sigma$
5. It is uni modal curve
6. Maximum coordinates at $X = \mu$ and it is $\frac{1}{\sigma\sqrt{2\pi}}$
7. The total area of the normal curve is unity.
8. $P(\mu - \sigma < x < \mu + \sigma) = 0.6826$
9. $P(\mu - 2\sigma < x < \mu + 2r) = 0.9544$
10. $P(\mu - 3\sigma < x < \mu + 3r) = 0.6673$

Standard Normal distribution

A normal distribution is said to be a standard normal distribution if it has mean μ=0 and variance $\sigma^2 = 1$

The probability density function of standard normal distribution is

$$\emptyset(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}, -\infty < Z < \infty$$

where, $z = \frac{x-\mu}{\sigma} = standard\ normal\ variare.$

and, it is denoted by $Z \sim N(0,1)$

Illustration No. 17

A normal random variable X having mean 50 and standard deviation 10. Calculate the following probability by standardizing.

i. Probability of X is less than or equal to 100
ii. Probability of X is less than or equal to 80
iii. Probability of X is lie between 65 and 100.
iv. P(|X-80|≤10)

Solution

Given mean μ=80, standard deviation σ =10 $Z = \frac{X-\mu}{\sigma} = \frac{X-80}{10}$

i. **P(X≤100) = P (Z ≤ 2)**

= P(-∞ < Z < 2)
= P (-∞< Z <0) + P (0 < Z < 2)
= 0.5 + 0.4772 (from the table) = 0.9772

ii. **P(X ≤ 80) = P (Z ≤ 0)**

= 0.5

iii. **P (65 ≤ X ≤ 100)**

= P (-1.5 ≤ X ≤ 2)

= P (-1.5 ≤ Z ≤ 0) + P (0 ≤ Z ≤ 2)
=0.4332 + 0.4772 = 0.9104

iv. **P(|X-80|≤10)**

= P(-10 ≤ X-80 ≤10)

= P (70 ≤ X ≤ 90)
= P (-1 ≤ Z ≤ 1)
= 2P (0 ≤ Z ≤1)
= 2 (0.3413) = 0.6826.

Illustration No. 19

The remuneration paid for 100 laborers in a poultry firm are normally distributed with mean ₹ 700 and standard deviation of ₹ 50. Estimate the number of laborers whose remuneration will be

I. P(700 < X < 720)
II. P (X > 750)

Solution

Let X be a random variable represent the remuneration
Given mean μ=700
Standard deviation $\sigma = 50$

$$Z = \frac{X-\mu}{\sigma} = \frac{X-700}{50}$$

(i) P(700 < X < 720) = P (0 < Z < 0 .4)

= 0.1554

Number of labourers whose remuneration is between 700 and 720

=100× 0.1554=15.54

≈16 laborers

(i) P(x >750) = P (Z > 1)

=P (1< Z <)

= P (0 < Z <) - P (0 < Z <)
= 0.5 - 0.3413
= 0.1587

Number of laborers whose remuneration is more than $\begin{cases} = 100 \times 0.1587 \\ = 15.87 \\ \approx 16\ Labourers. \end{cases}$

3.3 Estimation and its types

Estimation, in statistics any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population. The 18th-century English theologian and mathematician Thomas-Bayes was instrumental in the development of Bayesian estimation to facilitate revision of estimates on the basis of further information.

Definition

The method of obtaining the most likely value of the population parameter using statistic called estimation.

Estimator

Any sample statistic which is used to estimate an unknown population parameter is called an estimator.

Estimate

When we observe a specific numerical value of our estimate, we call that value is an estimate. i.e. an estimate is a specific observed value of a statistic. (i.e.) an estimate is a specific observed value of a statistic.

3.4 Confidence Interval

A confidence interval is a type of estimate computed from the Statistic of the observed data this process a range of plausible Values for an unknown parameter. The interval has an associated Confidence level that the true parameter is in the proposed range.

Confidence Interval for Population Mean for Large Samples

If we take repeated independent random sample of size n from a population with an unknown mean μ but known standard deviation σ, then the probability that the true population mean μ will fall in the interval is 1-α.

So, the confidence interval for population mean (μ) is given by

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

3.5 Point and Interval Estimation

There are two types of estimation (shown in fig 3.13)

i) Point estimation
ii) Interval estimation

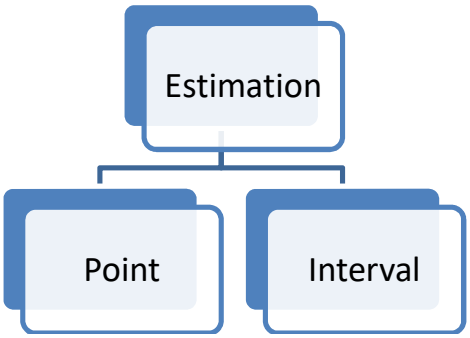


Fig 3.12 Types of Estimation

Point Estimation

1. When a only one specified value is used as an estimate the estimate is called a Point estimate of the population parameter i.e. An estimate of a population parameter given by a single value is called as point Estimation
2. The sample mean μ is the sample statistic used as an estimate Population mean.

Interval Estimation

In statistic interval estimation is the use of sample data to calculate an interval of possible values of an unknown population Parameter.

Normal Probability Table

Critical value Z_α	Level of significance(α)			
	1%	2%	5%	10%
Two tailed test	$\|z_\alpha\| = 2.58$	$\|z_\alpha\| = 2.33$	$\|z_\alpha\| = 1.96$	$\|z_\alpha\| = 1.645$
Right tailed test	$\|z_\alpha\| = 2.33$	$\|z_\alpha\| = 2.055$	$\|z_\alpha\| = 1.645$	$\|z_\alpha\| = 1.28$
Left tailed test	$\|z_\alpha\| = -2.33$	$\|z_\alpha\| = -2.055$	$\|z_\alpha\| = -1.645$	$\|z_\alpha\| = -1.28$

Illustration No. 19

A small part of 100 measurements at breaking strength of copper wire gave a mean of 7.4 and a standard deviation of 1.2gms Find 95% Confidence limits for the average breaking strength of copper wire.

Solution

Given Sample size n= 100 $\bar{x} = 7.4$ $\sigma = s = 1.2$ $Z_{\frac{\alpha}{2}} = 1.96$

$$S.E = \frac{\sigma}{\sqrt{n}} = \frac{S}{\sqrt{n}} = \frac{1.2}{\sqrt{100}} = 0.12$$

95% confidence limits for the population mean are

$$\left(\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\frac{\sigma}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

$$= (7.4 - (1.96 \times 0.12) < \mu < 7.4 + (1.96 \times 0.12))$$

$$= (7.165 < \mu \leq 7.635)$$

Illustration No. 20

Find the 95% confidence limits for population mean if $n = 31$, $\overline{X} = 80.0$, $\sigma = 2.0$

Solution

Confidence Interval $= \left(\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\frac{\sigma}{2}}\frac{\sigma}{\sqrt{n}}\right)$

$= \overline{X} \pm 1.96.\frac{\sigma}{\sqrt{n}} = 80 \pm 1.96\left(\frac{2}{\sqrt{31}}\right) = 80 \pm 0.7$

$= (79.3, 80.7)$

Illustration No. 21

A Service Channel monitored for an hour was found to have an estimated mean of 20 transitions transmitted per minute. The variance Is known to be 4. Find the point estimation.

Solution

Given $\sigma^2 = 4 \rightarrow \sigma = 2$

N = 1 hour = 60mins $\overline{X} = 20/min$

Point estimation = $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{60}}$

= 0.2582

Illustration No. 21

A die Is tossed 900 times and a toss of 4 or 5 Is observed 3240 times. Find the point estimation of the population for an unbiased die.

Solution

Here the occurrence of 4 or 5 on the die is called a success, then

Sample size =9000.

Number of success =3240.

Sample proportion p = $\frac{3240}{9000}$ = 0.36

Population proportion P = P (getting 4 or 5 when a die is thrown)

$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3} = 0.33$

P=0.33 and Q = 1-P = 0.67

∴ The point estimation for the sample proportion

$= \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.33 \times 0.67}{9000}}$

= 0.00496

Summary

This chapter provides an explanation of probability for processes with a finite number of possible outcomes. It explains the meaning of probability, as well as how to calculate probability and odds. It also examines the relationship between complementary events. This chapter makes the reader to understand that random variables in probability have a defined domain and can be continuous or discrete. Probability distributions summarize the relationship between possible values and their probability for a random variable. Probability density or mass functions map values to probabilities and cumulative distribution functions map outcomes less than or equal to a value to a probability. In statistics, estimation is a data analysis framework that uses a combination of effect sizes, confidence intervals, precision planning, and meta-analysis to plan experiments, analyze data and interpret results. A thorough explanation of point and interval estimation is discussed. Four important steps to understand interval estimation are explained.

Multiple Choice Questions

1. Two events A and B are mutually exclusive if

a) $P(A \cap B)=1$

b) b) $P(AUB)=0$

c) c) $P(A \cap B)=0$

d) d) $P(AUB)=1$

2. Let S be the sample space of an experiment, and $S=\{A_1,A_2,\text{-----}A_n\}$ then $\sum_{i=1}^{n} P(A) =$

a) 1

b) 0

c) $\frac{1}{2}$

d) $\frac{1}{3}$

3. Probability of an impossible event is

a) 1

b) 0.2

c) 0.5

d) 0

4. A letter is taken at random from the letters of the word "PROBABILITY" The probability that the with is a vowel is

a) $\frac{2}{11}$ b) $\frac{3}{11}$ c) $\frac{4}{11}$ d) $\frac{4}{11}$

5. If E and F and any two events then the probability that exactly one of them occurs is

a) $P(E \cap \bar{F}) + P(\bar{E} \cap F)$

b) $P(E \cup \bar{F}) + P(\bar{E} \cup F)$

c) $P(E) + P(F) - P(E \cap F)$

d) $P(E) + P(F) + 2\,P(E \cap F)$

6. In tossing a coin, until head appear, the sample space is

a) a null Set

b) a Countable finite Set

c) a count ably infinite Set

d) an uncountable Set

7. Probability of not getting 4 when die is thrown

a) $\frac{1}{3}$ b) $\frac{5}{6}$ c) $\frac{1}{6}$ d) $\frac{1}{4}$

8. If A and B are two events with P (A/B) = 0.3, P(B/A) = 0.2 then P(B) is

a) $\frac{1}{10}$ b) $\frac{2}{10}$ c) $\frac{7}{10}$ d) $\frac{3}{10}$

9. If A and B be two events such that P(A/B) = 1/2, P(B/A) = 1/3 and P(A$\cap$B) = 1/6 then P(AUB) is

a) $\frac{1}{3}$ b) $\frac{2}{5}$ c) $\frac{1}{6}$ d) $\frac{2}{3}$

10. If A and B are independent events such that P(A) = 0.25, P(AUB)= 0.75 then P(B) is

a) $\frac{4}{13}$ b) $\frac{1}{13}$ c) $\frac{2}{3}$ d) $\frac{4}{3}$

11. If P(A) = 0 then A is called

a) sure event

b) impossible event

c) possible event

d) mutually exclusive event

12. If A and B two events with P (AUB) = 2/3 then P $(\bar{A}\cap\bar{B})$ =

a) $\frac{2}{3}$ b) 0 c)1 d)$\frac{1}{3}$

13. The probability that a non- leap year selected at random will contain 53 Mondays is

a) $\frac{2}{7}$ b) $\frac{1}{7}$ c) $\frac{3}{7}$ d) $\frac{5}{7}$

14. Two events A and B independent then P(A/B) =

a) P(A) b) P(B) c) P(A$\cap$B) d) P(AUB)

15. P $(\bar{A})$ = ---------

a) 1-P(A) b) P(A)-1 c) P(A) d) 1

16. The mean and variance of binomial distribution one

a) npq, np b) np, npq c) np, $\sqrt{npq}$ d) $\sqrt{npq}$, np

17. An experiment succeeds twice as often as its fails. The chance that in the next six trials there shall be at least four success is

a) 489/729 b) 496/729 c) 240/729d) 251/729

18. If in 6 trials; x is a binomial variation which follows the relation 9P(X=4)=P(X=2) then the probability of success is

0.25 b) 0.375 c) 0.75 d) 0.125

19. Poison distribution is applied for

a) repeated two alternatives b) impossible events c) rare events d) certain events

20. In a Poisson distribution

a) mean= variance b) mean< variance c) mean> variance d) mean= 1/variance

21. If the mean of Poisson variant is 2 then P(X<1) is

$a)\ e^2$b) $1-e^{-2}$ c)$1-\frac{5}{2e^2}$d) $1-\frac{5}{2e^2}$

22. The normal distribution is a limiting from of binomial distribution if

a) n → ∞ and P→0 b) n → ∞ and P → n

c) n → ∞ and neither p nor q Is Small d) n → ∞ and p → q

23. If X ~N (5,36) the standard normal variate Z is

a) Z = $\frac{x-36}{5}$ $b)\ z=\frac{z-5}{36}$ $c)\ z=\frac{x-5}{36}$ $d)\ z=\frac{z-36}{25}$

24. The parameters of the normal distribution$f(x)=\frac{1}{\sqrt{72\pi}}e^{-\frac{(x-10)^2}{72}}, -\infty < x < \infty$

a) (10,6) b) (10,36) c) (6,10) d) (36,10)

25. Using the standard normal table, the sum of the probability to the right of Z= 2.18 and the left of Z= -1.75 is

a) 0.4854 b) 0.4599 c) 0.0146 d) 0.0547

Answers for MCQ

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
c	a	d	c	a	c	b	b	d	c	b	d	b	a	a
16	17	18	19	20	21	22	23	24	25					
b	b	a	c	a	b	c	c	b	d					

Problems

1) A number is chosen at random from the numbers 1 to 30 find the probability that it is

(i)a prime number (ii) multiple of 2 $Ans\ \ i)\frac{1}{3}\ ii)\frac{1}{2}$

2) In a village among the 260 farmers, 90 cultivated brinjal, 120 cultivated onion and 50 cultivated both onion and brinjal find the probability that a farmer cultivated onion or brinjal.

   i. Either onion or brinjal
   ii. Neither of the two varieties
   iii. Brinjal only
   iv. Onion only
   v. Exactly one of the varieties $Ans\ i)\frac{8}{13}\ ii)\frac{5}{13}\ iii)\frac{7}{26}\ \ iv)\frac{2}{13}\ \ v)\frac{11}{26}$

3) Define:

   a) Event b) Random Experiment c) Sample Space d) Mutually Exclusive Events
   e) Independent Events

4) State Axioms of Probability.

5) State Bayes' Theorem.

6) A purse contains 5 two rupee and 7 five-rupee coins. Another purse contains 6 two rupee and 9 five-rupee coin. if a coin drawn from any one of two purses, find the probability that the coin is 5-rupee coin $Ans:\ \frac{71}{120}$

7) If $P(A_1) = 0.35$, $P(A_2) = 0.73$ and $P(A_1 \cap A_2) = 0.14$ find $P(\bar{A}_1 U \bar{A}_2)$ $Ans: 0.86$

8) When the pain of dice is rolled, what is the probability of getting the sum

   6
   6 or 8
   6 or 11 $Ans\ i)\ \frac{5}{36}\ \ ii)\frac{5}{38}\ iii)\ \frac{7}{36}$

9) What is the probability that,

   a leap year
   a non -leap year should have 53 Mondays

   $Ans\ i)\frac{2}{7}\ ii)\ \frac{1}{7}$

10) The probability of an event A occurring is 0.6 and B occurring is 0.3, A and B are mutually exclusive events then find the probability of

    P(AUB)

    $P(A \cap \bar{B})$

    $P(\bar{A} \cap B) Ans\ i)\ 0.9\ ii)\ 0.6$

11) If A and B are independent events then prove that

    A and $\bar{B}$ are independent
    $\bar{A}$ and B are independent

    $\bar{A}$ and $\bar{B}$ are independent

$\bar{A}$ and $\bar{B}$ are independent

12) There are 4000 people living in a village including 1500 males. Among the people in the village, the age of 1500 people is above 25 years including 400 males. Suppose a person is chosen at random and you are told that the chosen person is a male, what is the probability that his age is above 25 years?

$Ans: \frac{32}{45}$

13) Given three identical bags B_1, B_2, B_3 each containing two fruits. In bag B_1, B_2 both the fruits are apple, in bag B_3 there is an apple and an orange. A person picks a bag at random at takes out a fruit. If the fruit is apple, what is the probability that the other fruit in the bag is also an apple?

$Ans: \frac{2}{3}$

14) A farmer has two cows C_1 and C_2. C_1 give 40% of daily consumption of milk and C_2 give 60% of daily consumption. Further 4% of the milk given by C_1 and 5% of the milk given by C_2 goes to calf. Milk is drowning at random if the draw milk is calf milk, Find the probability it was given by C_2. $Ans: \frac{15}{23}$

15) A product development officer at a car engine factory needs to estimate the average life time of a Car engine made at the factory. The standard deviation of lifetimes is known to be 100 hours. A random sample of 64 engines from the factory results in a sample mean lifetime of $\overline{X}$ = 350 hours.
(a) Find a 95% confidence interval for the mean lifetime for the complete products in a factory.
(b) Suppose that the standard deviation was 80 alternative that 100 hours.
Ans: (a) (325.5, 374.5)(b) (330.4, 369.6)

16) In a countrywide random sample of 250 BBA students who are graduating from college this semester, we observe that 80 have not found a job. Find a 95% confidence interval for the proportion of all graduating BBA students who have not found a job.
Ans: (0.2622, 0.3778)

17) 12 owners of Maruthi petrol pickups are randomly selected and asked to report their highway fuel mileage. The resulting data have a sample average of $\overline{X}$ = 17.2 mpg with a standard deviation of s = 1.4 mpp. Assuming that national highway mileage values follow a normal distribution, find a 90% confidence interval for the mean highway fuel mileage for all Maruthi petrol pickups
Ans: (16.474, 17.926)

18) Determine the standard error of sample proportion for a random sample of 500 vegetables was taken from a large consignment and 65 were found to be bad.
Ans: 0.015

19) The mean life time of a sample of 169 water heater manufactured by a company is found to be 1350 hours with a standard deviation of 100 hours. Establish 90% confidence limits within which the mean life time of water heater is expected to lie.
Ans: (1337.35, 1362.65)

20) Two students A and B are playing tennis, in which their chance of winning is in their ratio 3:2. Find

A's chance of winning at least three rounds out of five rounds are played.
Ans: 0.6826

21) Fair coin is tossed 6times Find the probability that exactly 2 tail turns up.
Ans : 15/64

22) Verify the following statement. The mean of a Binomial distribution is 16 and its standard deviation is 5
Ans: Given statement is wrong, since mean is less than variance.
23) If the average rain falls on 9 days in every thirty days find the probability thatRain will fall on atleast two days of given week.

Ans : 0.6706

24) Out of 750 farmers with 4 children each How many farmers would be expected to have
i. at least one boy
ii. at most 2 girls
iii. and children of both sexes? Assume equal probability for boys and girls
Ans : 703,516,656

25) A vegetable mart having 260 Kg of onions ,195kg of good quality. Assuming Poisson law for the number of good qualities per Kg, Find the probability that a random sample of 5 Kilograms of onion will contain good quality.
Ans : $e^{-3.75}$

References

1. Sharma, J.K. (2014). *Business Statistics – Problems and Solutions*. New Delhi : Vikas Publishing House Pvt Ltd.
2. Pillai, R.S.N. &Bagavathi, V. (1999). *Statistics*. New Delhi :S.Chand& Company Ltd.
3. Gupta, S.P. (2010). *Statistical Methods.* New Delhi :S.Chand& Company Ltd.
4. Beri, G.C. (2011). *Business Statistics.* New Delhi : Tata McGraw Hill Educations Pvt Ltd.
5. Foster, D. & Stine, E.R. (2010). *Statistics for Business : Decision Making and Analysis.* New Delhi : Pearson Publishers.
6. Gupta, S.C. & Kapoor, V.K. (2006). Fundamentals of Mathematical *Statistics.* New Delhi :S.Chand& Company Ltd.
7. Srivastava, S.C & Srivastava, S. (2003). Fundamentals of Statistics. New Delhi : Anmol Publications Pvt. Ltd.

Chapter 4 Sampling Theory and Tests of Significance

Introduction

This chapter is to teach you the concepts of sampling theory. The importance of sampling theory lies on the fact that it is neither easy nor necessary to collect data from each and every member of the population, when the population size is very large. For example, a grain merchant decides the quality of grain of whose assignment he is willing to purchase based on the examination of quality of a handful of grain. This sampling theory finds more useful for destructive items like bulbs, crackers, eggs, etc. Similarly this concept is more effective to determine parameters of any large population, like economy, living standard, purchasing power of a large group of people. Sampling theory gives the relationship between a population and a sample drawn from it. A sample investigation gives some results with which decisions are made on the population. But such decisions involve an element of uncertainty causing wrong decisions. Hypothesis is an assumption which may or may not be true about a population parameter.

Objectives

- To understand the concept of testing of null and alternative hypothesis
- To discuss the various tests of significance
- To test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution
- To be able to carry out the parametric and non-parametric tests

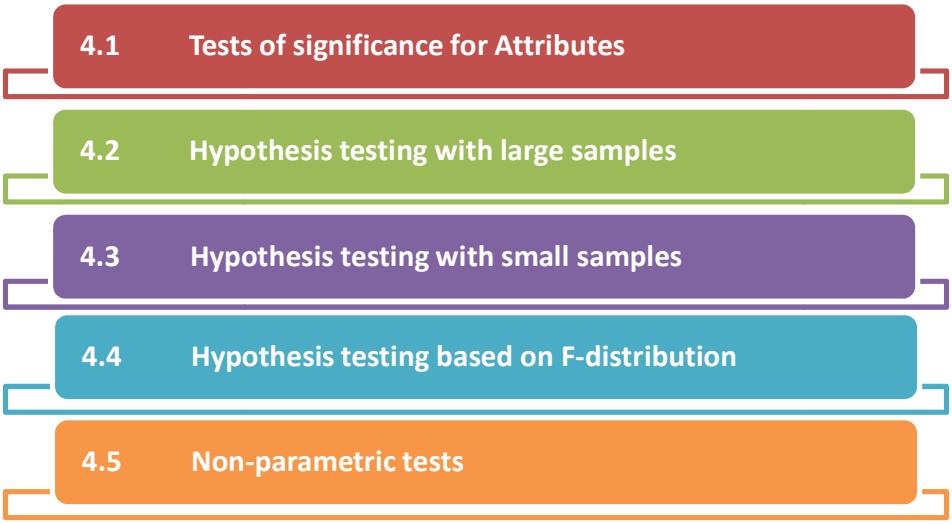


Fig 4.1 Chapter Flow

4.1 Tests of Significance for Attributes

Before getting into the concept of testing of significance for attributes or variables, let us learn the elements of testing of significance shown in Fig. 4.2

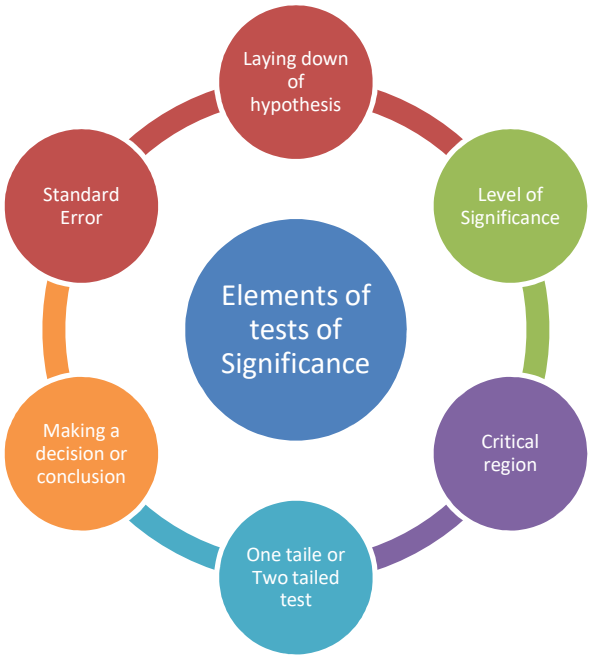


Fig. 4.2 Elements of Tests of Significance

a) Laying Down of Hypothesis

Any hypothesis concerning a population is called a statistical hypothesis. There are two hypotheses to test the significance of difference between the sample value and the population. They are :

i. Null Hypothesis: If the statement reveals that the value of sample and the value of population under study do not show any difference, then the hypothesis is said to be null hypothesis. In simple words, null hypothesis implies that the sample is representing the character of whole population. A statistical hypothesis is a null hypothesis if it is accepted. It is denoted by H_0.

Example of H_0 – Average monthly income of a particular village is Rs.5000

ii. Alternative Hypothesis : Rejection of H_0 leads to the acceptance of alternative hypothesis, which is denoted by H_1

Example of H_1 – Average monthly income of a particular village is not equal to Rs.5000

When there are two hypotheses set-up, the acceptance or rejection of a null hypothesis is based on the outcome of the sample study. Thus, it may lead to two wrong conclusions similar to two right conclusions. These two wrong conclusions are termed as two types of errors in sampling theory. They are given in the following table 4.1

Table 4.1 : Type I & II errors

Status of Hypothesis	Decision based on sample study	
	Accept H_0	Reject H_0
H_0 true	Correct	Wrong (Type I error)
H_0 false (i.e. H_1 true)	Wrong (Type II error)	Correct

Type I error : Rejecting Null Hypothesis when it is true
Type II error : Accepting Null Hypothesis when it is false

b) Level of Significance

It is the maximum probability of committing type I error in a test. It is generally fixed at 5% in statistical tests. 5% level of significance implies that we have 95% of confidence in accepting a null hypothesis.

c) Critical Region

The range of variation has two regions namely, acceptance region and rejection/critical region. If the sample statistics falls in rejection region, then we have to reject the null hypothesis.

d) One-Tailed and Two-Tailed Tests

The critical region under a normal curve can be stated in two ways, namely a) two sides under a normal curve b) one side under a normal curve, i.e both are either at the right tail or at the left tail of the curve.

e) Making a Decision or Conclusion

The decision should be on the basis of computed value if it lies in the acceptance region or rejection region. When the computed value is less than the rejection value, the null hypothesis is accepted otherwise not. The table values for acceptance region or rejection region in a normal distribution curve are given in table 4.2.

Procedure to Calculate Z-Table Value from Normal Probabilities Table (table 4.2)

For example, if we are interested to find Z-table value for 5% level of significance, the following procedure is to be followed.

Step 1 : Find confidence level by subtracting level of significance (in fraction) from the value of 1. For 0.05 level of significance, the confidence level is 0.95

Step 2 : Divide the confidence level by 2 (if we use single tailed values). For 0.95 confidence level, take confidence level of 0.475 for single tailed values

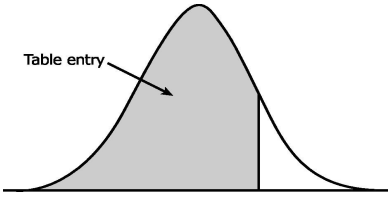
Step 3 : Add the value determined with 0.5 if the values are from the area of '0' (shaded from one tail end). When we use the table with the values from median i.e from mid of the curve, we need not add the value with 0.5. Table 4.2 is with values from tail end and hence 0.5 is to be added. Hence 0.475 is added with 0.5 to get 0.975

Step 4 : Lookup the value (0.975) determined in step 3 in Z- table. The Z-value that has an area of 0.975 is 1.96

Similarly for 1% significance level, the procedure follows :

1. Confidence level = 99% = 0.99
2. For one tailed test, it is 0.495
3. For table from '0' area, it is 0.995
4. Z- value for 0.995 area is 2.575 (i.e mid of 2.57 and 2.58 in the table)

Table 4.2 Table of Standard Normal Probabiliti



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

For quick reference, the Z-table values are given in table 4.3 for a few important confidence levels.

Table 4.3 Z-values for a few important confidence levels

Z - value	2.575	2.33	2.05	1.96	1.75	1.645	1.44	1.28	1.15	1.04
Confidence level (in fraction)	0.99	0.98	0.96	0.95	0.92	0.90	0.85	0.80	0.75	0.70
Significance level (in %)	1	2	4	5	8	10	15	20	25	30

f) Standard Error (S.E)

The standard deviation of the sampling distribution is called standard error. For example, when 20 samples are taken from a population and if X_1, X_2, X_3,.... X_{20} represent their mean values, then the mean

of these mean values would be close to the mean of the population and the standard deviation of these values will be called standard error.

Tests of Significance for Attributes

It is the process of drawing a sample from a population whose members consist of presence or absence of a particular characteristic. For example, in the study of literacy, a sample may be drawn and its members are classified as literate and illiterate. The presence of attribute, literacy may be represented by p and the absence of attribute may be represented by q. The various tests of significance are studied under three following heads,

1) Test for number of successes (for which S.E = $\sqrt{n\ p\ q}$)
2) Test for proportion of successes (for which S.E = $\sqrt{\frac{p\ q}{n}}$)
3) Test for difference between proportions (for which S.E = $\sqrt{pq\ (\frac{1}{n_1}+\frac{1}{n_2})}$

Let us learn the tests of significance in the following illustrations.

Category 1 : Test for Number of Successes

Illustration No. 1

Two coins are tossed at a time for 100 times and they turn up head 110 times. Discuss if the coin is unbiased one at 1% level of significance.

Solution

Null Hypothesis: Tossing up of coin is unbiased

Probability of an unbiased coin turns up head, p = Total no. of heads/Total no. of sides = ½

Hence q = 1-p = ½ & n = 2 x 100 = 200

Expected number of heads in tosses of 2 coins 100 times = 200 x½ (for unbiased condition)

= 100 (hypothesised value)

But the observed number of heads = 110 (sample statistic)

Standard Error, S.E = $\sqrt{n\ p\ q}$ = $\sqrt{200\ x\frac{1}{2}x\frac{1}{2}}$

= 7.07

Deviation from actual = Value of sample statistic – Value of hypothesised population parameter

= 110 – 100 = 10

Z (observed/test statistic) = Deviation / S.E

= 10/7.07 = 1.41

Z (critical/rejection/table value) at 1 % level of significance = 2.575 (obtained from normal table)

Since observed value is less than critical value, the null hypothesis is accepted.

Hence tossing of coin is not biased.

Illustration No. 2

In 500 throws of a six faced dice, odd points are obtained 300 times. Check whether the throw of dice is fair at 5% level of significance.

Solution

Let us take the null hypothesis that the throw of dice is fair.

In a fair throw of dice, we expect a minimum of 250 odd points in 500 throws

Hence, p = ½ and q = ½ and n=500

$$S.E = \sqrt{n\,p\,q} = \sqrt{500\,x\,\frac{1}{2}x\,\frac{1}{2}}$$

$$= 11.18$$

Deviation from actual = Value of sample statistic – Value of hypothesised population parameter

= 300 - 250 = 50

Test statistic, Z = Deviation / S.E

= 50 / 11.18

= 4.47

Z (critical/rejection/table value) at 5 % level of significance = 1.96 (obtained from normal table)

Since the observed value is more than the 1.96 i.e critical value at 5% level of significance, the null hypothesis of the throw of dice is fair, is rejected. Hence the throw of dice is not fair.

To Do Activity

Develop a hypothesis to test the number of vegetarians and non-vegetarians in your class testing a few sample of your classmates.

Category 2 : Test for Proportion of Successes

Illustration No.3

A vegetables-wholesaler claims that only 2 % of the vegetables supplied by him are defective. A random sample of 100 vegetables is tested and 3 vegetables are found to be defective. Test the claim of the wholesaler at 5% level of significance.

Solution

Hypothesis (claim of wholesaler) = Only 2 % vegetables are defective

Given, q (no. of defectives) = 3/100 = 0.03

Hence, p = 1-q = 1-0.03 = 0.97

$$S.E = \sqrt{\frac{p\,q}{n}}$$

$$= \sqrt{\frac{0.97\,x\,0.03}{100}}$$

$$= 0.017$$

At 95% confidence limit (i.e. 5% level of significance) = p ± 1.96 S.E

= 0.97 ± 1.96 (0.017)

= 0.93668 to 1.00332

i.e. Out of 100 vegetables, good vegetables are found between 94 and 100

0.93668 x 100 = 93.668 = 94 (rounded off)and

1.00332 x 100 = 100.332 = 100 (cannot be more than 100)

Therefore the number of defective vegetables is expected between 0 and 6 out of 100 vegetables i.e 0% to 6%

Hence the claim of wholesaler of only 2% of defective vegetables is not accepted. It can be more than

2% also, some times upto 6%.

Illustration No. 4

A sample size of 600 persons was selected at random from a village shows that the percentage of males in the sample is 53. It is believed that the ratio of males to total population of the village is 50 %. Test if the belief is confirmed by the observation.

Solution

Hypothesis (Belief) = The ratio of males to total population = 50% = 0.5

= No. of male persons in the total population will be 300 (i.e. 0.5 x 600)

Given, p = 53/600 = 0.53

Hence q = 1 – p = 1 – 0.53 = 0.47

$$S.E = \sqrt{\frac{p\,q}{n}}$$

$$= \sqrt{\frac{0.53\ x\ 0.47}{600}} \quad = 0.0204$$

At 95% confidence limit (i.e. 5% level of significance) = p ± 1.96 S.E

= 0.5 ± 1.96 (0.0204)

= 0.460016 to 0.539984

i.e. Out of 600 persons, male persons are found between 94 and 100

0.460016 x 600 = 276 and

0.539984 x 600 = 324

Therefore the number of male persons is expected between 276 and 324 out of 600

Hence the belief of 300 male persons in the population is accepted as the value lies within the expected limits.

To Do Activity

Develop a hypothesis to test if the pass percentage of the class in a particular subject is more than 75% in first semester exams, taking a sample values of your friends in the class.

Category 3 : Test for Difference between Proportions

Illustration No.5

In a village A, out of a random sample of 1000 persons, 90 are found to be females while in another village B, out of a random sample of 1200 persons, 110 are found to be females. Do you find any difference between the two villages in the proportion of females at 5% level of significance?

Solution

Null Hypothesis: There is no difference between the proportions of females in two villages

p_1 = 90/1000 = 0.09

p_2 = 110/1200 = 0.0917

$$p = \frac{p1}{n1} + \frac{p2}{n2} = \frac{90}{1000} + \frac{110}{1200} \text{ (or } p_1 + p_2)$$

= 0.1817 i.e. 18.17 % females

q = 1 – p

= 1 – 0.1817 = 0.8183 i.e 81.83 % males or others

S.E = $\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

= $\sqrt{0.1817\ x\ 0.8183\left(\frac{1}{1000}+\frac{1}{1200}\right)}$

= 0.0165

Test statistic, Z = Difference between proportions / S.E

= (0.0917 – 0.09)/ 0.0165 = 0.103

Since observed value of Z is less than the critic value of 1.96 at 5% level of significance, the null hypothesis is accepted.

That is, there is no difference between the proportions of females in both the two villages.

Illustration No. 6

500 sample mangoes are inspected for quality from a mango farm – A and 12 are found to be not good. Similarly 800 sample mangoes are inspected for quality from another mango farm – B and again 13 are found to be not good. Can it be concluded that at 5% level of significance, the quality of mangoes in farm-B is better?

Solution

Null hypothesis : There is no difference in quality of mangoes in both farms

p_1 = 12/500 = 0.024

p_2 = 13/800 = 0.01625

p = p_1 + p_2

= 0.04025 i.e 4.025 % defective mangoes

q = 1 – p

= 1 – 0.04025 = 0.95975 i.e 95.98 % quality mangoes

S.E = $\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

= $\sqrt{0.04025\ x\ 0.95975\left(\frac{1}{500}+\frac{1}{800}\right)}$

= 0.0112

Test statistic, Z = Difference between proportions / S.E

= (0.024 – 0.01625)/ 0.0112 = 0.692

Since observed value of Z is less than the critic value of 1.96 at 5% level of significance, the null hypothesis is accepted.

That is, there is no difference between the qualities of mangoes in both the two farms.

To Do Activity

Develop a hypothesis to test if the quality vegetables produced in two farms is same or not, taking a sample values of quality of vegetables from two farms.

4.2 Hypothesis Testing with Large Samples

Though it is difficult to differentiate large sample from small sample, the sample with sample size greater than 30 is regarded as large sample. Since assumption we make for the two samples are not the same, the test of significance for these two samples also will be different.

The assumptions made for large samples are:

i) The random sampling distribution of statistics is approximately normal

ii) Sampling values are sufficiently close to the population value and can be used for the calculation of standard error of estimate

The following formulae are used to determine the standard error:

i) When standard deviation of the population is known,

$$S.E = \frac{\sigma\ of\ population\ x\ p}{\sqrt{n}}$$

ii) When standard deviation of the population is not known,

$$S.E = \frac{\sigma\ of\ sample}{\sqrt{n}}$$

iii) When it is required to test the difference between means of two samples,

$S.E = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$ or $\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}$ when both the samples are drawn from same population

iv) When it is required to test the difference between standard deviations of two samples,

$S.E = \sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}$ or $\sqrt{\frac{\sigma^2}{2}(\frac{1}{n_1} + \frac{1}{n_2})}$ when both the samples are drawn from same population and combined σ is known

Let us learn the hypothesis testing with large samples in detail through the following illustrations.

Illustration No. 7

A sample of 100 students in MGNCRE was taken and their mean age of students was found to be 21 years with a standard deviation of 2 years. Could the mean age of students in the whole Institution be 24?

Solution

Hypothesis : There is no difference between mean age of students in sample and population

$$S.E = \frac{\sigma\ of\ sample}{\sqrt{n}}$$

$$= \frac{2}{\sqrt{100}} = 0.2$$

Test Statistic, Z = Difference / S.E

= (24-21)/0.2 = 15

Since observed value of Z is more than the critic value of 1.96 at 5% assumed level of significance, the null hypothesis is rejected.

That is, there is a significance difference between the mean age of students in sample and population.

Illustration No. 8

In the exams conducted on a common subject for two classes A and B consisting of 40 & 50 students, the average mark and standard deviation was 74 and 8 for class A and 78 and 7 for class B respectively. Check if there is any significance difference between the performances of the two classes at a level of significance of 1%.

Solution

Null Hypothesis : There is no difference in the performance of two classes

$$S.E = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

$$= \sqrt{\frac{8^2}{40} + \frac{7^2}{50}}$$

$$= 1.6062$$

Test Statistic, Z = Difference in averages / S.E

= (78 - 74)/1.6062 = 2.4903

Since observed value of Z is less than the critic value of 2.575 at 1 % level of significance, the null hypothesis is accepted.

Illustration No. 9

The mean production of rice from a sample of 100 fields is 100 kg per acre with a standard deviation of 5 kg. Another sample of 150 fields gives the mean production at 110 kg per acre with a standard deviation of 6 kg. The standard deviation of all fields is assumed to be 5.5 kg. Check if the two sample results are consistent at 99% of confidence level.

Solution

Null Hypothesis : The production of rice in both the samples is consistent

Note : When both mean and standard deviation data are available for both the samples, it is advisable to test the differences between standard deviations of the samples.

$$S.E = \sqrt{\frac{\sigma^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{\frac{5.5^2}{2}\left(\frac{1}{100} + \frac{1}{150}\right)}$$

$$= 0.502$$

Test Statistic, Z = Difference in standard deviations / S.E

= (6 - 5)/0.502

= 1.992

Since observed value of Z is less than the critic value of 2.575 at 1 % level of significance, the null hypothesis is accepted.

To Do Activity

Take a sample of 50 students in any class of MGNCRE and predict the mean height of students of entire institution, by finding out the mean height and standard deviation of the selected 50 students and carrying out Z-test at 99% confidence level.

4.3 Hypothesis Testing with Small Samples

When the sample size is less than 30, the sample is regarded as a small sample. When the sample size is small and the population standard deviation is known, we can use t- distribution. The methods and rules of small samples are applicable to large samples whereas the reverse wise is not applicable. The formula for t-distribution is given by

i) When one sample is taken from a population

$t = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$ & σ = $\sqrt{\frac{\Sigma(X-\bar{X})^2}{(n-1)}}$ when σ is required to be determined

ii) When two samples are taken from a population

$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ & combined$\sigma = \sqrt{\frac{(n_1-1)\sigma_1{}^2+(n_2-1)\sigma_2{}^2}{n_1+n_2-2}}$

t-distribution table values (two-tailed) are given in Table 4.4

Table 4.4 t-distribution values (two- tailed)

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Illustration No. 10

A random sample of 16 water filters has 53 days as mean life period. The sum of the squares of the deviations taken from mean is 135. Can this sample be regarded as taken from the population having 56 days as mean at 5 % level of significance?

Solution

Null hypothesis : Sample mean represents population mean

$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{(n-1)}}$

$= \sqrt{\frac{135}{(16-1)}}$

= 3

Applying t-test, $t = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$

$= \frac{53-56}{3}\sqrt{16}$

= - 4

= 4 ignoring sign

t – critical value is obtained from for df = n-1 = 15 and 5 % level of significance

Thus t-critical = 2.131

Since observed t-value is more than t-critical value, the null hypothesis is rejected.

Illustration No. 11

Two salesmen A and B are assigned with two different territories for selling same product of a company. From a sample survey conducted by the Sales manager on the performance of the two salesmen, the following results were obtained. Check if there is any significant difference between the performances of these two salesmen at 5% level of significance.

	Sales man	
	A	B
No. of samples	20	18
Average monthly sales (in thousand Rs.)	170	205
Standard deviation (in thousand Rs.)	20	25

Solution

Since the sample size is less than 30, we need to use t-test.

Let us find out the combined σ of the population as follows:

$$\sigma = \sqrt{\frac{(n_1-1){\sigma_1}^2 + (n_2-1){\sigma_2}^2}{n_1+n_2-2}}$$

$= \sqrt{\frac{(20-1)20^2+(18-1)25^2}{20+18-2}}$ = 22.5

Applying t-test, $t = \frac{\bar{X}_1-\bar{X}_2}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$

$= \frac{170-2}{22.5}\sqrt{\frac{20x1}{20}}$

= - 4.788 = 4.788 ignoring sign

t – critical value is obtained from for df = $n_1 + n_2 - 2$ = 36 (α for more than 30) and 5 % level of significance

Thus t-critical = 1.96

Since observed t-value is more than t-critical value, the null hypothesis is rejected.

To Do Activity
Check if your school performance is better than your friend's performance taking mean and standard deviation of marks in any 10 subjects of 10th and 12th standards with t-test at 95% confidence level.

4.4 Hypothesis testing based on F-distribution

When independent random samples are drawn from two different populations following normal distribution, the ratio of square of mean to square of standard deviation of one same sample to other sample follows F-distribution as follows:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

with degree of freedom at $df_1 = n_1-1$ and $df_2 = n_2-1$.
where s_1^2 and s_2^2 are variances of two samples.

If two normal populations have equal variances, then $\sigma_1^2 = \sigma_2^2$ and $F = \frac{s_1^2}{s_2^2}$; $s_1 > s_2$

The F-table values are given in table 4.5 for 1% and 5% levels of significance.

Let us learn F-distribution through the following illustrations.

Illustration No. 12

A study was carried out to learn if women have a greater variation in attitude on political issues than men. Two independent samples of 21 men and 31 women were picked up for the study. The sample variances so calculated were 100 and 70 for women and men respectively. Test if the difference in attitude towards political issues between women and men is significant at 5 % level of significance.

Solution

Null Hypothesis : There is no difference in attitude towards political issues between women and men i.e $\sigma_w^2 = \sigma_m^2$
Women sample is treated as sample-1

The F-test statistic is given by $F = \frac{s_1^2}{s_2^2}$

$$= \frac{100}{70} = 1.429$$

The F-critical value at 5% significant level with $df_1 = 21-1 = 20$ and $df_2 = 31-1 = 30$ is 1.9317
Since observed t-value is less than t-critical value, the null hypothesis is accepted.

Illustration No. 13

The following data relate to the number of units produced in two different shifts by two employees A and B for a number of days.

A	17	18	21	22	19	20	22	22	19		
B	45	28	33	40	25	33	32	35	24	42	37

Can it be inferred that employee A is more stable compared to worker B? Test it at 5% level of

significance.

Solution

Null Hypothesis : Both the employees are equally stable. i.e $\sigma_A^2 = \sigma_B^2$

Let us compute the values of variance for samples.

Employee A			Employee B		
X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
17	-3	9	45	11	121
18	-2	4	28	-6	36
21	1	1	33	-1	1
22	2	4	40	6	36
19	-1	1	25	-9	81
20	0	0	33	-1	1
22	2	4	32	-2	4
22	2	4	35	1	1
19	-1	1	24	10	100
			42	8	64
			37	3	9
Sum = 180		Sum = 28	Sum = 374		Sum = 454

$\bar{X}_1 = \frac{\sum X_i}{n_1} = \frac{180}{9} = 20$; $S_A^2 = \frac{\sum(Xi - \bar{X})2}{n_1 - 1} = \frac{28}{9-1} = 3.5$

$\bar{X}_2 = \frac{\sum X_i}{n_1} = \frac{374}{11} = 34$; $S_B^2 = \frac{\sum(Xi - \bar{X})2}{n_2 - 1} = \frac{454}{11-1} = 45.4$

The F-test statistic is given by $F = \frac{s_1^2}{s_2^2}$

$= \frac{45.4}{3.5}$; here $S_B^2 > S_A^2$ hence S_B is taken as first sample

$= 12.97$

The F-critical value at 5% significant level with $df_1 = 11-1 = 10$ and $df_2 = 9-1 = 8$ is 3.347

Since observed t-value is more than t-critical value, the null hypothesis is rejected. Hence, both the employees are not equally stable. Since $S_B^2 > S_A^2$, we can say that employee A is more stable than employee B.

Table 4.5 F-values for α = 1 % & 5% levels of significance (presented below)

Degrees of Freedom Denominator	α	Degrees of Freedom for Numerator: 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	30	50	100	1000
	0.05	161	199	216	225	230	234	237	239	241	242	243	244	245	245	246	250	252	253	254
1	0.01	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143	6157	6260	6302	6334	6363
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.46	19.48	19.49	19.49
2	0.01	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.41	99.42	99.42	99.43	99.43	99.47	99.48	99.49	99.50
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.62	8.58	8.55	8.53
3	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92	26.87	26.50	26.35	26.24	26.14
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.75	5.70	5.66	5.63
4	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20	13.84	13.69	13.58	13.47
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.50	4.44	4.41	4.37
5	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72	9.38	9.24	9.13	9.03
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.81	3.75	3.71	3.67
6	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56	7.23	7.09	6.99	6.89
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.38	3.32	3.27	3.23
7	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31	5.99	5.86	5.75	5.66
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.08	3.02	2.97	2.93
8	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.20	5.07	4.96	4.87
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.86	2.80	2.76	2.71
9	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96	4.65	4.52	4.41	4.32
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.70	2.64	2.59	2.54
10	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.25	4.12	4.01	3.92
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.57	2.51	2.46	2.41
11	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25	3.94	3.81	3.71	3.61
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.47	2.40	2.35	2.30
12	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.70	3.57	3.47	3.37
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.38	2.31	2.26	2.21
13	0.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.51	3.38	3.27	3.18

(Source : Understanding Business Research Edited by Bart L Weathington, Christopher J L Cunningham and David J Pittenger – Published by John Wiley & Sons)

Degrees of Freedom Denominator																				
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.31	2.24	2.19	2.14
14	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.35	3.22	3.11	3.02
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.25	2.18	2.12	2.07
15	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.21	3.08	2.98	2.88
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.19	2.12	2.07	2.02
16	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.10	2.97	2.86	2.76
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.15	2.08	2.02	1.97
17	0.01	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.00	2.87	2.76	2.66
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.11	2.04	1.98	1.92
18	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23	2.92	2.78	2.68	2.58
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.07	2.00	1.94	1.88
19	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15	2.84	2.71	2.60	2.50
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.04	1.97	1.91	1.85
20	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09	2.78	2.64	2.54	2.43
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.01	1.94	1.88	1.82
21	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.72	2.58	2.48	2.37
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	1.98	1.91	1.85	1.79
22	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.67	2.53	2.42	2.32
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	1.96	1.88	1.82	1.76
23	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.62	2.48	2.37	2.27
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	1.94	1.86	1.80	1.74
24	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.58	2.44	2.33	2.22
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	1.92	1.84	1.78	1.72
25	0.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85	2.54	2.40	2.29	2.18
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	1.90	1.82	1.76	1.70
26	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81	2.50	2.36	2.25	2.14
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	1.88	1.81	1.74	1.68
27	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78	2.47	2.33	2.22	2.11

(Continued)

Degrees of Freedom Denominator	α	Degrees of Freedom for Numerator																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	30	50	100	1000
	0.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	1.87	1.79	1.73	1.66
28	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75	2.44	2.30	2.19	2.08
	0.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	1.85	1.77	1.71	1.65
29	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73	2.41	2.27	2.16	2.05
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.84	1.76	1.70	1.63
30	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70	2.39	2.25	2.13	2.02
	0.05	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.11	2.08	2.05	2.03	2.00	1.83	1.75	1.68	1.62
31	0.01	7.53	5.36	4.48	3.99	3.67	3.45	3.28	3.15	3.04	2.96	2.88	2.82	2.77	2.72	2.68	2.36	2.22	2.11	1.99
	0.05	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.01	1.99	1.82	1.74	1.67	1.60
32	0.01	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93	2.86	2.80	2.74	2.70	2.65	2.34	2.20	2.08	1.97
	0.05	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	2.09	2.06	2.03	2.00	1.98	1.81	1.72	1.66	1.59
33	0.01	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91	2.84	2.78	2.72	2.68	2.63	2.32	2.18	2.06	1.95
	0.05	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99	1.97	1.80	1.71	1.65	1.58
34	0.01	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89	2.82	2.76	2.70	2.66	2.61	2.30	2.16	2.04	1.92
	0.05	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04	2.01	1.99	1.96	1.79	1.70	1.63	1.57
35	0.01	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.80	2.74	2.69	2.64	2.60	2.28	2.14	2.02	1.90
	0.05	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03	2.00	1.98	1.95	1.78	1.69	1.62	1.56
36	0.01	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86	2.79	2.72	2.67	2.62	2.58	2.26	2.12	2.00	1.89
	0.05	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96	1.94	1.76	1.68	1.61	1.54
38	0.01	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83	2.75	2.69	2.64	2.59	2.55	2.23	2.09	1.97	1.85
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.74	1.66	1.59	1.52
40	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52	2.20	2.06	1.94	1.82
	0.05	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.03	1.99	1.96	1.94	1.91	1.73	1.65	1.57	1.50
42	0.01	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78	2.70	2.64	2.59	2.54	2.50	2.18	2.03	1.91	1.79
	0.05	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.95	1.92	1.90	1.72	1.63	1.56	1.49
44	0.01	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75	2.68	2.62	2.56	2.52	2.47	2.15	2.01	1.89	1.76
	0.05	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04	2.00	1.97	1.94	1.91	1.89	1.71	1.62	1.55	1.47
46	0.01	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73	2.66	2.60	2.54	2.50	2.45	2.13	1.99	1.86	1.74

Degrees of Freedom Denominator																							
	0.05	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.99	1.96	1.93	1.90	1.88	1.70	1.61	1.54	1.46
48	0.01	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.71	2.64	2.58	2.53	2.48	2.44	2.12	1.97	1.84	1.72
	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.69	1.60	1.52	1.45
50	0.01	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.63	2.56	2.51	2.46	2.42	2.10	1.95	1.82	1.70
	0.05	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.97	1.93	1.90	1.88	1.85	1.67	1.58	1.50	1.42
55	0.01	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53	2.47	2.42	2.38	2.06	1.91	1.78	1.65
	0.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.65	1.56	1.48	1.40
60	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.44	2.39	2.35	2.03	1.88	1.75	1.62
	0.05	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98	1.94	1.90	1.87	1.85	1.82	1.63	1.54	1.46	1.38
65	0.01	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61	2.53	2.47	2.42	2.37	2.33	2.00	1.85	1.72	1.59
	0.05	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.62	1.53	1.45	1.36
70	0.01	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.51	2.45	2.40	2.35	2.31	1.98	1.83	1.70	1.56
	0.05	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.05	2.00	1.95	1.91	1.88	1.84	1.82	1.79	1.60	1.51	1.43	1.34
80	0.01	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.48	2.42	2.36	2.31	2.27	1.94	1.79	1.65	1.51
	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.57	1.48	1.39	1.30
100	0.01	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.43	2.37	2.31	2.27	2.22	1.89	1.74	1.60	1.45
	0.05	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.96	1.91	1.87	1.83	1.80	1.77	1.75	1.55	1.45	1.36	1.26
125	0.01	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.66	2.55	2.47	2.39	2.33	2.28	2.23	2.19	1.85	1.69	1.55	1.39
	0.05	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.79	1.76	1.73	1.54	1.44	1.34	1.24
150	0.01	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.37	2.31	2.25	2.20	2.16	1.83	1.66	1.52	1.35
	0.05	3.89	3.04	2.65	2.42	2.26	2.14	2.05	1.98	1.93	1.88	1.84	1.80	1.77	1.74	1.72	1.52	1.41	1.32	1.21
200	0.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27	2.22	2.17	2.13	1.79	1.63	1.48	1.30
	0.05	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.74	1.72	1.69	1.49	1.38	1.28	1.15
400	0.01	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.29	2.23	2.17	2.13	2.08	1.75	1.58	1.42	1.22
	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.80	1.76	1.73	1.70	1.68	1.47	1.36	1.26	1.11
1000	0.01	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.27	2.20	2.15	2.10	2.06	1.72	1.54	1.38	1.16

Illustration No. 14

The average production of mangoes produced by two mango farms A and B is assumed to be same but the standard deviations may vary. For a sample of 22 mango trees produced in farm A, the standard deviation is 2.9 mangoes and for a sample of 16 mango trees in farm B, the standard deviation is 3.8 mangoes. Check if the variability of production of mangoes in both the farms are same at 5% level of significance.

Solution

Null hypothesis : There is no difference in variability of mangoes produced in both farms A & B. i.e $\sigma_A^2 = \sigma_B^2$

Given :$\sigma_A = 2.9$, $\sigma_B = 3.8$, $n_1 = 22$, $n_2 = 16$

$S_1^2 = \frac{n_1}{n_1 - 1}\sigma_A^2 = \frac{22}{22-1}2.9^2 = 8.810$

$S_2^2 = \frac{n_2}{n_2 - 1}\sigma_B^2 = \frac{16}{16-1}3.8^2 = 15.403$

The F-test statistic is given by $F = \frac{S_1^2}{S_2^2}$

$= \frac{15.403}{8.81}$; here $S_2^2 > S_1^2$ hence S_2 is taken as first sample

$= 1.748$

The F-critical value at 5% significant level with $df_1 = 16-1 = 15$ and $df_2 = 22-1 = 21$ is 2.175

Since observed t-value is less than t-critical value, the null hypothesis is accepted.

Hence, the variability in production of mangoes in both the farms A and B are same.

4.5 Non-Parametric Tests

Non-parametric tests do not depend on the form of the underlying population distribution from which a sample is drawn for testing. Therefore, it is called distribution free test.

A non-parametric test should satisfy atleast one of the following criteria.

1) The test does not consider any population parameter such as mean, standard deviation
2) The test is applied only on categorical data that are non-numerical and frequency of categories for one or more variables.
3) The test does not depend on the form of the underlying population distribution, especially the requirement of normal distribution.

Chi-square test belongs to non-parametric category of methods to test a hypothesis. Chi-square test is used to test the goodness of fit to verify the distribution of observed data with expected theoretical data. Hence, it is a measure to study the divergence of actual and expected frequencies. Chi-square is denoted by the letter χ^2.

χ^2 test for goodness of fit is given by

$\chi^2 = \sum\left\{\frac{(O-E)^2}{E}\right\}$; where O = observed frequencies and E = expected frequencies

E values for contingency table is given by

$E = \frac{Row\ total\ x\ Column\ total}{Grand\ total}$

Degree of freedom is given by

df = n – k where n = number of frequency classes and k = number of independent constraints

For contingency table like c x r table,
df = (c-1) (r-1) where c = number of columns and r = number of rows

Characteristics of χ^2test :

1. It is based on events or frequencies
2. It is useful only to test the hypothesis, not for estimation
3. It is used between entire set of observed and expected frequencies
4. For every increase in the number of degree of freedom, a new χ^2 distribution is formed

Assumptions in χ^2test :

1) All the observations must be independent
2) All the events are mutually exclusive
3) There must be large observations

The values of χ^2 are given in table 4.6

Table 4.6 Percentage points of Chi-Square distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Let us learn the procedure to use the χ^2 test through the following illustrations.

Illustration No. 15

Out of a sample of 140 persons in a village, 80 persons were administered a new drug for preventing a particular disease and out of them 30 persons were found attacked by the disease. Out of those who were not administered the new drug, 15 were not found attacked by the disease. Test if the new drug is effective in preventing the targeted disease at 5% level of significance.

Solution

Let us first form the table for observed frequencies as given below.

Given table values :

	No. of persons attacked by the disease	No. of persons not attacked by the disease	Total
No. of persons administered with the new drug	30		80
No. of persons not administered with the new drug		15	
Total			140

The above table is filled with other values for all the blank cells as given below in the ascending order of (i), (ii), …(v)

	No. of persons attacked by the disease	No. of persons not attacked by the disease	Total
No. of persons administered with the new drug	30	50 (ii)	80
No. of persons not administered with the new drug	45 (v)	15	60 (i)
Total	75 (iv)	65 (iii)	140

Let us find out the expected frequencies using the formula E = $\frac{Row\ total\ x\ Column\ total}{Grand\ total}$

	No. of persons attacked by the disease	No. of persons not attacked by the disease	Total
No. of persons administered with the new drug	$\frac{80\ x\ 75}{140}$ = 42.86	$\frac{80\ x\ 65}{140}$ = 37.14	80
No. of persons not administered with the new drug	$\frac{60\ x\ 75}{140}$ = 32.14	$\frac{60\ x\ 65}{140}$ = 27.86	60
Total	75	65	140

χ^2 table is formed as follows :

O	E	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
30	42.86	- 12.86	165.38	3.859
45	32.14	12.86	165.38	5.146
50	37.14	12.86	165.38	4.453
15	27.86	-12.86	165.38	5.936
				Sum = 19.394

Test statistic, $\chi^2 = \sum\left\{\frac{(O-E)^2}{E}\right\}$ = 19.394 (from above table)

χ^2 –table value for df = (c-1)(r-1) = (2-1)(2-1) = 1 for 5% level of significance is 3.84

Since χ^2 observed value is more than χ^2-critical (table) value, the null hypothesis is rejected.

Therefore the drug is not effective in controlling the targeted disease.

To Do Activity
Check the effectiveness of any awareness program conducted in your institution taking a survey to collect feedback on a sample of 25 participants and 10 non-participants of the program.

Illustration No. 16

4 coins are tossed 160 times and the following results are obtained. Under the assumption that coins are balanced, find the expected frequencies of getting 0, 1,…4 heads and the test the goodness of fit of observed data.

No. of heads	0	1	2	3	4
Observed frequencies	11	35	57	42	9

Solution

Null hypothesis : Tossing of coins is unbiased

Let us first find out the expected frequencies.

x	Expected frequency $n\ {}^4C_x\ p^4$ where p = ½
0	$160\ {}^4C_0\ (0.5)^4 = 10$
1	$160\ {}^4C_1\ (0.5)^4 = 40$
2	$160\ {}^4C_2\ (0.5)^4 = 60$
3	$160\ {}^4C_3\ (0.5)^4 = 40$
4	$160\ {}^4C_4\ (0.5)^4 = 10$

χ^2 table is formed as follows :

O	E	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
11	10	1	1	0.1
35	40	-5	25	0.625
57	60	-3	9	0.15
42	40	2	4	0.1
9	10	-1	1	0.1
				Sum = 1.075

Test statistic, $\chi^2 = \sum\left\{\frac{(O-E)^2}{E}\right\}$ = 1.075 (from above table)
χ^2 –table value for df = n-1 = 5-1 = 4 for 5% level of significance (assumed) is 9.49
Since χ^2 observed value is less than χ^2-critical (table) value, the null hypothesis is accepted.
Therefore the tossing of coin is unbiased.

Illustration No. 17
A dice is thrown for 120 times with the following results. Test if the dice is unbiased at 10% level of significance.

No. turned up	1	2	3	4	5	6	Total
Observed frequencies	19	22	21	20	18	20	120

Solution
Null hypothesis : Throw of dias is unbiased
The expected frequency for each number of dice for 120 throws = 120 x $\frac{1}{6}$ = 20
Applying χ^2 test,

O	E	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
19	20	-1	1	0.05
22	20	2	4	0.2
21	20	1	1	0.05
20	20	0	0	0
18	20	-2	4	0.2
20	20	0	0	0
				Sum = 0.5

Test statistic, $\chi^2 = \sum\left\{\frac{(O-E)^2}{E}\right\}$ = 0.5 (from above table)
χ^2 –table value for df = n-1 = 6-1 = 5 for 10% level of significance (assumed) is 9.24
Since χ^2 observed value is less than χ^2-critical (table) value, the null hypothesis is accepted.
Therefore the throw of dice is unbiased.

To Do Activity
Throw a coin for 50 times and check whether the coin biased based on its outcomes with the support of chi-square test at 10% level of significance.

The Sign Test for Paired Data

The sign test is also known as paired-sample test and based on the sign of difference in paired observations (x,y) where x is the value of observation from population 1 and y for population 2. The test assumes that the pairs are independent and the measurement scale within each pair is at least ordinal. Let us discuss a problem to learn the procedure for sign test.

Illustration No. 18

The average runs of a cricket batsman in one day match is claimed to be 35 based on last 5 years history of cricket of all batsmen. The following data gives the runs scored by a particular batsman during last 20 matches. Test if the batsman score is above or below the average score of all batsmen.

Average score	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35
Batsman score	27	42	10	78	35	56	28	32	37	86	27	56	45	48	57	62	74	54	0	12
Sign	-	+	-	+	0	+	-	-	+	+	-	+	+	+	+	+	+	+	-	-

Solution

Null hypothesis : Batsman mean score is similar to average score of all batsmen.

Given : 20, number of (+) sign, x = 12, p = ½ , q = ½

Thus, $\mu = np = 20\ (0.5) = 10$

$\sigma = \sqrt{n\ p\ q} = \sqrt{20\ x\ 0.5\ x\ 0.5} = 2.236$

Applying z-test statistics, we get,

$z = \frac{\bar{X} - \mu}{\sigma} = \frac{7 - 10}{2.236} = -1.342$

where $\bar{X}$ = number of (-) sign.

Since Z_{cal}(= -1.342) is greater than Z_{tabe}(= -1.96) at 5% level of significance, null hypothesis is accepted.

Runs Test for Randomness

A 'run' is a sequence of identical occurrences of elements (number, symbol, character, etc) preceded and followed by different occurrences of elements or by no element at all.

Run test helps to check if the order or sequence of observations (number, symbol, character, etc) in a sample is random. Runs test examines the number of 'runs' of each of two possible characteristics that sample elements may have.

Let us go through the following illustrations to understand the procedure of Runs test.

Illustration No. 19

The following is the pattern of items, non-defective or defective produced by a machine in a batch of production. The quality engineer wants to determine if the sequence of defective(D) versus non-defective(N) items is random. Test if the distribution of defective and non-defective items is random at

5% level of significance.

NNNNN, DDDD, NNN, DD, NNNNN, DDDD, NNNNNN, DD, NNNNNN, DDD

Solution

Null hypothesis : Distribution of defective and non-defective items is random

Given, r= 10 runs n_1 = 25(number of Non-defective) n_2 = 15 (defective)

$\mu = \frac{2\,n_1 n_2}{n_1+n_2}+1 = \frac{2\,x\,25\,x1}{25+15}+1 = 17.75$

$\sigma = \sqrt{\frac{2\,n_1\,n_2\,(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} = \sqrt{\frac{2x\,25\,x\,15(2\,x\,25\,x\,15-25-15)}{(25+15)^2(25+15\quad)}} = 2.921$

Applying z-test statistics, we get,

$z = \frac{r-\mu}{\sigma} = \frac{10-17.75}{2.921} = -2.653$

Since Z_{cal}(= -2.653) is greater than Z_{tabe}(= -1.96) at 5% level of significance, null hypothesis is rejected. That is distribution of defective and non-defective items is not random

Summary

In this chapter, we have learnt the elements of tests of significance in the introduction. The difference between null hypothesis and alternative hypothesis has been explained with examples. The two types of errors arising out of decision maker's mistakes have been explained. The procedure to take values from normal distribution table has been explained with examples. Variour parameters tests such Z-test, t-test and F-test have been explained in detail with numerous examples. Hypothesis testing with small samples, large samples, samples from different populations have been detailed in this chapter. Non-parameter tests such Chi-square test, Sign test, runs test have been explained with good number of illustrations.

Model Questions

Problems

1. Out of a consignment of 1 Lac vegetables, 400 are selected at random for testing. 20 are found to be poor quality. How many poor quality vegetables we can expect to be in the whole assignment at 5% level of significance.
   Ans : 5000±1.96(68.9)
2. In a sample of 1000 fruits obtained from a large consignment, 200 are found to be defective. Estimate the percentage of defective fruits expected in the consignment.
   Ans : 20% ± 3(1.26)
3. In a survey of 1000 house-managers in a village, 23% of them preferred a particular brand of gas stove. Find at 1% level of significance limits, for the percentage of all house-managers in the village preferring same brand gas stove.
   Ans : 19.6 and 26.4
4. A random sample of 900 coconuts is found to have a mean weight of 65.3 g. Can it be regarded as a sample taken from a population with mean weight of 66.2 g and standard deviation of 5 g. Use 95% confidence level.
   Ans : Z = - 5.4

5. 5 dices are thrown 30 times. The number of 6 is obtained 23 times in the throws. Check if the difference between observed frequency and expected frequency is significant.
   Ans : Z = 0.431 < 1.96
6. An automatic cone ice-cream making machine is designed to fill ice-creams in the cone for 20 g. A sample of 100 cones are examined and found to have filled the cones for a mean weight of 19.4 g with standard deviation of 1 g. Is the machine working properly?
   Ans : Z = 6 > 3
7. In a village A, there are 450 vegetarians out of 1000 persons surveyed. In another village B, there are 400 vegetarians out of 00 persons surveyed. Is there any significance difference between the two villages as far as vegetarians are concerned? Use 5% level of significance.
   Ans : Z = 2.8 > 1.96
8. A typist claims that he can type the words at the rate of 120 words per minute. During testing his capability, he types the letter with a mean of 116 words and with a standard deviation of 15 words. Is it right to reject his claim at 5% level of significance.
   Ans : Z = 2.67 > 1.96
9. 500 students in a school were categorized according to their academic performance and economic conditions of their family as mentioned below. Check if there is any association between their academic performance and their economic conditions.

	Academic performance	
Economic conditions	Good	Bad
Poor	10. 85	11. 75
Rich	12. 165	13. 175

   Ans : $\chi^2 = 0.92$
9. The number of late comers to a class per weak is as follows.
   2, 6, 4, 12, 8, 20, 14, 10, 15, 9
   Are these frequencies in agreement with the belief that late comers are same during this 10 week period?
   Ans : $\chi^2 = 16.92$
10. 24 cattles were given immunization vaccine from anthrax and the following results were obtained. Check if the vaccine is effective in treating the disease.

	Died	Survived	Total
Inoculated with vaccine	2		12
Not inoculated with vaccine		6	
Total			24

    Ans : $\chi^2 = 7.796 > 3.84$
11. Suppose some cool drinkers are randomly tested to determine if they prefer brand A or brand B cool drinks. The sequence of sampled cool drink choices is given below. Is this sequence of cool drinks evidence that the sample is taken at random.

    BAAAAABAABAAAABABAAABBBAAA

Ans : R = 12

12. A cluster of 8 customers were asked about their perception of a product before and after they used the product on an ordinal scale as given below.

Customer	A	B	C	D	E	F	G	H
Before	9	4	6	4	5	8	7	6
After	10	3	4	1	6	8	8	1

Theoretical Questions

1. What are the elements of tests of significance?
2. Write down the differences between null hypothesis and alternative hypothesis with good examples
3. Define standard error
4. Explain Type- I and Type-II errors
5. What is meant by students-t test?
6. What is the condition to use F-test?
7. What do you mean by non-parametric tests?
8. What are the conditions to be fulfilled to categorize a test to non-parametric test?
9. State any two applications of sign test
10. State any two applications of runs test

Multiple Choice Questions

1. A hypothesis defining the population distribution is
   a) Simple Hypothesis
   b) Null Hypothesis
   c) Statistical Hypothesis
   d) Composite Hypothesis
2. An observed set of the population that has been selected at random for analysis is called
   a) Sample
   b) Forecast
   c) Process
   d) Descriptive statistics
3. Type 1 error occurs when
   a) we reject H0 if it is False
   b) we reject H0 if it is True
   c) we accept H0 if it is True
   d) we accept H0 if it is False
4. Alternative Hypothesis is also called as
   a) Composite Hypothesis
   b) Research Hypothesis
   c) Simple Hypothesis
   d) Null Hypothesis

5. What does a significant result in a chi-square test imply?
   a) That the sample is not representative of the population
   b) That there is a significant negative relationship
   c) That there is a significant difference between the categorical variables
   d) That there is a significant positive relationship
6. What symbol is used to represent chi-square?
   a) χ^2
   b) Ψ
   c) F
   d) *Π*
7. Null and alternative hypotheses are statements about
   a) sample parameters
   b) sample statistics
   c) sometimes population parameters and sometimes sample statistics
   d) population parameters
8. To perform a runs test for randomness, the data should
   a) a quantitative
   b) a qualitative
   c) be divided into exactly two classification
   d) divide into atleast two classification
9. The sign test assumes that the
   a) Samples have the same mean
   b) Samples are dependent
   c) Samples are independent
   d) Samples are not consider
10. A t- test is used to compare
   a) Two mean
   b) three mean
   c) four mean
   d) five mean
11. The null hypothesis can be described as
   a) The same as the research hypothesis
   b) A statement that there is no different
   c) A statement of the expected result
   d) A statement of probability
12. Which of the following distributions is used to compare two variances?
   a) F – Distribution
   b) T – Distribution
   c) Normal Distribution
   d) Poisson Distribution
13. Setting the p level at 0.01 increases the chances of making
   a) Type I error
   b) Type II error
   c) Type III error

d) Both a and b

14. Type I error is also called as
    a) False negative
    b) False positive
    c) Double negative
    d) Double positive
15. The two forms of t-test are
    a) One way and two way
    b) bivariate and multiple
    c) Factorial and interaction
    d) independent and dependent

Answers for MCQ

1	2	3	4	5	6	7	8	9	10
a	a	b	b	a	a	d	c	b	a
11	**12**	**13**	**14**	**15**					
b	a	b	b	d					

References

1. Sharma, J.K. (2014). *Business Statistics – Problems and Solutions*. New Delhi : Vikas Publishing House Pvt Ltd.
2. Pillai, R.S.N. &Bagavathi, V. (1999). *Statistics*. New Delhi :S.Chand& Company Ltd.
3. Gupta, S.P. (2010). *Statistical Methods.* New Delhi :S.Chand& Company Ltd.
4. Beri, G.C. (2011). *Business Statistics.* New Delhi : Tata McGraw Hill Educations Pvt Ltd.
5. Foster, D. & Stine, E.R. (2010). *Statistics for Business : Decision Making and Analysis.* New Delhi : Pearson Publishers.
6. Gupta, S.C. & Kapoor, V.K. (2006). Fundamentals of Mathematical *Statistics.* New Delhi :S.Chand& Company Ltd.
7. Srivastava, S.C & Srivastava, S. (2003). Fundamentals of Statistics. New Delhi : Anmol Publications Pvt. Ltd.

Chapter 5 Analysis of Variance, Correlation, Regression and Time Series

Introduction

The purpose of this chapter is to introduce you the basics of "Design of Excrement" in statistics. The details of steps taken to make sure a scientific analysis leading to valid inferences about the hypothesis is called Design of Experiment This design of experiment originated from agriculture research and all credits goes to Prof R.A.Fisher. For example, to verify the claim that particular manure causes increase in the yield of paddy, we may conduct agricultural experiment. In this experiment the amount of manure used and quantity of yield are two variables directly involved, these are experimental variables. The other factors affect the yield are fertility of soil, the amount of rainfall, the quality of seed these variables are extraneous variables and independent experimental variables are called factors. The statistical analysis related to the study of two or more variables and degree of relationship is the correlation. After knowing the relationship between two variables, we may be interested to estimate the value of one variable given the value of another. Regression analysis is the study of average relationship between the variables. A businessman is interested in finding out his likely sales in the future he may adjust his production accordingly to meet the demand. In this connection he arranges the collected sales details in specific period in some order. Generally, such data are known as time series.

Objectives of the Chapter

- To provide insights on meaning, definition, and uses of correlation
- To understand correlation coefficient for different type of measurement.
- To provide insights on the concept of regression, definition, types and uses
- To familiarise with difference between correlation and regression
- To familiarise with the concept of time series, estimate the trend values
- To understand the concept of interpolation and extrapolation
- To interpolate the values of a given data by using Newton's interpolation and LaGrange's interpolation

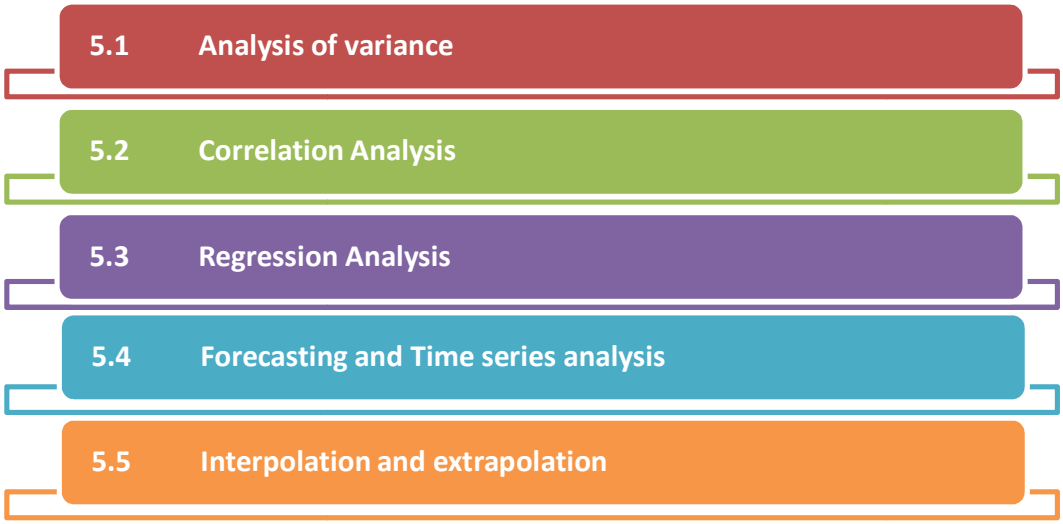


Fig 5.1 Chapter Flow

5.1 Analysis of variance

The powerful statistical tool for test of hypothesis is called analysis of variance. This concept is introduced by Prof R.A.Fisher to deal with the problems in agriculture In students t test we are testing the difference between means of two samples. Suppose we want to test the more than two samples , the alternative procedure is needed to test the hypothesis that all the samples are drawn from the same population(samples have same mean).For example x number of fertilizers applied to x number of plots each of paddy and yield of paddy from each plot . Now our aim is to estimate the effect of these fertilizers on the yield is significantly different or the samples are drawn from the same normal population. This answer is given by the technique of analysis of variance.

The Analysis of variance or shortly ANOVA refers widely to a collection of experimental situations and statistical procedures for the analysis of quantitative responses from experimental units. Variation is characteristic attribute in nature. Due to number of causes, it is classified in to two causes

(i) Assignable cause (ii) Chance cause

A process is that operating with only assignable cause is said to be in out of control and a process that operates only chance of cause is in statistical control. Detailed explanation of this topic s is summarized in fig. 5.2

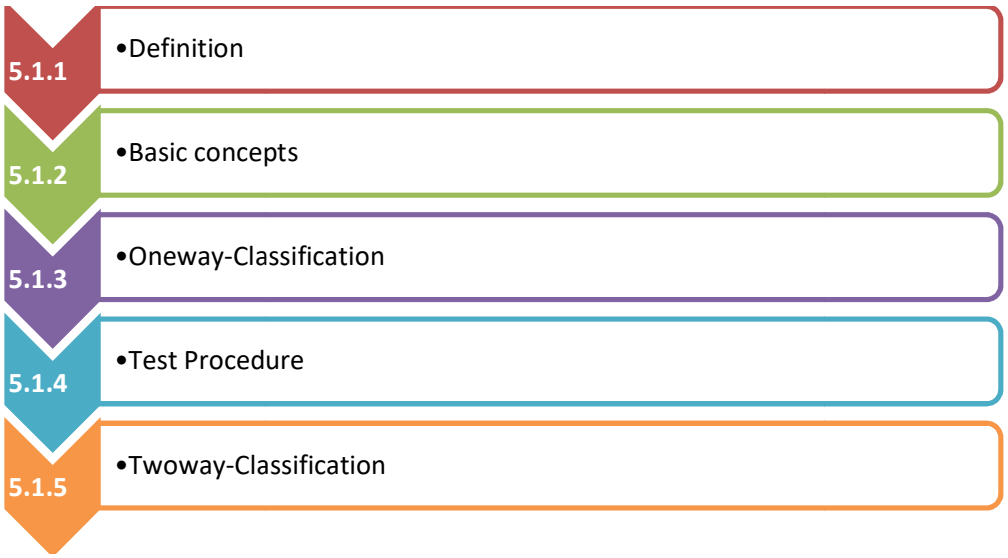


Fig. 5.2 Summary of ANOVA

Definition

According to R.A.Fisher, ”Analysis of variance is the separation of variance classification in show ascribable to one group of causes from the variance ascribable to the other group”. By this technique the total variation in the sample data is expressed as the sum of the possible components where each of these components is a measure of the variation due to some specific source or factor or cause.

Basic Concepts

For using analysis of variance, we should follow the following assumptions.

(i) All samples drawn from normal population

(ii) All the samples have same variances.

(iii) The samples are independently drawn from these populations.

Note: Suppose in any problem the assumption is not made we are not able to use ANOVA technique. In such cases we must use non-parametric tests.

Analysis of variance is a technique of partitioning the total sum of squared deviations of all sample values from the grand mean in to two parts.

(i) Sum of squares between samples (SSB)

(ii) Sum of squares within samples (SSW)

SSB is due to assignable cause and SSW is due to chance cause. This also called residual random variation (or error).

Under the null hypothesis there is no difference between means of populations. We apply F-test to see any significant difference between the two variances exist or not.

We consider the following types of ANOVA

(i) One way classification (One factor ANOVA) for CRD (Completely Randomised Design).

(ii) Two-way classification(or two factor ANOVA) for RBD(Randomised Block Design)

(iii) Three factor ANOVA for LSD(Latin squared Design)

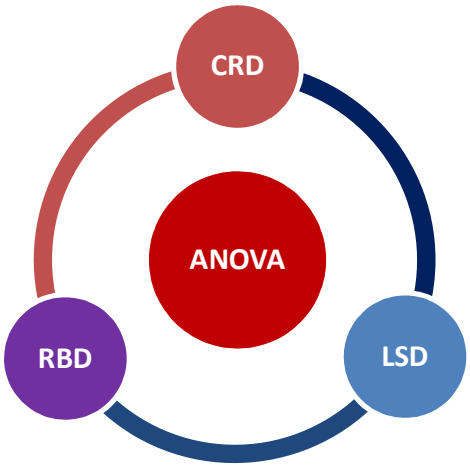


Fig.5.3 Types of ANOVA

One-way Classification (Completely randomised Design)

In one-way classification the observations or experimental units are classified according to one factor of interest. For example, the yields of several plots of land may be classified according to the type of fertilisers used. Here the factor is treatment namely the type of fertilisers. Suppose we have k independent random samples of sizes $n_1,n_2,n_3.....n_k$ from k normal population whose means are $\mu_1,\mu_2,.....\mu_k$ so that $\sum n_i = N$ .Then the null hypothesis H_0: $\mu_1=\mu_2=.....=\mu_k$. The sample means are $\overline{x_1},\overline{x_2},.....\overline{x_k}$

The grand mean = $\dfrac{\overline{x_1}+\overline{x_2}+.....+\overline{x_k}}{k}$

Sum of the squares between samples (SSB) = $\sum_j \left(\overline{x} - \bar{x}\right)^2$

Sum of the squares within samples (SSW) = $\sum_j \sum_i \left(x_{ij} - \overline{x_j}\right)^2$

Total sum of squares (SST) = $\sum_j \sum_i \left(x_{ij} - \overline{x}\right)^2$ = SSW + SSB

There are k samples, so the number of digress of freedom= k – 1

If this sum of squares is divided by the corresponding degrees of freedom, we get the mean sum of squares. In computation there are several steps to calculate SSB, SSW, SST and mean sum of squares these calculations can be summarised below the following table 5.1

Table 5.1 ANOVA One Way Classification

Source of variation	**Degrees of freedom(df)**	**Sum of squares(SS)**	**Mean sum of squares(MS)**	**F-Ratio**
Between samples	K - 1	SSB	MSB = SSB_1/k-1	F= $\left(\frac{MSB}{MSW}\right)^{\pm 1}$
Within samples	N - k	SSW	MSW = SSW/N-k	F distribution with (k-1,N-k)df
Error	---	----	------	
Total	N -1	SST	-------	

To use this ANOVA table the following computation is very useful.

1) SST = $\sum_j \sum_i \left(x_{ij}\right)^2 - \frac{T^2}{N}$ (Correction factor)

Where $\frac{T^2}{N}$ = Correction factor.

$T = \sum_j \sum_i x_{ij}$ = Grand total

N = Total number of units.

2) SSB = $\sum_{i=1}^{k} \frac{T^2{}_i}{k_i} - \frac{T^2}{N}$

3) SSW = SST – SSB

After the calculation of F-ratio compare the calculated value of F for the given degrees of freedom at α % level of significance.

(i) Calculated value of F < tabulated value null hypothesis H_0 is accepted and we conclude that there is no significant difference between the treatments.

(ii) Calculate value of F > tabulated value, H_0 is rejected.

Illustration No. 1

Four machines A,B,C and D are used to produce a certain kind of silk fabric samples of size 4 with each unit as 100 square meters are selected from the outputs of the machines at random and the number of flaws in each 100 square meters is counted with the following results.

A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is a significant difference in the performance of the four machines? Test at 5% level of significance.

Solution

Setup the null hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

i.e.) there is no significant difference between the performances of the machines.

Alternate hypothesis: H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

A	B	C	D
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23
$T_1 = 40$	$T_2 = 28$	$T_3 = 53$	$T_4 = 90$

Here N = 16 and

$$T = \sum_j \sum_i x_{ij} = 40 + 28 + 53 + 90 = 211$$

Correction factor = $\frac{T^2}{N}$

$$= \frac{(211)^2}{16} = 2782.56$$

Total sum of squares (SST) = $\sum_j \sum_i (x_{ij})^2 - \frac{T^2}{N}$

$$where \quad \sum_j \sum_i x^2{}_{ij} = (8)^2 + (9)^2 + \ldots\ldots\ldots + (23)^2 \quad = 3409$$

SST = 3400 – 2782.56 = 626.44

$$SSB(\text{between machines}) = \sum_{i=1}^{k} \frac{T^2{}_i}{k_i} - \frac{T^2}{N}$$

$$= \frac{40^2}{4} + \frac{28^2}{4} + \frac{53^2}{4} + \frac{90^2}{4} - 2782.56$$

= 540.69

Error = SSW = SST – SSB

= 626.44 – 540.69 = 85.75

ANOVA Table

Source of variation	Degrees of freedoim	Sum of squares	Mean sum of squares	F-Ratiio
SSB	3	540.69	540.69/3=180.23	$F = \frac{180.23}{7.15}$
Error(ssw)	12	85.75	7.15	= 25.207
Total	15	626.44	--	-----

Table value of F = F(3,12) at 5% level of significance = 2.61

Conclusion: Calculated value of F = 25.207 > table value 2.61.

H_0 is rejected. We conclude that there is a significant difference between the performance of the machines.

Illustration No. 2

There are three main brands of a certain medicine. A group of 120 sample values is examined and found to be allocated among four groups G_1, G_2, G_3, G_4 and three brands A,B,C as shown below:

Brand	Groups			
	G_1	G_2	G_3	G_4
A	0	4	8	15
B	5	8	13	6
C	8	19	11	13

Test whether any significant in brand preference? Test at 5% level.

Solution

Setup the null hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$

i.e.) there is no significant difference between the preference of the brands.

Alternate hypothesis: H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

A	B	C	Sum of squares		
X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
0	5	8	0	25	64
4	8	19	16	64	361
8	13	11	64	169	121
15	6	13	225	36	169
$T_1 = 27$	$T_2 = 32$	$T_3 = 51$	105	204	715

Here N = 12 and $T = \sum_j \sum_i x_{ij} = 27 + 32 + 51$

= 110

Correction factor = $\frac{T^2}{N}$

$$= \frac{110^2}{12} = 1008.33$$

Total sum of squares (SST)= $\sum_j \sum_i (x_{ij})^2 - \frac{T^2}{N}$

$$where \quad \sum_j \sum_i x^2{}_{ij} = 105 + 204 + 715 = 1114$$

SST = 1114– 1008.33 = 105.67

$$SSB\,(between\,machines) = \sum_{i=1}^{k} \frac{T^2{}_i}{k_i} - \frac{T^2}{N}$$

$$= \frac{27^2}{4} + \frac{32^2}{4} + \frac{51^2}{4} - 1008.33 = 80.17$$

Error = SSW = SST – SSB

= 105.67 – 80.17 = 25.5

ANOVA Table

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-Ratiio
SSB	3-1=2	80.17	80.17/2=40.085	$F = \frac{40.085}{2.83}$ 14.15
Error(ssw)	12-3=9	25.5	25.5/9=2.833	
Total	15	105.67	--	-----

Table value of F = F(2,9) at 5% level of significance = 4.26

Conclusion: Calculated value of F = 14.15 > table value 4.26.

H_0 is rejected. We conclude that there is a significant difference between the preferences of the brands.

To Do Activity
Does reasoning about personal problems improve the Psychological distance? Construct one way ANOVA with planned comparisons.

Two-way Classification

In two factor analysis of variance we consider one classification along column wise and the other is row wise. For example the yield of a crop in many plots of land may be classified according to different varieties of fertilizers. So seeds and fertilizers are the two factors. Let the N values (x_{ij}) represent the yield according to the two factors. Let there be m rows (blocks) representing one factor of classification

(different varieties of seeds) and n columns representing the other factor (different fertilizers) so that N = mn. Now we wish to test there is no difference in yield between various rows and between various columns.

Here the total variation SST consists of three parts,

(i) Sum of squares between columns(SSC)

(ii) Sum of squares between rows(SSR)

(iii) Sum of squares for residual (errors) (SSE)

(ie) SSE = SST-SSC-SSR

In two-way classification residual is the measuring tool for testing significance of difference. Residual represents the magnitude of variations due to factors called "chance".

Table 5.2 ANOVA Two way classification

Source of variation	Degrees of freedom(df)	Sum of squares(SS)	Mean sum of squares(MS)	F-Ratio
Between treatments (columns)	n – 1	SSC	MSC = SSC/n-1	F= $\left(\frac{MSC}{MSW}\right)^{\pm 1}$ F distribution with (n-1,(m-1)(n-1))df
Between blocks (Rows)	m-1	SSR	MSR = SSR/m-1	F= $\left(\frac{MSR}{MSE}\right)^{\pm 1}$ F distribution with (m-1,(m-1)(n-1))df
Residual(Error)	(m-1)(n-1)	SSE	MSE = SSE/(m-1)(n-1)	------
Total	mn-1	SST	-------	----=-

To use this ANOVA table the following computation is very useful:

1) SST = $\sum_j \sum_i \left(x_{ij}\right)^2 - \frac{T^2}{N}$ (Correction factor)

where $\frac{T^2}{N}$ = Correction factor.

$T = \sum_j \sum_i x_{ij}$ = Grand total

N = Total number of units.

2) SSC = $\sum_{j=1}^{n} \frac{T^2{}_j}{n_j} - \frac{T^2}{N}$

3) SSR = $\sum_{i=1}^{m} \frac{T^2{}_i}{n_i} - \frac{T^2}{N}$

4) SSW = SST – SSB -SSR

After the calculation of F-ratio compare the calculated value of F for the given degrees of freedom at α% level of significance.

(iii) Calculated value of F < tabulated value null hypothesis H_0 is accepted and we conclude that there is no significant difference between the treatments.

(iv) Calculate value of F > tabulated value, H_0 is rejected.

Illustration No. 3

The following data represent the number of units of production per day turned out by different workers using four types of machines.

Workers	Type of Machines			
	A	B	C	D
1	44	8	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	40	39

(i) Test whether five men differ with respect to mean productivity

(ii) Test whether the mean productivity is the same for the four different machine types.

Solution

Setup the null hypothesis: H_{01}: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (Columns)

: H_{02}: $\mu_1 = \mu_2 = \mu_3 = \mu_4$(Rows)

	A	B	C	D					
	X_1	X_2	X_3	X_4	Row total(y_i)	X_1^2	X_2^2	X_3^2	X_4^2
y_1	4	-2	7	-4	5	16	4	49	16
y_2	6	0	12	3	21	36	0	144	9
y_3	-6	-4	4	-8	-14	36	16	16	64
y_4	3	-2	6	-7	0	9	4	36	49
y_5	-2	2	9	-1	8	4	4	81	1
Column total(x_j)	5	-6	38	-17	**20**	101	28	326	139

N=20 and T = 20

Correction factor = $\frac{T^2}{N}$

= $(20)^2/20$ = 20

SST = 101 + 28 + 326 + 139 -20

= 584

$$SSC = \frac{5^2}{5} + \frac{(-6)^2}{5} + \frac{38^2}{5} + \frac{(-17)^2}{5} - 20$$

= 5 + 7.2 + 288.8 + 578 – 20

= 338.8

$$SSR = \frac{5^2}{5} + \frac{(21)^2}{5} + \frac{(-14)^2}{5} + 0 + \frac{8^2}{4} - 20$$

= 161.5

SSE = SST – SSC – SSR = 584- 161.5 – 338.8 = 83.7

Source of variation	Degrees of freedom(df)	Sum of squares(SS)	Mean sum of squares(MS)	F-Ratio
Between treatments (columns) (Machines)	n – 1= 4-1 = 3	SSC = 338.8	MSC = 33.8/3 = 112.93	F= $\left(\frac{112.93}{8.98}\right)$ = 16.18 F distribution with (3,12))df
Between blocks (Rows) (Workers)	m-1= 5-1 = 4	SSR = 161.5	MSR = SSR/m-1 = 161.5/4 = 40.38	F= $\left(\frac{40.38}{8.98}\right) = 5.79$ F distribution with (4,12) df
Residual(Error)	(m-1)(n-1) = 12	SSE= 83.7	MSE = SSE/(m-1)(n-1) = 83.7/12= 8.98	------
Total	19	584	-------	----=-

Conclusion: Caculated value of Fc = 16.18 > F_T (3,12) table value 3.49

Calculated value of FR = 5.79 > F_T (4,12) table value 3.26

H_0 is rejected in both the cases. We conclude that there is a significant difference between the productivity of machines and workers.

5.2 Correlation Analysis

Correlation analysis is the study of measuring the degree of linear relationship between two variables. If the quantities (X,Y)vary in such a way that change in one variable corresponds to change in the other variable, then the variables X and Y are said to be correlated.

Examples: (i) Price and demand

(ii) Rainfall and yield.

(iii)Income and expenditure

The summay of correlation analysis is given in fig. 5.4

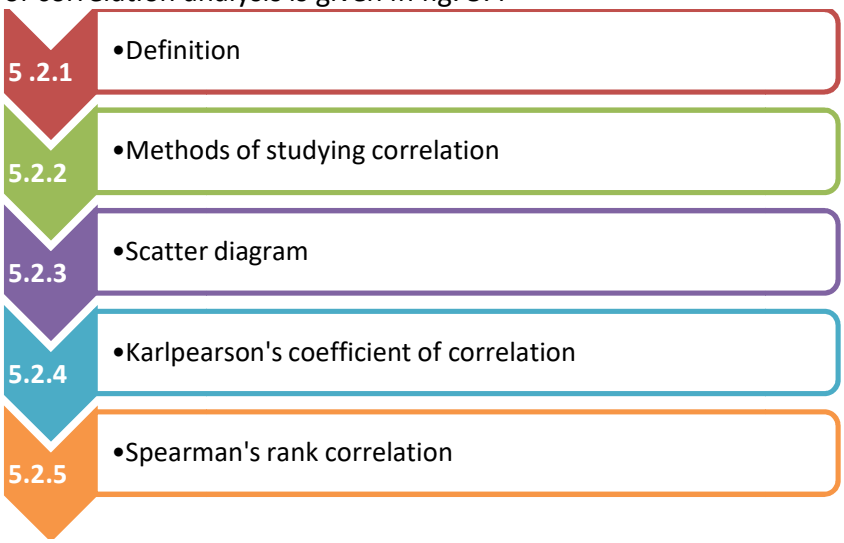


Fig.5.4 Summary of correlation analysis

Various methods to study correlation are mentioned in fig. 5.5

(i) Scatter diagram

(ii) Karl Pearson's coefficient of correlation

(iii) Spearman's rank correlation coefficient

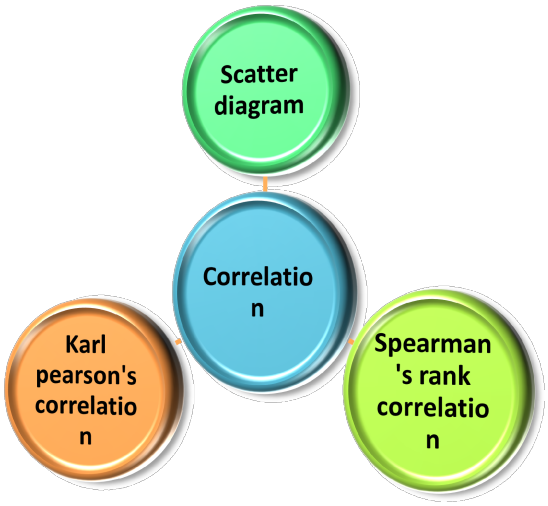


Fig.5.5 Methods of studying correlation

Scatter Diagram

A simple method to study correlation between two variables is a type of dot chart called scatter diagram. In this method we are plotting the points in the graph in the form of dots. For each pair of x and Y values we put dots and thus we obtain as many points as the number of observations. By observing to the scatter of various points we can form an idea as to whether the two variables are related or not. The plotted points scatter over a chart, the significance is the degree of relationship between the two variables. The dotted points are on the line, there is a higher degree of relationship. If the plotted points lie on the indiscriminate manner it shows the absence any relationship between the variables. The three types of scatter diagrams are given in fig. 5.6, 5.7, 5.8 & 5.9

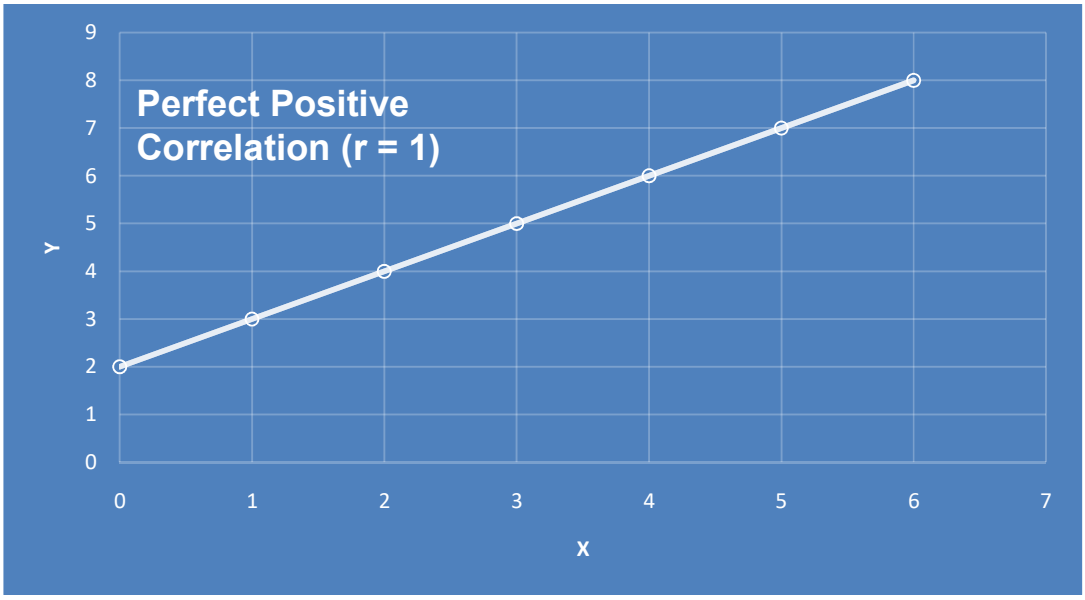


Fig.5.6 Scatter Plot-Perfect Positive Correlation

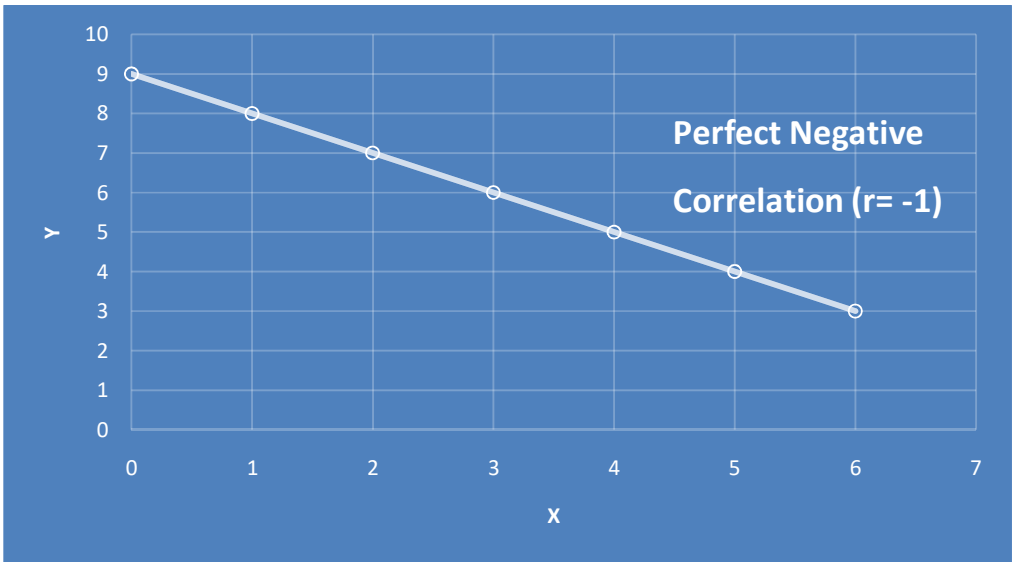


Fig.5.7 Scatter Plot-Perfect Negative Correlation

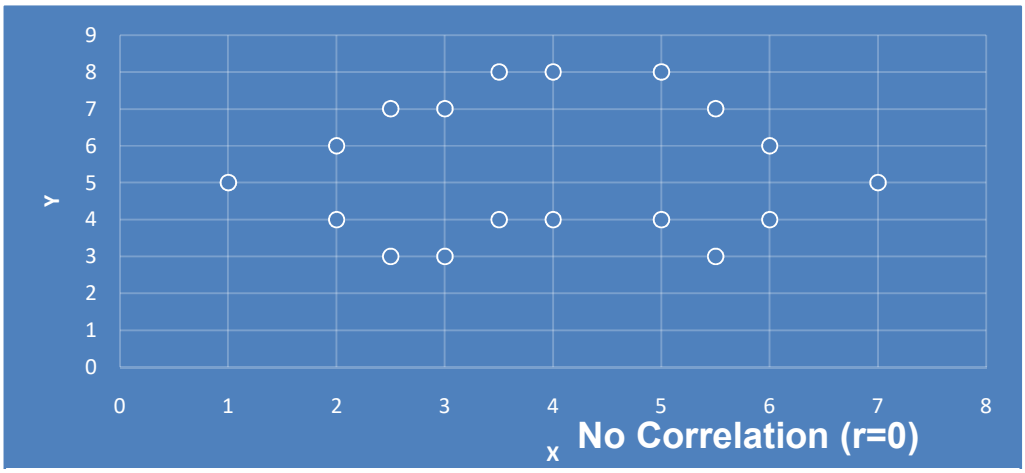


Fig.5.8 Scatter Plot-No Correlation

Properties of correlation coefficient:

(i) Correlation coefficient is lie between -1 and +1

(ie) $-1 \le r \le 1$

(ii) The coefficient of correlation is independent of change of scale and origin of the variables X and Y.

(iii) Two independent variables are uncorrelated.

Note:

(i) There is a perfect positive correlation if r=1

(ii) There is a perfect negative correlation if r=-1

(iii) No correlation if r=0

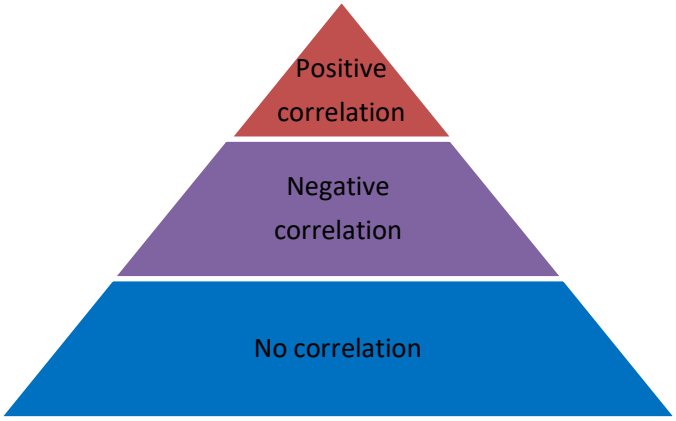


Fig.5.9 Types of Correlation

Karl Pearson's Coefficient of Correlation

The correlation coefficient between X and Y is defined as

$$r(X,Y) = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{N\sigma_x\sigma_y} \quad \text{(or)}$$

$$r(X,Y) = \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - \left(\sum x\right)^2}\sqrt{N\sum y^2 - \left(\sum y\right)^2}}$$ (or)

$$r(X,Y) = \frac{N\sum dxdy - \sum dx \sum dy}{\sqrt{N\sum dx^2 - \left(\sum dx\right)^2}\sqrt{N\sum dy^2 - \left(\sum dy\right)^2}}$$ where d represents deviation from mean

(dx = x- $\bar{x}$,dy = y - $\bar{y}$)

Illustration No. 4

Calculate the correlation coefficient for the following heights (in inches) of fathers(X) and their sons(Y)

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution

X	Y	dx=x-67	dy=y-68	dx^2	dy^2	dxdy
65	67	-2	-1	4	1	2
66	68	-1	0	1	0	0
67	65	0	-3	0	9	0
67	68	0	0	0	0	0
68	72	1	4	1	16	4
69	72	2	4	4	16	8
70	69	3	1	9	1	3
72	71	5	3	25	9	15
Total		8	8	44	52	32

$$r(X,Y) = \frac{N\sum dxdy - \sum dx \sum dy}{\sqrt{N\sum dx^2 - \left(\sum dx\right)^2}\sqrt{N\sum dy^2 - \left(\sum dy\right)^2}},$$

$$= \frac{8(32) - (8)(8)}{\sqrt{8(44) - (8)^2}\sqrt{8(52) - (8)^2}}$$

$$= \frac{256 - 64}{\sqrt{352 - 64}\sqrt{416 - 64}}$$

$$= \frac{192}{\sqrt{288}\sqrt{352}} = \frac{192}{318.3959}$$

= 0.603

Direct Method

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544	552	37028	38132	37560

$$\sum X = 544, \sum Y = 552, \sum X^2 = 37028, \sum Y^2 = 38132, \sum XY = 375609$$

N = 8

$$r(X,Y) = \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - \left(\sum x\right)^2}\sqrt{N\sum y^2 - \left(\sum y\right)^2}}$$

$$r(X,Y) = \frac{8(37560) - (544)(552)}{\sqrt{8(37028) - (544)^2}\sqrt{8(38132) - (552)^2}}$$

$$= \frac{300480 - 300288}{\sqrt{296224 - 295936}\sqrt{305056 - 304704}}$$

$$= \frac{192}{\sqrt{288}\sqrt{352}}$$

= 0.603

Illustration No. 5

Calculate Karl Pearson's coefficient of correlation from the following data:

(a) Sum of deviation of X = 5
(b) Sum of deviations of Y = 4
(c) Sum of squares of deviation of X= 40
(d) Sum of squares of deviation of y = 50
(e) Sum of squares of deviation X and Y = 32
(f) Number of pair of observation = 10

Solution

Given

$\sum dx = 5, \sum dy = 4, \sum dx^2 = 40, \sum dy^2 = 50, \sum dxdy = 32$ and N = 10

$$r(X,Y) = \frac{10(32) - (5)(4)}{\sqrt{10(40) - (5)^2}\sqrt{10(50) - (4)^2}}$$

$$= \frac{300}{\sqrt{375}\sqrt{484}}$$

= 0.704

Spearman's Rank correlation

This method is developed by the British psychologist Charles Edward Spearman in 1904.This method is useful when quantitative measures of certain factors (such as Obtaining marks in the semester exam, Judgement in beauty competition) cannot be fixed, but we arrange the performance of the individuals the number indicating the rank in the group. The rank correlation is applied to a set of normal ranks numbers, according to quantity or quality and so on.

Spearman's Rank correlation coefficient

$P = 1 - \frac{6\sum d^2}{N(N-1)}$, where ρ → Coefficient of Rank correlation

d→ deviation of the ranks between two observations

Repeated Ranks

In some problems it may be two or more ranks are same. In those cases the average rank is given to those individuals. In that case we calculate the correction factor and then find the rank correlation. The formula can be rewritten as

$$P = 1 - \frac{6\{\sum d^2 + C.F\}}{N(N^2-1)}$$

where correction factor C.F = $\left\{\frac{1}{12}\left(m(m^2-1)\right)+\frac{1}{12}\left(m(m^2-1)\right).....\right\}$

Illustration No. 6

alculate rank correlation coefficient from the following data:

X	48	33	40	10	16	16	65	24	16	57
Y	12	13	24	5	15	4	20	9	6	19

Solution

X	Y	Rx	Ry	d=Rx-Ry	d^2
48	12	3	6	-3	9
33	13	5	5	0	0
40	24	4	1	3	9
10	5	10	9	1	1
16	15	7	4	3	9
16	4	7	10	-3	9
65	20	1	2	-1	1
24	9	6	7	-1	1
16	6	7	8	-1	1
57	19	2	3	-1	1
Total				$\sum d^2$	41

Since the rank 7 is repeated 3 times in X series we find C.F(m = 3)

$$C.F = \frac{1}{12}\left(m(m^2-1)\right) = \frac{1}{12}\left(3(9-1)\right) = 2$$

$$P = 1 - \frac{6\{\sum d^2 + C.F\}}{N(N^2-1)} = 1 - \frac{6(41+2)}{10(100-1)}$$

= 1- 0.2606 = 0.7393

To Do Activity

Rank your favourite type of music with a 1.Continue with a ranking 2 for your second favourite and an 8 for your least favourite. Ties are not allowed. Once complete share your data with your partner. Does there seem to be an association (relation) between your ranking and partner's ranking. Explain.

5.3 Regression Analysis

The statistical tool to estimate or predict the unknown values of one variable from known values of another variable is called regression. Also using regression we are able to measure the average relationship between two or more variables in terms of original unit of data. For example the two variables price(X) and demand Y are closely related, using regression we are able to estimate the value of Y when X is given or to estimate the value of X when Y is known. To estimate the relationship between the stock market index and its relationship between the macro economic variables we use regression.

Relation between correlation and regression is given in fig. 5.10

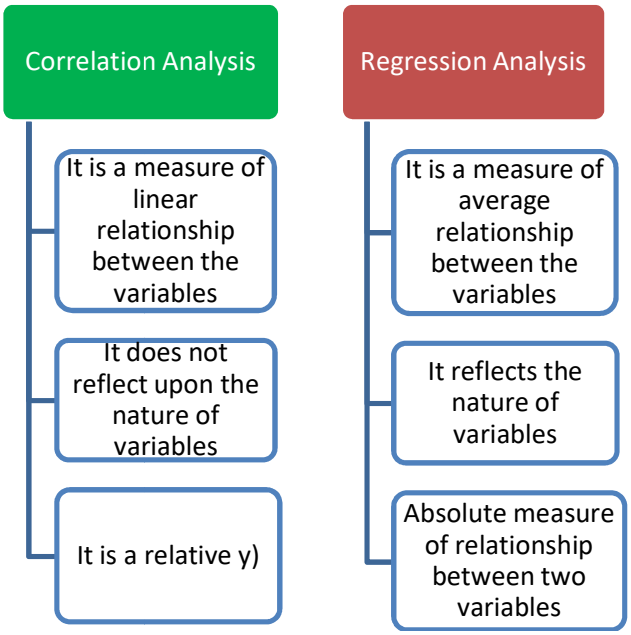


Fig.5.10 Relation between correlation and regression

Regression Lines

We have two regression lines (i) regression line of X on Y and (ii) regression line of Y on Using regression line of Y on X, we are able to estimate the probable value of Y when X is given and vice versa.

(i) The regression line of X on Y is $X - \overline{X} = r\frac{o_x}{\sigma_y}\left(Y - \overline{Y}\right)$

(ii) The regression line of Y on X is $Y - \overline{Y} = r\frac{o_y}{\sigma_x}\left(X - \overline{X}\right)$

where $r\frac{\sigma_x}{\sigma_y}$ and $r\frac{\sigma_y}{\sigma_x}$ are the regression coefficient of X on Y and Y on X

Regression Equations

The regression equation of Y on X is given as

Y = a + bX, where the constants a and b are determined by the normal equations

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

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

The above method is called method of least squares.

The regression equation of X on Y is given as

X = a + bY, where the constants a and b are determined by the normal equations

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

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

Properties of Regression Coefficients

(i) The regression lines are passes through their means $(\overline{X}, \overline{Y})$

(ii) Correlation coefficient r is the geometric mean of regression coefficients.

(i.e.) r = $\pm\sqrt{b_{xy}.b_{yx}}$ where $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

(iii) One of the regression coefficient, is greater than unity and other must be less than unity.

(iv) Both the regression coefficients have same sign.

(i.e.) the correlation coefficient r is positive if both the coefficients are positive and r is negative if both the regression coefficients are negative.

Note

(i) In case of numerical data we use the following formula to estimate regression coefficients.

$$b_{xy} = \frac{n\sum xy - \sum x\sum y}{n\sum y^2 - \left(\sum y\right)^2} \text{ and } b_{yx} = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

(ii) Angle between regression lines are given by $\tan\theta = \frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$

(iii) If r = 0 then the two regression lines are perpendicular.

(iv) If r = ±1 then the two regression lines are parallel or coincident.

Illustration No. 7

The following data relate to marketing expenditure in lakhs of rupees and the corresponding sales of a product in crores of rupees. Estimate the marketing expenditure to attain a sale of 40 crores.

Marketing Expenditure	10	12	15	20	23
Product Saless	14	17	23	21	25

Solution

Let x → Marketing expenditure

Y → Product sales

	x	y	x^2	y^2	xy
	10	14	100	196	140
	12	17	144	289	204
	15	23	225	529	345
	20	21	400	441	420
	23	25	529	625	575
Total	**80**	**100**	**1398**	**2080**	**1684**

$$\sum x = 80, \sum y = 100, \sum x^2 = 1398, \sum y^2 = 2080, \sum xy = 1684$$

and n = 5

The regression coefficient of X on Y is $b_{xy} = \dfrac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2}$

$$= \frac{5(1684) - (80)(100)}{5(2080) - (100)^2} = 1.05$$

$$\overline{X} = \frac{\sum x}{n} = \frac{80}{5} = 16$$

$$\overline{Y} = \frac{\sum y}{n} = \frac{100}{5} = 20$$

The regression line of X on Y is $X - \overline{X} = b_{xy}\left(Y - \overline{Y}\right)$

$$X - 16 = 1.05(Y - 20) \text{ or } X = 1\text{-}05\ Y - 5$$

∴ Marketing expenditure to attain a sale of 40 crores is the value of X when Y = 40

= 1.05(40) - 5 = 37 lakhs

Illustration No. 8 :

Given the following data

	X	Y
Arithmetic mean	36	85
Standard deviation	11	8

And correlation coefficient between x and y is 0.66

(i) Find the two regression lines.

(ii) Estimate the value of X when Y = 75 and estimate the value of Y when X= 40.

Solution

Given mean of X $= \overline{X}$ = 36

and mean of Y = $\overline{Y}$ = 85

Standard deviation of X = $\sigma_x = 11$

Standard deviation of Y = $\sigma_y = 8$

The regression line of X on Y is $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$

$$X - 36 = (0.66)\frac{11}{8}(Y - 85)$$

X = 0.9075Y - 41.1375

The regression line of Y on X is $Y - 85 = (0.66)\frac{8}{11}(X - 36)$

Y = 0.48X +

ii) When Y = 75 , X = 26.925 and when X = 40 , Y = 86.92

To Do Activity
Consider two variables BMI and DBP. The question is "Is body mass index a good predictor of diastolic blood pressure? Add a trend line your plot, and write down both the equations of your best fit line and correlation coefficient. Is the relationship between these variables are strong or weak? Positive or negative? What is the real meaning of r^2?

5.4 Forecasting and Time series Analysis

The most important task is estimate for the future. However, the first step in making estimates for the future consists of obtaining information from the past. In this connection one usually deals with statistical data which are called observed or recorded data at successive intervals of time. These data are referred to as "Time series "When we consider numerical data at different point of time the set of observations is known as time series. For example, if we collect the data for production, population, sales, imports, exports etc. At different point of time say yearly or 5 years or some period.

Definition

An arrangement of statistical data in accordance with time of occurrence is called a time series. According to Patterson," A time series consists of statistical data which are collected, recorded or observed over successive increments" According to Ya-Lun-Chou,"A time series may be defined as a collection of magnitudes belonging to different time periods, of some variables or composite of variables, such as production of steel, per capita income, gross national product, price of tobacco or index of industrial production".

Utility of Time Series

Time series analysis very helpful for economists, businessman, scientist, astronomist, geologist, sociologist and research worker, etc for the following reasons:

(i) It is very helpful to understand the past behaviour.
(ii) Helps to plan the future operations
(iii) Helps in evaluating current accomplishments.
(iv) It helps to make comparative studies.

Components of time series

There are four types of components of time series which is given in fig. 5.11

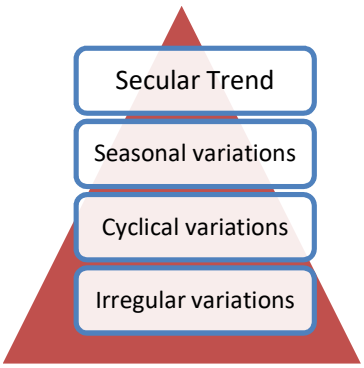


Fig 5.11 Components of Time Series

Secular Trend

The long-term trend is also called secular trend, is the basic tendency of a series to grow or decline over a period of time. It is a concept of trend does not include short range oscillations but rather the steady movement over the long term. The long period of time depends on the nature of problem under study, the result is better in the long period.

Seasonal Variations

If we study the irregular curve period by period, we see that in each period the curve starts with a low figure and reaches a peak about the middle of the year and then decreases again. This type of fluctuation, which completes the whole sequence of change within a span of period and has a same pattern period after period, is called a Seasonal variation. Seasonal variation occurs due to the following reasons.

(i) Climate and weather condition

Agriculture is influenced very much by the climate. The effect of the climate is that there are generally two seasons in agriculture, one is growing season and the other is harvesting season. These two are directly affecting the farmers' income.

(ii) Customs, tradition and habits

On the first of every month there are heavy withdrawals and the banks have to keep lots of money to meet the possible demand on the basis of last month experience and most of the students buy books in the first few months of the opening of schools and colleges and thus the sale of books, stationary, etc shows seasonal variations.

Cyclical Variations

Most of the time series relating to economics and business show some kind of cyclical variations and whose durations are more than one year. In spite of the importance of measuring cyclical variations, they are very difficult to measure due to the following reasons. A typical cyclical variation is given in fig. 5.12

(i) Business cycles do not show regular periodicity.

(ii) The cyclical variations are associated with erratic, random or irregular forces to make it impracticable to isolate separately the effect of cyclical and irregular forces.

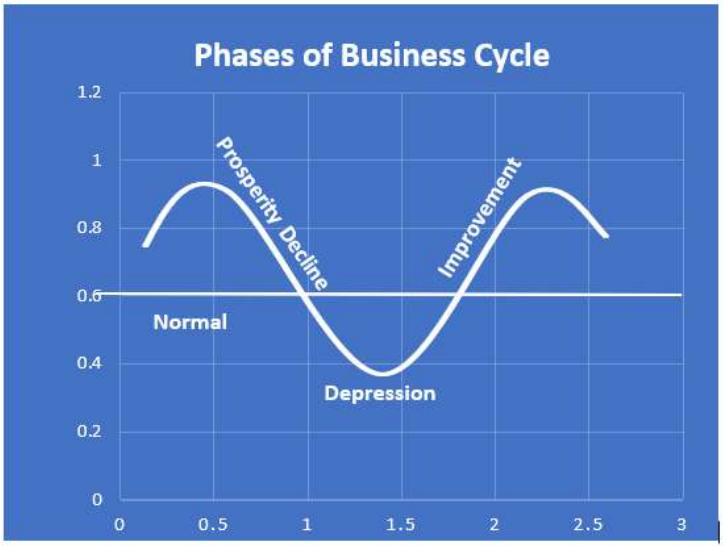


Fig 5.12 Cyclical Variations-Phases of a business cycle

Irregular Variations

These variations occur due to sudden causes and are unpredictable. In business activity these variations do not repeat in a definite pattern. Irregular movements are considered to be largely random. Irregular variations are caused by some special occurrences as floods, earthquakes, strikes and wars. Sudden change in demand is also included in this category.

Mathematical model for time series:

There are two models for analysing the time series.

(i) Additive model

(ii) Multiplicative model

Additive model

If Y_t is the time series value at time t, T_t, S_t, C_t and R_t are the trend value, seasonal, cyclic and random fluctuations at time t respectively. According to the additive model a time series can be expressed as

$Y_t = T_t + S_t + C_t + R_t$

This model assumes that four components of the time series act independently each other.

Multiplicative model

The multiplicative model assumes that various components in a time series operate proportionately to each other. According to this model

$Y_t = T_t \times S_t \times C_t \times R_t$

The differences between the two models of additive model and multiplicative model are given in fig. 5.13

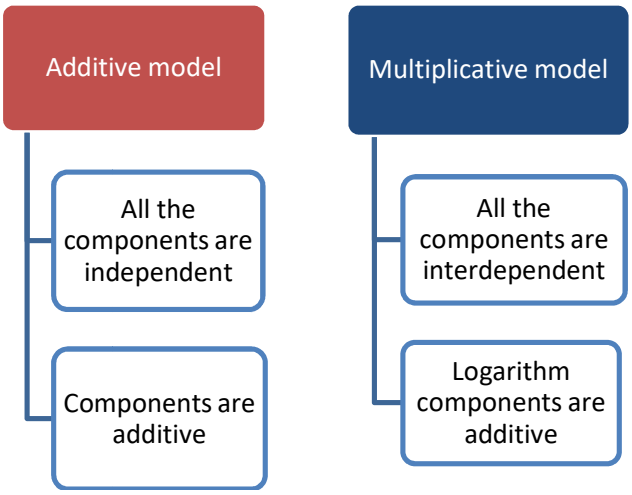


Fig 5.13 Differences between Additive & Multiplicative Models

Secular Trend

Trend can be calculated by the following methods and given in fig. 5.14

(a) Free hand or graphic method

(b) Method of semi averages

(c) Method of moving averages

(d) Method of least squares

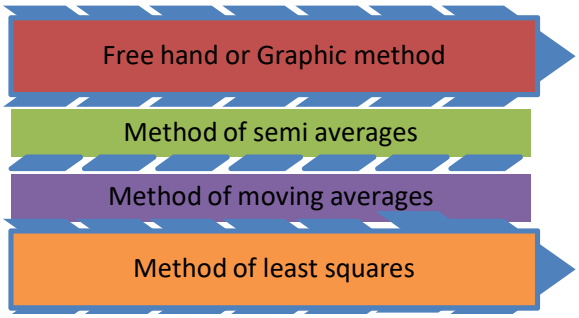


Fig 5.14 Methods to calculation of Trend

Free hand method or Graphic method

It is a simple method to estimate trend. To obtain the straight-line trend we follow the following procedure:

(i) Plot the points on the graph

(ii) Draw a straight line to best fit the data.

(iii) The curve should be smooth.

(iv) The number of points is equal above and below the line.

(v) The sum of the vertical deviations above and below curve are equal

(vi) The sum of the squares of vertical deviations from the curve is minimum.

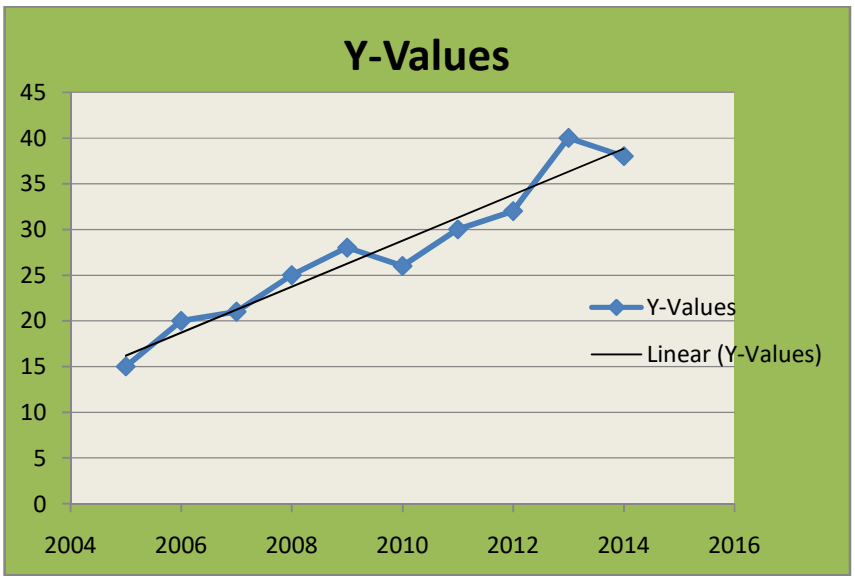
Illustration No. 9

Annual water consumption per household in a certain locality was reported below:

Years	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Water Used(units)	15	20	21	25	28	26	30	32	40	38

Fit a free hand smooth curve for the above data.

Solution



Method of Semi Averages

In this method the given data is divided in to two equal parts and averages of these two parts are calculated. These average values are plotted against the Midvale of the each part.The two points are joined by the lines which can be extended downward and upward.This line is called the trend line From this line we get trend values for other period.

Illustration No.10

Use the method of semi averages determine the trend values for the following data.

Year	2010	2011	2012	2013	2014	2015	2016
Sales(Units)	102	105	114	116	108	116	112

Solution

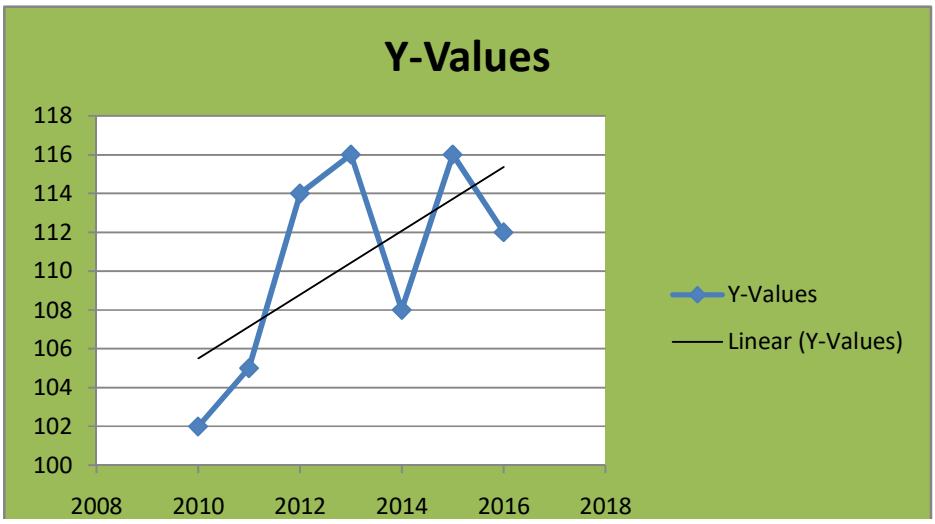
Since, seven years are given the middle year can be left out and average of first three years shall be obtained.

Average of first three years = $\frac{102+105+114}{3}=107$

Average of second three years = $\frac{108+116+112}{3}=112$

Now we get two points 107 and 112.These can be plotted corresponding to their respective middle year and joining these two points we get the required trend line.

The actual data and trend line are shown in the following graph.



Methods of Moving Averages

Another simple method to obtain trend values with a fair degree of accuracy by eliminating fluctuations. Here the average value of year or weak or day is taken for trend value and placed in the middle period of the period of moving average.

Odd Period Moving Average

Suppose the period is odd say 3 then we take the total of the first 3 items and place it against the second year. Then we take the total of second set of 3 items and place it against the 3 year. Continue this process till the last 3 years have been taken in to account.

Even Period Moving Averages

If the period of moving average is even (say 4 years) then the total of the first 4 items will be placed between the second and third year. The total of second set of 4 items will be placed between 3 and 4th year and so on. Then calculate moving total (i.e. add first two total placed against 3rd year. The total divided by 8 we get 4 year moving average. Continue the process till 4-year total.

Illustration No. 11

Draw a graph to represent the following data showing the number of farmers in a village:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
No of farmers	705	685	703	637	705	689	715	685	725	730

Calculate five yearly moving averages of the above data and plot them on the same chart.

Solution

Calculation of five yearly moving averages

Year	No of farmers	5-yearly moving total	5-yearly moving average
2009	705	--	----
2010	685	--	-----
2011	703	3485	697
2012	637	3469	693.8

2013	705	3499	699.8
2014	689	3481	696.2
2015	715	3519	703.8
2016	685	3544	708.8
2017	725	----	----
2018	730	-----	-----

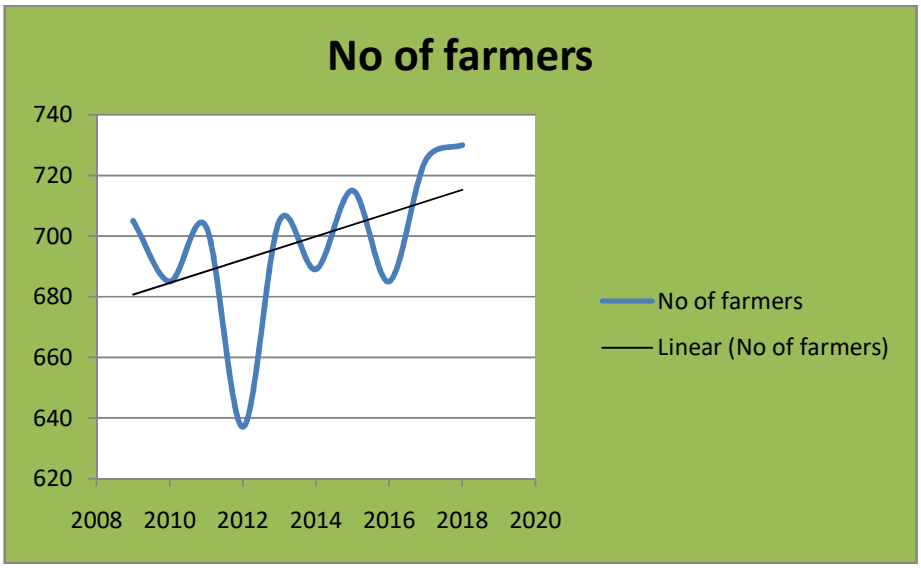


Illustration No. 12

Compute the trend by the method of moving averages, assuming 4 year cycle is present in the following series.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Sales	154	140.5	147	148.5	142.9	142.1	136.6	142.7	145.7	145.1	137.8

Solution

The four year moving averages are shown in the last column.

Year	Sales	4-year moving total	Centred moving total	4 Year moving average
2008	154			
2009	140.5			
		590		
2010	147		1168.9	146.11
		578.9		
2011	148.5		1159.4	144.93
		580.5		
2012	142.9		1150.6	143.83
		570.1		
2013	142.1		1134.4	141.8
		564.3		
2014	136.6		1131.4	141.43

		567.1		
2015	142.7		1137.2	142.15
		570.1		
2016	145.7		1141.4	142.68
		571.3		
2017	145.1			
2018	137.8			

Method of Least Squares

This method is most widely used in practice. For finding the trend line in this method the given data must satisfy the following conditions.

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

(i.e.) $\sum(Y - Y_c) = 0$

(ii) The sum of the squares of the deviations of the actual value and computed value is least.

(i.e.) $\sum(Y - Y_c)^2$.

The straight-line trend is represented by the equation Y = a + b X.

For evaluating the constants 'a' and' b' we use the following normal equations.

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

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

Suppose the time variable is measured as a deviation from mean, then ∑x = 0

Then the above two normal equations become,

$$\sum y = na \quad \& \sum xy = b\sum x^2$$

(i.e.) a = $\frac{\sum y}{n}$ and b= $\frac{\sum xy}{\sum x^2}$

Illustration No. 13

Fit a straight-line trend by the method of least squares to the following data. Use this trend line to estimate the earning for the year 2018.

Year	2009	2010	2011	2012	2013	2014	2015	2016
Earnings(lakhs)	38	40	65	72	69	60	87	95

Solution

Year	Earnings(Y)	d = X-2012.5	X=2d	XY	X^2
2009	38	-3.5	-7	-266	49
2010	40	-2.5	-5	-200	25
2011	65	-1.5	-3	-195	9
2012	72	-0.5	-1	-72	1
2013	69	0.5	1	69	1
2014	60	1.5	3	180	9
2015	87	2.5	5	435	25
2016	95	3.5	7	665	49
n=8	$\sum y = 526$		$\sum x = 0$	$\sum xy = 616$	$\sum x^2 = 168$

The trend line is Y = a + b X.

$$a = \frac{\sum y}{n} = \frac{526}{8} = 65.75$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{616}{168} = 3.67$$

The trend line is Y = 65.75 + 3.67 X

For 2018, X will be 11. When X = 11, Y = 106.12

Thus, estimated earnings for the year 2018 is **Rs.** 106.12 lakhs.

To Do Activity
Collect Olympic games data that relates to men's and women's performance in the 100m and high jump and display the data set graphically and verify random fluctuations on the data and the long term trend in 100m and also fit a trend line by using method of least squares.

Seasonal Variations

The variations that occur regularly and periodically with a period of less than one year are called as seasonal variations. Since seasonal variations affect economic and business and so knowledge of seasonal variations is necessary. For the following two purposes we study seasonal variation.

(a) To find out the effect of seasonal forces on a time series

(b) To eliminate the effect of seasonal forces from the time series.

The methods of seasonal variation are given in fig. 5.15

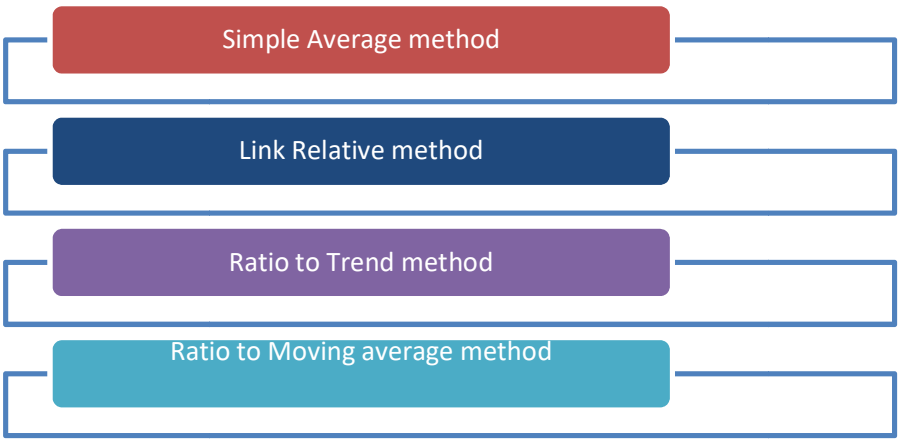


Fig 5.15 Methods of Seasonal Variation

Simple Average Method

This method is very simple to calculate the seasonal variations. For calculating seasonal indices, the following steps to be followed.

(i) Find the total of each period (month or Quarter)
(ii) Find the average of each period.(seasonal average)
(iii) Find the general average(average of average)
(iv) Calculate seasonal index by using the formula,

Seasonal index = $\frac{Seasonal\ average}{General\ average} x100$

Illustration No. 14

From the following data calculate quarterly seasonal indices assuming the absence of any type of trend.

Year	I	II	III	IV
2015	---	---	127	134
2016	130	122	122	132
2017	120	120	118	128
2018	126	116	121	130
2019	127	118	--	--

Solution

Calculation of quarterly seasonal indices.

	Year	I	II	III	IV
	2015	---	---	127	134
	2016	130	122	122	132
	2017	120	120	118	128
	2018	126	116	121	130
	2019	127	118	--	--
Total		503	476	488	524
Seasonal Averages		125.75	119	122	131
Seasonal index		101.05	95.6	98.04	105.03

General average = $\frac{125.75+119+122+131}{4} = \frac{497.75}{4} = 124.4375$

Seasonal index = $\frac{Seasonal\ average}{General\ average} x100$

Seasonal index for first quarter = $\frac{125.75}{124.4375} x100$ = 101.05

Seasonal index for second quarter = $\frac{19}{124.4375} x100$ = 95.6

Seasonal index for third quarter = $\frac{122}{124.4375} x100$ = 98.04

Seasonal index for fourth quarter = $\frac{131}{124.4375} x100$ = 105.03

Link relative method

In this method the following steps to be followed for calculating seasonal indices. Here we calculate the link relatives of seasonal figures.

$$\text{Link relative} = \frac{Current\ season\ figure}{\text{Pr}\ evious\ season\ figure} x100$$

Calculate link relative for each season. Convert these averages in to chain relatives on the basis of first season. Calculate the chain relative for the first season base of the last season. There will be some difference between the chain relative of the first season s and the chain relative calculated by the previous method. This difference will be due to effect of long term changes. For correction the chin relative of the first season calculated by first method is deducted from the chain relative calculated by second method. Then express corrected chain relatives as percentage of averages.

Illustration No. 15

Calculate the seasonal indices by the Link relative method for the data given below.

Year	I	II	III	IV
2014	60	65	62	69
2015	62	68	65	68
2016	65	70	64	62
2017	70	75	68	67
2018	72	80	70	78

Solution

The calculations are made by using the formula

Link relative of the quarter = $\frac{Current\ season\ figure}{\Pr evious\ season\ figure} x100$

The total of link relatives is found for each quarter and the mean is found by dividing the total for each quarter by the number of values.

The chain relative for a quarter is given by

$$\frac{Link\ relative\ of\ the\ quarter\ x\ Chain\ relative\ of\ previous\ quarter}{100}$$

The chain relative for the first quarter is taken to be 100.

Year	I	II	III	IV
2014	---	108.3	95.4	111.3
2015	89.9	109.7	95.6	104.6
2016	95.6	107.7	90.7	98.5
2017	112.9	107.1	90.7	98.5
2018	107.5	111.1	87.5	111.4
Total of link relatives	405.9	543.9	460.6	522.7
Arithmetic mean of link relatives	101.48	108.78	92.12	104.54
Chain relatives	100	108.78	100.216	104.716
Adjusted chain relatives	100	107.2	97.05	100.02
Seasonal index	98.94	106.07	96.02	98.96

Ratio to Trend Method

This method based on multiple methods of time series. In this we use the following steps. We calculate trend values for various time duration (Monthly or quarterly) with the help of least square method. Then we express all original data as the percentage of trend on the basis of the following formula.

$$\frac{Original\ data}{Trend\ value} x100$$

Remaining process is same as moving average method.

To Do Activity
Collect the data of average daily sales of soft drink in a milky bar for four seasons in the last four years. Calculate the seasonal indices by using simple average method.

5.5 Interpolation and Extrapolation

Many times, in practical work we come across situations where we have to estimate value which is not available in the given series or predict a future value.

The techniques of interpolation and extrapolation are extremely helpful in estimating the missing values or projecting the future values. Interpolation refers to the insertion of an intermediate value in a series of items whereas extrapolation refers to projecting a value for the future. Interpolation supplies us with the missing link whereas extrapolation helps in forecasting.

5.5 Interpolation and Extrapolation

The tools of interpolation and extrapolation are of great practical use. Their uses shall be listed below.

(1) It often happens that a particular type of information is being collected at regular intervals such as the census data. Now suppose we need the population figure for 2018 0r 2019 it would be impossible to conduct a census for these years. The only alternative is to make use of the technique of interpolation.

(2) The technique of interpolation is also used where a part of the data is destroyed or missing. The records may either be missing or may be lost. Such a figure may be obtained with the help of interpolation. Interpolation is thus helpful in filling up the gaps in available data.

The accuracy of interpolation depends upon

(1) Knowledge of the possible fluctuations of the figures to be obtained by a general inspection of the fluctuations at dates for which they are given.

(2) On knowledge of the course of events with which the figures are connected.

Assumptions

The following assumptions are made while making use of the techniques of interpolation and extrapolation.

(1) There are no sudden jumps in the series from one period to another.

(2) Another assumption that we make while interpolating or extrapolating values is that *the rate of change of figure from the period to another in uniform*.

Methods of Interpolation are shown in fig. 5.16

The various methods of interpolation can be divided under two heads:

(1) Graphic Method and

(2) Algebraic Methods

Under the head algebraic methods, we have several formulae. The following are some of the important and more popular methods:

(1) Binomial Expansion Methods
(2) Newton's Methods
(3) Lagrange's Method
(4) Parabolic Curve Method

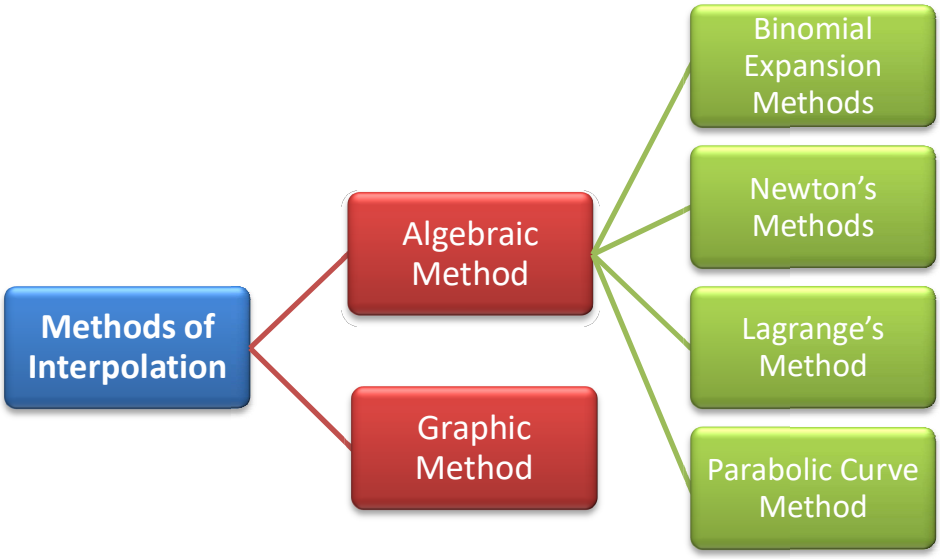


Fig 5.16 Methods of Interpolation

Graphic Method

This is the simplest of all the methods of interpolation. When this method is used the given data are plotted on a graph paper and the plotted points are joined. When there are only two values, we shall get a straight line otherwise a curve shall be obtained. On the X-axis we take the years and on the Y-axis the values of the variable. For the period for which the value is to be interpolated a perpendicular is drawn on the line (or curve). The point where it meets the line another perpendicular is drawn on the Y-axis. The corresponding value of the variable is read which the required value is.

Binomial Expansion Method

This method is applicable only in those situations where the following two conditions are satisfied,

(1) The x-variable advances by equal intervals.
(2) The value of x for which y is to be interpolated id one of the class limits of x series.

Newton's Method

A number of formulae were given by Newton to be applied in different situations. Some of these formulae are:

(1) Newton's Advancing Difference Method.
(2) Newton's Gauss (Forward) Method.
(3) Newton's Gauss (Backward) Method.
(4) Newton's Divided Difference Formulae.

Newton's Advancing Difference Method

This method is applicable in those cases where the independent variable x increases by equal intervals. The formula for interpolation is:

$$y_x = y_0 + x\Delta^1_0 + \frac{x(x-1)}{2!}\Delta^2_0 + \frac{x(x-1)(x-2)}{3!}\Delta^3_0 + \frac{x(x-1)(x-2)(x-3)}{4!}\Delta^4_0 + \ldots\ldots\ldots\ldots\ldots$$

where y_0, represents the value of *y* at origin.

y_x represents the figure to be interpolated, Δ's are the differences. The value of *x* is obtained as follows,

$$x = \frac{The\ Value\ to\ b\ e\ interpolated\ -\ The\ value\ at\ origin}{Difference\ between\ the\ two\ adjoining\ values}$$

Newton's Gauss (Forward) Method

This method is to be used when the following conditions are satisfied:

(1) When the independent variable (*X*) advances by equal intervals.

(2) When the value of dependent variable (*Y*) is to be interpolated for such value of *X* which is in the *middle*.

The formula used is:

$$y_x = y_0 + x\Delta^1 y_0 + \frac{x(x-1)}{2!}\Delta^2 y_{-1} + \frac{x(x-1)(x-2)}{3!}\Delta^3 y_{-2} + \frac{x(x-1)(x-2)(x-3)}{4!}\Delta^4 y_{-3} + \ldots\ldots\ldots\ldots\ldots$$

$$x = \frac{Interpolation\ item - Preceding\ item}{Difference\ between\ adjoining\ items}$$

Newton's Gauss (Backward) Method

The method is to be used when the value of the independent variable *X* advances by unequal intervals. The formula is

$$y_x = y_0 + (x - x_0)\,\Delta^1_0 + (x - x_0)(x - x_1)\,\Delta^2_0 + (x - x_0)(x - x_1)(x - x_2)\,\Delta^3_0 + \ldots\ldots\ldots\ldots\ldots$$

Δ^1_0 ,$\Delta^2_0$and $\Delta^3_0$are the first, second and third leading divided differences respectively.

Lagrange's Method

This method given by famous French Mathematician Lagrange is to be applied in those cases where x series advances by unequal intervals. The formula given by him is as follows:

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3)\ldots\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\ldots\ldots(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)\ldots\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)\ldots\ldots(x_1-x_n)}$$
$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3)\ldots\ldots(x-x_n)}{(x_2-x_0)(x_2-x_1)(x_1-x_3)\ldots\ldots(x_2-x_n)} + \ldots\ldots\ldots\ldots\ldots y_n \frac{(x-x_0)(x-x_1)(x-x_2)\ldots\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)(x_n-x_2)\ldots\ldots(x_n-x_{n-2})}$$

where x_0, x_1, x_2, etc., are the given values of x variable and y_0, y_1, y_2, etc., are corresponding values of y variable; y_n is the figure to be interpolated.

Parabolic Curve Method

This method of interpolation has the advantage of universal application, i.e., it can be applied to all types of problems on interpolation. When this method is applied one variable is taken as dependent and

another independent. The dependent variable is denoted by *y* and the independent one by *x*.
The equation of this curve is

$$y = a + bx + cx^2 + dx^3 + ...nx^n$$

This is curve of the n^{th} order.

Extrapolation

Extrapolation refers to estimating a value for future period. In order to extrapolate a particular value the various methods discussed above for interpolation can be adopted. The choice of a particular method would depend upon:

(a) Requirement of the question, and
(b) The nature of the given data.

Illustration No. 16

If $y_3=2, y_4=-6, y_5=8, y_6=9$ and $y_7=17$ then find $\Delta^4 y_3$.

Solution

Given $y_3=2, y_4=-6, y_5=8, y_6=9$ and $y_7=17$

$\Delta^4 y_3 = (E-1)^4 y_3$

Using Pascal triangle expand $(E-1)^4$.

$\Delta^4 y_3 = (E^4-4E^3+6E^2-4E^1+1)y_3$

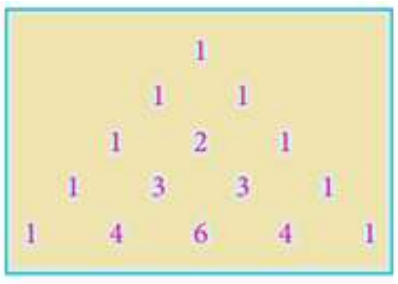
$= E^4 y_3-4E^3 y_3+6E^2 y_3-4E^1 y_3+ y_3$

$= y_7-4y_6+6y_5-4y_4+y_3$

$= 17-4(9)+6(8)-4(6)+2 = 55$

Illustration No. 17

From the following table find the number of farmers who yield the earnings less than 45 units:



Solution

Earnings	30-40	40-50	50-60	60-70	70-80
No of farmers	31	42	51	35	31

Let x ,y be the earnings and number of farmers respectively. The difference table is given below.

X	Y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
Less than 40	31				
		42			
50	73		9		
		51		-25	
60	124		-16		37
		35		12	
70	159		-4		
		31			
80	190				

$$y_x = y_0 + n\Delta^1 y_0 + \frac{n(n-1)}{2!}\Delta^2 y_0 + \frac{n(n-1)(n-2)}{3!}\Delta^3 y_0 + ...$$

To find y at x = 45, $x_0+nh=45$, $x_0 = 40$, h = 10 , n→1/2

$$y_{x=45} = 31+ \frac{1}{2}(42) + \frac{\frac{1}{2}(\frac{1}{2}-1)}{2!}(9) + \frac{\frac{1}{2}(\frac{1}{2}-1)(\frac{1}{2}-2)}{3!}(-25) + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)\left(\frac{1}{2}-2\right)\left(\frac{1}{2}-3\right)}{4!}(37)$$

$$= 31+21-\frac{9}{8}-\frac{25}{16}-\frac{555}{128} = 47.867 \approx 48$$

Illustration No. 18 :

Using Lagrange's interpolation formula find y(8) and y(12) from the following data.

X	5	6	9	11
Y	12	13	14	16

Solution

Here the intervals are unequal, so we use Lagrange's interpolation formula,

Given $X_0=5, x_1=6, x_2=9, x_3=11$ and $y_0=12, y_1=13, y_2=14, y_3=16$

$$y_x = y_0 \frac{(x-x_1)(x-x_2)(x-x_3).....(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3).....(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3).....(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3).....(x_1-x_n)}$$

$$+ y_2 \frac{(x-x_0)(x-x_1)(x-x_3).....(x-x_n)}{(x_2-x_0)(x_2-x_1)(x_1-x_3).....(x_2-x_n)} + ... y_n \frac{(x-x_0)(x-x_1)(x-x_2).....(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)(x_n-x_2).....(x_n-x_{n-2})}$$

$$y = \frac{(x-6)(x-9)(x-11)}{(5-6)(5-9)(5-11)}(12) + \frac{(x-5)(x-9)(x-11)}{(6-5)(6-9)(6-11)}(13) + \frac{(x-5)(x-6)(x-11)}{(9-5)(9-6)(9-11)}(14) + \frac{(x-5)(x-6)(x-9)}{(11-5)(11-6)(11-9)}(16)$$

Put x= 8 and x=12, we get

$$Y_{x=8} = \frac{(8-6)(8-9)(8-11)}{(5-6)(5-9)(5-11)}(12) + \frac{(8-5)(8-9)(8-11)}{(6-5)(6-9)(6-11)}(13) + \frac{(8-5)(8-6)(8-11)}{(9-5)(9-6)(9-11)}(14) + \frac{(8-5)(8-6)(8-9)}{(11-5)(11-6)(11-9)}(16)$$

$$= \frac{72}{-24} + \frac{117}{15} - \frac{252}{24} - \frac{96}{60} = -7.3$$

$$Y_{x=12} = \frac{(12-6)(12-9)(12-11)}{(5-6)(5-9)(5-11)}(12) + \frac{(12-5)(12-9)(12-11)}{(6-5)(6-9)(6-11)}(13) + \frac{(12-5)(12-6)(12-11)}{(9-5)(9-6)(9-11)}(14) + \frac{(12-5)(12-6)(12-9)}{(11-5)(11-6)(11-9)}(16)$$

$$= \frac{216}{-24} + \frac{273}{15} + \frac{588}{24} + \frac{2016}{60} = 67.3$$

Summary

In this chapter, we have learnt the concept of Analysis of variance and its and its application. After studying time series, you are able to predict the future values based on the previously observed values and able to identify the fluctuations in economics and business. Using correlation analysis, you are able

to study the economic behaviour and to study the progressive development in the methods of science has been increased the knowledge of correlation. Using regression, you are able to interpolate any value.

Model Questions

1. Explain the basic principles of experimental design.
2. When do you apply the analysis of variance technique?
3. What is meant by a completely randomized design?
4. State two differences between C.R.D and R.B.D
5. Write down the format of the ANOVA table for two factors of classification.
6. What is a scatter diagram?
7. Distinguish between correlation and regression.
8. What is time series? Give two examples.
9. Name any three forecasting methods used in time series analysis.
10. Four different, though supposedly equivalent forms of a standardized reading achievement test were given to each of 5 students, and the following are the scores which they obtained.

	Student 1	Student 2	Student 3	Student 4	Student 5
Form A	75	73	59	69	84
Form B	83	72	56	70	92
Form C	86	61	53	72	88
Form D	73	67	62	79	95

Perform a two-way analysis of variance to test at the level of significance $\alpha = 0.01$.

Ans= $F_{0.001}(12,3) = 27.1$ $F_{0.001}(4,12) = 5.41$

Conclusion:

$F_1 > F_{0.01}(12,3)$. Hence accept H_{01}. There is no significant difference between forms.

$F_2 > F_{0.01}(4,12)$. Hence reject H_{02}, we conclude that there is significant difference between students.)

11. Analyze the R.B.D at 5% level of significance.

	Variety		
Treatment	1	2	3
1	8	10	12
2	2	6	7
3	4	10	9
4	3	5	9

Ans = $F_{0.05}(3,6) = 4.76$ $F_{0.05}(2,6) = 5.41$

Conclusion:

$F_1 > F_{0.05}(3,6)$. Hence accept H_{01}. There is no significant difference between treatments.

$F_2 > F_{0.05}(2,6)$. Hence reject H_{02}, we conclude that there is significant difference between varieties.)

12. An experiment was designed to study the performance of 4 different detergents for cleaning fuel injectors. The following "cleanness" readings was obtained with specially designed equipment for 12 tanks of gas distributed over 3 different models of engines.

Detergent	Engine 1	Engine 2	Engine 3	Engine 4
A	45	43	51	139

B	47	46	52	145
C	48	50	55	153
D	42	37	49	128
Total	182	176	207	565

Test at the 0.01 level of significance whether there are differences in the detergents or in the engines.

Ans = $F_{0.01}(3,6) = 9.78$ $F_{0.01}(2,6) = 10.9$

Conclusion:

F_1>9.78. Hence there is significant difference between detergents.

F_2>10.9. Hence there is significant difference between engines.)

13. Using Karl Pearson's method, calculate the coefficient of correlation between density of population and death rate.

City	Area in Sq. Kilometers	Population in'000	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

Ans : r = 0.9875, i.e., there is very high positive correlation between density of population and death rate.

14. From the following data, calculate the coefficient of correlation between X and Y series.
Mean of X series = 75
Assumed mean of X series = 70
Mean of Y series = 126
Assumed mean of Y series = 113
Standard deviation of X series = 13.5
Standard deviation of Y series = 15.8
Sum of products of corresponding deviations of X and Y series = 2186
No. of pairs = 8

Ans r = 0.9763

15. A computer, while calculating the correlation coefficient between x and y from 25 pairs of observations, obtained the following:

$n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508$

It was however, later discovered at the time of checking that they had copied down two pairs as (6,14) and (8,6). Obtain the correct value of the correlation coefficient.

Ans: $r_{xy} = 0.667$

16. The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.

Marks\Age in Years	18	19	20	21	Total
10 - 20	4	2	2	-	8
20 - 30	5	4	6	4	19
30 - 40	6	8	10	11	35
40 - 50	4	4	6	8	22
50 - 60	-	2	4	4	10
60 - 70	-	2	3	1	6
Total	19	22	31	28	100

Calculate the correlation.

Ans: r = 0.2566

17. Find the two regression equations from the following data:

X:	57	58	59	59	60	61	62	64
Y:	77	78	75	78	82	82	79	81

Also estimate the value of Y when X is 65.

Ans: (i) X = 0.545Y + 16.945 and Y = 0.667X + 28.98 (ii) Y = 82

18. Fit a linear trend equation to the following data and estimate the value of sales for the year 2020.

Year :	2014	2015	2016	2017	2018
Sales : (in lakhs of Rs.)	100	120	140	160	180

Ans: Y_e= 140 + 20X, Y_{2020} = 220 Lakhs

19. Find seasonal variations by Ratio to trend method from the data given below.

	Quarters			
Year:	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
2014:	60	80	72	68
2015:	68	104	100	68
2016:	80	116	108	96
2017:	108	152	136	124
2018:	160	184	172	164

Multiple Choice Questions

1. Cyclical movements are due to
   a) Ratio to trend b) seasonal c) trend d) trade cycle
2. A time series is a set of data recorded
   a)Periodically b) at equal time intervals c) at successive point of time d) all the above
3. An additive model of time series with components T,S,C and I is
   a) Y = T x S x C x I b) Y = T+ S+ C+ I c) Y = T + S x C + I d) Y = T x S + C x I
4. In the least square theory, the sum of squares of residuals is
   a)Zero b) minimum c) constant d)maximum

5. In one-way classification the total variation can be split in to
   a) Two components b) Three components c) Four components d) Only one component
6. In two-way classification the total variation is TSS
   a)SST+SSE+SSB b) SST-SSE+SSB c) SST+SSE-SSB d) SST-+SSB
7. In two way classification with five treatments and four blocks the degrees of freedom due to error is a) 12 b) 19c) 16 d) 15
8. -------- measures the degree of relationship between two variables
   a)Standard deviation b) correlation coefficientc) mode d) median
9. The linear equation Y = a + bX is called a regression line of
   a) Y on X b) X on Y c) between X and Y d) ' a' on 'b'
10. Using regression co efficient we can calculate
    a)Cov(X,Y) b) SD of (X,Y) c) Correlation coefficient d) coefficient of variation
11. E = a) 1+ Δ b) 1-Δ c) 1+ ∇ d) 1- ∇
12. Lagrange's linear interpolation formula can be used for
    a)Equal intervals only b) unequal intervals only
    c) both equal and unequal intervals d) none of these

Answers for MCQ

1	2	3	4	5	6	7	8	9	10	11	12
b	b	b	b	b	a	b	b	a	c	a	b

References

1. Sharma, J.K. (2014). *Business Statistics – Problems and Solutions*. New Delhi : Vikas Publishing House Pvt Ltd.
2. Pillai, R.S.N. & Bagavathi, V. (1999). *Statistics*. New Delhi :S.Chand& Company Ltd.
3. Gupta, S.P. (2010). *Statistical Methods.* New Delhi :S.Chand& Company Ltd.
4. Beri, G.C. (2011). *Business Statistics.* New Delhi : Tata McGraw Hill Educations Pvt Ltd.
5. Foster, D. & Stine, E.R. (2010). *Statistics for Business : Decision Making and Analysis.* New Delhi : Pearson Publishers.
6. Gupta, S.C. & Kapoor, V.K. (2006). Fundamentals of Mathematical *Statistics.* New Delhi :S.Chand& Company Ltd.
7. Srivastava, S.C & Srivastava, S. (2003). Fundamentals of Statistics. New Delhi : Anmol Publications Pvt. Ltd.

Editors' Profile

Dr W G Prasanna Kumar

Dr.W. G. Prasanna Kumar, PhD in Education with basic degree in Social Work and Master's Degrees in Sociology, Public Administration and Political Science has professional education in Environmental Economics, Public Relations, Communication and Training and Development. Presently Chairman, Mahatma Gandhi National Council of Rural Education (MGNCRE) under the Ministry of Human Resource Development, in Government of India strives to promote resilient rural India through Higher Education interventions. The national initiative of reviving Mahatma Gandhi's ideas of NaiTalim, spearheaded by Dr. W G Prasanna Kumar, has met unprecedented success at both national and state levels. The primary objective of this initiative is to promote Gandhiji's ideas on Experiential Learning, NaiTalim, Work Education and Community Engagement, and mainstreaming them in School Education and Teacher Education Curriculum & Pedagogy. As Professor and Head Centre for Climate Education and Disaster Management in Dr MCR HRD Institute, conducted several capacity building and action research programmes in climate education, disaster management and crowd management. He has handled many regional, national and international environmental education programmes and events including UN CoP11 to Convention on Biological Diversity and Media Information Management on Environmental Issues.

He was Director in National Green Corps in the State Government for over 11 years and Senior Social Scientist in State Pollution Control Board for 6 years. Conducted various curriculum and non- curriculum related training programmes in environmental education. He was a Resource Person for AP Judicial Academy, AP Police Academy, AP Forest Academy, EPTRI, Commissonerate of Higher Education and Intermediate Education, State Council for Educational Research and Training and National Council for Educational Research and Training New Delhi, CCRT, BharathiyaVidyapeet University Pune, CPR Environmental Education Centre Chennai and Centre for Environment Education Ahmedabad. Dr W G Prasanna Kumar was trained in Community Consultation for Developmental Projects in EPA Victoria Australia in 1997 trained as State Chief Information Officer by IIM Ahmedabad and MCRHRDI Government of Andhra Pradesh in 2004 and trained in Environmental Education and Waste Management Technique by JICA, Japan in 2011.

He was awarded Best State Nodal Officer of National Green Corps Award from Centre for Science and Environment, New Delhi, 2008, Jal Mithra Award from Earthwatch Institute of India and Water Aid New Delhi, 2014 and Certificate of Commendation for the services in UN Conference of Parties to Convention for Biodiversity conducted at Hyderabad from 1-20 October 2012 by the Government of Andhra Pradesh 2012.

Dr K N Rekha

Dr K N Rekha, is a PhD Graduate from IIT Madras. She has 14 years of experience in training and education Industry. She works at Mahatma Gandhi National Council of Rural Education (MGNCRE), Hyderabad as Senior Faculty. She is involved in curriculum development on Rural Management and Waste Management. Prior to this, she worked as a researcher at Indian School of Business, Hyderabad, a short stint at Centre for Organisation Development (COD), Hyderabad. She has co-authored a book on "Introduction to Mentoring", written book chapters, peer reviewed research papers, book reviews, Case studies, and caselets in the area of HR/OB. She also presented papers in various national and

international conferences. Her research areas include Mentoring, Leadership, Change Management, and Coaching. She was also invited as a guest speaker at prominent institutions like IIT Hyderabad.

Authors' Profile

Dr. K. Duraiveluis a Professor of Mechanical Engineering and Dean of Faculty of Engineering & Technology, Vadapalani Campus, SRM Institute of Science and Technology, Chennai. He received his Ph.D from Faculty of Mechanical Engineering, College of Engineering-Guindy, Anna University, Chennai. He is also qualified with B.E degree in Mechanical Engineering from Institute of Road Transport Technology, Bharathiar University, Coimbatore, M.E degree in Quality Engineering & Management from Birla Institute of Technology, Mesra, M.B.A degree from Indira Gandhi National Open University. He has served as a Principal in a few private engineering colleges affiliated to Anna University. He has served as a Head, Department of Engineering at Ibri College of Technology, Oman. He has around 26 years of experience in industry and in academia, in India and abroad. He has also authored textbooks on Basic Mechanical Engineering, Engineering Mechanics, Operations Research and Engineering Metrology & Measurement. He has published a good number of research articles in refereed journals.

Dr.R.Manimaranis an Assistant Professor in Mathematics at SRM Institute of Science and Technology,Vadapalani campus. He received his PhD (Stochastic Process and Queuing Theory) in 2014. He received his bachelor's degree from Madurai Kamaraj University in 1986 and post graduate degree in 1988. He received his M.Phil degree from Alagappa University in 2004. He has more than 25 years of experience in teaching. He has published a book for Indira Gandhi National Open University on Production Management. He is specialised in mentoring the students in the most moral possible way by trying to teach the subject from his heart and guiding the students in the right way to utilize the concept aptly in their respective fields. He is an active member of the London Mathematical Association, The Association of Mathematics Teachers of India, Indian Society of Industrial and Applied Mathematics and the Indian Mathematical Society.

सत्यमेव जयते

Mahatma Gandhi National Council of Rural Education (MGNCRE)

Department of Higher Education

Ministry of Human Resource Development, Government of India

Where there is Rural Wellbeing
there is Universal Prosperity

Mahatma Gandhi National Council of Rural Education

Department of Higher Education

Ministry of Education, Government of India

O40 - 2321 2120

admin@mgncre.in
www.mgncre.in

#5-10-174, Shakkar Bhavan, Fateh Maidan Lane
Band Colony, Basheer Bagh,
Hyderabad-500004