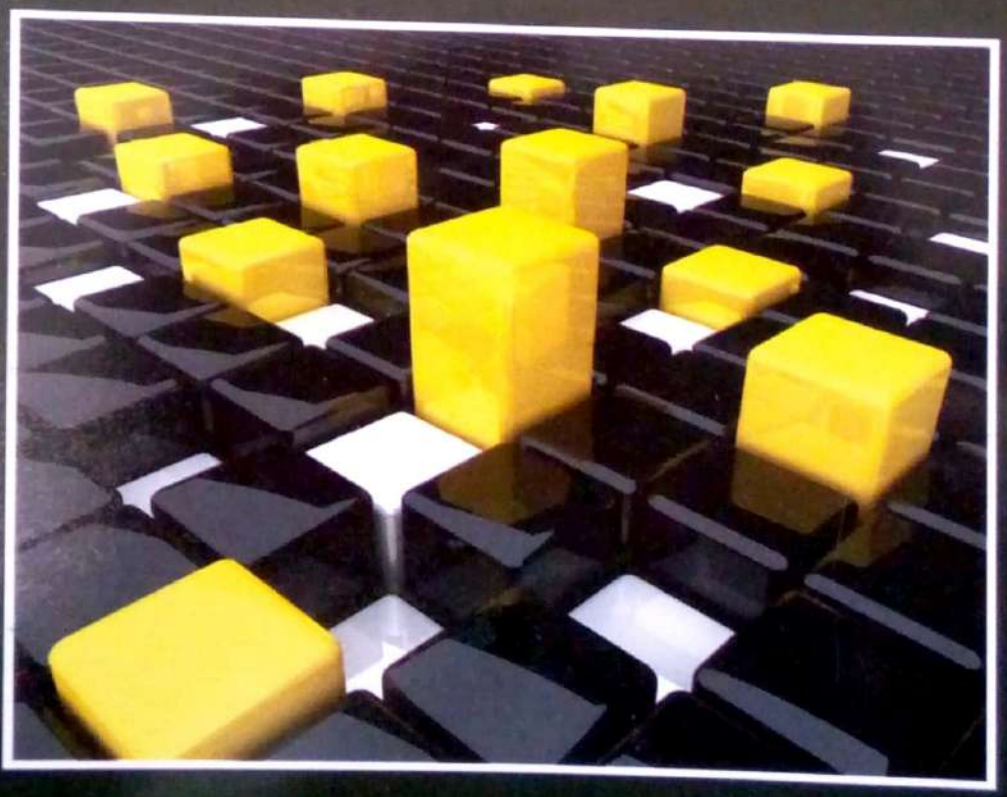
New Edition

STATISTICS-I

B.Sc. CSIT Second Semester



- Prof. Vikash Raj Satyal, Ph.D
- Prof. Ram P. Khatiwada, Ph.D
- Bijay Lal Pradhan, Ph.D
- Nabaraj Poudyal
- Prakash Bahadur Amatya

Scanned by CamScanner

Table of Contents

Chapter 1

INTRODUCTION

Basic Concept of Statistics ---- 1
Definitions ---- 2
Descriptive Statistics ---- 4
Inferential Statistics ---- 4
Limitations of Statistics ---- 7
Application of Statistics in the Field of CSIT ---- 8
Scales of Measurement ---- 10
Nominal Scale ---- 10
Ordinal Scale ---- 10
Interval Scale ---- 10
Ratio Scale ---- 11
Variable ---- 11
Qualitative Variable ---- 11
Quantitative Variable ---- 12
Discrete Variable ---- 12
Continuous Variable ---- 12
Data ---- 12
Primary Data ---- 12
Secondary Data ---- 13
Cross Section Data ---- 14
Time Series Data ---- 14
Failure Time Data ---- 14
Panel Data ---- 14
Population ---- 14
Finite Population ---- 15
Infinite Population ---- 15
Homogeneous Population ---- 15
Heterogeneous Population ---- 15
Notion of Statistical Population ---- 15
Exercise – 1 ---- 16
Introduction to Statistical Programs ---- 17
Introduction to Excel ---- 17

Scanned by CamScanner

Introduction to SPSS ---- 20
Introduction to STATA ---- 26

Chapter 2

DESCRIPTIVE STATISTICS

Central Tendency ---- 31
Measures of Central Tendency ---- 32
Arithmetic Mean ---- 32
Weighted Arithmetic Mean ---- 39
Median ---- 46
Partition Values ---- 51
Mode ---- 58
Choice of an Average ---- 62
Dispersion ---- 65
Absolute and Relative Measure of Dispersion ---- 66
Range ---- 66
Quartile Deviation (Semi Inter Quartile Range) ---- 68
Mean Deviation (MD) or Average Deviation ---- 71
Standard Deviation ---- 74
Coefficient of Variation ---- 78
Skewness ---- 82
Karl Pearson's coefficient of skewness (relative measure) ---- 84
Bowley'sMeasure of Skewness ---- 87
Bowley's coefficient of skewness (relative measure) ---- 87
Kurtosis ---- 90
Leptokurtic Curve ---- 90
Mesokurtic Curve ---- 90
Measure of Kurtosis ---- 90
Percentile Coefficient of Kurtosis (Kelly's Coefficient of Kurtosis) ---- 90
Moments ---- 93
Central Moments ---- 93
Raw Moments ---- 96
Measure of Skewness based upon Moments ---- 98
Measure of Kurtosis based upon Moments ---- 99
Stem Leaf Display ---- 101
Five Number Summary ---- 104
Box Plot ---- 105
Residual Plot ---- 111
Normal Probability Plot ---- 111
Using Excel, SPSS and STATA for Descriptive Statistics ---- 114
Exercise - 2 ---- 124

Scanned by CamScanner

Chapter 3

INTRODUCTION TO PROBABILITY

Introduction ---------- 133
Deterministic Experiment ---------- 133
Random Experiments ---------- 133
Trial and Event ---------- 134
Exhaustive Cases (N) ---------- 134
Favourable Cases or Events (m) ---------- 134
Mutually Exclusive Events or Cases ---------- 134
Sample Space ---------- 134
Equally Likely Cases ---------- 134
Impossible Event ---------- 134
Sure event ---------- 134
Independent Events ---------- 135
Dependent Events ---------- 135
Definitions of Probability ---------- 136
Statistical or Empirical or Posteriori or Relative Frequency Approach ---------- 137
Subjective Definition of Probability ---------- 138
Laws of Probability ---------- 138
Conditional Probability ---------- 141
Bayes Theorem, Prior and Posterior Probabilities ---------- 143
Bayes' Theorem (Laws for Inverse Probability) ---------- 143
Baye's Theorem for Future Events ---------- 146
Exercise – 3 ---------- 147

Chapter 4

SAMPLING

Population ---------- 151
Sample Survey verses Census Survey ---------- 151
Sample ---------- 152
Sampling Unit ---------- 152
Sampling Frame ---------- 152
Parameter and Statistic ---------- 153
Factors Affecting Sample Size ---------- 154
Errors in Statistics ---------- 155
Sampling Error ---------- 155
Non-sampling Error ---------- 156

Scanned by CamScanner

Basic Steps in Sampling ---------- 156
Fundamental characteristics ---------- 158
Advantages of sampling ---------- 158
Limitations of sampling ---------- 159
Types of Sampling Methods ---------- 159
Exercise - 4 ---------- 171

Chapter 5

RANDOM VARIABLE AND MATHEMATICAL EXPECTATION

Random Variable ---------- 173
Types of Random Variable ---------- 173
Discrete Random Variable ---------- 173
Continuous Random Variable ---------- 173
Probability Distribution ---------- 174
Probability Mass Function (Probability Distribution of Discrete Random Variable) ---------- 174
Distribution Function ---------- 176
Properties of Distribution Function ---------- 176
Mathematical Expectation ---------- 178
Properties of Expectation ---------- 181
Addition Theorem of Expectation ---------- 182
Multiplication Theorem of Expectation ---------- 182
Variance of univariate random variable ---------- 184
Covariance of two dimensional random variable ---------- 184
Properties of Variance ---------- 185
Conditional Expectation ---------- 185
Properties of Conditional Expectation ---------- 185
Exercise 5 ---------- 188

Chapter 6

PROBABILITY DISTRIBUTIONS

Introduction ---------- 191
Probability Distribution ---------- 191
Joint Probability Distribution of Two Random Variable ---------- 192
Joint Probability mass function (joint pmf) ---------- 192
Marginal Probability Mass Function (Marginal pmf) ---------- 192
Conditional probability mass function (conditional pmf) ---------- 193
Joint probability density function (Joint pdf) ---------- 197
Marginal Probability Density Function (Marginal pdf) ---------- 198

Scanned by CamScanner

Conditional probability density function(Conditional pdf) ---198
Joint Probability Distribution Function ---200
Marginal probability distribution function ---200
Conditional probability distribution function---201
Independence of random variables ---205
Correlation coefficient ---207
Moments about mean of bivariate distribution ---207
Discrete Distribution ---210
Bernoulli Distribution---210
Binomial Distribution---210
Probability Mass Function of Binomial Distribution. ---210
Mean and Variance of Binomial Distribution ---213
Fitting of Binomial Distribution---215
Poisson Distribution---216
Mean and variance of Poisson distribution ---220
Fitting of Poisson Distribution ---221
Normal Distribution ---222
Normal Probability Curve---223
Standard Normal Distribution ---223
Normal distribution as approximation of binomial distribution---223
Normal distribution as approximation of Poisson distribution ---223
Mean and variance of normal distribution---225
The Exponential Distribution---234
Gamma Distribution ---235
Exercise – 6 ---244

Chapter 7

CORRELATION AND LINEAR REGRESSION

Bivariate Data ---253
Bivariate Frequency Distribution ---253
Correlation---255
Simple Correlation---256
Scatter Diagram Method ---256
Karl Pearson's Correlation Coefficient ---257
Spearman Rank Correlation Coefficient ---266
Regression Analysis---271
Lines of Regression---271
Simple Linear Regression---272
Estimation of Coefficients using Least Square Method (OLS) ---273
Linear Relationship---273
Properties of Regression Coefficients---275

Scanned by CamScanner

Measures of Variation ---------- 280
Standard Error of the Estimate ---------- 280
Coefficient of Determination ---------- 281
Residual ---------- 282
Application of Correlation and Regression in Excel, SPSS and STATA ---------- 283
Exercise 7 ---------- 296

Bibliography ---------- 303
Standard Normal (Z) Table ---------- 304
Micro-Syllabus ---------- 305
Model Question ---------- 307

Scanned by CamScanner

Chapter 1

INTRODUCTION

Basic Concept of Statistics

Simple forms of statistics have been used since the beginning of civilization, when pictorial representations or other symbols were used to record numbers of people, animals, and inanimate objects on skins, slabs, or sticks of wood and the walls of caves. Before 3000 BC, the Babylonians used small clay tablets to record tabulations of agricultural yields and commodities bartered or sold. The Egyptians analyzed the population and material wealth of their country before beginning to build the pyramids in the 31st century bc. Numerical records existed in China before 2000 BC. The ancient Greeks held censuses to use as the bases for taxation as early as 594 BC.

The Roman Empire was the first government to gather extensive data about the population, area, and wealth of the territories that it controlled. During the Middle Ages in Europe, few comprehensive censuses were made. Registration of deaths and births was begun in England in the early 16th century, and in 1662, the first noteworthy statistical study of population, Observations on the London Bills of Mortality, was written. A similar study of mortality made in Brussels, Germany, in 1691, was used by the English astronomer Edmond Halley as a basis for the earliest mortality table. In the 19th century, with the application of the scientific method to all phenomena in the natural and social sciences, investigators recognized the need to reduce information to numerical values to avoid the ambiguity of verbal description.

At present, statistics is a reliable means of describing accurately the values of economic, political, social, psychological, biological, and physical data and serves as a tool to correlate and analyze such data. The work of the statistician is no longer confined to gather and tabulate the data but is chiefly a process of interpreting the information. The development of the theory of probability increased the scope of statistical applications. Much data can be approximated accurately by certain probability distributions, and the results of probability distributions can be used in analyzing statistical data. Probability can be used to test the reliability of statistical inferences and to indicate the kind and amount of data required for a particular problem.

Origin and Growth of Statistics

The word 'Statistics' and 'Statistical' are all derived from the Latin word Status, means a politic state. The theory of statistics as a distinct branch of scientific method is of comparatively rece

Scanned by CamScanner

growth. Research particularly into the mathematical theory of statistics is rapidly proceeding and fresh discoveries are being made all over the world.

In the modern world of computers and information technology, the importance of statistics is very well recognized by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on. As on date there is no other human walk of life, where statistics cannot be applied.

Meaning of Statistics

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis. Statistics is concerned with the systematic collection of numerical data and its interpretation. The word 'statistics' is used to refer to

1. Numerical facts, such as the number of people living in particular area.
2. The study of ways of collecting, analyzing and interpreting the facts.

Definitions

Statistics is defined differently by different authors over a period of time. In the olden days statistics was confined to only state affairs but in modern days it embraces almost every sphere of human activity. Therefore, a number of old definitions, which was confined to narrow field of enquiry were replaced by more definitions, which are much more comprehensive and exhaustive.

Secondly, statistics has been defined in two different ways – Statistical data and statistical methods. The following are some of the definitions of statistics as numerical data.

1. Statistics are the classified facts representing the conditions of people in a state. In particular they are the facts, which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.
2. Statistics are measurements, enumerations or estimates of natural phenomenon usually systematically arranged, analyzed and presented as to exhibit important interrelationships among them.

In plural sense

According to Webster "Statistics is classified facts representing the condition of people in a state".

According to Bowley "Statistics is numerical statement of facts in any department of inquiry placed in relation to each other".

According to Yule and Kendall "Statistics is quantitative data affected to a marked extent by multiplicity of causes".

According to Horace and Secrist "Statistics are aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in systematic manner for pre determined purpose and placed in relation to each other". It is complete definition of Statistics in plural sense.

Statistics are aggregate of facts: Statistics are number of facts. A single fact cannot be called Statistics. E.g. A person aged 20 years is not statistics. Number of persons aged 20, 22, 23 etc is statistics.

Statistics are affected to a marked extent by multiplicity of causes: The figure should be affected by a number of causes. e.g. production of rice is affected by amount of irrigation , quality of seed , quality and quantity of fertilizer etc.

Scanned by CamScanner

Statistics are numerically expressed: The figures should be expressed numerically. For the qualitative expressions statistical analysis is not possible unless numerical equivalent is assigned to the expressions. e.g. Population of a village was 3560 in 2068 B.S. and it was only 2795 in 2058 B.S.

Statistics are enumerated or estimated according to reasonable standard of accuracy: Numerical statement can be enumerated in which case, they are supposed to be accurate and when the area of inquiry is wide figures are estimated. The estimated figures can not be absolutely accurate.

Statistics are collected in systematic manner: The statistical information should be collected in planned manner. It should not be collected in haphazard manner.

Statistics are collected for pre determined purpose: The purpose of collecting information should be decided in advanced. The purpose should be specific and well defined.

Statistics are placed in relation to each other: The statistics should be comparable. The valid comparison can be made if numerical statements are related to some process.

Singular sense

According to Bowley "Statistics is science of counting".

According to Bowley "Statistics is science of average".

According to Boddington "Statistics is the science of estimates and probability".

According to Croxton and Cowden "Statistics is the collection, presentation, analysis and interpretation of numerical data". It is complete definition of Statistics in singular sense.

Collection: Data must be collected carefully so that no fault occurs. Data are collected by different sources.

Presentation: The collected data are organized by editing and classifying then are presented in tabular form, diagrammatic form and graphic form.

Analysis: Data are analyzed by using statistical tools such as central tendency, dispersion, correlation, significance test etc. The analysis depends upon nature of data and purpose of enquiry.

Interpretation: Interpret the results obtained from analysis. High level of skill and experience is required to interpret result.

Statistics can be classified into two parts

i) Mathematical Statistics
ii) Applied Statistics

Mathematical Statistics

It is application of Mathematics to Statistics. It involves collection of facts and analysis of facts using mathematical techniques.

It is divided into two parts.

a. Descriptive Statistics
b. Inferential Statistics

Scanned by CamScanner

Descriptive Statistics

Statistics which deals with collection of data, presentation of data using table, diagram, graph etc and summarize data using measure of central tendency, dispersion, skewness, kurtosis etc is called descriptive statistics. Hence summarized result is obtained from descriptive statistics which can describe the data but cannot be used to generalize. Descriptive Statistics is very important because if we simply presented raw data it is difficult to identify what the data was showing. Descriptive Statistics present data in meaningful way so that interpretation of data becomes easier.

Inferential Statistics

Statistics which deals with sample selection from population and statistical techniques used to draw conclusion about population on the basis of statistical measures obtained from sample. With the help of inferential statistics results can be generalized for whole population by computing measures from samples. It is used in estimation of parameters and statistical testing of hypothesis.

Applied Statistics

Statistics which deals with the application of statistical methods to specific problems is called applied statistics. It has application of mathematical statistics on real world problems. It prescribes analysis method depending upon nature of data or nature of problem.

Functions of Statistics

There are many functions of statistics. Let us consider the following five important functions.

Condensation

By the word 'to condense', we mean to reduce or to lessen. Condensation is mainly applied at embracing the understanding of a huge mass of data by providing only few observations. If in a particular class in a Higher Secondary School, only marks in an examination are given, no purpose will be served. Instead if we are given the average mark in that particular examination, definitely it serves the better purpose. Similarly, the range of marks is also another measure of the data. Thus, Statistical measures help to reduce the complexity of the data and consequently to understand any huge mass of data.

Comparison

Classification and tabulation are the two methods that are used to condense the data. They help us to compare data collected from different sources. Grand totals, measures of central tendency measures of dispersion, graphs and diagrams, correlation coefficientetc. provide ample scope for comparison. If we have one group of data, we can compare within itself. If the rice production (in Tons) in Chitwan district is known, then we can compare one region with another region within the district.Or if the rice production (in Tons) of two different districts within Narayani is known, then also a comparative study can be made. As statistics is an aggregate of facts and figures, comparison is always possible and in fact comparison helps us to understandthe data in a better way.

Forecasting

By the word forecasting, we mean to predict or to estimate beforehand. Given the data of the last ten years connected to rainfall of a particular district in Bagmati, it is possible to predict or forecast the rainfall for the near future. In business also forecasting plays a dominant role in connection with production, sales, profits etc. The analysis of time series and regression analysis plays an important role in forecasting.

Scanned by CamScanner

Estimation

One of the main objectives of statistics is drawn inference about a population from the analysis for the sample drawn from that population. The four major branches of statistical inference are

1. Estimation theory
2. Tests of Hypothesis
3. Non-Parametric tests

Sequential analysis

In estimation theory, we estimate the unknown value of the population parameter based on the sample observations. Suppose we are given a sample of heights of hundred students in a school, based upon the heights of these 100 students, it is possible to estimate the average height of all students in that school.

Testing of Hypothesis

A statistical hypothesis is some statement about the probability distribution, characterizing a population on the basis of the information available from the sample observations. In the formulation and testing of hypothesis, statistical methods are extremely useful. Whether crop yield has increased because of the use of new fertilizer or whether the new medicine is effective in eliminating a particular disease are some examples of statements of hypothesis and these are tested by proper statistical tools.

Empirical Research

Research begins with a question. How many of these are also those? Why do so many of those end up in these circumstances? How often does this happen? Why does it seem to happen mainly to this type of person? etc. Some questions can be studied empirically because they are questions about people or events or circumstances in the world that do or do not happen. Other questions can be the topic of heated debate, but they can never be resolved by looking to see what actually the case is. Other questions seem to fall partly in both categories. It is important to understand the kinds of questions that can be studied and to see how they can be framed in ways that make answers possible to find.

Empirical research is the research in which the questions are about things that exist or happen, and in which the answers are obtained by somehow observing things in the world. The answer of the research question cannot rely on reasoning, speculation, moral judgment or subjective preference. It is based on information that can be verified by using direct experience. In example of "Are women paid less than men for the same types of work?", we can get the observation on it because it is of about the people which is exist. But if question is "is racial equality good for society?" the observation is not feasible for it. Statistics is the science and practice of developing human knowledge using empirical data expressed in quantitative form. Statistics is a range of procedures for gathering, organizing, analyzing and presenting quantitative data. 'Data' is the term for facts that have been obtained and subsequently recorded, and, for statisticians, 'data' usually refers to quantitative data that are numbers. Essentially therefore, statistics is a scientific approach to analyzing numerical data in order to enable us to maximize our interpretation, understanding and use. This means that statistics helps us turn data into information; that is, data that have been interpreted, understood and are useful to the recipient. So, statistics is the systematic collection and analysis of numerical data, in order to investigate or discover relationships among phenomena so as to explain predict and control their occurrence. The possibility of confusion comes from the fact that not only is statistics the techniques used on quantitative data, but the same word is also used to refer to the numerical results from statistical analysis. Statistical techniques are valuable tools that help us to be thorough and systematic in asking and answering questions that involve data, observations or information available from the world around us.

Scanned by CamScanner

Scope of Statistics

Statistics is not a mere device for collecting numerical data, but as a means of developing sound techniques for their handling, analyzing and drawing valid inferences from them. The history of statistics shows in the ancient times the scope of statistics was limited. Censuses of population and wealth were conducted in those days to determine the man power and materials wealth for the purpose of waging wars.

With the passage of time the scope of statistics became wider and wider.

Statistics is applied in every sphere of human activity - like Physical and Biological science, Commerce, Education, Planning, Business Management, Information Technology, etc. It is almost impossible to find a single department of human activity where statistics cannot be applied. The medical research worker often must rely upon statistics to determine the significance of his result. It should of course be added that the musician, the artist, the actor and the writer to fiction would rarely have occasion to statistics, but even here certain data sales, box-office receipts and trends of popular taste might be apropos. "Statistics methods are applied to the result of physical chemistry and biological experiments and observation as well to result to obtain in Social and economics investigations.

It is clear from the above that, statistical analysis includes in its fold all quantitative analysis to whatever field of inquiry they might relate. The scope of statistical methods is stretched over all those branches of human knowledge in which a grasp of the significance of large numbers is looked for. The scope of statistical methods therefore, is wide the limiting factor being its applicability to studies of quantitative character only.

Probability theory is the attempt to extract knowledge about what event has happened or will happen in presence of this uncertainty. It tries to quantify as best as possible the risks and benefits involved. Apart from the earliest applications of probability in gambling, numerous others exist: in science, where we make experiments and interpret them, in finance, in insurance and in weather reports etc. In real life we do not have the probabilities, we have the data. And so we want tools that allow us to estimate the probabilities given the data that we have. This is what statistics is about. The difference is therefore merely what is known and what is not. In science it is the data we have and we want to know about the probabilities.

Statistics is the art of guessing the model and its parameters. It is based on probability theory. Probability theory shows us why the particular formula by means of which we guess the model is good. Most prominently, most of the scientific research probability theory underlies statistical inference. Statistical inference involves making generalizations or inferences about unknown population parameters. After selecting a sample from the population of interest, we measure the characteristic under study, summarize this characteristic in our sample and then make inferences about the population based on what we observe in the sample. The concept of probability theory allows us to have better understanding of statistical inference and to quantify the strength or confidence in our conclusion.

The use of statistics is to conduct research effectively. Without the use of statistics it would be very difficult to make decisions based on the data collected from a research project. Statistics, however, provides us with a tool to make an educated decision. We will be able to decide which of the two possibilities is more likely to be true. We will base this decision on our knowledge of probability and inferential statistics.

It is extremely important for a researcher to know what statistics they want to use before they collect their data. Otherwise data might be collected that is uninterruptable. Unfortunately, when this happens it results in a loss of data, time, and money.

Scanned by CamScanner

Limitations of Statistics

Despite its power, essential usefulness and universal applicability statistics has its own limitation and imperfections, which are as under,

1. Statistics does not study individual.
2. Statistics does not study qualitative phenomena
3. Statistics data do not reveal the story
4. Statistical results are not always unquestionable
5. Statistical laws are true on the average or in the long run
6. Statically result might leads to fallacious conclusions if they are quote without their context or If they are manipulated
7. Statistical data are liable to be misused easily.

Each of these limitations is discussed in detail below:

Statistics does not study individuals

Individuals facts and figures are of importance to individuals only statistics does not deals with them. Statistics deals with mass phenomena and therefore, throws light on the whole of a given group. Statistical deals with aggregates, though for purpose of analysis these aggregates are very often reduced to single figures like average, percentages.

Statistics does not study qualitative phenomenon

The statically methods can be applied to the study of those problems only that are capable of quantitative expression, i.e., it is not applicable to the study of those facts that are not quantitatively measurable, for example attributes like honesty, healthy skills etc be measured weight in figures.

Statistical data not reveal the entire story

Statistical data collected relating to problem under study do not reveal the entire story of the problems. This is because many problems are affected by factors which are in capable of statistical analysis. Statistician should always keep in mind that the data do not tell everything of the facts it expressed in their summarized, reduced form, they leave out a lot of information that may be importance.

Statistical results are not always unquestionable

Statistical results are more in the nature of estimate and probabilities than exact statement. The statistical methods are one of the several methods of studying problems. It helps us in studying trends, in forming an idea of the probabilities, in knowing how a given phenomenon has been behaving generally necessarily unquestionable and reliable.

Statistical laws are true on the average or in the long run

That is, they are not true in their entirety-they are statements of long-term tendencies, and, therefore, cannot be considered stables. Statistical laws are not 100% correct because statistics deals with phenomena which are affected by a multiplicity of causes. And it is not possible to study statistically the effects of each of those causes to factors separately, as it is possible under the experimental methods. Because of this limitation, conclusion arrived at are not perfectly accurate and consequently the same conclusions cannot be arrived at under similar condition at all times.

Statistical results might lead to fallacious conclusions they are quoted without their context or if they are manipulated

Statistical results are true only in a particular context, if the context is absent, we cannot be sure of their validity. E.g. if the average profits earned by two firms are over the past few years are quoted, and incidentally, m the averages happen to be identical quantities, the conclusions may be drawn that both the firms are doing equally well. But if the actual profits earned by each of the two firms

Scanned by CamScanner

during the period were examined, it may be found that the profits of one firms is going up year by year while those of the other are going down by down.

Statistics liable to be misused easily

Any person can use statistics and draw any type of conclusion he likesthey may be used by anyone to make a worse case appears to be a better case. Like medicines in the hands of quacks, statistics can be easily misused by the inexpert, statistical data can be scientifically handled only by those who have expert knowledge of statistical methods.

Application of Statistics in the Field of CSIT

Statistics respond appropriately to the new demands of research and development in various areas of experimental sciences. With the advances in computer technology, many innovative techniques for use in applied statistics have recently been developed. In medicine, agriculture, business and government, these applications have solved important problems. Performing calculations almost at the speed of light, the computer has become one of the most useful research tools in modern times. Computers are ideally suited for data analysis concerning large research projects. Researchers are essentially concerned with huge storage of data, their faster retrieval when required and processing of data with the aid of various techniques. In all these operations, computers are of great help. Their use, apart expediting the research work, has reduced human drudgery and added to the quality of research activity.

The computers can perform many statistical calculations easily and quickly. Computation of means, standard deviations, correlation coefficients, 't' tests, analysis of variance, analysis of covariance, multiple regression, factor analysis and various nonparametric analyses are just a few of the programs and subprograms that can be solve very easily using computer. Similarly using computer, linear programming, multivariate analysis, Monte Carlo simulation etc. are also can be done very easily. Software packages are readily available for the various simple and complicated analytical and quantitative techniques of which researchers generally make use of. The only work a researcher has to do is to feed in the data he/she gathered after loading the operating system and particular software package on the computer.

The output, or to say the result, will be ready within seconds or minutes depending upon the quantum of work. The storage facility of computers provides statistician the immense help to use the stored data whenever it required. Innumerable data can be processed and analyzed with greater ease and speed. Moreover, the results obtained are generally correct and reliable. Not only this, even the design, online data collection, pictorial graphing and report are being developed with the help of computers.

Hence, researchers should be given computer education and be trained in the line so that they can use computers for their research work.

Researchers interested in developing skills in computer data analysis, must be aware of the following steps:

(i) data organization and coding;

(ii) storing the data in the computer;

(iii) selection of appropriate statistical measures/techniques;

(iv) selection of appropriate software package;

(v) execution of the computer program.

First of all, researcher must pay attention toward data organization and coding prior to the input stage of data analysis. If data are not properly organized, the researcher may face difficulty while analyzing their meaning later on. For this purpose, the data must be coded. Categorical data need to be given a number to represent them. Once the data is coded, it is ready to be stored in the computer.

Scanned by CamScanner

Input devices may be used for the purpose. After this, the researcher must decide the appropriate statistical measure(s) he will use to analyze the data. He will also have to select the appropriate program to be used. SPSS, SAS, STATA etc. are the special statistical packaged program whereas Microsoft Excel can be used for simple statistical analysis.

The above description indicates clearly the usefulness of computers to researchers in statistical data analysis. Researchers, using computers, can carry on their task at faster speed and with greater reliability.

Recently developed statistical procedures such as those of data mining, jackknifing, bootstrap, pattern recognition and image analysis are essentially computational. These procedures have contributed significantly to the solution of many practical problems. The developments now taking place in computer technology will further enhance and facilitate the use of computers for researchers. Programming knowledge would no longer remain an obstacle in the use of a computer.

In spite of all this sophistication we should not forget that basically computers are machines that only compute, they do not think. The human brain remains supreme and will continue to be so for all times. As such, researchers should be fully aware about the following limitations of computer-based analysis:

1. Computerized analysis requires setting up of an elaborate system of monitoring, collection and feeding of data. All these require time, effort and money. Hence, computer-based analysis may not prove economical in case of small projects.
2. Various items of detail which are not being specifically fed to computer may get lost sight of.
3. The computer does not think; it can only execute the instructions of a thinking person. If poor data or faulty programs are introduced into the computer, the data analysis would not be worthwhile. The expression "garbage in, garbage out" describes this limitation very well.

In the field of statistics, technology is commonly thought of computers and the associated statistical computing packages, which together have become every statistician's indispensable working tool. There is no doubt that the impact of such technology has produced striking changes in the nature of both statistical research and statistical practice. To calculate the different statistical tools from the huge mass of data is too much tedious, computer software helps us to make it simple and easy.

Besides this the statistical there might be the problems specially related to computer science and information technology which can be address by statistics. For example if we have to identify the motherboards which has the lowest rate of failure according to industry studies or to get the answer of question of "How long should we expect these LCD monitors to last before failure?" then we must has to take help of statistics.

Below are some examples which is to be solved using statistics:

Which department is increasing their disk-space usage unusually fast? Is that part of the normal variations or is something new going on?

We have been testing three different antispam products over the last six months; is any one of them obviously better than the others?

Which of our programming teams has the lowest rate of coding errors per thousand lines of production code? Is the difference between their error rate and those of the other team significant enough to warrant a lecture from them, or is it just one of those random things?

Can you identify the callers responsible for 80% of our helpdesk calls? Is there anything in common among them that would help us identify areas for improvement through increased training?

Can we distinguish between the characteristics of normal software and those of malware based on what they do, how often, and how fast?

Scanned by CamScanner

Scales of Measurement

Measurement consists of counting the number of units or parts of units displayed by objects and phenomena. Measurement is a process of assigning numbers or symbols to any facts or objects or products or items according to some rule. It is a tool by which individuals are distinguished on the variables of area under study. Scale is simply a range of levels or numbers used for measuring something. It is a set of all the different levels of symbols or numerals or something so constructed, from the lowest to highest, that these can be assigned by rule to objects or to items or to the individuals or to their behavior to whom it is applied.

Different measurement scales are used on the basis of nature of data. These measurement scales of variables under study are (a) Nominal scale (b) Ordinal scale (c) Interval scale (d) Interval scale

Nominal Scale

It is the simplest type of scale, also known as categorical scale. It is lowest level of measurement scale. It is simply a system of assigning number or the symbols to objects or events to distinguish one from another or in order to label them. The symbols or the numbers have no numerical meaning. The arithmetic operations cannot be used for these numerals. The orders of the symbol have no mathematical meaning.

For example, gender, occupation, religion are measured in nominal scale. If we use scale 1 for male programmer and 2 for female programmer for measuring the gender of programmer, then 1 and 2 have no numeric meaning. It is used to distinguish male and female. In this case the number of a set of objects is not comparable to the other set. Mathematical operations such as addition, subtractions, multiplication, and division cannot be performed for the variable having nominal scale. Frequency count is possible in this case so that percentage, mode can be obtained, and chi square test can be performed for test of significance for variable having nominal scale.

Ordinal Scale

The second and the lowest level of ordered scale is the ordinal scale. It is the quantification of items by *ranking*. In this scale, the numerals are arranged in some order but the gaps between the positions of the numerals are not made equal. It is used to rate preference of respondents. It indicates relative extent to which object possess certain characteristic. It represents qualitative values in ascending or descending order. The rank orders represent ordinal scale and mostly useful in scaling the qualitative phenomena.

For example, qualification levels, preference to different statistical software, preference to different made of Laptops are measured in ordinal scale. If we are going to study the computer literacy and we took the qualification level of people as primary, lower secondary, secondary , higher secondary , bachelor , master and Ph. D then we can use 1 , 2 , 3, , 4, , 5, 6 , 7 to represent primary , lower secondary , secondary , higher secondary , bachelor , master and Ph.D. respectively in ascending order. Mathematical operations such as addition, subtraction, multiplication, division cannot be performed for variable having ordinal scale. Frequency count is possible and due to ranking partition values can be determined for variable having ordinal scale. In this case median, mode, rank correlation also can be obtained.

Interval Scale

In addition to ordering the data, this scale uses equidistant units to measure the difference between scores. It assumes data have equal intervals. This scale does not have absolute zero but only arbitrary zero. Interval scale is the developed form of the ordinal scale. The intervals between the ordered numerals are adjusted in terms of some rule.

For example, scale of temperature is an example of ordinal scale. In an increase in temperature from 320F to 420F and from 640F to 740F, we can say the increases are equal of 100F; but one cannot say

Scanned by CamScanner

that the temperature of 640F is twice as warm as the temperature of 320F. The 00C or 320F are the arbitrarily set points for the freezing point of water so the temperatures 320F and 640F are not viable to express in ratio because the zero is not a true zero but it is an arbitrary point. Arithmetic mean, standard deviation, correlation can be obtained for variable having ordinal scale. To test significance t-test and F test can be used.

Ratio Scale

Ratio scale is the ideal scale and an extended form of Interval Scale. It is most powerful scale of measurement. It possesses the characteristics of nominal, ordinal and interval scale. Ratio scale has an absolute zero or true zero or natural zero of measurement. The true zero point or the initial point indicates the completely absence of that property of an object what is being measured. It can be expressed as relative level of measurement.

For example, the absolute zero or the natural zero in the centimeter scale indicates the absence of the length. Numbers on the scale indicates the actual amount of property being measured. The ratio involved in the ratio scale possesses measure property and it facilitates comparison, which is not possible in interval scale. Mathematical operations like addition, subtraction, multiplication and division all can be performed.

It represents actual number of variables and it is used to measure the physical dimensions. Some examples of ratio scale variables are disk space, amount expenses in IT department in different year, age of programmers, etc. Geometric mean, Harmonic mean, coefficient of variation along with all other measures can be obtained for variable having ratio scale. All statistical techniques can be applied to the variable present in ratio scale.

The following is the properties of categories, rank, equal interval and true zero point of four scales.

Level of measurement	Property			
	Categories	Ranks	Equal intervals	True zero points
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Variable

An element which changes or varies is called variable. It is a character under study. It is generally denoted by x, y, z etc. The values of variable vary from one to another. E.g. height of persons, marks secured by students in an exam, religion of people etc. Generally variable is classified into two types

i. Qualitative variable

ii. Quantitative variable

Qualitative Variable

The variable which varies in kind rather than in magnitude is called qualitative variable. It can be divided into different categories. It is also called categorical variable. Qualitative variables are present in nominal and ordinal scales. E.g. hair color, eye color, gender, smoking habit, academic achievement, social status etc.

Scanned by CamScanner

Quantitative Variable

The variable which varies in magnitude and can be expressed numerically is called quantitative variable. Quantitative variables are present in interval and ration scales. e.g. age, height, weight, income etc. Quantitative variable can be further divided into two types

a. Discrete variable

b. Continuous variable

Discrete Variable

A variable which can take whole number or countable number of values is called discrete variable. E.g. number of rooms in building, number of road accidents, family size, household size etc.

Continuous Variable

A variable which can take all possible values (whole number and fractions) within certain range is called continuous variable. e.g. height , weight, temperature, age etc.

Entities

Something that exist apart from other thing having own independent existence. E.g. age, sex, color of hair etc.

Data

It is a set of values obtained on one or more characters under study. E.g. marks secured by students in a test. In any statistical investigation first attempt is to collect data. It is the collected information which is ready to use for statistical analysis. On the basis of collection procedure data is divided into two types.

i. Primary data

ii. Secondary data

Primary Data

The data which are originally collected by investigator or researcher for the first time for the purpose of statistical enquiry is called primary data. It is collected by government, an individual, institution and research bodies. It needs more fund, time and manpower. It is more reliable and suitable. Different methods used for collecting primary data are as follows:

a. Direct personal interview

b. Indirect oral interview

c. Mailed questionnaire

d. Information through telephone

e. Information through correspondents

f. Schedule sent through enumerator

Direct personal interview

According to this method, investigator obtains data by the personal observation of the objects under study. The investigator presents himself personally before the informant and obtains first hand

Scanned by CamScanner

information. This procedure is adopted when the field of inquiry is small and there is desire for a greater accuracy

Indirect oral interview

According to this method information is collected from third person when respondent or informants hesitate to give their answer or information desired in complex. Under this method, instead of directly approaching the informants, the investigator interviews several third persons, who are directly in touch with information sought.

Mailed questionnaire

According to this method a questionnaire containing a number of questions pertaining to the object under investigation is prepared. The questionnaires are sent to informants requesting to answer the questions and return the questionnaire. It is cheap method of collecting information. It has high chance of non response.

Information through telephone

According to this method information is obtained through telephone instead of face to face contacting informants. It is quick and accurate method of getting information. It is cheap method of collecting information.

Information through correspondents

According to this method information can be collected from local agents or correspondents appointed in different areas under investigation. The agent provides information of different incidence to the investigator or central office where data are processed.

Schedule sent through enumerator

According to this method enumerator carries list of questions with space to write answer and fill with their own handwriting after getting answer of questions from informants. It can be carried out for both literate and illiterate. It has very low chance of non response. It requires skilled enumerators. It is expensive method of getting information.

Secondary Data

The data that has been already collected for a particular purpose and used for next purpose is called secondary data. Hence one purpose primary data is another purpose secondary data. It is not new and original data. This type of data is generally published in newspapers, magazines, bulletins, reports, journals, website, radio etc. When investigators find impracticable to collect firsthand information on related issues, secondary data is used. It saves time, money and manpower.

Before using secondary data following precautions should be used.

- The data should be suitable for the purpose of inquiry or investigation.
- The data should be reliable and adequate

Source of secondary data

Secondary data source is broadly classified into two types

i. Published source

ii. Unpublished source

Scanned by CamScanner

Published Source

Different published source of secondary data are national organizations and international agencies. The national organizations are governmental, semi governmental and private organizations.

i. International agencies such as World Health Organization, World Bank, International Labor Organization

ii. Governmental organizations such as Central Bureau of Statistics, Ministry of Commerce and Industry, Nepal Rastra Bank, Ministry of Finance

iii. Semi-Governmental organizations such as Nepal Food Corporation, Nepal Electricity Authority

iv. Private organizations such as Nepal Chamber of Commerce, Federation of Nepal Chamber of Commerce and Industry (FNCCI), Publications

Unpublished Source

All the statistical information may not be in published form.. Some of Governmental and private organizations do not publish statistical information. Records of (i) individual business (ii) schools (iii) colleges (iv) village development committee (v) trade centre etc are not published.

Cross Section Data

Data for a single point or single factor is called cross section data. It is a snapshot of information at a particular point. This type of data cannot describe changes over time or cause and effect relationships in which one variable affects other variable. E.g population of children in census year 2068.

Time Series Data

The data which can be recorded over different period of time is called time series data. In this case same measurements are recorded on regular basis i.e daily, weekly, monthly, quarterly, half yearly, yearly etc. E.g. population of Nepal in census years 2048, 2058, 2068.

Failure Time Data

The data of each unit is recorded for each follow up time till the occurrence of event or till the unit fails is called failure time data. The unit is taken with similar characteristic and follows up till the occurrence of event. It is also called time to event data. It is data obtained from clinical studies, cancer studies, biomedical sciences etc.

Panel Data

It is longitudinal or cross sectional time series data. It is data related to behavior of entities observed across time. Data of individuals is recorded repeatedly over number of years. e.g. income of persons X and Y in years 2013, 2014 and 2015 according to age and qualification.

Population

It is totality of units or items under study belonging to a particular a class or group.e.g. children in a school, patients in a hospital, fruits in a tree, fishes in a pond etc. Census survey is conducted to enumerate all the population units. Population can be divided into finite and infinite population according to number of individuals belonging to the group.

Scanned by CamScanner

Finite Population

Population containing countable number of individuals is called finite population. E.g. vehicles in workshop, customers in shopping mall, passengers in vehicle etc.

Infinite Population

Population containing uncountable number of individuals is called infinite population. E.g. fishes in an ocean, stars in the sky, living beings in earth etc.

Population can be further divided into homogeneous and heterogeneous according to type of individuals in population.

Homogeneous Population

Population consisting of individuals of same type is called homogeneous population. e.g. population of graduate passed out student.

Heterogeneous Population

Population consisting of individuals of different types is called heterogeneous population. It contains sub population of different types. E.g. population of United States.

Notion of Statistical Population

Data collection is a crucially important step in Statistics. We use the collected and observed sample to make statements about a much larger set — the population.

Population is the entire pool from which a statistical sample is drawn. In statistics, population can be referred as people, objects, events, measurements etc. A population also can be said as the aggregate observation of subjects grouped together by a common feature. A population can be vague or specific. Computer users of Nepal, Tech startups in Asia, Students of Tribhuvan University are vague population. Population can also be defined more specifically like Nepalese Dell computer users, Tech startups in Asia that failed in less than two years,CSIT students of Tribhuvan University.

Most of the time people want to know the characteristics of every entity in a population, so as to draw valid conclusion. However, most of the time it is impossible due to different causes (time, cost and nature of study). For example, if a company wants to know whether its 50,000 customers serviced during the year were satisfied, it might be challenging, costly, and impractical to call each of the clients on the phone to conduct its survey. Since the characteristics of every individual in a population cannot be measured due to constraints of time, resources and accessibility, a sample of the population taken.

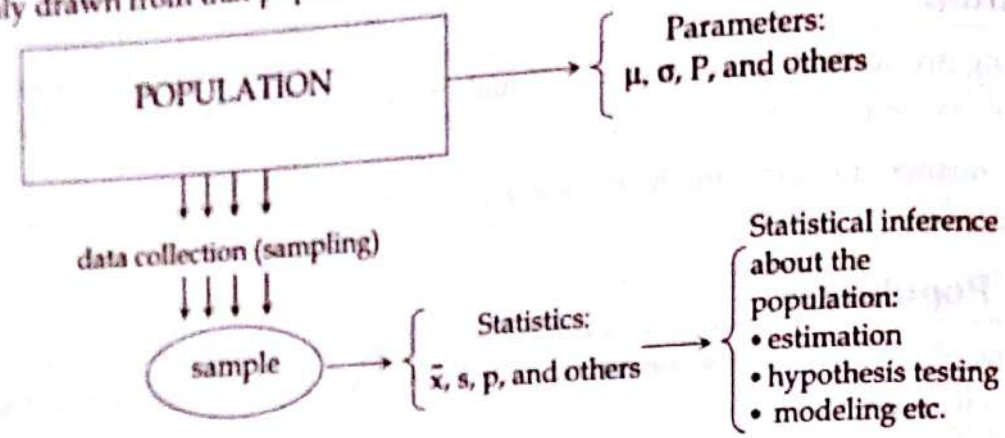
A sample is a random selection of members of a population. It is a smaller group drawn from the population that has the characteristics of the entire population. The observations and conclusions made against the sample data is attributed to the population. The information obtained from the statistical sample allows statisticians to develop hypothesis about the larger population. In statistical sense, the number of items in population generally denoted by uppercase 'N', whereas the sample generally denoted with lowercase 'n'.

For example, let's say a laptop manufacture wants to check the quality of fabrication on its lot before shipping them. It is not cost effective to examine every single item it manufactured. Instead, the manufacturer looks at 50 laptops to draw a conclusion about whether all 500 laptops are likely to have fabricated correctly.

A parameter is data based on an entire population. Any numerical characteristic of a population is a parameter. It may be averages and standard deviations. The population mean and population

Scanned by CamScanner

standard deviation are represented by the Greek alphabets μ and σ, respectively. In same way the numerical characteristics of a sample is referred as statistic. Inferential statistics enables us to make an educated guess about a population parameter based on a statistic computed from a sample randomly drawn from that population.



Exercise – 1

1. What do you mean by statistics? Describe.
2. Discuss limitations of statistics.
3. Describe the scope of statistics
4. Discuss the application of statistics in of computer science and information technology.
5. State with suitable examples the role played by computer technology in applied statistics.
6. What is measurement scale? Describe different types of measurement scale.
7. What do you mean by variable? Mention its types.
8. What do you mean by data? Differentiate between primary data and secondary data.
9. Describe different methods of collecting primary data.
10. Describe different sources of secondary data.
11. What is population? Describe different types of population
12. Write short note on
    i. Failure time data
    ii. Cross section data
    iii. Sample
    iv. Entities
    v. Ratio scale
    vi. Ordinal scale
    vii. Time series data
    viii. Panel data
13. Define statistical population.
14. Statistics is the most necessary component for 'Big Data Analysis'. Explain.

□□□

Scanned by CamScanner

Chapter 2

DESCRIPTIVE STATISTICS

Central Tendency

One of the most important objectives of statistical device is to get one single value that represents the characteristic (property) of the entire data. Such a value is called the central value or an average or the expected value of the variable. The word average is commonly used in day to day conversation. For example, we often talk of average boy in a class it means that he is neither very good nor very bad just a mediocre type of student.

It is clear from the above definitions that an average is a single value that represents a group of values. Such a value is of great significance because it depicts the property of the whole group. Its average value lies somewhere in between the two extremes i.e. the largest and the smallest items. For this reason, it is also known as central value. The process for obtaining an average value from the entire data is known as measure of central tendency.

Objectives

There are two main objectives:

1. **To get a single value that represents the characteristic of the entire data**

   One value can represent thousand lakhs and even millions of values. For examples, it is impossible to remember the individual incomes of millions of earning people of Nepal and even if one could do it there is hardly any use. But if the average income is obtained by dividing the total national income by total population, we get one single value that represent the entire population such a figure would throw light on the standard of living of an average Nepali.

2. **To facilitate comparison**

   Comparison between two or more than two data can be made either at a same time or different time period or different places. For example, we can compare the percentage pass results of two colleges for 2062 and thereby conclude which college the best.

Requisites of a Good Average

(i) It should be rigidly defined

Scanned by CamScanner

(ii) It should be easy to understand
(iii) It should be simple to compute
(iv) It should be based on each observation
(v) It should be suitable for further algebraic treatment
(vi) It should not be affected by extreme items
(vii) It should be least affected by fluctuations of sampling

If an average satisfies these properties it is known as an ideal average.

Measures of Central Tendency

The following are the measures of central tendency.

1. Arithmetic Mean
2. Geometric Mean
3. Harmonic Mean
4. Median and
5. Mode

Arithmetic Mean

The most popular widely used measure of representing the entire data by one value is the arithmetic mean. Its value is obtained by adding together all the items and by dividing this total by the number of items. Arithmetic mean may be either.

1. Simple arithmetic mean
2. Weighted arithmetic mean

Calculation of Simple Arithmetic Mean (Individual Observation)

1. Direct method
2. Deviation method
3. Step deviation method

1. Direct method

The process of computing mean in case of individual series is very simple. Add together the various values of the variable and divide the total by the number of items. Mathematically let be the value of n observation then arithmetic mean denoted by $\bar{x}$ (x-bar) or A.M. is given by

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{\Sigma x}{n}$$

Where Σx= total of the numerical values of each and every observation N is the number of observation.

Example 1. The following table gives the yearly income(000Rs) of the 10 computer programmers working in a software company.

Income (000'Rs.) 1780, 1760, 1680, 1750, 1830, 1940, 1100, 1800, 1060, 1950. Find average income of programmers.

Solution: Calculation of the arithmetic mean of Income.

Let, average income is denoted by, $\bar{x}$ then

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{1780 + 1760 + 1680 + 1750 + 1830 + 1940 + 1100 + 1800 + 1060 + 1950}{10} = 1665$$

Therefore, the average income is Rs. 1665 (000 rs).

Scanned by CamScanner

2. Deviation method (Change of origin method)

When the number of items is very large, the calculation of mean by direct method is quite tedious and time consuming.In such a case we take the deviations of the items from any arbitrary number for calculation of arithmetic mean.

Symbolically,

$$\bar{x} = A + \frac{\Sigma d}{n}$$

WhereA = assumed mean

d = x-A (also called the change of origin; deviation of the item taken from assumed mean)

n = No. of observations

Note:Any value whether existing in the data or not can be taken as assumed mean (A). However, the nearer the assumed mean is to the actual mean.

Example 2: Calculate the average income by short cut method from example 1.

Solution

Let assumed mean A = 1800

Calculation of A.M.

x(000'Rs)	d = x-A = x-1800
1780	-20
1760	-40
1680	-120
1750	-50
1830	30
1940	140
1100	-700
1800	0
1060	-740
1950	150
	Σd= –1350

$$\bar{x} = A + \frac{\Sigma d}{n}$$

$$= 1800 + \frac{-1350}{10} = 1665 \text{ (000' Rs)}$$

3. Step Deviation Method (Change of origin and scale method)

This method is more suitable if the difference between the items is same or multiple of same number. In this method we take the deviation of the items from an assumed mean A. It is given by,

$$\bar{x} = A + \frac{\Sigma d'}{n} \times h$$

Where A = Assumed mean, h = a common factor in the data, n = Number of item

This method is also called the method of change of origin & scale.

Example 3: Calculate the average income by step deviation method of example 1.

Solution

Let Assumed mean A = 1700

Scanned by CamScanner

Calculation of arithmetic mean

x (000' Rs)	d = x-A = x-1700	$d' = \frac{x - A}{h} = \frac{x - 1700}{10}$
1780	80	8
1760	60	6
1680	-20	-2
1750	50	5
1830	130	13
1940	240	24
1100	-600	-60
1800	100	10
1060	-640	-64
1950	250	-25
	Σd'	-35

We have

$$\bar{x} = A + \frac{\Sigma d'}{n} \times h = 1700 + \frac{-35}{10} \times 10 = 1665 \text{ (000'Rs)}$$

Calculation of ArithmeticMean (Discrete Series)

Arithmetic mean can be computed by applying

1. Direct method
2. Deviation method
3. Step deviation method

1. Direct method

In discrete series observation $x_1$occurs f_1 times, observation x_2 occurs f_2 times, observation $x_3$occurs f_3 times, ... observation $x_n$occurs f_n times ,then the arithmetic mean denoted by $\bar{x}$ is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{N}$$

$$= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N} \quad ..... (4)$$

Wheref = frequency, x = the variable in problem, N = total number of observations

Example 4: From the following data of marks obtained by 60 students in an exam of C programming Calculate A.M.

Marks	20	30	40	50	60	70
No. of Students	8	12	20	10	6	4

Solution

Let the marks be denoted by x and the number of student by f

Calculation of A.M. using direct method

Marks (x)	No. of Student (f)	fX
20	8	160
30	12	360
40	20	800
50	10	500

Scanned by CamScanner

60	6	360
70	4	280
	N = Σ f = 60	Σfx = 2460

Hence the average marks $\bar{x} = \frac{\Sigma fx}{N} = \frac{2460}{60} = 41$

2. Deviation method

The formula for calculating A.M. by this method is defined by

$$\bar{x} = A + \frac{\Sigma fd}{N} \quad .... (5)$$

Where, A = assumed mean (d) = x – A, N = Σf = total number of observations

Example 5: Calculate A.M. by the short-cut method from example 4.

Solution

Calculation of A.M.

Marks(X)	No. of Student(f)	d = x– A= x– 40	fd
20	8	–20	–160
30	12	–10	–120
40	20	0	0
50	10	10	100
60	6	20	120
70	4	30	120
	N = 60		Σfd= 60

$$\bar{x} = A + \frac{\Sigma fd}{N} = 40 + \frac{60}{60} = 41$$

3. Step-deviation method

In order to simplify the calculations, we can divide the deviations by common factor i.e. calculate $\frac{x - A}{h}$ and then multiply by C in the formula for getting A.M.

$$\bar{x} = A + \frac{\Sigma fd'}{N} \times h \qquad .....(3)$$

Where A=assumed mean (d') = $\frac{x - A}{h}$ = step deviation, h = common factor

Example 6. Calculate arithmetic mean by step deviation method from example 4

Solution: Calculation of A.M.

Marks X	No. of Student f	d = x - A= x– 40	$d' = \frac{x - A}{h} = \frac{x - 40}{10}$	fd'
20	8	- 20	-2	-16
30	12	- 10	-1	- 12
40	20	0	0	0
50	10	10	1	10
60	6	20	2	12
70	4	30	3	12
	N = 60		Σfd'	6

Scanned by CamScanner

By using formula,

$$\bar{x} = A + \frac{\Sigma fd'}{N} \times h = 40 + \frac{6}{60} \times 10 = 41$$

Hence the average marks = 41.

Calculation of A.M. (Group Data/Continuous Series)

In continuous series; arithmetic mean may be computed by following methods.

1. Direct method
2. Deviation Method
3. Step deviation method

1. Direct method

Arithmetic mean by this method is defined,

Where, m = Mid-point of various classes (m) = $\frac{\text{Lower limit + Upper limit}}{2}$

f = frequency of each class

N = Σf = total frequency

Example 7. The following data represents waiting time of students in bus stand to be seated. Calculate average waiting time by direct method.

Waiting time(min)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of students	5	10	25	30	20	10

Solution: Calculation of A.M. by direct method.

Waiting time	No of students(f)	Mid point (m)	fm
0 – 10	5	5	25
10 – 20	10	15	150
20 – 30	25	25	625
30 – 40	30	35	1050
40 – 50	20	45	900
50 – 60	10	55	550
	N = 100		Σfm = 3300

We have

$$\bar{x} = \frac{\Sigma fm}{N}$$

$$= \frac{3300}{100} = 33$$

Hence average waiting time is 33minute.

2. Deviation Method

Arithmetic mean by this method is given by

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

Where A = Assumed mean (d) = m – A (Deviation of mid-point from A)

N = total number of frequencies = Σf

Example 8 Calculate the A.M. by the short cut method from the data of Example 7.

Scanned by CamScanner

Waiting time	No. of students(f)	Mid-point(m)	d = m – A= m – 35	fd
0 – 10	5	5	-30	-150
10 – 20	10	25	-20	-200
20 – 30	25	25	-10	-250
30 – 40	30	35	0	0
40 – 50	20	45	10	200
50 – 60	10	55	20	200
	N = 100			Σfd = -200

Solution

Calculate AM using short cut method

$$\bar{X} = A + \frac{\Sigma fd}{N} = 35 + \frac{-200}{100} = 35 - 2 = 33$$

Hence,average waiting time is 33 years.

3. **Step-deviation Method**

The A.M. by this method is given by

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

Where A = Assumed mean, $d' = \frac{m - A}{h}$ (step deviation)

h= Common factor or class size

m =Mid-point of class interval

Note:In fact, step deviation method should be adopted where ever possible because it minimizes the calculations.

Example 9. Calculate the A.M. by the step-deviation method from the data of Example 7.

Solution

Calculation of A.M. by step deviation method

Waiting time	No of students	Mid-point (m)	m – A = m – 35	$d' = \frac{m - A}{h} = \frac{m - 35}{10}$	fd'
0 – 10	5	5	-30	-3	-15
10 – 20	10	15	-20	-2	-20
20 – 30	25	25	-10	-1	-25
30 – 40	30	35	0	0	0
40 – 50	20	45	10	1	20
50 – 60	10	55	20	2	20
	N = 100				Σfd' = –20

$$\bar{x} = A + \frac{\Sigma fd'}{N} \times h = 35 + \frac{-20}{100} \times 10 = 35 - 2 = 33$$

Hence, Average waiting time is 33 years.

Example 10. The mean marks of 100 students were found to be 40. Latter on it was discovered that the score of 53 was misread as 83. Find the correct mean corresponding to the correct score. Also find the mean after ignoring wrong score.

Solution

Here N = 100, $\bar{x} = 40$

Scanned by CamScanner

Wrong $\Sigma x = n\bar{x} = 100 \times 40 = 4000$

Wrong score = 83, Correct score = 53

Thus correct (Σx)= Wrong Σx– Wrong score + Correct score

$= 4000 - 83 + 53 = 3970$

$\therefore$ Correct mean $(\bar{x}) = \frac{\Sigma x}{n} = \frac{3970}{100} = 39.70$

Ignoring wrong score

Correct $\Sigma x = 4000 - 83 = 3917$, Correct $n = 100 - 1 = 99$

Correct mean $(\bar{x}) = \frac{\Sigma x}{n} = \frac{3917}{99} = 39.5656$

Calculation of A.M. in Case of Open End Classes

Open end classes are those in which lower limit of the first class and the upper limit of the last class are unknown. In such a case we cannot find out the A.M. unless we make an assumption about the unknown limits. The assumption would naturally depend upon the class interval following the first class and preceding the last class. For example, observe the following data.

Example: i

Marks	Below 10	10 – 20	20 – 30	30 – 40	Above 50
No of students	4	6	10	7	3

In above case since the class interval is uniform the appropriate assumption would be that the lower limit of the first class is zero and the upper limit of the last class is 60. The first class thus would be 0-10 and last class 50-60.

Example: ii

Marks	Below 10	10 – 30	30 – 60	60 – 100	Above 100
No of students	4	6	10	7	100

In the above example classes are not uniform. Here class interval is 20 in the second class, 30 in the third class, 40 in the fourth class i.e. it is increasing by 10. The appropriate assumption would be that lower limit of the first class is 0 and the upper limit of last class is 150. In other words first class is 0-10 and last class is 100-150.

Example: iii

Marks	Below 10	10 – 20	20 – 40	40 – 50	Above 50
No of students	2	5	6	3	1

If the class interval are of varying width. The appropriate assumption would be taking smallest interval in above example lower limit of the first class is 0 and the upper limit of last class is 60.

Properties of A.M.

1. $\bar{x} = \frac{\Sigma x}{n}$, $n\bar{x} = \Sigma x$

   This property is of great practical value. If we know any two values then we can calculate remaining third value. For example if we know $\bar{x} = 20$ and $N = 50$. We can compute total expenditure. $\Sigma x = n\bar{x} = 50 \times 20 = 1000$.

2. The sum of the deviations of the items from the A.M. is always zero.

   i.e. $\Sigma(x - \bar{x}) = 0$, $\Sigma f(x - \bar{x}) = 0$

3. The sum of the squared deviations of the items from A.M. is less than the sum of the squared deviations of the items from any other value. i.e. $\Sigma(x - \bar{x})^2 < \Sigma(x - A)^2$

   $\Sigma(f(x - \bar{x})^2 < \Sigma(f(x - A)^2$

Scanned by CamScanner

4. If we have the arithmetic mean and number of items of two or more than two related group, we can compute **combined average** of these groups by applying the following formula.

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Similarly, if we have to find out the combined mean of three sub groups the above formula can be extended as follows:

$$\bar{x}_{123} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3}$$

Where $\bar{x}_{12}$ = combined mean of two groups
$\bar{x}_1$ = A.M. of the first group
$\bar{x}_2$ = A.M. of the second group,
$\bar{x}_3$ = A.M. of the third group,
$\bar{x}_{123}$ = Combined mean of three groups
n_1 = Number of items in the first group
n_2 = Number of items in the second group
n_3 = Number of items in the third group

Example 11. The average number of files in 20folders of drive C is 60 and the average number of files in 10 folders of drive D in the same computer is 57 find the average number of files in total of 30 folders of the computer.

Solution

Here Number of folders in drive C(n_1) = 20 Number of folders in drive D(n_2)= 10

Average number of files in folder of drive C($\bar{X}_1$) = 60

Average number of files in folder of driveD($\bar{X}_2$)=57

Average number files for 30 folders denoted by$\bar{X}_{12}$is given by

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{20 \times 60 + 10 \times 57}{20 + 10} = \frac{1200 + 570}{30} = \frac{1770}{30} = 59$$

Thus, the average no. of files in computer with 30 folders is 59.

Weighted Arithmetic Mean

One of the limitations of the A.M. discussed above is that it gives equal importance to all the items. But there are cases where the relative important of the difference items is not the same. When this is so, we compute weighted arithmetic mean. The term weight stands for the relative importance of the different items. The formula for calculating weighted arithmetic mean, denoted by$\bar{X}_W$ is given by

$$\bar{x}_W = \frac{\Sigma wx}{\Sigma w}$$

Where X = value of variable i.e. $x_1, x_2, \ldots x_n$
w = weight attached to variable values i.e. $w_1, w_2, \ldots w_N$ respectively.

Example 12. Six internet service provider provides internet volume with price are as follows in certain period of time;

Internet Service Provider	Volume(GB)	Priced/GB(Rs)
Vianet	1800	80
Subisu	77000	70
Wimax	24100	60
Classic Tech	16000	50
Worldlink	1000	45
Broadlink	500	40

Scanned by CamScanner

Find out the average price per GB for internet service provider.

Solution

Calculation of weighted A.M.

Here $\bar{x}_w = \frac{\Sigma wx}{\Sigma w}$

Where, $\bar{x}_w$ = Weighted A.M., x=Price per GB; W = Weight

Internet Service Provider	Volume(w)	Price/GB(x)	wx
Vianet	1,800	80	144,000
Subisu	77,000	70	5,390,000
Wimax	24100	60	144,6000
Classic Tech	16,000	50	800,000
Worldlink	1,000	45	45,000
Broadlink	500	40	20,000
	Σw = 120,400		Σwx = 78,45,000

We have

$$\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{7845000}{120400} = \text{Rs } 65.157.$$

Therefore, the average price per GB is Rs 65.157.

Example 13. Comment on the performance of the students of the three universities given below using simple and weighted average:

University course of study	TU		Pokhara		Purvanchal	
	Pass %	No of student (in 0'0)	Pass %	No. of student (in 0'0)	Pass %	No of student (in 0'0)
MA	71	3	82	2	1	2.0
MBS	83	4	76	3	76	3.5
BA	73	5	73	6	74	4.5
BBS	74	2	76	7	58	2.0
BSc	65	3	65	3	70	7.0
MSc	66	3	60	7	73	2.0

Solution

Calculation simple and weighted A.M.

University Course of study	TU			Pokhara			Purvanchal		
	Pass %	No. of Stud. (in 000)		Pass %	No. of Stud. (in 000)		Pass %	No. of Stud. (in 000)	
	x	w	wx	x	w	wx	x	w	wx
MA	71	3	213	82	2	164	81	2	162
MBS	83	4	332	76	3	228	76	3.5	266
BA	73	5	365	73	6	438	74	4.5	333
BBS	74	2	148	76	7	532	58	2	116
B SC	65	3	195	65	3	195	70	7	490
M SC	66	3	198	60	7	420	73	2	146
Total	Σx = 432	Σw = 20	Σwx = 1451	Σx = 432	Σw = 28	Σwx = 1977	Σx = 432	Σw = 21	Σwx =1513

Scanned by CamScanner

For TU: $\bar{x} = \frac{\Sigma x}{n} = \frac{432}{6} = 72$ $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{1451}{20} = 72.55$

For Pokhara: $\bar{x} = \frac{\Sigma x}{n} = \frac{432}{6} = 72$ $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{1977}{28} = 70.61$

For Purwanchal: $\bar{x} = \frac{\Sigma x}{n} = \frac{432}{6} = 72$ $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{1513}{21} = 72.05$

The arithmetic mean is the same for all the three universities. Hence, it shows that the performance of students is a like. But this will be a wrong conclusion because what we should compare here is the weighted AM on comparing the weighted AM we find that for TU the mean value is the highest and hence we can say that the performance of students is best in TU.

Merits and Demerits of AM

AM is most widely used in practice because of the following reason.

Merits

(i) It is the simplest average to understand and easiest to compute.
(ii) It is based on each and every observation.
(iii) It is rigidly defined.
(iv) It is suitable for further algebraic treatments.
(v) It is a calculated value and not based on position in the series.

Demerits

(i) It is affected by very small and very large items.
(ii) It is not suitable in case of open end classes.
(iii) It cannot be obtained by inspection or by graphical method.
(iv) It cannot be used in a qualitative analysis e.g., knowledge, intelligence, beauty etc.

Geometric Mean

Geometric mean of N non-negative and non-zero items is defined as the Nth root of their product. If there are two non-negative items, the geometric mean is the square roots of their product and if there are three non-negative items, the cube root of their product gives geometric mean.

Let $x_1, x_2, x_3 \ldots x_n$ be N non zero and non-negative values (items) of the series, the geometric mean of the these items, denoted by G.M. is given by,

$$GM = (x_1x_2x_3 \ldots x_n)^{1/n}$$

When number of items is three or more we used logarithm for calculation of G.M.

Taking log both sides in (1)

$$\log GM = \frac{1}{n}\log(x_1x_2x_3 \ldots\ldots x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \log x_3 + \ldots\ldots + \log x_n)$$

$$= \frac{1}{n}\Sigma \log x$$

$$GM = \text{antilog}\left[\frac{1}{n}\Sigma \log x\right]$$

Example 14. Following are discount rate (%) offered by different makes of computer in CAN InfoTech.

Computer makes	Dell	Lenovo	Acer	Samsung	Sony	Apple
Discount rate(%)	15	17	18	12	10	8

Find the average rate of discount.

Scanned by CamScanner

Solution: Calculation of G. M.

x	15	17	18	12	10	8	
logx	1.1761	1.2304	1.2553	1.0792	1.000	0.9031	Σlogx = 6.6441

$$GM = \text{antilog}\left[\frac{1}{n}\Sigma\log x\right] = \text{antilog}\left(\frac{6.6441}{6}\right) = 12.8.$$

Thus average rate of discount is 12.8%.

Calculation of G.M. (Discrete Series)

If $f_1, f_2, f_3 \ldots f_n$ be the corresponding frequencies of n non-zero and non-negative values $x_1, x_2, x_3 \ldots x_n$ then their geometric mean denoted G.M. is given by,

$$GM = \text{antilog}\left[\frac{1}{n}\Sigma f\log x\right]$$

Continuous Series (Group Data)

$$GM = \text{antilog}\left[\frac{1}{n}\Sigma f\log x\right]$$

Wherem = midpoint (value) of the class interval.

Example 15. Find the average profit percentage per shop from the selling of Dell computers from following data.

Profit (%)	4–8	8 – 12	12–16	16-20	20–24	24–28	28–32	32–36	36–40
No of shop	6	10	18	30	15	12	10	6	2

Solution

Calculation of G.M.

Profit (%)	No of shop (f)	Mid point (m)	Log m	f log m
4 – 8	6	6	0.7782	4.6692
8 – 12	10	10	1.0000	10.0000
12 – 16	18	14	1.1461	20.6298
16 – 20	30	18	1.2533	37.6590
20 – 24	15	22	1.3424	20.1360
24 – 28	12	26	1.4150	16.9800
28 – 32	10	30	1.4771	14.7710
32 – 36	6	34	1.5315	9.1890
36 – 40	2	38	1.5798	3.1596
	N = 109			Σf logm=137.193

$$GM = \text{antilog}\left[\frac{1}{n}\Sigma f\log m\right] = \text{antilog}\left[\frac{137.1936}{109}\right] = 18.14.$$

Thus average profit per shop = 18.14%.

Weighted Geometric Mean

In situations in which all items are not of equally importance then we attach weight. The t weight refers the relative importance given to various items.

Let, observation x_1 takes w_1 weight, observation x_2 takes w_2 weight observation x_3 take weight,..... observation x_n takes w_n weight then the weighted geometric mean denote GM_w is given by

Scanned by CamScanner

$$GM = (x_1^{w_1} \times x_2^{w_2} \times x_3^{w_3} \times \ldots\ldots\ldots\ldots\ldots x_n^{w_n})^{1/\sum_{i=1}^{n} w_i}$$

$$\text{Or } GM = \text{Antilog}\left(\frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \log x_i\right)$$

Example 16. Following data represents interest rate of a bank in fixed deposit according to amount deposited.

Amount (million Rs)	5	10	15	20
Interest rate(%)	6	6.5	7	8

Find average rate of interest.

Solution

Amount(w)	Interest rate (x)	logx	wlogx
5	6	0.778	3.89
10	6.5	0.812	8.12
5	7	0.845	4.225
20	8	0.903	18.06
$\sum w = 40$			$\sum wlogx = 34.295$

Average rate of interest

$$GM_w = \text{Antilog}\frac{\sum \text{wlogx}}{\sum \text{w}}$$

$$= \text{Antilog}\frac{34.295}{40} = \text{Antilog } 0.857 = 7.2$$

Hence, average rate of interest = 7. 2%

Combined geometric mean

If sizes and geometric means of two component series are known then we can find the combined geometric mean of these series.

If n_1 and n_2 are sizes and GM_1 and GM_2 are their respective arithmetic means of two series then combined arithmetic mean is $GM_{12} = \text{Antilog}\left(\frac{n_1 \log GM_1 + n_2 \log GM_2}{n_1 + n_2}\right)$

Example 17. Two series of sizes 10 and 12 have geometric means of 7.89 and 6.47 respectively, find combined geometric mean

Solution

Here; $n_1 = 10$, $n_2 = 12$, $GM_1 = 7.89$, $GM_2 = 6.47$

$$GM_{12} = \text{Antilog}\left(\frac{n_1 \log GM_1 + n_2 \log GM_2}{n_1 + n_2}\right)$$

$$= \text{Antilog}\left(\frac{10 \times \log 7.89 + 12 \times \log 6.47}{10 + 12}\right)$$

$$= \text{Antilog}\left(\frac{10 \times 0.897 + 12 \times 0.81}{22}\right)$$

$$= \text{Antilog } (0.849) = 7.07$$

Hence, combined GM = 7.07

Scanned by CamScanner

Merits of GM

(i) It is based on each and every item of the series.

(ii) It rigidly defined.

(iii) It is suitable average while dealing with ratios and percentages and in determining rates increase and decrease.

(iv) It is capable of further algebraic manipulation. For example, if the GM of two or more series a their number of item is known, a combined GM can be easily calculated by apply the formula

$$GM_{12} = \text{antilog}\left\{\frac{n_1 \log GM_1 + n_2 \log GM_2}{n_1 + n_2}\right\}.$$

Demerits of GM

(i) It is difficult to understand.

(ii) It is difficult to calculate and to interpret.

(iii) It cannot be computed when there is zero or negative values.

Harmonic Mean

It is the reciprocal of the A.M. of the reciprocal of the individual items. Thus, by definition

$$HM = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots\ldots\ldots + \frac{1}{x_N}} = \frac{n}{\Sigma\frac{1}{x}}$$

Where $x_1, x_2, x_3, \ldots\ldots x_n$ are non zero value of variable (items).

In short: No. of item divided by sum of the reciprocal of the values of variable is known as Harmonic mean.

Calculation of Harmonic Mean

The formulae of computing HM are:

(i) Individual series $HM = \frac{n}{\Sigma\frac{1}{x}}$.

(ii) Discrete series $HM = \frac{N}{\Sigma f\frac{1}{x}}$

(iii) Group Series (Continuous Series): $HM = \frac{N}{\Sigma f\frac{1}{m}}$

Where m = mid-point of the class interval.

Example 18. Find average from the following:

a) Speed of port in Gbps: 6,14,21,36

b)

Speed	10	20	25	40	50
Number	20	30	50	15	5

c)

Speed	60-62	63-65	66-68	69-71	72-74
Number	15	118	142	127	18

Solution

a. Calculation of HM

$$HM = \frac{n}{\Sigma\frac{1}{x}} = \frac{4}{\frac{1}{6} + \frac{1}{14} + \frac{1}{21} + \frac{1}{30}} = \frac{4}{\frac{35 + 15 + 10 + 7}{210}} = \frac{210 \times 10}{67} = 12.54$$

Hence, average speed is 12.54 Gbps.

Scanned by CamScanner

b. Calculation of HM

Speed(X)	Number (f)	1/X	f/X
10	20	1/10	20/10 = 2.000
20	30	1/20	30/20 = 1.500
25	50	1/25	50/25 = 2.000
40	15	1/40	15/40 = 0.375
50	5	1/50	5/10 = 0.100
	N = Σf = 120		Σf/X = 5.975

$$\text{H.M.} = \frac{N}{\Sigma(1/x)} = \frac{120}{5.975} = 20.08$$

Thus average speed is 20.08.

Solution

c. Calculation of HM

Speed	Number (f)	Mid value(m)	1/m	$f.\frac{1}{m}$
60 – 62	15	61	1/61	15/73 = 0.24590
63 – 65	118	64	1/64	118/6 = 1.84375
66 – 68	142	67	1/67	142/67 = 2.11940
69 – 71	127	70	1/70	127/70 = 1.81429
72 – 74	18	73	1/73	18/73 = 0.24657
	N = 420			$\Sigma f\frac{1}{m} = 6.26991$

$$\text{H.M.} = \frac{N}{\Sigma f\frac{1}{m}} = \frac{620}{6.26991} = 66.987$$

Thus average speed is 66.987.

Weighted harmonic mean

Let, observation x_1 takes w_1 weight, observation x_2 takes w_2 weight observation x_3 takes w_3 weight,..... observation x_n takes w_n weight then the weighted geometric mean denoted by HM_w is given by

$$HM_w = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Example 19. A person travels a distance of 10 km by a bus at speed of 25 km/hr, after then travels a distance of 5 km by taxi at speed of 38 km/hr and a distance of 8 km by motorbike at speed of 45 km/hr. Find average speed of entire trip.

Solution

vehicle	Speed in km/hr (x)	Distance in km (w)	w/x
Bus	25	10	0.4
Taxi	38	5	0.1315
Motorbike	45	8	0.1777
		Σw = 23	Σw/x = 0.709

Scanned by CamScanner

$$\text{Average speed } (HM_w) = \frac{\sum w}{\sum w/x}$$

$$= \frac{23}{0.709} = 32.44$$

Hence, average speed of entire trip = 32.44 km/hr.

Merits of HM

(i) Its value is based on every item of the series.

(ii) It tends itself to algebraic manipulation.

(iii) It gives better results than other averages in problem relating to time and rates.

Demerits of HM

(i) It is not easily understood.

(ii) It is not very useful for the analysis of economic data.

(iii) Its value cannot be computed when there are both positive and negative items in a series or when one or more items are zero.

(iv) It is difficult on calculation.

Relationship between Averages

In any distribution when the original items differ in size, the value of AM, GM and HM would also differ and will be in the following order: AM ≥ GM ≥ HM i.e. AM is greater than GM and GM is greater than HM. The equality signs hold only if all the number $X_1, X_2 .....X_n$ are identical.

Median

The middle of a set of numbers is the median. It divides an ordered set into two equal parts half (50%) of the number will have values less than the median and half (50%) will have values greater than the median. Thus, if data are arranged according to their magnitude in an ascending or a descending order an (array) the median value is half way through it. The middle most value of the data which arrange in ascending or descending order is known as median.

For example, if we have 7 student with marks 28, 31, 27, 45, 43, 52, 38, median of the marks obtained by those students can be found by ranking the marks and taking the 4th figure in the order of magnitude. The mark thus ordered will be 27, 28, 31, 38, 43, 45, 52 and the median will be 38. If the observations are even number say 10, the median value is the average of the 5th and 6th observations in the order of magnitude. Thus if the annual rainfall of a city, arranged in an array are Rs. 28, 29, 30, 31, 32, 33, 33.50, 34, 34.50 and 35 then the 5th and 6th values in the order of their magnitude will be Rs. 32 and Rs. 33 respectively. Median is a positional value denoted by Med. Their median will be Md =(32+33)/2 = 32.50.

Calculation of Median-Individual Data

Steps

(i) Arrange the data in ascending or descending order of magnitude

(ii) Find number of the items in the data

(iii) Find the position of median by using the formula size

Thus, item give the value of median.

Note: If n is an odd no. the median is the middle value. If n is even number the median is the average of the two middle values.

Scanned by CamScanner

Example 20. Obtain the value of median from the following data:

(a) 39, 38, 59, 40, 67, 52, 77, 75, 24

(b) 391, 384, 591, 407, 672, 522, 777, 753, 2488, 1490

Solution

(a) Calculation of Median

Arranging the data in ascending order, we get

24, 38, 39, 40, 52, 59, 67, 75, 77, Here n = 9

Size of Md. $= \frac{n+1}{2} = \frac{9+1}{2} = 5^{th}$

Md. = 5^{th}item = 52

(b) Calculation of Md.

Arranging the data in ascending order we get

348, 391, 407, 522, 591, 672, 753, 777, 1490, 2488

Here, n= 10

Size of Md. $= \frac{n+1}{2} = \frac{10+1}{2} = 5.5^{th}$item

Md. = 5^{th}item + 0.5 (6th item - 5th item)

= 591 + 0.5 (672-591) = 591 + 40.5 = 631.5

Alternatively, Md$= \frac{5^{th} \text{ item} + 6^{th} \text{ item}}{2} = \frac{591 + 672}{2} = 631.5$

Calculation of Median- Discrete Data

Step

(i) Arrange the data in ascending order of magnitude

(ii) Find out cumulative frequencies (cf.)

(iii) Find N = (Σf).

(iv) Apply the formula Size of Md. $= \frac{N+1}{2}$

(v) Now look at the cumulative frequency column and find that total which is either equal to $\frac{N+1}{2}$ or next higher to that and determine the value of the variables corresponding to it. That gives the value of median.

Example 21. Obtain the median income from the data given below:

Income (Rs.)	1000	1500	800	2000	2500	1800
No. of person	24	26	16	20	6	30

Solution

Calculation of Median

Income (arranging in ascending order)	No of person (f)	c.f.
800	16	16
1000	24	40
1500	26	66
1800	30	96
2000	20	116
2500	6	122
	N = 122	

Scanned by CamScanner

Size of Md. $= \frac{N+1}{2}$

$= \frac{122+1}{2} = 61.5^{th}$ item.

The c.f. just greater than 61.5 is 66.

Hence, Md.income = Rs. 1500.

Calculation of Median-Continuous Series

Step

(i) Arrange the data in ascending order

(ii) Find out cumulative frequencies (cf.)

(iii) Find N = (Σf)

(iv) Apply the formula Size of Md. $= \frac{N}{2}$

(v) Now look all the cumulative frequency column and find that total which is either equal to $\frac{N}{2}$ or next higher to that and determine the class of the variable for exact value of Med. and use the following formula:

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h$$

Where

L = Lower limit of median class

cf = Cumulative frequency of the class preceding the median class

f = Simple frequency of the median class

h = The range of median class

Example 22. Calculate the Median height for the following frequency distribution.

a)

Height (in cm)	161-167	167-173	173-179	179-185	185-187	187-197
No. of students	79	92	60	22	5	2

b)

Height (in cm)	60-62	63-65	66-68	69-71	72-74
No. of students	25	108	142	127	18

Solution

(a) Calculation of Md

Height	No. of students(f)	c.f.
161 – 167	79	79
167 – 173	92	171
173 – 179	60	231
179 – 185	22	253
185 – 191	5	258
191 – 197	2	260
	N = 260	

Size of Md $= \frac{N}{2} = \frac{260}{2} = 130.$

The c.f. just greater than 130 is 171

Hence, Median lies in class (167-173)

Scanned by CamScanner

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h = 167 + \frac{130 - 79}{92} \times 6 = 170.32$$

Thus, Median height is 170.32 cm

(b) Calculation of Md

Height	No of student(f)	c.f.
60 – 62	25	25
63 – 65	108	133
66 – 68	142	275
69 – 71	127	402
72 – 74	18	420
	N = Σf = 430	

Size of $M_d = \frac{N}{2} = \frac{420}{2} = 210.$

The cf just greater than 210 is 275.

Hence, the M_d lies in class 66-68

But real class is 65.5 – 68.5 (in exclusive type)

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h = 65.5 + \frac{620 - 133}{108} \times 3 = 67.13 \text{ cm}$$

Thus, median height is 67.13cm

Note:

(i) The calculation of Med. data should be exclusive. Here data is inclusive, we should convert it to the exclusive one by deducting 0.5 from the lower limit and adding 0.5 to the upper limit

(ii) No need to change the all classes into exclusive only change the class where median lies.

Calculation of Median in Case of Open end Classes

In such a case no need to write the lower limit of the 1st class and upper limit of the last class. Find median using same rule as we have used in group data. For that reason this median is suitable in case of open end classes.

Example 23. Find the average of the data given below:

Marks	Below 10	10-20	20-30	30-40	40-50	50 and above
No. of students	4	6	10	15	8	7

Solution

Calculation of Median

Marks (X)	No of student(f)	c. f.
Below 10	4	4
10 – 20	6	10
20 – 30	10	20
30 – 40	15	35
40 – 50	8	43
50 and above	N = ∑f = 7	50

Size of $Md = \frac{N}{2} = \frac{50}{2} = 25$. The cf just greater than 25 is 35.

Hence, Median lies in class 30 - 40

Scanned by CamScanner

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h = 30 + \frac{25 - 20}{15} \times 10 = 33.33$$

∴ Median marks =33.33.

Calculation of Median When Class Intervals are Unequal

When the class intervals are unequal, the frequencies need not be adjusted to make the class intervals equal and the same formula for interpolation can be applied as discussed below:

Example 24. Calculate median marks from the following data:

Marks	0 – 10	10 – 30	30 – 60	60 – 80	80 – 90
No of students	5	15	30	8	2

Solution

Since class intervals are unequal let us first convert it to a distribution with equal class intervals on the assumption that the frequencies equally distributed throughout a class.

Marks	No of students (f)	c. f.
0 – 10	5	5
10 – 30	15	20
30 – 60	30	50
60 – 80	8	58
80 – 90	2	60
	N =∑f = 60	

Size of Md= $\frac{N}{2} = \frac{60}{2} = 30.$

The c.f. just greater than 30 is 50

Hence, median lies in class (30 – 60)

$$Md = L + \frac{\frac{N}{2} - cf}{f} \times h = 30 + \frac{30 - 20}{30} \times 20 = 40$$

Alternative method:

If we make the class intervals equal the same answer will be obtained.

Marks	No of students (f)	c.f.
0 – 10	5	5
10 – 20	7.5	12.5
20 – 30	7.5	20
30 – 40	10	30
40 – 50	10	40
50 – 60	10	50
60 – 70	4	54
70 – 80	4	58
80 – 90	2	60
	N =Σf = 60	

Size of Md= $\frac{N}{2} = \frac{60}{2} = 30$. The c.f. just greater than or equals to 30 is 30.

Hence,Median lies in class 30 - 40

Scanned by CamScanner

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h = 30 + \frac{30 - 20}{10} \times 10 = 40$$

Thus Md marks = 40.

Merits of Median

(i) It is especially useful in case of open end classes. The median is also recommended if the distribution has unequal classes.

(ii) Extreme values do not affect the median.

(iii) It is the most appropriate average in dealing with qualitative data.

(iv) Median can be graphically obtained.

(v) It is rigidly defined.

(vi) It is simple to calculate and easy to remember.

Demerits of Median

(i) It is necessary to arrange the data.

(ii) It is not based on each and every item.

(iii) It is not capable of algebraic treatment.

(iv) If number of items is even, median cannot be computed exactly.

(v) It is erratic if the number of item is small.

Partition Values

Besides median, there are other measures which divide a series into more than two equal parts. Important amongst these are quartiles, deciles and percentiles.

Quartiles: Quartiles are those values of the data which divide the series into four equal parts. There are three quartiles, known as lower quartile, Median and upper quartile, denoted by Q_1, Q_2 (med) and Q_3 respectively. Lower quartile Q_1 is the value of the data which has 25 percent of the values of the data below it and 75 percent of the values are greater than it. Second quartile Q_2, also known as med is the value for which 50 percent of the values of the data will be below it and 50 percent of the values will above it. Upper quartile Q_3 has 75 percent of the values below it and 25 percent of the value above it.

Deciles: Deciles are the those values of the data which divide the total frequency in ten equal parts. There are 9 deciles denoted by $D_1, D_2, \ldots D_9$letters. Here $D_1 < D_2 < D_3 < D_4 < D_5 \ldots < D_9$.

Percentile: Percentiles are those values of the data which divide the total frequency into hundred equal parts. There are 99 percentiles denoted by $P_1, P_2 \ldots P_{99}$ letter. Here $P_1 < P_2 < P_3 \ldots < P_{99}$.

Here ($M_d = Q_2 = D_5 = P_{50}$).

Calculation of Quartiles, Deciles and Percentiles

Case I: Individual Data

Steps

(i) Arrange the data according to their magnitude in ascending order.

(ii) Find No. of item (n)

(iii) Apply the formula for Calculation of these value.

(a) Size of $Q_i = \left(\frac{n+1}{4}\right) \times i$, where $i = 1, 2, 3$

Scanned by CamScanner

Thus, $\frac{n+1}{4}$ i^{th} item, give the value of Q_i

(b) Size of $D_i = \left(\frac{n+1}{10}\right) \times i$ where i= 1, 2, ... 9

Thus, $\frac{n+1}{10}$ i^{th} item, give the value of D_i

(c) Size of $P_i = \left(\frac{n+1}{100}\right) \times i$ where i = 1, 2, ... 99

Thus, $\frac{n+1}{100}$ i^{th} item, give the value of P_i.

Example 25. Find Q_1, D_7 and P_{60} from the following data:

a) 20, 12, 8, 17, 22, 25, 30

Solution

a) Calculation of Q_1, D_7, P_{60}

Arrange the data in ascending order

8, 12, 17, 20, 22, 25, 30

(i) Size of $Q_1 = \frac{n+1}{4} \times 1 = \frac{7+1}{4} \times 1 = 2^{nd}$ item = 12

(ii) Size of $D_7 = \frac{n+1}{10} \times \frac{7+1}{10} \times 7 = 5.6^{th}$ item

$D_7 = 5^{th}$ item + 0.6 (6th – 5th item)=22 + 0.6 (25 –22)= 23.8

(iii) Size of $P_{60} = \frac{n+1}{100} \times 60 = \frac{7+1}{100} \times 60 = 4.8^{th}$ item

$P_{60} = 4^{th}$ item + 0.8 (5th – 4th item)=20 + 0.8 (22 – 20) = 21.6

∴ $Q_1 = 12$, $D_7 = 23.8$ and $P_{60} = 21.6$

Case II: Discrete Data

Step

(i) Arrange the data in ascending order of magnitude

(ii) Find out cumulative frequencies (cf)

(iii) Find N (= Σf)

(iv) Apply the formula

a) Size of $Q_i = \frac{N+1}{4} \times i$, where i = 1, 2, 3

b) Size of $D_i = \frac{N+1}{10} \times i$, where i = 1, 2, ... 9

c) Size of $P_i = \frac{N+1}{100} \times i$, where i = 1, 2, ... 99

(v) Now look at the cumulative frequency column and find that the cf which is either equal to or next higher to that size (size determined in step iv) and determine the value of variable corresponding to it. That gives the values of quartiles (deciles or percentiles).

Example 26. Calculate Q_3, D_6, and P_{80} from the data.

Respiratory rate	10	15	20	25	30	35	40	45	50
No of person	8	12	36	25	28	18	9	12	6

Scanned by CamScanner

Solution

Calculation of Q_3, D_6 and P_{80}

Respiratory rate (X)	No of persons (f)	c.f.
10	8	8
15	12	20
20	36	56
25	25	81
30	28	109
35	18	127
40	9	136
45	12	148
50	6	154
	N = 154	

Size of Q_3 = Size of $Q_3 = \frac{N+1}{4} \times 3 = \frac{154+1}{4} \times 3 = 116.25$

The c.f. just greater than 116.25 is 127. $\therefore Q_3 = 35$

Size of $D_6 = \frac{N+1}{10} \times 6 = \frac{154+1}{10} \times 6 = 93$

The c.f. just greater than 93 is 109 $\therefore D_6 = 30$

Size of $P_{80} = \frac{N+1}{100} \times 80 = \frac{154+1}{100} \times 80 = 124$

The c.f. just greater than 124 is 127 $\therefore P_{80} = 35$

Thus, $Q_3 = 35$, $D_6 = 30$ and $P_{80} = 35$

Case III: Continuous Series

Steps

(i) Arrange the data in ascending order

(ii) Find out cumulative frequencies (cf)

(iii) Find N (Σf)

(iv) Apply the formula

a) Size of $Q_i = \frac{N}{4} i$, where i = 1, 2, 3

b) Size of $D_i = \frac{N}{10} i$, where i = 1, 2, ... 9

c) Size of $P_i = \frac{N}{100} i$, where i= 1, 2, ... 99

(v) Now look at the cumulative frequency column and find that total which is either equal to $\frac{N}{4} i$, $\frac{N}{10} i$, $\frac{N}{100} i$ or next higher to that and determine the Class of variable. For exact value of the quartiles, deciles and percentiles used the following formula

a) i^{th} Quartile $\quad Q_i = L + \frac{\frac{N}{4} i - cf}{f} \times h$

Scanned by CamScanner

b) i^{th} Deciles $\quad D_i = L + \dfrac{\frac{N}{10} i - cf}{f} \times h$

c) i^{th} Percentile $\quad P_i = L + \dfrac{\frac{N}{100} i - cf}{f} \times h$

Where

L = Lower limit of quartile, decile or percentile classes

cf = Cumulative frequency of class preceding the quartile, decile or percentile class

f = frequency of the quartile, decile or percentile class

h = the interval of the quartile, decile or percentile class

Example 27. From the data given below calculate the value of Q_1, Q_3, D_2, D_9 and P_6 and P_{57}.

Height (cm)	161 – 167	167 – 173	173 – 179	179 – 185	185 – 191	191 – 197
No of persons	79	92	60	22	5	2

Solution

Calculation of Q_1, Q_3, D_2, D_9, P_{45} and P_{67}

Height (cm)	No of person (f)	c f
161 – 167	79	79
167 – 173	92	171
173 – 179	60	231
179 – 185	22	253
185 – 191	5	258
191 – 197	2	260
	N = ∑f = 260	

Size of $Q_1 = \frac{N}{4} = \frac{260}{4} = 65^{th}$item

The c.f. just greater than 65 is 79. Q_1 lies in class (161-167)

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h = 161 + \frac{65 - 0}{79} \times 6 = 165.93$$

Size of $Q_3 = \frac{N}{4} \times 3 = \frac{260}{4} \times 3 = 195^{th}$item

The c.f. just greater than 195 is 231. Q_3lies in class(173 – 179)

$$Q_3 = L + \frac{\frac{N}{4} 3 - cf}{f} \times h = 173 + \frac{195 - 171}{60} \times 6 = 175.40$$

Size of $D_2 = \frac{N}{10} \times 2 = \frac{260}{10} \times 2 = 52^{th}$item

The c.f. just greater than 52 is 79.D_2 lies in class (161-167)

$$D_2 = L + \frac{\frac{N}{10} 2 - cf}{f} \times h = 161 + \frac{\frac{52 - 0}{79} - cf}{f} \times 6 = 164.95 \text{ cm}$$

Size of $D_9 = \frac{N}{10} \times 9 = \frac{260}{10} \times 9 = 234^{th}$ item

Scanned by CamScanner

The c.f. just greater than 234 is 253. D_9 lies in class (179-185)

$$D_9 = L + \frac{\frac{N}{10}9 - cf}{f} \times h = 179 + \frac{234 - 231}{22} \times 6 = 179.82$$

$\therefore$ $D_9 = 179.82$ cm

Size of $P_{45} = \frac{N}{100} \times 45 = \frac{260}{100} \times 45 = 117^{th}$ item

The c.f. just greater than 117 is 171. P_{45} = Lies in class (167-173)

$$P_{45} = L + \frac{\frac{N}{100}45 - cf}{f} \times h = 161 + \frac{117 - 79}{92} \times 6 = 169.48$$

$\therefore$ $P_{45} = 169.48$ cm

Size of $P_{57} = \frac{N}{100} \times 57 = \frac{260}{100} \times 57 = 148.2^{th}$ item

The c.f. just greater than 148.2 is 171. P_{57} lies in class (167-173)

$$P_{57} = L + \frac{\frac{N}{100}57 - cf}{f} \times h = 161 + \frac{148.2 - 79}{92} \times 6 = 171.51$$

$\therefore$ $P_{57} = 171.51$ cm

Thus,

$Q_1 = 165.93$ cm, $Q_3 = 175.40$ cm, $D_2 = 164.95$ cm

$D_9 = 179.82$ cm, $P_{45} = 169.48$cm and $P_{57} = 171.51$cm

Example 28. The following table give the frequency distribution of the marks of 800 candidates in a class test.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No of students	10	40	80	140	170	130	100	70	40	20

(i) If 30% of the students fails, find the highest mark obtained by fail students.

(ii) If the minimum marks required for passing are 35, what percentage of the students pass the test.

(iii) Find the limits of marks of central 30% student and range marks of central 30% students.

(iv) Find the percentage of the students who get the mark between 75 and 95.

(v) Find the percentage of students who get marks less than 55.

(vi) Find the median of the marks.

Solution

Calculate of median and percentage

Marks	No. of students(f)	cf
0-10	10	10
10-20	40	50
20-30	80	130
30-40	140	270
40-50	170	440
50-60	130	570
60-70	100	670
70-80	70	740
80-90	40	780
90-100	20	800
	N = 800	

Scanned by CamScanner

(i) The highest mark obtained by fail 30% of the students is given by P_{30}

Size of $P_{30} = \frac{30\,N}{100} = \frac{30 \times 900}{100} = 240^{th}$ item

Thus, P_{30} lies in class (30 – 40)

$$P_{30} = L + \frac{\frac{30\,N}{100} - cf}{f} \times h = 30 + \frac{240 - 130}{140} \times 10 = 37.86$$

(ii) No. of students who secured more than 35 marks

$$= \frac{40 - 35}{10} \times 140 + 170 + 130 + 100 + 70 + 40 + 20 = 600$$

Percentage $= \frac{600}{800} \times 100 = 75\%$

(iii) The limit of marks of central 30% students a given by P_{35} to P_{65}

Size of $P_{35} = \frac{35}{100} = \frac{35 \times 800}{100} = 280^{th}$ item

Thus P_{35} lies in class (40 – 50)

$$P_{35} = L + \frac{\frac{35\,N}{100} - cf}{f} \times h = 40 + \frac{280 - 270}{170} \times 10 = 40.58$$

Size of $P_{65} = \frac{65N}{100} = \frac{65 \times 800}{100} = 520^{th}$ item

Thus P_{65} lies in class (50-60)

$$P_{65} = L + \frac{\frac{65\,N}{100} - cf}{f} \times h = 50 + \frac{520 - 440}{130} \times 10 = 56.15$$

The limit of marks of central 30% student is P_{35} to P_{65} is 40.58 to 56.15.

The range marks of central 30% student = $P_{65} - P_{35} = 56.15 - 40.58 = 15.57$

(iv) Number of student who got the marks greater than 75 and less than 95.

$$= \frac{80 - 75}{10} \times 70 + 40 + \frac{95 - 90}{10} \times 20 = 85$$

Percentage age $= \frac{85}{800} \times 100 = 10.625\ \% = 10.63\%$

(v) No. of students who get the marks less than 55

$$= \frac{55 - 50}{10} \times 130 + 170 + 140 + 80 + 40 + 10 = 505$$

Percentage age $= \frac{505}{800} \times 100 = 63.125\ \% = 63.13\ \%$

(vi) Size of md. $= \frac{N}{2} = \frac{800}{2} = 400^{th}$ item

∴ Median lies in class 40 - 50

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h = 40 + \frac{400 - 270}{170} \times 10 = 47.65$$

Determination of Partition Values Graphically

Median can be determined graphically by applying any one of the following two methods:

Scanned by CamScanner

1. **Draw two ogives**

   One by less than method and other by 'more than' method. From the point where both these cures intersect each other draw a perpendicular on x-axis. The point where the perpendicular from the point of intersection touches on x-axis gives the value of median.

2. **Draw only one ogives by less than method**

   Take the variable on x-axis and frequency on the y- axis. Determine the median value by the formula.

   Median = Size of $N/2^{th}$ item. Locate this value on the y- axis and from it draw a perpendicular on the cumulative frequencies curve. From the point where it meets the ogive, draw another perpendicular on x-axis. The point where it meets the x- axis is the median.

   The other position values like quartiles, deciles etc can also be determined graphically.

Example 29. Draw an ogive for the following data: How many worked earned wage between Rs. 1365 and Rs. in 1430? Also locate the median wage.

Wages	No of persons
1000 – 1100	6
1100 – 1200	10
1200 – 1300	22
1300 – 1400	30
1400 – 1500	16
1500 – 1600	14
1600 – 1700	12

Solution:

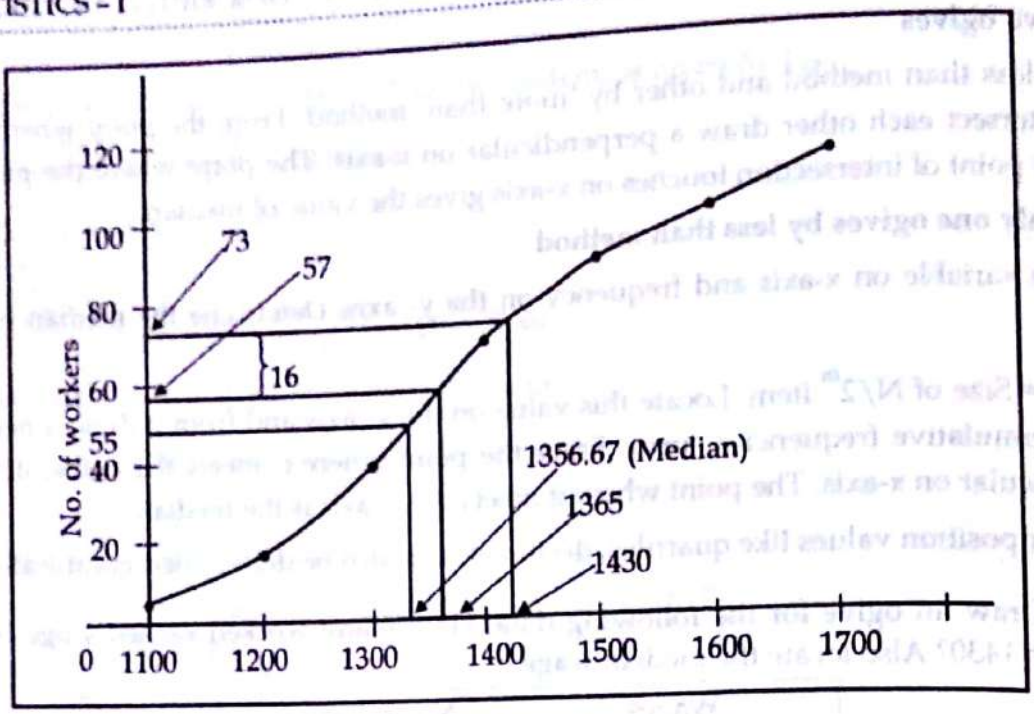
Calculation of Less than ogives

Wages less than Rs.	c.f.
1100	6
1200	16
1300	38
1400	68
1500	84
1600	98
1700	110

No. of worker whose wages are between Rs. 1365 and Rs. 1430. is 16 and median is Rs. 1356.67 shown below.

Workers getting wages between Rs. 1365-1430

Scanned by CamScanner



No. of workers whose wages are less then Rs. 1430 = 73

No. of workers whose wages are less then Rs. 1365 = 57

No. of workers whose wages are between 1365 and 1430=73- 57=16

Size of Md. $= \frac{N}{2} = \frac{110}{2} = 55^{th}$ item

Median value = 1356.67.

Mode

The mode or modal value is that value in a series of observations which occurs maximum number of times. The mode of distribution is the value at the point around which the items tend to be most heavily concentrated. The value having maximum frequency in the data is known as mode. In certain cases maximum frequency may exist in more than one class. In such a case two or more would be the modal classes and the distribution would be bi-modal or multimodal. There are many situations in which A.M. and median fail to reveal the true characteristic of data. For example, when we talk of most common wage, most common income, most common height, most common size of timber we have in mind mode and not the A.M. or Median.

Moreover, mode is the easiest to compute since it is the value corresponding to the highest frequency. For example,

Length of timber	5	6	7	8	9	10	11
No of timber	10	20	25	40	22	15	7

The model size is '8' since it appears maximum number of times in the series.

Calculation of Mode - individual Series

For finding mode find the value occurring maximum number of times, is the modal value.

Example 30. Find the mode from the data of marks obtained by 10 students

Marks 10, 27, 24, 12, 27, 20, 18, 15, 30, 27

Solution

Here the value 27 occurs the maximum no. of times i.e. 3, hence the modal marks are 27.

Note: If the data is bi-modal (two mode) or multi-modal (more than 2 mode), mode is ill defined. Then its value is calculated by a relation known as Empirical relation which is

Mode = 3 Median – 2 Mean

Scanned by CamScanner

Calculation of Mode-discrete Series

In this case mode can be determined just by inspection i.e. by looking to that value of the variable which has maximum frequency.

For Example:

Size of garment	28	29	30	31	32	33	34
No of persons	10	20	40	65	50	15	10

Here the value 31 has maximum frequency so mode = 31.

Note

(i) If the data is bi-model or multimodal, use empirical formula

(ii) By inspection method an error of judgment is possible in those cases where the difference between the maximum frequency and frequency preceding it or succeeding it is very small and the items are heavily concentrated on either side (at the beginning or at the end).

In such cases it is desirable to prepare a grouping table and an analysis table. These tables help us in calculation of modal class (value). The procedure of preparing grouping table and analysis table shall be clear from the following example:

Example 31. Calculate the value of mode for the following data:

Marks	10	15	20	25	30	35	50
No of person	8	12	36	25	28	18	9

Solution

Since it is difficult to say by inspection as to which is the modal value, we prepare grouping and analysis tables.

a) Grouping table

Calculation of Mode

X	I	II	III	IV	V	VI
10	8					
15	12	20		56		
20	(36)		48		(83)	
25	35	(71)				(99)
30	28		(63)	(81)		
35	18	46			55	
40	9		27			

Explanation:

Column I:	f_1, f_2, f_3, f_4.....	(Individual Frequency)
Column II:	f_1+f_2, f_3+f_4	(Add two)
Column III:	f_2+f_3, f_4+f_5	(Leave 1 add two)
Column IV:	$f_1+f_2+f_3, f_4+f_5+f_6$	(Add 3)
Column V:	$f_2+f_3+f_4, f_3+f_4+f_5$.....	(Leave 1 and add 3)
Column VI:	$f_3+f_4+f_5, f_6+f_7+f_8$.....	(Leave 2 and Add 3)

Scanned by CamScanner

b) **Analysis table**

Column	Value (X)						
	10	15	20	25	30	35	40
I			1				
II			1	1			
III				1	1		
IV				1	1	1	
V		1	1	1			
VI			1	1	1		
Total		1	4	5	3	1	

Corresponding to maximum total 5, the value of variable is 25. Hence the modal value is 25 marks.

Calculation of mode-Continuous Series

Steps:

(i) Find the modal class by inspection or by preparing grouping and analysis table

(ii) Use the following formula to get the value of mode:

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

Where,

L = Lower limit of the modal class

f_1 = Frequency of modal class

f_0 =Preceding frequency of modal class (frequency of pre-model class)

f_2 =Succeeding frequency of modal class (frequency of the post modal class)

h = class interval

or $\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$ where $\Delta_1 = f_1 - f_0$, $\Delta_2 = f_1 - f_2$

If any of Δ_1 or Δ_2 is negative then

$$\text{Mode} = L + \frac{|\Delta_1|}{|\Delta_1| + |\Delta_2|} \times h$$

Note:

(i) If mode is ill-defined. Its value may be obtained by empirical formula, i.e, Mode = 3 Median – 2 Mean

(ii) Class should be of equal range and exclusive.

Example 32. Calculatemode from the data:

Marks	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No of students	9	18	31	17	16	9

Solution: Calculation of mode

Marks	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No of students	9	18	31	17	16	9

By Inspection Mode lies in class 30 - 40

$L = 30, f_1 = 31, f_0 = 18, f_2 = 17, h = 10$

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$= 30 + \frac{31 - 18}{62 - 18 - 17} \times 10 = 30 + \frac{13}{27} \times 10 = 34.81$$

Scanned by CamScanner

Calculations of Mode in case of unequal classes

If the class intervals are unequal, then we must make them equal before we start the computation of mode. The class interval should be made equal and frequencies must be adjusted on the assumption that they are equally distributed throughout the class.

Example 33. Calculate Mode from the data:

Value	0 – 10	10 – 20	20 – 40	40 – 50	50 – 70
Frequency	5	15	40	32	28

Solution

Since the classes are unequal so adjustment is to be made before calculating mode by making class intervals equal.

Value	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50-60	60-70
Frequency	5	15	20	20	32	14	14

By Inspection Mode lies in class 40 – 50

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Here $L = 40, f_1 = 32, f_0 = 20, f_2 = 14, h = 10$

$$= 40 + \frac{32 - 20}{64 - 20 - 14} \times 10 = 40 + 4 = 44$$

$\therefore$ Mode = 44

Calculation of Mode by Graphically

In a frequency distribution the value of mode can also be determined graphically. The steps in calculation are:

(i) Draw a histogram of the given data.

(ii) Draw two lines diagonally in the inside of the modal class bar, starting from each upper corner of the bar to the upper corner of the adjacent bar.

(iii) Draw a perpendicular line from the intersection of the two diagonal lines to the x-axis which gives us the modal value.

Example 34. Draw a histogram for the following distribution and find Mode and check the value by direct calculation.

Life expectancy (months)	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
No of fishes	60	140	110	150	120	100	50

Solution

The histogram of this data is given below locating the mode graphically.

Here, Mode = 27 (By inspection)

Direct Calculation

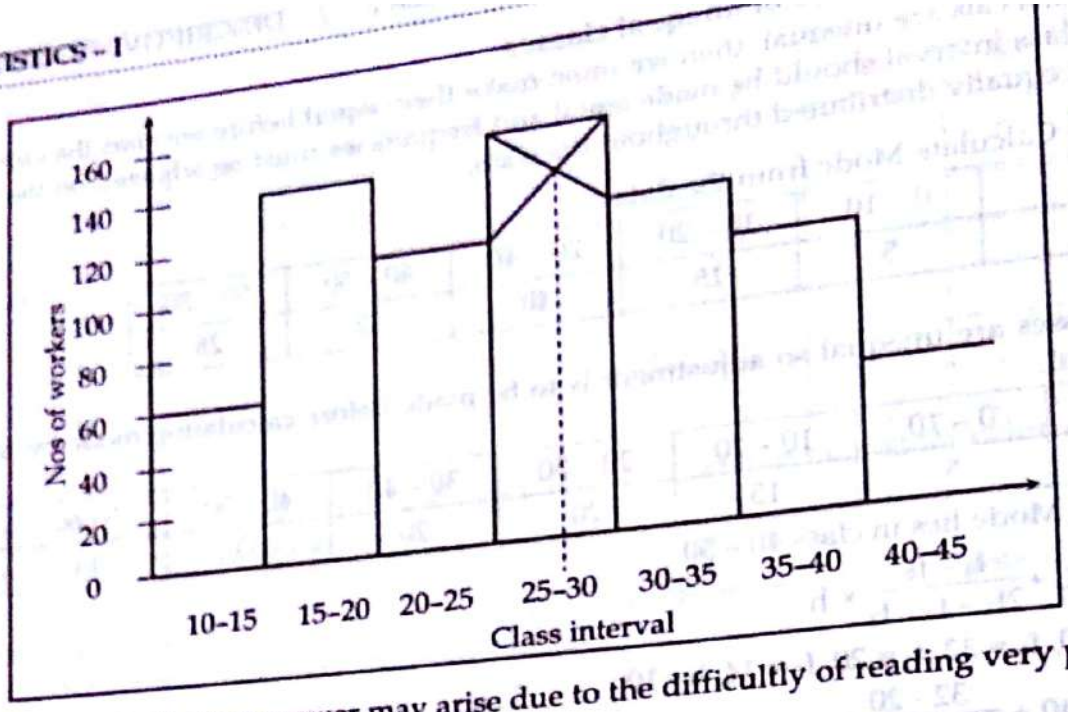
By inspection, mode lies in class (25 – 30)

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h, \quad \text{Where } L = 25, f_1 = 150, f_0 = 110, f_2 = 120, h = 5$$

$$= 25 + \frac{150 - 110}{300 - 120 - 110} \times 5 = 25 + 2.86 = 27.86$$

$\therefore$ Mode = 27.86

Scanned by CamScanner



The slight difference in the two answer may arise due to the difficultly of reading very precisely from the group.

Note:

1. If the distribution is bi-modal or multimodal graphical method fails to be used.
2. In case of Multimodal data, calculating mode using SPSS, it provides result of lowest mode with (*) indication.

Merits of Mode

(i) It is easy to calculate.
(ii) It is not affected by extreme items.
(iii) It can be found even in open end classes.
(iv) It can be determined graphically.

Demerits of Mode

(i) It is not based on all items
(ii) It is not rigidly defined
(iii) It is not capable of further algebraic treatments.
(iv) It is affected to great extent by sampling fluctuations.

Choice of an Average

There are many averages each of them has its own characteristics and usefulness. A right choice of average, to suit the purpose in hand is necessary. If a proper choice is not made, the average may not truly represent the particular characteristic of the data.

The following points may also be noted while choosing an average:

1. A.M. cannot be calculated for open end classes.
2. A.M. may not be a possible value of the variable.
3. G.M and H.M. cannot be calculated if any item is zero.
4. G.M. is not defined if any value is negative.

Arithmetic Mean

A.M. is based on all observations. We use it for all the averages. It is used in quantitative data such as production, income, price exports, imports etc. When the data are properly spread out and do not show wide variations their average is weighted average.

In the following cases A.M. should not be used:

Scanned by CamScanner

1. in highly skewed data.
2. in open end classes.
3. When there are very large and very small items.
4. In case of average ratio and rate of change.

Median

It is best average in open end classes and qualitative, analysis such beauty, knowledge intelligence, honest etc.

Mode

It is also used for open end classes. It is suitable for calculation of average capacity of Hard drive, RAM, shoes, most repeated value.

G.M.

It is suitable in averaging ratios and percentage rate of increase or decrease.

H.M.

It is suitable in averaging rate per unit and speed problems.

Calculation of A.M. in Case of Incorrect Values

In calculation of correct mean in case of incorrect values, the following steps are taken.

(i) Find incorrect Σx using the formulas $\sum x = n\bar{X}$

(ii) Subtract wrong (incorrect) values from wrong ΣX and add correct values.
i.e. correct ΣX= Wrong ΣX-Wrong items + Correct items

(iii) Divide the correct ΣX by n to get the correct $\bar{X}$

Example 35. Modal mark for a group of 94 students is 54. Find the no. of student getting the marks between 20-40 and 60-80. From data if the maximum mark of the examination were 100.

Marks	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100
No of students	10	-	30	-	14

Solution

Let the no. of students getting the marks between 20-40 and 60-80 are, a and b respectively.

Here, N = 94, Mode = 54

$94 = 54 + a + b \quad a + b = 94 - 54 = 40$

$N = 10 + a + 30 + b + 14$

$b = 40 - a$..... (i)

Calculation of missing frequencies a and b

Marks	No of students(f)
0 – 20	10
20 – 40	A
40 – 60	30
60 – 80	b = 40 – a
80 – 100	16
	N = 94

Mode = 54 (given) Lies in class 40-60

Here $L = 40, f_1 = 30, f_0 = a, f_2 = b = 40 - a, i = 20$

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

or, $$54 = 40 + \frac{30 - a}{60 - a - 40 + a} \times 20$$

or, $$54 - 40 = \frac{30 - a}{20} \times 20$$

Scanned by CamScanner

or, $14 = 30 - a$

or $14 - 30 = -a$

$a = 16$

$a = 16$ put in (i) we get, $b = 40 - 16 = 24$

Thus Missing frequencies are 16 and 24.

Example 36. The median and mode of the data are known to be Rs. 33.5 and Rs. 34 respectively. Three frequencies from the data are however, missing. Find these missing frequencies if N = 230

Wages (in Rs.)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No of worker	4	16	-	-	-	6	4

Solution

Let the missing frequencies of the classes 20-30, 30-40 and 40-50 and a, b, and c respectively.

$N = 4 + 16 + a + b + c + 6 + 4$

or $230 = 30 + a + b + c$

or $c = 200 - a - b$ ..... (i)

Calculation a, b and c

Wages in Rs.	Frequency (f)	c.f.
0 – 10	4	4
10 – 20	16	20
20 – 30	a	20 + a
30 – 40	b	20 + a + b
40 – 50	c = 200 – a – b	220
50 – 60	6	226
60 – 70	4	230

Since median and mode are 33.5 and 34, they both lie in the class 30-40

We have

$$Md. = L + \frac{\frac{N}{2} - cf}{f} \times h$$

or, $33.5 = 30 + \frac{115 - (20 + a)}{b} \times 10$

or, $33.5 - 30 = \frac{95 - a}{b} \times 10$

or, $3.5b = (95 - a)10$

or, $950 - 10a = 3.5b$

or, $10a + 3.5b = 950$ ..... (ii)

Again,

$$Mode = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

or, $34 = 30 + \frac{b - a}{2b - a - (200 - a - b)} \times 10$

or, $34 - 30 = \frac{b - a}{2b - a - 200 + a + b} \times 10$

or, $4(3b - 200) = 10b - 10a$

or, $12b - 10b + 10a = 800$

or, $2b + 10a = 800$ ..... (iii)

Scanned by CamScanner

Subtract equation (ii) from equation (iii)

$$2b + 10a = 800 \quad \text{..... (iii)}$$
$$\underline{3.5b \pm 10a = 950} \quad \text{..... (ii)}$$
$$-1.5b = -150$$
$$b = 100$$

Put, b= 100 in (iii) we get,

$$2(100) + 10a = 800$$

or $10a = 800 - 200 = 600$

$\therefore$ $a = 60$,

Put, $a = 60, b = 100$ in eq^n. (i) we get

$$c = 200 - 60 - 100 = 40$$

Thus, the missing frequencies are 60,100 and 40.

Dispersion

Two distributions may have the same mean, median, and mode but they may differ in formations. For example, consider the following series:

	Data					Mean	Median	Mode
Series X :	10	10	10	10	10	10	10	10
Series Y :	8	9	10	10	13	10	10	10
Series Z :	2	6	10	10	22	10	10	10

All the three series X, Y, and Z have the same mean, median and mode (i.e. 10) but they differ entirely in formations. It is clear that the measures of central tendency are inadequate to describe the distribution. Thus, the measures of central tendency must be supported and supplemented by some other measures. One such measure is *dispersion* or *variability*.

Dispersion means scatter or spread or variation. Dispersion provides an idea of the homogeneity (compactness) or heterogeneity (scatter) of the distribution. In the above example, we say the series X is stationary or constant and show no variability. Series Y is slightly dispersed and series Z is relatively more dispersed. We say that series Y is more homogeneous (or uniform) as compared to series Z i.e. series Z is more heterogeneous than series Y.

Objective of Measuring Dispersion

The objective of measuring variability are as follows

(a) To determine the reliability of an average

When dispersion is small, the average is said to be reliable. It is said to be a good estimate of the average in the corresponding distribution.

(b) To serve as a basis for the control of variability

For example, variations in body temperatures, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. Similarly, in industrial production, efficient operation is necessary to control quality variation of the products.

(c) To compare two or more series with regard to their variability

The greater the variation, the lesser will be the uniformity or consistency. The lesser the variation, the greater will be the uniformity or consistency.

(d) To facilitate the use of other statistical measures

Many analytical tools in statistics such as correlation and regression analysis are based on measures of dispersion.

Properties (Requisites) of Good Measure of Dispersion

As far as possible, a good measure of dispersion should have the following *properties*:

Scanned by CamScanner

(a) It should be simple to understand.
(b) It should be easy to compute.
(c) It should be rigidly defined.
(d) It should be based on all the observations.
(e) It should be capable of further algebraic treatment.
(f) It should be least affected by fluctuations of sampling.
(g) It should not be unduly affected by extreme items.

Absolute and Relative Measure of Dispersion

An *absolute measure of dispersion* expresses the variation in the same units of measurement as the original data. It is used to compare the variability of two distributions having the same units and almost the same average size. If the two distributions are expressed in different units (e.g. kilograms of sugar verses tones of sugarcane) or in different average size (e.g. managers' salary verses workers' salary), we use relative measure of dispersion to compare the variability of two distributions.

A *relative measure of dispersion* is defined by

$$\text{Relative dispersion} = \frac{\text{Absolute dispersion}}{\text{Suitable average}}$$

Here suitable average means the average used in computing absolute dispersion. Relative dispersion is also called *coefficient of dispersion* because it is a pure number independent of original units of measurement.

Note that the relative measure of dispersion can also be used to compare the variability of two distributions with different units.

Methods of Studying Dispersion

The important methods of studying dispersion are:

(a) Range
(b) Quartile Deviation
(c) Mean Deviation
(d) Standard Deviation
(e) Coefficient of variation

The first two methods are *positional* measures. The third and fourth methods are *computational* measures.

Range

A simple but crude measure of dispersion is the *range* which is defined as the difference between maximum (or largest) and minimum (or smallest) values in a given data (or distribution).

$\therefore$ Range = L – S

Where L = Largest item, S = Smallest item.

For a frequency distribution we determine the range either (i) by taking the difference between the upper limit of the highest class and the lower limit of the lowest class or, (ii) by taking the difference between mid point of the highest class and the mid point of the lowest class.

The range is an absolute measure of dispersion and it has the same units of measurement as that of the given data. The relative measure corresponding to the range, called the *coefficient of range*, is defined by

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

If we compare two distributions having almost the same mean, the distribution with smaller range is said to have less dispersion and thus its mean is more typical of the distribution.

Example 37. From the calories consumption of ten person given below calculate the range and the coefficient of range.

Scanned by CamScanner

Program	Run time in Second	Program	Run time in Second
1	1450	6	6190
2	3670	7	2800
3	2680	8	1150
4	730	9	8700
5	1850	10	3150

Solution

Largest item(L) = Rs 8700

Smallest item (S)= Rs 730

Range = L – S= 8700 – 730 = 7970

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{8700-730}{8700+730} = \frac{7970}{9430} = 0.845.$$

Example 38. The following table gives the age distribution of a group of 50 individuals.

Age (years)	No. of persons
16–20	10
21–25	15
26–30	17
31–35	8

Calculate the range and its coefficient.

Solution

If we convert the given inclusive classes into exclusive classes, we get the first class as 15.5 – 20.5 and the last class as 30.5 – 35.5.

Largest item (L)= 35.5

Smallest item(S) = 15.5

∴ Range = L – S = 35.5 – 15.5 =20 years.

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{35.5-15.5}{35.5+15.5} = \frac{20}{51} = 0.39$$

Note that we convert inclusive classes into exclusive classes by using the correction factor

$$CF = \frac{\text{Lower limit of 2}^{nd}\text{ class} - \text{Upper limit of 1}^{st}\text{ class}}{2} = \frac{21-20}{2} = 0.5$$

Add 0.5 to upper limits and subtract 0.5 to lower limits.

Merits and Demerits of Range

Merits

(a) Of all the methods of studying dispersion, range is the simplest to understand and easiest to compute.

(b) Its computation is very fast. Hence, if we want to know a quick rather than a very accurate picture of variability, we may compute range.

Demerits

(a) It is not based on all the items of the distribution.

(b) It is very much affected by fluctuations of sampling i.e. its value varies widely from sample to sample.

(c) It cannot be computed for open-ended class.

(d) It is not capable of further algebraic treatment.

The range cannot be regarded as a reliable measure of dispersion.

Scanned by CamScanner

Quartile Deviation (Semi Inter Quartile Range)

The *inter-quartile range*, based on the first and the third quartiles, includes the middle 50% of the items of a distribution and is defined by

$$\text{Inter-quartile Range} = Q_3 - Q_1$$

Quartile deviation (Q.D.) is the half of the inter-quartile range and is written as:

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

Q.D. is also called *semi-interquartile range*. It is clear that Q.D. gives the average amount by which the two quartiles differ from the median. Smaller the Q.D., smaller the variation of central 50% items and vice versa. The greater the Q.D., the greater will be the variation of central 50% items and vice versa. The relative measure corresponding to Q.D., called the *coefficient of Q.D.*, is defined by:

$$\text{Coefficient of Q. D} = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 39. Calculate quartile deviation, and coefficient of quartile deviation in Example 34.

Solution

Arrange the calories consumption in increasing order

730, 1150, 1450, 1850, 2680, 2800, 3150, 3670, 6190, 8700

$$Q_1 = \text{size of } \left(\frac{n+1}{4}\right)^{th} \text{item} = \text{Size of } \left(\frac{10+1}{4}\right)^{th} \text{item} = \text{size of } 2.75^{th} \text{ item}$$

$$= 2^{nd} \text{ item} + \frac{3}{4}(3^{rd} \text{ item} - 2^{nd} \text{ item})$$

$$= 1150 + \frac{3}{4}(1450 - 1150) = 1375$$

$$Q_3 = \text{Size of } 3\left(\frac{n+1}{4}\right)^{th} \text{item} = 8.25^{th} \text{ item}$$

$$= 8^{th} \text{ item} + \frac{1}{4}(9^{th} \text{ item} - 8^{th} \text{ item})$$

$$= 3670 + \frac{1}{4}(6190 - 3670) = 4300$$

$$\text{Q. D.} = \frac{Q_3 - Q_1}{2} = \frac{4300 - 1375}{2} = \frac{2925}{2} = 1462.5$$

$$\therefore \quad \text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{4300 - 1375}{4300 + 1375} = \frac{2925}{5675} = 0.5$$

Example 40. Compute Q.D. and its coefficient from the following table.

Size of item	2	4	6	8	10	12
Frequency	3	5	10	12	6	4

Solution

Size	f	c.f.
2	3	3
4	5	8
6	10	18
8	12	30
10	6	36
12	4	40

Scanned by CamScanner

$\therefore\ Q_1 = \left(\frac{N+1}{4}\right)^{th}\text{item} = \left(\frac{40+1}{4}\right)^{th}\text{item}$

$= (10.25)^{th}\text{item} = 6$ (i.e. size corresponding to c.f. = 18)

$Q_3 = 3\left(\frac{N+1}{4}\right)^{th}\text{item} = 30.75^{th}$ item = 10 (i.e. size corresponding to cf 36)

$\therefore\ Q.D = \frac{Q_3 - Q_1}{2} = \frac{10-6}{2} = 2$

Coefficient of Q.D $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{10-6}{10+6} = \frac{4}{16} = 0.25.$

Example 41. Calculate coefficient of range and coefficient of quartile deviation from the following data:

Humidity	No of days	Humidity	No of days
50–60	8	90–100	18
60–70	13	100–110	12
70–80	30	110–120	9
80–90	35	120–130	5

Solution

Class	f	c.f.
50–60	8	8
60–70	13	21
70–80	30	51
80–90	35	86
90–100	18	104
100–110	12	116
110–120	9	125
120–130	5	130

Coefficient of range $= \frac{L-S}{L+S} = \frac{130-50}{130+50} = \frac{80}{180} = 0.44$

For Q_1: $\frac{N}{4} = \frac{130}{4} = 32.5.$

$\therefore$ Q_1 lies in the class 70 – 80.

$\therefore\ Q_1 = L + \frac{N/4 - c.f.}{f} \times h = 70 + \frac{32.5-21}{30} \times 10 = 73.83$

For Q_3: $\frac{3N}{4} = 3 \times 32.5 = 97.5$

$\therefore$ Q_3 lies in 90 –100.

$\therefore\ Q_3 = L + \frac{3N/4 - c.f.}{f} \times h = 90 + \frac{97.5-86}{18} \times 10 = 96.39$

$\therefore$ Coefficient of Q.D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{96.39 - 73.83}{96.39 + 73.83} = \frac{22.56}{170.22} = 0.1325.$

Example 42. Calculate the Q.D. from the following data:

Rainfall (mm)	below 35	35–37	38–40	41–43	over 43
No of days	14	60	95	24	7

Scanned by CamScanner

Solution

Class (inclusive)	Class (exclusive)	No of workers f	c.f.
Below 35	Below 34.5	14	14
35–37	34.5–37.5	60	74
38–40	37.5–40.5	95	169
41–43	40.5–43.5	24	193
Over 43	Over 43.5	7	200

For Q_1, $\frac{N}{4} = \frac{200}{4} = 50$

$\therefore$ Q_1 lies in 35–37.

But the real limit of this class is 34.5 – 37.5

$$\therefore \quad Q_1 = L + \frac{N/4 - c.f.}{f} \times h$$

$$= 34.5 + \frac{50 - 14}{60} \times 3 = 36.3$$

For Q_3: $\frac{3N}{4} = 3 \times \frac{200}{4} = 150$

$\therefore$ Q_3 lies in 38 – 40

The real limit is 37.5 – 40.5

$$\therefore \quad Q_3 = L + \frac{3N/4 - c.f.}{f} \times h = 37.5 + \frac{150 - 74}{95} \times 3 = 39.9$$

$$\therefore \quad Q.D. = \frac{Q_3 - Q_1}{2} = \frac{39.9 - 36.3}{2} = 1.8$$

Merits and Demerits of Q.D.

Merits

(a) It is easy to understand and compute.

(b) It can be computed for a distribution with open ended classes.

(c) It is not affected at all by the presence of extreme items.

Demerits (or Limitations)

(a) It is not based on all the observations because it ignores 50% items (the first 25% and the last 25% items).

(b) It is not capable of further algebraic treatment.

(c) Its value is very much affected by fluctuations of sampling.

(d) It is in fact not a measure of dispersion because it does not indicate the scatter around an average rather a distance on a scale.

(e) It has a serious drawback that all the series having the same quartiles will have the same quartile deviations and hence the same measure of variability whatever may be the distribution of items in between two quartiles or outside them.

Percentile Range

Like semi-inter-quartile range, the *percentile range* and *semi-percentile range* are also used as a measure of dispersion. They are defined as follows:

$$\text{Percentile Range} = P_{90} - P_{10}$$

$$\text{Semi-percentile range} = \frac{P_{90} - P_{10}}{2}$$

Where P_{10} = 10th percentile and P_{90} = 90th percentile.

Scanned by CamScanner

Mean Deviation (MD) or Average Deviation

Range and Quartile Deviation are not a measure of dispersion in strict sense because they do not indicate the dispersion around an average. The two other measures of dispersion, mean deviation and standard deviation, show the dispersion around an average. The *mean deviation* (denoted by M.D) is the mean of the difference between the items in a series and the median or mean of that series. In theory, it is suitable to take deviations from median since "the sum of the deviations of items from median is least". In practice, we usually take deviations from the mean while computing the mean deviation.

Thus for individual observations,

$$\text{M.D. from mean} = \frac{1}{n}\Sigma|X - \bar{X}|$$

$$\text{M.D. from median} = \frac{1}{n}\Sigma|X - M_d|$$

$$\text{M.D. from mode} = \frac{1}{n}\Sigma|X - M_o|$$

For discrete or continuous frequency distribution

$$\text{M.D. from mean} = \frac{1}{N}\Sigma f|X - \bar{X}|$$

$$\text{M.D. from median} = \frac{1}{N}\Sigma f|X - M_d|$$

$$\text{M.D. from mode} = \frac{1}{N}\Sigma f|X - M_o|$$

Note that the vertical bars indicate the absolute value of deviation i.e. we ignore the algebraic sign.

$$\text{Coeff. of M.D. from mean} = \frac{\text{M.D. from mean}}{\text{Mean}}$$

$$\text{Coeff. of M.D. from median} = \frac{\text{M.D. from median}}{\text{Median}}$$

$$\text{Coeff. of M.D. from mode} = \frac{\text{M.D. from Mode}}{\text{Mode}}$$

Example 43. Find the mean deviation (i) from the mean (ii) from the median in the following data:

Height of wheat plant (in c.m.): 25, 28, 32, 44, 50

Also calculate the coefficient of mean deviation.

Solution

X	$X-\bar{X}$	$\lvert X-\bar{X}\rvert$	$X-M_d$	$\lvert X-M_d\rvert$
25	-10.8	10.8	-7	7
28	-7.8	7.8	-4	4
32	-3.8	3.8	0	0
44	8.2	8.2	12	12
50	14.2	14.2	18	18
$\Sigma X = 179$		$\Sigma\lvert X-\bar{X}\rvert = 44.8$		$\Sigma\lvert X-M_d\rvert = 41$

$$\text{Mean } (\bar{X}) = \frac{\Sigma X}{n} = \frac{179}{5} = 35.8 \text{ c.m.}$$

$$\text{Median, } M_d = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Size of } \left(\frac{5+1}{2}\right)^{th} \text{ item} = \text{Size of } 3^{rd} \text{ item} = 32$$

(i) $$\text{M.D from mean} = \frac{1}{n}\Sigma|X - \bar{X}| = \frac{1}{5} \times 44.8 = 8.96$$

Scanned by CamScanner

$$\text{Coeff. of M.D.} = \frac{\text{M.D. from mean}}{\bar{X}} = \frac{8.96}{35.8} = 0.25$$

(ii) $$\text{M.D from median} = \frac{1}{n}\Sigma|X - M_d| = \frac{41}{5} = 8.2$$

$$\text{Coeff. of M.D.} = \frac{\text{M.D.from median}}{M_d} = \frac{8.2}{32} = 0.256$$

Example 44. Calculate the mean deviation from mean of the following series :

Marks	No. of Students	Marks	No. of Students
0-10	5	30-40	16
10-20	18	40-50	6
20-30	15		

Solution

Class	Mid value (X)	f	fX	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$
0-10	5	5	25	20	100
10-20	15	18	270	10	180
20-30	25	15	375	0	0
30-40	35	16	560	10	160
40-50	45	6	270	20	120
Total		N = 60	$\sum fX = 1500$		$f\lvert X-\bar{X}\rvert = 560$

$$\text{Mean, } \bar{X} = \frac{\Sigma fX}{N} = \frac{1500}{60} = 25$$

$$\text{M.D. from mean} = \frac{\Sigma f|X - \bar{X}|}{N} = \frac{560}{60} = 9.33 \text{ marks.}$$

Example 45. Calculate the median and mean deviation for the following frequency distribution:

Age (years)	No. of persons	Age (years)	No. of persons
1-5	7	26-30	18
6-10	10	31-35	10
11-15	16	36-40	5
16-20	32	41-45	1
21-25	24		

Solution

Class	Mid value (X)	f	c.f	$\lvert X-M_d\rvert$	$f\lvert X-M_d\rvert$
1-5	3	7	7	17	119
6-10	8	10	17	12	120
11-15	13	16	33	7	112
16-20	18	32	65	2	64
21-25	23	24	89	3	72
26-30	28	18	107	8	144
31-35	33	10	117	13	130
36-40	38	5	122	18	90
41-45	43	1	123	23	23
		N=123			$\sum f\lvert X-M_d\rvert = 874$

Scanned by CamScanner

For median, $\frac{N}{2} = \frac{123}{2} = 61.5$.

$\therefore$ M_d lies in 16–20. But the real median class is 15.5 – 20.5.

$\therefore$ $M_d = L + \frac{N/2 - c.f.}{f} \times h$

$= 15.5 + \frac{61.5-33}{32} \times 5 = 15.5 + \frac{28.5}{32} \times 5 = 19.95 = 20$ (approx)

M.D. from median $= \frac{\Sigma f|X - M_d|}{N} = \frac{874}{123} = 7$ years

Example 46. Calculate the mean deviation and its co-efficient for the following data, using the mean:

Age(year)	25–29	30–34	35–39	40–44	45–49	50–54
No. of person	8	40	74	44	20	6

Solution

Class (years)	X(Mid value)	f	fX	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$
25–29	27	8	216	11.2	89.6
30–34	32	40	1280	6.2	248.0
35–39	37	74	2738	1.2	88.8
40–44	42	44	1848	3.8	167.2
45–49	47	20	940	8.8	176.0
50–54	52	6	312	13.8	82.8
		192= N	7334= ΣfX		852.4=$\Sigma f\lvert X-\bar{X}\rvert$

Mean $(\bar{X}) = \frac{\Sigma fX}{N} = \frac{7334}{192} = 38.198 = 38.2$

M.D. $= \frac{\Sigma f|X - \bar{X}|}{N} = \frac{852.4}{192} = 4.44$

Coeff. of M.D. $= \frac{M.D.}{\bar{X}} = \frac{4.44}{38.2} = 0.116$

Merits and Demerits of Mean Deviation

Merits

(a) It is easy to understand and compute.

(b) It is based on every item of the series.

(c) It is less affected by extreme items than the standard deviation.

(d) Since mean deviation is based on the deviations about an average, it gives a better measure for comparison about the formation of different distributions.

Demerits

(a) The greatest drawback of this method is that algebraic signs are ignored while taking deviations of items. Mathematically, this is wrong.

(b) It is not capable of further algebraic treatment.

(c) It cannot be computed for distributions with open ended classes.

(d) It is affected by fluctuations of sampling. The value of mean deviation tends to increase with the size of the sample.

(e) It may not give accurate results.

Scanned by CamScanner

(f) It is rarely used in sociological studies.

Thus, its use is limited because of the above demerits.

Standard Deviation

The standard deviation is the most important and widely used measure of dispersion or variability. The standard deviation is the square root of the mean of the squared deviations from the arithmetic mean and is denoted by S.D. or σ. The standard deviation is also called "root-mean-squared-deviation" (R.M.S.D.).

Thus, for *individual* observations,

$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}} = \sqrt{\frac{\Sigma X^2}{n} - \left\{\frac{\Sigma X}{n}\right\}^2}$$

For *discrete or continuous* distribution

$$\sigma = \sqrt{\frac{1}{N}\Sigma f(X-\bar{X})^2}$$

$$\sigma = \sqrt{\frac{1}{N}\Sigma fX^2 - \left(\frac{\Sigma fX}{N}\right)^2}$$

The standard deviation is an absolute measure of dispersion. The greater the S.D., the greater the amount of dispersion of a distribution. A small S.D. indicates a high degree of uniformity or homogeneity of the data, a large standard deviation indicates just the opposite.

Computing S.D. for Individual Observations (Ungrouped data)

(a) **Direct Method or Actual Mean Method:** In this method we take deviations of observations from the A.M. and then apply relation (9).

$$\sigma = \sqrt{\frac{1}{n}\Sigma(X-\bar{X}^2)} = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$$

(b) **Deviation Method:** In this case we take deviations of observations from an assumed mean and then apply the

Formula: $\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$

Where d = X – A and A = assumed mean

Computing S.D. for Discrete or Continuous Distribution (Grouped data)

(a) **Direct Method:** In this method we take deviations of items from the arithmetic mean and then use the formula (10) i.e.

$$\sigma = \sqrt{\frac{1}{N}\Sigma f(X-\bar{X})^2} = \sqrt{\frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2}$$

(b) **Deviation Method:** In this method we take deviations of items from the assumed mean and then use the formula:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \quad \text{where } d = X - A$$

(c) **Step-deviations Method:** In this case, we take deviations of items from the assumed mean and the divide by a common factor from the given data and then use the formula:

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$

Where $d' = \frac{d}{h} = \frac{X-A}{h}$ = step deviation

h = common factor

Note that X denotes mid points of classes in the case of continuous distribution.

Scanned by CamScanner

Example 47. Find the standard deviation of bacteria given below: by (i) direct method (ii) deviation method:

Sample No.:	A	B	C	D	E	F	G	H	I	J
No of Bacteria:	120	110	115	122	126	140	125	121	110	131

Solution

Bacteria No. (x)	x –125 = d	d^2	$x - \bar{x}$	$(x - \bar{x})^2$
120	–5	25	–2	4
110	–15	225	–12	144
115	–10	100	–7	49
122	–3	9	0	0
126	1	1	4	16
140	15	225	18	324
125	0	0	3	9
121	–4	16	–1	1
110	–15	225	–12	144
131	6	36	9	81
$\Sigma x = 1220$	$\Sigma d = -30$	$\Sigma d^2 = 862$		$\Sigma(x-\bar{x})^2 = 772$

$$\text{Mean } \bar{x} = \frac{\Sigma x}{n} = \frac{1220}{10} = 122$$

(i) S.D. by direct method:

$$\sigma = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2} = \sqrt{\frac{772}{10}} = \text{Rs. } 8.786$$

(ii) S.D. by deviation method

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{862}{10} - \left(\frac{-30}{10}\right)^2} = \sqrt{86.2 - 9} = 8.786.$$

Example 48. Calculate the standard deviation from the data:

Size	06	09	12	15	18
Frequency	07	12	19	10	02

Solution

Size(X)	f	d' = (X–12)/3	fd'	fd'²
06	07	–2	–14	28
09	12	–1	–12	12
12	19	0	0	0
15	10	1	10	10
18	02	2	4	8
Total	N = 50		–12	58

$$\text{We have S.D., } \sigma = \sqrt{\frac{1\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$

$$= \sqrt{\frac{58}{50} - \left(\frac{-12}{50}\right)^2} \times 3 = \sqrt{1.16 - 0.0576} \times 3 = 3.15$$

Example 49. The following data relate to a study of increment/decrement of blood sugar of 237 person due to specific food.

Scanned by CamScanner

Increment of Blood sugar	Nos of person
-10 to 0	19
0 to 10	24
10 to 20	49
20 to 30	87
30 to 40	31
40 to 50	27

Calculate the mean and standard deviation increment of blood sugar.

Solution

Class	Mid value (x)	frequency (f)	(x-25)/10 =d'	fd'	fd'²
-10 to 0	-5	19	-3	-57	171
0 to 10	5	24	-2	-48	96
10 to 20	15	49	-1	-49	49
20 to 30	25	87	0	0	0
30 to 40	35	31	1	31	31
40 to 50	45	27	2	54	108
		N = 237		Σf d' = -69	Σf d'² = 455

$$\text{Mean} = \bar{x} = A + \frac{\Sigma fd'}{N} \times h$$

$$= 25 + \frac{-69}{237} \times 10 = 22.09$$

$$\text{S.D.}, \sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$

$$= \sqrt{\frac{455}{237} - \left(\frac{-69}{237}\right)^2} \times 10 = \sqrt{1.92 - 0.08} \times 10 = 13.56$$

Mathematical Properties of S.D.

(a) **Combined standard Deviation:** The ***combined standard deviation*** of two or more group of values of a variable can be computed on the basis of their individual standard deviations. If σ_1 be the S.D. of 1st group consisting n_1 values and σ_2, the S.D. of 2nd group consisting n_2 values then the combined S.D. denoted by σ_{12} of two groups is given by the formula.

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

Where $d_1 = \bar{X}_1 - \bar{X}_{12}$, $d_2 = \bar{X}_2 - \bar{X}_{12}$

$$\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \text{combined mean.}$$

Note that this formula can be extended for three or more groups.

(b) **Deviation from the mean:** The sum of square of deviation from the mean is minimum i

$\Sigma(X - \bar{X})^2 < \Sigma(X - A)^2$ where $\bar{X} \neq A$. This is the reason that we always take deviations from t arithmetic mean while computing S.D.

(c) **Standard deviation is independent of change of origin but not of scale.**

Thus $\sigma_x = \sigma_d$

$\sigma_x = h\sigma_d'$

Scanned by CamScanner

Where $d = X - A$

$$d' = \frac{X - A}{h}$$

(d) Relation between S.D. and other measures of dispersion.

In moderately symmetrical (or asymmetrical) distribution, the following empirical relationships hold good.

M.D. = 0.7979 (σ); Q.D. = 0.6745 (σ)

Example 50. The mean of two samples of sizes 50 and 100 respectively are 54.1 and 50.3 and the S.D.'s are 8 and 7. Find the mean and S.D. of the sample size 150 obtained by combining the two samples.

Solution

Here $n_1 = 50$, $n_2 = 100$, $\bar{X}_1 = 54.1$, $\bar{X}_2 = 50.3$, $\sigma_1 = 8$, $\sigma_2 = 7$

Hence the combined mean is

$$\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{50 \times 54.1 + 100 \times 50.3}{50 + 100} = \frac{7735}{150} = 51.57$$

Now $d_1 = \bar{X}_1 - \bar{X}_{12} = 54.1 - 51.57 = 2.53$;

$d_2 = \bar{X}_2 - \bar{X}_{12} = -1.27$

The combined S.D.,

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$= \sqrt{\frac{50(64 + 6.4) + 100(49 + 1.6)}{50 + 100}} = \sqrt{\frac{3520 + 5060}{150}} = 7.56$$

Variance

The variance of the values of a variable X is defined as the square of the standard deviation.

Hence, Var (X) = σ^2

$\therefore\ \sigma = \sqrt{\text{Var}(X)}$

It is clear that variance is always positive and measured in terms of square units of given data. The concept of variance is very much useful in anfurther statistical work.

Sample standard deviation:

A standard deviation of a sample which is used to estimate the standard deviation of a population based on random sample is said to be sample standard deviation.

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_n$ be n sample observations then sample standard deviation denoted by s is given by positive square root of sum of square of deviation of sample observations from their mean divided by sample size minus one.

$$s = \sqrt{\frac{1}{n-1}\sum(x - \bar{x})^2} = \sqrt{\frac{1}{n-1}\left(\sum x^2 - n(\bar{x})^2\right)}$$

Sample variance:

It is square of sample standard deviation and is given by

$$S^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 = \frac{1}{n-1}\left(\sum x^2 - n\,\bar{x}^2\right)$$

Example: Given the following set of data from a sample of size 7, 12 , 7 , 4 , 9 , 0 , 7 , 3 . Find standard deviation and variance.

Scanned by CamScanner

Solution

x	x^2
12	144
7	49
4	16
9	81
0	0
7	49
3	9
$\Sigma x = 42$	$\Sigma x^2 = 348$

Now

$$\bar{x} = \frac{\Sigma x}{n} = \frac{42}{7} = 6$$

$$\text{Sample sd (s)} = \sqrt{\frac{1}{n-1}\Sigma(x^2 - n\bar{x}^2)} = \sqrt{\frac{1}{7-1}(348 - 7 \times 6^2)} = 4$$

Sample variance $(S^2) = 4^2 = 16$

Merits and Demerits of Standard Deviation

Merits

(a) The standard deviation is rigidly defined and is based on all the observations.

(b) The standard deviation is capable of further algebraic treatment. For example we can compute combined standard deviation of two or more groups.

(c) Of all the measures of dispersion, the standard deviation is affected least by fluctuations of sampling.

(d) The coefficient of variation, based on the standard deviation, is the most suitable for comparing the variability of two or more distributions.

(e) The standard deviation is useful in further statistical work.

(f) The standard deviation is most useful in judging the representativeness of mean i.e. the distribution with the smallest standard deviation is said to have the most representative mean.

(g) It is the ideal measure of dispersion.

Demerits (limitations)

(a) It is not easy to compute the standard deviation as compared to other measures of dispersion.

(b) It gives more weight to extreme items and less to those near the mean because of squaring the deviations.

Coefficient of Variation

The *co-efficient of variation* (C.V.) is the relative measure based on the standard deviation and is defined as the ratio of the standard deviation to the mean expressed in percent.

In symbols, $\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100\%$

The ratio $\sigma/\bar{X}$ is called the *coefficient of standard deviation*. It is obvious that C.V. is independent of units. Hence the coefficient of variation is a suitable measure for comparing variability of two series (or set of data or distributions) with the same or different units. A series with smaller C.V. is said to be less variable (or more consistent or more homogeneous or more uniform or more stable) than other and the series with greater C.V. is said to be more variable (or less consistent or less homogeneous or less stable) than the other.

Scanned by CamScanner

Note:
When samples are given coefficient of variation is given by

$CV = \frac{s}{\bar{x}} \times 100\%$ where s is sample standard deviation and $\bar{x}$ is sample mean.

Example 51. A factory produces two types of electric bulbs A and B. In an experiment relating to their life, the following results were obtained.

Length of life (in hours)	No. of bulbs A	No. of bulbs B
500–700	5	4
700–900	11	30
900–1100	26	12
1100–1300	10	8
1300–1500	8	6

Compare the variability of the life of the two types of bulbs using coefficient of variation.

Solution

Let d' = (X–1000)/200.

Mid point	Bulb A				Bulb B		
X	d'	f	fd'	fd'²	f'	f'd'	f'd'²
600	–2	5	–10	20	4	–8	16
800	–1	11	–11	11	30	–30	30
1000	0	26	0	0	12	0	0
1200	1	10	10	10	8	8	8
1400	2	8	16	32	6	12	24
Total		60 = Σf	5 = Σfd'	73 = Σfd'²	60 = Σf'	–18 = Σf'd'	78 = Σf'd'²

Mean for bulb A, $\bar{X}_A = A + \frac{\Sigma fd'}{N} \times h$

$$= 1000 + \frac{5}{60} \times 200 = 1016.67 \text{ hours}$$

S.D. for A, $\sigma_A = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$

$$= \sqrt{\frac{78}{60} - \left(\frac{5}{60}\right)^2} \times 20$$

$$= \sqrt{1.2167 - 0.0069} \times 200 = \sqrt{1.21} \times 200 = 220 \text{ hours}$$

C.V. for A $= \frac{\sigma_A}{\bar{X}_A} \times 100\% = \frac{220}{1016.67} \times 100 = 21.64\%$

Mean for bulb B, $\bar{X}_B = A + \Sigma\frac{f'd'}{N} \times h = 1000 + \left(\frac{-18}{60}\right) \times 200 = 1000 - 60 = 940$ hrs

S.D. for B, $\sigma_B = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$

$$= \sqrt{\frac{78}{60} - \left(\frac{-18}{60}\right)^2} \times 200$$

$$= \sqrt{1.3 - 0.09} \times 200 = \sqrt{1.21} \times 200 = 1.1 \times 200 = 220 \text{ hours}$$

Scanned by CamScanner

C.V. for B $= \frac{\sigma_A}{\bar{X}_B} \times 100\% = \frac{220}{940} \times 100\ \% = 23.4\%$

C.V. for A is less than C.V. for B. Hence bulb A is more stable than bulb B i.e. the bulb A has longer life than bulb B.

Example 52. The mean and standard deviation of 200 items are found to be 60 and 20 respectively. If at the time of calculations, two items were wrongly taken as 3 and 67 instead of 13 and 17, find the correct mean and standard deviation. What is the correct coefficient of variation?

Solution

Here $n = 200, \bar{X} = 60, \quad \sigma = 20 \qquad \therefore \sigma^2 = 400$

We have

$$\bar{X} = \frac{\Sigma X}{n} \qquad \therefore \quad \Sigma X = n\bar{X} = 200 \times 60 = 12000$$

We have

$$\sigma^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

$$\therefore \quad \Sigma X^2 = n(\sigma^2 + \bar{X}^2) = 200\,(400 + 3600) = 8{,}00{,}000$$

If the wrong items 3 and 67 are replaced by the correct values 13 and 17 respectively, we get

$$\text{Correct } \Sigma X = 12000 - 3 - 67 + 13 + 17 = 11960$$

$$\text{Correct } \Sigma X^2 = 800000 - 3^2 - 67^2 + 13^2 + 17^2$$

$$= 800000 - 9 - 4489 + 169 + 289 = 795960$$

$$\therefore \quad \text{Corrected mean } \bar{X} = \frac{\text{Correct } \Sigma X}{n} = \frac{11960}{200} = 59.8$$

$$\text{Correct s.d., } \sigma = \left[\frac{\text{Correct } \Sigma X^2}{n} - (\text{Correct mean})^2\right]^{1/2}$$

$$= \sqrt{\frac{795960}{200} - (59.8)^2}$$

$$= \sqrt{3979.8 - 3576.04}$$

$$= \sqrt{403.76}$$

$$= 20.09$$

$$\therefore \quad \text{Correct C.V.} = \frac{\text{Correct } \sigma}{\text{Correct } \bar{X}} \times 100\% = \frac{20.09}{59.8} \times 100\% = 33.60\%$$

Example 53. Sample of polythene bags from two manufacturers, A and B are tested by a prospective buyer for bursting pressure and the results are as follows:

Bursting Pressure (lb.)	Number of Bags	
	A	B
5.0 – 9.9	2	9
10.0 – 14.9	9	11
15.0 – 19.9	29	18
20.0 – 24.9	54	32
25.0 – 29.9	11	27
30.0 – 34.9	5	13

Which set of bags has more uniform pressure? If prices are the same which manufacturer's bag would be preferred by the buyer? Why?

Scanned by CamScanner

Solution

Let X be the mid value so that $d' = \frac{X - 17.45}{5}$. Let f and f' be the number of bags of manufacturers A and B respectively.

Class	m.v. (X)	d' = (X–17.45)/5	f	f'	f.d'	f d'²	f ' d'	f ' d'²
5–9.9	7.45	–2	2	9	–4	8	–18	36
10–14.9	12.45	–1	9	11	–9	9	–11	11
15–19.9	17.45	0	29	18	0	0	0	0
20–24.9	22.45	1	54	32	54	54	32	32
25–29.9	27.45	2	11	27	22	44	54	108
30–34.9	32.45	3	5	13	15	45	39	117
		3	110= N	110= N'	78	160	96	304

For manufacturer A

$$\text{Mean, } \bar{X}_A = A + \frac{\Sigma fd'}{N} \times h$$

$$= 17.45 + \frac{78}{110} \times 5 = 21$$

$$\text{S.D., } \sigma_A = h\sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$$

$$= 5\sqrt{\frac{160}{110} - \left(\frac{78}{110}\right)^2}$$

$$= 5\sqrt{1.45 - 0.50}$$

$$= 5 \times 0.975 = 4.875$$

$$\text{C.V. (A)} = \frac{\sigma_A}{X_A} \times 100\% = \frac{4.875}{21} \times 100\%$$

$$= 23.2\%$$

For manufacturer B

$$\text{Mean, } \bar{X}_B = A + \frac{\Sigma f'd'}{N} \times h$$

$$= 17.45 + \frac{96}{110} \times 5 = 21.8$$

$$\text{S.D., } \sigma_B = h\sqrt{\frac{\Sigma f'd'^2}{N} - \left(\frac{\Sigma f'd'}{N}\right)^2}$$

$$= 5\sqrt{\frac{304}{110} - \left(\frac{96}{110}\right)^2}$$

$$= 5\sqrt{2.76 - 0.76}$$

$$= 5 \times 1.41 = 7.05$$

$$\text{C.V. (B)} = \frac{\sigma_B}{X_B} \times 100\% = \frac{7.05}{21.8} \times 100\% = 32.34\%$$

Since the coefficient of variation for A is less than that for B, the set of bags manufactured by A has more uniform pressure. Hence the buyer would prefer to buy the bags manufactured by A.

Example 54. The board of directors of Gamma Products is considering acquiring one of two companies and is closely examining the management of each company in regard to their inclinations towards risk. During the past 5 years, the first company had an average return on investment of 28 percent, with a standard deviation of 5.3 percent. The second company had an average return of 37.8 percent, with a standard deviation of 4.8 percent. If we consider risk to be associated with greater relative dispersion, which of these two companies has pursued a riskier strategy?

Solution

First Company

$\bar{X}_1 = 28\%$

$\sigma_1 = 5.3\%$

Second company

$\bar{X}_2 = 37.8$

$\sigma_2 = 4.8\%$

$$\text{C.V. for the first company} = \frac{\sigma_1}{\bar{X}_1} \times 100\% = \frac{5.3}{28} \times 100\% = 18.93\%$$

$$\text{C.V. for the second company} = \frac{\sigma_2}{\bar{X}_2} \times 100\% = \frac{4.8}{37.8} \times 100\% = 12.7\%$$

Scanned by CamScanner

Since CV for 1st company (18.93%) is greater than CV for the second company (12.7%); the first company pursued the riskier strategy.

Example 55. Following are sample of ages of 6 male and 6 female enrolled in hardware training program of computer

Age of male	23	30	18	25	32	40
Age of female	20	35	44	27	41	38

Find which gender has uniform age.

Solution

Age of male (x)	x^2	Age of female (y)	y^2
23	529	20	400
30	900	35	1225
18	324	44	1936
25	625	27	729
32	1024	41	1681
40	1600	38	1444
$\Sigma x = 168$	$\Sigma x^2 = 5002$	$\Sigma y = 205$	$\Sigma y^2 = 7415$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{168}{6} = 28$$

$$s_x = \sqrt{\frac{1}{n-1}\left(\Sigma x^2 - n\bar{x}^2\right)} = \sqrt{\frac{1}{6-1}(5002 - 6 \times 28^2)} = \sqrt{59.6} = 7.72$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{205}{6} = 34.16$$

$$s_x = \sqrt{\frac{1}{n-1}\left(\Sigma y^2 - n\bar{y}^2\right)} = \sqrt{\frac{1}{6-1}(7415 - 6 \times (34.16)^2)} = \sqrt{82.71} = 9.09$$

$$CV_x = \frac{s_x}{\bar{x}} \times 100\% = \frac{7.72}{28} \times 100\% = 27.57\%$$

$$CV_y = \frac{s_y}{\bar{y}} \times 100\% = \frac{9.09}{34.16} \times 100\% = 26.61\%$$

Here $CV_y < CV_x$

Hence female have uniform age.

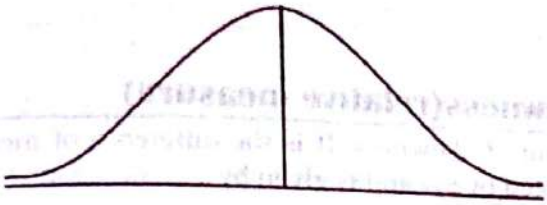
Skewness

The measure of central tendency tells us about the concentration of the items about the middle of the distribution and the measure of dispersion gives us an idea about the scatterness of the items about the measure of central tendency. But the distributions may differ in nature and composition though they have same central tendency and dispersion.

Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. So skewness shows the lack of symmetry in a distribution. It measures the degree of departure from symmetry. It is used to determine the nature and extent of the concentration of the observations toward higher or lower values of the variable. If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution. If in a distribution mean ≠ median ≠ mode, then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

Scanned by CamScanner

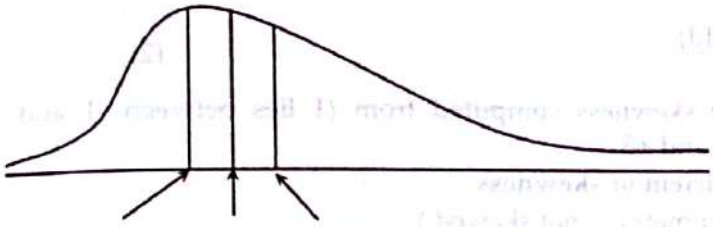
a) Symmetrical distribution



Mean = Median =Mode

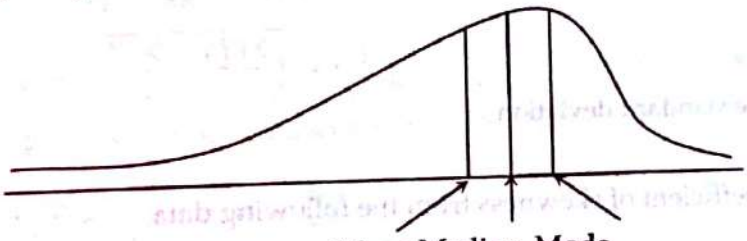
It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve.

b) Positively skewed distribution $\bar{X} > M_d > M_0$.



It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

c) Negatively skewed distribution $M_0 > M_d > \bar{X}$



It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

Properties of Skewness

The distribution is skewed if,

1. The value of mean, median and mode do not coincide i.e. $\bar{x} \neq M_d \neq M_0$
2. Q_1 and Q_3 are not equidistance from median. $M_d - Q_1 \neq Q_3 - M_d$
3. The sum of deviations from median is not equal to zero.
4. Frequencies are not equal in either side of the mode.
5. The curve of frequency distribution is not symmetrical (bell-shaped).

Measures of Skewness

The important measures of skewness are

(i) Karl - Pearason' s coefficient of skewness

(ii) Bowley' s coefficient of skewness

(iii) Measure of skewness based on moments

Scanned by CamScanner

Karl Pearson's absolute measure of skewness

Skewness = Mean - Mode

Skewness = Mean - Median

Karl Pearson's coefficient of skewness(relative measure)

It is also called Karl Pearson's coefficient of skewness. It is the difference of mean and mode divided by standard deviation. It is denoted by S_{KP} and is given by

$$S_{KP} = \frac{\text{Mean - Mode}}{\text{Standard deviation}} = \frac{\bar{X} - M_0}{\sigma} \quad \text{.....(1)}$$

If mode is ill defined then

Mode = 3 Median - 2 Mean

Hence,

$$S_{KP} = \frac{3\,(\text{Mean - Median})}{\text{Standard deviation}} = \frac{3\,(\bar{X} - Md)}{\sigma} \quad \text{.....(2)}$$

The Karl Pearson's coefficient of skewness computed from (1 lies between -1 and +1 but computed from (2) lies between -3 and +3

Interpretation of Karl Pearson's Coefficient of skewness

i. If $S_{KP} = 0$, the distribution is symmetrical(not skewed)

ii. If $S_{KP} > 0$,the distribution is positively skewed(right skewed)

iii. If $S_{KP} < 0$, the distribution is negatively skewed(left skewed)

Note:

When sample data are given

$$S_{KP} = \frac{\bar{x} - M_o}{s}$$

If mode is ill defined

$$S_{KP} = \frac{3\,(\bar{X} - Md)}{s}$$, where s is sample standard deviation

Example 56. Find Karl Pearson's coefficient of skewness from the following data.

2 3 6 9 15

Solution

X	X^2
2	4
3	9
6	36
9	81
15	225
$\sum X = 35$	$\sum X^2 = 355$

Now,

$\sum X = 35, \sum X^2 = 355, n = 5$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{35}{5} = 7$$

$$\sigma = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{355}{5} - \left(\frac{35}{5}\right)^2} = \sqrt{71 - 49} = \sqrt{32} = 5.656$$

Scanned by CamScanner

Here, mode is ill defined.

Hence, median $= \left\{\frac{(n+1)}{2}\right\}^{th}$ item $= \left\{\frac{5+1}{2}\right\}^{th}$ item $= 3^{rd}$ item $= 6$

$$S_{KP} = \frac{3(\bar{X} - Md)}{\sigma} = \frac{3(7-6)}{5.656} = 0.53$$

Here, $S_{Kp} = 0.53 > 0$. Hence the distribution is positively skewed.

Example 57. Find coefficient of skewness from the following data.

X	1	2	3	4	5	6	7	8
f	2	7	9	13	8	4	2	1

Solution

X	f	fX	fX^2
1	2	2	4
2	7	14	28
3	9	27	51
4	13	52	208
5	8	40	200
6	4	24	144
7	2	14	98
8	1	8	64
$\sum X = 36$	$N = \sum f = 46$	$\sum fX = 181$	$\sum fX^2 = 797$

Now,

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{181}{46} = 3.934$$

$$\sigma = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{797}{46} - \left(\frac{181}{46}\right)^2} = \sqrt{17.326 - 15.482} = \sqrt{1.844} = 1.357$$

Mo = 4 {since 4 is repeated maximum times i.e. 13}

$$S_{Kp} = \frac{\bar{X} - M_0}{\sigma} = \frac{3.934 - 4}{1.357} = \frac{-0.066}{1.357} = -0.048$$

Here, $S_{Kp} = -0.048 < 0$.

Hence the distribution is negatively skewed.

Example 58. In a distribution $\bar{X} = 45$, Md = 50, $S_{KP} = -0.5$, find Mo and C.V.

Solution

We know

$$S_{Kp} = \frac{3(\bar{X} - Md)}{\sigma}$$

or $-0.5 = \frac{3(45 - 50)}{\sigma}$

or $\sigma = \frac{-15}{-0.5} = 30$

Now, $M_0 = 3Md - 2\bar{X} = 3 \times 50 - 2 \times 45 = 150 - 90 = 60$

$$C.V. = \frac{\sigma}{\bar{X}} \times 100\% = \frac{30}{45} \times 100\% = 66.66\%$$

Scanned by CamScanner

Hence, $M_0 = 60$ and C.V.= 66.66%.

Example 59. For a group of 10 items we found that $\sum X = 452$, $\sum X^2 = 24270$, $M_0 = 43.7$. Find Karl Pearson's coefficient of skewness and coefficient of variation.

Solution

$$\bar{X} = \frac{\Sigma X}{n} = \frac{452}{10} = 45.2$$

$$\sigma = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{24270}{10} - \left(\frac{452}{10}\right)^2} = \sqrt{2427 - 2043.04} = \sqrt{383.96} = 19.594$$

$$\text{Now, } S_{Kp} = \frac{\bar{X} - M_0}{\sigma} = \frac{45.2 - 43.7}{19.594} = 0.076$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100\% = \frac{19.594}{45.2} \times 100\% = 43.349\%$$

Here $S_{Kp} = 0.076 > 0$.Hence distribution is positively skewed.

Example 60. For a symmetrical distribution arithmetic mean is 20.33, find the values of median and mode of the distribution.

Solution

For a symmetrical distribution $\bar{X} = Md = Mo$

Hence $Md = \bar{X} = 20.33$ and $Mo = \bar{X} = 20.33$

Example 61. Calculate Karl Pearson's coefficient of skewness from the following data.

Profit(000Rs)	10 - 20	20 - 30	30 - 40	40 -50	50 - 60
No. of companies	5	18	35	20	12

Solution

Profit(000Rs)	No of companies(f)	Mid value (X)	d' = (X-35)/10	fd'	Fd'²
10 - 20	5	15	-2	-10	20
20 - 30	18	25	-1	-18	18
30 - 40	35	35	0	0	0
40 – 50	20	45	1	20	20
50 - 60	12	55	2	24	48
	$N = \sum f = 90$			$\sum fd' = 16$	$\sum fd'^2 = 106$

Here $N = 90$, $A = 35$, $h = 10$, $\sum fd' = 16$ and $\sum fd'^2 = 106$

$$\text{Now, } \bar{X} = A + \frac{\sum fd}{N} \times h$$

$$= 35 + \frac{16}{90} \times 10 = 36.777$$

$$\sigma = \sqrt{\frac{1}{N}\sum fd'^2 - \left(\frac{\sum fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{106}{90} - \left(\frac{16}{90}\right)^2} \times 10 = \sqrt{1.177 - 0.031} \times 10 = \sqrt{1.146} \times 10 = 10.705$$

Modal class = (30 – 40) so that $L = 30$, $h = 10$ $f_0 = 18$, $f_1 = 35$, $f_2 = 20$

$\Delta_1 = f_1 - f_0 = 35 - 18 = 17$

$\Delta_2 = f_1 - f_2 = 35 - 20 = 15$

Scanned by CamScanner

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = 30 + \frac{17}{17+15} \times 10 = 30 + 5.312 = 35.312$$

$$\text{Hence, } S_{KP} = \frac{\bar{X} - M_0}{\sigma} = \frac{36.777 - 35.312}{10.705} = 0.136$$

$S_{Kp} = 0.136 > 0$. Hence the distribution is negatively skewed.

Bowley'sMeasure of Skewness

It is measure of skewness based upon quartiles. It can be used for qualitative data and data present in open end class. It is based upon only 50% of the data.

Bowley's Absolute Measure of Skewness

Skewness = $(Q_3 - Md) - (Md - Q_1) = Q_3 - 2Md + Q_1 = Q_3 + Q_1 - 2Md$

Bowley's coefficient of skewness(relative measure)

It is also called Bowley's coefficient of skewness. As it is based upon quartiles,it is also called quartiles coefficient of skewness. It is denoted by S_{KB} and is given by

$$S_{KB} = \frac{(Q_3 - Md) - (Md - Q_1)}{(Q_3 - Md) + (Md - Q_1)} = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1}$$

It lies between -1 and +1.

Interpretation of Bowley's Coefficient of skewness:

i. If $S_{KB} = 0$, the distribution is symmetrical(not skewed)
ii. If $S_{KB} > 0$,the distribution is positively skewed(right skewed)
iii. If $S_{KB} < 0$, the distribution is negatively skewed(left skewed)

Example 62. Compute Bowley's coefficient of skewness from following data.

7 , 3 , 14 , 11 , 9 , 18 , 20 , 15

Solution: Arranging given data in ascending order

3 , 7 , 9 , 11 , 14 , 15 , 18 , 20

Here n = 8

Q_1 = (n+1)/4th item = (8+1)/4 = 2.25th item
= 2nd item + (3rd item – 2nd item) × 0.25
= 7 + (9-7) × 0.25 = 7.5

Md = 2(n+1)/4th item = 2(8+1)/4 = 2×2.25 = 4.5th item
= 4th item + (5th item – 4th item) × 0.5
= 11 + (14 -11) × 0. 5 = 12.5

Q_3 = 3(n+1)/4th item = 3(8+1)/4 = 3×2.25 = 6.75th item
= 6th item + (7th item – 6th item) × 0.75
= 15 + (18 -15) × 0.75 = 17.25

Now,

$$S_{KB} = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1} = \frac{17.25 + 7.5 - 2 \times 12.5}{17.25 - 7.5} = \frac{-0.25}{9.75} = -0.025$$

$S_{KB} = -0.025 < 0$, hence the distribution is negatively skewed.

Example 63. Calculate the coefficient of skewness based upon quartiles from the following data.

No of children per family	0	1	2	3	4	5	6
No of families	7	10	16	25	18	11	8

Scanned by CamScanner

Solution

No of children per family(X)	No of families(f)	cf
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95
	N = ∑f = 95	

$Q_1 = \frac{(n+1)^{th}}{4}$ item $= \frac{95+1}{4} = \frac{96}{4} = 24^{th}$ item

To find 24th item, cf just greater than 24 is 33 for which corresponding value is 2,so that $Q_1 = 2$.

$Md = \frac{2(n+1)^{th}}{4}$ items $= \frac{2(95+1)}{4} = \frac{2 \times 96}{4} = 48^{th}$ item.

To find 48th item, cf just greater than 48 is 58 for which corresponding value is 3,so that Md = 3

$Q_3 = \frac{3(n+1)^{th}}{4}$ items $= \frac{3(95+1)}{4} = \frac{3 \times 96}{4} = 72^{th}$ item.

To find 72th item, cf just greater than 72 is 76 for which corresponding value is 4,so that $Q_3 = 4$

Now,

$$S_{KB} = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1} = \frac{4 + 2 - 2 \times 3}{4 - 2} = \frac{0}{2} = 0$$

$S_{KB} = 0$, Hence the distribution is symmetrical i.e. there is no skewness.

Example 64. Determine the coefficient of skewness from following data.

Daily expenditure(Rs)	Below20	20 -40	40 -55	55 -65	65-75	75-80	80 andover
No of students	18	25	37	30	18	12	10

Solution

It is open end class hence we can use Bowley's measure of skewness.

Daily expenditure(Rs)	No of students(f)	cf
Below 20	18	18
20 – 40	25	43
40 – 55	37	80
55 – 65	30	110
65 – 75	18	128
75 – 80	12	140
80 and over	10	150
	N = ∑f =150	

Now,

$Q_1 = \left(\frac{N}{4}\right)^{th}$ item $= \frac{150}{4} = 37.5^{th}$item

To find 37.5^{th} item, cf just greater than N/4 is 43 for which corresponding class is (20 – 40)

Hence Q_1 class is (20 – 40)

L = 20,h = 20, f = 25, cf = 18

Scanned by CamScanner

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times h$$

$$= 20 + \frac{37.5 - 18}{25} \times 20 = 35.6$$

$$Md = 2\left(\frac{N}{4}\right)^{th} \text{ item} = \frac{2 \times 150}{4} = 75^{th} \text{ item}$$

To find 75th item, cf just greater than2N/4 is 80 for which corresponding class is (40 – 55)

Hence Md class is (40 – 55).

L = 40, h = 15, f = 37, cf = 43

$$Md = L + \frac{\frac{2N}{4} - cf}{f} \times h = 40 + \frac{75 - 43}{37} \times 15 = 52.972$$

$$Q_3 = 3\left(\frac{N}{4}\right)^{th} \text{ item} = \frac{3 \times 150}{4} = 112.5^{th} \text{ item}$$

To find 112.5th item, cf just greater than $\frac{3N}{4}$ is 128 for which corresponding class is(65 – 75).

Hence Q_3 class is (65 – 75)

L = 65,h = 10, f = 18, cf = 110

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times h = 65 + \frac{112.5 - 110}{18} \times 10 = 66.388$$

Now,

$$S_{KB} = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1} = \frac{66.388 + 35.6 - 2 \times 52.972}{66.388 - 35.60} = \frac{-3.956}{30.788} = -0.128$$

$S_{KB} = -0.128 < 0$.

Hence the distribution is negatively skewed.

Example 65. In a distribution, the difference of the two quartiles is 15 and their sum is 35 with median 20. Find the coefficient of skewness.

Solution

Here, $Q_3 - Q_1 = 15$, $Q_3 + Q_1 = 35$, Md = 20

$$S_{KB} = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1} = \frac{35 - 2 \times 20}{15} = \frac{-5}{15} = -0.333$$

The coefficient of skewness $S_{KB} = -0.333 < 0$, hence the distribution is negatively skewed.

Example 66. For a symmetrical distribution, first quartile is 45 and semi-inter quartile range is 5, find the median.

Solution

Here, $S_{KB} = 0$, $Q_1 = 45$, $Q_3 - Q_1 = 5$, Md= ?

$Q_3 - Q_1 = 5$

or $Q_3 - 45 = 5$

or $Q_3 = 5 + 45 = 50$

We know

$$S_{KB} = \frac{Q_3 + Q_1 - 2Md}{Q_3 - Q_1}$$

Scanned by CamScanner

or $0 = \frac{Q_3 + Q_1 - 2Md}{Q_3 - Q_1}$

or $0 = \frac{50 + 45 - 2Md}{5}$

or $0 = 95 - 2Md$

or $2Md = 95$

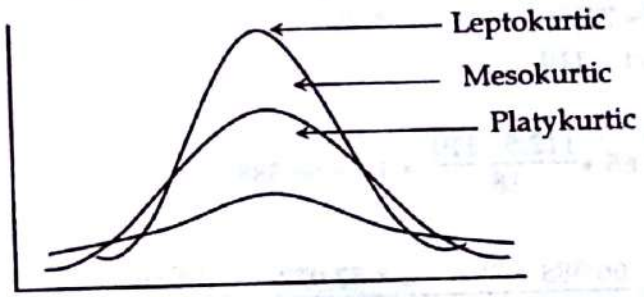
or $Md = \frac{95}{2} = 47.5$

Kurtosis

It is the measure of peakedness or flatness of curve of the given distribution.It gives the shape and nature of the middle part of a symmetrical distribution as compared to normal distribution. Kurtosis is characterized with reference of moderate type of curve(normal).

Types of Kurtosis

i. Leptokurtic curve
ii. Mesokurtic curve
iii. Platykurtic curve



Leptokurtic Curve

Frequency curve of a distribution is more peaked as compared to mesokurtic(normal) curve is called leptokurtic curve.It has very peaked top.

Mesokurtic Curve

Frequency curve of a distribution is neither peaked nor flat is called mesokurtic curve. It is also called normal curve i.e. no kurtosis. Kurtosis is determined with reference to this curve.

Platykurtic Curve

Frequency curve of a distribution is more flat as compared to mesokurtic(normal) curve is called platykurtic curve. It has very flat top.

Measure of Kurtosis

It measures peakness or flatness of the distribution. Different measures of kurtosis are

i. Percentile coefficient of kurtosis
ii. Coefficient of kurtosis based upon moments

Percentile Coefficient of Kurtosis (Kelly's Coefficient of Kurtosis)

It is measure of kurtosis based upon partition values(quartiles and percentiles).It is used f qualitative data and data present in open end class. It is denoted by K and is given by

$$K = \frac{\frac{(Q_3 - Q_1)}{2}}{P_{90} - P_{10}} = \frac{Q_3 - Q_1}{2\,(P_{90} - P_{10})} = \frac{P_{75} - P_{25}}{2\,(P_{90} - P_{10})}$$

Scanned by CamScanner

If K = 0.263, the distribution is mesokurtic (normal)
If K > 0.263, the distribution is leptokurtic
If K < 0.263, the distribution is platykurtic

Example 67. Find percentile coefficient of kurtosis from following data:
3 , 14 , 22 , 7 , 16 , 25 , 11 , 19 , 27

Solution

Arranging given data in ascending order;3 , 7 , 11 , 14 , 16 , 19 , 22 , 25 , 27

n=9

P_{10} = 10(n+1)/100th item = 10(9+1)/100 = 1st item = 3

P_{90} = 90(n+10/100th item = 90(9+1)/100 = 9th item = 27

Q_1 = (n+1)/4th item = (9+1)/4 = 2.5th item = $\frac{2^{nd}\text{ item} + 3^{rd}\text{ item}}{2} = \frac{7+11}{2} = 9$

Q_3 = 3(n+1)/4th item = 3 × (9 + 1)/4 = 7.5th item = $\frac{7^{nd}\text{ item} + 8^{th}\text{ item}}{2} = \frac{22+25}{2} = 23.5$

Now,

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} = \frac{23.5 - 9}{2(27 - 3)} = 0.302$$

Here, K =0.302 >0,263, hence the distribution is leptokurtic.

Example 68. The following data gives the height of a batch of 100 students. Compute the percentile coefficient of kurtosis and comment.

Marks	59	61	63	65	67	69	71	73	75
No of students	0	2	8	20	40	20	8	2	0

Solution

Marks	No of students(f)	cf
59	0	0
61	2	2
63	8	10
65	20	30
67	40	70
69	20	90
71	8	98
73	2	100
75	0	100
	N = ∑f = 100	

Here,

$$P_{10} = P_{25} = \left(\frac{10(N+1)}{100}\right)^{th}\text{ item} = \left(\frac{10(100+1)}{100}\right)^{th} = 10.1^{th}\text{ item}$$

Cf just greater than10 (N+1)/100 is 30 for which corresponding value is 65, hence P_{10}= 65.

$$P_{25} = \left(\frac{25(N+1)}{100}\right)^{th}\text{ item} = \left(\frac{25(100+1)}{100}\right)^{th}\text{ item} = 25.25^{th}\text{ item}$$

Cf just greater than25 (N+1)/100 is 30 for which corresponding value is 65, hence P_{25}= 65.

$$P_{75} = \left(\frac{75(N+1)}{100}\right)^{th}\text{ item} = \left(\frac{75(100+1)}{100}\right)^{th} = 75.75^{th}\text{ item}$$

Cf just greater than75 (N+1)/100 is 90 for which corresponding value is 69, hence P_{75}= 69.

Scanned by CamScanner

$P_{90} = \left(\frac{90(N+1)}{100}\right)^{th}$ item $= \left(\frac{90(100+1)}{100}\right)^{th} = 90.9^{th}$ item

cf just greater than 90 (N+1)/100 is 98 for which corresponding value is 71, hence $P_{90} = 71$.

Now, $K = \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})} = \frac{69 - 65}{2(71 - 65)} = 0.333$

Here, K = 0.333 >0.263, hence the distribution is leptokurtic.

Example 69. Calculate coefficient of kurtosis from following frequency distribution and interpret the result.

Hourly Remuneration(Rs)	No of workers
Below 150	8
Below 200	22
Below 250	40
Below 300	64
Below 350	80
Below 400	92
Below 450	100

Solution

Hourly Remuneration(Rs)	No of workers (f)	cf
Below 150	8	8
150 – 200	14	22
200 – 250	18	40
250 – 300	24	64
300 – 350	16	80
350 – 400	12	92
400 – 500	8	100
	$N = \sum f = 100$	

It is open end class, hence use the percentile coefficient of kurtosis.

Here $P_{10} = \left(\frac{10(N+1)}{100}\right)^{th}$ item $= 10 \times \frac{100}{100} = 10$

Cf just greater than 10N/100 is 22 for which corresponding class is (150 – 200), hence P_{10} class (150 – 200).

L = 150, h = 50 f = 14, cf =8

$$P_{10} = L + \frac{\frac{10N}{100} - cf}{f} \times h$$

$$= 150 + \frac{10 - 8}{14} \times 50 = 157.142$$

$P_{25} = 25N/100^{th}$ item $= 25 \times \frac{100}{100} = 25$

Cf just greater than 25N/100 is 40 for which corresponding class is (200 – 250), hence P_{25} clas (200 – 250).

L = 200, h = 50 f = 18, cf =22

$$P_{25} = L + \frac{\frac{25N}{100} - cf}{f} \times h = 200 + \frac{25 - 22}{18} \times 50 = 208.333$$

Scanned by CamScanner

$P_{75} = \frac{75N}{100}$ th item $= 75 \times \frac{100}{100} = 75$

Cf just greater than 75N/100 is 80 for which corresponding class is (300 - 350), hence P_{75} class is (300 - 350).

L = 300, h = 50 f = 16, cf =64

$$P_{75} = L + \frac{\frac{75N}{100} - cf}{f} \times h = 300 + \frac{75 - 64}{16} \times 50 = 334.375$$

$P_{90} = 90N/100^{th}$ item $= 90 \times \frac{100}{100} = 90$

Cf just greater than 90N/100 is 92 for which corresponding class is (350 - 400), hence P_{75} class is (350 - 400)

L = 350, h = 50 f = 12, cf =80

$$P_{90} = L + \frac{\frac{90N}{100} - cf}{f} \times h = 350 + \frac{90 - 80}{12} \times 50 = 391.666$$

Percentile coefficient of kurtosis (K) $= \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})} = \frac{334.375 - 208.333}{2(391.666 - 157.142)} = \frac{126.042}{469.048} = 0.268.$

K = 0.268 > 0.263, hence the distribution is platykurtic.

Example 70. Determine the coefficient of kurtosis of a distribution having tenth percentile 10, twenty fifth percentile 14.17, seventy fifth percentile 22.4 and ninetieth percentile 26.25.

Solution

Here $P_{10} = 10$, $P_{25} = 14.17$, $P_{75} = 22.4$, $P_{90} = 26.25$

Coefficient of kurtosis (K) $= \frac{P_{75} - P_{25}}{2(P_{90} - P_{10})} = \frac{22.4 - 14.17}{2(26.25 - 10)} = 0.253.$

K = 0.253 < 0.263, hence the distribution is platykurtic.

Moments

Moments are arithmetic average of various power of deviations of observations in which deviations are taken from mean(AM) or assumed mean(A).Moments are used to determine characteristics of frequency distribution i.e., central tendency, dispersion, skewness and kurtosis.

Types of Moments

i. Central moments
ii. Raw moments

Central Moments

The arithmetic average of various power of deviations of observations in which deviations are taken from mean(AM) are called central moments. Central moments are also called moments about mean. The rth central moment is denoted by μ_r.

For individual series

$$\mu_r = \frac{\Sigma(X - \bar{X})^r}{n} = r = 1, 2, 3, 4$$

Where n = number of observations, $\bar{X}$ = arithmetic mean

When r = 1, first central moment and is given by $\mu_1 = \frac{\Sigma(X - \bar{X})^1}{n}$

Scanned by CamScanner

When r = 2, second central moment and is given by $\mu_2 = \frac{\Sigma(X - \bar{X})^2}{n}$

When r = 3, third central moment and is given by $\mu_3 = \frac{\Sigma(X - \bar{X})^3}{n}$

When r = 4, fourth central moment and is given by $\mu_4 = \frac{\Sigma(X - \bar{X})^4}{n}$

For discrete/continuous series

$$\mu_r = \frac{\Sigma f(X - \bar{X})^r}{N} = 1, 2, 3, 4$$

Where N = $\sum f$, x = given values of variable for discrete series and mid value for continuous series, $\bar{X}$ = arithmetic mean

When r = 1, first central moment and is given by $\mu_1 = \frac{\Sigma f (X - \bar{X})^1}{N}$

When r = 2, second central moment and is given by $\mu_2 = \frac{\Sigma f (X - \bar{X})^2}{N}$

When r = 3, third central moment and is given by $\mu_3 = \frac{\Sigma f (X - \bar{X})^4}{N}$

When r = 4, fourth central moment and is given by $\mu_4 = \frac{\Sigma f (X - \bar{X})^4}{N}$

Example 71. Calculate the first four central moments for the following set of numbers 12, 14, 16,18 and 20.

Solution

X	$(X-\bar{X})$	$(X-\bar{X})^2$	$(X-\bar{X})^3$	$(X-\bar{X})^4$
12	-4	16	-64	256
14	-2	4	-8	16
16	0	0	0	0
18	2	4	8	16
20	4	16	64	256
$\sum X = 80$	$\sum(X-\bar{X}) = 0$	$\sum(X-\bar{X})^2 = 40$	$\sum(X-\bar{X})^3 = 0$	$\sum(X-\bar{X})^4 = 544$

Here $\bar{X} = \frac{\Sigma X}{n} = \frac{80}{5} = 16$

Now first four central moments are

$$\mu_1 = \frac{\Sigma(X - \bar{X})}{n} = \frac{0}{5} = 5 \qquad \mu_2 = \frac{\Sigma(X - \bar{X})^2}{n} = \frac{40}{5} = 8$$

$$\mu_3 = \frac{\Sigma(X - \bar{X})^3}{n} = \frac{0}{5} = 0 \qquad \mu_4 = \frac{\Sigma(X - \bar{X})^4}{n} = \frac{544}{5} = 108.88$$

Example 72. Compute first four moments about mean from following data of daily wages of 125 workers.

Income (000Rs)	0	1	2	3	4	5	6	7	8
No of workers	5	10	15	20	25	20	15	10	5

Scanned by CamScanner

Solution

Income (000Rs) (X)	No of workers (f)	fX	$(X-\bar{X})$	$f(X-\bar{X})$	$f(X-\bar{X})^2$	$f(X-\bar{X})^3$	$f(X-\bar{X})^4$
0	5	0	-4	-20	80	-320	1280
1	10	10	-3	-30	90	-270	810
2	15	30	-2	-30	60	-120	240
3	20	60	-1	-20	20	-20	20
4	25	100	0	0	0	0	0
5	20	100	1	20	20	20	20
6	15	90	2	30	60	120	240
7	10	70	3	30	90	270	810
8	5	40	4	20	80	320	1280
	N =∑f = 125	∑fX= 500		$\sum f(X-\bar{X})= 0$	$\sum f(X-\bar{X})^2$ =500	$\sum f(X-\bar{X})^3$ =0	$\sum f(X-\bar{X})^4$ =4700

Here $\bar{X} = \frac{\Sigma fX}{N} = \frac{500}{125} = 4$

First four moments about mean

$$\mu_1 = \frac{\Sigma f(X-\bar{X})}{N} = \frac{0}{125} = 0 \qquad \mu_2 = \frac{\Sigma f(X-\bar{X})^2}{N} = \frac{500}{125} = 4$$

$$\mu_3 = \frac{\Sigma f(X-\bar{X})^3}{N} = \frac{0}{125} = 0 \qquad \mu_4 = \frac{\Sigma f(X-\bar{X})^4}{N} = \frac{4700}{125} = 37.6$$

Example 73. Obtain first four central moments from following data.

No of hrs worked	1 – 3	3 – 5	5 – 7	7 – 9
No of days	3	5	1	1

Solution

No of hrs worked	No of days(f)	Mid hrs worked(X)	fX	$X-\bar{X}$	$f(X-\bar{X})$	$f(X-\bar{X})^2$	$f(X-\bar{X})^3$	$f(X-\bar{X})^4$
1 – 3	3	2	6	-2	-6	12	-24	48
3 – 5	5	4	20	0	0	0	0	0
5 – 7	1	6	6	2	2	4	8	16
7 – 9	1	8	8	4	4	16	64	256
	N =∑f =10		∑fX =40		$\sum f(X-\bar{X})$ =0	$\sum f(X-\bar{X})^2$= 32	$\sum f(X-\bar{X})^3$ =48	$\sum f(X-\bar{X})^4$ = 320

Here $\bar{X} = \frac{\Sigma fX}{N} = \frac{40}{10} = 4$

First four moments about mean

$$\mu_1 = \frac{\Sigma f(X-\bar{X})}{N} = \frac{0}{10} = 0 \qquad \mu_2 = \frac{\Sigma f(X-\bar{X})^2}{N} = \frac{32}{10} = 3.2$$

$$\mu_3 = \frac{\Sigma f(X-\bar{X})^3}{N} = \frac{48}{10} = 4.8 \qquad \mu_4 = \frac{\Sigma f(X-\bar{X})^4}{N} = \frac{320}{10} = 32$$

Scanned by CamScanner

Raw Moments

The arithmetic average of various power of deviations of observations in which deviations are taken from assumed mean (A) are called raw moments. Raw moments are also called moments about an arbitrary point. The rth raw moment is denoted by μ_r'

For individual series

$$\mu_r' = \frac{\sum f(X-A)^r}{N} = \frac{\sum fd^r}{n}, \quad r = 1, 2, 3, 4$$

Where n = number of observations, A = assumed mean, d = X – A

When r= 1, first raw moment and is given by $\mu_1' = \frac{\sum f(X-A)^1}{n} = \frac{\sum fd}{n}$

When r= 2, second raw moment and is given by $\mu_2' = \frac{\sum f(X-A)^2}{n} = \frac{\sum fd^2}{n}$

When r= 3, third central moment and is given by $\mu_3' = \frac{\sum f(X-A)^3}{n} = \frac{\sum fd^3}{n}$

When r= 4, fourth raw moment and is given by $\mu_4' = \frac{\sum f(X-A)^4}{n} = \frac{\sum fd^4}{n}$

For discrete series:

$$\mu_r' = \frac{\sum f(X-A)^r}{N} = \frac{\sum fd^r}{N}, \quad r = 1, 2, 3, 4$$

Where N = $\sum f$, A = assumed mean, d = X– A

When r= 1, first raw moment and is given by $\mu_1' = \frac{\sum f(X-A)^1}{N} = \frac{\sum fd}{N}$

When r= 2, second raw moment and is given by $\mu_2' = \frac{\sum f(X-A)^2}{N} = \frac{\sum fd^2}{N}$

When r= 3, third central moment and is given by $\mu_3' = \frac{\sum f(X-A)^3}{N} = \frac{\sum fd^3}{N}$

When r= 4, fourth raw moment and is given by $\mu_4' = \frac{\sum f(X-A)^4}{N} = \frac{\sum fd^4}{N}$

For continuous series:

$$\mu_r' = \frac{\sum f(X-A)^r}{N} = \frac{\sum f(hd')^r}{N} = \frac{h^r \sum f(d')^r}{N}, \quad r = 1, 2, 3, 4$$

Where N = $\sum f$, X = mid value, A = assumed mean, $d' = \frac{X-A}{h}$, X – A = hd', h = Class interval

When r= 1, first raw moment and is given by $\mu_1' = \frac{\sum f(X-A)^1}{N} = \frac{h \sum fd'}{N}$

When r= 2, second raw moment and is given by $\mu_2' = \frac{\sum f(X-A)^2}{N} = \frac{h^2 \sum fd'^2}{N}$

When r= 3, third central moment and is given by $\mu_3' = \frac{\sum f(X-A)^3}{N} = \frac{h^3 \sum fd'^3}{N}$

When r= 4, fourth raw moment and is given by $\mu_4' = \frac{\sum f(X-A)^4}{N} = \frac{h^4 \sum fd'^4}{N}$

Example 74: Find first four moments about origin for the set of numbers 2, 3, 4, 5 and 6.

Scanned by CamScanner

Solution

X	d = X-0	d^2	d^3	d^4
2	2	4	8	16
3	3	9	27	81
4	4	16	64	256
5	5	25	125	625
6	6	36	216	1296
	$\sum d = 20$	$\sum d^2 = 90$	$\sum d^3 = 440$	$\sum d^4 = 2274$

Now,

$$\mu'_1 = \frac{\Sigma d}{n} = \frac{20}{5} = 4 \qquad \mu'_2 = \frac{\Sigma d^2}{n} = \frac{90}{5} = 18$$

$$\mu'_3 = \frac{\Sigma d^3}{n} = \frac{440}{5} = 88 \qquad \mu'_4 = \frac{\Sigma d^4}{n} = \frac{2274}{5} = 454.8$$

Example 75. Compute first four moments about arbitrary point 5 from following information.

Age (year)	3	4	5	6	7
No of children	12	7	15	11	5

Solution

Age (X)	No of children (f)	d= X- 5	fd	fd^2	fd^3	fd^4
3	12	-2	-24	48	-96	192
4	7	-1	-7	7	-7	7
5	15	0	0	0	0	0
6	11	1	11	11	11	11
7	5	2	10	20	40	80
	$N = \sum f = 50$		$\sum fd = -10$	$\sum fd^2 = 86$	$\sum fd^3 = -52$	$\sum fd^4 = 290$

Now,

$$\mu'_1 = \frac{\Sigma fd}{N} = \frac{-10}{50} = -0.2 \qquad \mu'_2 = \frac{\Sigma fd^2}{N} = \frac{86}{50} = 1.72$$

$$\mu'_3 = \frac{\Sigma fd^3}{N} = \frac{-52}{50} = -1.04 \qquad \mu'_4 = \frac{\Sigma fd^4}{N} = \frac{290}{50} = 5.8$$

Example 76. Calculate first four moments about an arbitrary point from the following data.

Class interval	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100
frequency	5	12	20	7	6

Solution

Let arbitrary point A = 75

Class	Frequency (f)	Mid value(X)	$d' = \frac{X-75}{10}$	fd'	fd'^2	fd'^3	fd'^4
50 – 60	5	55	-2	-10	20	-40	80
60 – 70	12	65	-1	-12	12	-12	12
70 –80	20	75	0	0	0	0	0
80 – 90	7	85	1	7	7	7	7
90 – 100	6	95	2	12	24	48	96
	$N = \sum f = 50$			$\sum fd' = -3$	$\sum fd'^2 = 63$	$\sum fd'^3 = 3$	$\sum fd'^4 = 195$

Now,

Scanned by CamScanner

$$\mu'_1 = h\frac{\Sigma fd'}{N} = 10 \times \frac{-3}{50} = -0.6$$

$$\mu'_2 = h^2\frac{\Sigma fd'^2}{N} = 10^2\frac{63}{50} = 128$$

$$\mu'_3 = h^3\frac{\Sigma fd'^3}{N} = 10^3\frac{63}{50} = 126$$

$$\mu'_4 = h^4\frac{\Sigma fd'^4}{N} = 10^4\frac{195}{50} = 39000$$

Relationship between Central Moments and Raw Moments

If the arithmetic mean(AM) is in fraction,it becomes tedious to compute moments about mean by direct method. In such case we compute raw moments and convert into central moments by using the relationship of raw moments and central moments. The relationship are as follows;

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$
$$\mu_2 = \mu'_2 - \mu'^2_1$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$$
$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$$

Example 77. The first four moments about arbitrary point 4 are 1, 4, 10 and 45.Find first four moments about mean.

Solution

Here, $A = 4$, $\mu'_1 = 1$, $\mu'_2 = 4$, $\mu'_3 = 10$ and $\mu'_4 = 45$

Now first four moments about mean are

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$
$$\mu_2 = \mu'_2 - \mu'^2_1 = 4 - 1^2 = 3$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = 10 - 3\times 4 \times 1 + 2 \times 1^3 = 10 - 12 + 2 = 0$$
$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 45 - 4 \times 10 \times 1 + 6 \times 4 \times 1^2 - 3 \times 1^4 = 45 - 40 + 24 - 3 = 26$$

Measure of Skewness based upon Moments

Karl pearson defined the coefficient of skewness based upon second central moment and third central moment.Karl pearson defined following coefficient of skewness based upon moments.

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

Since μ_3^2 and μ_2^3 are always non negative, so that $\beta_1 \geq 0$. It is the limitation of the measure of skewness based upon moments. To remove the limitation Karl pearson defined coefficient of skewness as positive square root of β_1 and is denoted by γ_1.

$$\text{Hence } \gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\mu_2^{3/2}}$$

If $\mu_3 = 0$; $\beta_1 = 0$ and $\gamma_1 = 0$, the distribution is symmetrical.

{Use either β_1 or γ_1}

If $\mu_3 > 0$; $\beta_1 > 0$ and $\gamma_1 > 0$, the distribution is positively skewed.

{Use either β_1 or γ_1}

If $\mu_3 < 0$;$\gamma_1 < 0$, the distribution is negatively skewed.

{Use γ_1}

Scanned by CamScanner

Example 78. The first four moments about mean for a distribution are 0, 5.97, –0.397 and 89.22 respectively. Compute the coefficient of skewness.

Solution

Here $\mu_1 = 0$, $\mu_2 = 5.97$, $\mu_3 = -0.397$ and $\mu_4 = 89.22$

Coefficient of skewness$(\gamma_1) = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-0.397}{(5.97)^{3/2}} = \frac{-0.397}{14.586} = -0.027$

$\gamma_1 = -0.27 < 0$, hence the distribution is negatively skewed.

Measure of Kurtosis based upon Moments

Karl pearson defined coefficient of kurtosis based upon second central moment and fourth central moment.Karl pearson defined following coefficient of kurtosis based upon moments.

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

and $\gamma_2 = \beta_2 - 3$

If $\beta_2 = 3$, $\gamma_2 = 0$; the distribution is mesokurtic (normal)

{Use either β_2 or γ_2}

If $\beta_2 > 3$, $\gamma_2 > 0$; the distribution is leptokurtic

{Use either β_2 or γ_2}

If $\beta_2 < 3$, $\gamma_2 < 0$; the distribution is platykurtic

{Use either β_2 or γ_2}

Example 79. In a certain distribution the first four moments about arbitrary point 4 are 1.5, 17, –30 and 108 respectively. Compute the measure of kurtosis and comment.

Solution

Here $A = 4$, $\mu'_1 = 1.5$, $\mu'_2 = 17$, $\mu'_3 = -30$ and $\mu'_4 = 108$

To compute measure of kurtosis

$\mu_2 = \mu'_2 - \mu'^2_1 = 17 - (1.5)^2 = 17 - 2.25 = 14.75$

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$

$= 108 - 4 \times (-30) \times 1.5 + 6 \times 17 \times (1.5)^2 - 3 \times (1.5)^4$

$= 108 + 180 + 229.5 - 15.187$

$= 502.313$

Now, coefficient of kurtosis $(\beta_2) = \frac{\mu_4}{\mu_2^2} = \frac{502.313}{(14.75)^2} = 2.308$

$\beta_2 = 2.308 < 3$, hence the distribution is platykurtic.

Example 80. The standard deviation of a symmetrical distribution is 5. What must be the value of the fourth moment about mean so that the distribution is (i) leptokurtic (ii) mesokurtic (iii) platykurtic.

Solution

Here, $\sigma = 5$

or $\mu_2 = \sigma^2 = 5^2 = 25$

(i) For leptokurtic distribution $\beta_2 > 3$

or $\frac{\mu_4}{\mu_2^2} > 3$

or $\frac{\mu_4}{25^2} > 3$

or $\mu_4 > 625 \times 3$

Scanned by CamScanner

or $\mu_4 > 1875$

(ii) For mesokurtic distribution, $\beta_2 = 3$

(iii) For platykurtic distribution, $\beta_2 < 3$

$\frac{\mu_4}{\mu_2^2} = 3$ or $\frac{\mu_4}{\mu_2^2} < 3$

or $\frac{\mu_4}{25^2} = 3$ or $\frac{\mu_4}{25^2} < 3$

or $\mu_4 = 625 \times 3$ or $\mu_4 < 625 \times 3$

or $\mu_4 = 1875$ or $\mu_4 < 1875$

Characteristics of Frequency Distribution using Moments

Using the moments characteristics of frequency distribution i.e. central tendency, dispersion, skewness and kurtosis can be determined as

Central tendency = AM = $\bar{X} = A + \mu_1'$

A = Assumed mean, μ'_1 is first raw moment

Dispersion = SD = $\sqrt{\mu_2}$

Skewness = $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$

Kurtosis = $\beta_2 = \frac{\mu_4}{\mu_2}$

Characteristics of frequency distribution using partition values

Using the partition values characteristics of frequency distribution i.e. central tendency, dispersion, skewness and kurtosis can be determined as

Central tendency = Md

Dispersion = QD = $\frac{Q_3 - Q_1}{2}$; Skewness = $S_{KB} = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1}$; Kurtosis = $K = \frac{Q_3 - Q_1}{2\,(P_{90} - P_{10})}$

Example 81. The marks obtained by 70 students in Statistics are given below.

31	56	63	64	33	18	92	48	23	63
27	78	66	43	45	75	85	81	80	78
40	68	45	49	44	30	29	78	5	74
82	9	5	9	75	11	58	24	76	92
47	94	23	23	80	50	65	63	95	86
57	70	66	18	28	77	3	12	80	2
9	47	43	73	31	5	22	69	48	42

Find:

i) Mean, median, mode, quartiles, standard deviation, and coefficient of variation, Karl Pearson coefficient of skewness, Bowley's coefficient of skewness, Kelly's coefficient of skewness, and coefficient of kurtosis for the given individual data.

ii) Present the given data in frequency distribution with appropriate class interval.

iii) Present the given data in frequency distribution with class size of 10.

iv) Also calculate Mean and standard deviation. Are they different from that obtained using individual data?

Solution

(Calculation is left as exercise)

Here, for given data

Scanned by CamScanner

n	(Number of observation)	70
mean	(Arithmetic mean)	50.04286
median		49.5
mode		63
Q_1	First Quartile	27.75
Q_3	Third Quartile	74.25
P_{10}	Tenth Percentile	9
P_{90}	90th percentile	81.9
Q.D.	Quartile deviation	23.25
σ	Standard deviation	26.65142
C.V.	Coefficient of variation	53.25719
Range		93
max	Maximum value	95
min	Minimum value	2
S_k	Karl Pearson's coefficient of skewness	-0.48617
S_b	Bowley's coefficient of skewness	0.064516
Kelly sk	Kelly's coefficient of skewness	-0.11111
K	Kelly's coefficient of kurtosis	0.318

For construction of frequency distribution we can use the relation,

$$h = \frac{\text{Range}}{\text{Number of classes}} = \frac{\text{Max} - \text{Min}}{1 + 3.322 \log_{10}(n)}$$

The required appropriate class width is h = 13.1 for number of classes taken as $1 + 3.322 \log_{10}(n) = 8$ (approx.)

The frequency distribution with class width of 12.5 will be as follows:

Class	0-12.5	12.5-25	25-37.5	37.5-50	50-62.5	62.5-75	75-87.5	87.5-100	Total
Frequency	10	6	7	13	3	16	11	4	70

Similarly, the frequency distribution of class size 10 is as follows:

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	8	4	8	4	12	3	12	11	4	4

For the grouped data

Mean	49.57143
Standard Deviation	53.38367

The difference occurred due to mid value is used to calculate the mean and standard deviation

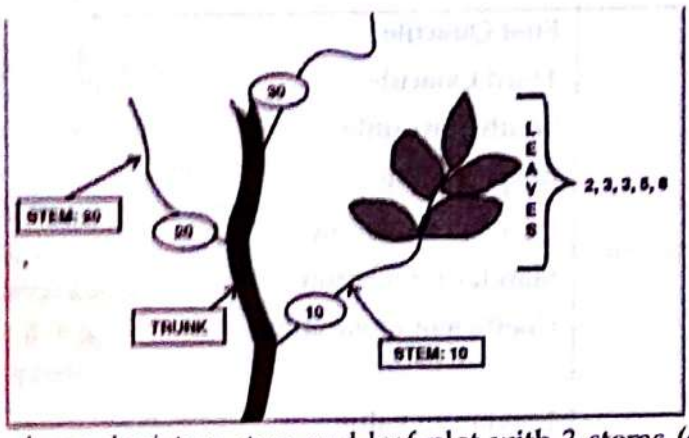
Stem Leaf Display

Raw data refers to data that has been collected or recorded by the researcher and it is usually in no particular order. It is difficult to see any pattern or read 'information' from the data. To obtain the meaningful information, we can sort categorical data according to the its characteristics such as gender, computer makes etc., but for quantitative data it is necessary to display charts after sorting the data.

For example, the following is the raw data showing the number of MP3 players sold daily.
30, 11, 29, 34, 54, 36, 49, 31, 42, 45, 25, 25, 15, 18, 13, 25, 13, 55, 55, 38, 31, 43, 38, 22, 37, 20, 36

Scanned by CamScanner

From above data, it is difficult to see any pattern or read 'information' from the data.
One of the easiest way to display the number data after sorting is Stem leaf plot (Display).
Stem leaf consisting of stems as the main groups while the leaves are elements within each m group/stem.



For example, the picture above depicts a stem-and-leaf plot with 3 stems (or main groups): 1; 2; The stem,1 (whose value is 10), has leaves 2; 3; 3, 5; 8 representing the numbers 12; 13; 13; 15; 18.. I not necessary to 'draw' a stem-and-leaf picture as above. The sorted data may be presented in a ta as illustrated in the following table.
Identification of stem and leaf from the data: (Lets take first five data from example)

Stem=3 Leaf = 0	Stem=1 Leaf = 1	Stem=2 Leaf = 9	Stem=3 Leaf = 4	Stem=5 Leaf = 5
30,	11,	29,	34,	54

Normally 5 to 20 categories or steam is used while construction of stem leaf display. Too few st gives very little information while more stem gives scrappy and disjointed information tha difficult to read. So in this example the appropriate choice of stem would be 1,2,3,4 and 5 hav values 10, 20, 30, 40 and 50. Arranging these information in tabular form gives the pictorial as wel numerical information.

Stem (tens)	Leaves (units)									Read out as
5	4	5	5							54, 55, 55
4	2	4	5	9						42, 44, 45, 49
3	0	1	1	4	6	6	7	8	8	30, 31, 31, 34, 36, 36, 37, 38, 38
2	0	2	5	5	5	9				20, 22, 25, 25, 29
1	1	3	3	5	8					11, 13, 13, 15, 18

From the above table one can get the information about the range of values from lowest to hig and the main group where more frequency occurs. In above example the biggest single group of 3 Between 30 and 39 MP3 players were sold on nine days out of 27 days.

Steps to follow while constructing stem leaf display

(1) Divide each number into two parts: a stem, consisting of one or more of the leading digits, a a leaf, consisting of the remaining digit.
(2) List the stem values in a vertical column.
(3) Record the leaf for each observation beside its stem.
(4) Write the units for stems and leaves on the display.

Scanned by CamScanner

Sometimes stem-and-leaf plots are used to compare two samples. We can put two leaves on the same stem to compare two samples.

Example samples of round-trip transit times (known as pings) received from two locations.

Location I: 0.015, 0.039, 0.035, 0.048, 0.041, 0.033, 0.054, 0.035, 0.028, 0.021, 0.030, 0.032, 0.021, 0.043, 0.048 seconds

Location II: 0.029, 0.067, 0.038, 0.078, 0.046, 0.071, 0.004, 0.016, 0.066, 0.010, 0.019, 0.003 seconds

Choosing a leaf unit of 0.001, a stem unit of 0.01, we construct the following two stem-and-leaf plots, one to the left and one to the right of the stem.

Stem Leaf Display

Location I Leaves (0.001)						Stem (0.01)	Location II Leaves (0.001)		
						0	3	4	
					5	1	0	6	9
			1	1	8	2			
0	2	3	5	5	9	3	8		
		1	3	8	8	4	6		
					4	5			
						6	1	6	7
						7	8		

On observing the stem leaf display, one can say that the average ping is almost same but the variability is more in case of second location.

Example 82. The following scores represent the final examination grades for statistics course:

23 60 79 32 57 74 52 70 82 36 80 77 81 95 41 65 92 85 55 76 52 10 64 75 78 25 80 98 81 67 41 71 83 54 64 72 88 62 74 43 60 78 89 76 84 48 84 90 15 79 34 67 17 82 69 74 63 80 85 61.

Construct a stem-and-leaf plot for the examination grades in which the stems are 1,2,3,...,9.

Stem (tens)	Leaves (units)	Frequency
1	0 5 7	3
2	3 5	2
3	2 4 6	3
4	1 1 3 8	4
5	2 2 4 5 7	5
6	0 0 1 2 3 4 4 5 7 7 9	11
7	0 1 2 4 4 4 5 6 6 7 8 8 9 9	14
8	0 0 0 1 1 2 2 3 4 4 5 5 8 9	14
9	0 2 5 8	4

Example 83. The following are historical data on staff salaries for 30 schools sampled in the early 1970s.

3.79, 2.99, 2.77, 2.91, 3.10, 1.84, 2.52, 3.22, 2.45, 2.14, 2.67, 2.52, 2.71, 2.75, 3.57, 3.85, 3.36, 2.05, 2.89, 2.83, 3.13, 2.44, 2.10, 3.71, 3.14, 3.54, 2.37, 2.68, 3.51, 3.37

Scanned by CamScanner

Construct a stem-and-leaf display of the data.

Stem (Unit)	Leaves (units)											Frequency
1U	(84)											1
2L	(05)	(10)	(14)	(37)	(44)	(45)						6
2U	(52)	(52)	(67)	(68)	(71)	(75)	(77)	(83)	(89)	(91)	(99)	11
3L	(10)	(13)	(14)	(22)	(36)	(37)						6
3U	(51)	(54)	(57)	(71)	(79)	(85)						6

Note: In stem unit U indicates Upper half (i.e. .50 and above upto .99) and L indicates Lower half (i.e. .01 and above up to .49)

Five Number Summary

The five number summaries is a descriptive statistical tool that provides information about a set of observations and includes the five descriptive measures of given data set. Three quartiles together with the low and high data values give us a very useful look at the data and their spread.

Therefore, the five number summary of a set of observations on a single variable consists of the following statistics:

Maximum (max) – the largest observation

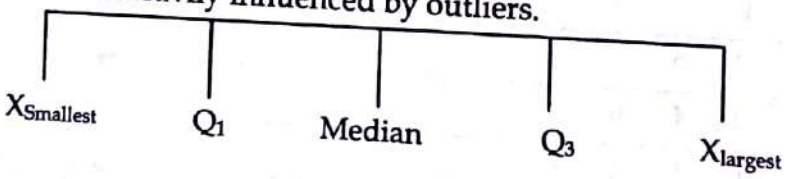
Upper Quartile (Q_3) – a value that separates the largest 25% of the observations from the smallest 75%

Median (M_d) – a value that separates the largest 50% of the observations from the smallest 50%

Lower Quartile (Q_1) – a value that separates the largest 75% of the observations from the smallest 25%.

Minimum (min) – the smallest observation

The mean and standard deviation can also be used together to convey the center and the spread of a set of data. However, both of these statistics are susceptible to *outliers*. The median, first quartile and third quartile are not as heavily influenced by outliers.



The five number summary is very helpful to determine the shape of the distribution, that is to recognize symmetry or not in the data.

The distribution is positively skewed if

i) The distance from the median (m_d) to the third quartile(Q_3) is greater than the distance from the median (m_d) to the first quartile (Q_1).

ii) The distance from the median (m_d) to the largest value is greater than the distance from the median (m_d) to the smallest value of the data.

iii) The distance from the third Quartile (Q_3) to the largest value is greater than the distance from the first quartile (Q_1) to the smallest value.

The distribution is negatively skewed if

i) The distance from the median (m_d) to the third quartile(Q_3) is less than the distance from the median (m_d) to the first quartile (Q_1).

ii) The distance from the median (m_d) to the largest value is less than the distance from the median (m_d) to the smallest value of the data.

Scanned by CamScanner

iii) The distance from the third quartile(Q_3) to the largest value is less than the distance from the first quartile (Q_1) to the smallest value.

The distribution is perfectly symmetrical if

i) The distance from the median (m_d) to the third quartile (Q_3) is equal to the distance from the median (m_d) to the first quartile (Q_1).

ii) The distance from the median (m_d) to the largest value is equal to the distance from the median (m_d) to the smallest value of the data.

iii) The distance from the third quartile (Q_3) to the largest value is equal to the distance from the first quartile (Q_1) to the smallest value.

Example 84. Calculate the five number summary from the following data.

1, 2, 2, 3, 4, 6, 6, 7, 7, 7, 8, 11, 12, 15, 15, 15, 17, 17, 18, 20

There are a total of twenty points in the data set. The median is thus the average of the tenth and eleventh data values or:

Here size of median $=\left(\frac{N+1}{2}\right)^{th}$ item$=\left(\frac{21}{2}\right)^{th}$ item $= 10.5^{th}$ item

Or, Median $= 10^{th} + (11^{th} - 10^{th}$ item$) \times 0.5 = 7 + (8 - 7) \times 0.5 = 7.5$.

Similarly, size of $Q_1=\left(\frac{N+1}{4}\right)^{th}=\left(\frac{21}{4}\right)^{th} = 5.25^{th}$ item

$Q_1 = 5^{th} + (6^{th} - 5^{th}$ item$) \times 0.25 = 4 + (6-4) \times 0.25 = 5$.

And, size of $Q_3=\left(3\frac{N+1}{4}\right)^{th}=\left(\frac{63}{4}\right)^{th} = 15.75^{th}$ item

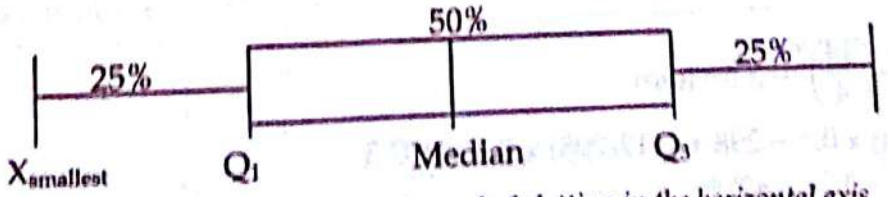
$Q_1 = 15^{th} + (16^{th} - 15^{th}$ item$) \times 0.75 = 15 + (15 - 15) \times 0.75 = 15$.

We assemble all of the above results together, and report that the five number summaries for the above set of data is 1, 5, 7.5, 12, 20.

Box Plot

Box Plot use a Five-Number Summary to create a graphic sketch of the data. Box Plot is also known as Box and Whisker Plot. Box Plot gives a graphical representation of the data set contained between the upper and lower limits. This plot determines the degree of symmetry (or skewness) based on the distances that separate the five numbers. To compare these distances effectively, we can create a box plot the vertical line within the box represent the location of median value in the data. The vertical line at the left side of the box represent the location of Q1 and the vertical line at the right side of the box represent the location of Q3. The box extends from Q1 to Q3. The box contains the middle 50% of the observations in the distribution. The lower 25% of the data are represented by the line connecting to the left side of the box to the smallest value and the upper 25% of the data are represented by the by the line connecting the right side of the box to the largest value and are known as whiskers.

Box plot



Note: The box plot can also be plotted in the vertical axis instead of plotting in the horizontal axis

Scanned by CamScanner

Outliers

An outlier is an observation point that is distant from concentration of other observation. It may be caused due to variability of the data, experimental errors, it could be unusual in respect of the combination of values of several variables etc.

They have to be detected since they have strong influence on the estimates of the parameters of a model that is being fitted to the data. This could lead to inaccurate prediction and conclusions.

Outliers can be determined using box plot, regression analysis, scatter plot etc.

Outliers in data using Box plot.

While using Box plot data values outside 1.5 times the inter quartile range below lower quartile and above upper quartile is considered as outliers i.e., the values less than $Q_1 - 1.5 \times$ I.R. and larger than $Q_3 + 1.5 \times$ I.R.

Example 85. Find outliers of the following data:

3 4 4 4 7 10 11 12 14 16 17 39

Here, minimum value = 3, $Q_1 = 4$, Median = 10.5, $Q_3 = 15.5$ and Maximum value is 39

But Inter quartile range = I.R. = $Q_3 - Q_1 = 11.5$

Then, $1.5 \times \text{I.R} = 1.5 \times 11.5 = 17.25$

$Q_1 - 1.5 \times \text{I.R.} = 4 - 17.25 = -13.25$

$Q_3 + 1.5 \times \text{I.R.} = 15.5 + 17.25 = 32.75$

Since, value 39 is larger than 32.75, it is considered as outliers.

The following box plot gives the value 39 away from its total range of [3, 17], such values have to be eliminated before estimating population parameters.

Example 86. A manufacturer of battery looks a sample of 13 batteries from a days production and used them continuously until they were drained. Number of hours they were used until failure were. 342, 426, 317, 545, 264, 451, 1049, 631, 512, 266, 492, 562, 298.

a) List the five number summary
b) Construct a box plot for the data
c) Are the data skewed? If so how?

Solutions:

Arranging the given data in ascending order as follows:

264, 266, 298, 317, 342, 426, 451, 492, 512, 545, 562, 631, 1049

(a) Calculation of Five number summary

$X_{smallest} = 264$ $\qquad$ $X_{largest} = 1049$

Here size of median $= \left(\frac{13+1}{2}\right)^{th}$ item $= \left(\frac{14}{2}\right)^{th}$ item $= 7^{th}$ item

7th item is Median = 451

Similarly,

Size of $Q_1 = \left(\frac{N+1}{4}\right)^{th} = \left(\frac{14}{4}\right)^{th} = 3.5^{th}$ item

$Q_1 = 3^{rd} + (4^{th} - 3^{rd}$ item$) \times 0.5 = 298 + (317-298) \times 0.5 = 307.5$

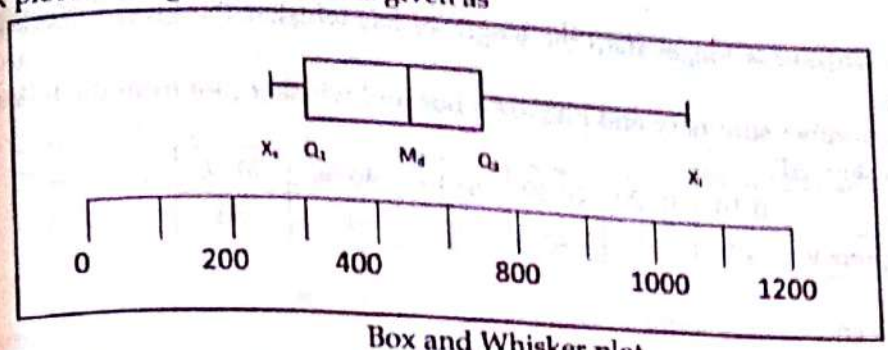
And, size of $Q_3 = \left(3\frac{N+1}{4}\right)^{th} = \left(\frac{42}{4}\right)^{th} = 10.5^{th}$ item

$Q_3 = 10^{th} + (11^{th} - 10^{th}$ item$) \times 0.5 = 545 + (562 - 545) \times 0.5 = 553.5$

The five number summary for the above set of data is {264, 307.5, 451, 553.5, 1049}

Scanned by CamScanner

b) The box plot for the given data set is given as



Box and Whisker plot

c) The distance from the smallest value to median = 451 - 264 = 187
The distance from median to the largest value = 1049 - 451 = 598
The distance from Q_3 to the largest value = 1049 - 553.5 = 495.5
The distance from Q1 to the smallest value = 307.5 - 264 = 435

598 > 187 & 495.5 > 435 and also since the length of right whisker is longer than the length of left whisker the given distribution is positively (right) skewed.

Example 87. Calculate the five number summary and construct box plot from the following data. Also describe the shape of the distribution.

Income 000 Rs	40	45	50	55	60
No. of family	2	2	5	3	2

Solutions:

Income (000 Rs.)	f	cf
40	2	2
45	2	4
50	5	9
55	3	12
60	2	14
	N=14	

Here, Largest value $X_L = 60$, Smallest value $X_s = 40$

Here size of median $= \left(\frac{14+1}{2}\right)^{th}$ item $= \left(\frac{15}{2}\right)^{th}$ item = 7.5th item

The c f just greater than 7.5 is 9 and median = 50

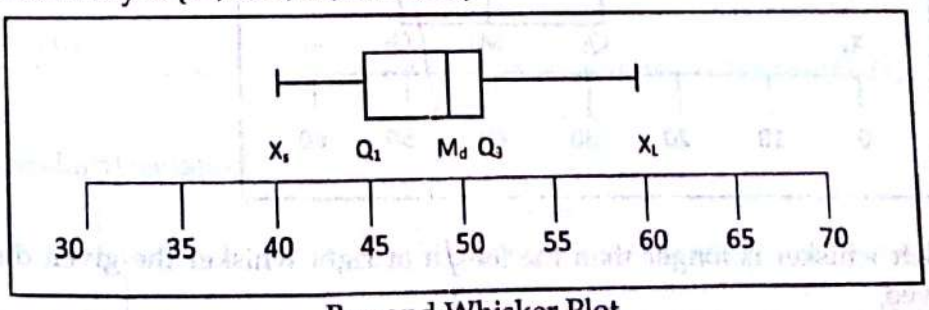
Similarly, size of $Q_1 = \left(\frac{N+1}{4}\right)^{th} = \left(\frac{15}{4}\right)^{th}$ = 3.75th item

The c f just greater than 3.75 is 4 and $Q_1 = 45$

And, size of $Q_3 = \left(3\frac{N+1}{4}\right)^{th} = \left(\frac{45}{4}\right)^{th}$ = 11.25th item

Q_3 = 11th + (12th – 11th item) × 0.25 = 50 + (55 – 50) × 0.25 =51.25

The five number summary is {40, 45.5, 50, 51.25, 60}



Box and Whisker Plot

Scanned by CamScanner

Since the length of right whisker is longer than the length of left whisker the given distribution is positively (right) skewed.

Example 88. List the five number summary and prepare a box and whisker plot from the following information. Are the data skewed?

Size	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	10	12	25	35	40	50

Solution

Here, Largest value $X_L = 60$

Smallest value $X_s = 0$

Size	f	c f
0-10	10	10
10-20	12	22
20-30	25	47
30-40	35	82
40-50	40	122
50-60	50	172
	N= 172	

Size of $M_d = \frac{N}{2} = \frac{172}{2} = 86^{th}$ item

The c.f. just greater than 86 is 122. Hence the corresponding class 40 - 50 is the median class.

$$M_d = L + \frac{N/2 - c.f.}{f} \times h = 40 + \frac{86 - 82}{40} \times 10 = 40+1 = 41$$

Size of $Q_1 = \frac{N}{4} = \frac{172}{4} = 43$

The c.f. just greater than 43 is 47. Hence, Q_1 lies in class 20 – 30

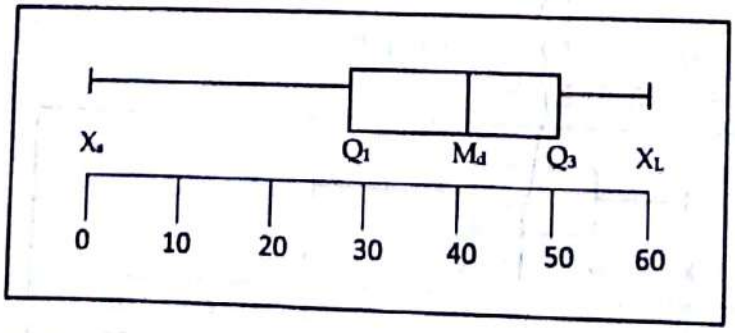
$$Q_1 = L + \frac{N/4 - c.f.}{f} \times h = 20 + \frac{43 - 22}{25} \times 10 = 20 + 8.4 = 28.4$$

Size of $Q_3 = \frac{3N}{4} = \frac{3 \times 172}{4} = 129$

The c.f. just greater than 129 is 172. Hence, Q_3 lies in class 50 – 60.

$$Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} \times h = 50 + \frac{129 - 122}{50} \times 10 = 50 + 1.4 = 51.4$$

The five number summary is {0, 28.4, 41, 51.4, 60}



Since the length of left whisker is longer than the length of right whisker the given distribution is negatively (left) skewed.

Scanned by CamScanner

Example 89. The following data set represents the number of new computer accounts registered during ten consecutive days.

43	37	50	51	58	105	52	45	45	10

(a) Compute the mean, median, quartiles, and sample standard deviation.
(b) Check for outliers using the 1.5(IQR) rule.
(c) Delete the detected outliers and compute the mean, median, quartiles, and sample standard deviation again.
(d) Make a conclusion about the effect of outliers on basic descriptive statistics.

Solution:

Arranging data in ascending order of magnitude and calculation

X	10	37	43	45	45	50	51	52	58	105	$\sum x = 496$
X^2	100	1369	1849	2025	2025	2500	2601	2704	3364	11025	$\sum x^2 = 29562$

$$\text{Mean} = \frac{Sx}{n} = \frac{496}{10} = 49.6$$

$$\text{Median} = \frac{n+1}{2}\text{ th item} = 5.5\text{ th item} = \frac{45+50}{2} = 47.5$$

$$Q_1 = \frac{n+1}{4}\text{ th item} = 2.75\text{ th item} = 2^{nd}\text{ item} + (3^{rd}\text{ item} - 2^{nd}\text{ item}) \times 0.75 = 37 + (43-37) \times 0.75 = 41.5$$

$$Q_3 = \frac{3 \times (n+1)}{4}\text{ th item} = 8.25^{th}\text{ item} = 8^{th}\text{ item} + (9^{th}\text{ item} - 8^{th}\text{ item}) \times 0.25 = 52 + (58-52) \times 0.25 = 53.5$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right)} = \sqrt{\frac{1}{10-1}(29562 - 10 \times 49.6^2)} = 24.47$$

Now to find out ouliers

$IQR = Q_3 - Q_1 = 53.5 - 41.5 = 12$

$IQR \times 1.5 = 12 \times 1.5 = 18$

Upper outlier limit $= Q_3 + 1.5 \times IQR = 53.5 + 18 = 71.5$

Lower outlier limit $= Q1 - 1.5 \times IQR = 41.5 - 18 = 23.5$

So the two observations are out of the range (23.5 to 71.5)

So deleting the two outlier observation 10 and 105 the remaining observations are

X	37	43	45	45	50	51	52	58	$\sum x = 381$
X^2	1369	1849	2025	2025	2500	2601	2704	3364	$\sum x^2 = 18437$

$$\text{Mean} = \frac{\sum x}{n} = \frac{381}{8} = 47.625$$

$$\text{Median} = \frac{n+1}{2}\text{ th item} = 4.5\text{ th item} = \frac{45+50}{2} = 47.5$$

$$Q_1 = \frac{n+1}{4}\text{ th item} = 2.25\text{ th item} = 2^{nd}\text{ item} + (3^{rd}\text{ item} - 2^{nd}\text{ item}) \times 0.25 = 43 + (45-43) \times 0.25 = 43.5$$

$$Q_3 = \frac{3x(n+1)}{4}\text{ th item} = 6.75\text{ th item} = 6^{th}\text{ item} + (7^{th}\text{ item} - 6^{th}\text{ item}) \times 0.75 = 51 + (52-51) \times 0.75 = 51.75$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right)}$$

$$= \sqrt{\frac{1}{8-1}(18437 - 8 \times (47.625)^2)} = \sqrt{41.696} = 6.46$$

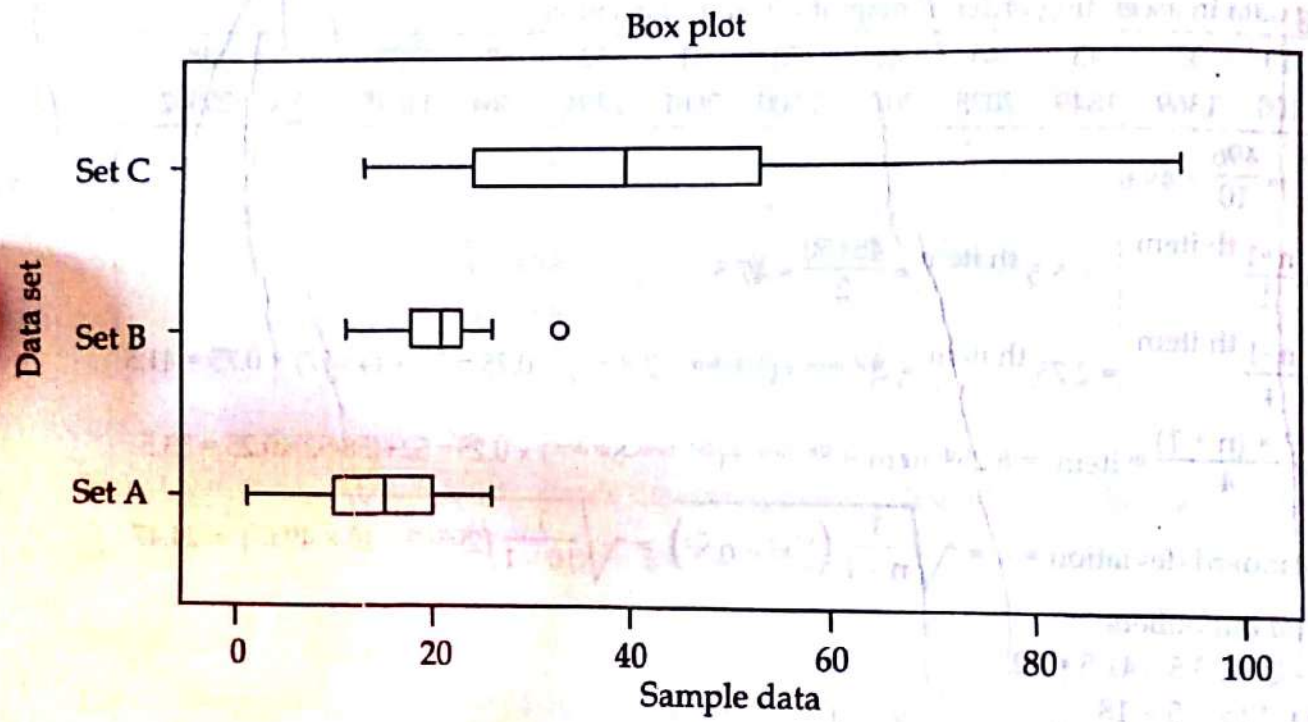
Since outliers from both side are omitted, the median was not changed, besides it outliers has significant effects on other descriptive statistical tools.

Scanned by CamScanner

Example 90. The following table represents descriptive statistics computed from three different independent sample data (X).

Data	Sample size (n)	$\sum X$	Minimum	Q1	Median	Q3	Maximum	$\sum(X-\bar{X})^2$
Data Set A	30	439	1	10	15	20	26	1348.97
Data Set B	30	625	11	18	21	23	33	540.17
Data Set C	30	1239	13	24	39.5	53	94	12836.3

a) Compare sample mean, median and explain about the shape of the data distribution for each dataset. Also compare variability of the three set of dataset.
Box-plots have been generated through SPSS for each dataset as follows



b) Do these box-plots support your findings obtained in a) about the shape of the distribution? Explain

Solution

For dataset A

$$\bar{x} = \frac{\sum x}{n} = \frac{439}{30} = 14.633$$

$$s = \sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2} = \sqrt{\frac{1}{29} \times 1348.97} = 6.82$$

$$C.V. = \frac{s}{\bar{x}} \times 100\% = \frac{6.82}{14.633} \times 100\% = 46.6\%$$

Mean is less than median, so the data is negatively skewed.

For dataset B

$$\bar{x} = \frac{\sum x}{n} = \frac{625}{30} = 20.833$$

$$s = \sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2} = \sqrt{\frac{1}{30-1} \times 540.17} = 4.31$$

$$C.V. = \frac{s}{\bar{x}} \times 100\% = \frac{4.31}{20.833} \times 100\% = 20.68\%$$

Mean is less than median, so the data is negatively skewed.

Scanned by CamScanner

For dataset C

$$\bar{x} = \frac{\sum x}{n} = \frac{1239}{30} = 41.3$$

$$s = \sqrt{\frac{1}{30-1} \times 12836.30} = 21.038$$

$$C.V. = \frac{s}{\bar{x}} \times 100\% = \frac{21.038}{41.3} \times 100\% = 50.9\%$$

Mean is greater than median, so the data is positively skewed.

On comparing three dataset the variability is more in dataset C and less in dataset B.

On the basis of SPSS output of Box-plot, since the left whisker of dataset A and B are larger than right whisker, these two dataset are negatively skewed where as data set C has longer right whisker so it is positively skewed. The shape of the distribution on the basis of box-plot is found to be same as calculation done in a).

Residual Plot

The difference between observed value of variable and estimated value is known as residuals. It is very important to analyze the residuals during regression analysis. It is used to test the adequacy of the fit of a regression model and check underlying assumptions. One of the most commonly used residual plot is normal probability plot.

Normal Probability Plot

It is used to test the normality of the residuals obtained from the regression analysis. Slight difference in the normality of the residual does not change in the assumption related to the regression analysis but it affects in the calculation of different test statistics like t, F, confidence intervals etc. these are all based on the normality assumption of the residuals.

Normal probability plot is used to check the normality assumption. It uses the cumulative probability of ranked residuals in ascending order of magnitude.

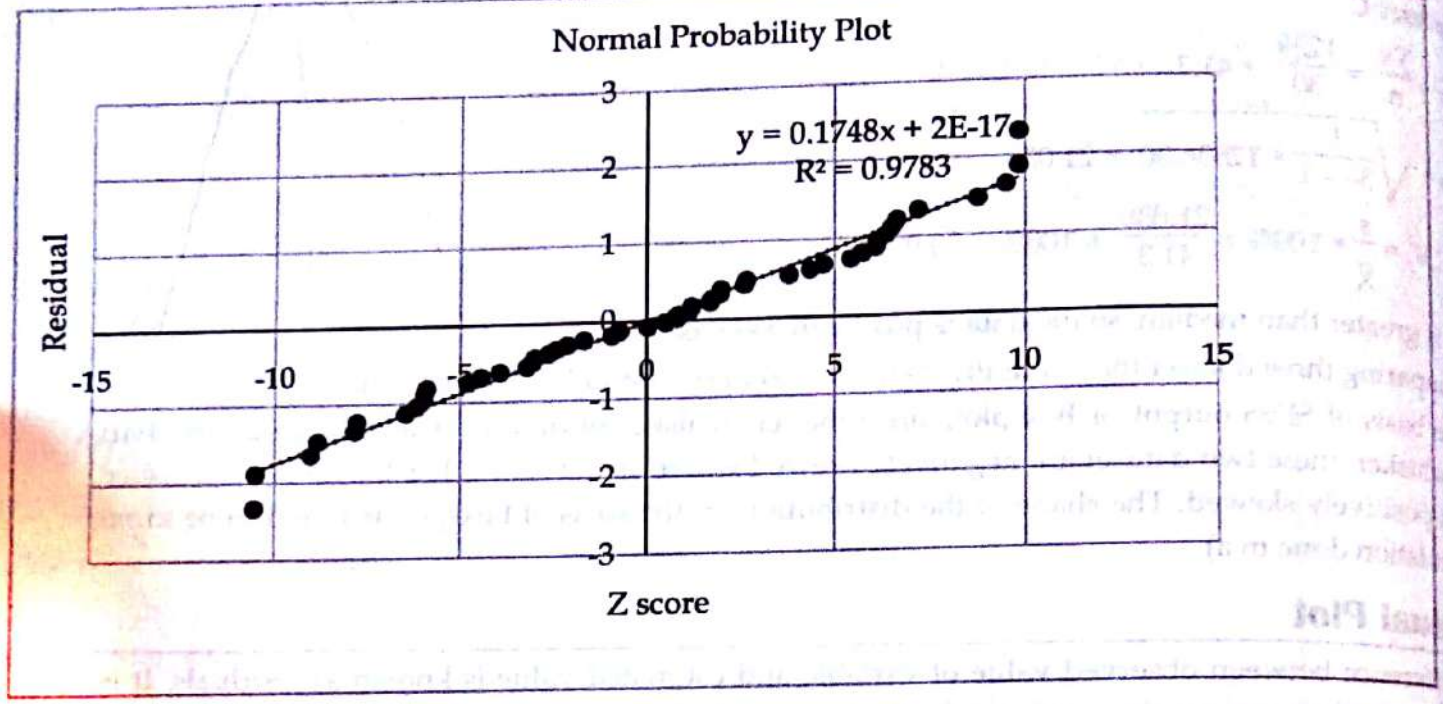
Steps:

1. Calculate the residuals from the regression analysis.
2. Rank them in ascending order of magnitude and sort them in terms of rank.
3. The probability associated with the rank is obtained using the relation $p_i = \frac{(i - 0.5)}{n}$

   where i = rank order, n = number of observation.
4. Z score value corresponding to the p_i is obtained.
5. The scattered diagram is plotted with value of residuals in x axis and corresponding z values in y axis.
6. Interpretation:
   a. It the graph is straight line the data is perfectly normally distributed.
   b. If it is curved upwards it is positively skewed.
   c. If it is curved downwards it is negatively skewed.

The example of normal probability plot

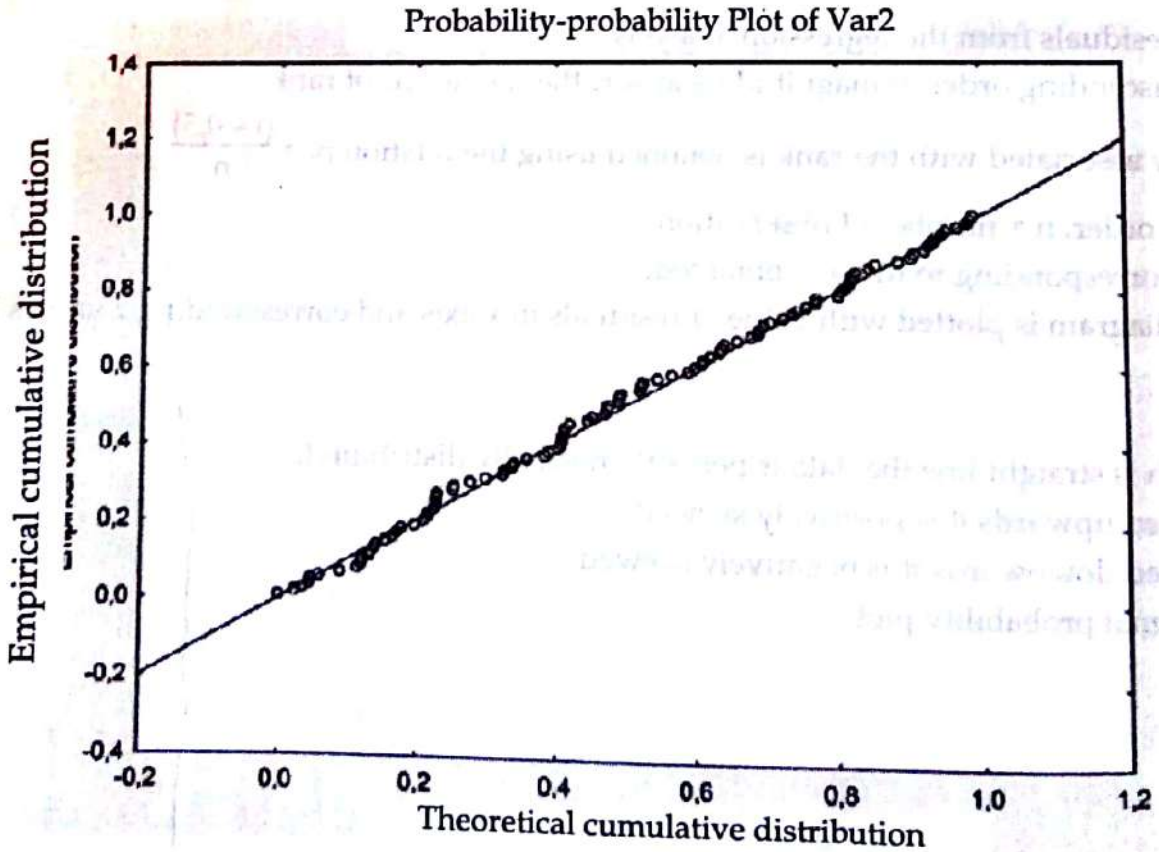
Scanned by CamScanner



P-P Plot

It is also known as probability-probability plot. It is used to see if a given set of data follows some specified distribution. It should be approximately linear if the specified distribution is the correct model. The P-P plot is constructed using the theoretical cumulative distribution function, F(x), of the specified model. The values in the sample of data, in order from smallest to largest are denoted by x_1, x_2, ... x_n.

P–P plots are sometimes limited to comparisons between two samples, rather than comparison of a sample to a theoretical model distribution. However, they are of general use, particularly where observations are not all modelled with the same distribution.one only plots points, plotting the observed kth observed points (in order: formally the observed kth order statistic) against the k/(n + 1) quantile of the theoretical distribution.



Scanned by CamScanner

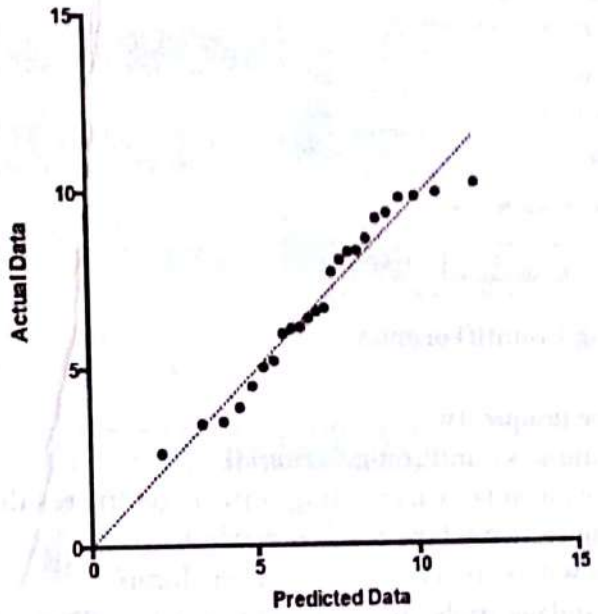
Q-Q plot

It is similar to normal probability plot. Q-Q stands for Quantile-quantile. 20th, 40th, 60th and 80th gives the point of division of five quantiles of the data. It is the as a percentile, but expressed as fraction. The 95th percentile is equivalent to a quantile of 0.95. The median or 50th percentile is 0.50 expressed as a quantile.

Steps

1. Compute a percentile for each value in the data set.
2. Compute a percentile using relation 100(i-0.5)/n. Where i is the rank and n is the sample size.
3. Calculate z score from normal table or using software.
4. Compute corresponding values of z scores with standard deviation of data and add with mean of the observed data to determine predicted values.
5. The graph between predicted values is obtained using those two value sets of observed and predicted values.

The example of Q-Q plot



Steps in excel

1. Place or load your data values into the first column. Leave the first row blank for labeling the columns. Sort the data in ascending order (look under the Data menu).
2. Label the second column as **Rank**. Enter the ranks, starting with 1 in the row right below the label. Each following row will be one more than the last (note: you can use an expression, copy and then paste to save you time)
3. Label the third column as **Rank Proportion**. This column shows the rank proportion of each value. Use this expression for the first data value = **(b2 - 0.5) / count(b$2:b$N)** where N should have the row number of the last cell. Finish the column by copying the first data expression to the remaining rows. Check to make sure your percentiles look like they are correct!
4. Label the fourth column as **Rank-based z-scores**. Excel provides these values with the **normsinv** function. Use this function to create the values in the fourth column.
5. Copy the first column to the fifth column. The Excel chart wizard works better if the x-axis values are just to the left of the y-axis values.
6. Select the fourth and fifth column. Select the chart wizard and then the scatter plot. The default data values should be good, but you should provide good labels.

Scanned by CamScanner

Exercise – 2

Theoretical Questions

1. What are the desirable characteristics of a good measure of central tendency?
2. What do you mean by average in statistics? With suitable examples, point out situation in which the different measure of averages are used.
3. Discuss the various measures of dispersion pointing out their relative merits and demerits. Explain their utility.
4. Define standard deviation. Explain why the standard deviation is regarded as superior to other measures of dispersion.
5. Distinguish between absolute and relative measures of dispersion.
6. Distinguish between 'variance' and 'coefficient of variation'.
7. What do you mean by skewness? Write the different measures of skewness.
8. How does skewness differ from dispersion?
9. What do you mean by kurtosis? Write the different measures of kurtosis
10. What do you mean by moments? Describe central moments and raw moments.
11. Write down the relationship between central moments and raw moments.

Numerical Questions

1. Compute mean, median and mode of the following data

55	39	45	55	41	35	60
40	55	35	37	55	55	65

[Ans: 48, 50, 55]

2. The number of telephone calls received at an exchange for 200 successive one-minute intervals are given below.

No of calls	0	1	2	3	4	5	6	Total
Frequency	15	22	28	35	42	34	24	200

Compute the mean, median and mode.

[Ans: 3.325, 3, 4]

3. The length in meter of 100 VGA Cable used in a company are measured to the nearest 0.01 meter and the results are given below.

Length in meter	Frequency	Length in meter	Frequency
3.80-3.89	3	4.20-4.29	28
3.90-3.99	8	4.30-4.39	18
4.00-4.09	14	4.40-4.49	10
4.20-4.19	19	4.40-4.50	8

Find the value of mode and median.

[Ans: 4.24, 4.23]

4. Compute the mean, median and mode for the following frequency distribution of I.T.I.Q. score for 100 CSIT Students.

Scanned by CamScanner

I.T.I.Q. score	No. of Students	I.T.I.Q. score	No. of Students
50-59	2	100-109	25
60-69	4	110-119	20
70-79	8	120-129	2
80-89	15	130-139	2
90-99	21	140-149	1

[Ans: 97.8, 99.5, 103.94]

5. Compute the geometric mean, for the series of data: 38, 45, 50, 68, 82, 40

[Ans: 51.68]

6. Compute the geometric mean for the following frequency distribution.

Weight of students (kg)	No. of students
50-54	30
55-59	25
60-64	32
65-69	20
70-74	13
Total	120

[Ans: 60]

7. A boy travels by a cycle for 5 days. Each day he travels for 5 hrs. His speeds for the five days were respectively 40km/hr, 38 km/hr, 35km/hr, 36km/hr and 39km/hr. Compute the harmonic mean.

[Ans: 37.508]

8. The data relating to space left in gigabyte of 100 computer of a company were given below.

Free disk space	No of computer	Free disk space	No of computer
0-5	3	20-25	20
5-10	7	25-30	8
10-15	18	30-35	5
15-20	35	40-45	4

Compute the first and third quartiles, 9th decile, 15th and 90th percentile.

[Ans: 14.16, 23, 29.37, 11.39, 29.3]

9. A random sample survey was conducted by a hardware manufacturing company to identify the size of the pen drive that company should produce in order to satisfy majority of the user. Following information was obtained from the survey about the preference of the size of pen drive the user expect to have with them.

Size of the Pen drive (in GB) prefer	No. of user
4-5	10
5-6	35
6-7	70
7-8	35
8-9	12
9 -10	6

Which measures of central tendency should be computed in this problem so that the company would be benefitted?

[Ans: 6.5]

Scanned by CamScanner

10. A software company has four sections employing 50, 35, 40 and 25 IT workers. The average monthly salary for these four sections are respectively Rs. 30000, 35000, 40000 and 50000. Compute the average monthly salary for the whole institute.

[Ans: 37166.67]

11. The percentage age distribution of urban male population of Nepal from 2011 census is given below:

Age group	Male population	Age group	Male population
0-4	11.8	35-39	6.2
5-9	12.9	40-44	4.7
10-14	12.5	45-49	4.0
15-19	11.2	50-54	2.9
20-24	10.7	55-59	2.3
25-29	8.9	60 and above	4.7
30-34	7.2		

Compute the first and third quartiles, 8th decile and 70th percentile.

[Ans: 9.62, 34.36, 38.37, 30.88]

12. The percentage of plastic used for the production of different makes of laptop is given below:

Makes of Laptop	A	B	C	D	E	F	G
Percentage of plastic	49.19	49.25	50.34	45.122	49.87	58.27	83.58

Compute the geometric mean.

[Ans: 48.717]

13. The marks distribution of 100 students of a college is as follows.

Marks	10 - 20	20 - 40	40 - 70	70 - 90	90 - 100
No of students	15	20	30	20	15

i) Find the highest mark of the weakest 30% of the students.

ii) Find the lowest mark of top 40% of the students.

iii) Find the lowest marks of top 20% of the students.

iv) Find the limits and range of marks of middle 50% of students.

[Ans: 35, 65, 85, 30 to 80 and 50]

14. Compute the quartile deviation of the following distribution giving the screen size of Laptop available in Nepalese Laptop Market.

Size of Screen (cm)	No. of Laptop	Size of Screen (Cm)	No. of Laptop
9.5	1	13.5	200
10.0	8	14.0	250
10.5	20	14.5	280
11.0	30	15.0	245
11.5	50	15.5	80
12.0	95	16.0	40
12.5	110	16.5	35
13.0	150	17.0	5

[Ans: 1]

Scanned by CamScanner

15. The following frequency distribution represents the weight of 200 laptops.

Weight in lbs	Frequency	Weight in lbs	Frequency
4-5	20	8-9	32
5-6	24	9-10	24
6-7	35	10-11	8
7-8	48	11-12	2

Compute the first three quartiles and quartile deviation.

[Ans: 6.12, 7.36, 8.55, 1.22]

16. The scores obtained by 10 students in Statistics I of an IT College are given below. Compute range and standard deviation;

55 35 60 55 55 65 40 45 35 42

[Ans: 30, 10.109]

17. The following data gives on temperature of Kathmandu for a week in summer. Compute the range and quartile deviation.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Temp.(^{0}C)	34	35	32	35	36	34	35

[Ans: 4, 0.5]

18. The time in minutes for solving an assignment given to 10 students of CSIT Students are given below:

10 12 11 15 18 20 24 23 26 16

Compute the mean deviation from mean and standard deviation.

[Ans: 4.7, 5.37]

19. To obtain information on the distribution of customer service time a sample of 100 customer's service times were recorded and presented in the following table.

Customer service time (in minutes)	No. of customers	Customer service time (in minutes)	No. of customers
0-5	2	15-20	30
5-10	8	20-25	28
10-15	26	25-30	6

Compute the standard deviation.

[Ans: 5.64]

20. The frequency distribution oftime required to open the operating system of 200 computers is given below.

Time in seconds	No. of Computers	Time in seconds	No. of Computers
0-4	2	20-24	48
5-9	20	25-29	32
10-14	35	30-34	18
15-19	40	35- 39	5

Compute the standard deviation.

[Ans: 7.87]

21. The average marks secured in CSIT exam in the year 2072 by College A and College B are 78 and 80 with variances 100 and 81 respectively. The number of students appeared in the CSIT exam from College A is 100 and College B is 150. Compute the combined mean and combined variance of marks secured by the two groups.

[Ans: 79.2, 89.5]

Scanned by CamScanner

22. The number of runs scored by two group of cricket players in a test match are

Group A	10	25	85	72	115	80	52	45	30	10
Group B	120	15	30	35	42	65	80	34	25	15

Test which group is more consistent.

[Ans: Group A]

23. The following data represents the scores made in an intelligence test by two groups of students from section A and section B of a college.

Student no.	Section A	Section B	Student no.	Section A	Section B
1	9	10	6	8	8
2	8	8	7	5	7
3	10	6	8	6	8
4	6	8	9	7	5
5	7	9	10	8	8

Test which group is more consistent.

[Ans: Section B]

24. The following data gives the number of minutes required for 15 boy and 15 girl students of a class to complete a task.

Time required for male students	Time required for female students
5.70 6.80 7.25 8.20 8.10	7.52 8.20 8.32 6.90 6.80
7.20 6.88 7.20 7.35 7.45	8.30 7.45 9.00 10.50 7.20
6.90 7.22 6.85 6.40 6.20	10.20 8.26 8.50 8.32 10.00

Test which group is more consistent regarding the time required to complete a task.

[Ans: Male]

25. For a group of 10 items $\sum X = 552$, $\sum X^2 = 3170$ and Mo = 46.2 . Find Karl Pearson's Coefficient of skewness

[Ans: 0.81]

26. The sum of 20 observations is 300 and it's sum of squares is 5000 and median is 15. Find the Coefficient of skewness and coefficient of variation

[Ans: 0, 33.33%]

27. In a distribution mean = 65, median = 70 and coefficient of skewness =- 0.6. Find mode and coefficient of variation.

[Ans: 80, 38.46%]

28. For a moderately skewed data, the arithmetic mean is 200, the coefficient of variation is 8 % and Karl Pearson's coefficient of skewness is 0.3. Find mode and median.

[Ans: 195.2, 198.4]

29. Calculate mean, S.D, and coefficient of skewness from following data.

Size	6	9	12	15	18
Frequency	7	12	19	10	2

[Ans: 11.28, 3.14, -0.23]

30. From following grouped frequency distribution compute Karl Pearson's coefficient of skewness.

Mid value of income in '00' Rs.	150	250	350	450	550	650	750	850
No of staff in IT company	80	105	120	165	100	90	60	40

[Ans: 0.08]

31. In a certain distribution, coefficient of skewness is 0.6. If the sum of the two extreme quartiles is 100 and median is 38. Find the value of the upper quartile.

[Ans: 70]

Scanned by CamScanner

32. The coefficient of skewness for a distribution is – 0.8. If the lower and upper quartilesare 44.1 and 56.6 respectively, calculate median

[Ans: 55.35]

33. Calculate Bowley's coefficient of skewness from following data;

X	20	23	24	26	28	30	40
f	3	6	10	12	11	9	4

[Ans: 0]

34. From the following frequency distribution, calculate Bowley's coefficient of skewness.

Hourly wages(Rs)	23 – 27	28 – 32	33 – 37	38 – 42	43 – 47	48 – 52
No of workers	22	16	9	4	3	1

[Ans: 0.173]

35. Compute the appropriate measure of skewness for download speed from the following data.

Download speed (Mbps)	Below 100	100 -150	150-200	200-250	250-300	300 and above	Total
Time in minutes	10	25	145	220	70	30	500

[Ans: - 0.102]

36. Compute percentile coefficient of kurtosis from following data;

X	2	4	6	8	10
f	1	5	10	3	1

[Ans: 0.25]

37. Calculate percentile coefficient of kurtosis and interpret the result.

Hourly Renemuration (Rs)	100 – 110	110 – 120	120 – 130	130 – 140	140 – 150	150 – 160	160 – 170
No of persons	10	14	18	24	16	12	6

[Ans: 0.268]

38. Compute first four central moments from following series:

Class interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
frequency	5	7	9	7	5

[Ans: 0, 163.63, 0, 52727.27]

39. Compute first four moments about arbitrary point 4 from following distribution:

X	2	3	4	5	6
f	1	3	7	2	1

[Ans: - 0.071, 0.928, -0.071, 2.642]

40. Calculate first four moments about mean from following distribution;

Marks	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No of students	7	10	20	18	7

[Ans: 0, 179.23,-523.69,55347.97]

41. The first four moments about an arbitrary origin of a frequency distribution are -2 , 14 , -20 , 50. Find first four central moments.

[Ans: 0, 10, 48, 178]

42. If the first four moments about mean are 0, 2.8, –2 and 24.5 respectively. Compute coefficient of skewness and kurtosis and comment upon result.

[Ans: - 0.427, 3.125]

Scanned by CamScanner

43. The first three moments of a distribution about value 5 of a distribution are 2, 20and 40 respectively. Compute measure of central tendency, dispersion and skewness.

[Ans: 7, 4, -1]

44. If the S.D. of a symmetrical distribution is 11, what must be the value of fourthmoment about the mean in order that the distribution be i) mesokurtic ii) platykurtic iii) leptokurtic

[Ans: 43923, Less than 43923, More than 43923]

45. Calculate the first four moments about the mean from following distribution and then findthe measure of skewness and kurtosis and comment upon the distribution.

X	0	1	2	3	4	5	6	7	8	9	10
f	5	10	30	70	140	200	140	70	30	10	5

[Ans: 0, 3.49]

46. Distribution of successful project by two different streams of students in computer are given below:

Successful projects conducted	20	22	23	25	26
Nos. of B Sc CSIT students	5	7	10	8	5
Nos. of BCA students	3	7	15	8	2

Compute measure of central tendency, dispersion, skewness and kurtosis based upon moments and compare result. [Ans: For A: 23.26, 8.18%, - 0.18, 2.044, For B: 23.17, 6.65%, - 0.10, 2.72]

47. Compute the measure of central tendency, dispersion skewness and kurtosis from following distribution.

Wage (Rs)	Below 97	98 – 102	103 – 107	108 – 112	113 – 117	118 – 122	123 – 127	128 and above
No of employees	2	5	12	17	14	6	3	1

[Ans: 110.735, 4.94, 0.007, 0.255]

48. Drive Dof a computer having 25 folders contains the following number of files.
4,0,5,2,3,1,4,3,2,3,4,3,1,6,3,2,3,4,2,1,0,3,2,5,4
a) List the five number summary
b) Construct a box plot for the data
c) Are the data skewed?

[Ans: {0, 2, 3, 4, 6}, symmetrical

49. Calculate the five number summary and construct box plot from the following data. Also describe the shape of the distribution.

Marks	35	38	45	60	72	80	85
No. of students	3	5	10	8	4	2	1

[Ans: {35, 38, 45, 72, 85}, +vely skewed

50. List the five number summary and prepare a box plot from the following information about the number of vehicle arrived during past 100 days in a certain tolling station. Are the data skewed?

Vehicles	0-10	10-20	20-30	30-40	40-50
No. of days	3	14	53	20	10

[Ans: {0, 21.51, 26.23, 32.5, 50}, +vely skewed

Scanned by CamScanner

51. The numbers of blocked intrusion attempts on each day during the first two weeks of the month were

60 51 53 41 42 64 54 47 47 63 54 60 58 62

After the change of firewall settings, the numbers of blocked intrusions during the next 20 days were

60 28 39 56 52 45 51 40 39 50 60 53 43 55 46 42 44 43 46 52

For comparison the number of blocked intrusions before and after the change, construct side-by-side stem-and-leaf plots;

52. A network provider investigates the load of its network. The number of concurrent users is recorded at fifty locations (thousands of people),

17.2, 22.1, 18.5, 17.2, 18.6, 14.8, 21.7, 15.8, 16.3, 22.8, 24.1, 13.3, 16.2, 17.5, 19.0, 23.9, 14.8, 22.2, 21.7, 20.7, 13.5, 15.8, 13.1, 16.1, 21.9, 23.9, 19.3, 12.0, 19.9, 19.4, 15.4, 16.7, 19.5, 16.2, 16.9, 17.1, 20.2, 13.4, 19.8, 17.7, 19.7, 18.7, 17.6, 15.9, 15.2, 17.1, 15.0, 18.8, 21.6, 11.9

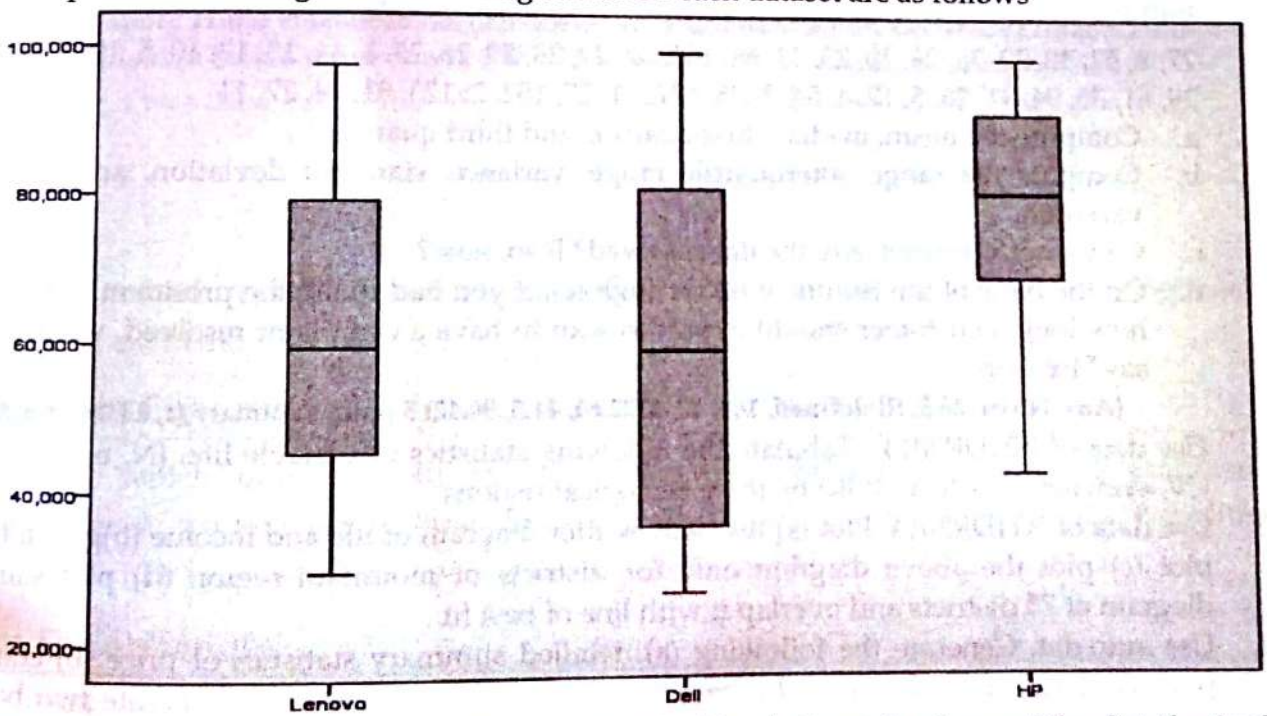
Summarize these data using stem and leaf display.

53. The following table represents descriptive statistics regarding price in Rupees of different model of computer made by three manufacturers.

Data	Sample size (n)	sum	Minimum	Q1	Median	Q3	Maximum	$\sum(X-\bar{X})^2$
Lenovo	20	1226700	29600	43675	59200	79325	97400	8174025500
Dell	20	1180900	26500	33875	58700	80800	99300	10736249500
HP	20	1527300	42100	68150	80000	92700	98700	6435905500

a) Compare sample mean, median, and explain about the shape of the data distribution for each dataset. Also compare variability of the price of three types of computers.

Box-plots have been generated through SPSS for each dataset are as follows



b) Do these box-plots support your findings obtained in a) about the shape of the distribution? Explain

[Ans: Mean:61335,59045,76365, CV: 33.82, 40.26, 24.10]

Scanned by CamScanner

54. The following data represents number of laptops passed from a checkpoint in a week

52	47	35	40	28	95	62

Compute

a) Mean, three quartiles, sample standard deviation from above data
b) Check whether there are outliers or not
c) If outliers are presents then delete the outliers and compute mean, three quartiles and sample standard deviation.

[Ans: 54.14,35,47,62,29, outlier:115, 44,33.25,43.5,54.5,12.25]

55. Use Normal Probability Plot to determine the normality of the residual calculated. Comment on the data.

Residual	6.106843	-6.4583	6.343267	4.343267	-0.93629	1.759133	-6.15975
	6.504608	-2.99841	-9.02947	-2.40824	2.603826	9.846284	9.510642
	-5.89919	1.131871	-3.06053	-10.5204	0.840249	-10.5515	0.510642
	-4.78098	9.840249	5.722038	-7.83707	-5.96735	1.945502	6.634888
	-4.43327	-7.77495	1.976563	1.684942	3.784161	5.417458	-0.71886
	-2.12266	-8.83707	-3.20981	8.759133	-5.99841	-1.67573	7.193993
	1.193993	-2.6136	0.815222	4.69701	-3.93025	2.653881	6.069748
				0.032652			

Practical Questions

1. One of the major measures of the quality of service provided by an organization is the speed with which it responds to customer complaints. An internet service provider, had undergone a major improvement by recruiting well trained installation crews, supervisors and office staffs. The business objective of the company was to reduce the time between when the complaint it received and when it is resolved. During a recent month, the company received 50 complaints concerning internet installation. The data from the 50 complaints, collected by ISP, represent the number of hours between the receipt and the resolution of the complaint:

   27, 4, 52, 30, 22, 36, 26, 20, 23, 33, 68, 165, 32, 29, 28, 29, 26, 25, 1, 14, 13, 13, 10, 5, 19, 126, 110, 110, 29, 61, 35, 94, 31, 26, 5, 12, 4, 54, 5, 35, 137, 31, 27, 152, 2, 123, 81, 74, 27, 11

   a. Compute the mean, median, first quartile, and third quartile.
   b. Compute the range, interquartile range, variance, standard deviation, and coefficient of variation.
   c. Construct a boxplot. Are the data skewed? If so, how?
   d. On the basis of the results of (a) through (c), if you had to tell the president of the company how long a customer should expect to wait to have a complaint resolved, what would you say? Explain.

   [Ans: (43.04, 28.5, ill defined, 164, 42, 1722.64, 41.5, 96.42) 5 point summary (1, 13.75, 28.5, 55.75, 165)]

2. Use data of NHDR2014. Tabulate the following statistics of variable life, (N, mean, std, range, CV, skewness, kurtosis, IQR) by three ecological regions.
3. Use data of NHDR2014. Plot (a) two way scatter diagram of life and income (b)give a title to the plot (c) plot the above diagram only for districts of mountain region (d) plot same scatter diagram of 75 districts and overlap it with line of best fit
4. Use auto.dat. Generate the following (a) detailed summary statistics of price (b) compare the price of domestic and foreign by generating summary statistics also generate two boxplots price to compare between domestic and foreign (c)plot steam-and-leaf diagram of milage (mp and comment on the skewness of the figure.

❑❑❑

Scanned by CamScanner

Chapter 3

INTRODUCTION TO PROBABILITY

Introduction

Most of the phenomena in physical and chemical sciences are of a deterministic nature. However, there exist a number of phenomena as generated by category *(b)* where the results cannot be predicted with certainty and are known as *unpredictable* or *probabilistic* phenomena. Such phenomena are frequently observed in computer, information technologies, engineering, economics, business and social sciences or even in our day-to-day life.

The word probability is used in many day to day cases. Like the probability of happening some events viz, success of project, failure of motherboard, accidents, misprints, passing exams etc.

Hence the probability is the chance of occurrence of an event. It is a number which ranges from 0 to 1. 0 for the event which is not possible and 1 for the event which is certain or sure.

In the beginning it was used in the field of game but now it is used in every field of life.

Terminology

Experiment

It is scientific procedure undertaken to make discovery. E.g. tossing a coin, playing card, appearing in an exam etc.

Deterministic Experiment

Experiment in which outcome can be predicted with certainty is called deterministic experiment. E.g. A student performing best in exam can be predicted pass with certainty, selecting a white ball from a bag containing white balls is certain.

Random Experiments

An experiment is called a random experiment ifconducted repeatedly under essentially homogeneous conditions, the result is not unique but may be any one of the various possible outcomes. For example, if you developed an apps and tested for launch, the apps may have coding error or it may be correct. Here the experiment is the testing of apps.

Scanned by CamScanner

Trial and Event

Performing a random experiment is called a trial and outcome or combination of outcomes are termed as events. For example: If a coin is tossed repeatedly, the result is not unique. We may get any of the two faces, head or tail. Thus tossing of a coin is a trial of a random experiment and getting of a head or tail is an event.

Event is called simple if it corresponds to a single possible outcome of the experiment or trial otherwise it is known as a compound or composite event. Thus, in tossing of a single die, the event of getting '5' is a simple event but the event 'getting an even number', is a composite event.

Exhaustive Cases (N)

The total number of possible outcomes of a random experiment is called the exhaustive cases for the experiment. Thus, in toss of a single coin, we can get head (H) or tail (T). Hence exhaustive number of cases is 2, viz., (H, T). If two coins are tossed, the various possibilities are HH, HT, TH, TT where HT means head on the first coin and tail on second coin, and TH means tail on the first coin and head on the second coin and so on. Thus, in case of toss of two coins, exhaustive number of cases is 4.

Favourable Cases or Events (m)

The number of outcomes of a random experiment which entail (or result in) the happening of an event are termed as the cases favourable to the event. For example: In a toss of two coins, the number of cases favourable to the event 'exactly one head' is 2, viz., HT, TH and for getting 'two heads' is one viz., HH.

Mutually Exclusive Events or Cases

Two or more events are said to be mutually exclusive if the happening of any one of them excludes the happening of all others in the same experiment. For example, in toss of a coin, the events 'head' and 'tail' are mutually exclusive because if head comes, we can't get tail and if tail comes we can't get head. Similarly, in the throw of a die, the six faces numbered 1,2, 3,4,5 and 6 are mutually exclusive. Thus, events are said to be mutually exclusive if no two or more of them can happen simultaneously.

Sample Space

The set all possible outcomes of a random experiment is called sample space. For example, the sample space of a random experiment of tossing of a coin is {H, T}

Equally Likely Cases

The outcomes are said to be equally likely or equally probable if none of them is expected to occur in preference to other. Thus, in tossing of a coin (dice), all the outcomes, viz., H, T (the faces 1, 2, 3, 4, 5, 6) are equally likely if the coin (die) is unbiased.

Impossible Event

An event which can not happen in random experiment is called impossible event. e.g. getting a face numbered 7 on rolling a die.

Sure event

An event which is certain to happen in random experiment is called certain event. E.g. getting a black ball on selecting ball from a bag containing black balls only, getting head or tail on tossing a coin.

Scanned by CamScanner

Independent Events

Events are said to be independent of each other if happening of any one of them is not affected by and does not affect the happening of any one of others. For example: In tossing of a die repeatedly, the event of getting '5' in 1st throw is independent of getting '5' in second, third or subsequent throws.

Dependent Events

Events are said to be dependent of each other if happening of any one of them is affected by happening of others. For example a card is drawn from a pack of cards and without replacing the drawn card if next card is drawn then second draw is affected by first draw.

Counting

When an experiment is conducted in two or many stages, there may be large number of possible outcomes at each stage and in this case, the counting and listing of the all possible outcomes becomes a difficult. It can be made simple by rules called fundamental rules of counting or combinatorial analysis.

It consist of addition and multiplication rule. When two event A and B are mutually exclusive they can occur in n + m ways. And when A and B occur simultaneously they can occur in nm ways.

There are two basic methods of combinatorial counting namely, Permutation and combination.

Permutation

Permutation is the arrangement of objects.

A permutation of n different objects taken r at a time, denoted by ${}^nP_r$ is an ordered arrangement of only r objects of the n objects.

It is given by

$$P(n, r) = {}^nP_r = \frac{n!}{(n-r)!};\ r \le n$$

For repeated cases the permutation of n unit in which different units are repeated p, q and r times then number of arrangement $= \frac{n!}{p!q!r!}$.

Combination

Combination is the selection of objects.

A combination of n different objects taken r at a time, denoted by ${}^nC_r$ or C(n,r) is a selection of only r objects out of the n objects, without any regard to the order of arrangement.

It is given by

$$C(n, r) = {}^nc_r = \frac{n!}{(n-r)!\ r!}$$

And C (n, r) = C (n, n-r)

Example 1. In how many ways can 3 person be seated in 5 chairs?

Solution

Here, n = 5, r = 3

Then 3 persons can be seated in 5 chairs in following ways $= P(n, r) = P(5, 3) = \frac{5!}{(5-3)!} = 60.$

Example 2. In how many ways can the letters in the word "STATISTICS" can be arranged?

Solution

Here, n=10, p= number of s =3, q = number of t= 3, r= number of i = 2

Scanned by CamScanner

Then permutation $= \frac{11!}{3!\,3!\,2!} = 50400.$

Example 3. In how many ways 2 computers can be selected out of 10 computers in Lab.

Solution

Here, n=10, r = 2

The possible number of ways that 2 computers can be selected out of 10 in C (10, 2) ways

$$C(10, 2) = \frac{10!}{2!\,(10-2)!} = 45$$

Definitions of Probability

Mathematical or Classical or a Priori probability

If a random experiment results in N exhaustive, mutually exclusive and equally likely outcomes out of which m are favourable to the happening of an event say A, then the probability of occurrence of A, denoted by P(A) is given by

$$P(A) = \frac{\text{Favourable number of cases to A}}{\text{Total number of cases}} = \frac{m}{N}$$

Remarks

1) The number of cases favourable to the complementary event $\bar{A}$ i.e. not happening of event A are N-m then

$$P(\bar{A}) = \frac{\text{Favourable number of cases to A}}{\text{Total number of cases}} = \frac{N-m}{N} = 1 - \frac{m}{N}$$

$P(\bar{A}) = 1 - P(A).$

2) Since, m and N are non-negative, $P(A) \geq 0$. Hence, the probability is a real number lying between 0 and 1.

$0 \leq P(A) \leq 1$

3) Total probability of the events is unity.

4) This definition of probability is applicable if the events are equally likely and exhaustive.

5) The probability an event is presented as odd in favour of event A is a: b.

i.e. $P(A) = \frac{a}{a+b}$ and $P(\bar{A}) = \frac{b}{a+b}$

odd against an event B is c:d i.e. $P(B) = \frac{d}{c+d}$ and $P(\bar{B}) = \frac{c}{c+d}$

Limitations of Mathematical or Classical or a Priori Definition of Probability

The definition of Mathematical probability does not work in the following conditions;

i) The various outcomes of experiment are not equally likely.

ii) Pre information about total number of outcomes of an experiment is not known.

iii) The exhaustive number of cases in a trial is infinite.

Example 4. What is the probability of 7 or 11 in throwing two dice?

Solution

7 can happen in following ways (2, 5), (5, 2), (1, 6), (6, 1), (3, 4) and (4, 3)

And 11 can happen in following ways (5, 6) and (6, 5)

Total number of outcomes in throwing two dice = 6×6 = 36

Then the probability of 7 or 11 = $P(7) + P(11) = \frac{4}{36} + \frac{4}{36} = \frac{8}{36} = \frac{2}{9}.$

Scanned by CamScanner

Example 5. What is the probability that a leap year selected at random contains 53 Sunday?

Solution

A leap year has 366 days

A year has 52 complete weeks, i.e. 52 × 7 = 364 days

Hence remaining days = 366 – 364 = 2

The two remaining days may be{Sun-Mon, Mon-Tue, Tue-Wed, Wed-Thu, Thu-Fri, Fri- Sat, Sat-Sun}

Here total number of cases (N) = 7

Favourable number of cases for Sunday (m) = 2

The Probability of 53 Sundays (P) = $\frac{m}{N} = \frac{2}{7}$

Example 6. A bag cartoon contains 6 Dell Laptops , 9 Lenovo Laptops and 5 Acer Laptops. Two Laptops are selected at random, find the probability that (i) both Laptops are Lenovo (ii) Dell and Acer Laptops.

Solution

Let, Dell Laptop = D, Lenovo Laptop = L and Acer Laptop = A

Total Laptops = 6D + 9L + 5A = 20 Laptops

Laptop selected = 2

2 Laptops can be selected from 20 balls in C(20, 2) ways

Hence, N = C(20, 2) = 190

2 Lenovo Laptop can be selected from 9 Lenovo Laptop in C(9, 2) ways

Hence, m (2L) = C(9, 2) = 36

1 Acer and 1 Dell Laptop can be selected from 5 Acer and 6 Dell Laptop in C(5, 1) × C(6, 1) ways.

Hence, m(AD) = C(5, 1) × C(6, 1) = 5 × 6 = 30

Now, $P(2\,L) = \frac{m(2L)}{N} = \frac{36}{190} = 0.1894$

$P(AD) = \frac{m(AD)}{N} = \frac{30}{190} = 0.1578$

Statistical or Empirical or Posteriori or Relative Frequency Approach

If an experiment is repeated a large number of times under essentially homogeneous and identical condition, then the limiting value of the ratio of the number of times the event occurs to the number of trails, as the number of trials becomes indefinitely large, is called the probability of happening of the event, it being assumed that the limit is finite and unique.

Suppose an event A occurs m times in N repetition of a random experiment. Then the ratio m/N, known as relative frequency, tend to a unique number and N becomes indefinitely large that number is known as probability of that event.

Mathematically,

$$P(A) = \lim_{N \to \infty} \frac{m}{N}$$

Remarks:

1. Large number repetition of random experiment may not be identical and homogeneous.
2. However large the number of repetition of experiment the ratio may attain an unique value.

Example 7. 220 heads are obtained on tossing a coin 450 times. What is the probability of getting tail?

Solution

Let H denotes head and T denotes tail

Here, N = 450, n(H) = 220 , n(T) = 450 – 220 = 230

Scanned by CamScanner

$$P(T) = \frac{n(T)}{N} = \frac{230}{450} = 0.511$$

Subjective Definition of Probability

It is probability of occurrence of an event by personal judgment, belief, intuition and expertise. It is used when both definition of probability namely mathematical and statistical definition of probability are not applicable. It reflects degree of personal belief in likelihood of occurrence of an event. It differs from person to person. It contains high degree of personal bias. It contains no calculation and reflects the personal opinion on basis of past experience.

Axiomatic definition of Probability

Given a sample space of a random experiment, the probability of the occurrence of an event A is defined as a set function P(A) satisfying the following axioms.

1. P(A) is defined , is real and non-negative. i.e. $P(A) \geq 0$.(axiom of non-negativity)
2. $P(\Omega) = 1$. (Axiom of certainty) , Ω is the sample space.
3. If A_i, i=1, 2, ...n are any finite or infinite sequence of disjoint events of Ω , then (axiom of additivity)

Note:

Probability of impossible event is zero.

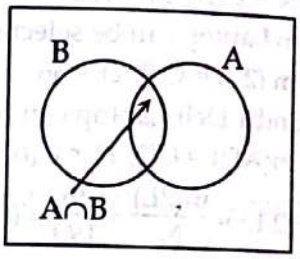
Probability of complementary event $\bar{A}$ of A is given by

$P(\bar{A}) = 1 - P(A)$

For any two events A and B;

$P(\bar{A} \cap B) = P(B) - P(A \cap B)$

Similarly, $P(A \cap \bar{B}) = P(A) - P(A \cap B)$



Laws of Probability

Addition Theorem of Probability

1. The probability of occurrence of at least one of two mutually exclusive events is given by sum their individual probabilities.

   If A and B are two mutually exclusive events, then the probability of happening of either A or B or both is given by $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$

   For three mutually exclusive events A, B and C

   $P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C)$

   In general, n mutually exclusive events

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = \sum_{i=1}^{n} P(A_i)$$

2. If A and B are not mutually exclusiveevents then the probability of happening to at least one of the event is given by

   $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

For three events A,B and C

$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

Note that: $P(A \text{ or } B) = P(A \cup B) = 1 - P(\overline{A \cup B}) = 1 - P(\bar{A} \cap \bar{B})$

Here A and B should be independent of each other.

Example 8. A card is drawn from a pack of 52 cards. What is the probability that selected card is (i) King or Jack (ii) Queen or Spade

Scanned by CamScanner

Solution

Let us consider, King = K, Jack = J, Queen = Q and Spade = S

Total number of cards (N) = 52

Favorable number of cards for King m(K) = 4

Favorable number of cards for Jack m(J) = 4

Favorable number of cards for Queen m(Q) = 4

Favorable number of cards for Spade m(S) = 13

Favorable number of card for Queen and Spade $m(Q \cap S) = 1$

Now,

i) $P(K \cup J) = P(K) + P(J)$

$$= \frac{m(K)}{N} + \frac{m(J)}{N} = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

ii) $P(Q \cup S) = P(Q) + P(S) = P(Q \cap S)$

$$= \frac{m(Q)}{N} + \frac{m(S)}{N} - \frac{m(Q \cap S)}{N} = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Example 9. The probability that A can solve a problem in programming is 4/5, B can solve it is 2/3 and C can solve it is 3/7. If all of them try independently, find the probability that problem will be solved?

Solution

Here, P(A) = 4/5, P(B) = 2/3 and P(C) = 3/7

$P(A \cup B \cup C) = ?$

We know, $P(A \cup B \cup C) = 1 - P(A \cup B \cup C) = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$

$$= 1 - P(\bar{A})\, P(\bar{B}) P(\bar{C})$$

$$= 1 - (1 - 4/5)\,(1 - 2/3)\,(1 - 3/7)$$

$$= 1 - \frac{1}{5} \times \frac{1}{3} \times \frac{4}{7} = \frac{101}{105}$$

Example 10. The odds against X solving a problem on C programming are 8 to 6 and odds in favour of Y solving the same problem are 7 to 8. What is the probability that i) problem will be solved ii) neither solves the problem?

Solution

For X, odds against 8 to 6

$$P(X) = \frac{6}{8+6} = \frac{6}{14}$$

For Y, odds in favour 7 to 8

$$P(Y) = \frac{7}{7+8} = \frac{7}{15}$$

i) $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) = P(X) + P(Y) - P(X)\,P(Y)$

$$= \frac{6}{14} + \frac{7}{15} - \frac{6}{14} \times \frac{7}{15} = \frac{146}{210} = \frac{73}{105}$$

ii) $P(\overline{X \cup Y}) = 1 - P(X \cup Y)$

$$= 1 - \frac{73}{105} = \frac{32}{105}$$

Example 11. For any two events A and B, P(A) = P(B), $P(A \cup B) = 0.7$, $P(A \cap B) = 0.1$. Find (i) P(B) (ii) $P(\bar{A})$ (iii) $P(\bar{A} \cap B)$

Scanned by CamScanner

Solution

We know

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

or $0.7 = P(A) + P(A) - 0.1$

or $0.8 = 2P(A)$

or $P(A) = 0.4$

Now, $P(B) = P(A) = 0.4$

$P(\bar{A}) = 1 - P(A) = 1 - 0.4 = 0.6$

$P(\bar{A} \cap B) = P(B) - P(A \cap B) = 0.4 - 0.1 = 0.3$

Multiplicative Law of Probability (Law of Compound Probability)

1. The probability of simultaneous happening two independent events A and B is given by product of their individual probabilities.

   $P(A \cap B) = P(A \text{ and } B) = P(A)\,P(B)$

   For three independent events A, B and C

   $P(A \cap B \cap C) = P(A \text{ and } B \text{ and } C) = P(A)\,P(B)\,P(C)$

2. The probability of simultaneous happening of two dependent events, A and B, is given by the product of probability of event A and conditional probability of event B given that A has already happened or the product of probability of event B and conditional probability of event A given that B has already happened. Mathematically,

   $P(A \cap B) = P(A \text{ and } B) = P(A)\,P(B/A)$

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

   For three dependent events A, B and C

   $P(A \cap B \cap C) = P(A \text{ and } B \text{ and } C) = P(A)\,P(B/A)\,P(C/A \cap B)$

Note:

If A and B are independent $P(A/B) = P(A)$ and $P(B/A) = P(B)$

If A, B and C are independent, $P(B/A) = P(B)$, $P(C/A \cap B) = P(C)$ and so on.

Note:

If A and B are independent then A and $\bar{B}$ are also independent

$P(A \cap \bar{B}) = P(A)\,P(\bar{B})$

If A and B are independent then $\bar{A}$ and $\bar{B}$ are also independent

$P(\bar{A} \cap \bar{B}) = P(\bar{A})\,P(\bar{B})$

Example 12. A person completes 3 projects out of 4. Another person doing the similar project completes 2 out of 3 shots. Find the probability of completing project by both when both try independently.

Solution

Let A = first person completes the project

B = second person completes the project

Then $P(A) = \frac{3}{4}$, $P(B) = \frac{2}{3}$

Probability of completing the project $= P(A \cap B) = P(A) \cdot P(B) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$

Example 13. A lab contains 6 desktops and 8 laptops. Two computers are selected randomly one after other without replacement. Find the probability that both computers are laptop.

Solution

Let Desktop = D and laptop = L

Total computers = 6D + 8L = 14

Computers selected = 2

Scanned by CamScanner

I be the first selection and II be the second selection respectively.

Now,

$$P(L_I \cap L_{II}) = P(L_I)\, P(L_{II}/L_I)$$

$$= \frac{8}{14} \times \frac{7}{13} = \frac{4}{13}$$

Example 14. The odd that Boy speaks truth is 3:2 and odd that Girls speaks the truth is 5:3. Find the probability that they contradict each other.

Solution

Let B = Boy speaks truth, G = Girl speaks truth

$P(B) = 3/5$, $P(\bar{B}) = 2/5$

$P(G) = 5/8$, $P(\bar{G}) = 3/8$

The event that they can contradict each other in following ways $B \cap \bar{G}$ or $G \cap \bar{B}$

Hence, probability they contradict each other = $P(B \cap P(B \cap \bar{G}$ or $G \cap \bar{B})$

$$= P(B)\, P(\bar{G}) + P(G)\, P(\bar{B})$$

$$= \frac{3}{5} \times \frac{3}{8} + \frac{5}{8} \times \frac{2}{5} = 0.475$$

Hence Boy and Girl can contradict each other with probability 0.475 i.e. 47.5% of cases.

Conditional Probability

For any two dependent events the probability of an event given that other has already happened is called conditional probability

The conditional probability of event A given that B has already happened is denoted by P(A/B) and given by $P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$ provided that $P(B) > 0$.

The conditional probability of event B given that A has already happened is denoted by P(B/A) and given by $P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$ provided that $P(A) > 0$.

Example 15. The IT trainee will remain in the company with probability 0.6. The probability that an employee earns more than Rs. 52000 per month is 0.5. The probability that an employee is a IT trainee who remain with the company or who earns more than Rs. 10000 per month is 0.7. What is the probability that an employee earns more than Rs. 10000 per month, given that he is anIT trainee who stayed in the company?

Solution

Let us define, A= IT trainee will remain with the company

B = An employee earns more than Rs. 52,000 per month

Then $P(A) = 0.6$

$P(B) = 0.5$

Now, P(IT trainee remains with the company or earns more than Rs. 52000 per month) = 0.7

Or $P(A \cup B) = 0.7$

Or $P(A) + P(B) - P(A \cap B) = 0.7$

Or $0.6 + 0.5 - P(A \cap B) = 0.7$

Or $P(A \cap B) = 0.4$

Then the probability is given by

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.4}{0.6} = \frac{2}{3}$$

Scanned by CamScanner

Example 16. The staff body of a IT company is composed of 60% male and 40% female. 40% of male and 60% of female shows interest in sports. What is the probability that a staff selected at random shows interest in sports given that the staff is female?

Solution

Let M = staff is male, F = staff is female and S = staff show interest in sports

Here, P(M) = 60% = 0.6

P(F) = 40% = 0.4

$P(M \cap S)$ = 40% of 0.6 = 0.24

$P(F \cap S)$ = 60% of 0.4 = 0.24

Now, $P(S/F) = \frac{P(S \cap F)}{P(F)} = \frac{0.24}{0.4} = 0.6$

Example 17. The following table shows the survey result regarding the purchase behavior of TV's and DVD players in the last six months of 300 house hold.

Purchase TV	Purchase DVD		
	Yes	No	Total
Yes	38	42	80
No	70	150	220
Total	108	192	300

a. Find the probability that a randomly selected household that purchased a TV.

b. Find the probability that a randomly selected household that purchased a TV and a DVD player.

c. What is the probability that he/she purchased a TV or DVD player?

d. What is the probability that a household has purchased a DVD player given that household purchased a TV?

Solution:

Purchase TV	Purchase DVD		
	Yes	No	Total
Yes	38	42	80
No	70	150	220
Total	108	192	300

a. Probability of household that purchased a TV = $P(TV) = \frac{80}{300} = 0.267$.

b. Probability of household that purchased a TV and a DVD player

$P(TV \& DVD) = \frac{38}{300} = 0.127$

c. Probability that he/she purchased a TV or DVD player = P(TV or DVD)

$= P(TV) + P(DVD) - P(TV\&DVD) = \frac{80}{300} + \frac{108}{300} - \frac{38}{300} = \frac{150}{300} = 0.5$

d. Probability of purchase DVD given he or she has TV = $P(DVD/TV) = \frac{38}{80} = 0.475$.

Example 18. A sample of 500 respondents was selected in a large metropolitan area to determine various information concerning consumer behaviour. Among the question asked was "Do you enjoy shopping for clothing?" Of 240 males 136 answered yes. Of 260 females, 224 answered yes. What is the probability that a respondent chosen at random

a. Enjoys shopping for clothing?

Scanned by CamScanner

b. Is a female and enjoy shopping for clothing?
c. Is a female or enjoy shopping for clothing?
d. Suppose the respondent chosen enjoys shopping for clothing, what then is the probability that the individual is a male?

Solution

Keeping these information in table form

Bank credit card	Enjoying shopping Clothing		
	Yes	No	Total
Male	136	104	240
Female	224	36	260
Total	360	140	500

a. Probability of Enjoys shopping for clothing = P(e)=360/500=0.72

b. Probability of female and enjoy shopping for clothing = $P(f\&e)=\frac{224}{500}=0.448$.

c. Probability of female or enjoy shopping for clothing

$= P(f \text{ or } e)=P(f)+P(e) - P(f\&e)=\frac{260}{500}+\frac{360}{500}-\frac{224}{500}=\frac{396}{500}=0.792$

d. Probability that the individual is a male given that he enjoyed for clothing

$= P(m/e)=\frac{136}{360}=0.378.$

Bayes Theorem, Prior and Posterior Probabilities

It is theorem given by Baye's. It gives prior probabilities on basis of information obtained from an experiment. Prior probabilities are the probabilities of some events that have been known before we gain information from the experiment. The probabilities obtained after new information obtained from the experiment are called posterior probabilities.

The conditional probability allows us to calculate the probability of some event given that some other event has already occurred. In many situations however we are given the outcome of a random experiment and are interested in determining the probability that the outcome was due to a particular cases or it was the result of some other events in such case Baye's theorem is used.

Bayes' Theorem (Laws for Inverse Probability)

The conditional probability allows us to calculate the probability of some event given that some other event has already occurred. In many situations however we are given the outcome of a random experiment and are interested in determining the probability that the outcome was due to a particular cases or it was the result of some other events. Such probability are known as inverse probability. The formula used to calculate the inverse probability is called Bayes formula.

Let Ai, i= 1, 2 ...n, are mutually exclusive and exhaustive events defined in the sample space Ω. B be any other event defined in Ω such that P (B)>0. Then,

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

Example 19. A company has two plants to manufacture computer and its accessories. Plant I manufactures 80% of computers and plant I manufactures 20%. At plant I, 85 out of 100 computers are rated standard quality or better. At plant II, 65 out of 100 are rated standard or better. What is the probability that computer selected at random came from plant I if it is of standard quality? What is the probability that computer selected is of standard quality?

Scanned by CamScanner

Solution

Let us define, A_1 = computer is from plant I

A_2 = computer is from plant II

B = event that the computer is of standard quality

Then we have,

$$P(A_1) = 0.8,\ P(A_2) = 0.2$$
$$P(B/A_1) = 0.85,\ P(B/A_2) = 0.65$$

Using Bayes theorem

Probability that computer is produced from plant I given that it is of standard quality

$$= P(A_1/B)$$
$$= \frac{P(A_1)\,P(B/A_1)}{P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2)}$$
$$= \frac{0.8 \times 0.85}{0.8 \times 0.85 + 0.2 \times 0.65} = 0.16$$

Probability that the computer selected at random is of standard quality

$$P(B) = P(A_1)\,P\left(\frac{B}{A_1}\right) + P(A_2)\,P\left(\frac{B}{A_2}\right)$$
$$= 0.8 \times 0.85 + 0.2 \times 0.65 = 0.81$$

Example 20. The members of a consulting firm rent cars from three rental agencies: 60% from agency A, 30% from agency B and 10% from agency C. If 9% of the cars from agency A need a tune up, 20% of the cars from agency B need a tune up and 6% of the cars from agency C need a tune up. What is the probability that a rental car delivered to the firm will need a tune up?

Solution

Let A_1, A_2 and A_3 be the events that car comes from agencies A , B and C respectively. Also let B be the event that car need tune up.

Here, $P(A_1) = 60\% = 0.6$

$P(A_2) = 30\% = 0.3$

$P(A_3) = 10\% = 0.1$

$P(B/A_1) = 9\% = 0.09$

$P(B/A_2) = 20\% = 0.2$

$P(B/A_3) = 6\% = 0.06$

Now, Probability that a rental car delivered to firm need tune up

$$P(B) = \sum_{i=1}^{n} P(A_1)\,P(B/A_1)$$
$$= 0.6 \times 0.09 + 0.3 \times 0.2 + 0.1 \times 0.06 = 0.12$$

Example 21. Urn A contains 2 white, 1 black and 3 red balls, urn B contains 3 white, 2 black and 4 red balls and urn C contains 4 white, 3 black and 2 red balls. One urn is chosen at random and 2 balls are drawn, they happen to be red and black. What is the probability that both balls came from urn B?

Solution:

Let,

Red ball = R, Black ball = B, White ball = W

Urn A : 2W + 1B + 3R = 6

Urn B : 3W + 2B +4R = 9

Urn C : 4W +3B + 2R = 9

Let us consider, selection of urn A, B and C are A_1, A_2 and A_3 respectively

Two balls drawn happen to be red and black is B

Scanned by CamScanner

Since one urn is chosen at random, $P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$

$$P(B/A_1) = \frac{C(3,1) \times C(1,1)}{C(9,2)} = \frac{3}{15} = \frac{1}{5}$$

$$P(B/A_2) = \frac{C(4,1) \times C(2,1)}{C(9,2)} = \frac{4 \times 2}{36} = \frac{2}{9}$$

$$P(B/A_3) = \frac{C(2,1) \times C(3,1)}{C(9,2)} = \frac{2 \times 3}{36} = \frac{1}{6}$$

$$P(B) = \sum_{i=1}^{n} P(A_1)\, P(B/A_1)) = \frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{2}{9} + \frac{1}{3} \times \frac{1}{6} = 0.159$$

Now,

$$P(A_2/B) = \frac{P(A_2)\, P\left(\frac{B}{A_2}\right)}{P(B)} = \frac{\frac{1}{3} \times \frac{2}{9}}{0.159} = \frac{0.074}{0.159} = 0.465$$

Example 22. A factory produces a certain type of output by three types of machines. The respective daily production figures are MachineI:3000 units, Machine II: 2500 units, Machine III: 4500 units. Past experience shows that 3% of the output produced by machine I is defective. The corresponding fraction of defectives for the other two machines are 1.2% and 2% respectively. An item is drawn at random from the day's production and is found to be defective. What is the probability that it comes from the output of (i)Machine I (ii) Machine II (iii) Machine III?

Solution

Let A_1, A_2 and A_3 be the production from machine I, II and III respectively. Also, D be the defective output.

Now

$$P(A_1) = \frac{3000}{3000 + 2500 + 4500} = 0.3$$

$$P(A_2) = \frac{2500}{3000 + 2500 + 4500} = 0.25$$

$$P(A_3) = \frac{4500}{3000 + 2500 + 4500} = 0.45$$

Also,

$$P(D/A_1) = 3\% = 0.03$$
$$P(D/A_2) = 1.2\% = 0.012$$
$$P(D/A_3) = 2\% = 0.02$$
$$P(D) = \sum_{i=1}^{n} P(A_i)\, P(D/A_i) = 0.3 \times 0.03 + 0.25 \times 0.012 + 0.45 \times 0.02 = 0.021$$

i) $P(A_1/D) = \frac{P(A_1)\, P(D/A_1)}{P(D)} = \frac{0.3 \times 0.03}{0.021} = 0.4285$

ii) $P(A_2/D) = \frac{P(A_2)\, P(D/A_2)}{P(D)} = \frac{0.25 \times 0.012}{0.021} = 0.1428$

iii) $P(A_3/D) = \frac{P(A_3)\, P(D/A_3)}{P(D)} = \frac{0.45 \times 0.02}{0.021} = 0.4285$

Scanned by CamScanner

Baye's Theorem for Future Events

Let A_i, i = 1, 2 ... n, are mutually exclusive and exhaustive events defined in the sample space Ω. B be any other event defined in Ω such that P (B)>0 then probability of happening of A_i when B has occurred and C is future event in regard to B then

$$P(C/B) = \frac{\sum_{i=1}^{n} P(A_i)\, P(B/A_i)\, P\{C/(A_i \cap B)\}}{\sum_{i=1}^{n} P(A_i)\, P(B/A_i)}$$

Example 23. Urn A contains 2 red and 3 green marbles, urn B contains 3 red and 4 green marbles, urn C contains 4 red and 2 green marbles. Au urn is selected at random and marble is drawn from it. If drawn marble is found to be green, then what is the probability of drawing next green marble if first marble is not replaced?

Solution

Let, A_1, A_2 and A_3 be the event of selecting urn A , B and C respectively.

Let B be the event of drawing green marble and C be the future event of drawing green marble.

Here, $P(A_1) = P(A_2) = P(A_3) = 1/3$

$P(B/A_1) = 3/5$, $P(B/A_2) = 4/7$, $P(B/A_3) = 2/6$

$P(C/A_1 \cap B)=1/2$, $P(C/A_2 \cap B)= 1/3$, $P(C/A_3 \cap B)= 1/5$

$$P(C/B) = \frac{\sum_{i=1}^{3} P(A_i)\, P(B/A_i)\, P\{C/(A_i \cap B)\}}{\sum_{i=1}^{3} P(A_i)\, P(B/A_i)}$$

$$= \frac{\frac{1}{3}\times\frac{3}{5}\times\frac{1}{2}+\frac{1}{3}\times\frac{4}{7}\times\frac{1}{3}+\frac{1}{3}\times\frac{2}{6}\times\frac{1}{5}}{\frac{1}{3}\times\frac{3}{5}+\frac{1}{3}\times\frac{4}{7}+\frac{1}{3}\times\frac{2}{6}}$$

$$= \frac{\frac{63+40+14}{630}}{\frac{126+120+70}{630}}$$

$$= \frac{117}{316}$$

$$= 0.37$$

Hence, probability of getting next green marble if drawn green marble is not replaced = 0.37

Scanned by CamScanner

Exercise – 3

Theoretical Questions

1. Explain the terms with examples: Sample space, Events, Mutually exclusive events.
2. What do you mean by probability? Give mathematical or classical definition of probability. Mention its limitations.
3. Explain the terms with examples: Trial, Exhaustive number of cases, Favorable number of cases, equally likely cases, dependent cases, independent cases, Bays theorem.
4. Discuss the axiomatic definition of probability.
5. What do you mean by conditional probability? Give examples.

Numerical Questions

1. If two dice are thrown, what is the probability that the sum is i) greater than 9 ii) neither 8 nor 10. **[Ans: 1/6, 7/9]**
2. A secretary types 5 letters and 5 envelopes and puts the letters into the envelopes at random. What is the probability that letters are not put into correct envelop? **[Ans: 119/120]**
3. What is the probability that a leap year selected at random contain 53 Saturdays? **[Ans: 2/7]**
4. A problem in Statistics is given to three students A, B & C whose chances of solving are 1/5, 2/5 &3/5 respectively. What is the probability that problem will be solved? **[Ans: 101/125]**
5. The odds against solving a problem by B is 4 to 3.The odds in favor of solving the same problem by A is 7 to 6. What is the probability that problem will be solved if both of them try independently? **[Ans: 67/91]**
6. From a pack of 52 cards 2 cards are drawn at random .Find the probability that (i) one is king & other is queen (ii) both are face cards. (iii) One is spade and other is club. **[Ans: 8/663, 11/221, 13/102]**
7. Probability that a student passes an exam in Numerical method is 0.75 and in Computer graphics is 0.85. Find the probability that student passes in i) both subjects ii) at least one subject. **[Ans: 0.637 , 0.962]**
8. If $P(A) = 0.4$, $P(B) = 0.3$ and $P(A \cup B) = 0.58$. Are A & B independent? **[Ans: indep]**
9. A and B play a game of throwing a die. A begins the game and the game terminates with one of them throwing 2 first for a win. Find the probability of A's and B's winning the game. **[Ans: 6/11, 5/11]**
10. The probability that an integrated circuit chip will have defective etching is 0.12, the probability that it will have a crack defect is 0.29, and the probability that it has both defect is 0.07. What is the probability that a newly manufactured chip will have either an etching or crack defect. **[Ans: 0.34]**
11. The probability that a 50 yrs. old man will alive at 60 is 0.83 and the probability that a 45 yrs. old woman will be alive at 55 is 0.87. What is the probability that a man who is 50 and woman who is 45 (i) both will alive 10 yrs. hence (ii) at least one will alive 10 yrs. hence. **[Ans: 0.722, 0.977]**
12. A card is drawn at random from a pack of 52 cards. What is the probability of getting (i) King or Queen (ii) Jack or Club? **[Ans: 2/13 , 4/13]**
13. Two cards are drawn at random one after another from a pack of 52 cards. What is the probability of getting (i) First black card and second red card (ii) First Ace and second Ace , if first draw is not replaced before second draw. **[Ans: 13/51 ,1/221]**

Scanned by CamScanner

14. A book case contains 6 digital logic books and 9 microprocessor books. Four books are selected at a time. Find the probability for the first selection to give 4 digital logic and second to give 4 microprocessorbooks if (i) the books are replaced before the second selection. (ii) the books are not replaced before the second selection. [Ans: 6/5915 , 3/715]

15. The student body of a college is composed of 60% girls and 40% boys. 40% of girls and 60% of boys show interest to indulge in mobile Apps. What is the probability that a student who shows interest in mobile Apps given that she is a girl? [Ans: 0.4]

16. A and B toss a coin alternately on the understanding that the first who obtains the head will be Microsoft student partner. If A starts show that their respective chances to become student partner are 2/3 and 1/3.

17. A person is known to hit a target in four out of five shots. Another person hits the target three out of four shots. Find the probability of the target being hit at all when they both try? [Ans: 3/5]

18. A six faced dice is so biased that it is twice as likely to show an even number as an odd number when it is thrown. What is the probability that the sum of two numbers thrown is even? [Ans: 5/9]

19. In tossing a coin three times, find the probability of getting (i) no heads (ii) two heads (iii) at least two heads. [Ans: 1/8, 3/8, 1/2]

20. There are three traffic light on your way home. As you arrive at each light assume that it is either red (R) or green (G) and that it is green with probability 0.7, construct the sample space by listing all possible eight sample events. Assign the probability to each sample event. Are the events equally likely? What is the probability that you stop no more than one time. [Ans: 0.78]

21. The odds that a book of digital logic will be favorably reviewed by three independent critics are 3 to 2, 4 to 3 and 2 to 3 respectively. What is the probability that of the three reviews a majority will be favorable? [Ans: 94/175]

22. In the past several years, credit card companies have made an aggressive effort to solicit new accounts from college students. Suppose that a sample of 200 students at your college indicated the following information as to whether the student possessed a bank credit card &/or a travel and entertainment credit card.

Bank credit card	Travel & Entertainment credit card	
	Yes	No
Yes	60	60
No	15	65

a. What is the probability that the student has a bank credit card?

b. What is the probability that the student has a bank credit card and a travel and entertainment card?

c. What is the probability that the student has a bank credit card or a travel and entertainment card?

d. Known that the student has a bank credit card, what is the probability that he orshe has a travel and entertainment card? Ans: (a) 0.60 (b) 0.30 (c) 0.675 (d) 0.50

23. A company has made available to its employees (without charge) extensive health club facilities that may be used before work, during lunch hour, after work, and on weekends. Records for the last year indicate that of 250 employees, 110 used the facilities at sometime of 170 males employed by the company, 65 used the facilities:

Scanned by CamScanner

Set up the (2×2) table to evaluate the probabilities.

a. What is the probability that an employee chosen at random is a female and has used the health club facilities?

b. What is the probability that an employee chosen at random is a male?

c. What is the probability that an employee chosen at random is a male or has not used the health club facilities?

Ans: (b) 0.18 (c) 0.68 (d) 0.82

24. We are given a box containing 5000 IC chips , of which 1000 are manufactured by company X and the rest by company Y. Ten percent of the chips made by company x and five percent of the chips made by company Y are defective. If a randomly chosen chip is found to be defective, find the probability that it came from company X. **Ans: 0.33**

25. In a certain assembly plant, three machines A_1, A_2& A_3 make 30%, 45% and 25% respectively of the product. It is known from past experience that 2%, 3% and 2% of the products made by each machine A_1, A_2 & A_3 respectively are defective. If a product was chosen randomly and found to be defective, what is the probability that it was made by machine A_3. **Ans: 0.204**

26. In a college the geographical distribution of students is as follows: 50% come from central development region, 30% come from eastern, 10% come from western, 8% come from mid-western and 2% come from far western. The following proportion of students have black hair; 80% of central, 70% of eastern, 60% of western, 65% of mid-western and 75% of far western. Find the probability that a student who has black hair come from the central region.

Ans: 0.542

27. A consulting firm rents car from three agencies, 20% from agency D, 20% from agency E and 60% from agency F. If 10% of cars from D, 12% of cars from E and 4% of cars from F have bad tires, what is the probability that the firm will get a car with bad tire? **Ans: 0.068**

28. Urn A contains 2 white, 1 black and 3 red balls. Urn B contains 3 white, 2 black and 4 red balls and urn C contains 4 white,3 black and 2 red balls. One urn is chosen at random and two balls are drawn. They happen to be red and black. What is the probability that both balls came from urn B? **Ans: 0.377**

29. The chances of X, Y, Z becoming manager of a Worldlink Pvt. Ltd. are4:2:3 .The probabilities that bonus scheme to the staff will be introduced if X,Y,Z become managers are 0.3, 0.5 and 0.8 respectively. If the bonus scheme has been introduced, what is the probability that X is appointed as the manager? **Ans: 0.26**

30. A television manufacturing company has three assembly line A, B and C manufactures 25%, 35% and 40% of total television. Of their output 5%, 4% and 2% are defective. A television is drawn at random from the product and is found to be defective. What is the probability that it was manufactured by assembly line A, B and C?

Ans: 0.362, 0.405, 0.231

31. A new computer program consists of two modules. The first module contains an error with probability 0.2. The second module is more complex; it has a probability of 0.4 to contain an error, independently of the first module. An error in the first module alone causes the program to crash with probability 0.5. For the second module, this probability is 0.8. If there are errors in both modules, the program crashes with probability 0.9. Suppose the program crashed. What is the probability of errors in both modules? **Ans: 0.18**

Scanned by CamScanner

32. Three urns are given each containing red and white chips as indicated.

Urn 1: 6 red and 4 white

Urn 2: 2 red and 6 white

Urn 3: 1 red and 8 white

(i) An urn is chosen at random and a chip is drawn from the urn, the chip is red. Find the probability that the chosen urn was urn 1

(ii) An urn is chosen at random and two chips are drawn without replacement from the urn. If both chips are red, find the probability that urn 1 was chosen. Under the condition what is the probability that urn 3 was chosen. **Ans: 0.624, 0.903, 0**

❑❑❑

Scanned by CamScanner

Chapter 4

SAMPLING

Population

In an investigation procedure, an investigator usually deals with the *general magnitude* and the *study of variation* of one or more characteristics of individual belonging to a group. A group or aggregate of all objects/units/individuals/items etc. related to the topic of the study is called a population. The number of units in the population is called population size. A population may be finite or infinite.

Finite Population

A population is said to be a finite population if the number of units in the population is countable in certain time. For example,

(a) Total number of students in TU

(b) Total number of households in Kathmandu

(c) Total number of employees in NTC

(d) Total number of districts in Nepal

Infinite Population

A population is said to be infinite population if the number of units in the population is not countable in certain time. For example,

(a) Total number of fish in Pacific ocean

(b) Total number of trees in a forest

(c) Total number of stars in the sky

Complete enumeration of all units of the population is called census. Census is possible when the population is finite. Due to the lack of human resource for data collection or due to the lack of sufficient resources or due to the case of vanishable (perishable) items census is not possible in all case. So we have to draw sample to study.

Sample Survey verses Census Survey

A study conducted on all units of the population is called census. In other words, a complete enumeration of all elements of the population is called census survey. For example,

- Population census survey

Scanned by CamScanner

- Industrial census survey
- Agricultural census survey
- National Economic census

Merits of census survey

- It is suitable for small population.
- It is appropriate for historical study.
- If information about all units of population is required, it is only an alternative method.

Demerits of census survey

- It is an expensive method.
- It requires large number of resources.
- It needs longer time to complete the study.
- For infinite population, it is not possible.

Sample

Some units selected from population which is representative of population is called sample. Sampl survey is conducted to enumerate all the selected sample units. Sample is divided into two type's (random sample (ii) non random sample.

i. **Random sample**

On selecting units from population each time unit selected has equal chance to appear in th sample is called random sample.

ii. **Non random sample**

On selecting units from population each time unit selected has no chance to appear in th sample is called non random sample.

Sampling Unit

The population units selected as a sample from the population is called sampling unit. It i individual item in the sample .For example a house, a college, a district, a city, an individual etc.

Sampling Frame

It is a list of all the elements that are in the population. When the population is defined according t the characteristic under study it is known as frame consisting of the sampling units in th population. It is used to select sample units from the population as a sample. Frame or the list ma not be readily available in practice so it has to be prepared for the selection of the sample befor conducting the main survey.

It should be exhaustive, up-to-date and should consist of all the information about the units so tha an appropriate sample design may be planned. It should be suitable for the coverage of the stud and should be restored by reliable sources. It should be free of errors like repeated units, missin elements, non-coverage, incompleteness, clustering of elements, blanks, foreign elements etc. th effective way of solving the frame problems are to correct entire population list or redefine th population to fit the frame or ignore and discard the problem if it cannot be fixed.

Scanned by CamScanner

Sampling

When one by one study of all units of a population is not possible due to some factors like time, cost, manpower, recourses and destructive nature of study, we take a small representative part from the population for the study. This small representative part selected for the study from the population is called sample.

The process of selecting a sample from a population is called *sampling*. For example,

- A housewife takes only two or three grains of rice as a sample from the cooking pan to know whether the rice is property cooked or not.
- A pathologist takes a syringe of blood as a sample to find out a disease.
- Selection of word during speech recognization with maximum probability

Merits of sample survey

- It is less expensive.
- It requires less time to get result.
- It needs smaller number of resources.
- It is the only method for large population.

Demerits of sample survey

- It is not suitable method if information about all units of population is required.
- It is not suitable for rare events.
- It is not appropriate for historical data.

Parameter and Statistic

There are various statistical measures in statistics such asmean, median, mode, standard deviation, proportion, correlation coefficient etc. These measures can be computed from both population data and sample data.

Any statistical measurecomputed from population data is called parameter. Let $X_1, X_2, \ldots\ldots X_N$ be the values in the population of size N, then

(a) Population Mean $(\mu) = \dfrac{X_1+X_2+\ldots.+X_N}{N}$

(b) Population Variance $(\sigma^2) = \dfrac{\Sigma(X-\mu)^2}{N}$

Here, population mean (μ) and population variance (σ^2) are called population parameters.

Any statistical measure computed from sample data is called statistic. . Let $x_1, x_2, \ldots\ldots x_n$ be the values in the sample of size n, then

(a) Sample Mean $(\bar{x}) = \dfrac{x_1+x_2+\ldots.+x_n}{n}$

(b) Sample Variance $(s^2) = \dfrac{\Sigma(x-\bar{x})^2}{n-1}$ = sample mean square

Here, sample mean $(\bar{x})$, sample variance (s^2) are called statistic.

The value of every statistic varies from one sample to another whereas the value of a parameter is always constant for a population. There is a natural correspondence between statistic and parameter. For each population parameter there is a sample statistic computed from data that represents the best information about the unknown population parameter.

Scanned by CamScanner

Sample Size

The size of the sample is an important factor. It has direct bearing on the accuracy, estimation and administration of the survey. Large sample has low sampling error where as small sample higher sampling error. To avoid unnecessary cost small sample should be selected. Hence optimum sample size should be selected to fulfill the requirement of efficiency, representativeness, reliability and flexibility. Some of the factors affecting the sample size are nature of study, nature of reaction respondent towards the subject under study, nature of population i.e. composition of population under study, number of classes in the population, types of sampling used during study etc.

Sampling Distribution

When a sample of size n is to be selected from a population of size N, the possible number different samples is ${}^{N}C_{n}$. Hence, there can be ${}^{N}C_{n}$ different estimates of the population parameter. Such estimates of the population (statistics) follows a distribution. The distribution followed by the sample statistics is known as sampling distribution of that statistics. Thus, we can have sampling distribution of sample mean, sample variance etc.

The standard deviation of the sampling distribution of sample statistic is known as standard error of that statistics. It is denoted by S.E. some its examples are as follows:

S. No.	Statistic	Standard Error
1	Sample mean ($\bar{x}$)	$\sqrt{\frac{\sigma}{n}}$ for SRSWR $\sqrt{\frac{(N-n)}{Nn}}$ for SRSWOR
2	Sample proportion (p)	$\sqrt{\frac{PQ}{Nn}}$ for SRSWR $\sqrt{\frac{(N-n)\,PQ}{Nn}}$ for SRSWOR
3	Standard deviation (s)	$\sqrt{\frac{\sigma^2}{2n}}$
4	Sample variance (s^2)	$\sigma^2\sqrt{\frac{2}{n}}$

Factors Affecting Sample Size

An element or sampling unit from which information is collected is called a sample. A sample should be optimum, effective, representative, reliable and flexible. The term sample size refers to the number of items to be selected from the universe to constitute a sample. This is the number of respondents or units in the population included in a sample for studying the population. Some of the factors affecting the sample size are nature of study, nature of reaction of respondent towards the subject under study, nature of population i.e. composition of population under study, number of classes in the population, types of sampling used during study etc The following are the factors affecting the size of sample under study.

- Population parameter to be estimated
- Cost of sampling
- The information known before sampling

Scanned by CamScanner

- Variability of the population
- Practicality of the data collection (how tough to collect)
- Precision required for the research

There must be a statement about what is expectation of sample. The researcher must determine what is to be estimate, how should be the precision of the estimate and what is the application of estimate so determined which is easily determined through objective of the study. The cost of sampling issue helps us determine how precise our estimates should be. This is where the trade-offs usually occurs. Researcher has to take enough observations to obtain reasonably precise estimates of the parameter of interest but researcher want to do this within a practical resource budget. When choosing sample sizes we need to select risk values. If the decision made from sampling activity are very valuable, then researcher should use low risk values and hence the size of sample will be increases. Prior information about the population also helps to reduce the sample size. Stratification can be done to reduce the total size of sample. Similarly variation in population also affects the size of sample. If there is homogeneity in the population small size of sample is sufficient to represent the population. Practicality is the next issue to determine the sample size, if it is hard to take the sample then the size should be taken less.

Errors in Statistics

The errors involved in collection, processing and analysis of data in a survey may be classified as: Sampling error and Non-sampling error.

Sampling Error

Usually, different samples selected from the same population will give different results because they contain different elements. The result obtained from any one sample will generally be different from the result obtained from the corresponding population. The difference between the values of the sample statistic obtained from a sample and the value of corresponding population parameter obtained from the population is called the sampling error.

Sampling error is the difference between the value of a sample statistic and the value of the corresponding population parameter. In case of mean,

$$\text{Sampling error} = \bar{X} - \mu$$

The sampling error occurs because of chance and is inherent and unavoidable in any and every sampling scheme. A sample with the smallest sampling error will always be considered a good representative of the population. This error can be reduced by increasing the size of the sample. The sampling error is inversely proportional to the square root of the sample size and the relationship can be examined graphically as shown in the figure below. When the sample survey becomes a census survey, the sampling error becomes zero.

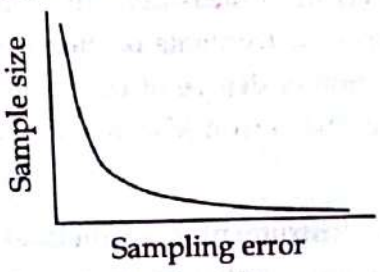


Figure: Relationship of sampling error with sample size.

Increasing sample size, cross check, used of unbiased sampling, using appropriate sampling design and making clear questionnaire are required to reduce the sampling error.

Scanned by CamScanner

Non-sampling Error

The errors that occur in the collection, recording, and tabulation of data are called non-sampling errors. The non-sampling error is likely to increase with increase in sample size, while sampling errors decreases with increase in sample size. Non-sampling errors can be attributed to many factors such as inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of the questions, inability or unwillingness on the part of the respondents to provide correct information, inability of recall information, errors made in collection of data, errors made in processing of data, errors made in estimating value of the parameter etc.

Non sampling error is occurred in research from other sources than the sample, which can be minimize by checking the process, preparing the questionnaire properly, performing pilot survey, fixing procedure, by using competent manpower and experts etc.

Basic Steps in Sampling

The basic steps involved in the planning & execution of a sample survey may be grouped somewhat arbitrarily under the following heads.

1. Objective of the survey

The first step in sampling is to define the objective of the survey in clear & lucid term. The sponsor of the survey should take care that these objective are commensurate with the available resources in terms of money, manpower and the time limit required for availability of the results of the survey.

2. Defining population to be sampled

The population i.e. the aggregate of objects from which sample is chosen should be define in clear and unambiguous terms. The population to be sampled (the sampled population) should be coincide with the population about which information is wanted (The target population). The definition of the population when sampling a batch of electric light bulb in order to estimate the average length of life of a bulb is clear but in sampling a population of farms, rules must be setup to define a farm regarding shape, size etc. keeping in mind the boarder-line cases so as to enable the investigator to decide in the field without much hesitation whether or not to include a given farm in the population.

3. Data to be collected

It is well to verify that all the data are relevant to the purpose of the survey and that no essential data are omitted. There is frequently a tendency, particularly with human population to ask too many question, some of which are never subsequently analyzed. An overlong questionnaire lowers the quality of the answer to important as well as unimportant questions.

4. Degree of precision desired

The result of sample surveys are always subject to some uncertainty because only part of population has been measured and because of errors of measurement. This uncertainty can be reduced by taking larger samples and by using superior instruments of measurement. But this usually cost time and money. Consequently, the specification of degree of precision wanted in the results is an important step. This step is the responsibility of the person who is going to use the data.

5. Method of measurement

There may be choice of measuring instrument & of method of approach to the population. The survey may employ a self-administered questionnaire, an interviewer who reads a standard set of questions with no discretion or an interview process that allows much latitude in the form & ordering of the questions (Schedule). The approach may be by mail, by telephone, by personal visit

Scanned by CamScanner

or its combinations.

6. Frame

Before selecting the sample, the population must be divided into parts that are called sampling units or units. These units must cover the whole of the population and they must not overlap; in the sense that every element in the population belongs to one and only one unit. For example in socio-economic survey for selecting people in a town the sampling unit might be individual person, a family, a household or a block of locality. In order to cover the population decided upon, there should be some list, map or other acceptable materials, called frame, which serves as a guide to the population to be covered. The construction of the frame is often one of the major practical problem, since it is a frame which determines the structure of the sample survey. The list which have been routinely collected for some purpose, are usually found to be incomplete or partially eligible or often contains an unknown amount of duplication such list should be carefully scrutinized & examining to ensure that they are free from those defects and are up to date.

7. Selection of proper sampling design

There is variety of plans by which the sample may be selected. For each plan that is considered rough estimates of the size of sample can be made from knowledge of the degree of precision desired. The relative cost and time involved for each plan are also compared before making decision.

8. Organization of field work

It is essential that the personnel should be thoroughly trained in locating the sample units, recording the measurements, the methods of collection of required data before starting the field work. The success of a survey to a great extent depends upon the reliable field work. It is very necessary to make provisions form adequate supervisory staff for inspection after field work.

It has been found useful to try out the questionnaire and the field method on small scale called pretest. It is always helps to decide upon effective method of asking questions and results in the improvement of the questionnaire. Moreover it might disclose certain problems and troubles that will otherwise be quite serious on a large-scale survey

9. Summary and Analysis of data

The analysis of the data may be broadly classify as

a) Scrutiny & editing of the data

The first step is to edit & scrutiny the completed questionnaire & schedule thoroughly, in the hope of amending recording errors, or at least of deleting data that are obviously erroneous. Decision about computing procedure are needed in cases in which answer to certain questions were omitted by some respondents or were deleted in editing process.

b) Tabulation of data

Depending upon the quantity of data, tabulation can be done with hand tabulation or Machine-tabulation. For large scale survey Machine tabulation using personnel computer will obviously much quicker and economical. For qualitative data coding is essential. The answer of the questions should be processed to mechanical equipment which lead to form table.

c) Statistical Analysis

After the data has been properly scrutinized, edited and tabulated a very careful statistical analysis is to be made. Different methods of estimation may be available for the same data. Appropriate formulae should then be used to provide final estimates of the required estimate.

Scanned by CamScanner

d) Reporting and Conclusion

Finally, a report incorporating detailed statement of the different stages of the survey should be prepared in the presentation of results, it is good practice to report the technical aspects of design viz the type of estimators used along with the amount of error to be expected in the most important estimate.

10. Information gained for future surveys

The more information we have initially about a population, the easier it is to device a sample that will give accurate estimate. Any completed sample is potentially a guide to improved future sampling. In the data it supplies about the means, standard deviations and nature of the variability of the principal measurements and about the cost involved in getting the data.

Fundamental characteristics

For the computation of the characteristics with the help of computers, the literature offers easily programmable expressions. In computing center, standard subroutines facilitating the statistical analysis of the sets of data are very often available. The assessment of the whole set of characteristics derived from the same population involves relatively complex problems.

The most significant are as follows:

(1) the relationship of the characteristics to parameters,
(2) the effect of the number of elements in a sample on the properties of parameter estimates, and
(3) the distribution of the characteristics.

The first of these problems is most conveniently dealt with by comparing the curve of the characteristics with the respective parameter.

Advantages of sampling

The following are the principal advantages of sampling as compared with census.

1. **Reduce Cost**

   If data are secured from only a small fraction of the aggregate, expenditure are smaller than if a complete census is attempted. With large population, results accurate enough to be useful can be obtained from samples that represent a small fraction of the population.

2. **Greater Speed**

   For some reason, the data can be collected & summarized more quickly with a sample that with complete count. This is a vital consideration when information is urgently needed.

3. **Greater Scope**

   In certain type of inquiry highly trained personnel or highly specialized equipment limited in availability, must be used to obtain the data. A complete census is impracticable. The choice lies between obtaining the information by sampling or not at all. Thus surveys that rely on sampling have more scope & flexibility regarding the types of information that can be obtained.

4. **Greater Accuracy**

   Because of higher quality personnel can be employed and given intensive training and because more careful supervision of the field work and processing of results become feasible when volume of work is reduced, a sample may produce more accurate result that the kind of complete enumeration that can be taken.

5. If population is too large, as for example, of trees in a jungle, we are left with no way but to

Scanned by CamScanner

resort to sampling

6. If testing is destructive, i.e. if the measurement can be done only by destroying the article the complete enumeration is impracticable and sampling technique is the only method to be used in such cases, where the process may continue indefinitely, sampling method is the only method of estimating the parameters of the universe.

7. If the population is hypothetical in nature, for example rolling of dice, tossing of coin etc.

From practical point of view, it has been seen that the method of random sampling with suitable adaptation of stratification of the universe if it is heterogeneous.

The technique of multistage random sampling if there is a clear demarcated stages, gives fairly good results, often better than those obtain by complete census.

Limitations of sampling

The advantages of sampling over complete enumeration works only if the sample units are selected using scientific manner, appropriate sampling technique and with adequate sample size.

Some of limitations of sampling over complete census are as follows:

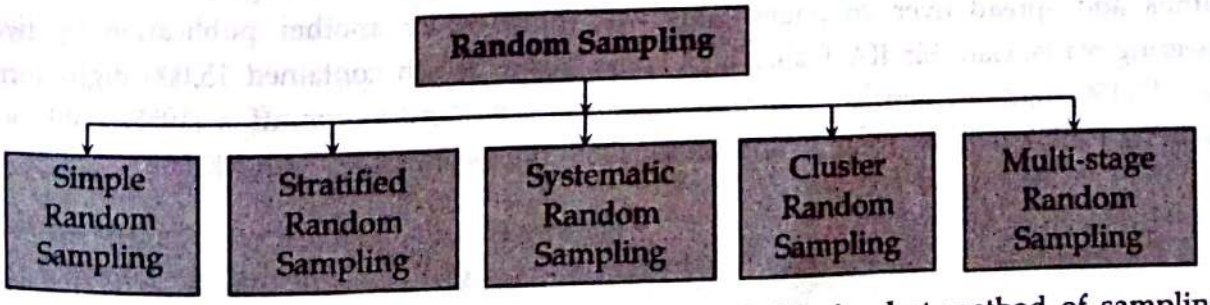
1. Proper care should be taken in the planning and execution of the sample survey, otherwise the results obtained might be inaccurate and misleading.
2. Sampling theory requires the services of trained and qualified personnel and sophisticated equipment for its planning, execution and analysis. In absence of these, the results of the sample survey are not dependable.
3. If the information is required about each and every units of the universe, there is no way but to use complete enumeration. The complete enumeration may be better than any sampling technique of time, money and labour.

Types of Sampling Methods

The sampling method or technique usually depends upon the nature of the population under investigation. The sampling techniques that are commonly used can be classified as:

1. Random or Probability Sampling Technique
2. Non-random or Non-probability Sampling Technique

(1) **Random or Probability Sampling Technique:** It is defined as the method of sampling technique in which each unit of the population has some fixed probability of being selected in the sample. There are several random sampling techniques for selecting samples from a population:



(a) **Simple random sampling:** It is the most common and the simplest method of sampling in which each sample unit is selected from a population with equal probability. There are two types simple random sampling:

Scanned by CamScanner

(i) **Simple Random Sampling with Replacement:** A simple random sampling process which a unit is selected from a population, noted it and then returned back to population before selecting a next unit from the population is called a simple rand sampling with replacement (SRSWR). This process is repeated 'n' times to get a sample 'n' units.

(ii) **Simple Random Sampling without Replacement:** A simple random sampling process which a unit is selected from a population, noted it and not returned back to population before selecting a next unit from the population is called a simple rand sampling without replacement (SRSWOR). This process is repeated 'n' times to get sample of 'n' units.

Selection of Simple Random Sampling

Simple random sampling is the purest form of probability sampling and is commonly used research. There are two aspects of performing simple random sampling viz SRS with replacement (the item which is selected already also has a chance to select in new draw) and SRS without replacement.

SRS can be done by any of the following method

a) **Lottery method**

This method involves three steps. First step is constructing the sampling frame, i.e. list of units the target population, e.g. students' list, member list of FNCCI etc in alphabetical order and numbered accordingly. Second step is writing numbers listed in the sampling frame on small piece of papers and placing these papers in some vessel/drum/jar etc. Third step is mixing all papers well and tacking out one piece of paper from the jar. This process is continued until the required number of respondents is reached.

b) **Mechanical Randomization or Random Number Method**

A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern where each position is filled with one of these digits. A table of random numbers is so constructed that all numbers 0, 1, 2,.....9 appear independent of each other. Some random tables in common use are

a) Tippett's random number tables
b) Fisher and Yate's table
c) Kendall and Smith table
d) A million random number

Random number tables are the tables of digit 0,1,2,...9 each digit having equal chance of selection at any draw. In 1927, Tippett produced 41,600 digits (from 0 to 9) arranged in set of 4 in several columns and spread over 26 pages. This was followed by another publication by two great pioneering statistician, Sir RA Fisher and Frank Yates, which contained 15,000 digits formed by listing 15-19th digits in some 20 figure logarithm tables. Rand Corporation (1955) Published table containing 1 million digits. Kendall and smith (1938) published tables with 100,000.

Scanned by CamScanner

TABLE: A Random Number Table

.68587	.25848	.85227	.78724	.05302	.70712	.76552	.70326	.80402	.49479
.73253	.41629	.37913	.00236	.60196	.59048	.59946	.75657	.61849	.90181
.84448	.42477	.94829	.86678	.14030	.04072	.45580	.36833	.10783	.33199
.49564	.98590	.92880	.69970	.83898	.21077	.71374	.85967	.20857	.51433
.68304	.46922	.14218	.63014	.50116	.33569	.97793	.84637	.27681	.04354
.76992	.70179	.75568	.21792	.50646	.07744	.38064	.06107	.41481	.93919
.37604	.27772	.75615	.51157	.73821	.29928	.62603	.06259	.21552	.72977
.43898	.06592	.44474	.07517	.44831	.01337	.04538	.15198	.50345	.65288
.86039	.28645	.44931	.59203	.98254	.56697	.55897	.25109	.47585	.59524
.28877	.84966	.97319	.66633	.71350	.28403	.28265	.61379	.13886	.78325
.44973	.12332	.16649	.88908	.31019	.33358	.68401	.10177	.92873	.13065
.42529	.37593	.90208	.50331	.37531	.72208	.42884	.07435	.58647	.84972
.82004	.74696	.10136	.35971	.72014	.08345	.49366	.68501	.14135	.15718
.67090	.08493	.47151	.06464	.14425	.28381	.40455	.87302	.07135	.04507
.62825	.83809	.37425	.17693	.69327	.04144	.00924	.68246	.48573	.24647
.10720	.89919	.90448	.80838	.70997	.98438	.51651	.71379	.10830	.69984
.69854	.89270	.54348	.22658	.94233	.08889	.52655	.83351	.73627	.39018
.71460	.25022	.06988	.64146	.69407	.39125	.10090	.08415	.07094	.14244
.69040	.33461	.79399	.22664	.68810	.56303	.65947	.88951	.40180	.87943
.13452	.36642	.98785	.62929	.88509	.64690	.38981	.99092	.91137	.02411
.94232	.91117	.98610	.71605	.89560	.92921	.51481	.20016	.56769	.60462
.99269	.98876	.47254	.93637	.83954	.60990	.10353	.13206	.33480	.29440
.75323	.86974	.91355	.12780	.01906	.96412	.61320	.47629	.33890	.22099
.75003	.98538	.63622	.94890	.96744	.73870	.72527	.17745	.01151	.47200

Notations:

Let $Y_1, Y_2, \ldots Y_N$ be the items of a population having N units and let $y_1, y_2, \ldots y_n$ be the n samples drawn from the N population. We can represent sample by y_i (i = 1, 2,....., n). Here capital letter represents the population and small letter represents the sample. Notation for mean and total

Population total $Y = (Y_1 + Y_2 + \ldots + Y_N) = \sum_{i=1}^{N} Y_i$

Sample total $y = (y_1 + y_2 + \ldots + y_n) = \sum_{i=1}^{N} y_i$

Population Mean $\bar{Y} = \dfrac{(y_1 + y_2 + \ldots + y_N)}{N} = \dfrac{\sum_{i=1}^{N} Y_i}{N}$

Sample Mean $\bar{y} = \dfrac{(y_1 + y_2 + \ldots + y_n)}{n} = \sum_{i=1}^{N} y_i$

We will be confined to measure the population total = Y and population mean = $\bar{Y}$. These parameters are estimated by the following statistic. Population mean is estimated by sample mean $\bar{y}$ and population total $\hat{Y}$ is estimated by $N\bar{y}$ (i.e. product of population size and sample mean)

Scanned by CamScanner

Similarly we may take interest on population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})$; as a matter of rule results are presented in terms of a slightly different expression, in which the divisor (N-1) is used instead of N. We take $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$; whose estimator will be $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Properties of simple random sampling:

1. In SRS the sample mean is unbiased estimator of population mean i.e $E(\bar{y}) = \bar{Y}$.
2. In SRS the sample mean square is an unbiased estimator of population mean square i.e. $E(s^2) = S^2$
3. In SRSWOR (Simple random sampling without replacement), $V(\bar{y}) = \frac{N-n}{Nn} = \frac{S^2}{n}(1-f)$ where the notations have their usual meaning.

   Therefore $Var(\bar{y})$ is also given by

   $$V(\bar{y}) = \frac{N-n}{Nn}\frac{N}{N-1}\sigma^2 = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

   If N is too large then $\frac{N-n}{N-1}$ term will be approximately one then $Var(\bar{y}) = \frac{\sigma^2}{n}$. So $\frac{N-n}{N-1}$ is said to be finite population multiplier (FPM).
4. **In SRSWR (Simple random sampling with replacement)**

   $var(\bar{y}) = var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var(y_i)$ the covariance terms vanishes since in SRSWR all the draws are independent and consequently y_i's are independently and identically distributed (i.i.d.) with the same variance σ^2.

   $$\therefore\ V(\bar{y}) = \frac{\sigma^2}{n}$$
5. **In SRSWR the sample mean square is unbiased estimator of population variance $E(s^2) = \sigma^2$.**

Standard error:

SRSWOR: The standard error of sampling distribution of $\bar{y}$ denoted by $S.E.(\bar{y})$ is given by

$$S.E.(\bar{y}) = \sqrt{V(\bar{y})}$$
$$= \sqrt{\frac{(N-n)S^2}{Nn}}$$
$$= \frac{S}{\sqrt{n}}(1-f), \text{ where } f = \frac{N-n}{N} = 1 - \frac{n}{N}$$

Usually S^2 is not known in that case we replace S^2 by its unbiased estimate s^2 and we get

$$\text{Estimated } S.E.(\bar{y}) = \frac{S}{\sqrt{n}}(1-f)$$

SRSWR: The standard error of the sampling distribution of $\bar{y}$ denoted by $S.E.(\bar{y})$ is given by

$$S.E.(\bar{y}) = \sqrt{V(\bar{y})}$$
$$= \sqrt{\frac{(N-1)S^2}{Nn}}$$

Scanned by CamScanner

$$= \frac{S}{\sqrt{n}}\sqrt{\frac{N-1}{N}}$$

$$= \frac{S}{\sqrt{n}}\sqrt{1-\frac{1}{N}}$$

$$= \frac{S}{\sqrt{n}} \text{ for large N}$$

If S^2 is not know then Estimated S.E.$(\bar{y}) = \frac{S}{\sqrt{n}}$

Example 1. In a population the values of a characteristics Y_i are 6, 3, 2, 1, 11 and 5 . i = 1, 2, 3, 4, 5, 6. Random sample of size two are drawn without replacement. Verify that $E(\bar{y}) = \bar{Y}$ and $E(s^2) = S^2$. Also calculate the $Var(\bar{y})$. The notations have the usual meanings.

Solution

Here the sample of size 2 from population of size 6 can be drawn in C(6,2) ways = 15 ways

Hence the samples are;

(6,3) (6,2) (6,1) (6,11) (6,5) (3,2) (3,1) (3,11) (3,5) (2,1) (2,11) (2,5) (1,11) (1,5) (11,5)

Here $\bar{Y} = \frac{\sum Y}{N} = \frac{6+3+2+1+11+5}{6} = 4.66$

$$S^2 = \frac{1}{N-1}\sum(Y_i - \bar{Y})^2 = \frac{1}{N-1}\{\sum Y_i^2 - N\bar{Y}^2\} = \frac{1}{5}\{6^2+3^2+2^2+1^2+11^2+5^2-6(4.66)^2\} = 13.14$$

Sample No.	Sample values (y_i)	Sample mean ($\bar{y}_i$)	$s_i^2 = \frac{1}{n-1}\{\sum y_i^2 - n\bar{y}^2\}$
1	(6,3)	4.5	4.5
2	(6,2)	4	8
3	(6,1)	3.5	12.5
4	(6,11)	8.5	12.5
5	(6,5)	5.5	0.5
6	(3,2)	2.5	0.5
7	(3,1)	2	2
8	(3,11)	7	32
9	(3,5)	4	2
10	(2,1)	1.5	0.5
11	(2,11)	6.5	40.5
12	(2,5)	3.5	4.5
13	(1,11)	6	50
14	(1,5)	3	8
15	(11,5)	8	18
Total		71	196

$E(\bar{y}) = \frac{1}{C(N,n)}\sum_i y_i = \frac{71}{15} = 4.73$. Hence $E(\bar{y}) = \bar{Y}$

Scanned by CamScanner

$E(s^2) = \frac{1}{C(N,n)} \Sigma_i s_i^2 = \frac{196}{15} = 13.06$. Hence, $E(s^2) = S^2$

Now, $V(\bar{y}) = \frac{(1-f)S^2}{n} = \left(1 - \frac{2}{6}\right) \times \frac{13.14}{2} = 4.38$

Example 2. A population consists of four numbers 2, 5, 6, and 7.(i) Enumerate all possible samples of the size two which can be drawn from the population with replacement. (ii) Calculate the mean and variance of the population. (iii) Show that the mean of the sampling distribution of the sample means is equal to the population mean. (iv) Calculate the variance of the sampling distribution of sample mean. (v) Calculate the standard error of the sample mean.

Solution: Here, Population size (N) = 4, Sample size (n) = 2.

(i) Number of possible samples of size 2 that can be drawn from the population of size N = 4 by using with replacement technique is

Number of possible samples = $N^n = 4^2 = 16$

Thus the possible samples are: (2,2), (2,5), (2,6), (2,7), (5,2), (5,5), (5,6), (5,7), (6,2), (6,5), (6,6), (6,7), (7,2), (7,5), (7,6), (7,7).

(ii) Calculation of population mean and population variance:

Y_i	$Y_i - \bar{Y}$	$(Y - \bar{Y})^2$
2	–3	9
5	0	0
6	1	1
7	2	4
$\Sigma Y = 20$		$\Sigma(Y - \bar{Y})^2 = 14$

Population mean (μ or $\bar{Y}$) $= \frac{\Sigma Y_i}{N} = \frac{20}{4} = 5$

Population variance (σ^2) $= \frac{\Sigma(Y_i - \bar{Y})^2}{N} = \frac{14}{4} = 3.5$

(iii) Calculation of sample means and variance of the sampling distribution of means:

S.No.	Sample Values (Y_i)	Sample Means ($\bar{y}_i$)	$(\bar{y}_i - \bar{y})$	$(\bar{y}_i - \bar{y})^2$
1	(2, 2)	2	–3	9
2	(2, 5)	3.5	–1.5	2.25
3	(2, 6)	4	–1	1
4	(2, 7)	4.5	–0.5	0.25
5	(5, 2)	3.5	–1.5	2.25
6	(5, 5)	5	0	0
7	(5, 6)	5.5	0.5	0.25
8	(5, 7)	6	1	1
9	(6, 2)	4	–1	1
10	(6, 5)	5.5	0.5	0.25
11	(6,6)	6	1	1
12	(6, 7)	6.5	1.5	2.25
13	(7,2)	4.5	–0.5	0.25
14	(7,5)	6	1	1
15	(7,6)	6.5	1.5	2.25
16	(7,7)	7	2	4
		80		28

Scanned by CamScanner

Mean of the sample means $(\bar{y}) = \frac{\Sigma \bar{y}_i}{\text{No. of samples } (N^n)} = \frac{80}{16} = 5$

Since the mean of the sample means $(\bar{y}) = 5$ is equal to the population mean $(\bar{Y}) = 5$, so we can conclude that the mean of the sampling distribution of the sample means is equal to the population mean.

(iv) Variance of the sample means is

$$V(\bar{y}) = \frac{\Sigma(\bar{y}_i - \bar{y})^2}{\text{No. of samples} = N^n} = \frac{28}{16} = 1.75 \text{ or } V(\bar{y}) = \frac{\sigma^2}{n} = \frac{3.5}{2} = 1.75$$

(v) The standard error of sample mean is given by

$$S.E.(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{1.75} = 1.3229$$

Merits and Demerits of Simple Random Sampling

Merits

1. Since the sample unit are selected at random giving each unit an equal chance of being selected, the element of subjectivity or personal bias is completely eliminated. As such a simple random sample is more representative of population as compared to the judgment or purposive sampling.
2. The statistician can ascertain the efficiency of the estimate of the parameters by considering the sampling distribution of the statistics (estimate) e.g.,$\bar{Y}n$ as an estimate of $\bar{Y}N$ becomes more efficient as sample size n increases.

Demerits:

1. The selection of SRS requires an up to date frame, i.e. a completely catalogued population from which samples are to be drawn. Frequently it is virtually impossible to identify the units in the population before the sample is drawn and this restricts the use of simple random sampling technique.
2. Administrative inconvenient: A simple random sample may result in the selection of the sampling units which are widely spread geographically and in such case the cost of collecting the data may be much in terms of time and money.
3. At times, a simple random sample might give most non-random looking results. For example if we draw a random sample of size 13 from a pack of card, we may get all the cards of same suit. However the probability of such an outcome is extremely small.
4. For a given precision, simple random sampling usually requires larger sample size as compared to stratified random sampling.

(b) **StratifiedRandom Sampling:** When units in the population are not similar i.e. population under study is heterogeneous in nature, the population is divided into sub groups called strata before the sample is drawn. Then a simple random sample is drawn from each stratum in proportion to its size. The stratification of the population should be done as follows:

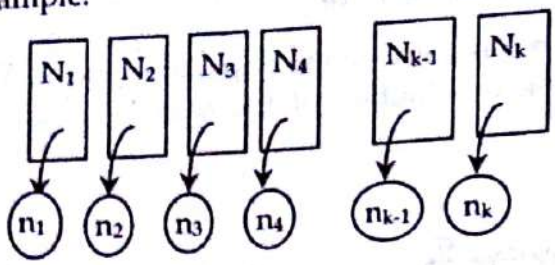
- The strata should be non-overlapping and should together comprise the whole population.
- The strata should be as possible as homogeneous within groups and heterogeneous between the groups.

Consider a heterogeneous population of size N and it is divided into k sub-population called strata.

Let $N_1, N_2, \ldots\ldots, N_k$ be size of $1^{st}, 2^{nd}, \ldots\ldots k^{th}$ stratum respectively such that $N_1 + N_2 + \ldots\ldots + N_k = N$. Then random samples of size $n_1, n_2, \ldots\ldots, n_k$ are selected randomly and independently from

Scanned by CamScanner

1st, 2nd, ... kth stratum respectively such that $n_1 + n_2 + \ldots\ldots + n_k = n$ where n is the size of required sample.



Purpose of stratification

i) To make more representative

In un-stratified random sample some group may be over representative other may be under representative even some may be excluded too. Stratified sampling ensures any desired representation in the sample of the various group in the population. It overrules the possibility of any essential group of population being completely excluded in the sample. Stratified sampling thus provide a more representative cross section of the population and is frequently regarded as the more efficient system of sampling.

ii) Greater accuracy

Stratified sampling provide estimate with increased accuracy at given cost. Moreover stratified sampling enable us to obtain the results of known precision of each of the group.

iii) Administrative convenience

As compare to SRS, the stratified samples would be more concentrated geographically Accordingly, the time and money involved in collecting the data and interviewing th individuals may be considerably reduce the supervision of the field work could be allotted wit greater ease and convenience.

Notation and Terminology

$N = \sum_{i=1}^{K} N_i$ Total number of sampling unit in the population

n_i = The number of sampling units selected with SRSWOR from the ith stratum.

$N = \sum_{i=1}^{K} n_i$ Total number of sample from all strata

Let Y_{ij}(i=1,2,3.....k and j = 1,2,3,.....N_i) be the value of the jth unit in the ith strata

y_{11}	y_{21}	y_{31}		y_{k1}
y_{12}	y_{22}	y_{32}		y_{k2}
y_{13}	y_{23}	y_{33}		y_{k3}
y_{1N1}	y_{2N2}	y_{3N1}		y_{kNk}

From each strata samples of size $n_1, n_2, n_3, \ldots\ldots n_k$ are drawn in SRS technique. Then the total samp size will be $n = \sum_{i=1}^{k} n_i$ from the total population $N = \sum_{i=1}^{k} N_i$.

$W_i = \frac{N_i}{N}$ weight of the ith Stratum

$C_i = \frac{N_i}{n_i}$ Sampling fraction in the stratum

Scanned by CamScanner

$\bar{Y}_{Ni}$ = Population mean of the ith stratum = $\frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$; i = 1, 2, ..., k

$\bar{Y}_N$ = Population mean = $\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} Y_{ij} = \frac{1}{N}\sum_{i=1}^{k} N_i\,\bar{Y}_{N\beta} = \sum_{i=1}^{k} W_i\,\bar{Y}_{Ni}$

S_i^2 = Population mean square of the ith stratum = $\frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_{Ni})^2$

Let y_{ij} be the value of jth sampled unit from i^{th} strata 9 I = 1, 2, 3.....k and j = 1,2,3.....n_i)

$\bar{y}_{ni}$ = Sample mean of the ith stratum = $\frac{1}{n_i - 1}\sum_{j=1}^{n_i} y_{ij}$, i = 1,2,.....k

$\bar{y}_n$ = Sample mean = $\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\,\bar{y}_{n\beta} = \sum_{i=1}^{k}\frac{n_i}{n}\,\bar{y}_{n\beta}$

s_i^2 = Sample mean square of the ith stratum = $\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{ni})^2$

An estimator $\bar{y}_{st}$ for the population mean $\bar{Y}_N$ is given by $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i\,\bar{y}_{ni} = \sum_{i=1}^{k} W_i\,\bar{y}_{ni}$

Note:

1. $\bar{y}_{st}$ is an unbiased estimator of population mean $\bar{Y}_N$.
2. $var(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i\,(N_i - n_i)\,\frac{s_i^2}{n_i}$

Allocation of sample size

As we have $var(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i\,(N_i - n_i)\,\frac{s_i^2}{n_i}$

By above expression we can say that $var(\bar{y}_{st})$ depends on n_i (the sample size for the ith stratum) which can be fixed at our will. The allocation of the sample size for different strata can be done in the following two different ways

a) Proportional allocation
b) Optimum allocation

Proportional Allocation:

Allocation of n_i's to various strata is called proportional if the sample fraction is constant for each stratum i.e.,

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \ldots\ldots\ldots\ldots = \frac{n_k}{N_k} = \frac{n_i}{N_i} = \frac{n}{N} = C$$

Note:

$$Var\,(\bar{y}_{st})\ prop = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} w_i\,S_i^2$$

Optimum allocation:

Another method of determination of the n_i's is to choose them so as to (a) minimize the variance of the estimate for (i) fixed sample size n and ii) fixed total cost c (b) Minimize the total cost for fixed desire precision. The allocation of n_i's to various strata in accordance with the above principle is known as optimum allocation

Scanned by CamScanner

Cost function

In stratified sampling it may be cost more to obtain information about a sample in one stratum than in other. For example interviewing people in rural area is going to be more costly because of travel expenses than interviewing people in urban areas. Thus in it's simplest form the cost c in stratified sampling may be given by the linear model $c = a + \sum_{i=1}^{k} c_i n_i$

Where, c = total cost, a = overhead cost, c_i = cost per unit in the ith strata

Note:

1. $Var(\bar{y}_{st})$ is minimum for fixed total size of sample n if $n_i = \frac{n N_i S_i}{\sum N_i S_i}$ $(n_i \alpha N_i S_i)$

2. $Var(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k} w_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_i S_i^2$

Example 3. A population of size 800 is divided into 3 strata .their sizes and standard deviations are given below.

Strata no	I	II	III
Size	200	300	300
S.D.	6	8	12

A stratified random sample of size 120 is to be selected from the population. Determine the sizes of samples from the three strata in case of (i) Proportional allocation (ii) Neyman's optimum allocation. Explain why the stratum sample sizes are different in two cases.

Solution:

Here, N = 800, N_1= 200, N_2 = 300 , N_3 = 300, S_1= 6, S_2= 8, S_3 = 12, n = 120

$N_1S_1 = 200 \times 6 = 1200$, $N_2S_2 = 300 \times 8 = 2400$, $N_3S_3 = 300 \times 12 = 3600$, $\sum N_i S_i = 1200+2400+3600 = 7200$

i) Under proportional allocation

$$\frac{n_i}{N_i} = \frac{n}{N}$$

$$n_i = \frac{n}{N} N_i, \; i = 1,2,3$$

$$n_1 = \frac{120}{800} \times 200 = 30$$

$$n_2 = \frac{120}{800} \times 300 = 45$$

$$n_3 = \frac{120}{800} \times 300 = 45$$

ii) Under Neyman's optimum allocation

$$\frac{n_i}{n} = \frac{N_i S_i}{\sum N_i S_i}$$

$$N_i = n \frac{N_i S_i}{\sum N_i S_i}$$

$$n_1 = \frac{120 \times 1200}{7200} = 20$$

Scanned by CamScanner

$$n_2 = \frac{120 \times 2400}{7200} = 40$$

$$n_3 = \frac{120 \times 3600}{7200} = 60$$

In the proportional allocation the sample fraction n_i/N_i is constant for each stratum

$$\frac{n_i}{N_i} = \frac{n}{N} = C, \quad n_i \text{ proportional to } N_i$$

In Neyman's optimum allocation $V(\bar{y}_{st})$ is minimum for fixed n if n_i proportional to N_iS_i. Hence in former one n_i depends on N_i but in later one n_i also depends on S_i

(c) **Systematic Random Sampling:** A sampling technique in which only first unit is selected with the help of random numbers and the rest get selected automatically according to some pre-designed pattern is known as systematic random sampling. This technique of selecting samples is usually recommended if the complete and up-to-date list of the sampling units is available and the units are arranged in some systematic order such as alphabetical, chronological, geographical order etc. Suppose N units of the population are numbered from 1 to Nin order and we want to draw a sample of size n such that $N = nk \Rightarrow k = \frac{N}{n}$ where k is called sample interval. Systematic sampling consists in selecting any unit at random from the first kunits numbered from 1 to k and then selecting every k^{th} unit in succession subsequently. For example suppose that we want to select 10 families from a list of families containing 80 families arranged systematically. Here, n = 10 and N = 80 then $k = \frac{N}{n} = \frac{80}{10} = 8$. We select any number from 1 to 8 at random, suppose selected number is 2. Then systematic sample of will consists of 10 families in the list as: 2, 10, 18, 26, 34, 42, 50, 58, 66, and 74. Here rest of the sample is obtained on adding k = 8 in succession for subsequent samples.

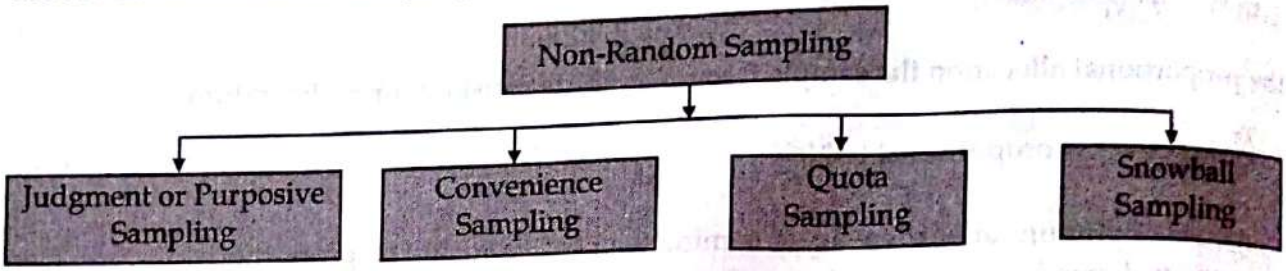
(d) **Cluster random sampling:** In *cluster random sampling*, we divide the population into discrete groups prior to sampling. The groups are termed as clusters and then we select some clusters as sample from all clusters using simple random sampling. After selecting the clusters in the sample, all the elements within the clusters are enumerated for the study. Clusters differ from strata in a manner that elements within a cluster are relatively heterogeneous and between clusters are relatively homogeneous. Thus, when dividing the population into different clusters, it should be kept in mind that there is greater variation within the clusters and less variation between the clusters.

This method of sampling is useful when the population consists of vary large number of similar groups which are geographically distant.

(e) **Multi-stage random sampling:** Multi-stage random sampling, sometimes called multi-stage cluster sampling, is a development of cluster sampling. In cluster sampling, clusters are considered as sampling units and all the elements in the selected clusters are enumerated completely. Instead of selecting all units of the selected clusters, a sample of desirable size can be drawn from selected clusters which may increase the efficiency is called multi-stage random sampling. In this technique, clusters which form the units of sampling at the first stage are called the *first stage units (fsu) or primary sampling units (psu)* and the elements within clusters are called *second-stage units (ssu)*. For example, in crop surveys for estimating yield of a crop in a district, a block may be considered a primary sampling unit, the villages the second stage units, the crop fields the third stage units, and a plot of fixed size the ultimate unit of sampling.

Scanned by CamScanner

(2) **Non-random or Non-probability Sampling Technique:** It is defined as the method of sampling technique in which each unit in a sample is selected on the basis of personal judgment. There are several non-random sampling methods for selecting samples from a population:



(a) **Judgment Sampling:** A sampling process in which the sample units are selected according to researchers or investigators personal judgment is called the *purposive or judgment sampling*. In other words a sampling process the investigators include only those items or units in the sample from the population which he thinks most convenient or appropriate to study. Only the average units are considered and extreme units are omitted in choosing the purposive samples. The objective of the inquiry has great importance in the selection of the samples, so selection of units has to be adjusted according to the objective of the study so that no significant units may be ignored. This method is suitable:

- Only a small number of sampling units is in the population or universe.
- In solving everyday business problems and making public policy decisions or to the urgent problems.

(b) **Convenience sampling:** A sampling process in which the units are selected neither by probability nor by personal judgment but by convenience is called the convenience sampling. A sample obtained from easily available lists such as telephone directories, automobile registrations, etc. is an example of convenient samples and it is not a random sample even if the sample is drawn randomly from lists. The results obtained from convenient samples can hardly be representative of the population because they are not free from biasness or unsatisfactory. Moreover, convenience sampling is often used for making pilot studies.

(c) **Quota sampling:** Quota sampling may be considered as a special of stratified sampling in which any investigator is told to examine or enumerate a fixed number of units called quotas from each stratum. The sampling quotas may be fixed according to some specified characteristics such as income, sex, occupation, religion etc. The selection of units or individuals for investigations is left to the investigator's choice rather than being decided by probability or random methods. Quota sampling is very popular in market survey and public opinion polls because it is cheaper than random sampling. It occasionally provides satisfactory results if the interviewers are carefully trained.

(d) **Snowball Sampling**

Snowball sampling technique is a non-random sampling technique generally used in the case where it is difficult to reach into exact population or population is hidden. These population may be hard to reach because they exhibit some kind of social stigma, illicit or illegal behavior or sometimes population is rare and prone to hidden. In this method survey subjects are selected on referral from other survey respondents. A respondent is identified according to the objective of study and other respondents are identified according to the referral from the respondent. For example, populations would be drug users or sex workers. As sample members are not selected from a sampling frame.

Scanned by CamScanner

Exercise – 4

Theoretical Questions

1. What do you mean by random sampling?
2. Distinguish between parameters and statistic.
3. What are the advantages of using sampling method than census?
4. What are some random sampling methods?
5. Differentiate between sampling and non-sampling error.
6. Define standard error.
7. How non sampling errors can be minimized, explain.
8. Describe the steps of sampling.
9. Give the fundamental characteristics of sample.
10. Discuss the advantage and limitation of sampling.
11. What are the advantages and disadvantages of using simple Random sampling?
12. Define biased in sampling.
13. Define the following:
    Stratified sampling, Cluster sampling, judgment sampling, Quota sampling, stratified sampling,
14. Differentiate between cluster sampling and stratified sampling.
15. When a non-random sampling is the best choice for some research?
16. "Parameter is always unknown" comment.

Practical Questions

1. The five numbers shown here are the first digits taken from the first row of the RANDS random number table which are: {1, 0, 9, 3, 8}. Suppose, they represents the number of computers (PC or laptops) in five randomly selected urban household. Assume these number represents a small population from which we are generating the sample distribution. Perform the following tasks:
   a) Calculate mean and variance of the given population of computers
   b) Construct all possible samples from this population of size 2 using sampling without replacement (WOR)
   c) Calculate from each possible samples, sample mean ($\bar{x}$) and sample variance (s^2)
   d) Calculate variance of sample means, Var($\bar{x}$).
   e) Test if sample mean is unbiased estimate of the population mean?
   f) Verify if the direct formula to calculate the variance of sample means viz. $Var(\bar{x})_{WoR} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$ gives the same value as calculated from (d) above.
2. The five numbers shown here are the first digits taken from the first row of the RAND's random number table which are: {1, 5, 7, 3, 0}. Suppose they represent the number of mobiles in five randomly selected rural household. Assume these number represents a small population from which we are generating the sample distribution. Perform the following tasks:
   a) Calculate the mean and variance of the given population of mobiles
   b) Construct all possible samples from this population of size 2 using sampling with replacement (WR)
   c) Calculate from each possible samples, sample mean ($\bar{x}$) & sample variance (s^2)

Scanned by CamScanner

d) Calculate variance of sample means, $Var(\bar{x})$

e) Test if sample mean is unbiased estimate of the population mean?

f) Verify if the direct formula to calculate the variance of sample means viz. $Var(\bar{x})_{WR} = \frac{\sigma^2}{n}$ gives the same value as calculated from (d) above.

3. A population of size 800 is divided into 3 strata. Their sizes and standard deviations are given below.

Strata	I	II	III
Size	200	300	300
S.D.	6	8	12

A stratified random sample of size 120 is to be drawn from the population. Calculate the variance of the estimate of the population mean

i) Under proportional allocation

ii) Under Neyman's optimum allocation.

❑❑❑

Scanned by CamScanner

Chapter 5

RANDOM VARIABLE AND MATHEMATICAL EXPECTATION

Random Variable

It is a rule which assigns one and only one real value to each outcome of a random experiment. It is also called a real valued function defined on a sample space of the random experiment. Random variables are denoted by capital letters X, Y, Z etc. and value taken by the random variables are denoted by small letters x, y, z and so on.

For example,

In tossing a coin, the number of heads is a rule which assigns 1 to the outcome head and 0 to the outcome tail. Hence the number of head is a random variable X taking value 1 and 0 with probabilities $\frac{1}{2}$ and $\frac{1}{2}$ respectively.

Types of Random Variable

Random variable is divided into two types:

(i) Discrete random variable

(ii) Continuous random variable

Discrete Random Variable

A random variable is called discrete if it takes integer values. It is also called real valued function defined on discrete sample space. e.g. number of printing mistakes in a page of a book, number of students enrolled in a college, number of defective items in a sample of certain size, number of patients admitted in a hospital , number of files in folders etc.

Continuous Random Variable

A random variable is called continuous if it takes all possible values within a certain interval. e.g. amount of rainfall in rainy season, height of an individual, temperature recorded in a particular day, etc.

Scanned by CamScanner

Properties of Random Variable

i. Scalar multiple of a random variable is also a random variable. If X is a random variable the aX is also a random variable

ii. Sum or difference of random variables is also a random variable. If X and Y are random variables then X+Y , X-Y , aX+bY , aX – bY are also random variables

iii. Multiple or quotient of random variables is also a random variable. If X and Y are random variables then XY, X/Y are also random variables.

iv. Different power of random variable is also a random variable. If x is a random variable then X^2, X^{-1} are also random variables.

Probability Distribution

It is a formula that specifies the probabilities associated with the random variables .A probability distribution must satisfy two conditions

(i) Probability associated with each random variable should be greater than or equal to zero

(ii) Total of all the probability values associated with random variables should be 1.

Probability Mass Function (Probability Distribution of Discrete Random Variable)

Let X be a discrete random variable taking values x_1 , x_2 , x_3, ...x_n with probabilities $p_i = P(X = x_i)$, i = 1,2,3 .....n. Then $p_i = P(X = x_i)$ is called probability mass function(pmf) if it satisfies the conditions:

i. $p_i = P(X = x_i) \geq 0$

ii. $\sum_{i=1}^{n} P_i = 1$

Example 1. A random variable x has the following probability distribution

x	0	1	2	3	4	5	6
P(x)	a	3a	2a	5a	7a	4a	8a

(i) Determine a (ii) Find P(x > 3), P(x ≤ 4) and P(1 < x < 5) (iii) Find k if P(x ≤ k) > 0.6.

Solution

x	P(x)	
0	a	1/30
1	3a	3/30
2	2a	2/30
3	5a	5/30
4	7a	7/30
5	4a	4/30
6	8a	8/30
	ΣP(x) = 30a	

(i) We know, ΣP(x) = 1

or 30a = 1

or $a = \frac{1}{30}$

(ii) P(x >3) =P(x = 4) + P(x=5) + P(x=6)

$= 7a + 4a + 8a = 19a = \frac{19}{30} = 0.63$

Scanned by CamScanner

$P(x \le 4) = 1 - p(X > 4) = 1 - \{P(x=5) + P(x=6)\}$

$= 1 - \{4a + 8a\} = 1 - 12a = 1 - \frac{12}{30} = \frac{18}{30} = 0.6$

(iii) $P(x \le 4) = P(x=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4)$

$= a + 3a + 2a + 5a + 7a = 18a = \frac{18}{30} = 0.6$

(iv) $P(1<x<5) = P(x = 2) + P(x = 3) + P(x = 4)$

$= 2a + 5a + 7a = 14a = \frac{14}{30} = 0.46$

(v) $P(x \le k) = 0.6$

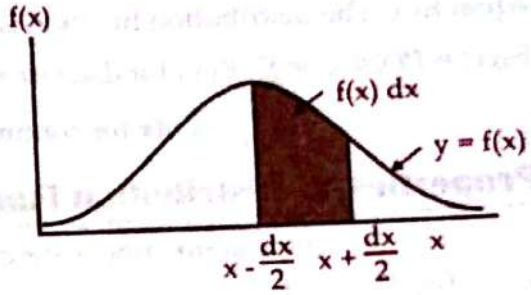
By trial method $k = 4$.

Probability Density Function

Let x be a continuous random variable and f(x) is continuous function of x, the probability that x lies in interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$ of width dx is given by

$$f(x)dx = P[x - \frac{dx}{2} \le x \le x + \frac{dx}{2}]$$

f(x) dx represents the probability that x lies in the infinitesimal small interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$.



The probability that x lies in interval [a, b] is given by $P[a \le x \le b] = \int_a^b f(x)dx$. Then the function f(x) is called probability density function (pdf) if it satisfies the conditions

i. $f(x) \ge 0$, for every x in the given range

ii. $\int_{-\infty}^{\infty} f(x)\,dx = 1, -\infty \le x \le \infty$

Example 2. A continuous random variable x has a pdf $f(x) = k\,x^3$, $0<x<1$. Find (i)k (ii) a such that $P(x \le a) = P(x > a)$(iii) b such that $P(x>b) = 0.5$

Solution

We know

(i) $\int_0^1 f(x)dx = 1$

or $\int_0^1 kx^3dx = 1$

or $k\int_0^1 x^3dx = 1$

or $K\left[\frac{X^4}{4}\right]_0^1 = 1$

or $k = 4$

(ii) $P(x \le a) = P(x > a)$

or $\int_0^a f(x)dx = \int_a^1 f(x)\,dx$

or $\int_0^a 4x^3dx = \int_a^1 4x^3dx$

or $[x^4]_0^a = [x^4]_a^1$

or $a^4 = 1 - a^4$

Scanned by CamScanner

or $a^4 + a^4 = 1$

or $2a^4 = 1$

or $a^4 = 0.5$

or $a = (0.5)^{1/4} = 0.84$

(iii) $P(x > b) = 0.5$

or $\int_b^1 4x^3 dx = 0.5$

or $[x^4]_b^1 = 0.5$

or $1 - b^4 = 0.5$

or $1 - 0.5 = b^4$

or $b^4 = 0.5$

or $b = (0.5)^{1/4}$

or $b = 0.84$

Distribution Function

Let x be a random variable then the cumulative distribution function or simply distribution function of the random variable is the probability that random variable takes values less than or equal to x. The distribution function is denoted by $F_X(x)$ and is given by

$F_X(x) = P(X \leq x) = \sum_x P(x)$ for discrete random variable having pmf P(x)

$= \int_{-\infty}^{x} f(x)dx$ for continuous random variable having pdf f(x)

Properties of Distribution Function

i. The distribution function is defined for every real number x.

ii. $0 \leq F(x) \leq 1$

iii. $F(-\infty) = 0$ and $F(\infty) = 1$

iv. $P(a \leq x \leq b) = F(b) - F(a)$

v. $F'_X(x) = f(x)$

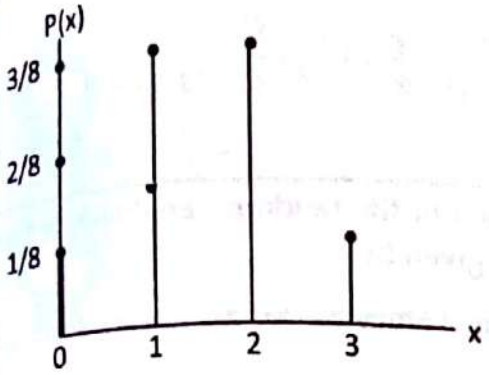
Example 3. A coin is tossed 3 times then (i) find probability distribution (ii) draw graph of probability distribution (iii) Find distribution function(iv) draw graph of distribution function.

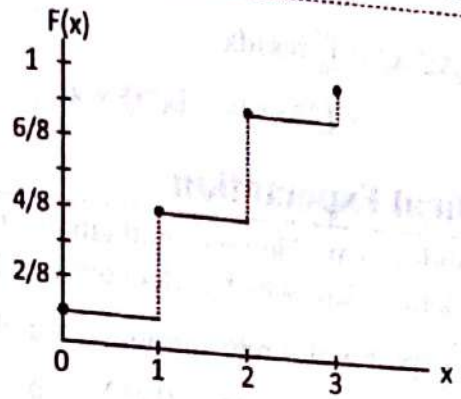
Solution

Let H denotes outcome head and T denotes outcome tail on tossing a coin. On tossing a coin 3 times the sample space are {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT} Let X = number of heads on tossing coin X = 0, 1, 2, 3.

No of heads (X)	P(X=x)	F(x)
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	1

Scanned by CamScanner





Example 4. The number of mobile phone calls received by a person per hour is found to be a random variable with probability mass function $p(x) = \begin{cases} k(5-x) & \text{for } x = 0, 1, 2, 3, 4, 5 \\ 0; & \text{otherwise} \end{cases}$. Find (i) constant k (ii) probability distribution of X (iii) Distribution function of X

Solution

Since, $\sum_{x=0}^{n} p(x) = 1$

or $\sum_{x=0}^{5} k(5-x) = 1$

or $k\{5+4+3+2+1+0\} = 1$

or $k = 1/15$

Hence, $p(x) = \frac{1}{15}(5-x)$

Now probability distribution P(x) and distribution function F(x)

x	P(x)	F(x)
0	5/15	5/15
1	4/15	9/15
2	3/15	12/15
3	2/15	14/15
4	1/15	1
5	0	1

Example 5. A random variable x has pdf $f(x) = \begin{cases} 5x^4; & 0 < x < 1 \\ 0; & \text{otherwise} \end{cases}$. Find (i) $P(0 < X < 0.5)$ (ii) $P(0.5 < X < 0.75)$ (iii) F(x).

Solution

(i) $P(0 < X < 0.5) = \int_0^{0.5} f(x)dx$

$= \int_0^{0.5} 5x^4 dx$

$= [x^5]_0^{0.5} = 0.03125$

(ii) $P(0.5 < X < 0.75) = \int_{0.5}^{0.75} f(x)dx$

$= \int_{0.5}^{0.65} 5x^4 dx$

$= [x^5]_{0.5}^{0.75} = 0.2373 - 0.03125 = 0.206$

Scanned by CamScanner

(iii) $F(x) = P(X \leq x) = \int_0^x f(x)dx$

$= \int_0^x 5x^4 dx = [x^5]_0^x = x^5$

Mathematical Expectation

Let X be a random variable then mathematical expectation of the random variable X is the average value (Arithmetic Mean) which is denoted by E(X) and is given by

$E(X) = \sum_{i=1}^{n} x_i p_i$, for discrete random variable X taking values $x_1, x_2, x_3, \ldots\ldots x_n$ with respective probabilities $p_1, p_2, p_3, \ldots, p_n$ such that $\sum_{i=1}^{n} p_i = 1$.

$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$, for continuous random variable x, $-\infty < x < \infty$ having pdf f(x) such that $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example 6. On tossing two coins find the expected number of heads.

Solution

On tossing coin let H be the head and T be the tail. When two coins are tossed the outcomes are HH, HT, TH, TT with probabilities $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$ respectively. Let X be the number heads then

No of head (x)	P(x)	xP(x)
0	1/4	0
1	2/4	2/4
2	1/4	2/4
		ΣxP(x) = 1

Expected number of heads $E(x) = \Sigma xP(x) = 1$.

Example 7. A coin is tossed until a head appears, find the expected number of tosses required?

Solution

On tossing a coin let H denotes head and T denotes tail. Let X be the number of toss required to get first head then head can occur first in following ways; H, TH, TTH, TTTH .....

Outcome	No of toss(x)	P(x)	xP(x)
H	1	1/2	1/2
TH	2	1/4	2/4
TTH	3	1/8	3/8
TTTH	4	1/16	4/16
:	:	:	:

Now,

$$E(X) = \Sigma xP(x) = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \cdots\cdots\cdots \quad \text{(i)}$$

Let $$S = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \cdots\cdots\cdots \quad \text{(ii)}$$

Dividing equation (ii) by 2

$$\frac{S}{2} = \frac{1}{4} + \frac{2}{8} + \frac{3}{16} + \quad \ldots\ldots\text{(iii)}$$

Scanned by CamScanner

From (ii) and (iii)

$$S - \frac{S}{2} = \frac{1}{2} + \left(\frac{2}{4} - \frac{1}{4}\right) + \left(\frac{3}{8} - \frac{2}{8}\right) + \left(\frac{4}{16} - \frac{3}{16}\right)$$

Or $\frac{S}{2} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots\ldots$

Or $\frac{S}{2} = \frac{1/2}{1-(1/2)}$

Or $\frac{S}{2} = \frac{1/2}{1/2}$

Or $S = 2$

Substitute S then E(X) = 2

Hence the expected number of toss required until a head appears is 2.

Example 8. Three lab contain 4 Dell and 3 Acer, 4 Dell and 5 Acer,3 Dell and 4 Acer computers respectively. One computer is selected from each lab. Find the expected number of Dell computer selected.

Solution

Let D be the Dell, A be the Acer and I, II and III indicates three labs

Lab I: 4D + 3A

Lab II: 4D + 5A

Lab III: 3D+4A

Let X be the number of Dell computers

One computer from each lab can be selected as

All acer: $A_I A_{II} A_{III}$

One Dell: $D_I A_{II} A_{III}$, $A_I D_{II} A_{III}$, $A_I A_{II} D_{III}$

Two Dell: $D_I D_{II} A_{III}$, $D_I A_{II} D_{III}$, $A_I D_{II} D_{III}$

All Dell: $D_I D_{II} D_{III}$

With probabilities:

All Acer: $\frac{3}{7} \times \frac{5}{9} \times \frac{4}{7}$

One Dell: $\frac{4}{7} \times \frac{5}{9} \times \frac{4}{7}, \frac{3}{7} \times \frac{4}{9} \times \frac{4}{7}, \frac{3}{7} \times \frac{5}{9} \times \frac{3}{7}$

Two Dell: $\frac{4}{7} \times \frac{4}{9} \times \frac{4}{7}, \frac{3}{7} \times \frac{4}{9} \times \frac{3}{7}, \frac{4}{7} \times \frac{5}{9} \times \frac{3}{7}$

All Dell: $\frac{4}{7} \times \frac{4}{9} \times \frac{3}{7}$ respectively.

Number of Dell Computer(x)	P(x)	xP(x)
0	$\frac{60}{441}$	0
1	$\frac{80}{441} + \frac{48}{441} + \frac{45}{441} = \frac{173}{441}$	$\frac{173}{441}$
2	$\frac{64}{441} + \frac{36}{441} + \frac{60}{441} = \frac{160}{441}$	$\frac{320}{441}$
3	$\frac{48}{441}$	$\frac{144}{441}$
		$\Sigma xP(x) = \frac{637}{441}$

Scanned by CamScanner

Hence the expected number of Dell computers $E(x) = \Sigma x P(x)$

$$= \frac{637}{441} = 1.44$$

Example 9: A lot contains 9 computers in which 5 are laptop and 4 are desktop. A sample of computers is taken for inspection, then find the expected number of laptop.

Solution

Let x = number of laptop

$$P(x) = \frac{{}^5C_x \times {}^4C_{4-x}}{{}^9C_4}$$

$$E(x) = \sum_{x=0}^{4} X\,P(X) = 0 \times \frac{{}^5C_0 \times {}^4C_{4-0}}{{}^9C_4} + 1 \times \frac{{}^5C_1 \times {}^4C_{4-1}}{{}^9C_4} + 2 \times \frac{{}^5C_2 \times {}^4C_{4-2}}{{}^9C_4} + 3 \times \frac{{}^5C_3 \times {}^4C_{4-3}}{{}^9C_4} + 4 \times \frac{{}^5C_4 \times {}^4C_{4-4}}{{}^9C_4}$$

$$= 0 + \frac{20}{126} + \frac{120}{126} + \frac{120}{126} + \frac{20}{126} = \frac{280}{126} = 2.22$$

Example 10. Time taken to boot a computer is a continuous random variable x having the

$$\text{pdf } f(x) = \begin{cases} kx(1-x), & 0 < x < 1 \\ 0, & \text{Otherwise} \end{cases}.$$

Find:(i) Expected value of x and (ii)$P(0.25 < x < 0.5)$.

Solution

We know

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Or $\int_0^1 kx(1-x)dx = 1$

Or $k\{\int_0^1 xdx - \int_0^1 x^2dx\} = 1$

Or $K\{[\frac{x^2}{2}]_0^1 - [\frac{x^3}{3}]_0^1\} = 1$

Or $k\left[\frac{1}{2} - \frac{1}{3}\right] = 1$

Or $k = 6$

Substitute k then $f(x) = 6x(1-x)$

Now,

$$E(x) = \int_0^1 xf(x)dx$$

$$= \int_0^1 x \times 6x(1-x)dx$$

$$= 6\int_0^1 (x^2 - x^3)dx$$

$$= 6\left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1$$

$$= 6\left[\frac{1}{3} - \frac{1}{4}\right] = \frac{6}{12} = 0.5$$

$$P(0.25 < x < 0.5) = \int_{0.25}^{0.5} f(x)dx = \int_{0.25}^{0.5} 6x(1-x)dx$$

$$= 6\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_{0.25}^{0.5}$$

$$= 6\left[\frac{1}{2}\{(0.5)^2 - (0.25)^2\} - \frac{1}{3}\{(0.5)^3 - (0.25)^3\}\right]$$

$$= 6(0.093 - 0.036) = 0.342$$

Scanned by CamScanner

Properties of Expectation

If X is a random variablea and b are constants then

i. $E(a) = a$
ii. $E(ax) = aE(x)$
iii. $E(ax+b) = aE(x) + b$
iv. $E(ax^2 + bx) = aE(x^2) + bE(x)$

Example 11. A continuous random variable x has the pdf $f(x) = ke^{-x}$, $0<x<\infty$. Find $E(x)$, $E(x^2)$ and $E(x^2+2x+5)$.

Solution

We know

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

or $\int_0^{\infty} ke^{-x}\,dx = 1$

or $k\int_0^{\infty} ke^{-x}\,dx = 1$

or $k(1-1)! = 1$ since $\int_0^{\infty} e^{-x}x^{n-1}\,dx = \Gamma n = (n-1)!$

or $k = 1$

Substitute k in f(x) then $f(x) = e^{-x}$

Now,

$$E(x) = \int_0^{\infty} x\,f(x)\,dx$$

$$= \int_0^{\infty} x\,e^{-x}\,dx = \int_0^{\infty} e^{-x}x^{2-1}\,dx = (2-1)! = 1$$

Again, $E(x^2) = \int_0^{\infty} x^2 f(x)\,dx$

$$= \int_0^{\infty} x^2 e^{-x}\,dx = \int_0^{\infty} e^{-x}x^{3-2}\,dx = (3-1)! = 2$$

Also,

$$E(x^2 + 2x + 5) = E(x^2) + 2\,E(x) + 5$$
$$= 2 + 2 \times 1 + 5 = 9$$

Example 12. Let X be a random variable with following probability distribution:

x	-3	6	9
P(x)	1/6	1/2	1/3

Find $E(x)$, $E(x^2)$, $E(3x+4)$.

Solution

x	X^2	P(x)	xP(x)	$X^2P(x)$
-3	9	1/6	-0.5	1.5
6	36	1/2	3.0	18
9	81	1/3	3.0	27
			$\Sigma xP(x) = 5.5$	$\Sigma x^2P(x) = 46.5$

Scanned by CamScanner

Now $E(x) = \Sigma x P(x) = 5.5$

$E(x^2) = \Sigma x^2 P(x) = 46.5$

$E(3x + 4) = 3E(x) + 4 = 3 \times 5.5 + 4 = 16.5 + 4 = 20.5$

Addition Theorem of Expectation

The mathematical expectation of the sum of random variables is equal to the sum of the expectations.

Let X and Y be any two random variables then

$$E(X+Y) = E(X) + E(Y)$$

For three random variables X,Y and Z

$$E(X+Y+Z) = E(X) + E(Y) + E(Z)$$

Multiplication Theorem of Expectation

The mathematical expectation of the product ofindependent random variables is equal to the product of their expectations.

Let X and Y be any two independent random variables then $E(XY) = E(X)\,E(Y)$

For three independent random variables X, Y and Z

$$E(XYZ) = E(X)\,E(Y)\,E(Z)$$

Moments about origin

Let x be a random variable then first four moment about origin(Raw moment with assumed mean zero) is expected value of different power of X which is denoted by μ'_r and is given by

$\mu'_r = E(x^r) = \Sigma_x x^r p(x)$, for discrete random variable X having pmf p(x)

$= \int_0^\infty x^r f(x)\,dx$, for continuous random variable having pdf f(x)

$r = 1, 2, 3, 4$

Moments about Mean

Let x be a random variable then first four moment about mean(central moment) is expected value of different power of deviation of X from expected value of X which is denoted by μ_r and is given by

$\mu_r = E[x - E(x)]^r$

$= \sum_x [x - E(x)]^r p(x)$, for discrete random variable X having pmf p(x)

$= \int_0^\infty [x - E(x)]^r f(x)\,dx$, for continuous random variable having pdf f(x)

$r = 1, 2, 3, 4$

Relationship between Central Moments (Moments about Mean) and Raw Moments(Moments about Origin)

$\mu_1 = 0$

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

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

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

Measure of Central Tendency, Dispersion, Skewness and Kurtosis

Central tendency $= \mu_1'$

Dispersion $= \mu_2$

Skewness $= \beta_1 = \frac{\mu_3^2}{\mu_2^3}$ if μ_3 is positive only $= \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$ if μ_3 is positive or negative.

Kurtosis $= \beta_2 = \frac{\mu_4}{\mu_2^2}$

Scanned by CamScanner

Example 13. The number of failure of a computer system in a week of operation has the following

No of failures	0	1	2	3	4	5	6
Probability	0.18	0.28	0.25	0.18	0.06	0.04	0.01

Find first four moments about origin, first four moments about mean, measure of central tendency, measure of dispersion, measure of skewness and measure of kurtosis.

Solution

x	P(x)	xP(x)	$x^2P(x)$	$x^3P(x)$	$x^4P(x)$
0	0.18	0	0	0	0
1	0.28	0.28	0.28	0.28	0.28
2	0.25	0.5	1.0	2	4
3	0.18	0.54	1.62	4.86	14.58
4	0.06	0.24	0.96	3.84	15.36
5	0.04	0.2	1	5	25
6	0.01	0.06	0.36	2.16	12.96
		$\Sigma xp(x)= 1.82$	$\Sigma x^2p(x)=5.22$	$\Sigma x^3p(x)= 18.14$	$\Sigma x^4p(x) = 72.18$

Here, first four moments about origin

$\mu_1' = E(x) = \Sigma xP(x) = 1.82$

$\mu_2' = E(x^2) = \Sigma x^2P(x)=5.22$

$\mu_3' = E(x^3) = \Sigma x^3P(x) = 18.14$

$\mu_4' = E(x^4)= \Sigma x^4P(x) = 72.18$

Now first four moments about mean

$\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 5.22 - (1.82)^2 = 1.907$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 18.14 - 3\times 5.22 \times 1.82 + 2 \times (1.82)^2 = -3.736$

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

$= 72.18 - 4\times 18.14 \times 1.82 + 6 \times 5.22 \times (1.82)^2 - 3 \times (1.82)^4 = 10.949$

Again,

Measure of central tendency, $E(x) = 1.82$

Measure of dispersion, $\sigma = \sqrt{\mu_2} = \sqrt{1.907} = 1.38$

Measure of skewness, $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-3.736}{(1.907)^{3/2}} = \frac{-3.736}{2.633} = -1.418$

Measure of kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{10.949}{(1.907)^2} = \frac{10.949}{3.636} = 3.011$

Example 14. Find first four moments about origin, first four moments about mean, measure of central tendency, measure of dispersion, measure of skewness and measure of kurtosis from following distribution

$$f(x) = \begin{cases} e^{-x}, & x>0 \\ 0, & \text{otherwise} \end{cases}$$

Solution

First four moments about origin

$\mu_1' = E(x) = \int_0^\infty x\, f(x)dx = \int_0^\infty xe^{-x} = \int_0^1 e^{-x} x^{2-1}\, dx = (2-1)! = 1$

Scanned by CamScanner

$\mu_2' = E(x^2) = \int_0^\infty x^2 f(x)dx = \int_0^\infty x^2 e^{-x} = \int_0^1 e^{-x} x^{3-1}\, dx = (3-1)! = 2$

$\mu_3' = E(x^3) = \int_0^\infty x^3 f(x)dx = \int_0^\infty x^3 e^{-x} = \int_0^1 e^{-x} x^{4-1}\, dx = (4-1)! = 6$

$\mu_4' = E(x^4) = \int_0^\infty x^4 f(x)dx = \int_0^\infty x^4 e^{-x} = \int_0^1 e^{-x} x^{5-1}\, dx = (5-1)! = 24$

Now first four moments about mean

$\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 2 - 1 = 1$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 6 - 3\times 2\times 1 + 2\times 1 = 6 - 6 + 2 = 2$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 24 - 4\times 6\times 1 + 6\times 2\times 1 - 3\times 1 = 24 - 24 + 12 - 3 = 9$

Again,

Measure of central tendency, $E(x) = 1$

Measure of dispersion, $\sigma = \sqrt{\mu_2} = \sqrt{1} = 1$

Measure of skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{2^2}{1^3} = 4$

Measure of kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{9}{1^2} = 9$

Variance of univariate random variable

Let X be a random variable then expected value of the square of deviation of random variable X from it's expected value E(X) is called variance of X.

It is denoted by V(X) and is given by $v(X) = E[X - E(X)]^2$

$= E[X^2 - 2X\,E(X) + \{E(X)\}^2]$

$= E(X^2) - 2E(X)\,E(X) + \{E(X)\}^2$

$= E(X^2) - 2\{E(X)\}^2 + \{E(X)\}^2$

$= E(X^2) - \{E(X)\}^2$

Covariance of two dimensional random variable

Let (X,Y) be two dimensional random variable then the expected value of the product of deviation of random variable X and Y from their respective expected values E(x) and E(Y) respectively is called covariance of random variables X and Y.

It is denoted by Cov(X,Y) and is given by $Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$

$= E[XY - XE(Y) - YE(X) + E(X)E(Y)]$

$= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)$

$= E(XY) - E(X)E(Y)$

Note:

If X and Y are independent random variables $E(XY) = E(X)E(Y)$ then

$Cov(X,Y) = E(XY) - E(X)E(Y)$

$= E(X)E(Y) - E(X)E(Y) = 0$

Here, $Cov(X,X) = V(X)$

$Cov(Y,Y) = V(Y)$

Scanned by CamScanner

Properties of Variance

i. $V(a) = 0$

ii. $V(aX+b) = a^2V(X)$

iii. $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2abCov\,(X,Y)$

Conditional Expectation

Let X and Y be any two random variables then the conditional variance of X given $Y = y_j$ is given by

$V(X/Y = y_j) = E\{X^2/Y = y_j\} - \{E(X/Y = y_j)\}^2$

$= \sum_i x_i^2 P(X = x_i/Y = y_j) - (\sum_i x_i P(X = x_i/Y = y_j))^2$; for discrete random variables

Let X and Y be any two random variables then the conditional variance of X given Y is given by

$V(X/Y) = E\{X^2/Y\} - \{E(X/Y)\}^2$

$= \int_{-\infty}^{\infty} x^2 f\left(\frac{x}{y}\right) dx - \left(\int_{-\infty}^{\infty} x f\left(\frac{x}{y}\right) dx\right)^2$; for continuous random variables

Similarly,

$V(Y/X = x_i) = E\{Y^2/X = x_i\} - \{E(Y/X = x_i)\}^2$

$= \sum_j y_j^2 P(Y = y_j/X = x_i) - (\sum_j y_j P(Y = y_j/X = x_i))^2$; for discrete random variables

$V(Y/X) = E\{Y^2/X\} - \{E(Y/X)\}^2$

$= \int_{-\infty}^{\infty} y^2 f\left(\frac{y}{x}\right) dy - \left(\int_{-\infty}^{\infty} y f\left(\frac{y}{x}\right) dy\right)^2$; for continuous random variables

Properties of Conditional Expectation

1. $E(a/x) = a$, where a is a constant
2. $E\{aX + b)/Y\} = aE(X/Y) + b$
3. $E\{E(X/Y)\} = E(X)$
4. $E\{\sum_{i=1}^{n} a_i X_i/Y\} = \sum_{i=1}^{n} a_i E(X_i/Y)$
5. $E\{V(X/Y)\} + V\{E(X/Y)\} = V(X)$

Example 15. Let $P(x, y) = \frac{x + 3y}{24}$, $x = 1, 2$ and $y = 1, 2$. $p(y) = \frac{2y + 1}{8}$, find $E\left(\frac{X}{Y} = 2\right)$ and $V\left(\frac{X}{Y} = 2\right)$.

Solution

$E(X/Y = 2) = \sum_{x=1}^{2} x\, P(X/y = 2)$

$= \sum_{x=1}^{2} x\, P\frac{P(x, y = 2)}{P(y = 2)}$

$= \sum_{x=1}^{2} x \frac{\frac{x + 6}{24}}{\frac{5}{8}}$

Scanned by CamScanner

$$= \sum_{x=1}^{2} x \frac{x+6}{15}$$

$$= \frac{1}{15} \sum_{x=1}^{2} (x^2 + 6x)$$

$$= \frac{1}{15}(1 + 6 + 4 + 12) = \frac{23}{15}$$

$$V(X/Y=2) = \sum_{x=1}^{2} x^2 P(X/y = 2) - \{E(X/Y = 2)\}^2$$

$$= \sum_{x=1}^{2} x^2 \frac{P(x, y = 2)}{P(y = 2)} - \left(\frac{23}{15}\right)^2$$

$$= \sum_{x=1}^{2} x^2 \frac{\frac{x+6}{24}}{\frac{5}{8}} - 2.351$$

$$= \frac{1}{15} \sum_{x=1}^{2} x^2 \frac{x+6}{15} - 2.351$$

$$= \frac{1}{15} \sum_{x=1}^{2} (x^3 + 6x^2) - 2.351$$

$$= \frac{1}{15}(1 + 6 + 8 + 24) - 2.351$$

$$= \frac{39}{15} - 2.351$$

$$= 2.6 - 2.351$$

$$= 0.249$$

Example 16. Let $f(x, y) = \begin{cases} 2, & 0 < X < Y < 1 \\ 0, & \text{Otherwise} \end{cases}$ and $f(x) = \begin{cases} 2(1-x); & 0 < x < 1 \\ 0; & \text{otherwise} \end{cases}$. Find E(Y/X) and V(Y/X).

Solution

$$E(Y/X) = \int_{-\infty}^{\infty} y\, f\left(\frac{y}{x}\right) dy$$

$$= \int_{x}^{1} y \frac{f(x, y)}{f(x)} dy$$

$$= \int_{x}^{1} y \frac{2}{2(1-x)} dy$$

$$= \frac{1}{1-x} \int_{x}^{1} y\, dy = \frac{1-x^2}{2(1-x)} = \frac{1+x}{2}$$

$$V(Y/X) = \int_{-\infty}^{\infty} y^2 f\left(\frac{y}{x}\right) dy - \left\{\int_{-\infty}^{\infty} y f\left(\frac{y}{x}\right) dy\right\}^2$$

$$= \int_{x}^{1} y^2 \frac{2}{2(1-x)}\, dy - \left(\frac{1+x}{2}\right)^2$$

$$= \frac{1}{1-x} \int_{x}^{1} y^2\, dy - \left(\frac{1+x}{2}\right)^2$$

Scanned by CamScanner

$$= \frac{1-x^3}{3(1-x)} - \left(\frac{1+x}{2}\right)^2$$

$$= \frac{1+x+x^2}{3} - \frac{1+2x+x^2}{4}$$

$$= \frac{4+4x+4x^2-3-6x-3x^2}{12}$$

$$= \frac{x^2-2x+1}{12}$$

$$= \frac{(x-1)^2}{12}$$

Computer Application

Example 17. Find the expected value and its variance of the random variable with value of maximum or the value that is same in both of dices when two dice are thrown.

Let X be the r.v. denoting with value of maximum or the value that is same in both of dices when two dice are thrown, X can take values 1, 2, 3, 4, 5 and 6 as follows:

	A	B	C	D	E	F	G
4		1	2	3	4	5	6
5	1	1	2	3	4	5	6
6	2	2	2	3	4	5	6
7	3	3	3	3	4	5	6
8	4	4	4	4	4	5	6
9	5	5	5	5	5	5	6
10	6	6	6	6	6	6	6

Note: In this case two dice are thrown and max value is the random variable or equal value in both case have that value as value of random variable. First table shows the possible sum of the outcomes of two dice. Second table is the table to calculate the E(X), E(X²) and V(X) when we calculate as in theory. Third part is the calculation of the values using excel relations. Formula column in column E is the calculation using usual method while formula column in Column G is the formula using SUMPRODUCT formula in excel.

	A	B	C	D
15	X	P(X=x)	x.P(X=x)	x^2*P(X=x)
16		=COUNTIF($B$6:$G$11,A17)/COUNT($B$6:$G$11)	=A17*B17	=(A17^2)*B17
17	1	0.0278	0.0278	0.0278
18	2	0.0833	0.1667	0.3333
19	3	0.1389	0.4167	1.2500
20	4	0.1944	0.7778	3.1111
21	5	0.2500	1.2500	6
22	6	0.3056	1.8333	11.0000
23	Total	1	4.472222	21.972222
24	Formula	=SUM(B17:B22)	=SUM(C17:C22)	=SUM(D17:D22)

Scanned by CamScanner

	A	B	C	D	E	F	G
25							
26				From Table	Formula	From formula	Formula
27	Mean	E(X)	$\Sigma x*P(X=x)$	4.4722222	=C23	4.47222222	=SUMPRODUCT(A17:A22,B17:B22)
28		$E(X^2)$	$\Sigma x^2*P(X=x)$	21.972	=D23	21.972	=SUMPRODUCT(A17:A22,A17:A22,B17:B22)
29	Variance	V(X)	$E(X^2)-(E(X))^2$	1.971	=D28-(D27)^2	1.971	=SUMPRODUCT(A17:A22,A17:A22,B17:B22)-(SUMPRODUCT(A17:A22,B17:B22)^2)

Exercise 5

Theoretical Questions

1. What is random variable? Describe discrete and continuous random variables with examples.
2. What do you mean by probability mass function? Write down the properties of it.
3. What do you mean by probability density function? Write down it's properties.
4. State addition and multiplication theorem of expectation
5. Differentiate between moments about origin and moments about mean.
6. Define the terms; expectation, probability distribution, random variable

Numerical Questions

1. A random variable x has the following probability distribution.

X	0	1	2	3	4	5	6	7	8
P(X)	a	3a	5a	7a	9a	11a	13a	15a	17a

(i)Determine value of a (ii) Find P(X<3), P(X≥3), P(0<X<5). **[Ans: 1/81, 1/9, 8/9, 8/27]**

2. A coin is tossed until a tail appears. What is the expected number of toss required? **[Ans: 2]**
3. Two unbiased dice are thrown. Find the expected values of the sum of numbers of pointson them. **[Ans: 7]**
4. A random variable has the following probability function.

X	-2	-1	0	1	2
P(X)	0.2	k	0.4	2k	k

Find:(i)k (ii) $P(x \geq 0)$ (iii)E(x) iv)E(4X+5). **[Ans: 0.1 , 0.7 , -0.1 , 4.6]**

5. Find mean and standard deviation of random variable having following probability distribution

x	0	1	2	3	4
P(x)	0.05	0.35	0.2	0.25	0.15

[Ans : 2.1, 1.179]

6. An attendant receives an amount of Rs 700, 900, 1100, 1300, 1500 and 1700 with probabilities 1/12, 1/12, 3/12, 3/12,2/12 and 2/12 respectively between 9 AM to 10 AM in a particular day. What is attendant expected earning during this period? **[Ans : 1266.67]**
7. An importer is offered a shipment of machine tools for Rs. 140,000 and the probability that he will be able to sell them for Rs. 180,000 Rs 170,000 or Rs 150,000 are 0.32, 0.55 and 0.13 respectively. What is the importers expected gross profit? **[Ans : 30600]**

Scanned by CamScanner

8. If a man purchases a raffle ticket he can win a first prize of Rs.5000, second prize of Rs. 3000 and a third prize of Rs. 1000 with probabilities 0.001, 0.003 and 0.005 respectively. What should be the fairy price to pay for a ticket? **[Ans: 19]**

9. Three bags contain 3 video cards and 2 network cards, 5 video cards and 6 network cards, 2 video cards and 4 network cards respectively. One card is drawn from each urn. Find the expected number of video cards. **[Ans:1.61]**

10. A jewelry dealer is interested in purchasing gold necklace for which probabilities are 0.18, 0.22, 0.33 and 0.27 respectively that it will be able to sell it for a profit of Rs 5000, Rs 8000, breakeven and sell for a loss of Rs 3000. Find expected profit? **[Ans: 1850]**

11. Suppose a continuous random variable X has the density function

$$f(x) = \begin{cases} k(1-x)^2 & \text{for } 0<x<1 \\ 0; & \text{elsewhere} \end{cases}$$

Find: (i) value of constant k (ii) $P(0 < X < 0.5)$ (iii) $E(x)$ **(Ans: 3 , 0.875, ¼)**

12. A continuous random variable x has the following density function

$$f(x) = \begin{cases} kx^2 & \text{for } 0<x<1 \\ 0, & \text{elsewhere} \end{cases}$$

Find the value of i) k ii) $P(0< x <0.5)$ and iii) $E(x)$. **[Ans: 3, 0.125, ¾]**

13. The time in minutes it takes to reboot a certain system is a continuous variable with the density function $f(x) = \begin{cases} c(10-x)^2, & 0 < x < 10 \\ 0; & \text{otherwise} \end{cases}$. Compute c then find the probability that it takes between 1 and 2 minutes to reboot.

[Ans:3/1000, 0.217]

14. Find the mean of the random variable having probability density function $f(x) = xe^{-x}$, $x \geq 0$ **[Ans: 2]**

15. The pdf of a continuous random variable is given by

$$f(x) = \begin{cases} 12\,x(1-x)^2, & 0 \leq x \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

Find $E(x)$, $E(x^2)$ **[Ans: 2/5, 1/5]**

16. For an assignment if an equipment is working the density function of outcome x is given by

$$f(x) = \begin{cases} 2(1-x), & 0< x<1 \\ 0, & \text{otherwise} \end{cases}$$

Find (i) mean (ii) Variance and (iii) SD

[Ans: 1/3, 1/18, 0.235]

17. A continuous random variable X has a pdf $f(x)=3x^2$, $0 \leq x \leq 1$. Find a and b such that (i) $P(X \leq a)= P(X>a)$ (ii) $P(X>b)=0.05$ **[Ans: 0.79, 0.98]**

18. Find mean and S.D. of a distribution having pdf $f(x) = ke^{-x/\sigma}$, $0 < x < \infty$ and $\sigma > 0$.

[Ans: σ, σ]

19. X is a random variable taking values –3, –2, –1,0,1,2,3 with following probability:

x	–3	–2	–1	0	1	2	3
P(x)	1/9	1/9	1/9	1/3	1/9	1/9	1/9

Find (i) first four moments about origin (ii) first four moments about mean (iii)Measure of central tendency, dispersion, skewness and kurtosis.

[Ans:0, 3.11, 0, 21.77,0, 3.11, 0, 21.77,0, 1.76, 0, 2.25]

Scanned by CamScanner

Chapter 6

PROBABILITY DISTRIBUTIONS

Introduction

The statistical measures like the averages, dispersion, Skewness, kurtosis, correlation, etc., for the sample frequency distributions, not only give us the nature of the data but also help us in formulating certain ideas about the characteristics of the population. More appropriate method of studying the data is its distribution, in the form of theoretical distribution.

The random variable, mathematical expectation, probability function and distribution function, moments, mean and variance in terms of probability function provide us the necessary tools for the study of theoretical distributions

Probability Distribution

It is a formula that specifies the probabilities associated with the random variables. A probability distribution must satisfy two conditions.

(i) Probability associated with each random variable should be greater than or equal to zero

(ii) Total of all the probability values associated with random variables should be 1.

Distribution Function

Let x be a random variable then the cumulative distribution function or simply distribution function of the random variable is the probability that random variable takes value less than or equal to x. The distribution function is denoted by F(x) and is given by

$F(x) = P(X \le x) = \sum_x P(x)$ for discrete random variable having pmf P(x)

$= \int_{-\infty}^{x} f(x)\, dx$ for continuous random variable having pdf f(x)

Properties of distribution function

i. The distribution function is defined for every real number x.

ii. $0 \le F(x) \le 1$

iii. $F(-\infty) = 0$ and $F(\infty) = 1$.

iv. $P(a \le x \le b) = F(b) - F(a)$

Scanned by CamScanner

v. $F'(x) = f(x)$

Joint Probability Distribution of Two Random Variable

Let us consider a random experiment in which we measure two numerical characteristics say X and Y which represents the outcome of the experiment and is denoted by (x,Y).eg Height and weight of an individual in a group, Marks secured by students in Statistics and Mathematics, age of husbands and age of their wives etc are two dimensional random variables.

Two dimensional random variable is a rule in which we assign real value to each outcome of random experiment in pair (X, Y). It is also discrete and continuous as univariate random variable. The probability distribution obtained by using two randomvariable is called joint probability distribution.

Joint Probability mass function (joint pmf)

Let(X, Y) be two dimensional discrete random variable taking values (x_i, y_j), i = 1 , 2 ,...., n and j = 1, 2, 3, ..., m then probability of ordered pair denoted by $p_{ij} = P(x_i, y_j) = P(x = x_i \text{ and } Y = y_j)$ is called joint probability mass function if it satisfies

i. $p_{ij} = P(X = x_i, Y = y_j) \geq 0$

ii. $\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1$

The point probability mass function of random variable (X,Y) taking value (x_i, y_j) can be presented in following tabular form.

Y \ X	x_1	x_2.....	x_i.....	X_n	Total
y_1	p_{11}	p_{21}	p_{i1}	P_{n1}	$p_{.1}$
y_2	p_{12}	p_{22}	p_{i2}	P_{n2}	$p_{.2}$
.					
y_j	p_{1j}	p_{2j}	p_{ij}	P_{nj}	$P_{.j}$
.					
Y_m	p_{1m}	p_{2m}	p_{im}	P_{nm}	$p_{.m}$
Total	$p_{1.}$	$p_{2.}$	$p_{i.}$	$P_{n.}$	$P_{..}$

Marginal Probability Mass Function (Marginal pmf)

Let (X,Y) be two dimensional discrete random variable taking values (x_i, y_j) with probability mass function $p_{ij} = P(x_i, y_j)$, i = 1 , 2, 3,.... n and j = 1 , 2, 3,.... m. then the probability mass function of one discrete random variable obtained by summing the joint pmf over other discrete random variable is called marginal probability mass function.

The probability of discrete random variable X denoted by $p_i = P(x_i) = P(X=x_i)$

$= P(X=x_i \text{ and } Y = y_1) + P(X=x_i \text{ and } Y = y_2) + P(X=x_i \text{ and } Y = y_3) + \ldots + P(X=x_i \text{ and } Y = y_m)$

$= p_{i1} + p_{i2} + p_{i3} + \ldots p_{im} = \sum_{j=1}^{m} p_{ij} = p_{i.}$ is called marginal probability mass function of random variable X if it satisfies

i. $p_i \geq 0$

ii. $\sum_{i=1}^{n} p_i = 1$

Similarly,

The probability of discrete random variable Y denoted by $p_j = P(y_j) = P(Y=y_j)$

$= P(X=x_1 \text{ and } Y = y_j) + P(X=x_2 \text{ and } Y = y_j) + P(X=x_3 \text{ and } Y = y_j) + \ldots + P(X=x_n \text{ and } Y = y_j)$

Scanned by CamScanner

$= P_{1j}+P_{2j}+P_{3j}+ \ldots P_{nj} = \sum_{i=1}^{n} p_{ij} = p.j$ is called marginal probability mass function of random variable Y if it satisfies.

i. $p.j \geq 0$

ii. $\sum_{i=1}^{m} p.j = 1$

Conditional probability mass function (conditional pmf)

Let (X,Y) be two dimensional discrete random variable taking values (x_i, y_j) with probability mass function $P_{ij} = P(x_i, y_j)$, i = 1 , 2, 3,... n and j = 1 , 2, 3,... m. $P_i = p(x_i)$ and $p_j = P(y_j)$ be marginal probability mass functions of random variables X and Y respectively then the probability mass function of one discrete random variable given other discrete random variable is called conditional probability mass function.

The probability of discrete random variable X for given Y denoted by $p_{i/j} = P(x_i/y_j) = \frac{P(x_i \text{ and } y_j)}{P(y_j)}$

$= \frac{P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p.j}$ is called conditional probability mass function of random variable X given Y if it satisfies

i. $\frac{p_{ij}}{p.j} \geq 0$ ii. $\sum_{i=1}^{n} \frac{p_{ij}}{p.j} = 1$

Similarly,

The probability of discrete random variable Y for given X denoted by

$p_{j/i} = P(y_j/x_i) = \frac{P(x_i \text{ and } y_j)}{P(x_j)} = \frac{P(X = x_i \text{ and } Y = y_j)}{P(X = x_j)} = \frac{p_{ij}}{p_i}$ is called conditional probability mass function of random variable Y given X if it satisfies

i. $\frac{p_{ij}}{p_i} \geq 0$

ii. $\sum_{i=1}^{m} \frac{p_{ij}}{p_j} = 1$

Example 1: From following bivariate probability distribution of X and Y find the (i)marginal probability distribution of X and Y (ii) conditional probability distribution of X for given Y

X \ Y	0	1	2
1	1/20	2/20	3/20
2	2/20	3/20	4/20
3	1/20	2/20	2/20

Solution

Marginal probability distribution

X \ Y	0	1	2	P(X=x)
1	1/20	2/20	3/20	6/20
2	2/20	3/20	4/20	9/20
3	1/20	2/20	2/20	5/20
P(Y=y)	4/20	7/20	9/20	1

Scanned by CamScanner

Marginal probability of X = 1

$$P(X=1) = P(X = 1, Y = 0) + P(X = 1, Y = 1) + P(X = 1, Y = 2) = \frac{1}{20} + \frac{2}{20} + \frac{3}{20} = \frac{6}{20}$$

And so on

Conditional probability distribution of X given Y

X \ Y	0	1	2
1	1/4	2/7	3/9
2	2/4	3/7	4/9
3	1/4	2/7	2/9

Conditional probability of X=1 given Y = 0

$$P(X = 1/Y = 0) = \frac{P(X = 1, Y = 0)}{P(Y = 0)} = \frac{1/20}{4/20} = \frac{1}{4}$$

and so on

Example 2: A lab contains 5 desktop and 3 laptop computers; two computers are selected at random from the lab. Let X denotes number of desktop computers and Y denotes number of laptop computers. Find (i) joint probability distribution (ii) marginal probability distribution (iii) conditional probability distribution.

Solution

Desktop computers = 5, Laptop computers = 3, Total computers = 5+3 = 8, Computers selected = 2

$$P(X=0,Y=2) = \frac{C(5, 0)\, C(3, 2)}{C(8, 2)} = \frac{3}{28}$$

$$P(X=1,Y=1) = \frac{C(5, 1)\, C(3, 1)}{C(8, 2)} = \frac{15}{28}$$

$$P(X=2,Y=0) = \frac{C(5, 2)\, C(3, 0)}{C(8, 2)} = \frac{10}{28}$$

Now,

X \ Y	0	1	2	P(X = x)
0	0	0	3/28	3/28
1	0	15/28	0	15/28
2	10/28	0	0	10/28
P(Y=y)	10/28	15/28	3/28	1

Conditional probability distribution of X given Y

$$P(X=0/Y=0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)} = \frac{0}{\frac{10}{28}} = 0 \text{ and so on.}$$

X \ Y	0	1	2
0	0	0	1
1	0	1	0
2	1	0	0

Scanned by CamScanner

Conditional distribution of Y given X

$$P(Y=0/X=1) = \frac{P(X=1, Y=0)}{P(X=1)} = \frac{0}{\frac{15}{28}} = 0 \text{ and so on}$$

X \ Y	0	1	2
0	0	0	1
1	0	1	0
2	1	0	0

Example 3: A computer shop has 3 Apple,4 Lenovo and 1 Samsung brand computers. A helper in shop selects two computers at random to explain features of computers to a customer. Let X be the no of Apple brand computers and Y be the no Lenovo brand computers. Obtain (i) joint probability distribution of X and Y (ii) Marginal probability distribution of X and Y(iii) conditional probability distribution of Y given X.

Solution

Here X = No of Apple brand computers, Y = No Lenovo brand computers

Apple brand computers = 3, Lenovo brand computers = 4,Samsung brand computers = 1,

Total computers = 3+4+1=8, Computers selected = 2

$P(X=0,Y=0) = 0$

$$P(X=0,Y=1) = \frac{C(3,0) \times C(4,1) \times C(1,1)}{C(8,2)} = \frac{4}{28}$$ [Probability of 1 Apple and 1 Lenovo computers]

$$P(X=0,Y=2) = \frac{C(3,0) \times C(4,2) \times C(1,0)}{C(8,2)} = \frac{6}{28}$$

$$P(X=1,Y=0) = \frac{C(3,1) \times C(4,0) \times C(1,1)}{C(8,2)} = \frac{3}{28}$$ [Probability of 1 Apple and 1 Lenovo computers]

$$P(X=1,Y=1) = \frac{C(3,1) \times C(4,1) \times C(1,0)}{C(8,2)} = \frac{12}{28}$$

$P(X=1,Y=2) = 0$

$$P(X=2,Y=0) = \frac{C(3,2) \times C(4,0) \times C(1,0)}{C(8,2)} = \frac{3}{28}$$

$P(X=2,Y=1) = 0$

$P(X=2,Y=2) = 0$

X \ Y	0	1	2	P(X= x)
0	0	4/28	6/28	10/28
1	3/28	12/28	0	15/28
2	3/28	0	0	3/28
P(Y=y)	6/28	16/28	6/28	1

Conditional probability distribution of Y given X

Conditional probability distribution of Y = 1 given X =0

$$P(Y=1/X=0) = \frac{P(X=0, Y=1)}{P(X=0)} = \frac{4/28}{10/28} = \frac{4}{10} \text{ and so on}$$

Scanned by CamScanner

X \ Y	0	1	2
0	0	4/10	6/10
1	3/15	12/15	0
2	1	0	0

Example 4: From the following joint probability distribution find (i) $P(X = 3)$ (ii) $P(Y \le 2)$ (iii) $P(X = 1, Y = 3)$ (iv) $P(X = 2, Y \ge 2)$ (v) $P(Y < 3/X = 2)$ (vi) $P(X > 2/Y = 1)$

X \ Y	1	2	3
1	1/25	2/25	6/25
2	3/25	4/25	2/25
3	2/25	1/25	4/25

Solution

X \ Y	1	2	3	P(X=x)
1	1/25	2/25	6/25	9/25
2	3/25	4/25	2/25	9/25
3	2/25	1/25	4/25	7/25
P(Y=y)	6/25	7/25	12/25	1

i) $P(X=3) = P(X=3,Y=1) + P(X=3,Y=2) + P(X=3,Y=3) = \frac{2}{25} + \frac{1}{25} + \frac{4}{25} = \frac{7}{25}$

ii) $P(Y \le 2) = P(Y = 1) + P(y = 2) = \frac{6}{25} + \frac{7}{25} = \frac{13}{25}$

iii) $P(X=1,Y=3) = \frac{6}{25}$

iv) $P(X=2, Y \ge 2) = P(X = 2, Y = 2) + P(X=2,Y=3) = \frac{4}{25} + \frac{2}{25} = \frac{6}{25}$

v) $P(Y < 3/X=2) = \frac{P(Y < 3, X = 2)}{P(X = 2)} = \frac{P(X = 2, Y = 1) + P(X = 2, Y = 2)}{P(X = 2)} = \frac{\frac{a}{25} + \frac{4}{25}}{\frac{9}{25}} = \frac{7}{9}$

vi) $P(X > 2/Y = 1) = \frac{P(X > 2, Y = 1)}{P(Y = 1)} = \frac{P(X = 3, Y = 1)}{P(Y = 1)} = \frac{2/25}{6/25} = \frac{1}{3}$

Example 5: The joint pmf of random variables X and Y be

$$P(x, y) = \begin{cases} C(x + y), & x = 1, 2, 3; y = 1, 2 \\ 0 & \text{otherwise} \end{cases}.$$

Find (i) C (ii) Marginal pmf of X (iii) Marginal pmf of Y (iv) Conditional pmf of X given Y (v) Conditional pmf of Y given X

Solution

We know, $\sum_{x=1}^{3} \sum_{y=1}^{2} P(x, y) = 1$

or $\sum_{x=1}^{3} \sum_{y=1}^{2} C(x + y) = 1$

or $C[(1+1) + (1+2) + (2+1) + (2+2) + (3+1) + (3+2)] = 1$

Scanned by CamScanner

$21C = 1$

$C = \frac{1}{21}$

Hence, $P(x, y) = \frac{1}{21}(x + y)$

Now, Marginal probability distribution

Marginal pmf of X, $P(x) = \sum_{y=1}^{2} P(x, y) = \frac{1}{21}\sum_{y=1}^{2}(x + y)$

$= \frac{1}{21}[(x + 1) + (x + 2)] = \frac{2x + 3}{21}$

Marginal pmf of Y, $P(y) = \sum_{x=1}^{3} P(x, y) = \frac{1}{21}\sum_{x=1}^{3}(x + y)$

$= \frac{1}{21}[(1 + y) + (2 + y) + (3 + y)]$

$= \frac{3y + 6}{21} = \frac{1}{7}(y + 2)$

Conditional probability distribution

Conditional pmf of X given Y, $P(x/y) = \frac{P(x, y)}{P(y)} = \frac{\frac{x + y}{21}}{\frac{3y + 6}{21}} = \frac{x + y}{3y + 6}$

Conditional pmf of Y given X, $P(y/x) = \frac{P(x, y)}{P(x)} = \frac{\frac{x + y}{21}}{\frac{2x + 3}{21}} = \frac{x + y}{2x + 3}$

Joint probability density function (Joint pdf)

Let(X, Y) be two dimensional continuous random variable taking values $(-\infty \le X \le \infty, -\infty \le Y \le \infty)$ and f(x, y) is continuous function of (X, Y) the probability that X lies in interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$ and Y lies in the interval $y - \frac{dy}{2}$ and $y + \frac{dy}{2}$ of area dx dy is given by

$$f(x, y)\,dxdy = P\left[x - \frac{dx}{2} \le X \le x + \frac{dx}{2}, y - \frac{dy}{2} \le Y \le y + \frac{dy}{2}\right]$$

f(x, y)dx dy represents the probability that(x, y) lies in the infinitesimal small interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$, $y - \frac{dy}{2}$ and $y + \frac{dy}{2}$.

The probability that X lies in interval [a, b] and Y lies in interval [c, d] is given by

$$P(a \le X \le b, c \le y \le d) = \int_a^b\int_c^d f(x, y)\,dx\,dy = \int_a^b\left\{\int_c^d f(x, y)\,dy\right\}dx$$

Then the function f(x, y) is called joint probability density function if it satisfies the conditions

i. $f(x, y) \ge 0$, for every x and y in the given range

ii. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1, \quad -\infty \le x \le \infty, -\infty \le y \le \infty$

Scanned by CamScanner

Marginal Probability Density Function (Marginal pdf)

Let (X,Y) be two dimensional continuous random variable taking values $(-\infty \le X \le \infty, -\infty \le Y \le$ with joint probability density function f(x,y) then probability function of only one continuous random variable obtained by integrating the joint pdf with respect to other continuous random variable is marginal pdf.

The probability function of continuous random variable x denoted by $f(x) = \int_{-\infty}^{\infty} f(x, y)dy$ is called marginal pdf of X if it satisfies

i. $f(x) \ge 0$

ii. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Similarly,

The probability function of continuous random variable y denoted by $f(y) = \int_{-\infty}^{\infty} f(x, y)dx$ is called marginal pdf of Y if it satisfies

i. $f(y) \ge 0$

ii. $\int_{-\infty}^{\infty} f(y)\, dy = 1$

Conditional probability density function(Conditional pdf)

Let (X,Y) be two dimensional continuous random variable taking values $(-\infty \le X \le \infty, -\infty \le Y \le \infty)$ with joint probability density function f(x, y). Also f(x) and f(y) be marginal pdf of X and Y respectively then the density function of one continuous random variable given other continuous random variable is called conditional pdf.

The density function of a random variable x given y is denoted by $f(x/y) = \frac{f(x, y)}{f(y)}$ is called conditional pdf of x given y if it satisfies

i. $f(x/y) \ge 0$

ii. $\int_{-\infty}^{\infty} f(x/y)\, dx = 1$

Similarly,

The density function of a random variable y given x is denoted by $f(y/x) = \frac{f(x, y)}{f(x)}$ is called conditional pdf of y given x if it satisfies

i. $f(y/x) \ge 0$

ii. $\int_{-\infty}^{\infty} f(y/x)\, dy = 1$

Example 6: Joint pdf of two dimensional random variable (X, Y) is given by

$$f(x, y) = \begin{cases} k(x + y); 0<x<1 \text{ and } 0<y<1 \\ 0, \text{ otherwise} \end{cases}$$

Find (i) k (ii) f(x) (iii) f(y) (iv) f(x/y) (v) f(y/x)

Solution.

We know, $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$

or $\int_{0}^{1}\int_{0}^{1} k(x + y)\, dx\, dy = 1$

Scanned by CamScanner

$$k\int_0^1\left\{\int_0^1 (x+y)\,dy\right\}dx = 1$$

or $$k\int_0^1\left[xy+\frac{y^2}{2}\right]_0^1 dx = 1$$

or $$k\int_0^1\left[x+\frac{1}{2}\right] dx = 1$$

or $$k\left[\frac{x^2}{2}+\frac{x}{2}\right]_0^1 = 1$$

or $$k\left[\frac{1}{2}+\frac{1}{2}\right] = 1$$

or $$k = 1$$

Hence, $f(x,y) = (x+y)$

Now, $f(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$

$$= \int_0^1 (x+y)\,dy = \left[xy+\frac{y^2}{2}\right]_0^1 = x+\frac{1}{2} = \frac{2x+1}{2}$$

$$f(y) = \int_{-\infty}^{\infty} f(x, y)\,dx$$

$$= \int_0^1 (x+y)\,dx = \left[\frac{x^2}{2}+xy\right]_0^1 = \frac{1}{2}+y = \frac{2y+1}{2}$$

$$f(x/y) = \frac{f(x,y)}{f(y)} = \frac{2(x+y)}{2y+1}$$

$$f(y/x) = \frac{f(x,y)}{f(x)} = \frac{2(x+y)}{2x+1}$$

Example 7: Let X and Y have joint pdf given by $F(x, y) = \begin{cases} 2-x-y; & 0<x<1, 0<y<1 \\ 0 & \text{otherwise} \end{cases}$.

Find (i) $P(x>1/4, y<1/2)$ (ii) $f(x)$ (iii) $f(y)$ (iv) $P(x>1/2)$ (v) $P(y<1/2)$.

Solution

$$P(x>1/4, y<1/2) = \int_{1/4}^{1}\int_0^{1/2} f(x, y)\,dx\,dy$$

$$= \int_{1/4}^{1}\left\{\int_0^{1/2} (2-x-y)\,dy\right\}dx$$

$$= \int_{1/4}^{1}\left[2y-xy-\frac{y^2}{2}\right]_0^{1/2} dx$$

$$= \int_{1/4}^{1}\left(1-\frac{x}{2}-\frac{1}{8}\right)dx$$

$$= \int_{1/4}^{1}\left(\frac{7}{8}-\frac{x}{2}\right)dx$$

$$= \left[\frac{7x}{8}-\frac{x^2}{4}\right]_{1/4}^{1} = \left[\frac{7}{8}-\frac{1}{4}-\frac{7}{32}+\frac{1}{64}\right] = \frac{39}{64}$$

$$f(x) = \int_0^1 f(x, y)\,dy = \int_0^1 (2-x-y)\,dy$$

$$= \left[2y-xy-\frac{y^2}{2}\right]_0^1$$

Scanned by CamScanner

$$= 2 - x - \frac{1}{2}$$

$$= \frac{3}{2} - x$$

$$f(y) = \int_0^1 f(x, y)\, dx \quad = \int_0^1 (2 - x - y)\, dx$$

$$= \left[2x - \frac{x^2}{2} - xy\right]_0^1 = 2 - \frac{1}{2} - y = \frac{3}{2} - y$$

$$P(x > 1/2) \quad = \int_{1/2}^1 f(x)\, dx = \int_{1/2}^1 \left(\frac{3}{2} - x\right) dx = \left[\frac{3x}{2} - \frac{x^2}{2}\right]_{1/2}^1 = \frac{5}{8}$$

$$P(y < ½) \quad = \int_0^{1/2} f(y)\, dy = \int_0^{1/2} \left(\frac{3}{2} - y\right) dy = \frac{5}{8}$$

Joint Probability Distribution Function

Let (X,Y) be two dimensional random variable then the cumulative distribution function or simply distribution function of the two dimensional random variable is the probability that random variable X takes value less than or equal to x and random variable Y takes value less than or equal to y. The joint probability distribution function is denoted by F(X,Y) and is given by

$F(X,Y) = P(X \le x, Y \le y) = \sum_x \sum_y P(x_i, y_j)$ for discrete random variable having joint pmf $P(x_i, y_j)$

$= \int_{-\infty}^{x} \left\{ \int_{-\infty}^{y} f(x, y)\, dy \right\} dx$ for continuous random variable having pdf f(x,y)

Properties of joint probability distribution function

i. The distribution function is defined for every pair of real number(X,Y)

ii. $0 \le F(x, y) \le 1$

iii. $F(-\infty, -\infty) = 0$

iv. $F(+\infty, +\infty) = 1$

v. $\dfrac{\partial^2 F(x, y)}{\partial x\, \partial y} = f(x, y)$

Marginal probability distribution function

Let(X, Y) be two dimensional random variable with joint distribution function F(x,y).The distribution function of X or Y alone calculated from joint distribution function F(X,Y) is called marginal probability distribution function.

The marginal probability distribution function of random variable X is denoted by F(x) and is given by

$F(x) = P(X \le x, Y \le \infty) \sum_x P(x_i)$ for discrete random variable X

$= \int_{-\infty}^{x} f(x)\, dx$ for continuous random variable X

The marginal probability distribution function of random variable Y is denoted by F(y) and is given by $F(y) = P(X \le \infty, Y \le y) = \sum_y P(y_j)$for discrete random variable Y

$= \int_{-\infty}^{y} f(y)\, dy$ for continuous random variable y

Properties of marginal probability distribution function

i. The distribution function is defined for every real number

ii. $0 \le F(x) \le 1$

$0 \le F(y) \le 1$

Scanned by CamScanner

iii. $F(-\infty) = 0$

iv. $F(+\infty) = 1$

v. $F(a) \le F(b)$ f or real numbers $a \le b$

Conditional probability distribution function

Let(X,Y) be two dimensional random variable with joint distribution function F(x, y).The distribution function of X when Y has already assumed a linear value y is called conditional distribution function of x when Y is fixed is denoted by F(x/y) and is given by

$F(x/y) = \sum_x P(X \le x/Y = y)$ for discrete random variable

$= \int_{-\infty}^{x} f(x/y)\, dx$

The distribution function of Y when X has already assumed a linear value x is called conditional distribution function of Y when X is fixed is denoted by F(y/X) and is given by

$F(y/x) = \sum_y P(Y \le y/X = x)$ for discrete random variable

$= \int_{-\infty}^{y} f(y/x)\, dy$ for continuous random variable.

Example 8: From following bivariate probability distribution find F(1,1), F(2,1), F(x=1), F(X=2), F(y=2), F(x/y=1), F(x=1/y=2), F(y=2/x=1).

Y \ X	0	1	2
0	1/15	2/15	1/15
1	3/15	2/15	1/15
2	2/15	1/15	2/15

Solution

$F(1,1) = P(x \le 1, y \le 1) = P(0, 0) + P(0, 1) + P(1, 0) + P(1,1)$

$$= \frac{1}{15} + \frac{3}{15} + \frac{2}{15} + \frac{2}{15} = \frac{8}{15}$$

$F(2,1) = P(x \le 2, y \le 1) = P(0,0) + P(0,1) + P(1,0) + P(1,1) + P(2, 0) + P(2,1)$

$$= \frac{1}{15} + \frac{3}{15} + \frac{2}{15} + \frac{2}{15} + \frac{1}{15} + \frac{1}{15} = \frac{10}{15} = \frac{2}{3}$$

Y \ X	0	1	2	P(y)
0	1/15	2/15	1/15	4/15
1	3/15	2/15	1/15	6/15
2	2/15	1/15	2/15	5/15
P(x)	6/15	5/15	4/15	1

$$F(x = 1) = P(x \le 1) = P(x = 0) + P(x = 1) = \frac{6}{15} + \frac{5}{15} = \frac{11}{15}$$

$$F(x = 2) = P(x \le 2) = P(x = 0) + P(x = 1) + P(x = 2) = \frac{6}{15} + \frac{5}{15} + \frac{4}{15} = 1$$

$$F(y = 2) = P(y \le 2) = P(y = 0) + P(y = 1) + P(y = 2) = \frac{4}{15} + \frac{6}{15} + \frac{5}{15} = 1$$

Scanned by CamScanner

$$F(x/y=1) = P(X \le x/y = 1) = P(x=0/y=1) + P(x=1/y=1) + P(x=2/y=1) = \frac{P(0,1)}{P(y=1)} + \frac{P(1,1)}{P(y=1)} + \frac{P(2,1)}{P(y=1)}$$

$$= \frac{\frac{3}{15}}{\frac{6}{15}} + \frac{\frac{2}{15}}{\frac{6}{15}} + \frac{\frac{1}{15}}{\frac{6}{15}} = 1$$

$$F(x=1/y=2) = P(X \le 1/y = 2) = P(x=0/y=2) + P(x=1/y=2) = \frac{P(0,2)}{P(y=2)} + \frac{P(1,2)}{P(y=2)}$$

$$= \frac{\frac{2}{15}}{\frac{5}{15}} + \frac{\frac{1}{15}}{\frac{5}{15}} = \frac{3}{15} = \frac{1}{5}$$

$$F(y=2/x=1) = P(y \le 2/x = 1) = P(y=0/x=1) + P(y=1/x=1) + P(y=2/x=1)$$

$$= \frac{P(1,0)}{P(x=1)} + \frac{P(1,1)}{P(x=1)} + \frac{P(1,1)}{P(x=1)}$$

$$= \frac{\frac{2}{15}}{\frac{6}{15}} + \frac{\frac{2}{15}}{\frac{6}{15}} + \frac{\frac{1}{15}}{\frac{6}{15}} = \frac{5}{6}$$

Example 9. Two dimensional random variable (X,Y) has pdf

$$f(x, y) = \begin{cases} e^{-(x+y)}, & 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find (i) Joint distribution function (ii) Marginal distribution functions (iii) Conditional distribution functions.

Solution

Joint distribution function $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)\, dx\, dy$

$$= \int_0^x\int_0^y e^{-(x+y)}\, dx\, dy$$

$$= \int_0^x\int_0^y e^{-x} e^{-y} dy\, dx$$

$$= \int_0^x e^{-x}\left\{\int_0^y e^{-y}\, dy\right\} dx$$

$$= -\int_0^x e^{-x} [e^{-y}]_0^y\, dx$$

$$= \int_0^x e^{-x} (1 - e^{-y})\, dx$$

$$= (1 - e^{-y}) \int_0^x e^{-x}\, dx$$

$$= -(1 - e^{-y}) [e^{-x}]_0^x$$

$$= (1 - e^{-y})(1 - e^{-x})$$

$$= 1 - e^{-x} - e^{-y} + e^{-(x+y)}$$

Now, $f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^{\infty} e^{-(x+y)} dy$

Scanned by CamScanner

$$= \int_0^{\infty} e^{-x} e^{-y} dy$$

$$= e^{-x} \int_0^{y} e^{-y} e^{1-1} dy$$

$$= e^{-x} (1-1)!$$

$$= e^{-x}$$

Also, $f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^{\infty} e^{-(x+y)}\, dx$

$$= \int_0^{\infty} e^{-x} e^{-y}\, dx$$

$$= e^{-y} \int_0^{\infty} e^{-x} x^{1-1}\, dx$$

$$= e^{-y} (1-1)$$

$$= e^{-y}$$

Marginal distribution function of X, $F(x) = \int_{-\infty}^{x} f(x)\, dx$

$$= \int_0^{x} e^{-x}\, dx$$

$$= [e^{-x}]_0^x$$

$$= (1 - e^{-x})$$

Marginal distribution function of Y, $F(y) = \int_{-\infty}^{y} f(x)\, dy$

$$= \int_0^{y} e^{-y} dy$$

$$= [e^{-y}]_0^y$$

$$= (1 - e^{-y})$$

Again, $f(x/y) = \dfrac{f(x, y)}{f(y)} = \dfrac{e^{-(x+y)}}{e^{-y}} = e^{-x}$

$$f(y/x) = \frac{f(x, y)}{f(x)} = \frac{e^{-(x+y)}}{e^{-x}} = e^{-y}$$

Conditional distribution function of x given y, $F(x/y) = \int_{-\infty}^{x} f\left(\frac{x}{y}\right) dx$

$$= \int_0^{x} e^{-x}\, dx$$

$$= 1 - e^{-x}$$

Conditional distribution function of y given x, $F(y/x) = \int_0^{y} f\left(\frac{y}{x}\right) dy$

$$= \int_0^{y} e^{-y} dy$$

$$= 1 - e^{-y}$$

Scanned by CamScanner

Example 10. A two dimensional random variable (X,Y) has joint pdf

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2); 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 \qquad \text{otherwise} \end{cases}$$. Find F(0.5, 0.5), F(x = 0.25), F(y = 0.75).

Solution

$$F(0.5, 0.5) = P(x \le 0.5, y \le 0.5) = \int_0^{0.5}\int_0^{0.5} f(x, y)\, dx\, dy$$

$$= \int_0^{0.5}\int_0^{0.5} \frac{6}{5}(x + y^2)\, dx\, dy$$

$$= \frac{6}{5}\int_0^{0.5}\left\{\int_0^{0.5}(x + y^2)\, dy\right\} dx$$

$$= \frac{6}{5}\int_0^{0.5}\left[xy + \frac{y^3}{3}\right]_0^{0.5} dx$$

$$= \frac{6}{5}\int_0^{0.5}\left(0.5x + \frac{0.125}{3}\right) dx$$

$$= \frac{6}{5}\left[0.5\frac{x^2}{2} + \frac{0.125x}{3}\right]_0^{0.5}$$

$$= \frac{6}{5}\left(\frac{0.125}{2} + \frac{0.125 \times 0.5}{3}\right)$$

$$= \frac{6}{5} \times \frac{0.5}{6} = 0.1$$

$$f(x) = \int_0^1 f(x, y)\, dy = \int_0^1 \frac{6}{5}(x + y^2)\, dy = \frac{6}{5}\int_0^1 (x + y^2)\, dy$$

$$= \frac{6}{5}\left[xy + \frac{y^3}{3}\right]_0^1$$

$$= \frac{6}{5}\left(x + \frac{1}{3}\right) = \frac{2}{5}(3x + 1)$$

$$f(y) = \int_0^1 f(x, y)\, dx = \int_0^1 \frac{6}{5}(x + y^2)\, dx = \frac{6}{5}\int_0^1 (x + y^2)\, dx$$

$$= \frac{6}{5}\left[\frac{x^2}{2} + xy^2\right]_0^1$$

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

$$F(x = 0.25) = P(x \le 0.25) = \int_0^{0.25} f(x)\, dx = \int_0^{0.25} \frac{2}{5}(3x + 1)\, dx$$

$$= \frac{2}{5}\int_0^{0.25}(3x + 1)\, dx$$

$$= \frac{2}{5}\left[\frac{3x^2}{2} + x\right]$$

$$= \frac{2}{5}\left\{\frac{3}{2} \times (0.25)^2 + 0.25\right\} = 0.1375$$

$$F(y = 0.75) = P(y \le 0.75) = \int_0^{0.75} f(y)\, dy = \int_0^{0.75} \frac{3}{5}(1 + 2y^2)\, dy$$

$$= \frac{3}{5}\int_0^{0.75}(1 + 2y^2)\, dy$$

Scanned by CamScanner

$$= \frac{3}{5}\left\{0.75 + \frac{2}{3}(0.75)^3\right\}$$

$$= \frac{3}{5}(0.75 + 0.281)$$

$$= 0.618$$

Independence of random variables

Let (X, Y) be two dimensional random variable with joint distribution function F(x,y) .Let F(x) and F(y) be marginal distribution functions of random variables X and Y respectively. Random variables X and Y are independent if and only if F(x,y) = F(x) F(y) for every pair of real numbers (X, Y).

Alternately;

If two dimensional discrete random variable (X, Y) has joint pmf $P(x_i, y_j)$, $P(x_i)$ and $P(y_j)$ be marginal pmf of X and Y respectively, then X and Y are independent if and only if $P(x_i, y_j) = P(x_i)$ $P(y_j)$

Alternately;

If two dimensional continuous random variable (X, Y) has joint pdf f(x, y), f(x) and f(y) be marginal pdf of X and Y respectively, then X and Y are independent if and only

if f(x,y) = f(x) f(y)

Example 11. Let two dimensional random variable (X, Y) has joint pmf

$$P(x, y) = \begin{cases} \frac{2x+y}{42}; & x = 0, 1, 2 \text{ and } y = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}$$. Are X and Y independent?

Solution

Now,

$$P(x) = \sum_y P(x, y) = \sum_y \frac{2x+y}{42} = \frac{4x+3}{21}\frac{2x+y}{42}$$

$$P(y) = \sum_x p(x, y) = \sum_x \frac{2x+y}{42} = \frac{(2\times 0 + y) + (2\times 1 + y) + (2\times 2 = y)}{42}$$

$$= \frac{6+3y}{42} = \frac{2+y}{14}$$

Here, $p(x) \times p(y) = \frac{4x+3}{21} \times \frac{2+y}{14} = \frac{8x+4xy+6+3y}{294}$

$$P(x,y) = \frac{2x+y}{42} \neq p(x) \times p(y) = \frac{8x+4xy+6+3y}{294}$$

Hence X and Y are not independent.

Example 12. Let (X,Y) be two dimensional random variable with joint pdf $f(x,y) = 4xy;\ 0 \le y,\ y \le 1$. Are X and Y independent?

Solution

Here, f(x,y) = 4xy

Now, $f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^1 4xy\, dy$

$$= 4x\int_0^1 y\, dy = 4x\left[\frac{y^2}{2}\right]_0^1 = 4x \times \frac{1}{2} = 2x$$

Also , $f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^1 4xy\, dy$

Scanned by CamScanner

$$= 4y \int_0^1 x\,dx = 4y\left[\frac{x^2}{2}\right] = 2y$$

Now $f(x)f(y) = 2x \times 2y = 4xy$

Here, $f(x,y) = f(x)\,f(y) = 4xy$

Hence X and Y are independent.

Expectation (Mean) of two dimensional random variable

For univariate random variable X the expected value is

$E(X) = \sum_x xP(x)$ for discrete random variable X having pmf $p(x)$, $x = 0, 1, 2, 3, \ldots\ldots\ldots\ldots$

$= \int_{-\infty}^{\infty} x\,f(x)\,dx$ for continuous random variable having pdf $f(x)$, $-\infty \le x \le \infty$.

For bivariate random variable(X,Y) when the function is in sum form ,the expected value is

$E(X+Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i + y_j)\,P(x_i, y_j)$ for two dimensional discrete random variable (X,Y)having joint pmf $p(x_i, y_j)$, $i = 1, 2, 3, \ldots\ldots n$ and $j = 1, 2, 3, \ldots\ldots m$

$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x + y)\,f(x, y)\,dx\,dy$ for continuous random variable having joint pdf $f(x,y)$,

$-\infty \le x \le \infty$ and $-\infty \le y \le \infty$.

For bivariate random variable(X,Y) when the function is in product form ,the expected value is

$E(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x_i, y_j)$ for two dimensional discrete random variable (X,Y)having joint pmf $p(x_i, y_j)$, $i = 1, 2, 3\ldots\ldots\ldots\ldots n$ and $j = 1, 2, 3\ldots\ldots\ldots\ldots m$

$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)\,f(x, y)\,dx\,dy$ for continuous random variable having joint pdf $f(x,y)$,

$-\infty \le x \le \infty$ and $-\infty \le y \le \infty$.

Variance of univariate random variable

Let X be a random variable then expected value of the square of deviation of random variable X from it's expected value E(X) is called variance of X.

It is denoted by V(X) and is given by $v(X) = E[X - E(X)]^2 = E(X^2) - \{E(X)\}^2$

Variance of two dimensional random variable

Let (X,Y) be two dimensional random variable and function is in sum form, the variance is

$$V(X+Y) = E(X + Y)^2 - \{E(X + Y)\}^2$$

When the function is in product form for two dimensional random variable, the variance is

$$V(XY) = E(XY)^2 - \{E(XY)\}^2$$

Covariance of two dimensional random variable

Let (X,Y) be two dimensional random variable then the expected value of the product of deviation of random variable X and Y from their respective expected values E(x) and E(Y) respectively is called covariance of random variables X and Y.

It is denoted by Cov(X,Y) and is given by $Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$

$$= E(XY) - E(X)\,E(Y)$$

If X and Y are independent random variables $E(XY) = E(X)\,E(Y)$ then

$Cov(X,Y) = E(XY) - E(X)\,E(Y)$

$= E(X)\,E(Y) - E(X)\,E(Y) = 0$

Scanned by CamScanner

Theorems on variance

i. $V(a) = 0$

ii. $V(aX + b) = a^2V(X)$

iii. $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab\ Cov(X, y)$

Correlation coefficient

Let (X, Y) be two dimensional random variable then correlation coefficient is the relationship between these two variables denoted by r and is given by

$$r = \frac{Cov\ (X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{E(XY) - E(X)\ E(Y)}{\sqrt{E(X)^2 - \{E(X)\}^2}\sqrt{E(Y^2) - \{E(Y)\}^2}}$$

R lies between -1 and +1.

Moments about origin of bivariate distribution

Let (X,Y) be two dimensional random variable then moments about origin of the bivariate distribution is denoted by μ'_{rs} and is given by

$\mu'_{rs} = E(X^r\ Y^S)$

$= \sum_x \sum_y x^r y^s P(x, y)$, for discrete random variable (x, y) having joint pmf P(x, y)

$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^r y^s\ f(x, y)\ dx\ dy$, for continuous random variable having joint pdf f(x,y)

$\mu'_{00} = E(X^0\ Y^0) = 1$

$\mu'_{10} = E(X^1\ Y^0) = E(X)$

$\mu'_{01} = E(X^0\ Y^1) = E(Y)$

$\mu'_{20} = E(X^2\ Y^0) = E(X^2)$

$\mu'_{02} = E(X^0\ Y^2) = E(Y^2)$

Moments about mean of bivariate distribution

Let (X, Y) be two dimensional random variable then moments about mean of the bivariate distribution is denoted by μ'_{rs} and is given by

$\mu'_{rs} = E[\{X - E(X)\}^r\{Y - E(Y)\}^S]$

$= \sum_x \sum_y \{x - E(X)\}^r\ \{y - E(Y)\}^S\ P(x, y)$, for discrete random variable (x, y) having joint pmf P(x,y)

$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{x - E(X)\}^r\ \{y - E(Y)\}^S\ f(x, y)\ dx\ dy$, for continuous random variable having joint pdf f(x,y)

$\mu_{00} = E[\{X - E(X)\}^0\ \{Y - E(Y)\}^0] = 1$

$\mu_{10} = E[\{X - E(X)\}^1\ \{Y - E(Y)\}^0] = 0$

$\mu_{01} = E[\{X - E(X)\}^0\ \{Y - E(Y)\}^1] = 0$

$\mu_{11} = E[\{X - E(X)\}^1\ \{Y - E(Y)\}^1] = Cov\ (X, Y)$

$\mu_{20} = E[\{X - E(X)\}^2\ \{Y - E(Y)\}^0] = V(X)$

Example 13: The joint probability distribution of X and Y is given in the following table

X \ Y	1	3	9
2	1/8	1/24	1/12
4	1/4	1/4	0
6	1/8	1/24	1/12

Find:(i) E(X) (ii)E(Y) (iii)E(XY) (iv) E(X+Y)(v) V(X) (vi)V(Y)
(vii)V(XY) (viii)Cov(X,Y) (ix) V(2X – 3Y) (x) r(X,Y)

Scanned by CamScanner

Solution

X \ Y	1	3	9	P(X = x)	xP(x)	$x^2P(x)$
2	1/8	1/24	1/12	6/24	12/24	24/24
4	1/4	1/4	0	2/4	8/4	32/4
6	1/8	1/24	1/12	6/24	36/24	216/24
P(Y=y)	4/8	8/24	2/12	1	ΣxP(x)=4	$\Sigma x^2P(x)=18$
yP(y)	4/8	24/24	18/12	Σy p(y)=3		
$y^2P(y)$	4/8	72/24	162/12	Σy^2 p(y)=17		

$E(X) = \sum_x x\,P(x) = 4$

$E(Y) = \sum_y y\,P(y) = 3$

$E(XY) = \sum_x \sum_y xyP(x, y)$

$$= 2 \times 1 \times \frac{1}{8} + 2 \times 3 \times \frac{1}{24} + 2 \times 9 \times \frac{1}{12} + 4 \times 1 \times \frac{1}{4} + 4 \times 3 \times \frac{1}{4} + 4 \times 9 \times 0 + 6 \times 1 \times \frac{1}{8} + 6 \times 3 \times \frac{1}{24} + 6 \times 9 \times \frac{1}{12} = \frac{2}{8} + \frac{6}{24} + \frac{18}{12} + \frac{4}{4} + \frac{12}{4} + 0 + \frac{6}{8} + \frac{18}{24} + \frac{54}{12} = \frac{286}{24} = 11.91$$

$E(X+Y) = \sum_x \sum_y (x + y)\,P(x, y)$

$$= (2 + 1) \times \frac{1}{8} + (2 + 3) \times \frac{1}{24} + (2 + 9) \times \frac{1}{12} + (4 + 1) \times \frac{1}{4} + (4 + 3) \times \frac{1}{4} + (4 + 9) \times 0 + (6 + 1) \times \frac{1}{8} + (6 + 3) \times \frac{1}{24} + (6 + 9) \times \frac{1}{12}$$

$$= \frac{3}{8} + \frac{5}{24} + \frac{11}{12} + \frac{5}{4} + \frac{7}{4} + 0 + \frac{7}{8} + \frac{9}{24} + \frac{15}{12}$$

$$= \frac{9 + 5 + 22 + 30 + 42 + 0 + 21 + 9 + 30}{24} = \frac{168}{24} = 7$$

$E(X^2) = \Sigma x^2P(x) = 18$

$E(Y^2) = \Sigma y^2P(y) = 17$

$V(X) = E(X^2) - \{E(X)\}^2 = 18 - 16 = 2$

$V(Y) = E(Y^2) - \{E(Y)\}^2 = 17 - 9 = 8$

$E(X^2Y^2) = \sum_x \sum_y x^2 y^2\,P(x, y)$

$$= 2^2 \times 1^2 \times \frac{1}{8} + 2^2 \times 3^2 \times \frac{1}{24} + 2^2 \times 9^2 \times \frac{1}{12} + 4^2 \times 1^2 \times \frac{1}{4} + 4^2 \times 3^2 \times \frac{1}{4} + 4^2 \times 9^2 \times 0 + 6^2 \times 1^2 \times \frac{1}{8} + 6^2 \times 3^2 \times \frac{1}{24} + 6^2 \times 9^2 \times \frac{1}{12}$$

$$= \frac{4}{8} + \frac{36}{24} + \frac{324}{12} + \frac{16}{4} + \frac{144}{4} + 0 + \frac{36}{8} + \frac{324}{24} + \frac{2916}{12} = \frac{7920}{24} = 330$$

$V(XY) = E(X^2Y^2) - \{E(XY)\}^2$

$= 330 - (11.91)^2 = 188.15$

$Cov(X,Y) = E(XY) - E(X)E(Y)$

$= 11.91 - 4 \times 3 = -0.09$

$V(2X-3Y) = 2^2\,V(X) + 3^2\,V(Y) - 2 \times 2 \times 3\,Cov(X, Y)$

$= 4 \times 2 + 9 \times 8 - 12 \times (-0.09)$

$$r(X,Y) = \frac{Cov\,(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{-0.09}{\sqrt{2}\sqrt{8}} = -0.022$$

Scanned by CamScanner

Example 14. A two dimensional random variable (X,Y) has pdf $f(x, y) = \begin{cases} 2x\, e^{-y}; & 0 \le x \le 1, y \ge 0 \\ 0; & \text{otherwise} \end{cases}$.

Find: (i) E(X) (ii) E(Y) (iii) E(4x + 1) (iv) E(3Y − 2) (v) E(2X+3Y) (vi) E(X+Y) (vii) V(X) (viii) V(3Y) (ix) V(2X+3) (x) V(4Y − 5) (xi) Cov(X, Y) (xii) r(X,Y) (xiii) V(2X+3Y)

Solution

Here,

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^{\infty} 2x\, e^{-y} dy = 2x \int_0^{\infty} e^{-y}\, y^{1-1} dy = 2x\,(1-1)! = 2x$$

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_0^1 x\, 2x\, dx = 2\int_0^1 x^2\, dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

$$f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^1 2x\, e^{-y}\, dx = 2\, e^{-y}\int_0^1 x\, dx = 2e^{-y}\left[\frac{x^2}{2}\right]_0^1 = e^{-y}$$

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\, dy = \int_0^{\infty} y e^{-y} dy = \int_0^{\infty} e^{-y}\, y^{2-1} dy = (2-1)! = 1$$

$$E(4X+1) = 4E(X) + 1 = 4 \times \frac{2}{3} + 1 = \frac{11}{3}$$

$$E(3Y-2) = 3E(Y) - 2 = 3 \times 1 - 2 = 1$$

$$E(2X+3Y) = 2E(X) + 3E(Y) = 2 \times \frac{2}{3} + 3 \times 1 = \frac{13}{3}$$

$$E(X+Y) = E(X) + E(Y) = \frac{2}{3} + 1 = \frac{5}{3}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_0^1 x^2\, 2x\, dx = 2\int_0^1 x^3\, dx = \frac{1}{2}$$

$$V(X) = E(X^2) - \{E(X)\}^2 = \frac{1}{2} - \left[\frac{2}{3}\right]^2 = \frac{1}{18} = 0.055$$

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y)\, dy = \int_0^1 y^2 e^{-y} dy = \int_0^{\infty} e^{-y}\, y^{3-1} dy = (3-1)! = 2$$

$$V(Y) = E(Y^2) - \{E(Y)\}^2 = 2 - 1^2 = 1$$

$$V(3Y) = 3^2 V(Y) = 9 \times 1 = 9$$

$$V(2X+3) = 2^2 V(X) = 4 \times \frac{1}{18} = \frac{2}{9}$$

$$V(4Y-5) = 4^2 V(Y) = 16 \times 1 = 16$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\, dx\, dy = \int_0^1\int_0^{\infty} xy\, 2xe^{-y}\, dx\, dy$$

$$= \int_0^1 2x^2\left\{\int_0^{\infty} e^{-y}\, y^{1-1}\, dy\right\} dx$$

$$= \int_0^1 2x^2\,(1-1)!\, dx = \int_0^1 2x^2\, dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{2}{3} - \frac{2}{3} \times 1 = 0$$

$$r(X,Y) = \frac{Cov\,(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{0}{\sqrt{0.055}\sqrt{1}} = 0$$

$$V(2X+3Y) = 4V(X) + 9V(Y) + 12\, Cov(X,Y) = 4 \times 0.055 + 9 \times 1 + 12 \times 0 = 9.22$$

Scanned by CamScanner

Discrete Distribution

Probability distribution of discrete random variable having probability mass function is called discrete probability distribution. Different discrete probability distributions are Bernoulli distribution, binomial distribution, poisson distribution etc.

Bernoulli Distribution

A trail having two mutually exclusive and exhaustive outcomes is called Bernoulli trial. These two outcomes are generally known as success and failure. A random variable X that takes values 0 and 1 with probability of failure q and probability of success p respectively is called Bernoulli variate and the distribution, it followed is called is Bernoulli distribution. Here $P(X=1) = p$ and $P(X=0) = q$.

Mean of Bernoulli distribution is $E(X) = 0.q+1.p = p$ and

Variance is given by $V(X) = E(X^2) - \{E(X)\}^2$

$$= \{0^2 \times q + 1^2 \times p\} = p^2$$

$$= p - p^2$$

$$= p(1-p) = pq$$

Binomial Distribution

This distribution can be used under the following conditions:

(i) The random experiment is performed repeatedly a finite and fixed number of times. In other words n, the number of trials, is finite and fixed.

(ii) The outcome of the random experiment (trial) results in the dichotomous classification of events. In other words, the outcome of each trial may be classified into two mutually disjoint categories, called success (the occurrence of the event) and failure (the non-occurrence of the event).

(iii) All the trials are independent, i.e., the result of any trial, is not affected in any way, by the preceding trials and doesn't affect the result of succeeding trials.

(iv) The probability of success (happening of an event) in any trial is p and is constant for each trial is $q=1-p$, is then termed as the probability of failure (non-occurrence of the event) and is constant.

Probability Mass Function of Binomial Distribution.

A discrete random variable following Binomial distribution is called Binomial variate. If X denotes the number of successes in n satisfying the above conditions, X is a random variable which can take the values 0, 1,2, ... n ;

The trials satisfying the above four conditions are also known as Bernoulli trials. For determining the probability associated with 0, 1, 2 ... n successes, the probability of x success is defined as follows.

$P(X = x) = p(x) = c(n, x)\, p^x q^{n-x}$

Here, x assumes the non-negative values associated with X.'s probability mass function is given by

$P(X=x) = p(x) = c(n, x)\, p^x q^{n-x}$

Here n and p are parameters of the distribution. A random variable x following Binomial distribution is denoted by $x \sim B(n, p)$.

Example 15: Probability that a student submit assignment before deadline is 0.7. What is probability that exactly 3 students out of 5 submit assignment before the deadline?

Solution

Let x = number of students submit assignment before deadline

Probability that student submit assignment before deadline (p) = 0.7 $\therefore$ $q = 1 - p = 0.3$, $n = 5$

Scanned by CamScanner

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

$P(x=3) = C(5,3)\,(0.7)^3\,(0.3)^2 = 10 \times 0.343 \times 0.09 = 0.308$

Hence probability that 3 students submit assignment before deadline is 0.308

Example 16: Ten unbiased coins are tossed simultaneously. Find the probability of obtaining

i) Exactly 6 heads ii) At least 8 heads iii) No heads

Solution

Let X = number of heads

Probability of getting head $(p) = \frac{1}{2}$ then $q = \frac{1}{2}$, $n = 10$

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

$$P(\text{exactly 6 heads}) = p(x = 6) = C(10, 6)\frac{1}{2}^{6}\frac{1}{2}^{10-6} = \frac{210}{1024} = \frac{105}{512} = 0.205$$

i) $P(\text{at least 8 heads}) = P(X \geq 8) = p(x = 8) + p(x = 9) + p(x = 10)$

$$= C(10, 8)\frac{1}{2}^{8}\frac{1}{2}^{10-8} + C(10, 9)\frac{1}{2}^{9}\frac{1}{2}^{10-9} + C(10, 10)\frac{1}{2}^{10}\frac{1}{2}^{10-10}$$

$$= \frac{45}{1024} + \frac{10}{1024} + \frac{1}{1024} = \frac{56}{1024} = \frac{7}{128} = 0.054$$

ii) $P(\text{no heads}) = P(x = 0) = C(10, 0)\frac{1}{2}^{10}\frac{1}{2}^{10-10} = \frac{1}{1024} = 0.0009.$

Hence probability of getting exactly six heads is 0.205

Probability of getting at least 8 heads is 0.054 and probability of getting no heads is 0.0009

Example 17. Out of 20 computers in a shop 6 are Dell make. A customer selects 5 computers to purchase then find the probability that among selected 2 are Dell make.

Solution

Let x = Number of Dell make computer.

Probability of Dell make computer $(p) = \frac{6}{20} = 0.333$ then $q = 1 - 0.333 = 0.667$, $n = 5$

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

$P(x=2) = c(n, x)\, p^x q^{n-x}$

$= c(5,2)\,(0.333)^2\,(0.667)^3$

$= 10 \times 0.11 \times 0.296 = 0.326$

Hence probability of Dell make computers is 0.326

Example 18: In a bombing attack there is 50% chance of any bomb hitting the target. If one hit is required to destroy the target, how many bombs should be dropped so that there is 99% or more chance of destroying the target?

Solution

Let x = number of bombs hitting the target. Also n be the number of bombs required to drop in order to destroy target

Probability of bomb hitting target (p)= 50% = 0.5 ∴ $Q = 1 - p = 1 - 0.5 = 0.5$

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

According to question

$P(x \geq 1) \geq 99\%$

or $[1 - P(X < 1)] \geq 0.99$

or $[1 - p(x = 0)] \geq 0.99$

Scanned by CamScanner

or $\quad 1 - \binom{n}{0}(0.5)^n \geq 0.99$

Then $n \geq 6.64 \approx 7$

Hence at least 7 bombs have to be dropped.

Example 19: Assume that the sex ratio is 1:1. What is the probability that a family with five children has (i)no girl (ii)at least one boy

Solution

Let x = number of boy

Probability that a children is boy $(p) = \frac{1}{1+1} = 0.5$

$$q = 1 - p = 1 - 0.5 = 0.5$$
$$n = 5$$

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

(i) Five children has no girl means all are boys

$P(x=5) = c(n,x))\, p^x q^{n-x} = c(5,5)(0.5)^5\,(0.5)^0 = 0.031$

ii) $P(x \geq 1) = 1 - P(x < 1)$

$= 1 - P(x=0)$

$= 1 - c(n,x)\, p^x q^{n-x} = 1 - c(5,0)(0.5)^0\,(0.5)^5$

$= 1 - (0.5)^5 = 1 - 0.031 = 0.968$

Example 20: About twenty percent of users do not close Windows properly. Suppose that Windows is installed in a computer laboratory of a campus that is used by random students in a random order.

a. What is the probability that exactly 8 of the next 10 student will close Windows properly?

b. What is the probability that at least 2 students will not close Windows properly out of 7 students?

Solution.

Let, x = number of students close windows properly

Probability that student close window properly (p) = 1 - 20% = 1 - 0.2 = 0.8 then q = 0.2

n = 10

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

a) $P(x=8) = c(n, x)\, p^x q^{n-x}$

$= C(10, 8)\,(0.8)^8\,(0.2)^2 = 45 \times 0.1677 \times 0.04 = 0.301$

Let y = number of students do not close windows properly

Probability that students do not close window properly (p) = 20% = 0.2 the q = 0.8, n = 7

b) $P(y \geq 2) = 1 - P(y < 2) = 1 - \{P(y = 0) + P(y = 1)\}$

$= 1 - \{C(7, 0)\,(0.2)^0\,(0.8)^7 + C(7, 1)\,(0.2)^1\,(0.8)^6\}$

$= 1 - (1 \times 0.209 + 7 \times 0.0524)$

$= 1 - 0.576 = 0.423$

Hence probability that out of 10 students 8 students close windows properly is 0.301

And probability of at least two students out of 7 students do not close windows is 0.423

Moments about origin (Raw moments) of Binomial distribution

Let x be a random variable following Binomial distribution with parameter n and p i.e. $x \sim B(n, p)$ then r^{th} moment about origin is denoted by μ_r' and is given by

$$\mu_r' = E[x^r] = \sum_{x=0}^{n} x^r\, p(x),\ r = 1, 2, 3, 4$$

Scanned by CamScanner

$= \sum_{x=0}^{n} x^r x^r c(n, x) p^x q^{n-x}$

When $r = 1$

$\mu_1' = np$ .... (i)

When $r = 2$

$\mu_2' = n(n-1)p^2 + np$ .... (ii)

When $r = 3$

$\mu_3' = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$ .... (iii)

When $r = 4$

$\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$ .... (iv)

Mean and Variance of Binomial Distribution

Let x be a random variable following Binomial distribution with parameter n and p i.e. $x \sim B(n, p)$ then

Mean = E(x) = np

Variance = V(x) = npq

Example 21: The mean and variance of Binomial distribution are 3 and 2 respectively. Find the number of cases out of 10000 repetitions of the experiments have less than or equal to 2 success.

Solution

Here mean $E(x) = np = 3$

Variance $V(x) = npq = 2$

Then

$$\frac{npq}{np} = \frac{2}{3} \Rightarrow q = \frac{2}{3} \Rightarrow p = 1 - \frac{2}{3} = \frac{1}{3} \text{ and } np = 3 \Rightarrow n = 9$$

Then $n \sim B\left(9, \frac{1}{3}\right)$ and N = 10,000

It's pmf is $p(x) = c(n, x)\, p^x q^{n-x}$

Then the expected Number of cases with number of success less than or equals to 2

$= N \times P(X \le 2)$

$= 10{,}000 \times \{P(0) + P(1) + P(2)\}$

$= 1000 \times \left\{ {}^9C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{9-0} + {}^9C_1 \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^{9-1} + {}^9C_2 \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^{9-2} \right\}$

$= 9586.5 \approx 9587$

Example 22. It is believed that 80% of Nepalese do not have health insurance .Suppose this is true and let x be the number of health insurance in a random sample of 12 Nepalese.

i) What is the probability of no health insurance?

ii) Find the mean and variance of x.

iii) Find P(x>2)

Solution

Since x = number of no health insurance

i) Probability of no health insurance (p) = 80% = 0.8

$q = 1 - p = 1 - 0.8 = 0.2$

Number of Nepalese people (n) = 12

Scanned by CamScanner

ii) Mean $E(x) = np = 12 \times 0.8 = 9.6$

Variance $V(x) = npq = 12 \times 0.8 \times 0.2 = 1.92$

iii) $P(x>2) = 1 - P(x \le 2) = 1 - \{P(x=0) + P(x=1) + P(x=2)\}$

$= 1 - \{c(12, 0)\ (0.8)^0\ (0.2)^{12} + c(12, 2)\ (0.8)^2\ (0.2)^{10}\}$

$= 1 - 0 \quad = 1$

Hence probability of no health insurance is 0.8

Mean is 9.6, variance is 1.92 and $P(X > 2)$ is 1

Example 23. If mean of a Binomial distribution is 3 and standard deviation is 2.Explain whether the information is correct or not.

Solution

Let $x \sim B(n, p)$

Then $E(x) = np = 3$ .... (i)

S.D. $= \sqrt{V(x)} = 2$

$V(x) = npq = 4$ .... (ii)

Substitute value of np in ii

$3q = 4$

or $q = \frac{4}{3}$

It is not possible. Because q lies between 0 and 1

Hence the given information is not correct.

Properties (Characteristics) of Binomial distribution

(i) It is a discrete distribution assuming nonnegative values of a random variable.

(ii) The parameters of Binomial distribution are n and p. so this distribution is also known as bi parametric i.e., having two parameters.

(iii) Mean = E(X) = np

(iv) Variance = V(X) = npq with maximum value of variance $= \frac{n}{4}$ when p = q.

(v) Mean ≥ Variance

(vi) Coefficient Skewness $= \beta_1 = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq}$ and $\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}}$

(vii) Coefficient of Kurtosis $= \beta_2 = 3 + \frac{1-6pq}{npq}$

(viii) The first four raw moments are

$\mu_1' = E(X) = np$

$\mu_2' = E(X^2) = n(n-1)\,p^2 + np$

$\mu_3' = E(X^3) = n(n-1)(n-2)\,p^3 + 3n(n-1)\,p^2 + np$

$\mu_4' = E(X^4) = n(n-1)(n-2)(n-3) + 6n(n-1)(n-2)\,p^3 + 7n(n-1)\,p^2 + np$

(ix) The first four central moments are

$\mu_1 = 0$

$\mu_2 = npq$

$\mu_3 = npq\,(q-p)$

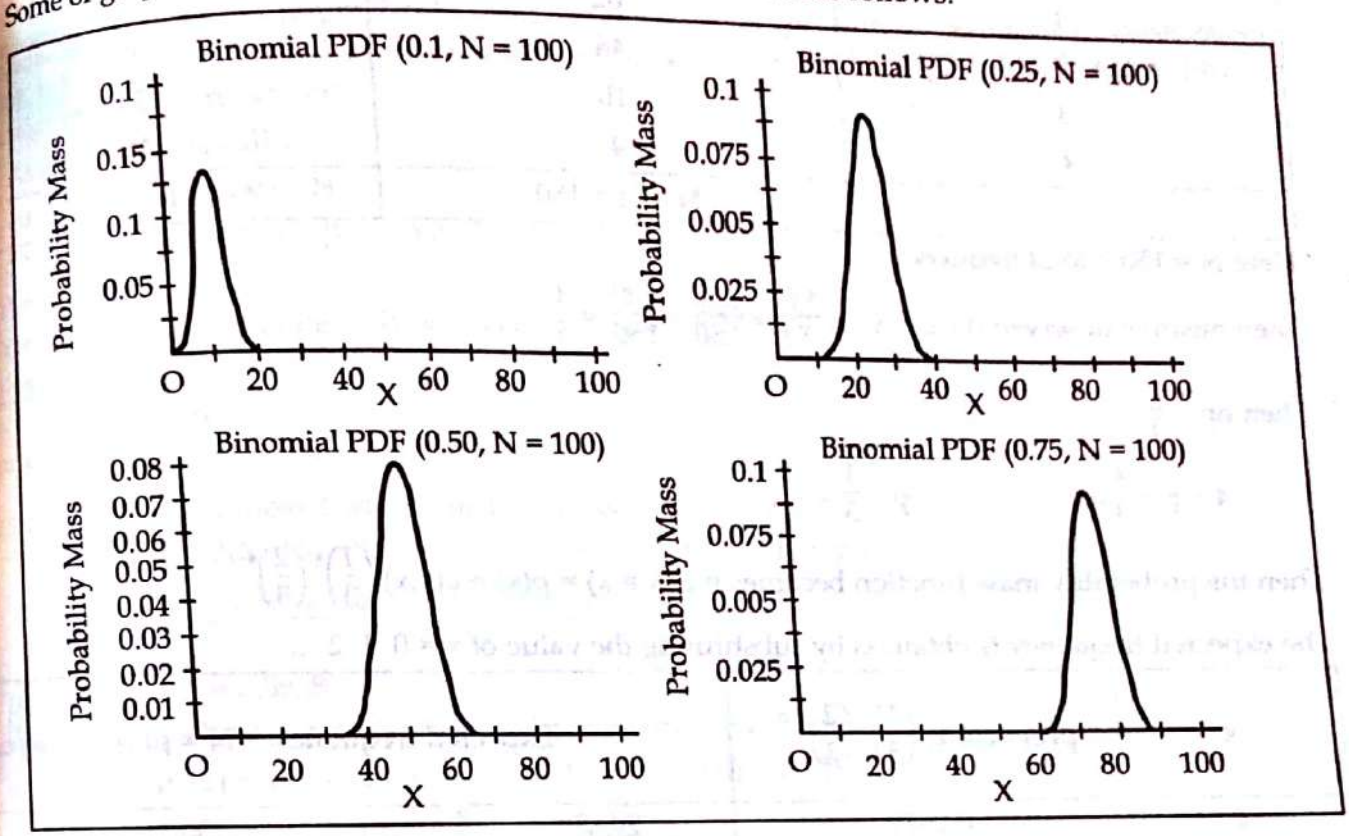
$\mu_4 = npq\,[1 + 3(n-2)\,pq]$

Scanned by CamScanner

(v) When probability of success are same for both binomial variate the sum of two binomial variate is also a binomial variate. i.e., $X_1 \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$ then $X_1 \pm X_2 \sim B(n_1 \pm n_2, p)$.

(vi) Binomial distribution is applicable if the trials are independent and have two mutually exclusive outcomes that can be taken as success and failure.

(vii) Some of graphical presentation of Binomial distribution are as follows.



Conditions for use of Binomial Distribution

i. The variable should be discrete.
ii. A dichotomy exists.
iii. Number of trials should be small and finite.
iv. Trials should be independent.
v. Trials must be repeated under identical conditions.

Fitting of Binomial Distribution

The random variable following Binomial distribution can be used to find the expected frequencies of each values of the random variable using relation $N \times P(X=x)$, where N is the total frequency of the observed data.

Since mean = np, the mean of observed data can be used to find the value of parameter p as follows

$$\bar{x} = np$$

Example 24: Fit the binomial distribution to the data given below

X	0	1	2	3	4
f	28	62	46	10	4

Scanned by CamScanner

Solution

Let X be the random variable following Binomial distribution with parameter n=4 and p.

x	f	fx
0	28	0
1	62	62
2	46	92
3	10	30
4	4	16
	N = $\sum f$ = 150	$\sum fx$ = 200

Here N = 150 = total frequency

Then mean of observed data = $\bar{x} = \frac{\sum fx}{\sum f} = \frac{200}{150} = \frac{200}{150} = \frac{4}{3}$

Then $np = \frac{4}{3}$

$4 \times p = \frac{4}{3}$ or $p = \frac{1}{3}$

Then the probability mass function becomes = $P(X = x) = p(x) = c(4,x)\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{4-x}$.

The expected frequency is obtained by substituting the value of x = 0, 1, 2 ...

x	$p(x) = c(4,x)\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{4-x}$	Expected frequency = N × p(x)	
0	0.1979	29.689	30
1	0.395	59.289	59
2	0.296	44.4	44
3	0.0985	14.78	15
4	0.0123	1.845	2
Total	1	150	150

Poisson Distribution

Poisson distribution as aLimiting Case of Binomial Distribution

Poisson distribution was derived in 1837 by a French mathematician Simeon D. Poisson. Poisson distribution may be obtained as a limiting case of Binomial probability distribution under following conditions:

i) n, the number of trials is indefinitely large i.e., $n \to \infty$.

ii) p, the constant probability of success for each trial is indefinitely small i.e., $p \to 0$.

iii) $np = \lambda$ (say) is finite.

Under the above three conditions the Binomial probability function tends to the probability function of the Poisson distribution;

The probability of x success in n independent trial is

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

Scanned by CamScanner

The discrete random variable X is said to follow Poisson distribution if its probability mass function is given by $P(X = x) = P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$; $X = 0, 1, 2, \ldots\ldots$

Here λ is the parameter of Poisson distribution and a random variable following Poisson distribution is denoted by $X \sim P(\lambda)$.

Example 25: Between the hours 2 P.M. and 4 P.M. the average number of phone calls per minute coming into the switch board of a software company is 2.35. Find the probability that during one particular minute there will be

i) At most 2 phone calls.

ii) At least 2 telephone calls

iii) Exactly one phone call. [Given $e^{-0.25} = 0.095374$]

Solution

Let X = number of telephone calls between 2 P.M. to 4 P.M. in a software company

$\lambda = 2.35$

Its pmf is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

The probability of at most 2 telephone calls in particular minute

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = p(0) + p(1) + p(2)$$

$$= \frac{e^{-2.35}\,2.35^0}{0!} + \frac{e^{-2.35}\,2.35^1}{1!} + \frac{e^{-2.35}\,2.35^2}{2!}$$

$$= 0.5828$$

The probability of at least 2 telephone calls in particular minute

$$= P(x \ge 12) = 1 - \{P(X = 0) + P(X = 1)\} = 1 - \{P(0) + P(2)\}$$

$$= 1 - \left\{\frac{e^{-2.35}\,2.35^0}{0!} + \frac{e^{-2.35}\,2.35^1}{1!}\right\}$$

$$= 1 - 0.5828 = 0.4172$$

The probability of one telephone calls in particular minute $= P(X = 1) = P(1) = \frac{e^{-2.35}\,2.35^1}{1!} = 0.2241$

Hence the probability of at most 2 telephones calls per minute in particular minute is 0.5828.

Probability of at least 2 telephones calls per minute in particular minute is 0.4172.

Probability of one telephones calls per minute in particular minute is 0.2241.

Example 26: Network breakdowns are unexpected rare events that occur every 3 weeks, on the average. Compute the probability of more than 4 breakdowns during a 21-week period.

Solution

Let x = number of network breakdown

Average no of network breakdown every 3 weeks = 1

Average number of network breakdown during 21 weeks $= \frac{1}{3} \times 21 = 7$

Hence, $\lambda = 7$

It's pmf is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$$P(x > 4) = 1 - P(\le 4) = 1 - \{P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)\}$$

$$= 1 - \left\{\frac{e^{-7}\,7^0}{0!} + \frac{e^{-7}\,7^1}{1!} + \frac{e^{-7}\,7^2}{2!} + \frac{e^{-7}\,7^3}{3!} + \frac{e^{-7}\,7^4}{4!}\right\}$$

Scanned by CamScanner

$$= 1 - e^{-7}\left\{1 + 7 + \frac{49}{2} + \frac{343}{6} + \frac{2403}{24}\right\}$$

$$= 1 - e^{-7} \times 189.791$$

$$= 1 - 0.1734 = 0.826$$

Hence probability of more than 4 network breakdown during 21 week is 0.826

Example 27: It is found that 5% of the keyboards produced by a computer manufacturing company is defective. They are sold in a box of 100 and guarantees that not more than 10 keyboards will be defective. What is the probability that the box selected at random fail to meet the guarantee?

Solution

Let x be the number of defective keyboards

Hints: total number of keyboards in a box (n) = 100

Probability of keyboard being defective (p) = 0.05

Then mean = λ = np = 100X0.05 = 5

Its pmf is $p(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, 3, ...

Probability of not meeting guarantee

$$= P(X > 10) = P(X > 10)$$

$$= 1 - P(X \le 10)$$

$$= 1 - \sum_{x=0}^{10} \frac{e^{-5}\,5^x}{x!}$$

$$= 1 - \left\{\frac{e^{-5}\,5^0}{0!} + \frac{e^{-5}\,5^1}{1!} + \frac{e^{-5}\,5^2}{2!} + \frac{e^{-5}\,5^3}{3!} + \frac{e^{-5}\,5^4}{4!} + \frac{e^{-5}\,5^5}{5!} + \frac{e^{-5}\,5^6}{6!} + \frac{e^{-5}\,5^7}{7!} + \frac{e^{-5}\,5^8}{8!} + \frac{e^{-5}\,5^9}{9!} + \frac{e^{-5}\,5^{10}}{10!}\right\}$$

$$= 1 - e^{-5}\left[1 + 5 + \frac{25}{2} + \frac{125}{6} + \frac{625}{24} + \frac{3125}{120} + \frac{15625}{720} + \frac{78125}{5040} + \frac{390625}{40320} + \frac{1953125}{362880} + \frac{9765625}{3628800}\right]$$

$$= 1 - 0.0067 \times 146.38$$

$$= 0.0136$$

Hence the probability that box fail to meet guarantee is 0.0136

Example 28: Microsoft Company producing Motherboard, there is a small chances of 1/500 being defective. The Motherboards are supplied in a packet of 10. Assuming Poisson distribution find the approximate number of packet containing no defective, one defective, two defective, three defective respectively in a consignment of 20000 packets.

Solution

Let x = number of defective motherboard

Total number of packets of motherboard (N) = 20000

Number of motherboard in a packet (n) = 10

Probability of a defective motherboard (p) = $\frac{1}{500}$

Then,

Average(λ) = np=10 × $\frac{1}{500}$ = 0.02

The pmf is $p(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$; x = 0, 1, 2, ...

The number packets containing x defective lenses is given by N × p(x)

Scanned by CamScanner

Then, number of packets with no defective = $20000 \times p(0) = 20000 \times \frac{e^{-0.02}\, 0.02^0}{0!} = 19604$

Number of packets with one defective = $20000 \times p(1) = 20000 \times \frac{e^{-0.02}\, 0.02^1}{1!} = 392$

Number of packets with two defective = $20000 \times p(2) = 20000 \times \frac{e^{-0.02}\, 0.02^2}{2!} = 4$

Number of packets with three defective = $20000 \times p(3) = 20000 \times \frac{e^{-0.02}\, 0.02^3}{3!} = 0.$

Example 29: Messages arrive at an electronic message center at random order, with an average of 10 messages per hour.

(a) What is the probability of receiving at least three messages during the next hour?

(b) What is the probability of receiving exactly five messages during the next three hours?

Solution

Let x = Number of message

Average number of message per hour (λ) = 10

It's pmf is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

a) Probability of getting at least three message in next hr

$$P(x \geq 3) = 1 - P(x < 3) = 1 - \{P(x=0) + P(x=1) + P(x=2)\}$$

$$= 1 - \left\{\frac{e^{-10}\, 10^0}{0!} + \frac{e^{-10}\, 10^1}{1!} + \frac{e^{-10}\, 10^2}{2!}\right\}$$

$$= 1 - e^{-10}\{1 + 10 + 50\}$$

$$= 1 - 61 \times 0.0000045 = 0.999$$

Average no of message per three hours (λ) = 30

Probability of getting exact 5 messages in next three hours

$$P(x = 5) = \frac{e^{-30}\, 30^5}{5!} = 0$$

Example 30: Telephone calls enter a college switchboard on the average of two every 3 minutes. If one assumes an approximate Poisson Process what is the probability that three or more calls arrive in a 9 minute interval.

Solution.

Let x = number of telephone calls enter a college switch board

Average number of calls per 3 minute = 2

Hence, average number of calls per 9 minute (λ) = 6

It's pmf is $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Probability of 3 or more calls in 9 minute interval is $P(x \geq 3) = 1 - P(x < 3)$

$$= 1 - \{P(x=0) + P(x=1) + P(x=2)\}$$

$$= 1 - \left\{\frac{e^{-6}\, 6^0}{0!} + \frac{e^{-6}\, 6^0}{1!} + \frac{e^{-6}\, 6^2}{2!}\right\}$$

$$= 1 - e^{-6}\{1 + 6 + 18\}$$

$$= 1 - 0.061 = 0.938$$

Hence probability of 3 or more call in 9 minutes interval is 0.938

Scanned by CamScanner

Moments about origin (Raw moments) of Poisson distribution

Let x be a random variable following Poisson distribution with parameter i.e. $x \sim P(\lambda)$ then r^{th} moment about origin is denoted by μ_r' and is given by

$$\mu_r' = E[x^r] = \sum_{x=0}^{\infty} x^r p(x), r = 1, 2, 3, 4.$$

$$= \sum_{x=0}^{\infty} x^r \frac{e^{-\lambda}\lambda^x}{x!}$$

When r = 1

$\mu_1' = \lambda$ when r = 2

$\mu_2' = \lambda^2 + \lambda$

When r = 3

$\mu_3' = \lambda^3 + 3\lambda^2 + \lambda$

When r= 4

$\mu_4' = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$

Mean and variance of Poisson distribution

Let x be a random variable following Poisson distribution with parameter i.e $x \sim P(\lambda)$ then

Mean = $E(x) = \mu_1' = \lambda$ from equation no (i)

Variance = $V(x) = \sigma = \mu_2 = \mu_2' - \mu_1'^2$

$= \lambda^2 + \lambda - \lambda^2$ from equation (i) and (ii)

$= \lambda$

Example 31. If a Poisson variate X is such that P(X=1)=2P(X=2). Find P(X=0), mean and variance.

Solution

Here, $x \sim P(\lambda)$

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

According to question,

P(X=1) = 2P(X=2)

or $\frac{e^{-\lambda}\lambda^1}{1!} = 2\frac{e^{-\lambda}\lambda^2}{2!} \Rightarrow \lambda = 1$

Now, $P(X=0) = \frac{e^{-1}1^0}{0!} = 0.367$

Mean $E(x) = \lambda = 1$

Variance $V(x) = \lambda = 1$

Some examples of Poisson distribution are as follows.

- Number death from a disease.
- Number of suicides in particular interval of time in particular locality.
- Number of misprints in a page of a book.
- Number of telephone calls in given interval of time.
- Number of accidents in certain section of road in given time interval etc.
- Number of cars passing through a certain street in given time interval
- Number of air accidents in given time interval.
- Number of defective materials in a packing materials.

Scanned by CamScanner

- Number of scores in a game.

Properties of Poisson distribution

i) Poisson distribution is discrete distribution.

ii) It has only one parameter λ, hence it is also known as uni-parametric.

iii) Mean = λ

iv) Variance = λ

v) Mean = variance

vi) For non-integer λ, it is the largest integer less than λ. For integerλ, $x = \lambda$ and $x = \lambda - 1$ are the two modes.

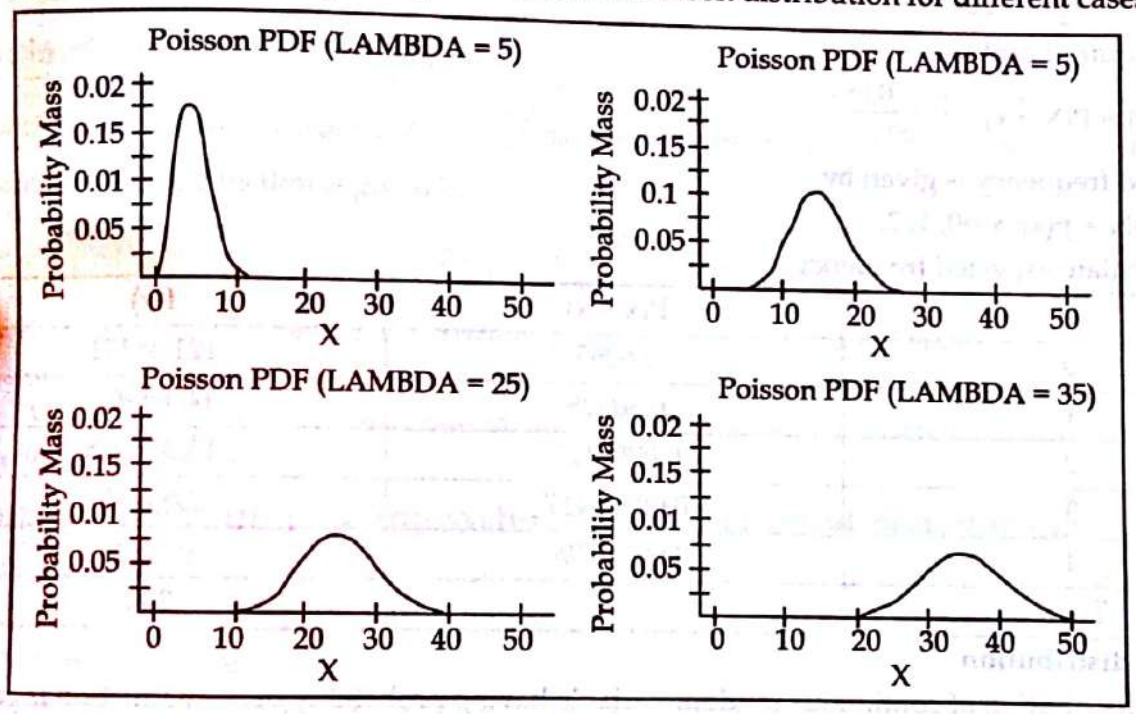
vii) Coefficient of Skewness$\beta_1 = 1/\lambda$, $\gamma_1 = 1/\sqrt{\lambda}$.

viii) Coefficient of kurtosis$\beta_2 = 1/\lambda$, $\gamma_2 = 1/\sqrt{\lambda}$.

ix) Sum of two Poisson variate is Poisson variate i.e. if $X \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$ then $X_1 \pm X_2 \sim P(\lambda_1 \pm \lambda_2)$.

x) It is used in the case of waiting time analysis where $np < 5$ and probability of success is very low i.e. $p \to 0$ and $n \to \infty$.

xi) Following are some graphical presentation of Poisson distribution for different cases of λ.



Conditions for the use of Poisson distribution:

i. The variable should be discrete.

ii. A dichotomy exists.

iii. Number of trials should be very large ($n \to \infty$)

iv. Probability of success is very small ($p \to 0$)

v. Trials should be independent.

Fitting of Poisson Distribution

The random variable following Poisson distribution can be used to find the expected frequencies of each values of the random variable using relation $N \times P(X = x)$, where N is the total frequency of the observed data.

Since mean = λ, the mean of observed data.

Scanned by CamScanner

Example 32: Fit a Poisson distribution.

X	0	1	2	3	4
f	123	59	14	3	1

Solution

Let X be the random variable following Poisson variable with parameter λ.

Then the probability mass function of random variable X is

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

We know, for Poisson distribution, mean$\lambda = \bar{x} = \frac{\sum fx}{\sum f}$

X	0	1	2	3	4	Total
f	123	59	14	3	1	200
fx	0	59	28	9	4	100

Then, mean $\bar{x} = \frac{\sum fx}{\sum f} = \frac{100}{200} = 0.5$

Then, the required pmf is

$$p(x) = P(X = x) = \frac{e^{-0.5}\, 0.5^x}{x!}$$

The expected frequency is given by

$$f(x) = N \times p(x); x = 0, 1, 2, \ldots.$$

Table to calculate expected frequency

X	P(x = x)	f(x)
0	0.6065	121.3≈121
1	0.30325	60.65≈61
2	0.0758125	15.325≈15
3	0.00505417	2.552≈3
4	0.00015794	0.32≈0
Total		200

Continuous distribution

Probability distribution of continuous random variable having probability density function is called continuous distribution. Different continuous distributions are Normal distribution, exponential distribution, gamma distribution etc.

Normal Distribution

Binomial distribution and Poissondistribution, are discrete probability distributions since the variables under study were discrete random variables. Normal probability distribution or commonly called the normal distribution is one of the most important continuous theoretical distributions in Statistics. Most of the data relating to economic and business statistics or even in social and physical sciences conform to this distribution.

The normal distribution was first discovered by English Mathematician Dc-Moivre in 1733 who obtained the mathematical equation for this distribution while dealing with problems arising in the game of chance. Normal distribution is also known as Gaussian distribution (Gaussian Law of Errors) after Karl Friedrich Gauss who used this distribution to describe the theory of accidental

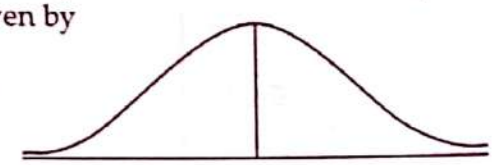
Scanned by CamScanner

errors of measurements involved in the calculation of orbits of heavenly bodies. Today, normal probability model one of the most important probability models in statistical analysis.

Normal Probability Curve

If X is a continuous random variable following normal probability distribution with mean μ and standard deviation σ, then its probability density function is given by

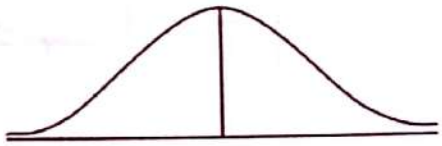
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad -\infty < x < \infty, \ -\infty < \mu < \infty, \ 0 < \sigma < \infty$$



Definition

A continuous random variable X is said to follow normal distribution if its probability density function is given by

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



With parameters μ and σ^2. It is written as $X \sim N(\mu, \sigma^2)$

Standard Normal Distribution

A continuous random variable $Z = \frac{x-\mu}{\sigma}$ is said to follow standard normal distribution if its probability density function is given by

$$\varnothing(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2}; \quad -\infty < z < \infty$$

It is written as $Z \sim N(0, 1)$ and Z is known as standard normal variate (SNV).

Its mean is 0 and variance is 1.

Note: While using normal distribution for approximation of discrete distribution continuity correction of 0.5 should be used.

Normal distribution as approximation of binomial distribution

Normal distribution is limiting case of Binomial distribution under following conditions

i) n, the number of trials are indefinitely large, i.e., $n \to \infty$.

ii) Neither p nor q are small

We know, mean = E(X) =np and variance = Var (X) = npq

Then, $Z = \frac{X - E(X)}{\sigma} = \frac{x - np}{\sqrt{npq}}$, tends to normal distribution as n tends to infinity. i.e., Z is standard normal variate.

Normal distribution as approximation of Poisson distribution

Normal distribution is limiting case of Poisson distribution under following conditions

N, the number of trials are indefinitely large, i.e., $n \to \infty$.

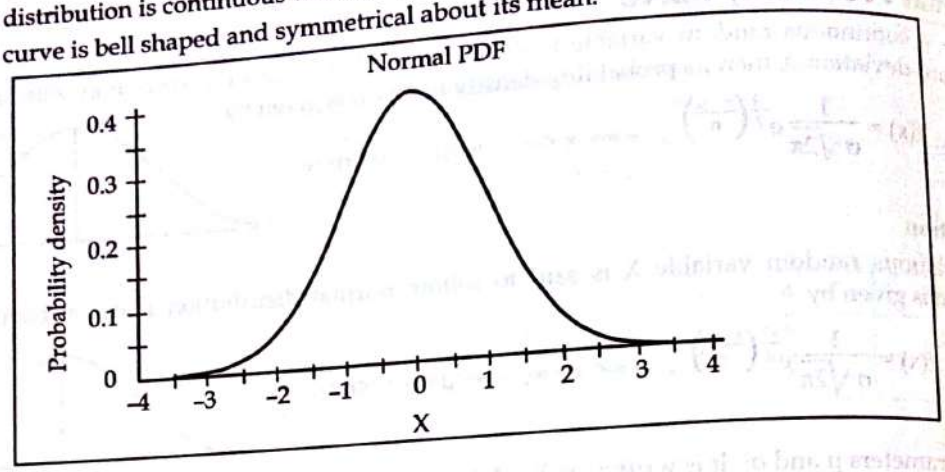
We know, mean = E(X) =λ and variance = Var (X) = λ

Then, $Z = \frac{X - E(X)}{\sigma} = \frac{x - \lambda}{\sqrt{\lambda}}$ tends to normal distribution as n tends to infinity. i.e., Z is standard normal variate.

Scanned by CamScanner

Properties of Normal Distribution are as follows:

1. The distribution is continuous with two parameters μ and σ.
2. The curve is bell shaped and symmetrical about its mean.



3. Mean =median = mode = μ.
4. The distribution is unimodal and mode = μ.
5. Maximum amplitude of curve is $\frac{1}{(\sigma\sqrt{2\pi})}$ and occurs at $x=\mu$.
6. Coefficient of skewness = $\beta_1 = 0, \gamma_1 = 0$.
7. Coefficient of Kurtosis = $\beta_2 = 3, \gamma_2 = 0$.
8. Quartile deviation is $\frac{2}{3}\sigma$.
9. Mean deviation is $\sqrt{\frac{2}{\pi}}\sigma \simeq \sigma$.
10. Q.D.:M.D.:S.D. = 10: 12: 15.
11. The curve is asymptotic to x axis.
12. Odd ordered moments are zero i.e. $\mu_{2n+1} = 0$.
13. Even ordered moments are given by relation $\mu_{2n} = (2n-1)\sigma^2\mu_{2n-2}$.
14. The theoretical range is $-\infty$ to ∞but the workable right is -3σ to 3σ.
15. The linear combination of independent normal variate is also a normal variate.
16. The point of inflexion of normal curve is $\mu \pm \sigma$.
17. The area under normal curve is unity.
18. $P(\mu - \sigma \le X \le \mu + \sigma) = 0.6826 \Rightarrow P(|Z| \le 1) = 0.6826$

    $P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9544 \Rightarrow P(|Z| \le 2) = 0.9544$

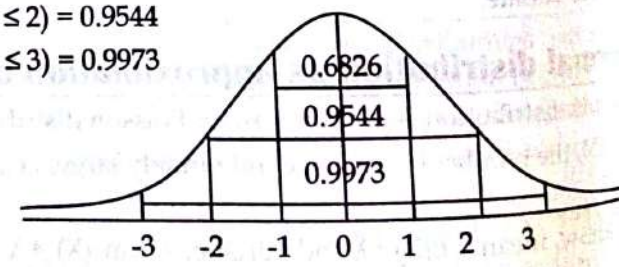
    $P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.9973 \Rightarrow P(|Z| \le 3) = 0.9973$

    $P(|Z| \le 1.645) = 0.90$

    $P(|Z| \le 1.96) = 0.95$

    $P(|Z| \le 2.33) = 0.98$

    $P(|Z| \le 2.58) = 0.99$



Importance of normal distribution

Normal distribution has great importance in statistics. Some of them are as follows:

Scanned by CamScanner

1. Normal distribution is the basis for entire large sample test.
2. It has is used widely in every part of research related to different branches of science.
3. Most of the discrete distribution of probability like binomial; Poisson, Negative binomial, hyper geometric etc. tends to normal for large sample size.
4. Many sampling distributions like chi square, student's t, F etc. tend to normal for large sample.
5. If the variable is not normal it can be transformed to normal by using central limit theorem which is fundamental theorem of statistics.
6. It is used in statistical quality control in industries.
7. The whole theory of exact test i.e. small sample test is based on the assumption that the parent population is normal.

Mean and variance of normal distribution

Let X be a random variable following Normal distribution with parameters μ and σ^2, then pdf is given by

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

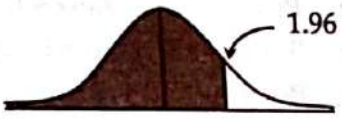
Mean = E(x) = μ

Variance = V(x) = σ^2

Example 33: Given a standard Normal distribution , determine the following probabilities: (i) P(Z<1.96) (ii) P(Z>1.64) (iii) P(0<Z<2.34) (iv) P(Z<-1.64) (v) P(Z>-0.34)(vi) P(-1.25<Z<0) (vii) P(-1.64<Z<1.64) (viii) P(0.17<Z<1.64) (ix) Probability that Z is less than – 0.84 or greater than 2.08 (x) What is the value of Z if and only if 31.87% of all possible Z values are smaller.

Solution

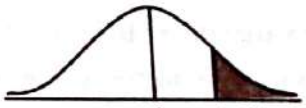
Here, Z is a standard normal variate.



(i) P(Z < 1.96)

= P(-∞< Z < 0) + P(0 < Z < 1.96)

= 0.50 + 0.475 = 0.975

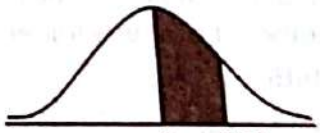


(ii) P(Z > 1.64)

= P(1.64 < Z <∞)

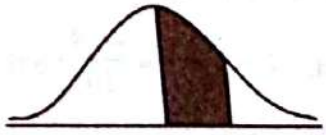
= P(0 < Z <∞) – P(0 < Z < 1.64)

= 0.50-0.4495 = 0.0505



(iii) P(0<Z<2.34)

= 0.4904

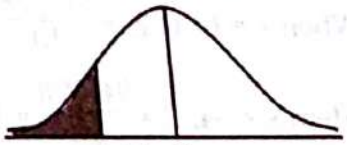


(iv) P(Z < -1.64)

= P(-∞< Z < 0) - P(-1.64 < Z < 0)

= 0.50 – P(0 < Z < 1.64)

= 0.50-0.4495 = 0.0505



(v) P(Z> -0.34)

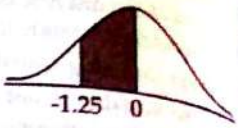
= P(-0.34<Z<0) + P(0<Z<∞)

= P(0<Z<0.34)+0.50 [By symmetry]

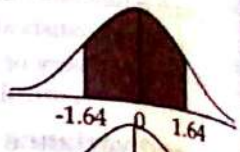
= 0.1331 + 0.50 = 0.6331

Scanned by CamScanner

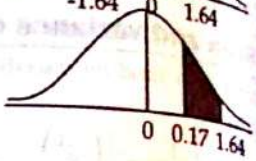
(vi) $P(-1.25 < Z < 0)$
$= P(0 < Z < 1.25)$ [By symmetry]
$= 0.3944$



(vii) $P(-1.64 < Z < +1.64) = P(-1.64 < Z < 0) + P(0 < Z < 1.64)$
$= P(0 < Z < 1.64) + P(0 < Z < 1.64)$
$= 2 \times P(0<Z<1.64)$
$= 2 \times 0.4495$
$= 0.899$

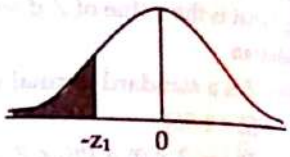


(viii) $P(0.17 < Z < 1.64)$
$= P(0 < Z < 1.64) - P(0 < Z < 0.17)$
$= 0.4495 - 0.0675 = 0.382$



(ix) The probability of Z is less than -0.84 or greater than +2.08 is
$= P(Z < -0.84) + P(Z > 2.08)$
$= [p(-\infty < Z < 0) - P(-0.84 < Z < 0)] + [p(o < z < \infty) - P(0 < Z < 2.08)]$
$= [0.50 - P(0 < Z < 0.84)] + [0.50 - 0.4812]$
$= 0.50 - 0.2995 + 0.50 - 0.4812 = 0.2193$

(x) Let the value of Z be $-Z_1$ such that only 31.87% of all values of Z are smaller than $-Z_1$, i.e.,
$P(Z < -Z_1) = 0.3187$
or, $P(-\infty<Z<0) - P(-Z_1<Z<0) = 0.3187$
or, $0.50 - P(0<Z<z_1) = 0.3187$
or, $P(0 < Z < Z_1) = 0.50 - 0.3187$
or, $P(0 < Z < Z_1) = 0.1813$



The value of probability closer to 0.1813 in Z table is 0.1808 at $Z = 0.47$. $Z_1 = 0.47$
Hence the required value of Z is -0.47.

Example 34 Number of files in folders of a computer programmer is normally distributed with average number of files in a folder is 78 with standard deviation of 10. Find the probability that number of files in a folder lies between 61 and 94.

Solution

Here $X \sim (2)$
$\Rightarrow X \sim (78, 100)$

Let, $Z = \frac{x - \mu}{\sigma} = \frac{-78}{10}$ be the standard normal variate then

$P(61<x<94) = ?$

When $x = 61.6$, $Z = \frac{61 - 78}{10} = -1.7$

When $x = 94$, $Z = \frac{94 - 78}{10} = 1.6$

Then,

$P(63<x<93) = P(-1.7<Z<1.6)$
$= P(-1.7<Z<0) + P(0<Z<1.6)$
$= P(0<Z< 1.7) + P(0<Z<1.6) = 0.4554 + 0.4452 = 0.9$

Scanned by CamScanner

Example 35 From a sample of 800 keyboards mean life is found to be 66 months with standard deviation of 5 months, find the probability that number of keyboards having life

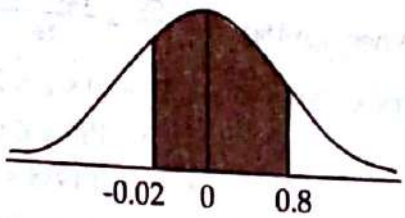
i) Between 65 and 70

ii) Greater than or equal to 72.

Solution

Let X = Number of keyboards, be random variable following normal distribution with parameters Mean = μ = 66 and standard deviation = σ = 5. i.e. $X \sim N(66, 25)$.

Let $Z = \frac{x-\mu}{\sigma} = \frac{-66}{5}$ be the standard normal variate.



i) P(between 65 and 70)

$= P(65<X<70)$

When x=65, $z = \frac{(65-66)}{5} = -0.2$

$x = 70, z = \frac{(70-66)}{5} = 0.8$

Then $P(65<X<70) = P(-0.02< Z < 0.8)$

$= P(-0.2< Z < 0) + P(0 < Z < 0.8)$

$= P(0<Z< 0.2) + P(0 <Z < 0.8)$

$= 0.0793 + 0.2881$ (from normal table)

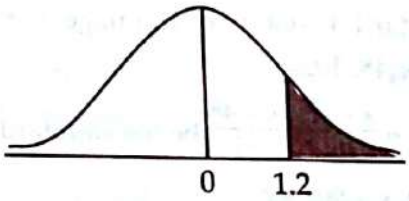
$= 0.3674$

The number of keyboards having life between 65 and 70 months = 800 × 0.3674 = 293.92 ≈294

ii) P(greater than or equal to 72)

$= P(X \geq 72)$

When X = 72, $z = \frac{(72-66)}{5} = 1.2$



$P(X \geq 72) = P(Z \geq 1.2)$

$= 0.5 - P(0 \leq Z \leq 1.2)$

$= 0.5 - 0.3849$

$= 0.1151$

The number of keyboards having life greater than or equal to 72 months = 800× 0.1151

= 92.08≈ 92.

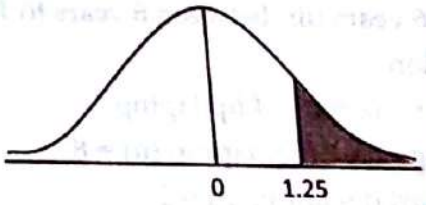
Example 36 The number of emails sent per week from email address of employees of a software company are normally distributed with mean 80 and standard deviation 16. (i) What is the probability of sending at least 100 emails during a week? (ii) What is probability of sending at most 120 emails during a week?

Solution

Let x be the random variable denoting the life of electric equipment. X follows normal distribution with mean 80 and standard deviation 16.

Let $Z = \frac{x-\mu}{\sigma} = \frac{x-80}{16}$

When x = 100, $Z = \frac{100-80}{16}$



i. The probability of sending at least 100 emails

$P(X \geq 100)$

Scanned by CamScanner

$= P(Z \geq 1.25)$

$= 0.5 - P(0 \leq Z \leq 1.25)$

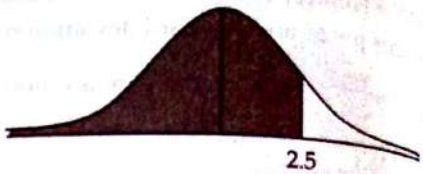
$= 0.5 - 0.3944$

$= 0.1056$

ii. Probability of sending atmost 120 emails

$P(X \leq 180)$

When $x = 180$, $Z = \frac{-80}{16} = \frac{120-80}{16} = 2.5$



$P(X \leq 120) = (X \leq 120) = (0 \leq Z \leq 2.50)$

$= 0.5 + P(0 \leq Z \leq 2.50)$

$= 0.5 + 0.4938 = 0.9938$

Hence probability of sending atleast 100 emails during a week is 0.1056

Probability of sending atmost 120 emails during a week is 0.9938

Example 37: An average scanned image occupies 0.6 megabytes of memory with a standard deviation of 0.4 megabytes. If you plan to publish 80 images on your web site, what is the probability that their total size is between 47 megabytes and 50 megabytes?

Solution

Let x = memory occupied by scanned image

Average memory occupied by an image = 0.6

Standard deviation of memory occupied by image = 0.4

Number of image = 80

Average memory of all image (μ) = $0.6 \times 80 = 4.8$

Standard deviation of all image(σ) = $0.4 \times 80 = 32$

$X \sim N(48, 1024)$.

Let $Z = \frac{x-\mu}{\sigma} = \frac{x-48}{32}$ be the standard normal variate.

$P(47 \leq x \leq 50) = ?$

When $x = 47$, $Z = \frac{47-48}{32} = -0.031$

When $x = 50$, $Z = \frac{50-48}{32} = 0.062$

$P(47 \leq x \leq 50) = P(-0.031 \leq Z \leq 0.062)$

$= P(-0.031 \leq z \leq 0) + P(0 \leq z \leq 0.062)$

$= P(0 \leq z \leq 0.031) + P(0 \leq z \leq 0.062)$

$= 0.012 + 0.023$

$= 0.035$

Example 38: Life time of hp laptop is normally distributed with mean life of 8 years with standard deviation of 2 years. What is probability that a hp laptop works for (i) more than 10 years (ii) less than 6 years (iii) between 6 years to 11 years

Solution

Let X = life time of hp laptop

Average life of hp laptop (μ) = 8

Standard deviation (σ) = 2

$X \sim N(8, 4)$

Scanned by CamScanner

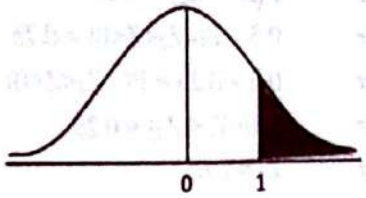
(i) $P(X \geq 10) = ?$

(ii) $P(X \leq 6) = ?$

(iii) $P(6 \leq X \leq 11) = ?$

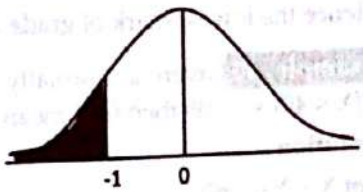
Let, $Z = \frac{x-\mu}{\sigma} = \frac{x-8}{2}$ be the standard normal variate.

When $x = 10$, $Z = \frac{10-8}{2} = 1$



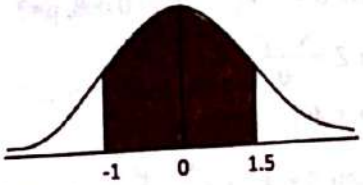
(i) $P(X \geq 10) = P(Z \geq 1)$
$= 0.5 - P(0 \leq Z \leq 1)$
$= 0.5 - 0.3413$
$= 0.158$

When $x = 6$, $Z = \frac{6-8}{2} = -1$



(ii) $P(X \leq 6) = P(Z \leq 6) = P(Z \leq -1)$
$= 0.5 - P(-1 \leq Z \leq 0)$
$= 0.5 - P(0 \leq Z \leq 1)$
$= 0.5 - 0.3413 = 0.158$

When $X = 11$, $Z = Z = \frac{11-8}{2} = 1.5$



(iii) $P(6 \leq X \leq 11) = P(-1 \leq Z \leq 1.5)$
$= P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 1.5)$
$= P(0 \leq Z \leq 1) + P(0 \leq Z \leq 1.5)$
$= 0.3413 + 0.4332 = 0.7745$

Example 39: Marks secured by students of a college in board exam is approximately normally distributed with mean 75 and standard deviation 5. If the top 5% of the students get grade A and the bottom 25% get grade F, what mark is the lowest A and what is the highest F?

Solution

Let x denotes the marks obtained by the students. X follows normal distribution with mean 75 and standard deviation 5. i.e. $X \sim N(75, 25)$

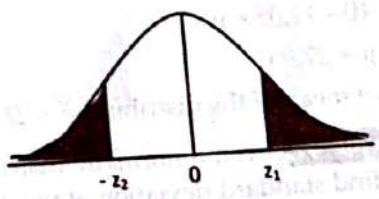
Let x_1 and x_2 be the lowest of grade A and highest of grade F respectively. Then we have,

$P(X > x_1) = 0.05$ and $P(X < x_2) = 0.25$

Let, $Z = \frac{x-\mu}{\sigma}$ be the standard normal variate.

When $X = x_1$, $Z = \frac{x_1 - 75}{5} = Z_1$ ...(i)

and $X = x_2$, $Z = \frac{x_2 - 75}{5} = -Z_2$ .....(ii)



Here, $P(X > x_1) = 0.05$

or $P(Z < Z_1) = 0.05$

or $0.5 - P(0 < Z < Z_1) = 0.05$

or $0.5 - 0.05 = P(0 < Z < Z_1)$

or $P(0 < Z < Z_1) = 0.45$

or $Z_1 = 1.65$

Scanned by CamScanner

Substitute value of Z_1 in equation (i) $\frac{x_1 - 75}{5} = 1.645$

$\Rightarrow$ $x_1 = 1.65 \times 5 + 75 = 83.25$

Also, $P(X<x_2) = 0.25$

or $P(Z< -Z_2) = 0.25$

or $0.5 - P(-Z_2<Z<0) = 0.25$

or $0.5 - 0.25 = P(- Z_2<Z<0)$

or $P(0<Z < Z_2) = 0.25$

or $Z_2 = 0.65$

Substitute value of Z_2 in equation (ii) $\frac{x_2 - 75}{5} = - 0.65$

$\Rightarrow$ $x_2 = - 5 \times 0.65 + 75 = 71.75$

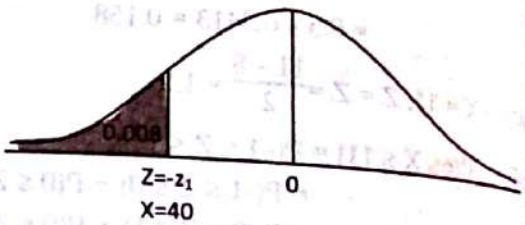
Hence the lowest mark of grade A is 83.25 and highest mark of grade F is 71.75.

Example 40: Given a normally distributed random variable X with standard deviation 5 and $P(X < 40) = 0.008$ then find mean of the distribution.

Solution

Let $X \sim N(\mu, \sigma^2)$

Here $\sigma = 5$, $P(X < 40) = 0.008$, $\mu = ?$



Let $Z = \frac{x - \mu}{\sigma}$

$P(X < 40) = 0.008$

When $X = 40$, $z = \frac{40 - \mu}{5} = - z_1$ (say)

Hence,

$P(X< 40) = 0.008$

or $P(Z<- z_1) = 0.008$

or $0.5 - P(-z_1< z < 0) = 0.008$

or $0.5 - P(0<Z < z_1) = 0.008$

or $0.5 - 0.008 = P(0<Z <z_1)$

or $P(0 < Z < z_1) = 0.492$

From normal table, $z_1 = 2.41$

Substitute $z_1 \frac{40 - \mu}{5} = - 2.41$

$\Rightarrow$ $40 - \mu = 12.05$

or $40 - 12.05 = \mu$

or $\mu = 27.95$

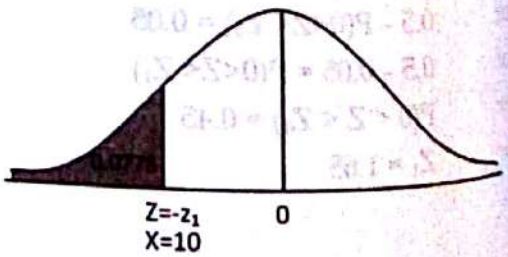
Hence mean of the distribution = 27.95

Example 41: Given a normally distributed random variable X with mean 25 and $P(X< 10) = 0.0778$ then find standard deviation of the distribution.

Solution

Let $X \sim N (\mu, \sigma^2)$

Here $\mu = 25$, $P(X < 0) = 0.0778$, $\sigma = ?$



Scanned by CamScanner

Let $Z = \frac{X-\mu}{\sigma}$ be a standard normal variate

$P(X < 10) = 0.0778$

When $X = 10$, $z = \frac{10-25}{\sigma} = \frac{-15}{\sigma} = -z_1$ (say)

Hence,

$P(X < 10) = 0.0778$

or $P(0 < Z < -z_1)$

or $0.5 - P(-z_1 < z < 0) = 0.0778$

or $0.5 - P(0 < Z < z_1) = 0.0778$

or $0.5 - 0.0778 = P(0 < Z < z_1)$

or $P(0 < Z < z_1) = 0.422$

From normal table, $z_1 = 1.42$

Substitute z_1

$\frac{-15}{\sigma} = -1.42$

or, $\sigma = \frac{15}{1.42} = 10.56$

Hence, standard deviation of the distribution = 10.56.

Example 42: In the normal distribution, 31%of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution.

Solution

Let X follows normal distribution with mean μ and standard deviation σ. i.e. $X \sim N(\mu, \sigma^2)$.

$P(X>64) = 0.08$ and $P(X<45) = 0.31$

Let $Z = \frac{X-\mu}{\sigma}$ be the standard normal variate.

When $X=64$, $Z = \frac{64-\mu}{\sigma} = z_1$ ...(i)

and $X=45$, $Z = \frac{45-\mu}{\sigma} = -z_2$ ...(ii)

$P(X>64) = 0.08$

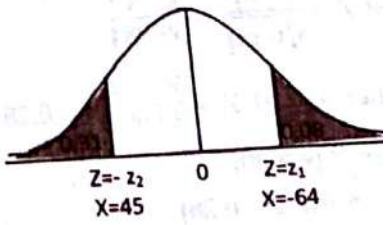
or $P(Z > Z_1) = 0.08$

or $0.5 - P(0<Z< Z_1) = 0.08$

or $0.5 - 0.08 = P(0<Z< Z_1)$

or $P(0<Z< Z_1) = 0.42$

or $Z_1 = 1.41$



$P(X<45) = 0.31$

or $P(Z< Z_2) = 0.31$

or $0.5 - P(-Z_2< Z < 0) = 0.31$

or $0.5 - 0.31 = P(-Z_2< Z < 0)$

or $0.19 = P(0 <Z< Z_2)$

or $Z_2 = 0.5$

Substitute value of Z_1 and Z_2 in equation (i) and (iii) respectively

Scanned by CamScanner

$$\frac{64-\mu}{\sigma} = 1.41 \quad \text{.... (iii)}$$

$$\frac{45-\mu}{\sigma} = -0.5 \quad \text{.... (iv)}$$

Divide (iv) by (iii)

$$\frac{45-\mu}{64-\mu} = \frac{-0.5}{1.41}$$

or $45 - \mu = -0.354\,(64 - \mu)$

or $45 + 22.69 = 0.354\mu + \mu$

or $\mu = \frac{67.69}{1.354} = 49.9$

Substitute μ in (iii)

$$\sigma = \frac{64-\mu}{1.41} = \frac{64-49.90}{1.41} = 10$$

The required mean is 49.9 and standard deviation is 10.

Fitting of Normal Distribution

In normal distribution pdf if $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty.$

Find standard normal variate corresponding to lower limit of each class intervals using relation $Z_i = \frac{x_i - \mu}{\sigma}$, where x_i' is lower limit of ith class interval.The area to the left of ordinate are obtained from table. Finally area for successive intervals are obtained by $\Delta\psi(Z) = \psi(Z_{i+1}) - \psi(Z_i)$. The expected frequencies are obtained by multiplying the area by N.

Example 43: A computer has 2000 files in a folder. The probability of their use is 0.02 independently is a day. Find the probability that more than 30 files are used in a day.

Solution. Let x be the random variable denoting no. of files used. X follows binomial distribution with n = 2000 and p = 0.02. Then

$$\text{Mean} = n\,p = 2000 \times 0.02 = 40$$

and standard deviation $= \sqrt{n\,p\,q} = \sqrt{2000 \times 0.02 \times (1 - 0.02)} = 39.984$

Applying central limit theorem.

Without continuity correction, $P\{x \geq 30\} = ?$

Let $Z = \frac{x - np}{\sqrt{n\,pq}} = \frac{x - 40}{39.984}$

When $x = 30$, $Z = \frac{30 - 40}{39.984} = -0.25$

Then $P\{x \geq 30\}$

$= P\{Z \geq -0.25\}$

$= 0.598745$

Note: In such case continuity correction can be applied to calculate probability with interval $[30.5, \infty]$ and probability

$P\{X \geq 30.5\}$

$= P\{Z \geq -0.2375\}$

$= 0.593902$

Scanned by CamScanner

The question is which one is correction.
1st case has to be considered incorrect.
2nd case has to be used because the range [29.5, ∞] and [30, ∞] are providing other result as seen above.

When a continuous distribution is approximated by another continuous distribution, continuity correction is not necessary but when discrete distribution is used for continuous probability, continuity correction is necessary.

Example 44: Fit normal distribution to the following data and find expected frequencies.

Class	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 – 70	70 – 80	80 - 90
frequency	3	15	37	52	61	30	14	8

Solution

Class	Frequency(f)	Mid value(x)	fx	fx^2
10 – 20	3	15	45	675
20 – 30	15	25	375	9375
30 – 40	37	35	1295	69825
40 – 50	52	45	2340	105300
50 – 60	61	55	3355	184525
60 – 70	30	65	1950	126750
70 – 80	14	75	1050	78750
80 – 90	8	85	680	57800
	$N = \sum f = 220$		$\sum fx = 11090$	$\sum fx^2 = 633000$

Now, $\mu = \bar{x} = \frac{\sum fx}{N} = \frac{11090}{220} = 50.409$

$$\sigma = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} = \sqrt{\frac{633000}{220} - \left(\frac{11090}{220}\right)^2} = 18.335$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{18.335\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-50.409}{18.335}\right)^2}$$

Class	Lower limit (X_L)	$Z_i = \frac{x_i' - \mu}{\sigma}$	$\Psi(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{\frac{-z^2}{2}} dz$	$\Delta\Psi(Z) = \Psi(Z_{i+1}) - \Psi(Z_i)$	Expected frequency $= N\,\Delta\Psi(Z)$
Below 10	$-\infty$	$-\infty$	0	0.0139	3.058≈ 3
10 – 20	10	-2.203	0.0139	0.0356	7.83≈ 8
20 – 30	20	-1.658	0.0495	0.114	25.08≈25
30 – 40	30	-1.113	0.1635	0.1242	27.32≈27
40 – 50	40	-0.567	0.2877	0.2003	44.06≈44
50 – 60	50	-0.022	0.488	0.2139	47.05≈47
60 – 70	60	0.523	0.7019	0.1535	33.77≈ 34
70 – 80	70	1.068	0.8554	0.0909	19.99≈ 20
80 – 90	80	1.613	0.9463	0.0379	8.33≈ 8
90 & over	90	2.159	0.9842		

Scanned by CamScanner

The Exponential Distribution

A continuous random variable X assuming non negative values is said to follow an exponential distribution with parameter $\theta > 0$, if its probability density function is given by

$$f(x) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

A random variable following exponential distribution with parameter θ is denoted by $X \sim \exp(\theta)$

Moments of exponential distribution

The moments about origin of the exponential distribution is denoted by μ_r' and is given by

$$\mu_r' = \int_0^\infty x^r f(x)\, dx = \int_0^\infty x^r \theta e^{-x}\, dx$$

When r = 1

$$\mu_1' = \frac{1}{\theta} = \frac{1!}{\theta}$$

When r = 2

$$\mu_2' = \frac{2}{\theta^2} = \frac{2!}{\theta^2}$$

When r = 3

$$\mu_3' = \frac{6}{\theta^3} = \frac{3!}{\theta^3}$$

When r = 4'

$$\mu_4' = \frac{24}{\theta^3} = \frac{4!}{\theta^4}$$

Hence $\mu_r' = \frac{r!}{\theta^r}$

Mean and variance of exponential distribution

Mean of the distribution is

$$E(x) = \mu_1' = \frac{1}{\theta}$$

Variance of the distribution is

$$V(x) = \mu_2' - \mu_1'^2 = \frac{1}{\theta^2}$$

Properties

1) Mean $= \mu_1' = \frac{1}{\theta}$

2) Variance $= \mu_2 = \frac{1}{\theta^2}$.

3) The rth raw moment is given by $\mu_r' = \frac{r!}{\theta^r}$; r = 0, 1, 2, ...

4) If $X_1, X_2, \ldots, X_n$ are independent random variables, Xi having an exponential distribution with parameter θ_i; i = 1, 2, ..., n then $Z = \min(X_1, X_2, \ldots, X_n)$ has exponential distribution with parameter $\sum_{i=1}^{n} \theta_i$.

5) The exponential distribution is also known as lack of memory distribution i.e., if X has a exponential distribution, then for every constant $a \geq 0$, $P(Y \leq x \mid X \geq a) = P(Y \leq x)$ for all x, where $Y = X - a$.

Scanned by CamScanner

Uses of exponential distribution:

i) It is used in reliability theory and queuing theory.

ii) It is used in studying decay time of radioactive particle.

Example 45: The life time(yrs) of a large no of components of an equipment has negative exponential distribution with parameter 3. Find the probability that the life time of component selected at random is (i) at least 4 months (ii)4 months to 6 months.

Solution

Let x = Life time of a component of an equipment

$f(x) = \theta e^{-\theta x} = 3e^{-3x}, x > 0$

i) $P(X \geq 4) = P(X \geq 4/12) = P(X \geq 1/3)$

$$= \int_{1/3}^{\infty} f(x)\,dx = \int_{1/3}^{\infty} 3e^{-3x}\,dx$$

$$= \frac{3}{-3}[e^{-3x}]_{1/3}^{\infty} = -e^{-\infty} + e^{-1} = e^{-1} = 0.367$$

ii) $P(4 \leq X \leq 6) = P(4/12 \leq X \leq 6/12) = P(1/3 \leq X \leq 1/2)$

$$= \int_{1/3}^{1/2} f(x)\,dx$$

$$= \int_{1/3}^{1/2} 3e^{-3x}\,dx$$

$$= \frac{3}{-3}[e^{-3x}]_{1/3}^{1/2} = -\{e^{-3/2} - e^{-1}\} = e^{-1} - e^{-1.5} = 0.143$$

Example 46. Arv X has exponential distribution with pdf $f(x) = 4e^{-4x}, x > 0$. Find the(i) probability that X is not less than 4 (ii) mean and variance

Solution.

Here, $\theta = 4$

(i) $P(X \geq 4) = \int_4^{\infty} f(x)\,dx = \int_4^{\infty} 4e^{-4x}\,dx = \frac{4}{-4}[e^{-4x}]_4^{\infty} = -(e^{-\infty} - e^{-16}) = e^{-16}$

(ii) Mean $E(X) = \frac{1}{\theta} = \frac{1}{4}$ and Variance $V(X) = E(X^2) - \{E(X)\}^2 = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{2}{16} - \frac{1}{16} = \frac{1}{16}$

Example 47. Jobs are sent to a printer at an average rate of 3 jobs per hour. (a) What is the expected time between jobs? (b) What is the probability that the next job is sent within 5 minutes?

Solution.

Let X = time of job, $\theta = 3$ /hr

a) $E(X) = 1/\theta = 1/3$ hours = 20 minutes between jobs

b) 5 min = (1/12) hrs

$$P\left\{X < \frac{1}{12}\text{hrs}\right\} = \int_0^{1/12} 3e^{-3x}\,dx = \left[\frac{3e^{-3x}}{-3}\right]_0^{1/12} = [-e^{-3x}]_0^{1/12} = -e^{-3/12} + 1 = 1 - e^{-1/4} = 0.22$$

Gamma Distribution

A continuous random variable x is said to follow gamma distribution with parameter $\alpha > 0$, if it's probability density function is given by

$$f(x) = \frac{e^{-x} x^{\alpha-1}}{\Gamma\alpha}, 0 < x < \infty$$

Scanned by CamScanner

A random variable following gamma distribution with parameter α is denoted by $x \sim G(\alpha)$.

Note: A continuous random variable X is said to follow gamma distribution with parameters α and β if its probability density function is given by $f(x) = \frac{\beta^{\alpha}}{\Gamma\alpha} e^{-\beta x} x^{\alpha-1}$; $0 < x < \infty$, $\alpha > 0$, $\beta > 0$

Moments of gamma distribution

The moments about origin of the gamma distribution is denoted by μ_2' and is given by

$$\mu_r' = \int_0^{\infty} x^r f(x)\, dx$$

$$= \int_0^{\infty} x^r \frac{e^{-x} x^{\alpha-1}}{\Gamma\alpha} dx$$

$$= \int_0^{\infty} \frac{e^{-x} x^{r+\alpha-1}}{\Gamma\alpha} dx$$

$$= \frac{\Gamma r + \alpha}{\Gamma\alpha}$$

When r = 1

$\mu_1' = \alpha$

When r = 2

$\mu_2' = \alpha(\alpha + 1)$

When r = 3

$\mu_3' = \alpha(\alpha + 1)(\alpha + 2)$

When r = 4

$\mu_4' = \alpha(\alpha + 1)(\alpha + 2)(\alpha + 3)$

Mean and variance of gamma distribution

Mean of the distribution is

$E(x) = \mu_1' = \alpha$

Variance of the distribution is

$V(x) = \mu_2' - \mu_2'^2 = \alpha(\alpha + 1) - \alpha 2 = \alpha$

Properties

i. Mean = α

ii. Variance = α

iii. It tends to normal distribution as α tend to infinity

iv. It follows additive property.

Uses of gamma distribution

i. It is used in reliability theory.

ii. It is used to find life time of electronic equipments.

Example 48. Life time in yrs of an electronic equipment has gamma distribution with parameter α=2. Find the probability that life of the equipment is (i) more than two years (ii) Between 1 to years.

Solution.

Let x = life of electronic machine

$X \sim G(\alpha)$

$$f(x) = \frac{e^{-x} x^{\alpha-1}}{\Gamma\alpha} = \frac{e^{-x} x^{2-1}}{\Gamma 2} = \frac{e^{-x} x}{\Gamma 2}$$

Scanned by CamScanner

Now, $P(X>2) = \int_2^{\infty} f(x)\,dx$

$$= \int_2^{\infty} \frac{e^{-x}\,x}{\Gamma 2}\,dx = \frac{1}{\Gamma 2}\int_2^{\infty} x\,e^{-x}\,dx$$

$$= \frac{1}{1!}\left[x\int e^{-x}\,dx - \int\left\{\frac{dx}{dx}\int e^{-x}\,dx\right\}dx\right]_2^{\infty}$$

$$= \frac{1}{1!}\left[x\frac{e^{-x}}{-1} - \int\frac{e^{-x}}{-1}\,dx\right]_2^{\infty}$$

$$= \frac{1}{1!}\left[x\frac{e^{-x}}{-1} + \frac{e^{-x}}{-1}\right]_2^{\infty}$$

$$= \frac{1}{1!}\left[-xe^{-x} - e^{-x}\right]_2^{\infty}$$

$$= \{-\infty\, e^{-\infty} - e^{-\infty} + 2e^{-2} + e^{-2}\}$$

$$= 3e^{-2}$$

Also

$P(1<x<2)$ $= \int_1^2 f(x)\,dx = \int_1^2 \frac{e^{-x}\,x}{\Gamma 2}\,dx = \frac{1}{\Gamma 2}\int_1^2 xe^{-x}\,dx$

$$= \frac{1}{1!}\left[x\int e^{-x}\,dx - \int\left\{\frac{dx}{dx}\int e^{-x}\,dx\right\}dx\right]_1^2$$

$$= \frac{1}{1!}\left[x\frac{e^{-x}}{-1} - \int\frac{e^{-x}}{-1}\,dx\right]_1^2$$

$$= \frac{1}{1!}\left[x\frac{e^{-x}}{-1} - \frac{e^{-x}}{-1}\right]_1^2$$

$$= \frac{1}{1!}\left[-xe^{-x} - e^{-x}\right]_1^2$$

$$= \{-2\,e^{-2} - e^{-2} + e^{-1} + e^{-1}\}$$

$$= -3e^{-2} + 2e^{-1}$$

Scanned by CamScanner

Exercise – 6

Theoretical Question

1. What do you mean by joint probability mass function? How can you differentiate it from marginal probability mass function?
2. Define the terms; joint probability density function,marginal probability density function, conditional probability density function.
3. Write down the properties of joint pdf, marginal pdf and conditional pdf.
4. What do you mean by conditional probability distribution function?
5. What do you mean by joint probability distribution function? Write down it's properties.
6. What do you mean by independence of random variables? Explain.
7. What do you mean by marginal probability distribution function? Write down it's properties.
8. Describe Binomial distribution with its properties.
9. Describe Poisson distribution with its properties
10. When is Poisson distribution more preferred to Binomial distribution?
11. What do you mean by Poisson variate?Under what conditions Binomial distribution tends to poisson distribution.
12. What do you mean by parameters? Why is it important?
13. What do you mean by moments about origin? What are moments about origin of Binomial and Poisson distribution?
14. Describe the characteristics of Binomial distribution.
15. Describe the characteristics of Poisson distribution.
16. What is normal distribution?Describe characteristics of normal distribution.
17. What is standardized normal distribution? Write down its importance.
18. What is exponential distribution?Describe characteristics of exponential distribution.
19. What is gamma distribution?Write down characteristics of gamma distribution.

Numerical Questions

1. If two random variables have the joint probability density function

$$f(x,y)=\begin{cases} 6\dfrac{(x+y^2)}{5} & \text{for } 0<x<1,\ 0<y<1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that (i) $0.2<x<0.5$ and $0.4<y<0.6$ (ii) $x>0.4$ and $y<0.5$.

[Ans: 0.43, 0.2

2. Suppose that x & y have joint density function

$$f(x,y)=\begin{cases} (x+y), & \text{if } 0<x,y<1 \\ 0 & \text{elsewhere} \end{cases}$$

Find:(i)Marginal density function of x & y (ii)Covariance between x & y.

[Ans: (x+1)/2, (y+1)/2, -0.0

3. If two random variables X_1 and X_2 have the joint probability density function

$$f(x_1,x_2)=\begin{cases} \dfrac{2}{3}(x_1+2x_2) & \text{for } 0<x_1<1,\ 0<x_2<1 \\ 0 & \text{elesewhere} \end{cases}$$

Find the conditional density of X_1 given X_2 and X_2 given X_1

Scanned by CamScanner

$\left(\text{Ans: } \frac{2(x_1+2x_2)}{1+4x_2}, \frac{x_1+2x_2}{x_1+1}\right)$

4. Let (X,Y) be two dimensional random variable with joint pmf $P(x,y)=\frac{x-y+3}{48}$; x=0,1,2,3; y=0,1,2,3. Find: (i) marginal pmf of X and Y (ii) conditional distribution of X for given Y=1 (iii) conditional distribution of Y for given X=2.

$\left(\text{Ans: } \frac{2x+3}{24}, \frac{9y-2}{24}, \frac{x+2}{14}, \frac{5-y}{14}\right)$

5. Given the following bivariate probability distribution of X & Y.

Y \ X	-1	0	1
0	1/15	2/15	1/15
1	3/15	2/15	1/15
2	2/15	1/15	2/15

Find:(i) $P(X=1,y\le1)$ (ii)$P(Y\le1)$ (iii) $P(X=-1)$

[Ans: 2/15 , 2/3 , 2/5]

6. A company has 3 hardware engineering consultants and 4 software engineering consultants. From these 2 engineering consultants are selected at random. Let X denote the number of hardware engineering consultant and Y denotes the number of software engineering consultant. Find the (i)Joint probability distribution of X & Y. (ii)Marginal distribution of X and that of Y (iii)Conditional distribution of X for given Y.

7. A two-dimensional discrete random variable (XY) has the joint probability distribution as given below:

Y \ X	0	1	2	3
-1	0.03	0.02	0.05	0
0	0.01	0.25	0.45	0.01
1	0	0.03	0.05	0.10

Determine:(i) marginal distribution of X & Y (ii)Conditional distribution of X for given Y and that of Y for given X (iii) $P(X=0,Y\le2)$(iv) $P(X\ge0,Y<3)$ (v)$P(X\le0/Y=1)$

8. Suppose two-dimensional random variable (XY) has pdf given by

$$f(x,y)=\begin{cases} x^2+kxy; & 0\le x\le 1;\ 0\le y\le 2 \\ 0 & \text{elsewhere} \end{cases}$$

Determine: (i) constant k (ii) marginal pdf of X and that of Y(iii) conditional pdf of X given Y and that of Y given X (iv)Are X and Y independent.

$\left(\text{Ans: } \frac{1}{3}, \frac{6x^2+2x}{3}, \frac{2+y}{6}, \frac{2(3x^2+xy)}{2+y}, \frac{3x^2+xy}{6x^2+2x}, \text{dependent}\right)$

9. The joint pdf of two dimensional random variable (X,Y) is $f(x,y)=\begin{cases} 2x; & 0\le x\le 1,\ 0\le y\le 1 \\ 0 & \text{elsewhere} \end{cases}$

(i)Find marginal pdf of X & Y (ii) Check independence of X & Y.

[Ans: 2x, 1, independent]

10. Let two dimensional random variable (X,Y) have joint pdf $f(x,y)=\begin{cases} k(6-x-y); & 0\le x\le 2,\ 2\le y\le 4 \\ 0 & \text{elsewhere} \end{cases}$

Determine: (i) Constant k (ii) $P(X\le1\cap Y<3)$ (iii)$P(Y\le3)$ (iv) $P(X\le1/Y\le3)$

Scanned by CamScanner

[Ans: 1/8, 3/8, 5/8, 3/5]

11. Joint distribution of X & Y are given by $f(x,y)=\begin{cases}4xye^{-(x^2+y^2)}; & x\geq 0,\ y\geq 0\\ 0 & \text{elsewhere}\end{cases}$. (i) Find conditional pdf of X given Y ii) test whether X and Y are independent.

[Ans: $2x\,e^{-x^2}$, independent]

12. If $f(x,y)=\begin{cases}k(2x+3y): & 0\leq x\leq 1;\ 0\leq y\leq 1\\ 0 & \text{otherwise}\end{cases}$. Find (a)Constant k (b)Conditional pdf of x given y & that of y given x (c)x & y are independent or not (d) V(X), V(Y)(e) V(XY), COV(X,Y), r(X,Y) (f) E(3X+2Y), V(3X-2Y)

[Ans: 2/5, $\frac{2x+3y}{1+3y}$, $\frac{4x+6y}{4x+3}$, dependent, 0.078, 0.073, 0.055, -1/150, -0.087, 2.9, 1.08]

13. If $P(x,y)=\begin{cases}k(x+y): & x=1,2,3;\ y=1,2\\ 0 & \text{Otherwise}\end{cases}$. Find: (a) Constant k (b) Conditional pmf of x given y & that of y given x (c) x & y are independent or not (d) V(X), V(Y) (e) V(XY), COV(X,Y), r(X,Y) (f) E(2X-3Y), V(2X+3Y)

[Ans: 1/21, $\frac{x+y}{3y+6}$, $\frac{x+y}{2x+3}$, dependent, 0.63, 0.24, 2.81, -0.017, -0.043, -0.33, 4.56]

14. If $f(x,y)=\begin{cases}k\,e^{-(x+2)}: & 0\leq x\leq\infty;\ 0\leq y\leq\infty\\ 0 & \text{otherwise}\end{cases}$. Find: (a) Constant k (b) Conditional pdf of x given y & that of y given x (c) x & y are independent or not (d) V(X), V(Y) (e) V(XY), COV(X,Y), r(X,Y) (f) E(3X-2Y), V(3X+2Y)

[Ans: 1, e^{-x}, e^{-y}, independent, 1, 1, 3, 0, 0, 1, 13]

15. If $f(x,y)=\begin{cases}k(x+2)e^{-y}: & 0\leq x\leq 1;\ y\geq 0\\ 0 & \text{otherwise}\end{cases}$. Find (a) Constant k (b) Conditional pdf of x given y & that of y given x (c) x & y are independent or not (d) V(X), V(Y) (e) V(XY), COV(X, Y), r(X, Y)

[Ans: 2/5, $\frac{2(x+y)}{5}$, e^{-y}, independent, 0.08, 1, 0.44, 0, 0]

16. Let (X,Y) be jointly distributed r.v. with pdf given by $f(x,y)=\begin{cases}2; & 0<x<y<1\\ 0; & \text{Otherwise}\end{cases}$. Find the marginal pdf of X and that of Y, V(X), V(Y), Cov(XY)

[Ans: 2(1 - x), 2y, 1/18, 1/18, 5/18]

17. The joint pdf of two random variable X and Y is given by $f(x,y)=\frac{9(1+x+y)}{2(1+x)^4(1+y)^4}$; $0<x<\infty$, $0<y<\infty$. Find the marginal pdf of x and that of y, conditional pdf of y given x.

$\left(\text{Ans: } \frac{3(2x+3)}{4(1+x)^4}, \frac{3(2y+3)}{4(1+y)^4}, \frac{6(1+x+y)}{4(1+y)^4}\right)$

18. Comment on the following that mean and variance of Binomial distribution are 7 and 11 respectively.

[Ans: incorrect]

19. Find p if n = 6 and 9 P (X = 4) = P(X=2).

[Ans:1/4]

20. In a Binomial distribution with 6 independent trials, the probability of 3 and 4 success is found to be 0.2457 and 0.08189 respectively. Find the parameters, mean, variance.

[Ans: 4/13, 1.84, 1.14]

21. The mean and variance of Binomial distribution are 3 and 2 respectively. Find the probability of (i) less than or equal to 2 (ii) greater than or equal to 7.

[Ans: 0.37, 0.008]

Scanned by CamScanner

22. 12% of the chips produced by a machine are defective. What is the probability that out of a random sample of 20 chips produced by the machine, 5 are defective?

[Ans: 0.056]

23. The average number of defective components on repairing a computer is found to be 1 in 10. Find the probability of getting exactly 3 defective components among 10 repaired computers selected at random.

[Ans: 0.057]

24. A quality control engineer tests the quality of produced computers. Suppose that 5% of computers have defects, and defects occur independently of each other.Find the probability of exactly 3 defective computers in a shipment of twenty.

[Ans: 0.095]

25. In a bombing campaign there is the probability of hitting the target is 50%. How many bombs should be dropped so there is 99.9% chance of hitting the target if (i) one hit is enough to destroy the target? (ii) at least two hits are required to destroy the target.

[Ans: 10, 14]

26. The probability of a program developer developing a software is 1/3. How many times must he try so that the probability of developing at least one software is more than 90%.

[Ans: 6]

27. The probability that a student will graduate is 0.4. Determine the probability that out of 5 students:(i) none(ii)one (iii) at least one and (iv) all will graduate.

[Ans: 0.077, 0.259 , 0.922, 0.01]

28. At a particular university it has found that 20% of the students withdraw without completing the Statistics Course. Assume that 18 students have registered for the course; (i)what is the probability that none will withdraw? (ii) what is the probability that at least one will withdraw (iii)what is the probability that at most 2 will withdraw.

[Ans: 0.01, 0.99, 0.27]

29. A programmer succeeds twice as often as it fails while developing a specific program. Find the chance that in the next six attempt, there will be at least four successes.

[Ans: 0.342]

30. The incidence of occupation disease in an industry such that workers have 20% chance of suffering from it. What is the probability that out of six workers 4 or more will catch the disease?

[Ans: 0.016]

31. From the past experience a stock broker finds that 70% of the telephone calls he receives during the business hours, are orders and the remaining are for other business. What is the probability that out of first 8 telephone calls during a day –
    (i) exactly 5 calls are order calls?
    (ii) at least 6 are order calls ?

[Ans:0.254 , 0.55]

32. A discrete random variable X has mean equal to 6 and variance equal to 2. If it is assumed that the underlying distribution of X is binomial, what is the probability that $5 < X < 7$?

[Ans: 0.273]

33. A software package consists of 12 programs, five of which must be upgraded. If 4 programs are randomly chosen for testing,
    (a) What is the probability that at least two of them must be upgraded?
    (b) What is the expected number of programs, out of the chosen four, that must be upgraded?

Scanned by CamScanner

[Ans: 0.554, 2.216]

34. An exciting computer game is released. Fifty percent of players complete all levels. Forty percent of them will then buy advanced version of the game. Among 12 users what is expected number of people who will buy the advanced version? What is probability that (i) at least two people will buy it (ii) at most three people will buy it.

[Ans: 2.4, 0.725, 0.794]

35. A large chain retailer purchases a certain kind of electronic device from a manufacturer. The manufacturer indicates that the defective rate of the device is 3%.

(a) The inspector randomly picks 20 items from a shipment. What is the probability that there will be at least one defective item among these 20?

(b) Suppose that the retailer receives 10 shipments in a month and the inspector randomly tests 20 devices per shipment. What is the probability that there will be exactly 3 shipments each containing at least one defective device among the 20 that are selected and tested from the shipment?

[Ans: 0.45 , 0.16]

36. A car hire firm have two cars which it hires out day by day at the average of 1.5. Calculate the proportion of days on which none of the cars are used and some demands are refused.

[Ans:0.22 , 0.19]

37. An automatic machine makes paper clips from coils of wire. On the average, 1 in 400 paper clips is defective. If the paper clips are packed in boxes of 100, assuming that the process follows Poisson distribution, what is the probability that any given box of clips will contain, (i) no defective, *(ii)* one or more defectives, *(iii)* less than two defectives?

[Ans:0.77 , 0.22 , 0.96]

38. The chance of traffic accidents in a day in a street of Nepal is assumed to be 0.0005. On how many days can you expect i)no accidents ii) more than three accidents if there are 1000 such streets.

[Ans: 221, 43 days per year]

39. An office switch board receives telephone calls at the rate of 3 calls per minute on average. What is the probability of receiving (i) no calls on one minute interval (ii)at most 4 calls in 3 minute interval.

[Ans: 0.049, 0.945]

40. The average number of network error experienced in a day on a local area network(LAN) is distributed with an average 2.4. What is the probability that in any given day (i) Zero network error will occur (ii) at least one network error will occur (iii) exactly one network error will occur.

[Ans: 0.09, 0.91, 0.21]

41. Messages arrive at an electronic message center at random order, with an average of 9 messages per hour.

(a) What is the probability of receiving at least three messages during the next hour?

(b) What is the probability of receiving exactly five messages during the next two hours?

[Ans: 0.997, 0.066]

42. A hospital have an average of 4 emergency calls in 10 minutes interval. What is the probability that (i) there are at most 2 emergency calls in a minute (ii) there are exactly 3 telephone calls in 10 minutes?

[Ans: 0.992, 0, 0.434]

43. If a random variable X follows Poisson distribution such that $P(X = 1) = P(X = 2)$, find the mean and variance of the distribution.

Scanned by CamScanner

44. Calculate mean and variance of a Poisson variable x, if P(x = 4) = P(x = 5). **(Ans : 2,2]**

45. A manufacturer of pen drives knows that 3% of his product is defective. If he sells in boxes of 200 and guarantees that not more than 2 pen drives will be defective, what is the probability that a box will fail to meet the guaranteed quality? **[Ans: 5, 5]**

46. A dangerous computer virus attacks a folder consisting of 250 files. Files are affected by the virus independently of one another. Each file is affected with the probability 0.032. What is the probability that more than 5 files are affected by this virus? **[Ans: 0.93]**

47. A manufacturer produces IC chips, one percent of which are defective. Find the probability that in a box containing 100 chips, no defective are found? **[Ans: 0.808]**

48. The probability of error in the transmission of a bit over a communication channel is 0.001. What is the probability of more than three errors in transmitting a block of 1,000 bits. **[Ans: 0.36]**

49. Messages arrive at an electronic message center at random times with an average of 9 messages per hour. What is probability of **[Ans: 0.08]**
    a) receiving at least five messages during the next hour?
    b) receiving exactly seven messages during the next hour? **[Ans: 0.945, 0.117]**

50. On the average,1 computer in 1000 crashes during a severe thunderstorm. A certain company had 5,000 working computers when the area was hit by a severe thunderstorm.
    (a) Compute the probability that less than 3 computers crashed.
    (b) Compute the probability that exactly 10 computers crashed. **[Ans: 0.124, 0.018]**

51. Suppose that the number of telephone calls between 10 A.M. and 11 A.M., denoted by X, is Poisson variate with mean 2 and the number of telephone calls between 11 A.M. and 12noon, denoted by Y, is Poisson variate with mean 4. If X and Y are independent what is the probability that there are at least 5 telephone calls between 10 A.M. to 12 noon? **[Ans: 0.55]**

52. A source of liquid is known to contain bacteria with the mean no of bacteria per cubic centimeter equal to 3. Ten 1 cubic centimeter test tubes are filled with the liquid. Assuming the Poisson distribution is applicable. Calculate the probability that (i) All 10 test tube will show growth(i.e. at least 1 bacteria each) (ii)Exactly 7 test tubes will show growth. **[Ans: 0.6008, 0.215]**

53. If Z is the standard normal variable then calculate the following probabilities: (a) Z is between 0 and 1.34 (b) Z is less than 1.34 (c) Z is more than 1.96 (d) Z is between - 1.34 and 0 (e) Z is less than - 1.34 (f) Z is greater than - 2.04 (g) Z is between 1.96 and 2.84 (h) Z is between -1.96 and 2 (i) Z is between - 2.84 and - 1.96 (j) Z is less than -1.96 or more than 1.96 **[Ans: 0.4, 0.9, 0.025, 0.409, 0.09, 0.097, 0.02, 0.95, 0.022, 0.05]**

56. Suppose X follows the normal distribution with mean 100 and standard deviation 10, find the probability that (a) X is between 100 and 110 (b) X is more than 120 (c) X is less than 115 (d) X is between 80 and 100 (e) X is less than 80 (f) X is more than 90 (g) X is between 112 and 130 (h) X is between 85 and 125 (i) X is between 75 and 95.

Scanned by CamScanner

[Ans: 0.341, 0.022, 0.833, 0.477, 0.028, 0.841, 0.113, 0.926, 0.302]

57. The monthly production of certain types of computer parts of acompany was found to be normally distributed with mean number of computer parts 100,000 with a standard deviation of 20,000. Find the probability that the monthly production of computer parts is (i) more than 120,000; (ii) less than 125,000; (iii) more than 60,000; (iv) less than 80,000; (v) between 105000and 130,000.

[Ans: 0.158, 0.894, 0.977, 0.658, 0.295]

58. The mean yield for one-acre plot is 662 kilos with st.dev. 32 kilos. Assuming normal distribution. How many one-acre plots in a batch of 1000 plots would you expect to have yield (i) over 700 kilos, (ii) below 650 kilos and(iii) what is the lowest yield of the best 100 plots?

[Ans: 117, 352, 702.96]

59. Installation of SPSS software package requires downloading 95 files. On the average, it takes 16 secs to download one file, with a s.d. of 5 sec. What is the probability that the software is installed in less than 25 minutes?

[Ans: 0.341]

60. The local authorities in a certain city install 1000 pen drives to provide in different organizations of city. If these pen drives have an average life of 1000 hours with a standard deviation of 200 hours, assuming normality, what number of pen drives might be expected to fail (i) in the first 800 hours ? (ii) between 800 and 1,200 hours ? After what period of time would you expect that (a) 10% of the pen drives would fail? (b) 10% of the pen drives would be still working.

[Ans:159, 683, 744, 1256]

61. A training program is designed to upgrade the skills of computer operators. Because the program is self administered, computer operators take different number of hours to complete the program. Study shows that average time spent in training is normally distributed with 500 hours with s.d. 100 hours. What is the probability that a candidate selected at random will take time less than 420 hours?

[Ans: 0.986]

62. The life time of a certain electronic component in a normal random variate with the expectation of 5000 hours and a standard deviation of 100 hours. Compute the probabilities under the following conditions

a) Life time of components is less than 5012 hours

b)Lifetime of components between 4000 to 6000 hours

c)Lifetime of components more than 7000 hours

[Ans: 0452, 0.998, 0]

63. A machine at certain steel corporation makes iron rods. However, the machine does not make all rods of exactly the same length. It is known that the probability distribution of the lengths of rods made on this machine is normal with a mean length of 50 inches and a standard deviation of 0.06 inch. The rods that are either shorter than 49.85 inches or longer than 50.15 inches are rejected. What percentage of the rods made on this machine (a) are accepted? (b) are rejected?

[Ans: 98.76%, 1.24%]

64. Suppose that life of battery used in laptop of a software company is normally distributed with a mean of 40 months and a standard deviation of 5 months. If at a time, 1000 batteries are issued, how many will need the replacement after 35 months?

[Ans: 841]

65. Incomes of a group of 10,000 computer operators were found to be normally distributed with mean Rs. 15,200 and standard deviation Rs. 1600, find (i) highest income of poorest 2000 computer operators (ii) lowest income of richest 1000 computer operators.

[Ans: 1385.6, 1724.8]

Scanned by CamScanner

66. It is known that the life time of a calculator manufactured by Casio Company Ltd. has a normal distribution with mean of 54 months and a standard deviation of *8* months. If the company wants to replace just 22% calculator which starts not working properly, what should be warranty period?

[Ans: 47.84]

67. The marks obtained by 5000 students in an examination are normally distributed with mean 65 and variance 400. Estimate (i) lowest mark of top 10% students; (ii) highest marks of poorest 500 students; (iii) the limits of marks of middle 80% students.

[Ans: 73.09, 56.91, 56.91-73.09]

68. In an examination it is laid down that a student passes if he secures 30 per cent or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% and 60% marks und marks between 30% and 45% respectively. He gets a distinction in case he secures 80% or more marks. It is noticed from the results that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the second division. (Assume normal distribution of marks.)

[Ans: 34%]

69. The mean I.Q. test scores of IT student was 100 and the standard deviation 16. Assuming that the distribution was normal, find,
    (i) What percent of the student had I.Q. under 80?
    (ii) Between what limit the I.Q. of middle 40% of the student lay?
    (iii) What percent of the student had I.Q. with the range $\mu \pm 1.96\sigma$?

[Ans: 10.56%, 95-105, 95%]

70. The life of printers follows a normal distribution with mean *H* hours and a standard deviation of 200 hours. It is known that 10% of the printers have a life less than 400 hours. Find the mean life of the printers.

[Ans: 656]

71. If the daily sales of computer series books of a publisher is normally distributed with standard deviation of Rs. 5000. Calculate the mean value of the daily sales if the probability that the daily sales is more than Rs. 40,000 is 90%.

[Ans: 46400]

72. Sacks of grain packed by an automatic machine loader follow normal distribution having an average weight of 114 kgs. It is found that 15% of bags are over 115 kgs. Find the standard deviation.

[Ans: 0.96]

73. In examination, it was observed that 10% of the candidates got first division and 30% failed. The pass marks is 40 out of 100 and first division marks is 60 out of 100. Assuming that the marks follow normal distribution, find the mean and standard deviation of the distribution.

[Ans: 45.78, 11.11]

74. Of a large group of computer science students, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming a normal distribution, find the mean height and standard deviation.

[Ans: 65.43, 3.3]

75. If the skulls arc classified as A, B and C according as the length breadth index is under 75, between 75 and 80, or over 80, find approximately (assuming that the distribution is normal) the mean and standard deviation of a series in which A are 58%, B are 38% and C are 4%.

[Ans:74.4, 3.2]

Scanned by CamScanner

76. Suppose that waiting time (hrs) for bus in a bus station has a negative exponential distribution with parameters θ= 5 hours. What is the probability that a man has to wait at least 15 minutes? Also find expected waiting time for bus. [Ans: 0.28, 1/5hr]

77. The daily consumption of electricity in a city has an exponential distribution with mean 2000 kilowatt. Find the probability that electricity consumption on a particular day is (i) at least 1500 kilowatt (ii) at most 2500 kilowatt. [Ans: 0.47, 0.713]

78. A printer can print 15 pages per minute or average. Using normal distribution as approximation of Poisson distribution find probability that printer can print (i) more than 25 pages per minute (ii) less than 10 pages per minute. [Ans: 0.242, 0.359]

79. The life time of Lenovo cell phone has gamma distribution with parameter 2. Find the probability that cell phone has life (i) more than 2 years (ii) Between 3 years to 5 years [Ans:0.405, 0.159]

Practical Questions

1. Fit the Binomial distribution and find the expected frequencies for the following data.

X	0	1	2	3	4	5	6	total
f	7	6	19	35	23	7	1	98

[Ans: 2, 11, 25, 30, 21, 8, 1]

2. Five fair coin are tossed 100 times. From the following data fit the appropriate distribution and find the expected frequencies.

No. of heads	0	1	2	3	4	5
Observed frequency	2	10	24	38	18	8

[Ans: 2, 12, 28, 33, 20, 5]

3. 192 families had following distribution of girl child among the first three children. Find the expected frequencies on the basis that (i) both sex are equally probable (ii) the probability may vary.

No. of girl child	0	1	2	3	Total
No. of families	77	90	20	5	192

[Ans:(i) 24, 72, 72, 24, (ii) 40, 74, 85, 22]

4. Fit the Poisson distribution and find mean and variance.

Mistakes per page	0	1	2	3	4	5
No. of pages	142	156	69	27	5	1

[Ans: 147, 147, 25, 6]

5. Fit Poisson distribution and find the expected frequencies.

X	0	1	2	3	4	5	6	7
f	71	112	117	57	27	11	3	1

[Ans: 68, 120, 106, 63, 28, 10, 3, 1]

6. Fit normal distribution to the following data and find expected frequencies.

Class	20 – 30	30 –40	40 – 50	50 – 60	60 - 70	70 -80	80 - 90	90- 100
frequency	1	3	16	34	28	14	3	1

[Ans:0.67, 04.61, 16.39, 30.27, 29.1, 14.57, 3.97, 0.51]

□□□

Scanned by CamScanner

Chapter 7

CORRELATION AND LINEAR REGRESSION

Bivariate Data

When data are collected according to a single variable is called univariate data. For example, data on age of individuals. When data is collected according to two variables then it is called bivariate data. For example, data on height and weight of students.

Bivariate Frequency Distribution

Frequency distribution involving only one variable is called univariate frequency distribution. In all situation univariate distribution is not sufficient to represent data. In many situations simultaneous study of two variables become necessary. For example, we want to classify data relating to the weight and height of a group of individuals, income and expenditure of a group of individuals, age of husbands and wives etc.

The data so classified on the basis of two variables give rise to the so called bivariate frequency distribution and it can be summarized in the form of a table is called bivariate(two-way) frequency table. While preparing a bivariate frequency distribution, the values of each variable are grouped into various classes (not necessarily the same for each variable). If the data corresponding to one variable, say X is grouped into m classes and the data corresponding to the other variable, say Y is grouped into n classes then the bivariate table will consist of mxn cells. By going through the different pairs of the values, (X,Y) of the variables and using tally marks we can find the frequency of each cell and thus, obtain the bivariate frequency distribution. The format of a bivariate frequency distribution is given below:

Scanned by CamScanner

Format of Bivariate Frequency distribution

X-series / Y-series		Class -Intervals / Mid Values	Marginal Frequency of Y
Class - Intervals	Mid Values	frequency of the pair f(x,y)	fy
Marginal Frequency of X		fx	Total Σfx=Σfy=N

Here f(x,y) is the frequency of the pair (x,y). The frequency distribution of the values of the variables x together with their frequency total (fx) is called the marginal distribution of x and the frequency distribution of the values of the variable Y together with the total frequencies is known as the marginal frequency distribution of Y. The total of the values of manual frequencies is called grand total (N)

For example: the data given below relate to the height and weight of 20 persons. Construct a bivariate frequency table with class interval of heightas 62-64,64-66 ... andweightas115-125,125-135,write down the marginal distribution of X and Y.

S.No.	Height	Weight	S.No.	Height	Weight
1	70	170	11	70	163
2	65	135	12	67	139
3	65	136	13	63	122
4	64	137	14	68	134
5	69	148	15	67	140
6	63	121	16	69	132
7	65	117	17	65	120
8	70	128	18	68	148
9	71	143	19	67	129
10	62	129	20	67	152

Bivariate frequency table showing height and weight of persons.

Height(X) / Weight(Y)	62-64	64-66	66-68	68-70	70-72	Total
115-125	II (2)	II (2)				4
125-135	I	(1)	I (1)	II (2)	I (1)	5
135-145		III (3)	II (2)		I (1)	6
145-155			I (1)	II (2)		3
155-165					I (1)	1
165-175					I (1)	1
Total	3	5	4	4	4	20

Scanned by CamScanner

Correlation

The term correlation is used by a common man without knowing the meaning of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only one variable such as height, weight, age, marks, wage etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-variate Analysis. Sometimes the variables may be inter-related. We study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure etc. The nature and strength of relationship may be examined by correlation.

Thus correlation refers to the relationship of two variables or more. For example, relation between height of father and son, yield of crop and rainfall, wage and price index etc. Correlation is statistical Analysis which measures and analyze the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. If two or more variables are so related that the change of one variable brings re change in the value of other variable, then the variables are said to be correlated. This relationship between the variables is called Correlation. Hence two variables are said to be correlated if change in one variable is accompanied by change in other variable.

For Example: The change in quantity of irrigation brings the changes in production of rice. So irrigation and production are correlated. The measure of Correlation is called the Coefficient of Correlation. It measures the degree and direction of relationship between the variables but it does not say which is cause and which is effect.

Types of Correlation

It can be classified by following three types

1. Positive and Negative Correlation
2. Linear and Non-Linear Correlation
3. Simple, Multiple and Partial Correlation

1. Positive and Negative Correlation

If both the variables move in the same direction i.e. if the increase or decrease in the value of one variable results the increase or decrease in the value of other variable, then the two variables are said to Positive Correlated. For example:

Age(year)	5	8	10	13	15
Weight(kg)	20	28	32	40	45

Similarly, if both variables move in opposite direction, then correlation between the two variables is called Negative Correlation. In this correlation, if the value of one increases then the value of other variable decrease & vice versa. For example:

x	15	20	30	40	50
y	25	20	10	8	2

The relationship between two variables can be measured either by graphical method (Scatter diagram) or by numerical calculated method.

2. Linear and Non-linear Correlation:

The correlation between two variables is said to be linear if corresponding to a unit change in the other variable over the entire range of the value

For example,

Scanned by CamScanner

X	6	7	8	9	10
Y	5	7	9	11	13

In the other hand, if corresponding to a unit change in one variable, there is no constant change in other variable then the correlation is said to be Non-linear.

X	1	2	3	4	5
Y	7	8	15	20	23

3. **Simple, Multiple and partial correlation**

The relationship between two (i.e. one dependent and one independent) variables is called Simple Correlation.

The study of relationship among three or more variables simultaneously (at the same time) is known as multiple correlations. In this correlation all the given variables are studied at one time by taking one variable as dependent and the entire remaining variable as independent and the study of relationship between two variables keeping the effect of all remaining variables constant when there are three or more variables is called partial correlation.

Simple Correlation

It is the relationship between two variables. Two variables are correlated if change in one variable is accompanied by change in other variable.

These are the various method of various method of studying the correlation between two variables:

(a) Scatter Diagram Method

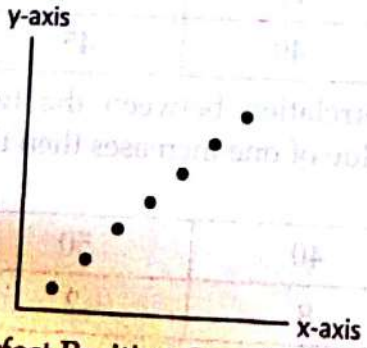
(b) Karl Pearson's Correlation coefficient

(c) Spearman's Rank Correlation

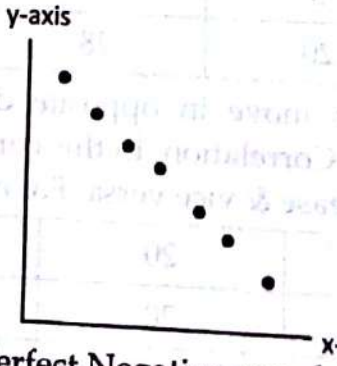
Scatter Diagram Method

Scatter diagram is one of the simplest ways of diagrammatic representation of bi-variate distribution and provides us one of simplest and attractive tools of ascertaining the correlation between two variables. In this method, the points are represented by dots by keeping the independent variable on x-axis and the dependent variable on y-axis. The graph formed by plotting these pairs of x and y is known as Scatter diagram. If the plotted dots show upward or downward then two variables under study are correlated. If the dots are close together and follows some trend of either increasing or decreasing then there is a strong relationship between them.

The following diagram of the scattered data depict different forms of correlation:

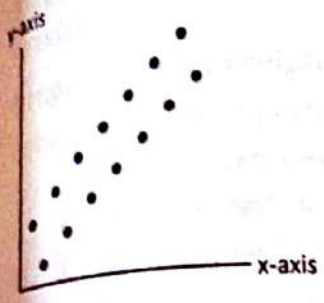


Perfect Positive Correlation

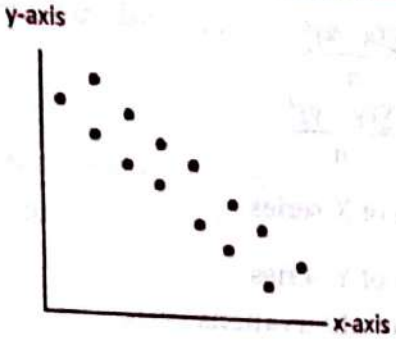


Perfect Negative correlation

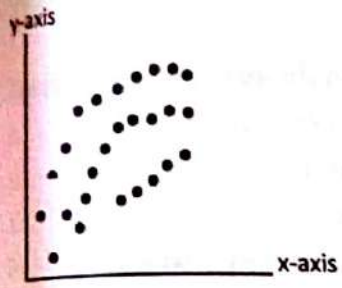
Scanned by CamScanner



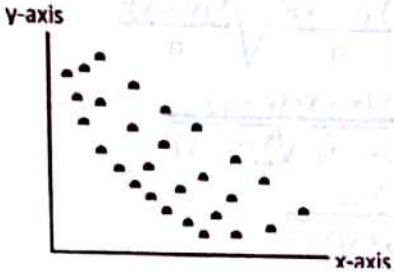
High Degree Positive Correlation



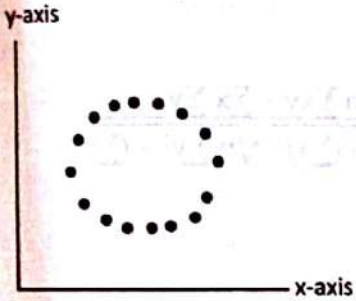
High Degree Negative Correlation



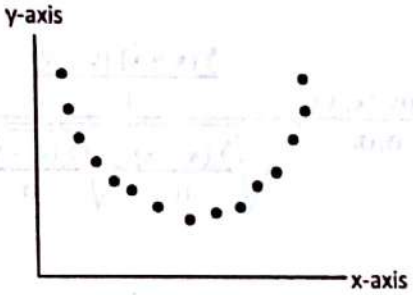
Low degree of positive correlation



Low degree of negative correlation



No correlation



No correlation

Merits

a. It is the simplest method of measuring correlation.

b. It is least affected by the size of extra values.

c. It can be easily understood by non-statistician person.

Demerits

It gives the only rough idea. So it cannot give the extra idea about the correlation between two variables.

Karl Pearson's Correlation Coefficient

The Karl Pearson's correlation coefficient measure the degree of association between the two variables. It is also known as Pearsonian correlation coefficient.

Let X and Y be two variables then Karl Pearson's correlation coefficient is denoted by r_{xy} or r_{yx} or simply r is define as ratio of Covariance between X and Y to multiple of SD of X and SD of Y

$$r = \frac{\text{Covariance (X,Y)}}{\sqrt{\text{VarianceX}}\sqrt{\text{Variance (Y)}}} = \frac{\text{Cov (X,Y)}}{\sigma_X \sigma_Y} \quad \ldots\ldots (1)$$

Where

$$\text{Cov}(X, Y) = \frac{1}{n}\sum(x - \bar{x})(y - \bar{y})$$

Scanned by CamScanner

$$\text{Variance } (x) = \sigma^2_x = \frac{\Sigma(x-\bar{x})^2}{n}$$

$$\text{Variance } (y) = \sigma^2_y = \frac{\Sigma(y-\bar{y})^2}{n}$$

$\bar{X}$ = Arithmetic mean of X-series

$\bar{Y}$ = Arithmetic mean of Y-series

n = Number of pair of observations

Now

$$r = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}$$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \qquad \text{..... (2)}$$

Where, $x = x-\bar{x}$, $y = y-\bar{y}$

Formula (2) is known as product moment method.

After simplification of (1)

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

Direct Method

$$r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{n\Sigma x^2-(\Sigma x)^2}\sqrt{n\Sigma y^2-(\Sigma y)^2}} \qquad \text{..... (3)}$$

Deviation Method (Change of Origin)

$$r = \frac{n\Sigma uv - \Sigma u\Sigma v}{\sqrt{n\Sigma u^2-(\Sigma u)^2}\sqrt{n\Sigma v^2-(\Sigma v)^2}}$$

Where $u = x - A$; $v = y - B$

A = assumed mean of X-series

B = assumed mean of Y-series

This is the suitable method of calculating the correlation coefficient of two variables consisting large numerical size.

Step Deviation Method(Change of Origin and Scale)

$$r = \frac{n\Sigma u'v' - \Sigma u'\Sigma v'}{\sqrt{n\Sigma u'^2-(\Sigma u')^2}\sqrt{n\Sigma v'^2-(\Sigma v')^2}}$$

$$u' = \frac{x-A}{h}, \; v' = \frac{y-B}{k}$$

Where h = common(scale) factor of X-series

k = common(scale) factor of Y-series

Scanned by CamScanner

This is the suitable method of calculating the correlation having common factor in x-series or y-series or both series.

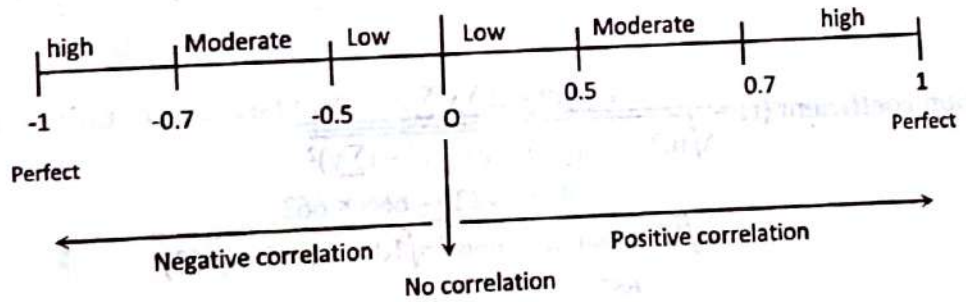
Properties of Correlation Coefficient

Following are the important properties of Karl Pearson's Correlation Coefficient

i. Correlation coefficient lies between -1 to 1 i.e., $-1 \le r \le +1$.
ii. Correlation coefficient is symmetrical i.e. $r_{xy} = r_{yx} = r$.
iii. Correlation coefficient is independent of change of origin and scale.
iv. Correlation coefficient is the geometric mean of two regression coefficient
$r = \pm\sqrt{b_{yx} \times b_{xy}}$
v. Correlation coefficient has no unit because of relative measure.

Interpretation of calculated value of r

i. If r = + 1, there is Perfect Positive Correlation between two variables.
ii. If r = –1, there is Perfect Negative Correlation between two variables
iii. If r = 0, there is no correlation between two variables or uncorrelated between two variables.
iv. If r lies between 0.001 to 0.499 there is low degree of positive correlation
v. If r lies between 0.5 to 0.699 there is moderate positive correlation
vi. If r lies between 0.70 to 0.999 High degree positive correlation
vii. If r lies between -0.499 to –0.001, there is Low degree of Negative Correlation
viii. If r lies between -0.699 to –0.500 there is moderate Negative Correlation.
ix. If r lies between -0.699 to –0.999 there is high degree of Negative Correlation.



Example 1. If the covariance between X and Y is 27 and the variance of X is 25 and the variance of Y is 64,find the Karl Pearson's correlation coefficient.

Solution

Here, Cov (x,y) = 27, Var(c) = 25 and Var(y) = 64

Karl Pearson's correlation coefficient (r) $= \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(y)}} = \frac{27}{\sqrt{25}\sqrt{64}} = \frac{27}{5 \times 8} = 0.675$

Example 2. The coefficient of correlation between two variable x and y is 0.38. Their covariance is 10.2. The variance of x is 16, find the standard deviation of y series.

Solution

Here, r = 0.38, Cov(x,y) = 10.2, Var(x) = 16, SD(y) = ?

We know

$$r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}\ \sqrt{\text{Var}(y)}}$$

Scanned by CamScanner

or $0.38 = \frac{10.2}{\sqrt{16}\sqrt{Var(y)}}$

or $0.38 = \frac{10.2}{4\sqrt{Var(y)}}$

or $\sqrt{Var(y)} = \frac{10.2}{4 \times 0.38} = 6.71$

or $SD(y) = 6.71$

Example 3: Calculate the Correlation Coefficient between x and y series from the following data:

	x series	y series
No of pair of observation	5	5
Arithmetic mean	6	8
Standard deviation	2.82	2
Sum of square of deviations from mean	40	20

Sum of the product deviation of x and y series from their respect mean = –26.

Solution

Here, $n = 5, \bar{x} = 6, \bar{y} = 8, \sigma_x = 2.82, \sigma_y = 2, \Sigma(x-\bar{x})^2 = 40, \Sigma(y-\bar{y}) = 20, \Sigma(x-\bar{x})(y-\bar{y}) = -26$

$$\therefore \quad r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{-26}{\sqrt{40}\sqrt{20}} = -0.9192$$

Example 4: For 10 pair of observations on two variables X and Y,the following information are obtained; $\Sigma x = 666, \Sigma y = 663, \Sigma x^2 = 44490, \Sigma y^2 = 44061, \Sigma xy = 44224$.Compute Karl Pearson's correlation coefficient.

Solution

$$\text{Karl Pearson's correlation coefficient (r)} = \frac{n\Sigma xy - n\Sigma x \Sigma y}{\sqrt{n(\Sigma x^2 - (\Sigma x)^2)}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10 \times 44224 - 666 \times 663}{\sqrt{10 \times 44490 - (666)^2}\sqrt{10 \times 44061 - (663)^2}}$$

$$= \frac{682}{36.66 \times 32.264} = 0.576$$

Example 5: Calculate the Karl Pearson's correlation coefficient between number of pages in a book and number of mistakes.

Number of pages of book X	6	2	10	4	8
Number of mistakes Y	9	11	5	8	7

Solution

X	Y	$x = X - \bar{x}$	$y = Y - \bar{y}$	x^2	y^2	xy
6	9	0	1	0	0	0
2	11	-4	3	16	9	-12
10	5	4	-3	16	9	-12
4	8	-2	0	4	0	0
8	7	2	-1	4	1	-2
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 40$	$\Sigma y^2 = 2$	$\Sigma xy = -26$

Scanned by CamScanner

Here,

$$\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6, \bar{y} = \frac{\sum y}{n} = \frac{40}{5} = 8, x = x - \bar{x}, y = y - \bar{y}$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{-26}{\sqrt{40}\sqrt{20}} = -0.919 > 0.$$

High degree negative correlation

Example 6: Calculate Karl Pearson's correlation coefficient between sales and repair of computers as given below.

Sales	50	55	55	60	65	70	65	60
Repair	11	13	14	16	16	15	15	20

Solution

Sales(X)	Repair(Y)	u = X – A	v = Y –B	uv	u^2	v^2
50	11	-15	-5	75	225	25
55	13	-10	-3	30	100	9
55	14	-10	-2	20	100	4
60	16	-5	0	0	25	0
65	16	0	0	0	0	0
70	15	5	-1	-5	25	1
65	15	0	-1	0	0	1
60	20	-5	4	-20	25	16
ΣX = 45	ΣY = 108	Σu = 597	Σv=285	Σuv=1356	Σu^2=500	Σv^2=56

Let A = 65, B = 16

$$\text{Correlation coefficient (r)} = \frac{n\Sigma uv - \Sigma u.\ \Sigma v}{\sqrt{n\,\Sigma u^2 - (\Sigma u)^2}\sqrt{n\,\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{8 \times 100 - (-40) \times (-8)}{\sqrt{8 \times 500 - (-40)^2}\sqrt{8 \times 56 - (-8)^2}}$$

$$= \frac{800 - 320}{48.990 \times 16.596}$$

$$= 0.50$$

Example 7. The following table gives the distribution of items and also defective items among them according to size groups. Find the correlation coefficient between size and defect in quality.

Size Group	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
No. of items	200	270	340	360	400	300
No of defective items	150	162	170	180	180	120

Solution:

Since correlation is to be found between size and defect in quality.

No. of items	No. of defect in items	Percentage of defect item
200	150	$\frac{150}{200} \times 100 = 75$
270	162	$\frac{162}{270} \times 100 = 60$

Scanned by CamScanner

340	170	$\frac{170}{340} \times 100 = 50$
360	180	$\frac{180}{360} \times 100 = 50$
400	180	$\frac{180}{400} \times 100 = 45$
300	120	$\frac{120}{300} \times 100 = 40$

Let X and Y be the mid value of age and percentage of number of players respectively.

Age group	Mid value (X)	(Y)	$u' = \frac{X - 27.5}{5}$	$v' = \frac{Y - 50}{5}$	$u'v'$	u'^2	v'^2
15 - 20	17.5	75	-2	5	-10	4	25
20 - 25	22.5	60	-1	2	-2	1	4
25 - 30	27.50	50	0	0	0	0	0
30 - 35	32.5	50	1	0	0	1	0
35 - 40	37.5	45	2	-1	-2	4	1
40 - 45	42.5	40	3	-2	-6	9	4
			$\Sigma u'=3$	$\Sigma v'=4$	$\Sigma u'v'=-20$	$\Sigma u^2=19$	$\Sigma v'^2=34$

Now,

$$r = \frac{n\Sigma u'v' - \Sigma u'.\Sigma v'}{\sqrt{n\,\Sigma u'^2 - (\Sigma u')^2}\sqrt{n\,\Sigma v'^2 - (\Sigma v')^2}}$$

$$= \frac{6 \times (-20) - 3 \times 4}{\sqrt{6 \times 19 - (3)^2}\sqrt{6 \times 34 - (4)^2}}$$

$$= \frac{-120 - 12}{\sqrt{105 \times 188}} = -0.94$$

Example 8: In order to find the correlation coefficient two variables x and y from 12 observations, the following calculations were made $\Sigma X = 30, \Sigma Y = 5, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 334$.

On checking data it was found that the pair (X = 11, Y = 4) was copied wrongly, the correct value being (X = 10, Y = 14). Find the correct value of correlation coefficient.

Solution

Here $n = 12, \Sigma X = 30, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 334$

Now Corrected $\Sigma X = 30 - 11 + 10 = 29$

Corrected $\Sigma Y = 5 - 4 + 14 = 15$

Corrected $\Sigma X^2 = 670 - 11^2 + 10^2 = 649$

Corrected $\Sigma Y^2 = 285 - 4^2 + 14^2 = 465$

Corrected $\Sigma XY = 334 - 11 \times 4 + 10 \times 14 = 430$

Then

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{12 \times 430 - 29 \times 15}{\sqrt{12 \times 649 - (29)^2}\sqrt{12 \times 465 - (15)^2}} = \frac{51 - 435}{\sqrt{7788 - 841}\sqrt{5580 - 225}} = \frac{4725}{\sqrt{6947}\sqrt{5355}} = 0.7757$$

High Degree of Positive Correlation.

Scanned by CamScanner

Goodness of Fit measure in terms of Correlation Coefficient

Probable error is a statistical measure for testing the reliability of the value of Correlation Coefficient. It is used to test the calculated correlation coefficient whether it is significant or not. If r is calculated correlation coefficient in a sample of n pair of observation then its standard error as $S.E.(r) = \frac{1-r^2}{\sqrt{n}}$ and Probable error is defined as

$$P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

Reason of taking the factor 0.6745 is that in a normal distribution 50% of the observations lie in the range.

- If $r < P.E.(r)$, then the value of r is not significant.
- If $r > 6\ P.E.(r)$, then the value of r is significant.

In other situation, nothing can be concluded certainly. The probable error of correlation is used to determine the limits within which the population correlation coefficient may be expected to lie limits of population correlation coefficient = $r \pm P.E.(r)$.

Example 9: If the correlation coefficient between X and Y is 0.7, the probable error of r is 0.0344, what will be the value of n?

Solution

Here, $r = 0.7$, $P.E.(r) = 0.0344$, $n = ?$

We know

$$P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

or $$0.0344 = 0.6745 \times \frac{1-(0.7)^2}{\sqrt{n}}$$

or $$\sqrt{n} = \frac{0.6745 \times 0.51}{0.0344}$$

or $$\sqrt{n} = 9.999$$

or $$n = 99.98 \approx 100$$

Example 10: The coefficient of correlation between 16 pairs of values of work hours and salary was found to be 0.7.Test the significance of the correlation coefficient.

Solution

Here, $n = 16$, $r = 0.7$

To test the significance of correlation coefficient,

$$P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

$$= 0.6745 \times \frac{1-0.7^2}{\sqrt{16}} = 0.6745 \times \frac{0.51}{4} = 0.085$$

Here $r = 0.7$ is not less than $P.E.(r) = 0.085$

$6\ P.E.(r) = 6 \times 0.085 = 0.515$

Now,

$r = 0.7 > 6\ P.E.(r) = 0.515$

The correlation coefficient between work hours and salary is found to be significant.

Scanned by CamScanner

Example 11. Find the correlation between x and y for given frequency as follows. And test consistency of x and y.

x	y	f
10	24	12
12	28	24
14	19	18
16	15	10
18	12	18
20	13	5

Solution:

x	y	f	fx	fx^2	fy	fy^2	fxy
10	24	12	120	1200	288	6912	2880
12	28	24	288	3456	672	18816	8064
14	19	18	252	3528	342	6498	4788
16	15	10	160	2560	150	2250	2400
18	12	18	324	5832	216	2592	3888
20	13	5	100	2000	65	845	1300
		Σf=87	Σfx = 1244	Σfx^2= 18576	Σfy = 1733	Σfy^2 = 37913	Σfxy = 23320

Here, Total frequency= N =87

$$\text{Mean of } x = \bar{x} = \frac{\sum fx}{\sum f} = \frac{1244}{87} = 14.299$$

$$\text{Mean of } y = \bar{y} = \frac{\sum fy}{\sum f} = \frac{1733}{87} = 19.919$$

$$\text{Standard deviation of } x = \sigma_x = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2} = 3.01$$

$$\text{Standard deviation of } y = \sigma_y = \sqrt{\frac{\sum fy^2}{\sum f} - \left(\frac{\sum fy}{\sum f}\right)^2} = 6.244$$

$$\text{Coefficient of variation of } x = C.V.(x) = \frac{\sigma_x}{\bar{x}} \times 100\% = 21.05\%$$

$$\text{Coefficient of variation of } y = C.V.(y) = \frac{\sigma_y}{\bar{y}} \times 100\% = 31.348\%$$

Karl Pearson's correlation coefficient

$$r = \frac{N\sum fxy - \sum fx \sum fy}{\sqrt{N\sum fx^2 - (\sum fx)^2}\sqrt{N\sum fy^2 - (\sum fy)^2}}$$

$$= -0.892$$

Example 12: A student calculates the value of correlation coefficient between study hour and marks secured is 0.795 when the number of item is 100 and concludes that r is highly significant. Is the conclusion correct? Also determine the limit of population correlation coefficient.

Scanned by CamScanner

Solution

Here, $r = 0.795$, $n = 100$

To test the significance of correlation coefficient,

$$P.E.(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

$$= 0.6745 \times \frac{1 - 0.795^2}{\sqrt{100}}$$

$$= 0.6745 \times \frac{0.3679}{10}$$

$$= 0.0248$$

Here $r = 0.795$ is not less than $P.E.(r) = 0.0248$

$6\ P.E.(r) = 6 \times 0.0248 = 0.1489$

$r = 0.795 > 6\ P.E.(r) = 0.1489$

Hence the conclusion that r is highly significant is correct.

Now, limit of population correlation coefficient $= r \pm P.E.(r) = 0.795 \pm 0.0248$

Taking – sign

$0.795 - 0.0248 = 0.7702$

Taking + sign

$0.795 + 0.0248 = 0.8198$

Hence, the limit of the population correlation coefficient is 0.7702 to 0.8198.

Example 13: Find whether there exists any relationship between the age of driver and no of accidents from the following table.

No of accidents	Age of driver				
	20 - 22	22 - 24	24 - 26	26 - 28	28 - 30
0	5	7	13	-	-
1	-	10	15	2	-
2	-	3	12	6	-
3	-	-	4	12	5
4	-	-	-	4	2

Solution

$$u = x - 2,\ v = \frac{y - 25}{2}$$

Scanned by CamScanner

Let X and Y represent age of driver and no of accidents

	Age of driver	20 - 22	22 - 24	24 - 26	26 - 28	28 - 30				
	Mid value (Y)	21	23	25	27	29				
No of accidents (X)	v / u	-2	-1	0	1	2	f	fu	fu^2	fuv
0	-2	5 [20]	7 [14]	13 [0]	- [ ]	- [ ]	25	-50	100	34
1	-1	- [ ]	10 [10]	15 [0]	2 [-2]	- [ ]	27	-27	27	8
2	0	- [ ]	3 [0]	12 [0]	6 [0]	[ ]	21	0	0	0
3	1	- [ ]	[ ]	4 [0]	12 [12]	5 [10]	21	21	21	22
4	2	- [ ]	[ ]	[ ]	4 [8]	2 [8]	6	12	24	16
	f	5	20	44	24	7	N = 100	Σfu =- 44	Σfu^2= 172	Σfuv = 80
	fv	-10	-20	0	24	14	Σfv =8			
	fv^2	20	20	0	24	28	Σfv^2 =92			
	fuv	20	24	0	18	18	Σfuv = 80			

$$r = \frac{n\Sigma fuv - \Sigma fu.\ \Sigma fv}{\sqrt{n\ \Sigma fu^2 - (\Sigma fu)^2}\sqrt{n\ \Sigma fv^2 - (\Sigma fv)^2}}$$

$$= \frac{100 \times 80 - 8 \times (-44)}{\sqrt{100 \times 92 - (8)^2}\sqrt{100 \times 172 - (-44)^2}}$$

$$= \frac{800 + 352}{\sqrt{9200 - 64}\sqrt{17200 - 1936}}$$

$$= \frac{8352}{95.58 \times 123.55} = 0.71$$

There exists high degree of positive correlation between age of driver and number of accident.

Spearman Rank Correlation Coefficient

Degree of relationship between two attributes is known as rank correlation coefficient and it is denoted by r_s. Rank correlation coefficient is also known as Spearman's rank correlation. It is used to find correlation coefficient between the qualitative variables such as beauty, knowledge,intelligence, honesty etc. which cannot be measured quantitatively directly. Rank correlation can be computed in the following three conditions:

(a) When ranks are given

(b) When ranks are not given and not repeated

(c) When ranks are not given and repeated

Scanned by CamScanner

(a) When ranks are given:

When ranks are given the following steps are to be considering as follows:

Find the difference between the ranks and denoted by $d = R_1 - R_2$ or $R_2 - R_1$

Find d^2 and obtained Σd^2

Then, find rank correlation coefficient by using following formula

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

Where, $\Sigma d = 0$ is always zero

$d = R_1 - R_2$ or $R_2 - R_1$

Example 14: Find the rank correlation from the following ranked given by the two judges in a beauty pageant.

Judge I	1	3	2	5	4
Judge II	2	1	4	3	5

Solution

Judge I (R_1)	Judge II (R_2)	$d = R_1 - R_2$	d^2
1	2	-1	1
3	1	-2	4
2	4	-2	4
5	3	2	4
4	5	-1	1
		$\Sigma d = 0$	$\Sigma d^2 = 14$

We have

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 14}{5\times(5^2-1)} = 1 - \frac{84}{5\times 24} = 1 - \frac{84}{120} = 1 - 0.7 = 0.3$$

(b) When rank are not given and not repeated

In this case, first of all we assigned the rank for given quantitative data by giving the highest value as 1 and second highest value '2' and so on. The same method is used for both variables under consideration. Rest procedure is same as earlier then the rank correlation coefficient

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

Where

$d = R_1 - R_2$, $\Sigma d = 0$

R_1 = Rank assigned by 1st attributes

R_2 = Rank assigned by 2nd attributes

Example 15: One of nature's pattern connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 6 colonies of sparrow-hawks. Find out the rank correlation between percent return and new adult.

Percent return	10	12	11.8	11.9	13	10.5
New adult	10	15	18	14	30	27

Scanned by CamScanner

Solution

Here, R_1 & R_2 be the rank of percent return and new adult assigned from smallest to largest

Percent return	New adult	Rank of percent return(R_1)	Rank of New adult (R_2)	$d = R_1 - R_2$	d^2
10	10	1	1	0	0
12	15	5	3	2	4
11.8	18	3	4	-1	1
11.9	14	4	2	2	4
13	30	6	6	0	0
10.5	27	2	5	-3	9
				$\Sigma d = 0$	$\Sigma d^2 = 18$

We have

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 18}{6(6^2-1)} = 0.48$$

Since, R = 0.48 so low degree of positive correlation between percent return and new adult.

(c) When rank are not given and repeated

Sometimes, it may happen that two or more variant values are equal. In such case it is necessary to rank two or more terms as equal or repeated or tie. In such case, each item should give an average or a common rank.

E.g., if the 2nd and 3rd items have the some value, then the rank for each of them is $\frac{2+3}{2} = 2.5$ and the rank of next item is 4. If 5th, 6th and 7th items have same value, then rank of them is $\frac{5+6+7}{3} = 6$ and so on.

Then, in this case, some adjustments are to be made in the Spearman's rank formula. The adjustment factor $\frac{m(m^2-1)}{12}$ is to be added to Σd^2.

Where m = the number of times a variate value repeats

After adjustment,

Spearman's rank correlation coefficient

$$R = 1 - \frac{\left\{\sum d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \ldots\right\}}{n(n^2-1)}$$

Example 16: Obtained Spearman's rank correlation for the following

x	10	10	10	15	14	15	20
y	50	80	70	70	80	80	100

Solution

Let R_1 and R_2 be rank of X and Y respectively

X	Y	Rank of X (R_1)	Rank of Y (R_2)	$d = R_1 - R_2$	d^2
10	50	2	1	1	1
10	80	2	5	-3	9

Scanned by CamScanner

10	70	2	2.5	- 0.5	0.25
15	70	5.5	2.5	3	9
14	80	4	5	1	1
15	80	5.5	5	0.5	0.25
20	100	7	7	0	0
				$\Sigma d = 0$	$\Sigma d^2 = 20.25$

$$R = 1 - \frac{\left\{\sum d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \ldots\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{20.25 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}\right\}}{7(7^2-1)}$$

$$= 1 - \frac{6(20.25 + 2 + 0.5 + 0.5 + 2)}{7 \times 48}$$

$$= 1 - \frac{6 \times 25.25}{336} = 1 - 0.45 = 0.55$$

Moderate positive correlation between x and y.

Example 17: The coefficient of rank correlation of marks obtained by 10 students in c programming and Introduction to IT and Economics was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the student was wrongly taken as 3 instead of 7. Find the correct rank correlation coefficient.

Solution

$n = 10, R = 0.5$, Wrong $d = 3$, Correct $d = 7$

We know

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

or $$0.5 = 1 - \frac{6\sum d^2}{10(10^2-1)}$$

or $$\frac{6\sum d^2}{990} = 1 - 0.5$$

or $$\sum d^2 = \frac{0.5 \times 990}{6} = 82.5$$

Correct $\Sigma d^2 = 82.5 - 3^2 + 7^2 = 122.5$

$$\text{Correct } r_s = 1 - 6\frac{\text{Correct } \sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 122.5}{10(10^2-1)} = 1 - 0.742 = 0.258$$

Hence correct value of correlation coefficient is 0.258.

Example 18 The coefficient of rank correlation between the marks in Mathematics and Statistics obtained by a group of students is 9/11 and the sum of the squares of the difference in ranks is 30. Find the number of students in the group.

Solution

Here,

$R = \frac{9}{11}, \Sigma d^2 = 30, n = ?$

Scanned by CamScanner

We know

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

or $\frac{9}{11} = 1 - \frac{6\times30}{n(n^2-1)}$

or $\frac{180}{n(n^2-1)} = 1 - \frac{9}{11}$

or $\frac{180}{n(n^2-1)} = \frac{2}{11}$

or $n^3 - n = 990$

or $n^3 - n - 990 = 0$

or $n^3 - 10n^2 + 10n^2 - 100n + 99n - 990 = 0$

or $n^2(n-10) + 10n(n-10) + 99(n-10) = 0$

or $(n-10)(n^2 + 10n + 99) = 0$

i) $(n-10) = 0$

ii) $(n^2 + 10n + 99) = 0$

From (i) $n = 10$

Example 19: Ten programmers in a presentation are ranked by three experts in the following order.

Programmers	1	2	3	4	5	6	7	8	9	10
Rank by expert I	1	6	5	10	3	2	4	9	7	8
Rank by expert II	3	5	8	4	7	10	2	1	6	9
Rank by expert III	6	4	9	8	1	2	3	10	5	7

Use the method of rank correlation to find which pair of expert have the nearest approach of ranking.

Solution

Competitors	Rank by expert I (R_1)	Rank by expert II (R_2)	Rank by expert III (R_3)	$d_{12} = R_1 - R_2$	d_{12}^2	$d_{13} = R_1 - R_3$	d_{13}^2	$d_{23} = R_2 - R_3$	d_{23}^2
1	1	3	6	-2	4	-5	25	-3	9
2	6	5	4	1	1	2	4	1	1
3	5	8	9	-3	9	-4	16	-1	1
4	10	4	8	6	36	2	4	-4	16
5	3	7	1	-4	16	2	4	6	36
6	2	10	2	-8	64	0	0	8	64
7	4	2	3	2	4	1	1	-1	1
8	9	1	10	8	64	-1	1	-9	81
9	7	6	5	1	1	2	4	1	1
10	8	9	7	-1	1	1	1	2	4
				$\Sigma d_{12}=0$	$\Sigma d_{12}^2 = 200$	$\Sigma d_{13} = 0$	$\Sigma d_{13}^2 = 60$	$\Sigma d_{23} = 0$	$\Sigma d_{23}^2 = 214$

Now, Spearman's rank correlation coefficient between expert I and expert II

$$R_{(I,II)} = 1 - \frac{6\Sigma d_{12}^2}{n(n^2-1)} = 1 - \frac{6\times200}{10(100-1)} = -0.212$$

Spearman's rank correlation coefficient between expert I and expert III

Scanned by CamScanner

$$R_{(I, III)} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)} = 1 - \frac{6\times 60}{10(100-1)} = 0.636$$

Spearman's rank correlation coefficient between expert II and expert III

$$R_{(II, III)} = 1 - \frac{6\sum d_{23}^2}{n(n^2-1)} = 1 - \frac{6\times 214}{10(100-1)} = -0.296$$

Here, $R_{(I, III)} = 0.636$ is positive correlation, hence expert I and III have common approach of ranking.

Merits and Demerits of Rank Correlation

Merits:

1. Spearman's rank correlation coefficient is easy to understand and apply as compared with Karl Pearson's formula.
2. This is the only the method for finding the correlation coefficient while dealing with qualitative characteristics such as beauty, general knowledge, honesty etc.
3. It can be used when data are highly skewed.

Demerits:

1. It is not practicable in the case of bi-variate frequency distribution.
2. There is always some loss of information due to the ranking is used.
3. It is not suitable in case of large number of observations.

Regression Analysis

The term regression means stepping back towards the average. It is a statistical tool used to determine the nature of relationship that exists among two or more variables and making estimate or prediction from that relationship. The unknown variable that we are going to estimate is called dependent variable or explained variable or regressed and the known variable is called independent variable or regressor or explanatory variable. For example let there is a high correlation between day temperature and sales of cold drinks then the salesman of cold drinks might wish to know the forecasted temperature for the next day to decide for the stock of cold drinks. This can be done with the help of regression. When two variable xand y are highly correlated one can find the best estimated value of x for given value of y or best estimated value of y for given value of x.

The principle difference of correlation analysis and regression analysis is that the correlation analysis indicates to what extent the variables are related while the regression analysis indicates how the variables are related.

Linear and Non-linear Regression

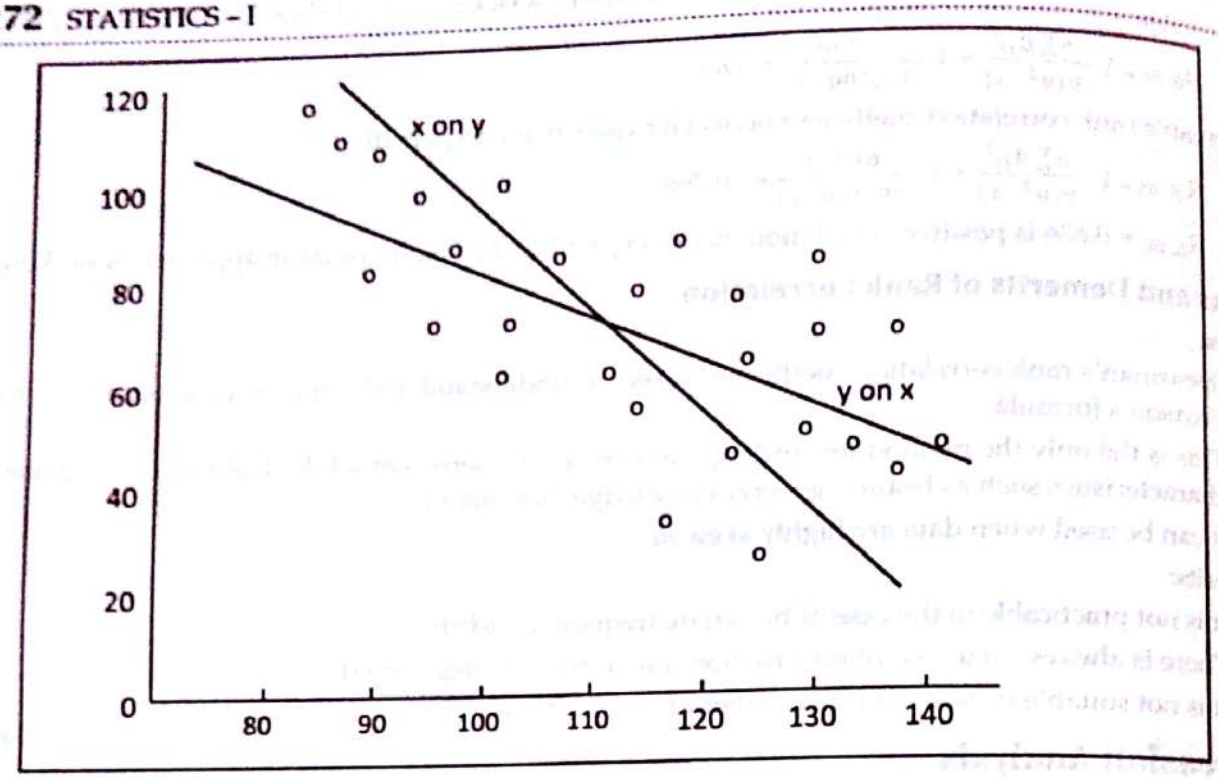
When the graph between dependent and independent variable is linear trend then the regression is called linear regression. When the graph between dependent and independent variable is not linear trend then the regression is called nonlinear regression.

Lines of Regression

In case of two variables x and y we have two regression lines. One is regression line of y on x and other is regression line of x on y. The regression line of y on x estimates the value of y for given value of x. The regression line of x on y estimates the value of x for given value of y.

If there is perfect correlation($r = \pm 1$) between two variables the two regression lines will coincide with each other. The regression lines are farther from each other as lesser degree of correlation between the variables. The regression lines are nearer to each other as high degree of correlation between the variables. In case of independent variables regression lines will be perpendicular to each other.

Scanned by CamScanner



Regression Equations

The algebraic expressions of the regression lines are called regression equations. For two variables in which one is independent and other is dependent variable there are two regression lines. Hence there are two regression equations.

i. Regression equation of y on x given by y = a+bx in which y is dependent variable and x is independent variable.

ii. Regression equation of x on y given by x = a+by in which x is dependent variable and y is independent variable.

Regression Equation of y on x

Let us suppose in bivariate distribution (x_i, y_i), i = 1,2,3...n dependent variable is y and independent variable is x. Regression equation of y on x is y = a+bx

Here a and b are constants and are estimated by using the principle of least square by minimizing error sum of square.

Simple Linear Regression

It is the linear function of a dependent variable with independent variable. With the help of independent variable the values of dependent variable can be predicted.It is statistical tool for finding the linear relationship of a dependent variable with independent variable. If the variables in bivariate distribution are related in linear form then the regression is linear otherwise curvilinear. It gives the best estimate to the value of one variable for any specific value of other variable.

Let us consider bivariate distribution(x_i, y_i); i = 1,2,3,.....n; y is dependent variable and x is independent variable then regression equation of y on x is given by y = a+bx, where a and b are constants and called y intercept and regression coefficient or slope respectively. b measures amount of change in y per unit change in x.

Assumptions of Linear Regression

Let us consider multiple regression modely = a + bx + ε

Scanned by CamScanner

There are certain assumptions about the model. The assumptions are based on relation between error ε and explanatory variables x

The following are the major assumptions on the random errors ε which is considers a serious problem in modeling if any of them is violated by the error term.

i. Regression model is linear in parameters.
ii. ε is random real variable
iii. The random errors ε have zero mean, i.e., $E(\varepsilon) = 0$
iv. The random errors ε has constant variance i.e.,$V(\varepsilon) = \sigma_\varepsilon^2$(No heteroscedaticity).
v. The random variable ε is normally distributed. i.e., $\varepsilon \sim N(0, \sigma^2)$
vi. The random errors ε are independent i.e., $E(\varepsilon_i\varepsilon_j) = 0 : i \neq j$. (No autocorrelation).
vii. x are uncorrelated to the error term ε, i.e., $E(X\varepsilon) = 0$ (uniformity of x over samples
viii. The explanatory variable x measured without error.

Estimation of Coefficients using Least Square Method (OLS)

The values of a and b are determined by using the principle of least square by minimizing error sum of square.

Here, error(e) = $y - \hat{y}$ so that $\Sigma e^2 = \Sigma(y - \hat{y})^2$

Let $S = \Sigma e^2 = \Sigma(y-\hat{y})^2 = \Sigma(y - a-bx)^2$

Differentiating both sides with respect to a

$$\frac{dS}{da} = \frac{d\Sigma(y-a-bx)^2}{da} = 2\Sigma(Y-a-bx)(-1) = -2\Sigma(y - a-bx)$$

For S to be minimum

$$\frac{dS}{da} = 0$$

or $-2\Sigma(y - a - bx) = 0$

or $\Sigma(y - a - bx) = 0$

or $\Sigma y - na - b\Sigma x = 0$

or $\Sigma y = na + b\Sigma x$ .....(i)

Differentiating both sides with respect to b

$$\frac{dS}{db} = \frac{d\Sigma(y-a-bx)^2}{db} = 2\Sigma(y - a - bx)(-x) = -2\Sigma(yx - ax - bx^2)$$

For S to be minimum

$$\frac{dS}{db} = 0$$

or $-2\Sigma(yx - ax - bx^2) = 0$

or $\Sigma(yx - ax - bX^2) = 0$

or $\Sigma yx - a\Sigma x - b\Sigma x^2 = 0$

or $\Sigma yx = a\Sigma x + b\Sigma x^2$ .....(ii)

Solving normal equations i and ii get a and b and substitute value to get the regression equation

$$y = a + bx$$

In equation $Y = a + bx$; a = intercept,

b = regression coefficient (change in y per unit change in x) = slope

Linear Relationship

A relationship of direct proportionality that, when plotted on a graph, traces a straight line. In linear relationships, any given change in an independent variable will always produce a corresponding change in the dependent variable.

Scanned by CamScanner

For example, a linear relationship between production hours and output in a factory means that a 10 percent increase or decrease in hours will result in a 10 percent increase or decrease in the output.

Deviation method

$u = x - A$ and $v = y - B$

Obtain regression equation of v on u i.e.,$v = a+bu$

Obtain a and b using principle of least square by minimizing error sum of square solving

$\sum v = na + b\sum u$ ..... (i)

$\sum uv = a\sum u + b\sum u^2$ ..... (ii)

Substitute value of a and b in $v = a+bu$ to get equation of v on u

Substitute value of $u = x - A$ and $v = y - B$ to get equation $y = a+bx$

Step Deviation Method

$u' = \frac{x-A}{h}$ and $v' = \frac{y-B}{k}$

Obtain regression equation of v' on u' i.e., $V = a+bu'$

Obtain a and b using principle of least square by minimizing error sum of square solving

$\sum v' = na + b\sum u'$ ..... (i)

$\sum u'v' = a\sum u' + b\sum u'^2$ ..... (ii)

Substitute value of a and b in $v' = a+bu$ to get equation of v' on u'

Substitute value of$u' = \frac{x-A}{h}$ and $V' = \frac{y-B}{k}$ to get equation $y = a+bx$

For regression equation of x on y

Equation be $x = a + by$

To estimate a and b using principle of least square by minimizing error sum of square

$\sum x = na + b\sum y$ .....(i)

$\sum xy = a\sum y + b\sum y^2$ .....(ii)

Solving (i) and (ii) get a and b and substitute in $x = a+by$

Deviation method

$u = x - A$ and $v = y - B$

Obtain regression equation of u on v i.e.,$u = a+bv$

Obtain a and b using principle of least square by minimizing error sum of square solving

$\sum u = na + b\sum v$ .....(i)

$\sum uv = a\sum v + b\sum v^2$ .....(ii)

Substitute value of a and b in $u = a+bv$ to get equation of u on v

Substitute value of $u = x - A$ and $v = y - B$ to get equation $x = a+by$

Step deviation method

$u' = \frac{x-A}{h}$ and $v' = \frac{y-B}{k}$

Obtain regression equation of u′ on v′ i.e $u' = a+bv'$

Obtain a and b using principle of least square by minimizing error sum of square solving

$\sum u' = na + b\sum v'$ .....(i)

$\sum u'v' = a\sum v' + b\sum v'^2$ .....(ii)

Substitute value of a and b in $u' = a+bv'$ to get equation of u′ on v′

Substitute value of$u' = \frac{x-A}{h}$ and $v' = \frac{y-B}{k}$ to get equation $x = a+by$

Scanned by CamScanner

Properties of Regression Coefficients

i) Correlation coefficient is the geometric mean between the regression coefficients.
$r = (b_{YX} \times b_{XY})^{1/2}$

ii) If one of the regression coefficients is greater than unity then other must be less than unity.

iii) Product of two regression coefficients must be less than or equal to 1.

iv) Arithmetic mean of regression coefficients is greater than the correlation coefficient.

v) Regression coefficients are independent of change of origin but not of scale.

vi) In regression equation of y on x, y = a+bx , b is slope of the line and is called regression coefficient of y on x. It measures change in dependent variable y per unit change in independent variable x. It is given by

$$b_{yx} = \frac{\text{Covariance}(x,y)}{\text{Variance }(x)} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Example 20 For the 10 observations, following information were obtained: $\Sigma X = 130$, $\Sigma Y = 220$, $\Sigma X^2 = 2238$, $\Sigma Y^2 = 5506$, $\Sigma XY = 3467$.

Obtain the line of regression of Y on X and estimate Y when X is 16?

Solution

Here, Y = ? When X = 16

We know

$\sum Y = na + b\sum X$

Or, $220 = 10a + 130b$ .... (i)

$\sum XY = a\,\Sigma X + b\,\Sigma X^2$

Or, $3467 = 130a + 2238b$ .... (ii)

Multiply (i) by 13 and subtract from (ii)

$3467 - 2860 = 130a + 2238b - 130a - 1690b$

Or, $607 = 548b$

Or, $b = 1.109$

Substitute b in (i)

$220 = 10a + 130 \times 1.109$

Or, $220 = 10a + 144.17$

Or, $10a = 220 - 144.17$

Or, $a = \frac{75.83}{10} = 7.583$

Regression equation of Y on X is Y = a+bX

Or, $Y = 7.583 + 1.109X$

When X= 16, $Y = 7.583 + 1.109 \times 16 = 25.32$

Example 21: Atutor believes that there is a linear relationship between verbal test score of students(Y) and the number of library books checked out(X). Following are the data collected on 10 students.

X	12	15	3	7	10	5	22	9	13	7
Y	77	85	48	59	75	41	94	65	79	70

Scanned by CamScanner

The above data reveals the following results;
$\Sigma X = 103$, $\Sigma Y = 693$, $\Sigma X^2 = 1335$, $\Sigma Y^2 = 50447$, $\Sigma XY = 7881$. Fit a simple linear regression model of Y on X. Interpret the slope regression coefficient

Solution

To fit Y = a+bX

$$\Sigma Y = na + b\,\Sigma X$$

or $693 = 10a + 103b$ ....(i)

$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2$$

or $7881 = 103a + 1335b$ ....(ii)

Coeff. of a	Coeff. of b	Constant
10	103	693
103	1335	7881

$$D = \begin{vmatrix} 10 & 103 \\ 103 & 1335 \end{vmatrix} = 13350 - 10609 = 2741$$

$$D_1 = \begin{vmatrix} 693 & 103 \\ 7881 & 1335 \end{vmatrix} = 925155 - 811743 = 113412$$

$$D_2 = \begin{vmatrix} 10 & 693 \\ 103 & 7881 \end{vmatrix} = 78810 - 71379 = 7431$$

Now, $a = \frac{D_1}{D} = \frac{113412}{2741} = 41.37$

$$b = \frac{D_2}{D} = \frac{7431}{2741} = 2.711$$

Regression equation of y on x

$$y = a + bx = 41.37 + 2.711x$$

Here b = 2.711 it means y changes by 2.711 per unit change in x.

Example 22. From the following data obtain the regression equation current in amperes and resistance in ohms using the method of least square and estimate the probable value of current if resistance is 4 ohms.

Resistance in ohms	2	3	5	6	8
Current in amperes	10	8	7	3	1

Solution

Resistance in ohms x	Current in amperes y	x^2	y^2	xy
2	10	4	100	20
3	8	9	64	24
5	7	25	49	35
6	3	36	9	18
8	1	64	1	8
$\Sigma x = 24$	$\Sigma y = 29$	$\Sigma x^2 = 138$	$\Sigma y^2 = 223$	$\Sigma xy = 105$

To fit Y = a+bX

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

or $29 = 5a + 24b$ .....(i)

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

or $105 = 24a + 138b$ .....(ii)

Scanned by CamScanner

Multiply equation i by 24 and ii by 5 and subtract

$$696 = 120a + 576b$$
$$\underline{_525 = _120a + _690b}$$
$$171 = 0 - 114b$$

$$\text{or, } b = \frac{-171}{114} = -1.5$$

Substitute b in i

$29=5a+24(-1.5)$

or, $29 + 36 = 5a$

$$\text{or, } a = \frac{65}{5} = 13$$

Regression equation of y on x is $y = a + bx$

or $y = 13 + (-1.5)x = 13 - 1.5x$

When $x = 4$, $y = 13 - 1.5\times4 = 7$

Hence value of Y is 7 when value of X is 4.

Example 23: The following table gives normal weight of a baby during the first six months of life.

Age in months	0	2	3	5	6
Weight in lbs	5	7	8	10	12

Determine the regression equation of weight on age using method of least square and estimate the weight of body at the age of 4 months.

Solution

Age in months(x)	Weight in lbs(y)	xy	x^2
0	5	0	0
2	7	14	4
3	8	24	9
5	10	50	25
6	12	72	36
$\Sigma x = 16$	$\Sigma y = 42$	$\Sigma xy = 160$	$\Sigma x^2 = 74$

To fit $Y = a+bx$

$\Sigma Y = na + b\Sigma x$

or $42 = 5a + 16b$ .....(i)

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

or $160 = 16a = 74b$ .....(ii)

From equation i and ii

Coeff of a	Coeff of b	Constant
5	16	42
16	74	160

$$\text{Now } D = \begin{vmatrix} 5 & 16 \\ 16 & 74 \end{vmatrix} = 114$$

$$D_1 = \begin{vmatrix} 42 & 16 \\ 160 & 74 \end{vmatrix} = 548$$

$$D_2 = \begin{vmatrix} 5 & 42 \\ 16 & 160 \end{vmatrix} = 128$$

Now,

$$a = \frac{D_1}{D} = \frac{548}{114} = 4.8 \quad b = \frac{D_2}{D} = \frac{128}{114} = 1.122$$

Scanned by CamScanner

Hence the regression equation is $y = a+bx = 4.8 + 1.122x$

When age $(x) = 4$

$y = 4.8 + 1.122 \times 4 = 9.288$

Hence estimated body weight is 9.288 lbs at age of 4months.

Example 24. From the following data between yearly turnover and profits. Find regression equation of profit and yearly turnover and estimate profit when yearly turnover is 30 million units.

Profit in thousand $	18	20	22	23	27	28	30
Yearly turnover in million units	23	25	27	30	32	31	35

Solution

Profit in thousand $ (Y)	Yearly turnover in million units(x)	v=y-23	u =x-30	uv	u^2
18	23	-5	-7	35	49
20	25	-3	-5	15	25
22	27	-1	-3	3	9
23	30	0	0	0	0
27	32	4	2	8	4
28	31	5	1	5	1
30	35	7	5	35	25
		$\sum v = 7$	$\sum u = -7$	$\sum uv = 101$	$\sum u^2 = 113$

First fit equation of v on u

$v = a+bu$

$\sum v = na + b\sum u$

or $7 = 7a - 7b$ .....(i)

$\sum uv = a\sum u + b\sum u^2$

or $101 = -7a + 113b$ .....(ii)

Add (i) and (ii)

$7+101 = 7a - 7b - 7a + 113b$

or $108 = 106b$

or $b = \frac{108}{106} = 1.018$

Substitute b in (i)

$7 = 7a - 7\times 1.018$

or $7 = 7a - 7.13$

or $a = \frac{14.13}{7} = 2.01$

Now,

$v = a+bu$

or $y-23 = 2.10 + 1.018(x - 30)$

or $y = 2.1 + 1.018x - 33.24 + 23$

or $y = -8.14 + 1.018x$

Regression equation of profit with yearly turnover is $y = -8.14 + 1.018x$

When yearly turnover(x) = 30,

$y = -8.14 + 1.018 \times 30 = 22.31 \approx 22$

Hence estimated profit is $22,000when yearly turnover is 30 million units.

Scanned by CamScanner

Example 25. Find the regression equation of X on Y from following data.

X	5	15	20	25	30
Y	50	60	80	110	130

Solution

X	Y	$u' = \frac{x-20}{5}$	$v' = \frac{y-80}{10}$	u'v'	v'^2
5	50	-3	-3	9	9
15	60	-1	-2	2	4
20	80	0	0	0	0
25	110	1	3	3	9
30	130	2	5	10	25
		$\Sigma u' = -1$	$\Sigma v' = 3$	$\Sigma u'v' = 24$	$\Sigma v'^2 = 47$

Now,

Regression equation of u' on v' is

$$u' = a + bv'$$

$$\sum u' = na + b\sum v'$$

or $-1 = 5a + 3b$ .....(i)

$$\sum u'v' = a\sum v' + b\sum v'^2$$

or $24 = 3a + 47b$ .....(ii)

Coeff. of a	Coeff. of b	Constant
5	3	-1
3	47	24

$$D = \begin{vmatrix} 5 & 3 \\ 3 & 47 \end{vmatrix} = 235 - 9 = 226$$

$$D_1 = \begin{vmatrix} -1 & 3 \\ 24 & 47 \end{vmatrix} = -47 - 72 = -119$$

$$D_2 = \begin{vmatrix} 5 & -1 \\ 3 & 24 \end{vmatrix} = 120 + 3 = 123$$

Now, $a = \frac{D_1}{D} = -\frac{119}{226} = -0.526$

$$b = \frac{D_2}{D} = \frac{123}{226} = 0.544$$

Regression equation of u' on v'

$$u' = a + bv'$$

$$u' = -0.526 + 0.554\, v'$$

or $\frac{x-20}{5} = -0.526 + 0.554\left[\frac{y-80}{10}\right]$

or $\frac{x-20}{5} = \frac{-5.26 + 0.554\,y - 44.32}{10}$

or $X - 20 = \frac{-49.58 + 0.554\,y}{2}$

or $2x - 40 = -49.58 + 0.554y$

or $2x = -9.58 + 0.554y$

or $x = -4.79 + 0.277y$

The required equation of X on Y is $X = -4.79 + 0.272Y$

Scanned by CamScanner

Difference between correlation and regression

Correlation	Regression
1. It measures the degree to which the variables are linearly related.	1. It measures the average relationship between variables whether variables are linearly related or not.
2. It is symmetric i.e. $r_{YX} = r_{XY}$	2. It is not symmetric. i.e. $b_{YX} \neq b_{XY}$.
3. It need not imply cause and effect relationship between the variables under study.	3. It clearly indicate the cause and effect relationship between the variables under study.
4. It is pure number independent of unit of measurement.	4. It is not pure number attached with the unit of measurement.
5. It has limited application as compared to regression.	5. It has wide applications.
6. It is relative measure.	6. It is absolute measure.
7. There may be non sense correlation.	7. There may not be non sense regression.

Measures of Variation

In regression model value of dependent variable are estimated on the basis of independent variables. In regression analysis total variation is divided into explained variation (sum of square due to regression) and unexplained variation (sum of square due to error). Hence according to Fisher total sum of square is decomposed into sum of square due to regression and sum of square due to error.

Total sum of square (TSS) = Sum of square due to regression (SSR) + Sum of square due to error (SSE)

For the regression model Y= a + bX, where Y is dependent variable and X is independent variable.

$TSS = \Sigma(y-\bar{y})^2 = \Sigma Y^2 - n\bar{y}^2$ = Total variation

$SSR = \Sigma(\hat{Y} - \bar{Y})^2 = \dfrac{(n\sum XY - \sum X \sum Y)^2}{n\{n\sum X^2 - (\sum X)^2\}}$ = Explained variation

(SSR is also given by $SSR = a\Sigma Y + b\Sigma XY - n\bar{Y}^2$)

SSE = TSS - SSR = Unexplained variation

(SSE also can be calculated by using the formula $SSE = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$)

Standard Error of the Estimate

Standard error is the Square root of the variance computed from sample data. The standard error of the estimate measures the average variation or scatterness of the observed data point around regression line. Standard error of the estimate is used to measure the reliability of the regression equation. Regression line having less standard error of estimate is more reliable than regression line having more standard error of estimate.

It is given by $S_e = \sqrt{\dfrac{SSE}{n-k-1}}$

SSE = sum of square due to error

k = number of independent variable in regression model

n = number of observations.

When $S_e = 0$, there is no variation of observed data around regression line .In such case regression line is perfect for estimating the dependent variable.

Scanned by CamScanner

Coefficient of Determination

It measures the proportion of variation in dependent variable that is explained by the set of independent variables .It is the measure based upon measure of variation and is used to determine the fitness of the data to the model. The regression line is reliable if the sum of square due to regression is much greater than sum of square due to error. It is the ratio of sum of square due to regression to the total sum of square. It is denoted by R^2 and is given by, $R^2 = \frac{SSR}{TSS}$.

For regression equation of y on x

$TSS = \Sigma(y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$

$SSE = \Sigma(y - \hat{y})^2 = \Sigma Y^2 - a\,\Sigma Y - b\,\Sigma XY$

$SSR = TSS - SSE$

It is also obtained by simply squaring the correlation coefficient.i.e. $R^2 = r^2$. Higher the value of R^2 the more reliable is the fitted equation .It lies between 0 and 1.

R^2 can never decrease when another independent variable is added to a regression. R^2 will usually increase with increase in number of independent variables.

Example 26: Calculate the coefficient of determination for following data.

X	50	55	55	60	65	70	65	60
Y	11	13	14	16	16	15	15	20

Solution

X	Y	X^2	Y^2	XY
50	11	2500	121	550
55	13	3025	169	715
55	14	3025	196	770
60	16	3600	256	960
65	16	3600	256	1040
70	15	4900	225	1050
65	15	4225	225	975
60	20	3600	400	1200
$\Sigma X = 480$	$\Sigma Y = 120$	$\Sigma X^2 = 29100$	$\Sigma Y^2 = 1848$	$\Sigma XY = 7260$

We have,

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{8\times7260 - 480\times120}{\sqrt{8\times29100 - (480)^2}\sqrt{8\times1848 - (120)^2}}$$

$$= \frac{480}{\sqrt{232800 - 230400}\sqrt{14787 - 14400}} = \frac{480}{48.989\times19.596} = 0.50$$

Now coefficient of determination $R^2 = r^2 = (0.5)^2 = 0.25$

Example 27. From 5 pair of observations it was found that $\Sigma X = 16$, $\Sigma Y = 42$, $\Sigma XY = 160$, $\Sigma Y^2 = 382$ and $\Sigma X^2 = 74$. Find coefficient of determination.

Solution

We know

Scanned by CamScanner

$$TSS = \Sigma(y-\bar{y})^2 = \Sigma y^2 - n\bar{y}^2 = 382 - 5\left(\frac{42}{5}\right)^2 = 382 - 352.8 = 29.2$$

To find a and b

$\Sigma Y = na + b\Sigma X$

or $42 = 5a + 16b$ .....(i)

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

or $160 = 16a + 74b$ .....(ii)

Coeff of a	Coeff of b	constant
5	16	42
16	74	160

$$D = \begin{vmatrix} 5 & 16 \\ 16 & 74 \end{vmatrix} = 370 - 256 = 114$$

$$D_1 = \begin{vmatrix} 42 & 16 \\ 160 & 74 \end{vmatrix} = 3108 - 2560 = 548$$

$$D_2 = \begin{vmatrix} 5 & 42 \\ 16 & 160 \end{vmatrix} = 800 - 672 = 128$$

Now, $a = \frac{D_1}{D} = \frac{548}{114} = 4.807$

$b = \frac{D_2}{D} = \frac{128}{114} = 1.122$

$$SSE = \Sigma(y-\hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy = 382 - 4.807 \times 42 - 1.122 \times 160$$
$$= 382 - 201.894 - 179.52 = 0.586$$

$$SSR = TSS - SSE = 29.2 - 0.586 = 28.614$$

Now $R^2 = \frac{SSR}{TSS} = \frac{28.614}{29.2} = 0.979$

Residual

The residual is the difference between observed and predicted value of dependent variable.

The OLS method is based on the assumption that errors are normally distributed.

These are based on the assumption of regression which are as follows:

Assumptions of linear regression

Theory of regression assumes that certain assumptions should hold for a reliable and acceptable regression analysiss. If one or more assumptions are not satisfied or violated the regression will have specific problem. The major assumptions are as described below.

Let us consider multiple regression model

$$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

There are certain assumptions about the model. The assumptions are based on relation between error ε and explanatory variables x_i's.

i. Regression model is linear in parameters.

ii. ε is random real variable

iii. The random errors ε have zero mean, i.e. $E(\varepsilon) = 0$

iv. The random errors ε has constant variance ie. $E(\varepsilon) = \sigma^2$ (Noheteroscedaticity).

v. The random variable ε is normally distributed. i.e. $\varepsilon \sim N(0, \sigma^2)$

vi. The random errors ε are independent i.e. $E(\varepsilon_i\varepsilon_j) = 0 : i \neq j$. (No autocorrelation).

vii. X are uncorrelated to the error term ε, ie. $E(X\varepsilon) = 0$ (uniformity of X over samples

viii. The explanatory variables x_i's are measured without error.

ix. The number of observations must be greater than the number of explanatory variables.

Scanned by CamScanner

x. The explanatory variables X_i's are not perfectly linearly correlated (No multicollinearity)

To test the normality of errors i.e., residual different types of method can be used. Some of them are residual plot like normal probability plot, P-P plot, Q-Q plot etc. (these are discussed earlier)

Application of Correlation and Regression in Excel, SPSS and STATA

Example 28: Calculate Karl Pearson's correlation coefficient test its significance and find the limits of population correlation coefficient. Find coefficient of determination.

Nutrition	Child mortality
12.1	9.5
9.1	9.2
26	11.8
6.4	6.4
9.5	7.3
18.5	20.3
22.8	24.4
17.4	21.1
13.9	10.7
3.2	3.5
30.2	11.8
15.7	12.3
8.7	11.8
5.6	9.4
11.2	8.3
9.8	9
8.4	4.7

Note: For calculation of Karl Pearson's correlation, theoretical formula and steps can be followed or the formula built in excel or data analysis too can be used.

Scanned by CamScanner

Source	SS	df	MS
Model	319328819	3	106442940
Residual	257468140	65	3961048.3
Total	576796959	68	8482308.22

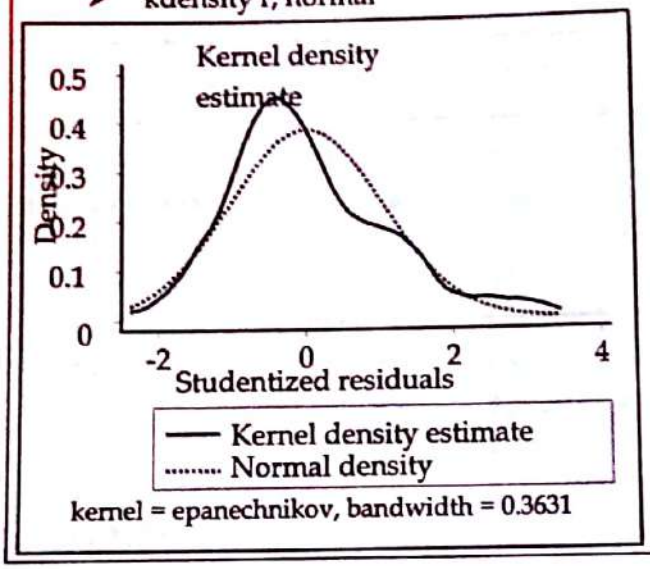
Number of obs	=	69
F(3, 65)	=	26.87
Prob > F	=	0.0000
R-squared	=	0.5536
Adj R-squared	=	0.5330
Root MSE	=	1990.2

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
foreign	3639.206	682.1513	5.33	0.000	2276.855	5001.556
length	-97.67348	33.28417	-2.93	0.005	-164.1466	-31.2004
weight	6.042855	.990284	6.10	0.000	4.065121	8.020588
_cons	5107.275	3776.33	1.35	0.181	-2434.577	12649.13

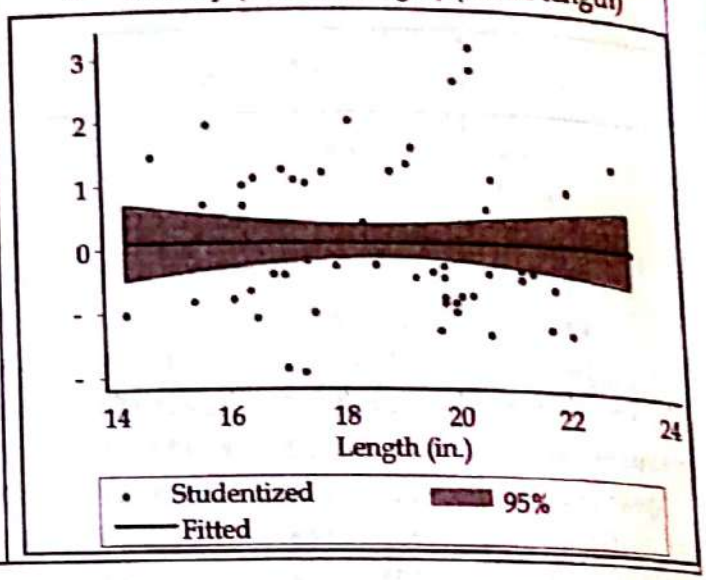
d) >predit p *generates **predictedvariable** as p
>predict r, rstudent *generates **studentizedresidual** variable as r

e) >list make if r<2 *generates only five auto names with r less than 2

➢ kdensity r, normal



➢ twoway (scatter r length) (lfitci r length)



Exercise 7

Theoretical Questions

1. What do you mean by correlation? Describe different measures of correlation.
2. Distinguish between positive correlation and negative correlation.
3. Explain how Karl Pearson's correlation coefficient differs from Spearman's rank correlation coefficient?
4. Explain the suitable conditions for applying Karl Pearson's correlation coefficient and Spearman's rank correlation coefficient.
5. Write down properties of correlation coefficient.

Scanned by CamScanner

6. What do you mean by significance of correlation? How can you interpret the correlation coefficient with the help of probable error?
7. What do you mean by regression? Interpret the regression coefficient.
8. What do you mean by regression coefficient? Write down its properties.
9. Differentiate between correlation and regression.
10. Describe the use of regression analysis in computer science and Information technology.
11. How linear regression is different from nonlinear regression?
12. What do you mean by coefficient of determination? How can you interpret it?

Numerical Questions

1. If the covariance between x and y variable is 36 and variance of X and Yare 36 and 100 respectively, find the coefficient of correlation between them. **[Ans: 0.6]**
2. Find the correlation coefficient between x and y series

	X	Y
Number of observation	10	10
Standard deviation	2.05	2.06

And $\Sigma(X-\bar{X})(Y-\bar{Y}) = 40$ [Ans: 0.947]

3. For 10 observations on two variables X and Y, the following information are as follows:
$\Sigma X = 666$, $\Sigma Y = 663$, $\Sigma X^2 = 44{,}490$, $\Sigma Y^2 = 44{,}061$, $\Sigma XY = 44{,}224$
Compute, Karl Pearson's coefficient of correlation. [Ans: 0.576]
4. From the following information find the total number of pair of observations given that $r = 0.8$, $\Sigma xy = 60$, S.D. of Y = 60 and $\Sigma x^2 = 90$, where x and y are deviations taken from their respective means. [Ans: 1636]
5. From the following data examine whether there exists any correlation between X and Y

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

[Ans: 0.95]

6. The following table gives the distribution of items and also defective items among them according to size groups. Find the correlation coefficient between size and defect in quality.

Size group	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
No of items	200	270	340	360	400	300
No of defective items	150	162	170	180	180	120

[Ans: –0.939]

7. Following figure give the age in years of newly married husbands and wives (25,17) (26,18) (27,19) (25, 17) (26,19) (28,20) (25,17) (25,17) (24,18) (26,18) (26,20) (27,18) (27,19) (28,19) (25,18) (25,19) (26,18) (25,18) (27,20). Find Karl Pearson's Correlation coefficient. Test its significance. [Ans: 0.654]
8. From 20 pairs of X and Y variables the following results obtained
$\Sigma X = 127$, $\Sigma Y = 100$, $\Sigma X^2 = 860$, $\Sigma Y^2 = 549$, $\Sigma XY = 674$
at the time of verification, the following wrong values of X and Ywere taken as X = 10, 8 and Y= 14,6 instead of correct values X = 8,6 and Y= 12, 8. Find correct value of correlation. [Ans: 0.47]
9. A student calculates the value of r as 0.795 when the number of items is 100 and concludes that r is highly significant. Is his conclusion correct? [Ans: Yes]

Scanned by CamScanner

10. Correlation coefficient between two variables with the pair of 10 observations is 0.81. Discuss if the value of r be significant or not. Also determine the limits of population correlation coefficient. [Ans: Significant,0.736 to 0.883]

11. The information given below are related with ages of husband(X) and wife(Y) for married couples living together in a sample survey. Calculate the coefficient of correlation between age of husband and that of his wife. Test the significance of the calculated r.

$N = 72, \Sigma fX = 3560, \Sigma fX^2 = 196800, \Sigma fY = 3260, \Sigma fY^2 = 168400, \Sigma fXY = 172000$

[Ans: 0.52,Significant]

12. The ranking of Ten Competitors in voice test given by two experts A and B are as follows:

A	3	5	8	4	7	10	2	1	6	9
B	6	4	9	8	1	2	3	10	5	7

Calculate correlation coefficient between the rank of A and B. [Ans: –0.296]

13. From the following data calculate Spearman's rank correlation.

Variable (X)	10	15	18	14	30	27
Variable (Y)	100	120	118	119	130	105

[Ans: 0.48]

14. Find rank correlation coefficient between age of husband and age of wife from following data:

Age of husband (yrs)	23	27	28	29	30	31	33	35
Age of wife (yrs)	21	22	23	24	25	26	28	29

[Ans: 1]

15. Calculate Spearman's rank correlation coefficient from the following

X	20	25	60	45	80	25	55	65	25	75
Y	52	50	55	50	60	70	72	78	80	63

[Ans: 0.187]

16. An examination of 10 applicants was taken by a company on skill and ability of candidates. From the marks obtained by the applicant in Skill and Ability Papers. Calculate the rank correlation coefficient.

Applicant	A	B	C	D	E	F	G	H	I	J
Marks in Skill	38	41	68	41	38	55	85	81	28	41
Marks in Ability	48	39	38	36	58	61	72	83	61	82

[Ans: 0.278]

17. Ten competitors in a competition are ranked by three experts in the following order.

1st expert	1	5	4	8	9	6	10	7	3	2
2nd expert	4	8	7	6	5	9	10	3	2	1
3rd expert	6	7	8	1	5	10	9	2	3	4

Use the rank correlation coefficient to discuss which pair of experts has the nearest approach to competitors. [Ans: 2nd and 3rd]

18. Following data represent the preference of 10 students studying B.Sc. CSIT towards two brands of computers namely DELL and HP.

Computer	Student preference									
DELL	5	2	9	8	1	10	3	4	6	7
HP	10	5	1	3	8	6	2	7	9	4

Scanned by CamScanner

Apply appropriate statistical tool to measure whether the brand preference is correlated. Also interpret your result. [Ans: -0.309]

19. The coefficient of rank correlation of the marks obtained by 10 students in mathematics and statistics was found to be 0.8. It was discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 9. Find the correct rank correlation coefficient. [Ans: 0.78]

20. Coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of the squares of differences in ranks is given to be 48, find the value of n. [Ans: 7]

21. A large company wants to measure the effectiveness of radio advertising media(x) on the sale promotion (y) of its products. A sample of 22 cities with approximately equal populations is selected for study. The sales of the product in thousand rupees and the level of radio advertising expenditure in thousand rupees are recorded for each of 22 cities. The sum, sum of square and sum of product of x and y are summarized below

$\sum y = 26953, \sum x = 950, \sum y^2 = 35528893, \sum x^2 = 49250, \sum xy = 1263940$

a) Fit a simple linear regression model of y on x using the least square method. Interpret the estimated slope coefficient

b) Compute R^2 and interpret. [Ans: y=699.957+12.162x, 0.4852]

22. The following measurements show the respective height in inches of 10 fathers and their eldest sons

Father	67	63	66	71	69	65	62	70	61	72
Son	68	66	65	70	67	67	64	71	62	63

Find the regression line of son's height on father's height and estimate the height of son for the given height of father as 70 inches. Also determine coefficient of determination and interpret. [Ans: y=40.43+0.388x, 67.62, 0.266]

23. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Operator	I	II	III	IV	V	VI	VII	VIII
Experience	16	12	18	4	3	10	5	12
Performance	87	88	89	68	78	80	75	83

Calculate the regression equation of performance on experience and hence estimate the probable performance if an operator has 8 years experiences. Interpret the regression coefficient. [Ans: y=69.669+1.133x,78.73]

24. A city council has gathered data on number of minor traffic accidents and the number of youth football games that occurred in town over the weekends.

X(football games)	20	30	10	12	15	25	34
Y(minor accidents)	6	9	4	5	7	8	9

(i) Develop the regression equation to predict minor accidents from football games.

(ii) Predict the number of minor traffic accidents that will occur at weekends during which X=30.

(iii) Calculate the value of coefficient of determination

(iv) Calculate the value of standard error of estimate. [Ans: y=2.732+0.198x, 9, 0.87, 0.77]

5. A chemical company wishing to study the effect of extraction time on the efficiency of an extraction operation obtained the data as follows

Extraction time in minute(X)	27	45	41	19	35	39	19
Extraction efficiency in %(Y)	57	64	80	46	62	72	52

Scanned by CamScanner

a) Fit a straight line to the given data by the method of least square and use it to predict the extraction efficiency one can expect when the extraction time is 35 minutes.

b) Determine the coefficient of determination and interpret its meaning.

[Ans: y=32.096+0.926x, 64.5, 0.843]

26. For the data given below i)Fit linear regression Y = a+bX by the method of least square and interpret regression coefficient ii) determine coefficient of determination and interpret.

X	0	5	10	15	20	25
Y	12	15	17	22	24	30

[Ans: y=11.29+0.697x, 0.97]

27. National Planning Commission(NPC) is performing preliminary study to determine the relationship between certain economic indicator and annual percentage change in Gross National Product(GNP). The concern is to estimate the percentage change in GNP. One of such indicator being examined is government's deficit. Data on 6 years are given below;

Percentage change in GNP	3	1	4	1	2	3
Government deficit in lakh Rs	50	200	70	100	90	40

a) Develop the estimating equation to predict percentage change in GNP from government deficit.

b) Interpret the estimated regression coefficient.

c) What percentage change in GNP would be expected in a year in which government deficit was Rs 110 lakh.

d) Compute the coefficient of determination and interpret.

[Ans: y=3.725 – 0.015x,2.055, 0.524]

28. The annual advertising expenditure (in lakh rs) and the corresponding annual sales (in crore rs) for the past 10 years of a company are presented in the following table.

Year	Annual advertising expenditure	Annual sales revenue
1	10	20
2	12	30
3	14	37
4	16	50
5	18	56
6	20	78
7	22	89
8	24	100
9	26	120
10	28	110

a. Find the correlation coefficient between annual advertising expenditure and annual sales revenue and comment the result

b. Develop the regression model of sales as a function of advertising expenditures.

c. Predict the value of annual sales while advertising expenditure was 27 lakh rupees.

[Ans: 0.985,y=-40.048+5.739x,114.915]

Career airline pilots face the risk of progressive hearing loss due to the noisy cockpits of most jet aircrafts. Much of the noise comes not from engines but from air roar which increases at

Scanned by CamScanner

high speeds. To assess this workplace hazard a pilot measured cockpit noise level(in decibels) and airspeed (knots indicated air speed). The data are shown in the given table

Speed	250	340	320	330	346	260	280	395	380	400
Noise level	83	89	88	89	92	85	84	92	93	96

a. Determine association between noise level and air roar which is increased due to high speed. Comment on strength of association

b. Develop a least square regression model to estimate the noise level with the help of speed of aircraft. Also interpret the regression coefficient.

[Ans: 0.957;y=64.191+0.075x]

30. A computer manager interested to know how efficiency of his/ her new computer program which depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. Applying the program to data sets of different sizes, the following data were gathered.

Data size(gigabytes)	6	7	7	8	10	10	15
Processed requests	40	55	50	41	17	26	16

a. Identify which one response variable and fit a simple regression line assuming that the relationship is linear

b. Interpret the regression coefficient with reference to your problem

c. Obtain the coefficient of determination and interpret this

d. Based on the fitted model predict the efficiency of new computer for data size 12(gigabites). Does it possible to predict efficiency for data size of 30 (gigabites)? Discuss. [Ans: y = 72.278- 4.142x, $R^2 = 0.661, 22.57, No$]

Practical Work

1. Omprakash Sharma, owner of the Kathmandu Precast Company, has hired you as a part-time analyst. He was extremely pleased when you uncovered a positive relationship between the number of building permits issued and the amount of work available to his company. Now he wonders if it's possible to use knowledge of interest rates on first mortgages to predict the number of building permits that will be issued each month. You collect a sample of data covering nine months.

Month	Building Permits (Y)	Interest rate (X)
1	786	10.2%
2	494	12.6
3	289	13.5
4	892	9.7
5	343	10.8
6	888	9.5
7	509	10.9
8	987	9.2
9	187	14.2

i. Calculate the correlation coefficient between building permits and interest rate and test its significance at 1%.

ii. Estimate the best fitting regression line and compute residual for month 9.

Scanned by CamScanner

Tribhuvan University
Institute of Science and Technology
Model Question

Bachelor Level/First Year/Second Semester/Science
Computer Science and Information Technology STA 164
(Statistics I)

Full Marks: 60
Pass Marks: 24
Time : 3 Hours

Candidates are required to give their answers in their own words as far as practicable.
All notations have the usual meanings.

Group A

Attempt any Two questions **(2×10=20)**

1. A new computer program consists of two modules. The first module contains an error with probability 0.2. The second module is more complex; it has a probability of 0.4 to contain an error, independently of the first module. An error in the first module alone causes the program to crash with probability 0.5. For the second module, this probability is 0.8. If there are errors in both modules, the program crashes with probability 0.9. Suppose the program crashed. What is the probability of errors in both modules?

2. Explain how box-plot is helpful to know the shape of the data distribution. The following data set represents the number of new computer accounts registered during ten consecutive days.

   43 37 50 51 58 105 52 45 45 10

   a) Compute the mean, median, quartiles, and sample standard deviation.
   b) Check whether there are outliers or not.
   c) If outliers are present, then delete the detected outliers and compute the mean, median, quartiles, and sample standard deviation again.
   d) Make your conclusion about the effect of outliers on descriptive statistical analysis.

3. A computer manager interested to know how efficiency of his/her new computer program which depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. Applying the program to data sets of different sizes, the following data were gathered.

Data size (gigabytes)	6	7	7	8	10	10	15
Processed requests	40	55	50	41	17	26	16

   a) Identify which one response variable, and fit a simple regression line, assuming that the relationship between them is linear.
   b) Interpret the regression coefficient with reference to your problem.
   c) Obtain coefficient of determination, and interpret this.
   d) Based on the fitted model in (a), predict the efficiency of new computer for data size 12 (gigabytes). Does it possible to predict efficiency for data size of 30 gigabytes? Discuss.

Group B

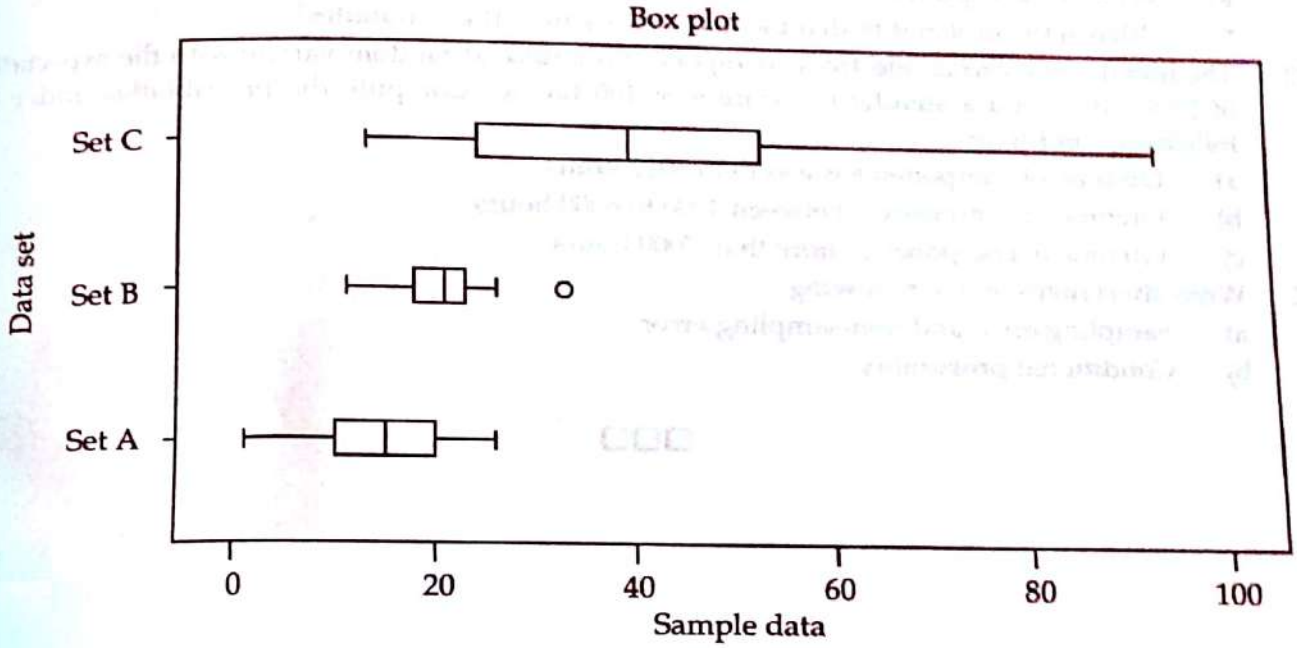
Attempt any Eight questions **(8×5 = 40)**

4. Explain the role of statistics in computer science and information technology.
5. Following table presents some descriptive statistics computed from three different independent sample dataset (X).

Scanned by CamScanner

Data	Sample size (n)	$\sum_{i=1}^{30} X_i$	Minimum	Q_1	Median	Q_3	Maximum	$\sum_{i=1}^{30} (X_i - \bar{X})^2$
Data set A	30	439	1	10	15	20	26	1348.97
Data set B	30	625	11	18	21	23	33	540.17
Data set C	30	1239	13	24	39.5	53	94	12836.3

a) Compare sample mean and median, and explain about the shape of the data dist for each dataset. Compare the variability of the three set of dataset. Box-plots have been generated through SPSS for each dataset as follows.



b) Do these box-plots support your findings obtained in a) about the shape of the distribution? Explain.

6. A large chain retailer purchases a certain kind of electronic device from a manufacturer. The manufacturer indicates that the defective rate of the device is 3%.
   a) The inspector randomly picks 20 items from a shipment. What is the probability that there will be at least one defective item among these 20?
   b) Suppose that the retailer receives 10 shipments in a month and the inspector randomly tests 20 devices per shipment. What is the probability that there will be exactly 3 shipments each containing at least one defective device among the 20 that are selected and tested from the shipment?

7. Messages arrive at an electronic message center at random times, with an average of 9 messages per hour.
   a) What is the probability of receiving at least five messages during the next hour?
   b) What is the probability of receiving exactly seven messages during the next hour?

8. The time, in minutes, it takes to reboot a certain system is a continuous variable with the density function:

$$f(x) = \begin{cases} C(10-x)^2, & 0 < x < 10 \\ 0, & \text{otherwise} \end{cases}$$

Compute C, and then compute the probability that it takes between 1 and 2 minutes to reboot.

9. Following data represent the preference of 10 students studying B.Sc .(CSIT) towards two brands of computers namely DELL and HP.

Scanned by CamScanner

Computer	Student preference									
DELL	5	2	9	8	1	10	3	4	6	7
HP	10	5	1	3	8	6	2	7	9	4

Apply appropriate statistical tool to measure whether the brand preference is correlated. Also interpret your result.

10. Define exponential distribution with parameter λ. The time required to reach to the printer after ordering in the computer follows exponential distribution at an average rate of 3 jobs per hour.
    a) What is the expected time between jobs?
    b) What is the probability that the next job is sent within 5 minutes?
11. The lifetime of a certain electronic component is a normal random variate with the expectation of 5000 hours and a standard deviation of 100 hours. Compute the probabilities under the following conditions
    a) Lifetime of components is less than 5012 hours
    b) Lifetime of components between 4000 to 6000 hours
    c) Lifetime of components more than 7000 hours
12. Write short notes on the following.
    a) Sampling error and non-sampling error
    b) Conditional probability

❑❑❑

Scanned by CamScanner