SUBJECT CODE : 217528

As per Revised Syllabus of

SAVITRIBAI PHULE PUNE UNIVERSITY

Choice Based Credit System (CBCS)

S.E. (AI and DS) Semester - IV

STATISTICS

Dr. Yogita M. Ahire
Ph.D. (Applied Mathematics), M.Sc., B.Ed, SET
Assistant Professor,
PVG's College of Engineering
and S.S. Dhamankar Institute of Management,
Nashik.

Dr. Vanita R. Daddi
Ph.D. (Mathematics),M.Phil., M.Sc., SET
Assistant Professor,
Dr. D.Y. Patil Institute of Engineering,
Management and Research, Akurdi,
Pune.

Mr. Sachin S. Jamadar
M.Sc., B.Ed., SET
Assistant Professor,
Dr. D.Y. Patil Institute of Engineering,
Management and Research,
Akurdi, Pune.

Dr. Indrayani Y. Sutar
Ph.D.(Numerical Analysis),
M.Phil. (Classical Mechanics),
Assistant Professor,
Dr. D.Y. Patil Institute of Technology,
Pimpri.

Dr. Sandhyatai D. Kadam
Ph.D. (Integral Equations and its Applications)
M. Sc. (Mathematics),
Assistant Professor,
Dr. D.Y. Patil Institute of Technology,
Pimpri.

TECHNICAL PUBLICATIONS
SINCE 1993
An Up-Thrust for Knowledge

SYLLABUS

Statistics - (217528)

Credit :	Examination Scheme :
03	Mid_Semester (TH) : 30 Marks End_Semester (TH) : 70 Marks

Unit I Introduction To Statistics And Sampling Theory

Statistics : Introduction, Origin and Development of Statistics, Definition, Importance and Scope, Limitations, Distrust of Statistics

Population and Sample : Sampling - Introduction, Types of Sampling, Purposive Sampling, Random Sampling, Simple Sampling, Stratified Sampling, Parameter and Statistic, Sampling Distribution, Statistical Inference, Sampling With and Without Replacement, Random Samples : Random Numbers, Population Parameters, Sample Statistics, Sampling Distributions. **(Chapter - 1)**

Unit II Descriptive Statistics : Measures Of Central Tendency

Frequency Distributions and Measures of central Tendency : Frequency Distribution, Continuous Frequency Distribution, Graphic Representation of a Frequency Distribution, Histogram, Frequency Polygon, Averages or Measures of Central Tendency or Measures of Location, Requisites for an Ideal Measure of Central Tendency, Arithmetic Mean, Properties of Arithmetic Mean, Merits and Demerits of Arithmetic Mean, Weighted Mean, Median, Merits and Demerits of Median, Mode, Merits and Demerits of Mode, Geometric Mean, Merits and Demerits of Geometric Mean, Harmonic Mean, Merits and Demerits of Harmonic Mean, Selection of an Average. **(Chapter - 2)**

Unit III Descriptive Statistics : Measures of Dispersion

Measures of Dispersion, Skewness and Kurtosis : Dispersion, Characteristics for an Ideal Measure of Dispersion, Measures of Dispersion, Range, Quartile Deviation, Mean Deviation, Standard Deviation and Root Mean Square Deviation, Coefficient of Dispersion, Coefficient of Variation, Skewness, Kurtosis.

Correlation and Regression : Bivariate Distribution, Scatter diagrams, Correlation, Karl Pearson's coefficient of correlation, Rank correlation, Regression, Lines of Regression, Regression Coefficients, Binomial and multinomial distributions, Poisson distribution, Uniform distribution, Exponential distribution, Gaussian distribution, Log-normal distribution, Chi-square distribution. **(Chapter - 3)**

Unit IV Random Variables And Probabilty Distributions

Random Variables and Distribution Functions :
Random Variable, Distribution Function, Properties of Distribution Function, Discrete Random Variable, Probability Mass Function, Discrete Distribution Function, Continuous Random Variable, Probability Density Function.

Theoretical Discrete Distributions : Bernoulli Distribution, Binomial Distribution, Mean Deviation about Mean of Binomial Distribution, Mode of Binomial Distribution, Additive Property of Binomial Distribution, Characteristic Function of Binomial Distribution, Cumulants of Binomial Distribution , Poisson Distribution, The Poisson Process, Geometric Distribution. **(Chapter - 4)**

Unit V Inferential Statistics : Hypothesis

Statistical Inference - Testing of Hypothesis, Non-parametric Methods and Sequential Analysis : Introduction, Statistical Hypothesis (Simple and-Composite), Test of a Statistical Hypothesis, Null Hypothesis, Alternative Hypothesis, Critical Region, Two Types of Errors, level of Significance, Power of the Test. **(Chapter - 5)**

Unit VI Inferential Statistics : Tests For Hypothesis

Steps in Solving Testing of Hypothesis Problem, Optimum Tests Under Different Situations, Most Powerful Test (MP Test), Uniformly Most Powerful Test, likelihood Ratio Test, Properties of Likelihood Ratio Test, Test for the Mean of a Normal Population, Test for the Equality of Means of Two Normal Populations, Test for the Equality of -Means of Several Normal Populations, Test for the Variance of a Normal Population, Test for Equality of Variances of two Normal Populations, Non-parametric Methods, Advantages and Disadvantages of Non-parametric Methods. **(Chapter - 6)**

TABLE OF CONTENTS

Unit - I

Unit - II

Unit - III

Unit - IV

Unit - V

Unit - VI

UNIT - I

1 Introduction to Statistics and Sampling Theory

Syllabus

Statistics : *Introduction, Origin and Development of Statistics, Definition, Importance and Scope, Limitations, Distrust of Statistics*

Population and Sample: *Sampling - Introduction, Types of Sampling, Purposive Sampling, Random Sampling, Simple Sampling, Stratified Sampling, Parameter and Statistic, Sampling Distribution, Statistical Inference, Sampling With and Without Replacement, Random Samples : Random Numbers, Population Parameters, Sample Statistics, Sampling Distributions*

Contents

1.1 Origin and Development of Statistics

- The word statistics might be derived from the latin word 'status' or 'statista' in Italian language. The meaning of this word is same that is political state. In ancient era the kings of different states used to survey the data of population in their states. The purpose of doing this was to ensure the manpower and later to regain the taxes. In the historical era during the wars also the kings were collecting this information to meet loss in war and implementation of new takes to balance the loss.
- The theory of probability gives a new face to statistics. It has been developed by many researchers from India, England, France and Germany. Indian scientists also contributed for development of statistics. Sample survey, design of experiments in agricultural, multivariate analysis are some of remarkable contributions to statistics.
- A theory of probability being a part of modern statistics came into play in the mid seventeenth century with introduction of theory of probability and theory of games and chance. A gambling game and theory of mathematics contributed for development of statistics which were from France, Germany and England. A problem of points solved by french mathematician Pascal and P. Fermat which was raised by gambler chevalier de-Mere. It became a foundation for theory of probability which is backbone of modern theory of statistics.
- This development of statistics was carried by notable mathematicians in their theories and research work. Among them Jame Bernoulli (1654 - 1705) who wrote the first treatise on the theory of probability. Next De-Moivre (1667 - 1754) contributed for probabilities and annuities and published his work in 'The Doctrine of chances' in 1718. Laplace (1749 - 1827) published his work in 1782 which was on theory of probability. Gauss (1777 - 1855) a notable contributer in consistent development of statistics gave 'Principle of least squares' and the 'Normal Law of Errors'. A modern theory of probability touched by many in 18^{th}, 19^{th} and 20^{th} centuries such as euler, lagrange, bayes, etc.
- Karl Pearson, a pioneer of 'Correlation Analysis' gave the first and more significant test in statistics known as Chi-square Test (X^2-test) of goodness of fit. W.S. Gosset found a t-test for exact (small) sample tests.
- R. A. Fisher (1890-1962) who used statistics in different fields like agriculture, genetics, education etc. known as 'Father of statistics'. His contribution in theory of statistics became the foundation for research workers in statistics. He is pioneer in estimation theory, exact sampling distributions, analysis of variance and design of experiments.

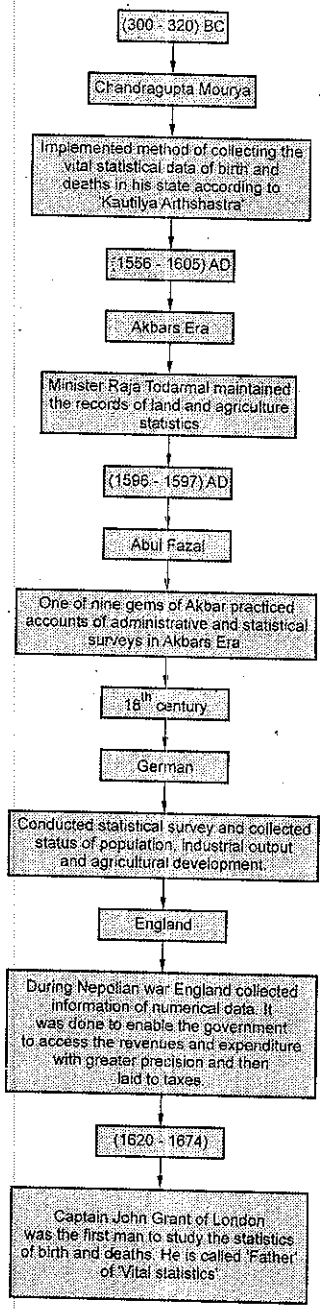


Fig. 1.1.1

- In his honour his contribution is described in the following manner.

  "R. A. Fisher is the real giant in the development of the theory of statistics".

- Indian mathematicians also contributed notably for development of statistics in diversified fields. Among them a few are Parthsarathi (Theory of probability) R.C. Bose, Panse, J.N. Srivastava (Design of experiments in agriculture), C.R. Rao (Statistical inference), P.C. Mahalanobis and P.V. Sukhatme (Sample survey) etc. shared distinguishable work in statistics.

Definition

- Statistics is a branch of mathematics which is the collection, description, analysis and interpretation of data.
- Statistics mainly deals with the information or any data which is on a large scale and needs to be estimated properly.

  Examples : Data mining, data compression, artificial intelligence and network, traffic modeling. Examples given above are the integral part of A.I.D.S.

1.2 Importance and Scope of Statistics

- In various field now a days this type of analysis and estimation of data become mandatory. Without doing statistical treatment one can't reach to conclusion of any technique, process or up and down of factors mainly seasonal. Some misunderstandings of people etc. No field remained untouched from statistics methods such as business, economics, science and engineering and social sciences.
- A statistical approach day-by-day becoming dominant almost in every field such as business, management and marketing. If we expect the good quality products and processes within business organizations, then statistical treatment becomes unavoidable.

1.2.1 Statistics in Business and Economics

- The economist and managers have to face and analyze the data, data gathering, management and data interpretation daily. A good businessman has to be kin about the desire of common customer and their needs. They have to rely on statistical techniques. A success of business depends entirely on statistical correct forecast which lead to satisfaction of supply and demand of a product. It need accuracy and precision and it is possible with good statistical method. Example : To start manufacturing of electronic components. Then it must considered that what is the scope of device in industry ? How many gazettes require such a device on routine basis ?

- How should one present this product in market ? To achieve this it must be prepared with well plan in advance which can not be done without quantitative facts.
- Economical trends are also based on statistical techniques. It is useful in finding the economical solution over price of product labor cost, transportations etc. The tools in statistics like data analysis, number indexing time series analysis demand and forecasting techniques are used to improve planning and economical development.
- Index number helps us to arrive at a single conclusion which represents general level of variation in factors like increase, decrease and rate of increase and rate of decrease. This method covers commodity prices, industrial production sales import and export, etc.

1.2.2 Statistics in Engineering and Technology

- The design, modeling and manufacturing of a product is an engineering. The process which is used for above mentioned is the use of technology. Which is suitable for design, modeling and manufacturing is the outcome of statistical study of principals, theories and mathematical treatment etc. statistical quality control techniques are based on theory of probability and sampling. It has been extensively used in industries like automobile textile, electronic component making, electric equipment, computer softwares, etc.
- The different updated skills are the most demanded part of development of engineering. It is not only to apply and design of product but also to keep balance between current technologies.
- Engineers job is to apply the knowledge and create a product statistical survey plays key role in this development.

1.2.3 Statistics and Medical or Biological Fields

- In medical field issues related to diseases are important aspects to study variation in growth of any virus. The conditions in surrounding like increase or decrease in temperature, moisture content of air also affect rate of growth or decrease in viruses. The collection and analysis of data on the basis of infections produced among people presented statistically always helps to reach to conclusion. On the basis of these conclusion medicines can be prepared for curing the patients. The statistics is not only a mathematical tool but also proved to be a useful tool for mankind. On the other side in the treatment of patients uses of drugs, its side effects and advantages are also studied statistically. This type of survey leads to finalize the drug and treatment of diseases.

- In biological fields cells of plants, any disease to plant as per atmospheric variation are studied by collecting species of different kind. The proper use of particular pesticide depends on survey of different kind of plants, soil quality of water, etc.
- This was first studied by Francis Galton in his work in 'Regression'. According to Prof. Karl Pearson the whole "Theory of heredity" rests on statistical basis. The efficiency of a manufactured drug or injection or medicine is tested by using the test of significance (t - test).

1.3 Importance of Statistics

1) Statistics is important mainly in data analysis.
2) It provides a single solution to multiple variational parametric data.
3) Statistics sometime used find and establish the relations between physical quantities.
4) Statistics also provides the prediction and estimation of certain relevant data.
5) It also provides dimensional nature of data like survey, spread of data, divergence of data.
6) Statistics support to reach to conclusion of a processed data.
7) Statistics itself provides a tool of comparison.
8) Uncertainty in data is reduced in statistical estimation.
9) Statistics is a tool of comparison, investigation of data independent of type of filed.
10) Statistics is important in business, economics, industry, technology science and many more.

1.4 Limitations of Statistics

1.4.1 Statistics does not Study Individuals

- Statistics mainly deals with data in bulk form and does not treat any individual part in separate. Individual data treated statistically becomes meaningless statistics is used only where the data is in the form of group.
- W. I. King states, "Statistics from their very nature of subject cannot and will never be able to take into account individual cases. When these are important other means must be used for their study".

1.4.2 Statistics does not Study Qualitative Phenomena

- It is applicable to quantitative data which can be expressed in numerical form. It does not provide the base to study qulitative data. Qualitative phenomena poverty, wisdom, intelligence, etc. cannot be expressed in numerical form. However it can be applied when qualitative data will be reduced to numerical form.

1.4.3 Statistical Laws are not Exact

- Probabilities and approximate values can be predicted by statistical method. It does not give exact nature of laws like scientific laws. There is always some uncertainty in predicted solution, but when it is averaged with certain details, uncertainty is reduced.

1.4.4 Statistical Results are True only on an Average

- It is always based on average values of any phenomena experiment or survey. It cannot produce a true factor for individual. It is in the form of average studied factor. It is not useful for substitution of any unit or event. Statistical average data when applied to an individual encounters error in observation. W. I. King states "Statistics largely deals with averages and these averages may be made up of individual items radically different from each other".

1.4.5 Statistical Relations do not Necessarily Bring Out the 'Cause and Effect' Relationship between Phenomena

- It is entirely depend on interpreters data. It reveals the association in relation of different quantities. It varies by the nature and type of quantities involved and depends on the view and expectation of interpreter.

1.5 Distrust of Statistics

- Many times applications and usefulness of statistics is ignored due to misunderstanding among some people. It may be due to their insufficient confidence in doing the proper analysis of data. It is often said that statistical analysis calculations and results are manipulated by the uses.

Causes of distrust :

1) Figures are always easily believed.
2) Statistics has also some limitations. These are ignored.
3) Figures are represented by manipulation. This is the misuse of figures.
4) Insufficient information or knowledge of the subject.

- Though the figures are accurate, some people manipulates the figures to hide practical data and try to represent expected data to achieve their selfish desires. The skilled speakers, writers force some wrong input in statistical data analysis which lead to lose faith of public in the subject.

- It should be rightly understand that statistics does not prove anything or disprove the same. It is only the tool to achieve some single conclusion, relation, dependence of quality, etc.
- According to bowley "Statistics only furnishes a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies".
- An example can be discussed here to illustrate the point. Medicines are used to cure the patient. If wrong medicine or excess dose is taken a person may die. We cannot blame the medicine for such a result.

1.6 Misinterpretation of Statistical Data

1) A survey report : Usually electronic gazettes are compared by customer on an online portal. Sometimes the opinions of customers if in favour of any particular company product, then it is retained. On the other side if the same data is believed to be true without comparison then it is wrong interpretation. It does not mean a product displayed is only the good.
2) The number of students taking the subject economics in institute increased twice as compared to last two years. Thus economics is most demanded subject among the students. This conclusion is faulty because the same time status of other subjects is not considered.

1.7 Population and Sample

- The word population is used in a wider sense. For example, In the study of industrial development all the industries under consideration is population. In titration experiment solution in breaker is a population. Thus population may be a group of employees, collection of books, a group of students, etc.

Definition :

- A set or group of observations relating to a phenomenon under statistical investigation called statistical population or population.
- The population is finite or infinite according to whether the set contains a finite or infinite number of observations.
- For any statistical analysis complete enumeration of the population is impracticable.
- If we want to have an idea of the average per capita (monthly) income of the people in India. We will have to enumerate all the earning individuals in the country which is a very difficult task.

- In order to study a group of large number of items we require to draw sample. We use technique of sampling several times in everyday life.

Example :

i) Purchasing food grains we inspect only handful of grains and draw conclusion about the whole sack.

ii) While examining blood of an individual few drops are enough for diagnosis.

Sampling is well accepted means of collecting information. Sampling plays very important role in statistical inference.

Sample : Definition : A finite subset of statistical individuals in a population is called a sample and number of individuals in a sample is called the sample size.

Types of sampling :

- Some of commonly known types of sampling are :

1) Purposive sampling
2) Random sampling
3) Simple sampling
4) Stratified sampling

1.7.1 Purposive Sampling

- Purposive sampling is one in which the sample units are selected with definite purpose in view. It is less costly and less time consuming.

Example :

1) To select candidates for debating competition, certainty deliberate selection of suitable candidates will be done. A survey may be carried out by investigator interested in opinions and views on certain specific issues.
2) Financial institutions can ensure representative character in a purposive selection of sick units financed by them.

1.7.2 Random Sampling

- In this case sample units are selected at random and the drawback of purposive sampling completely overcome. A random sample is one in which each unit of population has an equal chance of being included in it.
- Suppose we take a sample of size n from N size of finite population. Possibility of samples are ${}^{N}C_{n}$. A sampling technique in which each of ${}^{N}C_{n}$ samples has an equal

chance of being selected is known as random sampling and the technique is called as random sample.

- For example r candidates are selected from n candidates. Assigning 1 to n numbers to each candidate and make slips which are homogeneous and put all slips in a bag and thoroughly shuffled and then r slips are drawn one by one 'r' candidates corresponding to numbers on the slips drawn will constitute random sample.
- According to W. M. Harper "A random sample is one which is selected in such a way that every item in the population has an equal chance of being included".

1.7.3 Sample Sampling

- Under this sampling the whole population is taken as a single composite unit for purposes of sampling.
- This method is easiest and commonly used.
- If population consists of N elements then probability of selection of any element is $\frac{1}{N}$.

Examples :

1) To find diameter of a rod, we take reading at few points on a rod and then find the average of readings.
2) To conduct socio-economic survey of a certain village and find per capita income of a village.

- This method is widely used due to its simplicity and convenience. However, it suffers from some drawbacks such as, it may not be proper representative when population is heterogeneous, widely spread, etc.

1.7.4 Stratified Sampling

- If population is heterogeneous, simple sampling is not effective. Entire population is divided into many homogeneous groups called as strata (plural). The size of each stratum is same as a simple sample combining of all sampled observations formed stratified sampling.
- This method gives better results. This method finds suitability in administrative purposes.

Examples :

1) Estimation of annual income per family we divide population into groups such as families with yearly income below 20,000 between ₹ 20,000 to ₹ 50,000 between ₹ 50,000 to 1 Lakh and above.

2) To conduct health survey in a college we can use stratified sampling by considering strata as the facilities or classes or sex etc.

1.8 Parameter and Statistics

- A parameter is a descriptive measure of characteristic of the population. It's frequently possible to have a good idea of the population from a few descriptive measures. A measure which is calculated from all the compliances on the population is a parameter of the population. For illustration, if an average is calculated from a set of all observations on the population, the average will be a parameter of the population. The average is population average. A corresponding descriptive measure can be obtained from the observations contained in a sample of population. A descriptive measure computed from the observations in a sample is called a statistic. An average computed from observation in a sample is a statistics and is called sample average.
- In order to avoid verbal confusion with statistical constants of the population viz. mean (μ) and variance (σ^2) etc. which are usually referred to as parameters : Statistical measures calculated from the sample observations alone, e.g. mean ($\bar{X}$), variance (S^2) etc. referred as statistics.
- It may be noted that a parameter is constant for population the corresponding statistics may vary from sample to sample.

Remark :

1) μ and σ^2 will refer to the population mean and variance respectively while the sample mean and variance will be denoted by $\bar{X}$ and S^2 respectively.
2) Unbiased estimate : A statistic $t = t(x_1, x_2, \ldots, x_n)$ a function of the sample values $x_1, x_2, \ldots, x_n$ is an unbiased estimate of the population parameter F, if $E(t) = F$, i.e. E(statistic) = Parameter then statistic is said to be unbiased estimate of the parameter.

1.9 Sample Distribution of a Statistic

To draw a sample of size n from a given finite population of size N then total number of possible sample is,

$$^{N}C_{n} = \frac{N_1}{n!(N-n)!} = r(\text{say})$$

r samples we can compute some statistic $t = t(x_1, x_2, \ldots, x_n)$, in particular mean $\bar{x}$ and variance S^2 etc. given below

Sample number	Statistic t	$\bar{x}$	S^2
1	t_1	$\bar{x}_1$	S_1^2
2	t_2	$\bar{x}_2$	S_2^2
3	t_3	$\bar{x}_3$	S_3^2
:	:	:	:
r	t_r	$\bar{x}_r$	S_r^2

Mean and variance of the sampling distribution of the statistic t are given by,

$$\bar{t} = \frac{1}{r}(t_1 + t_2 + \ldots + t_r) = \frac{1}{r}\sum_{i=1}^{r} t_i$$

$$\text{var}(t) = \frac{1}{r}[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \ldots + (t_r - \bar{t})^2]$$

$$= \frac{1}{r}\sum_{i=1}^{r} (t_i - \bar{t})^2$$

1) **Standard error :** Standard deviation of sampling distribution of a statistics is known as its Standard Error (S.E.).

For large samples, where n is the sample size, σ^2 is population variance, p is population proportion and $q = 1 - p$, n_1 and n_2 represent sizes of two independent random samples reply drawn from given population (s).

Sr. No.	Statistics	Standard error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'P'	$\sqrt{pq/n}$
3.	Sample s.d : S	$\sqrt{\sigma^2/2n}$
4.	Sample variance : S^2	$\sigma^2\sqrt{2/n}$
5.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$
6.	Sample moment : μ_3	$\sigma^3\sqrt{96/n}$
7.	Sample moment : μ_4	$\sigma^4\sqrt{96/n}$

8.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
9.	Difference of two sample s.d : $(S_1 - S_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$

1.10 Statistical Inference

- It is method to draw the conclusion or what do we summarize from given analysis of data. This is the main objective of the statistics. There are two important problems in statistical inference - 1) Estimation 2) Testing of hypothesis.

1.10.1 Methods of Estimation

- For obtaining estimation commonly used methods are as follows :

1) Method of maximum likelihood estimation : This is the most general method of estimation.
2) Method of minimum variance.
3) Method of moments.
4) Method of least squares.
5) Method of minimum chi-square.
6) Method of inverse probability.

1.10.2 Testing of Hypothesis

- A test of a statistical hypothesis is a two action decision problem after the experimental sample values have been obtained, the two actions being the acceptance or rejection of the hypothesis under consideration.

Null hypothesis :

- A statistician should work without thinking of profit or loss which affects the acceptance of calculations. It should not be partial and should not influence the decision at any cost.

Example :

A writing pen is provided by two methods of manufacturing. There are different views like old method is more reliable or new one. It provides comparative observation of two methods. The zero difference method i.e. both methods are sufficient are accepted.

Alternative hypothesis :

- The acceptance of null hypothesis depends on any other testing method. Acceptance or rejection depends on the test against rival hypothesis.

1.11 Sampling with and without Replacement

- In statistical study of any data, samples are collected to reach to a single or average conclusion of any process. This population contains two types of sampling units.

  1) A finite number of sampling unit. 2) An infinite number of sampling units.

- But mostly population contains to found a finite number of sampling units. Sampling can be done by with replacement or without replacement. When the same unit of the population involves in each sample more than once, then it is sampling with replacement. In sampling without replacement same unit of the population may not be included in each sample more than once.
  - If the size of population is N and sample size in n then
  - If sampling is done with replacement, the total number of possible samples will amount to N^n.
  - If sampling is done without replacement the total number of possible samples will be ${}^{N}C_{n}$.
- In case of population of finite size sampling without replacement results in ${}^{N}C_{n}$ distinct ways where N is size of population and n is total number of possible samples.

Example 1.11.1 *If the population size is 5 and the sample size is 2. Find the number of possible samples 1) With replacement. 2) Without replacement.*

Solution :

1) Population size = N = 5

Sample size = n = 2

Samples = $N^n = 5^2 = 25$

Possible samples with replacement = 25

2) Population size = N = 5

Sample size = n = 2

Samples = ${}^{N}C_{n} = {}^{5}C_{2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2!3!} = \frac{4 \times 5}{1 \times 2} = 10$

Possible samples without replacement = 10

1.12 Random Sample

- Random samples are characterized by the way in which they are selected.
- Random sample is one where each item in the universe has an equal or known opportunity of being selected.
- According to C. H. Mayer "A sample is said to be random when each unit drawn has a probability identical to the probability of all the other units which might have been drawn in its place".

1) Sample is selected to study the concerned population. Sample is a miniature of population.
2) Sampling units should be independent.
3) Samples can be achieved by dividing population in homogeneous subgroups and selecting a random sample from each sub-groups.

- In random selection equal chance of selection is given to each sampling unit. It can be achieved by two ways.

1.12.1 Lottery Method

- In this method n elements are drawn by using N different slips which are made of the same size and shape. All slips are mixed properly. With the help of handle slips are thoroughly mixed and n elements of slips are drawn one by one.

1.12.2 Random Numbers

- A sample is selected for smooth operation purpose by use of random numbers. This can be done by using availble random tables. These tables are prepared by using the numbers from 0 to 9 with equal chance of finding or equal frequency of its occurence. Two digit numbers are selecting by formation of two rows and two columns. We form the numbers from 00 to 99. For 3 digit number 3 rows and 3 columns are formed to get the numbers from 000 to 999.
- For selection of a random sample any part of random numbers selected and numbers are chosen serially in row or column. If the number selected in this manner is in between 1 to N, the corresponding element is taken in the sample. Hence n elements are chosen.

Tables of random numbers :

1) **Tippet's random number tables :** This table consist of 10400 four digited numbers, gives in all 10400 × 4 i.e. 41600 digits. These tables proved to be fairly random in character.

2) **Fisher and yates tables :** Statistical tables for biological, agricultural and medical research. It is obtained by drawing numbers at random from the 10^{th} to 19^{th} digits of A.S.

3) **Kendall and babington smith table :** It consist of 1,00,000 digits grouped into 25000 sets of 4-digit random numbers.

4) **Rand corporation random table :** It consist of one million random digits consisting of 2,00,000 random numbers of 5 digit each.

1.13 Population Parameter

- It is a quantity or statistical measure that, for a given population is fixed and that is used as the value of a variable is same in general distribution or frequency function to make it descriptive of that population.
- These are not similar to parameters in mathematics which refers the constant value for a mathematical function.
- Parameters refer to the whole population whereas statistics refers a part of that population.

1.14 Sample Statistic

- Sample statistic is a piece of statistical information from the items under consideration and to be analyzed for the purpose of its characteristics.
- It is a fraction of data or a part of population which is small in comparison with population.
- Sample is a piece under the study taken from population which is on large scale.

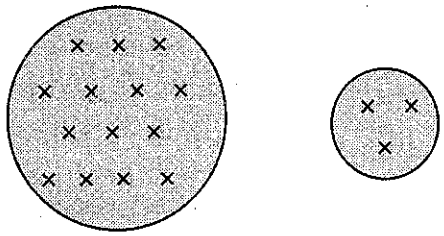


Fig. 1.14.1

1.15 Sampling Distribution

- Each and every unit of data collected by any means called as population is not considered for extracting the information. In short the whole population is not analyzed with its individual constituent elements.

- A few units selected from among the total lot in accordance with some specified procedure constitutes a sample.
- The various values assumed by statistics distributed which are computed from various samples of same size and drawn randomly from the population is called sampling distribution.
- Population of constant nature contains a large number of samples of same size. Each of these sample may contain different population members though it has same size. Each sample goes under treatment of statistics like mean, standard deviation. Variance etc. and series of such a sample are brought together at one place.
- The various values produced by this kind of method may be arranged as a frequency distribution which is known as sampling distribution.
- Sampling is done with replacement and without replacement.
- The samples are analyzed by taking their mean first. A mean of sample will be different because it has been derived from different units of same sample. The sample mean taken as random variable. Each sample mean will then constitute an observed value of this new random variable. It is given by a new set $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k$. These mean values may be used to form a frequency distribution. The frequency distribution of the statistic $\bar{X}$ is known as the sampling distribution of sample mean.

Exercise

1. *What is statistics ? Elaborate on the statistics with reference to business and economics.*
2. *Explain the scope of statistics in engineering and technology.*
3. *Give a brief note on the statistics in medical or biological fields.*
4. *Define statistics. Give importance of statistics. (any 5)*
5. *What are the limitations of statistics ? Explain one limitation in brief.*
6. *What are the reasons of distrust of statistics ? Give suitable example.*
7. *How statistical data can be misinterpreted ? Explain it with suitable examples.*
8. *What is population and sample ? Give difference between them.*
9. *Define sampling. Give its types. Explain any one.*

10. *Discuss the following in brief : 1) Purposive sampling, 2) Random sampling, 3) Simple sampling, 4) Stratified sampling.*
11. *What is parameter of population ? Discuss it in brief.*
12. *Give method of sample distribution of a statistics.*
13. *Mention methods of estimation. Discuss testing of hypothesis in short.*
14. *Give comparison between sampling with and without replacement.*
15. *What is random sample ? Explain lottery method and random numbers.*
16. *Explain random numbers and mention tables of random numbers.*
17. *What are the advantages and disadvantages of statistical analysis ?*

Multiple Choice Questions

Q.1 A pioneer of 'Corelation Analysis', Karl Pearson gave first and more significant test in statistics known as _____.

[a] Chi-square test [b] Bi-square test

[c] Analytic test [d] None

Q.2 Statistics involves which of the following process related to data _____.

[a] collection of data [b] description of data

[c] analysis and interpretation of data [d] all of the above

Q.3 The role of statistics in engineering and technology referred to as _____.

[a] to design [b] modeling of product

[c] manufacturing of product [d] all of the above

Q.4 In survey of a data some uncertainty is observed. The uncertainty can be reduced by _____.

[a] trignometric calculation [b] estimation of data

[c] algebraic calculation [d] none

Q.5 Statistical analysis requires a _____ data for analysis.

[a] individual data [b] single parameter data

[c] form of a group [d] very few no. of data

Q.6 Which of the following statement is wrong _____.

[a] statistics studies individual data [b] statistics refers qualitative phonomena

[c] statistical laws are exact [d] all of the above

Q.7 The cause of distrust of statistics is due to _____.

[a] collection of data [b] sampling of data

[c] interpretation of data [d] manipulation in collected data

Q.8 The number of individual in a sample is called _____.

[a] population [b] sample

[c] sample size [d] none

Q.9 Which of the following is not a type of sampling _____.

[a] purposive sampling [b] random sampling

[c] continuity in sampling [d] stratified sampling

Q.10 Each unit of population has an equal chance of being included in sample is the example of _____.

[a] purposive sampling [b] random sampling

[c] simple sampling [d] stratified sampling

Q.11 For a given data, a sample of size 'r' from a given finite population of size N, then total number of possible sample is _____.

[a] $^{N}C_{r} = \frac{N!}{r!(N-r)!}$ [b] $^{r}C_{N} = \frac{r!}{r\,(r-N)!}$

[c] $^{N}C_{r} = 1$ [d] none

Q.12 In measurement of height of a wall, following readings are taken 5.26, 5.28, 5.32, 5.24 and 5.31 respectively. In statistical analysis the mean and variance of data will be _____.

[a] 5.28 and 9×10^{-4} [b] 4.26 and 9×10^{-3}

[c] 3.21 and 9×10^{-2} [d] 2.26 and 9

Q.13 In sampling with replacement the total number of possible samples if the size of population is N and sample size is n will be _____.

[a] N^n [b] n^N

[c] N^{n-1} [d] n^{N-1}

Q.14 Formula for standard error of sample mean $\bar{x}$ based on sample of size n having variances σ^2 when population consisted of N items is _____.

[a] σ/n [b] $\sigma/\sqrt{n-1}$

[c] $\sigma/\sqrt{N-1}$ [d] $\frac{\sigma}{\sqrt{n}}$

Q.15 Which of the following statement is not true ?

[a] Standard error cannot be zero [b] Standard error cannot be 1

[c] Standard error can be negative [d] All the above

Q.16 If the sample values are 1, 3, 5, 7, 9 the standard error of sample mean is : _____.

[a] $S.E = \sqrt{2}$ [b] $S.E = \frac{1}{\sqrt{2}}$

[c] $S.E = 2.0$ [d] $S.E = \frac{1}{2}$

Q.17 If we have a sample of size n from a population of N units, the finite population correction is _____.

[a] $\frac{N-1}{N}$ [b] $\frac{n-1}{N}$

[c] $\frac{N-n}{N}$ [d] $\frac{N-n}{n}$

Q.18 If n units are selected in a sample from N population units, the sampling fraction is given as : _____.

[a] $\frac{N}{n}$ [b] $\frac{1}{N}$

[c] $\frac{1}{n}$ [d] $\frac{n}{N}$

Q.19 If sampling is done without replacement the total number of possible sample will be _____. If size of population is N and sampling size is n.

[a] ${}^{N}C_{n}$ [b] ${}^{n}C_{N}$

[c] ${}^{N-1}C_{n}$ [d] ${}^{n-1}C_{N}$

Q.20 If population size is 5 and the sample size is 2. The number of possible samples with replacement will be _____.

[a] 25 [b] 16

[c] 125 [d] 625

Q.21 If the sum of N observations is 630 and their mean is 42, then value of N is :

[a] 21 [b] 30

[c] 15 [d] 20

Q.22 If population size is 10 and the sample size is 4. The number of possible samples without replacement is _____.

[a] 210 [b] 215

[c] 110 [d] 115

Q.23 The method of sampling in which n elements are drawn by using N different slips made up of same size and shape is _____.

[a] lottery method [b] random numbers

[c] both (a) and (b) [d] can't say

Q.24 Which of the following is not a table of random numbers

[a] Tippet's Random Number Table [b] Fisher and Yates Table

[c] Newton's Number Table [d] Kendall and Bobington Smith Table

Q.25 The class interval of the continuous grouped data 10 - 19, 20 - 29, 30 - 39, 40 - 49, 50 - 59 is _____.

[a] 10 [b] 9

[c] 14.5 [d] 4.5

Q.26 Chi-square distribution is used for the test of : _____.

[a] goodness of fit [b] hypothetical value of population variance

[c] both (a) and (b) [d] neither (a) nor (b)

Q.27 Stratified sampling comes under the category of _____.

[a] unrestricted sampling [b] subjective sampling

[c] restricted sampling [d] purposive sampling

Q.28 If the observations recorded on five sampled items are 3, 4, 5, 6, 7 the sample variance is _____.

[a] 2 [b] 1

[c] 0 [d] 2.5

Q.29 Stratified sampling belongs to the category of _____.

[a] judgement sampling [b] subjective sampling

[c] controlled sampling [d] non-random sampling

Q.30 The magnitude of the standard error of an estimate is an index of its :

[a] accuracy [b] precision

[c] efficiency [d] all the above

Answer Keys for Multiple Choice Questions :

Q.1	a	Q.2	d	Q.3	d	Q.4	b	Q.5	c
Q.6	d	Q.7	d	Q.8	c	Q.9	c	Q.10	b
Q.11	a	Q.12	a	Q.13	a	Q.14	d	Q.15	d
Q.16	a	Q.17	c	Q.18	d	Q.19	a	Q.20	a
Q.21	c	Q.22	a	Q.23	a	Q.24	c	Q.25	a
Q.26	c	Q.27	c	Q.28	a	Q.29	c	Q.30	b

❑❑❑

Unit II

2 Descriptive Statistics : Measures of Central Tendency

Syllabus

Frequency Distributions and Measures of central Tendency : *Frequency Distribution, Continuous Frequency Distribution, Graphic Representation of a Frequency Distribution, Histogram, Frequency Polygon, Averages or Measures of Central Tendency or Measures of Location, Requisites for an Ideal Measure of Central Tendency, Arithmetic Mean, Properties of Arithmetic Mean, Merits and Demerits of Arithmetic Mean, Weighted Mean, Median, Merits and Demerits of Median, Mode, Merits and Demerits of Mode, Geometric Mean, Merits and Demerits of Geometric Mean, Harmonic Mean, Merits and Demerits of Harmonic Mean, Selection of an Average.*

Contents

2.1 Introduction

- In statistical exploration, once data is collected and edited... "The first task of the statistician is to organize of the figures in such a type that their significance for the aim in hand could also be appreciated, that comparison with masses of similar data could also be facilitated and that further analysis could also be possible.
- This is done through classification and tabulation.
- Classification is necessary and always precedes tabulation but after the determination of class categories the mode of presentation may take any form.

2.2 Classification

- Classification refers to grouping of data into homogeneous classes and categories.
- "Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts".
- Classification is the process of arrangement into groups or classes. A group or a class category has to be determined on the basis of the nature of data and purpose for which it is going to be used.
- For example : Income of 3000 individuals is given for analysis.
- It becomes essential to condense the data in suitable form classification that can be used as a tool.
- Entire process of making homogeneous and non-overlapping groups of observation according to similarities is called **classification.** The groups formed are called **classes** or **class intervals.**
- The objectives of classification can be summarized as follows :
  1) To condense the data.
  2) To prepare the data for tabulation.
  3) To provide ease in comparison with other data.
  4) To avoid unwanted details.
  5) To disclose leading characteristics of the data.
  6) To unlock further analysis like computation of averages, dispersion, etc.

2.3 Frequency Distribution

- A table containing class intervals with frequencies is called **frequency distribution,** frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude either individually or in groups. With their

corresponding frequencies side by side.

- To study how the observations are classified and a frequency distribution formed. Initially discrete variables will be considered for classification.

2.3.1 Frequency Distribution of Discrete Variable Procedure

1) Find smallest and largest observations.
2) Prepare a column of all possible values of variables from smallest to largest.
3) In the next column put a tally mark against values to which it relates.
4) Count the number of tally marks and place them in the next column in front of the corresponding value.

Example 2.3.1 *Row data for the survey of 100 families for studying the number of children in family are given as -*

3	2	5	2	3	1	2	4	6	4
3	1	2	2	4	1	1	3	2	3
2	4	3	6	0	2	5	3	4	2
3	3	3	2	5	0	4	3	4	3
4	3	2	1	2	3	2	2	3	4
2	2	2	3	4	2	3	8	5	1
0	2	4	0	3	3	4	2	4	6
2	4	2	5	2	3	4	3	3	3
4	6	3	1	0	5	4	3	3	3
3	2	3	1	5	0	2	3	4	3

Prepare a frequency distribution of the above data.

Solution : By observation we can locate the largest value as 8 and smallest as 0. We prepare the first column with entries from 0 to 8. We consider the first observation, it is 2, so we mark it as a vertical bar (|) called a **tally mark** and put a tally mark in the next column. If observations repeated three times we mark it as (|||) as a **tally mark** for fourth times repeated observation we mark it as (||||). But for fifth time repeated observations we mark as (/) on the earlier bunch of four tallies finally the structure looks like (𝍸). In this manner tally marks are arranged in a group of five. Finally, the number of tallies are counted and placed in the last column procedure given in Table 2.3.1.

Frequency distribution of number of the children in 100 families.

Number of children	Tally marks	Frequency
0	𝍸 |	6
1	𝍸 |||	8
2	𝍸 𝍸 𝍸 𝍸	25
3	𝍸 𝍸 𝍸 𝍸 𝍸 |	31
4	𝍸 𝍸 𝍸 |||	18
5	𝍸 ||	7
6	||||	4
7		0
8	|	1
Total		100

Table 2.3.1

Example 2.3.2 *Marks secured by 50 students in mathematics are given below :*

62	45	46	67	23	53	58	38	26	43
36	63	26	48	76	45	66	72	28	84
56	38	56	92	34	42	38	68	48	52
38	57	65	46	38	46	65	56	48	75
56	84	66	68	58	76	46	76	52	54

Prepare the frequency distribution table. Frequency distribution of marks secured by 50 students.

Solution :

Marks obtained	Tally marks	Frequency
23	|	1
26	||	2
28	|	1
34	|	1
36	|	1
38	𝍸	5

42	|	1
43	|	1
45	||	2
46	||||	4
48	|||	3
52	||	2
53	|	1
54	|	1
56	||||	4
57	|	1
58	||	2
62	|	1
63	|	1
65	||	2
66	||	2
67	|	1
68	||	2
72	|	1
75	|	1
76	|||	3
84	||	2
92	|	1
Total		50

2.3.2 Frequency Distribution of Continuous Variable

In this distribution the frequencies refer to groups of values.

The procedure of classification of continuous frequency distribution -

Procedure :

1) Find smallest and largest observations. Calculate the difference between them. Difference is called **'range'.**

2) Prepare a column of class intervals.

3) Classify observations one by one in appropriate class by putting tally marks in the second column.

4) Count tally marks and enter in the last column.

Example 2.3.3 *Height of 50 students to the nearest cm are given as below :*

151	147	145	153	156	152	159
153	157	152	144	151	157	147
150	157	153	151	149	147	151
147	155	156	151	158	149	147
153	152	149	149	153	150	152
154	150	152	151	151	151	154
155	152	154	152	156	155	154
150						

Construct a frequency distribution table.

Solution : In this problem the highest and lowest observations are 159 and 145 respectively. So we form classes as : 145-146, 147-148, 149-150, 151-152, 153-154, 155-156, 157-158, 159-160 and construct a table as - frequency distribution of heights of 50 students.

Class interval (Height in cm)	Tally marks	Frequency (No. of students having height)
145-146	II	2
147-148	𝍸	5
149-150	𝍸 III	8
151-152	𝍸 𝍸 𝍸	15
153-154	𝍸 IIII	9
155-156	𝍸 I	6
157-158	IIII	4
159-160	I	1
Total		50

Example 2.3.4 *Number of tools produced by 50 workers in a factory are given below :*

43	18	25	18	39	44	19	20	20	26
40	45	38	25	13	14	27	41	42	17
34	31	32	27	33	37	25	26	32	25
33	34	35	46	29	24	31	34	35	24
28	30	41	32	29	28	30	31	30	31

Construct a frequency distribution table.

Solution : Here the smallest observation is 13 and the largest observation is 45. So the range is 45 – 13 = 32. Since there are 50 observations we make classes each of size 5. Classes are from 13 – 17, 18 – 22, 23 – 27, 28 – 32, 33 – 37, 38 – 42, 43 – 47.

Frequency distribution of number of tools produced by 50 workers.

Class interval	Tally marks	Number of tools produced
13-17	III	3
18-22	𝍸	5
23-27	𝍸 𝍸	10
28-32	𝍸 𝍸 IIII	14
33-37	𝍸 III	8
38-42	𝍸 I	6
43-47	IIII	4
Total		50

2.4 Graphic Representation of a Frequency Distribution

- It is frequently useful to represent a frequency distribution via diagram which makes the unwieldy data and conveys to the eye the general run of the observations. It represents the comparison of two or more frequency distributions.

- There are two ways of graphical representations of frequency distributions are as follows :

  1) Histogram 2) Frequency polygon

2.4.1 Histogram

- It is one of the popularly used graphs for the representation of frequency distribution. Histogram distribution is drawn as follows :
  1) The class intervals are marked by taking a suitable scale along the X-axis.
  2) The rectangles with height depending on the frequency of corresponding class intervals are obtained such as to get the area of the rectangle proportioned to the frequency of the class.
  3) For unequal width the height of the rectangle depends on the ratio of frequencies to the width of the classes.
- The constituted construction of continuous rectangles called **histogram.**

Remark :

1) The histogram of an ungrouped frequency distribution of a variable is constructed by assuming the frequency with value of variable x between the interval $x - \frac{h}{2}$ to $x + \frac{h}{2}$ where h is the width.
2) In case of discontinuous grouped frequency distribution, it is converted into continuous distribution to construct the histogram.
3) There is no proportionality between the fractions of height and frequency of class. Hence histogram does not help to read the frequency over a fraction of class interval.
4) Histogram of the distribution of marks 100 students is obtained as follows :

Marks	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64
No. of students	9	18	12	15	21	05	10	7	2	1

- Since grouped frequency distribution is not continuous, we first convert it into a continuous distribution with exclusive type classes as given below :

Marks	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5	34.5-39.5	39.5-44.5
No. of students	9	18	12	15	21	5
Marks	**44.5-49.5**	**49.5-54.5**	**54.5-59.5**	**59.5-64.5**		
No. of students	10	7	2	1		

Histogram for frequency distribution

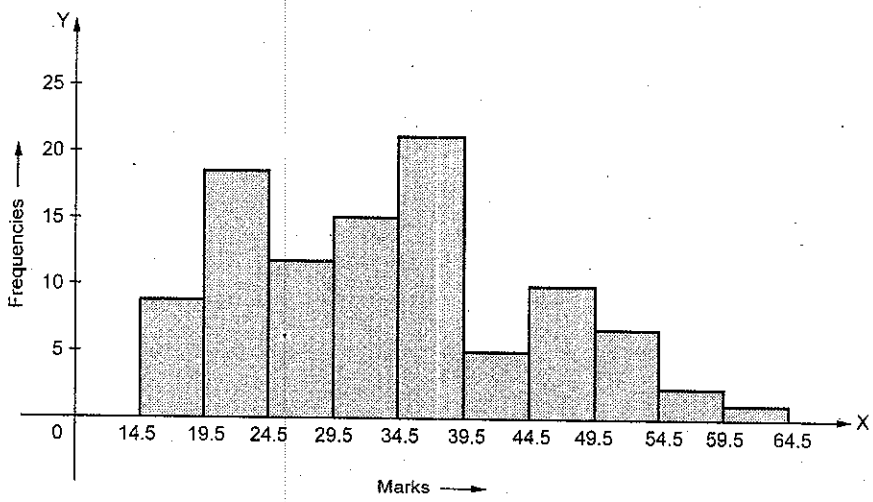


Fig. 2.4.1 Histogram

Note : Upper and lower class limits of the new exclusive types classes are known as class boundaries.

- Itch is gap between upper limit of any class and lower limit of the succeeding class, class boundaries for any class are given by,

$$\text{Upper class boundary} = \text{Upper class limit} + \frac{h}{2}$$

$$\text{Lower class boundary} = \text{Lower class limit} - \frac{h}{2}$$

Example 2.4.1 *Draw a histogram to represent the following frequency distribution.*

Marks obtained	0-10	10-20	20-30	30-40	40-50
No. of students	12	16	5	38	3

Solution :

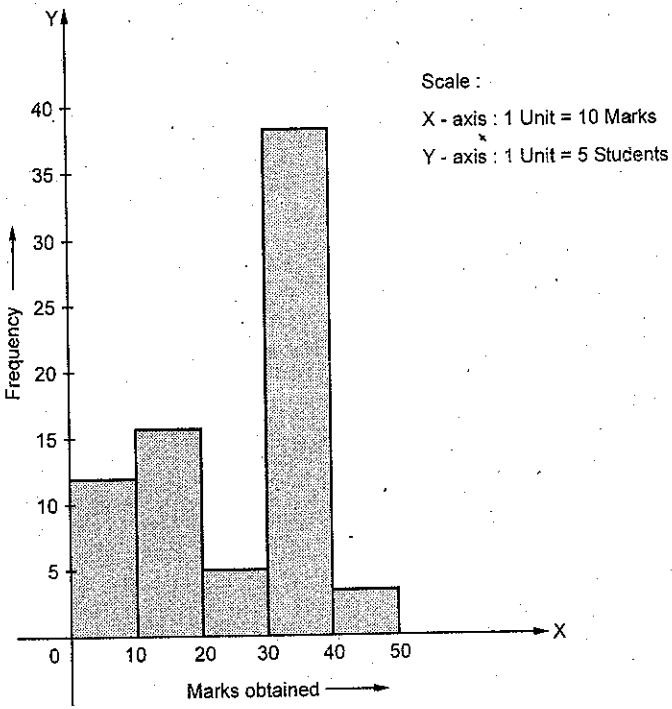


Fig. 2.4.2 Histogram

Example 2.4.2 *Draw a histogram to represent following frequency distribution -*

Age group (in years)	0-20	21-40	41-60	61-80	81-100	101-120
Population	500	2100	2200	2000	1600	400

Solution :

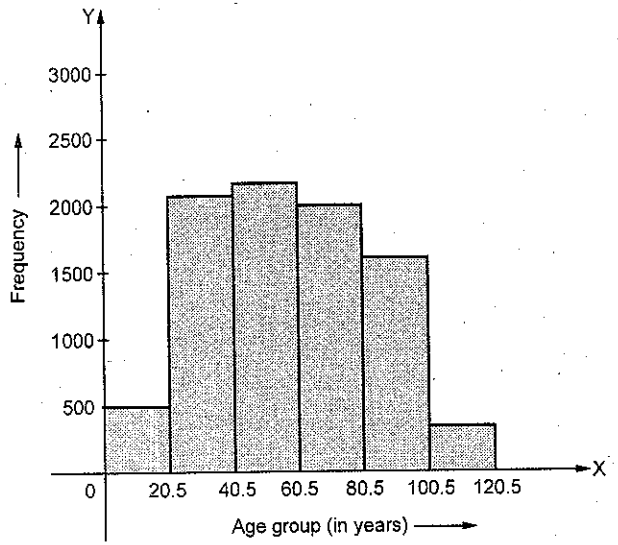


Fig. 2.4.3 Histogram

Example 2.4.3 *Draw a histogram for the following data -*

Age (in years)	2-5	5-11	11-12	12-14	14-15	15-16
No. of boys	6	6	2	5	1	3

Solution : Since classes are of unequal width, we need to compute frequency density.

Age (in years)	2-5	5-11	11-12	12-14	14-15	15-16
Frequency density	$\frac{6}{3}=2$	$\frac{6}{6}=1$	$\frac{2}{1}=2$	$\frac{5}{2}=2.5$	$\frac{1}{1}=1$	$\frac{3}{1}=3$

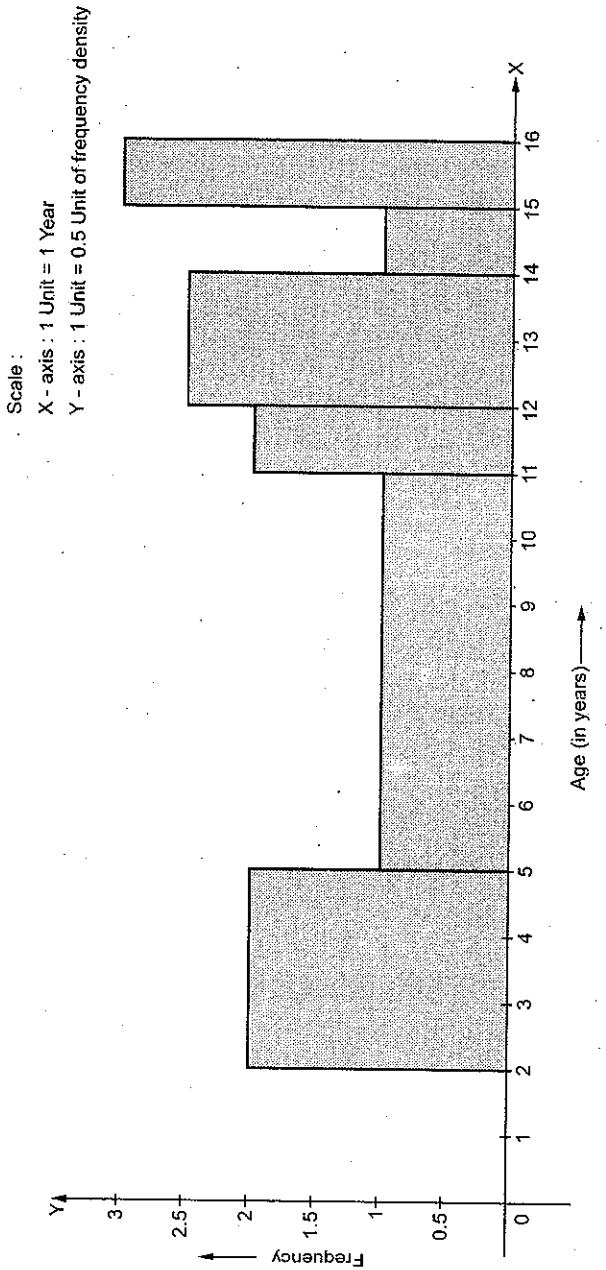


Fig. 2.4.4 Histogram

2.4.2 Frequency Polygon

- Graphical representation connects the variable values to corresponding base values. It must be a uniform and smooth curve. In histogram it does not satisfy these conditions. Therefore the alternate method of presentation of frequency distribution is frequency polygon. It provides clarity of understanding.
- Frequency polygon are constructed for two types of distribution,

  A) Ungrouped distribution B) Grouped distribution

A) Ungrouped distribution :

- Frequency polygon are constructed variate values along the X axis and corresponding frequencies along the Y axis. The nature of this plot is generally obtained by joining the points by a straight line.

B) Grouped distribution :

- In this case mid values of class intervals are along the X-axis. Polygon is obtained by joining the middle points by means of a straight line. Polygon can be drawn by a free hand curve for class intervals of small width.

Example 2.4.4 *Draw a frequency polygon for the following data :*

Monthly house rent	100-300	300-500	500-700	700-900	900-1100	1100-1300
No. of families	6	16	24	20	10	4

Solution : Along x-axis mid values of classes are taken and along y-axis frequency is taken. First point is (200, 6), second point will be (400, 16), third point will be (600, 24) and so on. Last point will be (1200, 4). To get a closed figure we take two more points (0, 0) and (1400, 0). Join all these points by line segments and we get frequency polygons. (Refer Fig. 2.4.5 on next page)

Example 2.4.5 *Draw a frequency polygon for following data :*

Marks obtained	0-10	10-20	20-30	30-40	40-50
No. of students	5	12	43	32	8

Solution : Mid values of classes are taken along X-axis and frequency along Y-axis. First point to plot is (5, 5), the second point will be (15, 12) and so on. Last point will be (45, 8). For a closed figure we take two more points (0, 0) and (50, 0). Join all points by line segments we and get frequency polygons. (Refer Fig. 2.4.6 on next page)

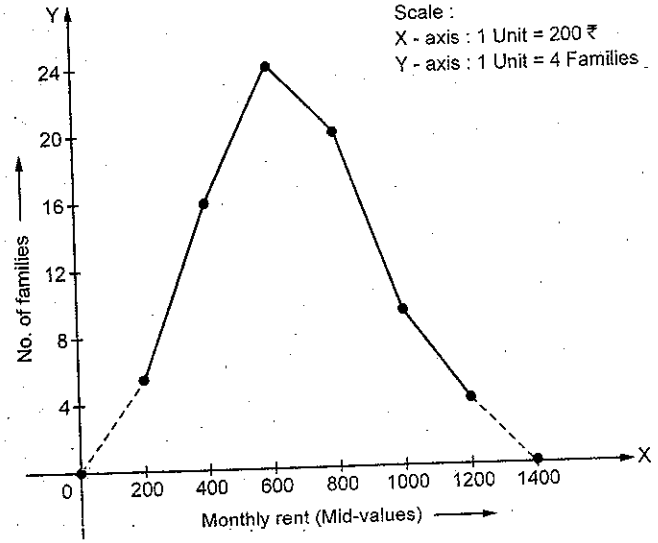


Fig. 2.4.5 Frequency polygon

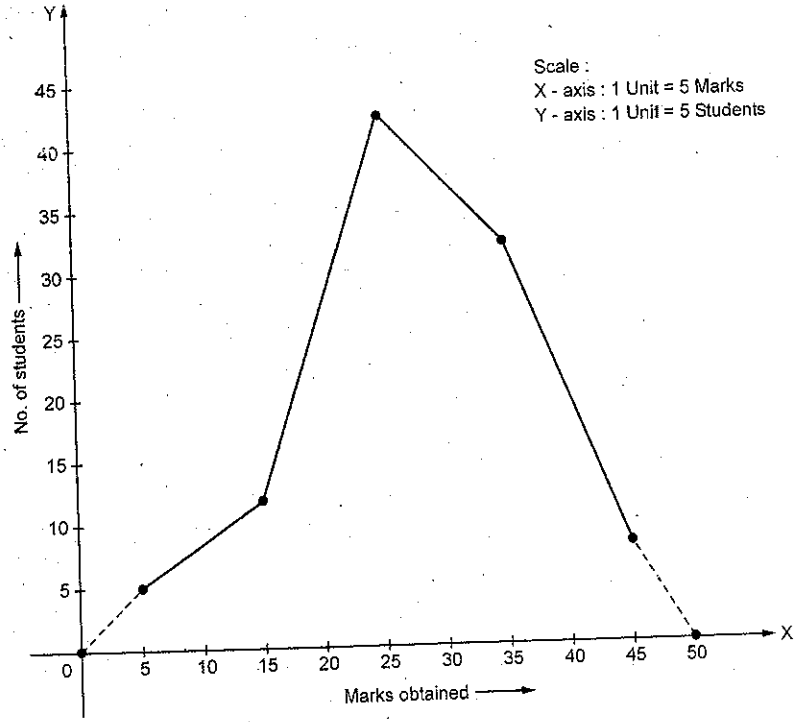


Fig. 2.4.6 Frequency polygon

2.5 Advantages and Limitations of Graphic Representation of Frequency Distribution

Advantages :

1) Graphical representation enables us to understand the resultant conclusion after analysis of data.
2) It is a collective presentation of information.
3) Graphs are more effective and impressive facts of tabulated values.
4) Everybody can read, analyse and understand the graph.
5) It is the understanding of data, concept for a longer time.
6) It helps to compare.

Limitations :

1) Exact results and values are not provided by the graph but it is in the form of approximate values.
2) Insufficient for analysis of statistical data but gives a general nature of the phenomenon.

2.6 Central Tendency

- On account of classification and frequency polygon we get an idea about the shape of frequency distribution. We observe that all class-frequencies are not the same. Initially the frequency is small in magnitude and then increases, it reaches maximum in the middle part and again falls down. We conclude that the observations are not uniformly spread. Most of the observations get clustered in the central part of data. This property of observations is called the **central tendency.**
- We select a representative observation from the central part. This is referred to as an average or measure of central tendency.

2.7 Average or Measure of Central Tendency

2.7.1 Requisites for an Ideal Measure of Central Tendency

- According to professor Yule, the following are the characteristics which are satisfied by an ideal measure of central tendency.
  1) It must include all the values of observations.
  2) It must be open to perform mathematical treatment repeatedly.
  3) There must be the least effect of extreme values on an average.

4) It must ensure the stability of sampling.

5) It must be defined rigidly.

6) It must be easy to calculate and simple to understand.

2.7.2 Types of Averages or Measures of Central Tendency

- Following are the measures of central tendency :
  1) Arithmetic mean
  2) Median
  3) Mode
  4) Geometric mean
  5) Harmonic mean.
- Among these arithmetic mean, geometric mean and harmonic mean are called **mathematical averages** and median mode are called **positional averages.**
- The type of average depends on the nature of data. Each method of averages has different advantages and disadvantages.

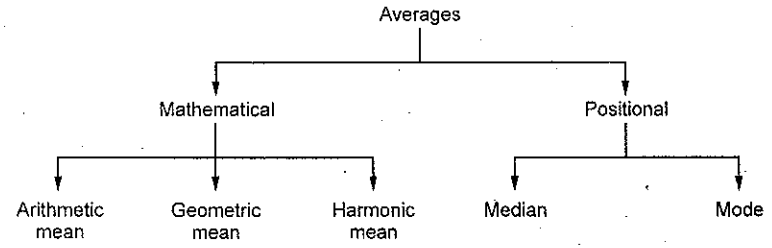


Fig. 2.7.1

2.8 Arithmetic Mean

- This is the method which is used in most computations.
- **Definition :** Sum of set of observations divided by the number of observations called as arithmetic mean

$$\text{A.M.} = \frac{\text{Sum of observations}}{\text{Number of observations}}$$

- There are different types of data analysis of A.M. different types are given as follows.

2.8.1 Row Data or Individual Observations

- Let $x_1, x_2, \ldots, x_n$ be n observations then arithmetic mean $\bar{X}$ is given by,

$$\text{A.M.} = \bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 2.8.1 *Calculate the arithmetic mean of marks scored by a student in 5 subjects given below : 45, 50, 40, 60, 55.*

Solution :

$$\text{A.M.} = \bar{X} = \frac{45 + 50 + 40 + 60 + 55}{5} = \mathbf{50}$$

Example 2.8.2 *Calculate arithmetic mean of weights of 10 students are 50, 46, 48, 51, 49, 52, 60, 32, 36, 42.*

Solution :

$$\text{A.M.} = \bar{X} = \frac{50 + 46 + 48 + 51 + 49 + 52 + 60 + 32 + 36 + 42}{10}$$

$$\bar{X} = \mathbf{46.6}$$

2.8.2 Ungrouped Data

- Let $x_1, x_2, \ldots, x_n$ be the observations and $f_1, f_2, \ldots, f_n$ be the frequencies then arithmetic mean is given by -

$$\text{A.M.} = \bar{X} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N} \qquad \left\{ \because \sum_{i=1}^{n} f_i = N \right\}$$

Example 2.8.3 *Calculate arithmetic mean of the frequency distribution*

x	1	2	3	4	5	6	7
f	5	9	12	17	14	10	6

Solution :

$$\text{A.M.} = \frac{5 \times 1 + 9 \times 2 + 12 \times 3 + 17 \times 4 + 14 \times 5 + 10 \times 6 + 6 \times 7}{5 + 9 + 12 + 17 + 14 + 10 + 6}$$

$$\text{A.M.} = \frac{299}{73} = \mathbf{4.09}$$

OR

x_i	f_i	$f_i x_i$
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Total	N = 73	299

$$\text{A.M.} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f_i x_i}{N}$$

$$= \frac{299}{73} = \mathbf{4.09}$$

Example 2.8.4 *Calculate arithmetic mean for the following frequency distribution :*

x	103	110	112	118	95
f	4	6	10	12	3

Solution :

x_i	f_i	$f_i x_i$
103	4	$103 \times 4 = 412$
110	6	$110 \times 6 = 660$
112	10	$112 \times 10 = 1120$
118	12	$118 \times 12 = 1416$
95	3	$95 \times 3 = 285$
Total	N = 35	$\Sigma f_i x_i = 3893$

$$\bar{X} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{3893}{35} = \mathbf{111.2286}$$

OR

- By using change of origin method we define u = x – a where a is any value and prepare the table :

x_i	$u_i = x_i - a$ $u_i = x_i - 100$	f_i	$f_i u_i$
103	3	4	12
110	10	6	60
112	12	10	120
118	18	12	216
95	– 5	3	– 16
Total		$\Sigma f_i = 35$	$\Sigma f_i u_i = 393$

$$\bar{u} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{393}{35} = 11.2286$$

As $u_i = x_i - 100$

$$\bar{u} = \bar{x} - 100$$

$$11.2286 = \bar{x} - 100$$

$$\bar{x} = 11.2286 + 100$$

$$\bar{x} = \mathbf{111.2286}$$

Example 2.8.5 *Calculate A.M. for the following frequency distribution :*

C.I.	0-8	8-16	16-24	24-32	32-40	40-48
f	8	7	16	24	15	7

Solution : For calculation of arithmetic mean, we find mid point of each class interval.

C.I.	Mid-values (x_i)	f_i	$f_i x_i$
0-8	4	8	$8 \times 4 = 32$
8-16	12	7	$7 \times 12 = 84$
16-24	20	16	$16 \times 20 = 320$
24-32	28	24	$24 \times 28 = 672$
32-40	36	15	$15 \times 36 = 540$

40-48	44	7	$7 \times 44 = 308$
Total		$\Sigma f_i = 77$	$\Sigma f_i x_i = 195.6$

$$\text{A.M.} = \bar{X} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{1956}{77} = \mathbf{25.4025}$$

Example 2.8.6 *Following is a distribution of weekly salaries of the employees of the firm :*

Salary (in ₹)	0-600	600-1200	1200-1800	1800-2400	2400-3000
No. of employees	3	7	14	27	23

Solution : For calculation of A.M., we calculate midpoint of each class interval. In most of the problems we get midpoints with equally spaced. We choose 'a' as possible closer to $\bar{X}$. Any midpoint in the central part can be taken as 'a'. We define $u_i = \frac{x_i - a}{h}$ where h is class width or any suitable number. From this it is observed that calculation of $\bar{X}$ seemed to be difficult than $\bar{u}$.

Class	**Mid-values (x_i)**	$u_i = \frac{x_i - 1500}{600}$	f_i	$f_i u_i$
0-600	300	– 2	3	– 6
600-1200	900	– 1	7	– 7
1200-1800	1500	0	14	0
1800-2400	2100	1	27	27
2400-3000	2700	2	23	46
Total	-	-	**74**	**60**

Here $$u_i = \frac{x_i - 1500}{600}$$

Therefore $$\bar{u} = \frac{\bar{X} - 1500}{600}$$

$$600\,\bar{u} = \bar{X} - 1500$$

$$\bar{X} = 600\,\bar{u} + 1500$$

As $$\bar{u} = \frac{\Sigma f_i u_i}{\Sigma f_i} = \frac{60}{74} = 0.81$$

$\therefore$ $$\bar{X} = 600 \times 0.81 + 1500$$

$$\bar{X} = 486 + 1500$$

$$\bar{X} = \mathbf{1986\ ₹}$$

2.8.3 Grouped Data

- In grouped or continuous frequency distribution arithmetic mean is given by,

$$\bar{X} = A + \frac{\Sigma fd}{N} \times h$$

where A = Assumed mean

d = Deviation $= \frac{x - A}{h}$

h = Class width

N = Σf

Example 2.8.7 *Calculation the mean for the following frequency distribution :*

C.I.	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	12	18	27	20	17	6

Solution : Here we take A = 35 and h = 10

C.I.	**Mid-values (x)**	**Frequency (f)**	$d = \frac{x - A}{h}$	**fd**
0-10	5	12	– 3	– 36
10-20	15	18	– 2	– 36
20-30	25	27	– 1	– 27
30-40	35	20	0	0
40-50	45	17	1	17
50-60	55	6	2	12
Total	-	**100**	-	**– 70**

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

$$= 35 + \frac{(-70)}{100} \times 10$$

$$= 35 - 7$$

$$= \mathbf{28}$$

Example 2.8.8 *Calculate A.M. for the following frequency distribution :*

C.I.	0-10	10-20	20-30	30-40	40-50
f	1	4	6	5	10

Solution : Here we take A = 25 and h = 10

C.I.	Mid-values (x_i)	Frequency (f_i)	$d = \frac{x - A}{h}$	fd
0-10	5	1	– 2	– 2
10-20	15	4	– 1	– 4
20-30	25	6	0	0
30-40	35	5	1	5
40-50	45	10	2	20
Total	-	**26**	-	**19**

$$\bar{X} = A + \frac{\Sigma fd}{N} \times h$$

$$= 25 + \frac{19}{26} \times 10$$

$$= 25 + 7.3076$$

$$= \mathbf{32.3076}$$

Example 2.8.9 *Find missing frequency from the following data :*

Marks	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Frequency	10	12	16	-	14	10	8

where the average mark is 16.82.

Solution : Here we consider missing frequency for the class 15-20 as f_4. We prepare a frequency distribution table.

Here we consider A = 17.5 and h = 5

C.I.	Mid-values (x_i)	Frequency (f_i)	$d = \frac{x - A}{h}$	fd
0-5	2.5	10	– 3	– 30
5-10	7.5	12	– 2	– 24
10-15	12.5	16	– 1	– 16
15-20	17.5	f_4	0	0
20-25	22.5	14	1	14
25-30	27.5	10	2	20
30-35	32.5	8	3	24
Total	-	**N = 70 + f_4**	-	**– 12**

As

$$\bar{X} = A + \frac{\Sigma fd}{N} \times h$$

$$16.82 = 17.5 + \frac{(-12)}{70 + f_4} \times 5$$

$$16.82 = \frac{17.5(70 + f_4) - 60}{70 + f_4}$$

$$16.82(70 + f_4) = 17.5(70 + f_4) - 60$$

$$1177.4 + 16.82f_4 = 1225 + 17.5f_4 - 60$$

$$1177.4 + 16.82f_4 = 1165 + 17.5f_4$$

$$1177.4 - 1165 = 17.5f_4 - 16.82f_4$$

$$12.4 = 0.68f_4$$

$$f_4 = \frac{12.4}{0.68} = 18.23$$

∴ Missing frequency = **18.23**

2.9 Properties of Arithmetic Mean

Example 2.9.1 *Sum of observations is equal to the product of arithmetic mean and number of observations i.e.* $\left(\sum_{i=1}^{n} x_i = n\bar{X}\right)$

Solution : Proof : Let $x_1, x_2, \ldots, x_n$ be n observations

By definition,

$$\bar{X} = \sum_{i=1}^{n} x_i$$

$$\therefore \quad n\bar{X} = \sum_{i=1}^{n} x_i$$

- **Note :** To calculate $\bar{X}$, n and $\sum_{i=1}^{n} x_i$ are sufficient instead of knowing individual observations.

Example 2.9.2 *Algebraic sum of deviations of observations from their arithmetic mean is zero i.e. $\Sigma(x_i - \bar{X}) = 0$*

Solution : Let $x_1, x_2, \ldots, x_n$ be n observations and deviations are $(x_1 - \bar{X}), (x_2 - \bar{X}), \ldots (x_n - \bar{X})$

$$\text{Sum of deviations} = \sum_{i=1}^{n} (x_i - \bar{X})$$

$$= \sum_{i=1}^{n} x_i - n\bar{X}$$

$$= n\bar{X} - n\bar{X} \qquad \ldots \{\text{Using property (1) } n\bar{X} = \Sigma x_i\}$$

$$= \mathbf{0}$$

Example 2.9.3 *Sum of squares of the deviations taken from arithmetic mean is minimum i.e. $\Sigma (x_i - \bar{X})^2 \leq \Sigma(x_i - a)^2$*

Solution : Let $Z = \Sigma(x_i - a)^2$ be the sum of the squares of the deviations of given values from any arbitrary point a.

- We have to prove Z is minimum when $a = \bar{X}$.
- By applying principle of maxima and minima, Z will be minimum for a if $\frac{\partial Z}{\partial a} = 0$ and $\frac{\partial^2 Z}{\partial a^2} > 0$

As $$Z = \Sigma(x_i - a)^2 \qquad \ldots(1)$$

$$\frac{\partial Z}{\partial a} = -2\Sigma(x_i - a) = 0 \qquad \ldots(2)$$

$$\Rightarrow \quad \Sigma(x_i - a) = 0$$

$$\Rightarrow \quad \Sigma x_i - \Sigma a = 0$$

$$\Rightarrow \quad n\bar{X} - na = 0 \qquad \ldots \{\text{Using property (1) } \Sigma x_i = n\bar{X}\}$$

$$\Rightarrow \quad \bar{X} = a$$

Again $$\frac{\partial^2 Z}{\partial a^2} = -2\Sigma(-1) \qquad \ldots(3)$$

Therefore, $$\frac{\partial Z}{\partial a} = 0 \text{ at } a = \bar{X} \text{ and}$$

$$\frac{\partial^2 Z}{\partial a^2} > 0 \text{ at } a = \bar{X}$$

Hence z is minimum at $a = \bar{X}$

$$= 2n > 0$$

Therefore, $\Sigma(x_i - \bar{X})^2 \leq \Sigma(x_i - a)^2$

OR

- Let 'a' be any arbitrary constant.
- Then $x_i - \bar{X}$ is deviation of x_i from $\bar{X}$
- $x_i - a$ is deviation of x_i from a

$$\Sigma(x_i - a)^2 = \Sigma(x_i - \bar{X} + \bar{X} - a)^2$$

$$= \Sigma[(x_i - \bar{X} + \bar{X} - a)]^2$$

$$= \Sigma[(x_i - \bar{X})^2 + 2(x_i - \bar{X})(\bar{X} - a) + (\bar{X} - a)^2]$$

$$= \Sigma(x_i - \bar{X})^2 + 2\Sigma(x_i - a)(x_i - \bar{X}) + \Sigma(\bar{X} - a)^2$$

$$= \Sigma(x_i - \bar{X})^2 + 0 + n(\bar{X} - a)^2 \qquad \ldots \{\because \Sigma(x_i - \bar{X}) = 0\}$$

$$= \Sigma(x_i - \bar{X})^2 + n(\bar{X} - a)^2$$

$$= \Sigma(x_i - \bar{X})^2 + \text{Non-negative quantity}$$

$$\therefore \quad \Sigma(x_i - a)^2 \geq \Sigma(x_i - \bar{X})^2$$

Example 2.9.4 *(Mean of combined groups) If $\bar{x}_1$ be the arithmetic mean of first group of size n_1 and $\bar{x}_2$ be the arithmetic mean of second group of size n_2 then,*

$$\text{combines mean} = \bar{x}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

Solution :

Let $x_{11}, x_{12}, \ldots, x_{1n_1}$ be n_1 members of the first series.

$x_{21}, x_{22}, \ldots, x_2n_2$ be n_2 members of the second series.

Then by definition,

$$\bar{x}_1 = \frac{x_{11} + x_{12} + \ldots + x_1 n_1}{n_1}$$

$$\bar{x}_1 = \frac{\text{Sum of observations in first group}}{n_1}$$

$$\bar{x}_2 = \frac{x_{21} + x_{22} + \ldots + x_{n2}}{n_2}$$

$$\bar{x}_2 = \frac{\text{Sum of observations in second group}}{n_2}$$

$\therefore$ Sum of observations in first group $= n_1\bar{x}_1$

Sum of observations in second group $= n_2\bar{x}_2$

$$\bar{x}_c = \frac{\text{(Sum of observations in first group) + (Sum of observations in second group)}}{\text{(Size of first group) + (Size of second group)}}$$

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Remark : Above result can be generalised to k ($k \geq 2$) groups as follows :

Let k groups of size of i^{th} group as n_i and the arithmetic mean as $\bar{x}_i (i = 1,2,3,\ldots,k)$. Then $\bar{x}_c$ is the arithmetic mean of all k groups combined together is given by,

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

Example 2.9.5 *Average weight of 70 boys is 55 kg and the average weight of 60 girls is 50 kg. Calculate the mean of a combined group of boys and girls.*

Solution : Let $\bar{x}_1$ be the average of boys with n_1 be the size of the group of boys.

Let $\bar{x}_2$ be the average of girls with n_2 be the size of the group of girls.

$\therefore$ $n_1 = 70$ $\quad \bar{x}_1 = 55$

$n_2 = 60$ $\quad \bar{x}_2 = 50$

$$\text{Combined mean} = \bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{70 \times 55 + 60 \times 50}{70 + 60}$$

$$= \frac{6850}{130} = \mathbf{52.6923}$$

Example 2.9.6 *Arithmetic mean of 40 items is 100. While checking, it was noticed that observation 76 was misread as 67. Find the correct value of mean.*

Solution :

$$\text{Incorrect mean} = 100 = \frac{\text{Incorrect sum}}{n}$$

$$\text{Incorrect sum} = 100 \times 40 = 4000$$

$$\text{Correct sum} = \text{Incorrect sum} + \text{Correct observation} - \text{Incorrect observation}$$

$$= 4000 + 76 - 67 = 4009$$

$$\text{Correct mean} = \frac{\text{Correct sum}}{n} = \frac{4009}{40} = \mathbf{100.225}$$

Example 2.9.7 *Average salary of male employees in a firm was ₹ 2500 and that of females was ₹ 2000. The mean salary of all employees was 2200. Find the percentage of male and female employees.*

Solution : Let n_1 and n_2 be the number of male and female employees in the firm respectively.

And $\bar{x}_1$, $\bar{x}_2$ be the average salary of male and female employees respectively.

Let $\bar{x}$ be the average salary of all workers in the firm.

Given :

$$\bar{x}_1 = 2500, \quad \bar{x}_2 = 2000, \quad \bar{x} = 2200$$

As $$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$2200 = \frac{n_1 \times 2500 + n_2 \times 2000}{n_1 + n_2}$$

$$2200\,(n_1 + n_2) = 2500n_1 + 2000n_2)$$

$$2200n_1 + 2200n_2 = 2500n_1 + 2000n_2$$

$$2200n_2 - 2000n_2 = 2500n_1 - 2200n_1$$

$$\frac{n_1}{n_2} = \frac{200}{300}$$

$$\frac{n_1}{n_2} = \frac{2}{3}$$

$\therefore$ Percentage of male employees in the firm,

$$= \frac{2}{2+3} \times 100 = \frac{2}{5} \times 100 = 40$$

and percentage of female employees in the firm

$$= \frac{3}{2+3} \times 100 = \frac{3}{5} \times 100 = \mathbf{60}$$

2.10 Merits and Demerits of Arithmetic Mean

Arithmetic means satisfies all the needs of a good average. Hence most commonly it is used in practice. Merits and demerits of arithmetic mean are as follows :

Merits :

1) It is a very simple and easy tool. It provides easy calculation.
2) Each and every reading is covered in this method.
3) It provides stability of sampling.
4) It is easy to understand at any level of analysis.
5) It is not limited to individual observation but also covers the mean of a group of observations.
6) It is defined rigidly.
7) It is independent of fluctuations in sampling. It does not affect the final result.

Demerits :

1) It cannot be located graphically nor can it be determined by inspection.
2) Arithmetic means can not deal with qualitative characteristics. It is limited to quantitative data only.
3) For this, continuity of observations is required. It is not applicable to observations with missing terms.
4) It is affected by extreme values.
5) Accuracy of arithmetic means depends on the source of data from which it is derived.
6) It is not suitable in extremely asymmetrical distribution.

2.11 Weighted Mean

- While determining the arithmetic mean of given values in a data, equal importance is given to all values contained in it. But some of the values may be more important as compared to other present values in a table. Average value is a common value for all readings in the data. According to the importance of that value in a data, some weights

must be given to it. For example, if cost of living of a particular group is surveyed then simple means will not solve the purpose. It contains the cost of commodities are used by them. But all commodities are not equally important. Food items are important but tea, coffee, confectionery, etc. are not so important.

- Let w_i be the weight attached to the item x_i where $i = 1,2,3,\ldots,n$. Then we define.

  Weighted arithmetic mean or weighted mean $= \dfrac{\sum_i w_i x_i}{\sum_i w_i}$

- Weighted arithmetic mean is the same as simple the mean with frequencies $f_i (i = 1,2,\ldots,n)$ replaced by $w_i (1 = 1,2,\ldots,n)$ the weights.
- Weighted mean is meaningful only if each value is equal. For the larger weight of a larger item, it will be a higher value. For smaller weights, it will be a small value. If larger items are given small weight and smaller items given larger weight then weighted mean results in smaller value than simple mean.

Example 2.11.1 *Find simple and weighted arithmetic mean of the first n numbers, the weights being the corresponding numbers.*

Solution : First n natural numbers are $1,2,\ldots,n$

We know that,

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\text{Simple A.M. } (\bar{X}) = \frac{\Sigma x}{n} = \frac{1 + 2 + 3 + \ldots + n}{n} = \frac{n(n+1)}{2} \times \frac{1}{n} = \frac{n+1}{2}$$

$$\text{Weighted A.M. } (\bar{X}_w) = \frac{\Sigma wx}{\Sigma w} = \frac{1^2 + 2^2 + 3^2 + \ldots + n^2}{1 + 2 + 3 + \ldots + n}$$

$$= \frac{n(n+1)(2n+1)}{6} \times \frac{2}{n(n+1)} = \frac{(2n+1)}{3}$$

2.12 Median

- Median of a distribution is the value of the variable which divides it into two equal parts. It is the central observation also called the **positional average.**

2.12.1 Ungrouped Data

- If the number of observations is odd then the median is the middle value after the values have been arranged in ascending or descending order of magnitude.

For example : The median of the values 25, 35, 10, 13, 5

- First of all arrange the observations either in ascending or descending order

  5, 10, 13, 25, 35

  Middle value is 13.

  ∴ Median is 13.

- If observations are even then there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

  For example, median of the values 25, 20, 15, 35, 18, 50.

  Ascending order : 15, 18, 20, 25, 35, 50

$$\text{Median} = \frac{20+25}{2} = \frac{45}{2} = \mathbf{22.5}$$

2.12.2 Discrete Frequency Distribution

- For discrete frequency distribution median is obtained by considering the cumulative frequencies. Steps for calculating median are as follows :

  1) Find $\frac{N}{2}$ where $N = \Sigma f$.

  2) Find cumulative frequency equal to or just greater than $\frac{N}{2}$.

  3) The corresponding value of x is median.

Example 2.12.1 *Obtain median for the following frequency distribution :*

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

Solution : Here $N = 120 \Rightarrow \frac{N}{2} = 60$

x	f	c.f.
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120

Cumulative frequency just greater than $\frac{N}{2}$ is 65.

Value of x corresponding to 65 is 5.

∴ Median = **5**

Example 2.12.2 *Obtain the median for the following frequency distribution :*

x	1	2	3	4	5	6	7	8	9
f	7	10	12	15	18	20	25	38	45

Solution : Here $N = 190 \Rightarrow \frac{N}{2} = 95$

x	f	c.f.
1	7	7
2	10	17
3	12	29
4	15	44
5	18	62
6	20	82
7	25	107
8	38	145
9	45	190

Cumulative frequency is just greater than $\frac{N}{2}$ is 107.

Value of x corresponding to 107 is 7.

∴ Median = **7**

2.12.3 Continuous Frequency Distribution

For continuous frequency distribution, the class corresponding to the cumulative frequency just greater than $\frac{N}{2}$ is called **median class** and the value of median is obtained by using following formula -

$$\text{Median} = l + \frac{h}{f}\left[\frac{N}{2} - C\right]$$

where

l = Lower limit of median class

f = Frequency of the median class

h = Class interval of median class

c = c.f. of the class preceding to the median class

$N = \Sigma f$

Graphical Explanation of Formula :

Median is along x-axis and less than c.f. is along the y-axis with $\frac{N}{2}$ where N is the total frequency.

We locate $\frac{N}{2}$ and median class. Let l be the lower boundary of the median class. h is the width and l + h be the upper boundary of class. f is the frequency of median class and c.f. be the less than type cumulative frequency of the preceding median class. Use less than the cumulative frequency curve.

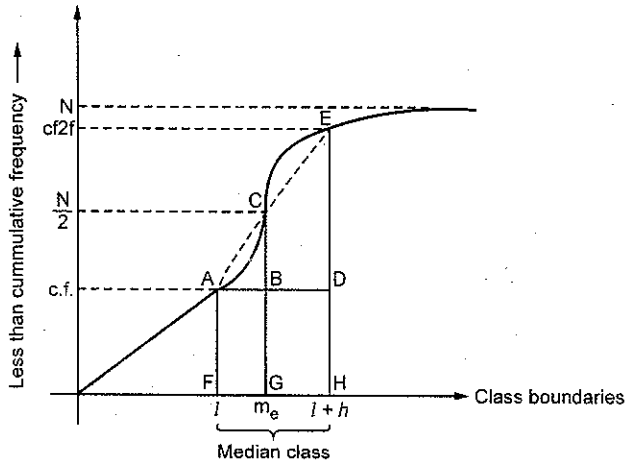


Fig. 2.12.1

Construction :

Let C $(M_e, \frac{N}{2})$ be point on less than c.f. with x-co-ordinate as median (M_e). Let CG ⊥ x-axis. Let f be the lower boundary and H be the upper boundary of the median class. Thus (F, H) is median class. Let AF ⊥ x-axis, EH ⊥ x-axis. Let AD ll^{el} to FH. Let us join ACE in a straight line. We assume over the median class less than c.f. is uniformly increasing.

$$\text{Median} = OG = OF + FG = l + FG$$

Now Δ ABC and Δ ADF are similar,

$$\frac{AB}{AD} = \frac{BC}{DE}$$

$$\frac{FG}{FH} = \frac{CG - BG}{HE - HD} \qquad \left\{\begin{array}{l}\because AB = FG \\ AD = FH\end{array}\right\}$$

$$\frac{M_e - l}{h} = \frac{N/2 - \text{c.f.}}{f} \qquad \ldots FH = \text{Width OR median class}$$

$$M_e - l = h\left(\frac{\frac{N}{2} - \text{c.f.}}{f}\right)$$

$$M_e = l + h\left(\frac{\frac{N}{2} - \text{c.f.}}{f}\right)$$

Example 2.12.3 *Obtain median for the following frequency distribution :*

Marks obtain	0-20	21-40	41-60	61-80	81-100
No. of students	1	9	32	16	7

Solution :

C.I.	f	Less than c.f.	
0-20.5	1	1	Here N = 65
20.5-40.5	9	10	$\frac{N}{2} = 32.5$
40.5-60.5	32	42	→ Median class
60.5-80.5	16	58	
80.5-100	7	65	

$l = 40.5,\ h = 20,\ f = 32,\ \text{c.f.} = 10$

$$\text{Median} = l + \left(\frac{\frac{N}{2} - \text{c.f.}}{f}\right)h$$

$$= 40.5 + \left(\frac{32.5 - 10}{32}\right) \times 20$$

$$= 40.5 + \frac{22.5}{32} \times 20$$

$$= \mathbf{54.5625}$$

Example 2.12.4 *Obtain median for the following frequency distribution*

C.I.	0-10	10-20	20-30	30-40	40-50
f	5	14	29	21	25

Solution :

C.I.	f	Less than c.f.	
0-10	5	5	Here N = 94
10-20	14	19	$\frac{N}{2} = 47$
20-30	29	48	→ Median class
30-40	21	69	
40-50	25	94	

$l = 20,\ h = 10,\ f = 29,\ c.f. = 19$

$$\text{Median} = l + \left(\frac{\frac{N}{2} - c.f.}{f}\right) \times h$$

$$= 20 + \left(\frac{47 - 19}{29}\right) \times 10$$

$$= 20 + 9.6551$$

$$= \mathbf{29.6551}$$

2.13 Merits and Demerits of Median

Merits :

1) Median gives positional average. In odd numbers the median is the middle value when numbers are arranged in ascending or descending mode. For even numbers it means of middle terms.
2) It can compute the value for unknown extreme values or missing.
3) The method of calculation is easy and able to find graphically also.
4) The median depends on the status of middle values but not on extreme values.
5) It is very useful tool in calculation of grouped data such as income of group average or weight etc.
6) It provides ease to study certain attributes which cannot be directly measured.
7) In case of grouped data, it is determined with open end intervals or under data.

Demerits :

1) It is not comfortable with long series data
2) It is quite tedious and time consuming for large values.
3) It is not suitable for algebraic manipulation.
4) It is not assumed a representative value in many situations.
5) It is based on the assumption that all frequency data are distributed uniformly with a class interval.

2.14 Mode

It is one of the measures in central tendency. To overcome drawbacks of arithmetic mean mode is used.

Definition :

- The observation with maximum frequency or the most repeated observations is called as **mode.**
- General nature of the frequency curve is bell shaped. Initially frequency is small, it increases and reaches maximum and then it declines value on the x-axis at which the maxima or peak of the frequency curve appears as a mode.

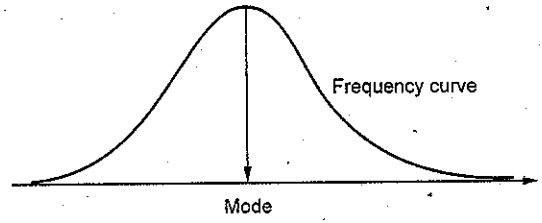


Fig. 2.14.1

- In the analysis of the voters percentages to different political parties are different. But in these results the political party having largest votes is assumed to be a representative. This is the mode and it is the appropriate average.

2.14.1 Discrete Frequency Distribution

In this case we can find observations with the largest frequency just by inspection.

Example 2.14.1 *Find mode of the following frequency distribution :*

x	5	6	7	8	9	10
f	15	7	9	21	18	13

Solution : Since maximum frequency is associated with observation 8. So mode is 8.

2.14.2 Continuous Frequency Distribution

In this case mode is given by the formula,

$$\text{Mode} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

where

l = Lower boundary of modal class
h = Class width of modal class
f_1 = Frequency of modal class
f_0 = Frequency of pre-modal class
f_2 = Frequency of post modal class

Graphical explanation of formula :

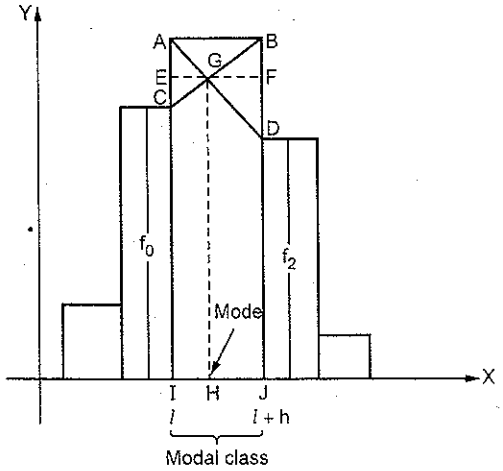


Fig. 2.14.2

- Let A and B be the vertices of modal class. Join the lines AD and CB. Let G be the point of intersection of AD and BC. Let GH $\perp^{ar}$ to X-axis, GF $\parallel^{el}$ to X-axis.
- Δ CGA and Δ DGB are similar. EG and GF are the altitudes of similar triangles.

Hence,

$$\frac{EG}{GF} = \frac{CA}{DB}$$

$$\frac{IH}{HJ} = \frac{CA}{DB}$$

$$\frac{IH}{IH + HJ} = \frac{CA}{CA + DB}$$

$$\frac{IH}{IJ} = \frac{CA}{CA + DB} \quad \ldots(1)$$

As

$$CA = IA - IC$$

$$= f_1 - f_0$$

$$DB = JB - JD$$

$$= f_1 - f_2$$

$$IH = \text{Mode} - l$$

$$IJ = h = \text{Width of modal class}$$

Hence, equation (1) becomes,

$$\frac{\text{Mode} - l}{h} = \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)}$$

$$\frac{\text{Mode} - l}{h} = \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

$$\text{Mode} - l = h\left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)$$

$$\text{Mode} = l + h\left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)$$

Example 2.14.2 *Calculate mode for the following distribution :*

C.I.	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	10	7	15	28	25	10	15

Solution :

C.I.	f
0-10	10
10-20	7
20-30	15
30-40	28
40-50	25
50-60	10
60-70	15

Here the maximum frequency is 28.

Thus class 30-40 is modal class.

where,

$l = 30$, $f_1 = 28$, $f_0 = 15$, $f_2 = 25$, $h = 10$

$\therefore$

$$\text{Mode} = l + h\left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)$$

$$= 30 + 10\left(\frac{28 - 15}{2 \times 28 - 15 - 25}\right)$$

$$= 30 + \frac{10 \times 13}{56 - 40}$$

$$= 30 + \frac{130}{16}$$

$$= \mathbf{38.125}$$

Example 2.14.3 *Calculate mode for the following distribution :*

Daily income	0-30	31-60	61-90	91-120	121-150	151-180
No. of persons	40	115	192	132	37	45

Solution :

C.I.	f
0-30.5	40
30.5-60.5	115
60.5-90.5	192
90.5-120.5	132
120.5-150.5	37
150.5-180	45

Here maximum frequency is 192.

So the modal class is 60.5 - 90.5.

where,

$l = 60.5$, $f_1 = 192$, $f_0 = 115$,

$f_2 = 132$, $h = 30$

$$\text{Mode} = l + h\left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)$$

$$= 60.5 + 30\left(\frac{192-115}{2 \times 192 - 115 - 132}\right)$$

$$= 60.5 + \frac{30 \times 77}{384 - 247}$$

$$= 60.5 + \frac{2310}{137}$$

$$= 60.5 + 16.8613 = \mathbf{107.3613}$$

Remark : Sometimes, mode is estimated from the mean and the median. For a symmetric distribution mean, median and mode coincide. If the distribution is moderately asymmetrical mean, median and mode obey the following relationship,

$$\text{Mean} - \text{Median} = \frac{1}{3}(\text{Mean} - \text{Mode})$$

2.15 Merits and Demerits of Mode

Merits :

1) As compared to other methods, it is very easy to understand.
2) The easiest tool for descriptive measure of average.
3) It is possible to locate in ungrouped data.
4) It does not depend on extreme values.
5) It is an important and useful average value.

Demerits :

1) Precise determination is missing.
2) Single mode does not exist in many calculations. Therefore it is not useful as an average in such calculation.
3) Algebraic manipulation is not possible.
4) It does not involve each value of a set.

Note : Mode is the average to be used to find the ideal size e.g. - In business forecasting, in the manufacture of ready-made garments, shoes etc.

2.16 Geometric Mean

Geometric mean of a set of n observations is the n^{th} root of their product. Thus the geometric mean G of n observations x_i, $i = 1, 2, \ldots, n$ is given by,

$$G = (x_1, x_2, \ldots x_n)^{1/n} \quad \ldots (2.16.1)$$

Computation is facilitated by using logarithms,

Taking logarithm of both, sides

$$\log G = \frac{1}{n}[\log x_1 + \log x_2 + \ldots. + \log x_n]$$

$$\log G = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$G = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right]$$

In case of frequency distribution (x_i, f_i) $i = 1, 2, \ldots, n$ then geometric mean G is,

$$G = \left(x_1^{f1} \cdot x_2^{f2} \cdot x_3^{f3} \ldots x_n^{fn}\right)^{1/N} \qquad \text{where } N = \sum_{i=1}^{n} f_i$$

Taking logarithm of both sides

$$\log G = \frac{1}{N}[f_1 \log x_1 + f_2 \log x_2 + \ldots. + f_m \log x_n]$$

$$\log G = \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i$$

$$\therefore \quad G = \text{Antilog}\left[\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right]$$

For grouped frequency distribution or continuous frequency distribution, x is taken to be the value corresponding to the midpoint of the class intervals.

Example 2.16.1 *Find the geometric mean of 24, 18, 15, 60, 50*

Solution :

$$G.M. = (24 \cdot 18 \cdot 15 \cdot 60 \cdot 50)^{1/5} = 28.6905$$

OR

x_i	$\log x_i$
24	1.3802
18	1.2552
15	1.1761
60	1.7782
50	1.6989

$$G = \text{Antilog}\left(\frac{1}{n}\sum \log x_i\right)$$

$$G = \text{Antilog}\,(1.45772)$$

$$= \mathbf{28.6893}$$

Example 2.16.2 *Show that in finding the arithmetic mean of a set of readings on a thermometer it does not matter whether we measure temperature in centigrade or Fahrenheit, but that in finding the geometric mean it does matter which scale we use.*

Solution : Let, $C_1, C_2, \ldots C_n$ be n readings on the centigrade thermometer.

Then the arithmetic mean is given by,

$$\bar{C} = \frac{1}{n}(C_1 + C_2 + \ldots + C_n)$$

If F and C be the readings in Fahrenheit and Centigrade respectively, then we have the relation.

$$\frac{F-32}{180} = \frac{C}{100} \Rightarrow F = 32 + \frac{9}{5}C$$

Thus Fahrenheit equivalents of $C_1, C_2, \ldots C_n$ are $\ldots 32 + \frac{9}{5}C_1, 32 + \frac{9}{5}C_2, \ldots, 32 + \frac{9}{5}C_n$ respectively.

Hence arithmetic mean of the reading in Fahrenheit is,

$$\bar{F} = \frac{1}{n}\left[(32 + \frac{9}{5}C_1) + (32 + \frac{9}{5}C_2) + \ldots + (32 + \frac{9}{5}C_n)\right]$$

$$= \frac{1}{n}\left[32\,n + \frac{9}{5}(C_1 + C_2 + \ldots + C_n)\right]$$

$$= 32 + \frac{9}{5}\left(\frac{C_1 + C_2 + \ldots + C_n}{n}\right)$$

$$= 32 + \frac{9}{5}\bar{C}, \text{ which is the Fahrenheit equivalent to } \bar{C}.$$

Hence in finding the arithmetic mean of a set of n readings on a thermometer, it is immaterial whether we measure temperature, it is centigrade or Fahrenheit.

Geometric mean G, of n readings in centigrade is,

$$G = (C_1 \cdot C_2, \ldots, C_n)^{1/n}$$

Geometric mean G, of Fahrenheit equivalent to $C_1, C_2, \ldots, C_n$ is

$$G = \left[(32 + \frac{9}{5}C_1) \cdot (32 + \frac{9}{5}C_2) \cdot (32 + \frac{9}{5}C_3) \ldots + (32 + \frac{9}{5}C_n)\right]^{1/n}$$

Which is not equal to Fahrenheit equivalent of G.

Hence in finding the geometric mean of the n reading on the thermometer, the scale (centigrade or equal to Fahrenheit) is important.

Note : Geometric mean of the combined group is given by,

$$\text{Log } G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}$$

where n_1 and n_2 are the sizes G_1 and G_2 be the geometric means of two series respectively.

2.17 Merits and Demerits of Geometric Mean

Merits

1) It involves all values contained in observations.
2) Geometric mean provides precise determination and each and every observations is considered.
3) It has algebraic properties and can be manipulated by a algebraic method.
4) It tries to keep the effect of extreme values to a minimum.
5) Geometric mean provides a representative average value in different large or small types of data.

Demerits

1) It refers to the positive value of a variable for a negative, geometric mean is undefined.
2) It uses logarithms.

2.18 Harmonic Mean

Harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of given values.

Let $x_1, x_2, \ldots x_n$ are observations then harmonic mean is given by,

$$H = \frac{1}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}\right)/n}$$

$$= \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

In case of frequency distribution (x_i, f_i), $i = 1,2,\ldots n$, then harmonic mean is given by,

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_i}{x_i}\right)} \text{ where } N = \sum_{i=1}^{n} f_i$$

Example 2.18.1 *Compute harmonic mean of 72, 80, 28, 35, 58.*

Solution :

$$\text{H.M.} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

$$= \frac{1}{\frac{1}{5}\left[\frac{1}{72}+\frac{1}{80}+\frac{1}{28}+\frac{1}{35}+\frac{1}{58}\right]}$$

$$= \frac{5}{0.1077} = \mathbf{46.4253}$$

Example 2.18.2 *Milk is sold at the rates of 8,10,12,15 rupees per liter in four different months. Assuming that equal amounts are spent on milk by a family in the four months, find the average price in rupees per month.*

Solution : Since equal amounts of money are spent by the family for each of the four months, the average price of milk per month is given by the harmonic mean of 8,10,12 and 15.

$$\therefore \text{Average price of milk per month} = \frac{1}{\frac{1}{4}\left(\frac{1}{8}+\frac{1}{10}+\frac{1}{12}+\frac{1}{15}\right)}$$

$$= \frac{4\times 120}{15+12+8+10} = \mathbf{10.67}$$

Example 2.18.3 *Reena drives a car from her house to her college at a speed of 10 km/h and back from the college to her house at 15 km/h find the average speed.*

Solution : Let the distance from house to college be x km. In going from house to college, the Distance (x km) is covered in $\frac{x}{10}$ hours, while in coming from college to house the distance is covered in $\frac{x}{15}$ hours.

Thus a total distance of 2x km is covered in $\left(\frac{x}{10}+\frac{x}{15}\right)$ hours.

Hence

$$\text{Average speed} = \frac{\text{Total distance travelled}}{\text{Total time taken}}$$

$$= \frac{2x}{\left(\frac{x}{10}+\frac{x}{15}\right)}$$

$$= \frac{2x}{x\left(\frac{1}{10}+\frac{1}{15}\right)} = \frac{2}{\left(\frac{1}{10}+\frac{1}{15}\right)} = \frac{2}{\left(\frac{15+10}{150}\right)}$$

$$= \frac{2\times 150}{(15+10)} = \frac{2\times 150}{25} = \mathbf{12\ km/h}$$

In this case average speed is given by the harmonic mean of 10 and 15 not by the arithmetic mean.

Remark :

1) If equal distances are covered (traveled) per unit of time with speed equal to $v_1, v_2, \ldots, v_n$ then average speed is given by,

$$\text{Average speed} = \frac{1}{\frac{1}{n}\left(\frac{1}{v_1}+\frac{1}{v_2}+\ldots+\frac{1}{v_n}\right)} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{v_i}\right)}$$

2) Instead of fixed distance being traveled with changing speed.
Let $S_1, S_2, \ldots, S_n$ are different distances and $v_1, v_3, \ldots, v_n$ are different speeds, the weights being the corresponding distances traveled is,

$$\text{Average speed} = \frac{S_1+S_2+\ldots+S_n}{\left(\frac{S_1}{v_1}+\frac{S_2}{v_2}+\ldots+\frac{S_n}{v_n}\right)} = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n}\left(\frac{S_i}{v_i}\right)}$$

3) If $x_1, x_2, \ldots x_n$ are n observations with weights $W_1, W_2, \ldots W_n$ respectively, then weighted harmonic mean is defined as -

$$H = \frac{\sum_{i=1}^{n} W_i}{\sum_{i=1}^{n} \left(\frac{W_i}{x_i}\right)}$$

2.19 Merits and Demerits of Harmonic Mean

Merits :

1) Harmonic means provide precise determination.
2) It is open to algebraic manipulation.
3) It depends on each and every value in observation.
4) It determines relative variation.

Demerits :

1) It is a complicated method to reach a final conclusion.
2) It gives importance to small values.
3) It is applicable in situations where more weightages are given to small values.

2.20 Selection of an Average

The selection of an average should be controlled by the ultimate purpose of investigation. The concrete conclusion and comparison with other series depends on proper choice. It can be justified by considering all the facts.

The selection of an average should be chosen by considering following point :

1) **Objective :** It depends on the nature of the data. Most arithmetic mean are preferred for keeping the importance of each value in a series. Mode is preferred for when items occur frequently. To indicate position of an average, median would be the choice. Geometric mean is considered when a small item is to be preferred and harmonic mean is used to give larger weights to smaller items.
2) **Representative :** An average must be representative of all values contained in a data.
3) **Nature and form of distribution :** The collected dada might be suitably distributed or non-uniform. The choice of an average depends on frequency distribution of data. If there is variation of items in collected data, then mode or median would be the right choice.
4) **Needs further analysis :** The proper selection of an average possesses many mathematical characteristics; hence it is suitable for further analysis.

5) **Sampling stability :** There must be stability in evaluated average values. It avoids sample fluctuations. The repeated samples will show minimum fluctuations in analysis.
6) **Weighting system :** It is applicable for assigning the importance to each value.
7) **Qualitative phenomena :** It is applicable to more stable data. It cannot be suitable where there is lack of stability.

Exercise

1. *What is the need of classification ?*
2. *Give procedure of frequency distribution of discrete variables.*
3. *Following are the scores in the unit test conducted for 50 students in a class.*

8	7	6	9	7	6	5	9	10	2
7	7	5	8	9	9	8	6	6	5
8	8	8	9	9	9	7	7	6	7
10	9	7	6	8	9	9	10	5	4
5	7	8	10	9	8	8	7	7	7

prepare a frequency distribution table.

4. *The data given below relates to the number of T.V. sets sold by a dealer on 25 working days of a certain month. Prepare a frequency distribution of the number of T.V. sets sold.*

   2, 4, 3, 0, 2, 1, 5, 3, 2, 0, 3, 4, 5, 1, 1, 4, 3, 2, 5, 4, 2, 2, 1, 3, 0.

5. *Heights in cm of 50 students in a class are given below :*

167.7	168.2	169.1	166.5	161.5	157.3	168.9
170.1	165.8	168.2	158.7	159.6	168.0	162.6
179.0	170.2	169.3	159.2	171.7	163.7	162.3
171.9	172.6	157.7	158.0	165.2	165.8	167.4
170.1	166.7	160.8	161.3	161.5	168.9	166.3
162.6	162.0	166.7	158.0	167.7	170.1	160.8
163.1	161.5	157.5	167.1	168.9	159.6	172.6
164.0						

Classify the above exclusive method of classification. Take the first class interval as 157-160.

6. *Prepare a frequency distribution for each of the following :*

Class mark	4	8	12	16	20
Frequency	24	45	20	10	1

7. *Give procedure of classification of continuous frequency distribution.*

8. Following is a frequency distribution of heights in cm

Classes	150-154	155-159	160-164	165-169	170-174
Frequency	2	17	29	21	1

Prepare frequency distribution table.

9. Prepare frequency distribution for the following

Marks obtained	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	4	6	20	10	7	3

10. Discuss the importance of graphic representation of a frequency distribution.
11. Explain the following terms : 1) Histogram, 2) Frequency polygon.
12. Draw a histogram of the frequency distribution given below :

Class interval	10-14	15-19	20-29	30-39	40-49	50-74	75-99
Frequency	4	12	20	18	14	25	10

13. Draw the histogram for the following data :

Monthly wages (₹)	10000-13000	13000-15000	15000-17000	17000-19000	19000-21000	21000-23000	23000-25000
No. of workers	6	53	85	56	21	16	8

14. Draw a histogram for the following data -

Age (in years)	2-5	5-11	11-12	12-14	14-15	15-16
No. of boys	6	6	2	5	1	3

15. Draw a frequency polygon for the following data

Marks	0-20	20-40	40-60	60-80	80-100
No. of students	2	18	42	28	5

16. Draw a frequency polygon for the following data :

Mid-values	25	35	45	55	65
Frequencies	5	12	33	13	7

17. Draw a frequency polygon for the following data :

I. Q.	60-69	70-79	80-89	90-99	100-109	110-119	120-129
Frequency	21	37	51	49	21	13	4

18. State the advantages and limitations of graphical representation of data.
19. What is central the tendency ? What are the requisites for an ideal measure of central tendency ?

20. What are the measures of central tendency ? Define each with necessary formulas.
21. Given below the distribution of marks obtaining for 140 students.

Marks obtained	10	20	30	40	50	60	70	80	90	100
No. of students	7	15	18	25	30	20	16	7	2	0

Calculate the mean of the distribution.

22. Calculate mean of the following distribution :

C.I.	5-7.5	7.5-12.5	12.5-17.5	17.5-22.5	22.5-27.5	27.5-32.5	32.5-37.5	37.5-42.5	42.5-47.5	47.5-52.5	52.5-57.5	57.5-62.5
Frequency	12	28	65	121	175	198	176	120	66	27	9	3

23. Calculate arithmetic mean of the group of students with weights (in kg) given below : 51, 52, 53, 51, 54, 55, 50, 53, 54, 51.
24. If $n = 10$ and $\Sigma(x - 5) = 90$ find the mean.
25. Age distribution of hundred life insurance policy holders as follows :

Age as on nearest birthday	17-19.5	20-25.5	26-35.5	36-40.5	41-50.5	51-55.5	56-60.5
Number	9	16	12	26	14	12	6

calculate arithmetic mean.

26. For a certain frequency table which has only been partly reproduced here. The mean was found to be 1.46.

No. of accidents	0	1	2	3	4	5
Frequency (No. of days)	46	-	-	25	10	5

$N = \Sigma f = 200$

Calculate missing frquecncy

27. Mean daily salary of 50 employees in a firm is ₹ 88.40. Frequency distribution of salaries of these employees in which some frequencies are missing is given below :

Salary	40-60	60-80	80-100	100-120	120-140
Frequency	6	-	17	-	5

Find missing frequency.

28. Give properties of arithmetic mean.
29. State and prove any two properties of arithmetic mean.
30. Find the weighted arithmetic mean of first n natural numbers with the same numbers and weights.

31. Scores of students along with weights are given below :

Test	Written	Viva-voce	Group discussion
Scores out of 100	75	60	65
Weights	2	1	2

Find the weighted arithmetic mean of scores.

32. Mean monthly salary of 77 workers in a certain factory is 1560/-. Mean salary of 32 of them is 1500/- and that of the next 25 of the remaining is 1640/-. What is the mean salary of the remaining 20 workers ?

33. Given :

Group 1	Group 2
$n_1 = 100$	$n_2 = 100$
$\Sigma(x - 19) = 68$	$\Sigma(y - 35) = 250$

Find $\bar{X}$, $\bar{Y}$ and combine arithmetic mean of two groups.

34. State the merit and demerits of A. M.

35. Compute the median of the following frequency distribution

x	1	2	3	4	5
y	2	7	15	5	2

36. Compute the median of the following frequency distribution :

Wages (in ₹)	above 30	above 40	above 50	above 60	above 70	above 80	above 90
No. of workers	520	470	399	210	105	45	7

37. Obtain median from the following table :

Class	0-100	100-200	200-300	300-400	400-500	500-600	600-700
Frequency	9	15	18	21	18	14	5

38. In a factory employing 3000 persons, in a day 5 percent work less than 3 hours, 580 work from 3.01 to 4.50 hours, 30 percent work from 4.5 to 6.00 hours. 500 work from 6.01 to 7.50 hours 20 percent work from 7.51 to 9.00 hours and the rest work 9.01 or more hours. What is the median hours of work ?

39. An incomplete frequency distribution is given as follows :

Variable	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	12	30	-	65	-	25	18

Total frequency = $\Sigma f = N = 229$

Given that the median value is 46. Determine the missing frequencies, using the median formula.

40. Given merits and demerits of median.

41. Daily expenditure of 100 families on transport is given below :

Expenditure	20-29	30-39	40-49	50-59	60-69
No. of families	14	-	27	-	15

If the mode of the distribution is 43.5, find the missing frequencies.

42. Obtain the mode from following frequency distribution.

Marks	0-4	4-8	8-12	12-14	14-18	18-20	20-24	24-28
No. of students	10	12	18	7	5	3	4	6

43. Age distribution of hundred life insurance policy holders is as follows :

Age	17-19	20-22	23-25	26-28	29-31	32-34	35-37	38-40
Number	9	16	12	26	14	12	6	5

Calculate mode.

44. Obtain mode from following frequency distribution :

x	1	2	3	4	5	6	7	8	9	10
f	3	15	45	57	50	36	25	12	9	15

45. What are the merits and demerits of mode.

46. Find the geometric mean 5, 10, 17, 0, 256.

47. Monthly consumption of electricity in units of a certain family in a year is given below : 210, 207, 315, 250, 240, 232, 216, 208, 209, 215, 300, 290. Calculate geometric mean.

48. Define geometric mean and harmonic mean and state the formula for each, in case of individual observations and frequency distributions.

49. A variable takes values $a, ar, ar^2, \ldots, ar^{n-1}$ find G.M.

50. Compute weighted A. M. of index numbers from the following table

Index number	300	200	250	150	250
Weight	62	4	6	12	16

51. Arithmetic mean and geometric mean of two items are 5 and 2.5 respectively. Find the harmonic mean.

52. What are the merits and demerits of G. M.

53. What are the merits and demerits of H. M.

54. Compare the method of G. M. and H. M. on the basis of their merits.

55. Calculate H.M. of the following series : 15, 250, 15.7, 157, 1.57, 105.7, 10.5, 1.06, 25.7 and 0.257.

56. *Calculate H.M. of the following series*

Values	*2*	*6*	*10*	*14*	*18*
Frequency	*4*	*12*	*20*	*9*	*5*

57. *From the following data, calculate H. M.*

Class-interval	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
Frequency	*30*	*75*	*70*	*135*	*220*

58. *A train runs 25 kms at a speed of 30 km/h another 50 km at a speed of 40 km/h, due to repairs to the track travels for 6 minutes at a speed of 10 km/h and finally covers the remaining distance of 24 kms at a speed of 24 km/h. What is the average speed in km/h ?*

Answers :

3

Marks scored	Tally marks	No. of students
1		0
2	𝍸	1
3		0
4	𝍸	1
5	𝍸	5
6	𝍸 𝍸	5
7	𝍸 𝍸 𝍸	12
8	𝍸 𝍸	10
9	𝍸 𝍸 𝍸	1
10	𝍸 𝍸	4
Total		50

4.

No. of T.V. sets sold	Tally marks	No. of days
0	𝍸	3
1	𝍸	4
2	𝍸 𝍸	6
3	𝍸 𝍸	5
4	𝍸 𝍸	4
5	𝍸	3
Total		25

5.

Class	Tally marks	Frequency
157-160	𝍸 𝍸	9
160-163	𝍸 𝍸	10
163-166	𝍸 𝍸	6
166-169	𝍸 𝍸 𝍸	14
169-172	𝍸 𝍸	8
172-175	𝍸	2
175-178	-	0
178-181	𝍸	1
Total		50

6.

Class Marks	Tally marks	Frequency
4	𝍸 𝍸 𝍸 𝍸	24
8	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	45
12	𝍸 𝍸 𝍸 𝍸	20
16	𝍸 𝍸	10
20	𝍸	1
Total		100

8.

Classes	Tally marks	Frequency
150-154	𝍸	2
155-159	𝍸 𝍸 𝍸 𝍸	17
160-164	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	29
165-169	𝍸 𝍸 𝍸 𝍸	21
170-174	𝍸	1
Total		70

9.

Marks obtained	Tally marks	No. of students
0-10	𝍸	4
10-20	𝍸 |	6
20-30	𝍸 𝍸 𝍸 𝍸	20
30-40	𝍸 𝍸	10
40-50	𝍸 ||	7
50-60	𝍸	3
Total		50

21. A. M. = $\bar{X}$ = 46.2143
22. A. M. = $\bar{\bar{X}}$ = 30.005
23. A. M. = 52.4
24. Mean = 14
25. A. M. = 40.7593
26. $f_1 = 76$, $f_2 = 38$
27. $f_2 = 10$, $f_4 = 12$
31. Weighted A. M. = 68
32. Mean salary of remaining 20 workers = ₹ 1556
33. $\bar{X}$ = 19.68, $\bar{Y}$ = 37.5, combined A. M. = 28.59
35. Median = 3
36. Median = 47.6170
37. Median = 338.0952
38. Median hours of work = 5.79
39. $f_3 = 33$, $f_5 = 46$
41. $f_2 = 27.77$, $f_4 = 28.23$
42. Mode = 9.41
43. Mode = 27.0769
44. Mode = 4
46. Geometric mean = 0
47. Geometric mean = 238.2609
49. Geometric mean = $ar^{(n-1)/2}$

50. Weighted A. M. = 267
51. Harmonic mean = 1.25
55. H. M. = 1.7374
56. H. M. = 7.2254
57. H. M. = 37.8571
58. Average speed = 31.4139 km/h

Multiple Choice Questions

Q.1 The class interval of the continuous grouped data

10 - 19

20 - 29

30 - 39

40 - 49

50 - 59 is

[a] 9 [b] 10

[c] 14.5 [d] 4.5

Q.2 Class interval is measured as -

[a] The sum of the upper and lower limit.

[b] Half of the sum of lower and upper limit.

[c] Half of the difference between upper and lower limit.

[d] The difference between upper and lower limit.

Q.3 A frequency distribution can be -

[a] Discrete [b] Continuous

[c] Both (a) and (b) [d] None of (a) and (b)

Q.4 Frequency of a variable is always _____.

[a] in percentage [b] a fraction

[c] an integer [d] none of above

Q.5 Following frequency distribution

x	12	17	24	36	45	48	52
f	2	5	3	8	9	6	1

is classified as _____.

[a] continuous distribution

[b] discrete distribution

[c] cumulative frequency distribution

[d] none of the above

Q.6 Following frequency distribution

Classes	Frequency
0 - 10	3
10 - 20	5
20 - 30	15
30 - 40	24
40 - 50	9

is of the type

[a] Discrete series [b] Continuous series

[c] Individual series [d] None of the above

Q.7 The data given is 12, 7, 25, 65, 87, 91 will be called as _____.

[a] continuous series [b] discrete series

[c] individual series [d] time series

Q.8 With the help of histogram we can prepare _____.

[a] frequency polygon [b] frequency curve

[c] frequency distribution [d] all of above

Q.9 Histogram can be used only when _____.

[a] class intervals are equal or unequal

[b] class intervals are all equal

[c] class intervals are unequal

[d] frequencies in class interval are equal

Q.10 Histogram is suitable for the data presented as _____.

[a] continuous grouped frequency distribution

[b] discrete grouped frequency distribution

[c] individual series

[d] all of above

Q.11 In a histogram with equal class intervals, height of bar are proportional to _____.

[a] mid-values of the classes [b] frequencies of respective classes

[c] either (a) or (b) [d] neither (a) nor (b)

Q.12 With the help of histogram, which of the following can be determine _____.

[a] median [b] deciles

[c] percentiles [d] mode

Q.13 Which of the following is not a measure of central tendency ?

[a] Mean deviation [b] Mean

[c] Median [d] Mode

Q.14 For n observations, harmonic mean is given by _____.

[a] $\frac{\Sigma 1/x}{n}$ [b] $\frac{n}{\Sigma 1/x}$

[c] $\frac{1/\Sigma x}{n}$ [d] $\frac{\Sigma 1/x}{1/n}$

Q.15 If $x_1, x_2, \ldots, x_n$ is a set of n observations, then harmonic mean of X is the reciprocal of _____.

[a] given observations and their arithmetic mean

[b] arithmetic mean of the given observations

[c] arithmetic mean of the reciprocals of the given observations

[d] reciprocal of mean

Q.16 Formula for geometric mean G is _____.

[a] $\frac{1}{n}\Sigma \log x_i$ [b] $\log\left[\frac{1}{n}\Sigma x_i\right]$

[c] $\text{antilog}\left[\frac{1}{n}\Sigma \log x_i\right]$ [d] $\text{antilog}\left[\frac{1}{n}\log \Sigma x_i\right]$

Q.17 Arithmetic mean of first n natural numbers is _____.

[a] $n(n+1)/2$ [b] $(n^2+1)/2$

[c] $n(n^2+1)/2$ [d] $(n^2+n)/2n$

Q.18 Which of the following relation is true

[a] $\text{Mean} = \frac{1}{2}(3 \text{ median} - \text{Mode})$

[b] Mean − 3(Mean − Median) = 2 mode

[c] $\text{Median} = \text{Mode} + \frac{2}{3}(\text{Mode} - \text{Mean})$

[d] Mode = 2 median − 3 mean

Q.19 Geometric mean of three numbers 7, 21, 63 is _____.

[a] 30.3 [b] $\sqrt{91}$

[c] $\sqrt{9251}$ [d] 21

Q.20 If n = 10, $\frac{\Sigma(x-5)}{5} = 18$ then mean is _____.

[a] 12 [b] 14

[c] 13 [d] 15

Q.21 Arithmetic mean of 5, 15, 21, 30, 34 is _____.

[a] 20 [b] 21

[c] 18 [d] 30

Q.22 Arithmetic mean of two numbers is 6.5 and their geometric mean is 6. Then two numbers are _____.

[a] 9, 6 [b] 9, 5

[c] 7, 6 [d] 4, 9

Q.23 Geometric mean of two numbers $\frac{1}{16}$ and $\frac{4}{25}$ is _____.

[a] $\frac{1}{10}$ [b] $\frac{1}{100}$

[c] 10 [d] 100

Q.24 Median of the values 11, 7, 6, 9, 12, 15, 19 is _____.

[a] 9 [b] 12

[c] 15 [d] 11

Q.25 If the two observations are 10 and – 10, then their harmonic mean is _____.

[a] 10 [b] 0

[c] 5 [d] ∞

Q.26 Mean of the following frequency distribution

x	7	12	16	22	25
f	4	5	8	3	2

is _____.

[a] 16.40 [b] 15.09

[c] 20.80 [d] none of above

Q.27 Median of the variate values 48, 35, 36, 40, 42, 54, 58, 60 is _____.

[a] 40 [b] 41

[c] 44 [d] 45

Q.28 Given n = 100, Σ(x – 19) = 68, then arithmetic mean is _____.

[a] 19.86 [b] 19.50

[c] 19.68 [d] 19.48

Q.29 Weighted mean gives a higher value than unweighted mean if _____.

[a] all the items have equal weights

[b] larger items have higher weights and small items have lower weights

[c] larger items have lower weights and smaller items have higher weights

[d] none of the above

Q.30 If for values of X, A.M = 25, H.M = 9, then the G.M is _____.

[a] 17 [b] 15

[c] 5.83 [d] 16

Q.31 Given the following frequency distribution of income of employees,

Income ₹/Month	No. of employees
0 - 250	12
250 - 500	20
500 - 750	23
750 - 1000	15
1000 - 1250	10
1250 - 1500	20

Median income of employees is _____.

[a] 625.00 [b] 760.00

[c] 695.65 [d] none of the above

Q.32 Histogram is useful to determine graphically the value of :

[a] Mean [b] Median

[c] Mode [d] All of the above

Q.33 Which of the following is not correct ?

[a] Extreme values affect the median less strongly than they do affect the mean.

[b] A median can be calculated for qualitative descriptions.

[c] The median can be calculated for every set of data, even for all sets containing open ended classes.

[d] The median is adaptable for further mathematical manipulations.

Q.34 If we add 15 in each observation of a set, then arithmetic mean is :

[a] 15 times the original data [b] not affected

[c] increased by 15 [d] decreased by 15

Q.35 The class intervals of the grouped data :

5 - 9	10 - 14	15 - 19	20 - 24

is of the type _____.

[a] inclusive class [b] discrete class

[c] exclusive class [d] none of above

Q.36 If 0.3, 0.5, 0.8, 0.7 and 1.5 are the respective weights of the values 10, 15, 20, 25, 30, then the weighted mean is _____.

[a] 20.0 [b] 23.42

[c] 16.58 [d] none of above

Q.37 When all the observations are same, then the relation between A.M, G.M and H.M is _____.

[a] A.M = G.M = H.M [b] A.M < G.M < H.M

[c] A.M < G.M = H.M [d] A.M > G.M > H.M

Answer Keys for Multiple Choice Questions :

Q.1	a	Q.2	d	Q.3	c	Q.4	c	Q.5	b
Q.6	b	Q.7	c	Q.8	d	Q.9	b	Q.10	a
Q.11	b	Q.12	d	Q.13	a	Q.14	b	Q.15	c
Q.16	b	Q.17	a	Q.18	d	Q.19	b	Q.20	b
Q.21	b	Q.22	d	Q.23	a	Q.24	d	Q.25	d
Q.26	b	Q.27	d	Q.28	c	Q.29	b	Q.30	b
Q.31	c	Q.32	c	Q.33	d	Q.34	b	Q.35	a
Q.36	b	Q.37	a						

❑❑❑

Notes

Unit III

3 Descriptive Statistics : Measures of Dispersion

Syllabus

Measures of Dispersion, Skewness and Kurtosis : *Dispersion, Characteristics for an Ideal Measure of Dispersion, Measures of Dispersion, Range, Quartile Deviation, Mean Deviation, Standard Deviation and Root Mean Square Deviation, Coefficient of Dispersion, Coefficient of Variation, Skewness, Kurtosis*

Correlation and Regression : *Bivariate Distribution, Scatter diagrams, Correlation, Karl Pearson's coefficient of correlation, Rank correlation, Regression, Lines of Regression, Regression Coefficients, Binomial and multinomial distributions, Poisson distribution, Uniform distribution, Exponential distribution, Gaussian distribution, Log-normal distribution, Chi-square distribution.*

Contents

3.1 Introduction

- In the previous chapter we saw that average condenses information into a single value. However, only the average is not sufficient to describe the distribution completely.
- There may be two or more distributions with the same means but distributions may not be identical. Runs by players X, Y, Z in 5 matches are as follows,

Players	Run					AM
X	94	98	100	102	106	100
Y	60	75	100	125	140	100
Z	30	90	100	110	170	100

- If we observe the average run scored by all students are the same but they are differ in variation. We can say that player X is more consistent than Y and y is more consistent than Z. That is runs are scattered from central value 100.
- A scatteredness of observation from a central value is called **dispersion.**

3.2 Characteristics for an Ideal Measures of Dispersion

An ideal measure of dispersion is to satisfy the following characteristics :

i) It should be based on all observations in the data set.

ii) It should be easy to understand.

iii) It should be capable of further mathematical treatment.

iv) It should be affected by fluctuations of sampling.

v) It should be affected by extreme observations.

3.3 Measure of Dispersion

- In this chapter we will discuss the following measures of dispersion i) Range ii) Quartile deviation iii) Mean deviation iv) Standard deviation.
- These measures have the same units as that of the observations e.g. cm, hours, etc. and the measures are called as **absolute measures of dispersion.**

Absolute and Relative Measure of Dispersion :

- It can be seen that absolute measures of dispersion possess units and hence create difficulty in comparison of dispersion for two or more frequency distributions. Absolute measures of dispersion use original units of data and are most useful for understanding the dispersion within the context of experiment and measurements.

- Relative measures of dispersion are calculated as ratios or percentages for example, one relative measure of dispersion is the ratio of the standard deviation to the mean.
- Relative measures of dispersion are always dimensionless and they are particularly useful for making comparisons between separate data sets or different experiments that might use different units. They are sometimes called **coefficients of dispersion.**

3.4 Range

- The range in statistics tells us the size of the data. Also for a given end datapoint of the data, we can find the other end data point with the help of range.

Definition : The range is the difference between the highest / largest value and the lowest / smallest value of the data.

L Largest / Highest value

S Smallest / Lowest value

Range = L – S

Steps to find range in statistics :

1. Arrange given data in ascending order or descending order.
2. Observe the smallest value (S) of the data and largest value (L) of the data
3. Subtract the smallest value (S) from largest value to obtain range of given data set.

e.g. The range of the data 2, 8, 10, 11, 12, 18, 21, 26, 27, 28, 33, 35, 38.

Smallest value (S) = 2

Largest value (L) = 38

$\therefore$ Range = L – S = 38 – 2 = **36**

Range is the most convenient metric to find but it has the following limitations :

a. The range does not tell us the number of data points.

b. The range cannot be used to find mean, median or mode.

c. The range is affected by extreme values.

d. The range cannot be used for open-ended distribution.

Note : Coefficient of range $= \frac{L - S}{L + S}$

3.5 Quartile Deviation

- In range we have only extreme observations of data. Hence, any change in the observations in between these two extreme observations is not going to affect the range. In many data sets extreme observations are away from remaining observations. Thus range does not gives actual picture of dispersion.
- To overcome this problem, range of middle 50 % items is calculated.
- The quartile deviations are convenient to assess the spread of a distribution. This distribution measures the spread from its central tendency.
- Inter quartile deviation can be defined as the difference between the first quartile and the third quartile in the frequency distribution table.
- When this difference is divided by two then it is called as **quartile deviation** or semi interquartile range.

Quartile deviation or semi interquartile range $= \frac{Q_3 - Q_1}{2}$

Note : Coefficient of quartile deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Formula to calculate Q_1 and Q_3 :

a) For row data or individual observations :

i) When n = Number of observations are odd

Q_1 = The value of $\left(\frac{n+1}{4}\right)^{th}$ observation

Q_3 = The value of $\left(\frac{3(n+1)}{4}\right)^{th}$ observation

ii) When n = Number of observations are even, suppose n = 10 even

Q_1 = The value of $\left(\frac{10+1}{4}\right)^{th}$ observation

= The value of $\frac{11^{th}}{4}$ observation = The value of 2.75^{th} observation

So, 2.75^{th} observation lies between 2^{nd} and 3^{rd} observation in the given data set

$Q_1 = 2^{nd}$ observation + 0.75 [3^{rd} observation – 2^{nd} observation]

Q_3 = The value of $\left(\frac{3(10+1)}{4}\right)^{th}$ observation

= The value of $\left(\frac{33}{4}\right)^{th}$ observation

= The value of 8.25^{th} observation

So, the 8.25th observation lies between the 8th and 9th observation in the given data set.

Q_3 = 8th observation + 0.25 (9th observation – 8th observation)

b) For grouped data :

N - Total frequency

C.F. - Cumulative frequency

f - Frequency of the particular class

l - Lower bound of the class which respective quartile lies

w - Class width of respective quartile lies,

Class of Q_1 = The value of $\left(\frac{N}{4}\right)^{th}$ observation.

This value provides the class where Q_1 lies,

$$Q_1 = l + \frac{w}{f}\left(\frac{\frac{N}{4} - C.F.}{f}\right)$$

Class of Q_3 = The value of $\left(\frac{3N}{4}\right)^{th}$ observation.

This value provides the class where Q_3 lies,

$$Q_3 = l + \frac{w}{f}\left(\frac{\frac{3N}{4} - C.F.}{f}\right)$$

Example 3.5.1 *The number of I-phones sold by the showroom was recorded for 15 days. Find the i) Range and coefficient of range ii) Quartile deviation and coefficient of quartile deviation for the following discrete distribution.*

Day	Frequency of iphones sold
1	29
2	14
3	21
4	100
5	26
6	38
7	108
8	19
9	100
10	74
11	69
12	104
13	62
14	92
15	22

Solution : i) To find range and coefficient of range

Smallest observation (S) = 14

Largest observation (L) = 108

Range = L – S = 108 – 14 = 94

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{94}{108+14} = \frac{94}{122} = \mathbf{0.7704}$$

ii) To find quartile deviation and coefficient of quartile deviation

We arrange data in ascending order.

14, 19, 21, [22], 26, 29, 30, 62, 69, 74, 92, [100], 100, 104, 108

Here n = Number of observations = 15

Q_1 = The value of $\left(\frac{n+1}{4}\right)^{th}$ observation

= The value of $\left(\frac{15+1}{4}\right)^{th}$ observation

= 4th observation = 22

Q_3 = The value of $\left(\frac{3(n+1)^{th}}{4}\right)$ observation

= The value of $\left(\frac{3(15+1)^{th}}{4}\right)$ observation

= 12th observation = 100

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2} = \frac{100-22}{2} = \frac{78}{2} = \mathbf{39}$$

$$\text{Coefficient of Q-D} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{100-22}{100+22} = \frac{78}{122} = \mathbf{0.6393}$$

Example 3.5.2 *Find the i) Range and coefficient of range ii) Quartile deviation and coefficient of quartile deviation for the following data :*

22 17 20 07 12 19 23 21 27 30

Solution : i) To find range and coefficient of range

Here,

Smallest observation (S) = 7

Largest observation (L) = 30

Range = L – S = 30 – 7 = 23

Coefficient of range = $\frac{L-S}{L+S} = \frac{23}{37} = \mathbf{0.6216}$

ii) To find quartile deviation, we arrange the observations in ascending order

07; 12, 17, 19, 20, 21, 22, 23 27, 30.

Here n = Number of Observation = 10

$\therefore$ Q_1 = The value of $\left(\frac{10+1}{4}\right)^{th}$ observation = 2.75^{th} observation.

2.75^{th} observation lies between 2^{nd} and 3^{rd} observation in given data

$\therefore$ Q_1 = 2^{nd} observation + 0.75(3^{rd} observation – 2^{nd} observation)

= $12 + 0.75 \times (17 - 12) = 12 + 3.75$

Q_1 = **15.75**

Q_3 = The value $\frac{3(n+1)^{th}}{4}$ observation

= The value of $\left(\frac{33}{4}\right)^{th}$ observation

= The value of 8.25^{th} observation

$\therefore$ Q_3 = 8^{th} observation + 0.25 (9^{th} observation – 8^{th} observation)

= 23 + 0.25(27 – 23) = 23 + 1

Q_3 = **24**

Quartile deviation = $\frac{Q_3 - Q_1}{2}$

= $\frac{24 - 15.75}{2}$ = **4.125**

Coefficient of QD = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

= $\frac{24 - 15.75}{24 + 15.75}$ = **0.2075**

Example 3.5.3 *Find the quartile deviation and coefficient of quartile deviation of the following frequency distribution*

Marks	< 10	10-20	20-30	30-40	40-50	50-60
No. of students	10	20	30	50	40	30

Solution : Given data :

Marks	No. of student's frequency	Cumulative frequency
< 10	10	10
10-20	20	30
20-30	30	60
30-40	50	110
40-50	40	150
50-60	30	180

Here N = 180

$\therefore$ Q_1 = The value of $\left(\frac{N}{4} = \frac{180}{4} = 45\right)^{th}$ observation

$\therefore$ Therefore (20 – 30) is Q_1 class

$\therefore$ $Q_1 = l + \frac{w}{f}\left(\frac{\frac{N}{4} - C.F.}{f}\right)$

$= 20 + \frac{10}{30}\left(\frac{\frac{180}{4} - 30}{30}\right)$

$= 20 + 5$

Q_1 = **25**

Q_3 = The value of $\left(\frac{3N}{4} = \frac{3 \times 180}{4} = 135\right)^{th}$ observation

= The value of $(135)^{th}$ observation

Therefore (40 - 50) is Q_3 class

$\therefore$ $Q_3 = l + \frac{w}{f}\left(\frac{\frac{3N}{4} - C.F.}{f}\right)$

$= 40 + \frac{10}{40}\left(\frac{\frac{3 \times 180}{4} - 110}{40}\right)$

Q_3 = **46.25**

Quartile deviation = $\frac{Q_3 - Q_1}{2}$

= $\frac{46.3 - 25}{2}$

= **10.625**

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{46.25 - 25}{46.25 + 25}$$

$$= \mathbf{0.2982}$$

3.6 Mean Deviation

- We have seen that in range, we calculate it by taking extreme observations. In quartile deviation 25 % extreme observations are left out while calculating Q_1 and Q_2.
- In both these cases only few observations are considered while calculation and remaining observations are ignored. These both terms cannot give the clear picture of all observations.
- Here we discuss the measure of dispersion which takes into account all the observations of the data set and can be calculated from an average. Such measures which are calculated from average are known as **"average deviation"**.

Definition : The arithmetic mean of absolute deviations from any average (mean or median or mode) is called Mean Deviation (M.D.) about the respective average.

i) Mean deviation (M.D.) about mean ($\bar{x}$) :

$$\text{M.D.} = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n} \qquad \text{(for row data of individual observation)}$$

$$\text{M.D.} = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{N} \qquad \text{(for frequency distribution)}$$

$$\text{Coefficient of M.D. about mean} = \frac{\text{M.D. about mean}}{\text{Mean}}$$

ii) Mean Deviation (M.D.) about median :

$$\text{M.D. (about median)} = \sum_{i=1}^{n} \frac{|x_i - \text{Median}|}{n} \qquad \text{(for row data or individual observation)}$$

$$= \sum_{i=1}^{n} f\frac{|x_i - \text{Median}|}{N} \qquad \text{(for frequency distribution)}$$

$$\text{Coefficient of M.D. about median} = \frac{\text{M.D. about median}}{\text{Median}}$$

iii) Mean Deviation (M.D.) about mode :

$$\text{M.D. (about mode)} = \sum_{i=1}^{n} \frac{|x_i - \text{Mode}|}{n} \qquad \text{(for row data or individual observation)}$$

$$= \sum_{i=1}^{n} f\frac{|x_i - \text{Mode}|}{N} \qquad \text{(for frequency distribution)}$$

Note : $\text{Coefficient of M.D. about mode} = \frac{\text{MD about mode}}{\text{Mode}}$

Example 3.6.1 *Calculate mean deviation about mean and median. Also calculate coefficient of mean deviation and median for following data*
70 72 69 68 75 70 73.

Solution : i) M.D. and coefficient of M.D. about mean

$$\bar{x} = \frac{\Sigma x}{n} = \frac{497}{7} = 71$$

x	$\lvert x_i - \bar{x}\rvert = \lvert x_i - 71\rvert$
70	1
72	1
69	2
68	3
75	4
70	1
73	2
Total	14

Here $\Sigma |x_i - \bar{x}| = 14$

$$\therefore \quad \text{M.D. about mean} = \frac{\Sigma|x_i - \bar{x}|}{n} = \frac{14}{7} = 2$$

$$\text{Coefficient of M.D. about mean} = \frac{\text{M.D.}}{\text{Mean}} = \frac{2}{71} = \mathbf{0.0281}$$

ii) M.D. and coefficient of M.D. about median

To obtain median we arrange data in ascending order

68 69 70 [70] 72 73 75

Here n = 7

$$\therefore \quad \text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

$$= 4^{th} \text{ observation}$$

$$= 70$$

x_i	$\lvert x_i - \text{Median}\rvert = \lvert x_i - 70\rvert$
70	0
72	2
69	1
68	2
75	5
70	0
73	3
Total	13

Here $\Sigma|x_i - \text{Median}| = 13$

$$\text{M.D. about median} = \frac{\Sigma|x_i - \text{Median}|}{n}$$

$$= \frac{13}{7}$$

$$= \mathbf{1.8571}$$

$$\text{Coefficient of M.D. about median} = \frac{\text{M.D.}}{\text{Median}} = \frac{1.8571}{70} = \mathbf{0.0265}$$

Example 3.6.2 *Compute M.D. about i) Mean ii) Median also find coefficient of M.D. for the following frequency distribution.*

Class internal	1-3	3-5	5-7	7-9
Frequency	3	4	2	1

Solution :

Class internal	Mid-values x_i	f_i	$f_i x_i$	Cumulative frequency
1-3	2	3	6	3
3-5	4	4	16	7
5-7	6	2	8	9
7-9	8	1	8	10
		10	38	

$$\text{Mean} = \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i = N} = \frac{38}{10} = \mathbf{3.8}$$

Median class : The $\left(\frac{N}{2}\right)^{th}$ observation i.e. 5^{th} observation

Median class : 5-7

$$\text{Median} = l + w\left(\frac{\frac{N}{2} - \text{C.F.}}{f}\right)$$

Here $l = 5, \frac{N}{2} = 5, \text{C.F.} = 7, w = 2$ Median = 3

x_i	f_i	$\lvert x_i - \bar{x} = 3.8\rvert$	$f_i\lvert x_i - \bar{x}\rvert$	$\lvert\bar{x} - \text{Median} = 3\rvert$	$f_i\lvert x_i - \text{Median} = 3\rvert$
2	3	1.8	5.4	1	3
4	4	0.2	0.8	1	4
6	2	2.2	4.4	3	6
8	1	4.2	4.2	5	5
Total	10		14.8		18

$$\text{M.D. about mean} = \frac{\Sigma f_i|x_i - \bar{x}|}{N}$$

$$= \frac{14.8}{10} = \mathbf{1.48}$$

$$\text{Coefficient of M.D. about mean} = \frac{\text{M.D. about mean}}{\text{Mean}}$$

$$= \frac{1.48}{3.8}$$

$$= \mathbf{0.3894}$$

$$\text{M.D. about median} = \frac{\Sigma f_i|x_i - \text{Median}|}{N}$$

$$= \frac{18}{10} = \mathbf{1.8}$$

Coefficient of M.D. about median

$$= \frac{\text{M.D. about median}}{\text{Median}}$$

$$= \frac{1.8}{3} = \mathbf{0.6}$$

3.7 Mean Square Deviation

- Suppose $u = x - a$ is a deviation taken from an arbitrary reference point 'a'. To remove negative sign from difference we either use $|u|$ of u^2. The measure of dispersion based on

|u| which means deviation. Using u^2 we can develop a measure of dispersion. Which is better than mean deviation.

The arithmetic mean of u^2 is known as mean square deviation in short M.S.D. and is given by,

i) For row data or individual observation

$$\text{M.S.D.} = \sum_{i=1}^{n} \frac{u_i^2}{n}$$

$$= \sum_{i=1}^{n} \frac{(x_i - a)^2}{n}$$

ii) For frequency distribution

$$\text{M.S.D.} = \sum_{i=1}^{n} \frac{f_i u_i}{N}$$

$$= \sum_{i=1}^{n} \frac{f_i(x_i - a)^2}{N}$$

For different values of 'a' we get different M.S.D. This creates a problem in measuring dispersion.

3.8 Standard Deviation and Root Mean Square Deviation

a) Variance : The arithmetic mean of squares of deviation taken from arithmetic mean is called as **variance** and is given by,

$$\text{Variance or var(x)} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \qquad \text{(for row data or individual observations)}$$

$$= \sum_{i=1}^{k} \frac{f_i(x_i - \bar{x})^2}{N} \qquad \text{(for frequency distribution)}$$

b) Standard deviation : The positive square root of squares of the deviations taken from arithmetic mean is called Standard Deviation (S.D.).

In other words the positive square root of variance is called **standard deviation** or root mean square deviation.

It is denoted by σ(sigma)

$$\sigma\,(\text{S.D.}) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}} \qquad \text{(for row data or individual observation)}$$

$$= \sqrt{\sum_{i=1}^{n} \frac{f_i(x_i - \bar{x})^2}{N}} \qquad \text{(for frequency distribution)}$$

Note : $\sigma\,(\text{S.D.}) = \sqrt{\text{var}(a)}$ is the relation between S.D. and variance.

Simplified formula of above formulae :

i) For row data or individual observations

Let, $$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

Squaring on both sides,

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} \frac{x_i^2}{n} - 2\sum_{i=1}^{n} \frac{x_i}{n}\bar{x} + \sum_{i=1}^{n} \frac{\bar{x}^2}{n}$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \quad \because \frac{\Sigma x_i}{n} = \bar{x} \quad \therefore \sum_{i=1}^{n} x_i = n\bar{x}$$

$$\sigma^2 = \sum_{i=1}^{n} \frac{x_i^2}{n} - (\bar{x})^2$$

$$\therefore \quad \sigma = \sqrt{\frac{\Sigma x_i^2}{n} - (\bar{x})^2}$$

ii) For frequency distribution

$$\sigma^2 = \sum_{i=1}^{k} \frac{f_i(x_i - \bar{x})^2}{N}$$

$$= \frac{1}{N}\left[\sum_{i=1}^{k} f_i(x_i^2 - 2x_i\bar{x} + (\bar{x})^2)\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{n} f_i x_i^2 - 2\bar{x}\sum_{i=1}^{n} f_i x_i + \sum_{i=1}^{k} f_i(\bar{x})^2\right]$$

$$= \frac{\Sigma f_i x_i^2}{N} - 2(\bar{x}^2) + \bar{x}^2 \qquad \left(\because \frac{\Sigma f_i x_i}{N} = \bar{x}\right)$$

$$\sigma^2 = \frac{\Sigma f_i x_i^2}{N} - (\bar{x})^2$$

$$\sigma = \sqrt{\frac{\Sigma f_i x_i^2}{N} - (\bar{x})^2}$$

Out of all measures of dispersion standard deviation which satisfies most necessary conditions of a good measure.

c) Coefficient of variation : The coefficient of variation represents the ratio of the standard deviation and the arithmetic mean and is given by,

$$\text{C.V.} = \frac{\text{S.D.}}{|\text{A.M.}|} \times 100 = \frac{\sigma}{\bar{x}} \times 100\ \%$$

Remarks :

1) Coefficient of variation always expressed in percentage.

2) In many cases, the value of the coefficient of variation is too small, for convenience it is multiplied by 100.

Example 3.8.1 *Prove that standard deviation of first n natural number is* $\sqrt{\frac{n^2-1}{12}}$

Solution : We know that,

$$\sigma = \sqrt{\frac{\Sigma x_i^2}{n} - (\bar{x})^2} \qquad \ldots(1)$$

We have to find standard deviation of first n natural numbers,

i.e. 1, 2, 3, ..., n

Addition of first n natural numbers = $\Sigma x_i = 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$

Addition of squares of first n-natural numbers $= \Sigma x_i^2$

$$= 1^2 + 2^2 + 3^2 + \ldots + n^2$$

$$= \frac{n(n+1)(2n+1)}{6}$$

$$\therefore \quad \bar{x} = \frac{\Sigma x_i}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$$

Equation (1) becomes

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{n \times 6} - \frac{(n+1)^2}{4}}$$

$$= \sqrt{\frac{(n+1)}{2}\left[\frac{2n+1}{3} - \frac{n+1}{2}\right]}$$

$$= \sqrt{\frac{(n+1)}{2}\left[\frac{4n+2-3n-3}{6}\right]}$$

$$= \sqrt{\frac{(n+1)(n-1)}{12}}$$

$$\sigma = \sqrt{\frac{n^2-1}{12}}$$

Which is required proof.

Example 3.8.2 *Calculate Standard Deviation (S.D.) and coefficient of variation (C.V.) for the data 26, 05, 15, 20, 24.*

Solution : We know that S.D. and C.V. is given by,

$$\sigma(\text{S.D.}) = \sqrt{\frac{\Sigma x_i^2}{n} - \bar{x}^2} \qquad \ldots(1)$$

$$\text{C.V.} = \frac{\sigma}{|\bar{x}|} \times 100 \qquad \ldots(2)$$

x_i	x_i^2
26	676
05	25
15	225
20	400
24	576
$\Sigma x_i = 90$	$\Sigma x_i^2 = 1902$

Here n = Number of observations = 5

$$\therefore \quad \bar{x} = \frac{\Sigma x_i}{n} = \frac{90}{5} = 18$$

$$\therefore \text{ From (1)} \quad \sigma(\text{S.D.}) = \sqrt{\frac{1902}{5} - (18)^2} = \sqrt{56.4} = \mathbf{7.5099}$$

$$\text{From (2)} \quad \text{C.V.} = \frac{7.5099}{18} \times 100 = \mathbf{41.7222\ \%}$$

Example 3.8.3 *Runs scored in 10 matches of T-20 by batsman A and B are obtained below :*

Batsman A	Batsman B
56	48
48	30
50	92

65	24
77	45
82	75
32	71
58	40
22	35
68	94

Identify which batsman is better and consistent.

Solution : 1) For batsman A

x_i	$u_i = x_i - 50$	u_i^2
56	6	36
48	– 2	4
50	0	0
65	15	225
77	27	729
82	32	1024
32	– 18	324
58	8	64
22	– 28	784
68	18	324
	$\Sigma u_i = 58$	$\Sigma u_i^2 = 3514$

$$\bar{x} = 50 + \frac{\Sigma u_i}{10} = 50 + 5.8 = \mathbf{55.8}$$

$$\sigma\text{(S.D.) for batsman A} = \sqrt{\frac{\Sigma u_i^2}{n} - \left(\frac{\Sigma u_i}{n}\right)^2}$$

$$= \sqrt{\frac{3514}{10} - \left(\frac{58}{10}\right)^2}$$

$$= \sqrt{317.76}$$

$$= \mathbf{17.82}$$

$$\text{Coefficient of variation for batsman A (C.V.)} = \frac{\sigma}{\text{A.M.}} \times 100$$

$$= \frac{17.82}{55.8} \times 100 = \mathbf{31.94\ \%}$$

2) For batsman B

x_i	$u_i = x_i - 46$	u_i^2
48	2	4
30	– 16	256
92	46	2116
24	– 22	484
45	– 1	1
75	29	841
71	25	625
40	– 6	36
35	– 11	121
94	48	2304
	$\Sigma u_i = 94$	$\Sigma u_i^2 = 6788$

$$\bar{x} = 46 + \frac{\Sigma u_i}{10} = 46 + 9.4 = \mathbf{55.4}$$

$$\therefore \sigma\text{(S.D.) for Batsman B} = \sqrt{\frac{\Sigma u_i^2}{n} - \left(\frac{\Sigma u_i}{n}\right)^2} = \sqrt{\frac{6788}{10} - \left(\frac{94}{10}\right)^2} = \sqrt{590.44} = \mathbf{24.30}$$

$$\text{Coefficient of variation for batsman B (C.V.)} = \frac{\sigma}{\text{A.M.}} \times 100$$

$$= \frac{24.30}{55.4} \times 100$$

$$= \mathbf{43.86\ \%}$$

C.V. of batsman A is less than that of batsman B.

$\therefore$ Batsman is more consistent.

Example 3.8.4 *Following data shows fluction of market rate of goods for 15 days : find S.D. and C.V. and compare C.V.*

Data A : *518, 519, 530, 530, 544, 542, 518, 550, 527, 531, 550, 550, 529, 528, 527*

Data B : *825, 830, 830, 819, 814, 814, 844, 842, 842, 826, 832, 835, 835, 840, 840*

Solution : Data A :

x_i	f_i	$u_i = x - 530$	f_iu_i	$f_iu_i^2$
518	2	– 12	– 24	288
519	1	– 11	– 11	121
527	2	– 3	– 6	18
528	1	– 2	– 2	4
529	1	– 1	– 1	1
530	2	0	0	0
531	1	1	1	1
542	1	12	12	144
544	1	14	14	196
550	3	20	60	1200
	N = 15		$\Sigma f_iu_i = 43$	$\Sigma f_iu_i^2 = 1973$

$$\text{A.M.} = 530 + \frac{\Sigma f_i u_i}{N}$$

$$= 530 + \frac{43}{15}$$

$$= \mathbf{532.866}$$

$$\sigma(\text{S.D.}) = \sqrt{\frac{\Sigma f_i u_i^2}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)^2}$$

$$= \sqrt{\frac{1973}{15} - \left(\frac{43}{15}\right)^2}$$

$$= \mathbf{11.105}$$

$$\text{C.V.} = \frac{\sigma}{\text{A.M.}} \times 100$$

$$= \frac{11.105}{532.866} \times 100$$

$$= \mathbf{2.0840\ \%}$$

Data B :

x_i	f_i	$u_i = x - 830$	f_iu_i	$f_iu_i^2$
814	2	– 16	– 32	512
819	1	– 11	– 11	121

825	1	– 5	– 5	25
826	1	– 4	– 4	16
830	2	0	0	0
832	1	2	2	4
835	2	5	10	50
840	2	10	20	200
842	2	12	24	288
844	1	14	14	196
	N = 15		$\Sigma f_iu_i = 18$	$\Sigma f_iu_i^2 = 1412$

$$\text{A.M.} = 830 + \frac{18}{15} = 831.2$$

$$\sigma(\text{S.D.}) = \sqrt{\frac{1412}{15} - \left(\frac{18}{15}\right)^2} = 9.628$$

$$\text{C.V.} = \frac{\sigma}{\text{A.M.}} \times 100$$

$$= \frac{9.628}{831.2} \times 100$$

$$= \mathbf{1.158\ \%}$$

C.V. of data A > C.V. of Data B

∴ Data B is more consistent than data A or Dates A has greater variability.

3.9 Skewness

3.9.1 Moments

- We have studied several aspects of frequency distribution such as coverage and dispersion. In order to study a few more aspects such as symmetry, shape of frequency distribution, a moments are useful.

The quantity $\frac{\Sigma(x_i - a)^r}{n}$ is called as r^{th} moment (or moment of order r about 'a')

a) Raw Moments

Moments about origin or about any point of observation differens from mean is known as raw moments.

The r^{th} moment or moment of order r is denoted μ'_r is given by,

i) About origin :

$$\mu'_r = \sum_{i=1}^{n} \frac{x_i^r}{n} \quad \text{(for individual observations)}$$

$$= \sum_{i=1}^{k} \frac{f_i x_i^r}{N} \quad \text{(for frequency distribution)}$$

ii) About any observation A :

$$\mu'_r = \sum_{i=1}^{n} \frac{(x_i - A)^r}{n} \quad \text{(for individual observation)}$$

$$= \sum_{i=1}^{k} \frac{f_i(x_i - A)^r}{n} \quad \text{(for frequency distribution)}$$

Substituting $r = 0, 1, 2, 3, 4$ in moments about any point A

we get, $\mu'_0 = 0$

$$\mu'_1 = \frac{\Sigma f_i(x_i - A)}{N} = \frac{\Sigma f_i x_i}{N} - \left(\frac{\Sigma f_i}{N}\right) A$$

$$\mu'_1 = \bar{x} - A$$

$$\mu'_2 = \frac{\Sigma f_i(x_i - A)^2}{N}$$

$$= \sigma^2 = \text{Mean square deviation}$$

$$\mu'_3 = \frac{\Sigma f_i(x_i - A)^3}{N}$$

and $$\mu'_4 = \frac{\Sigma f_i(x_i - A)^4}{N}$$

b) Central Moments :

Moments about the mean are known as **central moments.**

Central moment of order r (or r^{th} order moment) is denoted by μ_r and is given by,

$$\mu_r = \frac{\Sigma(x_i - \bar{x})^r}{n} \quad \text{(for individual observation)}$$

$$= \frac{\Sigma f_i(x_i - \bar{x})^r}{n} \quad \text{(for frequency distribution)}$$

Substituting $r = 0, 1, 2, 3, 4$ in moment about mean of frequency distribution formula.

$$\mu_0 = \frac{\Sigma f_i}{N} = 1$$

$$\mu_1 = \frac{\Sigma f_i(x_i - \bar{x})}{N}$$

$$= \frac{\Sigma f_i x_i}{N} - \left(\frac{\Sigma f_i}{N}\right)\bar{x} = \bar{x} - \bar{x} = 0$$

$$\mu_2 = \frac{\Sigma f_i(x_i - \bar{x})^2}{N} = \text{Variance} = \sigma^2$$

$$\mu_3 = \frac{\Sigma f_i(x_i - \bar{x})^3}{N}$$

$$\mu_4 = \frac{\Sigma f_i(x_i - \bar{x})^4}{N}$$

3.9.2 Relation between Raw and Central Moments

We know that,

$$\mu'_r = \frac{\Sigma f_i(x_i - A)^r}{N}$$

Let $\mu_i = x_i - A$

$$\therefore \quad \mu'_r = \frac{\Sigma f_i(\mu_i)^r}{N} \qquad \ldots(3.9.1)$$

Also, $$\mu_r = \frac{\Sigma f_i(x_i - \bar{x})}{N} = \frac{\Sigma f_i(x_i - A + A - \bar{x})}{N}$$

$$= \frac{1}{N}\Sigma f_i(\mu_i - \mu'_1) \quad (\because \mu'_1 = \bar{x} - A)$$

Using binomial expansion

$$\mu_r = \frac{1}{N}\Sigma f_i \left\{(\mu_i)^r - {}^rC_1(\mu_i)^{r-1}(\mu'_1) + {}^rC_2(\mu_i)^{r-2}(\mu'_i)^2 + \ldots + (-1)^r(\mu'_i)^r\right\}$$

from equation (3.9.1) $\mu'_r = \frac{\Sigma f_i(\mu_i)^r}{N}$

$$\mu_r = \mu'_r - {}^rC_1\mu'_{r-1}(\mu'_1) + {}^rC_2\mu_{r-2}(\mu'_1)^2 + \ldots + (-1)^r(\mu'_1)^r$$

Substituting $r = 2, 3, 4$ we get,

$$\mu_2 = \mu'_2 - {}^2C_1\mu'_1\mu'_1 + (\mu'_1)^2$$

$$= \mu'_2 - 2(\mu'_1)^2 + (\mu'_1)^2 \quad (\because {}^2C_1 = 2)$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - {}^3C_1\mu'_2\mu'_1 + {}^3C_2\mu'_1(\mu'_1)^2 - (\mu'_1)^3$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - {}^4C_1\mu'_3\mu'_1 + {}^4C_2\,\mu'_2\,(\mu'_1)^2 - {}^4C_3\,\mu'_1(\mu'_1)^3 + {}^4C_4(\mu'_1)^4$$

$$= \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\,(\mu'_1)^2 - 4\mu'_1(\mu'_1)^4 + (\mu'_1)^4$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\,(\mu'_1)^2 - 3(\mu'_1)^4$$

Thus

$$\mu_0 = 1$$
$$\mu_1 = 0$$
$$\mu_2 = \mu'_1\,(\mu'_1)^2$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$
$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$$

Note :

i) The expression μ_r contains r terms.

ii) First term in the expression of μ_r is positive and alternative terms are negative.

iii) Sum of the coefficient on R.H.S. of each of the relations is zero.

iv) The last term in the expression of μ_r is $(\mu')^r$.

3.9.3 Sheppard's Correction for Central Moments

- While computing arithmetic mean and S.D. of frequency distribution of continuous variables, we assume that all the items in the class interval are concentrated at the midpoint of the class. The same assumption is made for computing moments. This enables us to compute moments but it gives some amount of error.
- In case of odd order moments the sum of errors being small and it is negligible. Therefore, odd order moments need not be corrected.
- On the other hand, while computing even ordered central moments, errors are raised to even power, hence all errors are effectively positive. The sum of these errors will be considerably large. In this case W.F. Sheppard suggested the following corrections,

$$\mu_2\,(\text{corrected}) = \mu_2 - \frac{h^2}{12} \qquad h \text{ is class width}$$

$$\mu_3\,(\text{corrected}) = \mu_3$$

$$\mu_4\,(\text{corrected}) = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7h^4}{240}$$

3.9.4 Skewness

- A frequency distribution is symmetric about value 'a' if the corresponding frequency curve is symmetric about 'a'. i.e. the ordinate x = a divides frequency curve into two equal parts (Fig. 3.9.1(a)).
- Skewness is a lack of symmetry or departure from symmetry.

Depending of departure from symmetry, it is divided in two types :

i) Positive skewness ii) Negative skewness.

i) Positive skewness :

- If the mean lies to the right side of mode or the curve elongated (or stretched) towards the right side then distribution is said to be positively skewed Fig. 3.9.1(b).
- In this case we observe that mode < median < mean

ii) Negative skewness :

- If the mean lies on the left side of mode or the curve elongated (or stretched) towards the left side then distribution is said to be negatively skewed Fig. 3.9.1 (c).

In this case we observe then

Mean < Median < Mode

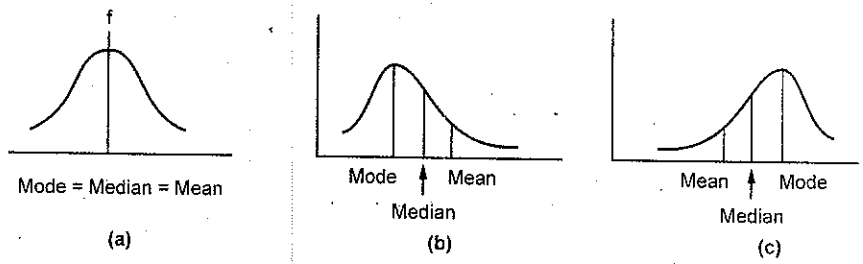


Fig. 3.9.1

Different measures of skewness are,

i) Skewness $= \dfrac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$

ii) Coefficient of skewness $= \dfrac{\mu_3^2}{\mu_2^3}$

3.10 Kurtosis

- We have studied various aspects of comparison of frequency distributions which are average, dispersion and symmetry. However these aspects are not enough for comparison.

- Two frequency distributions may have the same average, dispersion and same amount of skewness but they may differ in relative height of the curve.
- Observe the Fig. 3.10.1 containes three curves $C_1C_2C_3$ which are symmetrical about mean and have the same range.
- Kurtosis is the property of a distribution which express preakedness or flatness of curve.

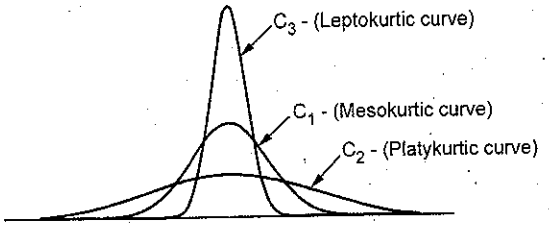


Fig. 3.10.1

Kurtosis is measured by the coefficient β_2 and is given by,

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} \text{ or } \gamma_2 = \beta_2 - 3$$

a) Mesokurtic curve (Normal curve) : The curve which is neither flat or peaked is called as **mesokurtic** curve or normal curve.

In this case $\beta_2 = 3$ or $\gamma_2 = 0$. The curve C_1 in Fig. 3.10.1 is mesokurtic.

b) Platykurtic curve : The curve which is flatter or it has less peak than that of normal curve than the curve is called as platykurtic curve.

In this case $\beta_2 < 3$ or $\gamma_2 < 0$. The curve C_2 in Fig. 3.10.1 is platykurtic.

c) Leptokurtic curve : The curve which is more peaked or curve having more peak than that of normal curve then the it is called as leptokurtic curve.

In this case $\beta_2 > 3$ or $\gamma_2 > 0$. The curve C_3 in Fig. 3.10.1 is leptokurtic curve.

Example 3.10.1 *The first four moments of four a distribution about the value 4 are 2, 20, 40 and 100 respectively.*

i) Obtain the first central moments ii) Find mean, standard deviation

iii) Find coefficients of skewness and kurtosis.

Solution : Given, $A = 4$,

First four raw moments $\mu'_1 = 2, \mu'_2 = 20, \mu'_3 = 40, \mu'_4 = 100$

i) To find the first four central moments.

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$= 20 - 2^2 = \mathbf{16}$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$

$$= 40 - 3 \times 20 \times 2 + 2 \times (2)^3 = \mathbf{-64}$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^2_1 - 3(\mu'_1)^4$$

$$= 100 - 4 \times 2 \times 40 + 6 \times 20 \times (2)^2 - 3 \times (2)^4 = \mathbf{212}$$

First four central moments are $\mu_1 = \mathbf{0}, \mu_2 = 16, \mu_3 = \mathbf{-64}, \mu_4 = \mathbf{212}$

ii) To find mean standard deviation

We know that,

$$\mu'_1 = \bar{x} - A$$

$$\therefore \quad \bar{x} = \mu'_1 + A$$

$$\therefore \quad \bar{x} = 2 + 4 = 6$$

$$\text{Mean} = \mathbf{6}$$

$$\text{S.D.} = \sqrt{\mu_2} = \sqrt{16} = \mathbf{4}$$

iii) To find coefficient of skewness and kurtosis

We know that,

Coefficient of skewness (β_1) and coefficient of kurtosis (β_2) is given by,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-64)^2}{(16)^3} = 1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{212}{(16)^2} = 0.8281$$

Example 3.10.2 *Find the first four moments about any arbitrary point for the following data*

Marks	29	30	31	32	33	34	35	36	37	38	39
No. of students	2	1	4	5	10	75	50	74	62	40	41

Solution : Also

i) Find four central moments. ii) Find coefficients of skewness and kurtosis.

Let us consider arbitrary point A = 34

Marks (x_i)	No. of students (f_i)	$u_i = x_i - 34$	f_iu_i	$f_iu_i^2$	$f_iu_i^3$	$f_iu_i^4$
29	2	– 5	– 10	50	– 250	1250
30	1	– 4	– 4	16	– 64	256

31	4	– 3	– 12	36	– 108	324
32	5	– 2	– 10	20	– 40	80
33	10	– 1	– 10	10	– 10	10
34	75	0	0	0	0	0
35	50	1	50	50	50	50
36	74	2	148	296	592	1184
37	62	3	186	558	1674	5022
38	15	4	60	240	960	3840
39	6	5	30	150	750	3750
	244		428	1426	3554	15766

i) To find four moments about any arbitrary point

We know that,

Raw moments is given by,

$$\mu'_r = \frac{1}{N}\Sigma f_i(x_i - A)^r = \frac{1}{N}\Sigma f_i u_i^r$$

$$\mu'_1 = \frac{\Sigma f_i \mu_i}{N} = \frac{428}{244} = \mathbf{1.75}$$

$$\mu'_2 = \frac{\Sigma f_i u_i^2}{N} = \frac{1426}{244} = \mathbf{5.84}$$

$$\mu'_3 = \frac{\Sigma f_i u_i^3}{N} = \frac{3554}{244} = \mathbf{14.56}$$

$$\mu'_4 = \frac{\Sigma f_i u_i^4}{N} = \frac{15766}{244} = \mathbf{64.61}$$

ii) To find four central moments

$$\mu_1 = \mathbf{0}$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 5.84 - (1.75)^2 = \mathbf{2.78}$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2(\mu'_1)^3$$

$$= 14.56 - 3 \times 5.84 \times 1.75 + 2(1.75)^3 = \mathbf{-9.98}$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu'^2_1 - 3(\mu'_1)^4$$

$$= 64.61 - 4 \times 1.75 \times 14.56 + 6 \times 5.84 \times 1.75 - 3(1.75)^4$$

$$= \mathbf{41.86}$$

iii) To find coefficient of skewness and kurtosis

We know that,

coefficient of skewness (β) is given by,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-9.98)^2}{(2.78)^3} = \mathbf{4.64}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{41.86}{(2.78)^2} = \mathbf{5.42}$$

$\beta_2 > 3$

$\therefore$ Distribution is leptokurtic.

3.11 Bivariate Distribution

The distribution for one variate x is knows as univariate distribution. The distribution which involves more than one variable is known as **bivariate distribution.**

Suppose X and Y are the two variables. Whenever variable X and Y are the measure on the same event, then they are likely correlated. For example the rainfall (X) and agricultural production (Y). We record the values of X and Y for each village of district. Suppose it gives a set of pairs $(x_1, y_1), (x_2,y_2), \ldots, (x_i,y_i), \ldots (x_n, y_n)$ where x_i is rainfall and y_i is the agricultural production of i^{th} village. This set of n pairs is a bivariate data.

3.12 Correlation

In real life we come across many situations where two variables are correlated. For example work and production, income and expense, price and demand, etc.

In all these cases we get idea about two variables. Such interrelated variables are called as correlated variables.

And linear relation between such two variables is called as **correlation.**

Types of correlation :

i) **Positive correlation :** If variable X increases, Y also increases or variable X decreases, variable Y also decreases then that type of correlation is called as positive correlation.

e.g. If work(X) in any factory increases production (y) increase and vice versa.

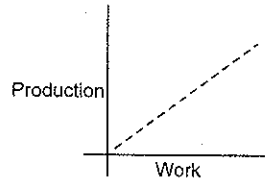


Fig. 3.12.1 Positive correlation

ii) Negative correlation : If variable X increases, variable Y decreases or variable X decreases then variable Y increases then that type of correlation is called as negative correlation.

e.g. If supply of any product is more, price decreases and if there is shortage of product then price increases.

Hence there is a negative correlation between supply of product and price of product.

There are three types of measures of correlation. i) Scatter diagram, ii) Karl Pearson's coefficient of correlation, iii) Rank correlation.

3.13 Scatter Diagram

When we plot correlated variables on XY plane using graph paper taking on variable on X axis and other on y axis that representation is called as scatter diagram.

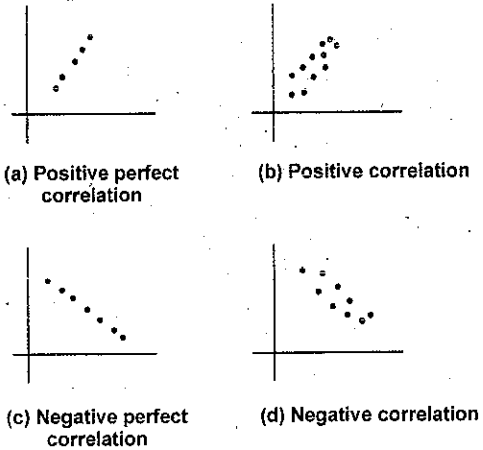


Fig. 3.13.1

3.14 Karl Pearson's Coefficient of Correlation

To measure the intensity or degree of linear relationship between two variables, Karl Pearson developed a formula called correlation coefficient.

a) Correlation coefficient between two variables x and y is denoted by r(x, y) and is defined as

$$r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

where $\text{cov}(x, y)$ = Co-variance of (x, y)

$$= \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{x} = \frac{\Sigma x_i}{n}$ $\bar{y} = \frac{\Sigma y_i}{n}$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}$$

Now
$$\text{cov}(x, y) = \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{n}\Sigma(x_i y_i - \bar{y}\cdot x_i - \bar{x}\cdot y_i + \bar{x}\,\bar{y})$$
$$= \frac{1}{n}(\Sigma x_i y_i - \bar{y}\,\Sigma x_i - \bar{x}\,\Sigma y_i + \Sigma\,\bar{x}\,\bar{y})$$
$$= \frac{1}{n}\Sigma x_i y_i - \bar{y}\left(\frac{\Sigma x_i}{n}\right) - \bar{x}\left(\frac{\Sigma y_i}{n}\right) + \frac{1}{N}\overline{xy}\,n$$
$$= \frac{1}{n}\Sigma x_i y_i - \bar{y}\,\bar{x} - \bar{x}\cdot\bar{y} + \bar{x}\cdot\bar{y}$$
$$= \frac{1}{n}\Sigma x_i y_i - \bar{x}\cdot\bar{y}$$

i.e.
$$\text{cov}(x, y) = \frac{1}{n}\Sigma x_i y_i - (\bar{x})(\bar{y})$$
$$\text{cov}(x, y) = \frac{1}{n}\Sigma x_i y_i - \left(\frac{\Sigma x_i}{n}\right)\left(\frac{\Sigma y_i}{n}\right)$$

Also
$$\sigma_x^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$
$$= \frac{1}{n}\Sigma(x_i^2 - 2\bar{x}\,x_i + (\bar{x})^2)$$
$$= \frac{1}{n}\Sigma x_i^2 - 2\bar{x}\left(\frac{\Sigma x_i}{n}\right) + (\bar{x})^2\frac{1}{n}n$$
$$= \frac{1}{n}\Sigma x_i^2 - 2\bar{x}(\bar{x}) + (\bar{x})^2$$
$$= \frac{1}{n}\Sigma x_i^2 - (\bar{x})^2$$

i.e.
$$\sigma_x^2 = \frac{1}{n}\Sigma x_i^2 - (\bar{x})^2$$
$$\sigma_x^2 = \frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2$$

Similarly
$$\sigma_y^2 = \frac{\Sigma y_i^2}{n} - \left(\frac{\Sigma y_i}{n}\right)^2$$

b) Method of step deviation

If $u_i = \frac{x_i - A}{n}$ and $v_i = \frac{y_i - B}{k}$

$$\text{cov}(u, v) = \frac{\Sigma u_i v_i}{n} - \bar{u} \cdot \bar{v}$$

then $\sigma_u^2 = \frac{\Sigma u_i^2}{n} = \bar{u}$, $\sigma_v^2 = \frac{\Sigma v_i^2}{n} - \bar{v}$

$$\bar{u} = \frac{\Sigma u_i}{n}, \quad \bar{v} = \frac{\Sigma v_i}{n}$$

and $r(u, v) = \frac{\text{cov}(u,v)}{\sigma_u \sigma_v}$

Note that $r(x, y) = r(u, v)$

Note :

1) If $r = 0$ then there is lack of relationship between x and y.

2) If $r = \pm 1$ then the relationship between x and y is very strong.

c) Correlation coefficient for bivariate frequency distribution

When a data is presented in bivariate frequency distribution then also

$$r(x, y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

where $\boxed{\text{cov}(x, y) = \frac{\Sigma f_i x_i y_i}{\Sigma f_i} - \bar{x}\,\bar{y}}$

$$= \frac{\Sigma f_i x_i y_i}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)\left(\frac{\Sigma f_i y_i}{\Sigma f_i}\right)$$

i.e. $\text{cov}(x, y) = \frac{\Sigma f_i x_i y_i}{N} - \left(\frac{\Sigma f_i x_i}{N}\right)\left(\frac{\Sigma f_i y_i}{N}\right)$

Also, $\boxed{\sigma_x^2 = \frac{\Sigma f_i x_i^2}{\Sigma f_i} - (\bar{x})^2}$

and $\boxed{\sigma_y^2 = \frac{\Sigma f_i y_i^2}{\Sigma f_i} - (\bar{y})^2}$

d) Method of step deviation

If $u_i = \frac{x_i - A}{n}$, $v_i = \frac{y_i - B}{k}$

$$\boxed{\begin{aligned} \text{cov}(u, v) &= \frac{\Sigma f_i u_i v_i}{\Sigma f_i} - \bar{u} \cdot \bar{v} \\ \sigma_u^2 &= \frac{\Sigma f_i u_i^2}{\Sigma f_i} - (\bar{u})^2 \\ \sigma_v^2 &= \frac{\Sigma f_i v_i^2}{\Sigma f_i} - (\bar{v})^2 \end{aligned}}$$

where $\bar{u} = \frac{\Sigma f_i u_i}{\Sigma f_i}$ $\bar{v} = \frac{\Sigma f_i v_i}{\Sigma f_i}$

and $r(x, y) = r(u, v) = \frac{\text{cov}(u, v)}{\sigma_u \cdot \sigma_v}$

Substituting all the above values we can write

$$r(x, y) = \frac{\frac{\Sigma f_i u_i v_i}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)\left(\frac{\Sigma f_i v_i}{n}\right)}{\sqrt{\frac{\Sigma f_i u_i^2}{N} - \left(\frac{\Sigma f_i u_i}{N}\right)^2}\sqrt{\frac{\Sigma f_i v_i^2}{N} - \left(\frac{\Sigma f_i v_i}{N}\right)^2}}$$

i.e. $$\boxed{r(x, y) = \frac{N \cdot \Sigma f_i u_i v_i - (\Sigma f_i u_i)(\Sigma f_i v_i)}{\sqrt{N\Sigma f_i u_i^2 - (\Sigma f_i u_i)^2}\sqrt{N\Sigma f_i v_i^2 - (\Sigma f_i v_i)^2}}}$$

3.15 Rank Correlation

One of the best measure of correlation is Karl pearson's coefficient of correlation but there is a difficulty in measuring the correlation between qualitative characteristics we can arrange the items in ascending or in descending order according to their merit this arrange is called as ranking.

The number indicating the position in ranking is called as rank.

The Karl pearson's correlation coefficient between these ranks is called as Spearman's Rank correlation coefficient and is denoted by R and is given by

$$R = 1 - \frac{\sigma\Sigma(x_i - y_i)^2}{n(n^2 - 1)}$$

Example 3.15.1 *Compute the coefficient of correlation for the following data.*

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Solution :

Let A = 22 $\quad \therefore u_i = \dfrac{x_i - A}{h} = \dfrac{x_i - 22}{4}$

and B = 24 $\quad \therefore v_i = \dfrac{y_i - B}{k} = \dfrac{x_i - 24}{6}$

Table

x_i	y_i	$u_i = \frac{x_i - 22}{4}$	$v_i = \frac{y_i - 24}{6}$	u_i^2	v_i^2	$u_i v_i$
10	18	– 3	– 1	9	1	3
14	22	– 2	– 2	4	4	4
18	24	– 1	0	1	0	0
22	6	0	– 3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		– 3	– 3	19	19	12

$$\bar{u} = \frac{\Sigma u_i}{n} = \frac{-3}{6} = \frac{-1}{2}$$

$$\bar{v} = \frac{\Sigma v_i}{n} = \frac{-3}{6} = \frac{-1}{2}$$

$$\text{cov}(u, v) = \frac{\Sigma u_i v_i}{n} - (\bar{u})(\bar{v})$$

$$= \frac{12}{6} - \left(-\frac{1}{2}\right)\left(-\frac{1}{2}\right)$$

$$= 2.0 - 0.25$$

$$= 1.75$$

$$\sigma_u^2 = \frac{\Sigma u_i^2}{n} - (\bar{u})^2 = \frac{19}{6} - \left(\frac{-1}{2}\right)^2$$

$$= 3.1666 - 0.25$$

$$= 2.9166$$

$$\sigma_v^2 = \frac{\Sigma v_i^2}{n} - (\bar{v})^2 = \frac{19}{6} - \left(-\frac{1}{2}\right)^2$$

$$= 2.9166$$

$\therefore$ $\quad r(x, y) = r(u, v)$

$$= \frac{\text{cov}(u, y)}{\sigma_u \sigma_v} = \frac{1.75}{\sqrt{(2.9166)(2.9166)}}$$

$$= \frac{1.75}{2.9166}$$

$$= 0.60$$

Example 3.15.2 *Following are the marks of 10 students in Maths III and strength of materials calculate the coefficient of correlation.*

Roll No.	1	2	3	4	5	6	7	8	9	10
SOM	78	36	98	25	75	82	90	62	65	39
M. III	84	51	91	60	68	62	86	58	53	47

Solution : Let x, y represents the marks in two subjects.

Arrange x in increasing order and write corresponding y in front of x.

Let $u_i = x_i - 65$, $v_i = y_i - 66$

x_i	y_i	$u_i = x_i - 65$	$v_i = y_i - 66$	u_i^2	v_i^2	$u_i v_i$
25	60	– 40	– 6	1600	36	240
36	51	– 29	– 15	841	225	435
39	47	– 26	– 19	676	361	494
62	58	– 3	– 8	9	64	24
65	53	0	– 13	0	169	0
75	68	10	2	100	4	20
78	84	13	18	169	324	234
82	62	17	– 4	289	16	– 68
90	86	25	20	625	400	500
98	91	33	25	1089	625	825
		0	0	5398	2224	2734

$$\bar{u} = \frac{\Sigma u_i}{n} = 0, \quad \bar{v} = \frac{\Sigma v_i}{n} = 0$$

$$\text{cov}(u, v) = \frac{\Sigma u_i v_i}{n} - \bar{u} \cdot \bar{v} = \frac{2734}{10} = 273.4$$

$$\sigma_u^2 = \frac{\Sigma u_i^2}{n} - (\bar{u})^2 = \frac{5398}{10} - 0 = 539.8$$

$$\sigma_v^2 = \frac{\Sigma v_i^2}{n} - (\bar{v})^2 = \frac{2224}{10} - 0 = 222.4$$

$$r(x, y) = r(u, v) = \frac{\text{cov}(u, y)}{\sigma_u \sigma_v}$$

$$= \frac{273.4}{\sqrt{(529.8)(222.4)}} = 0.787$$

Example 3.15.3 *The following marks have been obtained by a class of students in 2 papers of mathematics.*

Paper I	45	55	56	58	60	65	68	70	75	80	85
Paper II	56	50	48	60	62	64	65	70	74	82	90

Calculate the coefficient of correlation for the above data.

Solution : Let A = 68, B = 70

$u_i = x_i - 68$, $v_i = y_i - 70$

Consider the following table

x_i	y_i	$u_i = x_i - 68$	$v_i = y_i - 70$	u_i^2	v_i^2	$u_i v_i$
45	56	– 23	– 14	529	196	322
55	50	– 13	– 20	169	400	260
56	48	– 12	– 22	144	484	264
58	60	– 10	– 10	100	100	100
60	62	– 8	– 8	64	64	64
65	64	– 3	– 6	9	36	18
68	65	0	– 5	0	25	0
70	70	2	0	4	0	0
75	74	7	4	49	16	28
80	82	12	12	144	144	144
85	90	17	20	289	400	340
		– 31	– 49	1501	1865	1540

$$\bar{u} = \frac{\Sigma u}{n} = \frac{-31}{11} = 2.818,$$

$$\bar{v} = \frac{\Sigma v}{n} = \frac{-49}{11} = -4.455$$

$$\text{cov}(u, v) = \frac{\Sigma u_i v_i}{n} - \bar{u}\,\bar{v} = 140 - 12.554$$

$$\text{cov}(u, v) = 127.446$$

$$\sigma_u^2 = \frac{\Sigma u_i^2}{n} - (\bar{u})^2 = 128.512$$

$$\sigma_u = 11.336$$

$$\sigma_v^2 = \frac{\Sigma v_i^{-2}}{n} - \bar{v}^2 = 149.7, \quad \sigma_v = 12.2353$$

$$\therefore \quad r(x, y) = r(u, v) = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v}$$

$$= \mathbf{0.9188}$$

3.16 Regression

As we discussed there is a correlation between rainfall and agricultural production. If we want to calculate the agriculture production of this year when we know the rainfall of this year, regression helps us to calculate such value.

"A method of estimating the value of one variable when that of the value of other is known and when that two variables are correlated as called **regression".**

3.17 Lines of Regression

* ye to shurwat hai aage aur bhi majaa hai

If there are n observations in two variable X and Y, then there will be two regression line i) Regression line of X on Y ii) Regression line of Y on X.

When two variables are correlated then we can plot them on graph. And when these points are in nearly straight strip. Then this line represents an ideal variation and this line is called the line of best fit.

i) Regression line of X on Y :

The distance of points measured along X axis and if we minimise the deviation 'd' of these points from the line, we get a line of regression of X on Y.

It is given by X = a + by

Use of this line is to find the value of X when value of Y is given -

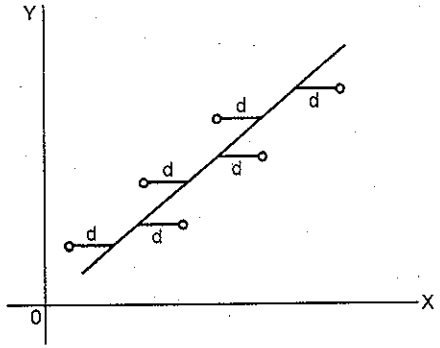


Fig. 3.17.1

ii) Regression line of Y on X :

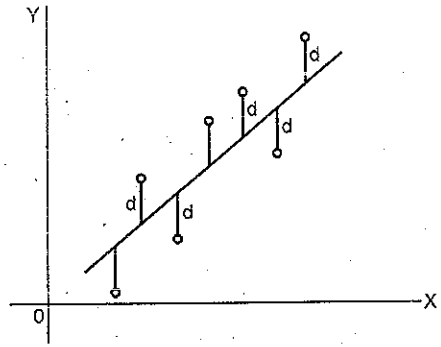


Fig. 3.17.2

The distance of the points measured along y axis and if we minimize the distance 'd' of these point from the line we get line of regression of Y on X.

It is given by $Y = aX + bX$

Use of this line is to find the value of y when value of X is given.

3.18 Regression Coefficients

i) Regression coefficient of X on Y :

The coefficient of Y in the equation of regression line of X on Y i.e. $x = a + by$ is called regression coefficient of X on Y.

It is denoted by b_{xy} and is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$\therefore$ The equation of the regression line of X on Y is

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

ii) Regression coefficient of Y on X :

The coefficient of x in the equation of regression line of y on x i.e. $y = a + bx$ is called regression coefficient of y on x.

It is denoted by b_{yx} and is given by,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$\therefore$ The equation of the regression line of y on x is

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

Note :

1) Regression coefficients can also be expressed as

$$b_{xy} = \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}, \quad b_{yx} = \frac{\frac{\Sigma xy}{n} - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

2) Coefficient of correlation is the geometric mean of regression coefficients bxy and byx

i.e. $r = \sqrt{b_{yx} \cdot b_{xy}}$

3) If one regression coefficient is greater than one then other must be less than one.

Example 3.18.1 *For the regression lines $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$ find i) $\bar{x}$ and $\bar{y}$ ii) Correlation coefficient r between x and y*

Solution : i) To find mean $\bar{x}$ and $\bar{y}$:

Mean $\bar{x}$ and $\bar{y}$ satisfies the equation of regression lines.

$\therefore$ When we solve given lines of regression we get values of $\bar{x}$ and $\bar{y}$.

$\therefore$ $3\bar{x} + 2\bar{y} = 26$ ... (1)

$6\bar{x} + \bar{y} = 31$ ... (2)

Solve equation (1) and (2) simultaneously, we get,

$\bar{x} = 4, \bar{y} = 7$

ii) To find correlation coefficient (r) :

Let equation $3x + 2y = 26$ be equation of regression line of x on y.

We convert it in $x = a + by$ form

$$\therefore \quad x = \frac{26}{3} - \frac{2y}{3}$$

In this equation coefficient of y gives regression coefficient b_{xy}

$$\therefore \quad b_{xy} = \frac{-2}{3}$$

Let equation $6x + y = 31$ be the equation of regression line of y on x.

$\therefore$ We convert it in $y = a + bx$ form

$$\therefore \quad y = 31 - 6x$$

In this equation coefficient of x gives regression coefficient b_{yx}

$$\therefore \quad b_{yx} = -6$$

We know that $r = \sqrt{b_{xy}\, b_{yx}} = \sqrt{\left(\frac{-2}{3}\right)(-6)} = 4 > 1$

But $-1 < r < 1$

$\therefore$ We consider $3x + 2y = 26$ be equation of regression lines of y on x

$$y = \frac{26}{2} - \frac{3x}{12}$$

$$b_{yx} = \frac{-3}{2}$$

We consider $6x + y = 31$ be equation of regression lines of x on y

$$\therefore \quad x = \frac{31}{6} - \frac{y}{6}$$

$$\therefore \quad b_{xy} = -\frac{1}{6}$$

$$r = \sqrt{\left(\frac{-3}{2}\right)\left(\frac{-1}{6}\right)}$$

$$= \sqrt{\frac{1}{4}}$$

$$\mathbf{r = 0.5}$$

Example 3.18.2 *Obtain the regression lines y on x and x on y for the data*

x	5	1	10	3	9
y	10	11	5	10	6

Also find i) Value of y when x = 12, ii) Value of x when y = 8

Solution : Consider the table

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
5	10	25	100	50
1	11	01	121	11
10	5	100	025	50
3	10	09	100	30
9	6	81	36	54
$\Sigma x_i = 28$	$\Sigma y_i = 42$	$\Sigma x_i^2 = 216$	$\Sigma y_i^2 = 382$	$\Sigma x_i y_i = 195$

Here number observation $n = 5$

$$\text{Mean of } x = \bar{x} = \frac{\Sigma x_i}{n} = \frac{28}{5} = 5.6$$

$$\text{Mean of } y = \bar{y} = \frac{\Sigma y_i}{n} = \frac{42}{5} = 8.4$$

Covariance between x and y :

$$\text{cov}(x, y) = \frac{\Sigma x_i y_i}{n} - \bar{x}\,\bar{y}$$

$$= \frac{195}{5} - (5.6)(8.4) = \mathbf{-8.04}$$

$$\sigma_x^2 = \frac{\Sigma x_i^2}{n} - \bar{x}^2 = \frac{216}{5} - (5.6)^2 = 11.84$$

$$\sigma_y^2 = \frac{\Sigma y_i^2}{n} - \bar{y}^2 = \frac{382}{5} - (8.4)^2 = 5.84$$

Regression coefficient of y on x :

$$b_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2} = \frac{-8.04}{11.84} = -0.679$$

Regression coefficient of x on y :

$$b_{xy} = \frac{\text{cov}(x, y)}{\sigma_y^2} = \frac{-8.04}{(5.84)^2} = -0.236$$

i) Regression line of y on x :

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

$$(y - 8.4) = -0.679(x - 5.6)$$

$$\mathbf{y = 12.202 - 0.679x} \qquad \ldots (1)$$

ii) Regression line of x on y :

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

$$(x - 5.6) = (-0.236)(y - 8.4)$$

$$\mathbf{x = 7.582 - 0.236y} \quad \ldots (2)$$

iii) Value of y when x = 12

Put x = 12 in (1)

We get, $y = 4.054$

iv) Value of x when y = 8

Put y = 8 in (2)

We get, $x = 5.964$

3.19 Binomial and Multinomial Distributions

Let us consider an experiment involving n trials performed under identical conditions each of which can result either success whose probability is p or failure whose probability is q. Such distribution is called by Binomial. As there are only two outcomes p + q = 1.

Let us consider r success out of n trials hence (n – r) are failure.

∴ Probability of r success in n trial is given by,

$$p(\text{r success}) = p\{(ssss\ldots ss)(\text{times})\} X\ p\{(fff\ldots ff)(n-r)\ \text{times}\}$$

$$= (ppp\ldots p)(\text{r times})\ X\ (qq\ldots q)\ (n-r)\ \text{times}$$

$$= p^r q^{n-r}$$

A r successes and (n – r) failures can occur in ${}^nC_r$ mutually exclusive cases.

$$\therefore \quad p[\text{r success}] = {}^nC_r p^r q^{n-r}$$

∴ Put $r = 0, 1, 2, \ldots, n$

We get,

r	0	1	2	... n – 1	n
p(r)	${}^nC_0 p^0 q^n$	${}^nC_1 p^1 q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	${}^nC_n p^{n-1} q$	${}^nC_0 p^n q^0$

Consider the binomial expansion

$$(q + p)^n = q^n + {}^nC_1 q^{n-1} p + {}^nC_2 p^{n-2} p^2 + \ldots + p^n$$

Terms of RHS given the probability of r = 0, 1, 2, …, n success. Hence above probability distribution to be called Binomial probability distribution. It is denoted by β(n, p, r)

$$\therefore \quad \beta(n, p, r) = {}^nC_r p^r q^{n-r}$$

Note : Detail study of binomial distribution will be in next unit.

Multinational distribution is generalisation of binomial distribution.

In the experiment if there are more than two mutually exclusive outcomes then it leads to multinomial distribution.

We use multinomial distribution on experiments in which the following conditions holds good.

- The experiment in which repeated trials performed. e.g. Rolling a die fire times instead of just once.
- Each trial must be independent of others.
- The probability of each outcomes be same across each trial of the experiment.

3.20 Poisson Distribution

Poisson distribution is a limiting case of the binomial distribution when 'n' the number of trials is infinitely large and 'p' the probability of success is very small and np is finite.

The probability of r successes in a series of n independent trials is

$$\beta(n, p, r) = {}^nC_r p^r q^{n-r},\ r = 0, 1, 2, \ldots, n$$

$$= \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$$= \frac{n(n-1)(n-2)\ldots(n-(r-1)(n-r))!}{(n-r)!r!} p^r q^{n-r}$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} p^r q^{n-r}$$

Let $z = np \quad \therefore \ p = \frac{z}{n}$

$$B(n, p, r) = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} p^r (1-p)^{n-r}$$

$$= \frac{\left[1 - \frac{1}{n}\right]\left[1 - \frac{2}{n}\right] - \left[1 - \frac{r-1}{n}\right]}{r!\left[1 - \frac{z}{n}\right]^r} \left(\frac{z}{n}\right)^r \left(1 - \frac{z}{n}\right)^n$$

As $n \to \infty$ and $np = z$

$$\lim_{n \to \infty} B(n, p, r) = \frac{z^r e^{-z}}{r!} \left(\therefore \lim_{n \to \infty} \left(1 - \frac{z}{n}\right)^n = e^{-z}\right)$$

3.21 Uniform Distribution

A discrete random variable x taking values 1, 2, 3, ..., n is said to have uniform distribution if its pmf is given by,

$$P(X = x) = \frac{1}{n}; 1, 2, \ldots, n$$

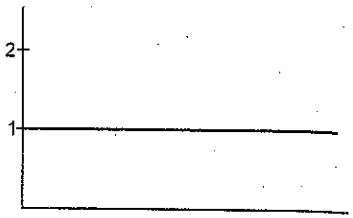


Fig. 3.21.1

Where 'n' is parameter of uniform distribution.

When $P(X = x) = 1$,

this distribution is called as standard uniform distribution shown in Fig. 3.21.1.

3.22 Exponential Distribution

A continuous random variable X taking non-negative values is said to have an exponential distribution with parameter $n > 0$, if its pmf is given by

$$P(X = x) = \begin{cases} ne^{-nx} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The exponential distribution is used when the amount of the time unit some event occurs.

Some example those can be modeled by using exponential distribution are

- Time duration between two customers those entering in a contain shop.
- Time duration between occurrance of earthquake.
- Time gap between customer calls in business.

3.23 Gaussian Distribution

Gaussian distribution is also known as normal distribution. Gaussian distribution is the continuous distribution of a variable x (known as random variable or normal variate)

Gaussian distribution can be derived from binomial distribution when n number of trials is very large and neither p nor q is very small.

Gaussian distribution interms of continuous random variable x with mean a and standard deviation σ^2 is given by

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{\sigma^2}} \quad \ldots (3.23.1)$$

When it is represented on graph, it is bell shaped curve. This curve is symmetric about ordinate x = a i.e. symmetric about mean as shown in Fig. 3.23.1.

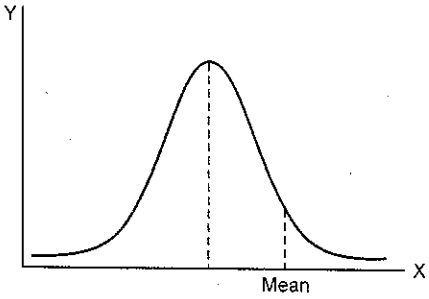


Fig. 3.23.1

In Gaussian curve mean, median and mode coinsides, therefore the gaussian curve has one maximum point. Two tails of the curve are extend to $-\infty$ and $+\infty$ towards the positive and negative directions of X axis i.e. curve is asymptotic to X axis. The line x = a(mean) divides the distribution curve in two equal parts above X axis.

Total area under this curve is 1.

i.e. $$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

When x lies between x_1 and x_2, then probability is represented by the area (shaded in Fig. 3.23.2) under the curve y = f(x) bounded by X axis and the lines x = 2y and $x = x_2$ and is given by

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2 dx}} \quad \ldots (3.23.2)$$

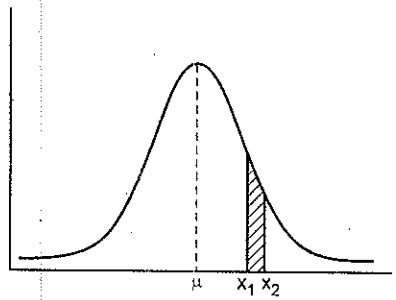


Fig. 3.23.2

Standard form :

Put $$z = \frac{x-a}{\sigma}; dx = dz$$

$\therefore$ $$z_1 = \frac{x_1 - a}{\sigma}$$

$$z_2 = \frac{x_2 - a}{\sigma}$$

in equation (3.23.2)

When mean (a) = 0 and Standard deviation (σ) = 1, random variable z is called as Standard normal random variable and equation (3.23.2) becomes

$$P(z_1 \le z \le z_2) = \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} dz \qquad \ldots (3.23.3)$$

It is free from two parameters a and σ. It is useful to compute areas under the normal probability curve by using standard table.

When curve is divided into two equal parts by z = 0 then LHS area = RHS area = 0.5

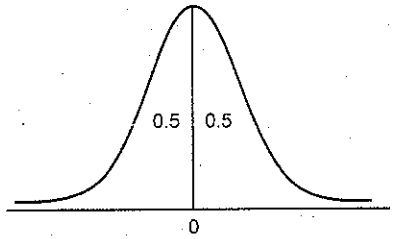


Fig. 3.23.3

Table of area :

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5259
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0,6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9494	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

1.8	0.9641	0.9649	0.9653	0.9664	0.9871	0.9678	0.9686	0.9693	0.9699	0.9708
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9883	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9634	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9809	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9988	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9998	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9898

Table 3.23.1 in each row and each column 0.5 to be substracted

Remarks :

1) $p(x_1 \le x \le x_2) = P(z_1 \le z \le z_2) = P(z_2) - P(z_1)$ = (Area under the normal curve from 0 to z_2) – (Area under the normal curve from 0 to z_1) = $A(z_2) - A(z_1)$

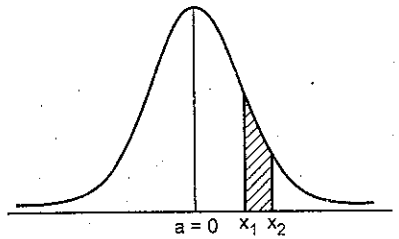


Fig. 3.23.4

2) $P(0 < x < x_1) = P(0 < z < z_1)$ = Area under the normal curve 0 to $z_1 = A(z_1)$

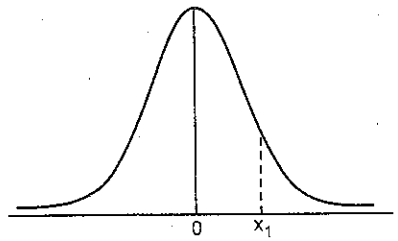


Fig. 3.23.5

3) $P(x > x_1) = P(z > z_1) = 0.5 - A(z_1)$

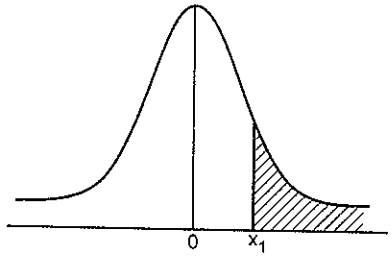


Fig. 3.23.6

4) $P(x < -x) = P(z < -z_1) = 0.5 - A(z_1)$

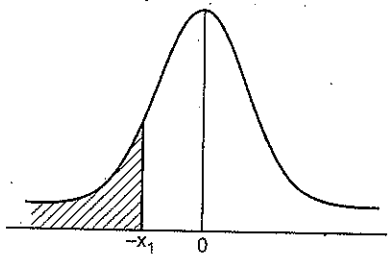


Fig. 3.23.7

5) $P(-x_1 < x < x_2) = P(-z_1 < z < -z_2) = A(z_1) - A(z_2)$

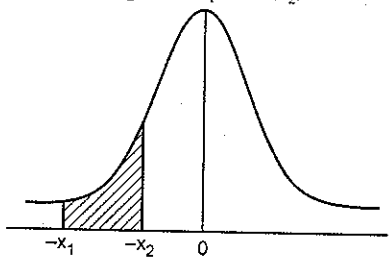


Fig. 3.23.8

6) $P(-x_1 < x < x_2) = P(-z_1 < z < z_2) = A(z_1) + A(z_2)$

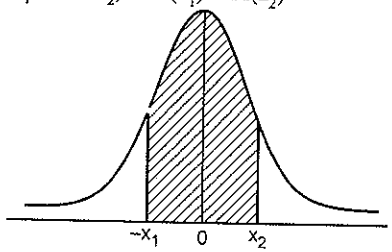


Fig. 3.23.9

7) Mean deviation from the mean of Gaussian distribution is $\frac{4}{5}\sigma$ (approximate)

3.24 Log-Normal Distribution

A continuous distribution of random variable X whose natural logarithm is normally distributed is known as log-normal distribution.

This distribution is also called as, Galton distribution.

For example if random variable X = exp(y) has log normal distribution then y = logx has normal distribution.

Most use of lognormal distribution in finance, in analysis of stock prices, milk production by cows.

The probability density function is given by then mean a and standard deviation σ,

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - a)^2}{2\sigma^2}\right); x > 0$$

The log normal distribution is represented by graph as shown in Fig. 3.24.1.

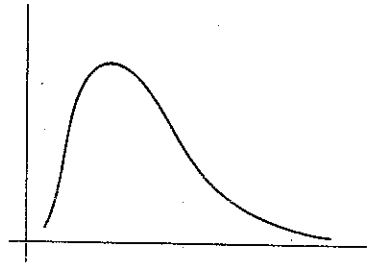


Fig. 3.24.1

Log normal distribution curve is positively skewed with long right tails due to low mean values and high variances in the random variables.

3.25 Chi-square Distribution

Chi-square distribution mainly used in

i) Testing hypothesis.

ii) Tests for the independence if two criteria of classification of qualitative data.

iii) Tests for goodness of an observed distribution to a theoretical one.

The chi-square distribution is one of the important distribution in statistical theory.

It is denoted by X_n^2 where 'n' is parameter of distribution called as "degrees of freedom".

The variate X_n^2 is the sum of squares of n independent standard normal variables [N(0,1)].

Let $X_1, X_2, X_3, \ldots, X_n$ be n independent N(0, 1) variables then the Chi-square distribution is given by,

$$Y = X_n^2 = \sum_{i=1}^{n} (X_i)^2$$

where n is degrees of freedom.

3.26 Additive Property of Chi-square Distribution

If Y_1 and Y_2 are two independent Chi-square (X^2) variates with n_1 and n_2 degrees of freedom respectively, then Y_1 and Y_2 has Chi-square (X^2) distribution with $(n_1 + n_2)$ degrees of freedom.

3.27 Hypothesis

Hypothesis is a claim about testing the population parameters such as mean variance

e.g.

1) Average of two vehical's.

2) Proportion of pass students in two subjects.

3) Proportion of placements of students of two branches.

4) Proportion of production of crop is two different fields.

These claims stated in terms of population parameters or distribution are called hypothesis.

Definition : Hypothesis is the statement or assertion about the statistical distribution or unknown parameter of statistical distribution.

3.28 Null and Alternative Hypothesis

Test begins by considering null and alternative hypothesis. These hypothesis containing opposite view points. i.e. If one rejected the other one is to be accepted.

Null Hypothesis (H_0) :

According to R.A. Fisher null hypothesis defined as a hypothesis of no difference and it is denoted by H_0.

e.g. $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$, This Hypothesis states that there is no difference between two population means.

Alternative Hypothesis (H_1) :

Alternative hypothesis is a hypothesis when null hypothesis rejected, alternative hypothesis accepted.

It other words Alternative hypothesis is a complementary hypothesis to null hypothes is and it is denoted by H_1.

e.g. If $H_0 : \mu_1 = \mu_2$ then alternative hypothesis may be

$H_1 : \mu_1 \neq \mu_2$ or

$H_1 : \mu_1 < \mu_2$ or

$H_1 : \mu_1 > \mu_2$

3.29 One-Sided or Two-Sided Hypothesis

Nature of the hypothesis decides the type of hypothesis whether it is one-sided Hypothesis or Two-sided Hypothesis.

One-sided : Hypothesis $H_1 : \mu_1 > \mu_0$, $H_1 : P < 0.5$, $H_1 : \mu_1 < \mu_2$ etc. are called as one-sided hypothesis.

Two-sided : Hypothesis of the type $H_1 : P_1 \neq 0.5$, $H_1 : \sigma_1 \neq \sigma_2$, $\sigma_1 : \mu \neq \mu_0$ etc. are called as two sided hypothesis.

3.30 Errors (Type-I and Type-II)

Type-I and Type-II are the two errors defined in testing hypothesis.

Type-I error : When H_0 is true and it is rejected.

e.g. Consider null hypothesis

H_0 product 'X' is a good.

In reality if product 'X' is a good product. i.e. H_0 is true, but if the quality analyst says that the product is bad hypothesis rejected then there is a mistake such a mistake is called as Type-I error.

Type-II error : When H_0 is false and it is accepted.

e.g. Consider null hypothesis

H_0 : Prcduct 'Y' is bad.

In reality is product 'Y' is good product. i.e. H_0 is false, but the quality analyst says that product is bad means hypothesis accepted then there is mistake such a mistake is called as Type-II error.

The following table will clear the idea.

Actual Situation	Decision	
	H_0 Rejected	H_0 is accepted
H_0 is true	Type I error	Correct decision
H_0 is false	Correct decision	Type-II error

3.31 Some Definitions

i) **Test of Hypothesis :** A rule which leads to the decision of acceptance of H_0 or rejection of H_0 on the basis of random samples is called test of hypothesis.

ii) **One sided and two sided tests :** The tests used for testing null hypothesis (H_0) are called as one sided or two sided tests according as the alternative hypothesis are one sided or two sided hypothesis.

iii) Test statistic : A function of random sample observation which is used to test null hypothesis H_0 is called as a test statistic.

iv) Critical region : Let $X_1, X_2, \ldots, X_n$ be the set of values for which the null hypothesis H_0 is rejected is called as critical region or rejection region.

Critical region is denoted by w and acceptance region is denoted by (w^c) as shown in Fig. 3.31.1.

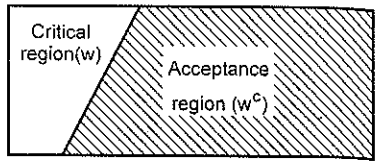


Fig. 3.31.1

v) Level of significance : Probability of rejecting the null hypothesis H_0 when it is true is called as level of significance.

That is it is the probability of committing type-I error. It is denoted by α.

If we try to minimize level of significance the probability of error of type-II increases, So level of significance cannot be made zero. However we can fix it in advance as 0.01 or 0.05 i.e. (1 % or 5 %). In most of the cases it is 5 %.

vi) Test for Goodness of fit of X^2 distribution : For a given frequency distribution are try to fit probability distribution.

We know that there are several probability distributions, among these distribution which will fit to given data use have to identify for that we have to test the appropriatness of the fit.

Let H_0 : Fitting of the probability distribution to given data is proper (good).

The test based on X^2 distribution used to test this H_0 is called X^2 test of goodness of fit.

Suppose $o_1, o_2, o_3 \ldots o_i \ldots o_k$ be the set of observed frequencies and $e_1, e_2, \ldots e_i \ldots e_k$ be corresponding expected frequencies or theoretical frequencies. There is no significance between observed and theoretical frequenceness.

Suppose P = Number of parameter estimated for fitting the probability distribution.

$$\therefore \quad \sum_{i=1}^{k} o_i = \sum_{i=1}^{k} e_i = N$$

If H_0 is true, then the statistic

$$X^2 = \sum_{i=1}^{k} \left(\frac{o_i - e_i}{e_i}\right)^2 = \sum_{i=1}^{k} \frac{o_i^2 - 2o_ie_i + e_i^2}{e_i}$$

$$X^2 = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - \sum_{i=1}^{k} 2o_i + \sum_{i=1}^{k} e_i$$

$$= \sum_{i=1}^{k} \frac{o_i^2}{e_i} - 2N + N \quad \left(\because \sum_{i=1}^{k} k = \sum_{i=1}^{k} e_i = N\right)$$

$$= \sum_{i=1}^{k} \left(\frac{o_i^2}{e_i}\right) - N$$

has X^2 distribution with $(k - p - 1)$ degrees of freedom and it is parameter of the chi-square distribution.

The critical region at level of significance α is

$$X^2_{k-p-1} \geq X^2_{k-p-1;\alpha}$$

where $X^2_{k-p-1;k}$ - Table value corresponding (critical value) to degree of freedom $k - p - 1$ and level of significance α, it is shown is Fig. 3.31.1(a).

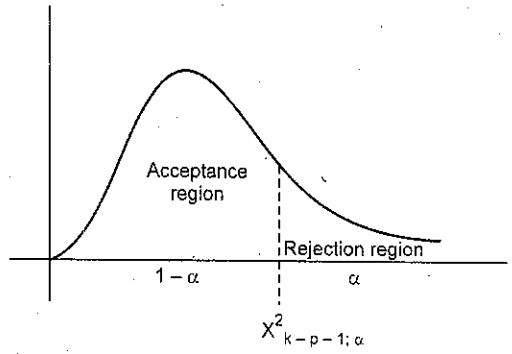


Fig. 3.31.1 (a)

Note :

1) If the total of cell frequencies is sufficiently large (greater than 50) and expected frequencies are greater than or equal to 5(i.e. $e_i \geq 5$) the we can apply this test.
2) While fitting a probability distribution or in finding expected frequencies if any parameter are not estimated then value of P is zero.
3) When expected frequency of a class is less than 5, the class is merged into neighbouring class along with its observed and expected frequencies until total of expected frequencies becomes greater than or equal to five. This procedure is called "Pooling the classes".

   In this case k is number of class frequencies after pooling.

Example 3.31.1 *A pair of dice is thrown 20 times. If getting a doublet is considered a success, find probability of i) 6 successes ii) No success.*

Solution :

Here $n = 10$

p = Probability of getting a doublet $= \frac{6}{36} = \frac{1}{6}$

q = Probability of not getting a doublet $= 1 - \frac{1}{6} = \frac{5}{6}$

By binomial distribution,

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

$$p(6) = {}^{20}C_6 \left(\frac{1}{2}\right)^6 \left(\frac{5}{6}\right)^{20-6}$$

$$p(6) = \frac{20!}{14!\,6!}\left(\frac{1}{6}\right)^6 \left(\frac{5}{6}\right)^{14}$$

Example 3.31.2 *10 % of articles from a certain machine are defective. What is the probability that there will be 4 defectives in a sample of 20.*

Solution :

Here n = No. of sample = 20

$$p = \mathbf{10\ \%} = \frac{10}{100} = \frac{1}{10}$$

$$q = \frac{1-1}{10} = \frac{9}{10}$$

We know that

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

$$p(4) = {}^{20}C_4 \left(\frac{1}{10}\right)^4 \left(\frac{9}{10}\right)^{21}$$

$$= \frac{20!}{16!\,4!}\left(\frac{1}{10}\right)^4 \left(\frac{9}{10}\right)^{21} = \frac{20\times 19\times 18\times 17}{4\times 3\times 2\times 1}\,\frac{(9)^{21}}{(10)^{25}}$$

$$= 0.05278$$

Example 3.31.3 *A dice is thrown 10 times. If getting an odd number is a success. What is the probability of i) 8 successes. ii) At least 6 success.*

Solution : A dice is thrown 10 times. odd numbers are 1, 3, 5, 6, 9

$\therefore$ Probability of an getting an odd number $= \frac{5}{10} = \frac{1}{2}$

i.e. $$p = \frac{1}{2}$$

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

We know that, $$p(r) = {}^{n}C_r\, p^r\, q^{n-r}$$

i) Probability of 8 successes :

$$p(8) = {}^{10}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2$$

$$= \frac{10!}{2!\,8!}\left(\frac{1}{2}\right)^{10} = \frac{10\times 9}{2}\left(\frac{1}{2}\right)^{10}$$

$$p(8) = \frac{45}{1024}$$

ii) Probability of getting at least 4 success :

$$p(r \geq 6) = p(6) + p(7) + p(8) + p(9) + p(10)$$

$$= {}^{10}C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^4 + {}^{10}C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^2$$

$$+ {}^{10}C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 + {}^{10}C_{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0$$

$$= \frac{10!}{4!\,6!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{2!\,8!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{1!\,9!}\left(\frac{1}{2}\right)^{10} + 1\left(\frac{1}{2}\right)$$

$$= \left(\frac{1}{2}\right)^{10} (210 + 45 + 10 + 1)$$

$$= \frac{266}{1024}$$

Example 3.31.4 *Assume that on average telephone number out of 15 called between 2 pm to 3 pm on a week days is busy. What is the probability that is 6 randomly selected telephone numbers called i) Not more than 3 busy ii) At least 3 of them is busy.*

Solution : The probability that the telephone number called between 2 pm and 3 pm is busy is

$$p = \frac{1}{15}$$

$$q = 1 - \frac{1}{15} = \frac{14}{15}$$

Hence probability of r no.s called out of 6 called are

$$p(r) = {}^{6}C_r\, p^r\, q^{6-r}$$

i) For not more than 3 calls are busy :

$$p(r \leq 3) = p(0) + p(1) + p(2) + p(3)$$

$$= {}^{6}C_0 \left(\frac{1}{15}\right)^0 \left(\frac{14}{15}\right)^6 + {}^{6}C_1 \left(\frac{1}{15}\right)^1 \left(\frac{14}{15}\right)^5$$

$$+ {}^{6}C_2 \left(\frac{1}{15}\right)^2 \left(\frac{14}{15}\right)^4 + {}^{6}C_3 \left(\frac{1}{15}\right)^3 \left(\frac{14}{15}\right)^3$$

$$= \frac{(14)^3}{(15)^6}\left[(14)^3 + 6(14)^2 + \frac{6\times 5}{1\times 2}(14) + \frac{6\times 5\times 4}{1\times 2\times 3}\right]$$

$$= 0.9997$$

ii) For at least 3 calls to be busy :

$$p(r \geq 3) = 1 - p(< 3)$$

$$p(r \geq 3) = 1 - [p(r = 0) + p(r = 1) + p(r = 2)]$$

$$= 1 - \left[{}^{6}C_0 \left(\frac{1}{15}\right)^0 \left(\frac{14}{15}\right)^5 + {}^{6}C_1 \left(\frac{1}{15}\right)^1 \left(\frac{14}{15}\right)^5 + {}^{6}C_2 \left(\frac{1}{15}\right)^2 \left(\frac{14}{15}\right)^4 \right]$$

$$= 0.0051$$

Example 3.31.5 *Out of 2000 families with 4 children each, how many would you expect to have i) At least a boy ii) Two boys iii) 1 or 2 girls iv) No girls.*

Solution :

Let p = Probability of having a boy $= \frac{1}{2}$

q = Probability of having a girl $= \frac{1}{2}$

We know that

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

i) Probability of at least a boy

$$p(\text{at least a boy}) = 1 - p(r < 1)$$

$$= 1 - p(r = 0)$$

$$= 1 - {}^{4}C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4$$

$$= 1 - \frac{1}{16}$$

$$= \frac{15}{16}$$

Hence expected no. of families having at least a boy,

$$2000 \, p(r \geq 1) = 2000 \times \frac{15}{16}$$

$$= 1875$$

ii) Probability of two boys :

$$p(\text{having 2 boys}) = p(r = 2)$$

$$= {}^{4}C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2$$

$$= \frac{4!}{2!\,2!} \left(\frac{1}{2}\right)^4$$

$$= \frac{3}{8}$$

Hence expected number of families having 2 boys

$$2000 \, p(r = 2) = 2000 \times \frac{3}{8}$$

$$= 750$$

iii) Probability of expected no. of families having 1 or 2 girls. Which is equivalent to finding probability of having 3 boys or 2 boys.

$$p(1 \text{ or } 2 \text{ girls}) = p(3 \text{ or } 2 \text{ boys})$$

$$= p(r = 3) + p(r = 2)$$

$$= {}^{4}C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 + {}^{4}C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2$$

$$= \frac{4!}{1!\,3!} \left(\frac{1}{2}\right)^4 + \frac{4!}{2!\,2!} \left(\frac{1}{2}\right)^4$$

$$= \frac{10}{16} = \frac{5}{8}$$

Hence expected number of families having 1 or 2 girls

$$2000 p[3 \text{ or } 2 \text{ boys}] = 2000 \times \frac{5}{8}$$

$$= 1250$$

iv) Probability of no girls i.e. probability of all children are boy's

$$p(\text{no girls}) = p(\text{having boys})$$

$$= p(r = 4)$$

$$= {}^{4}C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0$$

$$= \frac{1}{16}$$

Hence expected number of families having no girls

$$2000 \, [p(r = 4)] = 2000 \times \frac{1}{16}$$

$$= \mathbf{125}$$

Example 3.31.6 *In a certain factory producing tyres of bike, there is a small change of 1 in 500 tyres to be defective. The tyres are supplied 10 tyres of 10. Using poisson distribution, calculate the approximate number of lots containing i) No defective ii) One defective iii) In consignment of 1000 lots.*

Solution : Here 1 in 500 tyres to be defective.

Hence $p = \frac{1}{500}$ $n = 10$

$$z = np = 10 \times \frac{1}{500} = \frac{1}{50}$$

By poission distribution, we know that

$$p(z) = \frac{e^{-z} z^r}{r!}$$

i) Probability of no defective tyre :

$$p(r = 0) = \frac{e^{-0.02}(0.02)^0}{0!}$$

$$= 0.980$$

∴ Number of lots containing no defective

$$= 1000 \times 0.980$$

$$= 980 \text{ lots}$$

ii) Probability of one defective

$$p(r = 1) = \frac{e^{-0.02}(0.02)^1}{1!}$$

$$= 0.9802 \times 0.02 = 0.019$$

∴ Number of lots containing one defectives

$$= 1000 \times 0.019$$

$$= 19 \text{ lots}$$

Example 3.31.7 *Fit poissson's distribution to following data and calculate theoretical frequencies.*

x	0	1	2	3	4
f	122	60	15	2	1

Solution :

x_i	f_i	$x_i f_i$
0	122	0
1	60	60
2	15	30
3	2	6
4	1	4
	$N = \Sigma f_i = 200$	$N = \Sigma f_i x_i = 100$

Let z = Mean of the distribution

$$= \frac{\Sigma x_i f_i}{N = \Sigma f_i} = \frac{100}{200} = 0.5$$

By poisson distribution,

$$p(r) = \frac{e^{-z} z^r}{r!} ; r = 0, 1, 2, 3, 4$$

$$\therefore \quad p(0) = \frac{e^{-0.5} 0.5^0}{0!} = e^{-0.5} = 0.6065$$

$$p(1) = \frac{e^{-0.5} 0.5^1}{1!} = 0.6065 \times 0.5 = 0.3033$$

$$p(2) = \frac{e^{-0.5} 0.5^2}{2!} = 0.6065 \times 0.125 = 0.0758$$

$$p(3) = \frac{e^{-0.5} (0.5)^3}{3!} = 0.6065 \times 0.0208 = 0.0126$$

$$p(4) = \frac{e^{-0.5} 0.5^4}{4!} = 0.6065 \times 0.0026 = 0.0016$$

Theoretical frequencies are

$$p(0) \times 200 = 121$$

$$p(1) \times 200 = 161$$

$$p(2) \times 200 = 15$$

$$p(3) \times 200 = 3$$

$$p(4) \times 200 = 1$$

Example 3.31.8 *If the probability that an individual suffers a bad reaction from a certain injection is 0.001. Determine the probability out of 2000 peoples i) Exactly 3 ii) More than 2 will suffer a bad.*

Solution :

Here $p = 0.001$

$n = 2000$

$z = np = 0.001 \times 2000 = 2$

By poisson distribution,

$$p(r) = \frac{e^{-z} z^r}{r!}$$

i) Probability that exactly 3 suffers a bad reaction

$$p(r = 3) = \frac{e^{-2} (2)^3}{3!}$$

$$= 0.136 \times \frac{8}{6}$$

$$= 0.1813$$

ii) Probability that more than two will suffer a bad reaction

$$p(\text{more than } 2) = p(r=3) + p(r=4) + \ldots + p(r=2000)$$
$$= 1 - [p(r=0) + p(r=1) + p(r=2)]$$
$$= 1 - \left[\frac{e^{-2}(2)^0}{0!} + \frac{e^{-2}(2)^1}{1!} + \frac{e^{-2}(2)^2}{2!}\right]$$
$$= 1 - e^{-2}[1+2+2]$$
$$= 0.32$$

Example 3.31.9 *During working hours on an average 5 phone calls are coming in to a call centre within an Hr. Using poisson distribution, find the probability that during a particular Hr., there will be at most two phone calls.*

Solution : Given that 5 phone calls are coming into a call centre.

$$\therefore \quad z = \frac{5}{60} = 0.08$$

$\therefore$ The probability of getting at most two phone calls is

$$= p(r \le 2)$$
$$= p(r=0) + p(r=1) + p(r=2)$$
$$= \frac{e^{-z}(z)^0}{0!} + \frac{e^{-z}(z)^1}{1!} + \frac{e^{-z}(z)^2}{2!}$$
$$= e^{-z}\left[1 + z + \frac{z^2}{2}\right]$$
$$= e^{-0.08}\left[1 + 0.08 + \frac{(0.08)^2}{2}\right]$$
$$= 0.9999$$

Example 3.31.10 *When mean $a = 2$, standard derivation $\sigma = 4$ find the probabilities of the following intervals i) $4.43 \le x \le 7.29$ ii) $-0.43 \le x \le 5.39$*

Solution : i) $4.43 \le x \le 7.29$

Let $$z = \frac{x-a}{\sigma}$$

i) When $4.43 \le x \le 7.29$

$$z_1 = \frac{4.43-2}{4} = 0.607 \approx 0.61$$

$$z_2 = \frac{7.29-2}{4} = 1.322 \approx 1.32$$

$$p(4.43 \le x \le 7.29) = p(0.61 \le x \le 1.32)$$

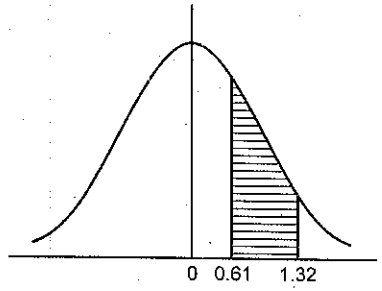


Fig. 3.31.2

A_1 = Area corresponding to $z_1 = 0.61$

= value of 0.61 in Normal table (0.6 row and 0.01 col.)

$= 0.7291 - 0.5$

$= 0.2291$

A_2 = Area corresponding to $z_2 = 1.32$

$= 0.9066 - 0.5 = 0.4066$

$\therefore$ Required probability $= A_2 - A_1$

i.e. $p(4.43 \le x \le 7.29) = 0.1775$

ii) $-0.43 \le x \le 5.39$

$$z_1 = \frac{-0.43-2}{4} = -0.61$$

$$z_2 = \frac{5.39-2}{4} = 0.85$$

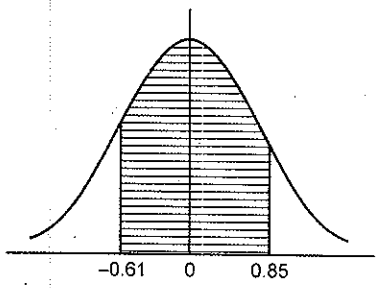


Fig. 3.31.3

$p(-0.43 \le x \le 5.39) = p(-0.61 \le z \le 0.85)$

A_1 = Area corresponding $z_1 = -0.61$ (≈ 0.61) due to symmetry

$= 0.2291$

A_2 = Area corresponding to $z_2 = 0.85 = 0.3023$

$\therefore$ Required probability $= A_1 + A_2$

$= 0.2291 + 0.3023 = 0.5314$

Example 3.31.11 *MNC company conducted 1000 candidates aptitude test. The average score is 45 and the standard deviation of score is 25. Assuming normal distribution for the result. Find i) The number of candidates whose scores exceed 60.*

ii) The number of candidates whose scores lies between 30 and 60.

Solution : Here mean a = 45 and standard deviation $\sigma = 25$

$\therefore \quad z = \frac{x - a}{\sigma} = \frac{x - 45}{25}$

i) Number of candidates whose score exceeds 60

When $x = 60, z = \frac{60 - 45}{25} = 0.6$

$\therefore \quad p(x > 60) = p(z > 0.6)$

$= 0.5 - p(z = 0.6)$

$= 0.5 - 0.2257 = 0.2743$

$\therefore$ No. of candidates who exceed 60 = 1000 × 0.274 = 274

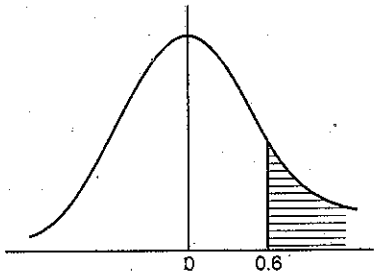


Fig. 3.31.4

ii) The number of candidates who score between 30 and 60 i.e. $p(30 < x < 60)$

$z_1 = \frac{x_1 - a}{\sigma} = \frac{30 - 45}{25} = -0.6$

$z_2 = \frac{x_2 - a}{\sigma} = \frac{60 - 45}{25} = 0.6$

$p(30 < x < 60) = p(-0.6 < z < 0.6)$

$= p(0 < z < 0.6) + p(0 < z < 0.6)$

$= 0.5257 + 0.5257$

$= 1.0514$

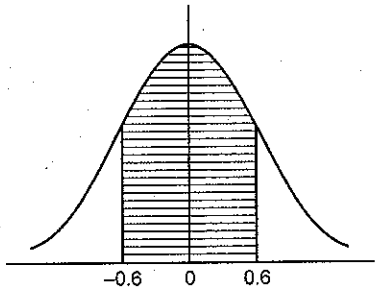


Fig. 3.31.5

No. of candidates who score between 30 and 60 = 1000 × 1.0514 = 1051.4

274 candidates score exceeds 60

1051 candidates score is between 30 and 60.

Example 3.31.12 *In a certain company install 2000 LED bulbs on each floor. If LED bulbs have average life of 1000 burning hours with standard deviation of 200 hours. Using normal distribution find What number of LED bulbs might be expected to fail in 700 Hours.*

Solution : Given : Mean a = 1000, Standard deviation $\sigma = 200$

$x = 700, z = \frac{700 - 1000}{200}$

$= -1.5$

$\therefore \quad p(x < 700) = p(z < -1.5)$

$= p(z > 1.5)$

$= 0.5 - 0.4332$

$= 0.0668$

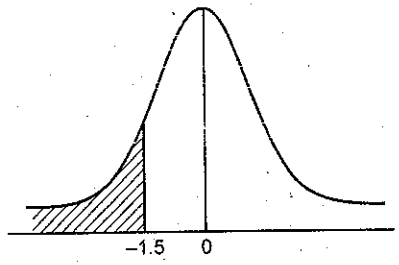


Fig. 3.31.6

Number of LED bulbs might be expected to fail in 700 Hrs. = 0.0668 × 2000 = 1336 bulbs

Example 3.31.13 *In a sample of 1000 cases the mean of a certain test is 14 and standard deviation is 2.5. Assuming the distribution to be normal find*

i) How many students score between 12 and 15 ? ii) How many score below 8 ? iii) How many score 16 ?

Solution : Here n = 1000, Mean a = 14, Standard deviation (σ) = 2

$z = \frac{x - a}{\sigma} = \frac{x - 14}{2.5}$

i) Students score between 12 and 15 :

$$x_1 = 12, z_1 = \frac{12-14}{2.5} = -0.8$$

$$x_2 = 15, z_2 = \frac{15-14}{2.5} = 0.4$$

$$p(2 < x < 13) = p(-0.8 < z < 0.4)$$

A_1 = Area corresponds to $(z_1 = -0.8) = 0.2881$

A_2 = Area corresponds to $(z_2 = 0.4) = 0.1554$

$$p(12 < x < 15) = A_1 + A_2$$

$$= 0.2881 + 0.1554 = 0.4435$$

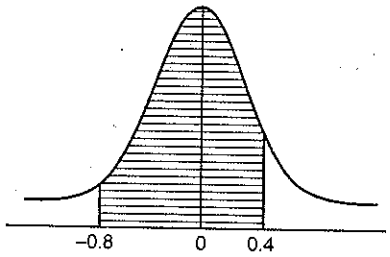


Fig. 3.31.7

Required number of students = 1000×0.4435

$= 443$

ii) Students score below 8

Here $x = 8$

$\therefore$ $$z = \frac{8-14}{2.5} = -2.4$$

$$p(x < 8) = p(z < -2.4)$$

$$= 0.5 - p(0 < z(2.4))$$

$$= 0.5 - 0.4918$$

$$p(x < 8) = 0.0082$$

$\therefore$ Required number of students = 1000×0.0082

$= 8.2$

$= 8$

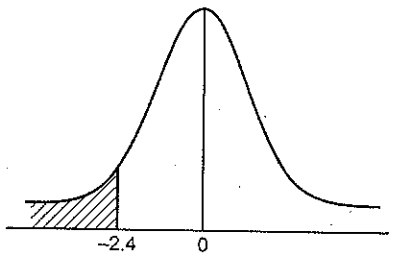


Fig. 3.31.8

iii) Students score 16 : 16 means between 15.5 and 16.5

$\therefore$ $$x_1 = 15.5, z_1 = \frac{15.5-14}{2.5} = 0.6$$

$x2 = 16.5,$

$$z_2 = \frac{16.5-14}{2.5} = 1$$

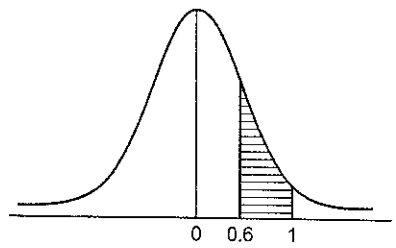


Fig. 3.31.9

$$p(15.5 < x < 16.5) = p(0.6 < z < 1)$$

$$= p(0 < z < 1) - p(0 < z < 0.6)$$

A_1 = Area corresponds to $(z_1 = 0.6)$

$= 0.2257$

A_2 = Area corresponds to $(z_2 = 1)$

$= 0.3413$

$$p(13.5 < x < 16.5) = A_2 - A_1$$

$$= 0.3413 - 0.2257$$

$$= 0.1156$$

$\therefore$ Number of students score 16 = 1000×0.1156

$= 115.6 = \mathbf{116}$

Example 3.31.14 *The table below gives the number of customers visit the certain company on various days of week.*

Days	Sun	Mon	Tue	Wed	Thurs	Fri	Sat
Number of Customers	6	4	9	7	8	10	12

Test at 5 % of level of significance whether customer visit are uniform distributed over the days.

Solution : Let H_0 : The distribution of customers over the days is uniform

$$p_i = \frac{1}{7}$$

$$N = \Sigma o_i = 56$$

$\therefore$ $E(X)$ = Expected customers

$$= N \cdot p_i = 56 \times \frac{1}{7} = 8$$

x_i	o_i	e_i	$O_i - e_i$	$(O_i - e_i)^2$
1	6	8	– 2	4
2	4	8	– 4	16
3	9	8	1	1
4	7	8	– 1	1
5	8	8	0	0
6	10	8	2	4
7	12	8	4	16
				$\Sigma = 42$

Here p = 0, k = 7.

$$X^2_{K-p-1} = \sum_{i=0}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$X^2_6 = \sum_{i=0}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{42}{8}$$

$$X^2_6 = 5.25 \text{ (observed value)}$$

$$X^2_{6.0.05} = 15.592 \text{ (Table value)}$$

$\therefore$ $X^2_6 < X^2_{6.0.05}$

i.e. Observed value < Table value

$\therefore$ We accept H_0

$\therefore$ The customer visit are distributed uniformly over the days of week.

Example 3.31.15 *An university library utilizes four windows to issue and return of book. On a particular day 800 students were observed inside the library. They were given service at the different windows as following.*

Window Number	1	2	3	4
Number of Students	150	250	170	230

Test whether the students are uniformly distributed over the window.

Solution : Let, H_0 : The distribution of students over 4 windows is uniform

As the distribution is even 1

$\therefore$ $p_i = \frac{1}{4}$

E(x) = Expected frequency

$= N \cdot p_i$

$$= 800 \times \frac{1}{4} = 200$$

x_i	o_i	e_i	$O_i - e_i$	$(O_i - e_i)^2$
1	150	200	– 50	2500
2	250	200	50	2500
3	170	200	– 30	900
4	230	200	30	900
				$\Sigma = 6800$

Here k = 41, p = 0

$\therefore$ $$X^2_{k-p-1} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{6800}{2000}$$

$$X^2_3 = 34 \text{ (observed value)}$$

$$X^2_{3,0.05} = 7.815 \text{ (Table value)}$$

As $X^2_3 > X^2_{3.0.05}$

Calculated value > Table value,

$\therefore$ We reject H_0

$\therefore$ The students are not uniformly distributed over the windows.

Exercise

1. *Calculate first four moments about the mean for the distribution*

x	1	2	3	4	5	6	7	8	9
f	1	6	13	25	30	22	9	5	2

Also find coefficients of skewness and kurtosis.

[Ans. : $\mu_1 = 0$, $\mu_2 = 2.49$, $\mu_3 = 0.68$, $\mu_4 = 18.26$, $\beta_1 = 0.029$, $\beta_2 = 2.95$]

2. *The first four moments of a distribution at x = 2 are 1, 2.5, 5.5 and 16. Calculate the first four moments about mean.* **[Ans. : $\mu_1 = 0$, $\mu_2 = 1.5$, $\mu_3 = 0$, $\mu_4 = 6$]**

3. *Obtain the regression line y on x for the data*

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Also estimate value of y for x = 10 **[Ans. : y = 1.3399 x – 1.9606, y = 11.4384]**

4. For the lines of regression $8x - 10y = -66$ and $4x - 18y = 214$. Find i) Mean of x and y, ii) Coefficient of correlation r. **[Ans. : $\bar{x} = 13$, $\bar{y} = 17$, $r = \pm 0.6$]**
5. The following equations represents regression lines equation
   $y = 0.516x + 33.73$
   $x = 0.512y + 32.52$
   Find i) Mean of x and y, ii) Coefficient of correlation (r)
   [Ans. : $\bar{x} = 67$, $\bar{y} = 68.6y$, $r = 0.514$]

Problems on Binomial Distribution

6. An unbaised coin is thrown 10 times. Find the probability of getting i) exactly 6 heads ii) at least 6 heads. **[Ans. : i) $\frac{105}{512}$ ii) $\frac{193}{512}$]**
7. Probability of man aged 60 years will live for 70 years is 1/10. Find the probability of 5 men selected at random 2 will live for 70 years. **[Ans. : 0.05467]**
8. Out of 800 families with 4 children each how many families would be expected to have i) 2 boys and 2 girls ii) At least one boy iii) No girl iv) At most two girl
   Assume equal probability for boys and girls. **[Ans. : i) 300 ii) 750 iii) 50 iv) 550]**
9. A group of 20 aeroplane are sent on an operational flight. The chances that the aeroplane fails to return from the fight is 5 %. Determine the probability then i) No plane returns ii) At most 3 planes do not return. **[Ans. : i) 0.3585 ii) 0.9841]**
10. Team A has a probability of $\frac{2}{3}$ of winning whenever the team plays a particular game.
    If team A plays 4 games, find the probability that the team coins : i) Exactly two games, ii) At least two games. **[Ans. : i) 0.2962 ii) 0.8889]**

Problems on Poisson Distribution

11. Find the probability that atmost 5 defective fuses will be found in a box of 200 fuses if 2 % of such fuses are defective. **[Ans. : 0.735]**
12. A car hire firm has 2 cars which it hires out day by day. The member of demands for the car on each day is distributed as poisson's distributikn with parameter 15.0. Calculate a) the probability of days on which neither car is used and b) for days on which demand is refused. **[Ans. : (a) 0.22 (b) 0.2025]**
13. A manufacturer knows that the razor blades he makes contain on the average 0.5 % of defectives. He picks them in packets of 5. What is the probability that a packet picked at random will contain 3 or more faulty blades ? **[Ans. : 0.00000255]**
14. There are 300 misprints are distributed randomly throughout a book of 500 pages. Find the probability that the gives page contains i) Exactly two mispoints, ii) Two or more mispoints. **[Ans. : i) 0.1 ii) 0.122]**
15. Using Poisson's distribution, find the probability that the ace of spades will be drawn from a pack of well shaffled carts at least once in 104 consecutive trials. **[Ans. : 0.864]**

Problems on Normal Distribution

16. In a certain examination test 200 students appeared in subject of statistics. Average marks obtained were 50 % with standard deviation 5 %. How many students do you expect to obtain more than 60 % of marks, supposing that marks are distributed normally. **[Ans. : 46 students approximately]**
17. In a distribution exactly normal 7 % of the items are under 35 and 89 % are under 63. Find the mean and standard deviation of the distribution
    [Ans. : mean a = 50.3, $\sigma = 10.33$]
18. A normal distribution has a mean 15.73 and a standard deviation of 2.08. Find % of cases that fall between 17.81 and 13.65. **[Ans. : 68.26]**
19. A fair coin is tossed 600 times. Using normal distribution. Find the probability of getting i) Number of heads less than 270, ii) Number of heads between 280 to 360.
    [Ans. : i) 0.00714, ii) 0.4845]
20. The average test marks in particular class is 59 and S.D. is 9. If the marks are normally distributed, how many students in a class of 70 received i) Marks below 50 ? ii) More than 70 ? **[Ans. : i) 24 ii) 8]**

Problems on Ch-square Distribution

21. The table below gives the number 07 accidents that occurred in certain country on various days of week

Days	Mon	Tue	Wed	Thurs	Fri	Sat
Accidents	120	130	110	115	135	110

    Test 5 % of level of significance whether the accidents are uniformly distributed over the days. **[Ans. : $X_5^2 = 4.58$, $X^2_{5,0.05} = 11.07$ H_0 accepted]**
22. The demand for a particular spare part in a factory was found to vary from day to day. In a sample study. The following information was obtained.

Days	Mon	Tue	Wed	Thurs	Fri	Sat
No. of parts demanded	1124	1125	1110	1120	1126	1115

    [Ans. : $X_4^2 = 3.9222$, $X^2_{4,0.05} = 9.488$ H_0 is accepted]

Multiple Choice Questions

Q.1 There are two types of measures of dispersion : _____.

[a] nominal measure of dispersion and real measure of dispersion

[b] nominal measure of dispersion and relative measure of dispersion

[c] absolute measure of dispersion and relative measure of dispersion

[d] real measure of dispersion and relative measure of dispersion

Q.2 The _____ is the easiest measure of dispersion to calculate.

[a] range [b] mean absolute deviation

[c] standard deviation [d] variance

Q.3 Quartile deviation is calculated by _____.

[a] $\frac{Q_3 + Q_1}{2}$ [b] $\frac{Q_3 - Q_1}{2}$

[c] $\frac{Q_3 + Q_2}{2}$ [d] $\frac{Q_3 - Q_2}{2}$

Q.4 The numerical value of the standard deviation can never be _____.

[a] none [b] zero

[c] negative [d] larger than the variance

Q.5 Standard deviation is also called _____.

[a] mean square deviation [b] mean deviation

[c] square deviation [d] root mean square deviation

Q.6 Square of standard deviation is called _____.

[a] Mean Square Deviation [b] Mean Deviation

[c] Variance [d] Root Mean Square Deviation

Q.7 Calculate mean deviation from median and its coefficient from the following data : 110, 150, 100, 90, 160, 200, 140 :

[a] 34.28 [b] 44.28

[c] 32.14 [d] 33.55

Q.8 The ratio of the standard deviation to the arithmetic mean expressed as a percentage is called : _____.

[a] Coefficient of standard deviation [b] Coefficient of skewness

[c] Coefficient of kurtosis [d] Coefficient of variation

Q.9 The moments about mean are called _____.

[a] raw moments [b] central moments

[c] moments about origin [d] all of the above

Q.10 If the distribution is negatively skewed, then the _____.

[a] mean is more than the mode [b] median is at right to the mode

[c] mean is at right to the median [d] mean is less than the mode

Q.11 In a frequency distribution of a variable x, mean = 34, median = 28. The distribution is _____.

[a] positively skewed [b] negatively skewed

[c] mesokurtic [d] platykurtic

Q.12 In a negatively skewed distribution _____.

[a] mean > mode > median [b] mode > median > mean

[c] mean > median > mode [d] mode < median > mean

Q.13 For a distribution, mean is 42, median is 42.8 and mode is 44. The distribution is _____.

[a] normal [b] positively skewed

[c] negatively skewed [d] mesokurtic

Q.14 The second moment (μ_2) and fourth moment (μ_4) about mean are 4 and 18 respectively. What is value of skewness β_2 ?

[a] 0.875 [b] 1.125

[c] 1.25 [d] 4.5

Q.15 The first four moments of a distribution about the origin are – 1.5, 17, – 30 and 108. The third moment about mean is :

[a] 39.75 [b] 41.75

[c] 40.75 [d] 42.75

Q.16 If μ_r be the r^{th} order central moments of a population then μ_0, μ_1 and μ_2 are given by ____.

[a] 0, 1, σ^2 [b] 1, 0, σ^2

[c] 1, 1, σ [d] 1, 0, σ

Q.17 The value of β_2 can be _____.

[a] less than 3 [b] greater than 3

[c] equal to 3 [d] All of the above

Q.18 The method of least squares find the best fit line that _____ the error between observed and estimated points on the line.

[a] maximizes [b] minimizes

[c] reduces to zero [d] approaches to infinity

Q.19 The slope of the regression line Y on X is also called _____.

[a] the correlation coefficient of X on Y

[b] the correlation coefficient of Y on X

[c] the regression coefficient of X on Y

[d] the regression coefficient of Y on X

Q.20 If the equation of the regression line X = 4 then _____.

[a] the line passes through the origin

[b] the line passes through (0, 4)

[c] the line is parallel to Y axis

[d] the line is parallel to X axis

Q.21 If the scatter diagram is drawn the scatter points lie on a straight line then it indicates _____.

[a] skewness [b] perfect correlation

[c] no correlation [d] none of the above

Q.22 If β_{yx} is positive then β_{xy} will be _____.

[a] Positive [b] Negative

[c] Zero [d] 1

Q.23 Relation between r, β_{xy} and β_{yx} is _____.

[a] $r = \beta_{xy} + \beta_{yx}$ [b] $r = \frac{\beta_{xy}}{\beta_{yx}}$

[c] $r = \sqrt{\beta_{xy} \cdot \beta_{yx}}$ [d] $r = \beta_{xy} \cdot \beta_{yx}$

Q.24 If r = 0.8, β_{yx} = 1.6 then β_{xy} = ?

[a] 0.2 [b] 0.4

[c] 1 [d] 0

Q.25 The value of coefficient of correlation _____.

[a] between 1 and 0 [b] between – 1 and 0

[c] between – 1 and + 1 [d] Equals to 1

Q.26 Line of regression y on x is _____.

[a] $y + \bar{y} = r\frac{\sigma x}{\sigma y}x + \bar{x}$ [b] $y - \bar{y} = r\frac{\sigma y}{\sigma x}(x - \bar{x})$

[c] $x - \bar{x} = r\frac{\sigma x}{\sigma y}(y - \bar{y})$ [d] $y - \bar{y} = r\frac{\sigma x}{\sigma y}(x - \bar{x})$

Q.27 Line of regression x on y _____.

[a] $x - \bar{x} = r\frac{\sigma y}{\sigma x}(y - \bar{y})$ [b] $x - \bar{x} = r\frac{\sigma x}{\sigma y}(y - \bar{y})$

[c] $x + \bar{x} = r\frac{\sigma x}{\sigma y}(y + \bar{y})$ [d] $x - \bar{x} = r\frac{\sigma y}{\sigma x}(y + \bar{y})$

Q.28 If the two regression coefficients are $-\frac{7}{16}$ and $-\frac{3}{4}$ then correlation coefficient is _____.

[a] – 0.862 [b] – 0.3281

[c] 0.5 [d] 0.4

Q.29 Line of regression y on x is $8x - 10y + 66 = 0$ and Line of regression x on y is $40x - 18y = 214$. Then values of $\bar{x}$ and $\bar{y}$ are _____.

[a] $\bar{x} = 12, \bar{y} = 15$ [b] $\bar{x} = 10, \bar{y} = 11$

[c] $\bar{x} = 13, \bar{y} = 17$ [d] $\bar{x} = 9, \bar{y} = 8$

Q.30 Line of regression y on x is $8x - 10y + 66 = 0$. Line of regression x on y is $40x - 18 = 214$. The value of variance of y is 16. The standard deviation of x is equal to _____.

[a] 3 [b] 2

[c] 6 [d] 7

Answer Keys for Multiple Choice Questions :

Q.1	c	Q.2	a	Q.3	b	Q.4	c	Q.5	d
Q.6	c	Q.7	b	Q.8	d	Q.9	b	Q.10	d
Q.11	a	Q.12	b	Q.13	c	Q.14	b	Q.15	a
Q.16	d	Q.17	d	Q.18	b	Q.19	d	Q.20	c
Q.21	b	Q.22	a	Q.23	c	Q.24	b	Q.25	c
Q.26	b	Q.27	b	Q.28	b	Q.29	c	Q.30	a

❑❑❑

Unit IV

4 Random Variables and Probability Distributions

Syllabus

Random Variables and Distribution Functions : *Random Variable, Distribution Function, Properties of Distribution Function, Discrete Random Variable, Probability Mass Function, Discrete Distribution Function, Continuous Random Variable, Probability Density Function*

Theoretical Discrete Distributions : *Bernoulli Distribution, Binomial Distribution, Mean Deviation about Mean of Binomial Distribution, Mode of Binomial Distribution, Additive Property of Binomial Distribution, Characteristic Function of Binomial Distribution, Cumulants of Binomial Distribution, Poisson Distribution, The Poisson Process, Geometric Distribution...*

Contents

4.1 Introduction

- Already in theory of probability we have studied about random experiments and its corresponding sample space. Now our interest is not to construct the sample space of random experiment but in some number obtained from the outcomes.
- For example, in case of the experiment of tossing a coin twice, our interest may by only in the number of heads when a coin are tossed two times. So in probability problems it is convenient to think of a variable and consider outcomes as the values, the variable takes such a variable is called a random variable.

4.2 Random Variable

- A random variable is a real valued function defined on the sample space of a random experiment. In other words, the random variable X can be considered as a function that maps all the elements in the sample space S into points on the real number.

  Thus $X : S \rightarrow R$ is a random variable.

  Example :

  Consider the experiment of tossing two fair coins.

  The sample space S = {HH, HT, TH, TT}. There are 4 possible outcomes.

  But if we are interested in the number of heads when two coins are tossed, then there are only 3 different possible values from 0 to 2.

  1) Getting no head, 2) Getting one head 3) Getting two heads.

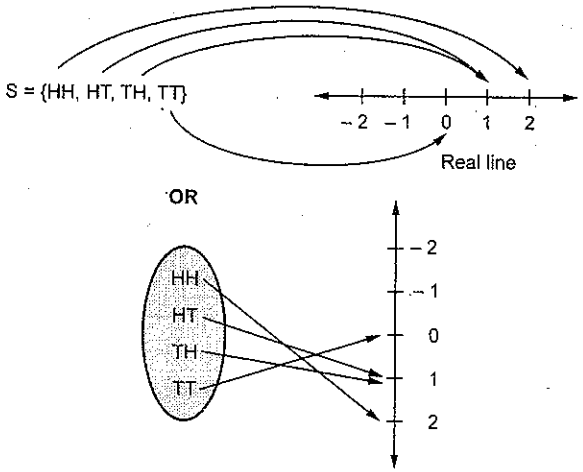


Fig. 4.2.1

- In simple words a variable used to denote the real value of the outcome of a random experiment is a random variable.
- A random variable X may be finite or countably finite or countably infinite or uncountably infinite.

4.2.1 Types of Random Variables

There are two types of random variables.

1) **Discrete random variable :**

   If the random variable X takes only a finite or countably infinite values, then it is called discrete random variable. Examples of discrete random variables are the number of children in a family, the number of cars sold by a dealer, number of stars in the sky and so on.

2) **Continuous random variables :**

   If the random variable X takes uncountably infinite values, then it is called continuous random variable.

   Examples of continuous random variables are height of person selected at random, the age of person selected at random, because x can take any value between a specified range.

4.3 Probability Mass Function (P.M.F)

- If a discrete random variable X takes values $x_1, x_2, x_3 \ldots x_n$ with corresponding probabilities $p_1, p_2, p_3, \ldots p_n$ respectively where $\sum_{i=1}^{n} p_i = 1$ then the distribution .

x	x_1	x_2	x_3	...	x_n
P(x)	p_1	p_2	p_3	...	p_n

is called discrete probability distribution.

In discrete probability distribution the probabilities P_i satisfying the conditions that

i) $0 \leq p_i \leq 1$ for all i ii) $\Sigma p_i = 1$

Then p_i is called probability mass function.

Example :

Suppose x denotes the outcome of the throw of a die. Then X takes values 1, 2, 3, 4, 5, 6 each with probability 1/6.

∴ The discrete probability distribution is

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

Properties : In p.m.f.

i) $0 \le P(x_i) \le 1$ for all i

ii) $\sum_{i=1}^{n} P(x_i) = 1$

iii) $P(x_i \le x \le x_{i+j}) = P(x_i) + P(x_{i+1}) + \ldots + P(x_{i+j})$

iv) Mean or expected value :

The mean or expected value of X is denoted by μ or E(X) and is defined as.

$$\mu = E(X) = \frac{\sum p_i x_i}{\sum p_i} = \sum p_i x_i$$

v) The variance of X is denoted by V(X) and $V(X) = \sigma^2 = \sum p_i (x_i - \mu)^2 = E(X^2) - [E(X)]^2$

where σ = Standard deviation $= +\sqrt{V(X)}$

vi) Cumulative distribution function (c.d.f.) of a discrete random variable X is denoted by F(x) and is defined as,

$$F(x) = P[X \le x] = \sum_{X_i \le x} P(X = x_i)$$

4.4 Probability Density Function (P.D.F.)

- If x is a continuous random variable and the probability of that x will be given in interval (a, b), then the probability distribution is called a continuous probability distribution.
- Usually a continuous probability distribution is given in the form of function f(x).
- The probability function f(x). The probability function f(x) which has the following properties.

i) $f(x) \ge 0$ for all x

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ is called a probability density function.

Example : For a continuous random variable x, the probability density function defined by

$$f(x) = \begin{cases} \frac{x^2}{9} & 0 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

Properties : In p.d.f.

i) $f(x) \ge 0$ for all x

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

iii) $P(a \le x \le b) = \int_{a}^{b} f(x)\,dx$

iv) Mean or Expected value :

$$E(x) = \mu = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

v) Variance $= V(x) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = E[x^2] - [E(x)]^2$

4.5 Solved Examples

Example 4.5.1 *The probability distribution of X is as follows*

X	0	1	2	3	4
P(X = x)	0.1	k	2k	2k	k

is probability mass function.

Find i) k ii) P [x < 2] iii) P[X ≥ 3] iv) P [1 ≤ x ≤ 3]

Solution : i) The table gives a probability distribution.

$\therefore P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) = 1$

$\therefore \quad 0.1 + k + 2k + 2k + k = 1$

$\therefore \quad 0.1 + 6k = 1$

$\therefore \quad 6k = 0.9$

$\therefore \quad k = 0.15$

ii) $P[X < 2] = P[X = 0] + P[X = 1] = 0.1 + k = 0.1 + 0.15 = 0.25$

iii) $P[X \geq 3] = P[X = 3] + P[X = 4] = 2k + k = 3k = 3(0.15) = 0.45$

iv) $P[1 \leq X \leq 3] = P[X = 1] + P[X = 2] + P[X = 3] = k + 2k + 2k$

$= 5k$

$= 5(0.15)$

$= 0.75$

Example 4.5.2 *A random variable X the following probability distribution.*

X	1	2	3	4	5	6	7
P(X)	k	2k	3k	k^2	$k^2 + k$	$2k^2$	$4k^2$

i) Find k ii) P(x > 5) iii) P (0 ≤ x ≤ 5)

Solution : i) From probability distribution

$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) = 1$

$\therefore k + 2k + 3k + k^2 + k^2 + k + 2k^2 + 4k^2 = 1$

$\therefore 7k + 8k^2 = 1$

$\therefore 8k^2 + 7k - 1 = 0$

$\therefore 8k^2 + 8k - k - 1 = 0$

$(k + 1)(8k - 1) = 0$

$\therefore k = -1$ or $k = 1/8$

Since $P(x = 1) = k \quad \therefore k > 0$

$\therefore$ The value of $k = 1/8$

ii) $P(X > 5) = P(X = 6) + P(X = 7)$

$= 2k^2 + 4k^2 = 6k^2$

$= \frac{6}{64}$

$= \frac{3}{32}$

iii) $P(0 \leq X \leq 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$

OR $P(0 \leq X \leq 5) = 1 - P(X > 5) = 1 - \frac{3}{32} = \frac{29}{32}$

Example 4.5.3 *Let a random variable x takes values –2, –1, 0, 1, 2 such that P(X = –2) = P(X = –1) = P(X = 1) = P(X = 2) and P(X < 0) = P(X = 0) = P(X > 0). Determine the probability mass function of X.*

Solution : Given random variable X = –2, –1, 0, 1, 2

By definition of probability mass function

$P(X = -2) + P(X = -1) + P(X = 1) + P(X = 2) + P(X = 0) = 1$ ... (I)

Given, $P(X = -2) = P(X = -1) = P(X = 1) = P(X = 2) = k$ (say)

Also given $P(X < 0) = P(X = 0)$

$\therefore P(X = 0) = P(X = -1) + P(X = -2)$

$P(X = 0) = 2K$ ...(II)

From (I) and (II)

$k + k + 2k + k + k = 1$

$\therefore 6k = 1$

$\therefore k = \frac{1}{6}$

Hence the probability distribution (p.m.f.) is

X	–2	–1	0	1	2
P(X)	1/6	1/6	2/6	1/6	1/6

Example 4.5.4 *Find the value of k if the following function f(x) is a p.d.f.*

$$f(x) = \begin{cases} k x^2 (1 - x^3), & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Solution : Since $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\therefore \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$$

$$\therefore 0 + \int_{0}^{1} kx^2 (1 - x^3)\,dx + 0 = 1$$

$$\therefore \quad k\int_0^1 (x^2 - x^5)\,dx = 1$$

$$\therefore \quad k\left[\frac{x^3}{3} - \frac{x^6}{6}\right]_0^1 = 1$$

$$\therefore \quad k\left[\frac{1}{3} - \frac{1}{6}\right] = 1$$

$$\therefore \quad k\left(\frac{1}{6}\right) = 1$$

$$\therefore \quad k = 6$$

Example 4.5.5 *In a continuous distribution density function $f(x) = k\,x\,(2-x)$, $0 < x < 2$. Find the value of k, mean and c variance.*

Solution : Given p.d.f.

$$f(x) = \begin{cases} k\,x\,(2-x), & 0 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

i) Since, $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\therefore \quad \int_0^2 k\,x\,(2-x)\,dx = 1$$

$$\therefore \quad k\int_0^2 (2x - x^2)\,dx = 1$$

$$\therefore \quad k\left[x^2 - \frac{x^3}{3}\right]_0^2 = 1$$

$$\therefore \quad k\left(4 - \frac{8}{3}\right) = 1$$

$$\therefore \quad k\left(\frac{4}{3}\right) = 1$$

$$\therefore \quad k = \frac{3}{4}$$

$$\therefore \quad f(x) = \frac{3}{4}x\,(2-x), \quad 0 < x < 2$$

ii) $$\text{Mean} = E(x) = \int_0^2 x\,f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 x \cdot x(2-x)\,dx$$

$$= \frac{3}{4}\int_0^2 (2x^2 - x^3)\,dx$$

$$= \frac{3}{4}\left[\frac{2x^3}{3} - \frac{x^4}{4}\right]_0^2$$

$$= \frac{3}{4}\left[\frac{16}{3} - \frac{16}{4}\right]$$

$$= \frac{3}{4}\left(\frac{4}{3}\right)$$

$$\therefore \quad E(x) = 1$$

iii) $$\text{Variance} = V(x) = E(x^2) - [E(x)]^2$$

$$\because \quad E(x^2) = \int_0^2 x^2\,f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 x^3\,(2-x)\,dx$$

$$= \frac{3}{4}\int_0^2 (2x^3 - x^4)\,dx$$

$$= \frac{3}{4}\left[\frac{2x^4}{4} - \frac{x^5}{5}\right]_0^2$$

$$= \frac{3}{4}\left[\frac{16}{2} - \frac{32}{5}\right]$$

$$E(x^2) = \frac{6}{5}$$

$$\therefore \quad V(x) = E(x^2) - [E(x)]^2 = \frac{6}{5} - 1 = \frac{1}{5}$$

Example 4.5.6 *The probability density function $f(x) = \lambda x e^{-\lambda x}$, $x > 0$. Find : i) the value of λ. ii) $P(2 < x < 5)$*

Solution : Given p.d.f.

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

i) Since, $f(x) \geq 0$ for all x

and $\int_0^\infty f(x)\,dx = 1$

$$= \int_0^\infty \lambda\, x e^{-x}\,dx = 1$$

$$= \lambda \int_0^\infty x e^{-x}\,dx = 1$$

$$= \lambda\sqrt{2} = 1$$ (Since by definition of Gamm function $\int_0^\infty e^{-x} x^{n-1}\,dx = \sqrt{n}$ and $\sqrt{2} = 1! = 1$)

$$= \lambda (1!) = 1$$

$$\lambda = 1$$

ii) $$P(2 < x < 5) = \int_2^5 f(x)\,dx$$

$$= \int_2^5 x e^{-x}\,dx$$

$$= \left[\frac{x e^{-x}}{-1} - e^{-x}\right]_2^5$$

$$= [(-5e^{-5} - e^{-5}) - (-2e^{-2} - e^{-2})]$$

$$= -6e^{-5} + 3e^{-2}$$

$\therefore$ $$P(2 < x < 5) = 3e^{-2} - 6e^{-5}$$

4.6 Introduction

- Frequency distributions can be classified as,

1) Observed frequency distributions
2) Theoretical or expected frequency distributions

- Observed frequency distributions are based on actual observations and experimentation.
- Theoretical distributions are derived mathematically with some assumptions.
- There are many types of theoretical distribution but we shall consider only three.

1) Bionomial (Bernoulli's) distribution
2) Poisson's distribution
3) Geometric distribution.

4.7 Binomial (Bernoulli's) Distribution

If (i) An experiment results only in two ways, success or failure , (ii) The probability of success is p and the probability of failure is q such that $p + q = 1$ (iii) The experiment is repeated n times then probability of r success is given by,

$$P(x = r) = {}^nC_r\, p^r \cdot q^{n-r}, \quad \text{Where } r = 0,1,2 \ldots\ldots$$

$${}^nC_0 = 1, \; {}^nC_n = 1 \quad \text{and} \quad {}^nC_r = \frac{n!}{r!\,(n-r)!}$$

- To find this probability we consider Binomial. Expansion, so the above probability distribution is called Binomial probability distribution. It is denoted by β (n, p, r), where n, p, q are called the parameters.

4.8 Mean of the Binomial Distribution

Mean of the Binomial distribution β (n, p, r) is denoted by $\bar{x}$ or μ and defined as,

$$\mu = \sum_{r=0}^{n} P(X = r) \cdot r = E(X = r)$$

$$= \sum_{r=0}^{n} r \cdot P(r)$$

$$\mu = np$$

4.9 Variance of the Binomial Distribution

Variance of the Binomial distribution is denoted by V and is defined as,

$$V = \sigma^2 = \sum_{r=0}^{n} r^2 P(r) - \left[\sum_{r=0}^{n} r \cdot P(r)\right]^2$$

$\therefore$ Variance $= \sigma^2 = npq$

and the standard deviation $= \sigma = +\sqrt{npq}$

4.10 Mode of the Binomial Distribution

- Depending on the values of the two parameters n and p, Binomial distribution may be unimodal or bi-modal.
- To find the mode of Binomial distribution, first we have to find the value of (n + 1) p.

1) If (n+1) p is a non-integer then mode is uni-modal and mode = the largest integer contained in (n+1) p

2) If (n+1) p is an integer then mode is bi-modal and mode = (n+1)p, (n+1) p – 1.

4.11 Additive Property of Binomial Distribution

- Let X and Y be two independent Binomial variables such that X follows Binomial distribution with (n, p) and Y follows Binomial distribution with (n_2, p), then X + Y follows Binomial distribution with ($n_1 + n_2$, p).

4.12 Characteristic Function of Binomial Distribution

The characteristic function of Binomial distribution is denoted by $\Phi_x(t)$ and is defined as

$$\Phi_x(t) = E[e^{itx}]$$
$$= (q+ pe^{it})^n = (1-p + pe^{it})^n$$

4.13 Cumulants of Binomial Distribution

The moment generating function of Binomial distribution is denoted by $M_X(t)$ and is defined as

$$M_X(t) = E[e^{tx}]$$
$$= (q+ pe)^n = (1 - p + pe^t)^n$$

The cumulant of Binomial variable X is $k_X(t) = \log_e (q + pe^t)^n \log_e(q+ pe^t)$

4.14 Solved Examples

Example 4.14.1 *Mean and variance of Binomial distribution are 6 and 2 respectively. Find : i) P(r = 0) ii) P(r ≤ 1)*

Solution : Given : Mean = μ, np = 6 and variance = v = npq = 2

$\therefore$ $6 \times q = 2$

$\therefore$ $q = 1/3$

Hence $p = 1 - q = 2/3$

As $np = 6 \Rightarrow n \times \frac{2}{3} = 6 \Rightarrow n = 9$

i) Since $P(X = r) = {}^nC_r\, p^r \cdot q^{n-r}$

$\therefore$ $P(r = 0) = {}^9C_0\, p^0 q^9 = 1\left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^9$

$= \left(\frac{1}{3}\right)^9 = 0.000058$

ii) Since $P(r \le 1) = P(r = 0) + P(r = 1)$

$= {}^9C_0\, p^0 q^9 + {}^9C_1\, p^1 q^8$

$= \left(\frac{1}{3}\right)^9 + 9\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^8$

$= \left(\frac{1}{3}\right)^9 (1 + 18)$

$= 19\left(\frac{1}{3}\right)^9$

$\therefore$ $P(r \le 1) = 0.00097$

Example 4.14.2 *From a box containing 100 transistors 20 of which are defective, 10 are selected at random. Find the probability that (i) All will be defective (ii) All are non-defective.*

Solution : Given : n = 10

Let p be the probability of defective transistor

$\therefore$ $p = \frac{20}{100} = 0.2$

$\therefore$ $q = 1 - p = 0.8$

Since p(r defective) = ${}^nC_r\, p^r \cdot q^{n-r} = {}^{10}C_r (0.2)^r (0.8)^{10-r}$

(i) p (All defective transistor) = $P(r = 10) = {}^{10}C_{10}(0.2)^{10} (0.8)^0 = 0.0000001$

(ii) p (All are non-defective) = $P(r = 0) = {}^{10}C_0(0.2)^0 (0.8)^{10} = 0.1074$

Example 4.14.3 *An unbiased coin is thrown 10 times. Find the probability that (i) Getting exactly 6 heads (ii) Getting at least 6 heads.*

Solution : Since an unbiased coin is thrown, therefore,

$$p = q = \frac{1}{2} = 0.5 \text{ and } n = 10$$

where p - the probability of getting head. Since p (getting r heads) = ${}^nC_r \cdot p^r \cdot q^{n-r}$

(i) p(getting exactly 6 heads)

$= P(r = 6)$

$= {}^{10}C_6\, p^6 \cdot q^4$

$$= \frac{10!}{6!\,4!}(0.5)^6(0.5)^4$$

$$= \frac{10.9.8.7}{4.3.21}(0.5)^{10}$$

$$= 210(0.5)^{10}$$

$$= 0.205$$

(ii) p (getting at least 6 heads)

$$= p(\text{getting 6 or 7 or 8 or 9 or 10 heads})$$

$$= P(r=6) + P(r=7) + P(r=8) + P(r=9) + P(r=10)$$

$$= {}^{10}C_6(0.5)^6(0.5)^4 + {}^{10}C_7(0.5)^7(0.5)^3 + {}^{10}C_8(0.5)^8(0.5)^2 + {}^{10}C_9 0.5)^9 (0.5) + {}^{10}C_{10}(0.5)^{10}(0.5)^0$$

$$= (0.5)^{10}\left[{}^{10}C_6 + {}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}\right]$$

$$= (0.5)^{10}\left[\frac{10.9.8.7}{4.3.2.1} + \frac{10.9.8}{3.2.1} + \frac{10.9}{2.1} + 10 + 1\right]$$

$$= (0.5)^{10}(386)$$

$$= 0.3769$$

Example 4.14.4 *Six dice are thrown 729 times. How many times do you expect at least three dice to show 5 or 6 ?*

Solution : In a toss of a single dice,

let p = Probability of getting 5 or 6 $= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

$\therefore$ $q = 1 - p = \frac{2}{3}$ and $n = 6$ (since 6 dice are thrown simultaneously)

Since $P(x = r) = {}^{n}C_r\, p^r \cdot q^{n-r}$

$\therefore$ p(at least 3 dice to show 5 or 6)

$$= 1 - [P(r=0) + P(r=1) + P(r=2)]$$

$$= 1 - \left[{}^6C_0\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^6 + {}^6C_1\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^5 + {}^6C_2\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^4\right]$$

$$= 1 - \left[\frac{64}{729} + \frac{192}{729} + \frac{240}{729}\right]$$

$$= 1 - \frac{496}{729}$$

$$= \frac{233}{729}$$

$\therefore$ Expected number = N·P(at least 3 dice to show 5 or 6)

$$= 729 \times \left(\frac{233}{729}\right)$$

$$= 233$$

Example 4.14.5 *Seven coins are tossed and the number of heads obtained is recorded. If the experiment is repeated 128 times. Find the theoretical frequencies of different number of heads.*

Solution : Given : $N = 128$, $p = \frac{1}{2}$, $q = \frac{1}{2}$

Since the theoretical frequencies are given by

$$N(p+q)^n = 128\left(\frac{1}{2} + \frac{1}{2}\right)^7$$

$$= 128\left[{}^7C_0\left(\frac{1}{2}\right)^7 + {}^7C_1\left(\frac{1}{2}\right)^7 + {}^7C_2\left(\frac{1}{2}\right)^7 + {}^7C_3\left(\frac{1}{2}\right)^7 + {}^7C_4\left(\frac{1}{2}\right)^7 + {}^7C_5\left(\frac{1}{2}\right)^7 + {}^7C_6\left(\frac{1}{2}\right)^7 + {}^7C_7\left(\frac{1}{2}\right)^7\right]$$

$$= 128\left[\frac{1}{128} + \frac{7}{128} + \frac{21}{128} + \frac{35}{128} + \frac{35}{128} + \frac{21}{128} + \frac{7}{128} + \frac{1}{128}\right]$$

Thus the theoretical frequencies are

No. of heads	0	1	2	3	4	5	6	7
Frequency	1	7	21	35	35	21	7	1

Example 4.14.6 *Find the mode of the distribution for which mean and variance are 12 and 6 respectively.*

Solution : Given :

$$\text{Mean} = \mu = np = 12$$

and $$\text{Variance} = v = npq = 6$$

$\therefore$ $$q = \frac{npq}{np} = \frac{6}{12} = \frac{1}{2}$$

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

Since $$np = n \times \frac{1}{2} = 12$$

Hence $$n = 24$$

To find the mode, first we find the value of $(n+1)p$

$\therefore$ $$(n+1)p = (25)\left(\frac{1}{2}\right) = 12.5$$

Since $(n + 1)p = 12.5$ is a non-integer, hence the given binomial distribution is uni-modal and its mode = largest integer contained in 12.5

$\therefore$ Mode = 12

4.15 Poisson Distribution

- Poisson distribution is the limiting case of Binomial distribution when p, the probability of success is very small and n, the number of trials is very large. i.e. $p \to 0$, $n \to \infty$ such than $np = m$ remaining finite.
- In some cases where an experiment results in two outcomes success or failure, the number of successes only can be observed and not the number of failures. e.g. we can observe how many persons die of corona (covid-19) but we can not observe how many do not die of corona. In such cases Binomial distribution can not be used. In such cases we use poisson distribution.
- A random variable X is said to follows Poisson distribution if the probability of X is given by

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

4.16 Poisson Process

- The poisson process is one of the most widely used counting process in probability. It is the limiting case of Binomial distribution such that $n \to \infty$, $p \to 0$ and $np = m$ – finite. Poisson distribution is uniparametric distribution as it is characterized by only one parameter m.

1) Mean of Poisson distribution is $\mu = E(X) = m$
2) Variance of Poisson distribution is $V = \sigma^2 = m$
3) Poisson distribution may be uni-modal or bi-modal depending upon the value of the parameter m,
   (i) If m is a non-integer, then distribution is uni-modal and
   Mode = The largest contained in m
   (ii) If m is an integer, then the distribution is bi-modal and mode = m, m – 1
4) Additive property of Poisson distribution : Let X and Y be the two independent Poisson variables such that X follows Poisson distribution having parameter m_1 and Y follows Poisson distribution having parameter m_2, then X + Y is also a Poisson distribution with parameter $m_1 + m_2$.

4.17 Solved Examples

Example 4.17.1 *In a Poisson distribution $P(x = 2) = P(x = 3)$ find m and $P(x = 0)$, $P(x = 4)$.*

Solution : For a Poisson distribution

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

Since $P(x = 2) = P(x = 3)$

$$\therefore \quad \frac{e^{-m} \cdot m^2}{2!} = \frac{e^{-m} \cdot m^3}{3!}$$

$$\therefore \quad m = 3$$

Now $$P(x = 0) = \frac{e^{-m} \cdot m^0}{0!} = \frac{e^{-3} \cdot 1}{1} = e^{-3} = 0.04978$$

and $$P(x = 4) = \frac{e^{-m} \cdot m^4}{4!} = \frac{e^{-3} \cdot (3)^4}{4!} = 0.5041$$

Example 4.17.2 *If the probability that an individual suffers a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals (i) Exactly 3 (ii) More than 1 will suffer a bad reaction.*

Solution : We are given $n = 2000$, $p = 0.001$

Since $m = np = 2000 \times (0.001) = 2$

for a Poisson distribution

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

$$\therefore \quad P(x) = \frac{e^{-2} \cdot 2^x}{x!}$$

i) P(exactly 3 will suffer a bad reaction)

$$= P(x = 3)$$
$$= \frac{e^{-2} \cdot (2)^3}{3!}$$
$$= 0.1804$$

ii) P(more than 1 will suffer a bad reaction)

$$= 1 - [P(x = 0) + P(x = 1)]$$
$$= 1 - \left[\frac{e^{-2} \cdot (2)^0}{0!} + \frac{e^{-2} \cdot (2)^1}{1!}\right]$$
$$= 1 - [0.1353 + 0.2707]$$
$$= 1 - 0.4060$$
$$= 0.5940$$

Example 4.17.3 *In a town, 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows Poisson distribution. Find the probability that there will be 3 or more accidents in a day ?*

Solution : We are given n = 50, p - probability of accident in a day $= \frac{10}{50} = 0.2$

$$m = np = 50 \times 0.2 = 10$$

For a Poisson distribution,

$$P(x) = \frac{e^{-m} \cdot m^x}{x!} = \frac{e^{-10} \cdot (10)^x}{x!}$$

$\therefore$ P(3 or more accidents in a day)

$$= 1 - [P(x = 0) + P(x = 1) + P(x = 2)]$$
$$= 1 - \left[\frac{e(10)}{0!} + \frac{e(10)}{1!} + \frac{e(10)}{2!}\right]$$
$$= 1 - [0.000045 + 0.0004539 + 0.002269]$$
$$= 1 - 0.002754$$
$$= 0.997246$$

Example 4.17.4 *Fit a Poisson distribution to the following data and calculate theoretical frequencies.*

x :	0	1	2	3	4	Total
f :	109	65	22	3	1	200

Solution : The mean of the distribution $= m = \frac{\Sigma x_i f_i}{\Sigma f_i}$

$$m = \frac{109 \times 0 + 65 \times 1 + 22 \times 2 + 3 \times 3 + 1 \times 4}{200} = \frac{122}{200} = 0.61$$

For Poisson distribution

$$P(x) = \frac{e^{-m} \cdot m^x}{x!} \text{ where } x = 0, 1, 2, 3, 4$$
$$= \frac{e^{-0.61}(0.61)^x}{x!}$$

$\therefore$ $P(x = 0) = \frac{e^{-0.61}(0.61)^0}{0!} = e^{-0.61} = 0.5434$

$P(x = 1) = \frac{e^{-0.61}(0.61)^1}{1!} = 0.3315$

$P(x = 2) = \frac{e^{-0.61}(0.61)^2}{2!} = 0.1011$

$P(x = 3) = \frac{e^{-0.61}(0.61)^3}{3!} = 0.02056$

$P(x = 4) = \frac{e^{-0.61}(0.61)^4}{4!} = 0.0031$

Now Theoretical frequency = N·P(x)

Putting x = 0, 1, 2, 3, 4 we get the theoretical frequencies as 109, 66, 20, 4, 1.

Example 4.17.5 *The number of breakdowns of a computer in a week is a Poisson variable with m = np = 0.3. What is the probability that the computer will operate : i) With no breakdown and ii) At the most one breakdown in a week.*

Solution : Given that

$$m = np = 0.3$$

For a Poisson distribution

$$P(x) = \frac{e^{-m}(m)^x}{x!}, \quad x = 0, 1, 2, \ldots$$

$\therefore$ $$P(x) = \frac{e^{-0.3}(0.3)^x}{x!}$$

i) P(the computer has no breakdown) $= P(0) = \frac{e^{-0.3}(0.3)^0}{0!} = e^{-0.3} = 0.7408$

ii) P(there is at most one breakdown) = P(0) + P(1)

$$= \frac{e^{-0.3}(0.3)^0}{0!} + \frac{e^{-0.3}(0.3)^1}{1!} = 0.7408 + 0.2222 = 0.9630$$

Example 4.17.6 *On an average, there are 2 printing mistakes on a page of a book. Using Poisson distribution, find the probability that a randomly selected page from the book has at least one printing mistake.*

Solution : Given that m = 2

For a Poisson distribution

$$P(x) = \frac{e^{-m} m^x}{x!}$$
$$= \frac{e^{-2} 2^x}{x!} \quad \text{where } x = 0, 1, 2, \ldots$$

P(a page from the book has at least one printing mistake) $= P(x \geq 1)$

$$= 1 - P(x < 1)$$
$$= 1 - P(x = 0)$$
$$= 1 - \frac{e^{-2} 2^0}{0!}$$
$$= 1 - 0.1353$$
$$= 0.8647$$

4.18 Geometric Distribution

- In a series of Bernoulli's trial (independent trials with constant probability P of success), let the random variable, X denote the number of trials until the first process, then X is a geometric random variable with parameter p such that $0 < p < 1$ and the probability mass function (p.m.f.) of X is

$$P(X = x) = q^{x-1} \cdot p, \text{ where } x = 1, 2, 3, \ldots$$

4.18.1 Properties of Geometric Distribution

1) The mean of the geometric distribution is $\mu = E(X) = \frac{1}{p}$

2) The variance of the geometric distribution is

$$V(X) = E(X^2) - [E(X)]^2 = \frac{1-p}{p^2} = \frac{q}{p^2}$$

3) The standard deviation of the geometric distribution is $\sigma = +\sqrt{V(X)} = +\sqrt{\frac{q}{p^2}}$

4) Moment generating function of geometric distribution is $\frac{p}{1-(1-p)e^t}$

Example 4.18.1 *A light bulb manufacturing factory finds 3 in every 60 light bulbs defective. What is the probability that the first defective light bulb will be found when 5th one is tested ?*

Solution : Given : For Geometric distribution

$$P(X) = q^{x-1} \cdot p \qquad \text{where } q = 1-p$$

Here $p = \frac{3}{60} = 0.05 \qquad \therefore q = 1 - p = 1 - 0.05 = 0.95$

$$\therefore P(X = 6) = (0.95)^{5-1}(0.05) = (0.95)^4 (0.05)$$
$$= 0.0406$$

Example 4.18.2 *Suppose you are playing a game of darts. The probability of success is 0.4. What is the probability that you will hit the bullseye on the third try ?*

Solution : Given : For Geometric distribution

$$P(X) = q^{x-1} \cdot p \qquad \text{where } q = 1-p$$

Here $p = 0.4 \qquad \therefore q = 1 - 0.4 = 0.6$

$$\therefore P(X = 3) = (0.6)^{3-1}(0.4)$$
$$= (0.6)^2 (0.4)$$
$$= 0.144$$

Exercise

1) *A random variable X has the following probability distribution.*

X	−2	−1	0	1	2	3
P(X)	0.1	k	0.2	2k	0.3	3k

Find : i) the value of k ii) Mean iii) Variance iv) $P(x < 2)$ v) $P(|X| < 2)$

2) *The probability mass function of a random variable is defined as $P(X = 0) = 3C^2$, $P(X = 1) = 4C - 10C^2$ and $P(X = 2) = 5C - 1$, where $C > 0$ then find the value of C.*

3) *If the random variable X takes the values 1, 2, 3, 4 such that $(2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4))$ then find the probability distribution and $P(2 \le X \le 4)$.*

4) *Find the value of k if the probability density function $f(x) = k x^4 e^{-x/2}$, $0 \le x < \infty$*

5) *Find the value k if the probability density function $f(x) = \frac{k}{1+x^2}$, $-\infty < x < \infty$ Also find $P(0 \le x \le 1)$..*

6) *Find probability that almost 5 defective fuses will be found in a box of 200 fuses if 2 % of such fuses are defective.* **[Ans. : 0.7350]**

7) *In a certain factory turning out razor blades there is a small chance of 0.002 for any blade to be defective. The blades are supplied in a packet of 10 calculate approximate number of packets containing no defective and 2 defective blades in a consignment of 10,000 packets.* **[Ans. : 2]**

8) *Fit a Poisson distribution to the following data*

X	0	1	2	3	4	5
f	142	158	67	27	5	1

9) *Using Poisson distribution, find the probability that ace of spade will be drawn from a pack of well shuffled cards at least once in 104 consecutive draws.* **[Ans. : 0.8647]**

10) *Accidents occur on a particular stretch of highway at an average rate 3 per week. What is the probability that there will be exactly two accidents in a given week ?* **[Ans. : 0.224]**

Multiple Choice Questions

Q.1 If the discrete random variable X has the following probability distribution

x	1	2	3	4
P(X = x)	3k	k	k	3k

then the value of k is _____.

a $\frac{1}{4}$ b $\frac{1}{8}$

c $\frac{1}{6}$ d $\frac{2}{3}$

Q.2 If the discrete random variable X has the following probability distribution

x	0	1	2	3	4
P(X = x)	0.1	k	2k	2k	k

then the value of k is _____.

a 0.15 b 0.25

c 0.10 d 0.35

Q.3 Three coins are tossed simultaneously, X is the number of heads, then P(X = 2) is _____.

a $\frac{1}{8}$ b $\frac{1}{4}$

c $\frac{3}{8}$ d $\frac{1}{2}$

Q.4 If the discrete random variable X has the following probability distribution

x	1	2	3	4
P(X = x)	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{2}{5}$

then the value of $P(2 < X \leq 4)$ = _____.

a $\frac{3}{5}$ b $\frac{3}{10}$

c $\frac{7}{10}$ d $\frac{9}{10}$

Q.5 The discrete random variable X has the following probability distribution

x	0	1	2	3
P(X = x)	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Then the value of $P(X \geq 2)$ is _____.

a 0.5 b 0.25

c 0.125 d 0.8

Q.6 If the discrete random variable X has the following probability distribution

x	0	1	2	3	4	5
P(X = x)	0.15	0.1	0.15	0.30	0.15	0.15

then the value of P(X < 2) is _____.

a 0.15 b 0.25

c 0.4 d 0.45

Q.7 If discrete random variable X has the following probability distribution

x	0	1	2	3	4	5
P(X = x)	2k	2k	3k	k^2	$2k^2$	$7k^2 + 2k$

a $\frac{1}{7}$ b $\frac{1}{8}$

c $\frac{1}{9}$ d $\frac{1}{10}$

Q.8 The mathematical expectation E(X) of the following probability distribution is _____.

x	0	1	2	3
P(X = x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

a $\frac{13}{8}$ b $\frac{3}{2}$

c $\frac{9}{8}$ d $\frac{2}{3}$

Q.9 If probability mass function of a discrete random variable X is

$$P(X = x) = \begin{cases} \frac{c}{x^3}, & \text{for } x = 1,2,3 \\ 0, & \text{otherwise} \end{cases}$$ then value of c is _____.

a $\frac{216}{251}$ b $\frac{251}{216}$

c $\frac{294}{216}$ d $\frac{294}{251}$

Q.10 If probability mass function of a discrete random variable X is

$$P(X = x) = \begin{cases} \frac{c}{x^3}, & \text{for } x = 1,2,3 \\ 0, & \text{otherwise} \end{cases}$$ then E(X) is _____.

a $\frac{343}{297}$ b $\frac{294}{251}$

c $\frac{297}{294}$ d $\frac{294}{297}$

Q.11 If the probability function of a discrete random variable X which assumes values x_1, x_2, x_3 such that $P(x_1) = 2P(x_2) = 3P(x_3)$ then $P(x_1) =$ _____.

a $\frac{3}{11}$ b $\frac{6}{11}$

c $\frac{8}{11}$ d $\frac{2}{11}$

Q.12 If the probability density function of a continuous random variable x is

$$f(x) = \begin{cases} k(x-1)^3, & 1 \le x \le 3 \\ 0, & \text{otherwise} \end{cases}$$ then the value of k is _____.

a $\frac{1}{4}$ b $\frac{1}{2}$

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

Q.13 If the p.d.f. of a continuous random variable X is defined by

$$f(x) = \begin{cases} k(1-2x^2), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$ then the value of k is _____.

a 3 b $\frac{1}{3}$

c 1 d 0

Q.14 If the probability density function of a continuous random variable X is defined by _____.

$$f(x) = \begin{cases} kx^2, & -3 < x < 3 \\ 0, & \text{otherwise} \end{cases}$$

a $\frac{1}{27}$ b $\frac{1}{28}$

c $\frac{1}{18}$ d $\frac{1}{9}$

Q.15 The probability density function of a continuous random variable X is

$$f(x) = \begin{cases} 6x(1-x), & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

If $P(X < k) = P(X > k)$ then the value of k is _____.

a 1 b $\frac{1}{2}$

c $\frac{1}{3}$ d $\frac{1}{4}$

Q.16 If the probability density function of a continuous random variable X is

$$f(x) = \begin{cases} \frac{3}{2}x^2, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$ then $P(0 \le x \le 1)$

a $\frac{1}{3}$ b $\frac{3}{2}$

c $\frac{2}{3}$ d $\frac{1}{2}$

Q.17 The probability density function of a continuous random variable X is

$$f(x) = \begin{cases} kx^2(1-x), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$ then the value of k is _____.

a 12 b $\frac{1}{12}$

c 6 d $\frac{1}{6}$

Q.18 The probability density function of a continuous random variable x is

$$f(x) = \begin{cases} e^{-x}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$ then the value of $P(0 < x < 2)$

a $\frac{1-e^2}{e^2}$ b $\frac{e^2-1}{e^2}$

c $\frac{e^{-2}-1}{e^{-2}}$ d $\frac{e^2-e^{-2}}{2}$

Q.19 The probability density function of x is $f(x) = \frac{k}{1+x^2}$, $-\infty < x < \infty$ then the value of k = _____.

a π b $\frac{\pi}{2}$

c $\frac{1}{\pi}$ d $\frac{2}{\pi}$

Q.20 Mean and variance are equal in _____.

a Binomial distribution b Poisson's distribution

c Normal distribution d Geometric distribution

Q.21 In a Binomial distribution, if n is the number of trials and p is the probability of success, then the mean value is given by _____.

[a] np [b] n

[c] p [d] $np(1-p)$

Q.22 In a Binomial distribution, if p, q and n are probability of success, failure and number of trials respectively then the variance is given by _____.

[a] np [b] npq

[c] np^2q [d] npq^2

Q.23 If X is a random variable taking X, probability of success and failure being p and q respectively and n be the number of trials then P(X = x) is _____.

[a] $nC_xp^xq^x$ [b] $nC_xp^xq^{n-x}$

[c] ${}^xC_np^xq^n$ [d] ${}^xC_np^xq^{n-x}$

Q.24 In a Binomial distribution, if p, q and n are probability of success, failure and number of trials respectively then the standard deviation is _____.

[a] $\sqrt{np}$ [b] $\sqrt{pq}$

[c] $\sqrt{npq}$ [d] $\sqrt{nq}$

Q.25 In a Binomial distribution which of the following is correct ?

[a] Mean = Vairance [b] Mean < Variance

[c] Mean > Variance [d] None of these

Q.26 In a Binomial distribution, n = 5, if P(X = 4) = P(X = 3) then p = _____.

[a] $\frac{4}{13}$ [b] $\frac{5}{13}$

[c] $\frac{9}{13}$ [d] $\frac{6}{13}$

Q.27 In a binomial distribution, if p = q, then P(X = x) is _____.

[a] ${}^nC_x(0.5)^n$ [b] ${}^nC_x(0.5)^x$

[c] ${}^nC_n(0.5)^x$ [d] ${}^nC_xp^{n-x}$

Q.28 The mean and variance of a Binomial distribution are 4 and 2 respectively. Then the P(X = 2) is _____.

[a] $\frac{128}{256}$ [b] $\frac{219}{256}$

[c] $\frac{37}{256}$ [d] $\frac{28}{256}$

Q.29 The mean and variance of Binomial distribution are 6 and 2 respectively then the number of trials n is _____.

[a] 14 [b] 18

[c] 10 [d] 12

Q.30 The mean and variance of Binomial distribution is 36 and 9 respectively, then the value of probability of success p = _____.

[a] 1/3 [b] 3/4

[c] 1 [d] 1/2

Q.31 The mode of the Binomial distribution for which mean and variance are 12 and 6 respectively, then the value of mode is _____.

[a] 12 [b] 12.5

[c] 13 [d] 2

Q.32 The mean and variance of Binomial probability distribution are 6 and 4 respectively then the value of mode is _____.

[a] 6.5 [b] 6.3

[c] 6 [d] 6.6

Q.33 In Poisson's probability distribution, if m = np, where n the number of trials is very large and p the probability of success at each trial then the probability of r successes is _____.

[a] $\frac{e^m m}{r!}$ [b] $\frac{e^{-m} \cdot m^r}{r}$

[c] $\frac{e^{-m} \cdot m^r}{r!}$ [d] $\frac{e^m \cdot m^r}{r!}$

Q.34 In a Poisson's distribution if n = 100, p = 0.01 then the value of P(X = 0) is _____.

[a] $\frac{1}{e}$ [b] $\frac{2}{e}$

[c] $\frac{3}{e}$ [d] $\frac{4}{e}$

Q.35 In a Poisson's probability distribution, which of the following is correct ?

[a] Mean = Variance [b] Mean > Variance

[c] Mean < Variance [d] None of these

Q.36 Poisson's probability distribution is _____.

[a] uni-modal [b] bi-modal

[c] uni-modal or bi-modal [d] None of the above

Q.37 In Poisson's probability distribution, m = np is an integer then the mode of distribution is _____.

[a] uni-modal and m [b] uni-modal and m – 1

[c] bi-modal and m, m–1 [d] bi-modal and m, m + 1

Q.38 In a Poisson's probability distribution, P(x = 2) = P(x = 3) then the value of m is _____.

[a] 1 [b] 3

[c] 2 [d] 4

Q.39 The mean and variance of geometric distribution are _____.

[a] $\frac{p}{q}$ and $\frac{p}{q}$ [b] $\frac{q}{p}$ and $\frac{q}{p}$

[c] $\frac{q}{p}$ and $\frac{q}{p^2}$ [d] $\frac{p}{q}$ and $\frac{p^2}{q}$

Q.40 The probability mass function of a geometric distribution is P(X = x) = _____.

[a] pq [b] $q^{x-1}p$

[c] $p^x q$ [d] $\frac{p}{q}$

Answer Keys for Multiple Choice Questions :

Q.1	b	Q.2	a	Q.3	c	Q.4	c	Q.5	b
Q.6	b	Q.7	d	Q.8	b	Q.9	a	Q.10	b
Q.11	b	Q.12	a	Q.13	a	Q.14	c	Q.15	b
Q.16	d	Q.17	a	Q.18	b	Q.19	c	Q.20	b
Q.21	a	Q.22	b	Q.23	b	Q.24	c	Q.25	c
Q.26	d	Q.27	a	Q.28	d	Q.29	b	Q.30	b
Q.31	a	Q.32	c	Q.33	c	Q.34	a	Q.35	a
Q.36	c	Q.37	c	Q.38	b	Q.39	c	Q.40	b

❑❑❑

Notes

Unit V

5 Inferential Statistics: Hypothesis

Syllabus

Statistical Inference - Testing of Hypothesis, Non-parametric Methods and Sequential Analysis: *Introduction, Statistical Hypothesis (Simple and-Composite), Test of a Statistical Hypothesis, Null Hypothesis, Alternative Hypothesis, Critical Region, Two Types of Errors, level of Significance, Power of the Test*

Contents

5.1 Introduction

- Statistical inference is the branch of statistics which deals the study of applications of probability theory for decision making of some proposition. Statistical inference problems are divided into two categories.

  1.Estimation

  2.Hypothesis testing

- When data is collected from a sample taken from the population then the problem is either to estimate the value of population parameter or to check the validity of the statement made about the population. Finding the value of population parameter leads to estimation theory.

1) **Examples of estimation :** Estimate the population mean weight using sample mean weight.

   While the process of decision making of statement is 'testing of Hypothesis'.

   For example - A data based decision made by an engineer about the efficiency of newly invented gadget over the existing gadgets.

2) **Hypothesis testing :** Use sample evidence to test the claim that population mean weight is 50 kg.

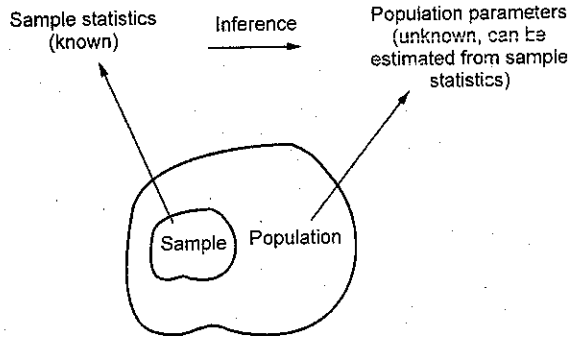


Fig. 5.1.1

Preliminaries

1) **Population :** A collection of individuals / objects under study is called population or universe. It is classified into two parts.
   a) **Finite population :** A population that contains finite number of individuals /objects is called finite population.
   b) **Infinite population :** A population containing infinite number of individuals / objects is called infinite population.

2) **Sampling :** A finite subset of universe is called "sample". The process of selecting a sample from the population is called sampling.
3) **Degree of freedom :** Degrees of freedom is number of independent values / variables that can vary in an analysis without violating any constraints.

   If data contains total "n" values of random variable and no. of contraints on the variable are "k" then

$$d.f. = n - k - 1$$

Parameter of statistics :

- The statistical constants of the population such as mean (μ), the variance (σ^2) etc. are known as the parameters.
- The statistical quantity computed from the members of the sample to estimate the parameters of the population from which the sample has been drawn are known as 'statistic'.

Standard Error (S.E.) :

- Standard deviation of the sampling distribution of a static is known as the **'Standard Error'**. It forms a basis of the testing of hypothesis.

  If t is any statistic.

  For given large samples,

$$Z = \frac{t - E(t)}{S \cdot E(t)}$$ is normally distributed with mean 0 and variances 1.

  Where E(t) is error in t and SE (t) is standard error in t.
- This Z is called **test statistic** which is used to test significant difference of sample and population proportion.
- In other words, test static is a number calculated from statistical test of hypothesis. This shows how well the observed data agrees with expected distribution under the null hypothesis.

5.2 Statistical Hypothesis

- In many situations one may be interested in drawing certain conclusions about the population based on sample information. In taking such decisions certain assumptions are made. These assumptions (statement / claim) are called 'Statistical Hypothesis'.
- Hypothesis is mainly categorised into : 1) Simple hypothesis 2) Composite hypothesis

1) Simple hypothesis (Definition) : If the statistical hypothesis specifies the population completely then it is called as simple hypothesis.

2) Composite hypothesis (Definition) : It is composed by many simple hypothesis and hence don't specify the population completely.

For example, If $x_1, x_2, \ldots, x_n$ is a random sample of size n from a population having mean μ and variance σ^2 then the hypotheses.

$H_0: \mu = \mu_0, \sigma^2 = \sigma_0^2$ is a simple hypothesis.

Where as

a) $H_0 : \mu = \mu_0$ (where σ^2 unknown)

b) $H_0 : \sigma^2 = \sigma_0^2$ (where μ unknown)

c) $H_0 : \mu < \mu_0, \sigma^2 = \sigma_0^2$

d) $H_0 : \mu > \mu_0, \sigma^2 = \sigma_0^2$

e) $H_0 : \mu = \mu_0, \sigma^2 < \sigma_0^2$

f) $H_0 : \mu = \mu_0, \sigma^2 > \sigma_0^2$

g) $H_0 : \mu < \mu_0, \sigma^2 > \sigma_0^2$

are all composite hypothesis.

5.3 Test of Statistical Hypothesis (Test of Significance)

- The method which one can decide whether to accept or reject a hypothesis is known as 'test of hypothesis' or hypothesis testing'. There are two types of hypotheses

  1) Null hypothesis 2) Alternate hypothesis

5.4 Null and Alternate Hypothesis

1) Null hypothesis :

- It is a hypothesis which is to be tested for possible rejection under the assumption that it is true.

For Example :

- If we want to find out whether extra coaching has benefited the students or not we shall set up a null hypothesis that " extra coaching has not benefited the students."
- The null hypothesis is very useful tool in testing the significance of difference. In other words, null hypothesis states that there is no real difference in the sample and the population in a particular characteristics under consideration. Therefore it is also defined as "hypothesis of no difference". Null hypothesis is denoted by H_0.

2) Alternate hypothesis :

- The hypothesis that is different from the null hypothesis is called the alternate hypothesis.

For Example :

If we want to test the null hypothesis that " the average percentile of students in a college have specified mean 70." then we have

$$H_0 : \mu = 70$$

then alternate hypothesis, denoted by H_1 or H_a is

$H_1 : \mu \neq 70$ ($\mu < 70$ or $\mu > 70$) (Two tailed alternate hypothesis)

$H_1 : \mu < 70$ Left / single tailed alternate hypothesis

$H_1 : \mu < 70$ Right / single tailed alternate hypothesis

Note :

1) H_0 always have an "equal sign" (and possibly 'less than" or "greater than" symbol depending on the alternative hypothesis).

2) The null and alternate hypotheses are stated together.

Solved Example

Example 5.4.1 *A researcher thinks that if expectant mother use vitamins, the birth weight of the babies will increase. The average birth weight of the population is 8.6 pounds. State the null and alternate hypothesis for this problem.*

Solution : We need to test that average birth weight is 8.6 pound.

$\therefore$ $H_0 : \mu = 8.6$

$H_1 : \mu > 8.6$

5.5 Types of Error

- Acceptance or rejection of null hypothesis depends on the sampling while making any conclusion about the null hypothesis. There may arise two types of error.

Type I error

Type I error happens when null hypothesis (H_0) is rejected even if it is true.

Type II error

Type II error happens when null hypothesis (H_0) is accepted when it is false.

These errors can be summarized in tabular form as :

	Decision	
Actual situation	Reject H_0	Accept H_0
H_0 is true	Type I error	Correct decision
H_0 is false	Correct decision	Type II error

- The aim of hypothesis testing is to minimize both types of errors. But in practice it is not possible to reduce both errors simultaneously. The reduction of one type of error usually increases the other type of error. Therefore, a compromise must be found to limit more serious errors.

5.6 Critical Region

- Hypothesis testing involves determining the region such that the hypothesis will be rejected if the sample points falls in the region and it will be accepted if the sample points are outside the region. This region of rejection is called "critical region". It is denoted by W and acceptance region is denoted $\overline{W}$.
- The value of a test statistic which separates the critical region from the acceptance region is called "critical value" or significance value.

One tailed test -

- If the critical region lies on the one side / tail of distribution of test value then such tests are called one tailed /one sided tests.
- If critical region lies on the left side of the distribution then it is known as left tailed test.
- For right tailed test critical region lies on the right side.

Two tailed test

- If critical region is located in both tails of the distribution then such tests are called two tailed tests.

Area of critical region

i) For one tail test :

Let Z be the test statistic

Z_α be the critical value at level of significance α

Then for right tailed test

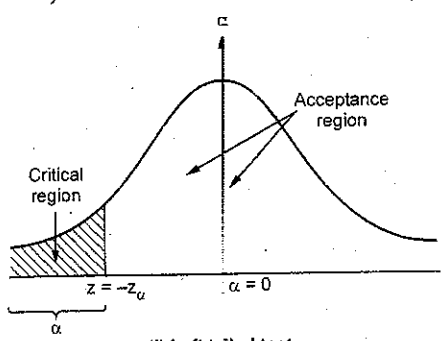
$$P(Z > Z_\alpha) = \alpha$$

and for left tailed test

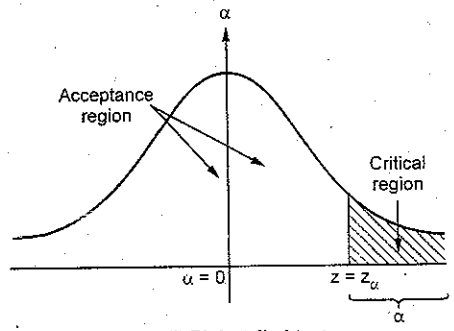
$$P(Z < -Z_\alpha) = \alpha$$

i.e. area of the critical region on either side of tail (left or right) is α.

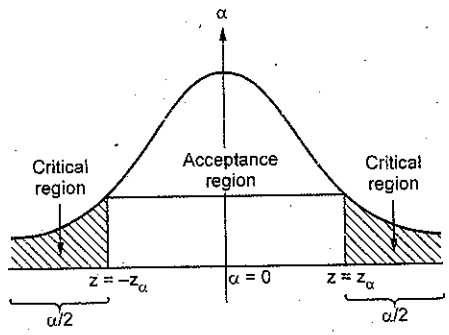
ii) For two tailed test



(i) Left tailed test



(ii) Right tailed test



(iii) Two tailed test

Fig. 5.6.7

$$P(|Z| - Z_\alpha) = \alpha$$

$$P(Z < -Z_\alpha) + P(Z > Z_\alpha) = \alpha \quad \ldots (5.6.1)$$

But normal curve is symmetric

$$\therefore \quad P(Z < -Z_\alpha) = P(Z > Z_\alpha) \quad \ldots (5.6.2)$$

$\therefore$ From (5.6.1) and (5.6.2)

$$2P(Z > Z_2) = \alpha$$

$$P(Z > Z_2) = \frac{\alpha}{2} = P(Z - Z_\alpha)$$

i.e. area of critical region on both tails is same and is equal to $\frac{\alpha}{2}$.

Thus total area of critical region for two tailed test is α.

Note :

1) Critical value Z_α ; For one tailed test (left or right) at significance level α is equal to the critical value Z_α for two tailed test at significance level 2α

2)

If hypothesis contains	Nature of test
≠	Two tailed test
<	Left tailed
>	Right tailed

3) In following table critical values Z_α, for one tailed and two tailed tests are listed at different levels of significance.

Level of significance α	Z_α		
	Two tailed test	Right tailed Test	Left tailed Test
0.01 (1 %)	± 2.58	2.33	– 2.33
0.05 (5 %)	± 1.966	1.645	– 1.645
0.10 (10 %)	± 0.645	1.28	– 1.28

5.7 Level of Significance

- The probability of Type I error is known as level of significance (l.o.s) and is denoted by "α". It is also known "size of test".
- In many hypotheses testing situations, consequences of type I error are more serious than type II error. Therefore the probability of type I error i.e. α is fixed at low level generally at 5 % or 1 %. which means probability of accepting correct hypothesis is 95 %.
  - Mathematically

$$\alpha = P[\text{Rejecting } H_0 \mid H_0 \text{ is true}]$$
$$= P[\text{Reject } H_0 \mid H_0]$$
$$= P[\text{Type I error}]$$

5.8 Power of Test

- Probability of type II error is denoted by β i.e. β is probability of accepting wrong null hypothesis. Therefore $1 - \beta$ gives the probability of rejecting null hypothesis when it is false.
- This probability (i.e. $1 - \beta$) is called power of the test.
- Power of the test helps to measure the goodness of test (i.e. how well test is working)

$$\beta = [\text{Accept } H_0 \mid H_0 \text{ is false}]$$
$$= [\text{Accept } H_0 \mid H_1 \text{ is true}]$$
$$= [\text{Accept } H_0 \mid H_1]$$
$$= P[\text{Type II error}]$$

Note :

1) Smaller the value of β, larger the value of $1 - \beta$ i.e. chances of rejecting false hypothesis are more, thus test will work quiet well.
2) Low value of $1 - \beta$ (near to zero) means that the test is working very poorly.

- Following factors primarily affect the power of a test.

I) Level of significance (α) : If l.o.s. (α) is increased, by keeping all other parameters constant, then power of test is also increased.

Since larger value of l.o.s. (α) means the larger critical region (rejection region) and thus probability of rejecting null hypothesis H_0 is greater.

II) Sample size (n) - The power of test increases as sample size increase.

III) Effect size - It is a number that measures the strength of relationship between two variables of population.

Effect size = Hypothetical value of parameter – Actual value of parameter

It is also known as "magnitude of effect".

Larger effect size enhances the power of test,

Examples on α and β

Example 5.8.1 *A sample of size one is drawn from Poisson distribution with parameter λ. To test $H_0 : \lambda = 1$ v/s $H_1 : \lambda = 2$. Test $\phi(x)$ of critical region $w = \{x : x > 3\}$ what is probability of Type - I error ?*

Solution : Given $X \sim P(\theta)$

Hypothesis testing problem is

$$H_0 : \lambda = 1 \quad v/s \quad H_1 : \lambda = 2$$

Test function :

$$\phi(x) = \begin{cases} 1 & \text{when } x > 3 \\ 0 & \text{when } x \leq 3 \end{cases}$$

We want to find the probability of Type I error

By using definition

$$\begin{aligned} \alpha &= P(\text{Type I error}) = E_{H_0}[\phi(x)] \\ &= P_{H_0}[x > 3] \\ &= P[x > 3 / x \sim P(\lambda = 1)] \\ &= 1 - P[x \leq 3 / x \sim P(\lambda = 1)] \\ &= 1 - \{P[x = 0, 1, 2, 3 / x \sim P(\lambda = 1)]\} \\ &= 1 - \{P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)\} \\ &= 1 - \left\{\frac{e^{-1}1^0}{0!} + \frac{e^{-1}1^1}{1!} + \frac{e^{-1}1^2}{2!} + \frac{e^{-1}1^3}{3!}\right\} \\ &= 1 - \frac{8}{3e} \end{aligned}$$

Example 5.8.2 *Let P be the probability that a coin will fall head in a single toss in order to test $H_0 : P = \frac{1}{2}$ against $P = \frac{3}{4}$. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.*

Solution : Given

$$H_0 : P = \frac{1}{2}$$

$$H_1 : P = \frac{3}{4}$$

$$n = 5$$

Let X denotes the no. of heads in n tosses. Then by binomial distribution

$$\begin{aligned} P(X - x) &= {}^nC_r\, p^x q^{n-x} \qquad q = 1 - p \\ &= {}^5C_x\, p^x (1-p)^{5-x} \end{aligned}$$

H_0 is rejected if more than 3 heads are obtained. Therefore critical region is given by

$$W = \{x / x \geq 4\}$$

$$\rightarrow \quad \bar{W} = \{x / x \leq 3\}$$

α = Probability of type I error = P [Reject H_0/ H_0]

$$\begin{aligned} &= p[X \geq 4 / H_0] \\ &= p\left[x = 4/P = \frac{1}{2}\right] + P\left[x = 5/P = \frac{1}{2}\right] \\ &= {}^5C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{5-4} + {}^5C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{5-5} \\ &= 5\left(\frac{1}{2}\right)^5 + 1\left(\frac{1}{2}\right)^5 \\ &= 6\left(\frac{1}{2}\right)^5 = \frac{3}{16} \end{aligned}$$

$$\therefore \quad \boxed{\alpha = \frac{3}{16}}$$

β = Probability of type II error

$$\begin{aligned} &= P[\text{Accept } H_0 / H_1] \\ &= P\left[x \geq 3 / P = \frac{3}{4}\right] \\ &= 1 - P\left[x \geq 4 / P = \frac{3}{4}\right] \\ &= 1 - P\left[x = 4 / P = \frac{3}{4}\right] - P\left[x = 5 / P = \frac{3}{4}\right] \\ &= 1 - {}^5C_4 \left(\frac{3}{4}\right)^4 \left(1 - \frac{3}{4}\right)^{5-4} + {}^5C_5 \left(\frac{3}{4}\right)^5 \left(1 - \frac{3}{4}\right)^{5-5} \\ &= 1 - 5\left(\frac{3}{4}\right)^4 \left(\frac{1}{4}\right)^1 - 1\left(\frac{3}{4}\right)^5 \left(\frac{1}{4}\right)^0 \\ &= 1 - 5\frac{3^4}{4^5} - \frac{3^5}{4^5} \\ &= \frac{47}{128} \end{aligned}$$

Then power of test $= 1 - \beta = 1 - \frac{47}{128} = \frac{81}{128}$

Example 5.8.3 *Let P is the probability that a given die shows even number. To test $H_0 : P = \frac{1}{2}$ v/s $H_1 : P = \frac{1}{3}$ following procedure is adopted. Toss the die twice and accept H_0 if both times it shows even number. Find the probabilities of type I and type II error.*

Solution :

$$H_0 : P = \frac{1}{2}$$

$$H_1 : P = \frac{1}{3}$$

$$n = 2$$

Let X is a r.v. that denotes even no.

Then by binomial distribution

$$P(X = x) = {}^nC_x P^x q^{n-x} = {}^nC_x P^x (1-p)^{n-x}$$

We reject H_0 if no. of times, that an even number is shown is less than 2.

$\therefore$ Critical region is,

$$W = \{x \mid x < 2\}$$

$$\overline{W} = \{x \mid x = 2\}$$

$$\alpha = P \cdot [\text{Reject } H_0 / H_0]$$

$$= P\left[x < 2 / P = \frac{1}{2}\right]$$

$$= 1 - P\left[x = 2 / P = \frac{1}{2}\right]$$

$$= 1 - {}^2C_2 \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^{2-2}$$

$$= 1 - \frac{1}{4}$$

$$= \frac{3}{4}$$

$\therefore$ $\boxed{\alpha = \frac{3}{4}}$

$$\beta = P[\text{Accept } H_0 | H_1]$$

$$= P\left[x = 2 / P = \frac{1}{3}\right]$$

$$= {}^2C_2 \left(\frac{1}{3}\right)^2 \left(1 - \frac{1}{3}\right)^{2-2}$$

$\boxed{\beta = \frac{1}{9}}$

$\therefore$ Power of test $= 1 - \beta$

$$= 1 - \frac{1}{9}$$

$$= \frac{8}{9}$$

5.9 General Procedure Followed in Testing of Statistical Hypothesis

Step I : Set up the null hypothesis H_0 clearly.

Step II : Specify an alternate hypothesis H_1 in an appropriate way.

Step III : Choose a level of significance (α).

Step IV : State and compute the appropriate test statistic (Z) under H_0

Step V : Compare the test statistic value (Z) with Z_α (the critical value) at level of significance α.

Draw a conclusion about the rejection or acceptance of H_0.

Note :
1. If $|Z| > Z_\alpha$ we reject H_0
2. If $|Z| < Z_\alpha$ we accept H_0

5.10 Test of Significance

Depending on the sample size. We categories testing of significance into two types.

1. Test of significance for large samples - (sample size $n > 30$)
2. Test of significance for small samples - (sample size $n < 30$)

5.10.1 Test of Significance for Large Samples

If the sample size $n > 30$, the sample is taken as large sample. For these samples we apply normal tests [as Binomial, Poisson, Chi-square etc. are approximated by normal distributions considering the population as normal.]

Following are the important tests to test the significance for large samples.

1. Testing of significance for single proportion.
2. Testing of significance for difference of proportions.
3. Testing of significance for single mean.
4. Testing of significance for a difference of means.

5.10.2 Test of Significance for Single Proportion

We use this test to check the significant difference between proportion of sample and the population.

Notations used :

X - no. of success in n independent trials

P - probability of success for each trail

Q - Probability of failure

$$Q = 1 - P$$

$$P = \frac{X}{n} \text{ - observed proportion of success}$$

$$E(X) = nP$$

$$V(X) = nPQ$$

Then $$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} nP$$

$$E(p) = P$$

$$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{1}{n^2} (nPQ) = \frac{PQ}{n}$$

$$\therefore \quad SE(P) = \sqrt{V(P)} = \sqrt{\frac{PQ}{n}}$$

Then test statistic is given by

$$Z = \frac{P - E(P)}{SE(P)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$$

Note :

1) The probable limit for the observed proportion of success are $p \pm Z_\alpha \sqrt{\frac{PQ}{n}}$, where Z_α is the significant value at level of significance "α".

2) If P is not known, the limits for the proportion in the population are $p \pm Z_\alpha \sqrt{\frac{PQ}{n}}$ where $q = 1 - p$

3) If α is not given, we can take safely 3σ limits.

Hence, confidence limits for observed proportion p are $p \pm \sqrt{\frac{pq}{n}}$

Solved Examples

Example 5.10.1 *A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased.*

Solution : Probability of getting head /T $= \frac{1}{2} = 0.5 = P$

$\therefore$ H_0 : The coin is unbiased (i.e. $\phi = 0.5$)

H_1 : The coin is not unbiased ($P \neq 0.5$)

Here $n = 400$

$X = 216$ (No. of success i.e. getting head)

$\therefore$ $p =$ Proportion of success in sample $= \frac{X}{n} = \frac{216}{400} = 0.54$

Population proportion $P = 0.5$

$\therefore$ $Q = 1 - P = 1 - 0.5 = 0.5$

Test statistic

$$Z = \frac{p - P}{\sqrt{PQ/n}}$$

$$|Z| = \left| \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{400}}} \right| = 1.6$$

since $|Z| = 1.6 < 1.96 = Z_\alpha$

i.e. $|Z| < Z_\alpha$, where Z_α, is l.o.s. at 5 %

$\therefore H_0$ is accepted i.e. coin is unbiased.

Example 5.10.2 *An certain cubical die was thrown 9000 times and 5 or 6 was obtained 3240 times. On the assumptions of certain throwing, do the data indicate an unbiased die ?*

Solution : Here, $n = 9000$, $X = 3240$

$P =$ Probability of success (i.e. getting 5 or 6)

$$= \frac{2}{6} = \frac{1}{3}$$

$\therefore$ $Q = 1 - P = \frac{2}{3}$

$$p = \frac{X}{n} = \frac{3240}{9000} = 0.36$$

Let H_0 : Die is unbiased i.e. $P = \frac{1}{3}$

H_1 : Die is biased i.e. $P \neq \frac{1}{3}$

$\therefore$ Tests statistic

$$Z = \frac{p - P}{\sqrt{PQ/n}}$$

$$|Z| = \left| \frac{0.36 - 0.33}{\sqrt{(1/3 \times 2/3)/9000}} \right|$$

$$= 0.03496 < 1.96$$

i.e. $|Z| < Z_\alpha$ (l.o.s at 5 %)

$\therefore$ We accept H_0 i.e. die is unbiased.

Example 5.10.3 *A manufacturer claims that 4 % of his products supplied by him are defective. A random sample of 600 products contained 36 defectives. Test the claim of manufacturer.*

Solution : Here, $n = 600$, $X = 36$

p = Probability of getting defective bolt $= \frac{36}{600} = 0.06$

P = Proportion of defectives in the population $= 4\ \% = 0.04$

$\therefore$ $Q = 0.96$

Let

H_0 : 4 % products are defective (i. p. = 0.04)

H_1 : $p \neq 0.04$

Test statistic

$$Z = \left[\frac{p - P}{\sqrt{PQ/n}}\right]$$

$$|Z| = \left|\frac{0.06 - 0.04}{\sqrt{\frac{0.04 \times 0.96}{600}}}\right| = 2.5 > 1.96 = Z_\alpha \qquad \text{(l.o.s. at 5 \%)}$$

Since $|Z| > Z_\alpha$

$\therefore H_0$ is rejected i.e. defective products are more than 4 %.

Example 5.10.4 *400 apples are taken at random from a large basket and 40 are found to be bad. Estimate the proportion of bad apples in the basket and assign limits within which the percentage most probably lies.*

Solution : Given, $X = 40$, $n = 400$

p = Proportion of bad apples $= \frac{X}{n} = \frac{40}{400} = 0.1$

$\therefore$ $q = 1 - p = 0.9$

Since confidence limits is not given, we assume it is 95 %

$\therefore$ Level of significance $\alpha = 5\ \%$

$\therefore$ $Z_\alpha = 1.96$

Also population proportion P is not provided

$\therefore$ The limits are,

$$p \pm Z_\alpha \sqrt{\frac{pq}{n}} = 0.1 \pm 1.96 \sqrt{\frac{0.1 \times 0.9}{400}}$$

$$= 0.1 \pm 0.0294$$

5.10.3 Testing of Significance for Difference Proportion

Let X_1 and X_2 be two samples taken from two different populations. n_1 and n_2 be the sizes of X_1 and X_2 respectively. If P_1 and P_2 are proportion on success in the samples X_1 and X_2 respectively. Then to test the significance of difference between P_1 and P_2 we use test statistic.

$$Z = \frac{P_1 - P_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$

and $Q = 1 - P$

Example 5.10.5 *In a batch of 500 articles, produced by a machine, 16 articles are found defective. After overhauling the machine, it is found that 3 articles are defective in a batch of 100. Has the machine improved ?*

Solution : Given : $n_1 = 500$, $n_2 = 100$

$$P_1 = \frac{16}{500} = 0.032$$

$$P_2 = \frac{3}{100} = 0.03$$

$$P = \frac{P_1 n_1 + P_2 n_2}{n_1 + n_2} = \frac{(0.032)(500) + (0.03 \times 100)}{500 + 100} = \frac{19}{600} = 0.032$$

$\therefore$ $Q = 1 - P = 1 - 0.032 = 0.968$

Since we have to test whether machine is improved or not.

$\therefore$ We state null hypothesis as,

H_0 : Machine has not improved after overhauling i.e. $p_1 = p_2$

Thus H_1 : $P_1 > P_2$ (right tailed)

under H_0, $$Z = \frac{P_1 - P_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.032 - 0.03}{\sqrt{0.032 \times 0.968 \times \left(\frac{1}{500} + \frac{1}{100}\right)}}$$

$$= 0.104 < 1.645 = Z_\alpha\ (\alpha = 5\ \%)$$

Conclusion : As $|Z| < Z_\alpha$ (at 5 % l.o.s.), H_0 is accepted i.e. machine is not improved.

Example 5.10.6 *Before increasing the excise duty, it is observed that out of a sample 1000 persons 800 were coffee drinkers. After an increase in the duty, out of 1200 persons, 800 were found to be coffee drinkers. Is there a significance decrease in coffee consumption after increasing the excise duty ?*

Solution : Given : $n_1 = 1000, n_2 = 1200$

$$p_1 = \frac{800}{1000} = 0.8 = \frac{4}{5} \; ; p_2 = \frac{800}{1200} = \frac{2}{3}$$

$$= \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{1600}{2200} = \frac{8}{11}$$

$$\therefore \quad Q = 1 - P = 1 - \frac{8}{11} = \frac{3}{11}$$

To test the difference in coffee consumption of people, we set the null hypothesis as H_0 : $p_1 = p_2$ i.e. there is no difference in the consumption of coffee before and after increase in excise duty.

H_1 : $p_1 > p_2$ (right tailed test)

under H_0 :

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{4}{5} - \frac{2}{3}}{\sqrt{\frac{8}{11} \times \frac{3}{11}\left(\frac{1}{1000} + \frac{1}{1200}\right)}}$$

$$= 6.992 > 1.645$$

$$= Z_\alpha \; (\alpha = 5\,\%)$$

Conclusion :

As $|Z| > Z_\alpha$ (at 5 % l.o.s.)

$\therefore$ H_0 is rejected i.e. there is significant decrease in the consumption of coffee after increasing the excise duty.

Example 5.10.7 *Random sample of 400 men and 600 women were asked whether they would have a school near their residence. 200 men and 325 women were in favour of proposal. Test the hypothesis that the proportion of men and women in favour of proposal is same at 5 % level of significance.*

Solution : Given : $n_1 = 400, n_2 = 600$

$$p_1 = \frac{200}{400} = \frac{1}{2}$$

$$p_2 = \frac{325}{600} = \frac{13}{24}$$

$$P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = 0.525$$

$$Q = 1 - P = 0.475$$

H_0 : Proportion of men and women in favour of proposal is same i.e. $p_1 = p_2$

H_1 : $p_1 \neq p_2$ (Two Tailed test)

Under H_0

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{1}{2} - \frac{13}{24}}{\sqrt{(0.525/0.475)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.29$$

$|Z| = 1.29 < 1.96 = Z_\alpha$

Conclusion :

As $|Z| < Z_\alpha$ at 5 % l.o.s.

$\therefore$ H_0 is accepted.

5.10.4 Test of Significance for Single Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample of sign n from a large population $X_1, X_2, \ldots, X_N$ of size N with mean μ and variance σ^2. Therefore the standard error of mean of a random sample of size n from a population with variance σ^2 is $\sigma/\sqrt{n}$.

To test whether the given sample of size n has been drawn from a population with mean μ that is to test whether the difference between the sample mean and population mean is significant or not. Under the null hypothesis that there is no difference between the sample mean and population mean.

The test statistic is

$$Z = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right|$$

Where σ - Standard deviation of the population.

Note : If σ is unknown, we use the test statistic

$$Z_N = \left| \frac{\bar{X} - \mu}{s / \sqrt{n}} \right|$$

Where s = Standard deviation of the sample.

Examples

Example 5.10.8 *A normal population has mean 6.8 and standard deviation of 1.5. A sample of 400 members gave a mean of 6.75. Is the difference significant ?*

Solution : Given data : $\mu = 6.8, \sigma = 1.5, \bar{x} = 6.75$ and $n = 400$

Define

H_0 : There is no significant difference between $\bar{x}$ (sample mean) and μ (population mean)

i.e. $\mu = \bar{x}$

H_1 : there is significant difference between $\bar{x}$ and μ i.e. $\mu \neq \bar{x}$.

Test statistic :

$$|Z| = \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| = \left|\frac{6.75-6.8}{1.5/\sqrt{900}}\right| = |-0.67|$$

$Z = 0.67$

$\therefore$ As H_0 is accepted $|Z| = 0.67 < Z_\alpha = 1.96$ at 5 % level of significance.

$\therefore$ There is no significant difference between the sample mean and the population mean.

Example 5.10.9 *A sample of 400 male students is found to have mean height of 160 cms. Can it be reasonably regarded as a sample from large population with height 162.5 cms and standard deviation 4.5 cms.*

Solution : Given data, $n = 400$, $\mu = 162.5$ cms., $\bar{x} = 160$ cms, $\sigma = 4.5$ cms.

Note :
σ : S.D. of population,
$\bar{x}$: Sample mean,
μ : Population mean,
n : Sample size)
H_0 : Sample drawn from the population with mean $\mu = 162.5$.
H_1 : $\mu \neq 162.5$

Here $$|Z| = \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| = \left|\frac{160-162.5}{4.5/\sqrt{400}}\right| = \left|\frac{-2.5}{4.5}\times 20\right| = 11.11$$

As $|Z| = 11.11 > Z_\alpha = 1.96$

At 5 % level of significance H_0 is rejected.

Example 5.10.10 *The height of college students in the city are normally distributed with S.D. 6 cms. A sample of 1000 students has mean height 158 cms. Test the hypothesis mean height of college students in the city is 160 cms ?*

Solution : Given data, $n = 1000$, $\sigma = 6$, $\mu = 160$, $\bar{x} = 158$

here, H_0 : The sample drawn from the population with mean $\mu = 160$.

H_1 : $\mu \neq 160$

Under H_0 : $$|Z| = \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| = \left|\frac{-158-160}{6/\sqrt{1000}}\right| = |-10.54| = 10.54$$

As $|Z| = 10.54 > Z_\alpha = 1.96$ at 5 % level

H_0 is rejected at both 1 % to 5 % level of significance.

Example 5.10.11 *The average marks in mathematics of a sample of 100 students was 51 with S.D. of 6 marks. Could this have a random sample from the population with average marks 50 ?*

Solution : Given data :

$n = 100$, $\bar{x} = 51$, $S = 6$, $\mu = 50$

σ - Unknown

H_0 : The sample is drawn from the population with mean $\mu = 50$.

H_1 : $\mu \neq 50$

Under H_0 : $$|Z| = \left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| = \left|\frac{51-50}{6/\sqrt{100}}\right| = \frac{10}{6} = 1.666$$

Since $|Z| = 1.666 < Z_\alpha = 1.96$

$\therefore$ H_0 is accepted

Therefore the sample is drawn from the population with mean $\mu = 50$

5.10.5 Test of Significance for Two Means

Let $\bar{x}_1$ be the mean of sample size n_1 from the population with μ_1 and variance σ_1^2. Let $\bar{x}_2$ be the mean of an independent sample of size n_2 from another population with mean μ_2 and variance σ_2^2. The test statistics is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Note : If σ_1, σ_2 are not known $\sigma_1 \neq \sigma_2$ then

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under the null hypothesis that the samples are drawn from the same population there may arise two cases.

Case I : $\sigma_1 = \sigma_2 = \sigma$

The test statistic is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Case II : If $\sigma_1 = \sigma_2$ and σ is unknown

We know

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s^2}{n_1 + n_2}$$

Test statistics, $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Example 5.10.12 *From the data given below. Intelligence tests of two groups of boys and girls gave the following results. Examine the difference is significant.*

	Mean	S.D.	Size
Girls	70	10	70
Boys	75	11	110

Solution : H_0 : There is no significant difference between means of intelligences of boys and girls i.e. $\bar{x}_1 = \bar{x}_2$

H_1 : $\bar{x}_1 \neq \bar{x}_2$

Under H_0 $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{70 - 75}{\sqrt{\frac{(10)^2}{70} + \frac{(11)^2}{110}}} = -3.144$$

As $|Z| = 3.144 < Z_\alpha = 1.96$ at 5 level of significance.

H_0 is accepted.

∴ There is no significant difference between, the means of the intelligence of boys and girls.

Example 5.10.13 *From the given data below compute the standard error of the difference of the two sample means and find out if the two means significantly differ at 5 % level of significance.*

	No of item	Mean	S.D.
Group I	50	181.5	3.0
Group II	75	179	3.6

Solution : Refer above example 5.10.12.

H_0 : rejected

Example 5.10.14 *For sample I : $n_1 = 1000$, $\Sigma x = 49000$, $\Sigma(x - \bar{x})^2) = 7,84,000$,*
For sample II : $n_2 = 1500$, $\Sigma x = 70500$, $\Sigma(x - \bar{x})^2) = 24,00,000$.
Discuss the significant difference between mean score.

Solution : H_0 : There is no significant difference between the sample means i.e. $\bar{x}_1 = \bar{x}_2$.

$H : \bar{x}_1 \neq \bar{x}_2$

We need to calculate sample variance

$$S_1^2 = \sum \frac{(x_1 - \bar{x}_1)^2}{n_1} = \frac{784000}{1000} = 784$$

$$S_2^2 = \sum \frac{(x_2 - \bar{x}_2)^2}{n_2} = \frac{2400000}{1500} = 1600$$

Now we calculate the sample mean

$$\bar{x}_1 = \frac{\sum x_1}{n_1} = \frac{49000}{1000} = 49$$

$$\bar{x}_2 = \frac{\sum x_2}{n_2} = \frac{70500}{1500} = 47$$

The test statistic $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{49 - 47}{\sqrt{\frac{784}{1000} + \frac{1600}{1500}}} = 1.470$$

As $|Z| = 1.47 < 1.96$

At 5 % level of significance H_0 is rejected i.e. there is no significance difference between the sample mean.

Example 5.10.15 *An examination was given to 50 students of college A and 60 students of college B. For A the mean grade was 75 with S.D. of 9 and for B, the mean grade was 79 with S.D. of 7. Is there any significant difference between the performance of the students of college A and those of college B ?*

Solution : For college A

$n_1 = 50$ $\bar{x}_1 = 75$ $S_1 = 9$

For college B

$n_2 = 60$ $\bar{x}_2 = 79$ $S_2 = 7$

Define H_0 and H_1 and solve !

Ans. : H_0 : rejected

5.10.6 Test of Significance of Small Samples

When the sample size is less than 30 then the sample is called small sample.

Following two types are enlisted for small samples to test the significance.

1) Student t-test (t - test) 2) F-Test (Snedecor's variance test)

t-test : (for mean of one sample)

t-test is a small sample test. It is developed by William Gosset in 1908. It is also called as students t-test. T test assesses whether the means of two groups are statistically different from each other when the population standard deviation is unknown.

If $X_1, X_2, \ldots X_n$ is the random sample from the normal population N (μ, σ^2) then t-statistics is defined as

$$t = \frac{\bar{X} - \mu}{S\sqrt{n}}$$

S is standard deviation of the sample.

Where $$\bar{X} = \frac{\sum x_i}{n}$$

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

with degrees of freedom (n–1).

Solved Examples

Example 5.10.16 *A random sample of 16 newcomers gave mean of 1.67 m and s.d. of 0.16 m. The mean height of the students of the previous year known to be 1.600 m. Is the mean height of the newcomers significantly different from the main height of the students population of the previous years ?*

Solution : H_0 : There is no significant difference between the mean height of the newcomers and the mean height of the student population of the previous year.

i.e. $H_0 : \mu = 1.60$, $H_1 : \mu \neq 1.60$

here $n = 16$, $\bar{X} = 1.67$, $\mu = 1.60$, $S = 0.16$

$$\therefore \quad t = \frac{\bar{X} - \mu}{S\sqrt{n}} = \frac{1.67 - 1.60}{0.16/\sqrt{16}} = 1.75$$

Here, degrees of freedom = n – 1 = 16 – 1 = 15 from the t table $|t| = 1.75 < t_{0.05,\,15} = 2.13$

$\therefore$ H_0 accepted at 5 % l.o.s.

Therefore we conclude that the mean height of the newcomers is not significantly different from the mean height of the student population of the previous year.

Example 5.10.17 *Suppose that sweets are sold in packages of fixed weight of the contents. The procedure of the packages is interested in testing the average weight of content in packages in 1 kg. Hence a random sample of 12 packages is drawn and their contents found (in kg) are as follows : 1.05, 1.01, 1.04, 0.98, 0.96, 1.01, 0.97, 0.99, 0.98, 0.95, 0.97, 0.95. Using the above data what should he conclude about the average.*

Solution : $H_0 : \mu = 1$, $H_1 : \mu \neq 1$

here $n = 12$ $$\bar{X} = \frac{\sum x}{n} = \frac{11.86}{12} = 0.9883$$

X	$X - \bar{X}$	$(X - \bar{X})^2$
1.05	0.0617	0.00387
1.01	0.0217	0.000471
1.04	0.0517	0.002673
0.98	– 0.0083	+ 0.0000688
0.96	– 0.0283	0.000801
1.01	0.0217	0.000471
0.97	– 0.0183	0.000335
0.99	0.0017	0.00000289
0.98	– 0.0083	0.0000689
0.95	– 0.0383	0.001467
0.97	– 0.0183	0.000335
0.95	– 0.0383	0.001467
		$\Sigma(X - \bar{X})^2 = 0.011967$

$$S = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} = \sqrt{\frac{0.011967}{11}} = \sqrt{0.001088} = 0.032983$$

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{0.9883-1}{0.032983/\sqrt{12}} = -1.228815$$

$\therefore$ $|t| = 1.228815 < t_{0.05} = 2.201$ d.f. 11

H_0 accepted at 5 % level of significance.

$\therefore$ The average weight of contents of package taken as 1 kg.

5.10.7 t - Test for Comparison of Mean of Two Samples

Let $X_1, X_2, \ldots X_n$ and $X_1', X_2', \ldots X_n'$ be the values of two random samples of size n_1 and n_2 respectively from the same normal population $N(\mu, \sigma^2)$ then the t-test statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Where, $\bar{X}_1 = \frac{\sum X_i}{n_1}$ and $\bar{X}_2 = \frac{\sum X_i}{n_2}$

n_1 n_2

and $$S^2 = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_i-\bar{X}_2)^2}{n_1+n_2-2}$$

with degrees of freedom $n_1 + n_2 - 2$

Note :

1) If the two sample standard deviations S_1, S_2 are given then we have

$$S^2 = \frac{n_1 S_1^2 + n_1 S_2^2}{n_1+n_2-2}$$

2) Let $X_1, X_2, \ldots X_n$ and $X_1', X_2', \ldots X_n'$ be the values of two random samples from two different normal population $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ having μ_1 and μ_2 but the same standard deviation σ then the t-test statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

With $n_1 + n_2 - 2$ degrees of freedom.

Examples

Example 5.10.18 *Fertilizers A and B are tried respectively on 10 and 8 randomly chosen experimental plots. The yields in the plots were as given below. Test whether in the effects of the fertilizers as reflected in the mean yields.*

Fertilizers	Yields									
A	8.0	7.6	8.2	7.8	8.3	8.4	8.2	7.8	7.7	8.0
B	7.4	8.1	7.6	8.1	7.5	7.6	7.3	7.2		

Solution : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

here $n_1 = 10$, $\bar{X}_1 = 8.0$, $\sum (X_i - \bar{X}_1)^2 = 0.66$

$n_2 = 8$, $\bar{X}_2 = 7.6$, $\sum (X_i - \bar{X}_2)^2 = 0.80$

$$S^2 = \frac{0.66+0.80}{16} = \frac{1.46}{16} = 0.09125$$

$\therefore$ $$t = \frac{8.0-7.6}{\sqrt{0.09125}\sqrt{\frac{1}{10}+\frac{1}{8}}} = \frac{\bar{X}_1-\bar{X}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = 2.79$$

Degrees of freedom $= n_1 + n_2 - 2 = 10 + 8 - 2 = 16$

For 16 degrees of freedom, the probability of getting a value of t as high as 2.79 is less than 0.05 Therefore the value : at 5 % level of significance rejects null hypothesis. There is significant difference in the mean yields.

Example 5.10.19 *Samples of size 10 and 14 were taken from two normal populations with S.D. 3.5 and 5.2 The sample means were found to be 20.3 and 18.6. Test whether the means of the two populations are at the same level.*

Solution : $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Here $\bar{X}_1 = 20$, $n_1 = 10$, $S_1 = 3.5$,

$\bar{X}_2 = 18.6$, $n_2 = 14$, $S_1 = 5.2$,

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2} = \frac{10(3.5)^2 + 14(5.2)^2}{10+14-2} = 22.775$$

$$S = 4.772$$

$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.3 - 18.6}{\left(\sqrt{\frac{1}{10} + \frac{1}{14}}\right) 4.772} = 0.8604$$

$|t| = 0.8604 < t_{0.05}$ For degrees of freedom 22, $|t| = 2.07$

$\therefore H_0$ accepted.

5.10.8 F Test

Let $X_1, X_2, \ldots X_n$ and $X_1', X_2', \ldots X_n'$ are two independent random samples from two normal population $N(\mu_1, \sigma_1^2)$ and (μ_2, σ_2^2) respectively then the statistics of F-distribution is

$$F = \frac{S_1^2}{S_2^2} \quad \text{where } S_1^2 > S_2^2$$

Where $$S_1^2 = \frac{\sum(X_i - \overline{X}_1)^2}{n_1 - 1}$$

$$S_2^2 = \frac{\sum(X_i' - \overline{X}_2)^2}{n_2 - 1}$$

(n_1 and n_2 be the sizes of two samples with the variances S_1^2 and S_2^2)

Note :

1. In F-test. It is very important to note that the population variances are equal i.e. $\sigma_1^2 = \sigma_2^2$
2) The degrees of freedom is $(n_1 - 1, n_2 - 2)$

Test result : If the calculated value of F exceeds $F_{0.05}$ for $n_1 - 1, n_2 - 1$ degree of freedom from the given F-value table. We conclude that the ratio is significant at 5 % level.

Remark : When readymade data is given instead of tabular long data. We use

$$S_1^2 = \frac{n_1 S_1^2}{n_1 - 1} \qquad S_1^2 = \frac{\sum x_1^2 - n_1(\overline{x}_1)^2}{n_1 - 1}$$

$$S_2^2 = \frac{n_2 S_2^2}{n_2 - 1} \qquad S_2^2 = \frac{\sum x_2^2 - n_2(\overline{x}_2)^2}{n_2 - 1}$$

Example 5.10.20 *From the following information, calculate F statistics and also state the degree of freedom.*

	Sample 1	Sample 2
Size (n)	10	13
Mean ($\overline{x}$)	50	52
Sum of squares of deviations $\Sigma(x - \overline{x})^2$ from mean	144	264

Solution :

$$S_1^2 = \frac{\sum(x_1 - \overline{x}_1)^2}{n_1 - 1} = \frac{144}{10 - 1} = 16$$

$$S_2^2 = \frac{\sum(x_2 - \overline{x}_2)^2}{n_2 - 1} = \frac{264}{13 - 1} = 22$$

F test statistics

$$F = \frac{S_2^2}{S_1^2} \; (S_2^2 > S_1^2) = \frac{22}{66} = 1.38$$

Degrees of freedom = $n_2 - 1, n_1 - 1 = 13 - 1, 10 - 1 = 12, 9$

Example 5.10.21 *Find the F-statistics from the following data.*

Sample	Size (n)	Total observations Σx	Sum of squares of observations Σx^2
1	8	9.6	61.52
2	11	16.5	73.26

Solution :

$$\overline{X}_1 = \frac{\sum x_1}{n_1} = \frac{9.6}{8} = 1.2$$

and $$X_2^2 = \frac{\sum x_2}{n_2} = \frac{16.5}{11} = 1.5$$

$\therefore$ $$S_1^2 = \frac{\sum x_1^2 - n_1(\overline{x}_1)^2}{n_1 - 1} = \frac{61.52 - 8(1.2)^2}{8 - 1} = 7.14$$

$$S_2^2 = \frac{\sum x_2^2 - n_2(\bar{x}_2)^2}{n_2 - 1} = \frac{73.26 - 11(1.5)^2}{11 - 1} = 4.85$$

$\therefore$ F test statistics is

$$F = \frac{S_1^2}{S_2^2} = \frac{7.14}{9.85} = 1.47$$

Example 5.10.22 *From the following information. Calculate F-statistics and also its degrees of freedom.*

	Sample 1	Sample 2
Size (n)	22	16
Standard deviations (S)	2.9	3.8

Solution :

$$S_1^2 = \frac{n_1 S_1^2}{n_1 - 1} = \frac{22(29)^2}{22 - 1} = 8.81$$

$$S_2^2 = \frac{n_2 S_2^2}{n_2 - 1} = \frac{16(3.8)^2}{16 - 1} = 15.4$$

Test statistics

$$F = \frac{S_1^2}{S_2^2} = \frac{15.4}{8.81} = 1.75$$

Degrees of freedom $= n_2 - 1, n_1 - 1 = 16 - 1, 22 - 1 = 15, 21$

Example 5.10.23 *In two independent samples of size 8 and 10 the sum of squares deviations of the sample values from the respective sample means were 84.4 and 102.6. Test whether the difference of variances of the population is significant or not.*

Solution :

H_0 : There is no significant difference between population variance i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$

H_1 : There is significant difference between population variance.

For sample I :

$$n_1 = 8 \qquad \sum(x_i - \bar{x}_1)^2 = 84.4$$

For Sample II

$$n_2 = 10 \qquad \sum(x'_i - \bar{x}_2)^2 = 102.6$$

$\therefore$ $$S_1^2 = \frac{84.4}{7} = 12.057$$

$$S_2^2 = \frac{102.6}{9} = 11.4$$

$\therefore$ $$F = \frac{S_1^2}{S_2^2} \quad (\because \sigma_1^2 = \sigma_2^2) \quad \text{(Sample drawn from the same population or from two population with same variance)}$$

$$= \frac{12.057}{11.4} = 1.0576$$

Here the tabulated value of F at 5.1 level of significance with degrees of freedom.

$(n_1 - 1, n_2 - 1) = (7, 9)$ is 3.29

here $|F| = 1.057 < F_{0.05} = 3.29$

$\Rightarrow H_0$ is accepted.

Therefore there is no significant difference between the variance of two populations.

Example 5.10.24 *The following table gives the number of units produced per day by 2 workers A and B for some days. Can we say that worker A is more stable than worker B ? Use 5 % level of significance.*

Worker A	40	30	38	41	38	35		
Worker B	39	38	41	33	32	49	49	34

Solution :

H_0 : The stability of both workers is same $\sigma_1^2 = \sigma_2^2$

H_1 : Worker A is more stable than worker B.

$$n_1 = 6 \quad n_2 = 8$$

$$\bar{x}_1 = \frac{\sum X_i}{n_1} = \frac{222}{6} = 37$$

$$\bar{x}_2 = \frac{\sum X'_i}{n_2} = \frac{315}{8} = 39.38$$

$$S_1^2 = \frac{\sum(x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{80}{6 - 1} = 16$$

Worker A

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_2)^2$
40	3	9
30	–7	49
38	1	1
41	4	16
38	1	1
35	–2	4
Sum 222		**80**

Worker B

x_2	$x_2 - \bar{x}_2$	$(x_1 - \bar{x}_2)^2$
39	–0.38	0.14
38	–1.38	1.90
41	1.62	2.62
33	– 6.38	40.7
32	– 7.38	54.61
49	9.62	92.54
49	9.62	92.54
34	– 5.38	28.94
Sum 315	-	**313.99**

$$S_2^2 = \frac{\sum(x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{313.99}{8 - 1} = 44.86$$

F test statistics $$F = \frac{S_2^2}{S_1^2} = \frac{44.86}{16} = 2.80$$

Degree of freedom = $n_2 - 1, n_1 - 1 = 8 - 1,\ 6 - 1 = 7, 5$

Now we will write table value of F

$$F_{t,\ 0.05,\ 7,5} = 4.88$$

$$F = 2.80 < F_{t,\ 0.05} = 4.88$$

$\therefore H_0$ is accepted.

i.e. The stability of both workers is same.

Case Study :

1. Dieters lose more fat than exercise ?

Diet only

Sample mean : 5.5 kg.

Sample standard deviation : 4.2 kg

Sample size = n = 46

Stand error for dieters = $\frac{4.2}{\sqrt{46}} = 0.62$

Exercise only

Sample mean = 4.2 kg

Sample standard deviation = 3.8 kg

Sample size = n = 52

Standard error for exercise = $\frac{3.8}{\sqrt{52}} = 0.53$

Measure of variability = $\sqrt{(0.62)^2 + (0.53)^2} = 0.8156$

Testing a hypothesis

Step I : Determination of null and alternate hypothesis

H_0 : No difference in average fat lost in population for two methods. Population mean difference is zero.

H_1 : There is difference in average fat lost in population for two methods. Population mean difference is non-zero.

Step II : Test statistic

The sample mean difference = 5.5 – 4.2 = 1.3 kg

and the standard error of the difference is = 0.8156 ≈ 0.82

So Z = Test statistic = $\frac{1.3 - 0}{0.82} = 1.5853 \approx 1.56$

Step III : Decision

$\alpha = 0.05$ i.e. 5 %

As $Z = 1.56 < 1.96 (= Z_\alpha)$

$|Z| = Z_\alpha$

We accept H_0 - null hypothesis and we conclude that there is no significant difference between the two methods.

5.11 Chi-square Test

When any experiment is performed, it is observed that theoretical consideration and observed data are not same. There is always some difference between observed data and theoretical data. χ^2 (Chi-square) test helps to measure the magnitude of this discrepancy between observed and expected data.

Chi-square test is most commonly used test to check the goodness of fit of theoretical distribution to actual distribution.

Test statistic for χ^2 - test is given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \ldots(5.11.1)$$

where O_i ($i = 1, 2, \ldots n$) are observed frequencies.

E_i ($i = 1, 2, \ldots n$) are expected frequencies.

Such that

$$\Sigma O_i = \Sigma E_i = N \text{ (Total frequencies)}$$

Simplifying RHS of equation (5.11.1), we get

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{O_i^2 - 2O_iE_i - E_i^2}{E_i}\right]$$

$$= \sum_{i=1}^{n} \left[\frac{O_i^2}{E_i}\right] - 2\sum_{i=1}^{n} O_i - \sum_{i=1}^{n} E_i$$

$$= \sum_{i=1}^{n} \frac{O_i^2}{E_i} - 2N - N$$

$$\chi^2 = \sum_{i=1}^{n} \frac{O_i^2}{E_i} - N \quad \ldots(5.11.2)$$

Which is another form to calculate χ^2 value.

Remarks

i) χ^2 value ranges from 0 to ∞

ii) If $\chi^2 = 0$ then it indicates that observed and theoretical frequencies matches exactly.

iii) If $\chi^2 > 0$, shows that there is difference between observed and theoretical frequencies. The greater the discrepancy, greater the χ^2 value.

Note :

1) If expected frequencies are not provided then it can be calculated by using the formula.

Expected frequency $= \frac{\text{Row total} \times \text{column total}}{\text{Total frequency}}$ [formula is used for contigency table]

2) χ^2 test can be used as a

a) Test of goodness of fit

b) Test of homogeneity of independent estimates of population variance

c) Test of hypothetical value of population variance

d) List to homogeneity of independent estimates of population correlation.

3) Once expected frequency for one cell of 2×2 contigency table is calculated, then remaining frequencies can be also calculated by using the formula either (Row total – calculated frequency) or (Column total – Calculated frequency).

Conditions for applying χ^2 test :

1) Total frequency (N) should be large ($N > 50$).

2) Expected frequencies should be greater than or equal to five i.e. $E_i \geq 5 \; \forall \; i = 1, 2, \ldots, n$. Since χ^2 value is over estimated for small values of expected frequencies.

3) When $E_i < 5$ for some class "i" then that class is merged with neighbouring class until the total value of expected frequency becomes greater than or equal to 5. [While merging the class merge both expected and observed frequencies].

This process of grouping the classes is known as "**pooling**" of classes.

4) Note that the number of degrees of freedom is to be determined with the number of classes after pooling them.

Examples

Example 5.11.1 *A coin is tossed 160 times and following are expected and observed frequencies for number of heads.*

No. of heads	0	1	2	3	4
Expected frequency	17	52	54	31	6
Observed frequency	10	40	60	40	10

Find the χ^2 value.

Solution :

Degree of freedom = 5 – 1 = 4

We have $$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(10-17)^2}{17} + \frac{(40-52)^2}{52} + \frac{(60-54)^2}{54} + \frac{(40-31)^2}{31} + \frac{(10-6)^2}{6}$$

$$\chi^2 = 11.597$$

Example 5.11.2 *A set of five similar coins is tossed 210 times and the result is given in following table.*

No. of heads	0	1	2	3	4	5
Frequency	2	5	20	60	100	31

Test the hypothesis that data follows a binomial distribution. **SPPU : Dec.-01**

Solution : From given data

No. of coin tossed n = 5

No. of times coin tossed N = 210

∴ By binomial distribution, expected frequency for getting r heads is

Expected frequency $(E_r) = NP(X = r) = 210\ {}^5C_r p^r q^{n-r}$

Probability of getting head $p = \frac{1}{2}$

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

∴ $$E_r = 210\ {}^5C_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{5-r}$$

$$= 210\ {}^5C_r \left(\frac{1}{2}\right)^5$$

No. of heads	0	1	2	3	4	5
Expected frequency $E_r = 210\ {}^5C_r\left(\frac{1}{2}\right)^5$	7	33	66	66	33	7
Observed frequency	2	5	20	60	100	31
$(O_i - E_i)^2$	25	784	2116	36	4489	576

Now, degrees of freedom = 6 – 1 = 5

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{25}{7} + \frac{784}{33} + \frac{2116}{66} + \frac{36}{66} + \frac{4489}{33} + \frac{576}{7} = 278.251$$

Tabular χ^2 value for 5 d.f. at 5 % l.o.s. is 11.07.

Conclusion -

Since calculated χ^2 value is much greater than tabular χ^2 value, therefore the hypothesis that "Given data follows binomial distribution" is rejected.

Example 5.11.3 *In an experiment on pea breeding, the following frequencies of seeds were obtained.*

Round and green	Wrinkled and green	Round and yellow	Wrinkled and yellow	Total
222	120	32	150	524

Theory predicts that the frequencies should be in proportion 8:2:2:1. Examine the correspondence between theory and experiment.

Solution : Given proportion 8 : 2 : 2 : 1

8 + 2 + 2 + 1 = 13

Total frequency = 524

Then expected frequencies E_i are

$$\frac{8}{13} \times 524 = 323$$

$$\frac{2}{13} \times 524 = 81$$

$$\frac{2}{13} \times 524 = 81$$

$$\frac{1}{13} \times 524 = 40$$

Thus we have

Seed type	Round and green	Wrinkled and green	Round and yellow	Wrinkled and yellow
Observed frequencies (O_i)	222	120	32	150
Expected frequencies (E_i)	323	81	81	40
$(O_i - E_i)^2$	101^2	39^2	49^2	110^2

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{101^2}{323} + \frac{39^2}{81} + \frac{49^2}{66} + \frac{110^2}{40} = 382.502$$

Calculated χ^2 value = 382.502

Tabular χ^2 value at 5 % *l*.o.s. for 3 degrees of freedom = 7.815

Example 5.11.4 *Following table shows number of books issued, on the various days of week, from a certain library. At 5 % l.o.s. test the null hypothesis that number of books issued independent of the day.*

Day	Mon	Tue	Wed	Thur	Fri	Sat
No. of books issued	120	130	110	115	135	110

Solution : Given H_0 = No. of books issued is independent of days.

Total books issues = 120 + 130 + 110 +115 + 135 +110 = 720

No. of days n = 6

∴ Expected frequency $= \frac{720}{6} = 120$

Thus we have

Observed frequencies (O_i)	120	130	110	115	135	110
Expected frequencies (E_i)	120	120	120	120	120	120
$(O_i - E_i)^2$	0	100	100	25	225	100

∴ $$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{0}{120} + \frac{100}{120} + \frac{100}{120} + \frac{25}{120} + \frac{225}{120} + \frac{100}{120} = \frac{550}{120} = 4.5833$$

Calculated χ^2 value = 4.5833

Tabular χ^2 value for 5 degree of freedom and at 5 % of *l*.o.s. = 11.071

Conclusion :

∵ Calculated χ^2 value < Tabular χ^2 value

∴ We accept H_0

i.e. No. of books issued do not depend on days.

Example 5.11.5 *The figures given below are a) Observed frequencies of a distribution and b) the frequencies of a normal distribution having same mean, standard deviation and the total frequency as in (a) Apply χ^2 test of goodness of fit.*

(a)	1	5	20	28	42	22	15	5	2
(b)	1	6	18	25	40	25	18	6	1

Solution : From the given data, we observe that expected frequencies in first and last classes are less than 5. Hence we need to apply pooling of classes.

We will merge these classes with their neighbouring classes.

Thus we get

O_i(a)	E_i(b)	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1 + 5 = 6	1 + 6 = 7	1	0.1429
20	18	4	0.2222
28	25	9	0.36
42	40	4	0.1
22	25	9	0.36
15	18	9	0.5
5 + 2 = 7	6 + 1 = 7	0	0
		Total	1.6851

∴ $$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1.6851$$

d.f. $= 7 - 1 = 6$

Tabular value of $\chi^2 = 12.592$, at 5 % *l*.o.s. As calculated χ^2 value is less than tabular value. Thus fit is good.

Example 5.11.6 *In an antimalarial campaign in a certain area, quinine was administered to 812 persons out of total population of 3248. The no. of fever cases is given in the following table. Discuss the usefulness of quinine in checking malaria.*

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

Solution : Let the null hypothesis be

H_0 : Quinine is not effective/useful to check malaria.

H_1 : Quinine is useful in checking malaria.

From given contigency table degrees of freedom $= (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$

and expected frequency can be calculated by using the formula

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Total frequency}}$$

Calculation of expected frequencies

Treatment	Fever		No fever		Row total
	Observed frequency	Expected frequency	Observed frequency	Expected frequency	
Quinine	20	$\frac{240 \times 812}{3248} = 60$	792	$\frac{3008 \times 812}{3248} = 752$	812
No Quinine	220	180	2216	2256	2436
Column total	240		3008		3248 ← Total Frequency

Now consider the table

O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
20	60	1600	26.667
220	180	1600	8.889
792	752	1600	2.128
2216	2256	1600	0.709
		Total	38.393

$$\therefore \quad \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 38.392$$

From χ^2 table, for 1 d.f., at 5 % *l*.o.s.

$$\chi^2 = 3.84$$

Conclusion : Since calculated χ^2 value is greater than tabular χ^2 value.

$\therefore$ We reject the null hypothesis.

$\therefore$ Quinine is useful in checking malaria.

Exercise

1. *A engineer hypothesises that the mean number of defects can be decreased in manufacturing process of compact discs by using robot instead of humans for certain tasks. The mean number of defective discs per 1000 is 18. State H_0 and H_1.*

   H_0 :

   H_1 :

2. *A psychologist feels that playing soft music during a test will change the result of test. The psychologist is not sure whether the grades will be higher or lower. In the past mean of the scores was 74.*

   State H_0 and H_1.

3. *Let P is the probability of getting head when a coin is tossed once. Suppose that the hypothesis $H_0 : P = 0.5$ is rejected in favour of $H_1 : P = 0.6$, if 7 or more heads are obtained in 10 trials. Calculate the probability of type I error and power of the test.*

   [Ans. : $\alpha = \frac{11}{64}$, $\beta = 0.6177$]

4. *In a Bernoulli's distribution with parameter P, $H_0 : P = \frac{1}{2}$, against $H_1 : P = \frac{2}{3}$ is rejected if more than 3 heads are obtained out of 6 throws of a coin. Find the probabilities of type I and type II error.* **[Ans. : $\alpha = 0.3437$, $\beta = 0.3196$]**

5. An urn contains 6 marbles of which θ are white and the others black. In order to test the null hypothesis $H_0 : \theta = 3$, v/s $H_1 : \theta = 4$, two marbles are drawn at random (without replacement) and H_0 is rejected if both marbles are white otherwise H_0 is accepted. Find the probabilities of committing type I and type II error.
[Ans. : α = 0.20, β = 0.60]

6. Let $X_1, X_2, \ldots X_9$ be a random sample from I $(\theta, 25)$. If for testing $H_0 : \theta = 20$ against $H_1 : \theta = 26$, the critical region W is defined by $W = \{x \mid \bar{x} > 23.266\}$. Then find the size of critical region and the power.
[Ans. : 0.2389]

7. 325 women out of 600 women chosen from a city were found to be working in different sectors. Test the hypothesis that majority of women in the city are working.
[Ans. : H_0 is rejected at 5 % l.o.s.]

8. A random sample of 500 bolts was taken from a large consignment and 65 were found to be defective. Find the percentage of defective bolts in the consignment ?
[Ans. : Between 17.51 and 8.49]

9. A bag contains defective articles, the exact number of which is not known. A sample of 100 from the bag gives 10 defective articles. Find the limits for the proportion of defective articles in the bag.
[Ans. : limits are (0.1588, 0.0412)]

10. In a city a sample of 1000 people were taken and out of them 540 are vegetarian and the rest are non- vegetarian. Can we say that both habits of eating are equally popular in the city at i) 1 % level of significance ii) 5 % level of significance.
[Ans. : H_0 rejected at 5 % of l.o.s, H_0 accepted at 1 % of l.o.s]

Test of significance for difference of proportion

11. A manufacturing firm claims that its brand A product out sells its brand B product by 8 %. It is found that 42 out of a sample of 200 person prefer brand A and out of 18 out of another sample of 100 person prefer brand B. Test whether the 8 % difference is valid claim.
[Ans. : H_0 accepted]

12. In a pole submitted to the students body at university, 850 men and 560 women voted. 500 men and 320 women voted "yes". Does this indicate a significant difference of opinion between men and women on this matter at 1 % of level of significance ?
[Ans. : H_0 accepted]

13. In a town A, there were 956 births of which 52.5 % was males while in towns A and B combined, this proportion in total of 1406 birth was 0.496 Is there any significant difference in the proportion of male births in two towns.
[Ans. : H_0 rejected]

Test of significance for single mean

14. A random sample of 900 members has a mean 3.4 cms. Can it be reasonably regarded as a sample from a large population of mean 3.2 cms and standard deviation 2.3 cms ?
[Ans. : H_0 accepted]

15. The mean weight obtained from a random sample size 100 is 64 gms the S.D. of the weight distribution of the population is 3 gms. Test the statement that the mean weight of the population is 67 gms at 5 % level of significance.
[Ans. : H_0 rejected]

16. Five measurement of the tar content of a certain kind of cigarette yielded 14.5, 14.2, 14.4, 14.3 and 14.6 mg/cigarette. Assuming that the data are a random sample from a normal population show that at the 0.05 level of significance the null hypothesis $\mu = 14$ must be rejected in favour of the alternative hypothesis $\mu \neq 14$. (Hint : yo calculate S.D. of sample.)

Use S.D. $= \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

[Ans. : H_0 rejected]

17. A sample of 1000 students from a university was taken and their average weight was found to be 112 pounds with S.D. of 20 pounds. Could the mean weight of students in the population be 120 pounds.
[Ans. : H_0 rejected]

18. The average income of person was Rs. 210 with a S.D. of Rs. 10 in sample of 100 people of a city. For another sample of 150 persons the average income was Rs. 220 with S.D. of Rs. 12. The S.D. of incomes of the people of the city was Rs. 11. Test whether there is any significant difference between the average income of the localities.
[Ans. : H_0 rejected]

19. Intelligence tests on two groups of boys and girls gave the following results. Examine the difference is significant ?
[Ans. : H_0 accepted]

20. A product developer is interested in reading the drying time of a primer paint. Two formulation of the paint tested formulation 1 is the standard chemistry and formulation 2 has new dying of ingredient that should reduce the drying time is 8 minutes and this inherent variability should be unaffected by the addition of the new ingradient 10 specimens are painted with formulation 1 and another 10 specimens are painted with formulations 2. Two samples average drying time are $\bar{x}_1 = 121$ min and $\bar{x}_2 = 112$ min respectively. What conclusion can the product developer draw about the effectiveness of the new ingredients ?
[Ans. : H_0 accepted]

21. A study of the number lunches that executes in the insurance and bank studies claim as deductible expenses per month was based random samples and yields the following results. $n_1 = 40$, $\bar{x}_1 = 9$, $S_1 = 1.9$, $n_2 = 50$, $\bar{x}_2 = 8.0$, $S_2 = 2.1$, Test the null hypothesis against the alternative hypothesis ?
[Ans. : H_0 rejected]

22. *To find out whether the in habitants of two /slands may be regarded as having racial ancestry an anthropologist determines the cephalic incides of six adult males from each island getting $\bar{x}_1 = 77.4$ and $\bar{x}_2 = 72.2$ the corresponding standard deviations $\sigma_1 = 3.3$ and $\sigma_2 = 2.1$ Test whether the difference between the two samples means can reasonably be attributed to chance.* **[Ans. : H_0 rejected]**

23. *A group of 10 boys fed on diet A another group of 8 boys fed on different diet B recorded the following increase in weight (kgs).*

Diet A	5	6	8	1	12	4	3	9	6	10
Diet B	2	3	6	8	10	1	2	8		

[Ans. : H_0 accepted]

24. *A random sample of 10 boys had the I.Q.'s 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100. Do these data support the assumption of population mean I .Q. of 160 ?*
[Ans. : H_0 accepted]

25. *A sample of 18 items has a mean 24 units and standard deviation 3 units Test the hypothesis that it is a random sample from a normal population with mean 27 units.*
[Ans. : H_0 rejected]

26. *The average number of articles produced by two machines per day are 200 and 250 with standard deviation 20 to 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 5 % level of significance ?* **[Ans. : H_0 rejected]**

27. *For filling each bottle with 170 tablets of a perticular machine, an automatic machine was installed from the production a sample of 9 bottles was taken the number of tablet found in these 9 bottles are as follows. Test whether the machine has been installed properly or not. X : 168, 164, 166, 167, 168, 169, 170, 170, 171.* **[Ans. : H_0 rejected]**

F-Test

28. *The following data shows the runs scored by 2 batsman. Can it be said the performance of batsman A is more consistent than the performance of batsman B ? Use 1% level of significance.*

Batsman A	40	50	35	25	60	70	65	55
Batsman B	60	70	40	30	50			

[Ans. : H_0 accepted]

29. *Two random samples are drawn from 2 normal populations are as follows :*

A	17	27	18	25	27	29	13	17
B	16	16	20	27	26	25	21	

[Ans. : H_0 accepted]

30. *The two random samples reveals the following data*

Sample no.	size	Mean	Variance
I	16	440	40
II	25	460	42

[Ans. : H_0 accepted]

31. *Two independent samples of size 8 and 9 had the following values of the variables.*

Sample I	20	30	23	25	21	22	23	24	
Sample II	30	31	32	34	35	29	28	27	26

Do the estimates of the population variance differ significantly ?

32. *In a sample study, following information regarding demand for a particular spare part on various days was obtained from given data , at 5% l.o.s. fest that demand of no. of spare part does not depend on day.*

Days	Mon	Tue	Wed	Thur	Fri	Sat
No. of parts demanded	1124	1125	1110	1120	1126	1115

[Ans. : $\chi^2 = 0.18029$, Accept H_0]

33. *The table below gives the number of accidents that occurred in the certain factory on the various days of particular week.*

Days	Sun	Mon	Tue	Wed	Thur	Fri	Sat
No. of accidents	6	4	9	7	8	10	12

Test whether the accidents are uniformly distributed over the different days.
[Ans. : $\chi^2 = 5.25$, Accept H_0]

34. *Using the data given in following contigency table, test whether the eye colour in son is associated with the eye colour in father.*

Eye colour in father	Eye colour is son	
	Not light	Light
Not light	23	15
Light	15	47

[Ans.: $\chi^2 = 13.2$, Reject H_0 at 5 % l.o.s]

35. *In an experiment on pea breeding, a scientist obtained the following frequencies of seeds : 316 round and yellow, 102 wrinkled and yellow, 109 round and green and 33 wrinkled and green. Theory predicts that the frequencies of seeds should be in the proportion 9 : 3 : 3 : 1. respectively. Set a proper hypothesis and test it at 5 % l.o.s.*

[Ans.: $\chi^2 = 0.3555$, Accept H_0]

36. *From the information given below, test whether the type of occupation and attitude towards the social laws are independent [Use 1 % l.o.s] Attitude towards social laws.*

Occupation	Favourable	Neutral	Opposite
Blue collar	29	26	37
White collar	25	32	56
Professional	34	21	42

37. *The no. of defects per unit in a sample of 330 units of a manufactured product was found as follows.*

No. of defects	0	1	2	3	4
No. of units	214	92	20	3	1

Fit a Poisson distribution to the data and test for goodness of fit.

[Ans.: $\chi^2 = 0.0292$, Accept H_0]

[Hint - Apply pooling of classes]

38. *A set of 5 coins is tossed 3200 times and number of heads appearing each time is noted. The results are given below.*

No. of heads	0	1	2	3	4	5
Frequency	80	570	1100	900	500	50

Test the hypothesis that coins are unbiased. **[Ans.: $\chi^2 = 58.80$, Reject H_0]**

Multiple Choice Questions

Q.1 An assertion made about a population for testing, is called _____.

[a] hypothesis [b] statistic

[c] test-Statistic [d] level of significance

Q.2 If the assumed hypothesis is tested for rejection considering it to be true is called ?

[a] Null hypothesis [b] Statistical hypothesis

[c] Simple hypothesis [d] Composite hypothesis

Q.3 In _____, conclusions about a statement is drawn on the basis of a sample.

[a] null hypothesis [b] statistical hypothesis

[c] simple hypothesis [d] alternate hypothesis

Q.4 A hypothesis that specifies the population distribution completely, is called _____.

[a] composite hypothesis [b] simple hypothesis

[c] null hypothesis [d] alternate hypothesis

Q.5 The probability of rejecting null hypothesis when it is true is called as _____.

[a] level of significance [b] confidence level

[c] power of test [d] critical value

Q.6 _____ is the point where the null hypothesis gets rejected.

[a] Sample size [b] Power of test

[c] Critical value [d] None of these

Q.7 A region where the null hypothesis gets rejected is called _____.

[a] critical region [b] acceptance region

[c] critical value. [d] none of these

Q.8 If the critical region lies on the one side of the distribution then the test is referred as,

[a] Zero tailed [b] One tailed

[c] Two tailed [d] Three tailed

Q.9 If the critical region lies on the both sides of the distribution then the test is referred as,

[a] Zero tailed [b] One tailed

[c] Two tailed [d] Three tailed

Q.10 The type of test (One tailed/ Two tailed) depends on,

[a] Null hypothesis [b] Alternative hypothesis

[c] Simple hypothesis [d] Composite hypothesis

Q.11 A rule / formula used to test a null hypothesis is called _____.

[a] population statistic [b] variance statistic

[c] test statistic [d] null statistic

Q.12 Consider a null hypothesis $H_0 : \mu_0 = 15$ against $H_1 : \mu_0 > 15$. The test is _____.

[a] Two tailed [b] Left tailed

[c] Right tailed [d] None of these

Q.13 Consider a hypothesis where $H_0 : \mu_0 = 2$ against $H_1 : \mu_0 < 2$. The test is ?

[a] Two tailed [b] Left tailed

[c] Right tailed [d] None of these

Q.14 Type I error occurs when ?

[a] We accept H_0 if it is true [b] We reject H_0 if it is false

[c] We accept H_0 if it is false [d] We reject H_0 if it is true

Q.15 Type II error occurs when ?

[a] We accept H_0 if it is true [b] We reject H_0 if it is false

[c] We accept H_0 if it is false [d] We reject H_0 if it is true

Q.16 The probability of Type I error is referred as _____.

[a] $1 - \alpha$ [b] β

[c] α [d] $1 - \beta$

Q.17 Alternative hypothesis is a type of _____.

[a] research hypothesis [b] composite hypothesis

[c] null hypothesis [d] simple hypothesis

Q.18 If a null hypothesis (H_0) is accepted then the value of test statistic (z) lies in the _____.

[a] rejection region [b] acceptance region

[c] left tail of the curve [d] right tail of the curve

Q.19 The range of level of significance is _____.

[a] $-\infty$ to 0 [b] $-\infty$ to ∞

[c] 0 to ∞ [d] 0 to 1

Q.20 ____ is known as confidence coefficient.

[a] α [b] $1 - \alpha$

[c] β [d] $1 - \beta$

Q.21 The independent values in a set of values of a test is called as _____.

[a] degrees of freedom [b] test statistic

[c] level of Significance [d] level of confidence

Q.22 A bank utilizes these teller windows to render service to the customer. On a particular day 600 customer were served. If the customers are uniformly distributed over the counters. Expected numbers of customer served on each counter is _____.

[a] 100 [b] 200

[c] 300 [d] 150

Q.23 200 digits are chosen at random from a set of tables. The frequencies of the digits are as follows :

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

The expected frequency and degree of freedom is _____.

[a] 20 and 10 [b] 21 and 9

[c] 20 and 9 [d] 15 and 8

Q.24 In experiment on pea breeding, the observed frequencies are 222, 120, 32, 150 and expected frequencies are 323, 81, 81, 40, then χ^2_3 has the value _____.

[a] 382.502 [b] 380.50

[c] 429.59 [d] 303.82

Q.25 If observed frequencies O_1, O_2, O_3 are 5, 10, 15 and expected frequencies e_1, e_2, e_3 are each equal to 10, then χ^2_2 has the value _____.

[a] 20 [b] 10

[c] 15 [d] 5

Q.26 Number of books issued on six days of the week, excluding sunday which is holiday are given as 120, 130, 110, 115, 135, 110 and expectation is 120 books on each day, then χ^2_5 is _____.

[a] 2.58 [b] 3.56

[c] 6.56 [d] 4.58

Q.27 A coin tossed 160 times and following are expected are observed frequencies for number of heads

No. of heads	0	1	2	3	4
Observed frequency	17	52	54	31	6
Expected frequency	10	40	60	40	10

Then χ^2_4 is _____.

[a] 12.72 [b] 9.49

[c] 12.8 [d] 9.00

Q.28 Among 64 offspring's of a certain cross between guinea pig 34 were red, 10 were black and 20 were white. Acceding to genetic model, these numbers should in the ratio 9:3:4. Expected frequencies in the order _____.

[a] 36, 12, 16 [b] 12, 36, 16

[c] 20, 12, 16 [d] 36, 12, 25

Q.29 A sample analysis of examination results of 500 students was made. The observed frequencies are 220, 170, 90 and 20 and the numbers are in the ratio 4:3:2:1 for the various categories. Then the expected frequencies are _____.

[a] 150, 150, 50, 25 [b] 200, 100, 50, 10

[c] 200, 150, 100, 50 [d] 400, 300, 200, 100

Q.30 In experiment on pea breeding, the observed frequencies are 222, 120, 32, 150 and the theory predicts that the frequencies should be in proportion 8:2:2:1. Then the expected frequencies are _____.

[a] 323, 81, 40, 81 [b] 81, 323, 40, 81

[c] 323, 81, 81, 40 [d] 433, 81, 81, 35

Q.31 A sample analysis of certain data of 1000 students was made. The observed frequencies are 220, 270, 390 and 120 and the numbers are in the ratio 5:2:2:1 for the various categories. Then the expected frequencies are

[a] 150, 150, 50, 25 [b] 200, 100, 50, 10

[c] 500, 200, 200, 100 [d] 400, 300, 200, 100

Answers Keys for Multiple Choice Questions :

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

Q.21	a	**Q.22**	d	**Q.23**	c	**Q.24**	a	**Q.25**	d
Q.26	d	**Q.27**	a	**Q.28**	a	**Q.29**	c	**Q.30**	c
Q.31	c								

❑❑❑

Unit VI

6 Inferential Statistics : Tests for Hypothesis

Syllabus

Steps in Solving Testing of Hypothesis Problem, Optimum Tests Under Different Situations, Most Powerful Test (MP Test), Uniformly Most Powerful Test, likelihood Ratio Test, Properties of Likelihood Ratio Test, Test for the Mean of a Normal Population, Test for the Equality of Means of Two Normal Populations, Test for the Equality of -Means of Several Normal Populations, Test for the Variance of a Normal Population, Test for Equality of Variances of two Normal Populations, Non-parametric Methods, Advantages and Disadvantages of Non-parametric Methods.

Contents

6.1 Pre-requisites

Parameter space

Parameter space is a set of all values that a parameter of population can take. It is denoted by Θ.

e.g. :

1) Consider normal distribution $N(\mu, \sigma^2)$. Here the parameter μ (mean) can take any value from $-\infty$ to ∞ ; therefore parameter space for μ is

$$\Theta = \{\mu / -\infty < \mu < \infty\}$$

And the parameter σ^2 (variance) can take only positive values, therefore parameter space for σ^2 is

$$\Theta_0 = \{\sigma^2/\sigma^2 > 0\}$$

2) Consider a null hypothesis

$$H_0 : \mu \leq 4$$

Here μ can take any value from $-\infty$ to 4

$$\Theta_0 = \{\mu/-\infty < \mu \leq 4\}$$

Θ_0 is parameter space w.r.t. H_0

Clearly $\Theta_0 \leq \Theta$

Likelihood function

It describes how likely a specific population can fit the observed data.

Definition : Consider a population with p.d.f.

$f(X, \theta_1, \theta_2, \ldots \theta_k)$; where $X : x_1, x_2, x_3, \ldots x_n$ is a random sample. Then likelihood function of given sample observation is

$$L = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2, \ldots, \theta_k)$$

6.2 Optimum Test Under Different Situations

In hypothesis testing, our main focus is to minimize the type II error (β) or to maximize power ($1 - \beta$) under fixed α. This can be done by proper choice of test statistics and critical region.

Henceforth, following methods help us to choose best critical region.

1. Most powerful test (MP)

2. Uniformly most powerful test (UMP)
3. Likelihood ratio test

6.3 Best Critical Region (BCR) / Most Powerful Critical region / Most Powerful Test (MP)

In testing simple null hypothesis (H_0) vs simple alternative hypothesis (H_1) the critical region which has maximum power ($1- \beta$) among the set of all critical region belongs to sample space is called most powerful test.

The region is called most powerful region.

While testing simple null hypothesis

$$H_0 : \theta = \theta_0$$

vs

$$H_1 : \theta = \theta_1 \text{ (Simple alternative hypothesis)}$$

The critical region W is said to be most powerful critical region of size α.

if $$P(x \in W/ H_0) = \int_W L_0 \, dx = \alpha$$

$$P(x \in W/H_1) \geq P(x \in W_1/H_1)$$

Where, W_1 is any other critical region.

6.4 Uniformly Most Powerful Critical Region or UMP Test

In testing simple/composite null hypothesis vs composite alternative hypothesis, the critical region which has maximum power among the set of all critical regions belongs to sample is called UMPCR or UMP.

UMP region may or may not exist. However, for $H_0 : \theta = \theta_0$ vs $H_1 : \theta_1 > \theta_0$ or $H_1 : \theta_1 < \theta_0$ a UMP test exist.

Note : BCR or UMPCR is find out with the help of Neymann-Pearsson Lemma.

Neyman - Pearson Fundamental Lemma (N-P-Lemma) Statement :

Let $k > 0$ and W be a critical region of size α such that

$$W = \left\{x \in s \; ; \frac{L_1}{L_0} > k\right\} \quad \ldots (6.4.1)$$

$$\bar{W} = \left\{x \in s \; ; \frac{L_1}{L_0} \leq k\right\} \quad \ldots (6.4.2)$$

Where L_1 and L_0 are the likelihood function of sample observations under H_1 and H_0 respectively then W is the most powerful critical region of size α to test hypothesis.

$$H_0 : \theta = \theta_0 \quad Vs \quad H_1 : \theta = \theta_1$$

Proof : Let W be the critical region of size α then

$$P(x \in W/H_0) = \alpha = \int_W L_0 \, dx \qquad \ldots (6.4.3)$$

$$P(x \in W/H_1) = 1 - \beta = \int_W L_1 \, dx \qquad \ldots (6.4.4)$$

Let, W_1 be any other CR of size $\alpha_1 \leq \alpha$.

$$P(x \in W_1/H_0) = \alpha_1 = \int_{W_1} L_0 \, dx \qquad \ldots (6.4.5)$$

$$P(x \in W_1/H_1) = 1 - \beta_1 = \int_{W_1} L_1 \, dx \qquad \ldots (6.4.6)$$

Consider ,

$$\alpha_1 \leq \alpha$$

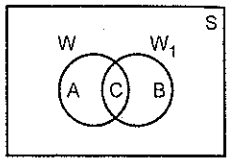


Fig. 6.4.1

from (6.4.5) and (6.4.6)

$$\int_{W_1} L_0 \, dx \leq \int_{W_1} L_0 \, dx$$

$$\int_{BUC} L_0 \, dx \leq \int_{AUC} L_0 \, dx$$

$$\Rightarrow \quad \int_B L_0 \, dx \leq \int_A L_0 \, dx$$

$$\therefore \quad \boxed{\int_A L_0 \, dx = \int_B L_0 \, dx}$$

Now, consider (6.4.1),

$$\frac{L_1}{L_0} > k \qquad \text{(on W)}$$

$$\Rightarrow \quad L_1 > k L_0$$

$$\Rightarrow \quad \int_W L_1 \, dx > k \int_W L_0 \, dx$$

$$\int_A L_1 \, dx > k \cdot \int_A L_0 \, dx \geq k \cdot \int_B L_0 \, dx$$

$$\int_A L_1 \, dx \geq k \cdot \int_B L_0 \, dx$$

$$\therefore \quad \boxed{k \cdot \int_B L_0 \, dx \leq \int_A L_1 \, dx}$$

Now consider (6.4.3),

$$\frac{L_1}{L_0} < k \qquad \text{(on } \bar{W}\text{)}$$

$$\Rightarrow \quad L_1 < k L_0$$

$$\int_B L_1 \, dx < k \int_B L_0 \, dx \leq \int_A L_1 \, dx$$

$$\int_B L_1 \, dx \leq \int_A L_1 \, dx$$

$$\int_{BUC} L_1 \, dx \leq \int_{AUC} L_1 \, dx$$

$$\int_{W_1} L_1 \, dx \leq \int_W L_1 \, dx \qquad \text{From (6.4.6) and (6.4.4)}$$

$$1 - \beta_1 \leq 1 - \beta$$

$\Rightarrow$ Power of $W_1 \leq$ Power of W

$\Rightarrow$ W is the best critical region.

Unbiased test :

If the power of a test $(1 - \beta)$ is never less than its size α i.e. $1 - \beta \geq \alpha$.

then the test is known as unbiased test.

Note : For MPCR or BCR, power is always greater than its size, so test is always unbiased.

Examples

Example 6.4.1 *A single observation is drawn from the distribution*

$f(x, \theta) = \frac{2x}{2\theta + 1}$; $0 \leq x \leq \theta + 1$,

$= 0$, $\theta \leq \theta + 1$.

It is require to test $H_0 : \theta = 1$ vs $H_1 : \theta = 1.5$, H_1 is rejected iff the observation is > 1.7. Calculate the probability the errors of two kinds.

Solution : Given :

$$f(x, \theta) = \frac{2x}{2\theta + 1} \; ; \theta \le x \le \theta + 1$$

$$= 0 \qquad \theta \le \theta + 1$$

$H_0 : \theta = 1$ Vs $H_1 = \theta = 1.5$

$\alpha = P[\text{Reject } H_0 / H_0]$

$= P[X > 1.7 / H_0]$

$= 1 - P[X < 1.7 / H_0]$

$$= 1 - \int_{\theta=1}^{1.7} \frac{2x}{3} dx$$

$$= 1 - \frac{2}{3}\left[\frac{x^2}{2}\right]_1^{1.7}$$

$$\alpha = 1 - \frac{2}{3}\left[\frac{(1.7)^2}{2} - \frac{1}{2}\right]$$

$$\boxed{\alpha = 0.37}$$

$\beta = P[\text{Accept } H_0 / H_1]$

$= P[X < 1.7 / H_1]$

$$= \int_{\theta=1.5}^{1.7} \frac{2x}{3} dx$$

$$= \frac{1}{2}\left[\frac{x^2}{2}\right]_{1.5}^{1.7}$$

$$= \frac{1}{4}[0.64]$$

$$\therefore \quad \boxed{\beta = 0.16}$$

Example 6.4.2 *Let X be a.r.v. with probability density function $f_0(x)$ under H_0 and $f_1(x)$ under H_1 as given below.*

X	1	2	3	4
$f_0(x)$	0.1	0.2	0.3	0.4
$f_1(x)$	0.4	0.3	0.2	0.1

Find

1) All the Critical Regions (CR) for which the probability of type I error is 0.2.

2) Find the C.R. in [1] which has maximum power.

3) Find all CR for which the probability of type I error is ≤ 0.2.

4) Find the C.R. in (3) which has maximum power.

5) IS the CR in (4) more powerful than CR in (2).

Solution. : $H_0 : f = f_0$ vs $H_1 : f = f_1$

1] $\alpha = 0.2$

We know that $\alpha = P[\text{Reject } H_0 / H_0]$

$0.2 = P[X \in W / H_0]$

where $W = \{1, 2, 3, 4\}$

$\therefore \quad X = 2$ is the critical region.

2] Power of the test $= 1 - \beta$

$= 1 - P[\text{Accept } H_0 / H_1]$

$= P[\text{Reject } H_0 / H_1]$

$= P[X = 2 / H_1]$

$$\boxed{1 - \beta = 0.3}$$

3] There are two critical regions for which $\alpha \le 0.2$

i) $W_1 = \{X = 2\}$

$\therefore$ Power of test $= P[\text{Reject } H_0 / H_1]$

$= P[X = 2 / H_1]$

$= 0.3$

ii) $W_2 = \{X = 1\}$

$\therefore$ Power of test $= P[\text{Reject } H_0 / H_1]$

$= P[X = 1 / H_1]$

$= 0.4$

4] W_2 is more powerful than W_1

5] Yes the C.R. $\{X = 1\}$ is more powerful than C.R. $\{X = 2\}$

Example 6.4.3 *Given the frequency function*

$$f(x, \theta) = \frac{1}{\theta} \; ; 0 \le x \le \theta$$

$= 0$; elsewhere

and that you are testing the null hypothesis $H_0 : \theta = 1$ vs $H_1 : \theta = 2$ by means of a single observed value of x. What would be the size of the type I and type II errors, if you

choose the interval i) $0.5 \le x$ ii) $1 \le x \le 1.5$ as the critical regions ? Also obtain the power function of the test.

Solution : $H_0 : \theta = 1$ vs $H_1 : \theta = 2$

Here C.R. $= W = \{x : 0.5 \le x\}$

$= \{x : x \ge 0.5\}$

$\alpha = P\{x \in W/H_0\}$

$= P\{x \ge 0.5/\theta = 1\}$

$= P\{0.5 \le x \le \theta/\theta = 1\}$

$= P\{0.5 \le x \le 1/\theta = 1\}$

$$= \int_{0.5}^{1} \frac{1}{\theta_{[\theta=1]}} dx$$

$$= \int_{0.5}^{1} 1 dx = 0.5$$

Similarly, $\beta = P[x \in \bar{W}/H_1]$

$= P[x \le 0.5/\theta = 2]$

$$= \int_{0}^{0.5} [(x, \theta)]_{\theta=2} dx$$

$$= \int_{0}^{0.5} \frac{1}{2} dx$$

$\beta = 0.25$

Therefore size of type I and type II errors are $\alpha = 0.5$ and $\beta = 0.25$ respectively.

ii) $W = CR = \{x : 1 \le x \le 1.5\}$

$\alpha = P\{x \in W/\theta = 1\}$

$$= \int_{1}^{1.5} [f(x, \theta)]_{\theta=1} dx$$

$= 0$

Since under $H_0 : \theta = 1$

$f(x, \theta) = 0$, for $1 \le x \le 1.5$

$\beta = P\{x \in \bar{W}/\theta = 2\}$

$= 1 - P\{x \in W/\theta = 2\}$

$$= 1 - \int_{1}^{1.5} [f(x, \theta)]_{\theta=2} dx$$

$\beta = 0.75$

$\therefore$ Power function $= 1 - \beta = 0.25$

Example 6.4.4 *If $x \ge 1$ is the critical region for testing $\Theta_0 : \theta = 2$ against the alternative $\theta = 1$ on the basis of the single observation from the population $f(x, \theta) = \theta e^{-\theta x}$ $0 \le x < \infty$, obtain the values of type I and type II error.*

Solution : Here,

$W = \{x : x \ge 1\}$ and $\bar{W} = \{x : x < 1\}$

$H_0 : \theta = 2$ and $H_1 : \theta = 1$

$\alpha =$ Size of the type I error

$= P[x \in W/H_0]$

$= P[x \ge 1/\theta = 2]$

$$= \int_{1}^{\infty} e^{-2x} dx$$

$$= 2\left[\frac{e^{-2x}}{2}\right]_{1}^{\infty}$$

$$= e^{-2} = \frac{1}{e^2}$$

$\beta =$ Size of type II error

$= P[x \in \bar{W}/H_1] = P[x < 1/\theta = 1]$

$$= \int_{0}^{1} e^{-x} dx = \left[\frac{e^{-x}}{-1}\right]_{0}^{1}$$

$$\beta = \frac{e-1}{e}$$

Example 6.4.5 *Let $X_1, X_2, \ldots, X_n$ be a random sample of a normal population $N(\mu, 16)$. Find BCR using UMP to test the null hypothesis $H_0 : \mu = 10$ vs $H_1 : \mu > 10$ (Given $n = 16$ and $\alpha = 0.05$)*

Solution : Pdf for normal distribution is

$$P.d.f = (2\pi\sigma_0^2)^{-1/2} \exp\left(-\frac{1}{2}\frac{(x_j - \mu_0)^2}{\sigma_0^2}\right)$$

Mean $= \mu_0$, variance $= \sigma_0^2$

The likehood function is

$$L(\mu, \sigma) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Therefore the ratio of likelihood function is

$$\frac{L(10)}{L(\mu)} = \frac{(32\pi)^{-16/2}\exp\left[\frac{-1}{32}\sum_{i=1}^{16}(x_i-10)^2\right]}{(32\pi)^{-16/2}\exp\left[\left(\frac{-1}{32}\right)\sum_{i=1}^{16}(x_i-\mu)^2\right]} \le K$$

on cancellation,

$$\exp\left\{-\frac{1}{32}\left(\sum_{i=1}^{16}(x_i-10)^2 - \sum_{i=1}^{16}(x_i-\mu)^2\right)\right\} \le K$$

$$\therefore \quad \exp\left\{-\frac{1}{32}\left(\Sigma x_i^2 - 20\,\Sigma x_i + 16\times(10)^2 - \Sigma x_i^2 + 2\mu\,\Sigma x_i - 16\,\mu^2\right)\right\} \le K$$

$$\Rightarrow \quad \exp\left\{-\frac{1}{32}\left(-2\,\Sigma x_i(\mu-10) + 16\,(\mu^2-10^2)\right)\right\} \le K$$

Taking natural log on both the sides

$$-2(\mu-10)\,\Sigma x_i + 16\,(\mu^2-10^2) \le 32\ln(k)$$

$$\Rightarrow \quad \frac{1}{16}\Sigma x_i \ge -\frac{1}{16(2(\mu-10))}\left(32\ln k - 16\,(\mu^2-10^2)\right)$$

$$= k^* \text{ say}$$

We know that

$$= \frac{L(10)}{L(\mu)} \le K \text{ iff } \bar{x} \ge k^*$$

Therefore BCR of size α for $H_0 : \mu = 10$ vs

$$H_1 : \mu > 10$$

is given by

$$CR = \{(x_1, x_2, \ldots, x_n) : \bar{x} \ge k^*\}$$

$$\therefore \quad \alpha = (P\,X \ge k^*) \text{ when } \mu = 10$$

$\therefore$ CR is a uniformly most powerful region of size α.

Example 6.4.6 *For distribution*

$$dF = \begin{cases} \beta\, e^{-\beta(x-\gamma)}\,dx, & x \ge \gamma \\ 0 & x < \gamma \end{cases}$$

Show that for $H_0 : \beta = \beta_0, \gamma = \gamma_0$ and $H_1 : \beta = \beta_1, \gamma = \gamma$ the best. Best critical region is given by $\bar{x} = \frac{1}{\beta-\beta_0}\left\{\gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n}\log k + \log\frac{\beta_1}{\beta_0}\right\}$. Provided $\gamma_1 \le \gamma_0, \beta_1 \ge \beta_0$

Solution :

$$f(x;\beta,\gamma) = B\exp\{-\beta(x-\gamma)\},\ x \ge \gamma$$
$$= 0, \quad \text{otherwise}$$

the likelihood function is given

$$L(x,\beta,\gamma) = \beta^n \exp\left\{-\beta\sum_{i=1}^{n}(x_i-\gamma)\right\}$$

using N-P lemma.

Ratio of likelihood function is given by,

$$\frac{L(\gamma_1)}{L(\gamma_0)} = \frac{\beta_1^n \exp\left\{-\beta_1\sum_{i=1}^{n}(x_i-\gamma_1)\right\}}{\beta_0^n \exp\left\{-\beta_0\sum_{i=1}^{n}(x_i-\gamma_0)\right\}} \ge k$$

$$\Rightarrow \quad \left(\frac{\beta_1}{\beta_0}\right)^n \exp\left\{-\beta_1\sum_{i=1}^{n}(x_i-\gamma_1) + \beta_0\, n\,(\bar{x}-\gamma_0)\right\} \ge k$$

Taking log,

$$n\ln\left(\frac{\beta_1}{\beta_0}\right) - n\bar{x}\,(\beta_1-\beta_0) + n\beta_1\gamma_1 - n\beta_0\gamma_0 \ge \ln k$$

$$\Rightarrow \quad \bar{x}\,(\beta_1-\beta_0) \le \left\{\gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n}\ln k + \ln\left(\frac{\beta_1}{\beta_0}\right)\right\}$$

$$\Rightarrow \quad \bar{x} \le \frac{1}{\beta_1-\beta_0} \le \left\{\gamma_1\beta_1 - \gamma_0\beta_0 - \frac{1}{n}\ln k + \ln\left(\frac{\beta_1}{\beta_0}\right)\right\} : \beta_1 > \beta_0$$

Example 6.4.7 *Check whether a BCR exists for testing the null hypothesis $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1 > \theta_0$ for the parameter θ of the distribution. $f(x,\theta) = \frac{1+\theta}{(x+\theta)^2}$; $1 \le x \le \infty$*

Solution :

For $$f(x,\theta) = \frac{1+\theta}{(x+\theta)^2}$$

likelihood function is defined as

$$\prod_{i=1}^{n} f(x_i, \theta) = L(x, \theta) = (1+\theta)^n \prod_{i=1}^{n} \frac{1}{(x_i+\theta)^2}$$

By N-P lemma

Ratio of likelihood function is given as

$$\frac{(1+\theta_1)^n \prod_{i=1}^{n} \frac{1}{(x_i+\theta_1)^2}}{(1+\theta_0)^n \prod_{i=1}^{n} \frac{1}{(x_i+\theta_0)^2}} \geq k$$

$$\Rightarrow \quad (1+\theta_1)^n \prod_{i=1}^{n} \frac{1}{(x_i+\theta_1)^2} \geq k (1+\theta_0)^n \prod_{i=1}^{n} \frac{1}{(x_i+\theta_0)^2}$$

Taking log

$$\Rightarrow \quad n \ln(1+\theta_1) - 2 \sum_{i=1}^{n} \ln(x_i+\theta_1) \geq \ln(k) + n \ln(1+\theta_0) - 2 \sum_{i=1}^{n} \ln(x_i+\theta_0)$$

$$\Rightarrow \quad 2 \sum_{i=1}^{n} \ln\left(\frac{x_i+\theta_0}{x_i+\theta_1}\right) \geq \ln k + n \ln\left(\frac{1+\theta_0}{1+\theta_1}\right)$$

here the test criterion is of the form

$$\sum_{i=1}^{n} \ln\left(\frac{x_i+\theta_0}{x_i+\theta_1}\right)$$

Which can't be put in the form of a function of the sample observations, not depending on hypothesis.

Hence no B.C.R. exist in this case.

6.5 Likelihood Ratio Test (L.R.T.)

N-P lemma is applicable only if both H_0 and H_1 are simple hypotheses. But if either H_0 or H_1 or both are composite then no such theorem is available. In such cases we can employ the likelihood ratio test (LRT), to find BCR and power of the test.

Statement -

Let Θ be a parametric space, Θ_0 and Θ_1 are parametric spaces w.r.t. null Θ_0 and alternative hypothesis Θ_1 respectively. Such that $\Theta = \Theta_0 \cup \Theta_1$.

Consider a random variable X with p.d.f. $f(X, \theta)$ where $\theta \in \Theta$

To test $H_0 : \theta \in \Theta_0$ V/s $H_1 : \theta \in \Theta_1$,

LRT statistics denoted by λ or $\lambda(x)$ is given by,

$$\lambda = \lambda(x) = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\max_{\theta \in \Theta_0} L(x,\theta)}{\max_{\theta \in \Theta} L(x,\theta)}$$

where $L(x, \theta)$ is likelihood function then LRT says

Reject H_0 if, $\lambda < C$

where C is constant choosen such that

$$P[\lambda < C / H_0] = \alpha$$

i.e. C is choosen s.t. probability of type I error is of size α.

Remark :

i) Clearly $\lambda \geq 0$

ii) Since $\Theta_0 \subset \Theta \Rightarrow \max_{\theta \in \Theta_0} L(x,\theta) \leq \max_{\theta \in \Theta} L(x, \theta)$

$$\Rightarrow \quad L(\hat{\Theta}_0 \leq L(\hat{\Theta}_0)$$

$$\Rightarrow \quad \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} \leq 1$$

$$\Rightarrow \quad \lambda \leq 1$$

Thus $0 \leq \lambda \leq 1$ (Using remark i)

In other words, smaller the value of λ, more the chances of rejecting H_0.

iii) If H_0 and H_1 both are simple hypothesis, then L.R.T. is equivalent to N-P lemma.

iv) If either one or both hypotheses (H_0 and H_1) are composite then LRT statistic λ is a function of every sufficient statistic for θ.

6.6 Properties of Likelihood Ratio Test

1) If null and alternative hypothesis are both simple then LRT leads to the N-P lemma.
2) If UMP test at all exists then LRT is generally UMP. In this case following asymptotic properties of LRT are satisfied.
   a) Under particular conditions, $-2 \log_e \lambda$ has asymptotic chi-square distribution.
   b) Under certain assumptions LRT is consistent.

6.7 Test for the Mean of Normal Population

Mean and variance both are unknown

Let $(x_1, x_2, \ldots, x_n)$ be a random sample of size n from normal population with mean μ and variance σ^2, where both μ and σ^2 are unknown.

Consider the null hypothesis (composite)

$$H_0 : \mu = \mu_0 ; \quad 0 < \sigma^2 < \infty$$

V/s an alternative hypothesis,

$$H_1 : \mu \neq \mu_0 ; 0 < \sigma^2 < \infty$$

For unknown μ and σ^2, parameter space over the entire region is

$$\Theta = \{(\mu, \sigma^2) / -\infty < \mu < \infty , 0 < \sigma^2 < \infty\}$$

and the parameter space determined by the null hypothesis is

$$\Theta_0 = \{(\mu, \sigma^2) / \mu = \mu_0, \ \sigma^2 > 0\}$$

Also we have, for $N(x, \sigma^2)$ distribution curve is given by the formula.

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Then likelihood function of the sample observations $x_1, x_2, \ldots, x_n$ is given by

$$L(\Theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{n}{2}} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2\right] \quad \ldots (6.7.1)$$

The maximum likelihood estimates (MLE) of μ and σ^2 are given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \quad \ldots (6.7.2)$$

$$\hat{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2 \text{ (say)}$$

Substituting of equation (6.7.2) in equation (6.7.1), we get

$$L(\hat{\Theta}) = \left[\frac{1}{2\pi s^2}\right]^{\frac{n}{2}} \exp\left[-\frac{ns^2}{2s^2}\right] \quad \ldots \left[\because \frac{1}{n}\Sigma(x_i - \bar{x}^2) = s^2\right]$$

$$= \left[\frac{1}{2\pi s^2}\right]^{\frac{n}{2}} e^{-n/2} \quad \ldots (6.7.3)$$

Which is maximum likelihood estimate of L in parameter space Θ.

Similarly in (Θ_0) (parameter space over H_0), as μ is specified, σ^2 is the only variable parameter

$\therefore$ Maximum likelihood estimate of σ^2 under H_0 is

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu_0)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - \bar{x})^2 + (\bar{x} - \mu_0)^2 + 2(x_i - \bar{x})(\bar{x} - \mu_0)\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{n} \sum_{i=1}^{n} (\bar{x} - \mu_0)^2 + \frac{2}{n}(\bar{x} - \mu_0) \sum_{i=1}^{n} (x_i - \bar{x})$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{(\bar{x} - \mu_0)^2}{n} \sum_{i=1}^{n} 1$$

$$+ 2(\bar{x} - \mu_0)\left(\frac{\sum x_i}{n} - \frac{\bar{x}}{n}\sum 1\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{1}{n}(\bar{x} - \mu_0)^2 (n) + 2(\bar{x} - \mu_0)\left(\bar{x} - \frac{\bar{x}}{n} n\right),$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 + (\bar{x} - \mu_0)^2 = s_0^2 \quad \ldots (6.7.4)$$

Substituting equation (6.7.4) in equation (6.7.1) we get,

$$L(\hat{\Theta}_0) = \left[\frac{1}{2\pi s_0^2}\right]^{\frac{n}{2}} e^{-n/2} \quad \ldots (6.7.5)$$

Thus from equation (6.7.3) and (6.7.4), LRT statistic is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left[\frac{s^2}{s_0^2}\right]^{n/2} \quad \ldots (6.7.6)$$

$$\lambda = \left[\frac{s^2}{s^2 + (\bar{x} - \mu_0)^2}\right]^{n/2} \quad \ldots \text{(from equations (6.7.2) and (6.7.4))}$$

$$\lambda = \left[\frac{1}{1 + \left[\frac{\bar{x} - \mu_0}{s}\right]^2}\right]^{n/2} \quad \ldots (6.7.7)$$

Now, according to students t - distribution with (n – 1) d.f., the statistics is

$$t = \frac{(\bar{x} - \mu_0)}{S/\sqrt{n}}$$

where $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{ns^2}{n-1}$$ (From equation (6.7.2))

$$\therefore \quad t = \frac{\bar{x} - \mu_0}{\left(\frac{n - S^2}{n-1}\right)^{1/2} / \sqrt{n}}$$

$$= \frac{\bar{x} - \mu_0}{S/\sqrt{n-1}} \sim t_{n-1}$$

$$\Rightarrow \quad \frac{t}{\sqrt{n-1}} = \frac{\bar{x} - \mu_0}{S}$$

Thus equation (6.7.7) reduces to

$$\lambda = \frac{1}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}} = \phi(t^2) \quad \text{(say)}$$

The critical region of the LRT viz $0 < \lambda < C$ is equivalent to

$$\left(1 + \frac{t^2}{n-1}\right)^{-n/2} \leq C$$

$$\Rightarrow \quad \left(1 + \frac{t^2}{n-1}\right)^{n/2} \geq \frac{1}{C}$$

$$\Rightarrow \quad 1 + \frac{t^2}{n-1} \geq C^{-2/n}$$

$$\Rightarrow \quad t^2 \geq (n-1)[C^{-2/n} - 1]$$

$$\geq K^2 \text{ (say)}$$

Thus the critical region may be will defined by

$$|t| = \left|\frac{\sqrt{n}(\bar{x} - \mu_0)}{S}\right| \geq K$$

where the constant K is derived such that size of critical region α i.e.

$$P(|t| \geq K \mid H_0) = \alpha$$

Since, under H_0, the statistic "t" follows students t-distribution with (n –1) degrees of freedom,

$$\therefore \quad K = t_{n-1}(\alpha/2)$$

Thus for a normal distribution where σ^2 is unknown, while testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, LRT converges to two tailed students t-test.

H_0 is rejected if

$$|t| = \left|\frac{\sqrt{n}(\bar{x} - \mu_0)}{S}\right| > t_{n-1}(\alpha/2)$$

and we accept H_0 if

$$|t| \leq t_{n-1}(\alpha/2)$$

Remark :

1) Consider, testing of $H_0 : \mu = \mu_0, 0 < \sigma^2 < \infty$ against $H_1 : \mu > \mu_0, 0 < \sigma^2 < \infty$

The parameter space is

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

and that w.r.t. H_0 is

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$$

For the parameter space Θ, the MLE (maximum likelihood estimate) of μ and σ^2 are

$$\hat{\mu} = \bar{x} \quad \text{if } \bar{x} \geq \mu_0$$

$$= \mu_0 \quad \text{if } \bar{x} < \mu_0$$

$$(\hat{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S^2 \quad \text{if } \bar{x} \geq \mu_0$$

$$= \frac{1}{n}\sum (x_i - \mu_0)^2 = S_0^2 \quad \text{if } \bar{x} < \mu_0$$

And for the space Θ_0 MLE is

$$\hat{\mu} = \mu_0$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2 = S_0^2$$

Then likelihood functions under Θ and Θ_0 are

$$L(\hat{\Theta_0}) = \left(\frac{1}{\sqrt{2\pi}\,S}\right)^n \exp\left(-\frac{1}{2S^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \begin{cases} \left(\frac{1}{\sqrt{2\pi S_0}}\right)^n e^{-n/2} & \text{if } \bar{x} < \mu_0 \\ \left(\frac{1}{\sqrt{2\pi S}}\right)^n e^{-n/2} & \text{if } \bar{x} \geq \mu_0 \end{cases}$$

and $$L(\hat{\Theta}_0) = \left(\frac{1}{\sqrt{2\pi S_0^2}}\right)^n \exp\left(-\frac{1}{2S_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi S_0}}\right)^n e^{-n/2}$$

Thus LRT statistic is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left(\frac{S}{S_0}\right)^n$$

$$= \left(\frac{S^2}{S_0^2}\right)^{n/2} \quad \text{if } \bar{x} \geq \mu_0$$

$$= 1 \quad \text{if } \bar{x} < \mu_0$$

Which is same as the LRT statistics derived in equation 6.7.5.

Therefore following the same procedure (as in the case of testing of $H_0 : \mu = \mu_0$ v/s $H_1 : \mu \neq \mu_0$), we get the critical region viz $0 < \lambda < \lambda_0$, equivalent to

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} \geq k \qquad \because \bar{x} \geq \mu_0$$

Here t is right tailed t–test and k is calculated as

$$P(t > k) = \alpha$$

$$k = t_{n-1}(\alpha)$$

Thus we reject $H_0 : \mu = \mu_0$ if

$$t \geq t_{n-1}(\alpha)$$

Otherwise H_0 is accepted

Similarly to test

$$H_0 : \mu = \mu_0, \sigma^2 > 0$$

against

$$H_1 : \mu < \mu_0, \sigma^2 > 0$$

The critical region is given by

$$t < -t_{n-1}(\alpha)$$

In this case we get left tailed t-test as

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} < -t_{n-1}(\alpha)$$

Thus we reject H_0 if $t < -t_{n-1}(\alpha)$ otherwise H_0 is accepted.

We summarize is in following table.

Hypothesis	Test type	Test statistic	Reject H_0 at l.o.s. α if
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Two tailed test	$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$	$\lvert t \rvert > t_{n-1}(\alpha/2)$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	Right tailed test	$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$	$t > t_{n-1}(\alpha)$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	Left tailed test	$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$	$t < -t_{n-1}(\alpha)$

Table 6.7.1 Normal population $N(\mu, \sigma^2)$ (σ^2 is unknown)

6.8 Test for Equality of Means of Two Normal Populations

Let X_1 and X_2 be two independent random variables from normal populations, having means μ_1 and μ_2 respectively and corresponding variances σ_1^2 and σ_2^2.

Let μ_1, μ_2, σ_1^2 and σ_2^2 all are unknown. Consider the problem of testing hypothesis $H_0 : \mu_1 = \mu_2 = \mu$; $\sigma_1^2, \sigma_2^2 > 0$ against an alternative hypothesis.

$$H_1 : \mu_1 \neq \mu_2 ; \sigma_1^2, \sigma_2^2 > 0$$

Case I : $\sigma_1^2 \neq \sigma_2^2$ (Population variances are not equal)

Parameter spaces w.r.t. entire region and H_0 are defined respectively as

$$\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, \sigma_i^2 > 0\} \quad i = 1, 2$$

$$\Theta_0 = \{(\mu, \sigma_i^2) : -\infty < \mu < \infty, \sigma_i^2 > 0\} \quad i = 1, 2$$

Let $(x_{11}, x_{12}, x_{13}, \ldots, x_{1m}$ and $x_{21}, x_{22}, x_{23}, \ldots, x_{2n}$ be the random samples of normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, with sizes m and n respectively.

Then likelihood function is given by

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i} - \mu_1)^2\right] \times \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{1/n}(x_{2j} - \mu_2)^2\right]$$

...(6.8.1)

Taking log on both sides we get

$$\log_e L = \frac{m}{2}\log\frac{1}{2\pi\sigma_1^2} + \frac{n}{2}\log\frac{1}{2\pi\sigma_2^2} - \frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i}-\mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{j=2}^{n}(x_{2j}-\mu_2)^2 \quad \ldots (6.8.2)$$

MLE for μ_i and σ_i^2 $(i = 1, 2)$ are given by

$$\left.\begin{aligned}
&\frac{\partial}{\partial\mu_1}\log L = 0 \Rightarrow \hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m} x_{1i} = \bar{x}_1\\
&\frac{\partial}{\partial\mu_2}\log L = 0 \Rightarrow \hat{\mu}_2 = \frac{1}{n}\sum_{j=1}^{n} x_{2j} = \bar{x}_2\\
&\frac{\partial}{\partial\sigma_1^2}\log L = 0 \Rightarrow \hat{\sigma_i^2} = \frac{1}{m}\sum_{j=1}^{m}(x_{1i}-\bar{x}_1)^2 = S_1^2 \text{ (say)}\\
&\frac{\partial}{\partial\sigma_2^2}\log L = 0 \Rightarrow \hat{\sigma_i^2} = \frac{1}{n}\sum_{j=1}^{n}(x_{2j}-\bar{x}_2)^2 = S_2^2 \text{ (say)}
\end{aligned}\right\} \quad \ldots (6.8.3)$$

Substituting these values from equation (6.8.3) in equation (6.8.1)

We get, $$L(\hat{\Theta}) = \left(\frac{1}{2\pi S_1^2}\right)^{m/2}\left(\frac{1}{2\pi S_2^2}\right)^{n/2} e^{-\left(\frac{m+n}{2}\right)} \quad \ldots (6.8.4)$$

Now, in the parameter space defined under H_0 (i.e. in Θ_0), we have $\mu_1 = \mu_2 = \mu$ and the likelihood function is given by,

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma_i^2}\right)m/2\left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2} e^{-\left[\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i-\mu)^2\right]} e^{-\left[1/2\sigma_2^2\sum_{j=1}^{n}(x_{2j}-\mu)^2\right]}$$

In (Θ), MLE of μ is obtained as the root of cubic equation

$$\frac{m^2(\bar{x}_1-\mu)}{\sum_{i=1}^{m}(x_{1i}-\hat{\mu})^2} + \frac{n^2(\bar{x}_2-\mu)}{\sum_{j=1}^{n}(x_{2j}-\hat{\mu})^2} \quad \ldots (6.8.5)$$

Which is a complicated function of sample observations. Apparently the LRT statistic and also become complex function of sample observations and its distribution become tedious. Therefore critical region for given l.o.s. α can ot be obtained.

Case II : Population variances σ_1^2 and σ_2^2 are equal i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say)

The parameter spaces are

$$\Theta = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu ; < \infty, \sigma^2 > 0\} \quad i = 1, 2$$

$$\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

then likelihood function when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is defined as

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{m+n}{2}} e^{\left[\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{m}(x_{1i}-\mu_1)^2 + \sum_{j=1}^{n}(x_{2j}-\mu_2)^2\right)\right]} \quad \ldots (6.8.6)$$

MLE of μ_1, μ_2 and σ^2 under Θ are

$$\left.\begin{aligned}
&\frac{\partial}{\partial\mu_1}\log L = 0 \Rightarrow \hat{\mu}_1 = \bar{x}_1\\
&\frac{\partial}{\partial\mu_2}\log L = 0 \Rightarrow \hat{\mu}_2 = \bar{x}_2\\
&\frac{\partial}{\partial\sigma^2}\log L = 0 \Rightarrow \hat{\sigma^2} = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_{1i}-\hat{\mu}_1)^2 + \sum_{j=1}^{n}(x_{2j}-\hat{\mu}_2)^2\right]\\
&\qquad = \frac{1}{m+n}[m s_1^2 + n s_2^2] = S^2 \text{ (say)}
\end{aligned}\right\} \quad \ldots (6.8.7)$$

In Θ_0, likelihood function becomes

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{m+n}{2}} e^{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{m}(x_{1i}-\mu)^2 + \sum_{j=1}^{m}(x_{2j}-\mu)^2\right]} \quad \ldots (6.8.8)$$

And MLE of μ and σ^2 are

$$\begin{aligned}
\frac{\partial \log L}{\partial\mu} = 0 &\Rightarrow 0 - \frac{1}{2\sigma^2}\left[-2\sum_{i=1}^{m}(x_{1i}-\mu) - 2\sum(x_{2j}-\mu)\right] = 0\\
&\Rightarrow \sum_{i=1}^{m} x_{1i} + \sum_{j=1}^{n} - m\mu - n\mu = 0\\
&\Rightarrow m\bar{x}_1 + n\bar{x}_2 = (m+n)\mu\\
&\Rightarrow \mu = \hat{\mu} = \frac{m\bar{x}_1 + n\bar{x}_2}{m+n} \quad \ldots (6.8.9)
\end{aligned}$$

Similarly, $$\frac{\partial \log L}{\partial\sigma^2} = 0$$

$$\Rightarrow -\frac{m+n}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{m}(x_{1i}-\mu)^2 + \sum_{j=1}^{n}(x_{2j}-\mu)^2\right] = 0$$

$$\sigma^2 = \hat{\sigma^2} = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_{1i}-\hat{\mu})^2 + \sum_{j=1}^{n}(x_{2j}-\hat{\mu})^2\right] \quad \ldots (6.8.10)$$

Consider

$$\sum_{i=1}^{m}(x_{1i}-\hat{\mu})^2 = \sum_{i=1}^{m}(x_{1i}-\bar{x}_1+\bar{x}_1-\hat{\mu})^2$$

$$= \sum_{i=1}^{m}(x_{1i}-\bar{x}_1)^2 + m(\bar{x}_1-\hat{\mu})^2$$

(See explanation in test for mean of normal population)

$$= m s_1^2 + m\left[\bar{x}_1 - \frac{m\bar{x}_1+n\bar{x}_2}{m+n}\right]^2$$

$$\left[\because \hat{\mu} = \frac{m\bar{x}_1+n\bar{x}_2}{m+n} \text{ and } ms_1^2 = \sum_{i=1}^{m}(x_{1i}-\bar{x}_1)^2\right]$$

$$= ms_1^2 + \frac{mn^2(\bar{x}_1-\bar{x}_2)^2}{(m+n)^2}$$

Similarly, $\sum_{j=1}^{n}(x_{2j}-\hat{\mu})^2 = s_2^2 + \frac{m^2n(\bar{x}_2-\bar{x}_1)^2}{(m+n)^2}$

Substituting these values in $\hat{\sigma^2}$ in equation (6.8.10) we get

$$\hat{\sigma^2} = \frac{1}{m+n}\left[ms_1^2+ns_2^2+\frac{mn}{m+n}(\bar{x}_1-\bar{x}_2)^2\right] \quad \ldots(6.8.11)$$

Thus from equations (6.8.8), (6.8.9) and (6.8.10) we get

$$L(\hat{\Theta}_0) = e^{-\frac{m+n}{2}} \cdot \left[\frac{m+n}{2\pi\left(ms_1^2+ms_2^2+\frac{mn}{m+n}(\bar{x}_1-\bar{x}_2)^2\right)}\right]^{\left[\frac{m+n}{2}\right]} \ldots(6.8.12)$$

From equation (6.8.6) and (6.8.7) we get,

$$L(\hat{\Theta}) = e^{-\frac{m+n}{2}} \cdot \left[\frac{m+n}{2\pi(ms_1^2+ns_2^2)}\right]^{\frac{m+n}{2}} \quad \ldots(6.8.13)$$

From equations (6.8.12) and (6.8.13)

LRT statistic is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta}_0)}$$

$$= \left[\frac{(ms_1^2+ns_2^2)}{(ms_1^2+ns_2^2)+\frac{mn}{m+n}(\bar{x}_1-\bar{x}_2)^2}\right]^{\frac{m+n}{2}}$$

$$= \left[\frac{1}{1+\left(\frac{mn(\bar{x}_1-\bar{x}_2)^2}{(m+n)(ms_1^2+ns_2^2)}\right)}\right]^{\frac{m+n}{2}} \quad \ldots(6.8.14)$$

For students t-distribution with $(m+n-2)$ degrees of freedom, the test statistic is given by

$$t = \frac{\bar{x}_1-\bar{x}_2}{s\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

Where $s^2 = \frac{ms_1^2+ns_2^2}{m+n-2}$

$$t = \frac{\sqrt{mn}(\bar{x}_1-\bar{x}_2)}{\sqrt{(m+n)(ms_1^2+ns_2^2)}} \times \sqrt{m+n-2}$$

$$\Rightarrow \quad \frac{t^2}{m+n-2} = \frac{mn(\bar{x}_1-\bar{x}_2)}{(m+n)(ms_1^2+ns_2^2)}$$

$\therefore$ equation 6.8.14 becomes

$$\therefore \quad \lambda = \left[1+\frac{t^2}{m+n-2}\right]^{-\left(\frac{m+n}{2}\right)} \quad \ldots(6.8.15)$$

The critical region of LRT is given by

$$0 < \lambda < C$$

$$\left[1+\frac{t^2}{m+n-2}\right]^{-\left(\frac{m+n}{2}\right)} < C$$

$$1+\frac{t^2}{m+n-2} > \lambda$$

$$\frac{t^2}{m+n-2} > \left[\frac{1}{\lambda^{2/m+n}}-1\right]$$

$$t^2 \geq (m+n-2)\left[\frac{1}{\lambda^{2/m+n}}-1\right]$$

$$t^2 \geq k^2 \text{ (say)}$$

$$\Rightarrow \quad |t| \geq k$$

where k is calculated as

$$P(|t| > k \mid H_0) = \alpha \quad \ldots(6.8.16)$$

Since, under H_0, the statistic t follows the student's t – distribution with degrees of freedom (m + n – 2), therefore from equation (6.8.1) we get,

$$k = t_{m+n-2}(\alpha/2) \qquad \ldots (6.8.17)$$

Thus LRT for testing $H_0 : \mu_1 = \mu_2 , \sigma_1^2 = \sigma_2^2 = \sigma^2$ v/s $H_1 : \mu_1 \neq \mu_2 , \sigma_1^2 = \sigma_2^2 = \sigma^2$ is equivalent to two tailed t-test where we reject H_0 if

$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \right| > t_{m+n-2}(\alpha/2)$$

Otherwise H_0 may be accepted.

Remarks :

1) LRT for testing $H_0 : \mu_1 = \mu_2 , \sigma_1^2 = \sigma_2^2 = \sigma^2$ against $H_1 : \mu_1 > \mu_2 , \sigma_1^2 = \sigma_2^2 = \sigma^2$ is same as right tailed t-test and its critical region is given by

$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \right| \geq t_{m+n-2}(\alpha)$$

2) To test $H_0 : \mu_1 = \mu_2, \sigma_1^2 = \sigma_2^2 = \sigma^2$ against $H_1 : \mu_1 < \mu_2, \sigma_1^2 = \sigma_2^2 = \sigma^2$; we get left tailed t-test whose critical region is given by

$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \right| \leq -t_{m+n-2}(\alpha)$$

3) Population having equal variances are known as "homoscedastic populations".

Hypothesis	Test type	Test statistic	Reject H_0 at l.o.s. α if
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	Two tailed test	$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$	$\lvert t \rvert > t_{m+n-2}(\alpha/2)$
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$	Right tailed test	$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$	$\lvert t \rvert > t_{m+n-2}(\alpha)$
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$	Left tailed test	$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$	$\lvert t \rvert < -t_{m+n-2}(\alpha)$

Table 6.8.1 Normal populations with means μ_1, μ_2 respectively.

Variances $\sigma_1^2 = \sigma_1^2 = \sigma^2$ (unknown).

6.9 Test of Equality of Means of Several Normal Populations

Consider there are k normal populations with means μ_i (i = 1, 2, ... k) and common variance σ^2. Which is unknown to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k = \mu$

against an alternate hypothesis $H_1 : \mu_2 \neq \mu_2 \neq \mu_3 \neq \ldots \neq \mu_k$

Let x_{ij} (i = 1, 2, ... k and j = 1, 2, ... n_i) be the j^{th} observation from i^{th} normal population having sample size n_i (i = 1, 2, ..., k) and $\sum_{i=1}^{k} n_i = n$.

In this case parameter spaces are defined a

$$\Theta = \{(\mu_i, \sigma^2) : -\infty < \mu_i < \infty, \sigma^2 > 0\} \quad i = 1, 2, \ldots k$$

and $$\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

The likelihood function w.r.t. Θ and Θ_0 are given respectively by

$$L(\Theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \mu_i)^2}$$

$$L(\Theta_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \mu)^2}$$

The MLE of μ_i (i = 1, 2, ..., k) and σ^2 under (Θ) are

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = \bar{x}_i \qquad i = 1, 2, \ldots, k$$

and $$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i^2) = \frac{S_w^2}{n}$$

where $$S_w^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Also, in Θ_0, MLE of μ and σ are

$$\hat{\mu}_i = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \bar{x}$$

and $$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = \frac{S_T^2}{n}$$

where $$S_T^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$$

Then $$L(\hat{\Theta}) = \left(\frac{n}{2\pi \cdot S_w^2}\right)^{n/2} e^{-n/2}$$

and $$L(\hat{\Theta}_0) = \left(\frac{n}{2\pi \cdot S_T^2}\right)^{n/2} e^{-n/2}$$

With these value, LRT statistic is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left(\frac{S_W^2}{S_T^2}\right)^{n/2} \quad \ldots (6.9.1)$$

Consider $$S_T^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\bar{x})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + \sum_{j=1}^{n_i} n_i(\bar{x}_i-\bar{x})^2$$

$$= S_W^2 + S_B^2 \quad \ldots (6.9.2)$$

where $$S_B^2 = \sum_{j=1}^{n_i} n_i(x_i-\bar{x})^2$$

From equation (6.9.2) and equation (6.9.1) becomes

$$\lambda = \left[\frac{1}{1+\left(\frac{S_B^2}{S_W^2}\right)}\right]^{n/2} \quad \ldots (6.9.3)$$

But under H_0

$$F = \frac{S_B^2/k-1}{S_W^2/n-k} \quad \ldots (6.9.4)$$

Follows F-distribution with degree of freedom $(k-1, n-k)$

$\therefore$ From equation (6.9.3) and (6.9.4) we can write

$$\lambda = \left[\frac{1}{1+\frac{k-1}{n-k}F}\right]^{n/2}$$

and the critical region is given by

$$\lambda \leq C$$

$$\lambda = \left[\frac{1}{1+\frac{k-1}{n-k}F}\right]^{n/2} \leq C$$

$$\Rightarrow \quad \lambda = \left[1+\frac{k-1}{n-k}F\right] \geq C^{-2/n}$$

$$\Rightarrow \quad F > \frac{n-k}{k-1}[C^{-2/n}-1]$$

$$F > A \text{ (Say)}$$

Where A is obtained as $P(F > A|H_0) = \alpha$

From F-table at 100 α % of l.o.s. and for $(k-1, n-k)$ degrees of freedom, we get

$$A = F_{k-1,\,n-k}(\alpha)$$

Thus, H_0 is rejected if

$$F = F_{k-1,\,n-k}(\alpha)$$

Otherwise it is accepted.

Note : To find MLE and ML function, follow the procedure as discussed in previous cases.

Remark :

In analysis of variance (ANOVA) the terminology.

i) S_W^2 is called Within Sample Sum of Squares (W.S.S.)

ii) S_T^2 is called Total Sum of Squares (T.S.S.)

iii) S_B^2 is called Between Samples Sum of Squares (B.S.S.)

iv) $S_B^2/k-1$ is called between samples Mean Sum of Squares (M.S.S.)

v) $S_W^2/n-k$ is called within samples M.S.S.

6.10 Test for Variance of Normal Population

Let $x_1, x_2, \ldots, x_n$ be random sample of size n, drawn from a normal population having mean μ and variance σ^2.

Assume that the variance σ^2 is known and is equal to σ_0^2 i.e. $\sigma^2 = \sigma_0^2$

To test $H_0 : \sigma^2 = \sigma_0^2$

against $H_1 : \sigma^2 = \sigma_0^2$

We have the parameter spaces.

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$$

and $$\Theta_0 = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 = \sigma_0^2\}$$

Likelihood functions in Θ and Θ_0 are viz

$$L(\Theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{(n/2)} e^{\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]}$$

and $$L(\Theta_0) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{(n/2)} e^{\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]}$$

MLE of μ and σ^2, in Θ are

$$\hat{\mu} = \bar{x}_i$$

and $$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = S^2$$ [Refer 6.7 for MLE]

and in Θ_0, $\sigma^2 = \sigma_0^2$ is known, therefore MLE of μ is,

$$\hat{\mu} = \bar{x}$$

Thus likelihood functions under Θ and Θ_0 reduces to

$$L(\hat{\Theta}) = \left(\frac{1}{2\pi S^2}\right)^{n/2} e^{-n/2}$$

and $$L(\hat{\Theta}_0) = \left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} e^{-ns^2/2\sigma_0^2}$$

Then LRT criterion become

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \left(\frac{S^2}{\sigma_0^2}\right)^{n/2} e^{\frac{-1}{2}\left(\frac{ns^2}{\sigma_0^2}-n\right)} \quad \ldots (6.10.1)$$

But under H_0 the statistic

$$\chi^2 = \frac{ns^2}{\sigma_0^2} \quad \ldots (6.10.2)$$

Follows chi-square distribution with (n–1) degrees of freedom

$\therefore$ From equations (6.10.1) and (6.10.2) we get

$$\lambda = \left(\frac{\chi^2}{n}\right)^{n/2} e^{\frac{-1}{2}(\chi^2-n)}$$

and the critical region is

$$\lambda < C$$

$$\left(\frac{\chi^2}{n}\right)^{n/2} e^{\frac{-1}{2}(\chi^2-n)} < C$$

$$(\chi^2)^{n/2} e^{(-\chi^2/2)+n/2} < C\, n^{n/2}$$

$$(\chi^2)^{n/2} e^{(-\chi^2/2)} < C\, n^{n/2} e^{-n/2}$$

$$< C\,(ne^{-1})^{n/2}$$

$$< B \text{ (say)}$$

As the statistic χ^2 follows the chi-square distribution with $(n-1)$ degrees of freedom. The critical region is obtained by pair of intervals $0 < \chi^2 < \chi_2^2$ and $\chi_1^2 < \chi^2 < \infty$, where χ_1^2 and χ_2^2 are determined such that

$$(\chi_1^2)^{n/2} e^{\frac{-1}{2}\chi_1^2} = (\chi_2^2)^{n/2} e^{\frac{-1}{2}\chi_2^2}$$

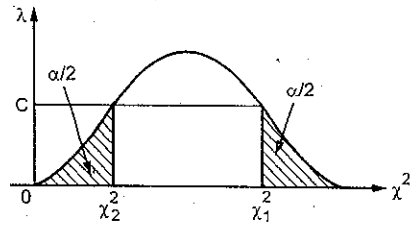


Fig. 6.10.1

i.e. X_1^2 and X_2^2 are defined as

$$P(\chi^2 > \chi_1^2) = \alpha/2$$

$$P(\chi^2 > \chi_2^2) = 1 - \frac{\alpha}{2}$$

in other words $$\chi_1^2 = \chi_{n-1}^2(\alpha/2)$$

$$\chi_2^2 = \chi_{n-2}^2\left(1-\frac{\alpha}{2}\right) \quad \ldots (6.10.3)$$

Where $\chi_n^2(\alpha)$ is upper α -point of the chi-square distribution with n degrees of freedom. Thus required critical region is two tailed region defined as

$$\chi^2 > \chi_{n-1}^2(\alpha/2)$$

$$\chi^2 < \chi_{n-2}^2\left(1-\frac{\alpha}{2}\right)$$

We reject H_0 if $\chi^2 > \chi_{n-1}^2(\alpha/2)$ or $\chi^2 < \chi_{n-2}^2\left(1-\frac{\alpha}{2}\right)$ otherwise we may accept it.

Test $\quad \lambda \quad \sigma_0^2$

Table 6.10.1 : Normal population N (μ, σ^2) where μ is unknown and σ^2 is known.

6.11 Test for Equality of Variances of Two Normal Populations

Let N (μ_1,σ_1^2) and N (μ_2,σ_2^2) be two normal populations where μ_1, μ_2, σ_1^2 and σ_2^2 are all unknown.

x_{1i}(i = 1, 2, ..., m) and x_{2j}(i = 1, 2, ..., n) be independent random samples from normal populations of sizes m and n respectively.

Suppose we want to test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$$

V/s $\qquad H_1 : \sigma_1^2 \neq \sigma_2^2$

the parameter spaces defined as

$$\Theta = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_1, \mu_2 < \infty ; \ \sigma_1^2, \sigma_2^2 < 0\}$$

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) : -\infty < \mu_1, \mu_2 < \infty ; \sigma^2 > 0\}$$

and the likelihood function is given by

$$L = \left(\frac{1}{2\pi\sigma_1^2}\right)^{m/2} \left(\frac{1}{2\pi\sigma_2^2}\right)^{n/2} e^{\frac{-1}{2\sigma_1^2}\sum_{i=1}^{m}(x_{1i}-\mu_1)^2} \cdot e^{\frac{-1}{2\sigma_2^2}\sum_{j=1}^{m}(x_{2j}-\mu_2)^2} \quad \ldots (6.11.1)$$

Calculating MLE of μ_k and σ_k^2 (k = 1, 2) (as discussed in previous topics) in Θ and Θ_0 and substituting them in equation (6.11.1) we get,

$$L(\hat{\Theta}) = \left(\frac{1}{2\pi S_1^2}\right)^{m/2} \left(\frac{1}{2\pi S_2^2}\right)^{n/2} e^{-(m+n)/2} \quad \ldots (6.11.2)$$

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{(m+n)^{\frac{m+n}{2}}}{m^{m/2} n^{n/2}} \left[\frac{(mS_1^2)^{m/2} (nS_2^2)^{n/2}}{[mS_1^2 + nS_2^2]^{\frac{m+n}{2}}}\right] \quad \ldots (6.11.4)$$

Under H_0, the statistic

$$F = \frac{\sum_{i=1}^{m}(x_{1i}-\bar{x}_1)^2/m-1}{\sum_{j=1}^{n}(x_{2j}-\bar{x}_2)^2/n-1} = \frac{S_1^2}{S_2^2} \quad \ldots (6.11.5)$$

Follows F distribution with (m – 1, n – 1) d.f. substituting equation (6.11.5) in equation (6.11.4) we get,

$$\lambda = \frac{(m+n)^{\frac{m+n}{2}}}{m^{m/2} n^{n/2}} \left\{\frac{\left(\frac{m-1}{n-1}F\right)^{m/2}}{\left[1+\frac{m-1}{n-1}F\right]^{(m+n)/2}}\right\} \quad \ldots (6.11.6)$$

But λ defined in (6.11.6) is a monotonic function of F. Therefore test can be carried out following F-distribution with test statistic as given in equation (6.11.5).

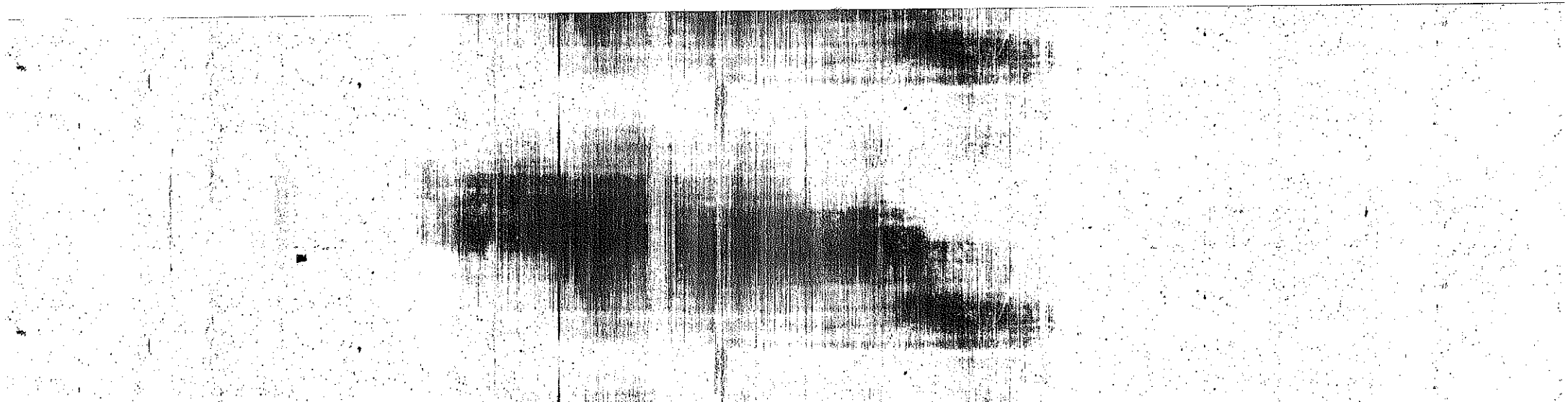
The critical region is given by a pair of intervals as

$$F \leq F_1 \text{ and } F \geq F_2$$

where F_1 and F_2 are obtained such that

$$P(F \geq F_1) = 1 - \frac{\alpha}{2}$$

$$P(F \geq F_2) = \frac{\alpha}{2}$$



Since under H_0, LRT is equivalent to F-test, thus from F-table,

$$F_1 = F_{(m-1, n-1)}\left(1-\frac{\alpha}{2}\right)$$

$$F_2 = F_{(m-1, n-1)}\left(\frac{\alpha}{2}\right)$$

where $F_{m,n}(\alpha)$ is upper α - point of F-distribution with (m, n) degrees of freedom.

Thus we get two tailed F-test with critical region.

$$F > F_{(m-1, n-1)}\left(\frac{\alpha}{2}\right)$$

and

$$F < F_{(m-1, n-1)}\left(1-\frac{\alpha}{2}\right)$$

Example 6.11.1 *Let X be a random sample having normal distribution $N(\mu, \sigma^2)$ where μ is unknown and $\sigma^2 = 2$. Derive LRT for H_0 : $\mu = 10$ against H_1 : $\mu \neq 10$ at 5% l.o.s.*

Solution : Given :

$$\Theta_0 : \mu = 10$$

$$\Theta_1 : \mu \neq 10$$

As variance σ^2 is known, therefore μ is the only parameter of distribution.

$\therefore$ We define parameter spaces as

$$\Theta = \{\mu \mid -\infty < \mu < \infty\}$$

$$\Theta_0 = \{\mu \mid \mu = 10\}$$

We know that p.d.f. for $N(\mu, \sigma^2)$ is

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\therefore$ Likelihood function becomes

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$L = \left(\frac{1}{4\pi}\right)^{n/2} e^{\frac{-1}{4}\sum_{i=1}^{n}(x_i-\mu)^2} \qquad (\because \sigma^2 = 2)$$

Replacing μ by its MLE, likelihood function under Θ and Θ_0 is

$$L(\hat{\Theta}) = \left(\frac{1}{4\pi}\right)^{n/2} e^{\frac{-1}{4}\Sigma(x_i-\bar{x})^2}$$

$$L(\hat{\Theta}_0) = \left(\frac{1}{4\pi}\right)^{n/2} e^{\frac{-1}{4}\Sigma(x_i-10)^2}$$

$$= \frac{1}{(4\pi)^{n/2}} e^{\frac{-1}{4}\left[\sum(x_i-\bar{x})^2 + n(\bar{x}-10)^2\right]}$$

then LRT statistic is given by

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = e^{-\frac{n}{4}(\bar{x}-10)^2}$$

$\therefore$ The critical region $\lambda \leq c$ becomes

$$e^{-\frac{n}{4}(\bar{x}-10)^2}$$

$$\frac{-n}{4}(\bar{x}-10)^2 \leq \ln C$$

$$(\bar{x}-10)^2 \geq \frac{-4}{n}\ln C$$

$$\Rightarrow \quad |\bar{x}-10^2| \geq \sqrt{\frac{-4}{n}\ln c}$$

Dividing both sides by $\sigma/\sqrt{n}$ we get

$$\frac{|(\bar{x}-10)|}{\sigma/\sqrt{n}} \geq \frac{\sqrt{\frac{-4}{n}\ln c}}{\sigma/\sqrt{n}} = k \text{ (say)}$$

But from SND ; M (0, 1),

$$Z = \frac{|(\bar{x}-\mu)|}{\sigma/\sqrt{n}} = \frac{|(\bar{x}-10)|}{\sigma/\sqrt{n}}$$

$\therefore$ We get $Z \geq k$

Where k is derived s.t. size of critical region is $\alpha = 5\%$ i.e.

$$P(Z \geq k / \Theta_0) = \alpha$$

$$\Rightarrow \quad k = Z_{\alpha/2} = 1.96$$

Thus we reject Θ_0 if $Z \geq 1.96$.

Example 6.11.2 *Consider a random samples $(x_1, x_2, \ldots, x_n)$ from an exponential density function defined as $f(x, \theta) = \theta e^{-\theta x}$, $x > 0$. Find likelihood ratio test for H_0 : $\theta \leq \theta_0$ against H_1 : $\theta > \theta_0$.*

Solution : Given

$$f(x, \theta) = \theta e^{-\theta x} \qquad \ldots (1)$$

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta \leq \theta_0$$

The parameter spaces are

$$\Theta = \{\theta \mid \theta > 0\}$$

$$\Theta_0 = \{\theta \mid \theta \leq \theta_0\}$$

Using equation (1), likelihood function can be written as

$$L = \prod_{i=1}^{\infty} \theta e^{-\theta x_i} = \theta^n e^{-\theta \Sigma x_i}$$

MLE of θ can be obtained by solving the equation

$$\frac{\partial L}{\partial \theta} = 0 \Rightarrow n\,\theta^{n-1} e^{-\theta \Sigma x_i} - \theta^n \Sigma x_i e^{-\theta \Sigma x_i} = 0$$

$$\Rightarrow \quad n\,\theta^{-1} = \Sigma x_i$$

$$\Rightarrow \quad \theta = \frac{n}{\Sigma x_i} = \hat{\theta}$$

With this MLE $\hat{\theta}$, the likelihood function under Θ and Θ_0 become

$$L(\hat{\Theta}) = \prod_{i=1}^{n} \left[\frac{n}{\Sigma x_i}\right] e^{-n}$$

$$= \left[\frac{n}{\Sigma x_i}\right]^n e^{-n}$$

and

$$L(\hat{\Theta}_0) = \begin{cases} \left(\frac{n}{\Sigma x_i}\right)^n e^{-n} & \text{if } \frac{n}{\Sigma x_i} \leq \theta_0 \\ \theta_0^n e^{-\theta_0 \Sigma x_i} & \text{if } \frac{n}{\Sigma x_i} > \theta_0 \end{cases}$$

Then LRT statistic becomes

$$\lambda = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})}$$

$$= \begin{cases} 1 & \text{if } \frac{n}{\Sigma x_i} \leq \theta_0 \\ \dfrac{\theta_0^n e^{-\theta_0 \Sigma x_i}}{\left(\frac{n}{\Sigma x_i}\right)^n e^{-n}} & \text{if } \frac{n}{\Sigma x_i} > \theta_0 \end{cases}$$

$$= \begin{cases} 1 & \text{if } \bar{x} \geq \frac{1}{\theta_0} \\ (\theta_0 \bar{x})^n e^{-(\theta_0 \Sigma x_i - n)} & \text{if } \bar{x} < \frac{1}{\theta_0} \end{cases}$$

$$= \begin{cases} 1 & \text{if } \theta_0 \bar{x} \geq 1 \\ (\theta_0 \bar{x})^n e^{-n(\theta_0 \bar{x} - 1)} & \text{if } \bar{x} | \theta_0 < 1 \end{cases}$$

Let $\theta_0 \bar{x} = t$

then $(\theta_0 \bar{x})^n e^{-n(\theta_0 \bar{x} - 1)} = t^n e^{-n(t-1)}$

Note that $t^n e^{-n(t-1)}$ has maximum value "1" at $t = 1$ i.e. when $\bar{x} = \frac{1}{\theta_0}$

$\therefore$ Clearly $\lambda \leq 1$

Assume that $0 < C < 1$,

Then we reject H_0 if $\lambda \leq C$

i.e. if $t^n e^{-n(t-1)} \leq C$

or if $\theta_0 \bar{x} = t < 1$

Thus for $0 < C < 1$;

$t < 1$ and $t^n e^{-n(t-1)} < C$,

Iff $t \leq k$, where k is a constant such that $0 < k < 1$.

If *l*.o.s. of critical region is α, then k is defined as

$$P(t \leq k) = \alpha$$

$$\Rightarrow \quad P(\theta_0 \bar{x} \leq k) = \alpha$$

$$\Rightarrow \quad P\left(\theta_0 \frac{\Sigma x_i}{n} \leq k\right) = \alpha$$

$$\Rightarrow \quad P(\theta_0 \Sigma x_i \leq nk) = \alpha$$

$$\Rightarrow \quad \int_0^{nk} \left(\frac{1}{\lfloor n}\right) e^{-y} y^{n-1}\, dy = \alpha$$

6.12 Non-Parametric Methods

All hypothesis testing methods, that we have studied so far, are based on an important assumption that the data, drawn from a sample is normally distributed. These methods are concerned about either estimating or testing hypothesis about population parameters. Thus the

tests which deals with drawing the conclusion about parameters of population are called "Parametric tests".

But in practice, many situations arises where the population under consideration does not follow any specific distribution. To deal with such problems, statisticians have derived several methods and tests, in which no assumptions about distributions as well as about parameters of population are made. These methods are known as "Non-parametric" methods.

As these methods are independent of population distribution, they are also known as "Distribution free" methods.

6.12.1 Advantages of Non-parametric Tests (N.P.)

N.P. methods have many advantages over the parametric methods.

1) N.P. methods are easy and simple to understand, easy and simple to implement when sample size is small.
2) N.P. methods do not require complicated computation, therefore these methods are less time consuming.
3) In N.P. methods no assumption, regarding distribution of parent population, from which sample is drawn is made.
4) They need only nominal or ordinal data.
5) N.P. methods not only make less rigorous assumptions but also work well even certain assumption are violated.

6.12.2 Disadvantages of Non-parametric Tests

1) Due to simplicity of N.P. tests, they are often used even if there exist an appropriate parametric method.

   In such cases (where parametric test is available and can be more powerful), applying N.P. is just waste of time and data.
2) All N. P. tests are not as simple as they are claimed to be.
3) It is not possible to determine the actual of a N.P. test.

Exercise

1. *Define i) Parameter space ii) Likelihood function iii) Likelihood Ratio test statistic*
2. *Consider a random sample $x_1, x_2, \ldots x_n$ of size n, from a normal distribution $N(\mu, \sigma^2)$. Develop LRT to test $H_0 : \mu = \mu_0$ (specified) against.*

   i) $H_1 : \mu > \mu_0$ ii) $H_1 : \mu < \mu_0$ iii) $H_1 : \mu \neq \mu_0$

   [Given σ^2 is known]

3. *Let $x_1, x_2, \ldots, x_n$ be a random sample of size n, from a normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 both are unknown. Show that LRT, used to test $H_0 : \mu = \mu_0$ v/s $H_1 : \mu \neq \mu_0$, $0 < \sigma^2 < \infty$, is two tailed t-test.*
4. *Prove that the LRT for testing the equality of variance of two normal populations is same as two tailed F-test.*
5. *Derive the LRT to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, based on a random sample of size n, from a Poisson distribution.*
6. *What is composite and simple hypothesis. Give examples.*
7. *Write short notes on -*

   1) Most powerful test 2) Uniformly most powerful test

   3) The best critical region 4) Level of significance.
8. *What is the importance of type I and type II errors in Neyman and Pearson, theory of testing.*
9. *State and prove Neyman pearson lemma for testing a simple hypothesis against a simple alternative hypothesis.*
10. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a p.d.f.*

$$f(x,\theta) = \begin{cases} \theta x^{\theta-1} & 0 \le x \le 1 \quad \theta > 0 \\ \theta & \text{otherwise} \end{cases}$$

Find on U.M.P. test of size α for testing $H_0 : \theta = 1$ against $H_1 : \theta > 1$.

Multiple Choice Question

Q.1 The set of all possible values of a parameter of some p.d.f. is called _____.

[a] null space [b] parameter space

[c] critical region [d] none of these

Q.2 The ratio of maximum likelihood function defined over Θ_0 (null parameter space) to that defined over Θ (complete parameter space) is called _____.

[a] likelihood ratio statistic [b] p-value

[c] maximum likelihood estimate [d] t-test statistic

Q.3 Likelihood ratio test statistic λ = _____.

[a] $\sup_{\theta \in \Theta} L(x,\theta)$ [b] $\sup_{\theta \in \Theta_0} L(x,\theta)$

[c] $\dfrac{\sup_{\theta \in \Theta_0} L(x,\theta)}{\sup_{\theta \in \Theta} L(x,\theta)}$ [d] $\dfrac{\sup_{\theta \in \Theta} L(x,\theta)}{\sup_{\theta \in \Theta_0} L(x,\theta)}$

Q.4 According to L.R.T., H_0 is rejected if $0 < \lambda \leq \lambda_0$ where, λ_0 is derived from _____.

[a] $P(\lambda < \lambda_0 | H_1) = \alpha$ [b] $P(\lambda < \lambda_0 | H_0) = \alpha$

[c] $P(\lambda > \lambda_0 | H_1) = \alpha$ [d] $P(\lambda > \lambda_0 | H_0) = \alpha$

Q.5 If $x_1, x_2, \ldots, x_n$ be a random sample from a population with p.d.f. $f(x, \theta_1, \theta_2, \ldots, \theta_k)$ then likelihood function is given by _____.

[a] $L = \sum_{i=1}^{n} f(x_i, \theta_1, \theta_2, \ldots, \theta_k)$ [b] $L = \prod_{i=1}^{n} f(x_i, \theta_1, \theta_2, \ldots, \theta_k)$

[c] $L = \dfrac{\partial f}{\partial x_i}$ [d] None of these

Q.6 The LRT for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, based on random sample of size n from a normal population with unknown μ and σ^2 is equivalent to _____.

[a] two tailed F-test [b] one tailed F-test

[c] one tailed t-test [d] two tailed t-test

Q.7 A uniformly most powerful test among the class of unbiased test is termed as

[a] Optimum biased test [b] Uniformly most-powerful unbiased test

[c] Minimax unbiased test [d] Minimax test

Q.8 The Neyman-Pearson lemma provides the best critical region for testing _____ null hypothesis against _____ alternative hypothesis.

[a] composite, simple [b] composite, composite

[c] simple, composite [d] simple, simple

Q.9 Suppose t test is applied to test the equality of means of two random variables X and Y, where X and Y are independent. Then :

[a] Distribution of both X and Y must be normal with equal variances.

[b] Distribution of both X and Y must be normal.

[c] Distribution of X and Y may not be normal.

[d] Both X and Y must have the same variance.

Answer Keys for Multiple Choice Questions :

Q.1	b	**Q.2**	a	**Q.3**	c	**Q.4**	b	**Q.5**	b
Q.6	d	**Q.7**	b	**Q.8**	d	**Q.9**	b		

❑❑❑

Tum to heavy driver ho ....

Best of luck.....

Notes

SOLVED MODEL QUESTION PAPER (In Sem)

Statistics

S.E. (AI and DS) Semester - IV (As Per 2020 Pattern)

Time : 1 Hour] **[Maximum Marks : 30**

N. B. :

i) Attempt Q.1 or Q.2, Q.3 or Q.4.

ii) Neat diagrams must be drawn wherever necessary.

ii) Figures to the right side indicate full marks.

iv) Assume suitable data, if necessary.

Q.1 **a)** *What is statistics ? Give importance and limitations of statistics.* **(Refer sections 1.1, 1.3 and 1.4)** [5]

b) *Define sampling. Explain random sampling.* **(Refer sections 1.7 and 1.7.2)** [4]

c) *What is population and sample ? Give difference between them.* **(Refer section 1.7)** [4]

OR

Q.2 **a)** *What are the methods of estimation ? Give brief on testing of hypothesis.* **(Refer sections 1.10.1 and 1.10.2)** [5]

b) *Give advantages and disadvantages of statistical analysis.* **(Refer sections 1.3 and 1.4)** [4]

c) *Explain the scope of statistics in engineering and technology.* **(Refer section 1.2.2)** [4]

Q.3 **a)** *What is histogram ? Draw the histogram for the following data -* [5]

Age (in years)	2 - 5	5 - 11	11 - 12	12 - 14	14 - 15	15 - 16
No. of boys	6	6	2	5	1	3

Ans. :

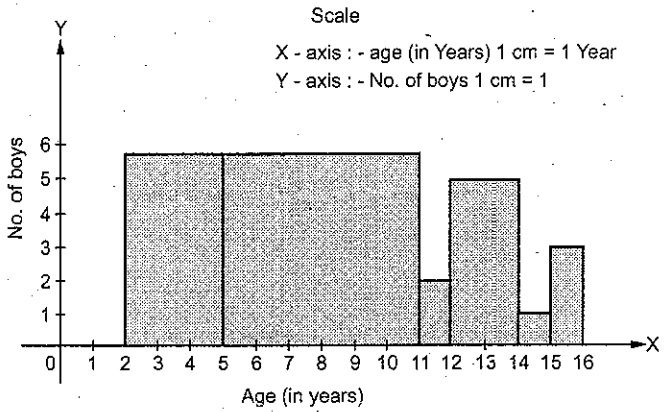


Fig. 1

b) *State merits and demerits of arithmetic mean (two each).* **(Refer section 2.10)** [4]

c) *Obtain the median from following table.* [4]

Class	*0 - 100*	*100 - 200*	*200 - 300*	*300 - 400*	*400 - 500*	*500 - 600*	*600 - 700*
Frequency	*9*	*15*	*18*	*21*	*18*	*14*	*5*

Ans. :

C.I.	f	Less than c.f.
0-100	9	9
100-200	15	24
200-300	18	42
300-400	21	63
400-500	18	81
500-600	14	95
600-700	5	100

N = 100

$\frac{N}{2} = 50$

Median class

l = 300, h = 100, f = 21, c.f. = 42

$$\text{Median} = l + \left(\frac{\frac{N}{2} - \text{c.f.}}{f}\right) \times h$$

$$= 300 + \left(\frac{50 - 42}{21}\right) \times 100 = 300 + 0.3809 \times 100$$

$$= 338.0952$$

d) *Define Geometric mean and harmonic mean. Compare them on the basis of merits.* **(Refer sections 2.17 and 2.19)** [4]

OR

Q.4 a) *Daily expenditure of 100 families on transport is given below :* [5]

Expenditure	*20 - 29*	*30 - 39*	*40 - 49*	*50 - 59*	*60 - 69*
No. of families	*14*	*?*	*27*	*?*	*15*

If mode of the distribution is 43.5 find the missing frequencies.

Ans. :

c.f.	f
20 - 29.5	14
29.5 - 39.5	f_2
39.5 - 49.5	27
49.5 - 59.5	f_3
59.5 - 69.5	15

Mode = 43.5

Here maximum frequency is 27

Thus class 39.5 - 49.5 is modal class.

where $l = 39.5, f_1 = 27, f_0 = f_2, f_2 = f_3, h = 10$

$$\text{Mode} = 39.5 + 10\left(\frac{27 - f_2}{54 - f_2 - f_3}\right)$$

$$43.5 = 39.5 + \frac{270 - 10f_2}{54 - f_2 - f_3}$$

$$43.5 = \frac{2133 - 39.5f_2 - 39.5f_3 + 270 - 10f_2}{54 - f_2 - f_3}$$

$$2349 - 43.5f_3 - 43.5f_2 = 2403 - 49.5f_2 - 39.5f_3$$

$$49.5f_2 + 39.5f_3 - 43.5f_3 - 43.5f_2 = 54$$

$$6f_2 - 4f_3 = 54 \quad \ldots (1)$$

As $N = \sum f = 100$

$$14 + f_2 + 27 + f_3 + 15 = 100$$

$$f_2 + f_3 = 100 - 56$$

$$f_2 + f_3 = 44 \quad \ldots (2)$$

Solving equation (1) and (2),

$\boxed{f_2 = 23}$ $\boxed{f_1 = 21}$

b) *Give merits and demerits of median.* **(Refer section 2.13)** **[4]**

c) *Calculate harmonic mean of the following series* **[4]**

Values	2	6	10	14	18
Frequency	4	12	20	9	5

Ans. :

w_i	x_i
2	4
6	12
10	20
14	9
18	5

$$\sum w_i = 50$$

$$\sum \frac{w_i}{x_i} = \frac{2}{4} + \frac{6}{12} + \frac{10}{20} + \frac{14}{9} + \frac{18}{5}$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1.5556 + 3.6$$

$$= 1.5 + 5.1556 = 6.6556$$

$$H = \frac{\sum w_i}{\sum\left(\frac{w_i}{x_i}\right)} = \frac{50}{6.6556} = 7.5124$$

d) *State the advantages and limitation of graphical representation of data.* **(Refer section 2.5)** **[4]**

SOLVED MODEL QUESTION PAPER (End Sem)

Statistics

S.E. (AI and DS) Semester - IV (As Per 2020 Pattern)

Time : 2 $\frac{1}{2}$ Hours] **[Maximum Marks : 70**

N. B. :

- *i) Attempt Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.*
- *ii) Neat diagrams must be drawn wherever necessary.*
- *ii) Figures to the right side indicate full marks.*
- *iv) Assume suitable data, if necessary.*

Q.1 **a)** *For the regression lines $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$ find i) $\bar{x}$ and $\bar{y}$ ii) Correlation coefficient r between x and y.* **(Refer example 3.18.1)** **[6]**

b) *Compute M.D. about i) Mean ii) Median also find coefficient of M.D. for the following frequency distribution.* **(Refer example 3.6.2)** **[5]**

Class internal	1-3	3-5	5-7	7-9
Frequency	3	4	2	1

c) *MNC company conducted 1000 candidates aptitude test. The average score is 45 and the standard deviation of score is 25. Assuming normal distribution for the result. Find i) The number of candidates whose scores exceed 60.*
ii) The number of candidates whose scores lies between 30 and 60.
(Refer example 3.31.11) **[6]**

OR

Q.2 a) *The following marks have been obtained by a class of students in 2 papers of mathematics* **[5]**

Paper I	45	55	56	58	60	65	68	70	75	80	85
Paper II	56	50	48	60	62	64	65	70	74	82	90

Calculate the coefficient of correlation for the above data. **(Refer example 3.15.3)**

b) *Find the first four moments about any arbitrary point for the following data* **(Refer example 3.10.2)** **[6]**

Marks	29	30	31	32	33	34	35	36	37	38	39
No. of students	2	1	4	5	10	75	50	74	62	40	41

c) *Out of 2000 families with 4 children each, how many would you expect to have i) At least a boy ii) Two boys iii) 1 or 2 girls iv) No girls.* **(Refer example 3.31.5)** **[6]**

Q.3 a) *An unbiased coin is thrown 10 times. Find the probability that (i) Getting exactly 6 heads (ii) Getting at least 6 heads.* **(Refer example 4.14.3)** **[6]**

b) *The probability distribution of X is as follows :* **[6]**

X	0	1	2	3	4
P(X = x)	0.1	k	2k	2k	k

is probability mass function.
Find i) k ii) P [x < 2] iii) P[X ≥ 3] iv) P [1 ≤ x ≤ 3]
(Refer example 4.5.1)

c) *In a continuous distribution density function f(x) = k x (2 – x), 0 < x < 2. Find the value of k, mean and c variance.* **(Refer example 4.5.5)** **[6]**

OR

Q.4 a) *Fit a Poisson distribution to the following data and calculate theoretical frequencies.* **(Refer example 4.17.4)** **[6]**

x :	0	1	2	3	4	Total
f :	109	65	22	3	1	200

b) *Let a random variable x takes values –2, –1, 0, 1, 2 such that* $P(X = -2) = P(X = -1) = P(X = 1) = P(X = 2)$ *and* $P(X < 0) = P(X = 0) = P(X > 0)$. *Determine the probability mass function of X.* **(Refer example 4.5.3)** **[6]**

c) *The probability density function* $f(x) = \lambda x e^{-x}$, $x > 0$.
Find : i) The value of λ, *ii)* $P(2 < x < 5)$ **(Refer example 4.5.6)** **[6]**

Q.5 a) *Let P be the probability that a coin will fall head in a single toss in order to test* $H_0 : P = \frac{1}{2}$ *against* $P = \frac{3}{4}$. *The coin is tossed 5 times and* H_0 *is rejected if more than 3 heads are obtained. Find the probability of type I error and power of the test.* **(Refer example 5.8.2)** **[6]**

b) *A normal population has mean 6.8 and standard deviation of 1.5. A sample of 400 members gave a mean of 6.75. Is the difference significant ?* **(Refer example 5.10.8)** **[5]**

c) *Suppose that sweets are sold in packages of fixed weight of the contents. The procedure of the packages is interested in testing the average weight of content in packages in 1 kg. Hence a random sample of 12 packages is drawn and their contents found (in kg) are as follows : 1.05, 1.01, 1.04, 0.98, 0.96, 1.01, 0.97, 0.99, 0.98, 0.95, 0.97, 0.95. Using the above data what should he conclude about the average.* **(Refer example 5.10.17)** **[6]**

OR

Q.6 a) *In an experiment on pea breeding, the following frequencies of seeds were obtained.* **[6]**

Round and green	*Wrinkled and green*	*Round and yellow*	*Wrinkled and yellow*	*Total*
222	120	32	150	524

Theory predicts that the frequencies should be in proportion 8:2:2:1. Examine the correspondence between theory and experiment. **(Refer example 5.11.3)**

b) *From the data given below. Intelligence tests of two groups of boys and girls gave the following results. Examine the difference is significant.* **(Refer example 5.10.12)** **[5]**

	Mean	S.D.	Size
Girls	70	10	70
Boys	75	11	110

c) *In two independent samples of size 8 and 10 the sum of squares deviations of the sample values from the respective sample means were 84.4 and 102.6. Test whether the difference of variances of the population is significant or not.* **(Refer example 5.10.22)** **[6]**

Q.7 a) *State and Prove N-P-Lemma.* **(Refer section 6.4)** **[8]**

b) *Check whether a BCR exists for testing the null hypothesis* $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta = \theta_1 > \theta_0$ *for the parameter* θ *of the distribution.* $f(x, \theta) = \frac{1+\theta}{(x+\theta)^2}$; $1 \le x \le \infty$. **(Refer example 6.4.7)** **[5]**

c) *Let X be a random sample having normal distribution* $N(\mu, \sigma^2)$ *where* μ *is unkown and* $\sigma^2 = 2$. *Derive likelihood ratio test for* $H_0 : \mu = 10$ *against* $H_1 : \mu = 10$ *at 5% level of significance.*

(Given that, maximum likelihood function under Θ *and* Θ_0.

$$L(\hat{\Theta}) = \left(\frac{1}{4\pi}\right)^{n\,2} e^{-1\,\mu \sum (x_i - \overline{X})^2}$$

$$L(\hat{\Theta})_0 = \left(\frac{1}{4\pi}\right)^{n\,2} e^{-1\,4 \sum [(x_i - \overline{X})^2 + n(\overline{x} - 10)^2]}$$

[5]

Ans. : Θ_0 - rejected if $Z \ge 1.96$

OR

Q.8 a) *If* $x \ge 1$ *is the critical region for testing* $\Theta_0 : \theta = 2$ *against the alternative* $\theta = 1$ *on the basis of the single observation from the population.* $f(x, \theta) = \theta e^{-\theta x}$ $0 \le x < \infty$, *obtain the values of type I and type II error.* **(Refer example 6.4.4)** **[5]**

b) *Show that the likelihood ratio test for testing the equally of variances of two normal distribution is the usual f-test.* **(Refer section 6.11)** **[8]**

c) *State advantages and disadvantages of Non-parametric Tests.* **(Refer sections 6.12.1 and 6.12.2)** **[5]**

❑❑❑

Notes