UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION

M.Com

(I SEMESTER)
(2019 Admn Onwards)

CORE COURSE : MCM1C03

QUANTITATIVE TECHNIQUES FOR BUSINESS DECISIONS

190603

QUANTITATIVE TECHNIQUES FOR BUSINESS DECISIONS.

STUDY MATERIAL FIRST SEMESTER

CORE COURSE : MCM1C03

M.Com
(2019 Admn Onwards)

UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION
Calicut University- PO, Malappuram,

Kerala, India - 673 635

QUANTITATIVE TECHNIQUES FOR BUSINESS DECISIONS

M.COM FIRST SEMESTER

CONTENTS

CHAPTER 1

INTRODUCTION TO QUANTITATIVE TECHNIQUES

Quantitative technique is a very powerful tool with the help of which, the business organizations can augment their production, maximize profits, minimize costs, and production methods can be oriented for the accomplishment of certain pre – determined objectives. Quantitative techniques are used to solve many of the problems that arise in a business or industrial area. A large number of business problems, in the relatively recent past, have been given a quantitative representation with considerable degree of success. All this has attracted the business executives, public administrators alike towards the study of these techniques more and more in the present times.

Managerial activities have become complex and it is necessary to make right decisions to avoid heavy losses. Whether it is a manufacturing unit, or a service organization, the resources have to be utilized to its maximum in an efficient manner. The future is clouded with uncertainty and fast changing, and decision- making – a crucial activity – cannot be made on a trial-and-error basis or by using a thumb rule approach. In such situations, there is a greater need for applying scientific methods to decision-making to increase the probability of coming up with good decisions. Quantitative Technique is a scientific approach to managerial decision-making. The successful use of Quantitative Technique for management would help the organization in solving complex problems on time, with greater accuracy and in the most economical way.

Definitions

Since Quantitative technique is a practical methodological technique, there is no precise definition for the term. Quantitative techniques are defined as "those statistical techniques which lead to numerical analysis of variables, affecting a decision situation, and evaluation of alternative strategies to attain objectives of organizations."

Quantitative techniques involves " transformation of a qualitative description of a decision situation, into quantitative format, identifying of variables, setting out alternative solutions and supplementing decision making, by replacing judgment and intuition."

Quantitative techniques may be described as those techniques "which provide decision maker, with a systematic and powerful tool of analysis, based on quantitative and numeric data relating to alternative option."

Thus quantitative techniques are a set of techniques involving numerical formulation of a decision situation and analysis of variables, so as to arrive at alternative solutions, leading to optimal decision.

Classification of quantitative techniques

Quantitative techniques are a set of methods used to quantitatively formulate, analyze, integrate and decide problems or issues. They are broadly classified into three – mathematical techniques, statistical techniques and programming techniques.

Mathematical techniques

They are quantitative techniques in which numerical data are used along with the principles of mathematics such as integration, calculus etc. They include permutations, combinations, set theory, matrix analysis, differentials integration etc.

Permutations and combinations

Permutation is mathematical device of finding possible number of arrangements or groups which can be made of a certain number of items from a set of observations. They are groupings considering order of arrangements.

Combinations are number of selections or subsets which can be made of a certain number of items from a set of observations, without considering order. Both combinations and permutations help in ascertaining total number of possible cases.

Set theory

It is a modern mathematical device which solves the various types of critical problems on the basis of sets and their operations like Union, intersection etc.

Matrix Algebra

Matrix is an orderly arrangement of certain given numbers or symbols in rows and columns. Matrix analysis is thus a mathematical device of finding out the results of different types of algebraic operations on the basis of relevant matrices. This is useful to find values of unknown numbers connected with a number of simultaneous equations.

Differentials

Differential is a mathematical process of finding out changes in the dependent variable with reference to a small change in the independent variable. It involves differential coefficients of dependent variables with or without variables.

Integration

It is a technique just reversing the process of differentiation. It involves the formula $\int f(x)\,dx$ where $f(x)$ is the function to be integrated

Statistical techniques

They are techniques which are used in conducting statistical inquiry concerning a certain phenomenon. They include all the statistical methods beginning from the collection of data till interpretation of those collected data. Important statistical techniques include collection of data, classification and tabulation, measures of central tendency, measures of dispersion, skewness and kurtosis, correlation, regression, interpolation and extrapolation, index numbers, time series analysis, statistical quality control, ratio analysis , probability theory, sampling technique, variance analysis, theory of attributes etc.

Programming techniques

These techniques focus on model building, and are widely applied by decision makers relating to business operations. In programming, problem is formulated in numerical form, and a suitable model is fitted to the problem and finally a solution is derived. Prominent programming techniques include linear programming, queuing theory, inventory theory, theory of games, decision theory, network programming, simulation, replacement non linear programming, dynamic programming integer programming etc.

Functions of Quantitative Techniques:

The following are the important functions of quantitative techniques:

1. To facilitate the decision-making process
2. To provide tools for scientific research
3. To help in choosing an optimal strategy
4. To enable in proper deployment of resources
5. To help in minimizing costs
6. To help in minimizing the total processing time required for performing a set of jobs

Quantitative and qualitative approaches

Decision making is the process of selecting optimal alternative from among several alternatives, subject to states of nature. While analyzing a situation for such a selection, two approaches can be adopted – quantitative approach and qualitative approach

Quantitative approach

This approach involves generation and analysis of data in numerical form. Data obtained as per quantitative approach can be subjected to rigorous quantitative analysis in a formal fashion. This will reveal almost all inherent characteristics of the variable under study.

Quantitative approach may further be subdivided into inferential, experimental and simulation approaches. The purpose of inferential approach is to form a data base to infer characteristics or relationships of variables. Required data would be usually obtained through field survey.

Experimental approach is characterized by much greater control over the study environment, and in this case variables are manipulated to observe their effect on other variables.

Simulation approach involves the construction of an artificial environment or model within which relevant information and data can be generated. This permits an observation of dynamic behavior of the system or sub system under modeled conditions. The term simulation, in the context of business, means building of a model, that represents the structure of a dynamic process or operation.

Qualitative approach

Qualitative approach is concerned with subjective assessment of attitudes, opinions and behavior. Decision making in such situations is a function of decision maker's insight and impressions. Such an approach generates results either in non-quantitative form or in a form which cannot be subjected to rigorous quantitative analysis. For example, opinion that a person may be good or bad

Basically, the techniques of focus group interviews, projective techniques and depth interviews use qualitative approach for decision making.

Generally there are four non quantitative techniques of decision making

Intuition – decision making on intuition is characterized by inner feelings of the decision maker. It is purely subjective

Facts –It follows the rule that decision should be based on facts, and not on feelings.

Experiences – Experience is the most valuable asset, if used logically. Decisions should be based on precedence.

Opinion – in decision making, expert opinions can be relied on. In fact, this is widely used by all levels of managers.

However, even qualitative approach may be transformed into quantitative form, in practical studies. This is achieved through measurement and scaling. Measurement is assigning numbers or values to concepts or phenomena. Scaling refers to placing a concept or characteristic on the appropriate position of a measured scale. For example, Marital status of a person may be : (single)1 , (married)2, (divorced)3 (widowed)4. Here qualitative or non quantitative data is logically converted into quantitative data.

Significance of quantitative decisions

Quantitative Techniques have proved useful in tackling managerial decision problems relating to business and industrial operations. Quantitative decisions are considered significant on the following grounds.

Simplifies decision making

Quantitative techniques simplify the decision making process. Decision theory enables a manager to select the best course of action. Decision tree technique refines executive judgment in systematic analysis of the problem, these techniques permit scientific decision making under conditions of risk and uncertainty. Decision problems such as manpower planning ,demand forecasting, selection suppliers, production capacities, and capital requirements planning can be more effective tackled using quantitative techniques.

Scientific analysis

It provides a basis for precise analysis of the cause and effect relationship. They make it possible to measure the risks inherent, in business by providing an analytical and objective approach. These techniques reduce the need for intuition and objective approach. In this way quantitative techniques enable managers to use logical thinking in the analysis of organizational problems,

Allocation of resources

They are very helpful in the optimum deployment of resources. For example, Programme Evaluation and Review Techniques enable a manager to determine the earliest and the latest times for each of the events and activities involved in a project. The probability of completing the project by a specified date can be determined. Timely completion of the project helps to avoid time and cost overruns. Similarly, linear programming technique is very useful in optimal allocation of scarce resources, production scheduling and in deciding optimal assignments.

Profit maximization

Quantitative techniques are invaluable in assessing the relative profitability of alternative choices and identifying the most profitable course of action. What should be the relative mix of different products, which site to choose for location out of alternative sites,

which arrangement of orders in terms of time and quantity, will give maximum profits. Such question can be answered with the help of quantitative techniques.

Cost minimization

Quantitative techniques are helpful in tackling cost minimization problems. For example waiting line theory enables a manager to minimize waiting and servicing costs. Their techniques help business managers in taking a correct decision through analysis of feasibility of adding facilities.

Forecasting

Quantitative techniques are useful in demand forecasting. They provide a scientific basis of coping with the uncertainties of future demand. Demand forecasts serve as the basis for capacity planning. Quantitative technique enables a manager to adopt the minimum risk plan.

Inventory control

Inventory planning techniques help in deciding when to buy and how much to buy. It enables management to arrive at appropriate balance between the costs and benefits of holding stocks. The integrated production models technique is very useful in minimizing costs of inventory, production and workforce. Statistical quality controls help us to determine whether the production process is under control or not.

Applications of quantitative techniques in business operations

Quantitative techniques are widely applied for solving decision problems of routine operations of business organizations. It is especially useful for business managers, economist, statisticians, administrators, technicians and others in the field of business, agriculture, industry services and defense. It has specific applications in the following functional areas of business organizations.

Planning

In planning, quantitative techniques are applied to determine size and location of plant, product development, factory construction, installation of equipment and machineries etc.

Purchasing

Quantitative techniques are applied in make or buy decisions, vendor development, vendor rating, purchasing at varying prices, standardization and variety reduction, logistics management.

Manufacturing

Quantitative techniques address questions like product mix, production planning, quality control, job sequencing, and optimum run sizes.

Marketing

Marketing problems like demand forecasting, pricing competitive strategies, optimal media planning and sales management can be solved through application appropriate quantitative techniques.

Human resource management

Quantitative techniques supports decision making relating to man power planning with due consideration to age, skill, wastage and recruitment , recruitment on the basis of proper aptitude, method study , work measurement, job evaluation, development of incentive plans, wage structuring and negotiating wage and incentive plan with the union.

Research and Development

Quantitative techniques are helpful in deciding research issues like market research, market survey, product innovation, process innovations, plant relocation, merger and acquisitions etc.

Classification of quantitative techniques

Quantitative techniques are a set of methods used to quantitatively formulate, analyze, integrate and decide problems or issues. They are broadly classified into three – mathematical techniques, statistical techniques and programming techniques.

Mathematical techniques

They are quantitative techniques in which numerical data are used along with the principles of mathematics such as integration, calculus etc. They include permutations, combinations, set theory, matrix analysis, differentials integration etc.

Permutations and combinations

Permutation is mathematical device of finding possible number of arrangements or groups which can be made of a certain number of items from a set of observations. They are groupings considering order of arrangements.

Combinations are number of selections or subsets which can be made of a certain number of items from a set of observations, without considering order. Both combinations and permutations help in ascertaining total number of possible cases.

Set theory

It is a modern mathematical device which solves the various types of critical problems on the basis of sets and their operations like Union, intersection etc.

Matrix Algebra

Matrix is an orderly arrangement of certain given numbers or symbols in rows and columns. Matrix analysis is thus a mathematical device of finding out the results of different types of algebraic operations on the basis of relevant matrices. This is useful to find values of unknown numbers connected with a number of simultaneous equations.

Differentials

Differential is a mathematical process of finding out changes in the dependent variable with reference to a small change in the independent variable. It involves differential coefficients of dependent variables with or without variables.

Integration

It is a technique just reversing the process of differentiation. It involves the formula $\int f(x)\,dx$ where $f(x)$ is the function to be integrated

Statistical techniques

They are techniques which are used in conducting statistical inquiry concerning a certain phenomenon. They include all the statistical methods beginning from the collection of data till interpretation of those collected data. Important statistical techniques include collection of data, classification and tabulation, measures of central tendency, measures of dispersion, skewness and kurtosis, correlation, regression, interpolation and extrapolation, index numbers, time series analysis, statistical quality control, ratio analysis , probability theory, sampling technique, variance analysis, theory of attributes etc.

Programming techniques

These techniques focus on model building, and are widely applied by decision

makers relating to business operations. In programming, problem is formulated in numerical form, and a suitable model is fitted to the problem and finally a solution is derived. Prominent programming techniques include linear programming, queuing theory, inventory theory, theory of games, decision theory, network programming, simulation, replacement non linear programming, dynamic programming integer programming etc.

Quantification of qualitative data

In most cases, information is born in the form of qualitative description of situations. This may be quantified. Such quantification leads to following favorable out comes

1. It attracts readers' attention to patterns in the information
2. It helps to memorize and stacking of information
3. It assists in timely retrieval of data.
4. It supports efficient decision making.

Limitations of Quantitative Techniques:

Even though the quantitative techniques are inevitable in decision-making process, they are not free from short comings. The following are the important limitations of quantitative techniques:

1. Quantitative techniques involves mathematical models, equations and other mathematical expressions
2. Quantitative techniques are based on number of assumptions. Therefore, due care must be ensured while using quantitative techniques, otherwise it will lead to wrong conclusions.
3. Quantitative techniques are very expensive.
4. Quantitative techniques do not take into consideration intangible facts like skill, attitude etc.
5. Quantitative techniques are only tools for analysis and decision-making. They are not decisions itself.

REVIEW QUESTIONS:

1. Define Quantitative Techniques.
2. Explain the classification of quantitative techniques.
3. Explain the significance of quantitative decisions.
4. What are the uses of quantitative techniques in Business?
5. Explain the qualitative approach in decision making.
6. What are the important limitations of quantitative techniques?

Chapter 2

CORRELATION ANALYSIS

Meaning and Definition of Correlation

Correlation analysis is an attempt to examine the relationship between two variables. It analyses the association between two or more variables. It is a bi-variate analysis.

According to Croxton and Cowden, "when the relationship is of quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation"

According to A.M.Tuttle, "Correlation is an analysis of the co-variation between two or more variables."

According to Ya-Lun-Chou, "Correlation analysis is an attempt to determine the degree of relationship between variables."

Correlation analysis helps to know the direction of relationship as well as the degree of relationship exists between two or more variables.

Types of Correlation

I. Positive and Negative Correlation

(a) Positive Correlation: If two variables move in the same direction, then the correlation is called positive. For example, price and supply are positively correlated. When price goes up, supply goes up and vice versa.

(b) Negative Correlation: If two variables move in the opposite direction, then the correlation is called negative. For example, price and demand are negatively correlated. When price goes up, demand falls down and vice versa.

II. Simple, Partial and Multiple Correlation

(a) Simple Correlation: In a correlation analysis, if there are only two variables, then the correlation analysis is called simple correlation. For example, the relationship between weight and height, price and demand, price and supply, etc.

(b) **Partial Correlation:** When there are more than two variables and we study the relationship between any two variables only, assuming other variables as constant, it is called partial correlation. For example, the study of the relationship between rainfall and agricultural produce, without taking into consideration the effects of other factors such as quality of seeds, quality of soil, use of fertilizer, etc.

(c) **Multiple Correlation:** When there are more than two variables and we study the relationship between one variable and all the other variables taken together, then it is the case of multiple

correlation. Suppose there are three variables, namely x, y and z. The correlation between x and (y & z) taken together is multiple correlation. Similarly, the relation between y and (x & z) taken together is multiple correlation. Again, the relation between z and (x & y) taken together is multiple correlation.

III. Linear and Non-linear Correlation

(a) Linear Correlation: When the amount of change in one variable leads to a constant ratio of change in the other variable, the relationship is called linear correlation. For example, if price falls down by 10%, it leads to a fall in supply by 12% each time, it is linear correlation. When we plot the data on graph, we will get a straight line. Here, the relationship between the variables may be expressed in the form of $y = ax + b$.

(b) Non-linear Correlation: When the amount of change in one variable does not lead to a constant ratio of change in the other variable, the relationship is called non-linear correlation. When we plot the data on graph, we never get a straight line. Therefore, non-linear correlation is also called curvi-linear correlation.

IV. Logical and Illogical Correlation

(a) Logical Correlation: When the correlation between two variables is not only mathematically defined but also logically sound, it is called logical correlation. For example, correlation between price and demand.

(b) Illogical Correlation: When the correlation between two variables is mathematically defined but not logically sound, it is called illogical correlation. For example, correlation between availability of rainfall and height of people. This type of correlation is also known as Spurious correlation or Non-sense correlation.

Methods of Studying Correlation

The various methods for studying correlation can be classified into two categories. They are:

I. Graphic Methods:
   (1) Scatter diagram method
   (2) Correlation Graph method

II. Mathematical Methods:
   (1) Karl Pearson's Product Moment Method
   (2) Spearman' Rank Correlation Method
   (3) Concurrent Deviation Method

Graphic Methods:

Scatter Diagram Method

This is a simple method for analysing correlation between two variables. One variable is shown on the X- axis and the other on the Y-axis. Each pair of values is shown on the graph paper using dots. When all the pairs of observations are plotted as dots, the relationship exists

between the variables is analysed by observing how the dots are scattered. If the dots show an upward or downward trend, then the variables are correlated. We may interpret the scatter diagram as follows:

(a) If all the dots are lying on a straight line from left bottom corner to the right upper corner, there is perfect positive correlation between variables.
(b) If all the dots are lying on a straight line from left upper corner to the right bottom corner, there is perfect negative correlation between variables.
(c) If all the dots are plotted on a narrow band from left bottom corner to the right upper corner, there is high degree of positive correlation between variables.
(d) If all the dots are plotted on a narrow band from left upper corner to the right bottom corner, there is high degree of negative correlation between variables.
(e) If all the dots are plotted on a wide band from left bottom corner to the right upper corner, there is low degree of positive correlation between variables.
(f) If all the dots are plotted on a wide band from left upper part to the right bottom part, there is low degree of negative correlation between variables.
(g) If the plotted dots do not show any trend, the variables are not correlated.

Correlation Graph Method

In correlation graph method, separate curves are drawn for each variable on the same graph. The relationship between the variables is interpreted on the basis of the direction and closeness of the curves. If both the curves move in the same direction, there is positive correlation and if they are moving in opposite directions, there is negative correlation between the variables.

Mathematical Methods:

Under mathematical methods, the correlation between variables is studied with the help of a numerical value obtained using an appropriate formula. This numerical value is called coefficient of correlation. Coefficient of correlation explains both the direction as well as degree of relationship exists between the variables.

Degree of Correlation:

The degree of correlation can be classified as follows:

(a) Perfect Positive Correlation: When coefficient of correlation is +1
(b) Perfect Negative Correlation: When coefficient of correlation is –1
(c) High Degree of Positive Correlation: When coefficient of correlation lies between + 0.75 and + 1
(d) High Degree of Negative Correlation: When coefficient of correlation lies between – 0.75 and –1
(e) Moderate Degree of Positive Correlation: When coefficient of correlation lies between + 0.5 and + 0.75

(f) Moderate Degree of Positive Correlation: When coefficient of correlation lies between – 0.5 and – 0.75

(g) Low Degree of Positive Correlation: When coefficient of correlation lies between 0 and + 0.33

(h) Low Degree of Negative Correlation: When coefficient of correlation lies between 0 and – 0.33

(i) No Correlation: When coefficient of correlation is zero.

Karl Pearson's Product Moment Method

This is the popularly used method for analysing correlation. This method is designed by a reputed Statistician, Prof. Karl Pearson and therefore, it is generally known as Pearsonian Coeffiecient of Correlation. Karl Pearson's coefficient of correlation is denoted by 'r' Under this method, coefficient of correlation is computed by usin any one of the following formulae:

$$r = \frac{\varepsilon xy}{\sqrt{\varepsilon x^2 \quad \varepsilon y^2}}$$

where x = deviations of X values from its actual mean
y = deviations of X values from its actual mean

OR

$$r = \frac{n\varepsilon dxdy - (\varepsilon dx \,.\, \varepsilon dy)}{\sqrt{n\varepsilon dx^2 - (\varepsilon dx)^2 \quad n\varepsilon dy^2 - (\varepsilon dy)^2}}$$

where n = number of pairs of observations
dx = deviations of observations (variable x) from assumed mean
dy = deviations of observations (variable y) from assumed mean

OR

$$r = \frac{n\varepsilon XY - (\varepsilon X \,.\, \varepsilon Y)}{\sqrt{n\varepsilon X^2 - (\varepsilon X)^2 \quad n\varepsilon Y^2 - (\varepsilon Y)^2}}$$

where n = number of pairs of observations
X = Given values of variable X
Y = Given values of variable Y

Qn: From the following data compute product moment correlation coefficient and interpret it:

X	57	42	40	38	42	45	42	44	40	46	44	43
Y	10	26	30	41	29	27	27	19	18	19	31	29

Sol:

Computation of Product Moment Correlation Coefficient						
X	Y	dx = (x --45)	dy=(y—25)	Dxdy	dx^2	dy^2
57	10	12	-15	-180	144	225
42	26	-3	1	-3	9	1
40	30	-5	5	-25	25	25
38	41	-7	16	-112	49	256
42	29	-3	4	-12	9	16
45	27	0	2	0	0	4
42	27	-3	2	-6	9	4
44	19	-1	-6	6	1	36
40	18	-5	-7	35	25	49
46	19	1	-6	-6	1	36
44	31	-1	6	-6	1	36
43	29	-2	4	-8	4	16
		εdx= -17	εdy= 6	εdxdy= -317	$εdx^2$=277	$εdy^2$ = 704

$$r = \frac{n\varepsilon dxdy - (\varepsilon dx \cdot \varepsilon dy)}{\sqrt{n\varepsilon dx^2 - (\varepsilon dx)^2 \quad n\varepsilon dy^2 - (\varepsilon dy)^2}}$$

$$r = (12 \times -317) - (-17 \times 6) / \sqrt{[(12 \times 277) - (-17)^2] [(12 \times 704) - (6)^2]}$$

$$= -3804 - {}^{-}102 / [\sqrt{3324 - 289] [8448 - 36]}$$

$$= -3702 / \sqrt{3035 \times 8412} = -14208 / \sqrt{25530420}$$

$$= -3702 / 5052.76 = \underline{\underline{-0.733}}$$

There is a high degree of negative correlation between x and y.

Properties of Coefficient of Correlation

1. Coefficient of correlation lies between – 1 and + 1
2. It is a pure numerical value independent of the units of measurement.
3. It does not change with reference to the change of origin or change of scale.
4. It is the geometric mean of the two regression coefficients
5. It is computed by using a well defined formula.
6. Coefficient of correlation between x and y and that between y and x are same.

Probable Error

Probable error is a statistical device used to measure the reliability and dependability of the value of correlation coefficient. When the numerical value of probable error is added to and

subtracted from the value of correlation coefficient, we get two limits within which the population parameter is expected to lie.

Probable Error (P.E) = 0.6745 x Standard Error Probable Error (P.E) = 0.6745 x $[(1 - r^2)/\sqrt{n}]$

where n= number of pairs of observations; r = correlation coefficient.

Uses of Probable Error

1. Probable error can be used to measure the reliability and dependability of coefficient of correlation
2. It helps to determine the limits within which population parameter is expected to lie.
3. With the help of P.E, coefficient of correlation can be interpreted more accurately:
   (a) If 'r' is less than P.E, there is no evidence of correlation.
   (b) If 'r' is more than 6 times of P.E, correlation is significant
   (c) If 'r' is 0.5 or more and the P.E is not much, the correlation is considered to be significant.

Qn: Find Standard Error (S.E) and Probable Error (P.E), if r = 0.8 and number of pairs of observations = 64. Also interpret the value of 'r'.

Sol:

Standard Error (S.E) = $1 - r^2/\sqrt{n}$

$= 1 - 0.8^2/\sqrt{64} = 1 - 0.64/8$

$= 0.36/8 = \underline{0.045}$

Probable Error (P.E) = 0.6745 x Standard Error

$= 0.6745 \times 0.045 = \underline{\underline{0.0304}}$

$r/P.E = 0.08/0.0304 = 26.32.$

Since 'r' is more than 26.32 times of P.E, the value of 'r' is highly significant.

Qn: Following table shows the marks obtained by students in two courses:

Course I	45	70	65	30	90	40	50	75	85	60
Course II	35	90	70	40	95	40	60	80	80	50

Find coefficient of correlation and P.E. Is 'r' significant?

Sol:

Computation of Product Moment Correlation Coefficient						
Course I (x)	Course II (y)	dx = (x --60)	dy=(y—60)	Dxdy	dx^2	dy^2

45	35	-15	-25	375	225	625
70	90	10	30	300	100	900
65	70	5	10	50	25	100
30	40	-30	-20	600	900	400
90	95	30	35	1050	900	1225
40	40	-20	-20	400	400	400
50	60	-10	0	0	100	0
75	80	15	20	300	225	400
85	80	25	20	500	625	400
60	50	0	-10	0	0	100
		$\varepsilon dx = 10$	$\varepsilon dy = 40$	$\varepsilon dxdy = 3575$	$\varepsilon dx^2 = 3500$	$\varepsilon dy^2 = 4550$

$$r = \frac{n\varepsilon dxdy - (\varepsilon dx \cdot \varepsilon dy)}{\sqrt{n\varepsilon dx^2 - (\varepsilon dx)^2 \quad n\varepsilon dy^2 - (\varepsilon dy)^2}}$$

$$r = (10 \times 3575) - (10 \times 40) \,/\, \sqrt{[(10 \times 3500) - (10)^2]\,[(10 \times 4550) - (40)^2]}$$

$$= 35750 - 400 \,/\, \sqrt{[35000 - 100]\,[45500 - 1600]}$$

$$= 35350 \,/\, \sqrt{34900 \times 43900} = 35350 \,/\, 39142.177$$

$$= \underline{\pm 0.903}$$

Probable Error (P.E) $= 0.6745 \times [(1 - r^2)/\sqrt{n}]$

$$= 0.6745 \times [1 - 0.903^2)/\sqrt{10}]$$

$$= 0.6745 \times [1 - 0.815)/\sqrt{10}]$$

$$= 0.6745 \times [0.185/3.162]$$

$$= 0.6745 \times [0.0585] = \underline{0.0395}$$

$r/P.E = 0.903/0.0395 = 22.86$

Since coefficient of correlation is more than 22.86 of P.E, 'r' is very significant.

Coefficient of Determination

Coefficient of Determination is defined as the ratio of the explained variance to the total variance. It denoted by r^2 and is usually expressed as percentage. Coefficient of Determination explains the percentage of the variation in the dependent variable that can be explained in terms of the independent variable.

Coefficient of Determination = r^2

Coefficient of Determination = Explained Variance/Total Variance

Coefficient of non-determination = $1 - r^2$.

Qn: If the coefficient of correlation between two variables is 0.85, what percentage of variation of dependent variable is explained? Also find the coefficient of non-determination.

Sol:

Coefficient of Determination
(Percentage of explained Variance) = r^2
= 0.85 x 0.85 = 0.7225 = 72.25%
Coefficient of non-determination = $1 - r^2$ = 1 – 0.7225= 2775 = 27.75%

Rank Correlation Method

When the variables cannot be measured in quantitative terms, the coefficient of correlation can be found out by using rank correlation method. Here ranks are to be assigned to the individual observations. Ranks may be assigned in either ascending or descending order. This method was designed by Charles Edward Spearman in 1904. He suggested two formulae for computing rank correlation coefficient. Rank correlation coefficient if denoted by 'R'

(1) When there is no equal rank:

$$R = 1 - \frac{6\varepsilon D^2}{(n^3 - n)}$$

where D = Rank difference

n = Number of pairs of observations

(2) When there are equal ranks:

$$R = 1 - \frac{6\{\varepsilon D^2 + [(m^3-m)/12] + [(m^3-m)/12] + ...\}}{(n^3 - n)}$$

where D = Rank difference

n = Number of pairs of observations

m = Number of times a particular rank repeats

Qn: The ranks of 6 students in two courses are given below:

Rank for Course I	6	1	5	2	4	3
Ran for Course II	3	1	4	2	5	6

Compute Spearman's Rank Correlation Coefficient.

Sol:

Here there is no equal rank.

$$R = 1 - \frac{6\varepsilon D^2}{(n^3 - n)}$$

Computation of Rank Correlation Coefficient			
Rank for Course I (R_1)	Rank for Course I(R_2)	D ($R_1 - R_2$)	D^2
6	3	3	9
1	1	0	0
5	4	1	1
2	2	0	0
4	5	-1	1
3	6	-3	9
			$\varepsilon D^2 = 20$

$$R = 1 - \frac{6\varepsilon D^2}{(n^3 - n)}$$

$$= 1 - [(6 \times 20)/(6^3 - 6)] = 1 - (120/210) = 1 - 0.5714 = \underline{\underline{0.4286}}$$

Qn: From the following data, compute Spearman's Rank Correlation Coefficient:

x	330	332	328	331	327	325
y	415	434	420	430	424	428

Sol:

Here ranks are not given. So, at first, we have to assign ranks to each observation:

$$R = 1 - \frac{6\varepsilon D^2}{(n^3 - n)]}$$

Computation of Rank Correlation Coefficient					
X	Y	R_1	R_2	D ($R_1 - R_2$)	D^2
330	415	4	1	3	9
332	434	6	6	0	0
328	420	3	2	1	1
331	430	5	5	0	0
327	424	2	3	-1	1
325	428	1	4	-3	9
					$\varepsilon D^2 = 20$

$$R = 1 - \frac{6\varepsilon D^2}{(n^3 - n)} = 1$$

$= 1 - [(6 \times 20)/(6^3 - 6)] = 1 - (120/210) = 1 - 0.5714 = \underline{0.4286}$

Qn: From the following data, compute Spearman's Rank Correlation Coefficient:

x	80	45	55	58	55	60	45	68	70	45	85
y	82	56	50	43	56	62	64	65	70	64	90

Sol:

This is the case of equal marks.

$$R = 1 - \frac{6\{\varepsilon D^2 + [(m^3 - m)/12] + [(m^3 - m)/12] + ...\}}{(n^3 - n)}$$

Computation of Rank Correlation Coefficient					
X	Y	R_1	R_2	D $(R_1 - R_2)$	D^2
80	82	10	10	0	0
45	56	2	3.5	-1.5	2.25
55	50	4.5	2	2.5	6.25
58	43	6	1	5	25
55	56	4.5	3.5	1	1
60	62	7	5	2	4
45	64	2	6.5	-4.5	20.25
68	65	8	8	0	0
70	70	9	9	0	0
45	64	2	6.5	-4.5	20.25
85	90	11	11	0	0
					$\varepsilon D^2 = 79$

$$R = 1 - \frac{6\{79 + [(3^3 - 3)/12] + [(2^3 - 2)/12] + [(2^3 - 2)/12] + [(2^3 - 2)/12]\}}{(11^3 - 11)}$$

$$= 1 - \frac{6\ [79 + (2 + 0.5 + 0.5 + 0.5)]}{1320}$$

$= 1 - [6(79 + 3.5) \div 1320] = 1 - (495/1320) = 1 - (0.375) = \underline{0.625}$

Concurrent Deviation Method

This is a simple method for computing coefficient of correlation. Here, we consider only the direction of change and not the magnitude of change. The coefficient of correlation is determined on

the basis of number of concurrent deviations. That is why this method is named as such. The coefficient of concurrent deviation is denoted by r_c.

The formula for computing coefficient of concurrent deviation is:

$$r_c = \pm\sqrt{\pm (2c - n) / n}$$

where c = number of concurrent deviations

n = number of pairs of signs (not the pairs of observations)

Qn: Calculate coefficient of concurrent deviation from the following data:

x	180	182	186	191	183	185	189	196	193
y	246	240	230	217	233	227	215	195	200

Sol:

$$r_c = \pm \sqrt{\pm (2c - n) / n}$$

Computation of Coefficient of Concurrent Deviation				
x	y	dx	Dy	Dxdy
180	246	...	...	...
182	240	+	-	-
186	230	+	-	-
191	217	+	-	-
183	233	-	+	-
185	227	+	-	-
189	215	+	-	-
196	195	+	-	-
193	200	-	+	-
				c = 0

Number of concurrent deviations = 0

$$r_c = \pm \sqrt{\pm (2c - n) / n}$$

$$r_c = \pm \sqrt{\pm (2 \times 0 - 8) / 8} = \pm\sqrt{\pm(0-8)/8} = -1$$

There is perfect negative correlation between x and y.

PARTIAL CORRELATION

When there are more than two variables and we study the relationship between any two variables only, assuming other variables as constant, it is called partial correlation. For example, the study of the relationship between rainfall and agricultural produce, without taking into consideration the effects of other factors such as quality of seeds, quality of soil, use of fertilizer, etc.

Partial correlation coefficient measures the relationship between one variable and one of the other variables assuming that the effect of the rest of the variables is eliminated.

Suppose there are 3 variables namely x_1, x_2 and x_3. Here, we can find three partial correlation coefficients. They are:

(1) Partial Correlation coefficient between x_1 and x_2, keeping x_3 as constant. This is denoted by $r_{12.3}$
(2) Partial Correlation coefficient between x_1 and x_3, keeping x_2 as constant. This is denoted by $r_{13.2}$
(3) Partial Correlation coefficient between x_2 and x_3, keeping x_1 as constant. This is denoted by $r_{23.1}$

The formulae for computing the above partial correlation coefficients are:

$$r_{12.3} = \frac{r_{12} - r_{13}\ r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{23}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{23}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\ r_{13}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{13}^2}}$$

Qn: If $r_{12} = 0.98$, $r_{13} = 0.44$ and $r_{23} = 0.54$, find (1) $r_{12.3}$, (2) $r_{13.2}$ and (3) $r_{23.1}$

Sol:

(1)

$$r_{12.3} = \frac{r_{12} - r_{13}\ r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}}$$

$$r_{12.3} = \frac{0.98 - (0.44 \times 0.54)}{\sqrt{1-0.44^2}\ \sqrt{1-0.54^2}}$$

$$= \frac{0.98 - 0.2376}{\sqrt{1-0.1936}\ \sqrt{1-0.2916}}$$

$$= 0.7424 / (0.898 \times 0.842) = 0.7424 / 0.7561 = \underline{0.982}$$

(2)

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{23}}{}$$

$$\sqrt{1-r_{12}^2}\ \sqrt{1-r_{23}^2}$$

$$r_{12.3} = \frac{0.44 - (0.98 \times 0.54)}{\sqrt{1-0.98^2}\ \sqrt{1-0.54^2}}$$

$$= \frac{0.44 - 0.5292}{\sqrt{1-0.9604}\ \sqrt{1-0.2916}}$$

$$= -0.0892 / (0.199 \times 0.842) = -0.0892 / 0.1676 = -\underline{0.5322}$$

3)

$$r_{23.1} = \frac{r_{23} - r_{12}\ r_{13}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{13}^2}}$$

$$r_{12.3} = \frac{0.54 - (0.98 \times 0.44)}{\sqrt{1-0.98^2}\ \sqrt{1-0.44^2}}$$

$$= \frac{0.54 - 0.4312}{\sqrt{1-0.9604}\ \sqrt{1-0.1936}}$$

$$= 0.1088 / (0.199 \times 0.898) = 0.1088 / 0.1787 = \underline{0.6088}$$

MULTIPLE CORRELATION

When there are more than two variables and we study the relationship between one variable and all the other variables taken together, then it is the case of multiple correlation. Suppose there are three variables, namely x, y and z. The correlation between x and (y & z) taken together is multiple correlation. Similarly, the relation between y and (x & z) taken together is multiple correlation. Again, the relation between z and (x & y) taken together is multiple correlation. In all these cases, the correlation coefficient obtained will be termed as coefficient of multiple correlation.

Suppose there are 3 variables namely x_1, x_2 and x_3. Here, we can find three multiple correlation coefficients. They are:

1. Multiple Correlation Coefficient between x_1 on one side and x_2 and x_3 together on the other side. This is denoted by $R_{1.23}$
2. Multiple Correlation Coefficient between x_2 on one side and x_1 and x_3 together on the other side. This is denoted by $R_{2.13}$
3. Multiple Correlation Coefficient between x_3 on one side and x_1 and x_2 together on the other side. This is denoted by $R_{3.12}$

The formulae for computing the above multiple correlation coefficients are:

$$R_{1.23} = \sqrt{[r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{23}^2]}$$

$$R_{2.13} = \sqrt{[r_{12}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{13}^2]}$$

$$R_{3.12} = \sqrt{[r_{13}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{12}^2]}$$

Qn: If $r_{12} = 0.6$, $r_{23} = r_{13} = 0.8$, find $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

Sol:

$$R_{1.23} = \sqrt{[r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{23}^2]}$$

$$= \sqrt{[0.6^2 + 0.8^2 - 2 \text{ x } 0.6 \text{ x } 0.8 \text{ x } 0.8] \div [1 - 0.8^2]}$$

$$= \sqrt{[0.36 + 0.64 - 0.768] \div [1 - 0.64]}$$

$$= \sqrt{0.232 \,/ 0.36} = \sqrt{0.6444} = \underline{0.8028}$$

$$R_{2.13} = \sqrt{[r_{12}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{13}^2]}$$

$$= \sqrt{[0.6^2 + 0.8^2 - 2 \text{ x } 0.6 \text{ x } 0.8 \text{ x } 0.8] \div [1 - 0.8^2]}$$

$$= \sqrt{[0.36 + 0.64 - 0.768] \div [1 - 0.64]}$$

$$= \sqrt{0.232 \,/ 0.36} = \sqrt{0.6444} = \underline{0.8028}$$

$$R_{3.12} = \sqrt{[r_{13}^2 + r_{23}^2 - 2\, r_{12}\, r_{13}\, r_{23}] \div [1 - r_{12}^2]}$$

$$= \sqrt{[0.8^2 + 0.8^2 - 2 \text{ x } 0.6 \text{ x } 0.8 \text{ x } 0.8] \div [1 - 0.6^2]}$$

$$= \sqrt{[0.64 + 0.64 - 0.768] \div [1 - 0.36]}$$

$$= \sqrt{0.512 / 0.64} = \sqrt{0.8} = \underline{0.8944}$$

REVIEW QUESTIONS:

1. What do you mean by correlation analysis?
2. Define correlation.
3. What is scatter diagram? What are its advantages?
4. What is coefficient of correlation? What are its properties?
5. What are the different types of correlation?
6. What do you mean by degree of correlation?
7. What is meant by positive and negative correlation?
8. What are the merits of Karl Pearson's Coefficient of Correlation?
9. What are the demerits of Karl Pearson's Coefficient of Correlation?
10. What is rank correlation?
11. What are the merits of rank correlation?
12. What are the demerits of rank correlation?
13. What is meant by correlation graph method?
14. What is concurrent deviation method?
15. What is the main drawback of concurrent deviation method?
16. What is coefficient of determination?
17. What do you mean by coefficient of non-determination?
18. What do you mean by linear and non-linear correlation?
19. What is multiple correlations?
20. What is partial correlation?
21. What is probable error?
22. What are the uses of probable error?
23. From the following data, find coefficient of correlation and give interpretation:

X	200	270	400	310	340
Y	150	162	180	180	170

24. Find n, if P.E = 0.034 and r = 0.917
25. Find coefficient of correlation using Edward Spearman's method:

Roll No.	1	2	3	4	5	6	7	8	9	10

Marks I	45	56	39	54	45	40	56	60	30	35
Marks II	40	56	30	44	36	32	45	42	20	36

26. Compute coefficient of concurrent deviation:

X	100	110	110	120	122	125
Y	120	140	160	160	130	110

27. If $r_{12} = 0.7$, $r_{13} = 0.61$, $r_{23} = 0.4$, find $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$

28. If $r_{12} = 0.98$, $r_{13} = 0.44$, $r_{23} = 0.54$, find $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

Chapter 3

REGRESSION ANALYSIS

Meaning and Definition of Regression Analysis

Correlation analysis helps to know whether two variables are related or not. Once the relationship between two variables is established, the same may be used for the purpose of predicting the unknown value of one variable on the basis of the known value of the other. For this purpose we have to examine the average functional relationship exists between the variables. This is known as regression analysis.

Regression analysis may be defined as the process of ascertaining the average functional relationship exists between variables so as to facilitate the mechanism of prediction or estimation or forecasting. Regression analysis helps to predict the unknown values of a variable with the help of known values of the other variable. The term regression was firstly used by Francis Galton.

Types of Regression

Regression may be classified as follows:

I. On the basis of number of variables:
   (a) Simple Regression
   (b) Multiple Regression
II. On the basis of Proportion of change in the variables:
   (a) Liner Regression
   (b) Non-liner Regression

1. Simple Regression

In a regression analysis, if there are only two variables, it is called simple regression analysis.

2. Multiple Regression

In a regression analysis, if there are more than two variables, it is called multiple regression analysis.

3. Linear Regression

In a regression analysis, if linear relation exists between variables, it is called linear regression analysis. Under this, when we plot the data on a graph paper, we get a straight line. Here, the relationship exists between variables can be expressed in the form of **y = a + bx.** In case of linear regression, the change in dependent variable is proportionate to the changes in the independent variable.

4. Non-linear Regression:

In case of non-linear regression, the relation between the variables cannot be expressed in the form of **y = a + bx.** When the data are plotted on a graph, the dots will be concentrated, more or less, around a curve. This is also called curvi-linear regression.

Regression Line (Line of Best Fit)

Regression line is a graphical method to show the functional relationship between two variables, namely dependent variable and independent variable. Since regression line helps to estimate the unknown values of dependent variable, based on the known values of the independent variable, it is also called estimating line (line of average).

According to Francis Galton, "The regression lines show the average relationship between two variables."

Regression Equations

Regression equations are algebraic expression of regression lines. As there are two regression lines, there are two regression equations. They are:

(a) Regression Equation of X on Y : It shows the change in the value of variable X for a given change in the value of variable Y.
(b) Regression Equation of Y on X **:** It shows the change in the value of variable Y for a given change in the value of variable X.

Methods of drawing Regression Lines

There are two methods for drawing regression lines. They are:

(a) Free hand curve method
(b) The method of least squares.

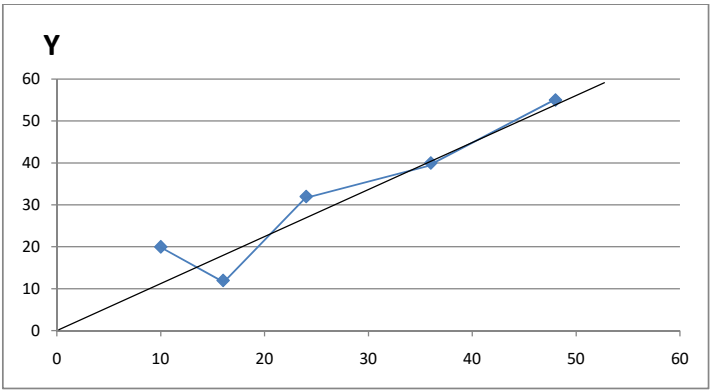
Free hand Curve Method:

This is a simple method for constructing regression lines. Under this method, the values of paired observations of the variable are plotted, by way of dots, on a graph paper. The X- axis represents the independent variable and Y-axis represents the dependent variable. After observing how the dots are scattered on the graph paper, we draw a straight line in such a way that the areas of the curve above and below the line are approximately equal. The line so drawn clearly indicates the tendency of the original data. Since there is subjectivity, this method is not commonly used in practice.

Qn: From the following data, draw a regression line of Y on X:

X	10	16	24	36	48
Y	20	12	32	40	55

Sol:



Method of Least Squares

Under method of least squares, the regression line should be drawn in such a way that the sum of the squares of the deviations of the actual Y-values from the computed Y-values is the least. In other words, $\varepsilon(y - y_c)^2$ = minimum. The line so fitted is called line of best fit.

Methods of Calculating Regression Equations

The following are the two important methods for calculating regression equations:

1. Normal equation method
2. Regression coefficient method

Normal Equation Method

The General form of regression equation is:

X on Y :	$X = a + bY$
Y on X :	$Y = a + bX$

For finding out the constants 'a' and 'b', we have to develop and solve certain equations, called normal equations. Therefore, this method is called normal equation method.

The normal equations computing 'a' and 'b' in respect of regression equation X on Y are:

$$\varepsilon X = Na + b\varepsilon Y, \text{ and}$$
$$\varepsilon XY = a\varepsilon Y + b\varepsilon Y^2$$

The normal equations computing ‘a’ and ‘b’ in respect of regression equation Y on X are:

$$\varepsilon Y = Na + b\varepsilon X, \text{ and}$$
$$\varepsilon XY = a\varepsilon X + b\varepsilon X^2$$

After computing the values of the constants ‘a’ and ‘b’, substitute them to the respective regression equations.

Qn: From the following data, fit the two regression equations:

x	4	5	8	2	1
y	5	6	7	3	2

Sol:

Regression Equation X on Y is:

$$X = a + bY$$

The normal equations to find the values of ‘a’ and ‘b’ are:

$$\varepsilon X = Na + b\varepsilon Y, \text{ and}$$

$$\varepsilon XY = a\varepsilon Y + b\varepsilon Y^2$$

Computation of Regression Equations				
x	y	Xy	x^2	y^2
4	5	20	16	25
5	6	30	25	36
8	7	56	64	49
2	3	6	4	9
1	2	2	1	4
$\varepsilon X=20$	$\varepsilon Y=23$	$\varepsilon XY=114$	$\varepsilon X^2=110$	$\varepsilon Y^2=123$

$$20 = 5a + 23\,b \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$114 = 23a + 123\,b \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

(1) x 23 : $460 = 115\,a + 529\,b$ ..... (1)
(2) X 5 : $570 = 115a + 615\,b$ ..... (2)
(3) -- (1) : $110 = 0 + 86\,b$

$$86\,b = 110$$

$$b = 110/86 = 1.28$$

Substitute b= 1.2 in equation number (1)

$$20 = 5a + 23 \times 1.28;$$

$20 = 5a + 29.44$; $5a = 20 - 29.44$; $5a = -9.44$

$a = -9.44/5 = -1.89$

Substitute the values of 'a' and 'b' in regression equation X on Y:

$X = -1.89 + 1.28y$

Regression Equation Y on X is:

$Y = a + bX$

The normal equations are:

$\varepsilon Y = Na + b\varepsilon X$, and

$\varepsilon XY = a\varepsilon X + b\varepsilon X^2$

$23 = 5a + 20b$ ..... (1)

$114 = 20a + 110b$ ..... (2)

(1) x 4 : $92 = 20a + 80b$ ..... (1)

$114 = 20a + 110b$ ..... (2)

(2) – (1): $22 = 0 + 30b$

$30b = 22$

$b = 22/30 = 0.73$

Substitute b=0.73 in equation (1)

$23 = 5a + 20 \times 0.73$; $23 = 5a + 14.6$

$5a = 23 - 14.6 = 8.4$; $a = 8.4/5 = 1.68$

Substitute a = 1.68 and b = 0.73 in the general form of Y on X

$Y = 1.68 + 0.73x$

Regression Equation X on Y : $X = -1.89 + 1.28y$

Regression Equation X on Y : $Y = 1.68 + 0.73x$

Regression coefficient method

Under regression coefficient method, regression equations are developed with the help of regression coefficients. Since there are two regression equations, two regression coefficients are to be computed.

The regression coefficient used to find the regression equation X on Y is "regression Coefficient of X on Y". It is denoted by b_{xy}

The regression coefficient used to find the regression equation Y on X is "regression Coefficient of Y on X". It is denoted by b_{yx}

The regression Equation X on Y is:

$$\mathbf{X - \bar{X} = b_{xy} (Y - \bar{Y})}$$

Where $\bar{X}$ = actual mean of X variable, $\bar{Y}$ = actual mean of Y variable

b_{xy} is computed by using any one of the following formula:

$$\mathbf{b_{xy} = r \cdot (\sigma_x / \sigma_y)}$$

where b_{xy} = Regression Coefficient of Regression equation X on Y
r = Coefficient of correlation
σ_X = Standard deviation of series X
σ_y = Standard deviation of series Y

OR

$$\mathbf{b_{xy} = \varepsilon xy / \varepsilon y^2}$$

where b_{xy} = Regression Coefficient of Regression equation X on Y
x = Deviation of X values from its actual mean
y = Deviation of X values from its actual mean

OR

$$\mathbf{b_{xy} = \frac{n\varepsilon dxdy - [(\varepsilon dx)(\varepsilon dy)]}{n\varepsilon dy^2 - (\varepsilon dy)^2}}$$

where b_{xy} = Regression Coefficient of Regression equation X on Y
dx = Deviation X values from its assumed mean
dy = Deviation Y values from its assumed mean
n = Number of paired observations

OR

$$\mathbf{b_{xy} = \frac{n\varepsilon XY - [(\varepsilon X)(\varepsilon Y)]}{n\varepsilon Y^2 - (\varepsilon Y)^2}}$$

where b_{xy} = Regression Coefficient of Regression equation X on Y
X = Given values of X variable
Y = Given values of Y variable

n = Number of paired observations

The regression Equation Y on X is:

$$\mathbf{Y - \bar{Y} = b_{yx} (X - \bar{X})}$$

Where $\bar{X}$ = Actual mean of X variable, $\bar{Y}$ = Actual mean of Y variable

B_{yx} is computed by using any one of the following formula:

$$\mathbf{b_{yx} = r \cdot (\sigma_y / \sigma_x)}$$

where b_{yx} = Regression Coefficient of Regression equation Y on X
r = Coefficient of correlation
σ_x = Standard deviation of series X
σ_y = Standard deviation of series Y

OR

$$\mathbf{b_{yx} = \varepsilon xy / \varepsilon x^2}$$

where b_{yx} = Regression Coefficient of Regression equation Y on X
x = Deviation of X values from its actual mean
y = Deviation of X values from its actual mean

OR

$$\mathbf{b_{yx} = \frac{n\varepsilon dxdy - [(\varepsilon dx)(\varepsilon dy)]}{n\varepsilon dx^2 - (\varepsilon dx)^2}}$$

where b_{yx} = Regression Coefficient of Regression equation Y on X
dx = Deviation X values from its assumed mean
dy = Deviation Y values from its assumed mean
n = Number of paired observations

OR

$$b_{yx} = \frac{n\varepsilon XY - [(\varepsilon X)(\varepsilon Y)]}{n\varepsilon X^2 - (\varepsilon X)^2}$$

where b_{yx} = Regression Coefficient of Regression equation Y on X
X = Given values of X variable
Y = Given values of Y variable
n = Number of paired observations

Qn: You are given the following bivariate data:

X	7	2	1	1	2	3	2	6
Y	2	6	4	3	2	2	8	4

Using regression coefficients:

(a) Fit the regression equation of Y on X and predict Y if X = 5
(b) Fit the regression equation of X on Y and predict X if Y = 20

Sol.

The regression Equation Y on X is:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

Assumed mean method is used to find b_{yx}

$$b_{yx} = \frac{n\varepsilon dxdy - [(\varepsilon dx)(\varepsilon dy)]}{n\varepsilon dx^2 - (\varepsilon dx)^2}$$

Computation of Regression Equations						
X	Y	dx (X–4)	dy (Y–3)	dxdy	dx^2	dy^2
7	2	3	-1	-3	9	1
2	6	-2	3	-6	4	9
1	4	-3	1	-3	9	1
1	3	-3	0	0	9	0
2	2	-2	-1	2	4	1
3	2	-1	-1	1	1	1
2	8	-2	5	-10	4	25
6	4	2	1	2	4	1
εX=24	**εY=31**	**εdx = -8**	**εdy = 7**	**εdxdy= -17**	**εdx^2 =44**	**εdy^2= 39**

$$b_{yx} = \frac{8 \times -17 - (-8 \times 7)}{8 \times 44 - (-8)^2}$$

$= (-136 - \bar{}56) / (352 - 64) = -80/288 = -\mathbf{0.278}$

$\bar{X} = \varepsilon X/n = 24/8 = 3$

$\bar{Y} = \varepsilon Y/n = 31/8 = 3.875$

The regression Equation Y on X is:

$Y - 3.875 = -0.278\,(X - 3)$

$Y = -0.278\,X + 0.834 + 3.875 = -0.278\,X + 4.709$

$Y = \underline{\mathbf{-0.278\,X + 4.709}}$

If X=5, Y = (-0.278 x 5) + 4.709 = -1.39 + 4.709 = **3.319**

The regression Equation X on Y is:

$$\mathbf{X - \bar{X} = b_{xy}\,(Y - \bar{Y})}$$

$$b_{xy} = \frac{n\varepsilon dxdy - [(\varepsilon dx)(\varepsilon dy)]}{n\varepsilon dy^2 - (\varepsilon dy)^2}$$

$$b_{xy} = \frac{8 \times -17 - (-8 \times 7)}{8 \times 39 - (7)^2}$$

$= (-136 - \bar{}56) / (312 - 49) = -80/263 = -\mathbf{0.3042}$

$\bar{X} = \varepsilon X/n = 24/8 = 3$

$\bar{Y} = \varepsilon Y/n = 31/8 = 3.875$

The regression Equation X on Y is:

$X - 3 = -0.3042\,(Y - 3.875)$

$X = -0.3042\,Y + 1.179 + 3 = -0.3042Y + 4.179$

$X = \underline{\mathbf{-0.3042Y + 4.179}}$

If Y=20, X = (-0.3042 x 20) + 4.179 = -6.084 + 4.179 = **-1.905**

Properties of Regression Coefficients

1. In a bivariate data, there will be two regression coefficients. They are b_{xy} and b_{yx}
2. b_{xy} is the regression coefficient of regression equation X on Y

3. b_{yx} is the regression coefficient of regression equation Y on X
4. Both the regression equations will have the same signs.
5. The sign of regression coefficients and correlation coefficient will be same.
6. The geometric mean of two regression coefficients is equal to coefficient of correlation.
   $\sqrt{b_{xy} \times b_{yx}} = r$
7. The product of two regression coefficients is equal to coefficient of determination.
   $b_{xy} \times b_{yx} = r^2$
8. When there is perfect correlation between X and Y, then b_{xy} and x b_{yx} will be reciprocals of each other.
9. When the standard deviations of both the variables are same, then the values of regression coefficients and correlation coefficient will be same.
10. Both the regression coefficients will not be greater than 1. In other words, one of them can be greater than 1; or both of them can be less than 1.

MULTIPLE REGRESSION

In multiple regression there are more than two variables. Here, we examine the effect of two or more x3ndependent variables on one dependent variable. Suppose there are three variables, namely, x_1, x_2 and x_3. Here we may find three regression equations. They are:

1. Regression equation of x_1 on x_2 and x_3
2. Regression equation of x_2 on x_1 and x_3
3. Regression equation of x_3 on x_1 and x_2

Equations of regression lines are generally termed as equations of planes of regression. Following are the formulae for computing the above 3 regression plane equations:

1. Regression equation of x_1 on x_2 and x_3:

$$(x_1 - \bar{x}_1) = b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3)$$

2. Regression equation of x_2 on x_1 and x_3:

$$(x_2 - \bar{x}_2) = b_{21.3}(x_1 - \bar{x}_1) + b_{23.1}(x_3 - \bar{x}_3)$$

3. Regression equation of x_3 on x_1 and x_2:

$$(x_3 - \bar{x}_3) = b_{31.2}(x_1 - \bar{x}_1) + b_{32.1}(x_2 - \bar{x}_2)$$

where $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$ are actual means of x_1, x_2 and x_3 respectively.

Yule's Notation

Yule suggested that, the above equations may be simplified by taking $(x_3 - \bar{x}_3) = X_1$, $(x_3 - \bar{x}_3) = X_2$ and $(x_3 - \bar{x}_3) = X_3$. Then the equations of planes of regression are:

1. Regression equation of x_1 on x_2 and x_3:

   $$\mathbf{X_1 = b_{12.3} X_2 + b_{13.2} X_3}$$

2. Regression equation of x_2 on x_1 and x_3:

   $$\mathbf{X_2 = b_{21.3} X_1 + b_{23.1} X_3}$$

3. Regression equation of x_3 on x_1 and x_2:

   $$\mathbf{X_3 = b_{31.2} X_1 + b_{32.1} X_2}$$

In the above three equations, we used six regression coefficients. Following are the formulae for computing regression coefficients:

$$b_{12.3} = (\sigma_1 / \sigma_2)\,[(r_{12} - r_{13}r_{23})/(1 - r_{23}^2)]$$

$$b_{13.2} = (\sigma_1 / \sigma_3)\,[(r_{13} - r_{12}r_{23})/(1 - r_{23}^2)]$$

$$b_{21.3} = (\sigma_2 / \sigma_1)\,[(r_{12} - r_{23}r_{13})/(1 - r_{13}^2)]$$

$$b_{23.1} = (\sigma_2 / \sigma_3)\,[(r_{23} - r_{12}r_{13})/(1 - r_{13}^2)]$$

$$b_{31.2} = (\sigma_3 / \sigma_1)\,[(r_{13} - r_{23}r_{12})/(1 - r_{12}^2)]$$

$$b_{32.1} = (\sigma_3 / \sigma_2)\,[(r_{23} - r_{13}r_{12})/(1 - r_{12}^2)]$$

Qn: If $r_{12} = 0.7$, $r_{31} = r_{23} = 0.5$, $\sigma_1 = 2$, $\sigma_2 = 3$ and $\sigma_3 = 3$, find the equation of plane of regression x_1 on x_2 and x_3.

Sol:

Here, means of the variables are not given, and therefore, it is convenient to write the equations of planes of regression using Yule's notation.

Equation of plane of regression x_1 on x_2 and x_3 is :

$X_1 = b_{12.3}X_2 + b_{13.2}X_3$

$b_{12.3} = (\sigma_1/\sigma_2)[(r_{12} - r_{13}r_{23})/(1 - r_{23}^2)]$

$= (2/3)[(0.7 - 0.5 \times 0.5)/(1 - 0.5^2)]$

$= (2/3)[(0.7 - 0.25)/(1 - 0.25)]$

$= (2/3)[0.45/0.75] = (2/3)(0.6) = \underline{0.4}$

$b_{13.2} = (\sigma_1/\sigma_3)[(r_{13} - r_{12}r_{23})/(1 - r_{23}^2)]$

$= (2/3)[(0.5 - 0.7 \times 0.5)/(1 - 0.5^2)]$

$= (2/3)[(0.5 - 0.35)/(1 - 0.25)]$

$= (2/3)[0.15/0.75] = (2/3)(0.2) = \underline{0.133}$

$\therefore \underline{X_1 = 0.4X_2 + 0.133X_3}$

Qn: In a trivariate distribution, $\bar{x}_1 = 53, \bar{x}_2 = 52$, $\bar{x}_3 = 51$, $\sigma_1 = 3.88$, $\sigma_2 = 2.97$, $\sigma_3 = 2.86$, $r_{23} = 0.8$, $r_{31} = 0.81$ and $r_{12} = 0.78$. Find the linear regression equation of x_1 on x_2 and x_3.

Sol:

Here means of the variables are given.

$\therefore$ Regression equation of x_1 on x_2 and x_3 is:

$(x_1 - \bar{x}_1) = b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3)$

$b_{12.3} = (\sigma_1/\sigma_2)[(r_{12} - r_{13}r_{23})/(1 - r_{23}^2)]$

$= (3.88/2.97)[(0.78 - 0.81 \times 0.8)/(1 - 0.8^2)]$

$= (3.88/2.97)[(0.78 - 0.648)/(1 - 0.64)]$

$= (3.88/2.97)\,[0.132/0.36] = (3.88/2.97)\,(0.367) = \underline{0.4794}$

$b_{13.2} = (\sigma_1/\sigma_3)\,[(r_{13} - r_{12}r_{23})/(1 - r_{23}^2)]$

$= (3.88/2.86)\,[(0.81 - 0.78 \times 0.8)/(1 - 0.8^2)]$

$= (3.88/2.86)\,[(0.81 - 0.624)/(1 - 0.64)]$

$= (3.88/2.86)\,[0.186/0.36] = 1.357 \times 0.517 = \underline{0.702}$

$\therefore (x_1 - 53) = 0.4794(x_2 - 52) + 0.702(x_3 - 51)$

$= 0.4794\,x_2 - 24.929 + 0.702\,x_3 - 35.8$

$x_1 = 0.4794\,x_2 + 0.702\,x_3 - 35.8 - 24.929 + 53$

$= 0.4794\,x_2 + 0.702\,x_3 - 7.729$

$\therefore x_1 = 0.4794\,x_2 + 0.702\,x_3 - 7.729$

$\underline{\underline{x_1 = 0.48x_2 + 0.7x_3 - 7.73}}$

Qn: In a trivariate distribution, $\bar{x}_1 = 28.02$, $\bar{x}_2 = 4.91$, $\bar{x}_3 = 594$, $\sigma_1 = 4.4$, $\sigma_2 = 1.1$, $\sigma_3 = 80$, $r_{23} = -0.56$, $r_{31} = -0.4$ and $r_{12} = 0.8$. Estimate the value of x_1 when $x_2 = 6$ and $x_3 = 650$.

Sol:

Here, to estimate the value of x_1, we have to find the regression equation of x_1 on x_2 and x_3.

Regression equation of x_1 on x_2 and x_3 is:

$(x_1 - \bar{x}_1) = b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3)$

$b_{12.3} = (\sigma_1/\sigma_2)\,[(r_{12} - r_{13}r_{23})/(1 - r_{23}^2)]$

$= (4.4/1.1)\,[(0.8 - -0.56 \times -0.4)/(1 - -0.56^2)]$

$= (4)\,[(0.8 - 0.224)/(1 - 0.314)]$

$= 4\,[0.576/0.6869] = 4 \times 0.84 = \underline{3.36}$

$b_{13.2} = (\sigma_1/\sigma_3)\,[(r_{13} - r_{12}r_{23})/(1 - r_{23}^2)]$

$= (4.4/80)\,[(-0.4 - 0.8 \times -0.56)/(1 - -0.56^2)]$

$= 0.055\,[(-0.4 - -0.448)/(1 - 0.314)]$

$= 0.055\,[(-0.4 + 0.448)/0.686] = 0.055 \times [0.048/0.868]$

$= 0.055 \times 0.07 = \underline{\underline{0.00385}}$

$\therefore (x_1 - 28.02) = 3.36(x_2 - 4.91) + 0.00385(x_3 - 594)$

$= 3.36\, x_2 - 16.498 + 0.00385\, x_3 - 2.287$

$x_1 = 3.36x_2 + 0.00385x_3 - 16.498 - 2.287 + 28.02$

$= 3.36\, x_2 + 0.00385x_3 + 9.235$

$\therefore x_1 = \underline{3.36x_2 + 0.00385x_3 + 9.235}$

REVIEW QUESTIONS:

1. What do you mean by regression analysis?
2. What are the different types of regression?
3. What do you mean by linear and non-linear regressions?
4. What do you mean by line of best fit?
5. What are the different methods for computing regression equations?
6. What do you mean by regression lines?
7. What are the important properties of regression coefficients?
8. What do you mean by multiple regressions?
9. Distinguish between correlation and regression analysis.
10. What do you mean by normal equation method for computing regression equations?
11. From the following data, fit the regression equations X on Y and Y on X:

X	102	80	100	88	84	82	90	96	97	83	79	88
Y	100	97	98	83	84	72	84	101	102	88	84	87

Also find the value of X, if Y 90 and Y if X = 105

12. In a trivariate distribution, $\bar{x}_1 = 10$, $\bar{x}_2 = 15$, $\bar{x}_3 = 12$, $\sigma_1 = 3$, $\sigma_2 = 4$, $\sigma_3 = 5$, $r_{23} = 0.4$, $r_{31} = 0.6$ and $r_{12} = 0.7$. Determine the regression equation of X_1 on X_2 and X_3.

CHAPTER 4

PROBABILITY DISTRIBUTIONS (THEORETICAL DISTRIBUTIONS)

Definition

Probability distribution (Theoretical Distribution) can be defined as a distribution obtained for a random variable on the basis of a mathematical model. It is obtained not on the basis of actual observation or experiments, but on the basis of probability law.

Random variable

Random variable is a variable who value is determined by the outcome of a random experiment. Random variable is also called chance variable or stochastic variable.

For example, suppose we toss a coin. Obtaining of head in this random experiment is a random variable. Here the random variable of "obtaining heads" can take the numerical values.

Now, we can prepare a table showing the values of the random variable and corresponding probabilities. This is called probability distributions or theoretical distribution.

In the above, example probability distribution is :-

Obtaining of heads (X)	Probability of obtaining heads P(X)
0	½
1	½
	∑ P (X) = 1

Properties of Probability Distributions:

1. Every value of probability of random variable will be greater than or equal to zero.
   i.e., $P(X) \leq 0$
   i.e., P(X) is always non-negative value

2. Sum of all the probability values will be 1
   $\sum P(X) = 1$

Question:

A distribution is given below. State whether this distribution is a probability distribution.

X:	0	1	2	3	4
P(X):	0.01	0.10	0.50	0.30	0.90

Solution

Here all values of P(X) are more than zero; and sum of all P(X) value is equal to 1

Since two conditions, namely $P(X) \leq 0$ and $\sum P(X) = 1$, are satisfied, the given distribution is a probability distribution.

MATHEMATICAL EXPECTATION
(EXPECTED VALUE)

If X is a random variable assuming values $x_1, x_2, x_3, \ldots\ldots\ldots, x_n$ with corresponding probabilities $P_1, P_2, P_3, \ldots\ldots\ldots, P_n$, then the Expectation of X is defined as $x_1p_1 + x_2p_2 + x_3p_3 + \ldots\ldots\ldots + x_np_n$.

$$E(X) = \sum [x.\ p(x)]$$

Expected Value [i. e, E (X)] = $\sum [x.\ p(x)]$

Qn:

A petrol pump proprietor sells on an average Rs. 80,000/- worth of petrol on rainy days and an average of Rs. 95.000 on clear days. Statistics from the meteorological department show that the probability is 0.76 for clear weather and 0.24 for rainy weather on coming Wednesday. Find the expected value of petrol sale on coming Wednesday.

There are three alternative proposals before a business man to start a new project:-

Proposal I:	Profit of Rs. 5 lakhs with a probability of 0.6 or a loss of Rs. 80,000 with a probability of 0.4.
Proposal II:	Profit of Rs. 10 laksh with a probability of 0.4 or a loss of Rs. 2 lakhs with a probability of 0.6
Proposal III:	Profit of Rs. 4.5 lakhs with a probability of 0.8 or a loss of Rs. 50,000 with a probability of 0.2

If he wants to maximize profit and minimize the loss, which proposal he should prefer?

Sol:

Here, we should calculate the mathematical expectation of each proposal.

Expected Value $E(X) = \sum [x.\ p(x)]$

Expected Value of Proposal I = (500000 x 0.6) + (80000 x 0.4) = 300000 – 32,000

= Rs. 2,68,000

Expected Value of Proposal II = (10,00.000 × 0.4) +(-2,00.000) = 400000 - 120000

= Rs: 2,80,000.

Expected Value of Proposal III = (450000 × 0.8) + (- 50000 x 0.2) = 360000 - 10000

= Rs: 3,50,000

Since expected value is highest in case of proposal III, the businessman should prefer the proposal III.

Classification of Probability Distribution

Following are the different types of probability distribution:

1. Binomial Distribution
2. Poisson Distribution
3. Uniform Distribution
4. Exponential Distribution
5. Normal Distribution

REVIEW QUESTIONS:

1. Define frequency distribution.
2. Define Random Variable.
3. What are the important properties of frequency distribution?
4. What is meant by Expected Value?
5. What are the different types of probability distributions?

CHAPTER 5

BIONOMIAL DISTRIBUTION

Meaning & Definition:

Binomial Distribution is associated with James Bernoulli, a Swiss Mathematician. Therefore, it is also called Bernoulli distribution. Binomial distribution is the probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure. In other words, it is used to determine the probability of success in experiments on which there are only two mutually exclusive outcomes. Binomial distribution is discrete probability distribution.

Binomial Distribution can be defined as follows: "A random variable r is said to follow Binomial Distribution with parameters n and p if its probability function is:

$$P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$$

Where, P = probability of success in a single trial

q = 1 – p

n = number of trials

r = number of success in 'n' trials.

Assumption of Binomial Distribution

(Situations where Binomial Distribution can be applied)

Binomial distribution can be applied when:-

1. The random experiment has two outcomes i.e., success and failure.
2. The probability of success in a single trial remains constant from trial to trial of the experiment.
3. The experiment is repeated for finite number of times.
4. The trials are independent.

Properties (Features) of Binomial Distribution

1. It is a discrete probability distribution.
2. The shape and location of Binomial distribution changes as 'p' changes for a given 'n'.
3. The mode of the Binomial distribution is equal to the value of 'r' which has the largest probability.
4. Mean of the Binomial distribution increases as 'n' increases with 'p' remaining constant.
5. The mean of Binomial distribution is np.
6. The Standard deviation of Binomial distribution is $\sqrt{npq}$
7. The variance of Binomial Distribution is npq
8. If 'n' is large and if neither 'p' nor 'q' is too close zero, Binomial distribution may be approximated to Normal Distribution.
9. If two independent random variables follow Binomial distribution, their sum also follows Binomial distribution.

Qn: Six coins are tossed simultaneously. What is the probability of obtaining 4 heads?

Sol: $P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

$r = 4$

$n = 6$

$p = ½$

$q = 1 - p = 1 - ½ = ½$

$\therefore p(r = 4) = {}^{6}C_{4} (½)^{4} (½)^{6-4}$

$$= \frac{6!}{(6-4)!4!} \times (½)^{4+2}$$

$$= \frac{6!}{2!\,4!} \times (½)^{6}$$

$$= \frac{6 \times 5}{2 \times 1} \times \frac{1}{64}$$

$$= \frac{30}{128}$$

$$= 0.234$$

Qn: The probability that Sachin scores a century in a cricket match is 1/3. What is the probability that out of 5 matches, he may score century in:

(1) Exactly 2 matches
(2) No match

Sol: Here p = 1/3 , n = 5, q = 2/3

$P(r) = {}^{n}C_{r}\, p^{r} q^{n-r}$

(1) Probability that Sachin scores centuary in exactly 2 matches is:

$$P(r = 2) = {}^{5}C_{2}\, 1/3^{2}\, 2/3^{5-2}$$

$$= \frac{5!}{(5-2)!2!} \times \frac{1}{9} \times \frac{8}{27}$$

$$= \frac{5 \times 4}{2 \times 1} \times \frac{1}{9} \times \frac{8}{27}$$

$$= \frac{160}{486}$$

$$= \frac{80}{243} = 0.329$$

(2) Probability that Sachin scores century in no match is:

$$P(r = 0) = {}^5C_0\, 1/3^{\,0}\, 2/3^{5-0}$$

$$\frac{5!}{(5-0)!\,0!} \quad X \quad 1 \quad x \quad 2/3^5$$

$$= 1 \times 1 \times (2/3)^5$$

$$= \frac{32}{243}$$

$$= \underline{0.132}$$

Qn: Consider families with 4 children each. What percentage of families would you expect to have :-

(a) Two boys and two girls
(b) At least one boy
(c) No girls
(d) At the most two girls

(a) P(having a boy) = ½

P (having a girl) = ½

n = 4

P (getting 2 boys & 2 girls) = p (getting 2 boys)

$$= p(r = 2) = {}^4C_2\, (½)^{\,2}\, (1/2)^{4-2}$$

$$= \frac{4!}{(4-2)!\,2!} \quad x \ (1/2)^2 \ x \ (½)^2$$

$$= \frac{4 \times 3}{2} \times (1/2)^4$$

$$= 6 \times 1/16 = 6/16 = \underline{3/8}$$

∴ Percentage of families with 2 boys and 2 girls = (3/8) x 100 = 37.5 % .

(b) Probability of having at least one boy:

= p (having one boy or having 2 boys or having 3 boys or having 4 boys)

= p (having one boy) + p (having 2 boys) + p (having 3 boys)

+ p (having 4 boys)

= p (r=1) + p (r = 2) + p(r = 3) + p (r = 4)

$$= 4/16 + 6/16 + 4/46 + 1/16 = 15/16$$

∴ Percentage of families with at least one boy = (15/16) x 100 = 93.75 %

(c) Probability of having no girls = Probability of having 4 boys

$P(r = 4) = 4C4\ (½)^4 (½)^{4-4} = 1 \times (½)^4 = 1/16$

∴ Percentage of families with at least one boy = (1/16) x 100 = 6.25 %

(d) Probability of having at the most 2 girls = P(having 2 or 1 or 0 girls)

= P(having 2 boys or 3 boys or 4 boys)

= 11/16.

∴ Percentage of families with at least one boy = (11/16) x 100 = 68.75 %

Qn: For a binomial distribution mean = 4 and variance = 12/9. Find n.

Sol. Mean np = 4 …….. (1)

Variance npq = 12/9 …… (2)

Divide (2) by (1):

We get $q = 12/9 \div 4 = 12/36 = 1/3$

$\therefore p = 1 - 1/3 = 2/3$

$\therefore n \times 2/3 = 4,\ n = 4 \times 3/2 = 6$

n = 6

Fitting a Binomial Distribution

Steps:

1. Find the value of n, p and q
2. Substitute the values of n, p and q in the Binomial Distribution function of $^nC_r p^r q^{n-r}$
3. Put r = 0, 1, 2, ... in the function $^nC_r p^r q^{n-r}$
4. Multiply each of such terms by total frequency (N) to obtain the expected frequency.

Qn: Eight coins were tossed together for 256 times. Fit a Binomial Distribution of getting heads. Also find mean and standard deviation.

Sol: p (getting head in a toss) = ½ , n = 8, q = ½

Binomial Distribution function is $p(r) = {}^nC_r p^r q^{n-r}$

Put r = 0, 1, 2, 3..... 8, then are get the terms of the Binomial Distribution.

Binomial Distribution		
No. of Heads (x)	P(x)	Expected Frequency = P(x) x 256
0	$8C_0 (1/2)^0 (1/2)^8 = 1/256$	1
1	$8C_1 (1/2)^1 (1/2)^7 = 8/256$	8
2	$8C_2 (1/2)^2 (1/2)^6 = 28/256$	28

3	$8C_3 (1/2)^3 (1/2)^5 = 56/256$	56
4	$8C_4 (1/2)^4 (1/2)^4 = 70/256$	70
5	$8C_5 (1/2)^5 (1/2)^3 = 56/256$	53
6	$8C_6 (1/2)^6 (1/2)^2 = 28/256$	28
7	$8C_7 (1/2)^7 (1/2)^1 = 8/256$	8
8	$8C_8 (1/2)^8 (1/2)^0 = 1/256$	1
Total		256

Mean = np = 8*1/2 = 4

S.D = $\sqrt{npq} = \sqrt{8*1/2*\ ½} = \sqrt{2} = 1.414$

REVIEW QUESTIONS:

1. Define Binomial Distribution.
2. What are the important properties of Binomial Distribution?
3. Examine whether the following statement is true:

   " For a Binomial Distribution, mean = 10 and S D = 4"
4. For a Binomial Distribution, mean = 6 and S D = $\sqrt{2}$. Find parameters. Write down all the terms of the distribution.

****

CHAPTER 6

POISSON DISTRIBUTION

Meaning and Definition

Poisson distribution is a limiting form of Binomial Distribution. In Binomial distribution, the total number of trials is known previously. But in certain real life situations, it may be impossible to count the total number of times a particular event occurs or does not occur. In such cases Poisson distribution is more suitable.

Poison Distribution is a discrete probability distribution. It was originated by Simeon Denis Poisson.

A random variable "r" said to follow Binomial distribution if its probability function is:

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

Where r = random variable (i.e., number of success in '**n**' trials)

e = 2.7183

m = mean of Poisson distribution.

Properties of Poisson Distribution

1. Poisson distribution is a discrete probability distribution.
2. Poisson distribution has a single parameter 'm'. When 'm' is known all the terms can be found out.
3. It is a positively skewed distribution.
4. Mean and Variance of Poisson distribution are equal to 'm'.
5. In Poisson distribution, the number of success is relatively small.
6. Standard deviation of Poisson distribution is $\sqrt{m}$.

Practical situations where Poisson distribution can be used

1. To count the number of telephone calls arising at a telephone switch board in a unit of time.
2. To count the number of customers arising at the super market in a unit of time.
3. To count the number of defects in Statistical Quality Control.
4. To count the number of bacteria per unit.
5. To count the number of defectives in a park of manufactured goods.

6. To count the number of persons dying due to heart attack in a year.
7. To count the number of accidents taking place in a day on a busy road.

Qn: A fruit seller, from his past experience, knows that 3 of apples in each basket will be defectives. What is the probability that exactly 4 apples will be defective in a given basket?

Sol. $m = 0.03$

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

$\therefore$ P (exactly 4 apples are defective) = $(e^{-3} \cdot 3^4) / 4!$

$= (0.0498 \times 81) / 24$

$= \underline{0.16807}$

Qn: It is known from the past experience that in a certain plant, there are on an average four industrial accidents per year. Find the probability that in a given year there will be less than four accidents. Assume Poisson distribution.

Sol:

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

$m = 4$

$\therefore$ P (exactly 4 apples are defective) = P ($r < 4$)

P ($r < 4$) = P (r = 0 or 1 or 2 or 3)

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

P (r = 0) = $(e^{-4} \cdot 4^0) / 0! = (0.0183 \times 1) / 1 = 0.0183$

P (r = 1) = $(e^{-4} \cdot 4^1) / 1! = (0.0183 \times 4) / 1 = 0.0732$

P (r = 2) = $(e^{-4} \cdot 4^2) / 2! = (0.0183 \times 16) / 2 = 0.1464$

P (r = 3) = $(e^{-4} \cdot 4^3) / 3! = (0.0183 \times 64) / 6 = 0.1952$

$\therefore$ P ($r < 4$) = 0.0183+ 0.0732 + 0.1464 + 0.1952 = $\underline{0.4331}$

Qn: Out of 500 items selected for inspection, 0.2% is found to be defective. Find how many lots

will contain exactly no defective if there are 1000 lots.

Sol:

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

$m = 500 \times 0.2\% = 1$

$\therefore P(r = 0) = (e^{-1}\ 1^0) / 0! = (0.3679 \times 1) / 1 = 0.3679$

$\therefore$ No. of lots having zero defective = $0.3679 \times 1000 = \underline{368}$

Qn: In a certain factory producing optical lenses, there is a small chance of 1/500 for any one lens to be defective. The lenses are supplied in packets of 10. Use P.D to calculate the approximate number of packets containing no defectives, one defective, two defectives and three defective lenses respectively in a consignment of 20,000 packets.

Sol:

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

$m = 10 \times 1/500 = 0.02$

$\therefore P(r = 0) = (e^{-0.02} \times 0.02^{\ 0}) / 0! = (0.9802 \times 1) /1 = 0.9802$

$\therefore$ No. of packets containing no defective lens = $0.9802 \times 20000 = \underline{19604}$

$P(r = 1) = (e^{-0.02} \times 0.02^{\ 1}) / 1! = (0.9802 \times 0.02) /1 = 0.0196$

$\therefore$ No. of packets containing no defective lens = $0.0196 \times 20000 = \underline{392}$

$P(r = 2) = (e^{-0.02} \times 0.02^{\ 2}) / 2! = (0.9802 \times 0.0004) /2 = 0.00019604$

$\therefore$ No. of packets containing no defective lens = $0.00019604 \times 20000 = \underline{4}$

$P(r = 3) = (e^{-0.02} \times 0.02^{\ 3}) / 3! = (0.9802 \times 0.000008) /6 = 0.0000013069$

$\therefore$ No. of packets containing no defective lens = $0.0000013069 \times 20000 = \underline{0}$

Qn: A Systematic sample of 100 pages was taken from a dictionary and the observed frequency distribution of foreign words per page was found to be as follows:

No. of foreign words per page (x) :	0	1	2	3	4	5	6

Frequency (f) : 48 27 12 7 4 1 1

Calculate the expected frequencies using Poisson distribution.

Sol: At first, we have to know the parameter of P.D, which is equal to the mean of the given distribution. So find the mean of the distribution:

Mean = (εfx) / εf

x	0	1	2	3	4	5	6	
f	48	27	12	7	4	1	1	N = εf = 100
fx	0	27	24	21	16	5	6	(εfx) = 99

Mean = 99/100 = 0.99

Calculation of Expected Frequencies		
X	P(x) = $(e^{-m} . m^x) / x!$	Expected Frequency = P(x) . N
0	$(e^{-0.99} . 0.99^0)/0!$ =(0.3716 x1)/1 = 0.3716	0.3716 x 100 = 37.16 = 37
1	$(e^{-0.99} . 0.99^1)1!$ =(0.3716 x 0.99)/1= 0.3679	0.3716 x 100 = 37.16 = 37
2	$(e^{-0.99} . 0.99^2)2!$ =(0.3716 x 0.98)/2= 0.1821	0.1821 x 100 = 18.21 = 18
3	$(e^{-0.99} . 0.99^3)3!$ =(0.3716 x 0.97)/6= 0.0601	0.0601 x 100 = 6.01 = 6
4	$(e^{-0.99} . 0.99^4)4!$=(0.3716 x 0.96)24= 0.0149	0.0149 x 100 = 1.49 = 2
5	$(e^{-0.99} . 0.99^5)5!$=(0.3716 x 0.95)120= 0.0029	0.0029 x 100 = 0.29 = 0
6	$(e^{-0.99} . 0.99^6)6!$=(0.3716 x 0.94)720= 0.0005	0.0005 x 100 = 0.05 = 0
	Total	**100**

REVIEW QUESTIONS:

1. Define Poisson distribution.
2. What are the important properties of P.D?
3. What are the situations under which P D can be applied?
4. Write down the probability function of P.D. whose mean is 2. What is its variance?
5. A machine is producing 4% defectives. What is the probability of getting at least 4 defectives in a sample of 50 =, using (a) BD and (b) PD?
6. The following table gives the number of days in a 50 day period during which automobile accidents occurred in a certain part of the city. Fit a Poisson distribution to the data:

No. of accidents	0	1	2	3	4
No. of days	19	18	8	4	1

****

CHAPTER 7

NORMAL DISTRIBUTION

Meaning and Definition

The normal distribution is a continuous probability distribution. It was first developed by De-Moivre in 1733 as limiting form of binomial distribution. Fundamental importance of normal distribution is that many populations seem to follow approximately a pattern of distribution as described by normal distribution. Numerous phenomena such as the age distribution of any species, height of adult persons, intelligent test scores of students, etc. are considered to be normally distributed.

Definition of Normal Distribution

A continuous random variable, 'X', said to follow Normal Distribution if its probability function is:

$P(x) =$	$\frac{1}{\sqrt{2\pi}\,.\,\sigma}\, e^{-\frac{1}{2}\left[(x-\mu)/\sigma\right]^2}$

Properties of Normal Distribution (Normal Curve)

1. Normal distribution is a continuous distribution.
2. Normal curve is symmetrical about the mean.
3. Both sides of normal curve coincide exactly.
4. Normal curve is a bell shaped curve.
5. Mean, Median and Mode coincide at the centre of the curve.
6. Quantities are equi-distant from median. $Q_3 - Q_2 = Q_2 - Q_1$
7. Normal curve is asymptotic to the base line.
8. Total area under a normal curve is 100%.
9. The ordinate at the mean divide the whole area under a normal curve into two equal parts. (50% on either side).
10. The height of normal curve is at its maximum at the mean.
11. The normal curve is unimodel, i.e., it has only one mode.
12. Normal curve is mesokurtic.
13. No portion of normal curve lies below the x-axis.
14. Theoretically, the range of normal curve is $-\infty$ to $+\infty$. But practically the range is $\mu - 3\sigma$ to $\mu + 3\sigma$.
15. Area under the normal curve is distributed as follows: (Area property)
    (a) $\mu \pm \sigma$ covers 68.27% area
    (b) $\mu \pm \sigma$ covers 95.45% area
    (c) $\mu \pm \sigma$ covers 99.73% area

Importance or Uses of Normal Distribution

The normal distribution is of central importance because of the following reasons:

1. The discrete probability distributions such as Binomial distribution and Poisson distribution tend to normal distribution as 'n' becomes large.
2. Almost all sampling distributions conform to the normal distribution for large values of 'n'.
3. Many tests of significance are based on the assumption that the parent population from which samples are drawn follows normal distribution.
4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. Normal distribution finds applications in Statistical Quality Control.
6. Many distributions in social and economic data are approximately normal. For example, birth, death, etc. are normally distributed.

Area under Standard Normal Curve

In case of normal distribution, probability is determined on the basis of area. But the area we have to calculate the ordinate of z – scale.

The scale to which the standard deviation is attached is called z-scale.

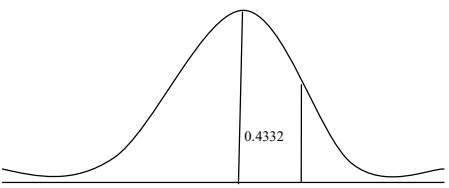
$Z = (x - \mu) / \sigma$

Qn: The variable, x, follows normal distribution with mean = 45 and S.D = 10. Find the probability that $x \geq 60$.

Sol: $\mu = 45$, $\sigma = 10$, $x = 60$

$Z = (x - \mu) / \sigma$

$Z = (60 - 45) / 10 = 15/10 = 1.5$



$P(x \geq 60)$ means $P(z \geq 1.5)$

$= 0.5 - 0.4332 = 0.0668$

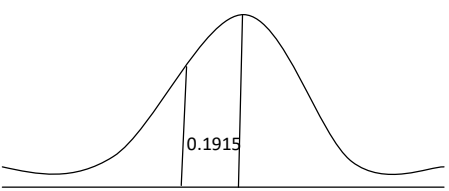
$P(x \geq 60) = \underline{0.0668}$

Qn: The variable, x, follows normal distribution with mean = 45 and S.D = 10. Find the probability that $x \leq 40$.

Sol: $\mu = 45$, $\sigma = 10$, $x = 40$

$Z = (x - \mu) / \sigma$

$Z = (40 - 45) / 10 = -5/10 = -0.5$



$P(x \leq 40)$ means $P(z \geq -0.5)$

$= 0.5 - 0.1915 = 0.3085$

P ($x \leq 40$) = 0.3085

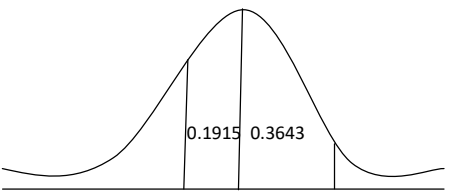
Qn: The variable, x, follows normal distribution with mean = 45 and S.D = 10. Find the probability that $40 \leq x \leq 56$.

Sol: $\mu = 45$, $\sigma = 10$, $x_1 = 40$, $x_2 = 56$

$$\mathbf{Z = (x - \mu) / \sigma}$$

When x = 40, Z = (40 -- 45) / 10 = -5 / 10 = - 0.5

When x = 56, Z = (56 -- 45) / 10 = 11 / 10 = 1.1



P ($40 \leq x \leq 56$) means P ($z - 0.5 \leq x \leq 1.5$)

= 0.1915 + 0.3643 = 0.5558

P ($40 \leq x \leq 56$) = 0.5558

Qn: The scores of students in a test follow normal distribution with mean = 80 and S D = 15. A sample of 1000 students has been drawn from the population. Find (1) probability that a randomly chosen student has score between 85 and 95 (2) appropriate number of students scoring less than 60.

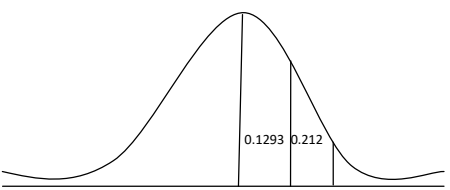
Sol.

(1) $\mu = 80$, $\sigma = 15$, $x_1 = 85$, $x_2 = 95$

$$\mathbf{Z = (x - \mu) / \sigma}$$

When x = 85, Z = (85 – 80) / 15 = 5/15 = 0.333

When x = 95, Z = (95 – 80) / 15 = 15/15 = 1



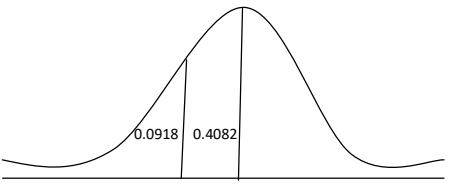
∴ P (85 ≤ x ≤ 95) = P (0.333 ≤ z ≤ 1) = 0.3413 – 0.1293 = 0.212

Probability that a student scores between 85 and 95 = 0.212

(2) P (Less than 60):

When x = 60,

Z = (x – μ)/σ = (60 -- 80)/15 = -20/15 = - 1. 333



P(x < 60) = P(z < -1.333) = 0.5 -- 0.4082 = 0.0918

∴ Number of students scoring less than 60 = 0.0918 x 1000 = 91.8 = 92 students

Computation of Z-value when Area is known

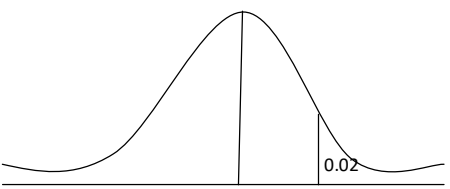
Qn: In a competitive examination, 5000 candidates have appeared. Their average mark was 62 and S.D was 12. If there are only 100 vacancies, find the minimum marks that one should score in order to get selection.

Sol: **μ** = 62, σ = 12

Number of vacancies = 100

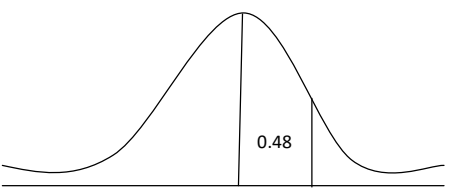
Percentage of vacancies to the total number of candidates = (100/5000) x 100 = 2% = 0.02

Area corresponds to the students who will get selection is shown in the following normal curve:



Therefore, the area to the left of the above area of 0.02 is:

Z = (40 – 45) / 10 = -5/10 = -- 0.5



Locate the area of 0.48 in the table and find the Z – value corresponds to it.
The table shows the area nearest to 0.48 is 0.4798, and the corresponding z-value is 2.05
$Z = 2.05$
$(x - \mu)/\sigma = 2.05$
$(x - 62)/12 = 2.05, \quad x - 62 = 2.05 \times 12$
$x - 62 = 24.6, \quad \therefore x = 24.6 + 62 = 86.6$
∴ The minimum marks one should score to get section = 86.6 marks

Construction of Normal Distribution

Procedure:

1. Find the mean and S.D of the given distribution and take them as μ and σ (parameters) of the normal distribution.
2. Take the lower limit of each class as the x values.
3. Calculate the z-value corresponding to each x-value by using formulae $z = (x—\mu)/\sigma$. Z-value of first and last values need not be computed.
4. Find the area corresponds to z-value from the standard normal distribution table. The area corresponds to the first and last z-values will be 0.5.
5. Find the area of each class using the area (probability) of respective class limits. (Take the difference in case of same signs; and take the total in case of opposite signs)
6. Multiply tye area of ech class by the total frequency to the frequency of the class.

   The new frequency distribution with theoretical frequencies will be a normal approximation to the given frequency distribution.

Qn: Fit a normal distribution to the following data:

X	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	4	22	48	66	40	16	4

Sol:

Computation of Mean and Standard deviation							
Class	**Mid point (m)**	**F**	**d (m-35)**	**d'**	**fd'**	**d'^2**	**fd'^2**
10-20	15	4	-20	-2	-8	4	16
20-30	25	22	-10	-1	-22	1	22
30-40	35	48	0	0	0	0	0
40-50	45	66	10	1	66	1	66
50-60	55	40	20	2	80	4	160
60-70	65	16	30	3	48	9	144
70-80	75	4	40	4	16	16	64
		200			180		472

$\bar{x} = A + [(\varepsilon fd')/N] \times C, \qquad \bar{x} = 35 + [(180/200)x10], \quad = 35+9 = 44$

$S.D = \sqrt{(\varepsilon fd'^2/N) - [(\varepsilon fd')/N]^2} \times 10 = \sqrt{1.55} \times 10 = 12.45$

$\therefore \mu = 44$ and $\sigma = 12.45$

Computation of Expected Frequencies				
Lower limit	$Z = (x—\mu)/\sigma$	Area	Area of class	Expected Frequency
10	-2.73	0.5000	0.0268	5
20	-1.93	0.4732	0.1046	21
30	-1.12	0.3686	0.2431	49
40	-0.32	0.1255	0.3099	62
50	0.48	0.1884	0.2171	43
60	1.29	0.4015	0.0802	16
70	2.09	0.4817	0.0183	4
80	2.89	0.5000		
			Total	**200**

Review Questions:

1. Define normal distribution.
2. What are the important properties of normal distribution?
3. Explain the importance of normal distribution.
4. Explain the procedure for construction of normal distribution.
5. If x follows a normal distribution with mean 12 and variance 16, find $P(x \geq 20)$.
6. The weekly wages of 1000 workers are normally distributed with mean of 70 and S.D of 5. Estimate the number of workers whose wages lie between 69 and 72.
7. In an aptitude test administered to 900 students, the mean score is 50 and S.D is 20. Find the number of students securing scores (a) between 30 and 70 (b) exceeding 65. Find the value of the score exceeded by the top 90 students.
8. Construct a normal distribution to the following data of marks obtained by 100 students:

Marks	60-62	63-65	66-68	69-71	72-74
No. of Students	5	18	42	27	8

****

Chapter 8

EXPONENTIAL DISTRIBUTION

Definition of Exponential Distribution

A continuous random variable, x, follows random variable if its probability density function is :

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x > 0 \text{ and } \lambda > 0$$

Parameter of Exponential Distribution

The single parameter of exponential distribution is λ. If we know λ, we can find out all the terms.

Properties of Exponential Distribution

1. The mean of Exponential distribution is = $1/\lambda$
2. The variance of Exponential distribution is = $1/\lambda^2$
3. The first four moments of Exponential distribution are:

$$\mu_1 = 1/\lambda, \quad \mu_2 = 1/\lambda^2, \quad \mu_3 = 2/\lambda^3, \quad \mu_4 = 9/\lambda^4$$

4. The measure of Skewness of Exponential distribution is = $\sqrt{\beta_1} = 2$
5. The measure of Kurtosis of Exponential distribution is = $\beta_2 = 9$
6. The Median of Exponential distribution is = $1/\lambda$

REVIEW QUESTIONS:

1. Define exponential distribution.
2. Write down the first four moments of exponential distribution.
3. What is the skewness of exponential distribution?
4. What about the median of an exponential distribution?
5. What are the important properties of exponential distribution?

Chapter 9

UNIFORM DISTRIBUTION

Definition of Uniform Distribution

A discrete random variable, x, follows uniform distribution if its probability density function is :

$$f(x) = 1/n \quad \text{for } x = x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n$$

For example, when a die is thrown, let x stands for the numbers obtained.

Then $f(x) = 1/6$ for $x = 1, 2, 3, 4, 5, 6$.

Mean of Uniform Distribution

Mean of Uniform Distribution $= \varepsilon x/n$

Variance of Uniform Distribution

Variance of Uniform Distribution $= [\varepsilon x^2/n] - [\varepsilon x/n]^2$

REVIEW QUESTIONS:

1. Define Uniform distribution.
2. What about the mean of uniform distribution?
3. What about the variance of uniform distribution?

CHAPTER 10
STATISTICAL INFERENCE

Basic Concepts

Population: In statistics, 'Population' refers to collection of all individuals or objects or items or things under consideration.

Finite Population: If a population contains a finite number of objects, it is called finite population. Eg: Students in a college.

Infinite Population: If a population contains a infinite number of objects, it is called infinite population. Eg: Stars in the sky.

Sample: A sample is a representative part of the population.

Sample size: Number of units in a sample group is called sample size. If sample size is too small, it may not represent the population. If it is very large, it may require more time and money for investigation. Hence, the size of a sample should be optimum.

Large Sample: If the size of a sample exceeds thirty, it is called as large sample.

Small Sample: If the size of a sample does not exceed thirty, it is called as small sample.

Parameter: It is a statistical measure derived from population elements. If the arithmetic mean is computed from all the elements of a population, it is a population parameter. Here it is called population mean. Population mean is denoted by the symbol μ. Population standard deviation is denoted by σ.

Statistic: It is a statistical measure derived from sample elements. If the arithmetic mean is computed from the elements of a sample group, it is a sample statistic. Here it s called sample mean. Sample mean is denoted by the symbol $\bar{x}$. Sample standard deviation is denoted by 's'.

Statistical inference

Statistical inference refers to the process of selecting samples and using sample statistic to draw inference or conclusion about the population parameter or population distribution. The two main branches of statistical inference are:

(a) Testing of Hypothesis
(b) Estimation

Testing of Hypothesis

Testing of hypothesis is the process under which a statistical hypothesis about a population is formulated and its validity is tested on the basis of a random sample drawn from that population. For testing the validity of a hypothesis, a number of tests are used. All these tests can be classified into two categories, namely (i) parametric tests and (ii) non-parametric tests. Z-test, t-test, Chi-square test, F-test, etc. are commonly used statistical tests.

Procedure for testing hypothesis

(1) Set up null hypothesis and alternative hypothesis
(2) Decide the test statistic (statistical test) applicable for testing hypothesis.
(3) Apply the appropriate formulae for computing the value of the test statistic.
(4) Specify the level of significance. If nothing is mentioned about the level of significance, take 5% level o significance.
(5) Fix the degree of freedom

(6) Locate the table value (critical value) of the test statistic at specified level of significance

(7) Compare the calculated value of the test statistic with the corresponding table value (critical value) and decide whether to accept or reject the null hypothesis. If calculated value of the test statistic is numerically less than the table value, the null hypothesis is accepted. If calculated value of the test statistic is numerically more than the table value, the null hypothesis is rejected.

Hypothesis

Hypothesis is a tentative solution or assumption or proposition about the parameter or nature of the population. It is a logically drawn conclusion about the population.

Null Hypothesis

This is the original hypothesis. A null hypothesis is a hypothesis which formulated for the purpose of rejection. The term "null" refers to 'nil' or 'no' or 'amounting to nothing'. This hypothesis is generally set up as there is no significant difference between the sample statistic and population parameter. A null hypothesis is denoted by H_0

Alternative Hypothesis

Any hypothesis other than null hypothesis is called alternative hypothesis. It is the hypothesis which is accepted when the null hypothesis is rejected. An alternative hypothesis is denoted by H_1 or H_a

Sampling Distribution

Sampling distribution is a distribution of sample statistic derived from various samples drawn from the same population. Since sample statistic is a random variable, sampling distribution is a probability distribution.

Standard Error

Standard Error (SE) of a statistic is the standard deviation of the sampling distribution of that statistic. For example, the Standard deviation of the sampling distribution of the sample mean is $\sigma/\sqrt{n}$, where σ = population S.D. and n = sample size. Therefore the Standard Error (SE) of sampling distribution of mean is $\sigma/\sqrt{n}$.

Uses of Standard Error

(1) Standard Error is used for testing a given hypothesis.

(2) Standard Error gives an idea about the reliability of a sample. The reciprocal of Standard Error is a measure of reliability of the sample.

(3) Standard Error can be used to determine the confidence limits for population values like mean, proportion and standard deviation.

Errors in Testing of Hypotheses

In any test of hypothesis is the decision is to accept or to reject a null hypothesis. The

decision is based on the information supplied by the sample data. The four possibilities of the decision are:

(1) Accepting a null hypothesis when it is true
(2) Rejecting a null hypothesis when it is false
(3) Rejecting a null hypothesis when it is true
(4) Accepting a null hypothesis when it is false

It is clear that the possibilities (1) and (2) are correct decisions. But the possibilities (3) and (4) are errors.

Type I Error:

The error which is committed by rejecting the null hypothesis even when it is true is called Type I error. It is denoted by alpha (α).

Type II Error:

The error which is committed by accepting the null hypothesis even when it is wrong is called Type II error. It is denoted by beta (β).

When we try to reduce the possibility for one error, the possibility for the other will be increased. Therefore, a compromise of these two is to be ensured. Type II error is more dangerous than Type I error.

Power of a Test

Probability for rejecting the null hypothesis when the alternative hypothesis is true is called power of a test.

Power of a test = 1 – P(Type II Error)

Level of Confidence

Level of confidence is the probability of accepting a true null hypothesis.

Level of Confidence = 1 – Level of significance.

If Level of significance is 5%, Level of Confidence = 95%.

Level of Significance

Level of Significance is the probability of rejecting a true null hypothesis. Level of Significance is denoted by alpha (α). If nothing is mentioned about the level of significance, it is taken as 5%.

Level of Significance (α) = 1 – level of acceptance

Acceptance Region

The area under the normal curve which represents the acceptance of a null hypothesis (i.e; level of confidence) is called the Acceptance Region or Acceptance Area.

Acceptance Region = 100% -- Rejection Region

Rejection Region (Critical Region)

The area under the normal curve which represents the rejection of a null hypothesis (i.e; level of significance) is called the Rejection Region or Critical Region.

Rejection Region = 100% -- Acceptance region

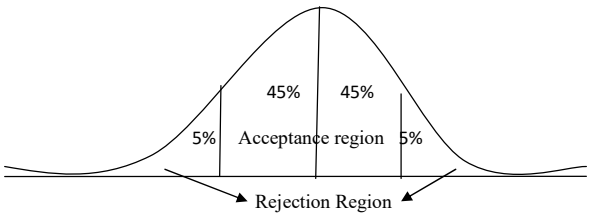
Degree of Freedom

Degree of freedom is defines as the number of independent observations which is obtained by subtracting the number of constraints from the total number of observations.

Degree of freedom (d.f) = Total No. observations – No. of constraints.

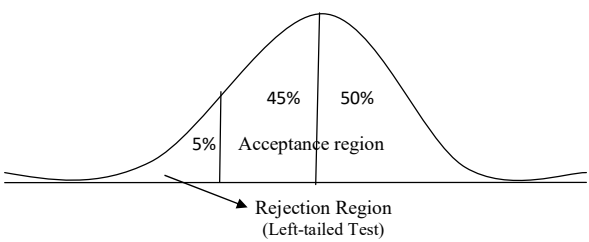
Two-tailed Test

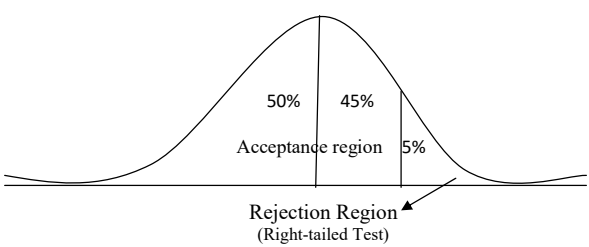
A two tailed test is one in which we reject the null hypothesis if the computed value of the test statistic is significantly greater than or lower than the critical value (table value) of the test statistic. Thus in two tailed tests the critical region is represented by both tails. If we test the hypothesis at 10% level of significance, the size of the acceptance region is 90% and the size of the rejection region is 10% on both sises together.



One-tailed Test

One tailed test is one in which the rejection region is located in only one tail of the normal curve. It may be at left tail or right tail, depending on the alternative hypothesis. If the alternative hypothesis is with '<' (less than) sign, the rejection region is placed on the left tail, and the test is called left-tailed test. If the alternative hypothesis is with '>' (more than) sign, the rejection region is placed on the right tail, and the test is called right-tailed test.





Critical Value (Table Value)

The critical value is the value of the test statistic which separates the rejection region from the acceptance region. It depends on the level of significance and degree of freedom. When the calculated value of the test statistic is numerically less than the critical value, the null hypothesis is accepted. When the calculated value of the test statistic is numerically more than the critical value, the null hypothesis is accepted.

Parametric Tests

When testing of hypothesis is done, if some assumptions are made about the nature of population distribution, then the test statistic applied there is called parametric test. There are number of parametric tests. Eg: t-test, Z test, F test, etc.

Non-Parametric Tests

When testing of hypothesis is done, if no assumptions are made about the nature of population distribution, then the test statistic applied there is called non-parametric test. There are number of non-parametric tests. Eg: Chi-square Test, Sign tests, Signed Rank Tests, Rank Sum Tests, Run Test, Kolmogrov Smirnov Test, etc. Since, no assumptions are made about the nature of population, non-parametric tests are also called distribution-free tests.

TESTING OF GIVEN POPULATION MEAN

This testing of hypothesis is used to test whether the given population mean is true or not. In other words, it is used to test whether there is significant difference between sample mean and population mean.

Procedure:

1. Set up H_0 and H_1

H_0 : There is no significant difference between sample mean and population mean

(i.e; $\mu = \mu_0$)

H_1 : There is significant difference between sample mean and population mean

(i.e; $\mu \neq \mu_0$)

2. Decide the test statistic:

The test statistic applicable here is Z-test or t-test.

If population S.D.(i.e; σ) is known, apply Z-test

If population S.D.(i.e; σ) is unknown but sample is large, apply Z-test

If population S.D.(i.e; σ) is unknown but sample is small, apply t-test

3. Apply the appropriate formula for computing the value of the test statistic:

Z / t = Difference/Standard Error

Difference = Difference between sample mean and the given population mean

Standard Error = $\sigma / \sqrt{n}$ (If population S.D is known)

Standard Error = $s / \sqrt{n}$ (If population S.D is unknown, but sample is large)

Standard Error = $s / \sqrt{n-1}$ (If population S.D is unknown and sample is small)

Where σ = population S.D, s = sample S.D, n = sample size

4. Specify the level of significance. If nothing is mentioned about the level of significance, take 5%.
5. Fix the degree of freedom:

For Z-test, d.f = infinity; For t-test, d.f = n-1

6. Locate the table value (critical value) of the test statistic at specified level of significance and fixed degree of freedom.
7. Compare the calculated value of test statistic with the table value and decide whether to accept or reject the null hypothesis. If calculated value of the test statistic is numerically less than the table value, the null hypothesis is accepted. If calculated value of the test statistic is numerically more than the table value, the null hypothesis is rejected.

Qn: The mean life of random sample of 100 tyres is 15269 km. The manufacturer claims that the average life of tyres manufactured by the company is 15200 km with SD of 1248 km. Test the validity of company's claim.

Sol:

H_0 : There is no significant difference between sample mean and population mean

(i.e; $\mu = 15200$)

H_1 : There is significant difference between sample mean and population mean

(i.e; $\mu \neq 15200$)

Since population S.D is known, the test statistic applicable here is Z-test

$Z = D/SE$

$D = \bar{x} - \mu = 15269-15200 = 69$

$SE = \sigma/\sqrt{n} = 1248/\sqrt{100} = 1248/10 = 124.8$

$Z = 69/124.8 = 0.553$

Level of significance = 5%

Degree of freedom = infinity (population S D is known)

Table value (Critical value) at 5 % level of significance and infinity degree of freedom is 1.96

Since calculated value of Z is less than the critical value, H_0 is accepted. That is, there is no significant difference between sample mean and population mean. μ = 15200. So, we may conclude that the claim of the company is valid.

Qn: A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from the normal population with mean = 100 ad S.D = 8 at 5% level of significance.

Sol:

H_0 : There is no significant difference between sample mean and population mean

(i.e; $\mu = 100$)

H_1 : There is significant difference between sample mean and population mean

(i.e; $\mu \neq 100$)

Since population S.D is known, the test statistic applicable here is Z-test

$Z = D/SE$

$D = \bar{x} - \mu = 100 - 99 = 1$

$SE = \sigma/\sqrt{n} = 8/\sqrt{400} = 8/20 = 0.4$

$Z = 1/0.4 = 2.5$

Level of significance = 5%

Degree of freedom = infinity (population S D is known)

Table value (Critical value) at 5 % level of significance and infinity degree of freedom is 1.96

Since calculated value of Z is more than the critical value, H_0 is rejected. H_1 is accepted. That is, there is significant difference between sample mean and population mean. So, we may conclude that $\mu \neq 100$

Qn: A random sample of 200 bottles of talcum powder gave an average weight of 49.5 gram with a S.D of 2.1 gram. Do we accept the hypothesis of weight per bottle is 50 gram at 1% level of significance?

Sol:

H_0 : There is no significant difference between sample mean and population mean

(i.e; $\mu = 50$)

H_1 : There is significant difference between sample mean and population mean

(i.e; $\mu \neq 50$)

Since sample is large, the test statistic applicable here is Z-test

$Z = D/SE$

$D = \bar{x} - \mu = 50 - 49.5 = 0.5$

$SE = s/\sqrt{n} = 2.1/\sqrt{200} = 2.1/14.142 = 0.148$

$Z = 0.5/0.148 = 3.378$ (Calculated value)

Level of significance = 1%

Degree of freedom = infinity (population is large)

Table value (Critical value) at 1 % level of significance and infinity degree of freedom is 2.58

Since calculated value of Z is more than the critical value, H_0 is rejected. H_1 is accepted. That is, there is significant difference between sample mean and

population mean. So, we may conclude that $\mu \neq 50$ gram

Qn: The average life of 26 bulbs were found to be 1200 hours with a S.D of 150 hours. Test whether these bubs could be considered as a random sample from a normal population with mean 1300 hours.

Sol: H_0 : There is no significant difference between sample mean and population mean

(i.e; $\mu = 1300$)

H_1 : There is significant difference between sample mean and population mean

(i.e; $\mu \neq 1300$)

Since sample is small, the test statistic applicable here is t-test

$t = D/SE$

$D = \bar{x} - \mu = 1300 - 1200 = 100$

$SE = s/\sqrt{n-1} = 150/\sqrt{26-1} = 150/5 = 30$

$t = 100/30 = 3.333$ (Calculated value)

Level of significance = 5%

Degree of freedom = 26-1 = 25 (sample is small)

Table value (Critical value) at 5% level of significance and 25 degree of freedom is 2.06

Since calculated value of Z is more than the critical value, H_0 is rejected. H_1 is accepted. That is, there is significant difference between sample mean and population mean. So, we may conclude that the bulbs could not e drawn from the normal population with mean 1300 hours (i.e; $\mu \neq 1300$).

Qn: A typist claims that he can type at a speed of more than 120 words per minute. Of the 12 tests given to him, he could perform an average of 135 words with a S.D of 40. Is his claim valid at 1% level of significance?

Sol: H_0 : There is no significant difference between sample mean and population mean

(i.e; $\mu = 120$)

H_1 : There is significant difference between sample mean and population mean

(i.e; $\mu > 120$)

Here, the test One-tailed test (Right tailed test)

Since sample is small, the test statistic applicable here is t-test

$t = D/SE$

$D = \bar{x} - \mu = 135 - 120 = 15$

$SE = s/\sqrt{n-1} = 40/\sqrt{12-1} = 40/\sqrt{11} = 40/3.32 = 12.05$

$t = 15/12.05 = 1.245$ (Calculated value)

Level of significance = 1%

Degree of freedom = 12-1 = 11 (sample is small)

Table value (Critical value) at 1% level of significance and 11 degree of freedom is 2.718

Since calculated value of t is less than the critical value, H_0 is accepted. $\mu = 120$. That is, there is no significant difference between sample mean and population mean. So, we may conclude that the claim of the typist that he can type at a speed of more than 120 words is not valid.

TESTING OF SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

This testing of hypothesis is used to test whether the difference between two sample means are significant or not. If the difference is not significant, they are treated as equal; or we may think that the two samples are drawn from the same population.

Procedure:

1. Set up H_0 and H_1

H_0 : There is no significant difference between two sample means(i.e; $\mu_1 = \mu_2$)

H_1 : is no significant difference between two sample means(i.e; $\mu_1 \neq \mu_2$)

2. Decide the test statistic:

The test statistic applicable here is Z-test or t-test.

If population S.D.(i.e; σ) is known, apply Z-test

If population S.D.(i.e; σ) is unknown but sample is large, apply Z-test

If population S.D.(i.e; σ) is unknown but sample is small, apply t-test

3. Apply the appropriate formula for computing the value of the test statistic:

Z / t = Difference/Standard Error

Difference = Difference between two sample means

Standard Error = $\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}$ (If population S.Ds are known)

Standard Error = $\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}$ (If population S.D is unknown, but samples are large)

Standard Error = $\sqrt{(n_1 s_1^2 + n_2 s_2^2)/n_1 + n_2 - 2} \times (1/n_1 + 1/n_2)$

(If population S.Ds are unknown, and samples are small)

[Where σ_1= population S.D of sample 1, s_1 = sample S.D of sample1, n_1 = sample size of sample 1; σ_2= population S.D of sample 2, s_2 = sample S.D of sample2, n_2 = sample size of sample 2]

4. Specify the level of significance. If nothing is mentioned about the level of significance, take 5%.
5. Fix the degree of freedom:

For Z-test, d.f = infinity; For t-test, d.f = $n + n_2 - 1$

6. Locate the table value (critical value) of the test statistic at specified level of significance and fixed degree of freedom.
7. Compare the calculated value of test statistic with the table value and decide whether to accept or reject the null hypothesis. If calculated value of the test statistic is numerically less than the table value, the null hypothesis is accepted. If calculated value of the test statistic is numerically more than the table value, the null hypothesis is rejected.

Qn: The mean yield of wheat from District I was 210Kg per acre from a sample of 100 plots. In another District II, the mean yield was 200 Kg per acre from a sample of 150 plots. Assuming that the S.D of yield of the entire State was 11 Kg, test whether there is any significant difference between the mean yields of the crop in the two districts.

Sol:

District I	District II
$n_1 = 100$	$n_2 = 150$
$\bar{x}_1 = 210$	$\bar{x}_2 = 200$
$\sigma_1 = 11$	$\sigma_2 = 11$

H_0 : There is no significant difference between two sample means(i.e; $\mu_1 = \mu_2$)
H_1 : is no significant difference between two sample means(i.e; $\mu_1 \neq \mu_2$)

Since population S.Ds are given, the test statistic applicable here is Z-test.
Z = Difference / S E
Difference = $\bar{x}_1 - \bar{x}_2 = 210 - 200 = 10$
S E = $\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}$ (Population S.Ds are known. For the entire State SD is 11).
= $\sqrt{(11^2 / 100) + (11^2 / 150)}$ = $\sqrt{1.21+0.81}$ = $\sqrt{2.02}$ = 1.42
Z = $10/1.42 = 7.04$

Level of significance = 5%

Degree of freedom = infinity

Table value of Z at 5% level of significance and infinity degrees of freedom = 1.96

Since the calculated value of Z is more than the table value, H0 is rejected. We accept H1. So we may conclude that there is significant difference in the mean yields of crops in two districts.

Qn: Electric bulbs manufactured by X Ltd. and Y Ltd. gave the following results:

Particulars	X Ltd	Y Ltd
Number of bulbs used	100	100
Mean Life in Hours	1300	1248
Standard Deviation	82	93

State whether there is any significant difference in the life of bulbs of the two makes.

Sol: H_0 : There is no significant difference between two sample means(i.e; $\mu_1 = \mu_2$)
H_1 : is no significant difference between two sample means(i.e; $\mu_1 \neq \mu_2$)

Since population S.Ds are unknown but samples are large, the test statistic applicable here is Z-test.

Z = Difference / S E

Difference = $\bar{x}_1 - \bar{x}_2 = 1300 - 1248 = 52$

S E = $\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}$ (Population S.Ds are unknown)

= $\sqrt{(82^2 / 100) + (93^2 / 100)}$ = $\sqrt{67.24+86.49}$ = $\sqrt{153.73}$ = 12.4

Z = 52/12.4 = 4.19

Level of significance = 5%

Degree of freedom = infinity

Table value of Z at 5% level of significance and infinity degrees of freedom = 1.96

Since the calculated value of Z is more than the table value, H_0 is rejected. We accept H_1. (i.e; $\mu_1 \neq \mu 2$). So we may conclude that there is significant difference in the mean life of bulbs of the two makes.

Qn: Two batches of same product are tested for their mean life. Assuming that lives of the two products follow a normal distribution, test the hypothesis that the mean life is same for both the batches, given the following information:

Batch	Sample Size	Mean life in hours	S.D
A	10	750	12
B	8	820	14

Sol: H_0 : There is no significant difference between two sample means(i.e; $\mu_1 = \mu_2$)
H_1 : is no significant difference between two sample means(i.e; $\mu_1 \neq \mu_2$)

Since population S.Ds are unknown and samples are small, the test statistic applicable here is t-test.

t = Difference / S E

Difference = $\bar{x}_1 - \bar{x}_2 = 820 - 750 = 70$

S E = $\sqrt{(n_1 s_1^2 + n_2 s_2^2)/n_1 + n_2 - 2 \text{ x } (1/n_1 + 1/n_2)}$ (Population S.Ds are unknown and samples are small)

= $\sqrt{(10*12^2) + (8*14^2) / 10+8-2) \text{ x } (1/10 + 1/8)}$

= $\sqrt{3008/16 \text{ x } 0.225}$ = $\sqrt{42.3}$ = 6.5

t = 70/6.5 = 10.77 (Calculated Value)

Level of significance = 5%

Degree of freedom = 10+8-2 = 16

Table value of t at 5% level of significance and 16 degrees of freedom = 2.12

Since the calculated value of t is more than the table value, H_0 is rejected. We accept H_1 (i.e; $\mu_1 \neq \mu 2$). So we may conclude that the lives of products produced in two batches are not same.

Qn: In a test given to 2 groups of students, the marks obtained were as follows:

Group I	18	20	36	50	49	36	34	49	41
Group II	29	26	28	35	30	44	46		

Test whether the group means are equal.

Sol: Here we have to find the Means and S.Ds of the two samples**.**

Computation of Mean and S D of two Groups					
Group I			Group II		
X	$X -- \bar{X}$	$(X -- \bar{X})^2$	X	$X -- \bar{X}$	$(X -- \bar{X})^2$
18	-19	361	29	-5	25
20	-17	289	26	-8	64
36	-1	1	28	-6	36
50	13	169	35	1	1
49	12	144	30	-4	16
36	-1	1	44	10	100
34	-3	9	46	12	144
49	12	144			
41	4	16			
$\varepsilon X = 333$		1134	$\varepsilon X = 238$		386

Mean of Group I = 333/9 = 37 Mean of Group II = 238/7 = 34

S.D of Group I = $\sqrt{\varepsilon (X -- \bar{X})^2 / n}$ S.D of Group I = $\sqrt{\varepsilon (X -- \bar{X})^2 / n}$

$= \sqrt{1134/9} = \sqrt{126}$ $= \sqrt{386/7} = \sqrt{55.14}$

H_0 : There is no significant difference between two sample means(i.e; $\mu_1 = \mu_2$)
H_1 : is no significant difference between two sample means(i.e; $\mu_1 \neq \mu_2$)

Since population S.Ds are unknown and samples are small, the test statistic applicable here is t-test.

t = Difference / S E

Difference = $\bar{x}_1 - \bar{x}_2 = 37 - 34 = 3$

S E $= \sqrt{(n_1 s_1^2 + n_2 s_2^2)/n_1 + n_2 - 2} \times (1/n_1 + 1/n_2)$ (Population S.Ds are unknown and samples are small)

$= \sqrt{(9*126) + (7*55.14) / 9+7-2} \times (1/9 + 1/7)$

$= \sqrt{1510.98/14} \times 0.254 = \sqrt{27.41} = 5.24$

$t = 3/5.24 = 0.573$ (Calculated Value)

Level of significance = 5%

Degree of freedom = 9+7-2 = 14

Table value of t at 5% level of significance and 14 degrees of freedom = 2.145

Since the calculated value of t is less than the table value, H_0 is accepted. (i.e; $\mu_1 = \mu 2$). So we may conclude that the difference in the group means are not significant. They are equal.

TESTING OF SIGNIFICANCE OF THE DIFFERENCE IN CASE OF DEPENDENT SMPLES (PAIRED OBSERVATIONS)

Here the observations in one sample are some way related to the observations in the other.

Therefore they are called paired observations. The test statistic applicable here is t-test.

Procedure:

1. Set up H_0 and H_1
   H_0 : There is no significant difference between samples
   H_1 : There is significant difference between samples
2. Decide test statistic:
   Since the paired data are comparatively less, the test statistic applicable here is always t-test.
3. Apply the appropriate formula for computing the value of the test statistic.

$$t = \overline{d}/SE$$

Where:

$\overline{d}$	Arithmetic mean of the difference between the values
SE	$S/\sqrt{n-1}$ [s = standard deviation of the difference]

4. Specify the level of significance. Take 5%, if nothing is mentioned in the question.
5. Fix the degree of freedom. d.f = n – 1 , where n= Number of pairs of observations.
6. Locate the critical value of the test statistic (t-test) at specified level of significance and fixed degree of freedom.
7. Compare the calculated value of test statistic with the table value and decide whether to accept or reject the null hypothesis. If calculated value of the test statistic is numerically less than the table value, the null hypothesis is accepted. If calculated value of the test statistic is numerically more than the table value, the null hypothesis is rejected.

Qn: The marks scored by 10 students, before and after providing special coaching, are given in the following table:

Before	67	24	57	55	63	54	56	68	33	43
After	70	38	58	58	56	67	68	72	42	38

Test whether there is any significant difference in their performance.

Sol: H_0 : There is no significant difference between samples

H_1 : There is significant difference between samples

Test statistic applicable here is t-test.

$$t = \bar{d}/SE$$

Computation of men and standard deviation of the difference between the values			
Score (Before)	Score (After)	Difference (d)	d^2
67	70	3	9
24	38	14	196
57	58	1	1
55	58	3	9
63	56	-7	49
54	67	13	169
56	68	12	144
68	72	4	16
33	42	9	81
43	38	-5	25
		$\varepsilon d = 47$	$\varepsilon d^2 = 699$

Arithmetic mean of d values = 47/10 = 4.7

S D of d values $= \sqrt{\varepsilon d^2/n - (\varepsilon d/n)^2} = \sqrt{699/10 - (47/10)^2} = \sqrt{47.81} = 6.91$

S E $= 6.91/\sqrt{10-1} = 6.91/3 = 2.3$

t $= 4.7/2.3 = 2.04$

Level of significance = 5%

Degree of freedom = 10—1 = 9

Table value (critical value) of t at 5% level of significance and 9 degree of freedom is 2.262.

Since the calculated value of t is less than the critical value, the null hypothesis is accepted. So, we may conclude that there is no significant difference in the performance of the students.

TESTING OF GIVEN POPULATION PROPORTION

This type of testing of hypothesis is used to test whether there is any significant difference between the sample proportion and the given population proportion.

Procedure:

1. Set up H_0 and H_1 :

   H_0 : There is no significant difference between sample proportion and population proportion (i.e; $H_0 : P = P_0$)

   H_1 : There is significant difference between sample proportion and population proportion (i.e; $H_0 : P \neq P_0$)
2. Decide the test statistic:

   The test statistic applicable here is Z-test
3. Apply appropriate formulae for computing the value of Z (i.e; calculated value):

   Z = Difference / S E ie; $Z = (p - P) / S E$

   Where p = sample proportion, P = Population proportion

   $S E = \sqrt{PQ / n}$
4. Decide the level of significance (Take 5%, if nothing is mentioned in the question).
5. Fix the degree of freedom (Infinity d.f)
6. Locate the table value of Z at specified level of significance and fixed degree of freedom.
7. Compare the calculated value of Z with the table value and decide whether to accept or reject the null hypothesis. If calculated value of Z is numerically less than the table value, the null hypothesis is accepted. If calculated value of Z is numerically more than the table value, the null hypothesis is rejected.

Qn: It is found that out of 500 units of a product produced by a machine, 30 are defectives. Test whether the machine produces 2% defective items on an average.

Sol:

H_0 : There is no significant difference between sample proportion and population proportion (i.e; $H_0 : P = 0.02$)

H_1 : There is significant difference between sample proportion and population proportion (i.e; $H_0 : P \neq 0.02$)

$Z = (p - P) / S E$

$P = 0.02$, $p = 30/500 = 0.06$, $Q = 1 - Q = 1 - 0.02 = 0.98$, $n = 500$

$S E = \sqrt{PQ / n} = \sqrt{0.02 \times 0.98 / 500} = \sqrt{0.0196/500} = \sqrt{0.0000392} = 0.0063$

$\therefore Z = (0.06 - 0.02) / 0.0063 = 0.04/0.0063 = \underline{6.349}$

Level of significance = 5%

Degree of freedom = infinity

Table value of Z at 5% level of significance and infinity degree of freedom is 1.96

Since the calculated value of Z is more than the table value, null hypothesis is rejected. We accept alternative hypothesis. $P \neq 0.02$. So, it is not possible to think that the machine produces 2% defective items.

TESTING OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

This testing of hypothesis is used to test whether the difference between two sample proportions are significant or not. If the difference is not significant, they are treated as equal; or we may think that the two samples are drawn from the same population.

Procedure:

1. Set up H_0 and H_1

   H_0 : There is no significant difference between two sample proportions(i.e; $p_1 = p_2$)

   H_1 : is no significant difference between two sample proportions(i.e; $p_1 \neq p_2$)
2. Decide the test statistic:

   The test statistic applicable here is Z-test.
3. Apply the appropriate formula for computing the value of the test statistic:

   Z = Difference/Standard Error

   i e; Z = $p_1 - p_2$ / SE

   where p_1 and p_2 are the proportions of two samples

   S E = $\sqrt{p_0 q_0 [(1/n_1) + (1/n_2)]}$, where $p_0 = (n_1p_1 + n_2p_2)/(n_1 + n_2)$, $q_0 = 1 - p_0$

4. Specify the level of significance. If nothing is mentioned about the level of significance, take 5%.
5. Fix the degree of freedom (d f = infinity)
6. Locate the table value (critical value) of the test statistic at specified level of significance and fixed degree of freedom.
7. Compare the calculated value of test statistic with the table value and decide whether to accept or reject the null hypothesis. If calculated value of the test statistic is numerically less than the table value, the null hypothesis is accepted. If calculated value of the test statistic is numerically more than the table value, the null hypothesis is rejected.

Qn: In a sample of 1000 people selected from District X, 450 were regular drinkers of coffee. In another sample of 800 people drawn from District Y, 400 were regular drinkers of coffee. Test whether there is significant difference between the two districts, regarding the coffee drinking habit of people.

Sol:

H_0 : There is no significant difference between two districts regarding the coffee drinking habits of people (i.e; $p_1 = p_2$)

H_1 : is There is no significant difference between two districts regarding the coffee drinking habits of people (i.e; $p_1 \neq p_2$)

The test statistic applicable here is Z-test.

Z = Difference/Standard Error

i e; Z = $p_1 - p_2$ / SE

$p_1 = 450/1000 = 0.45$, $p_2 = 400/800 = 0.5$

S E = $\sqrt{p_0 q_0 [(1/n_1) + (1/n_2)]}$, $n_1 = 1000$, $n_2 = 800$

$p_0 = (1000 \times 0.45 + 800 \times 0.5)/(1000 + 800) = 850/1800 = 0.472$

$q_0 = 1 - 0.472 = 0.528$

$\therefore SE = \sqrt{0.472 \times 0.528 [(1/1000) + (1/800)]} = \sqrt{0.249 (0.001 + 0.00125)}$

$= \sqrt{0.249 \times 0.00225} = \sqrt{0.00056} = 0.0237$

$\therefore Z = (0.5 - 0.45) / 0.0237 = 0.05/0.0237 = 2.11$

Level of significance = 5%.

Fix the degree of freedom = infinity

Table value (critical value) of Z at 5% level of significance and infinity fixed degree of freedom is 1.96

Since the calculated value of Z is more than the table value, null hypothesis is rejected. Alternative hypothesis is accepted. $p_1 \neq p_2$. So, we may conclude that there is significant difference between the two districts regarding the coffee drinking habits of people.

REVIEW QUESTIONS:

1. What do you mean by inferential analysis?
2. What do you understand by sampling distributions?
3. What are the two branches of inferential analysis?
4. What do you mean by hypothesis?
5. What is the difference between parameter and statistic?
6. What do you mean by Standard Error? What are its uses?
7. What are the differences between standard deviation and standard error?
8. What do you mean by parametric tests?
9. What do you mean by non-parametric tests?
10. What is type I error?
11. What is Type II error?
12. What do you mean by power of a test
13. What is meant by critical region and acceptance region?
14. What is one tailed test?
15. What is two-tailed test?
16. What do you mean by dependent sample?
17. Explain the general procedure for testing of hypothesis.
18. A random sample of 10 bottles of filled in by an automatic machine gave the following weights in kilogram:

    2.05, 2.01, 2.04, 1.96, 2.01, 1.98, 1.99, 1.98, 2.04, 2.02

    Can we accept at 5% level of significance, the claim that the average weight of the tin is 2 Kg.
19. From the following data, test whether there is significant difference between two samples:

Sample I	25	32	30	32	24	14	32			
Sample II	24	34	30	22	42	31	40	35	32	30

20. In a sample of 600 people in Bihar 336 are coffee drinkers and the rest are tea drinkers. Can we assume that both coffee and tea are equally popular in the State at 1% level of significance?

21. In a sample of 900 men from a certain large city 675 were found to be smokers. In a random sample of 1350 men from another large city 675 were found to be smokers. Do the data indicate that the cities are significantly different in respect of the prevalence of smoking among men?

22. A sample of size 50 has S.D of 10.5. Can you contradict the hypothesis that the population S.D. is 12?

CHAPTER 11

CHI-SQUARE TEST

What is Chi-Square Value?

The word "Chi-square" is denoted by the symbol, $\chi 2$. Chi-square is a value (quantity) which describes the magnitude of the difference between observed frequencies and expected frequencies.

Chi-Square Test

Chi-square test is a statistical test used to test the significance of the difference between observed frequencies and the corresponding theoretical frequencies (expected frequencies) of a distribution, without any assumption about the nature of distribution of the population. This is the most popular widely used non-parametric test. It was developed by Prof. Karl Pearson.

Uses of Chi-Square Test (Applications of Chi-Square Test)

Chi-Square test is mainly used for the following purposes:

1. **Used to test goodness of fit:** As a test for goodness of fit, $\chi 2$ test can be used to test how far the theoretical frequencies fit to the observed frequencies.
2. **Used to test independence:** As a test of independence, $\chi 2$ test is used to test whether the attributes of a sample are associated or not.
3. **Used to test homogeneity:** As a test of homogeneity, $\chi 2$ test is used to test whether different samples are homogeneous as far as a particular attribute is concerned.
4. **Used to test population variance:** Here, Chi-square test is used for testing the given population variance when the sample is small. In other words, it used to test whether there is any significant difference between sample variance and population variance. Here, the test statistic value (Chi-square value) is obtained by using the following formulae (ns^2/σ^2).

Conditions for applying Chi-square Test

1. The total frequencies(N) must be at least 50
2. Expected frequencies of less than 5 must be pooled with the preceding or succeeding frequency so that the expected frequency is 5 or more.
3. The distributions should be of original units. They should not be of proportions or percentages.

Testing of Goodness of Fit

Procedure:

1. Set up H_0 and H_1

   H_0 : There is goodness of fit between observed frequencies and expected frequencies.

   H_0 : There is no goodness of fit between observed frequencies and expected

frequencies.

2. Decide the test statistic. Here, the test statistic is Chi-Square test.
3. Apply the appropriate formula:

$$\chi 2 = \varepsilon [(O—E)^2/E]$$

where o = Observed frequencies and E = Expected frequencies

4. Specify the level of significance. If nothing is mentioned, take 5% level of significance.
5. Fix the degree of freedom. Degree of freedom = n – r – 1

Where n = number of pairs of observations

r = number of parameters computed from the given data to find the expected frequencies.

6. Obtain the table value of Chi-square at specified level of significance and fixed degree of freedom.
7. Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Qn: The numbers of road accidents per week in a certain city were as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that the accidents occurred were the same during the 10 week period?

Sol:

H_0 : There is goodness of fit between observed frequencies and expected frequencies.

H_0 : There is no goodness of fit between observed frequencies and expected frequencies.

$$\chi 2 = \varepsilon [(O—E)^2/E]$$

Here the Observed values (Actual values) are 12, 8, 20, 2, 14, 10, 15, 6, 9 and 4.

i.e; O = 12, 8, 20, 2, 14, 10, 15, 6, 9, 4

If accidents occurred are same, then the number of accidents per week which we may expect is 10 (i.e; the average of the given values).

i.e; E = 10

Now we can find the value of Chi-square as follows:

Computation of Chi-square Value			
Observed Values (O)	Expected Values (E)	$(O – E)^2$	$(O – E)^2 / E$
12	10	4	0.4
8	10	4	0.4
20	10	100	10.0

2	10	64	6.4
14	10	16	1.6
10	10	0	0.0
15	10	25	2.5
6	10	16	1.6
9	10	1	0.1
4	10	36	3.6
			$\chi 2 = 26.6$

Calculated Value of $\chi 2 = 26.6$

Level of significance = 5%

Degree of Freedom = n -- r -- = 10 – 0 -- 1 = 9

Table value of $\chi 2$ at 5% level of significance and 9 d.f is 16.919.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternative hypothesis. So we may conclude that the given figures do not agree with the belief that accident occurred were same during the 10 weeks period.

Qn: The principal of a college made a sample analysis of an examination result of 200 students. It was found that 24 students had got first class, 62 second class, 68 third class and the rest were failed. Are these figures commensurate with the general examination result which is in the ratio of 2:3:3:2 for various categories respectively.

Sol:

H_0 : There is goodness of fit between the given figures and the figures expected in general examination

H_0 : There is no goodness of fit between the given figures and the figures expected in general examination

$\chi 2 = \varepsilon [(O—E)^2/E]$

Here the Observed values (Actual values) for first, second, third and failed categoris of students are respectively 24, 62, 68 and 46.

If results are in the ratio of 2:3:3:2, then the number of students for above categories may be expected as follows:

First Class	200 x 2/10	40
Second Class	200 x 3/10	60
Third Class	200 x 3/10	60
Failed	200 x 2/10	40
Total		200

So the E Values are 40, 60, 60 and 40.

Now we can find the value of Chi-square as follows:

Computation of Chi-square Value			
Observed Values (O)	Expected Values (E)	$(O - E)^2$	$(O - E)^2 / E$
24	40	36	0.900
62	60	64	1.067
68	60	4	0.067
46	40	256	6.400
			$\chi 2 = 8.434$

Calculated Value of $\chi 2 = 8.434$

Level of significance = 5%

Degree of Freedom = n -- r -- = 4 – 0 -- 1 = 3

Table value of $\chi 2$ at 5% level of significance and 9 d.f is 7.815.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternative hypothesis. So we may conclude that the given figures do not commensurate with the general examination result which is in the ratio of 2:3:3:2.

Testing of Independence

Procedure:

1. Set up H_0 and H_1
   H_0 : There is independence between observed frequencies and expected frequencies.
   H_0 : There is no independence between observed and expected frequencies.
2. Decide the test statistic. Here, the test statistic is Chi-Square test.
3. Apply the appropriate formula:
   $\chi 2 = \varepsilon [(O—E)^2/E]$
   where o = Observed frequencies and E = Expected frequencies
   Here E values are obtained by using the following formula:
   E Value = $[(\text{Row Total x Column Total})/\text{Grand Total}]$
   E Values are computed by preparing a table called Contingency Table.
4. Specify the level of significance. If nothing is mentioned, take 5% level of significance.
5. Fix the degree of freedom. Degree of freedom = (r – 1) x (c – 1)
   Where r = number of rows; c = number of columns
6. Obtain the table value of Chi-square at specified level of significance and fixed degree of freedom.
7. Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Qn: From the following data, can you say that there is relation between the habit of smoking and literacy:

	Smokers	Non-smokes
Literates	83	57
Illiterates	45	68

Sol:

H_0 : There is independence between smoking habit and literacy.

H_0 : There is no independence between smoking habit and literacy.

$\chi 2 = \varepsilon [(O—E)^2/E]$

Here the Observed values (Actual values) are 83, 57, 45 and 68.

The E Values corresponding to the above 'O' values can be found out by preparing a 2 X 2 contingency table:

2 X 2 Contingency Table			
	Smokers	Non-smokes	Total
Literates	[(83+57) x (83+45)] /253 = 71	(140 x 125) / 253 = 69	140
Illiterates	(113 x 128) /253 = 57	(113 x 125) / 253 = 56	113
Total	128	125	253

So, the E values are 71, 69, 57 and 56.

Computation of Chi-square Value			
Observed Values (O)	Expected Values (E)	$(O – E)^2$	$(O – E)^2 / E$
83	71	144	2.03
57	69	144	2.09
45	57	144	2.53
68	56	144	2.57
			$\chi 2 = 9.22$

Calculated Value of $\chi 2 = 9.22$

Level of significance = 5%

Degree of Freedom = (2 – 1) x (2 – 1) = 1 x 1 = 1

Table value of $\chi 2$ at 5% level of significance and 1 d.f is 3.841.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternative hypothesis. So we may conclude that there is no independence between smoking habit and literacy. In other words, smoking habit and literacy are related.

Qn: In a sample study about the tea drinking habit in a town, following data are observed in a sample of size 200.

46%nwerenmale, 26% were tea drinkers and 17% were male tea drinkers.

Is there any association between gender and tea habits?

Sol:

H_0 : There is independence between gender and tea drinking habits.

H_0 : There is no independence between gender and tea drinking habits.

$\chi 2 = \varepsilon [(O—E)^2/E]$

Here all the Observed values (Actual values) are not directly given in the question. So, we have to find the missing figures with the help of a 2 x 2 contingency table:

2 X 2 Contingency Table ('O' values)			
	Tea drinkers	Non-tea drinkers	Total
Male	(200 x 17) / 100 = 34	58	(200 x 46)/100 = 92
Female	= 18	90	= 108
Total	(200 x 26) / 100 = 52	148	200

"O" values are 34, 58, 18 and 90.

The E Values corresponding to the above 'O' values can be found out by preparing a 2 X 2 contingency table:

2 X 2 Contingency Table ('E' values)			
	Tea drinkers	Non-tea drinkers	Total
Male	(92 x 52) / 200 = 24	(92 x 148) / 200 = 68	92
Female	(108 x 52)/200 = 28	(108 x 148)/200 = 80	108
Total	52	148	200

So, the 'E' values are 24, 68, 28 and 80.

Computation of Chi-square Value			
Observed Values (O)	Expected Values (E)	$(O – E)^2$	$(O – E)^2 / E$
34	24	100	4.17
58	68	100	1.47
18	28	100	3.57
90	80	100	1.25
			$\chi 2 = 10.46$

Calculated Value of $\chi 2 = 10.46$

Level of significance = 5%

Degree of Freedom = (2 – 1) x (2 – 1) = 1 x 1 = 1

Table value of $\chi 2$ at 5% level of significance and 1 d.f is 3.841.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternative hypothesis. So we may conclude that there is no independence between gender and smoking habit. In other words, gender and smoking habit are closely associated.

Testing of Homogeneity

Procedure:

1. Set up H_0 and H_1
   H_0 : There is homogeneity between the samples on the basis of the attribute.
   H_0 : There is no homogeneity between the samples on the basis of the attribute.
2. Decide the test statistic. Here, the test statistic is Chi-Square test.
3. Apply the appropriate formula:
   $\chi 2 = \varepsilon [(O—E)^2/E]$
   where o = Observed frequencies and E = Expected frequencies
   Here 'E' values are obtained by using the following formula:
   'E' Value = [(Row Total x Column Total)/Grand Total]
   'E' Values are computed by preparing a table called Contingency Table.
4. Specify the level of significance. If nothing is mentioned, take 5% level of significance.
5. Fix the degree of freedom. Degree of freedom = (r – 1) x (c – 1)
   Where r = number of rows; c = number of columns
6. Obtain the table value of Chi-square at specified level of significance and fixed degree of freedom.
7. Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Qn: In a diet survey the following results were obtained:

	Hindus	Muslim
No. of families drinking tea	124	16
No. families not drinking tea	56	10

Is there any difference between the communities in the matter of tea drinking?

Sol:

H_0 : There is homogeneity between communities in the matter of tea drinking.

H_0 : There is no homogeneity between communities in the matter of tea drinking.

$\chi 2 = \varepsilon [(O—E)^2/E]$

Here the Observed values (Actual values) are 124, 16, 56, and 10

The 'E' values corresponding to the above 'O' values can be found out by preparing a 2 X 2 contingency table:

2 X 2 Contingency Table			
	Smokers	Non-smokes	Total
No. of families drinking tea	(140 x 180) /206 = 122	(140 x 26)/206 = 18	140
No. of families not drinking tea	(66 x 180) /206 = 58	(66 x 26) / 206 = 8	66
Total	180	26	206

So, the 'E' values are 122, 18, 58 and 8.

Computation of Chi-square Value			
Observed Values (O)	Expected Values (E)	$(O - E)^2$	$(O - E)^2 / E$
124	122	4	0.033
16	18	4	0.222
56	58	4	0.069
10	8	4	0.500
			$\chi 2 = 0.824$

Calculated Value of $\chi 2 = 0.824$

Level of significance = 5%

Degree of Freedom = (2 – 1) x (2 – 1) = 1 x 1 = 1

Table value of $\chi 2$ at 5% level of significance and 1 degree of freedom is 3.841.

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is homogeneity between communities in the matter of tea drinking.

Testing of Variance

Procedure:

1. Set up H_0 and H_1

   H_0 : There is no significant difference between sample variance and population variance.

   H_1 : There is significant difference between sample variance and population variance.
2. Decide the test statistic. Here, the test applicable is Chi-square test.
3. Apply the appropriate formula for computing the value of test statistic.

   $\chi 2 = ns^2/\sigma^2$, where n = sample size, s^2 = sample variance, σ^2 = population variance.
4. Specify the level of significance. Take 5%, unless specified otherwise.
5. Fix the degree of freedom. d.f = n –1.
6. Locate the table value of Chi-square at specified level of significance and fixed degree of freedom.

7. Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Qn: A sample is drawn from a population which follows normal distribution. The size of sample and S.D are respectively 10 and 5. Test whether this is consistent with the hypothesis that the S D of the population is 5.3

Sol:

H_0 : There is no significant difference between sample S.D and population S.D. (i.e; H_0 : S.D of population = 5.3)

H_1 : There is significant difference between sample S.D and population S.D. (i.e; H_1 : S.D of population $\neq$ 5.3)

The test applicable is Chi-square test.

$\chi 2 = ns^2/\sigma^2$, where n = 10, $s^2 = 5^2$, $\sigma^2 = 5.3^2$

$= (10 \times 5^2)/5.3^2 = 250/28.09 = 8.899$

Specify the level of significance. Take 5%, unless specified otherwise.

Fix the degree of freedom. d.f = 10 –1 = 9

Table value of Chi-square at 5% level of significance and 9 degree of freedom is 16.919

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference between sample S.D and population S.D. The population S.D = 5.3

Qn: A sample group of 10 students are selected randomly from a class. Their weights (in K.g) are 49, 40, 53, 38, 52, 47, 48, 45, 55, and 43. Can we say that the population variance is 20 Kg?

Sol:

H_0 : There is no significant difference between sample variance and population variance. (i.e; H_0 : Variance of population = 20)

H_0 : There is significant difference between sample variance and population variance. (i.e; H_0 : Variance of population $\neq$ 20)

The test applicable is Chi-square test.

$\chi 2 = ns^2/\sigma^2$

Here, n = 10, $\sigma^2 = 20$, Sample variance is to be computed from the given data.

Computation of sample variance (s^2)		
Weight (X)	$(X -- \bar{x})$	$(X -- \bar{x})^2$
49	2	4
40	-7	49
53	6	36
38	-9	81
52	5	25
47	0	0
48	1	1
45	-2	4
55	8	64
43	-4	16
$\varepsilon X = 470$		$\varepsilon (X -- \bar{x})^2 = 280$

$\bar{x} = 470/10 = 47.$

Sample Variance $(s^2) = [\varepsilon (X -- \bar{x})^2]/n = 280/10 = 28.$

$\chi 2 = (10 \times 28)/20 = 280/20 = 14.$

Level of significance = 5%

Degree of freedom. (d.f) = 10 –1 = 9

Table value of Chi-square at 5% level of significance and 9 degree of freedom is 16.919

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference between sample variance and population variance. ∴ The population variance = 20 Kg.

REVIEW QUESTIONS:

1. What do you mean by Chi-square value?
2. What are the important uses of Chi-square test?
3. What do you mean by goodness of fit?
4. Explain the procedure for testing goodness of fit.
5. What do you mean by contingency table?
6. What are the important conditions for applying Chi-square test?
7. A die is thrown 150 times with the following results:

No. turned up	1	2	3	4	5	6
Frequency	19	23	28	17	32	31

Test the hypothesis that the die is unbiased.

8. Following data are given:

Gender	Education			
	Middle	High School	College	Total
Male	52	10	20	82
Female	44	12	26	82
Total	96	22	46	164

Can you say that education depends on gender?

Chapter 12

ANALYSIS OF VARIANCE

Meaning of Analysis of Variance

The testing of hypotheses so far discussed consists of different sample groups which do not exceed two. If there are three or more sample groups, the testing of equality of them cannot be done in any of the methods which have already been discussed. The testing of significance of the difference among three or more samples is generally done by using the technique of analysis of variance. In case of analysis of variance, as part of testing procedure, we have to prepare a separate statement called Analysis of Variance Table or ANOVA Table. Therefore, this type of testing of hypothesis is also called analysis of variance. The test statistic used for Analysis of Variance is F-test. F-test is a parametric test.

Types of Analysis of Variance

There are two types of Analysis of variance. They are:

1. One-way classification of data (One way analysis of variance)
2. Two-way classification of data (Two way analysis of variance)

One-way classification of data (One way analysis of variance)

In one way classification, observations are classified into different groups on the basis of a single criterion. Suppose we want to study about the yield of a particular crop. You know there are number of factors which influence the productivity of crops. If we undertake this study to know the effect of quality of seed on the yield of crop, it is called one- way analysis of variance. Here yield of crops based on different seed must be given in columns. In other words, in case of one way analysis of variance, the samples must always be in columns.

Types of variances in One-way ANOVA

1. **Variance between samples (Columns):** This is the net result of the variation different sample means from grand mean. Grand mean is the mean of all the observations coming under all sample groups.
2. **Variance within the sample:** This is the net result of variations different items of the sample from the respective sample means.
3. **Variance about the sample:** This is the sum of the variance between samples and the variance within the sample.

Proforma of One-way ANOVA Table

One-way ANOVA Table				
Source of variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-Ratio
Between Samples	SSC =	C – 1	MSC = SSC/(C–1)	[F= Larger variance ÷ Smaller variance]
Within Sample	SSE =	N – C	MSE = SSE/(N–C)	F = MSC ÷ MSE, or F = MSE ÷ MSC
Total	SST =	N –1		

SSC = Sum of Squares between Columns (Samples)

SSE = Sum of Square within Column (Sample)
SST = Sum of Square Total
MSC = Mean Sum of Squares between Columns (Samples)
MSE = Mean Sum of Squares within Column (Sample)
C = Number of Columns (Samples)

Procedure for carrying out One-way Analysis of variance

1. Set up H_0 and H_1.
   H_0 : There is no significant difference between samples
   H_1 : There is significant difference between samples
2. Decide the test statistic:

   Test statistic applicable here is F-test
3. Apply the appropriate formula for computing the value of F-test.

   F = Larger Variance ÷ Smaller Variance, i.e; MSC÷MSE or MSE÷MSC
   (i) Find SST.

   SST = Sum of square of all items – (T^2/N)

   Where T = Total of all observations, N = Total Number of observations

   (T^2/N) is generally called correction factor
   (ii) Find SSC.

   SSC = $[(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots - (T^2/N)$

   Where εx_1= sum of items in the first column

   εx_2= sum of items in the second column
   n_1 = number of items in the first column
   n_2 = number of items in the second column

   (iii) Draw one-way ANOVA Table and enter the values of SST and SSC
   (iv) Find the value of SSE. SSE = SST – SSC
   (v) Find the degree of freedom in the third column as indicated in the proforma.
   (vi) Find MSC. MSC = SSC ÷ (C–1)
   (vii) Find MSE. MSE = SSE ÷ (N–C)
   (viii) Find F–Ratio.
   F= Larger variance ÷ Smaller variance; [i.e; F= MSC÷MSE, or F= MSE÷MSC]
4. Specify the level of significance. Take 5% if nothing is mentioned.
5. Fix the degrees of freedom. Here we have to fix a pair of d.f.
   If 'F' is obtained by using F= MSC÷MSE, then pair of df is (d f of MSC, d f of MSE)

   If 'F' is obtained by using F= MSE÷MSC, then pair of df is (d f of MSE, d f of MSC)
6. Obtain table value of F at specified level significance and fixed degree of freedom.
7. Compare the Calculated value of F with the Table value, and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, H_0 is accepted. If calculated value is more than the table value, H_0 is rejected.

Qn: Four varieties of a crop was grown on 3 plots, and the following yield was obtained. You are required to test whether there is significant difference in the productivity of seeds:

Plot	Variety of Seeds			
	P	Q	R	S
I	10	7	8	5
II	9	7	5	4
III	8	6	4	4

Sol:

H_0 : There is no significant difference in the productivity of seeds.

H_1 : There is significant difference in the productivity of seeds.

Test Statistic applicable here is F-test

F = Larger Variance ÷ Smaller Variance, i.e; MSC÷MSE or MSE÷MSC

$SST = 10^2 + 7^2 + 8^2 + 5^2 + 9^2 + 7^2 + 5^2 + 4^2 + 8^2 + 6^2 + 4^2 + 4^2 - (T^2/N)$

$= 100+49+64+25+81+49+25+16+64+36+16+16 - (77^2/12)$

$= 541 - (5929/12) = 541 - 494 = \underline{47}$

$SSC = [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots - (T^2/N)$

$= [(10+9+8)^2 \div 3] + [(7+7+6)^2 \div 3] + [(8+5+4)^2 \div 3] + [(5+4+4)^2 \div 3] - (T^2/N)$

$= (729/3) + (729/3) + (729/3) + (729/3) - (77^2/12)$

$= (1587/3) - 494 = 529 - 494 = \underline{35}$

One-way ANOVA Table				
Source of variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-Ratio
Between Samples	SSC = 35	C – 1= 4 –1 = 3	MSC = SSC/(C–1) 35/3 = 11.67	[F= Larger variance ÷ Smaller variance]
Within Sample	SSE = 12	N–C = 12 – 4 = 8	MSE = SSE/(N–C) 12/8 = 1.5	F = MSC ÷ MSE, or = 11.67/1.5= 7.78
Total	SST = 47	N –1= 12-1=11		

Level of significance = 5%

Degree of freedom = (3,8)

Table value of F at 5% level of significance and (3,8) degrees of freedom is 4.07

Since calculated value of F is more than the table value, H_0 is rejected. We accept alternative hypothesis. So we may conclude that there is significant difference in the productivity of three varieties of seeds.

Qn: The following table shows the yield of 3 varieties. Perform analysis of variance and test whether there is significant difference between varieties:

Varieties	Plots				
	A	B	C	D	E
I	30	27	42		
II	51	47	37	48	42
III	44	35	41	36	

Sol:

Here, we are asked to test whether there is significant difference between varieties. But varieties are given in rows, not in columns. In one way ANOVA, the samples must be in columns. Therefore, we have to rearrange the given data so as to bring the samples in columns as shown below:

Plots	Varieties		
	I	II	III
A	30	51	44
B	27	47	35
C	42	37	41
D		48	36
E		42	

H_0 : There is no significant difference in the productivity of varieties.

H_1 : There is significant difference in the productivity of varieties.

Test Statistic applicable here is F-test

F = Larger Variance ÷ Smaller Variance, i.e; MSC÷MSE or MSE÷MSC

$SST = 30^2 + 51^2 + 44^2 + 27^2 + 47^2 + 35^2 + 42^2 + 37^2 + 41^2 + 48^2 + 36^2 + 42^2 - (T^2/N)$

$= 900+2601+1936+729+2209+1225+1764+1369+1681+2304+1296+1764 - (480^2/12)$

$= 19778 - 19200 = \underline{578}$

$SSC = [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots - (T^2/N)$

$= [(30+27+42)^2 \div 3]+[(51+47+37+48+42)^2 \div 5]+[(44+35+41+36)^2 \div 4] - (480^2/12)$

$= (9801/3) + (50625/5) + (24336/4) - (19200)$

$= 3267 + 10125 + 6084 - 19200 = 19476 - 19200 = \underline{276}$

One-way ANOVA Table				
Source of variation	Sum of Squares	Degree of freedom	Mean Sum of Squares	F-Ratio
Between Samples	SSC= 276	3-1=2	MSC=276/2=138	F=138/33.56 = 4.112
Within Sample	SSE=302	12-3=9	MSE=302/9=33.56	
Total	SST=578	12-1=11		

Level of significance = 5%

Degree of freedom = (2,9)

Table value of F at 5% level of significance and (2,9) degrees of freedom is 4.26

Since calculated value of F is more than the table value, H_0 is accepted. <u>So we may conclude that there is no significant difference in the productivity of three varieties.</u>

Two-way classification of data (Two way analysis of variance)

In two way classification, observations are classified into different groups on the basis of two criteria. Consider the example mentioned in one-way classification. If we study the effect of both the quality of seeds and the type of fertilizers on the productivity of crop, the data are to be classified on the basis of two criteria, namely type of seed and type of fertilizer. This is called two-way analysis of variance. In case of two-way analysis of variance, we need not make any kind of rearrangement in the given data. Since two criteria are considered, here, there will be two sets of hypotheses.

Types of variances in Two-way ANOVA

1. **Variance between samples (Columns):** This is the net result of the variation different sample means (in respect of columns) from grand mean. Grand mean is the mean of all the observations coming under all sample groups.
2. **Variance between rows:** This is the net result of the variation different sample means (in respect of rows) from grand mean. Grand mean is the mean of all the observations coming under all sample groups.
3. **Variance within the sample (Residual):** This is the net result of variations different items of the sample from the respective sample means.
4. **Variance about the sample:** This is the sum of the variance between columns, variance between rows and the variance within the sample (residual)

Proforma of Two-way ANOVA Table

One-way ANOVA Table				
Source of variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-Ratio
Between Columns	SSC =	c – 1	MSC = SSC/(c–1)	F_C = MSC ÷ MSE, or = MSE ÷ MSC
Between Rows	SSR =	r – 1	MSR = SSR/(r–1)	F_R = MSR ÷ MSE, or = MSE ÷ MSR
Within Sample	SSE =	(c – 1)x(r – 1)	MSE=SSE/(c-1)x(r–1)	
Total	SST =	N –1		

SSC = Sum of Squares between Columns
SSR = Sum of Squares between Rows
SSE = Sum of Square within Samples
SST = Sum of Square Total
MSC = Mean Sum of Squares between Columns
MSR = Mean Sum of Squares between Rows
MSE = Mean Sum of Squares within Samples
c = Number of Columns
r = Number of Rows

Procedure for carrying out Two-way Analysis of variance

1. Set up H_0 and H_1.
   H_0 : There is no significant difference between samples (in respect of columns)
   H_1 : There is significant difference between samples (in respect of columns)

   H_0 : There is no significant difference between samples (in respect of rows)
   H_1 : There is significant difference between samples (in respect of rows)
2. Decide the test statistic:

   Test statistic applicable here is F-test
3. Apply the appropriate formula for computing the values of F ratios.

   F_C = Larger Variance ÷ Smaller Variance. i.e; MSC÷MSE or MSE÷MSC

   F_R = Larger Variance ÷ Smaller Variance. i.e; MSR÷MSE or MSE÷MSR

(i)Find SST.

SST = Sum of square of all items – (T^2/N)

Where T = Total of all observations, N = Total Number of observations

(T^2/N) is generally called correction factor

(ii)Find SSC.

SSC = $[(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots - (T^2/N)$

Where εx_1= sum of items in the first column

εx_2= sum of items in the second column
n_1 = number of items in the first column
n_2 = number of items in the second column

(iii)Find SSR.

SSR = $[(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots - (T^2/N)$

Where εx_1= sum of items in the first row

εx_2= sum of items in the second row
n_1 = number of items in the first row
n_2 = number of items in the second row

(iv)Draw one-way ANOVA Table and enter the values of SST, SSC and SSR

(v)Find the value of SSE. SSE = SST – (SSC+SSR)

(vi)Find the degree of freedom in the third column as indicated in the proforma.

(vii)Find MSC. MSC = SSC ÷ (c–1)

(viii)Find MSR. MSR = SSR ÷ (r–1)

(ix)Find MSE. MSE = SSE ÷ [(c–1) x (r–1)]

(x)Find F–Ratios (i. e; F_C and F_R)

F_C= Larger variance ÷ Smaller variance; [i.e; F= MSC÷MSE, or F= MSE÷MSC]

F_R= Larger variance ÷ Smaller variance; [i.e; F= MSR÷MSE, or F= MSE÷MSR]

4. Specify the level of significance. Take 5% if nothing is mentioned.
5. Fix the degrees of freedom. Fix a pair of d.f. in respect of F_C and F_R.
6. Obtain table value of F_C and F_R at specified level significance and fixed degree of freedom.
7. Compare the Calculated value of Fc with the Table value, and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, H_0 is accepted. If calculated value is more than the table value, H_0 is rejected.
8. Compare the Calculated value of F_R with the Table value, and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, H_0 is accepted. If calculated value is more than the table value, H_0 is rejected.

Qn: Following table shows the yield of crops using 3 varieties of seeds:

Plots	Varieties of Seeds		
	P	Q	R
I	6	7	8
II	4	6	5
III	8	6	10
IV	6	9	9

Test whether there is significant difference in the productivity of varieties of seeds.

Also test the significance of the difference between plots.

Sol:

H_0 : There is no significant difference in the productivity of varieties of seeds.

H_1 : There is significant difference in the productivity of varieties of seeds.

H_0 : There is no significant difference in the productivity of plots.

H_1 : There is no significant difference in the productivity of plots.

Test statistic applicable here is F-test

F = Larger Variance ÷ Smaller Variance

F_C = Larger Variance ÷ Smaller Variance. i.e; MSC÷MSE or MSE÷MSC

F_R = Larger Variance ÷ Smaller Variance. i.e; MSR÷MSE or MSE÷MSR

SST = Sum of square of all items – (T^2/N)

$= 6^2 + 4^2 + 8^2 + 6^2 + 7^2 + 6^2 + 6^2 + 9^2 + 8^2 + 5^2 + 10^2 + 9^2 - (84^2/12)$

$=36+16+64+36+49+36+36+81+64+25+100+81 - (7056/12)$

$= 624 - 588 = \underline{36}$

$SSC = [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots - (T^2/N)$

$= (24^2/4) + (28^2/4) + (32^2/4) - 588$

$= (576/4) + (784/4) + (1024/4) - 588$

$= (2384/4) - 588 = 596 - 588 = \underline{8}$

$SSR = [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots - (T^2/N)$

$= (21^2/3) + (15^2/3) + (24^2/3) + (24^2/3) - 588$

$= (441/3) + (225/3) + (576/3) + (576/3) - 588$

$= (1818/3) - 588 = 606 - 588 = \underline{18}$

One-way ANOVA Table				
Source of variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-Ratio
Between Columns	SSC =8	3 – 1= 2	MSC = 8/2 = 4	F_C = MSC ÷ MSE = 4/1.67 = 2.395
Between Rows	SSR =18	4 – 1 = 3	MSR = 18/3 = 6	F_R = MSR ÷ MSE = 6/1.67 = 3.593
Within Sample	SSE =10	(3 – 1)x(4 – 1) = 6	MSE=10/6 = 1.67	
Total	SST =36	12 –1 = 11		

Between Columns:

Calculated value of $F_C = 2.396$

Level of Significance = 5%

Degrees of freedom = (2,6)

Table value of F_C at 5% level of significance and (2,6) degrees of freedom = 5.14

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference in the productivity of three varieties of seeds.

Between Rows:

Calculated value of $F_R = 3.593$

Level of Significance = 5%

Degrees of freedom = (3,6)

Table value of F_C at 5% level of significance and (3,6) degrees of freedom = 4.76

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference in the productivity of plots.

Coding Method

In analysis of variance, while preparing ANOVA table (both one-way and two-way), at first, we have to find the values of SST, SSC, SSR, etc. But, if the individual observations of the given data are of large values, the computation of SST, SSC, SSR, etc becomes a tedious task. So, as to avoid this complication, we may apply coding method. Coding method refers to the addition, subtraction, multiplication and division of individual observations of the given data by a constant. The addition, subtraction, multiplication or division of all the individual items by a constant will not affect the value of F.

Qn: The following table shows the number of units of a product produced by 5 workers using 4 different types of machines:

Workers	Machines			
	P	Q	R	S
I	44	38	47	36
II	46	40	52	43
II	34	36	44	32
IV	43	38	46	33
V	38	42	49	39

You are required to test:
(1) Whether there is significant difference in the mean productivity of machines.
(2) Whether there is significant difference in the mean productivity of workers.

Sol:

Let us apply coding method by subtracting 45 from each observation of the given data. Then we get;

Workers	Machines			
	P	Q	R	S
I	-1	-7	2	-9
II	1	-5	7	-2
III	-11	-9	-1	-13
IV	-2	-7	1	-12
V	-7	-3	4	-6

H_0 : There is no significant difference in the productivity of machines.

H_1 : There is significant difference in the productivity of machines.

H_0 : There is no significant difference in the productivity of workers.

H_1 : There is no significant difference in the productivity of workers.

Test statistic applicable here is F-test

F = Larger Variance ÷ Smaller Variance

F_C = Larger Variance ÷ Smaller Variance. i.e; MSC÷MSE or MSE÷MSC

F_R = Larger Variance ÷ Smaller Variance. i.e; MSR÷MSE or MSE÷MSR

SST = Sum of square of all items – (T^2/N)

$= -1^2 + 1^2 + -11^2 + -2^2 + -7^2 + -7^2 + -5^2 + -9^2 + -7^2 + -3^2 + 2^2 + 7^2 + -1^2 + 1^2 + 4^2 + -9^2 + -2^2 + -13^2 + -12^2 + -6^2 – (-80^2/20)$

$=1+1+121+4+49+49+25+81+49+9+4+49+1+1+16+81+4+169+144+36 – (6400/20)$

$= 894 – 320 = \underline{574}$

SSC = $[(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots – (T^2/N)$

$= (-20^2/5) + (-31^2/5) + (13^2/5) + (-42^2/5) – (-80^2/20)$

$= (400/5) + (961/5) + (169/5) + (1764/5) – 320$

$= (3294/5) – 588 = 658.8 – 320 = \underline{338.8}$

SSR = $[(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + [(\varepsilon x_1)^2/n_1] + \ldots\ldots\ldots\ldots\ldots – (T^2/N)$

$= (-15^2/4) + (1^2/4) + (-34^2/4) + (-20^2/4) + (-12^2/4) – (-80^2/20)$

$= (225/4) + (1/4) + (1156/4) + (400/4) + (144/4) – 320$

$= (1926/4) – 320 = 481.5 – 320 = \underline{161.5}$

One-way ANOVA Table				
Source of variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-Ratio
Between Columns	SSC =338.8	4 – 1= 3	MSC = 338.8/3 = 112.93	F_C = MSC ÷ MSE = 112.93/6.142 = 18.387
Between Rows	SSR =161.5	5 – 1 = 4	MSR = 161.5/4 = 40.375	F_R = MSR ÷ MSE = 40.375/6.142 = 6.574
Within Sample	SSE = 73.7	(4 – 1)x(5 – 1) = 12	MSE=73.7/12 = 6.142	
Total	SST =574	20 –1 = 19		

Between Columns:

Calculated value of F_C = 18.387

Level of Significance = 5%

Degrees of freedom = (3,12)

Table value of F_C at 5% level of significance and (3,12) degrees of freedom = 3.49

Since calculated value is more than the table value, null hypothesis is rejected. Alternative hypothesis is accepted. So we may conclude that there is significant difference in the mean productivity of machines.

Between Rows:

Calculated value of $F_R = 6.574$

Level of Significance = 5%

Degrees of freedom = (4,12)

Table value of F_C at 5% level of significance and (4,12) degrees of freedom = 3.26

Since calculated value is more than the table value, null hypothesis is rejected. Alternative hypothesis is accepted. So we may conclude that there is significant difference in the mean productivity of workers.

REVIEW QUESTIONS:

1. What do you mean by analysis of variance?
2. Explain the two types of analysis of variance.
3. What are the different types of variances in case of one way analysis of variance?
4. What are the different types of variances in case of two way analysis of variance?
5. Draw the proforma of one way ANOVA table.
6. Draw the proforma of two way ANOVA table.
7. Explain the hypothesis testing procedure in case of one way ANOVA.
8. Explain the hypothesis testing procedure in case of two way ANOVA.
9. What do you mean by coding method in analysis of variance?
10. Following table shows the scores attained by trainees under three different instructional methods:

Methods	Scores				
I	84	71	84	76	85
II	85	76	88	86	90
III	81	68	73	71	82

Test whether there is significant difference in the scores under three methods.

11. A company had 4 salesmen P,Q,R and S, each of whom was sent for a period of one moth to three types of areas, namely, urban area, rural area and semi-urban area. The sales (in thousand rupees) achieved by the salesmen are shown in the following table:

Area	Salesmen			
	P	Q	R	S
Urban	80	80	60	100
Rural	30	30	70	30
Semi-urban	70	40	50	80

Carry out an analysis of variance and interpret the results.

Chapter 13

NON-PARAMETRIC TESTS

Meaning:

A test which is not concerned with testing of parameters is called Non-parametric test. Non-parametric test does not make any assumption about the nature of distribution. Therefore, non-parametric tests are called distribution-free tests.

Situation where non-parametric tests are used

1. When hypothesis does not involve population parameter
2. When observations are not accurate as required for a parametric test.
3. When the researcher thinks that parametric test is not applicable.

Assumptions of Non-parametric tests

1. Samples are drawn randomly
2. Sample observations are independent
3. Observations are measured on ordinal or nominal scale
4. The variable is continuous
5. The probability density function of population is continuous

Advantages of Non-parametric tests

1. It is very simple and easy to apply the non-parametric tests
2. They can be applied when the observations are measured on ordinal or nominal scale
3. There is no assumption about the nature of population distribution
4. They can be used even if the sample is small
5. They have wide application in Psychometry, Sociology, Educational Statistics, etc.

Drawbacks of Non-parametric tests

1. They can be used only if the observations are measured on ordinal or nominal scale.
2. They cannot be used for estimating population parameters
3. The application of all non-parametric tests is not very simple.

Types of Non-parametric tests

1. Chi-square Test
2. Sign Tests
3. Signed Rank Test (Wilcoxon Matched Pairs Test)
4. Rank Sum Tests
5. One Sample Runs Test (Wald Wlfowitz' Runs Test)
6. Kolmogrov - Smirnov Test (K-S Test)

Sign tests

t-test is generally used when sample is small and there is an assumption that the population is normal. Therefore, when sample is small but it is not possible to make an assumption about the nature of population distribution, t-test cannot be applied. In such a case sign test is used. In sign test, to find the value of test statistic, we use the proportion of signs (+ve or -ve signs), not the numerical magnitude. That is why, the test is known as sign test. There are two types of sign tests. They are (a) One sample sign test and (b) Two sample sign test.

One sample sign test

One sample sign test is used to test whether the sample belongs to a particular population.

Procedure:

1. Set up null and alternative hypotheses:
   H_0: There is no significant difference between sample mean and population mean (i.e; $\mu = \mu_0$)
   H_1: There is significant difference between sample mean and population mean (i.e; $\mu = \mu_0$)
2. Decide the test statistic. The test statistic applicable is one sample sign test.
3. Use the appropriate formula for computing the value of test statistic
   Test statistic $= (p - P)/SE$, where $P = ½$
   p = proportion of + signs, (+ or – signs for each observation is determined by subtracting 5 from each of them)
   $SE = \sqrt{(PQ)/n}$ where n = Total number of signs, $Q = 1 - P$
4. Specify the level of significance. Take 5%, if not mentioned.
5. Degree of freedom = infinity
6. Locate the table value of t-test.
7. Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is rejected and otherwise, it is rejected.

Qn: Mr. A had to wait for following time (in minutes) for bus in 15 occasions:

9, 5, 6, 8, 3, 9, 8, 10, 7, 2, 6, 6, 7, 10 and 7 minutes. Use the sign test at 5% level of significance to test the claim of bus that on the average Mr A has to wait 5 minutes.

Sol:

H_0: There is no significant difference between sample mean and population mean (i.e; $\mu = 5$)

H_1: There is significant difference between sample mean and population mean (i.e; $\mu \neq 5$)

The test statistic applicable is one sample sign test.

Test statistic $= (p - P)/SE$ where p = proportion of + signs, $P = ½$
$SE = \sqrt{(PQ)/n}$

Computation of proportion of + signs (p):

Waiting time (x)	(X – 5)
9	+
5	.
6	+
8	+
3	–
9	+
8	+
10	+
7	+
2	–
6	+
6	+
7	+
10	+
7	+
No. of + signs =	12

Total number of signs (n) = 14
Total number of + signs = 12
Proportion of + signs (p) = (12/14) = 0.857
P = 0.5, Q = 1- 0.5 = 0.5
S E = $\sqrt{(0.5 \times 0.5)/14} = \sqrt{0.25/14} = \sqrt{0.01786} = 0.1336$
Test statistic = $(0.857 - 0.5)/0.1336 = 2.672$
Level of significance = 5%
Degree of freedom = infinity
Table value of 't' at 5% level of significance and infinity degrees of freedom is 1.96
Since calculated value is more than the table value, null hypothesis is rejected. We accept alternative hypothesis. So we may conclude that there is significant difference between sample mean and population mean (i.e; $\mu \neq 5$).

Two sample sign test (Paired sample sign test)

Two sample sign test is used to test whether two populations are identical. In case of two sample sign test, each pair is replaced by +ve or –ve sign. If first vale in a pair is larger, assign + ve sign to that pair, and otherwise assign – ve sign. Procedure the procedure is same as in the case of one sign test.

Qn: The following are the scores obtained by 2 students in different tests:

Student I	7	10	14	12	6	9	11	13	7	6	10
Student II	10	13	14	11	10	7	15	11	10	9	8

Use the sign test at 1% level of significance to test the null hypothesis that on an average the two students are identical.

Sol:

H_0: There is no significant difference between students (i.e; performance of the students are identical)
H_1: There is significant difference between students

The test statistic applicable is two sample sign test.
Test statistic $= (p - P)/SE$ where p = proportion of + signs, P = ½

$$S E = \sqrt{(PQ)/n}$$

Computation of proportion of + signs (p):

Student I (x)	Student II (y)	Sign for difference
7	10	+
10	13	+
14	14	.
12	11	–
6	10	+
9	7	–
11	15	+
13	11	–
7	10	+
6	9	+
10	8	–
	No. of + signs =	6

Total number of signs (n) = 10
Total number of + signs = 6
Proportion of + signs (p) = (6/10) = 0.6
P = 0.5, Q = 1- 0.5 = 0.5
$S E = \sqrt{(0.5 \times 0.5)/10} = \sqrt{0.25/10} = \sqrt{0.025} = 0.1581$
Test statistic = (0.6 – 0.5)/0.1581 = 0.1/0.1581 = 0.633
Level of significance = 1%
Degree of freedom = infinity
Table value of 't' at 1% level of significance and infinity degrees of freedom is 2.576
Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that the students are identical in their performance.

Signed Rank Test (Wilcoxon Matched Pairs Test)

Signed Rank Test is another important non-parametric test used to test whether matched paired samples are identical or not. Here we use the signed ranks for testing. Wilcoxon Matched Pairs Test is used differently depending upon following two situations:

(a) When the number of matched pairs ≤ 25, and
(b) When the number of matched pairs > 25.

Signed Rank Test (When the number of matched pairs ≤ 25)

Here, we find the difference of matched pairs and assign them ranks. Then ranks are classified into two categories based on their respective signs. Then take the sun of two categories of ranks. The minimum of the two is considered as the value of test statistic.

Procedure:

1. Set up null and alternative hypotheses:
   H_0: There is no significant difference between samples
   H_1: There is significant difference between samples
2. Decide the test statistic. The test statistic applicable here is Wilcoxon matched pairs test (i.e; Wilcoxon's T test)
3. Use the appropriate formula for computing the value of test statistic (Wilcoxon's T test)
   T = Sum of Positive Ranks or Sum of Negative Ranks, whichever is less.
4. Specify the level of significance. Take 5%, if not mentioned.
5. Degree of freedom = n-1
6. Locate the table value of Wilcoxon's T test.
7. Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is rejected and otherwise, it is rejected.

Qn: The following table shows the details of number of units of a product produced by two workers. Test whether there is significant difference between the performances of the workers using Wilcoxon matched pairs test.

Worker P	73	43	47	53	58	47	52	58	38	61	56	56	43	55	65	75
Worker Q	51	41	43	41	47	32	24	58	43	53	52	57	44	57	40	68

Sol:

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

Decide the test statistic. The test statistic applicable here is Wilcoxon matched pairs test (i.e; Wilcoxon's T test)

T = Sum of Positive Ranks or Sum of Negative Ranks, whichever is less.

orker P	orker Q	fference (d) – Q)		nk of |d|	nks of +ve values	nks of – ve values
73	51	22	22	13	13	
43	41	2	2	2.5	2.5	
47	43	4	4	4.5	4.5	
53	41	12	12	11	11	
58	47	11	11	10	10	
47	32	15	15	12	12	
52	24	28	28	15	15	
58	58	0	0	-	-	
38	43	-5	5	6	-	-6
61	53	8	8	8	8	

56	52	4	4	4.5	4.5	
56	57	-1	1	1		-1
34	44	-10	10	9		-9
55	57	-2	2	2.5		-2.5
65	40	25	25	14	14	
75	68	7	7	7	7	
Total of Signed Ranks					**101.5**	**.5**

The calculated value of T = 101.5 or 18.5 whichever is lower.

∴ T value = 18.5

Level of significance = 5%

Degree of freedom = n-1, (n= number of vlues who have either + or –ve sign)

N = 15

Table value of Wilcoxon's T test at 5% level of significance and 15 df = 25

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference in he performances of workers are P and Q.

Signed Rank Test (When the number of matched pairs > 25)

Here, we find the difference of matched pairs and assign them ranks. Then ranks are classified into two categories based on their respective signs. Then take the total of two categories of ranks. The test statistic is Z test.

Procedure:

1. Set up null and alternative hypotheses:
   H_0: There is no significant difference between samples
   H_1: There is significant difference between samples
2. Decide the test statistic. The test statistic applicable here is Wilcoxon matched pairs test (i.e; Z test)
3. Use the appropriate formula for computing the value of test statistic (Z test)
   $Z = [(T - \mu)/\sigma]$ where T = Sum of Positive Ranks or Sum of Negative Ranks, whichever is less; $\mu = [n(n+1)]/4$; $\sigma = \sqrt{[n(n+1)(2n+1)]/24}$
4. Specify the level of significance. Take 5%, if not mentioned.
5. Degree of freedom = infinity
6. Locate the table value of Z.
7. Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is rejected and otherwise, it is rejected.

Qn: The following are the marks obtained by 26 students before and after giving a special coaching to them:

Marks (before) : 70, 35, 21, 16, 75, 63, 70, 54, 77, 82, 68, 19, 13, 72, 78, 17, 24, 3, 45, 80, 15, 20, 58, 65, 35, 52.
Marks (after) : 79, 62, 90, 37, 35, 14, 26, 32, 90, 54, 85, 44, 83, 90, 92, 32, 34, 28, 34, 79, 35, 32, 62, 63, 30, 68.
Use the signed rank test to test at whether there is significant difference in the marks of students before and after providing special coaching ($\propto$ = 5%).

Sol:

H_0: There is no significant difference in the marks of students before and after giving special coaching.

H_1: There is no significant difference in the marks of students before and after giving special coaching.

The test statistic is Wilcoxon matched pairs test (i.e; Z test)

$Z = [(T - \mu)/\sigma]$ where T = Sum of Positive Ranks or Sum of Negative Ranks, whichever is less; $\mu = [n(n+1)]/4$; $\sigma = \sqrt{[n(n+1)(2n+1)]/24}$

arks (before)	arks (after)	fference (d)		nk of |d|	nks of +ve values	nks of – ve values
70	79	-9	9	6		6
35	62	-27	27	20		20
21	90	-69	69	25		25
16	37	-21	21	17		17
75	35	40	40	22	22	
63	14	49	49	24	24	
70	26	44	44	23	23	
54	32	22	22	18	18	
77	90	-13	13	10		10
82	54	28	28	21	21	
68	85	-17	17	14		14
19	44	-25	25	19		19
13	83	-70	70	26		26
72	90	-18	18	15		15
78	92	-14	14	11		11
17	32	-15	15	12		12
24	34	-10	10	7		7
35	28	7	7	5	5	

45	34	11	11	8	8	
80	79	1	1	1	1	
15	35	-20	20	16		16
20	32	-12	12	9		9
58	62	-4	4	3		3
65	63	2	2	2	2	
35	30	5	5	4	4	
52	68	-16	16	13		13
Total of Signed Ranks					**128**	**223**

T = 128 (128 or 223, whichever is low)

$\mu = [n(n+1)]/4, = 26(26+1)/4 = (26 \times 27)/4 = 702/4 = 175.5$

$\sigma = \sqrt{[n(n+1)(2n+1)]/24} = \sqrt{[26(26+1)(52+1)]/24} = \sqrt{(26 \times 27 \times 53)/24}$

$= \sqrt{37206/24} = \sqrt{1550.25} = 39.373$

$\therefore Z = (128 - 175.5)/39.373 = -47.5/39.373 = -1.21 = \underline{1.21}$ (numerically)

Level of significance = 5%

Degree of freedom = infinity

Table value of Z at 5% level of significance and infinity degree of freedom = 1.96

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference in the performances of students before and after giving special coaching.

Rank Sum Tests

Rank sum tests are another type of tests used for testing whether the populations are identical. Here, various samples are taken together and then ranks are assigned. There are two important types of rank sum tests. They are (a) Wilcoxon-Mann-Whitney Test (U- test), and (b) Kruskal – Wallis Test (H- test).

Wilcoxon-Mann-Whitney Test (U- test):

This method is used when there are two group of samples. The testing procedure is:

1. Set up null and alternative hypotheses:
   H_0: There is no significant difference between two samples
   H_1: There is significant difference between two samples
2. Decide the test statistic. The test statistic applicable here is Wilcoxon Mann Whitney test (i.e; U- test)
3. Use the appropriate formula for computing the value of test statistic (Z test)
   Test Statistic = $[(\mu - U)/SE]$
   where $\mu = (n_1.n_2)/2$

$U = U_1$ or U_2 whichever is less.

$U_1 = n_1.n_2 + [n_1(n_1+1)]/2 - R_1$

$U_2 = n_1.n_2 + [n_2(n_2+1)]/2 - R_2$

R_1 = Rank sum of Sample I

R_2 = Rank sum of Sample II

n_1 = Number of observations in Sample I

n_2 = Number of observations in Sample II

$SE = \sqrt{[n_1.n_2\,(n_1+ n_2+1)]/12}$

4. Specify the level of significance. Take 5% unless specified otherwise.
5. Fix the degrees of freedom. df = infinity
6. Locate the table value of test statistic (i.e; Z test) at specified level of significance and fixed degrees of freedom.
7. Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is rejected and otherwise, it is rejected.

Qn: Apply Wilcoxon- Mann-Whitney Test to test whether the following samples come from populations with same mean (i.e; they are identical):

Sample I	54	39	70	58	47	40	74	49	74	75	61	79
Sample II	45	41	62	53	33	45	71	42	68	73	54	73

Sol:

H_0: There is no significant difference between two samples (i.e; they are identical)

H_1: There is significant difference between two samples (i.e; they are not identical)

The test statistic is Wilcoxon Mann Whitney test (i.e; U- test)

Value of Test Statistic $= [(\mu - U)/SE]$

where $\mu = (n_1.n_2)/2$

$U = U_1$ or U_2 whichever is less.

$U_1 = n_1.n_2 + [n_1(n_1+1)]/2 - R_1$

$U_2 = n_1.n_2 + [n_2(n_2+1)]/2 - R_2$

R_1 = Rank sum of Sample I

R_2 = Rank sum of Sample II

$SE = \sqrt{[n_1.n_2\,(n_1+ n_2+1)]/12}$

Computation of Rank Sums			
lues of Samples together (ascending order)	Rank	Rank	
		sample I	sample II
33	1		1
39	2	2	

40	3	3	
41	4		4
42	5		5
45	6.5		6.5
45	6.5		6.5
47	8	8	
49	9	9	
53	10		10
54	11.5		11.5
54	11.5	11.5	
58	13	13	
61	14	14	
62	15		15
68	16		16
70	17	17	
71	18		18
73	19.5		19.5
73	19.5		19.5
74	21.5	21.5	
74	21.5	21.5	
75	23	23	
79	24	24	
Rank Sum		**R_1=167.5**	**R_2=132.5**

$U_1 = 12 \times 12 + [12(12+1)]/2 - 167.5 = 144 + 78 - 167.5 = 54.5$

$U_2 = 12 \times 12 + [12(12+1)]/2 - 132.5 = 144 + 78 - 132.5 = 89.5$

U = 54.5 or 89.5 whichever is lower, $\therefore U = 54.5$

$\mu = (12 \times 12)/2 = 144/2 = 72$

$SE = \sqrt{[(12 \times 12)(12 + 12 + 1)]/12} = \sqrt{(144 \times 25)/12} = \sqrt{300} = 17.32$

$\therefore$ Test Statistic $= (72 - 54.5)/17.32 = 17.5/17.32 = \underline{1.011}$

Level of significance = 5%.

Degrees of freedom = infinity

Table value of test statistic (i.e; Z test) at 5% level of significance and infinity degrees of freedom = 1.96

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference between two samples. Both the samples come from populations with the same mean.

Kruskal – Wallis Test (H- test):

Here, we tests whether three or more independent sample groups come from the population having the same mean. The testing procedure is:

1. Set up null and alternative hypotheses:

   H_0: There is no significant difference between samples

   H_1: There is significant difference between samples
2. Decide the test statistic. The test statistic applicable here is Kruskal – Wallis test (i.e; H- test)
3. Use the appropriate formula for computing the value of test statistic.

   Test Statistic $H = [12/n(n+1)] \times [\varepsilon R_1^2/n_{1+} \varepsilon R_2^2/n_2)+.....] - 3(n+1)$

   R_1 = Rank sum of Sample I

   R_2 = Rank sum of Sample II

   n_1 = Number of observations in Sample I

   n_2 = Number of observations in Sample II
4. Specify the level of significance. Take 5% unless specified otherwise.
5. Fix the degrees of freedom. df = c–1
6. Locate the table value of test statistic (i.e; Chi-square test) at specified level of significance and fixed degrees of freedom.
7. Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is rejected and otherwise, it is rejected.

Qn: the sales figures of 4 salesmen are given below:

Salesmen	Sales (Rupees in thousands)				
P	171	182	157	148	162
Q	152	175	202	168	176
R	160	155	139	146	166
S	179	142	197	170	158

Test whether 4 salesmen have performed equally. Use Kruskal-Wallis Test at 1%.

Sol:

H_0: There is no significant difference between salesmen

H_1: There is significant difference between salesmen

The test statistic applicable here is Kruskal – Wallis test (i.e; H- test)

Use the appropriate formula for computing the value of test statistic.

Test Statistic $H = [12/n(n+1)] \times [\varepsilon R_1^2/n_{1+} \varepsilon R_2^2/n_2)+.....] - 3(n+1)$

R_1 = Rank sum of Sample I

R_2 = Rank sum of Sample II

n_1 = Number of observations in Sample I

n_2 = Number of observations in Sample II

Computation of Rank Sums					
les figures of 4 salesmen (ascending order)	Rank	Ranks of Salesmen			
139	1			1	
142	2				2
146	3			3	
148	4	4			
152	5		5		
155	6			6	
157	7	7			
158	8				8
160	9			9	
162	10	10			
166	11			11	
168	12		12		
170	13				13
171	14	14			
175	15		15		
176	16		16		
179	17				17
182	18	18			
197	19				19
202	20		20		
Rank Sums		**$_1$=53**	**$_2$=68**	**$_3$=30**	**$_4$=59**

$H = [12/(n(n+1))] \times [(\varepsilon R_1^2/n_1) + (\varepsilon R_2^2/n_2) + \ldots\ldots\ldots] - 3(n+1)$

$= 12/(20 \times 21) \times [(53^2/5)+(68^2/5)+(30^2/5)+(59^2/5) - 3(20+1)$

$= (12/420) \times (561.8+924.8+180+696.2) - 63$

$= (0.02857 \times 2362.8) - 63 = 67.5052 - 63 = \underline{\underline{4.5052}}$

Level of significance = 1%

Degree of freedom = 4 – 1 = 3

Table value of Chi-square at 5% level of significance and 3 d.f = 11.341

Since calculate vale is less than table value, null hypothesis is accepted. So we may conclude that there is no significant difference between 4 salesmen. Their performance are equal.

One Sample Runs Test (Wald Wolfiwitz' Runs Test)

Runs test is used to test the randomness of a sample on the basis of the order in which the observations are taken. A 'run' is a succession of identical items. This test was designed by Wald Wolfowitz. The testing procedure is:

1. Set up null and alternative hypotheses:
   H_0: There is randomness
   H_1: There is no randomness
2. Decide the test statistic. The test statistic applicable here is Z-test.
3. Use the appropriate formula for computing the value of test statistic.
   $Z = (r - \mu)/\sigma$
   where $r =$ Number of runs; $\mu = [2n_1n_2/(n_1+n_2)] + 1$
   $\sigma = \sqrt{[2n_1n_2(2n_1n_2-n_1-n_2)]/ (n_1+n_2)^2(n_1+n_2-1)}$
   n_1 = Number of first item in all the runs together
   n_2 = Number of second item in all the runs together
4. Specify the level of significance. Take 5% unless specified otherwise.
5. Fix the degrees of freedom. df = infinity
6. Locate the table value of Z at specified level of significance and infinity degrees of freedom.
7. Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is rejected and otherwise, it is rejected.

Qn: Test the randomness of following arrangement of students (Boys and Girls) in a class:

B,G,B,G,B,B,B,G,B,G,B,B,B,G,G,B,B,B,B,G,G,B,G,B,B,B,G,B,B,B,G,G,G,B,G,B,B,B ,G,B,G,B,B,B,B,G,G,B

Sol:

H_0: There is randomness

H_1: There is no randomness

The test statistic applicable here is Z-test.

Use the appropriate formula for computing the value of test statistic.

$Z = (r - \mu)/\sigma$

r = Number of runs

$\mu = [2n_1n_2/(n_1+n_2)] + 1$

$\sigma = \sqrt{[2n_1n_2(2n_1n_2-n_1-n_2)]/ (n_1+n_2)^2(n_1+n_2-1)}$

n_1 = Number of first item in all the runs together

n_2 = Number of second item in all the runs together

B,/G,/B,/G,/B,B,B,/G,/B,/G,/B,B,B,/G,G,/B,B,B,B,/G,G,/B,/G,/B,B,B,/G,/B,B,B,/G,G, G,/B,/G,/B,B,B,/G,/B,/G,/B,B,B,B,/G,G,/B/

Number of runs (r) = 27

n_1= 30

n_2= 18

$\mu = [2n_1n_2/(n_1+n_2)] + 1 = [2 \times 30 \times 18/(30+18)] + 1$

$= (1080/48)+1 = 22.5 + 1 = 23.5$

$\sigma = \sqrt{[2n_1n_2(2n_1n_2-n_1-n_2)]/ (n_1+n_2)^2(n_1+n_2-1)}$

$= \sqrt{[2 \times 30 \times 18(2*30*18 - 30-18)]/ [(30+18)^2(30+18-1)]}$

$= \sqrt{1080(1080-30-18)/(48^2 * 47)} = \sqrt{(1080 \times 1032)/108288}$

$= \sqrt{1114560/108288} = \sqrt{10.2926} = 3.208$

$\therefore Z = (27 - 23.5)/3.208 = 3.5/3.208 = \underline{1.091}$

Level of significance = 5%

Degree of freedom = infinity

Table value of Z at 5% level of significance and infinity d f = 1.96

Since calculate vale is less than table value, null hypothesis is accepted. So we may conclude that the arrangement is made at random.

REVIEW QUESTIONS:

1. What do you mean by non-parametric tests?
2. What are the situations under which non-parametric tests are applied?
3. What are the important assumptions of non-parametric tests?
4. What are the important merits of non-parametric tests?
5. What are the important drawbacks of non-parametric tests?
6. What are the different types of non-parametric tests?
7. Distinguish between parametric tests and non-parametric tests.
8. What do you mean by one sample sign test?
9. Explain the hypothesis testing procedure under one sample sign test.

10. Explain the hypothesis testing procedure under two sample sign test.
11. What do you mean by Wilcoxon matched pairs test?
12. Explain the hypothesis testing procedure of Wilcoxon matched pairs test.
13. What is meant by Wilcoxon Mann Whitney U-test?
14. Explain the hypothesis testing procedure of Wilcoxon Mann Whitney U-test.
15. What is meant by Kruskal-Wallis H-test?
16. Explain the hypothesis testing procedure of H-test.
17. What do you mean by one sample runs test?
18. Explain the hypothesis testing procedure under one sample runs test.
19. The following are the measurements of the breaking strength of a certain commodity:

    173, 187,163, 172, 166, 163, 165, 160, 189, 161, 171, 158, 151, 169, 162, 163, 139, 172, 165 and 148. Use sign test to test the null hypothesis that mean breaking strength of the commodity is 160.
20. A driver buys petrol either at station X or at station Y. the following arrangement shows the order of the stations from which the driver bought petrol over a certain period of time:

    X, X, X, Y, X,Y, X,Y, X, Y, Y, Y, X, Y, Y, Y, X, Y, Y, X, Y, X, Y, X, Y, Y, X, Y, Y, X, X,Y, X, Y, Y, Y, X, Y, X, X, X, Y, X, X, Y, X, X, X, X, Y.

Chapter 14

SAMPLE SIZE DETERMINATION

Determination of size of sample is very important. If the sample size is very large, it will be very difficult to manage the data. But, if the size is too small, the sample will not represent the population, and the conclusion drawn may not be correct. Therefore, the size of sample must be optimum.

Following are some of the important formulae commonly used for determining sample size:

A. Sample Size Determination While Estimating Population Mean When Population is Infinite

Sample Size (n) = $(Z\sigma/e)^2$

where n = sample size, Z = table value, σ = S D of population

e = allowable difference between population mean and sample mean.

Qn: From the details given below, determine the sample size for estimating population mean:

(a) Population S D = 15
(b) Confidence level = 99%
(c) Estimate should be within 6 units of the population mean.

Sol:

$n = (Z\sigma/e)^2$

Population S D (σ) = 15

Allowable difference (e) = 6

Value of Z at 1% level of significance and infinity d f = 2.576

$\therefore n = [(2.576 \times 15)/6]^2 = (38.64/6)^2 = 6.44^2 = 41.474$

$= \underline{41}$

B. Sample Size Determination While Estimating Population Mean When Population is Finite

Sample Size (n) = $[Z^2N\sigma^2] / \{[(N-1)e^2] + [Z^2\sigma^2]\}$

where Z = table value of Z; N = Size of population; σ = S D of Population

e = allowable difference between population mean and sample mean.

Qn: From the details given below, determine the sample size for estimating population mean:

(a) Population size = 5000
(b) Variance of the population = 4
(c) Estimate should be within 0.4 units of the true value of the population mean

(d) Desired level of confidence = 99%

Sol:

$n = [Z^2N\sigma^2] / \{[(N-1)e^2] + [Z^2\sigma^2]\}$

Population Sixe (N) = 5000

Population S D $= \sqrt{4} = 2$

Allowable difference (e) = 0.4

Value o at 1% level of confidence and infinity df = 2.576

$\therefore n = [2.576^2 * 5000*2^2] / \{[(5000-1)0.4^2] + [2.576^2*2^2]\}$

$= (6.6358*5000*4) / (799.84 + 26.543) = 132716/826.383$

$= 160.599 = \underline{\underline{161}}$

C. Sample Size Determination While Estimating Population Proportion When Population is Infinite

Sample Size (n) = $[Z^2pq/ e^2]$

where Z = Table value of Z

p = sample proportion; q = (1 – p)

e = allowable difference between population proportion and sample proportion.

Qn: It is decided to draw a sample from a population to estimate the percentage of defectives within 2% of the true value with 95.5% confidence, on the basis of 3% defective in the sample. What should be the sample size?

Sol:

$n = [Z^2pq/ e^2]$

p = 3% = 0.03; q = 1 – 0.03 = 0.97

e = 2% = 0.02

Value of Z at 4.5% level of significance and infinity df = 2.005

$\therefore n = [(2.005^2 *0.03*0.97)/0.02^2] = 0.116983/0.0004 = 292.457$

$n = \underline{\underline{292}}$

D. Sample Size Determination While Estimating Population Proportion When Population is Finite

Sample Size (n) = $[Z^2Npq] / \{[(N-1)e^2] + [Z^2pq]\}$

Where Z = Table value of Z; p = sample proportion; q = (1 – p)

N = Size of population

e = allowable difference between population proportion and sample proportion.

Qn: It is decided to draw an optimal sample from a population of 5000 units to estimate the percentage of defectives on the basis of 3% defectives in the sample within 0.05 units of its true value. Level of confidence desired is 95%.

Sol:

Sample Size (n) $= [Z^2Npq] / \{[(N-1)e^2] + [Z^2pq]\}$

N = 5000

p = 3% = 0.03

q = (1 – 0.03) = 0.97

e = 0.05

Table Value of Z at 5% level of confidence and infinity df = 1.96

$\therefore n = [1.96^2*5000*0.03*0.97] / \{[(5000-1)0.05^2] + [1.96^2*0.03*0.97]\}$

$= 558.9528/(12.4975 + 0.111791) = 558.9528/12.6093$

$= 44.33 = \underline{44.}$

REVIEW QUESTIONS:

1. What do you mean by sample size?
2. What are the important formulae used for determining sample size while estimating population mean?
3. What are the important formulae used for determining sample size while estimating population proportion?

Chapter 15

STATISTICAL ESTIMATION

Statistical estimation is one of the important branches of Statistical inferences. It is concerned with estimation of population parameters with the help of samples drawn from that population. The accurate value of population parameter can be computed only by an exhaustive study of the population. But, it is infeasible to collect date from each and every element of the population. Therefore, we estimate that population parameters through sample. This is the actual process of statistical estimation.

Two types of estimates are generally used for estimating population parameter. They are (a) Point Estimate and (b) Interval Estimate.

Point Estimation

If a single statistic is used as an estimate of an unknown parameter, it is called point estimate of that parameter. Eg; when the particular value of the sample mean is called the "estimate", sample mean is called the "estimator".

Properties of a good Estimator:

1. An estimator should be unbiased
2. An estimator should be consistent
3. An estimator should be efficient
4. An estimator should be sufficient

Methods used for Point Estimation:

1. Method of maximum likelihood
2. Method of moments
3. Method of minimum variance
4. Method of least squares
5. Method of minimum chi-square
6. Method of inverse probability

Interval Estimation

An estimate which suggests the lowest and highest values within which population parameter is expected to lie, they are called the interval estimates. Here, the two limits (lower and upper) give an interval.

Qn: from the following data, find the limits within which population mean may lie:

Sample size = 100; Sample mean = 45; Sample S.D = 15

Sol:

$\bar{x} = 45;\ S = 15;\quad n = 100$

Here, since the sample is large, the test statistic = Z test

Z = Difference / SE

Degrees of freedom = infinity

The table value of Z at 5% level of significance and infinity d f = 1.96

$Z = [\bar{x} \pm \mu] / [s/\sqrt{n}]$

$1.96 = [45 \pm \mu] / [15/\sqrt{100}]$

∴95% Confidence limits of population mean = $45 \pm (1.96 \times 1.5)$

95% Confidence limits of population mean = 45 ± 2.94

95% Confidence limits of population mean = 42.06 and 47.94

Qn: Estimate the limits within which population mean lie at 95% level of confidence:

$n = 25, \quad \bar{x} = 4800, \quad s = 500$

Sol:

t = Difference / SE

Degrees of freedom = n-1 = 25-1 = 24

The table value of t at 5% level of significance and 24 d f = 2.064

$t = [\bar{x} \pm \mu] / [s/\sqrt{n-1}]$

$2.064 = [4800 \pm \mu] / [500/\sqrt{25-1}]$

$2.064 = [4800 \pm \mu] / [500/\sqrt{24}]$

$2.064 = [4800 \pm \mu] / 102.06$

∴95% Confidence limits of population mean = $4800 \pm (2.064 \times 102.06)$

95% Confidence limits of population mean = 4800 ± 210.65

95% Confidence limits of population mean = 4589.35 and 5010.65

Qn: Out of a sample of 500 items drawn from a population, 2% were found to be defective. Estimate the proportion of population defectives at 95% confidence level. Also find number of expected defectives in the daily production of 60,000 units.

Sol:

Sample proportion of defectives (p) = 2% = 0.02

Confidence level = 95%

Degrees of Freedom = infinity

95% Confidence limits of population proportion = $p \pm (Z \times SE)$

Here, SE $= \sqrt{(pq/n)} = \sqrt{(0.02 \times 0.98)/500}$

$= \sqrt{0.0196/500} = 0.006261$

95% Confidence limits of population proportion = $0.02 \pm (1.96 \times 0.006261)$

$= 0.02 \pm 0.012272$

= 0.007728 and 0.032272

Expected number of defectives = 0.007728x60000; 0.032272 x 60000

= (463.68 , 1936.32)

= (464, 1936)

REVIEW QUESTIONS:

1. What do you mean by statistical estimation?
2. What are the two types of estimation?
3. What do you mean by Point Estimation?
4. What are the various methods used for point estimation?
5. What is meant by Interval Estimation?
6. Distinguish between Estimate and Estimator.
7. What are the important characteristics (properties) of a good estimator?
8. Distinguish between point estimation and interval estimation.
9. A random sample of 50 people from a population showed incomes with a mean = 50000 and Standard Deviation = 6000. Estimate the population mean with 95% and 99% confidence level.
10. In a sample of 500 units of a commodity from a large consignment, 40 units were considered defective. Estimate the percentage of defective in the whole consignment and limits within which the percentage will probably lie.

.*********.

Chapter 16

SOFTWARES FOR QUANTITATIVE METHODS

Microsoft Excel: An Introduction

The aim of this chapter is to provide an introduction to using Microsoft Excel for quantitative data analysis within the context of a business and management research project. It covers some of the key features of Excel that are particularly useful when doing a research project. Further it gives information on the use of Excel to apply various analysis techniques discussed in various chapters. The information are presented here on the assumption that you are already familiar with the basics of using Excel such as how to create worksheets, enter data, use of formulae and functions, create charts (graphs), print and work, etc. If you have never used Excel, there are many textbooks to get you started. Alternatively, you may find that Excel training or support material is available in your institution. There are also various websites, including Microsoft's Office Support area (http://office.microsoft.com/en-001/support/?CTT=97) that offers advice to get you started.

Why use Excel?

With so many specialist software packages available, why use Excel for statistical analysis? Convenience and cost are two important reasons: many of us have access to Excel on our own computers and do not need to source and invest in other software. Another benefit, particularly for those new to data analysis, is to remove the need to learn a software program as well as getting to grips with the analysis techniques. Excel also integrates easily into other Microsoft Office software products which can be helpful when preparing reports or presentations.

What you can do with Excel?

As a spreadsheet, Excel can be used for data entry, manipulation and presentation but it also offers a suite of statistical analysis functions and other tools that can be used to run descriptive statistics and to perform several different and useful inferential statistical tests that are widely used in business and management research. In addition, it provides all of the standard spreadsheet functionality, which makes it useful for other analysis and data manipulation tasks, including generating graphical and other presentation formats. Finally, even if using customised statistical software, Excel can be helpful when preparing data for analysis in those packages.

Limitations of Excel

Even though it has wide applications and usage in data analysis, Excel is not free from limitations. It remains first and foremost a spreadsheet package. Inevitably it does not cover many of the more advanced statistical techniques that are used in research. More surprisingly, it lacks some common tools (such as box plots) that are widely taught in basic statistics. There is also concern amongst some statisticians over the format of specific output in some

functions. The extensive range of graph (chart) templates is also criticised for encouraging bad practice in data presentation through inappropriate use of colour, 3-D display, etc. Despite these limitations Excel remains a very valuable tool for quantitative data analysis as you will see.

Quantitative data analysis tools in Excel

Excel includes a large number of tools that can be used for general data analysis. Here our primary concern is those that are relevant to the statistical and related analysis techniques introduced in earlier chapters. Four sets of tools are particularly useful:

(1) Statistical functions:

Excel offers a broad range of built-in statistical functions. These are used to carry out specific data manipulation tasks, including statistical tests. An example is the AVERAGE 1 function that calculates the arithmetic mean of the cells in a specified range. A list of Excel functions referred to in this and other guides is included in Appendix A along with instructions on how to access them.

(2) Data Analysis Tool Pak:

The Data Analysis Tool Pak is an Excel add-in. It contains more extensive functions, including some useful inferential statistical tests. An example is the Descriptive Statistics routine that will generate a whole range of useful statistics in one go. An introduction to loading and using the Tool Pak add-in is included at Appendix B. The ToolPak is not available in Excel for Mac. See Appendix B for an alternative.

(3) Charts:

Excel's in-built charts (graphs) cover most of the chart types introduced in Chapter 13 and are invaluable in data exploration and presentation. We illustrate their use in Chapter 13 and also in the other guides.

(4) Pivot tables

Pivot tables provide a way of generating summaries of your data and organising data in ways that are more useful for particular tasks. They are extremely useful for creating contingency tables, cross-tabulations and tables of means or other summary statistics. A brief introduction to creating pivot tables is given in the guide Data exploration in Excel: univariate analysis.

Preparing Excel for analysis

Before starting, check that your Data Analysis ToolPak has been loaded. Do this by selecting the Data tab; the Data Analysis command should appear in Analysis group on the right-hand side of the ribbon.

Setting up your data for analysis

Typically there are two options for getting your data into Excel:

1. Import the data in a suitable format from, for example, an online survey tool.
2. Enter the data manually.

If you are going to enter your data manually use a single worksheet to hold all the data in your dataset and set up the worksheet with variables (questions) as the columns and the cases (e.g. respondents) as the rows. An individual cell, therefore, contains a respondent's answer to a specific question.

Allocate column headers

In the first row, give each column a simple, informative header that will be easy to understand when entering data or reviewing output. Avoid just using question numbers (e.g. Q1, Q2, etc.) as these can be confusing if you have a large number of questions. Instead, use a simple naming system. A variable measuring customer satisfaction, for example, could be headed CSat: easy to remember and not likely to be confused during analysis. Ensure each header is unique (this will facilitate subsequent analysis and avoid confusion when interpreting output).

Allocate each case a unique ID

If they do not have one already, allocate each case in the dataset a unique numerical identifier (ID). The easiest way to do this is simply to number them consecutively from 1 through to *n* (where *n* is the number of cases). For clarity, it is best to put the ID as the first column in the worksheet. Giving each respondent a unique ID aids in sorting and tracking individual responses when (for example) cleaning the data or checking outliers. A simple, consecutive number ID system also makes it easy to reorder the data if needed. If you are transferring data from paper copies of a questionnaire, it is useful to write the ID number onto the paper copy to make it easier to check any errors.

Entering your data

Once the spreadsheet is set up, simply enter the data into the appropriate cell as required. Numerical data can be entered as numbers, other data, such as Likert scale data, may need to be coded .With nominal data you have two options:

- Enter the values as words (e.g. male/female), appropriately abbreviated if required (e.g. m/f). Ensure you are consistent in spelling and format as Excel will treat each variation as a different value.

- Enter the re-coded numerical values (e.g. 0/1 for male/female), ensuring you keep a record in a code book (Chapter 13). A worksheet in the workbook is a useful place to record details of your variables and to store your code book as shown in Figure 3.

Which to do depends on your analysis needs. Some tools in Excel (e.g. pivot tables) work well with text and generate meaningful output but some analysis tasks may require numerically coded data. If you are exporting your data to another software package, check the format required by that package. In some cases, it may be helpful to have both formats. You can do this by creating a copy of the column containing the original data, then selecting the new column and using Home > Find & Select > Replace to replace the original values with the new ones. Ensure you give the new column a unique header.

Importing data

If you are importing the data from another electronic file, check that the layout is suitable (i.e. respondents as rows, variables as columns), add or modify variable names if required, add respondent ID if needed and check that the data has imported correctly.

Managing your data

Once you have created your dataset, ensure that you back it up in a secure place, not on your PC or laptop. If you make any changes to your master dataset, record those changes and create a duplicate back-up.

Give files a meaningful name. It is also helpful to date them as this makes it easier to track back if you need to do so. Worksheet tabs can also be named to help you manage your data.

Preparing your data

Once your data are entered you can follow the steps in Chapter 13 to prepare your data for analysis. If you need to carry out data transformation, such as recoding variables or calculating summated scores, do so now. (Hint: you can use functions such as SUM and

AVERAGE to help you with creating summated scales.) If you are creating new variables during data transformation ensure they are given unique column headers.

Example statistical functions

Function name	Description
AVERAGE	Returns the arithmetic mean (average) of the given numbers
CHISQ.DIST.RT	Returns the right-tailed probability for the chi squared distribution
CHISQ.TEST	Returns the *p*-value for the chi-squared test of association
CONFIDENCE.T	Returns the margin of error for a confidence interval for the mean
COUNT	Counts the number of cells in a range that contain numbers
COUNTIF	Counts the number of cells in a range that meet a given condition
KURT	Returns the kurtosis of a dataset
MAX	Returns the maximum value of the given numbers
MEDIAN	Returns the median of the given numbers
MIN	Returns the minimum value of the given numbers
MODE.SNGL	Returns the mode of the given numbers
PEARSON	Returns the Pearson correlation coefficient (*r*) of two variables
SKEW	Returns the skewness of a dataset
STDEV.P	Returns the standard deviation of the given numbers, based on the population
STDEV.S	Returns the standard deviation of the given numbers, based on a sample
VAR.P	Returns the variance of the given numbers, based on the population
VAR.S	Returns the variance of the given numbers, based on a sample

Using a function

We will introduce specific functions in the other guides but the following example of applying the AVERAGE function to calculate the mean age in the sample dataset in Figure 2 illustrates their use:

- Select the cell in which you wish the calculation to be placed (Hint: if you are using the same worksheet as your dataset, avoid cells that are immediately adjacent to your data).

- Select Formulas > More Functions > Statistical > AVERAGE to open the Function Argument dialogue box

SPSS (Statistical Package for Social Sciences)

SPSS (Statistical package for social sciences) is the set of software programs that are combined together in a single package. The basic application of this program is to analyze scientific data related with the social science. This data can be used for market research, surveys, data mining, etc.

With the help of the obtained statistical information, researchers can easily understand the demand for a product in the market, and can change their strategy accordingly. Basically, SPSS first store and organize the provided data, then it compiles the data set to produce suitable output. SPSS is designed in such a way that it can handle a large set of variable data formats.

How SPSS Helps in Research & Data Analysis?

SPSS is revolutionary software mainly used by researchers which help them process critical data in simple steps. Working on data is a complex and time consuming process, but this software can easily handle and operate information with the help of some techniques. These techniques are used to analyze, transform, and produce a characteristic pattern between different data variables. In addition to it, the output can be obtained through graphical representation so that a user can easily understand the result. Read below to understand the factors that are responsible in the process of data handling and its execution.

1. Data Transformation: This technique is used to convert the format of the data. After changing the data type, it integrates same type of data in one place and it becomes easy to manage it. You can insert the different kind of data into SPSS and it will change its structure

as per the system specification and requirement. It means that even if you change the operating system, SPSS can still work on old data.

2. Regression Analysis: It is used to understand the relation between dependent and interdependent variables that are stored in a data file. It also explains how a change in the value of an interdependent variable can affect the dependent data. The primary need of regression analysis is to understand the type of relationship between different variables.

3. ANOVA(Analysis of variance): It is a statistical approach to compare events, groups or processes, and find out the difference between them. It can help you understand which method is more suitable for executing a task. By looking at the result, you can find the feasibility and effectiveness of the particular method.

4. MANOVA(Multivariate analysis of variance): This method is used to compare data of random variables whose value is unknown. MANOVA technique can also be used to analyze different types of population and what factors can affect their choices.

5. T-tests: It is used to understand the difference between two sample types, and researchers apply this method to find out the difference in the interest of two kinds of groups. This test can also understand if the produced output is meaningless or useful.

This software was developed in 1960, but later in 2009, IBM acquired it. They have made some significant changes in the programming of SPSS and now it can perform many types of research task in various fields. Due to this, the use of this software is extended to many industries and organizations, such as marketing, health care, education, surveys, etc.

Advantages of SPSS:

The advantages of using SPSS as a software package compared to other are:

• SPSS is comprehensive statistical software.

• Many complex statistical tests are available as a built in feature.

• Interpretation of results is relatively easy.

• Easily and quickly displays data tables.

• Can be expanded.

Limitations of SPSS:

Following are the important limitations of SPSS:

• SPSS can be expensive to purchase for students.

• Usually involves added training to completely exploit all the available features.

• The graph features are not as simple as of Microsoft Excel.

SPSS Windows and Files

SPSS statistics has three main windows and a menu bar at the top. These allow to:

(1) See your data
(2) See your statistical output
(3) See any programming commands you have written.

Each window corresponds to a separate type of SPSS file.

Students are directed to acquaint with the application of SPSS in performing testing of hypotheses.

REVIEW QUESTIONS:

1. What is Microsoft Excel?
2. What are the important quantitative data analysis tools in Microsoft excel?
3. List the various statistical functions which can be performed in Microsoft excel.
4. What are the important limitations of Microsoft excel?
5. What is SPSS?
6. How does SPSS help in research and data analysis?
7. What are the advantages of SPSS?
8. What are the important limitations of SPSS?

.*********.