QUANTITATIVE TECHNIQUES FOR BUSINESS DECISIONS

Master of commerce

Semester I

Paper II

Study Material

2015 Admission onwards

UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION

CALICUT UNIVERSITY P.O., THENJIPALAM, MALAPPURAM-673635

2022

UNIVERSITY OF CALICUT

SCHOOL OF DISTANCE EDUCATION

Master of commerce

Study Material

2015 Admission onwards

Semester I

Paper II

QUANTITATIVE TECHNIQUES FOR BUSINESS DECISIONS

Prepared by Dr. Yakoob . c.
Reader and Research Guide,
SS COLLEGE, AREACODE.

Type settings and Lay out :
Computer Section, SDE

TABLE OF CONTENT

UNIT 1

QUANTITATIVE TECHNIQUES FOR MANAGERIAL DECISIONS

Introduction

Decision making is the most complex, but essential human activity. Several tools have been developed for facilitating decision making – whether in ordinary course of life or business. Good decisions are always applauded. Decisions determine the destiny of individuals or organizations.

Decisions can be based on both qualitative aspects and quantitative data. Certain situations warrant the introduction of elements of quantities which support judgment and intuition. Many of the decision circumstances of business organizations necessitate planning and evaluation of alternatives. Thus emerged the subject of quantitative techniques.

Business is becoming more and more complex and requires effective management to succeed. Managing complexity requires many professional skills including quantitative analysis. Business environment is being globalized, competitors are increasing, demand is getting diverse, and employees becoming more mobile and demanding. As a consequence, effective decision making is more crucial than ever before.

On the other hand, managers have more access to larger and more detailed data base that are potential sources of information. However, to achieve this potential, it is required that managers know how to convert data into information. This is one of the reasons why quantitative techniques are being more and more focused.

Definitions

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

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

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

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

Meaning and Nature

Quantitative techniques comprise those statistical and programming techniques which are helpful in managerial decision making. These techniques involve use of numbers, symbols and other mathematical expressions to express problems in quantitative terms. They serve as a useful supplement to human judgment and intuition. They prove a systematic means of analysis and choice for attaining predetermined objectives of an organization. Quantitative techniques involve use of scientific methods of experimentation, observation and modification, in managerial decision making process.

Features

Descriptions of quantitative techniques reveal following characteristics or features.

Measurement

Measurement is the basis of quantitative technique. Measurement is assigning numbers to concepts and phenomena. Measurement generates necessary data.

Numerical analysis

Another basic feature of quantitative techniques is numerical expression of variables and analysis thereto. Even qualitative characteristics or phenomenon can be transformed to numbers and symbols using quantitative techniques.

Scientific method

Quantitative techniques for decision making are examples for the use of scientific methods of management. It offers a systematic and objective experimentation, observation and evaluation of best strategies.

Decision making

It is a support system in decision making process. It provides decision makers with appropriate tools of evaluation and presentation.

Options

Quantitative techniques should evaluate and reveal alternative strategies or options. There is no scope for decision where there is a single option.

Improvement

Quantitative techniques should replace personal judgment and intuition. It should lead to improved and quality decisions.

Functions of quantitative techniques

Quantitative Techniques are those methods in which details of a problem or situation are expressed in numerical terms, so as to support decision making. Accordingly, following are the functions of Quantitative techniques

Quantification

Critical factors affecting a decision situation is transformed into quantitative or numerical form. It is easy to comprehend, understand and delegate an issue in numerical form.

Analysis

Quantitative techniques enable scientific and systematic study of any issue. It probes deep into the factors influencing the problem and helps to express the situation in a comprehensive form.

Decision making

Quantitative techniques facilitate the process of decision making. It sets out all possible alternatives and enables a feasibility study of each so that the optimal alternative can be chosen

Deployment of resources

Quantitative techniques, if properly applied, leads to optimal allocation of available limited resources. It avoids wastages and less efficient usage of resources, and leads to conservation of resources.

Sequencing

Certain projects may involve several complex activities , to be performed in a sequential order. Quantitative techniques aids in determining optimal sequence of performing a set of jobs, so as to minimize total process time.

Optimize service

In service sector, quantitative techniques is the only option for addressing questions like waiting time, service time, traffic intensity, idle time etc.

Role of quantitative techniques in decision making

Quantitative techniques have been increasingly used in decision making relating to solution of complex problems of business and industry. Quantitative techniques are now recognized as an effective tool for solving managerial problems. Its role is vital due to following reasons,

Better control

Management of bigger organizations find it much costlier to provide continuous executive supervision over routine decisions. A Quantitative techniques approach directs the executives to devote their attention to more pressing matters. For example quantitative techniques approach deals with production scheduling and inventory control.

Better coordination

Sometimes quantitative techniques have been very useful in maintaining the orderly situation out of chaos. For example, quantitative techniques based planning model becomes a vehicle for coordinating possibilities of marketing decisions with the limitations imposed on manufacturing capabilities.

Better system

Quantitative techniques study is also initiated to analyze a particular problem of decision making such as establishing a new warehouse. Later quantitative techniques approach can be further developed into a system to be employed repeatedly. Consequently the cost of undertaking the first application may improve the profits.

Better decisions

Quantitative techniques models frequently yield actions that do improve an intuitive decision making. Sometimes a situation may be so complicated that the human mind can never hope to assimilate all the important factors without the help of quantitative techniques and computer analysis.

Quantitative and qualitative approaches

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

Quantitative approach

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

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

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

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

Qualitative approach

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

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

Generally there are four non quantitative techniques of decision making

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

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

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

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

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

Significance of quantitative decisions

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

Simplifies decision making

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

Scientific analysis

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

Allocation of resources

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

Profit maximization

Quantitative techniques are invaluable in assessing the relative profitability of alternative choices and identifying the most profitable course of action. What should be the relative mix of different products, which site to choose for location out of alternative sites, which arrangement of orders in terms of time and quantity, will give maximum profits. Such question can be answered with the help of quantitative techniques.

Cost minimization

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

Forecasting

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

Inventory control

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

Applications of quantitative techniques in business operations

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

Planning

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

Purchasing

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

Manufacturing

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

Marketing

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

Human resource management

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

Research and Development

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

Classification of quantitative techniques

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

Mathematical techniques

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

Permutations and combinations

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

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

Set theory

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

Matrix Algebra

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

Differentials

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

Integration

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

Statistical techniques

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

Programming techniques

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

Quantification of qualitative data

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

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

Example : A carpet factory manufactures carpets of which minimum length is 15.1 mts and maximum is 16.9 mts. It produces carpets having length of 15.1 – 15.5 - 2 nos, 15.6 – 15.8 - 8 nos, 15.9 - 16.1 - 9 nos, 16.2 - 16.5 - 7 nos and 16.6 - 16.9 - 4 nos. it is convenient to present this information in the form of a frequency distribution as below:

Class	Frequency
15.1 - 15.5	2
15.6 - 15.8	8
15.9 - 16.1	9
16.2 - 16.5	7
16.6 - 16.9	4
Total	30

Review Questions and Exercises

1. Define Quantitative Technique.
2. Describe the various methods of classifying of Quantitative Techniques.
3. State the various Mathematical Quantitative Techniques.
4. State sources of the important Statistical techniques.
5. State various Operations Research Techniques.
6. Explain the role of Quantitative Techniques in business management
7. List out the important areas where Quantitative techniques have applications.
8. Discuss the Scope and limitations of Quantitative techniques
9. Explain the uses quantitative techniques in business

EX 1.1

An employment exchange gave following information about its registered candidates.

Level of education – not completed +2 = 35%, completed +2 31%, attended but not completed degree 16%, completed degree 9%, not completed PG 6% and completed PG 3%. Construct a relative frequency table and comment on the trend of registration.

Ex 1.2

The administrator of a hospital provided following information on waiting time in casually department. Construct a table on the waiting and comment on this.

Waiting time (minutes)	12	16	21	20	24	3	11	17	29	28
No of patients	26	4	7	14	25	1	27	15	16	5

UNIT II

INFERENTIAL ANALYSIS – POINT ESTIMATE

Introduction

One of the main objectives of statistical studies is to draw valid conclusion about the population on the basis of samples drawn from the population. Such a process of inferring about the population is called inferential analysis. Inferential analysis is often required and applied in business management.

Management is confronted with various practical problems like augmentation of production, maximization of profit, minimization of cost, introduction of innovations improvement of production methods etc. these problems lead to accomplishment of certain pre-determined objectives and goals.

There has been a growing tendency to turn to quantitative techniques as a means for solving many of these managerial decision problems that arise in a business or industrial enterprise. A large number of business problems have been given quantitative representation with considerable degree of success. Inferential analysis is such a quantitative e technique widely applied for managerial decision taking.

Inferential analysis

Inferential analysis is a prominent quantitative technique based on probability concept to deal with uncertainty in decision making it is a set of statistical methods to assume with reasonable accuracy, population characteristics on the basis of given en sample statistics.

Statistical inference can be defined as drawing inference from probabilistic sample, about unknown population parameters.

Types of statistical inference

Statistical inference may be focused either on examining hypotheses or on predicting probable values. Accordingly two types of statistical inferences are hypotheses testing and statistical estimation.

In hypotheses testing we examine the claims made about unknown population parameter using sample statistics. These claims are made using some past experience and logic.

Statistical Estimation means estimating unknown population parameters, with reasonable accuracy, using sample statistics. This unit focuses on statistical estimation.

Statistical Estimation

Everyone makes estimates. When we are ready to cross a road, we estimate the speed of any approaching car, the distance to the car, and our own speed. Having made these quick estimates, we decide whether to wait or to walk.

Business managers also estimate for various purposes. Estimation is the process of assessing characteristics of a phenomenon, on the basis of intuition, experience, statistics and other available information

When estimation is exclusively based on statistical methods, it is statistical estimation. Statistical estimation is a useful quantitative technique.

Significance of estimation in managerial decision making

Decision making is the most important and complex task of management. Estimation is inherent to decision making. Thus, in decision making process, estimation plays a significant role, in the following ways.

1. Long term - the outcome of estimation will affect organizational effectiveness, for a long time. Therefore estimates will be critical in the long run.
2. Accuracy - estimates are made basing on past experience and realistic projections in to the future. This will ensure reasonable accuracy in estimates.
3. Goal oriented - estimates are made , revolving around the objectives and goals of the organization. Goal orientation of estimates will improve decision making process.
4. Guidance - estimates are realistic projections into the future. They serve as milestones and guidance towards the attainment of vision and mission of organization.
5. scientific outlook - estimates and follow up will create a systematic and scientific environment within the organization. It will eliminate rule of thumb and intuition in managerial decisions.
6. Relationship - management will have to take decisions in situations of uncertainty and risk. Statistical estimates in such situations will rationalize decisions.

Types of estimates

Estimates mean rationally assessed values of populations on the basis of sample statistics. Such estimates may be specific single point values or range values. Accordingly there are two types of estimates - point estimate and interval estimate.

Point estimate

When the estimated value is a single specific value of the population, it is called point estimate. In point estimate we determine a value which may be taken as an estimate of the population parameter. Sample mean is popular point estimate of the population mean. Arithmetic mean is generally used to express the characteristics of a phenomenon.

For example, when a football fan says “the average age of Kerala Blasters team is 26”, it is point estimate. Other popular point estimates are population proportion, standard deviation and variance.

Properties of good estimator

Estimation enables prediction, with reasonable accuracy, of unknown value on the basis of known value. Such accuracy depends on following qualities.

Unbiasedness

while taking sample for population estimate, it must be done in an unbiased manner. Each item should be given equal opportunity of being taken as sample.

Consistency

Sample value should approach population value, when sample size is increased. This property is consistency. So sample size should be sufficiently large.

Efficiency

Variation between population estimate and sample value should be the least. When the variation is more, it leads to inefficiency.

Ease

Process of estimation should be simple . It should be understood and done with less calculation.

Merits of point estimation

Point estimates are valuable tools in analyzing complex decision problems. Following are their merits

1. It provides a single value as the estimate of population parameter. It is easy and simple to understand and calculate.
2. It gives an exact value for the parameter under investigation. There is no confusion as to which value to be selected.
3. It is considered unbiased and consistent, if the sample size is sufficiently large. It became more reliable with large samples.

Demerits of point estimate.

1. It does not consider uncertainty of estimation. A point estimate cannot ensure whether population parameter will come equal to sample statistics or not .
2. It does not consider the concept of standard error, which will purify estimation process. Standard error will rectify fluctuations in sample data.
3. Lack of confidence level will eliminate the confidence of the estimator in assessing unknown population values on the basis of known sample value.

Steps - point estimation

- Consider the given sample data – sample size and given sample statistic
- Obtain sample mean, variance, standard deviation or proportion as the case may be by dividing sum of quantities by number of elements within a sample.
- Apply the sample statistic over population
- Treat sample statistic as population parameter.

Ex . 2.1

An auditorium is considering its seating capacity. Following are the attendance in 9 days (in 000s). Find point estimates of mean, and variance of daily attendance of people for the coming days.

Attendance (000s)
8.8
14.0
21.3
7.9
12.5
20.6
16.3
14.1
13.0

Ans

X	X - $\bar{x}$ ($\bar{x}$ = 14.3)	$x-x^2$
8.8	-5.5	30.25
14.0	-0.3	0.09
21.3	7.0	49.00
7.9	-6.4	40.96
12.5	-0.8	0.64
20.6	6.3	41.58
16.3	2.0	4.00
14.1	-0.2	0.04
13.0	-1.3	1.69
Total 128.5		168.25

Estimated population mean = 14.3 Estimated population variance = 168.25

Ex 2.2

A carton of syringes contain 5 packets of 20 each. Following is the number of defectives in each packet. Estimate the proportion of defectives.

Packet No	Defectives
1	2
2	4
3	3
4	2
5	1

Solution

Packet No	Defectives
1	2
2	4
3	3
4	2
5	1
Total	12

Total number of items inspected = 5 x 20 = 100

Defectives observed = 12

Therefore, estimated proportion of defectives = 12/100 = .12 (it is likely that 12 out of every 100 items may be defectives.)

Ex 2. 3

Mamatha Bakery delivers shavarma @ Rs 60 and guarantees that it will be delivered within 30 minutes of order. If it takes more than 30 minutes, it will be given free and recorded as 30 minutes. Twelve random orders are delivered as below. Find the average estimate delivery time. What is the population and can the sample be used to estimate average delivery time correctly?

Sl No	Delivery time
1	15.0
2	25.0
3	30.0
4	10.0
5	30.0
6	19.0
7	10.0
8	12.0
9	14.0
10	30.0
11	22.0
12	23.0
TOTAL	240.0

Ans

Total delivery time = 240 minutes

Total number of days = 12

Therefore, average delivery time = 20 minutes

The population is the set of shavarmas delivered in the past, present and future. This sample cannot be used to estimate average delivery time correctly, because any delivery taking more 30 minutes is recorded as 30 minutes.

Proportion

Mean is considered a useful estimator. Proportion is another popular estimator of population values. Proportion is generally used to express parameters of populations in social studies. Using sample proportion, population proportion can be estimated statistically.

Ex2.3

Out of 60 executives in an IT company, 50 uses 4G Cell phones. Give a point estimate of proportion of 4G users.

Sample proportion = 50/60 or 5/6

Estimated populating proportion = 5/6 (that is out of every 6 persons, 5 will be using 4G cell phones.)

Estimator and estimate

Any sample statistic that is used to estimate a population parameter is called an estimator. An estimator is a sample statistic used for estimating an unknown population parameter. The sample mean can be an estimator of the population mean, and sample proportion can be used as estimator of the population proportion.

When we have obtained a specific numerical value of an estimator, we call that value an estimate. In other words, an estimate is a specific value of a statistic. We form an estimate by taking a sample and computing the value taken by our estimator in that sample. Suppose that we calculate the means of mileage of motor cars, by reading speedometers of 25 cars, is 90000 miles. If we use this specific value to estimate mileage for the whole motor cars, the value 90000 is the estimate. Table 2.1 illustrates several populations, parameters, estimators and estimates.

Population	**Parameter**	**Estimator**	**estimate**
Hospital workers	Employee turnover	Mean turnover/month	10% turnover/month
Applications for post of manager	Mean education level	Mean education of every 5^{th} applicant	Graduation
Bachelor college teachers	Proportion of males	Proportion out of a sample of 20 teachers	40% males or 0.4 proportion

Review Questions and Exercises

1. What is inferential analysis
2. State different types of Statistical 1nference
3. Explain estimation
4. What is significance of estimation in decision making?
5. What are two types of estimates?
6. State the properties of the good estimator
7. what is point estimate?
8. state the steps in point estimation
9. When a sample of 40 executives was surveyed regarding poor sales performance, 70% blamed poor weather. Find probabilistic percentage of blaming executives.
10. 500 articles were selected at random out of a batch consisting 10000 articles and 30 were found defective. How many defective articles would you reasonably expect to find in the whole batch?
11. In an office there are 176 male employees, and 24 female employees. What is the percentage of female employees, in general?

UNIT III

INTERVAL ESTIMATION

Point estimation enabled the estimation of single value, in order to assess values of population or phenomenon. It is rather limited in scope, In interval estimate, we calculate an interval of 2 values to include population parameter. It gives lowest and highest values within which population parameter will lie. An interval estimate describes a range of values within which population parameter is expected to lie.

Confidence level

Usually interval estimate are made at a desired level of confidence. Level of confidence means the probability that the interval values will contain the true population parameter. Generally the confidence level are fixed as 99%, 95% , 90% etc.

Confidence level is the expected level of accuracy of estimating the interval within which population parameter is expected to lie. It is the reciprocal of level of significance (alpha = α). So level of confidence is equal to 1-α.

Interval estimates are dependent on the standard error, level of confidence and degrees of freedom.

Standard error

When a sample and its values are given for estimation, the basis of calculation is the standard error. It is the standard deviation of sampling distribution. It is a relative measurement of variations between various samples.

Standard error plays a very important role in the large sample theory and forms the basis of testing of hypothesis and statistical estimation. It is used to test the reliability of samples.

Standard error takes various forms according to circumstances and given information

For large samples (30 or more) SE = $\frac{\sigma}{\sqrt{n}}$

Where, σ = given standard deviation , and n = number of items within a sample.

For small samples (less than 30) SE =$\frac{\sigma}{\sqrt{n-1}}$

Degrees of freedom

Interval estimates are made on either small samples or large samples. Small samples are samples with size 29 or less. Large samples are with size 30 or more.

For small samples Student's t value is taken with degrees of freedom n-1. That is number of observation minus one, which is the freedom a person has in selecting certain observations.

For large samples z value is taken for infinity sample size. At 95% level of confidence z value is 1.96, at 99% it is 2.58 and so on…

Merits of interval estimation

1. It considers the uncertainty or likely error of an estimation. It is more likely that a population parameter is either more or less than a point estimate.
2. It provides a confidence level to estimation process. The estimator is likely to fall within these intervals.
3. It is more realistic, reliable and efficient. It provides consistent results.
4. It provides an idea about, what risk is a decision maker is taking in estimating unknown values.

Steps Estimating population mean

Mean or arithmetical average is the most popular statistic which can be used to estimate population mean. Such estimation can be performed through the following steps.

1. Consider given mean, standard deviation and sample size.
2. Find standard error – for large sample - $\frac{\sigma}{\sqrt{n}}$, for small sample = $\frac{\sigma}{\sqrt{n-1}}$
3. Decide level of confidence and degree of freedom.
4. Ascertain Z value or t value.
5. Calculate upper confidence limit = Mean + (t x se) or Mean + (Z x se)
6. Calculate lower control limit = Mean - (t x se) or Mean - (Z x se)

Ex 3.1

In a survey of 26 consumers, average income was found as Rs 4800 with standard deviation 500. Estimate upper and lower confidence limits at 95%. Interpret.

Ans

n = 26 x = 4800 σ = 500 confidence level = 95 % small sample

Standard error = $\frac{\sigma}{\sqrt{n-1}}$ == $\frac{500}{\sqrt{26-1}}$ = 102.06

Student's t table value, at 95% level of confidence , for degree of freedom=25 = 2.06

Upper confidence limit = Mean + (t x se)

= 4800 + (2.06 x 102.06) = 5010. 65

Lower confidence limit = Mean - (t x se)

= 4800 – (2.06 x 102.6) = 4589.35

Population mean is expected to lie between 4589.35 and 5010.65

Ex 3.2

Average age of 100 college teachers is 45 with standard deviation 15. What would be the probable general average age of college teachers in Kerala.

Ans

Mean = 100 n = 45 standard deviation = 15

Standard Error = - $\frac{\sigma}{\sqrt{n}}$ = $\frac{15}{\sqrt{100}}$ = 1.5

Z value for 95 level of confidence = 1.96

Upper confidence limit = Mean + (Z x se)

= 45 + 1.96 x 1.5 = 47.94

Lower confidence limit = 45 - 1.96 x 1.5 = 42.06

Estimated General average of college teachers will be between 42 and 48 (approx.)

Ex 3.3

In 17 Grama Panchayaths, average age of presidents was 50 with standard deviation 3. Ascertain 95% confidence limits for the age of panchayath presidents in general.

Mean = 50, Standard deviation = 3 n = 17

Standard Error = $\frac{\sigma}{\sqrt{n-1}}$ = $\frac{3}{\sqrt{17-1}}$ = .75

t table value for 95% , and n = 16 = 2.12

Upper 95% limit = 50 + .75 x 2.12 = 51.59

Lower 95% limit = 50 - .75 x 2.12 = 48.41

Ex 3.4

Life Insurance Corporation found that mean age of death of 64 workers in a factory is 64 years with standard deviation 8 years. What are the 99% confidence and 95% confidence limits for mean age of workers in that factory for insuring them.

Mean = 64 n = 64 σ = 8

Z table value at 99% level of confidence = 2.58

Z table value at 95% level of confidence == 1.96

99% upper confidence limit = 64 + 2.58 x 8 = 84.64

99% upper confidence limit = 64 - 2.58 x 8 = 43.36

95% upper confidence limit = 64 + 1.96 x 8 = 79.68

95% Lower confidence limit = 64 - 1.96 x 8 = 48.32

Estimating population proportion

Just as mean is subjected for estimating population parameters, proportion also can be estimated for the population on the basis of given sample proportion. Proportion is commonly used to express parameters of populations in marketing studies. Using sample proportion, population proportion can be estimated statistically, as per following steps

1. Consider given sample proportion = p
2. Obtain 1 – p = q
3. Calculate standard error of proportion =
4. Calculate upper and lower confidence limits UCL = p + (z x SE), LCL = p – (Z x SE)

Ex 3. 5

When a sample of 40 executives was surveyed regarding poor sales performance, 70% blamed poor weather. Find probabilistic percentage of blaming executives using 95% level of confidence.

Solution

Sample proportion = p = 70% = .7 n == 40 q = 1-p = .3

Standard Error of proportion = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{.7 \, x \, .3}{40}}$ = .022

Upper 95% confidence limit = .7 + 1.96 x.022 = .743

Lower 95%confidence limit = .7 - 1.96 x .022 = .657

Ex 3.6

Syndicate Bank found that out of 200 loanies, 32% defaulted in payment of their loans.

(A) Estimate standard error of proportion

(B) Estimate upper and lower confidence interval @ 99%

Solution

Sample proportion = p = .32 n= 200 q = .68

Standard error of proportion = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{.32 \, x \, .68}{200}}$ = .033

Upper 99% confidence limit = .32 + 2.58 x .033 = .405

Lower 99% confidence limit = .32 - 2.58 x .033 = .235

Ex 3.7

AM motors states that out of 75 employees, 40% are females. Estimate with 99% confidence, proportion of females, to be recruited in future.

Solution

P = .4 q = .6 n = 75

Standard error = $\sqrt{\frac{.4 \times .6}{75}}$

99% upper confidence limit = .4 + 2.58 x .03 = .477

99% lower confidence limit = .4 - 2.58 x . 03 = .323

Ex 3.8

Dr Shivkumar, a psychologist, surveyed 70 executives and found that 66% of them could not add fractions.

(a) Estimate standard error of proportion

(b) Find lower and upper 95% confidence limits.

P = .66 q = .34 n= 70

Standard error proportion = $\sqrt{\frac{.66x .34}{70}}$ = .056

95% upper confidence limit = .66 + 1.96 x .056 = .771

95% lower confidence limit = .66 - 1.96 x .056 = .549

Review Questions and Exercises

1. What is interval estimation?
2. Distinguish between point estimates and interval estimate
3. State the steps in interval estimation
4. Determine 95% confidence interval for population mean when sample of size 'n' is drawn from that population given that the population is normal with variance σ^3
5. A random sample of 50 people from a population showed incomes with a mean = 50000 and standard deviation = 6000. Estimate the population mean with (a) 95% (b) 99% confidence interval.
6. A random sample of 100 articles selected from a batch of articles shows that the average diameter of the articles = 0.354 with a standard deviation = 0.048. Find 95% confidence interval for the average of this batch of articles.
7. A sample of 25 workers have an average wage of Rs. 45 with S.D = 10 Rs. Give 90% confidence interval for mean wage of the population from which the sample was taken. State the assumptions made.
8. The mean of a sample of size 16 from a normal population is 20. If it is known that variance of population is 4, find the standard error of the sample mean and 95% confidence interval for the population mean?
9. A random sample of 20 bullets produced by a machine shows an average diameter of 3.5 mm and a standard deviation 0.2 mm. Assuming that the diameter measurement follows normal distribution with mean μ and SD = σ obtain 95% interval estimate for the mean?

UNIT IV

SAMPLING AND SAMPLE SIZE

One of the critical factors influencing statistical estimation is sampling. Sampling is reliable to study the whole population. A housewife takes a few rice, from a boiling pot, to check its cooling. She is ensuring the cooking of the whole pot.

Sampling is a tool which helps to know the characteristics of the population. Sampling is defined as the process of drawing representative number of items for collecting information to infer about the population.

Principles of sampling

Sampling is reliable, because it is based on two universally accepted principles or theories.

Law of statistical regularity

The principle states that if samples are drawn at random from a population, it is likely to possess the characteristics of the population. In other words samples will be statistically regular, if samples are regular.

Law of inertia of large numbers

According to this principle, large numbers are relatively more stable than small numbers. It is difficult to move or change large numbers. Large numbers have consistency.

Sampling techniques

Several sampling techniques or methods are in use to get required data. They are broadly classified as random sampling techniques and non random sampling techniques. Random sampling techniques include simple random sampling, stratified sampling, systematic sampling and cluster sampling etc. non random sampling techniques include judgment sampling, multistage sampling, quota sampling, snowball sampling etc.

Simple random sampling.

This is the easiest method of sampling. In this technique every item get an opportunity of being selected. This technique is applied through taking lots or Random Number Tables.

Stratified random sampling

Here population is subdivided into several strata of homogenous items. Then samples are taken from each stratum so as to make up total number of samples..

Systematic sampling

In this technique a system is designed to pick samples. As per requirement, every k th item is picked to make the required sample.

Cluster sampling

Here population is located in convenient clusters where items concentrate. Required sample numbers are selected from such clusters, applying some random technique.

Judgment sampling

Here samples are taken according to judgment or purpose. We simply pick those items, which convenient to select.

Multistage sampling

Population is subdivided into several stages from top to bottom and the lowest stage is utilized for sample selection.

Quota sampling

Here quotas are fixed, and required sample numbers are picked according to the quota determined.

Snowball sampling

This is a type of convenient sampling technique, where initially a few items are selected as samples, and as the study proceeds, required sample numbers are added according to convenience.

Sample size

Sample size is the number of items included in a sample. This is a decision factor in accurately estimating population parameters. Sampling precision depends more on sample size, and not on proportion of population sampled.

In sampling analysis, vital questions are - how large the sample should be? If the sample size is too small, estimation may be inaccurate. If it is too large, heavy cost may be incurred.

Factors influencing sample size.

As a general rule, sample must be of an optimum size. Size of the sample is determined by following factors.

1. Nature of population – Population may be homogenous or heterogeneous. If it is homogenous, a small sample can serve the purpose. Otherwise large sample is required.

2. Nature of study – if intensive and focused study is required, small sample will do. For a general study, large samples may be undertaken.

3. Sampling technique – sampling technique plays an important role in sample size. A small but properly selected sample is better than a large but poorly selected sample.

4. Accuracy - If higher level of accuracy is required, relatively larger samples are required. For doubling the level of accuracy, sample size should be increased fourfold.

Approaches in sample size decision

There are two alternative approaches for determining size of sample. First approach is to specify the precision of estimation desired and then to determine sample size (n) necessary to ensure it. The second approach uses cost of additional information against expected value of additional information. The first approach is capable of solving a mathematical solution, and as such is a frequently used technique of determining sample size. The limitation of this technique is that it does not consider the cost of gathering information. The second approach is theoretically optimal, but is rarely used because of difficulty in measuring the value of information. Therefore, we shall concentrate here on the first approach.

Determining sample size –confidence level approach

When a sample study is made, sampling errors are bound to occur, and this can be controlled by selecting sample of adequate size. The precision level must be specified along with confidence level. Sample size can be determined considering such level of confidence, standard deviation and expected error.

Sample size for infinite population

In case of infinite populations, number of elements within the population in indefinite or uncertain. Normally populations belong to this group.

Sample size = $n = \frac{Z^2 \, x \, \sigma^2}{e^2}$

Where n = sample size

Z = value of significance level = 1.96 or 2.58

σ = standard deviation

e = expected error

Sample size for finite population

In case of finite population, number of elements within population (N) will be certain and given, along with standard deviation and expected error.

$$n = \frac{Z^2 \, x \, N\sigma^2}{(N-1)e^2 + Z^2 \, x \, \sigma^2}$$

Where n = sample size

Z = value of significance level = 1.96 or 2.58

σ= standard deviation

e= expected error

N = Number of items in population

If standard deviation is not given

In the above formulae, standard deviations of population were given. But in many cases it may not be given or is not available. Since we have not yet taken the sample and are in the stage of deciding the size of sample, we cannot calculate standard deviation of population. In such a situation, if we have an idea about the range (difference between highest value and lowest value) we can estimate standard deviation of population as below

Ex . 4.1

Suppose we are estimating wages of workers in a village, and it is learned that there is difference of Rs 40 between the highest wage 640 and lowest wage 600

we know that ± 1.96 σ covers 95% of items in a study. It means 3.92 σ covers the given range of population. Thus

Given Range = $3.92 \times \sigma \approx 40$

$\sigma = \frac{\text{Given Range}}{3.92} = \frac{40}{3.92} = 10.20$ ie, about 10 workers

The obtained standard deviation can be utilized for estimating sample size.

Ex 4.2

A Production controller estimates that average production per day is 100 with standard deviation 4.8. He expects an error of ±3. How many items he should include in a sample, at 95% level of confidence

σ = 4.8 e = 3 Z = 1.96 n = ?

$n = \frac{Z^2 x\ \sigma^2}{e^2} = \frac{1.96^2 x\ 4.8^2}{3^2} = 9.834$ Taken as 10

Ex 4.3

Determine sample size for estimating weight of 5000 milk packets with variance of weight 4 gms, and expected error of .8 gms, at 99% level of confidence .

N = 5000 Z = 2.58 $\sigma = \sqrt{4} = 2$ e = .8 n = ?

$n = \frac{Z^2 x\ N\sigma^2}{(N-1)e^2 + Z^2 \times \sigma^2} = \frac{2.58^2 x\ 5000 \times 2^2}{(4999) x\ .8^2 + 2.58^2 \times 2^2} = 40.95$

= Taken as 41

Ex 4.4

Determine the number of items to be included in a sample to obtain estimated mean weight of Frooti, at 99% level of confidence, with standard deviation 2 and expected error of .8

σ = 2 e = .8 Z = 2.58 n = ?

$n = \frac{Z^2 x\ \sigma^2}{e^2} = \frac{2.58^2 x\ 2^2}{.8^2} = 41.28$ Taken as 41

Sample size for estimating proportion

For estimating proportion, the number of elements in a sample is a decision factor which must be specified beforehand. Besides, the expected precision and the confidence level also should be considered. Then, number of items to be included in a sample for estimating proportion value , will be: $n = \frac{Z^2 x\ p \times q}{e^2}$

Where n = sample size

Z = value of significance level = 1.96 or 2.58

σ = standard deviation

e = expected error

Ex 4.5

What should be the size of sample, to be drawn from a population, to estimate per cent defective, within 2 % of true value, with 95% level of confidence? For this 100 items were selected, and obtained 2 defectives.

e = .02 Z =1.96 p = .02 q = .98 n = ?

$$n = \frac{Z^2 \times p \times q}{e^2} = \frac{1.96^2 \times .02 \times .98}{.02^2}$$

=188.24 taken as 188

Ex 4.6

in a hotel, 5 out of 100 visitors stay overnight. The management wants to be 95% confident that population percentage to be estimated with ±3% of true value. What should be minimum sample size?

e = .03 z =1.96 p = .05 q = .95 n = ?

$$n = \frac{Z^2 \times p \times q}{e^2} = \frac{1.96^2 \times .05 \times .95}{.03^2} = 202.75$$ or 203 visitors

Review Questions and Exercises

1. Define sampling
2. Explain non-random sampling methods
3. What are random sampling techniques?
4. What is sample size?
5. Explain optimum sample size
6. State the determining factors in sample size
7. Describe steps in determining sample size for mean test
8. What is the procedure in deciding proportion sample size?
9. Determine the size of sample for estimating true mean of the population of size 5ooo, on the basis of the following:
   Population variance = 4
   Level of confidence = 99%
   Estimated error = .4
10. what would be the size of sample for estimating mean of the population, if –
    Standard deviation = 15
    Estimated error = 6
    Level of confidence =99(%
11. Determine sample to be drawn from the population of 5ooo units to estimate the percentage of defectives on the basis of 3% defectives in the sample within .05 units of its true value. Level of confidence desired is 95%.
12. what should be the size of the sample drawn from a population to estimate the percent defective within 2% of the true value with 95% level of confidence , on the basis of 3% defective in the sample.

UNIT V
TESTS OF SIGNIFICANCE - CONCEPTS

One of the objectives of statistical investigation is to evaluate whether there is significant difference between the estimated parameter and true parameter after estimating population mean or proportion. Naturally a question arises –Does the estimated parameter conform to real parameter, or, is there any considerable difference between them, the answer leads us to the evaluation of difference. There are tests to aSSEss the significance of such difference, which are called significance tests.

The basis of statistical tests is hypothesis. First we form a hypothesis regarding the population. Then we conduct a test to ass whether there is any significant difference. The hypothesis will be accepted or rejected according the significance of difference revealed by the test. Therefore, significance test is also called hypothesis tests.

In social sciences where direct knowledge of population parameter is rare, significance or hypothesis testing is the often used strategy for deciding whether sample data supports population characteristics or not.

Basic concepts

Significance tests are amply supported by several theoretical basic concepts. In order to conduct tests, knowledge of following basic ingredients are essential.

Types of significance tests

There are numerous types of significance tests, according to situations and criteria of testing. Tests may be parametrical or non parametric, one tailed or two tailed, small sample or large sample etc.

Parametric and non parametric tests

On the basis of focus of the test, they can be classified as parametric and non parametric test. In certain tests, assumption about population distribution can be made. For example in large sample test or Z test we assume that samples are drawn from population following normal distribution. Such tests which are based on assumptions about population are called parametric tests. Mean Tests, Proportion Tests, Variance tests are parametric tests. These tests focus on means of samples or population, proportion, variance or standard deviation and accordingly, all mean tests, proportion tests or variance tests are parametric tests.

But in certain situations, it is not possible to make any assumption about population distribution, from which samples are drawn. Besides they do not focus on parameters like mean, proportion or variance. Such tests are called non parametric tests. Non parametric tests include Chi square test, Rank test, Sign test , Runs test etc.

Small sample and large sample tests

According to the number of items included in a sample, tests can be divided as small sample tests and large sample tests. If the test includes a sample of size less than 30, it is small sample test. If the size is 30 or more, it is large sample test.

Small sample tests follow student's t distribution. Large samples tests follow normal distribution. Mean tests may be conducted as large or small tests. But proportions are conducted as large sample tests only.

One tailed or two tailed tests

On the basis of location of rejection region, tests may be one tailed or two tailed . When a test examines the significance of difference of either more than a specific value or less than a specific value, rejection appears only on one side of the curve. It is called one tailed test.

When test examines both more than or lower than a specific value at the same time, rejection region appears on both sides of the curve, and such test is called two tailed test. Most of tests are two tailed tests.

Hypothesis

Hypothesis is the basis of all significance tests. For a researcher, it is a formal question that he intends to resolve. Usually we begin some assumptions about the population from which the sample is drawn. This assumption may be about the form of the population or about the parameters of the population. Such assumption is called hypothesis.

A hypothesis may be defined as "a tentative conclusion logically drawn concerning the parameter or the form of the distribution of the population. Example of hypothesis may be "the mean of the population will be 12000" or "the population proportion will be the same as sample proportion."

Types of hypothesis

According to the nature and situation, hypothesis may be simple or composite, parametric and non parametric, or null or alternative.

Simple and composite hypothesis

If a hypothesis is concerning sample statistic or population parameter only, it is called simple hypothesis. For example, "population standard deviation conforms to sample standard deviation" is a simple hypothesis."

If a hypothesis forms a statement about any sample statistic or parameter and form of distribution, it is called composite hypothesis. For example, "population follows normal distribution with mean = 25 " is a composite hypothesis."

Parametric and non parametric hypothesis

A hypothesis which specifies only the parameter or statistic of either the sample or population is called parametric hypothesis. If a hypothesis specifies only the form of the distribution, it is non parametric. For example, the hypothesis " Mean of the population is 2300" is a parametric hypothesis, while "population is normal" is non parametric.

Null and alternative hypothesis

A null hypothesis is statistical hypothesis which states that the difference between the sample statistic or population parameter is nil, or statistically insignificant. Usually null hypotheses are formed for significance testing. Any hypothesis other than null hypothesis is called an alternative hypothesis. The null hypothesis is denoted by Ho and alternative hypothesis by H1, H2 and so on.

Sampling distribution

From a population, several samples may be collected , from each sample group, some sample statistic like mean, median range or standard deviation can be ascertained. The distribution thus obtained from a sample statistic is called a sampling distribution. It is a probability distribution. Thus a sampling distribution is a list of certain sample statistics. A sample statistic is a random variable and it has a probability distribution, called sampling distribution of that statistic. Accordingly, there will be sampling distribution of means, sampling distribution of standard deviations etc.

Standard error

A sampling distribution has several sample means or other statistics. A grand mean of such a sampling distribution can be ascertained, and deviations calculated. The standard deviation of sampling distribution of a statistic is called standard error of that statistic. The standard error is very useful quantitative tool in hypothesis testing. Generally the standard error is $\frac{\sigma}{\sqrt{n}}$, where, σ is the standard deviation and n is the sample size.

Uses of standard error

Standard error plays dominant role in parametric significance testing. It forms the basis of hypothesis testing due to following utilities.

1. It gives an idea about the reliability of a sample. The reciprocal of standard error is a measure of reliability of the sample.
2. It is used to determine the confidence limits for population values like means, proportion and standard deviation.
3. It is used to examine given hypothesis. The test value is computed by dividing the actual difference tested with standard error. Thus the fate of hypothesis is determined.

Difference between Standard Error and Standard Deviation

1. Standard deviation is a measure of dispersion of statistical data. Standard error is a measure of dispersion of a sampling distribution.
2. Standard deviation measures the variability or consistency of a statistical series. Standard error determines the precision of reliability of an estimated value.
3. Standard deviation is calculated in relation to the means of a series . Standard error is calculated in relation standard deviation, and sample size.
4. Standard deviation is the basis of standard error, but not vice versa.

Level of significance

A null hypothesis when proved true is to be accepted. But there is a remote probability that it may be wrongly rejected. Such a probability is called level of significance. Level of significance is the probability with which a null hypothesis is rejected. Therefore, level of significance is the risk; a statistician is running in his decisions. Generally statistical hypothesis tests are conducted at a given level of significance. If level of significance is not specified, it is taken as 5 %. Level of significance is denoted as α (alpha) . It is usually determined before conducting the statistical test.

Type I and Type II errors

Null hypothesis is the starting point in significance testing. A null hypothesis may lead to four possibilities:

- Null hypothesis - proved true - Accepted
- Null Hypothesis - proved wrong - Rejected
- Null Hypothesis - proved true - Rejected
- Null hypothesis - proved wrong - Accepted

First and second possibilities are correct**.** But the third and fourth possibilities are errors. So type I error is committed by rejecting a true null hypothesis. Type II error is committed by accepting a hypothesis which is proved wrong. Of these two errors, Type II errors is more serious and far reaching than Type I error

Degree of freedom

While selecting items of statistical process, we have limited freedom. Degree of freedom is the number of independent choices in determining observations. In the case of individual observations, degree of freedom is total number of observations less the number of constraints. Usually is is equal to n-1.

In the case of 2 x 2 table, degree of freedom is equal to column – 1 x row - 1. Usually degree of freedom is denoted by υ (nu) .

Critical value

The critical value is the limit of the difference, beyond which it is assumed that difference is significant. It is the value of the test statistic which separates the rejection region from the acceptance region. It is determined by the level of significance and the nature of the test. For example, if it is a large sample test, the critical value @ level of significance = 0.05 will be 1.96. the calculated test value is compared with the critical value to decide the fate of H0. Critical value is also called Table Value.

Critical Region

Critical value separates total region into acceptance region and rejection region. Critical region corresponds to a predetermined level of significance, and acceptance region corresponds to 1 – level of significance.

Central limit theorem

Statistical significance tests are conducted under certain assumption. The basis of such assumption is central limit theorem.

When samples are drawn from a normal population, the means of such samples will form a normal distribution. When the population is not normal, the sampling distribution may be non normal. But as sample size increases, the distribution will tend to become normal. The theorem which explains this relation between shape of distribution and sample size is called central limit theorem. This is the most remarkable theorem in statistical inference and hypothesis testing.

The theorem states that given a sufficiently large sample size from a population, the means of all samples will be approximately equal to the means of population. It assumes that sampling distribution of means approach normality, as sample size increases.

The significance of central limit theorem is that it permits to use sample statistics to make inference about population parameter. The theory makes sampling processes and techniques reliable.

Assumptions of Central Limit Theorem

1. All samples drawn are random samples.
2. Sampling distribution is based on a random variable and is independent
3. Means and variance of samples and population exist
4. The mean of samples will approximate population mean
5. The means of samples will tend to be normal.

General procedure for testing hypothesis

To test hypothesis means to tell whether or not a hypothesis seems to be valid. In hypothesis testing, main question is whether to reject he null hypothesis or not to reject hypothesis. Generally steps involved in hypothesis testing are the following.

1. Formulate null hypothesis, conforming to the problem
2. Determine type of test – whether Z, t, one tailed, two tailed etc.
3. Decide level of significance and degree of freedom
4. Obtain Standard Error = = $\frac{\sigma}{\sqrt{n}}$, or = $\frac{\sigma}{\sqrt{n-1}}$
5. Compute test statistic t value or Z value, as the case may be
6. Ascertain table value at appropriate level of significance and degree of freedom
7. Compare calculated value with table value
8. If calculated value is less than table value, accept the hypothesis
9. If calculated value is more than table value, reject hypothesis.

Limitations of hypothesis testing

Important limitations of significance tests are:

1. Hypothesis testing should not be used in mechanical fashion. They are to be designed according to situation.
2. They simply indicates the magnitude of difference, and do not reveal reasons for difference.
3. Results of significance tests are based on probabilities, and as such cannot be expressed with full certainty.
4. It is not entirely correct. In the case of small samples, probability of erring inferences is higher.

All these limitations suggest that in problems of statistical significance, inference techniques may be combined with adequate basic knowledge of subject matter along with the ability of good judgment.

Significance tests come under two categories – parametric tests and non parametric tests. The following unit deals with parametric tests for means and proportions. Paired observations and variance tests are dealt with in the next unit. . Non parametric tests are discussed in later chapters.

Review Questions and Exercises

1. What are tests of significance?
2. Distinguish between statistics and [parameters
3. What is hypothesis testing?
4. What is hypothesis?
5. Give a classification of significant tests
6. Distinguish between null and alternative hypothesis
7. Explain sampling distribution
8. Explain standard error
9. Distinguish between standard deviation and standard error
10. What is level of significance?
11. What are type i and type ii errors?
12. Explain degree of freedom
13. What is critical region?
14. What are one-tailed and two-tailed tests?
15. In the following case, which test you will apply-whether 1 tailed or 2 tailed?

    A bowler claims that he can bowl at a speed of 120 km/hour. 5 test bowling sessions were conducted and found the average speed as 118, 132, 112, 134, 121.

16. Explain σ = standard deviation
    μ = population mean
    α = level of significance

UNIT VI

PARAMETRIC TESTS FOR MEANS & PROPORTIONS

Population parameters like the mean, proportions, variance etc. are of great importance in significance tests and economic applications. Test based on such parameters are called parametric tests. Parametric tests enable to specify the parameters of population and the form of a concerned probability sampling distribution.

Statistical tests in which hypothesis deals with population parameters or sample statistics are called parametric tests. For example, when we want to test given population mean or population proportion, or any sample statistic, the test applied is called parametric test.

Importance

Parametric tests like mean tests, proportion tests and variance tests are of great importance in business and economic applications. Standard test procedures are available to test various hypotheses regarding these parameters.

1. Most of the business and economic situations tend to be normal, conforming to normal probability distribution. This enables scientific testing and proving of hypothesis relating to industry and commerce.
2. Standard test procedures are available to test various hypotheses regarding these parameters. This makes significance testing powerful and reliable
3. Central Limit Theorem is the basis of parametric tests, which enables accurate estimation of values within which populating parameters will lie.
4. Level of confidence and level of significance plays a crucial role in making parametric tests useful and handy.

Features of parametric tests

Parametric tests are universally recognized as the most useful and reliable hypothesis testing technique. It exhibits following features.

1. Parametric – it is based on population parameters like mean, variance, proportion or standard deviation. Parametric tests make use of one or more statistics obtained from sample data or proportion to arrive at a probabilistic statement as hypothesis.

2. Distribution – a parametric test is based on the assumption that population from which samples is drawn follows normal or some other probability distribution. Normality of the distribution makes it possible to make statistical inference.

3. Randomness - it is also assumed that samples drawn from a population are random samples. Randomness makes sampling technique and testing powerful and reliable.

4. Level of measurement – parametric tests conform to higher level of measurement such as interval scale and ratio scale. Nominal o0r ordinal measurement levels do not apply in case of parametric tests.

Merits of parametric tests

Parametric tests are widely applied to solve decision making problems relating to business and industry. It is due to following reasons

1. Simple – parametric tests are the most simple to understand, explain and prove. Test results have a direct bearing with the procedure
2. Efficient – parametric tests are considered efficient and powerful. It is due to the fact that they are based on valid assumptions about population.
3. Realistic – since the parametric tests are based on parameters and statistics, which are true representatives of large mass of data , they are considered realistic and more or less accurate

4. Prediction – parametric tests are based on Central Limit theorem, and form of distribution. This enables valid predictions and projections.
5. Sharp – parametric tests do not make use of ranks or signs, unlike in parametric tests. So the results will be more sharp.

Assumptions in Parametric Tests

1. Sample Observations are independent.
2. Observations follow any sampling distribution.
3. Samples drawn are random samples.
4. Observations are made at least on interval scale

Types

Parameters include mean, proportion, standard deviation, variance etc. Accordingly, parametric tests are Mean Tests, Proportion Test, Variance Test, and Standard Deviation Test etc. Besides, mean tests may again be classified as mean test between population and sample or between sample and sample etc. This unit contains tests for means and test for proportion, under various situations. Most of the significance tests are based on some parameter like mean, proportion, variance or standard deviation. Therefore such tests are called parametric tests. Since these tests depend on the shape of a sampling distribution, they are also called distribution tests. Thus mean tests, proportion tests etc are parametric tests. They are useful to test hypothesis about data which are normal or tend to be normal.

Parametric Tests for Means

Arithmetic mean is the most widely used statistic in order to study population characteristics or parameters. The difference between sample mean and population mean can be subjected to hypothesis testing and, on this basis managerial decision may be taken. Sample mean may be compared with population mean or with another sample mean to reveal the significance of differences. Accordingly varied situations emerge, together with small samples and large samples.

Large Sample Mean Test

This test focus on sample mean or population mean, and examine the significance of difference between given sample mean and estimated population mean, and between one sample mean and another sample mean. Sample size is 30 or more.

Parametric test of significance is based on the assumption that population of data from which samples is drawn is normally distributed. It focuses on some probability distribution for arriving at a decision about the resemblances of an assertion or hypothesis. The technique makes use of one or more values obtained from sample data which is generally called test statistic, to arrive at a probability statement about the hypothesis.

Normality of population distribution forms the basis for making statistical inference about the samples drawn from population. When sufficient number of samples are drawn from a population, which may be normal or not normal, the distribution of means or proportions of such samples will tend to be normal and approximate a normal curve, with predictable properties.

Sample Vs population

Ex 6.1

A cell phone battery company claims that its batteries have a average life of 200 hrs. A consumer tested 49 batteries and found that they have an average life 191 hours, with standard deviation 21 hours. Is the claim valid?

Ho : No significant difference. Sample mean and population mean are the same. Claim is valid.

Given : Sample mean = $\bar{x}$ = 191 Population mean $\mu = 200$ $\sigma = 21$ $n = 49$ Standard Error = $\frac{\sigma}{\sqrt{n}} = \frac{21}{\sqrt{49}} = 3$

Z value = $= \frac{Difference}{standard\ Error} = \frac{200 - 191}{3} = 3$

Z table value @ 0.05, one tiled test = 1.645

Since calculated z value 3 is more than Z table value 1.645, difference is considered significant. Hypothesis is rejected. Therefore claim is not valid.

Ex. 6.2

A sample of 300 screws has a mean length of 3.4 cm with standard deviation of 2.61 cm. Can it be regarded as a sample from a population with mean length of 3.25 cm, at $\alpha = 0.01$?

Ho : No significant difference. It is sample from a population with mean length 3.25cm.

Given : Sample mean = $\bar{x}$ = 3.4 CM Population mean $\mu = 3.25$cm $\sigma = 2.61$cm $n = 300$

Standard Error = $\frac{\sigma}{\sqrt{n}} = \frac{2.61}{\sqrt{300}} = .15$

Z value = $= \frac{Difference}{standard\ Error} = \frac{3.4 - 3.25}{.15} = 1$

Z table value @ 0.01, = 2.58

Since calculated z value is less than Z table value, difference is considered insignificant. Hypothesis is accepted. Therefore sample comes from a population with mean length 3.25 cm

Ex. 6.3

It is guaranteed that A sample of 100 tyres with mean life of 15231 kms is drawn from a population of tyres with mean life 15300 kms and standard deviation 1248 km. test the validity of guarantee.

Ho : there is no significant difference. Guarantee is valid.

Sample mean = $\bar{x}$ = 15231 Population mean $\mu = 15300$ $\sigma = 1248$ $n = 100$

Standard Error = $\frac{\sigma}{\sqrt{n}} = \frac{1248}{\sqrt{100}} = 124.8$

Z value = $= \frac{Difference}{standard\ Error} = \frac{15300 - 15231}{124.8} = 0.55$

Z table value @ 0.05, one tiled test = 1.645

Since calculated z value is less than Z table value, difference is considered not significant. Hypothesis is accepted. Therefore guarantee is valid.

EX: 6.4

A soap manufacturing company was distributing a particular brand of soap through a number of retail shops. Before a heavy advertisement campaign, the mean sale per week per shop was 140 dozen. After the campaign, a sample of 20 shops was taken and mean sale was found by 147 dozen with standard deviation 16. Can you consider the advertisement effective?

Ans:

H_0 : No significant difference

The advertisement is not effective

$\mu = 140$, $\bar{x} = 147$, $\sigma = 16$, $n = 20$

$$S.E = \frac{\sigma}{\sqrt{n-1}} = \frac{16}{\sqrt{20-1}} = \frac{16}{\sqrt{19}} = \frac{16}{\sqrt{4.35}} = 3.67$$

$$t = \frac{Diff}{S.E} = \frac{147-140}{3.67} = \frac{7}{3.67} = 1.91$$

Degree of freedom = 20 - 1 = 19

Table value 't' @ 5% level of significance is 2.093

The calculated value is numerically less than the table value. So, we ACCEPT the null hypothesis. There is no significant difference between mean of sample and original mean.

So, the advertisement is not effective.

EX: 6.5

Prices of the shares of a company on different days in a month were found to be 66, 65, 69, 70, 69, 71, 70, 63, 64, 68. Discuss weather mean price of in month is 65.

Ans:

X		
66	-4	16
65	-5	25
69	-1	1
70	0	0
69	-1	1
71	1	1
70	0	0
63	-7	49
64	-6	36
68	-2	4
675	**-25**	**133**

H_0= There is no significant difference. The mean price of the share of the month is 65.

$\Sigma X = 675$, $n = 10$,

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{675}{10} = 67.5$$

$$S.D\ (s) = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{133}{10} - \left(\frac{-25}{10}\right)^2} = \sqrt{13.3 - 6.25}$$

$$= \sqrt{7.05} = 2.655$$

$$SE = \frac{s}{\sqrt{n-1}} = \frac{2.655}{\sqrt{10-1}} = \frac{2.655}{\sqrt{9}} = \frac{2.655}{3} = 0.885$$

$$t \text{ value} = \frac{Diff}{SE} = \frac{67.5-65}{0.855} = \frac{2}{0.855} = 2.92$$

Degree of freedom = n-1 = 9

Table value of 't' ' for 9 degree of freedom @ .05 level of significance =2.262, The calculated value 2.92 is more than the table value 2.262 numerically. Thus we reject hypothesis.

The mean price of the shares of the month is not 65.

One Tailed Test

In one tailed test, we examine whether a particular value is either more than or less than a given value. The rejection region appears only on one side of the curve. For example, if we wants to test whether population average is more than Rs 5000, we place the rejection on the right side only. In one tailed test, there are separate table values.

EX: 6.6

A Stenographer claims that she can take dictations at the rate of more than 135 words per minute. Of the 12 tests given to her she could perform an average of 120 words with standard deviation of 40. Is her claim valid? ($\alpha = .01$)

Ans:

This is one tailed test. Ho = There is no significant difference. Her claim that she can take dictations @ more than 135/minute is valid

$\bar{x} = 120$, $\mu = 135$ n = 12 s =40

$$t = \frac{Diff}{SE}$$

$$SE = \frac{s}{\sqrt{n-1}} = \frac{40}{\sqrt{12-1}} = \frac{40}{\sqrt{11}} = \frac{40}{3.316} = 12.06$$

$$t = \frac{Diff}{SE} = = \frac{135-120}{12.06} = \frac{15}{12.06} = 1.24$$

Degree of freedom = n-1 = 12-1 = 11 , Given level of significance = .01 Table value of t (for one tailed test) = 2.178

The calculated value 1.24 is less than the table value 2.178. So, we accept the Ho. *The stenographer can type at an average 120 words per hour.*

EX: 6.7

A factory was producing electric bulbs of average length of life 2000 hours. A new manufacturing process was introducing with the hope of increasing the length of life of bulbs. A sample of 25 bulbs produced by new process, was examined and average length of life was found to be 2200 hours. Examine whether the average length of the bulbs was increased assuming of lives of bulbs follow normal distribution with $\sigma = 300$. (Significant level 0.05)

Ans:

Ho: There is no significant difference. The new product has not increased the length of bulbs.

Here we want to test that the increase in the length of bulb's life is significant. So, it's *a one-tailed* test. Population SD is given. So, we can use Z-test for the calculation.

$$Z = \frac{Diff}{SE} \text{ , } SE = \frac{\sigma}{\sqrt{n}}$$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{300}{\sqrt{25}} = \frac{300}{5} = 60$$

$$Z = \frac{Diff}{SE} = \frac{2200-2000}{60} = \frac{200}{60} = 3.33$$

Table value @ 0.05 level of significance is 1.645 (Sine one tailed test).

The calculated value 3.3 is numerically greater than the table value 1.645.

So, we REJECT the H_0. The new product has increased length of life.

Mean Tests – Sample v/s Sample

Such test may use large sample or small sample. The test examines the significance of difference of one sample against another sample. There is no population value. The test focuses on the difference between samples, or whether they conform to each other. In this case two sample sizes, and the combined standard deviation, or two sample standard deviations must be given. Then standard errors are calculated as below

$SE = \sigma\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$ - (for large sample, where only the combined standard deviation is given)

$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ (for large sample. When two sample standard deviations are given)

$SE = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ (for small samples, when 2 standard deviations and 2 sample sizes are given)

Degree of freedom = $n_1 + n_2 - 2$

EX: 6.8

The mean yield of wheat from district A was 2010 lbs. with standard deviation 10 lbs. per acre from a sample of 100 plots. In another district B, the yield was 200 lbs. with standard deviation 12 lbs. from a sample of 150 plots. Assuming that standard deviation of yield in the entire state was 11 lbs., test whether there is any significant difference between the mean yields of the crops in the two districts

Ans:

H0 = There is no is no significant difference.

There is no difference between mean yield of crops between mean crops in the two districts. Here, the samples are large. So, we can use z-test. You note that here in this question population S.D is given. So, sample S.D not used for finding S.E

$Z = \frac{Diff}{SE}$, $SE = \sigma\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$ $n_1 = 100$, $n_2 = 150$.

$$= 11\sqrt{\left[\frac{1}{100} + \frac{1}{150}\right]} = 11\sqrt{[.01 + .007]} = 11 \times \sqrt{.017} = 11 \times .1304 = 1.43$$

$$Z = \frac{Diff}{SE} = \frac{210-200}{1.43} = \frac{10}{1.43} = 7$$

Degree of freedom is infinity

Table value @ 5% level of significance is 1.96

The calculated value (7) is numerically greater than the z-table value (1.96). So, we REJECT the hypothesis. *There is significant difference in the mean yield of crops between two states.*

EX: 6.9

50 Children were given special diet for a certain period and another control group of 50 children were given normal diet. Their average weight was found to be 4.2 lbs. and 5.7 lbs. respectively and the common standard deviation for gain in weight was 2 lbs. Assuming normality of distributions would you conclude that the special diet promoted weight?

Ans:

H_0 = There is no significant difference

The special diet did not promote the weight.

This is one tailed test because of testing of *promoting of weight* by special diet.

$Z = \frac{Diff}{SE}$, $SE = \sigma\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$ Here $n_1 = 50$, $n_2 = 50$

$SE = \sigma\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$

$SE = \sigma\sqrt{\left[\frac{1}{50} + \frac{1}{50}\right]} = 2 \times \sqrt{\frac{2}{50}} = 2 \times \sqrt{.04} = 2 \times .2 = .4$

$Z = \frac{Diff}{SE} = \frac{7.2-5.7}{.4} = \frac{1.5}{.4} = 3.75$

Level of significance = .05

Degree of freedom = infinity

Table value of Z = 1.645 (Since one tail test)

The calculated value (3.75) is numerically greater than table value (1.675). So, we REJECT the null hypothesis.

The special diet really promotes the weight.

EX: 6.10

Electric bulbs manufactured by X and Y Companies gave the following result

	X Company	Y Company
No. of Bulbs Used	100	100
Mean Life in Hours	1300	1248
Standard Deviation	82	93

Using standard error of the difference between mean, state whether there is any significant difference in the life of the two makes

Ans:

H_0= There is no significant difference. There is no significant difference in the life of 2 brands.

$Z = \frac{Diff}{SE}$, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $S_1 = 82$; $S_2 = 93$; $n_1 = 100$; $n_2 = 100$

$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{82^2}{100} + \frac{93^2}{100}} = \sqrt{\frac{6724}{100} + \frac{8649}{100}} = \sqrt{67.24 + 89.49} = \sqrt{153.73} = 12.4$

$Z = \frac{Diff}{SE} = \frac{1300-1248}{12.4} = \frac{52}{12.4} = 4.19$

Degree of freedom is infinity

The Z-Table value @ 5% level of significance is 1.96

The calculated value (4.19) is numerically greater than table value (1.96). So, we REJECT the H_0.

There is very significant difference in the life of 2 brands.

EX: 6.11

For a sample of 100 workers of from Kerala, the average daily wages are RS 10.50 with SD 1.50. For a sample of 150 workers from Tamil Nadu the corresponding figures are 8.00 and 1.00 respectively. Can you conclude that the average wages of workers in Kerala are more than workers of Tamil Nadu?

Ans:

H_0 = There is no significant difference average wages are not more

$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$; $S_1 = 1.5$; $S_2 = 1$; $n_1 = 100$; $n_2 = 150$

$= \sqrt{\frac{1.5^2}{100} + \frac{1^2}{150}} = \sqrt{\frac{2.25}{100} + \frac{1}{150}} = \sqrt{.0225 + .0067} = \sqrt{.0292} = .1708$

$Z = \frac{Diff}{SE}$; $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$Z = \frac{Diff}{SE} = \frac{10.50 - 8.00}{.1708} = \frac{2.50}{.1708} = 14.64$

Z-table value @ 5% level of significance for infinite degree of freedom is 1.645 (one tailed test). The calculated value (14.64) is numerically greater than the table value (1.645). So, we REJECT the H_0.

The average wages of workers of Kerala are higher than that of workers in Tamil Nadu.

EX: 6.12

A random sample of 200 villages was taken from district A and average proportion per village was 485 with SD 50. Another village random sample of 250 villages from the sample the same district gave an average population of 510 per village with SD of 40. Is this difference between the averages of these of the two sample statistically significant?

Ans:

Ho = There is no significant difference

Difference between average of two samples are not statistically significant.

SE = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$; S1 = 50, S2 = 40, n1 = 200, n2 = 250

$= \sqrt{\frac{50^2}{200} + \frac{40^2}{250}} = \sqrt{\frac{2500}{200} + \frac{1600}{250}} = \sqrt{12.5 + 6.4} = \sqrt{18.9} = 4.35$

Z $= \frac{Diff}{SE} = \frac{510 - 485}{4.35} = \frac{25}{4.35} = 5.75$

The value of Z @ 5% significant level is 1.96 The calculated value (4.35) is numerically more than the table value (1.96). So, we will reject the Ho.

The difference between the average of two samples are statistically significant.

EX: 6.13

The average number of articles produced by 2 machines per day are 200 and 250 with standard deviation 20 and 25 respectively on the basis of 25 day's production. Can you regard both the machines equally efficient @ 1% level of significance?

Ans:

Ho = There is no significant difference.

The machines are t equally efficient.

Here we will apply t-test.

t = $\frac{Diff}{SE}$; SE = $\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

n1 = 25 ; n2 = 25 ; S1 = 20 ; S2 = 25

$$SE = \sqrt{\frac{20X20^2+25X25^2}{25+25-2}\left(\frac{1}{25}+\frac{1}{25}\right)} = \sqrt{\frac{20X400+25X625}{50-2}(.04+.04)} = \sqrt{\frac{10000+15625}{48}(.08)}$$

$$= \sqrt{\frac{25625 \times .08}{48}} = \sqrt{\frac{2050}{48}} = \sqrt{45.708} = 6.535$$ t

$$= \frac{Diff}{SE} = \frac{250-200}{6.535} = \frac{50}{6.535} = 7.65$$ Degree of freedom = (25 + 25) – 2 = 48. Table value @ 1% of significant level is 2.58 The calculated value 7.65 numerically greater than the table value 2.58

so, we will REJECT the Ho. Both machjines are not equally efficient. **EX: 6.14**

In a test given to two groups of students the mark obtained were obtained were as follows:

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

Assuming that the groups' standard deviations are same, test the hypothesis that the group means are equal.

Ho = There is no significant difference The group means are equal.

here, we have to find out the mean and SD of both groups.

18	-19	361
20	-17	289
36	-1	1
50	+13	169
19	+12	144
36	-1	1
34	-3	9
19	+12	144
41	+4	16
333	**0**	**1134**

Group I

$$\bar{x}_1 = \frac{\Sigma x}{n} = \frac{333}{9} = 37 \qquad S_1 = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} = \sqrt{\frac{1134}{9} - \left(\frac{0}{9}\right)^2} = \sqrt{126 - 0^2} = 11.22$$

	d	d
29	-5	25
26	-8	64
28	-6	36
35	+1	1
30	-4	16
44	+10	100
46	+12	144
238	**0**	**386**

Group ii

$\bar{x}_2 = \frac{\Sigma x}{n} = \frac{238}{7} = 34$ $S_2 = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{386}{7} - \left(\frac{0}{7}\right)^2} = \sqrt{55.14 - 0^2} = 7.426$

SE = $\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{9x11.22^2 + 7x7.426^2}{9+7-2}\left(\frac{1}{9} + \frac{1}{7}\right)} = \sqrt{\frac{9x126 + 7x55.14}{16-2}(.111 + .143)}$ =

$\sqrt{\frac{1134+385.98}{14}\ x\ (.254)} = \sqrt{\frac{1519.98}{14}\ x\ (.254)} = \sqrt{108.57\ x\ .254}$ =

$\sqrt{25.577} = 5.25$ t value =

$\frac{Diff}{SE}$ or $\frac{\bar{x}_1 - \bar{x}_2}{SE}$ $= \frac{37-34}{5.25}$

$= \frac{3}{5.25} = .57$

Table value of t for 14 degree of freedom @ .05 level of significant is 2.145.

The calculated value .57 is numerically less than the t-table value 2.145.

Thus we will ACCEPT the Ho.

The group means are equal.

Ex 6.15 Average life of 26 bulbs is 1200 hours with standard deviation 15o hours. Test whether these bulbs could be considered as a random sample from a normal population with mean 1300 hours.

Ho : there is no significant difference. Bulbs could be considered a random sample from a normal population.

Small sample test.

Sample mean = $\bar{x}$ = 1200 Population mean μ = 1300 σ = 150 n = 26

Standard Error = = $\frac{\sigma}{\sqrt{n-1}}$ = $\frac{150}{\sqrt{25}}$ = 30

Z value = = $\frac{Difference}{standard\ Error}$ = $\frac{1300 - 1200}{30}$ = 3.33

t table value @ 0.05, degree of freedom 25 = 2.06

Since calculated t value is more than t table value, difference is considered significant. Hypothesis is rejected. Bulbs cannot be considered random sample from normal population.

Parametric Tests for proportions

Just as means can be compared and examined for determining significance of their differences, proportions can be subjected to significance testing. By testing the difference of two population proportions, or difference between one sample and population, we can decide whether sample proportion differs significantly from given sample proportion or whether 2 samples come from populations having the same proportion of success.

In parametric tests for proportions, the null hypothesis will be 'there is no significant difference between sample proportion and population proportion. Proportion tests follow normal distribution. Proportion tests are generally conducted as large sample tests.

Proportion test – population vs sample.

Steps

1. Form Null hypothesis
2. Consider given population proportion, sample proportion and sample size.
3. Find standard error = $\sqrt{\frac{pq}{n}}$
4. Obtain Z value = = $\frac{Difference}{standard\ Error}$
5. Compare with Z table value at significance level and degree of freedom.
6. Decide the fate of null hypothesis

Ex 6.16 A population survey indicates that out of 3232 births, 1705 were boys and 1527 girls. Do these figures confirm the hypothesis that the gender ratio is 50:50?

Null Hypothesis : No significant difference between sample proportion and population proportion. Gender ration is 50:50

Given p = 0.5 q = 0.5 n = 3232

Standard error = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{.5\ x\ .5}{3232}} = .0088$

Z value = $\frac{Difference}{standard\ Error} = \frac{.5275 - .5000}{.0088} = 3.125$

Z table value at α = .05 == 1.96. The calculated value 3.125, is more than Z table value. Hence the difference between proportions is significant. So, the hypothesis is rejected. Gender ratio is not 50;50.

Proportion test - Sample Vs Sample

Sometimes proportions of 2 samples may be given, instead of population proportion. In that case population proportion may be estimated, on the basis of sample proportion and sample size. And the n standard error obtained.

Steps

1. Form null hypothesis
2. Consider given sample proportions and sample size.
3. Obtain standard error of proportion = $\sqrt{PQ\frac{1}{n1} + \frac{1}{n1}}$ (Where P = population proportion and Q is 1-P
4. Calculate population proportion P = $\frac{n1p1 + n2p2}{n1+n2}$
5. Find Z value = = $\frac{Difference}{standard\ Error}$
6. Compare with Z table value and decide fate of Ho

Ex 6.17

In a city 400 out 500 men were smokers. After an awareness program, a random sample of 600 men revealed 400 smokers. Is the awareness program effective?

Ho = No significant difference between sample proportions. Awareness program is not effective.

Given = $\frac{n1 \quad =500 \quad p1 = .8 \quad n2 \quad 600 \quad p2 \quad = .67}{}$

P = = $\frac{n1p1 + n2p2}{n1+n2} = \frac{500\ x\ 0.8\ + 600\ x\ .67}{500 + 600}$ 0.73

Q = 1 - .73 = .27

Standard Error $= \sqrt{PQ\frac{1}{n1} + \frac{1}{n1}} = \sqrt{.73\ x\ .27\ (\frac{1}{500} + \frac{1}{600})} =).027$

Z value = $= \frac{Difference}{standard\ Error} = \frac{p1\ -\ p2}{standard\ Error} = \frac{.8\ -\ .67}{.027}$ 4.81

Z table value at α = .05, = 1.96 and calculated Z value is more than table value. Therefore the difference is considered significant. The hypothesis is rejected. Ultimately awareness program is effective.

Ex 6.18

A random sample of 16 men from Malappuram District had a mean height of 68inches and sum of squares from mean 132. 25 men from Amrithser district had the corresponding values 66.5 and 165 respectively. Do the samples belong to the same population?

Ho No significant difference between samples. They belong to the same population

Given . $\bar{x}1 = 68$ $\bar{x}2 = 66.5$ $\Sigma (x1 - \bar{x}1)^2 = 132 = n1s1^2$ $\Sigma (x2 - \bar{x}2)^2 = 165 = n2s2^2$

n1 = 16 n2 = 25

Standard Error = $\sqrt{\frac{n1s1^2 + n2s2^2}{n1 + n2 - 2}} (\frac{1}{n1} + \frac{1}{n2})$

$= \sqrt{\frac{132 + 165}{161 + 252 - 2}} - (\frac{1}{16} + \frac{1}{25}) = \sqrt{0.78} = 0.88$

t value = $\frac{Difference}{standard\ Error}$ $\frac{1.5}{.88} = 1.697$

t table value at α = 0.05 and υ = 39 , = 2.70

Since the calculated t value is less than t table value, the difference is insignificant. Hypothesis is rejected. Samples belong to the same population

Review Questions and Exercises

1. What are large samples and small samples?
2. Explain the steps in large samples mean tests
3. What is the procedure in small sample mean test?
4. What are
5. why mean test is popular
6. What are the steps in testing population against sample
7. State how two sample groups can be tested as to mean
8. State the steps in testing two large populations
9. State model hypothesis in a proportion test

10. It is claimed that a random sample of 10 tyres with mean life of 15400 km is drawn from a population of tyres which has a mean life of 16666 kms. With standard deviation of 1200 kms. Test the validity of the claim.
11. Average life of 50 bulbs were found to be 1500 hours with a standard deviation 150 hours. Test whether these bulbs could be considered as a random sample from a normal population with mean 1600 hours.
12. A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population with mean 100 and standard deviation 8 at 5% level of significance.
13. A sample of size 99 items is taken from a population with standard deviation 15. The mean of the sample is 25. Test whether the sample has come from a population with mean 26.8
14. A stenographer claims that she can take dictation at the rate of 120 wpm. Can we reject her claim on the basis of 100 trials in which she has mean of 116 wpm with a variance of 225 words?
15. A random sample of 200 tins of coconut oil gives an average weight of 4.95 kg with a standard deviation of .21kg. do we accept the hypothesis of net weight of 5kg per tin at 1% level of significance
16. Of two salesmen – X claims that le has made larger sales than Y. from following details, can you say that A is a better performer?

	No of sales	Mean sales	Standard deviation
X	10	6200	690
Y	17	5600	600

17. Two batches of same product are tested for their mean life. Test the hypothesis that mean life is better for Batch I?

Batch	Sample size	Mean life	Standard deviation
I	10	750	12
II	8	820	14

18. In a sample of 500 people in Kerala 280 are tea drinkers and the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in this State at 1% level of significance?
19. A company producing glass roads claims that there would be 25 defectives per 100. When a sample of 5oo was taken, number of defectives found was 150. Test the correctness of the company's claim.
20. A candidate in an election from a large city thinks that he will win with 45% of votes. He conducted a survey which covered 10000 voters and found that 4450 voted in his favor. Is his claim valid?
21. Out of 8oo literate people 480 were employed and out 600 illiterate, only 350 are employed. Is the difference between two proportions of employed persons significant?
22. Out of a sample of 600 men from a city 450 were smokers. In another sample of 900 men another city, 450 were smokers. Do the data indicate that the cities are significantly different in smoking?

UNIT VII

PARAMETRIC TESTS FOR VARIANCES AND PAIRED OBSERVATIONS

Two popular parametric tests are variance tests and paired observations tests, after mean tests and proportion tests. Variance tests focus on significance of difference between population variance and a sample variance , or between a sample variance and another sample variance. Paired observations tests examine significance of difference between two dependent sample groups.

Tests for variances

Significance tests may be designed to examine equality of variance of two samples or populations. Such tests examine whether one population is significantly differ from other , or two samples are randomly drawn from a population or not. In other word, we are able to determine whether two independent estimates of population variance are homogenous or not. The test can be also used to examine equality of standard deviation of two sample groups. In variance tests there is no need of standard error. The test statistic in tests of variances follows Snedcor's F distribution.

We are not always interested in means and proportion. In many situations responsible decision makers have to make inferences about its variability, within a population. A sociologist investigating effect of education on earning capacity may be eager to know whether income of college graduates is more variable than income of post graduates.

Steps

1. Formulate Null hypothesis
2. Consider given observations, standard deviations and sample size.
3. Find F Ratio = $\frac{n1s1^2}{n1-1} \div \frac{n2s2^2}{n2-1}$
4. Compare the obtained F value with F table value
5. Accept or reject null hypothesis , as the case may be

Assumptions

1. Populations from which samples are drawn is normally distributed
2. Samples are randomly drawn and independent of each other
3. Means of population or samples are taken to be equal.

Ex . 7.1

In a rat feeding experiment ,high protein was given to sample x containing 12 rats and low protein to sample y with 7 rats. Weights gained by them (in gms) are as below. Examine whether protein leads to weight gain.

X 13 14 10 11 12 16 10 8 11 12 9 12
Y 7 11 10 8 10 13 9

Ho : No significant difference . Protein does not lead to weight gain.

X^2	Y^2
169	49
196	121
100	100
121	64
144	100
256	169
100	81

	64	
	121	
	144	
	81	
	144	
Total	1640	684

$S1^2 = \frac{\Sigma x^2}{n1} - (\frac{\Sigma x}{n})^2 = = \frac{1640}{12} - (\frac{138^2}{12}) = 4.42$

$S2^2 = \frac{\Sigma y^2}{n1} - (\frac{\Sigma y}{n})^2 = = \frac{684}{7} - (\frac{68^2}{7}) = 3.43$

F value = $\frac{n1s1^2}{n1-1} \div \frac{n2s2^2}{n2-1} = \frac{12\ x\ 4.42}{11} \div \frac{7\ x\ 3.43}{6} = 1.205$

F table value $\alpha = 0.05$ and $\upsilon = 11, 6 = 4.03$.

Since the calculated F value is much less than the F table value, the difference is considered significant. Ho is accepted. Protein does not lead to weight gain.

Ex 7.2

An economist believes that income earned by graduates is more variable than non graduate employees. A sample of 21 graduates have earning with standard deviation of 17000, and income of 25 non graduates gave a standard deviation of 7500. Is his belief true?

Ho : No significant difference. Income variability is equal.

Given : n1 = 21 n2 = 25 s1 = 17000 s2 == 7500

F ratio = = $\frac{n1s1^2}{n1-1} \div \frac{n2s2^2}{n2-1} = \frac{21\ x\ 17000}{20} \div \frac{25\ x\ 7500}{24}$ 5.18

F table value $\alpha = 0.05$ and $\upsilon = 20, 24 = 2.74$.

Since the calculated F value is more than F table value, the null hypothesis is rejected, and the difference is significant. The income variability of graduates is more than non graduates.

Ex 7.3

Quality controller of Hero Honda is concerned with uniformity in the number of defects in bikes coming off two assembly lines. If there is significant variability in defects, he wants to re install the assembly lines. Test at a 0.05 level of significance.

	Defects in A	Defects in B
Mean	10	11
Variance	9	25
Sample size	20	16

HO ; no significant difference Variability in defects are equal.

Given : n1 = 20 n2 = 16 $s1^2$ = 9 $s1^2$ == 25

F ratio = $= \frac{n1s1^2}{n1-1} \div \frac{n2s2^2}{n2-1} = \frac{20\ x\ 9}{19} \div \frac{16\ x\ 25}{15} = \underline{3.55}$

F table value $\alpha = 0.05$ and $\upsilon = 19,15 = \underline{2.20}$

Since the calculated F value is more than F table value, the null hypothesis is rejected, and the difference is significant. The income variability of graduates is more than non graduates.

Tests for paired observations

Mean tests, proportion tests or variance tests relate to independent sample groups. The samples were drawn independently from a population, on random basis. However there are situations where samples are related in one way or other and they have to be tested as dependent sample groups.

Pre-post approach or before-after approach in research provide dependent samples. For example, number of consumer disputes before the Act of 1986 and disputes after the Act is Before-after approach. In such cases, we would like to test whether disputes have reduced or increased as an impact of the Act.

Generally data in this case of dependent samples will be given in pairs. Therefore the test of dependent samples is also called test for paired observations.

For example, disputes before and after Act will form 2 dependent samples. The variables in the two samples are the same. A value from one sample is given , with the corresponding value in the other sample, are together taken and forms a pair. Thus we get pairs of observations

Paired observation focuses on average deviations between pairs (d) and standard error. Here no parameter is employed or no distribution is adopted.

Assumptions

1. Samples are dependent on each other, in some way or other
2. Samples follow either t distribution or z distribution.
3. Samples are drawn randomly, from normal or approximately normal distribution

Steps:

1. Form null hypothesis
2. Ascertain d values = difference between pairs.
3. Find standard error SE = $\frac{S}{\sqrt{n-1}}$ where,

   s = standard deviation of difference = $\frac{\Sigma d^2}{n} - (\frac{\Sigma d}{n})^2$
4. Obtain t value = $\frac{\bar{d}}{SE}$ where $\bar{d}$ = average of differences ,
5. Compare with t table value at appropriate α and υ
6. Decide fate of hypothesis.

Ex 7.4

To test efficacy of sleeping pills, 5 person are selected. Time before sleep is given below in seconds. Test whether sleeping pills are effective.

Person	No pills	pills
A	65	45
B	35	15
C	80	61
D	40	31
E	50	20

Ans

Person	No pills	pills	d	d^2
A	65	45	20	400
B	35	15	20	400
C	80	61	19	361
D	40	31	9	81
E	50	20	30	900
		Total	98	2142

$$s = = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = \frac{2142}{5} - \left(\frac{98}{5}\right)^2$$

$$SE = \frac{S}{\sqrt{n-1}} = \frac{6.9}{\sqrt{9}} = 2.3$$

$$T \text{ value} = \frac{\bar{d}}{SE} = \frac{4.7}{2.3} = -2.04$$

Ex 7.5

Ten students scored following marks in two tests as below. Examine, if there is significant difference in their performance.

Test1 67 24 57 55 63 54 56 68 33 43

Test 2 70 38 58 58 56 67 68 72 42 38

Test 1	Test 2	d	d
67	70	-3	9
24	38	-14	196
57	58	-1	1
55	58	-3	9
63	56	+7	49
54	67	-13	169
56	68	-12	144
68	72	-4	16
33	42	-9	81
43	38	+5	25
		-47	699

$$s = = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = \frac{699}{10} - \left(\frac{-47}{10}\right)^2 = .69$$

$$SE = \frac{S}{\sqrt{n-1}} = \frac{6.9}{\sqrt{9}} = 2.3$$

$$t \text{ value} = \frac{\bar{d}}{SE} = \frac{4.7}{2.3} = -2.04$$

t table value at degree of freedom 9, and level of significance 5% = 2.26

Since calculated t value is less than t table value we accept the hypothesis that there is no difference in performance.

Ex 7.6

A certain stimulus administered to 12 chicken resulted in the following increase of weights. 5 2 8 -1 3 0 -2 1 5 0 4 6. Can it be concluded that the stimulus increases weight?

Ans Ho : No significant difference. Stimulus does not increase weight.

d	dType equation here.
5	25
2	4
8	64
-1	1
3	9
0	0
-2	4
1	1
5	25
0	0
4	16
6	36
31	185

$$s = = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = \frac{185}{12} - \left(\frac{31}{12}\right)^2 = 2.96$$

$$SE = \frac{S}{\sqrt{n-1}} = \frac{2.96}{\sqrt{11}} = 0.89$$

$$t\ value = \frac{\bar{d}}{SE} = \frac{2.58}{0.89} = 2.9$$

t table value at degree of freedom 11, and level of significance 5% = 1.79

Since calculated t value is more than t table value we reject the hypothesis that there is no increase in weight. There is surely increase in weight.

Review Questions and Exercises

1. What is a significant test for variance
2. What are the assumptions in variance test?
3. Give the formula for obtaining variance
4. What are the steps in dependent sample test
5. What is f-ratio?
6. What are dependent samples?
7. State the hypothesis in a paired observation test
8. State the steps in paired observation test
9. What are the assumptions in dependent sample test?

10. The standard deviations of two samples of sizes 10 and 14 from two normal populations are 3.5 and 3.0 respectively. Examine whether the standard deviations of the populations are equal.
11. A random sample of pigs fed on diet A over a period gave the following values – mean 1 = 6, standard deviation 1 = 3.08, n1 = 8 and another sample fed on diet B gave the following values - mean 2 = 8, standard deviation 2 = 4.15, n2 = 5. Test whether the diets A and B significantly differ in their means, and in their variances.
12. Two samples are drawn from two normal populations. From the following data test whether the two samples have the same variance at 5% level of significance

Sample 1 60 65 71 74 76 82 85 87
Sample 2 61 66 67 85 78 63 85 86 88 91

13. Ten soldiers visit a rifle range for two consecutive weeks. for the first week, their scores are 61, 26, 57, 55, 63, 54, 56, 68, 33, 43 and during the second week, they score in the same order 56, 36, 58, 58, 56, 67, 68, 72, 42, 38. Examine, if there is significant difference in their performance. Conduct paired observation test.
14. A certain medicine administered to each of the 12 patients resulted in the following increase of blood pressure 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. can it be concluded that stimulus will, in general be accompanied by an increase in blood pressure, by doing paired t test.
15. Two sets of 10 students selected at random from a college were taken, and were given memory test and their scores were;

set A:	10	8	7	9	8	10	9	6	7	8
set B:	12	8	8	10	8	11	9	8	9	9

Test whether there is a significant difference in mean scores. Do paired samples test.
(the value of t for 18 d.f = 2.101)

16. 10 accountants were given intensive coaching and four tests were conducted in a month. the scores of the tests 1 and 4 are given below:

Sl.no:	1	2	3	4	5	6	7	8	9	10
marks in 1 test:	50	42	51	42	60	41	70	55	62	38
marks in 4 test:	62	40	61	52	68	51	64	63	72	50

Examining the result of paired t test, does the score from test1 to test 4 show an improvement?

UNIT VIII

ANALYSIS OF VARIANCE

Analysis of variance , abbreviated as ANOVA, is an extremely useful technique concerning researchers in the field of economics, biology, education, psychology, sociology, business and industry and in researches of several other disciplines. This technique is used when multiple samples cases are involved. As stated earlier, the significance of the difference between the means of two samples can be judged through either Z test or the T test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher. Using this technique, one can draw inferences about whether sample have been drawn from populations having the same mean.

ANOVA technique

The ANOVA technique is important in the context of all those situations where we want to compare more than two populations such as comparing the yield of crop from several varieties of seeds, analyzing the gasoline mileage of four automobiles, studying the saving habits of five groups of university students and so on. In such circumstances, one generally does not want to consider all possible combinations of two populations at a time, because that would require a great number of tests before we would be able to arrive at a decision. This would also consume a lot of time and money, and even then certain relationships may be left unidentified , particularly the interaction effect. Therefore, one can quite often utilize the ANOVA technique and through it investigate the differences among the means of all the populations simultaneously.

Prof. R.A.Fisher was the first man to use the term Variance, and in fact, it was he who developed a very elaborate theory concerning ANOVA and its usefulness in practical field. Later on Professor Snedcor and many others contributed to the development of this technique. ANOVA is essentially, a procedure for testing the difference among different groups of data for homogeneity. The essence of ANOVA Is that total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes. There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes. Hence, it is a method of analyzing the variance to which a response is subject into its various components corresponding to various sources of variation. Through this technique one can explain whether various varieties of seeds or fertilizers or soils differ significantly so that a policy decision could be taken accordingly, concerning a particular variety in the context of agriculture researches. Similarly, the differences in various types of feed prepared for a particular class of animal or various types of drugs manufactured for curing a specific disease may be studied and judged to be significant or not , through the application of ANOVA technique. Likewise, a manager of a big concern can analyse the performance of various salesmen of his concern in order to know, whether their performances differ significantly.

Thus in general, through ANOVA technique, one can investigate any number of factors which are hypothesized or said to influence the dependent variable. One may as well investigate the differences amongst various categories within each of these factors which may have a large number of possible values. If we take on only one factor and investigate the differences amongst its various categories having numerous possible values, we are said to use one way ANOVA . And in case we investigate two factors at the same time, then we use two way ANOVA . In two way ANOVA, the interaction or inter relationship of two factors affecting the values of a variable can be studied for better decisions.

Basic principle of ANOVA

The basic principle of Anova is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variations between the samples. In terms of variation within the given population, it is assumed that the observations differ from the means of this population only because of random effects, ie, there are influences , which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the population and the grand mean is attributable to what is called a specific factor, or what is technically described as treatment effect. Thus while assuming these populations has the same variance, we also assume that all factors other than the one or more being tested are effectively controlled. This , in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factors to be studied.

The total variance in the joint sample is partitioned into two parts

- Variance between samples
- Variance Within samples.

Variance between samples is due to different treatments, while within samples variance is due to the random unexplained disturbance. Using these two variances, we define the test statistic as F value = variance between samples/variance within samples.

Using this method, we wish to test that all population means are the same or not

Test in Anova

Statistical test applied in ANOVA is f test. The test statistic is f value which is the ratio between variance between samples and variance within samples. If the f test statistic value is less than the corresponding table value of f, we accept null hypothesis. in that case, we conclude that samples do not differ significantly or both samples belong to population with same values.

Assumptions in ANOVA

1. Populations from which samples have been drawn are normally distributed.
2. Populations from which the samples are drawn have same variance.
3. The observations are non correlated random variables.
4. Any observation is the sum of the effects of the factors influencing it.
5. The random errors are normally distributed with mean 0 and a common variance σ^d

One way classification

ANOVA is done under two circumstances – one way classification and two way classification. Under one way ANOVA, we consider only one factor and there we assume that reason for variations in the factor may be due to variance between samples and variance within samples.

For example, suppose we want to study sales made by three salesman – A ,B, and C for 4 weeks. We would like to test whether sales made by 3 salesman significantly differ between each other. For this, variances between salesman and variances within salesman will be thoroughly analyzed and measured. This is done through one way ANOVA.

Variances in one way ANOVA

There are three types of variances in one way classification of ANOVA.

1. variances between sample - MSC

This is the most important variances in ANOVA. And is the net variation of different samples mean from the grand mean (mean of mean)

2. variance within samples - MSE

This is the net result of variation between observation of each sample and the sample mean. This is also called residue variance.

3. Total variance - this is the sum total of all variances for all observations taken together. It is also called variance about sample and is used to ascertain residue variance. (residue variance = total variance –between variance)

F ratio in one way ANOVA is computed as a ratio between variance between samples and variance within samples. thus,

F ratio $= \frac{VARIANCE\ BETWEEN\ SAMPLES}{VARIANCE\ WITHIN\ SAMPLES} = = \frac{MSc}{MSE}$

Steps in one way Anova

1. Form h_o that there is no significant difference between samples
2. Compute mean of each samples – $\bar{x}_1\ \ \bar{x}_2\ \ \bar{x}_3$ .....
3. Calculate grand mean = mean of sample means = $\bar{\bar{x}}$
4. Obtain sum of square between samples = SSC $= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2$
5. Ascertain Mean Square between samples = MSC $= = \frac{SSC}{C-1}$
6. Obtain sum of squares within samples = SSE $= n_1(x - \bar{x}_2)^2 + n_2(x - \bar{x}_2)^2 \ldots\ldots$
7. Compute mean square within samples = $MSE = \frac{SSE}{N-C}$
8. Where N = Total number of observations and C = number of columns
9. Calculate F ratio = $\frac{MSC}{MSE}$
10. Compare with F table value and decide fate of h_o
11. Present the result in ANOVA table

Ex. 8.1

Following are the sales by 3 salesmen. is A better performer than b and c?

	sales	
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

A. Ho – no significant difference. A is not a better performer. All are equal

	sales	
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4
Total 24	20	16

$\bar{x}_1 = \frac{24}{4} = 6$ $\bar{x}_2$ = 20/4 = 5 $\bar{x}_3$ = 16/= 4

Mean of means = $\frac{6+5+4}{3} = 5$

Sum of square between = SSC = $4(6-5)^2+4(5-5)^2+4(4-5)^2=8$

12. Mean sum of square = $\frac{SSC}{C-1} = \frac{8}{3-1}=4$

Sum of square within = SSE = $(6-6)^2,(7-6)^2,(3-6^2) \ldots\ldots\ldots\ldots = 24$

mean square within = MSE = $\frac{SSE}{N-C} = \frac{24}{12-3} = 2.67$

F ratio = MSC / MSE = 4/2.67 = 1.5

F table value = 4.26

Calculated f value 1.5 is less than f table value = 4.26

Difference is not significant. Ho is accepted, A is not a better performer. All are equal.

Short- cut method

The above explained method of ANOVA involves much calculation. To simplify calculation, the following short-cut method may be employed :

steps

1. Find correction factor = $T^2/n = \frac{(SUMOF\ ALL\ OBSERVATIONS)^2}{TOTAL\ NUMBER\ OF\ ITEMS}$
2. Find sum of squares total = SST = sum of squares of all observations - T^2/n
3. Obtain sum of square columns = SSC = $\frac{(COLUMN\ 1\ TOTAL)^2}{n} + \frac{(COLUMN21\ TOTAL)^2}{n}$.....
4. Find mean square column = MSC = $\frac{SSC}{C-1}$
5. Obtain sum of squares within = SSE = SST – SSC
6. find mean square within = $\frac{SSE}{N-C}$
7. find f ratio = $\frac{MSE}{MSE}$
8. present the result in ANOVA table
9. compare with f table value and decide the fate of ho

Ex 8.2

Following are the scores of three batsmen. Examine whether B is the best among the three.

A	B	C
30	51	44
27	47	35
42	37	41
	48	36
	42	
Total 99	225	156

T = Sum of all observations = 480 $T^2/N = \frac{480 \times 480}{12} = 19200$

SST = sum of squares of all observation = T^2/N = 578

SSC = $\frac{(99)2}{3} + + \frac{(225)2}{5} + \frac{(156)2}{4} - 19200 = 276$

SSE = SST - SSC = 578 - 276 = 302

MSC = $\frac{SSC}{C-1} = \frac{276}{2} = 138$

MSE = = $\frac{SSE}{N-C} = \frac{302}{9} = 33.56$

Source of variation	Sum of squares	Degree of freedom	Mean squares	F ratio
Between samples Within sample	SSC = 276 SSE = 302	C - 1 = 2 N - C = 9	138 33.56	Fc=$\frac{138}{33.56} = 4$ 11

Table value of F, at 5% level of significance, degree of freedom C-1 x N – C = 2,9, = 4.26. Calculated value is less than table value. Therefore we accept the hypothesis that the performance of batsmen are equal. Type equation here.

Two way classification

Two way ANOVA is used when data are classified on the basis of two factors simultaneously for example, sales made by salesman man be given vertically, and values of sales in 4 weeks may be given horizontally.

in two way ANOVA, 3 variables are analyzed – variance between columns (MSC), variance between rows (MSR) and variance within or residue (MSE). Since calculations are more complex in two way ANOVA, we use short-cut method.

Steps in two way classification

1. form ho that al column samples and row samples are equal
2. compute $T^2/n = \frac{(TOTAL\ OF\ ITEMS)^2}{TOTAL\ NUMBER\ OF\ ITEMS}$
3. find SST = sum of squares of all observations - T^2/n
4. find SSC = $\frac{(COLUMN\ TOTAL)^2}{NO.OF\ COLUMN\ ITEMS}$ + ... - T^2/n
5. obtain mean square column = MSC = $\frac{SSC}{C-1}$
6. Find SSR = $\frac{(ROW\ TOTAL)^2}{NO.OF\ ROW\ ITEMS}$ + ... - T^2/n
7. Obtain mean square row = MSR = $\frac{SSR}{R-1}$
8. Find SSE = SST – SSC + SSR

9. Obtain mean square residue = $\frac{SSE}{N-C}$
10. Find two F ratios = Fc = $\frac{MSE}{MSC}$ and Fr = $\frac{MSR}{MSE}$
11. Present the results in ANOVA table
12. compare with 2 F table values and decide fate of h_o

ANOVA table

The results obtained from analysis of variance – whether one way classification or two way classifications can be presented in a table called ANOVA table. It shows sum of variances, sum of squares, degrees of freedom, type of variance and f ratio, along with table values. it facilitates comprehension , comparison and analysis.

Ex. 8.3

From the following data on production by 4 machines by 5 workers, test

a) whether productivity differs between machines
b) whether workers differ in productivity

		machines		
weeks	a	b	c	d
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

Ans

Coding may be applied to reduce the data size. thus, deducting 41 from all the values, coded data will be as below:

Ho : Productivity does not differ between machines. Workers do not differ in productivity

Let us apply coding method and subtract 41from all the values

A	B	C	D	TOTAL
3	-3	6	-5	1
5	-1	11	2	17
-7	-5	3	-9	-18
2	-3	5	-8	-4
-3	1	8	-2	4
0	-11	33	-22	0

N=20

T=Sum of all the value=3+5-7.....+8-2=0 $\frac{T2}{N}$=0

SST=sum of squares of all values=[3^2+5^2+.....+(-2) 2]- 0^2=574

SSC=$\frac{(\varepsilon x1)2}{n1}+\frac{(\varepsilon x2)2}{n2}+\ldots\ldots-\frac{T2}{N}$ $=\frac{(0)^2}{5}+\frac{(-11)^2}{5}+\frac{(33)^2}{5}+\frac{(-22)^2}{5}-0$ = 338.8

SSR=$\frac{(\varepsilon x1)2}{n1}+\frac{(\varepsilon x2)2}{n2}+\ldots\ldots-\frac{T2}{N}$ $=\frac{(1)^2}{4}+\frac{(-17)^2}{4}+\frac{(-18)^2}{4}+\frac{(-4)^2}{4}+\frac{(4)^2}{4}$ -0 =161.5

SSE=SST-SSC-SSR = 574 - (338.8-161.5) =73.7

MSC = $\frac{SSC}{C-1}$ $=\frac{338.8}{3}$ = 112.93;

MSR = $\frac{SSR}{R-1}=$ $\frac{161.5}{4}$ = 40.38

MSE= $\frac{SSE}{(C-1)(R-1)}$ = $\frac{73.7}{12}$ = 6.14

Source of variation	Sum of squares	Degree of freedom	Mean squares	F ratio
Between columns	338.8	3	112.93	Fc=$\frac{112.93}{6.14}$=18.39
Between rows	161.5	4	40.8	Fr =$\frac{40.8}{6.14}$=6.58
Residual	73.7	12	6.14	
Total	574	19	-	

Between columns

Degree of freedom=(3,12) Table value of f=3.49

Calculated value of Fc=18.39 which is greater than the table value.

We reject the hypothesis. Mean of columns are not equal. Mean productivity is not the same for different machines.

Between rows

Degree of freedom(4,12) Table value of F=3.26

Calculated value of Fr=6.58 which is greater than the table value.

We reject the hypothesis . Mean of rows are not equal. Workers do differ in productivity..

Review Questions and Exercises

1. Describe Anova
2. What is mean by analysis of variance
3. Explain the uses of Anova
4. Discuss various assumptions in Anova techniques
5. What is the null hypothesis in Anova
6. What is mean by coding with reference to Anova
7. Discuss the technique of Anova one-way classification procedure
8. Explain the technique of Anova in two-way classification of data
9. Give a specimen of (i) Anova table for one-way analysis (ii) Anova table for two-way an analysis
10. Distinguish between one-way analysis and two way analysis of variance

11. State assumptions in analysis of variance
12. what is variance
13. Explain different types of variance
14. distinguish between one way and two way Anova
15. what is the null hypothesis in the analysis of variance?
16. What is meant -by coding with reference to analysis of variance
17. discuss the technique of analysis of variance of one way classification procedure.
18. explain the technique of analysis of variance in two way classification of data.
19. A special fertili;er was experimented on four fields a, b, c and d.in each field 4 beds were prepared and fertilizer was used. The yields of the beds of a, b,c and d fields are given below. Find out the difference between the means of the yields in fields is significant or not(the value of f at 5% level of significance for v2 = 3 and v1 = 12 is 8.74)

A	B	C	D
8	9	3	3
12	4	8	7
1	7	2	8
3	1	5	2

20. Three varieties Of wheat - A, B, and C were sown in 4 plots each and the following yields in quintals per acre were obtained.

	Varieties		
plots	a	b	c
1	10	9	4
2	6	7	7
3	7	7	7
4	9	5	6

Set up a table of analysis of variance and find out whether there is a significant difference between the mean yields of the three varieties.(the table value of f at 5% level is 4.26)

21. Set a table of analysis of variance for the following data

	variety			
plots	a	b	c	d
1	200	230	250	300
2	190	270	300	270
3	240	150	145	180

Test whether the varieties are different

22. Following figures relate to production in kilogram of three varieties a, b and c of wheat sown in 12 plots.

A	B	C
14	14	18
16	13	16
18	15	16
22	19	20

Is there significant difference in the production of three varieties?

UNIT IX

NON-PARAMETRIC TESTS - CONCEPTS

All the tests of significance discussed earlier were based on certain parameters. These tests were applied under certain assumptions about population, like normal distribution, t distribution, f distribution etc.

There are situations where population parameter is not available or assumption about population cannot be made. In such situations, various assumptions required for standard tests of significance, such as population is normal, samples are independent, standard deviation is known etc, cannot be made, then we can use non parametric methods. Moreover, they are easier to explain and easier to understand. This is the reason why such tests have become popular. But one should not forget the fact that they are usually less efficient or powerful as they are based on no assumptions, and we all know that the less one assumes, the less one can infer from a set of data. But then the other side must also be kept in view that the more one assumes, the more one limits the applicability of one's methods. Non parametric tests are quantitative techniques designed for such situations.

A statistical test is a formal technique, based on some probability distributions, for arriving at a decision about the reasonableness of an assertion or hypothesis. The test technique makes use of one or more values, obtained from sample data to arrive at a probability statement about the hypothesis. But such a test technique also makes use of some more assertions about the population from which the sample is drawn. For instance, it may be assumed that population is normally distributed, sample drawn is a random sample and similar other assumptions. The normality of the population distribution forms the basis for making statistical inferences about the sample drawn from the population. But no such assumptions are made in case of non parametric tests.

Features of Non parametric tests

Non-Parametric test do not assume any distribution, or is not based on any statistic or parameter. They exhibit following features:

1- Non parametric test is not based on standard error concept. They directly deal with the observations.
2- They do not suppose any particular distribution and consequential assumptions
3- They are rather quick and easy to use, ie, they do not require laborious computations since in many cases the observations are replaced by their rank or order and in many cases we simply use signs.
4- They are often not as efficient or sharp as tests of significance or the parametric tests. An interval estimate with 95% confidence may be twice as large with the use of non parametric tests as with regular standard methods.
5- When our measurements are not as accurate as is necessary for standard tests of significance, the non parametric methods come to our rescue which can be used fairly satisfactorily.
6- Parametric tests cannot apply to ordinal or nominal scale but non parametric tests do not suffer from any such limitation.
7- Parametric tests of difference like t test or F test make assumption about the homogeneity of the variance whereas this is not necessary for non parametric tests of difference.

Rationale for Non-Parametric Tests

Parametric tests are useful and handy for heavy practical studies. It helps to decide whether a hypothesis should be accepted or rejected. But there are certain constraints on the use of parametric test.

1. Parametric tests are based on Central Limit Theorem. It is applicable in the case of large samples only.
2. If the form of the distribution form which samples are drawn is skewed or is non-normal, the parametric tests will not yield meaningful result.
3. These tests are not applicable for ordinal or nominal measurement, because they are not appropriate for qualitative expression and analysis.
4. It may lead to misleading results, when samples drawn are very small or inadequate.

Situation When Non-Parametric Test Are Applied

Non parametric tests cannot be applied without discretion. They are situation specific in the following circumstances; non parametric tests can be applied with caution.

1. When no distribution cannot be relied upon
2. When hypothesis does not involve a parameter of the population.
3. When the observations are not as accurate as required for a parametric test.
4. When assumptions necessary for the validity of a parametric test are not clearly and correctly known.
5. When data re given on nominal or ordinal scale.

Shortcomings of parametric tests

In parametric tests two kinds of assertions are involved, viz., an assertion directly related to the purpose of investigation and other assertions to make a probability statement. The former is an assertion to be tested and is technically called a hypothesis, whereas the set of all other assertions is called the mode. When we apply a test to examine the hypothesis, without a parameter or statistic, it is known as distribution free test, or non parametric test. Non parametric tests do not make an assumption about parameters of the population and thus do not make use of the parameters of the distribution. In other words, under non parametric or distribution free tests, we do not assume that a particular distribution is applicable, or that a certain value is attached to a parameter of the population. For instance, while testing the two training methods, say A and B, for determining the superiority of one over the other, if we do not assume that the scores of the trainees are normally distributed or that the mean score of all trainees taking methods A would be a certain value, then the testing methods is known as a distribution free or non parametric methods.

In fact, there is a growing use of such tests in situations when the normality assumption is open to doubt. As a result, many distribution free tests have been developed that do not depend on the shape of the distribution or deal with the parameters of the underlying population.

Non-Parametric tests are useful to examine hypothesis about data and which are non-nominal for which meaningful sample studies cannot be calculated. Since these tests do not depend on the shape of the distribution, they are also called distribution free tests.

Instead of parameters, these tests directly assume the observation and values. The frequencies - whether they are observed or expected and their differences or ranks are subjected to examination, in non-parametric tests.

Difference Between Parametric and Non-Parametric Testing

Both parametric and non-parametric tests are significance tests, because they examine significance of differences between given values. But they have to be distinguished, on following basis

Parametric tests	**Non-Parametric tests**
1. Based on assumption about distribution	1. No assumption about distribution
2. Based on statistics and parameters	2. No statistics and parameters
3. Focus on SE Concept and Level of Significance	3. Does not focus on SE Concept and Level of Significance
4. Mostly interval scale or ratio scaled data	4. Mostly nominal scale or ordinal scale data
5. Precise mathematical analysis	5. No precise mathematical analysis
6. Handle variables	6. Handle mostly attributes

Merits of non parametric tests

Non-parametric tests have following advantages:

1. Assumption free - Assumption free - They do not require assumption that a population is distributed in the shape of probability distribution
2. Simple – They are simple to understand, explain and solve. No mean, variance, SD to be used
3. Measurement – They require only basic level of measurement. Such as ordinal or nominal. Ranks and signs form basis of non-parametric tests.
4. Flexible – Non-Parametric test allow considerable deviation from normal distribution. It need not follow the rigid rules of normal curve and central limit theorem.
5. Natural option- It's the only choice when hypothesis is to be tested, but no parameter is available.
6. Realistic – It's more realistic, because they consider observations themselves, rather than derivation like mean or standard deviation.

Assumptions in NP tests

Non-parametric tests belong to significant tests, designed for special situations. They come to our rescue , subject to following assumptions ..

1. Sample observations are independent.
2. Samples drawn are randomly selected.
3. Observations are measured at least on ordinal scale.

Classification of Non-Parametric Test

Non-parametric tests deals with data which are nominal (sick or not sick, increase or decrease etc.) or ordinal (I rank, II rank etc.), Such tests includes chi squire test, sign test, signed rank test, rank sum test, runs test, rank correlation etc. Next 3 units are devoted to discuss chi squire test, sign and signed rank tests, and other non-parametric tests respectively.

Limitations of non parametric tests

Non-parametric tests are useful in many occasions when statistic and parameter is not available. However, it's less powerful than parametric tests in following respects.

1. Non-parametric test does not use all the information provided by sample, and results will be less informative than parametric tests.
2. As no assumption are formed results are likely to be misleading and non-normal.
3. There is greater risk of accepting a wrong hypothesis and thus commit Type II error
4. They are difficult to be applied when the number of samples and sample size is very large.

Steps in non parametric tests

Since non parametric tests are a set of certain distribution free, parameter free testing procedures, there are no unified procedure for conducting such tests. Following are certain model steps to take up and complete non parametric tests, in general.

1. Make necessary assumptions and framework relating to the conduct of the test.
2. Form null or alternative hypothesis
3. Consider the sample data given and other details.
4. Decide the type and manner of conducting non parametric tests
5. Decide the test value and procedure of the test.
6. Decide the level of significance and degree of freedom.
7. Obtain table value and compare with the test value.
8. Decide the fate of hypothesis.

Important non parametric tests

Numerous non parametric tests have been developed, but in this book, we are discussing following non-parametric tests

Chi square test

Prominent differences between sampling distributions have been previously studied through constants like mean, standard deviation, proportion etc, which are the estimates of the parameters, but generally these do not give all the features of these distributions. This caused the necessity to have some index which can measure the degrees of difference between the actual frequencies of various groups and can thus compare all necessary features of them. An index of this type if Karl Pearson's Chi Square which is used to measure the deviations of observed frequencies in an experiment from the expected frequencies obtained from some hypothetical universe. In the next chapter, we are going to study a distribution called chi square distribution which enables us to compare a whole set of sample values with a corresponding set of hypothetical values.

Chi square distribution was discovered by Helmert in 1875 and was again discovered independently in 1900 by Karl Pearson who applied it as a test of goodness of fit.

If O and E denote observed and corresponding expected frequencies of class interval or cell, then chi square is defined by the relation

$$\text{Chi square value} = \Sigma \frac{(O-E)^2}{E}$$

Chi square test enables calculation of chi square value and this calculated value is compared with chi square table value given in the Appendix.

Sign tests

Sign tests are prominent non parametric tests which are used to test hypothesis on a population median. In the case of many of non parametric procedures, the mean is replaced by the median as the pertinent location parameter under the test. Given a random variable x, the population median, corresponding to the sample median can be estimated and evaluated for significance testing.

The appropriate test statistic for the sign test is the binomial random variable X, representing the number of plus signs in our random sample. If the null hypothesis that random median equals population median, the probability that a sample value results in either a plus sign or a minus sign is equal to 0.5. therefore, to test the null hypothesis that the two medians are equal, we are actually testing the null hypothesis that number of plus signs is a value of random variable having the binomial distribution with parameter p = 0.5. p values for both one sided and two sided alternatives can then be calculated using this binomial distribution. Cumulative binomial probabilities are available in the Table given at the Appendix.

Signed Rank tests

The sign test utilizes only the plus and minus signs of the differences between the observations and population mean in one sample case or the plus and minus signs of differences between the pairs of observations in the paired sample case. But it does not take not account the magnitudes of these differences. A test utilizing both direction and magnitude, proposed in 1945, by Frank Wilcoxon, is now commonly referred to as the Wilcoxon Signed rank test.

Rank sum tests

When one is interested in testing equality of means of two continuous distributions that are obviously non normal and samples are independent, ie, there is no pairing of observations, the Wilcoxon rank Sum Test or Wilcoxon Two Sample Test, is an appropriate alternative to the two sample t test.

For rank sum test, first we select a random sample from each of the populations, and rank them from the lowest to the highest. In the case of ties, identical observations, we replace the observations by the mean of the ranks that the observations would have if they were distinguishable. These ranks are evaluated using Wilcoxon table given at the Appendix.

Kruskell- Wallis test

The technique of analysis of variance is prominent as an analytical technique for testing equality of several means. However, the applicability of ANOVA or the F test is based on the premise of normality. In this section, we investigate a anon parametric alternative to ANOVA , which is called Kruskal – Wallis test. The test is also called H test, is a generalization of the rank sum test to the case of more than 2 sample groups. We combine all the samples and arrange observations in ascending order, substituting appropriate rank from 1, 2…. The sum of the ranks corresponding to the first group of observation is denoted by the random variable R_1 and so on …..

Runs Test (Wald Wolfowitz Test)

Applying the many statistical concepts discussed through out this chapter, it was always assumed that our sample data had been collected by some randomization procedure. The runs test, based on the order in which the sample observations are obtained, is a useful technique for testing the null hypothesis that the observations have indeed been drawn at random.

Next unit deals with Chi Square test. Sign tests and Signed Rank tests, and Rank Sum test and other non parametric tests are discussed in following two chapters.

Review Questions and Exercises

1. hat are non-parametric tests?
2. State the characteristics of non parametric tests
3. What are the assumptions in non parametric tests
4. Why non parametric tests are needed
5. State the merits of non parametric tests
6. What are the rationale behind the non parametric tests
7. Distinguish between p tests and non parametric tests
8. What are the disadvantages of non parametric tests
9. What are the important non parametric tests
10. What is the objective of non parametric tests
11. What is the basis of sign test/
12. How is chi square test conducted.
13. What is the procedure in rank sum test?

UNIT X

CHI-SQUARE TESTS

From a series of observation, different statistics are constructed to estimate population parameters. In general, sampling distribution of the statistic depends on parameter and form of population. The difference between distributions has been studied through constants such as mean, proportion, etc. They may not truly represent a distribution. This caused the necessity to have some index which can measure the degree of difference between actual frequencies and expected frequencies directly, without any representative value. Thus emerged chi-square test, which is used to measure deviation of observed frequencies from expected frequencies.

Chi-square distribution was discovered in 1875 , by Karl Pearson in 1900. Chi-squire test is a significant test, which is not based on any parameter like mean, variance or proportion. Therefore such tests are distinguished as non-parametric test.

Types of chi-square tests

On the basis of situation, nature and purpose of test, chi-square test may be classified as – test of independence of attributes, test of goodness of fit , test of homogeneity, and test for variance

Test of independence of attributes

In significant testing, difference or dependence between two attributes can be studied . when we wish to test the difference of more than two proportions in terms of two attributes, chi-square test is applied. It is similar to Anova when variance between several sample groups are analyzed at a time.

For the chi square test, actual frequencies will be given in the question. Expected or theoretical frequencies have to be calculated and compared with each other to obtain measure of deviation.

Steps

1. form null hypothesis
2. consider observed frequencies or actual frequencies = O
3. ascertain expected frequencies using the formula

   Expected Cell frequency E= $\frac{A \times B}{AB}$

   where

   E = Expected frequency

   A = column total

   B = row total

   AB = grand total
4. obtain $\Sigma \frac{(O-E)^2}{E}$ for each cell frequency
5. Summate to get total chi-square value. $\Sigma \frac{(O-E)^2}{E}$ or X^2 value
6. Compare with chi-square table value at required level of significance and degree of freedom
7. Decide the fate of null hypothesis

Ex .10.1

Assembly election are announced in Kerala, Tamil nadu and Karnataka. A major political party wants to test, if the proportion of supporters in these three states are the same or difference. the party conducted a sample survey of 1000 people in each states and found that they are 300,350 and 425 supporters. test if the supporters are the same or different significantly at 5% α.

supporters	Kerala	Tamil Nadu	Karnataka	total
yes	300	350	425	1075
no	700	650	575	1925
total	1000	1000	1000	3000

Expected frequency = $\frac{A \times B}{AB} = \frac{1000 \times 1075}{3000} = 358$ .....

yes	358	358	359
no	642	642	641

(Total of expected frequencies must come equal to total of observed frequencies. for this purpose, rounding of figures must be done.)

H0 = No significant difference. Supporters are the same

o	e		
300	358	58	9.39
350	358	8	.18
425	358	67	12.54
700	642	58	5.24
650	642	8	.10
575	642	67^2	6.99
			34.44

X^2 value $= \Sigma \frac{(O-E)^2}{E}$ = = 34.44

X^2 table value @ α = 0.05 ,and degree of freedom = c-1 x r- 1 = 3-1 x 2-1 = 2 = 5.99. H0 is rejected because , calculated value is greater than table value. So supporters are the same.

Ex. 10.2

In a hostel, there are 200 students, in 4 classes – A: 70, B: 60, C: 30 & D: 40. Following details are given on their meals – heavy, medium & light. Are meals independent of class?

meals				
class	heavy	medium	light	total
a	24	32	14	70
b	22	26	12	60
c	10	14	6	30
d	14	16	10	40
total	70	88	42	**200**

Ans : Ho = No significant difference . Meals are independent of class

Expected frequencies

meals				
class	heavy	medium	light	total
a	24	31	15	70
b	21	26	13	60
c	11	13	6	30
d	14	18	8	40
total	70	88	42	**200**

O	E	$(O-E)^2$	$(O-E)^2/E$
24	24	0	0
32	31	1	.03
14	15	1	.07
22	21	1	.05
26	26	0	0
12	13	1	.07
			.22

Chi square table value at α = 0.05, υ = 3-1 x 4 -1 = 2 x 3 = 6 = 12.59

Calculated value - 0.22 is much less than table value. Hypothesis is accepted. Meals do not dependent on class.

b) Tests for goodness of fit

If we have a set of frequencies of a distribution obtained by an experiment and if we are interested in knowing whether these frequencies are consistent with those which may be obtained based on some theory, then we can use chi square test of goodness of fit.

For example, if frequency distribution like Binomial or Poisson or Normal is applicable, the expected frequencies would be derived using that distribution.

Ex .10.3

A sample analysis of an examination result of 200 students were made. It was found that 46 students had failed, 68 secured III class, 62 second class and the rest were placed in the first division. Are these figure commensurate with the general examination results which is in the ratio 2: 3 : 3; 2 for various categories respectively.

H0 : No significant difference. There is goodness of fit. The sample of 200 students' result is commensurate with general exam result.

O	O - E	$(O-E)^2$	$(O-E)^2/E$
46	40	36	0.9
68	60	64	1.07
62	60	4	0.06
24	40	256	6.4
		Total	8.43

X^2 value = $\Sigma \frac{(O-E)^2}{E}$ =8.43

Table value at α = 0.05, υ = 4 -1 = 3 = 7.815

The calculated value is higher than table value. Hence the hypothesis is rejected. The sample result is not commensurate with general examination result.

c) Test of homogeneity

Here we have more than one sample unlike the test of independence where there is only one sample. We want to test whether these samples are homogeneous as far as a particular attribute is concerned .When there is homogeneity we conclude that the samples belong to the same population or identical population. The null hypothesis in these cases is that there is homogeneity.

The test is performed in the same manner as a test of independence. When the null hypothesis is accepted, we conclude that there is homogeneity.

Ex 10.4

From the adult population of four large cities, random samples were selected and the number of married and unmarried men were recorded.

Cities					
	A	B	C	D	Total
Married	137	164	152	147	600
Single	32	57	56	35	180
Total	169	221	208	182	780

Is there significant variation among the cities in the tendency of men to marry?

Ans: Ho : No significant difference. Cities show same tendency of men to marry. There is homogeneity of variances.

Table showing expected frequencies

	A	B	C	D	TOTAL
MARRIED	$\frac{169 \times 600}{780} = 130$	$\frac{221 \times 600}{780} = 170$	$\frac{208 \times 600}{780}$ =160	140	600
SINGLE	39	51	48	42	180
TOTAL	169	221	208	182	780

.

O	E	$(O-E)^2$	$(O-E)^2/E$
137	130	49	.4
32	39	49	1.3
164	170	36	.2
57	51	36	.7
152	260	64	.4
56	68	64	1.3
147	140	49	.4
35	42	49	1.2
			5.9

$x^2 = \sum \frac{(O-E)^2}{E} = 5.9$

Degree of freedom = (r-1) (c-1) = (2-1) (4-1) = 3

Table value = $\alpha = 0.05, \upsilon = 3 = 7.815$. The calculated value 5.9 is less than table value. Therefore the hypothesis is accepted. Cities show same tendency to marry.

(4) TEST FOR POPULATION VARIANCE .

X^2 test can be used for testing the given population variance, (to test whether there is any significant difference between sample variance and population variance).

The test statistic is obtained by the formula t= $\frac{ns^2}{\sigma^2}$, where n is the sample size s^2 is the sample variance and σ^2 is the populatin variance. Degree of freedom = n-1

Ex .10.5

The standard deviation of a sample of 10 observations from a normal population was found to be 5. Examine whether this is consistent with the hypothesis that the standard deviation of the population is 5.3?

Ans ; H0 = No significant difference . Standard deviation of the population is 5.3.

Chi square value = $\chi^2 = \frac{ns^2}{\sigma^2}$ $\frac{10\ x\ 5^2}{5.3^2}$ = 8.9

Degree of Freedom = n-1 = 10 – 1 = 9

Table value of χ^2 for 9 degrees of freedom at 0.05 level of significance = 16.9

Calculated χ^2 value 8.9 is then the table value. Therefore we accept the null hypothesis.

Population standard deviation is 5.3 .

Ex 10.6

A random sample of size 20 from a normal population gives a standard deviation of 6. Test the hypothesis that the population standard deviation is 9.

Ans ; H : there is no significant difference . Population standard deviation is 9.

$\chi^2 = \frac{ns^2}{\sigma^2}$ $\frac{20\ x\ 6^2}{9^2}$ = 8.88

Table value of χ^2 at 0.05 level of significance for 19 degrees of freedom is 30.14. The calculated value 8.88 is less than the table value. Hence we accept the null hypothesis. Population standard deviation can be taken 9.

Ex 10.7 Following data relates to inoculation against fever. Is inoculation effective?

	Attacked	Not attacked
Inoculated	31	469
Not inoculated	185	1315

Ho: the two attributes – inoculation and attack are independent. Inoculation is not effective.

O	E	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{E}$
31	54	529	9.80
469	446	529	1.19
185	162	36	3.27
1315	1338	529	0.40

Expected frequencies are = $\frac{216\times500}{2000}$ = 54 ……446, 162, 1338

Total 14.66

χ^2 Table value at 1 degree of freedom, and 5% level of significance = 3.84. The calculated value is much greater than table value. Hence the hypothesis is rejected. The inoculation is effective.

Ex 10. 8

Given following data relating to social status and state of intelligence. Test whether intelligence is related so social status.

	dull	intelligent	Brilliant	Total
Lower	22	35	23	80
Middle	38	70	32	140
Upper	60	20	20	100
Total	120	125	75	320

O	E	$(O-E)^2$	$(O-E)^2/E$
22	30	64	2.13
38	52.5	210.25	4.00
60	37.5	506.25	13.50
35	31.25	14.06	0.45
70	54.69	234.4	4.29
20	39.06	363.26	9.30
23	18.75	18.06	0.96
32	32.81	0.66	0.02
20	23.44	11.83	0.50
		Total	35.15

Expected frequencies ; $\frac{120\times80}{320}$ = 30 ……52.5 37.5, 31.25 , 54.69, 39.06, 18.75, 32.81, and 23.44

χ^2 Value = 35.15.

Calculated χ^2 value 35.15, whereas χ^2 table value 9.488. The difference is significant. The hypothesis is rejected . Intelligence is related to

Precautions for chi square test

The chi square test is no doubt a most frequently used test, but its correct application is an uphill task. It must be performed with special care and diligence.

1. It should be borne in mind that the test is to be applied only when the individual observations of sample are independent which means that the occurrence of one individual observation has no effect upon the occurrence of any other observation.
2. Small theoretical frequencies, if these occur in certain groups, should be dealt with ial care, and must be added along with appropriate frequency.
3. sum of observed frequencies and expected frequencies must be equalized by rounding off.
4. Level of significance and degrees of freedom must be properly ascertained.

Conditions in applying chi square test

1. The total frequencies must be reasonably large say, at least 50
2. Expected frequency of less than 5 is pooled with the preceding or succeeding frequency so that no expected frequency is less than 5.
3. Accordingly the degrees of freedom must be modified.
4. The distribution should not be of proportions or percentages. It should be of original units.

Contingency tables

A contingency table is a frequency table in which a a sample from th population is classified according to two attributes, which are divided into two or more classes. When there are only two divisions for each attribute the contingency table is known as 2x2 contingency table. The frequencies appearing in the table are known as cell frequencies. The independence of these two attributes can be tested by chi square test. Contingency table can have m rows and n columns.

Limitations of Chi square test

1. ***It is not as reliable as a parametric test. Hence it should be used only when parametric tests cannot be used.***
2. ***Its use is restricted by certain conditions like total of the frequencies should not be less than 50.***
3. ***It is not possible to complete the chi square values when the given values are in proportions or percentages.***
4. ***It does not distinguish between favorable deviation and unfavorable deviation. Therefore, the results may be misleading.***

Review Questions and Exercises

1. What is chi-square tests
2. State the assumptions of chi-square tests
3. What is chi-square value
4. Explain characteristics of chi-square tests
5. What are the uses of chi-square tests
6. State the types of chi-square tests
7. State the steps in chi-square tests
8. What is test of goodness of fit
9. What are the conditions of chi-square tests
10. State precautions while conducting chi-square test
11. What are expected frequencies
12. What is a contingency table
13. In an experiment on pea breeding Mental obtained the following frequencies of seeds : 315 round and yellow, 101 wrinkled and yellow, 108 round and green, 36 wrinkled and green. Theory predicts that the frequencies should be in the proportion 9:3:3:1. Examine the correspondence between theory and experiment.
14. A chemical extraction plant processes sea water to collect sodium chloride, magnesium and other elements in the ratio 62:4:34. A sample of 200 tons of sea water has resulted in 130 tones of sodium chloride and 6 tones of magnesium and 64 tons of other elements. Are these data consistent with the known composition of sea water at 5%level?
15. The number of road accidents per week in a certain area were as follows.

    12 8 20 2 14 10 15 6 9 4

    Are these frequencies in agreement with the belief that the accident occurred were uniform during the 10 week period.(hint: take all expected values as 10)
16. A die is thrown 180 times with the following results.

Number turned	1	2	3	4	5	6
Frequency	25	35	40	22	32	06

    Test whether the die is unbiased

    (hint: expected frequencies are 30 each)
17. The following figures show the distribution of digits in numbers chosen at random from a telephone directory.

digits	0	1	2	3	4	5	6	7	8	9
freq	1026	1107	997	996	1075	933	1107	972	964	853

    test at 5% level whether the digits may be taken to occur equally frequently in the directory.

    (hint: e values may be taken as 1000 each)

18. In an experiment with the immunization of cattle from tuberculosis, following results were obtained.

	Affected	Unaffected
Inoculated	12	26
Not-inoculated	16	6

Explain the effects of the vaccine in controlling susceptibility to tuberculosis.

19. In a course of anti malarial work in a certain city over a period of time quinine was administered to 606 adults out of a total population of 3540. The data regarding incidence of malarial fever is given below. Examine whether the quinine has the effect of preventing fever.

	Fever	no fever	total
Quinine	19	587	606
No quinine	193	1741	2934
Total	212	3328	3540

8. In an anti- malaria campaign, in a certain area, quinine was administered to 812 persons out of a total population of 3248. The number of fever cases is given below:

Treatment	fever	no fever
Quinine	20	792
No quinine	220	2216

Discuss the usefulness of quinine in checking malaria.(table value of chi-square at 5% probability level for one degree of freedom is 3.84)

9. From the following table test whether the color of the son's eye is associated with that of the fathers.

	son's eye color	
Father's eye color	light	not light
Light	471	151
Not light	414	230

10. Two sample polls of votes for two candidates - A and B for a public office are taken, one from among residents of rural areas and other from among residents from urban areas. The results are given below. Examine whether the nature of the area is related to voting preference in this election.

	candidates		
Area	a	b	total
Rural	620	480	1100
Urban	380	520	900
Total	1000	1000	2000

11. The results of a survey to know the educational attainment among 164 persons randomly selected in a locality are given below:

	Education			
	Middle	high school	college	total
Male	52	10	20	82
Female	44	12	26	82
Total	96	22	46	164

Can you say that education depends on sex?

UNIT 11

NON-PARAMETRIC TESTS – SIGN AND SIGNED RANK TESTS

When the nature of population distribution from which samples are drawn is not known to be normal, hypothesis test are called non-parametric. Prominent non-parametric tests include chi-square test, sign test, rank test, runs test, rank correlation test. Among these, chi-square test has been discussed in earlier unit. This unit deals with sign tests, and signed rank tests. Rank sum tests, Runs test, and other non parametric tests are dealt with in the coming unit.

SIGN TEST

Sign test is designed to examine hypothesis on a population median. In the case of many non-parametric procedures, the mean is replaced by the median as the location of parametric test.

Sign test is the non-parametric version of t test, where we examine -

1 if the population mean is the same as samples mean or not,

2 If two population means are the same

Sign test focuses on the + or – sign of each observation. In sign test, we do not give importance to the magnitude of observation. Usually the central point of a data set is arithmetic mean when and we give weight age to the magnitude of observations. It is not the case with sign test.

Since the location wise central point is median sign test consider median, in place of mean. Generally a value known as median value will be given in the question, and sign test examines how many + signs are there which come above the median value, and how many – signs, representing values less than median value.

Sign test may be one sample sign test or two sample sign test. Again such tests may be conducted in the context of small samples and large samples.

one sample sign test

All tests concerning means that you have studied are based on the assumptions that samples are taken from a population having the shape of a normal distribution. When this assumption is not possible, or any statistic is not available, vital questions still arise:

`. is there a significant difference between actual observations and theoretically

Expected observations

2. Is it reasonable to believe that samples have been taken from a probabilistic sampling distribution

3 is it reasonable to accept that the sample as a random sample from a population etc.

One sample sign test is applicable, when sample is taken from a population which is continuous. in this case, the probability that the sample value is less than mean, and the probability that sample value is greater than mean are both i.e, ½ and ½ . Here the sign test is used to test hypothesis as a population median. The median divide the distribution into two equal halves. Now we may examine whether two halves are equal or not. if these halves are equal or approximately equal, the distribution is and the same prediction is possible. This is the rationale behind one sample sign test.

In this test, each sample value greater than given median value is replaced by plus sign and each sample value less than the given value is replaced by negative sign. Then we test the hypothesis that these plus signs and minus signs are equal or approximately equal or there is no significant difference between them.

One sample Sign test for small sample

For small samples, sign test may be conducted, following Binomial probability terms. For this, a median value is to be identified, and compared with other values, and + or – signs separated and compared. Binomial probability value must be ascertained from the Binomial Table, at the required n and probability. This probability must be compared with the specified level of significance value, to decide whether to accept or reject null hypothesis.

Steps

1. Find + and – signs, by comparing with sample median value.
2. Take maximum of signs.
3. Find required Binomial probability value, at specified n and p, from the table.
4. Compare with level of significance value.
5. If the ascertained binomial value is less than level of significant value, Ho is accepted and vice versa.

Ex 11.1

A bank gives details of customer per day - 280, 282, 292,273, 283, 283, 275, 284, 282, 279, and 281. At 5% level of significance, examine whether average customers is 284.

Ans ;

Population median is equal to 284. So average customers is 284.

By comparing each observation with the median value 284

-, -, +, -, -, -, -, 0, -, -, -

Therefore No – signs = 9

Number of + signs = 1

Number of no signs = 1

Reduced Sample size = 10 n = 10, expected population Probability = .5

Required probability is 9 and above. = 1 - probability of 8 and less = 1 - .9892 = 0.0108

Ascertained binomial probability, from binomial Table is less than required level of significance value 0.05.therefore the hypothesis should be accepted, and average customers can be taken as 284.

Ex 11.2

In a post graduate class there are 11 students, and grade points secured points secured by them are following: 1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, and 1.7. Use sign test to examine the hypothesis that intelligence is a random function with median = 1.8 and α = 0.05.

Ans; H0 ; no significant difference between -+ signs and _ signs. intelligence is a random function with median = 1.8

Signs : - + - - + - 0 - + -

Number of + signs = 3

Number of – signs ; 7

Number of 0 signs = 1

Sample size = n = 10, p = .5 , r = p of 7 or more = 1 – 6 or less)

Binomial value for 6 or less = .8281. therefore required probability is 1 - .8281 = .1719. This is more than the specified α value .05.therefore, the hypothesis is rejected. There is significant difference between numbers of signs. On this basis, Intelligence is not a random function with median value 1.8.

Sign tests for two samples

This test is an analogous of paired t test. We use this test when we have two samples having paired observations like data the usual null hypothesis for this test is that there is no difference between the + signs and – signs.

Sign tests examine whether + signs and – signs differ significantly from the centre of location or the median value. For conducting this test, find the differences between the observations of the given two sample groups and ascertain the most occurring number of signs. The binomial probability of this value may be compared with expected level of significance value like 0.05, 0.01, 0.10 etc

Steps

1. Form null hypothesis that + signs and - signs are equal
2. Deduct the second score from the first score and get + signs and – signs
3. Discard the pairs having both observations as the same and reduce sample size..
4. Count numbers of + signs and – signs and take maximum signs.
5. Find P value referring the Binomial table.
6. Compare with specified level of significance value and decide fate of Ho.

Ex.11.3

Following are numbers of aircrafts dug up by two archeologists – A and B at an ancient cliff. Use sign test at 10% level of significance to test that two archeologists are equally good.

A	1	0	2	3	1	0	2	2	3	0	1	1	4
B	0	0	1	0	2	0	0	1	1	2	0	1	2

Ho = No significant difference. The two archeologists are equally good.

A	B	A - B
1	0	+
0	0	0
2	1	+
3	0	+
1	2	-
0	0	0
2	0	+
2	1	+
3	1	+
0	2	-
1	0	+
1	1	0
4	2	+

Number of + signs 8

Number of – signs 2

Number of no signs 3

Reduced sample size = n = 10

P value = probability of 8 or more = 1 - probability of 7 or less (B 10, .5)

= 1 - .9453 = .0547

The obtained p value is compared with expected level of significance value – 0.10, and it is seen that the difference is insignificant. Therefore the hypothesis is accepted that there is no significant difference. And it is considered that the two archeologists are equally good.

One sample sign test for large sample

One sample sign test may be conducted by computing and comparing with z value, for large sample. When the sample is large, the test value may be assumed to follow normal probability distribution, and therefore to compare the test value, z value can be considered. z value may be obtained as per the formula:

$$Z = \frac{S-np}{\sqrt{npq}}$$

where, s = no. of maximum sign (+ or -)

n = no. of total sign

p = proportion of population mostly 0.05)

q = 1 - p

steps

1. Form null hypothesis – no significant difference
2. Consider the observations given and the median value
3. ascertain values greater than median value and replace them with + sign
4. ascertain values less than median value and replace them with - sign
5. if any value happens to be equal to given median value, no sign is assigned and accordingly sample size is reduced
6. obtain number of maximum signs where = s
7. ascertain probability of population, as per null hypothesis
8. find z value as per the formula

   $$z = \frac{S-np}{\sqrt{npq}}$$

   where s = number of maximum sign (+ or -)

   p = proportion of population

   q = 1 – p.

   n = sample size
9. compare with z table value, and decide the fate of Ho

EX 11 .3

The increase in pulse rate of 24 patients measured before and after a drug- Atenolol is given below. Examine whether drug influence pulse rate by conducting a sample sign test. 18 is the normal increase rate.

18 24 20 26 23 17 24 21 22 20 16 27 25 14 20 15 18 22 21 24 26 27 28

You may use 1% level of significance.

A. h_o = no significant difference. drug does not influence pulse rate

signs : 0 + + + + - + + + + - + + - + - 0 + + + + + +

- signs = 5

+ signs = 17

No sign = 2

s = maximum signs = 17 (+)

p = proportion of population = 0.5

n = total net observations = 22

q = 1 – p = 0.5

$z = \frac{S-np}{\sqrt{npq}} = \frac{17-(22\times0.5)}{\sqrt{22\times0.5\times0.5}} \approx 2.56$

Z value @0.01 level of significance = 2.58. The calculated value 2.56 is less than 2.58. Hence, hypothesis is accepted. Drug do not influence pulse rate

Ex 11.4

A survey was conducted to study preference for fast food. A sample of 100 persons indicated that 54 do not prefer fast food, and 46 preferred fast food. By using sign test, examine the hypothesis that half of people prefer fast food.

A. Ho = no significant difference. half of persons prefer fast food

Denote those prefer by + sign, and those don't prefer by – sign

Then s = 54 (maximum), p = 0.5, q = 0.5 n = 100

$z = \frac{S-np}{\sqrt{npq}} = \frac{54-(100\times0.5)}{\sqrt{100\times0.5\times0.5}} = 0.8$

Z table value at α = 1.96, and calculated sign test value is 0.8. Hence the difference is considered insignificant. Thus the hypothesis is accepted and half of persons prefer fast-food.

Sign test as proportion test

Two sample test can be performed as proportion test also. The difference between given two sample groups form the sample proportion. Population proportion can be estimated as 0.5. Standard error can be used to compute test statistic. Test statistic is the ratio between the difference between population proportion, sample proportion , and standard error. The test statistic will be compared with z table value to reveal the significance of difference.

steps

1. Form h_o
2. Consider given sample values and sample size
3. Find the difference between pairs, and put appropriate signs
4. Take Maximum + or - signs as sample proportion
5. Reduce sample size considering no signs.
6. Find standard error = $\frac{\sqrt{PQ}}{n}$
7. Obtain test statistic, z = $\frac{DIFFERENCE}{STANDARD\ ERROR} = \frac{p-P}{\frac{\sqrt{PQ}}{n}}$
8. Compare with z table value

Ex 11.5

To examine effectiveness of traffic signal system, number of accidents before and after its installation at a junction is given below. Use the sign test at $\propto$ =0.01, to examine the effectiveness

9 and 5, 7 and 3, 3 and 4, 16 and 11, 12 and 7, 12 and 5, 5 and 5, 6 and 1

Ans

there are 6 + signs , 1 – sign, and 1 no sign

P = 0.5, p = 6/7 = 0.86

Standard error = $\frac{\sqrt{PQ}}{n} = \frac{\sqrt{0.5 \times 0.5}}{7} = 0.2$

Z value = $\frac{0.86-0.5}{0.2} = 1.8$

Z table value at $\propto = 0.01 = 2.58$

The calculated z value is less than z table value. Ho is accepted. The signal system is not effective.

Merits of sign tests

1. it is easy to calculate , understand and present
2. it considers the observations directly, without representation
3. it is the only solution, where any parameter, or statistic, or a distribution are not available
4. it does not depend on assumptions

Demerits

1. it may lead to misleading conclusion
2. it does not consider magnitude of observation
3. it depends only on the direction of data through signs
4. it recognizes some assumption relating to distribution

Wilcoxon signed rank test

As is evident, the sign test utilizes only the plus and minus signs of the differences between observations and the median value in one sample case, or the plus and minus signs of differences between pairs of observations in two sample cases. It does not take into account the magnitude of these differences. A test utilizing both direction and magnitude, proposed in 1945 by Frank Wilcoxon is now commonly referred to as Wilcoxon signed rank difference test

Normal sign test does not give any importance to magnitude of the observation. So, the results of sign test are much weaker. Ranking of the values is a solution for this drawback. . Wilcoxon signed rank tests are used for ordered categorical data when a numerical scale is inappropriate. Then, it is possible to rank observations.

In signed rank test, both direction and magnitude of observations are given due importance, along with the difference between observations. Signed rank tests are classified as Wilcoxon signed rank difference test (single sample) and Wilcoxon signed rank difference sum test or matched pairs test.

Wilcoxon signed rank sum test for single sample

Here, observations for single samples will be given and will be required to test whether a sample median comes more or less equal to population median. The test focuses on the difference of ranks and their sum. It considers magnitude of data in a primary form – ordinal.

Steps

1. Form null hypothesis - no significant difference between population median and sample median
2. Obtain difference of each observation from given median value
3. Assign ranks to absolute differences, ignoring signs, from smallest difference to largest difference
4. If there is equality in ranks, assign average value to the ranks

5. calculate sum of positive ranks (W +) and sum of negative ranks (W -), and take the least of them as W value
6. Check whether total of W + and W - = n(n + 1)/2
7. Refer
8. If the calculated W value is less than w table value, accept null hypothesis and vice versa.

EX. 11. 6

Following is the price of a share on 11 days. Test at 5% level of significance, that average price of the share is 284 or not.

280 282 292 273 283 283 275 284 282 279 281

Ho = No significant difference, the average price is 284

X	d (X - 284)	RANK	- SIGN	+ SIGN
280	-4	6	-6	
282	-2	3.5	-3.5	
292	+8	8		8
273	-11	10	-10	
283	-1	1.5	-1.5	
283	-1	1.5	-1.5	
275	-9	9	-9	
284	0	_		
282	-2	3.5	-3.5	
279	-5	7	-7	
281	-3	5	-5	
TOTAL			**47.0**	**8**

W value - the least = 8

(Self check) n(n+1)/2 = 10×11/2 = 55 =47+8

At 5% level of significance, W table value for reduced n = 10 = 8

Since calculated W value is less than or equal to W table value (also called T Table Value), hypothesis is accepted and the average price is taken as 284.

Wilcoxon signed rank test – Two samples

In the case of two related samples, we can determine both direction and magnitude of difference between matched values to find the significance of difference. We can use the non parametric test known as Wilcoxon signed Difference Rank Test. If this technique is employed, we first find the differences between each pair of values and assign ranks to the differences from the smallest to the largest without regard to the sign. The actual signs of each difference are then put to corresponding ranks and the test statistic T is calculated. T is the smaller of the two, namely, the total of the negative ranks and the total of the positive ranks. When the calculated T value is less than or equal to Wilcoxon T Table given at the end of the book, the null hypothesis is accepted and vice versa. This test is also called Wilcoxon Matched Pairs Test

Steps

1. Form null hypothesis
2. Considering the data as pairs, ascertain differences between pairs .
3. Rank the differences, from the smallest to the largest.
4. Divide the ranks into + signs and – signs and total them
5. Take the least sum as Wilcoxon T value
6. Compare with Wilcoxon T Table value, at the specified level of significance
7. Decide the fate of Null hypothesis.

Ex 11.7

Given below is 16 pairs of values showing the performance of two machines. Test whether there is difference between the performances. (Use Wilcoxon matched pairs test at 5% level of significance)

H0 = there is no difference

Machine A	Machine B	D (difference)	Rank of D	Rank with signs
73	51	22	13	+ 13 …
43	41	2	2.5	+ 2.5 …
47	43	4	4.5	+ 4.5 …
53	41	12	11	+ 11 …
58	47	11	10	+ 10 …
47	32	15	12	+ 12 …
52	24	28	15	+ 15 …
58	58	0	-	- -
38	43	-5	6	… - 6
61	53	8	8	+ 8 …
56	52	4	4.5	4.5 …
56	57	-1	1	… - 1
34	44	-10	9	… - 9
55	57	-2	2.5	… - 2.5
65	40	25	14	+ 14 …
75	68	7	7	+ 7 …
			total	101.5 -18.5

Calculated value of T = 18.5(smaller of 101.5 and 18.5)

Note : since d=0 for the 8th pair, we dropped the value. ∴ n = 16-1=15

- Two of the 'd' values namely 2 and 4 appear twice. So they are given average of ranks i.e 2 has repeated twice, and the average is (2+3)÷2 = 2.5 - given to 2nd rank and 3rd rank. Similarly for the two 4's ranks are 4 and 5. So the average (4+5) ÷2=4.5 - given to 4th value and 5th value respectively.

Wilcoxon Table value of T =25

Calculated value of T = 18.5

Since calculated Wilcoxon T table value is less than calculated T value, we accept the null hypothesis and conclude that there is no difference between the performance of the two machines. Both machines are equally efficient.

Wilcoxon Matched Pairs Test for larger samples

In Wilcoxon matched pairs test, data are dependent on each other, and are given as paired observations. When the sample is larger, it is better to approximate normal distribution and conduct the test as Z test. The obtained Z value may be compared with Z table value at specified level of significance.

In the above case, population mean and standard deviation can be estimated as

Population mean = $\mu = \frac{n(n+1)}{4}$

$Standard\ Deviation = \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

THE Z VALUE = $\frac{W-\mu}{\sigma}$ where W = least of signed rank total

Steps

1. Form null hypothesis – no difference between samples.
2. Find differences between values and rank them from small to large
3. Divide ranks into two groups ie, with + signs, and with - signs.
4. If there are ties, give average ranks
5. Total the ranks separately and take the least sum as W or T value.
6. Estimate population mean and standard deviation
7. Find Z value $= = \frac{W-\mu}{\sigma}$
8. Where w = least of rank total $\mu = = \frac{n(n+1)}{4}$, $\sigma = \sqrt{\frac{n(n+1(2n+1))}{24}}$
9. Compare with Z table value at specified level of significance, and decide the fate of Ho

Ex 11.8

A training program is conducted for 16 teachers, and 2 tests are conducted – one before and one after training and the results are given below:

sl. no	score before	score after
1	85	82
2	76	79
3	64	68
4	59	52
5	72	75
6	68	69
7	43	40
8	54	53
9	57	50
10	61	67
11	71	74
12	82	83
13	39	54
14	51	59
15	54	51
16	57	58

Is the training effective? Perform Wilcoxon Matched pairs signed difference rank test at 5% level of significance.

Ans : H0 : there is no significant difference. Training is not effective.

Sl No	SCORE BEFORE	SCORE AFTER	d	Rank without sign	- rank	+ rank
1	85	82	+ 3	7.5		7.5
2	76	79	- 3	7.5	7.5	
3	64	68	-4	11	11	
4	59	52	+ 7	13.5		13.5
5	72	75	- 3	7.5	7.5	
6	68	69	- 1	2.5	2.5	
7	43	40	+ 3	7.5		7.5
8	54	53	+ 1	2.5		2.5

9	57	50	+ 7	13.5		13.5
10	61	67	-6	12	12	
11	71	74	- 3	7.5	7.5	
12	82	83	- 1	2.5	2.5	
13	39	54	-15	16	16	
14	51	59	- 8	15	15	
15	54	51	+ 3	7.5		7.5
16	57	58	- 1	2.5	2.5	
TOTAL					- 52	+ 84

W = LEAST OF -52 AND +84 = -52

μ

$= \frac{n(n+1)}{4} = \frac{16 \times 17)}{4} = 68$, $\sigma = \sqrt{\frac{n(n+1(2n+1))}{24}} = \sqrt{\frac{16 \times 17 \times 33)}{24}} = 19.34$

$Z = \frac{W-\mu}{\sigma} = \frac{52-68}{19.34} = 0.83$

TABLE VALUE OF Z AT α =0.05 = 1.96

Since calculated Z value is less than Z table value, the difference is considered insignificant and the hypothesis is accepted. Training is not effective.

Review Questions and Exercises

1. What is the basis of sign test
2. What are the different types of sign test
3. What is one sample sign test
4. What is two sample sign test
5. Explain importance of median in sign tests.
6. What are the assumptions in non parametric one sample sign test
7. What is the rationale behind Wilcoxon Matched pairs sign test
8. State the steps in One sample sign test
9. Explain the procedure in Wilcoxon two sample test.
10. What are the merits of one sample sign tests
11. State the use of Wilcoxon Matched Pairs test.
12. On 15 occasions Mr.X had to wait 9,5,3,8,8,6,9,7,2,10,7,7,6,10,6 minutes for a particular bus. Use the sign test at 5% level of significance to test the claim that on the average, Mr.X has to wait 8 minutes daily .
13. Following are the measurements of the breaking strength of a certain kind of 20 cotton ribbons

    163 165 160 189 161 171 158 151 169 162 163

    139 172 165 148 166 172 163 187 173.

    Use the sign test to test the mull hypothesis that median breaking strength is = 160

14. The pulse rate of 12 patients are measured before and after administering a drug. Test the null hypothesis that drug has no effect on pulse rate.

Patient	1	2	3	4	5	6	7	8	9	10	11
Before	72	70	68	67	73	71	72	70	69	70	68
After	74	72	69	68	72	72	72	71	67	73	69

15. To determine the effectiveness of a new traffic control system, the number of accidents that occurred at 12 dangerous intersections during 4 weeks before and after the installation of the new system was observed and the following data were obtained.

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

Use the paired-sample sign test at 0.05 level of significance to test the null hypothesis that the new traffic control system is effective only as the old system.

16. Given the score of two groups of persons, the one under placebo and other under drug are as follows:

score under placebo :	10	13	12	15	16	8	6	13	16
score under drug	20	14	7	9	17	18	19	25	24

Test whether the distribution of scores under placebo and under rug are identical.

17. Following are 15 measurements of the octane rating of a certain kind of gasoline 97.5, 95.2, 97.3, 96.0, 96.8, 100.3, 97.4, 95.3, 93.2, 99.1, 96.1, 97.6, 98.2, 98.5, and 94.9. Use the signed rank test at 0.05 level of significance to test whether the mean octane rating of the given kind of gasoline is 98.5

18. Following are the number of employers absent from two government agencies on 25 days:

24-29	32-45	36-36	33-39	41-48	45-36	33-41	38-39	46-40	32-39
37-30	34-45	41-42	32-40	30-33	46-42	38-50	34-37	45-39	32-37
44-32	25-33	45-48		35-33	30-35				

Use the Wilcoxon Two sample sign test at 0.05 level of significance to test the hypothesis that absentees are uniform, all days.

UNIT XII

RANK SUM & OTHER NON PARAMETRIC TESTS

After chi square tests, sign tests, and ranked sign tests, next important non parametric test is Rank Sum test. The test is similar to t test for paired observations. Popular rank sum tests are Mann Whitney Wilcoxon U test, Khruskal Wallis H test etc. Besides these, the unit describes other non parametric tests like Wald Wolwofitz test etc.

Rank Sum Tests

Wilcoxen signed rank test examined the difference between samples focusing on the rank differences between pairs of observations and also plus or minus signs. Differences were converted into ranks, and were totaled in terms of +ve values and –ve values. The minimum of rank totals were taken as W value and compared with Wilcoxen table value to decide fate of Ho.

However, in the rank sum tests, values in the two samples are first mixed and then ranked as if in a single sample. Then ranks are separated under the respective sample groups and summated for the two samples separately. In rank sum tests, we are replacing values by ranks. All values are taken together and they are assigned ranks. Rank sum tests are applied to test whether populations are identical or two samples come from same population.

Two important rank sum tests are Mann Whitney Wilcoxon U test and Kruskal-wallis test called H test. Besides these two, this unit also describes Wald Wolfowitz Runs test, which examines the randomness of selected sample.

Mann – Whitney- Wilcoxon U Test

This test examines whether two samples groups come from same population, or if there is any significant variation between two populations. To perform this test we first of all rank the data jointly, considering them as belonging to a single sample, in either increasing or decreasing order. We start from low to high. If tie occurs, ranks will be distributed between them applying average principle.

Then ranks will be separated as belonging to first sample and second sample. Then we find the sum of ranks, assigned to 1st sample and call it R1. Sum of ranks in the second sample is ascertained to get R2. Then we work out a measure of the difference between ranked observations of the two samples – U value.

In applying U test, we take the hypothesis that two samples come from identical population. Values of U are found to approximate normal distribution. Therefore the U test is conducted as normal distribution test as below:

This test is analogous to t test for two independent samples. Here we test whether two sample means are identical or they come from same populations. This test is conducted on the basis of two independent samples drawn from continuous populations.

To perform this test, we usually adopt low to high ranking process which means we assign rank 1to an item with lowest value, rank 2 to the next higher item and so on. In case there are ties, then we would assign each of the tied observation, the mean of the ranks which they jointly occupy. For example, if sixth, seventh and eights values are identical, we would assign 7th rank to the three values. After this we compute U value as below:

$$U = n_{1\,x}\, n_2 + \frac{n1\,(n1+1)}{2} - R_1$$

Where n1, n2 are sample sizes.

R1 is the sum of ranks assigned to the values of the first sample.

(in practice, any rank sum which can be conveniently obtained , may be taken as R1.

In applying U test we take the null hypothesis that the two samples come from identical population. If this hypothesis is true, it seems reasonable to suppose that the means of the ranks assigned to the values of two samples should be more or less the same. Under the alternative hypothesis, the means of two populations are not equal and if this is so, then most of the smaller ranks will go to the values of one sample.

If n1 and n2 are sufficiently large, the sampling distribution of U can be approximated closely with normal distribution and limits of the acceptance region can be determined, in the usual way, at a given level of significance. But if either n1 or n2 is so small, then, the normal curve approximation to the sampling distribution of U cannot be used. In such a case, we use the exact distribution based on the W values given in the form of Wilcoxon Table. First we perform the U test when the samples are small.

Steps

1. Form null hypothesis
2. Mix the items in the two samples and assign ranks to each item.
3. Divide the ranks into two sample groups
4. Summate the rank values in the first sample and second sample separately and get R1 and R2.
5. Take the smaller sum as Ws and number of items in this sample as s.
6. Consider number of items in the other sample as l
7. Find the minimum sum of ranks from 1 through up to s which is Minimum Ws
8. Ascertain Wilcoxon value = Ws – Minimum Ws
9. Read from Wilcoxon unpaired Distribution Table value at specified s and l.
10. Compare with required level of significance value (mostly 0.05) and decide fate of H0.

Ex 12.1

Two samples with values 90, 94 36, and 44 in one case and the other with values 53, 39, 6, 24, and 33 are given. Test applying Wilcoxon test whether the two samples come from populations with the same mean at 5% level of significance that these samples come from population with different medians.
Ho : samples come from populations with same medians.

items	rank	Sample A or B
6	**1**	**B**
24	**2**	**B**
33	**3**	**B**
36	**4**	**A**
39	**5**	**B**
44	**6**	**A**
53	**7**	**B**
90	**8**	**A**
94	**9**	**A**

Sum of ranks in Sample A = 4 + 6 + 8 + 9 = 27
Number of items in Sample A = 4
Sum of ranks in Sample B = 1 = 2 + 3 + 5 + 7 = 18
Number of items in Sample B = 5
Ws = smaller sum = 18 s = 5 l = 4
Minimum Ws = Total of ranks 1 through 5 = 1 + 2 + 3 + 4 + 5 = 15

Wilcoxon value = Ws - minimum Ws = 18 - 15 = 3

Probability for Wilcoxon value 3 , s = 5, and l = 4, = 0.056

This probability 0.056 may be compared with specified level of significance value 0.05. As per Wilcoxon test, if the calculated value is greater than Wilcoxon table value, the null hypothesis should be accepted. Thus we may conclude that the two samples come from the same population.

Steps (if the samples are large)

1. Form null hypothesis
2. Mix the values in both the sample groups and rank them in the ascending order
3. Divide the ranks into two sample groups
4. Summate the rank values in the first sample and second sample separately.
5. Use the normal curve approximation formula $Z== \frac{U-\mu}{\sigma}$
6. Compare with Z table value and decide fate of H0

$U = n_1.n_2 + \frac{n1\,(n1+1)}{2} - R_1$ $\mu = \frac{n1 \times n2}{2} = \frac{12 \times 12}{2} = 72$ $\sigma = \sqrt{\frac{n1n2(n1+n2+1))}{12}}$

Ex. 12.2

The values in one sample are 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. In another sample, they are 44, 60, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. Test @ 10% level the hypothesis that they come from population with the same mean. Apply Mann Whitney u test.

Ans : Ho ; the two samples come from population with the same mean.

Items in ascending order	RANK	SAMPLE A or B
32	1	B
38	2	A
39	3	A
40	4	B
41	5	B
44	6.5	B
44	6.5	B
46	8	A
48	9	A
52	10	B
53	11.5	B
53	11.5	A
57	13	A
60	14	A
61	15	B
67	16	B
69	17	A
70	18	B
72	19.5	B
72	19.5	B
73	21.5	A
73	21.5	A
74	23	A
78	24	A

Sum of ranks assigned to sample A or R_1= 2+3+8+9+11.5+13+14+17+21.5+21.5+23+24=**167.5**

sum of ranks assigned to sample B or R_2 = 1+4+5+6.5+6.5+10+11.5+15+16+18+19.5+19.5 = **132.5**

We have $n_1 = 12$ and $n_2 = 12$. Hence, test statistic $U = n_1.n_2 + \frac{n1\,(n1+1)}{2} - R_1$

$= 12 \times 12 + \frac{12(12+1)}{2} - 167.5 = 144 + 78 - 167.5 =$ **54.5**

$\mu_u = \frac{n1 \times n2}{2} = \frac{12 \times 12}{2} =$ **72**

$\sigma = \sqrt{\frac{n1n2(n1+n2+1))}{12}} = \sqrt{\frac{12\ x\ 12(12+12+1))}{12}} =$ **17.32**

$Z == \frac{U-\mu}{\sigma} = \frac{54.5-72}{17.32} = -1.01$

Since the calculated U value -1.01 is less than Z table value 1.96 the difference is not significant and therefore hypothesis is accepted. The two samples come from the same population.

Khruskal Wallis Test (H Test)

This test is similar to one way analysis of variance, but it does not require the assumption that samples come from approximately normal population or the population to have the same standard deviation or variance.

Khruskal wallis test examines whether three or more samples come from identical populations or their means are approximately the same. The test is used to test the null hypothesis that three or more independent samples come from identical populations. The test focuses on sum of ranks. The test approximately follows chi square distribution and therefore can be compared with chi square value for comparing test result, and deciding the null hypothesis.

The test is conducted in a way similar to the U test described above. This test is used to test the null hypothesis that K independent random samples come from identical universes against the alternative hypothesis that the medians of these universes are not equal. This test is analogous to the one way analysis of variance, but unlike the latter it does not require the assumption that the samples come from approximately normal populations or the universes having the same standard deviation.

In this, test like the U test, the data are ranked jointly from low to high or high to low, as if they constituted a single sample. The test statistic is H for this test.

If the null hyphenise is true that there is no difference between sample medians and each sample has at least five items, then the sampling distribution H can be approximated with a chi square distribution with (k-1) degrees of freedom. As such we can reject the null hypothesis at a given level of significance if H value calculated, as stated above, exceeds the concerned table value of chi square. For small sample the critical values can be taken from the Table given in the appendix, for comparison and decision.

Steps;

1. Form null hypothesis that samples belong to identical population
2. Consider values in several sample groups
3. Mix the values and rank them jointly in increasing order from lowest to higher
4. If there are ties, assign mean of tied ranks.
5. Find sum of ranks for all samples separately and get R1, R2, R3,……
6. Obtain test statistic $H = \frac{12}{n^2+n} \Sigma \frac{Ri^{2\ldots}}{ni\ldots} - 3n + 3$
7. Compare with chi square Table value at specified degree of freedom (C-1) and level of significance

EX. 12.3

Use the Khruskal Wallis test at 5% level of significance to test the null hypothesis that a professional bowler performs equally well with the four bowling balls, given following results.

With Ball A	271	282	257	248	262
With Ball B	252	275	302	268	276
With Ball C	260	255	239	246	266
With Ball D	279	242	297	270	258

Bowling results (ordered)	Rank	Ball
302	1	B
297	2	D
282	3	A
279	4	D
276	5	B
275	6	B
271	7	A
270	8	D
268	9	B
266	10	C
262	11	A
260	12	C
258	13	D
257	14	A
255	15	C
252	16	B
248	17	A
246	18	C
242	19	D
239	20	C

Ball A	**Rank**	**Ball B**	**Rank**	**Ball C**	**Rank**	**Ball D**	**Rank**
271	**7**	**252**	**16**	**260**	**12**	**279**	**4**
282	**3**	**275**	**6**	**255**	**15**	**242**	**19**
257	**14**	**302**	**1**	**239**	**20**	**297**	**2**
248	**17**	**268**	**9**	**246**	**18**	**270**	**8**
262	**11**	**276**	**5**	**266**	**10**	**158**	**13**
n 1 = 5	**R 1 = 52**	**n 2 = 5**	**R 2 = 37**	**n 3 = 5**	**R 3 = 75**	**n 4 = 5**	**R 4 = 46**

$$H = \frac{12}{n^2+n} \Sigma \frac{Ri^{2\ldots}}{ni\ldots} - 3n + 3$$

$$H = \frac{12}{20^2+20} \Sigma \frac{52^2}{5} + \frac{37^2}{5} + \frac{75^2}{5} + \frac{46^2}{5} - 3 \times 20 + 3 = 4.15$$

H value follows chi-square distribution. Chi-square value for C-1 = 4-1 = 7.815. The calculated value 4.51 is less than chi-square table value. The difference is not significant and the hypothesis is accepted.

Ex 12.4

Sales effected by four salesmen in four districts are given below. Use Kruskal Wallis test at 1% level of significance to examine whether the four salesmen have performed equally in their sales drive.

	Sales figures				
Salesman A	171	182	157	148	162
Salesman B	152	175	202	168	176
Salesman C	160	155	139	146	166
Salesman D	179	142	197	170	158

Sales results (ordered)	Rank	Ball
202	1	B
197	2	D
182	3	A
179	4	D
176	5	B
175	6	B
171	7	A
170	8	D
168	9	B
166	10	C
162	11	A
160	12	C
158	13	D
157	14	A
155	15	C
152	16	B
148	17	A
146	18	C
142	19	D
139	20	C

A	Rank	B	Rank	C	Rank	D	Rank
271	**7**	**252**	**16**	**260**	**12**	**279**	**4**
282	**3**	**275**	**6**	**255**	**15**	**242**	**19**
257	**14**	**302**	**1**	**239**	**20**	**297**	**2**
248	**17**	**268**	**9**	**246**	**18**	**270**	**8**
262	**11**	**276**	**5**	**266**	**10**	**158**	**13**
n 1 = 5	**R 1 = 52**	**n 2 = 5**	**R 2 = 37**	**n 3 = 5**	**R 3 = 75**	**n 4 = 5**	**R 4 = 46**

$$H = \frac{12}{n^2+n} \Sigma \frac{Ri^{2\cdots}}{ni\cdots} - 3n + 3$$

$$H = \frac{12}{20^2+20} \Sigma \frac{52^2}{5} + \frac{37^2}{5} + \frac{75^2}{5} + \frac{46^2}{5} - 3 \times 20 + 3 = 4.15$$

For Kushkjal Wallis test, we follow chi-square Table. At 1% level of significance, Chi-square table value for C

-1 = 4-1 = 7.815. The calculated H value 4.51 is less than chi-square table value. The difference is Insignificant and hypothesis is accepted. Four salesmen have performed equally in sales drive.

Wald –Wolfowitz test or Runs Test

Named after Abraham Wald and Joseph Wolfowitz, This test is a non parametric statistical test and is used to examine the hypothesis that a sequence of numbers is random. Mathematically a run is a set of sequential values that are either all above or below the given mean.

Applying the many statistical concepts discovered throughout the above units , it was always assumed that all sample data had been collected by some randomization procedure.

The run test based on the order in which sample observation are obtained is a useful technique for testing the null hypothesis that observations are drawn at random.

For example suppose that 20 people are surveyed as to use of cosmetics. A researcher may be eager to test the randomness of the simple, or is there any bias towards men or women.

In this case men may be denoted as M and women as W. The sequence in which 20 men and women surveyed may be as below: WMMWWWMMWWWMMMM, WMWMW . First W is a run, next MM is another run, and next WWW is a run and so on. There are altogether 11 runs in this survey.

Here each groupings is a run. Thus a run is a subsequence of one or more identical symbols representing a common property of the data.

Run test is conducted to examine deviation from randomness of sequence of numbers or objects. The test statistic can be approximated to approach normal distribution and therefore the required value can be obtained as Z value. In this test, the null hypothesis is that there is randomness. That is, if the calculated Z value is less than Z table value, items are considered as selected in random manner, and vice versa.

$$Z = \frac{r-\mu}{\sigma} \quad \text{where } r = \text{number of runs}, \quad \mu = \text{population average} = 1 + \frac{2n1n2}{n1+n2}$$

$$\sigma = \frac{\sqrt{2n1n2(2n1n2-n1-n2)}}{(n1+n2)^2\,(n1+n2-1)}$$

Steps

1) Form null hypothesis that there is randomness.
2) Consider sequence of similar numbers or objects
3) Count the number of runs = r.

4) Find expected population mean of run $\mu = \frac{2n1n2}{n1+n2} + 1.$

5) Obtain $\sigma = \frac{\sqrt{2n1n2(2n1n2-n1-n2)}}{(n1+n2)^2\,(n1+n2-1)}.$

6) Calculate z value = $\frac{r-\mu}{\sigma}$

7) Compare with z table value at required α.

8) Decide the fate of null hypothesis.

Ex 12.5

20 men and 10 women queue before a bank. A man argues that men are neglected and women favored by the clerk. Is it true? Conduct run test at α=0.05. Presently the sequence of men and women are;

MM WW WWWWWWMMMMM WWW MMMM WWWWW MMM

H0; men and women were chosen randomly.

Number of runs r = 7

Number of men n_1 =20, women n_2=10

$\mu = \frac{2n1n2}{n1+n2} + 1. = \frac{2*20*10}{20+10} + 1 = 14.33$

$\sigma = \frac{\sqrt{2n1n2(2n1n2-n1-n2)}}{(n1n2)^2(n1+n2-1)}. = \frac{\sqrt{2*20*10(2*20*10-20-10)}}{(20*10)^2(n1+n2-1)} = 2.38$

Z value = $\frac{r-\mu}{\sigma} = \frac{7-14.33}{2.38} = -3.079$

The calculated value of Z for runs test is -3.079, and is more than Z table value at α=0.05 = 1.96. The difference is considered significant. The hypothesis is rejected. Men and women were not chosen randomly. The argument that men are neglected and women favored by clerk is true.

Review questions and Exercises

1. What is the basis of rank sum test
2. Distinguish between signed rank test and rank sum test.
3. What is the use of Mann Whitney U test
4. State the steps in Mann-Whitney U tests
5. What is Kruskal Wallis H test?
6. How it is H test conducted
7. What is the basis of runs test
8. State the steps in performing runs test
9. Explain the hypothesis in Mann Whitney U test.
10. There are two samples. First sample contains the observations ; (54,39,70,58,47,40,74,49,74,75,61 and 79). The second sample contains (45,41,62,53,33,45,71,42,68,73,54 and 73). Apply Rank Sum test to test at 5% level of hypothesis that they come from populations with the same mean

11.

Following are the kilometers per gallon which a test driver got for ten vehicles filling each of three kinds of gasoline.

Gasoline a	43	27	43	29	42	49	48	21	37	--
Gasoline b	39	40	35	26	34	45	32	22	23	18
Gasoline c	28	30	25	31	41	38	36	44	19	50

Use the Kruskel Wallis test at the level of significance α = 0.05 to test the null hypoyhesis that there is no difference in the average km. yield of three types of gasoline.

12. A driver buys diesel either at a Texaco station (t), or at a Mobile station (m), and the following arrangement shows the order of the stations from which she bought diesel over a certain period of time.

t	t	t	m	t	m	t	m	t	m	m	m
t	m	m	m	t	m	m	t	m	t	m	m
t	m	m	t	t	m	t	m	m	m	t	m
t	t	t	m	t	t	m	t	t	t	t	m

Test for randomness at the 0.05 level of significance.

13. Following are the speeds at which every 5th passenger car was timed in a certain check point.

46	58	60	56	70	66	48	54	62	41	39	52
45	62	53	69	65	65	67	76	52	59	59	67
51	46	61	40	43	42	77	67	63	59	63	63
72	57	59	42	56	45	62	67	70	63	66	69

Test the null hypothesis for randomness at 0.05 level of significance. Do sign test.

(Hint median = 59.5, take value less than 59.5 as negative sign , and values greater than 59.5 as positive signs. Any value equal to 59.5 is ignored)

14. Use the Kruskell Wallis H test at 5% level of significance to test the hypothesis that a professional bowler performs equally well with 4 bowling balls.

score ball a	271	282	257	248	262
score ball b	252	275	302	268	276
score ball c	260	255	239	246	266
score ball d	279	242	297	270	258

15.Following are the number of students absent from a college on 24 days. Test for randomness at α = 0.01

29	25	31	28	30	28	33	31	35	39	31	33
35	28	36	30	33	26	30	28	32	31	38	27

UNIT XIII

STATISTICAL QUALITY CONTROL - CONCEPTS

Introduction

Until recently industry worldwide was not much bothered about quality of products and services. However, increasing globalization of production, commerce and science, global industry had to respond to challenges in quality improvement. One of these challenges was a dedication to quality control and the management of quality in production. In response to this challenge, philosophy of quality in production and techniques of quality of control and management are becoming widespread.

Quality Control is one of the most useful statistical techniques in industries and service sector. It is applied largely for detecting malfunctioning that creeps into management process.

SQC refers to the use of statistical methods in controlling the quality of manufactured products. It is the means of establishing and achieving quality specification, which requires use of statistical tools and techniques. It is an application of the theory of sampling and the theory of probability. It is a specialized technique used to improve the technical efficiency of production processes under mass production system. The technique is extensively used in almost all industries such as consumer goods, aircraft, armament, automobile, textile, plastic, rubber, petroleum, electrical equipment, telephone, transportation, chemical l, medicine etc.

What Is Quality?

Quality implies fitness for use. Quality means conformance to requirements. Besides, quality includes consistency, reliability and lack of errors and defects. Things that are of high quality are those which work in the way that we expect them to.

Most of us do connect luxury with quality. But expensive and luxurious products don't exactly mean quality products, if they do not come up to expected standards. Quality is to be judged by consumers and producers jointly.

Variability and Quality

Variability or variation is the enemy of quality . When a craftsman is making something by hand he continuously check, measure, assess and rework. Quality control is not an issue when such artworks and unique products are made. But when mass production became common during the 19^{th} century, variations became an issue to be controlled. With too many variations, quality began to fade away. Now industries and managers are conscious of regaining quality of products and services.

Causes of Variations in quality

No production process is good enough to produce all items exactly alike. No two articles produced by the same machine are perfectly identical in measurable characteristics. Thus variations are inherent and inevitable in every repetitive process in industry. These variations are grouped into two classes – random variations and assignable variations.

Random Variations

Random variations may result from random combinations of circumstances that cause slight difference in the individual units produced. These variations may be considered as simply a

characteristic of the manufactured process. Even though same machine, materials, labout and manufacturing techniques are used, some variations may occur in the product. It is impractical to find the reason for each such variation since they are purely the result of chance. Random or chance variations are uncontrollable. They cannot be eliminated. If variability of the process is confined to assignable factors, the process is said to be under statistical control.

Assignable variation

Assignable variations comprise those variations that result from specified causes which can be identified. Mistakes of inexperienced workers, worn out tools, machines in need of adjustment, defective raw materials etc are some of the assignable causes for such variation. Assignable causes may creep in at any stage of the process, right from the arrival of the raw materials to the final delivery of goods. They will either change the average value or dispersion or both. The assignable causes can be identified and eliminated and have to be traced in any production process , before the production becomes defective. Actually, statistical control technique deal with such assignable variation.

Definitions of SQC

SQC is a "statistical method of arriving at an estimate of the quality of industrially produced goods and services, on the basis of sampling and probability."

SQC is "setting specifications for quality, and assuming whether a production process conforms to such specifications on the basis of periodical samples draw."

Objectives of Quality Control

In this highly competitive, cost aware and quality conscious world, need for SQC cannot be ignored. Following are the objectives of SQC

1. To locate and identify process faults, leading to variations.
2. To separate chance variations and assignable variations.
3. To take necessary steps to maintain quality of products.
4. To make adjustments and modifications in process to prevent rejections.
5. To increase and maintain quality level for consumer satisfaction.

Uses of SQC

SQC is useful and widely applied quantitative technique for assuring quality of industrially manufactured products and services.

1. It is used to assure whether production process is within control or out of control
2. It enables to check whether products meet global standards.
3. It isolates chance variations and assignable variations.
4. It reveals wear and tear of machines and processes.
5. It reduces scrap and rejections.
6. It gives easily and timely warning about possibility of occurrence of defects.
7. Its presence creates quality consciousness among workers.
8. It is the base of total quality management.

Techniques of SQC

Statistical Quality Control is done in two ways – process control and product control

Process control

The main objective of any production process is to control and maintain a satisfactory quality level for its product. It should be ensured that production conforms to specified quality standards. That is, it should not contain a large number of defective its. This is termed as process control

A process is said be in a state of statistical control if the variation is attributable to chance variations only and no special cause can be pointed out. But when the process is out of control, it should be possible to locate specific causes for the variation and it should be possible to remove them to improve the future performance of the process. The process control is achieved through the technique of control charts.

Product control

By product control we mean controlling the quality of the product by critical examination at strategic points which is achieved through sampling inspection plans.

When the products come to the market, the producer has to consider the requirements of the customers or the firms or companies who receive the end products. In product control, the producer wants to ensure himself that the manufactured goods are according to the specifications of the customers or receiving companies. Product control aims at guaranteeing a certain quality level to the consumer regardless of what quality level is being maintained by the producer. In other words it attempts to ensure that the product marketed for sale does not contain a large no of defective items.

Control charts

One of the most important tool of production management and control of quality is the control chart technique. control chart is graphical device for marking quality levels, and analyzing quality deviations during production process.

The essence of process control is to identify a parameter that is easy to measure and whose value is important for the quality of the process output. These are statistical tools which enable to recognize non-random variations and decide when to make adjustment to a process.

Technique of control charts

A control chart is a statistical device used for the study and control of repetitive processes. The basis process is not the quantum of the product but the quality of the product. The producer is basically interested to see that the product is of acceptable quality, That is, whether it conforms to certain prescribed standards or specifications. All items produced will not be identical in measurements. Manufacturing specifications take into account the variability of identical items and allow tolerance limits within which measurements fall .Items falling within the tolerance limits are judged to be of acceptable quality, and those falling outside the tolerance limit be reworked.

A control chart consists of a diagram with three horizontal lines-central line, upper control limit and lower control limit. Central line is the ideal quality of production specification. Maximum amount of tolerable variation in quality is called upper control limit, and the lowest tolerable limit of quality specification is called lower control limit.

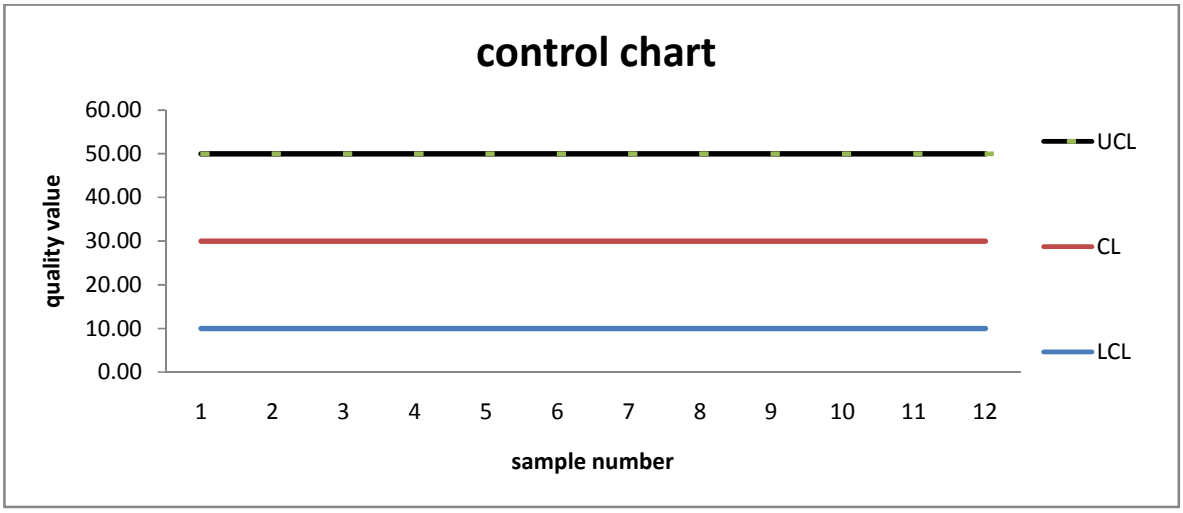
Periodically sample values are plotted on the control chart. If all the plots fall within the upper and lower control limit, we can ensure that quality is under control. If any point fell beyond

the control limit, the production is not under control or there are only chance variations. If any point falls in the region outside control limit, called out of control region, it indicates that there is some assignable cause of variation and the process is not under statistical control.

Steps in construction of control chart

1. Determine the type of control chart - for variables or for attributes.
2. An appropriate statistic may be selected- like mean, range, no of defectives etc.
3. Calculate the central line which is the ideal production specification. It is obtained by taking, for example, mean of several means.
4. Find lower control limit, which is three standard deviations lower than central line.
5. Find upper control limit, which is three standard deviations higher than central line.
6. Select an appropriate scale and draw x axis and y axis.
7. Draw Central Line, Upper Control Limit and Lower Control Limit.
8. Plot the values of the given samples on the chart.
9. If all the points fall within the limits, decide that process is under control.
10. If any point falls beyond the limits, decide that process is not under control, and therefore require correction.

Model of a control chart



Model control chart

Acceptance Sampling

Producers produce goods and services for consumption by consumers. Consumers and their acceptance of products cannot be ignored in SQC. A buyer draws a sample of a few units from a lot that is to be accepted by him if the number of defects and defective items are less. If the number of defective is more than expected, the lot is to be rejected. This technique of sampling inspection and decision making is called acceptance sampling.

Acceptance sampling is concerned with the decision to accept a mass of items or to reject them on the basis of they being conforming to specifications. When the products come to the market, producer has to see that the manufactured goods are according to the specifications of the customers.

Product control is exercised through samples. Lot by lot, sampling inspection is performed. A lot is accepted or rejected on the basis of information yielded by sampling. Cent per cent inspection is very costly. Even with 100% I nsp0ection, defective items seem to slip. In some cases quality can be tested only by destroying items. 100% inspection is out of question in such cases. For example , testing life of bulbs, which are perishable by inspection of the filament. For these reasons, acceptance sampling is adopted in modern manufacturing centers.

Following three types of acceptance sampling plans are commonly used –single sampling plan, double sampling plan and Multiple Sampling plan

Single sampling plan

This consists in arriving at a decision to accept or to reject the lot on the basis of inspecting a single sample only. Let us denote

N = Lot size

n = sample size

c = maximum number of allowable defectives in the sample (acceptance number)

d = number of defectives in the sample

Steps in single sampling plan

Single sampling plan is very popular in manufacturing and durable goods industries. It involves following steps:

1. Take a random sample of size n from the lot of size N
2. Inspect the sample 100% and count the number of defectives d in the sample
3. If d is less that c accept, the lot, replacing all the defective pieces found in the sample by good ones.
4. If d is greater than c, reject the lot.

In this case , we resort to 100% inspection of the lot and replace all defective pieces by standard ones. Single sampling plan is very simple to understand, design and carry out. The basic problem in a single sampling plan is the choice of n and c. The most economical single sampling inspection plan is obtained in minimizing the average total inspection by providing adequate protection to consumer and producer.

Double Sampling Plan

Double sampling inspection plan provides for taking a second sample, if we are not in a position to arrive at a decision, about accepting or rejecting a lot on the basis of a single sample.

Let us denote N for lot size

n1 for size of 1sr sample

C1 for acceptance number of first sample

n2 for size of second sample

C2 for acceptance number for both the samples combined

D1 for number of defective items in the first sample

D2 for number of defective items in the second sample.

Steps in double sampling plan

Double sampling plan involves following steps:

1.take a random sample of size n1 from the lot size N

2. if d1 is less than c1, accept the lot, replacing all the defective items by good ones.

3. If d2 is greater than c2, reject the lot. Resort to 100% inspection, replacing all defectives by good ones.

4. If c1 is less or equal to d1 which is less than or equal to c2, take a second sample of size n2 from the remaining lot of size N - n 1

5. If d1 + d2 ls less than or equal to c2, accept the lot, replacing all defective items by good ones

6. If d1 + d2 is greater than or equal to c2, reject the lot. Resort to 100% inspection and replace all defective items by good ones.

Dodge and Rooming obtained the most economical double sampling plans after providing adequate protection to producer and consumer such that

(1) Average number of total inspection is minimum, and

(11) Probability of acceptance on the basis of first sample is same as the probability of acceptance on the basis of second sample.

Double sampling inspection plans give a second chance of being accepted for the border line lots. The advantage of double sampling over single sampling is psychological as it appears psychologically more convincing to say that the lot was rejected after inspecting two samples.

Sequential sampling

It is an extension of double sampling plan. In multiple sampling inspection plans, we arrive at a decision to accept or reject after inspecting more than two samples. In multiple inspection plans, we get multiple chances of examining and accepting border line lots, and minor defective items,

Types of control charts

Control charts are broadly classified as chart for variables and chart for attributes, are the basis of nature of measurement and data used.

Control chart for variables

Variables are characteristics of products which are quantitatively measurable like thickness of screw, weight of can etc. Control chart for variables focus on maintain process quality level of such products. Observations on such unit can be expressed in specific units of measurement. Mean chart, range chart and are examples of control chart for variables.

Control chart for attributes

Attributes are properties the presence or absence of which, can be measured and expressed. For example products may be classified and counted as defective and non-defective. Control chart can be drawn on the basis of number of defective articles or number of defects. Such charts are called control charts for attributes. Examples are number of defective chart, proportion of defective chart, and number of defects chart etc.

The forthcoming unit contains control chart for variables- i.e., mean chart, range chart and standard deviation chart. Control chart for attributes are described in the next unit.

Review Questions and Exercises

1. What is the significance of statistical quality control
2. What is quality ? what is quality control?
3. Define quality control
4. Explain variability and quality.
5. What are the causes of variations in quality
6. Explain objectives quality control
7. What are the uses of quality control?
8. What are control charts?
9. Distinguish between process control and product control
10. What is the basis of quality control
11. Explain the technique of control chart
12. Give classification of control charts
13. What are control charts for variables
14. Explain co9ntrol chart for attributes
15. What is meant by an attribute
16. Give a model of a control chart
17. The value of a control chart are given below. Draw control chart
    Central Line = 32.4 Upper control line = 38.8 Lower control line = 26.0
18. The Upper Control Limit and Lower Control Limit of a certain control process are 40.6 and 30.4 respectively. Ascertain Central Line and draw the control chart.
19. Central Line of a control chart situates at point 52.6, and Upper control Line is at 55.0. What would be the point of Lower control Line?
20. The Upper and Lower Control Limits of a control chart are 17.6 and 11.6 respectively. Ascertain the Central Line.

UNIT XIV

CONTROL CHARTS FOR VARIABLES

A popular control chart which is widely used in controlling quality of industrially manufactured products is control chart for variable. Variables are characteristics of products which are quantitatively measurable like thickness of screw, weight of can etc. Control chart for variables focus on maintaining process quality level of such products through controlling mean values of certain given samples. Observations on such unit can be expressed in specific units of measurement. Mean chart, range chart and standard deviation charts are examples of control chart for variables

Mean chart

The essence of statistical process control is to identify a parameter which is easy to measure and whose value is important for the quality of the process output. Such a parameter is arithmetic mean or simply mean. Control chart drawn focusing on arithmetic mean of observations of samples, is called mean chart.

Mean chart is prepared to show the fluctuations of means of samples. It can be used to determine whether or not fluctuations are due to random causes or assignable causes.

Steps in drawing Mean chart

1. Ascertain means of given samples = $\bar{x}1, \bar{x}2, \bar{x}3$ ……
2. Calculate grand mean, ie., mean of means = Central Line = CL = $\bar{\bar{x}}$
3. Obtain UCL = $\bar{\bar{x}} + 3\sigma$ (if standard deviation is given)
   or UCL=$\bar{\bar{x}}$+A2 $\bar{R}$ (if ranges are given) .
4. Obtain LCL =$\bar{\bar{x}}$ - 3σx or $\bar{\bar{x}}$ - A2 $\bar{R}$. Where A2 = table value
5. Select appropriate scale and draw x axis and y axis
6. Draw CL, UCL and LCL on the graph.
7. Plot values of samples on the graph.
8. Examining the plots, decide whether process is under control or not.
9. If all the points fall within the control limits, the process is under control
10. If any point falls beyond any control line, the process is not under control

Ex. 14.1From the following details draw control chart for mean

Sample no	weight
1	43
2	49
3	37
4	44
5	45
6	37
7	51
8	46
9	43
10	47

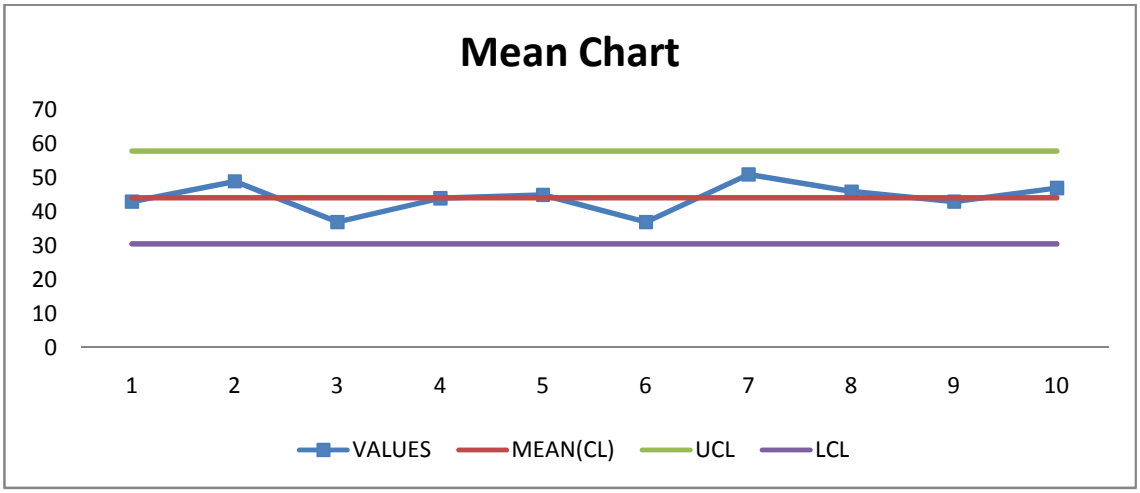
Standard deviation = 4.5

Ans

Central Line = Mean of means = 44.2

Upper control Line = Mean + 3 x σ = 44.2 + 3 x 4.57 = 57.9

Lower control Line = Mean - 3 x σ 44.2 - 3 x 4.57 = 30.5



Since all points fall within control limits, the process is under control.

Ex 14.2

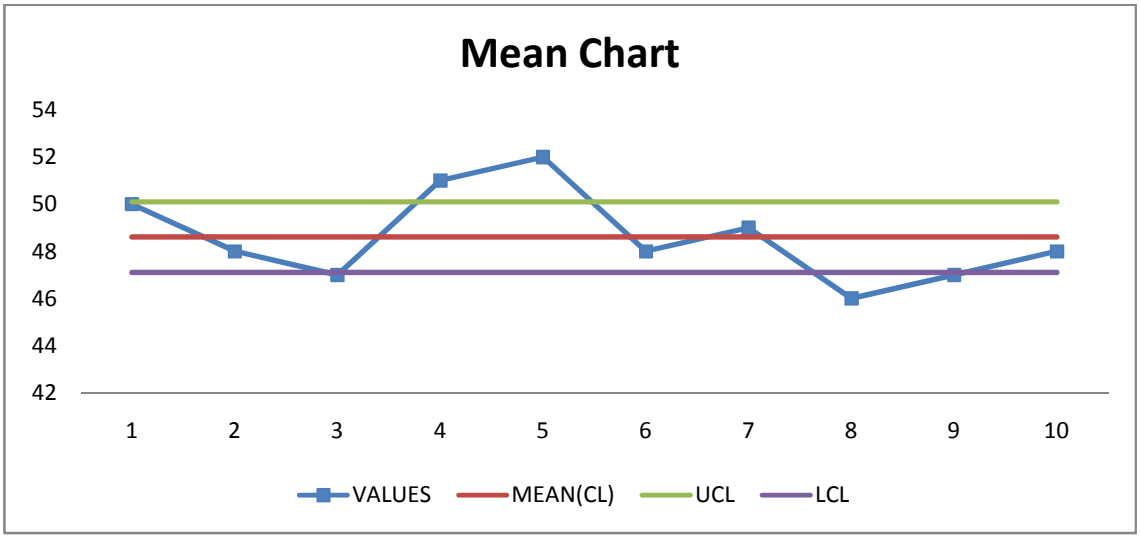
In the production of cement, 10 sample bags were selected for quality and weights are given below. Standard deviation of weight is .5kg.

Sample number	weight
1	50
2	48
3	47
4	51
5	52
6	48
7	49
8	46
9	47
10	48

Mean of sample mean=CL= $\frac{486}{10}$= 48.6

UCL =$\bar{\bar{x}} + 3\sigma X$ = 48.6+3×.5 = 50.1 KG

LCL = $\bar{\bar{x}}$- 3σx = 48.6 - 3×.5 = 47.1 kg



Some points fall beyond the two control limits, therefore the process is not under control

Ex 14.3

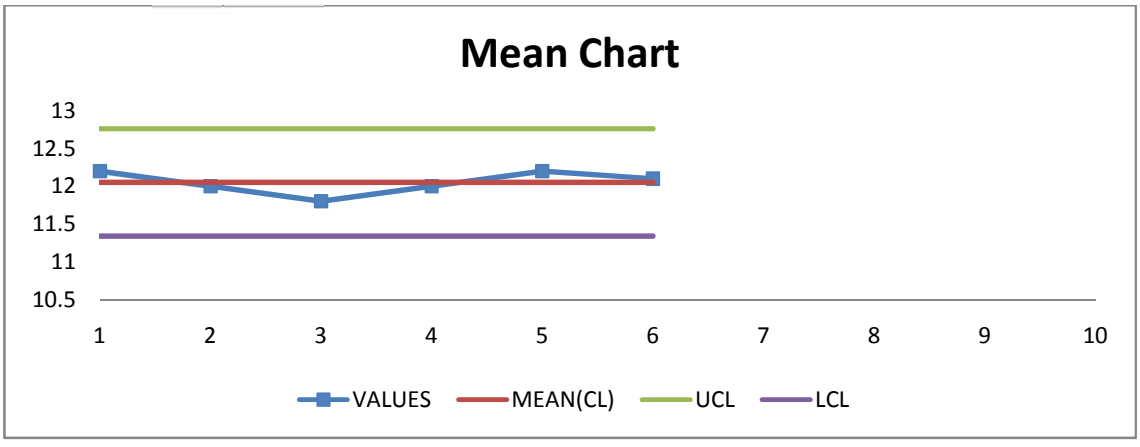
Construct a control chart for mean from the following sample size

Sample number	Mean length	Range
1	12.2	1.2
2	12.0	1.4
3	11.8	1.2
4	12.o	1.0
5	12.2	1.2
6	12.1	1.4

CL = $\frac{72.3}{6} = 12.05$ Range = $\frac{7.4}{6} = 1.23$

UCL=$\bar{\bar{x}}$+A2$\bar{R}$ = 12.05+0.577x1.23=12.76

LCL=$\bar{\bar{x}}$ -A2 $\bar{R}$. = 2.05 - 0.577 x1.23=11.34



As all points fall within the upper and lower control limits, quality control program is perfect.

Interpretation of mean chart

In the mean chart, the three lines are the Central line, Upper Control Line and Lower Control Line. The Central Line represents ideal quality of process. The Upper and Lower Control limits are maximum and minimum tolerable limits of process quality.

Reasonably all points should fall within control limits. If any point falls beyond the upper control limit or lower control limit, the process is said to be out of control. If all the points fall within the control limits, the process is under control.

Mean chart reveals undesirable variations between samples. If the variations are more than chance variations, such variations are assignable variations and the process is considered to be not under control. Such variations are assignable to specified causes and leading to lack of quality.

Range chart

When sample range values are given, range chart can be drawn. Range chart is used to show the fluctuations of the ranges of samples about the mean of ranges. The range chart is diagram on which ideal range value and upper and lower limits are drawn. Range values will be plotted, and quality of process assessed, on the basis of such plots. Similar to mean chart, when the values are plotted on the control chart, they are verified as to whether they fall within or beyond the control limits. If all the points fall within the control limits, the process is considered to be within control. If any or some of the points fall beyond the limits, the process is considered to be out of control, which requires special and immediate attention.

Normally, standard deviation will not be given in order to calculate upper and control limits. In such case, we make use of control chart table given at the end of this book. For the purpose of range chart , we make use of D4 and D3 values

Steps to draw range chart

1. Select scale and draw x axis and y axis.
2. Find value of mean of given ranges= $\bar{R}$. = CL
3. Find Upper Control Limit= UCL= $\bar{R}.+3\sigma$, or $D4\bar{R}.$.
4. Find Lower Control Limit = LCL = $\hat{R}$-3σ. Or D3$\bar{R}$.
5. D4 and D3 values are table values for control limits.
6. Draw the lines on the graph.
7. Plot values of given sample ranges.
8. Decide process quality on the basis of plots falling within or beyond the limits.

Ex. 14.4

Construct a control chart for Means and Ranges from the following sample size

Sample number	Mean length	Range
	12.2	1.2
2	12	1.4
3	11.8	1.2
4	12	1
	12.2	1.2
6	12.1	1.4

Control limit values for Means

CL = $\bar{\bar{x}} \frac{72.3}{6} = 12.05$ range = $.\frac{7.4}{6} = 1.23$

UCL= $\bar{\bar{x}}$+A2$\bar{R}$ = 12.05+0.577 x 1.23=12.76

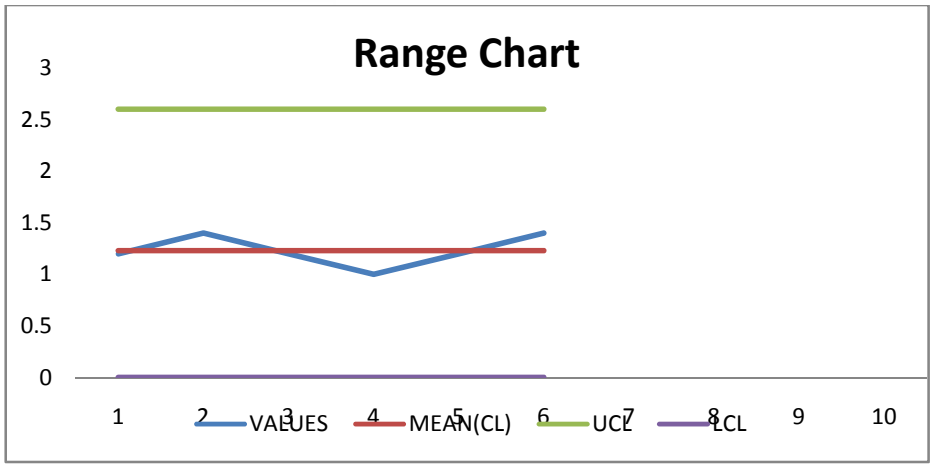
LCL= $\bar{\bar{x}}$ -A2 $\bar{R}$ = 12.05-0.577 x 1.23=11.34

Control limit values for Ranges

Central Line = $\bar{R} = \frac{7.4}{6}$ = 1.23

Upper Control Line = $D4\bar{R}$ = 2.115 x 1.23 = 2.60

Lower Control Line = $D3\bar{R}$ = 0 x 1.23 = 0



As per the range chart, all range values come within the upper and lower control limits, and the process is under control.

Ex. 14.5

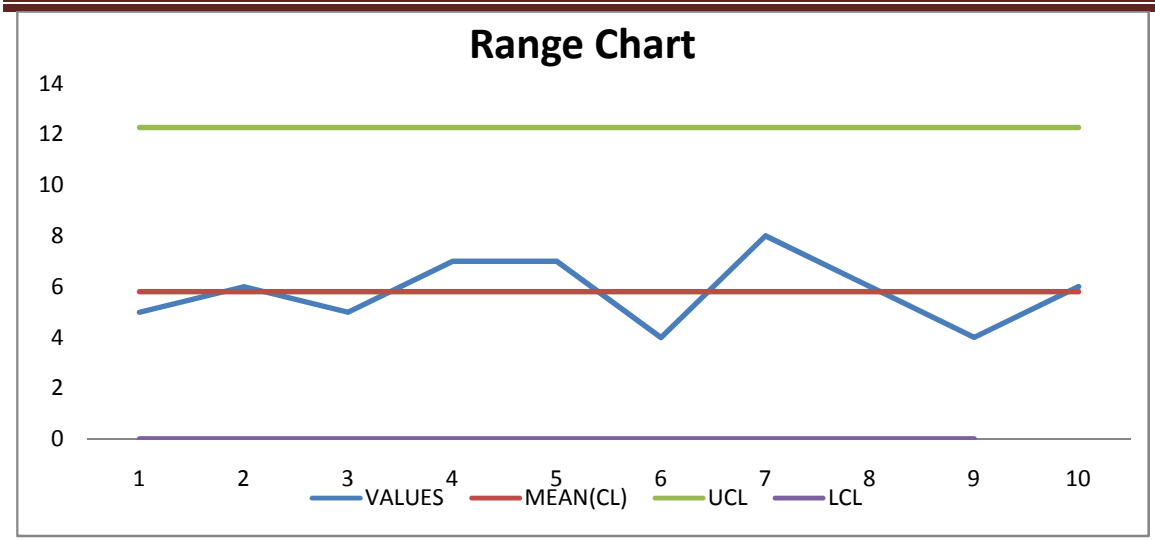
You are given following sample numbers and ranges relating to production of writing chalks. Construct range Chart and comment on the quality.

Sample	1	2	3	4	5	6	7	8	9	10
Range	5	6	5	7	7	4	8	6	4	6

Centrl Line = $.\frac{58}{10}$ = 5.8

Upper Control Line = $D4\bar{R}$ = 2.115 x 5.8 = 12.27

Lower Control Line = $D3\bar{R}$ = 0 x 5.8 = 0



Since all range values fall within the two control limits, the process is under control

Interpretation of range chart

Range chart reveals process quality on the basis of given range. Range is the difference between smallest value and largest value with a sample group.

Range chart contains three lines CL, UCL and LCL. Range chart focuses on variations with in samples groups and is more analytical than mean chart. Both mean chart and range chart should be studied together since the former reveals variations between samples and the latter reveals variations within sample groups.

Review Questions and exercises

1. Explain variable
2. State the causes of variations in quality
3. Distinguish between variable and attribute
4. State uses of control charts
5. What are important variable control charts
6. Explain control limits for a mean chart
7. How is a mean chart drawn
8. Draw a model control chart for mean
9. Explain range Chart
10. State the interpretation of a Mean Chart
11. What are the steps in drawing a Range Chart
12. How is Range chart interpreted
13. What is UCL and LCL

14. You are given the values of sample mean and the range for ten samples of size 5 each. Draw mean and range charts and comment on the state of control of the process.

Sample	1	2	3	4	5	6	7	8	9	10
X	43	49	37	44	45	37	51	46	43	47
R	5	6	5	7	7	4	8	6	4	6

You may use the following control chart constants.(For n = 5, A2 = 0.58,D3 = 0 and D4= 2.115)

15. A machine is set to deliver packets of a given weight. 10 samples of size 5 each were recorded. below are given relevant data:

Sample no.	1	2	3	4	5	6	7	8	9	10
Mean	15	17	15	18	17	14	18	15	17	16
Range	7	7	4	9	8	7	12	4	11	5

Calculate the values for the central line and the control limits for mean chart and then comment on the state of control. (Conversion factors of n=5, are A2=0.58, d3=0, D4=2.115)

16. Construct a control chart for mean and range for the following data on the basis of fuses, samples of 5 being taken every hour (each set of 5 has been arranged o ascending order of magnitude)

42	42	19	36	42	51	60	18	15	69	64	61
65	45	24	54	51	74	60	20	30	109	90	78
75	68	80	69	57	75	72	27	39	113	93	94
78	72	81	77	59	78	95	42	62	118	109	109
87	90	81	84	78	132	138	60	84	153	112	136

UNIT XV

CONTROL CHARTS FOR ATTRIBUTES

Mean chart and range chart are powerful statistical tool of diagnosis of sources of variations in production process. They are charts for variables.

As an alternative to mean chart and range chart, we have control charts of attributes which can be used for controlling quality characteristics. Quality characteristics can be either number of defectives, proportion of defectives or number of defects.

In the area of statistical process control a qualitative variable that can take on only specified values is called an attribute. In quality control, conformity to specification is a qualitative variable which can be either yes or no – ie, conforming to quality, or not conforming to quality.

If the products do not meet quality specification, they are called defectives. Defects are number of imperfection on a products. Control chart for attributes may be proportion defectives chart (P chart) number of defectives chart (nP chart), and number of defects chart (C chart).

Fraction Defectives Chart (P chart)

P chart is the control chart for fraction defectives or proportion of defectives among manufactured products. This is used when quality characteristics observed can be classified as defective or not defective. It is constructed to control the production of defective items.

The objective of this chart is to evaluate the quality of the items and to note the changes in quality over a period of time. It is constructed to control the production of defective items.

In fraction defective chart, the focus of quality control is on average defectives or $\bar{p}$. Proportion of defectives is obtained by dividing the total number of defectives by total number of items. The average fraction defective is taken as the Central Line. Thereon, data will be converted into fractions.

The UCL and LCL will be 3 standard deviations away from the central Line. Sample values will be in the form of fractions.

Steps in drawing fraction defective chart

1. Select appropriate scale and draw the two axes.
2. Find mean of defectives = $\bar{p}. = \frac{\Sigma d}{n}$ = CL.
3. Obtain UCL = $\bar{p} + 3 x \sqrt{\frac{pq}{n}}$
4. Obtain LCL = $\bar{p} - 3x \sqrt{\frac{pq}{n}}$
5. If the LCL falls below zero, take it as zero.
6. Draw the three lines on the control chart.
7. Ascertain and Plot fraction defectives (p) of each sample.
8. Decide process quality on the basis of plots.

Ex 15.1

In a certain sampling inspection, the number of defectives found in 10 samples 100 each are as given below:

16, 18, 11, 18, 21, 10, 20, 18, 17, and 21

Do these indicate that the quality characteristics under inspection is under statistical control?.

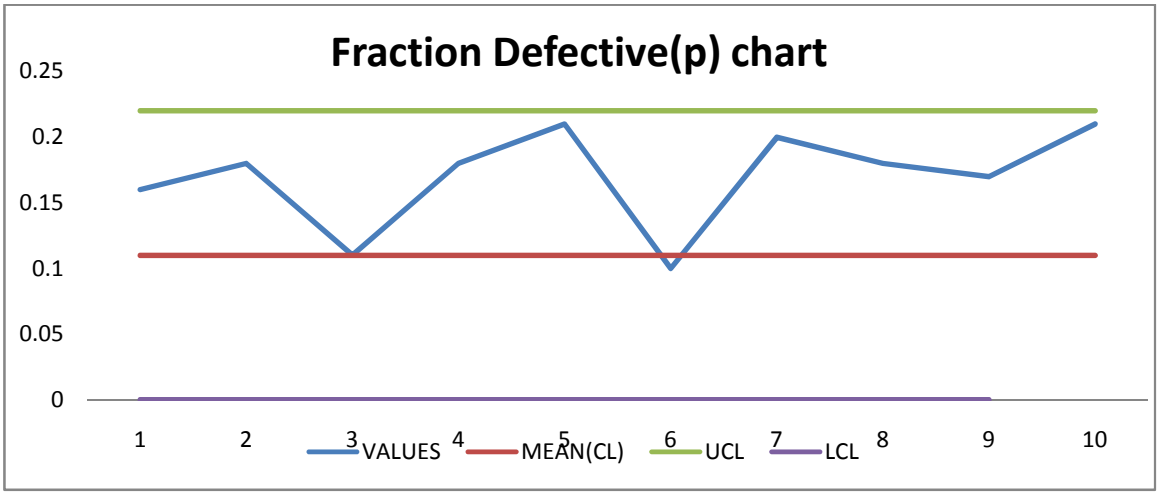
Mean of defectives = $\bar{p}. = \frac{\Sigma d}{n}$ = CL. = $\frac{170}{1000}$ = .17

Therefore $\bar{q}$ = 1- $\bar{p}.$ = 1- .17 = .83

UCL= $\bar{p}$ + 3 x $\sqrt{\frac{pq}{n}}$ = .11 + 3 x $\sqrt{\frac{.17 \, x \, .83}{100}}$ = .22

LCL.= $\bar{p}$ - 3x $\sqrt{\frac{pq}{n}}$ = .11 - 3 x $\sqrt{\frac{.11 \, x \, .89}{100}}$ = .00

P values = .16, .18, .11, .18, .21, .10, .20, .18, .17, and .21



All fraction defectives values fall within the control limits, and therefore the process is under control.

Ex 15.2

20 Samples of 100 batteries each are taken from a production process which gives number of defectives as below. Determine control chart limits for fraction defectives.

9,17, 8, 7, 12, 5, 11, 16, 14, 15, 10, 6, 7, 18, 16, 10, 5, 14, 7, 13

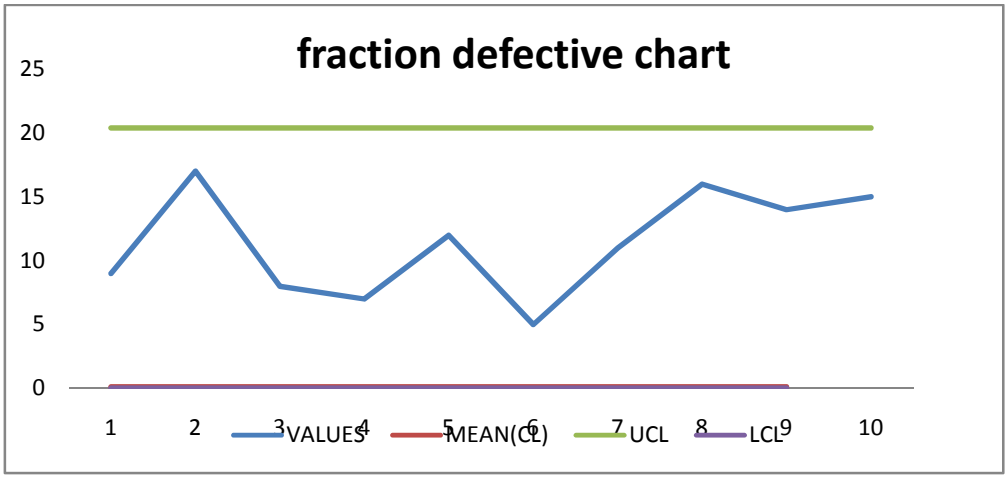
Mean of defectives = $\bar{p}. = \frac{\Sigma d}{n}$ = CL. = $\frac{220}{2000}$ = .11

therefore $\bar{q}$ = 1- $\bar{p}.$ = 1- .11 = .89

UCL= $\bar{p}$ + 3 x $\sqrt{\frac{pq}{n}}$ = .11 + 3 x $\sqrt{\frac{.11\ x\ .89}{100}}$ = .204

LCL.= $\bar{p}$ - 3x $\sqrt{\frac{pq}{n}}$ = .11 - 3 x $\sqrt{\frac{.11\ x\ .89}{100}}$ = .016

Fraction defectives .9, .17, .8, .7 . . . . . .



Since all fraction defective points come within the upper and lower control limits, the process is said to be under control.

Ex 15.3

The average number of defective in 22 sampled lots of 2000 rubber belts each was found to be 16%. Indicate how to construct the relevant control chart.

Here Fraction Defectives chart may be constructed.

Find CL = $\bar{p}$. =$\frac{\Sigma d}{n}$ = $\frac{16}{100}$ = .16

Obtain UCL = $\bar{p}$ + 3 x $\sqrt{\frac{pq}{n}}$ = .16 + 3 x $\sqrt{\frac{.16\ x\ .84}{100}}$ = .1846

Obtain LCL = $\bar{p}$ - 3x $\sqrt{\frac{pq}{n}}$ = .16 - 3 x $\sqrt{\frac{.16\ x\ .84}{100}}$ = .1354

For drawing the control chart for fraction defectives, defectives values under each sample must be given.

Ex 15.4

A daily sample of 30 cell phones was taken over a period of 14 days in order to establish attributes control limits. If 21 defectives were found, what should be the upper and lower control limits of the proportion of defectives?

CL = = $\bar{p}$. =$\frac{\Sigma d}{n}$ = $\frac{21}{30\ x\ 14}$ = .05

UCL = $\bar{p}$ + 3 x $\sqrt{\frac{pq}{n}}$ = ..05 + 3 x $\sqrt{\frac{.05\ x\ .95}{30}}$ = .169

LCL = $\bar{p}$ - 3x $\sqrt{\frac{pq}{n}}$ = .05 - 3 x $\sqrt{\frac{.05\ x\ .95}{30}}$ = -.069 taken as zero

Interpretation of P chart

From the P chart, a process is judged to be in statistical control if all the sample points fall within control limit. If one or more points fall beyond the two control limits-UCL and LCL, it shows deterioration in quality. Reasons for this should be traced out and eliminated. If a point goes below LCL or above UCL, it is an indication of bad quality. On the basis of fraction defectives chart, reasons for deviation from the ideal product quality specification can be studied and corrective measures adopted.

Control chart for number of defectives (d chart or np chart)

P chart is based on fraction defectives i.e. when the given data is on number of defectives out of a certain number of samples, with definite sample size. And fraction defectives are a little difficult to calculate and present.

Instead of fraction defectives, number of defectives can be directly subjected to quality control. From the given number of defectives, average defectives(đ) can be ascertained, and central line drawn. Then the control limits can be obtained at three standard deviations away from the central limit. Here it is to be noted that đ = $n\bar{p}$. From this $\bar{p}$ can be found as $\bar{p}$ = $\frac{n\bar{p}}{n}$, and $\bar{q}$ = 1 - $\bar{p}$, in order to ascertain standard deviation.

Thus CL= đ = $n\bar{p}$

UCL= $n\bar{p}$ + 3 $\sqrt{n\bar{p}\bar{q}}$

LCL= $n\bar{p}$ - 3 $\sqrt{n\bar{p}\bar{q}}$

Steps

1. Select scale and draw x and y axis.
2. Find mean of defectives= đ=$\frac{\varepsilon d}{n}$ = $n\bar{p}$=CL.
3. Ascertain $\bar{p}$ = $\frac{n\bar{p}}{n}$ and $\bar{q}$ = 1- $\bar{p}$
4. Ascertain UCL= $n\bar{p}$ + $^{3}\sqrt{n\bar{p}\,\bar{q}}$
5. Find LCL = $n\bar{p}$ - $^{3}\sqrt{n\bar{p}\,\bar{q}}$.
6. If LCL is less than zero, it is taken as zero.
7. Draw the three lines, plot the number of defectives and decide the quality.

Ex 15.5

A sample of 100 items was examined each hour from a production process. The number of defectives so found on a day is reproduced below:

16, 18, 12, 4, 10, 15, 13, 6, 7, 12, 10, 10, 2, 3, 13, 4, 1, 6, 5, 8, 4, 2, 5, and 6.

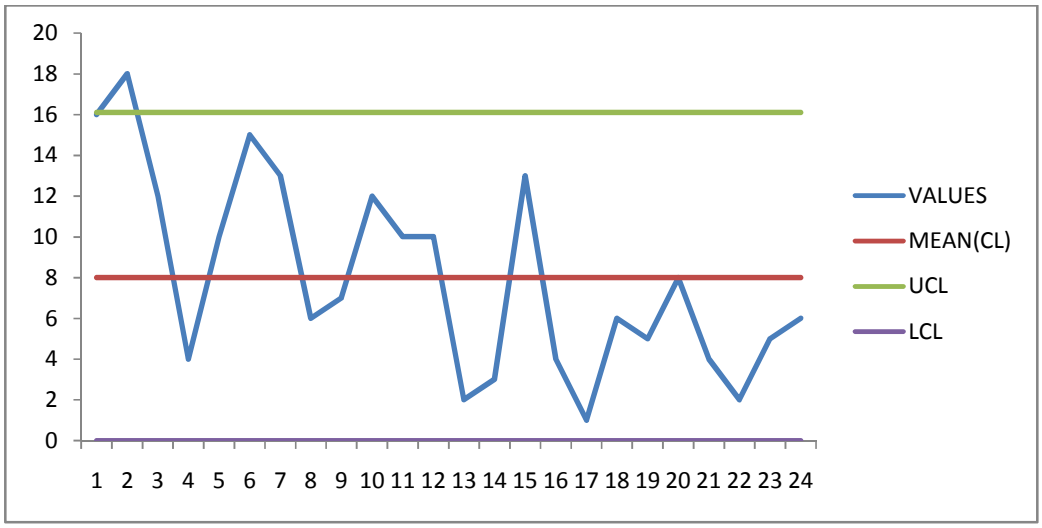
Find the control limits for number of defectives and comment on the state of control of the process.

mean of defectives= $\bar{d}=\frac{\varepsilon d}{n} = n\bar{p}$=CL = $\frac{192}{24}$. = 8

$\bar{p}=\frac{n\bar{p}}{n}$ = $\frac{8}{100}$ = .08 and therefore $\bar{q}$=1- $\bar{p}$ = ,92

UCL= $n\bar{p} + 3\sqrt{n\bar{p}\,\bar{q}}$ = 8 + $\sqrt[3]{100\,x\,.08\,x\,.92}$ = 16.1

LCL= $n\bar{p} - 3\sqrt{n\bar{p}\,\bar{q}}$. = 8 - $\sqrt[3]{100\,x\,.08\,x\,.92}$ = -.1 taken as 0. (.If LCL is less than zero, it is to be taken as zero.)



Ex. 15. 6

An inspection of 10 samples of size 100 each from 10 lots revealed the following number of defective units.

17, 15, 14, 6, 9, 4, 9, 17, 9, and 14.

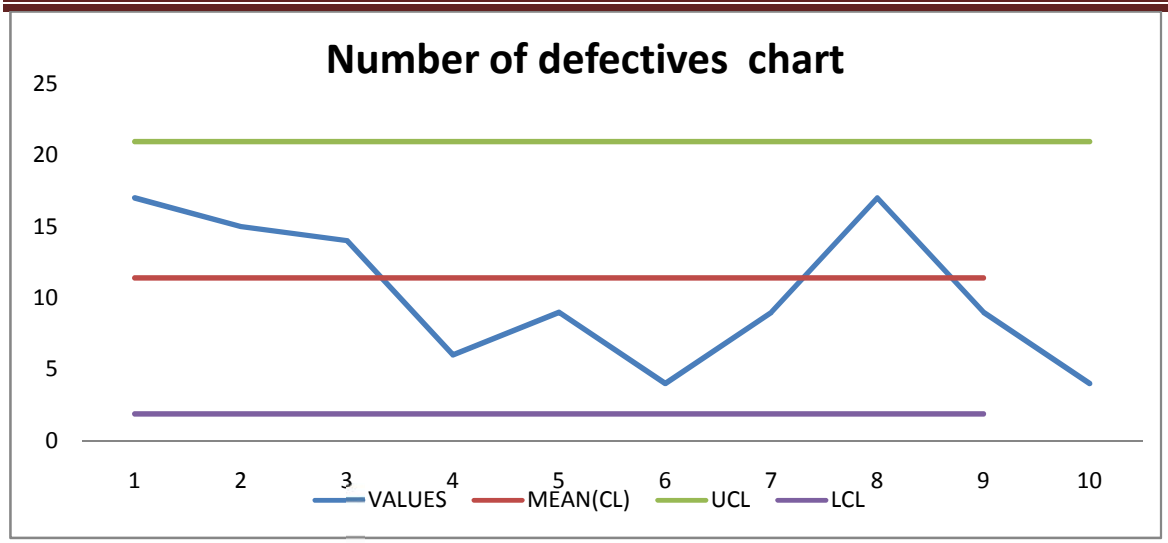
Calculate control limits for the number of defective units.

Mean of defectives= $\bar{d} = \frac{\varepsilon d}{n} = n\bar{p}$= CL = $\frac{114}{10}$. = 11.4

$\bar{p}=\frac{n\bar{p}}{n}$ = $\frac{11.4}{100}$ = .114 and therefore $\bar{q}$=1- $\bar{p}$ = . 886

UCL= $n\bar{p} + 3\sqrt{n\bar{p}\,\bar{q}}$ = 11.4 + $\sqrt[3]{100\,x\,.114\,x\,.886}$ = 11.4 + 9.54 = 20.94

LCL= $n\bar{p} - 3\sqrt{n\bar{p}\,\bar{q}}$. = 11.4 - $\sqrt[3]{100\,x\,.114\,x\,.886}$ = 11.4 - 9.54 = 1.86



No point falls beyond control limits, and therefore, the quality control of this firm is good.

Interpretation of or d or $n\bar{p}$ chart

D chart is based on number of defectives, instead of fraction defectives. For fraction defective chart, sample size may vary. It need not be uniform.

But in d chart sample size should be uniform. Number of defectives are plotted on the control chart, on the basis that, number of defectives were ascertained on the inspection of samples with equal number of items.

D chart reveal whether the number of defectives produced from the production process is within control or not. If one or more points fall beyond control limits, the process is said to be beyond the control. Necessary actions must be taken to control number of defectives, so as to fall with in UCL and LCL.

Number of Defects Chart (C chart)

C chart is constructed to control the number of defects per unit of the product. A defective article is one that in some way fails to conform to one or more given specifications. Every defective item may contain one or more defects. The C chart applies to the number of defects in sub groups of constant size. The C chart is used for the control of number of defects per unit.

Unlike p chart or d chart, which are applicable for defectives in samples, C chart applies to number of defects. Sample size for C chart may be single unit like glass, cloth, carpet etc. C chart focuses on examining the defects on products, rather than number of defectives.

Defectives and defects

Defectives and defects are very closely related topics, but they are different concepts. Defective items contain one or more defects. But all defects may not lead to defective items. A defect is an imperfection or lack of quality on a product. Even if there are certain defects on a product, still the product may not become defective. but when a product bears more than tolerable number of defects, such product may become defective. C chart focuses on number of defects.

For example, in car manufacturing industry, certain number of missing screws is defects, but due to a limited number of missing screws, a car need not be defective and discarded. By replenishing the missing screws, the car may be marketed.

Steps

1. Take appropriate scale and draw x, y axes.
2. Find mean of defects = $\frac{total\ defects}{total\ items}$ = $\bar{C}$ = CL
3. Ascertain Upper Control Limit = $\bar{C} + \sqrt[3]{\bar{C}}$
4. Find Lower Control Limit = $\bar{C} - \sqrt[3]{\bar{C}}$

Ex. 15.7

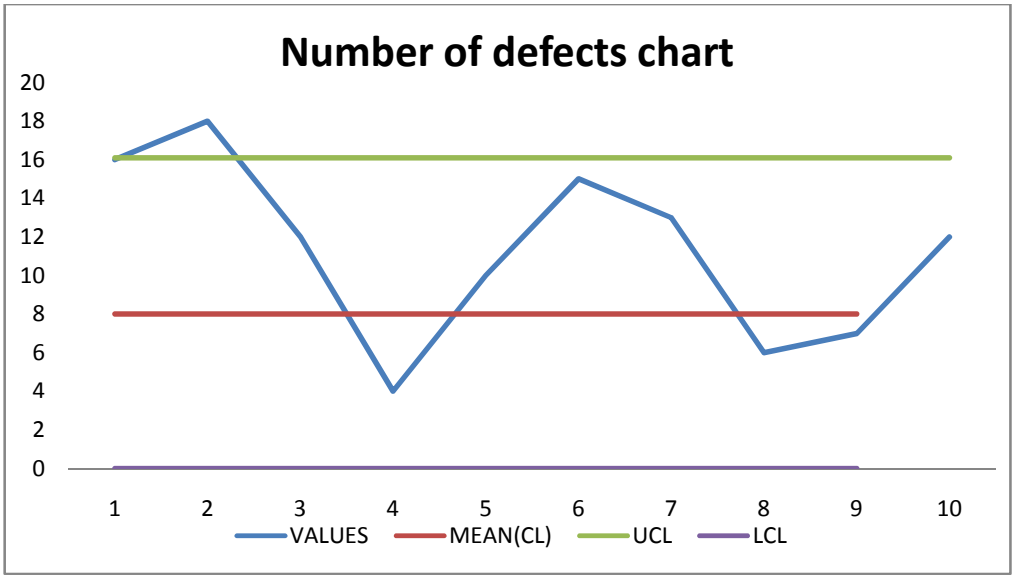
10 Pieces of cloth, out of different rolls, were inspected and following defects found. Draw control chart for number of defects and comment on the quality.

Defects : 1,3,5,0, 6, 0, 9, 4, 4, and 3

mean of defects = $\frac{total\ defects}{total\ items} = \frac{35}{10} = 3.5 = \bar{C}$ = CL

Upper Control Limit = $\bar{C} + \sqrt[3]{\bar{C}} = 3.5 + \sqrt[3]{3.5} = 9.11$

Lower Control Limit = $\bar{C} - \sqrt[3]{\bar{C}} = 3.5 - \sqrt[3]{3.5} = -2.11$, taken as zero



Since all defects values fall within control limits, the process is under quality control

Interpretation of C chart

In spite of wide applicability of Mean chart and Range chart, a number of practical situations exist in many industries where numbers of defects are to be focused in order to control quality of manufactured products.

Number of defects can be properly presented in control charts and quality of production processes can be assessed on this basis.

C chart is based on number of defects fund on inspection of unit of products such as cloth, carpet, paper, sheet, glass etc. they can be observed on inspection of sample units. When the number of defects exceeds the upper and lower control limits, the process is said to be beyond control, and needs corrective action.

Ready Reckenor for control charts

Chart	CL	UCL	LCL
Mean Chart	$\bar{\bar{x}}$	$\bar{\bar{x}} + A_2\bar{R}$	$\bar{\bar{x}} - A_2\bar{R}$
Range Chart	$\bar{R}$	$D_4\bar{R}$	$D_3\bar{R}$
P Chart	$\bar{p}$	$\bar{p} + 3\frac{\bar{R}}{d_2\sqrt{n}}$	$\bar{p} - 3\frac{\bar{R}}{d_2\sqrt{n}}$
Np Chart	$n\bar{p}$	$n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$	$n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$

Review questions & exercises

1. What is fraction defective chart
2. What is p chart
3. Explain control chart number of defectives
4. How is control chart for number of defectives constructed
5. State the steps in drawing number of defectives charts
6. How is quality assured using np charts
7. An inspection of 10 samples of size 20 each revealed following number of defective units.

   17 15 14 26 9 4 19 12 9 and 15

   Calculate control limits for the number of defective units. Plot the control limits and

   State whether the process is under control or not
8. 40 Samples of each 10 were inspected. The number of defective of each item is given below.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	15	17	18	19	20
0	1	0	3	9	2	0	7	0	1	1	0	0	3	1	0	0	2	1	0

   Construct Number of Defective Chart and establish quality control standard for future.
9. The following data refer to visual defects found during inspection of the first 10 samples of size 100 each. Use them to obtain upper and lower control limits for np chart.

Sample No. :	1	2	3	4	5	6	7	8	9	10
Defectives :	7	8	11	3	11	7	7	16	12	6

10. The table below gives the results of daily inspection of sewing machine needles for a particular quality characteristic. Compute the control limits and plot p-chart..

Number Inspected	Number of Defectives	Number Inspected	Number of Defectives
110	5	140	10
120	8	35	10
30	1	190	16
0	0	160	20
35	2	35	5
60	3	50	4
165	15	70	5
18	2		

11. The following table shows number of missing rivets observed at the time of inspection of 12 aircrafts. Find control limits for the number of defects and comment on the state of control.

Air Craft Number	1	2	3	4	5	6	7	8	9	10	11	12
missing rivets	7	15	13	18	10	14	13	10	20	11	22	15

12. A company produces sheet glass. The number of defects found on inspection of 12 sheets are given below. 2, 5, 4, 3, 7, 2, 4, 2, 1, 2, 4, 2, Construct appropriate control chart.
13. Draw a suitable control chart for the following data pertaining to the number of missing screws, found on inspection of 8 cars.

    7 12 3 20 21 5 4 3

14. During the examination of 10 equal length of cloth, the following imperfections were noted. draw c chart and comment on the quality.

    2 3 4 0 5 6 7 4 3 2

UNIT XVI

TOTAL QUALITY MANAGEMENT

Statistical quality control is very useful for continuous mass production processes such as spare parts industry, consumer goods, durable goods, capital goods, oil industry etc. however decision makers may feel that their businesses are so complex that, they cannot capture quality aspects through control charts alone. From this view point, control charts have limited scope that they focus on certain aspects only such as mean values, ranges, number of defectives, defects etc .Besides, service sector is one critical area where statistical quality control techniques cannot be applied because quality characteristic cannot be effectively represented through control charts. Total quality management emerged in such circumstances.

For example, take off delays in large airports is defective quality in aviation industry. Although delays are easy to identify, their causes are harder to locate. Take offs can be delayed by weather, equipment problems, late number of incoming crews, holiday traffic etc. How to measure and control delays is a tedious but critical task.

Total quality management is the appropriate technique for managing continuous and effective improvement of quality. In total quality management, quality management is performed in 3 stages – analyzing the causes leading to quality defects, assessing the contribution of each cause and its effect and thirdly making continuous quality improvement. Total quality management is integrated organizational approach in delighting customers by meeting their expectations on a continuous basis through everyone involved, for continuous improvement in all products, services and processes by proper problem solving methodology.

Total quality management involves assessment of quality process, zero defect management, and application of quantitative tools such as Fishbone Diagram, Paretto diagram, continuous quality improvement etc.

Quality process

Quality process begins by understanding of who the customer is, what his needs are, and how these needs can be satisfied. Thus, quality process comprises different steps through which complete satisfaction is provided to the customer.

In the early days of mass production, sorting out of defectives became the chief method of quality process. Quality inspectors tested goods at the end of a production process and released only some of them (good ones only) to the consumers. It was widely believed that cost of a few rejects did not amount to much, because the marginal cost of each unit was small. But by the late 1970s managers began to point out that cost of defectives were much higher than supposed. The large team of inspectors had to be paid, defectives had to be recycled, warranties costed much, and customers' good will was lost. These led to heavy costs and opportunity costs.

Arguments began to arise that it is simply cheaper to do things right the first time. They preached the concept of zero defects. Many industries and manufacturers practiced quality management to attain cent percent quality production. If we demand near perfect performance from power generating companies and airlines, we should expect no less from producers of goods and services. Preventing defects were related to workers' pride and mangers' efficiency. Everybody involved became conscious of total quality management.

Zero defect management

The basic approach behind zero defect management is to negate common belief that "to err is human". Instead it believes in perfection in management. Zero defect management refers to production system which results in cent percent quality production and zero defectives. Earlier the normal loss concept led to the belief that defectives could be upto 5% of total production. But total quality management cannot tolerate 5% or even 1% defectives. Global defectives standard is 1PPM (part per Millennium).

Adoption of total quality management techniques focuses on quality of inputs to each stage of operations, so as to be defective free. Manufacturers often have to accept raw materials and components from external suppliers. To ensure that results of their own operations are of perfect quality, they must often test inputs and ensure conformity to requirements and specifications.

As a part of zero defect management, reliance on sampling to ensure quality of inputs is replacing old time product inspection. Zero defect management is an effective way of motivating suppliers and workers to improve quality of their inputs.

Fishbone diagrams

How to identify and group causes of inferior quality? The question is difficult to answer. Total quality management approach to complex business operations began with the realization that all errors, defects and problems have causes, and that there is only a finite number of such causes.

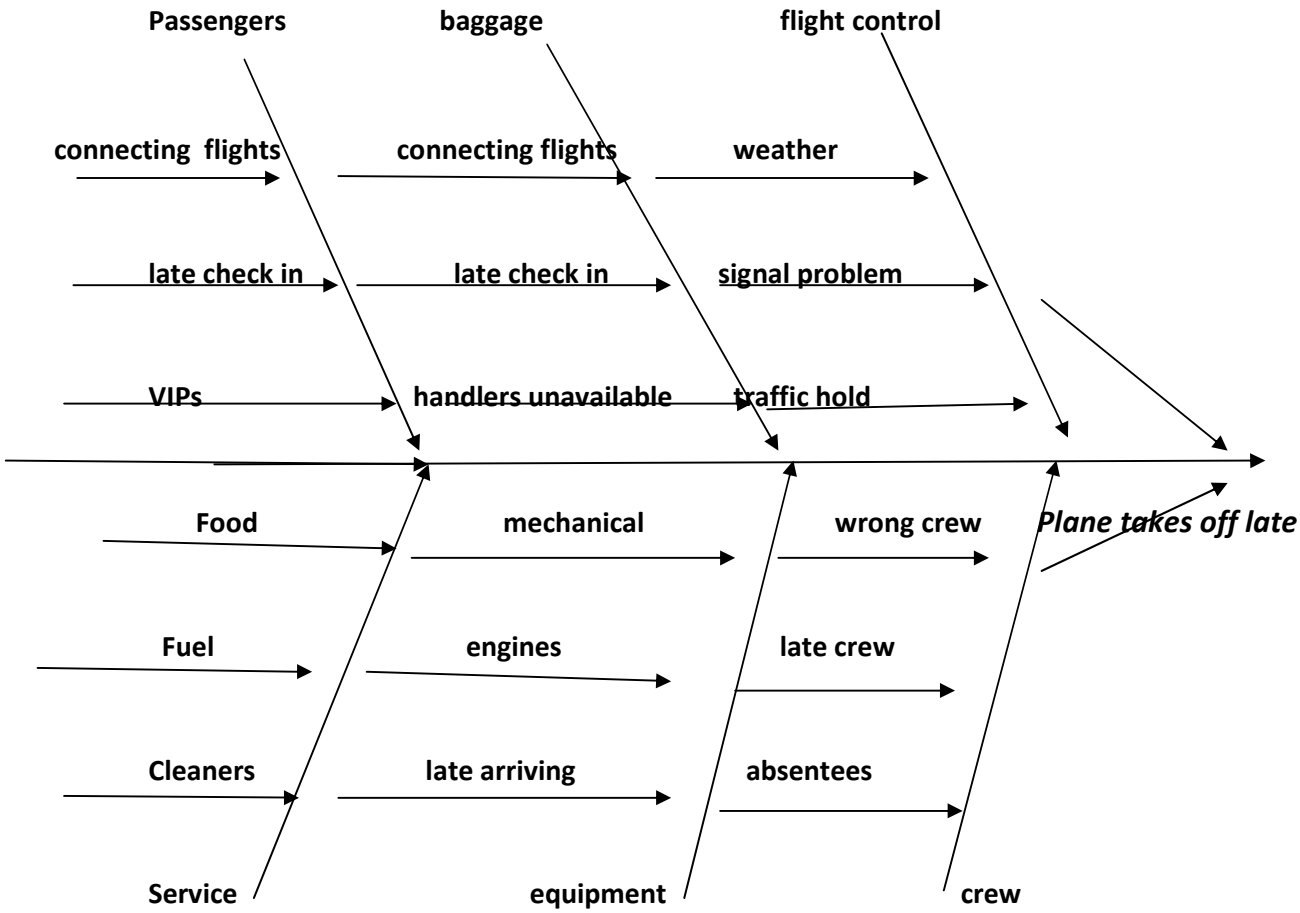
One crucial process in total quality management is to identify and discriminate between things gone right and things gone wrong. In our airport example, some of the planes do leave on time (things gone right) . When we observe late departures (things gone wrong), we can begin to build up a list of causes behind their delays.

Causes of problems can be gathered into logical groups. There will be various reasons for departure delays, and there will be cause and effect relationship among them. These relationships can be captured pictorially in a cause and effect diagram. Such diagrams are called Ishikawa diagrams or Fishbone diagrams. Fishbone diagrams takes an unstructured list of factors that contribute to delayed take-offs, and organizes that list in two major ways. First, it gathers the factors into logical groups. And then within the groups, it indicates how various factors feed into one another in cause and effect relationship.

Fishbone diagrams point out that employees at all levels must be involved in total quality management, to be successful. Baggage handlers are much more likely than top management to be able to identify a complete list of baggage problems that contribute to take off delays. Besides, they are also very likely to be able to suggest ways to improve baggage operations.

Thus, fishbone diagram is the prominent tool of analysis in total quality management. It sets out all the likely causes of departure from quality, along with their sources. Following is a Fishbone diagram presentation of causes and effects of relationships among various operations in a large airport. It identifies and presents things gone right and things gone wrong, which leads to

planes delayed plane takeoff. In this case , passengers, baggage, flight control,, services, equipment and crew play their parts.



FISHBONE DIAGRAM - AIRPORT TAKEOFF DELAYS

Slaying the dragons first

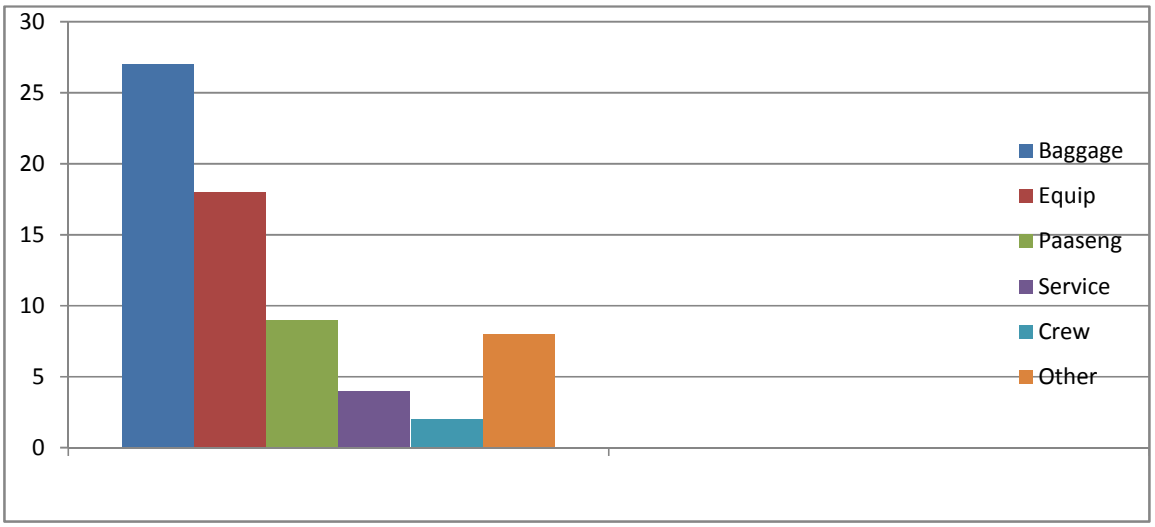
In any quality improvement process, there are likely to be a very large number of causes for defects and errors. Looking at all the possible things that can go wrong, even if they are organized into a neat fishbone diagram, can lead even well motivated people to despair at the complexities. The solution is to distinguish between vital few and trivial many. In our airport example, most of the delays are due to baggage handling, and only one delay a year is attributable to a freak hail storm. In total quality improvement, companies must slay the dragons first in working to improve the quality of their goods or service. A Parretto chart will assist in identifying major causes.

Parreto diagrams

Fishbone diagram reveals the source, cause and effect of each variation in quality. when such cause are revealed, they must be closely analyzed as to the contribution of each cause and its effect. The diagram enabling such an analysis is called Parreto diagram.

A parreto chart is a bar diagram showing groups of errors and their causes, arranged by the occurrence frequencies. It is constructed by simply counting data from observation. The results are ordered in a sequence from most common to least common. These charts are named after Vilfredo Parreto , an Italian economist.

Certain quality management studies revealed that 80 percent of defects and errors can be attributed to 20% of causes. The Parreto chart may show that majority of delays were caused by baggage handling. Total quality manager should begin improvements efforts by concentrating on this area.



PARETTO DIAGRAM - REASONS FOR AIRPORT DELAYS

Continuous quality improvement

Once the causes of errors and defects have been identified, the resources are denoted to making changes to improve the quality of goods and services. it requires re-configuration of system of production, and reallocation of resources.

As part of total quality management efforts, the leading cause will be given special attention, and scientific methods will be adopted for improvement. For baggage handling delays, scanners and computers can be installed for correct follow-up and handling.

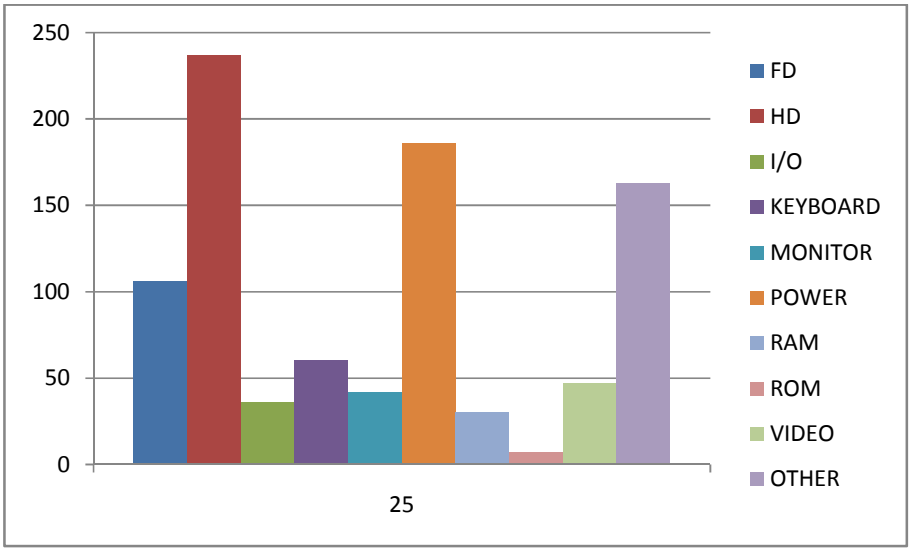
When total quality management efforts are successful, the leading causes of error is likely to drop to zero. This means that major cause will be revealed. And many become dragon. Management attention will now shift to the dragon, and attempts will be taken to slay the dragon. Thus constant attention for identification of problems, assessment of impact, and continuous improvement effects are commonly called total quality management.

Ex 16.1

Northway Computers has just begun a total quality management program to manage quality of personal computers and their manufacturing defects. A careful analysis of 25000 computer systems located following facts.

component	Number of faults
CPU	25
Floppy Discs	106
Hard Discs	237
I/O ports	36
Keyboard	60
Monitor	42
Power supply	186
RAM memory	30
ROM BIOS	7
Video adaptor	47
Others	163

Construct a Paretto chart for Northway computers, who are going to analyze the causes of faults, to identify major ones.



Producer's risk

Any person or firm that manufactures or assembles goods to be supplied to others is called a producer. Any sampling inspection plan of a lot has the disadvantage of rejecting a lot of satisfactory quality. The problem of rejecting a lot under sampling inspection plan is called producer's risk. thus producer's risk stands for the problem of rejecting a good lot.

Consumer's risk

Any person or form that buys and uses a product or services is called the consumer. the consumer has the risk of accepting a lot of unsatisfactory goods, on the basis of sampling inspection plan. the problem of accepting a lot which may be defective is called consumer's risk. Thus consumer's risk stands for the probability of accepting a poor lot.

1. Define total quality management
2. What are the causes of failures of control charts
3. What is the basis of total quality management
4. Explain reasons for variations in quality

5. What is the use of fishbone diagram
6. Why dragons must be slayed first
7. What is the use of Parreto diagram
8. What is AQL
9. What is acceptance sampling
10. What is the role of acceptance sampling in total quality management?
11. Draw a specimen fishbone diagram for a car manufacturing company
12. Best India computers have began total quality management program and located following facts on an analysis of 100 computers

Components	Faults
Cpu	5
Hdd	10
1/0 ports	10
Key-board	50
Power supply	20
Ram	5
Rom	12
Adaptor	5
others	6

Construct Parreto chart

13. What is AQL
14. Explain producer's risk
15. What is consumer's risk
16. Draw a Fishbone diagram from the data given relating to \Media One Print Media Press.

Problem	department	Occurrences
Omitted adv	Classified	18
Wrong insertion	Classified	32
Typo error	Reporting	15
Factual error	Reporting	8
Late finish	Printing	3
Print error	Printing	12
Incorrect date	Advertising	6
Fault message	Advertising	5
Wrong edit	Editing	4
Misquoting	Editing	2
Wrong delivery	Dispatch	3
No vehicle	dispatch	6

UNIT XVII

CORRELATION ANALYSIS

Managers make personal and professional decision that is based on prediction of future events. To make predictions, they rely on relationships between variables. Correlation helps us to identify nature of relationship between variables, and then to determine the degree of such relation.

There are situations where there is relation between two variables and statistical analysis is necessary to study such relation. for example, a manager may want to know whether there is any relation between amount spent on research and development and sales. in this case, after identifying whether there is any relation, he may also want to know how much is the relation, what is the type of relation, etc. The quantitative technique that can be used to study such relation is the correlation analysis.

Definitions

Correlation is defined as "tendency of two or more variables to vary together directly or inversely" (Boddington). He also states that "whenever definite connection exists between two or more variables, there is said to be correlation"

Bowley defined correlation as "when two quantities are so related that fluctuations in one are in sympathy with the fluctuations of the other, or that increase or decrease of one is found in connection with increase or decrease of the other, such quantities are said to be correlated"

According to M M Turtle, correlation is "an analysis of the association between two or more variables"

Thus correlation analysis is the quantitative tool used to describe the degree to which one variable is related to another variable.

Correlation and causation

The word correlation usually implies cause and effect relationship, that is, mutual interdependence. For example, a change in the price is the cause for a change in demand. That is, there exists causal connection between the two variables. Causal connection means cause and effect relationship. A positive correlation between the increase in cigarette smoking and the increase in cancer may prove that one causes the other. But correlation does not always imply cause and effect relationship.

When two variables are correlated there need not be cause and effect relationship. It is just possible that a high degree of correlation between the variables may be due to the same cause affecting each variable. For example, a high degree of correlation between the yield per acre of rice and tea may be due to the fact that both are related to the amount of rainfall. But none of the two variables is the cause of the other.

There may be high degree of correlation between the variables but it may be difficult to pi point as to which is the effect. For example, demand and supply, price and production, etc. it is a well known fact principle of Economics that as the price of a commodity increases its demand goes down and so change in price is the cause and change in demand is the effect. But it is also possible that increased demand of commodity is due to growth f population or other reasons.

Two series may show a high degree of correlation which may result purely from chance also. For example, during the last decade, there has been a significant increase in the sales of new papers and in the number of crimes. That does not mean that there exists any correlation between sales of new papers and crimes. But we can try to establish correlation between the two variables. Such illogical correlation is referred to as nonsensical or spurious correlation.

Types of correlation

According to the nature of relation between variables, correlation may by positive and negative, linear or non linear, or simple, partial or multiple correlations.

Positive or negative correlation

When the values of two variables move in the same direction, correlation is said to be positive. That is, an increase in the value of one variable results in an increase in the value of the other variable also. Similarly a decrease in the value of one variable results in a decrease in the other variable.

When the value of two variables move in opposite direction, so that an increase in the value of one variable results in a decrease in the value of the other variable or vice versa, correlation is said to be negative. Generally price and supply are positively correlated, and correlation between price and demand is said to be negative.

Linear or Non Linear correlation

When the amount of change in one variable leads to a constant ratio of change in the other variable, correlation is said to be linear. For example, if price goes up by 10%, and It leads to a rise in the supply by 15% each time, there is a linear relation between price and supply. When there is linear correlation, the points plotted on a graph will give a straight line

Correlation is said to be non linear when the amount of change in one variable is not in constant ratio to the change in the other variable. Here the ratio of change fluctuates and is never constant.

Simple, partial or multiple correlations

When there are only two variables, the correlation is said to be simple. For example, the correlation between price and demand is simple.

When one variable is related to a number of other variables, correlation is not simple. When there are three or more variables under study, at the same time, it may be multiple correlation or partial correlation.

In multiple correlations, we measure the degree of association between one variable on one side and all other variables together on the other side. The relation between yield with both rainfall and temperature is case of multiple correlations.

In partial correlation we study the relationship of one variable with one of the other variables presuming that the third or other variables remain constant. For example, we may study relation between yield and rainfall, keeping constant the effect of temperature.

Methods of studying correlation

Correlation between two variables can be measured by both graphic and algebraic method, scatter diagram and correlation graph are the two graphic methods, while coefficient of correlation is algebraic method for measuring correlation

Scatter diagram

In this method, one of the variables is shown on X axis and the other on the Y axis. Each pair of values is plotted on the graph by means of a dot mark. After, all the items are plotted, we get as many dots on the graph paper as the number of points.

The scatter diagram is a visual aid to show the presence or absence of correlation between two variables. A line of best fit can be drawn using the method of least square. This line will be as close to the points as possible.

Correlation graph

Under this method, separate curves are drawn for the X variable and Y variable on the same graph paper. The values of the variables are taken as ordinates of the points plotted. From the direction and closeness of the two curves we can infer whether the variables are related or not. If both the curves move in the same direction, upward or downward, correlation is said to be positive. If the curves move in the opposite direction, correlation is said to be negative.

Coefficient of correlation

Coefficient of correlation is an algebraic method of measuring correlation. Under this method, we measure correlation by finding a value known as the coefficient of correlation using an appropriate formula. Correlation coefficient is a numerical value, which shows the degree or extent of correlation between two variables.

Coefficient of correlation is a pure number lying between -1 and +1. When the correlation is negative, it lies between -1 and 0. When the correlation is zero, it indicates that there is no correlation between the variables. When the correlation coefficient is 1, there is perfect positive correlation. Between no correlation and perfect correlation, there are varying degrees of correlation.

Coefficient of correlation can be computed by applying Karl Pearson's formula or Spearman's formula. Spearman's formula is related to Rank correlation which is discussed later.

Karl Pearson's coefficient of correlation

This is also known as Pearsonian coefficient of correlation, and it is denoted by the symbol r. The formula for computing Pearsonian Coefficient of correlation is

$$r = \frac{n\sum XY - \sum X \times \sum y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

where X, Y are values of two series of observations and n is the number of pairs of observations.

Steps in finding Karl Person's r

1. consider the 2 series of values
2. squire the values and find $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$
3. obtain products of x and y and get $\sum x\, y$
4. ascertain $r = \frac{n\sum XY - \sum X \times \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$

In this formula, individual observations are directly taken, without taking deviations from their means. This is feasible when the values are small in size. Instead, when observations are large, deviations from original means of the two series can be taken. Then the formula is modified as below.:

$$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum X^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$ where x, y are deviations from original means.

The formula can still be modified and simplified as $$r = \frac{\sum xy}{\sqrt{\sum X^2}\sqrt{\sum y^2}}$$

If deviations are taken from assumed means, instead of original means, the formula will be as below;

$$r = \frac{n\sum dxdy - \sum dx \times \sum dy}{\sqrt{n\sum dX^2 - (\sum dx)^2}\sqrt{n\sum dy^2 - (\sum dy)^2}}$$ where dx, dy are deviations from assumed means.

Ex 17.1

Following are data on price and supply of a commodity. Find the correlation between price and supply and comment on the degree of correlation between them.

X	22	26	29	30	31	33	34	35
Y	19	21	22	29	27	24	27	31

X (30)	Y (25)	X^2	Y^2	xy
-8	-6	64	36	48
-4	-4	16	16	16
-1	-3	1	9	3
0	4	0	16	0
1	2	1	4	2
3	-1	9	1	-3
4	2	16	4	8
5	6	25	36	30
0	0	132	122	104

$$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum X^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{\sum xy}{\sqrt{\sum X^2}\sqrt{\sum y^2}} = \frac{104}{\sqrt{132}\sqrt{122}} = .82$$

There is high degree of positive correlation. That is when price increases supply also increases, to a great extant.

Ex 17.2

Calculate Karl Pearson's coefficient of correlation, from the following data on price and demand for a product.

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

X	Y	X^2	Y^2	XY
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
36	60	266	706	193

$$r = \frac{n\sum xy - \sum x \times \sum y}{\sqrt{n\sum X^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = r = \frac{6 \times 293 - 36 \times 60}{\sqrt{6 \times 266 - 36^2}\sqrt{6 \times 706 - 60^2}} \quad \textbf{-.92}$$

There is high degree of negative correlation. That is, when price increases, demand decreases, almost in the same manner.

Ex 17.3

Calculate correlation coefficient in the short cut method, ie, using assumed means.

X 70 92 80 74 65 83

Y 74 84 63 87 78 90

dx (80)	dy (84)	dx^2	dy^2	dxdy
-10	-10	100	100	100
12	0	144	0	0
0	21	0	441	
-6	3	36	9	0
-15	-6	225	36	18
3	6	9	36	90
-16	-28		100	18

$$r = \frac{n\sum dxdy - \sum dx \times \sum dy}{\sqrt{n\sum dX^2 - (\sum dx)^2}\sqrt{n\sum dy^2 - (\sum dy)^2}} \quad r = \frac{6 \times 190 - -16 \times -28}{\sqrt{6 \times 514 - 16^2}\sqrt{6 \times 622^2 - 28^2}} = \textbf{.2397}$$

There is very low degree of positive correlation. That is, when variable X increases, variable Y also increases, but only slightly.

Ex. 17.4

Find the coefficient of correlation between age and playing habit of students.

Age	14.5-15.5	15.5-16.5	16.5-17.5	17.5-18.5	18.5-19.5	19.5-20.5
Students	250	200	150	120	100	80
Players	200	150	90	48	30	12

X = Age $\qquad$ Y = $\frac{\text{Players}}{\text{Students}} = \frac{200}{250} \times 100 = 80$ .....

X	Y	dx (18)	dy (50)	dx^2	dy^2	dxdy
15	80	-3	30	9	900	-90
16	75	-2	25	4	625	-50
17	60	-1	10	1	100	-10
18	40	0	-10	0	100	0
19	30	1	-20	1	400	20
20	15	2	-35	4	1225	4
		-3	0	19	3350	-240

$$r = \frac{n\sum dxdy - \sum dx \times \sum dy}{\sqrt{n\sum dX^2 - (\sum dx)^2}\sqrt{n\sum dy^2 - (\sum dy)^2}} \quad r = \frac{6\,x - 240 - -3\,x\,-0}{\sqrt{6\,x\,19 - -3^{\,2}}\sqrt{6\,x3350 - 0^2}} = -.99$$

There is very high degree of negative correlation between age and playing habit. That is, as age increases, playing habit also increases, in almost perfect manner.

Interpretation of r

Karl Pearson's coefficient of correlation = r has become the most popular measure of correlation. It represents the degree or intensity of relation between two variables. The interpretation of r may take following forms:

1. when the value of r is +1, it represents perfect positive correlation, and when it is -1, it is negative correlation
2. when the value is < 0.3 it is low level of correlation
3. when the value is between 0.3 and 0.7, it shows moderate level of correlation
4. when the value is > 0.7, it is high degree of correlation

Properties of correlation coefficient

1. Value of r lies between -1 and +1
2. Value of r is a pure, and is independent of its individual observation
3. Karl Pearson's r will not change, due to a change of origin or scale
4. Karl Pearson's r between x and y will be the same as between y and x
5. It does not represent causal relation

Probable error

Probable error of the coefficient of correlation is a s statistical measure which measures reliability and dependability of the value of coefficient of correlation. If probable error is added to or subtracted from the coefficient of correlation , it would give two such limits within which we can reasonably expect the value of coefficient of correlation, to vary. Usually the coefficient of correlation is calculated from samples. For different samples drawn from the same population, the coefficient of correlation may vary. But the numerical value of such variation is expected to be less than the probable error.

The formula for finding probable error is PE = $\frac{.6745\,(1-r^2}{\sqrt{n}}$

Where r = coefficient of correlation n = number of pairs of observations

Interpretation of probable error

1. If coefficient of correlation, is less than probable error , it is not at all significant.
2. If coefficient of correlation is more than six times its probable error it is significant
3. If the probable error is not much and if the coefficient of correlation is .5 or more it is generally considered to be significant.

Coefficient of determination

Coefficient of correlation indicates the nature and extent of relation between 2 variables. Now a measure is required to explain the strength of relation and strength of variation. It is coefficient of determination.

Coefficient of determination gives the percentage variation in the dependent variable in relation with independent variable.

An effective way of interpreting strength of correlation is coefficient of determination. It is derived by r^2, and is the ratio between explained variation and total variation. The more is this value of determination, the better is the strength of correlation.

Coefficient of determination = $r^2 = \frac{EXPLAINED\ VARIANCE}{TOTAL\ VARIANCE}$

Similarly, coefficient of non-determination can be ascertained which is the complement of coefficient of determination. It shows the ratio between unexplained variation to total variation.

Coefficient of non determinant = $K^2 = 1 - r^2 = \frac{UNEXPLAINED\ VARIATION}{TOTAL\ VARIATION}$

Lag and lead correlation

Usually, two variables may be correlated and they signify cause and effect relationship. This does not mean that change in one variable will immediately cause a change in the other variable. There can be some time lag or difference of time between cause and effect. For instance, we suppose that a rise in price leads to rise in supply of goods. If price changes today, the supply may not change immediately. It may take five or six months to adjust itself to the changes in prices. Such a difference in the period of change is called time lag..

When two series are given, one may be leading and the other may be lagging. In other words, forward moving series are known as the leading series and following or later moving series are known as lagging series. Thus we can estimate lag or lead by plotting the data on a graph.

Review Questions and Exercises

1. Define term correlation
2. Does correlation is significantly cause affect with relationship?
3. Why is study of correlation important?
4. What are the uses of correlation?
5. Explain the various methods of studying correlation?
6. Write notes on Scatter Diagram
7. What is Scattered Diagram? Form the Scatter Diagram and how do you infer the nature of relationship of variables?
8. Explain term Coefficient of Correlation
9. Distinguish between Positive and Negative Correlation
10. How do you interpret the sign of coefficient of correlation?
11. Explain Linear and Non-Linear correlation
12. Enumerate the different types of correlation
13. What is simple, multiple and partial correlation?
14. What you meant by perfect correlation?
15. Even a high degree of correlation does not mean that a relationship of cause and effect exists between variables. Discuss
16. Write notes on Karl Pearson's Correlation?
17. What are the properties of correlation coefficient?
18. What are the properties of Karl Pearson's correlation coefficient?
19. How do you measure correlation?
20. What you mean by probable error?
21. What you mean by Coefficient of determination?
22. For the data given below obtain the correlation coefficient between the average price and demand of a particular commodity in a region

Average Price	11	19	15	13	17
Demand (Kgs)	30	18	24	29	24

23. The data given below relates to the price and demand of a commodity over a period. Compute the correlation of coefficient between price and demands

Price in RSs	80	75	60	90	70
Demand in KGs	12	15	13	9	4

There is any correlation between x and y?

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

24. Compute the coefficient of correlation between heights and weights of 10 persons and command

Heights (Inches)	62	72	78	58	65	70	66	63	60	72
Weights (Kgs)	50	65	63	50	54	60	61	55	54	65

25. Find Karl Pearson's coefficient of correlation between x and y from following data giving test scores of 10 candidates in mathematics and statistics and interpret

Scores in Mathematics :	98	70	40	20	85	75	95	80	10	5
Scores in Statistics :	85	65	32	30	80	60	61	55	54	65

26. Find the coefficient of correlation from the following data give below

X:	12	20	15	18	33	24	30	12	15	22
Y:	30	35	28	36	29	39	30	25	30	38

27. Find the coefficient of correlation between sales and expenses of following 10 firms (Figures omitting 000s)

Firms	:	1	2	3	4	5	6	7	8	9	10
Sales	:	50	50	55	60	65	65	65	60	60	50
Expenses :		11	13	14	16	16	15	15	14	14	13

28. Compute the coefficient of correlation for the following data and interpret it

Year	1925	1926	2927	1928	1929	1930	1931	1932	1933	1934
Labors	368	384	385	361	347	384	395	503	400	385
bales	22	20	21	24	20	22	26	26	29	28

29. Find Karl Pearson's coefficient of correlation between two variables of x and y from given below. Also find probable error and interpret

X:	78	89	96	69	59	79	68	61
Y:	121	137	156	112	107	136	223	108

30. Find coefficient of correlation between sales and price from following data. Also find Probable Error (P. E)

Price (RS)	100	90	85	92	90	84	88	90	93	95
Sales (Units)	600	610	700	630	760	800	800	750	700	680

31. Establish correlation between the following pair of series and find out probable error and also interpret your findings

X:	17	19	20	22	24	27	29	30	33	35
Y:	87	85	80	78	75	72	70	65	62	60

UNIT XVIII

RANK, PARTIAL AND MULTIPLE CORRELATION

Karl Pearson's coefficient of correlation is obtained when variables are measured in quantitative form. Usually, only two variables are involved and the resultant coefficients of correlation are called simple correlations. However, there are situations, where proper quantitative measurements are not available, or several variables – three or more –have to be simultaneously considered. Rank correlation takes care of situations where ranks of observations are given. When there are more than two variables involved, partial or multiple correlations can be calculated.

Rank Correlation

Correlation is a measure of relation between two variables and observations are obtained in ratio level of measurement. However, there are situations when data is available in ordinal scale or interval scale, or when data is irregular. Rank correlation is a method of ascertaining relation between two series of observations, especially when data is not in the form of variables, or irregular or when they are ranked. Rank correlation is introduced by Prof. Charles spearman, a British psychologist. He focused on rank differences. Rank correlation is useful in developing with qualitative characteristics such as beauty, intelligence, effectiveness, morality, etc.

The formula for spearman's rank correlation coefficient is

r = 1 -- $\frac{6\sum D^2}{N^3-N}$ where, D = difference between ranks N = number of pairs

Steps in rank correlation

1. Consider the given sets of observations
2. If ranks are not available, rank the data.
3. Find difference between ranks and square them = D^2
4. Summate the squares of differences and get $\sum D^2$
5. Find spearman's rank correlation **R = 1 -** $\frac{6\sum D^2}{N^3-N}$

Uses

1. Rank correlation coefficient R is used to measure the degree of association between two attributes
2. It is used in cases where exact or reliable measurement are not available like beauty, attitude, intelligence, etc.
3. It gives a quick estimate of the degree of association between two attributes, avoiding heavy calculation of Karl Pearson's coefficient of correlation

Ex. 18.1.the ranking 10 beauty contestants by two judges X and Y are given below. Calculate Spearman's Rank correlation coefficient.

Contestant	A	b	C	D	E	F	G	H	I	J
Judge A	1	6	3	9	5	2	7	10	8	4
Judge B	6	8	3	2	7	10	5	9	4	1

Judge A	Judge B	Rank diff D	D^2
1	6	5	25
6	8	2	4
3	3	0	0
9	2	7	49
5	7	2	4
2	10	8	64
7	5	2	4
10	9	1	1
8	4	4	16
4	1	3	9
		Total	176

Spearman's rank correlation R = **1 -** $\frac{6\sum D^2}{N^3-N}$ = 1 - $\frac{6 \times 176}{10 \times 10 \times 10 - 10}$ = -0.07

Ex 18.2

Data on intelligence and income are given. Find Rank correlation coefficient

Intelligence	17	13	15	16	6	11	14	9	7	12
Income		36	46	35	24	12	18	27	22	2	8

X value	X rank	Y value	Y rank	D	D^2
17	1	36	2	1	1
13	5	46	1	4	16
15	3	35	3	0	0
16	2	24	5	3	9
6	10	12	8	2	4
11	7	18	7	0	0
14	4	27	4	0	0
9	8	22	6	2	4
7	9	2	10	1	1
12	6	8	9	3	9

Spearman's rank correlation **R = 1 -** $\frac{6\sum D^2}{N^3-N}$ **= 1 -** $\frac{6 \times 44}{10 \times 10 \times 10 - 10}$ **= 0.733**

Partial and Multiple correlation

When two variables move together, we say they are correlated. Thus, two variables are correlated if the change in one variable results in a corresponding change in the other variable. For example, when price of a commodity changes, the demand of that commodity also changes. So both variables move together or they move in sympathy. Hence price and demand are correlated.

In the study of relation between variables, if there are only two variables, the correlation is said to be simple. For example, correlation between height and weight is simple. But when one variable is related to a number of variables, simultaneously, the correlation is not simple. For example, yield of crop is related to rainfall and temperature , at the same time. In this example, correlation is not simple. The study of relationship between three or more variables, is done with the help of partial or multiple correlations.

The principle advantage of partial and multiple correlation is that it allows us to use more than two variables for studying relations between them. Sometimes, the study between two variables only may be insufficient to determine a reliable estimating equation. Partial and multiple correlation enable an analysis of relation between 3 or more variables.

Partial Correlation

When there are more than two variables and if we want to study relationship between only two of them, hen we have partial correlation. So in partial correlation, we consider only two variables and others are treated as normal or having no effect and so ignored.

For example, consider three variables – yield of wheat, rainfall and temperature. Here the correlation between yield, and rainfall ,treating temperature as normal, is partial correlation. In partial correlation, two variables are studied as to relationl, keeping constant the impact of a third variable.

Partial Correlation Coefficient

Partial correlation coefficient measures the relationship between one variable, and one of the other variables , assuming that the effect of the rest of the variables is eliminated. Here there will be three or variables, but correlation is analyzed between any two of them, keeping constant the effect of other variables.

Let x_1 x_2 and x_3 three variables, given, $r_{12.3}$ is the partial correlation coefficient between x_1 and x_2 treating x_3 as constant or normal. Similarly we can have $r_{13..2}$ and $r_{23..1}$. These partial correlation coefficients can be computed using the given simple correlation coefficients, as shown below.

$$r12.3, = \frac{r12 - r13 \, x \, r23}{\sqrt{1 - r13^2}\sqrt{1 - r23^2}}$$

$$r23.1 = \frac{r23 - r21 \, x \, r31}{\sqrt{1 - r21^2}\sqrt{1 - r31^2}}$$

$$r31.2 = \frac{r31 - r32 \, x \, r12}{\sqrt{1 - r32^2}\sqrt{1 - r12^2}}$$

Where $r\,12,\ r23,\ r13$ *are simple coefficient correlations.*

Ex. 18.3

If r12 = 0.7, r13 = 0.61, r23 = 0.4, find **r12.3, r23.1, r13.2.**

$$r12.3, = \frac{r\,12 - r13\,x\,r\,23}{\sqrt{1-r\,13^2}\sqrt{1-r23^2}} = \frac{.7-.61\,x\,.4}{\sqrt{1-.61^2}\sqrt{1-.4^2}} = \mathbf{.63}$$

$$r23.1 = \frac{r\,23 - r21\,x\,r31}{\sqrt{1-r\,21^2}\sqrt{1-r\,31^2}} = \frac{r\,.4 - .7\,x\,.61}{\sqrt{1-.7^2}\sqrt{1-.61^2}} = \mathbf{-.048}$$

$$r31.2 = \frac{r31 - r32\,x\,r12}{\sqrt{1-r\,32^2}\sqrt{1-r\,12^2}} = \frac{.61-.7\,x\,.4}{\sqrt{1-.7^2}\sqrt{1-.4^2}} = \mathbf{.50}$$

Properties of partial correlation

1. it measures the relation between any 2 variables, keeping constant the effect of the third variable
2. the value of partial correlation may come between -1 and +1
3. it study the net correlation between a dependent variable and one of the two independent variables

Multiple correlation

When there are three or more variables, and we want to study relation of one variable with all the other variables taken together, the correlation obtained is called multiple correlation. For eg. if the variables are yield, rainfall and temperature, and we want to study the relation of yield, with both rainfall and temperature taken together, we find the multiple correlation. So, in multiple correlation, one variable is studied or taken on one side and all other variables together on other side.

In multiple correlation, we have at least three variables. Of these, 2 variables may influence a third variable. The combined impact of two variables on the third variable can be studied in multiple correlation. To calculate multiple correlation, three simple correlation coefficients must be given. these multiple correlation coefficient can be calculate as below :

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2}}$$

$$r_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{13}^2}}$$

$$r_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{12}^2}}$$

Where r12, r13, r23 are simple correlation coefficients.

Ex.18.4

Given following simple correlation coefficients : $r_{12=0.93}$, $r_{13=0.89}$, $r_{23=0.96}$. *Calculate* $r_{1.23}$, and $r_{2.13}$

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{r_{23}^2}} = \sqrt{\frac{.93^2 + .89^2 - .93 \times .89 \times .96}{0.96^2}} = \mathbf{0.92}$$

$$r_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{13}^2}} = \sqrt{\frac{r_{93}^2 + .96^2 - .93\,x\,.89\,x\,.96}{1-.89^2}} = \mathbf{0.84}$$

EX: 18.5

If r12=0.98, r13=0.44, r23=0.54, find r1.23, r2.13, r 3.12.

ANS:

$$R1.23=\sqrt{\frac{r^2 12+r^2 13-2r12r13r23}{1-r^2 23}}=\sqrt{\frac{(.98)^2+(.44)^2-2(.98)(.44)(.54)}{1-(.54)^2}}$$

$$=\sqrt{\frac{.9604+.1936-.4657}{.7084}}=\sqrt{.972}=\underline{.986}$$

$$R2.13=\sqrt{\frac{r^2 12+r^2 13-2r12r13r23}{1-r^2 13}}=\sqrt{\frac{(.98)^2+(.54)^2-2(.98)(.44)(.54)}{1-(.44)^2}}$$

$$=\sqrt{\frac{.9604+.2916-.4657}{.8064}}=\sqrt{.9751}=\underline{.9875}$$

$$R3.12=\sqrt{\frac{r^2 12+r^2 13-2r12r13r23}{1-r^2 12}}=\sqrt{\frac{(.44)^2+(.54)^2-2(.98)(.44)(.54)}{1-(.98)^2}}$$

$$=\sqrt{\frac{.1936+.2916-.4657}{.0396}}=\sqrt{.4924}=\underline{0.7017}$$

EX: 18.6

If x_1, x_2, and x_3, are three variables measured from their means with N=10, $\varepsilon x_1{}^2$=90, $\varepsilon x_2{}^2$ =160, $\varepsilon x_3{}^2$ =40, $\varepsilon x_1 x_2$ =60, $\varepsilon x_2 x_3$ =60, $\varepsilon x_1 x_3$ =40, calculate the partial correlation coefficient $r_{31.2}$,and multiple correlation of coefficient r1.23.

Ans: $r_{12.}=\frac{\varepsilon x1x2}{\sqrt{\varepsilon x_1{}^2}\sqrt{\varepsilon x_2{}^2}}=\frac{60}{\sqrt{90}\sqrt{160}}$ =.5

$r_{13.}=\frac{\varepsilon x1x3}{\sqrt{\varepsilon x_1{}^2}\sqrt{\varepsilon x_3{}^2}}=\frac{40}{\sqrt{90}\sqrt{40}}$ = .67

$r_{23.}=\frac{\varepsilon x2x3}{\sqrt{\varepsilon x_2{}^2}\sqrt{\varepsilon x_3{}^2}}=\frac{60}{\sqrt{160}\sqrt{40}}$ = .75

$r_{31.2}=\frac{r_{13}-r_{23}r_{12}}{\sqrt{1-r_{23}{}^2}\sqrt{1-r_{12}{}^2}}=\frac{67-(.75*.5)}{\sqrt{1-.5625}\sqrt{1-.25}}=\frac{.295}{.573}$ = .515

$$r1.23=\sqrt{\frac{r^2 12+r^2 13-2r12r13r23}{1-r^2 23}}=\sqrt{\frac{.25+.4489-2*.5*.67*.75}{1-.5625}}=67$$

Properties of multiple correlation coefficients

1. multiple correlation studies the combined impact of two or more variables upon a single variable
2. multiple correlation enables calculation of expected values of a variable, on the basis of two other variables, influencing the former one
3. the value of multiple correlation will range between 1 and 1
4. if the value of multiple correlation is 1, the impact of two independent variables on dependent variable is perfect
5. If the value of multiple correlation is 0, the dependents variable is completely uncorrelated with all other variables

Limitation of correlation analysis

i. In correlation, it is assumed that there is a straight line relation between variables. A small value of r , therefore indicates poor correlation. but really the relation may be strong not indicated by r
ii. A high degree of correlation may be misunderstood as if there is cause and effect relation. But actually such close association may not exists.
iii. It may happen that two variables show a high degree of correlation, but there is no logical basis. (eg. no. of baby births and no. of lectures)
iv. if the ata id not reasonable homogeneous, coefficient of correlation gives a misleading picture. data may be in dusters, but coefficient of correlation may be very high.

Dummy variables in regression analysis

In regression analysis, the dependent variable is generally metric in nature, and it is most influence by other metric variables. However, there could be situations where dependents variables may be influence by qualitative variables like gender, marital status, religion, color, etc. For example,. demand for cosmetics is not only influenced by price of cosmetics and consumer's income but also by gender of respondents. Therefore, its inclusion in the regression model as one of the regressors is required.

But how to quantify a qualitative variable like gender? In such a situation, dummy variable comes to our rescue. They are used to quantify qualitative data.

Review Questions and Exercises

1. What is simple correlation
2. Explain partial correlation
3. What is multiple correlation
4. Distinguish between partial and multiple correlation
5. Compare simple with multiple correlation
6. what is the use of partial and multiple correlation
7. Explain and distinguish between simple, partial and multiple correlation
8. Explain partial Correlation. What information does it seek to convey?
9. What are multiple regression? Write down the multiple regression equation of x_1 on x_2 and x_3
10. What are residuals?
11. On the basis of observations made on 30 cotton plants, the simple correlation of yield of cotton (X_1) and, number of bales (X_2), and height (X_3) are found to be r1 = .8, r13 = .65 and r3 = .7. Compute the partial coefficients for $r_{1.13}$, $r_{13.2}$ and $r_{23.1}$
12. r_{12} = .98, r_{13} = .44, r_{31} = .54. Find values of $r_{1.23}$, $r_{2.13}$, $r_{3.12}$
13. If r_{12} = 0.6; r_{23} = r_{31} = .8. Find values of $R_{1.23}$, $R_{2.13}$, $R_{3.12}$
14. If r12 = 0.80; r13 = -.40; r23 = -0.56, and find $r_{13.2}$.Show that $R^2_{1.23} = r^2_{12} + r^2_{13}$ if $r_{23} = 0$
15. Calculate the value of $r_{1.23}$ when
    i. a – r_{12} = r_{13} = 0. b – r_{12} = r_{23} = r_{13} = r
16. If r_{12} = .7, r_{23} = r_{31} = 5. σ_1 = 2, σ_2 = σ_3 = 3. Find equation of plane of regression for x_1 on x_2 and x_3
17. In a trivariate distribution, $\bar{x}_1$ = 53, $\bar{x}_2$ = 52, $\bar{x}_3$ = 51. σ_1 = 3.88, σ_2 = 2.97, σ_3 = 2.68. r_{23} = .8, r_{31} = .81, r_{12} =.78. Find Linear regression equation of x_1 on x_2 and x_3.
18. σ_1 = 3.88, σ_2 = 2.97, σ_3 = 2.86. r_{12} = 0.78, r_{23} = r_{31} = -0.8. Find $b_{12.3}$
19. Find equation of plane of regression of x_1 on x_2 and x_3. If, σ_1 = 3,σ_2 = 2.97, σ_3 = 2.86.r_{12} = .5, r_{23} = .7, r_{13} = .2. Find $\sigma_{12.3}$
20. If σ1 = 3, r_{12} = .3, r_{13} = .4, r_{23} = .5. Find $\sigma_{1.23}$
21. In a trivariate distribution, r_{12} = .7, r_{23} = r_{31} = .6.σ_1 = 3, σ_2 = σ_3 = 5.Find (a) $r_{12.3}$, (b) $R_{1.23}$

UNIT XIX

REGRESSION ANALYSIS

The term regression was first used by sir Francis Galton, who made a study and proved that height of children born to tall parents will tend to move back or regress towards the mean height of population. He designated the word regression to the name of the process of predicting one variable form the other variable. He coined the term multiple regression to describe the process by which several variables are used to predict another. Therefore, when there is a well-established relation between variables, it is possible to make use of the relation in making the estimates and to forecast the value of one variable on the basis of the other variable.

A banker, for example, could predict deposits, on the basis of per capita income (pci) in the trading area of bank. A marketing manager may plan his advertising expenditure on the basis of expected effect on total sales revenue of the change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such prediction may be made by using regression analysis.

An investigator may employ regression analysis to test his theory having cause and effect relationships. All these explain that regression is an extremely useful tool specially in problems of business and industry, involving prediction and forecasting.

Definitions

Literally regression means 'going back'. Certain variables are found to move back and conform to progression of a related variable. In regression variables, reduce variances between them.

Regression is defined as "*statistical device used to study the relationship between 2 or more variables that are related.*"

According to M.M. Blair, regression is "*mathematical measure of average relationship between two or more variables in terms of original units of data.*"

Regression analysis in the general sense is "*estimation or prediction of unknown value of one variable, from known value of another variable.*"

Assumptions

While making use of regression technique, for making predictions, it is always assumed that:

1. There are 2 variables – one is dependent and another is independent and there is a relation between them.
2. values of dependent variables are random, and independent variable is fixed
3. There is a clear indication of direction of relationship. That is dependent variable is a function of independent variable.
4. The analysis is made to predict values of dependent variable on the basis of independent variable.

Types of regression

Regression can be classified as according to the number of variables involved and proportion of changes in the variables. Accordingly there can be simple and multiple regression, or linear and non linear regression.

Simple and multiple regressions

On the basis of number of variables, regression can be classified into simple and multiple regressions. When there are only two variables, the regression equation obtained is called simple regression equation. In multiple regression analysis, there are more than two variables and we try to find out the effect of two or more independent variables on one dependent variable. Let X, Y, and Z be three variables and let X and Y be the independent variables and Z be the dependent on them, then we use multiple regression analysis to study the relative movement of Z , for a unit movement in X and Y.

For example, if there are three variables yield, rainfall and temperature, then suppose yield is dependent on rainfall and temperature , then we get the regression equation of Z on X and Y where Z is yield, X rainfall and Y temperature.

A multiple regression equation is an equation for estimating the value of a depending variable say Z, from values of the independent variables X and Y and it is called a regression equation of Z on X and Y.

Linear and non Linear Regression

On the basis of proportion of changes in the variables, the regression can be classified into Linear and Non Linear regressions. If the given bivriate data are plotted on a graph, the points so obtained on the scatter diagram will more or less concentrate around a curve called curve of regression.

If the regression curve is a straight line, we say that there is linear regression between the variables under study. The equation of such a curve is the first degree equation in the variables X and Y. mathematically, the relation between X and Y in a linear regression, can be expressed in the form $Y = a + bX$. In linear regression, the change in the dependent variable is proportionate to the changes in the independent variable.

If the curve of regression is not a straight line, then the regression is termed as curved or non-linear regression. The regression equation in such cases is not of first degree. In this case the dependent variable does not change by a constant amount of change in the independent variable.

Methods of studying regression

In regression, the objective is to predict the value of one variable on the basis of another variable, using the average relation. This is done through following methods.

Scatter diagram method

Scatter diagram is a diagram representing two series with independent variable on the X axis and dependent variable on the Y axis. The scatter diagram by itself is not sufficient for predicting value of dependent variable. Some formal expression of relationship between the two variables is necessary for prediction purposes. For this purpose, one may simply draw a straight line through the points in the scatter diagram, with the help of which one can predict values.

It is a diagram with two series, with independent variable on the x axis and dependent variable on the y axis. The scatter diagram by itself is not sufficient for predicting values of dependent variable. Some formal expression of relationships between the two variances is necessary for prediction purposes. For this purpose, one may simply draw a straight-line through the point in the scatter diagram, with the help of which are can predict values.

Method of Least Squares

This method involves drawing regression line of best fit, by applying the principle of least squares. The line of regression should be drawn in such a manner that sum of squares of differences between values of dependent variables and independent variables should be least. This line of least squares can be used for prediction.

Regression Equations

For each regression line, there will be a regression equation. Regression equation is a mathematical relation between the dependent variable and independent variable. There are two regression equations - Y on X and X on Y.

Y on X = a + bX

To solve the equation and get values of a and b

$y - \bar{y} = byx(x - \bar{x})$

where, x , y are question values and $\bar{x}, \bar{y}$ = means of 2 series

byx = $\frac{n\sum XY - \sum X \times \sum Y}{n\sum X^2 - (\sum X)^2}$, where X , Y are observations directly taken.

Steps in regression

1. Consider the given values
2. Find X^2, Y^2, and XY , where values are in original form.
3. Instead of original values, deviation from actual means or assumed means may be taken.
4. Find regression coefficients using the formula byx = $\frac{n\sum XY - \sum X \times \sum Y}{n\sum X^2 - (\sum X)^2}$
5. solve the formula using regression coefficient $y - \bar{y} = byx(x - \bar{x})$
6. obtain the value of y = a + bX

Ex:19.1

From the following data of values of x and y, find the regression equation of y on x

x :	2	3	3	5	6
y :	3	5	4	8	9

Ans:

2	3	6	4
3	5	15	9
4	4	16	16
5	8	40	25
6	9	54	36
20	29	131	90

The equation of the regression line y on x is = $y - \bar{y} = b_{yx}(x - \bar{x})$

where $b_{yx} = \frac{n\sum XY - \sum X \times \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5\times131 - (20\times29)}{5\times90 - (20)^2} = \frac{655-580}{450-400} = \frac{75}{50} = 1.5$

$\bar{x} = \frac{\sum X}{n} = \frac{20}{5} = 4$; $\bar{y} = \frac{\sum Y}{n} = \frac{29}{5} = 5.8$

the equation is $y - 5.8 = 1.5\ (x-4)$

$y - 5.8 = 1.5x - 6$

$y = 1.5x - 6 + 5.8$

$\underline{y = 1.5x - .2}$

Similarly, x on y → x = a = by

to solve this equation, $x - \bar{x} = byx\ (y - \bar{y})$

where, $\bar{x}, \bar{y}$ = means of 2 variables and x , y are question values

$$bxy = \frac{n\sum XY - \sum X \times \sum Y}{n\sum Y^2 - (\sum Y)^2}$$

$$x \text{ on } y = x - \bar{x} = b_{xy}\ (y - \bar{y})$$

EX 19. 2:

From the following data of age of husband and age of wife, form the two regression equations and calculate (i) – Husband's age when wife's age is 16

(ii) – Wife's age when husband's age is 40.

Husband's Age	36	23	27	28	28	29	30	31	33	35
Wife's Age	29	18	20	22	27	21	29	27	29	28

Ans:

36	29	1044	1296	841
23	18	414	529	324
27	20	540	729	400
28	22	616	784	484
28	27	756	784	729
29	21	609	841	441
30	29	870	900	841
31	27	837	961	729
33	29	957	1089	841
35	28	980	1225	784
300	250	7623	9138	6414

$\bar{x} = \frac{\sum X}{n} = \frac{300}{10} = 30$; $\bar{y} = \frac{\sum Y}{n} = \frac{250}{10} = 25$

$byx = \frac{n\sum XY - \sum X \times \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{10\times7632 - (300\times250)}{10\times9138 - (300)^2} = \frac{76230 - 75000}{91380 - 90000} = \frac{1230}{1380} = 0.89$

$bxy = \frac{n\sum XY - \sum X \times \sum Y}{n\sum Y^2 - (\sum Y)^2} = \frac{10\times7623 - (300\times250)}{10\times6414 - (250)^2} = \frac{76230 - 75000}{64140 - 62500} = \frac{1230}{1640} = 0.75$

Equation of the line of regression of Y on X

$(y - \bar{y}) = b_{yx}\ (x - \bar{x})$ y - 25 = .089 (x-30)

y = 0.89x – 26.7 + 25

y = .89x – 1.7

Equation of line of regression on Y on X

$x - \bar{x} = byx\ (y - \bar{y})$

x – 30 = 0.75 (y – 25) x = 0.75y – 18.75 + 30

x = 0.75y + 11.25

(i) When wife's age (y) is 16, the husband's age (x) is obtained by putting y = 16 in the equation X on Y.

x = 0.75 (16) + 11.25 = 12 + 11.25 = 23.25

Thus, husband's age is 23.25

(ii) When husband's (x) age is 40, wife's age (y) is obtained by putting x = 40 in the equation Y on X.

y = .89 (40) – 1.7 = 35.6 – 1.7 = 33.9

Thus wife's age is 33.9

EX 19 .3

A panel of 2 judges, P and Q graded by 7 dramatic performances by independently awarding marks as follows:

Performance	1	2	3	4	5	6	7
Marks by P	46	42	44	40	43	41	45
Marks by Q	40	38	36	35	39	37	41

The eighth performance, for which judge could not attend, was awarded 37 marks by judge P. if judge Q has been also present, how many marks would be expected to have been awarded by him to eighth performance?

Ans:

Let x stand for marks by P and y for marks by Q

ks by x	and y y	d_x=mark 42 x —	d_y= −39	d_x^2	d_y^2	d_xd_y
46	40	4	1	16	1	4
42	38	0	-1	0	1	0
44	36	2	-3	4	9	-6
40	35	-2	-4	4	16	8
43	39	1	0	1	0	0
41	37	-1	-2	1	4	2
45	41	3	2	9	4	6
301	266	7	-7	35	35	14

Here, only regression equation is required, i.e., Y on X

$$\bar{x} = \frac{\sum X}{n} = \frac{301}{7} = 43 \; ; \; \bar{y} = \frac{\sum Y}{n} = \frac{266}{7} = 38$$

$$B_{yx} = \frac{n\sum d_x d_y - (\sum d_x \times \sum dy)}{n\sum d_x^2 - (\sum d_x)^2} = \frac{7\times14-(7\times-7)}{7\times35-7^2} = \frac{98-(-49)}{245-49} = \frac{98+49}{196} = \frac{147}{196} = 0.75$$

The equation of line of regression of Y on X is

$$y - \bar{y} = byx(x - \bar{x})$$

$$y - 38 = .75\,(x - 43) = 0.75x - 32.25$$

$$\therefore y = 0.75x - 32.25 + 38$$

$$y = 0.75x + 5.75$$

When x = 37, the estimate of y is given by

$$y = (0.75\times37) + 5.75 = 27.75 + 5.75 = 33.50$$

Hence, if the judge Q where present at eighth performance, he would have assigned the score 33.50

When deviations taken from actual mean $b_{yx} = \frac{\sum xy}{\sum x^2}$ and $b_{xy} = \frac{\sum xy}{\sum y^2}$

Relation between Correlation and Regression

In correlation, we study the presence and strength of relation between variables and in regression we predict the value of one variable on the basis of another variable. Both are closely related. Using the relation between their coefficients, following equations are obtained.

$byx = r\frac{\sigma y}{\sigma x}$

$bxy = r\frac{\sigma x}{\sigma y}$

Where byx and bxy = given regression coefficients.

r = correlation between x and y

$\sigma x, \sigma y$ = Standard deviations of x, y variables.

EX 19.4:

You are given the following data:	x	y
Arithmetic Mean	36	85
Standard Deviation	11	8

Correlation coefficient between x and y = 0.66

(a)-Find regression equation

(b)-Estimate value of x when y = 75.

Ans:

Given $\bar{x} = 36$; $\bar{y} = 85$ | $\sigma_x = 11$; $\sigma_y = 8$ | r = 0.66

$byx = r\frac{\sigma y}{\sigma x} = \frac{0.66 \times 8}{11} = \frac{5.28}{11} = 0.48$

$bxy = r\frac{\sigma x}{\sigma y} = \frac{0.66 \times 11}{8} = \frac{7.26}{8} = 0.91$

Equations of regression of Y on X

$y - \bar{y} = byx(x - \bar{x})$

$\therefore y - 85 = 0.48\ (x\text{-}36) = .048x - 17.28$

$\therefore$ <u>$y = 0.48x + 67.72$</u>

Equations of regression of X on Y

$x - \bar{x} = byx\ (y - \bar{y})$ $\therefore$

$x - 36 = 0.91\ (y - 85) = 0.91y - 77.35$ $\therefore$

<u>$x = 0.91y - 41.35$</u> To estimate the value of x when value of y = 75, we use regression equation x on y. $x = 0.91$ (75) – 41.35 = 68.25 – 41.35 = <u>26.9</u>

EX 19 5

From the following result, estimate the yield of crops when the rainfall is 22 cm and rainfall when yield is 600 kgs.	Yields in KG (Y)	Rainfall in CMs (X)
Arithmetic Mean	508.4	26.7
Standard Deviation	36.8	4.6

Coefficient correlation between yield and rainfall is 0.52

Ans:

Given $\bar{x} = 26.7$; $\bar{y} = 508.4$ | | $\sigma_x = 4.6$; $\sigma_y = 36.8$ | $r = 0.52$

Equations of regression of Y on X

$$y - \bar{y} = byx(x - \bar{x})$$

$byx = r\frac{\sigma y}{\sigma x} = \frac{0.52 \times 36.8}{4.6} = \frac{19.14}{4.6} = 4.16$ $\quad$ y - 508.4 = 4.16 (x – 26.7) = 4.16x – 111.072

y = 4.16x + 397.328

The most likely yield for crop (y) when rainfall (x) = 22 CM is given by,

y = (4.16 × 22) + 397.328 = 91.52 + 397.328 = 488.848 KGs.

Equations of regression of X on Y

$$x - \bar{x} = byx\,(y - \bar{y})$$

$bxy = r\frac{\sigma x}{\sigma y} = \frac{0.52 \times 4.6}{36.8} = \frac{2.39}{36.8} = 0.065$ $\quad$ x - 26.7 = 0.065 (y − 508.4) = 0.065y – 33.046 + 26.7

x = 0.065y – 6.364

The most likely rainfall (x) corresponding to the yield (y)= 600KGs is given by, x = 0.065 × 600 – 6.346 = 32.654 CMs.

Coefficient of correlation can be estimated on the basis of regression coefficients.

$$r = \sqrt{byx \times bxy}$$

EX 19.6

Find r if $b_{yx} = -0.2$; $b_{xy} = -0.7$

Ans:

$r = \sqrt{byx \times bxy}$ $\quad r = \sqrt{-.2 \times -.7} = \sqrt{.14} =$ -.37

EX 19 . 7:

$b_{yx} = .83$; $\sigma_x = 10$; $\sigma_y = 12$. Find r

Ans:

$b_{yx} = r\frac{\sigma y}{\sigma x}$

$.83 = r\frac{12}{10}$

$.83 = r \times 1.2$

$\therefore r = \frac{.83}{1.2} = .7$

Ex:19.8 comment on the following result.

For a bivariate distribution,

1. Coefficient of regression of y on x is 4.2 and coefficient of regression of x on y is 0.50
2. bxy= -.82 and byx =.25

Ans: 1. $r^2 = byx \times bxy = 4.2 \times 0.5 = 2.10$

But r square cannot be greater than 1. Hence the given statement is wrong.

2 byx and bxy can not have opposite signs. So the statement is wrong.

Uses of Regression

The study of regression is very useful in business, economics and researches.

1. Regression helps to obtain most probable values of one variable for given values of other variable.
2. It helps to study the effect of price on supply or demand of a commodity.
3. It is widely applied in physical science where the relation is functional.
4. It is used to describe the nature of relation between 2 or more variable.
5. It reveals rate of change in one variable based on change in other variable.

Properties of Regression Coefficient

1. The sign of both regression and coefficient will be the same. That is , both will be positive or both will be negative.
2. Product of regression coefficient will be squire of correlation coefficient.
3. b_{yx} and b_{xy} will have same squire of correlation coefficient.
4. $\frac{byx}{r} = \frac{\sigma y}{\sigma x}$ and $\frac{bxy}{r} = \frac{\sigma x}{\sigma y}$
5. Both b_{yx} and b_{xy} cannot be greater than one

Distinguish Between Correlation and Regression

Correlation	***Regression***
Several degrees of relationship	Several nature of relationship
No dependent or independent variables	One variable is dependent other independent
Does not enable prediction	Does enable prediction
Means relation between two variable	Means … to average value
Need not imply cause and effect relationship	Indicates cause and effect relationship between variables
Relative measure	Absolute measure

Multiple Regressions

Regression can be simple or multiple according to the number of variables involved. In simple regression, there are only two variables, and we are concerned with establishing linear relationship between them. It ignores the possibility of variations in the dependent variable being explained in terms of 2 or more independent variables.

In multiple regression, 3 or more variables are involved and we study the average relation between one dependent variable and two or more independent variables.

For example – The price of a product depends on both demand and supply for it. Here, if the average relation between price on the one hand, and supply and demand on the other hand, is calculated, it can be used to estimate volume of price, as a combined impact of supply and demand.

Multiple regression equation of X_1 on X_2 X_3 = X_1 = $b_{12.3}$ X_2 + $b_{13..2}$ X_3

Multiple regression equation of X_2 on X_1 X_3 = X_2 = $b_{21.3}$ X_1 + $b_{23.1}$ X_3

Multiple regression equation of X_3 on X_1 X_2 = X_3 = $b_{31.2}$ X_1 + $b_{32.1}$ X_3

Where $b_{12.3} = \frac{\sigma 1}{\sigma 2}\left\{\frac{r12 - r13\, r23}{1 - r23^2}\right\}$ and $b_{13.2} = \frac{\sigma 1}{\sigma 3}\left\{\frac{r13 - r12\, r23}{1 - r23^2}\right\}$

$b_{21.3} = \frac{\sigma 2}{\sigma 1}\left\{\frac{r12 - r13\, r23}{1 - r13^2}\right\}$ $b_{23.1} = \frac{\sigma 2}{\sigma 3}\left\{\frac{r23 - r12\, r13}{1 - r13^2}\right\}$

$b_{31.23} = \frac{\sigma 3}{\sigma 1}\left\{\frac{r13 - r12\, r23}{1 - r12^2}\right\}$ $b_{32.1} = \frac{\sigma 3}{\sigma 2}\left\{\frac{r23 - r13\, r12}{1 - r12^2}\right\}$

Dummy variables in regression analysis

In regression analysis, the dependent variable is generally metric in nature, and it is most influence by other metric variables. However, there could be situations where dependent variables may be influenced by qualitative variables like gender, marital status, religion, colour,etc

For eg. demand for cosmetics is not only influence by price of cosmetics and consumers' income but also by gender of respondents. Therefore, its inclusion in the regression model as the regressor is required.

But how to quantify an qualitative variable like gender, in such a situation, dummy variable comes to our rescue. They are used to quantify qualitative data.

Unit 19

1. Explain concept of regression
2. Explain utility of regression analysis
3. Distinguish between Correlation and Regression
4. Distinguish between linear and non-linear regression
5. What are two regression lines?
6. Write notes on regression lines
7. Where will regression line meet?
8. What are regression coefficient?
9. What are regression coefficients?
10. What do you by standard error of estimate?
11. What are the limitations of regressions?
12. Find regression equation of x on y

X:	2	4	6	8	10
Y:	5	7	9	8	11

13. From the data give below, obtain regression equation x on y

X:	2	3	7	8	10
Y:	10	9	11	8	12

14. The following give the monthly income and expenditure on 10 families. Calculate the linear regression of expenditure of food (y) on income (x)

Income	120	90	83	150	130	140	110	95	75	105
Expenditure	40	36	40	45	40	44	45	38	50	35

15. The price index number of wheat (x) and cereals (y) at twelve successive seasons (quarters) are given below.
    (a) – Fit a line of regression of X on Y
    (b) – Suggest wheat value of Y will be when X is expected to be 110?

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

16. The following table shows the number of motor registration in a certain territory for the same period. Find the regression equation to estimate the sale of motor tyres when motor registration is known. Estimate sale of tyres when registration is 850

Year	Motor Registration	No. of Tyres Sold
1	600	1250
2	630	1100
3	720	1300
4	750	1350
5	800	1500

17. The following data shows the Maximum and Minimum temperature on a certain day at 10 important cities throughout India.

Maximum Temperature :	29	23	25	15	27	29	24	31	32	35
Minimum Temperature :	8	3	7	5	8	19	10	7	5	8

(a) Fit regression lines for X on Y and Y on X
(b) Estimate maximum temperature when minimum temperature is 12
(c) Estimate minimum temperature when maximum temperature is 40

18. On the basis of following data, estimate
(a) Sales for advertising expenditure for RS 90 Lakh
(b) Advertising expenditure for sales target of RS 25 Crore

Sales (RS Crore) :	10	11	13	15	16	19	14
Adv. Expenditure (Lakhs):	60	62	65	70	73	75	71

19. The following calculations have been made for the closing price 12 stocks (x) on the Bombay stock exchange on a certain day along with the volume of thousands of shares (y). From these calculations, find regression equations
$\sum x$= 580, $\sum y$ = 370, $\sum xy$ = 11494, $\sum x^2$= 41685, $\sum y^2$ = 17062
20. For 10 observations on price (x) and supply (y) the following data were obtained (in appropriate units).
21. $\sum x$ = 130, $\sum y$ = 220,$\sum x^2$= 2288, $\sum y^2$ = 5506, , $\sum xy$ = 3467. Obtain line of regression on Y on X and estimate supply when price is 16 units
22. For a bivariate data, following constraints are known. Mean of x values = 700, Mean of y values = 1300. Sum of the products of deviations form mean from x and y values are = 41500. Sum of the squired deviations of x values from mean = 27800, Sum of squired deviations of squired deviations of y values from mean = 85000.Find regression equation.
[Hint: Take $\sum d_x = 0, \sum y_x = 0$]
23. The following table gives the relative values of two variables. From these values find,
(a) Regression equation&(b) Correlation coefficient

X:	42	44	58	55	89	98	66
Y:	56	49	53	58	65	76	58

24. From the data give below find:
(a) Two regression equation
(b) Correlation coefficient between marks in Economics and Statistics
(c) The most likely marks in statistics when mark in Economics is 30.

Marks in Economics	25	28	35	32	31	36	29	38	38	32	30
Marks in Statistics	43	46	49	41	36	32	31	30	33	39	

UNIT XX

SOFTWARE FOR QUANTITATIVE ANALYSIS

Managers have to their advantage a wide array of statistical programs to assist them in both data management and data analysis. In this unit we will briefly describe only the most frequently used statistical analysis packages.

MS EXCEL

The simplest and most widely used method of presenting and tabulating is MS Excel. Basic mathematical functions can be calculated here. Secondly, the software is easy to understand and used by most computer users. The data entered on Excel can be transported to most of the statistical packages, for a wider level analysis.

MINITAB

Minitab Inc. was developed more than 25 years ago at the Pennsylvania state university. It can be used both each … and effectiveness in all business areas. It was originally used by statisticians and professionals. However, today it is used for multiple applications – especially quality control six sigma and design of experiments. The URL for Minitab is *http://www.minitab.com*. A manager or researcher can utilize the products and help the research guide to undertake quantitative research analysis.

System for Statistical Analysis (SAS)

SAS was created in the late 1960s at North Carolina State University. It has been actively and extensively used in managing, storing and analyzing information. It has the advantage of being able to manage really bulky data sets which considerable ease. Linear models like Regression, ANOVA, ANOCOVA and all standard techniques for statistical analysis are enabled with SAS.

SPSS

Amongst the student community as well as with most professional managers and researchers, this is the most widely used package. It is adaptable to most business problems and is extremely user friendly. The URL for SPSS is *"http://www-01.ibm.com/software/analytics/spss/".* This unit contains application guidelines for SPSS borrowers.

Besides theses, there are a number of statistical soft wares like LISREL (Linear Structural Relation), LINDO etc…

SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others. The original SPSS manual (Nie, Bent & Hull, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis.[4] In addition to statistical analysis, and data management..

Statistics included in the base software:

- Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics
- Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate,partial, Nonparametric tests
- Prediction for numerical outcomes: Linear regression
- Prediction for identifying groups: Factor analysis, cluster analysis (two-step, K-means, hierarchical), Discriminant

The many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL *command syntax language*. Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses. Additionally, some complex applications can only be programmed in syntax and are not accessible through the menu structure. The pull-down menu interface also generates command syntax: this can be displayed in the output, although the default settings have to be changed to make the syntax visible to the user. They can also be pasted into a syntax file using the "paste" button present in each menu. Programs can be run interactively or unattended, using the supplied Production Job Facility.

The graphical user interface has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS Statistics window. The 'Data View' shows a spread sheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. The 'Variable View' displays the metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax. This may be sufficient for small datasets. Larger datasets such as statistical surveys are more often created in data entry software, or entered during computer-assisted personal interviewing, by scanning and using optical character recognition and optical mark recognition software, or by direct capture from online questionnaires. These datasets are then read into SPSS.

SPSS Statistics can read and write data from ASCII text files (including hierarchical files), other statistics packages, spreadsheets and databases. SPSS Statistics can read and write to external relational database tables via ODBC and SQL.

SPSS for Windows has the same general look a feel of most other programmes for Windows. Virtually anything statistic that you wish to perform can be accomplished in combination with pointing and clicking on the menus and various interactive dialog boxes. You may have noted that the examples in the Howell textbook are performed/analyzed via code. That is, SPSS, like many other packages, can be accessed by programming short scripts, instead of pointing and clicking. We will not cover any programming in this tutorial.

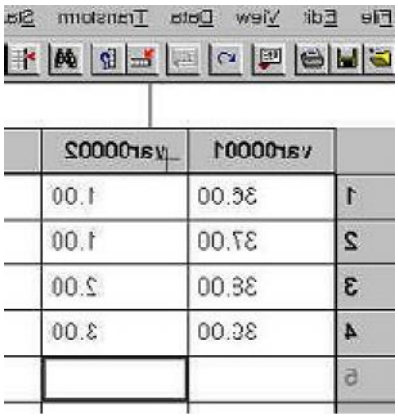
SPSS offers a large number of possible formats, including their own. A list of the available formats can be viewed and selected by clicking on the **Save as type:** , on the **Save As** dialog box. If your intention is to only work in SPSS, then there may be some benefit to saving in the *SPSS(*.sav)* format. I assume that this format allows for faster reading and writing of the data file. However, if your data will be analyzed and looked by other packages (e.g., a spreadsheet), it would be advisable to save in a more universal format (e.g., Excel(*.xls), 1-2-3 Rel 3.0 (*.wk3).

Once the type of file has been selected, enter a filename, minus the extension (e.g., sav, xls). You should also save the file in a meaningful directory, on your hard drive or floppy. That is, for any given project a separate directory should be created. You don't want your data to get mixed-up.

Data Entry

To begin the process of adding data, just click on the first cell that is located in the upper left corner of the datasheet. It's just like a spreadsheet. You can enter your data as shown. Enter each data point then hit [Enter]. Once you're done with one column of data you can click on the first cell of the next column.

These data are taken from table studying reaction time of eyes. The first column represents



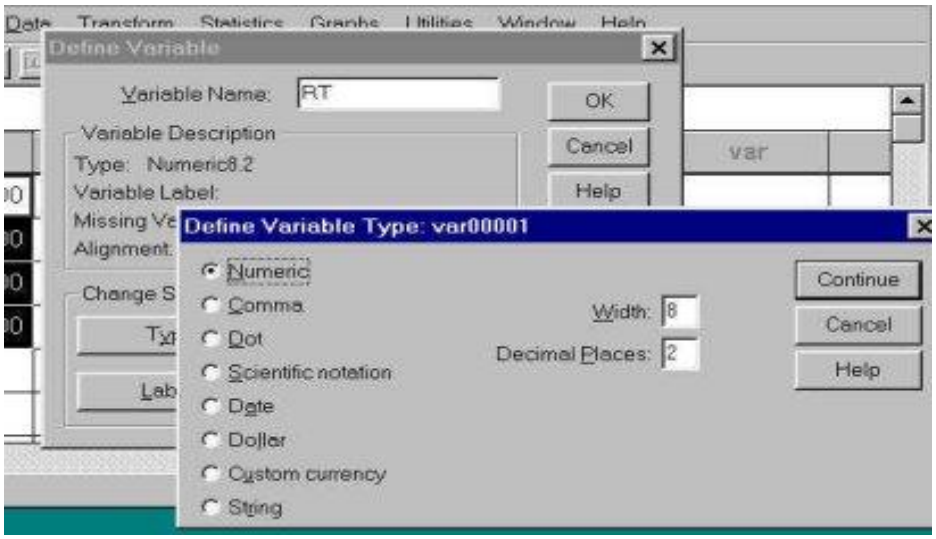
"Reaction Time in 100ths of a second" and the second

Column indicates "Frequency".

``If you're entering data for the first time, like the above example, the variable names will be automatically generated (e.g., var00001, var00002,....). They are not very informative. To change these names, click on the variable name button. For example, double click on the "var00001" button. Once you have done that, a dialog box will appear. The simplest option is to change the name to something meaningful. For instance, replace "var00001" in the textbox with "RT" (see figure below).

In addition to changing the variable name one can make changes specific to **[Type], [Labels], [Missing Values],** and **[Column Format].**

- **[Type]** One can specify whether the data are in numeric or string format, in addition to a few more formats. The default is numeric format.



Labels

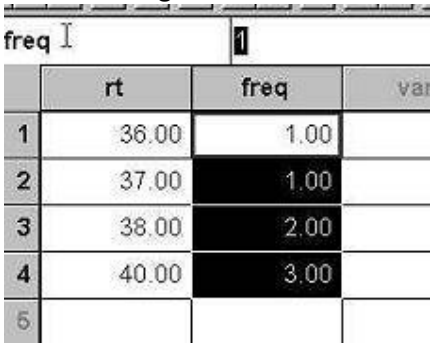
Using the labels option can enhance the readability of the output. A **variable name** is limited to a length of 8 characters, however, by using a **variable label** the length can be as much as 256 characters. This provides the ability to have very descriptive labels that will appear at the output.

Often, there is a need to code categorical variables in numeric format. For example, **male** and **female** can be coded as **1** and **2**, respectively. To reduce confusion, it is recommended that one uses value labels . For the example of gender coding, **Value:**1 would have a corresponding **Value label:** male. Similarly, **Value:**2 would be coded with **Value Label:** female. (click on the **[Labels]** button to verify the above)

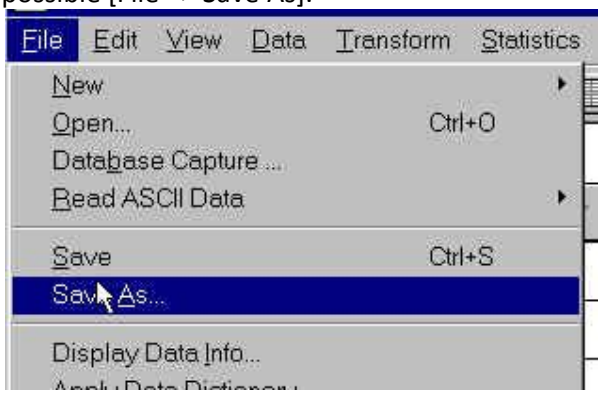
[Missing Values] See the accompanying help. This option provides a means to code for various types of missing values.

[Column Format] The column format dialog provides control over several features of each column (e.g., width of column).

The next image reflects the variable name change.

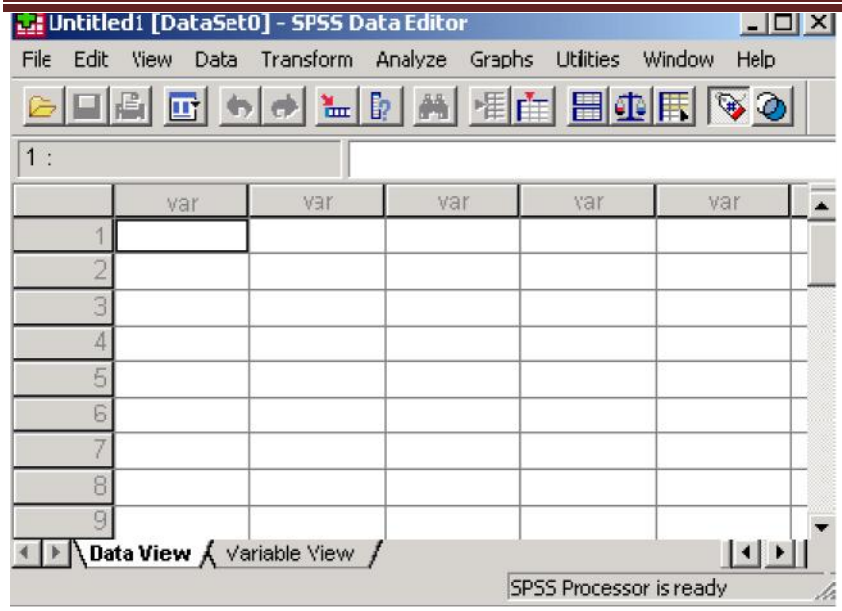


Once data has been entered or modified, it is advisable to save. In fact, save as often as possible [File => Save As].



Data View

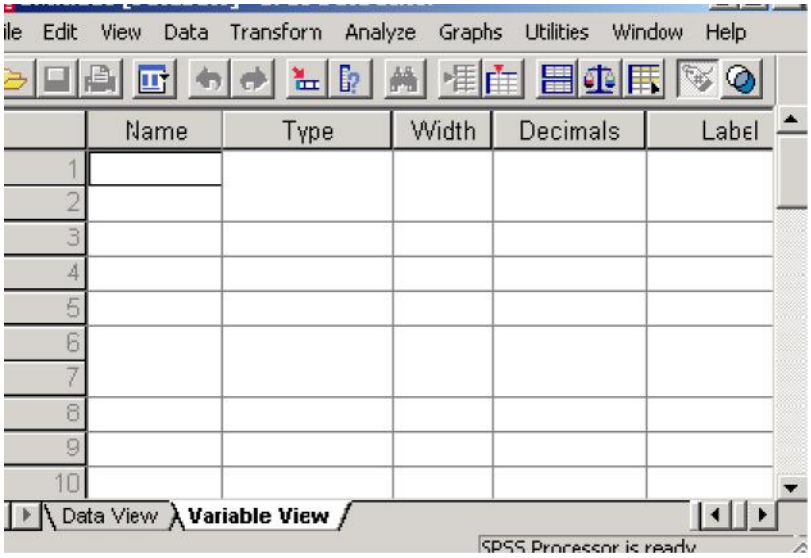
The **Data View** window is simply a grid with rows and columns. The rows represent subjects (cases or observations) and columns represent variables whose names should appear at the top of the columns. In the grid, the intersection between a row and a column is known as a cell. A cell will therefore contain the score of a particular subject (or case) on one particular variable. This window displays the contents of data file. You create new data files or modify existing ones in this window. This window opens automatically when you start an SPSS session. See Figure 1 for a brief annotation of this window.



The **Variable View** window is also a simple grid with rows and columns. This window contains descriptions of the attributes of each variable that make up your data set. Window, rows are variables and columns are variable attributes. You can make changes to variable attributes in this window such as add, delete and modify attributes of variables. There are eleven columns altogether namely: **Name**, **Type**, **Width**, **Decimal**, **Label**, **Value**, **Missing**, **Columns**, **Align**, **Measure** and **Role**. See Fig. 2 for more information. As you define variables in this window, they are displayed in the **Data View** window. The number of rows in the **Variable view** window corresponds to the number of columns in the **Data view** window.

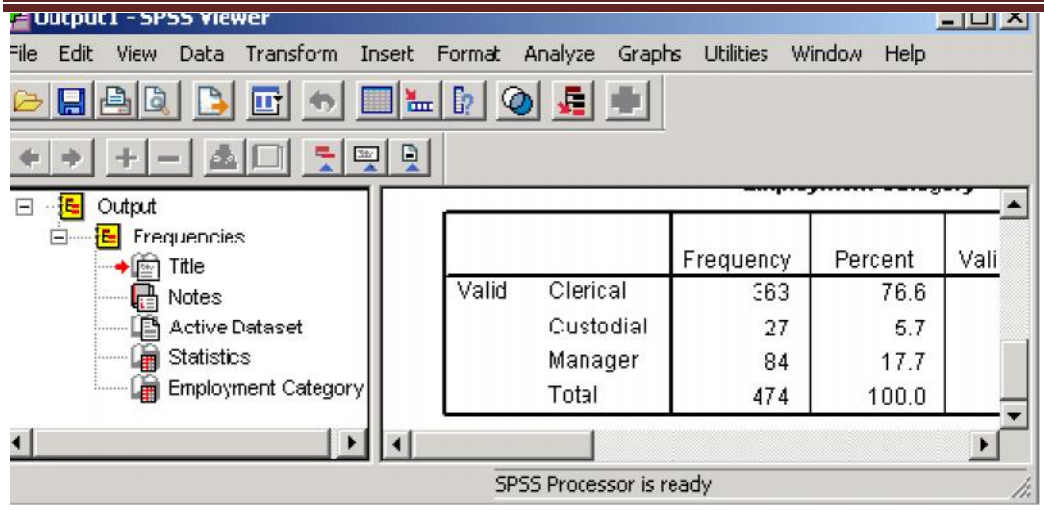
Columns represent attributes of variables

Rows



Variable View Window

The **Viewer** window is where results are displayed after a statistical procedure has been performed. It is divided into two main sections: the left pane contains an outline view of the output contents and the right pane contains statistical tables, charts, and text output. You can edit the output in this window and save it for later use. This window opens automatically the first time you run a procedure that generates output. See Fig. 3 for details.

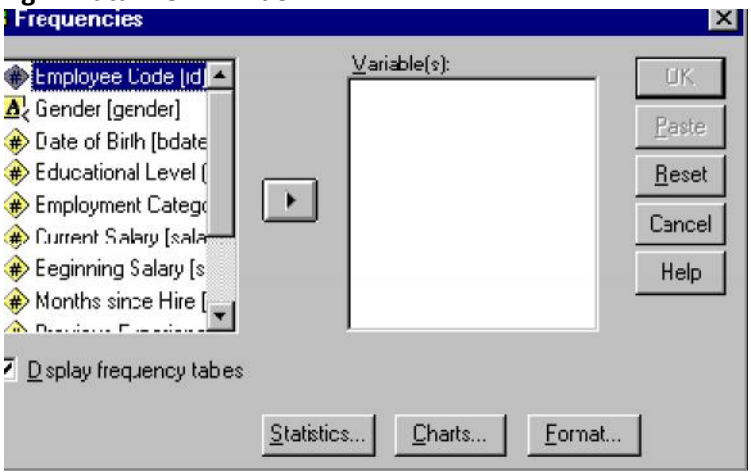


Right pane contains statistical tables, charts and text output
Left pane contains outline
view of the output contents

Dialogue boxes

You use dialogue boxes to select variables and options for statistics and charts. You select variables for analysis from the **source** list. And you use the arrow button to move the variables into the **target** list. Dialogue box buttons with an ellipsis (...) open sub dialogue boxes for optional selections. There are five standard buttons on most dialogue boxes (**OK**, **PASTE**, **RESET**, **CANCEL**, and **HELP**). You see some diagrams of some dialogue boxes as you progress through this document. The Frequency dialogue box is shown in Fig. 11.

Fig. 1 Data View Window



The OK button is not available because no variable has been transferred to the target list yet A single click on any of these buttons will open sub dialogue boxes Click this arrow to transfer variable(s) Target variable list Source variable list. The variables can be selected

Variable names

Always give meaningful names to all your variables. If you do not, SPSS will name the variables for you, calling the first variable var00001, the second var00002 and so on. There are six specific rules that you should follow when selecting variable names. A variable name:

1. must not exceed 32 characters. (A character is simply a letter, digit or symbol).
2. must begin with a letter.
3. could have a mixture of letters, digits and any of the following symbol: @, #, _, $.
4. must not end with a full stop.
5. must not contain any of the following: a blank, !, ?, *.
6. must not be one of the keywords used in SPSS (e.g. AND, NOT, EQ, BY, and ALL)

Value labels

` With **Value labels** you assign names to arbitrary code numbers. For example, you may want to perform a statistical procedure on two groups that have been given arbitrary code numbers of 1 and 2. You can give **Value labels** to these code numbers such as:

1="group 1"
2="group 2"

Data entry using the keyboard

When the **Data Editor** window is accessed for the first time, the top cell of the leftmost column will be highlighted (i.e. thickened black borders round the cell). This is the active cell. You can make any cell active by moving your mouse to the required cell and then clicking the left mouse button. Notice that as you change the active cell, the cell editor on the left, track the location of the active cell. A value typed in from the keyboard will appear in the cell editor and can be transferred to the active cell by pressing **return** or **enter** key on the keyboard. You can change position of the active cell in grid by using the cursor keys (i.e. the up, down, right and left arrows on the keyboard). You can now enter data into any cell.

Editing data on the grid

The editing functions found in most applications are available in SPSS for Windows. You can copy, cut, and paste in SPSS. The block-and-paste technique can also be used. To delete the values in a cell (or block), highlight the required area and press **shift delete** or the **back space** key. To delete the values of an entire row, click on the grey area 8 containing the row number followed by delete. Similarly, to delete the values of an entire column, click on the grey area containing the name of the column followed by delete.

How to conduct an Exploratory Data Analysis - Quantitative Variable

Now that we have successfully entered and saved data into SPSS, it is time to perform some statistical data analysis procedures. However, it is advisable to conduct an Exploratory Data Analysis (EDA) before carrying out any formal data analysis. Why not attempt some Exploratory Data Analysis using the following: **Explore**, **Descriptive**, and **Frequencies**. Follow these instructions:

The Explore Procedure

Start SPSS by selecting **Start -> All programs -> Statistical software -> IBM SPSS Statistics -> IBM SPSS Statistics 19**. Click on **Cancel** to cancel the displayed dialogue box. From the menu bar select **File -> Open -> Data**. Under **File name:** type **\\campus\software\dept\spss** and click **Open.** Select **Employee data** and click **Open**. Study this data file.

Select **Analyze -> Descriptive Statistics -> Explore...**. The **Explore** dialogue box will appear on the screen. Highlight the variable *Current Salary [salary]* by clicking on it once using your mouse left button and transfer it to the **Dependent List** box by clicking the top arrow. Highlight the variable *Employment Category [jobcat]* and transfer it to the **Factor List** box by clicking the middle arrow.

Click on **Plots…** to open the **Explore:Plots** dialogue box and deselect the **Stem-and-Leaf** check box in the **Descriptive** group. If **Stem-and-Leaf** is already deselected click on **Continue**.

5. Click on **OK** to run the procedure. The result of this procedure will be displayed on the **Output Viewer** window. Examine and try to interpret the result.

The Descriptives Procedure

With Descriptive you can quickly generate summary statistical measures such as *mean*, *standard deviation*, *variance*, *maximum* and *minimum* values, *range* and *sum* for a given variable. Follow these instructions:

1.From the menu bar, select Analyze -> Descriptive Statistics -> Descriptives.... The Descriptives dialogue box will appear on the screen.

2. Transfer the variable *Current Salary [salary]* into the Variable(s) box.

3. Select the Options pushbutton. The Descriptives: Options dialogue box will appear on the screen. Notice that Mean, Std. deviation, Minimum and Maximum have already been selected for you. These are the default statistics.

4. Also select these statistical measures: Variance, Range, Sum, and S.E mean. To select an item click on the check box once. To deselect it click on it again once.

5. Select Continue to return to the Descriptives dialogue box.

6. Select OK to run the procedure.

Examine and attempt to interpret the output.

What are the main differences between the output from the **Descriptives Procedure** compare to the output from the **Explore Procedure**?

The Frequencies Procedure

With the Frequencies procedure you can also generate summary statistical measures for a given variable. **Frequencies** gives frequency distributions for all types of data (nominal, ordinal and interval). This example concentrates on the quantitative variable *Current Salary [salary]*. An example involving qualitative

Application of spss programs

To start practical situational applications of SPSS, program, we have to enter data or open entered data. Data is directly entered in the Data Editor in any sequence. Data can be entered either by case or by variable, especially for selected areas or for individual cells.

Usually, a variable is to be defined or selected and related data is entered under the variable. Then data is subjected to analysis, and finally necessary analysis report is generated.

for data entry, data may be classified as numeric or non-numeric.

1. <u>to enter non- numeric data</u>
   - select data view or variable view
   - select type and opt variable
   - click ok
   - double lick row number or click data view tab
   - enter data in the selected variable
     we can also use value label for data entry
2. <u>to enter numeric data</u>

   numeric data take the form of numbers, symbols and other mathematical expressions.
   - select data view or variable view
   - elect a cell
   - enter data value
   - press enter or select another value, to record

3. to use value labels
    - select vies-value labels
    - click the cell in which value is to be entered
    - from the drop down list, select required value label

Data value restrictions

when a variable is defined, its width also must be determined. accordingly, following data value restrictions will operate:

- data will be accepted only allowed by the width
- only acceptable characters permitted by the variable will be entered
- for numeric variables, integer value and symbols are accepted
- for entering more characters, variable width can be changed
- wider and unacceptable values will be displayed in brackets( )

To edit data

There are numerous options available in SPSS that permit the user to modify window displays and also output format. With the data editor, user can modify data values in data view using following options:

- change data values
- cut, copy and paste data values
- add or delete values
- add or delete variables
- change order of variables

Using the pull down menu

- select cases-delete or edit
- select case
- select – sort
- select-merge-add cases-add variables
- select-weight cases
- select-transform-combine – variable

Editing output in spas

spss produces discreet output objects rather than ordinary ascii test. These output objects are immensely formatted tables and are called pivot tables or charts which can be easily edited.

- table output viewer and select one or more objects
- select, hide or delete
- drag the objects to new positions in the outline
- double-click on any object
- to activate toolbar, select view→ toolbar
- Insert new lines, new titles, headings etc.

Descriptive statistics

The most primary objective of statistical study is describing the properties like central tendency, dispersion, normality, spread of distribution, percentages, properties, etc.

spas enables such descriptive statistics, which provide necessary information about description of variables. to compute descriptive statistics:

- enter data editor
- select analyze
- select descriptive analysis
- opt any of the descriptive, tables, reports, correlation, regression, etc

Descriptives help to compute univariate statistics for numeric variables including the mean, standard deviation, minimum and maximum values. Because it does not sort values into a frequency table, hence it is considered an efficient way of computing descriptive statistics for continuous variables. Other measures that display descriptive statistics include frequencies, means and examine.

Under descriptive option, you can specify the display order options by ticking the option as per your requirement. Following options are available under display order

- Variable list : this is the default for this option. This arranges the items in the same order as found in the data editor.
- Alphabetic: names of variables are arranged alphabetically.
- Ascending means : this orders the means from smallest mean value to highest mean value in the output.
- Descending means: this orders the means from largest mean value to smallest mean value in the output.

Data analyses tools in SPSS

Codebook

Codebook reports the dictionary information, such as variable names, variable labels, value labels, missing values and summary statistics for all, or specified variables and multiple response sets in the active dataset. For normal and ordinal variables and multiple response sets, summary statistics include counts and percents. for sale variables, summary statistics include mean, standard deviation, and quartiles. Codebook ignores split file status. this include split file groups created for multiple imputation of missing values available in the missing value add-ons option. Select the spss table and then form the main menu select **analyze→reports , codebook** . The following screen will appear. Make the selections as per your requirements.

Frequencies

The frequency method provides statistics and graphical displays and helps to describe many types of variables. the frequencies procedure is a good place to start looking at your data. for a frequency report and bar chart, the distinct values can be arranged in ascending or descending order or you can organize the categories as per their frequencies. the frequency reports can be labeled with frequencies (the default) or percentages.

Explore

The explore method generate summary statistics and graphical displays for all of the cases or it can individually generate it for groups of cases. there are various causes to use the explore procedure data screening, outlier identification, description, assumption checking and characterizing differences among subpopulations (groups of cases). data screening can illustrate the unusual values, extreme values, gaps in the data or other peculiarities that you have. exploring the data helps to conclude that the statistical techniques that are being considered for data analysis are correct. the exploration may possibly specify that the data must be transformed if the technique needs a standard normal distribution or the user can use the appropriate non-parametric tests.

Means

The means method evaluate subgroup means and correlated univariate statistics for dependent variables within the various types of one or more independent variables. Alternatively, a one-way analysis of variance and tests for linearity can be obtained

Arithmetic analysis using SPSS

One of the usual requirements is finding arithmetic mean of certain observations. Once data is entered into SPSS, arithmetic mean is found out following the procedure below.

Enter variables (Example ; Years, Income etc.)

Select statistics viewer or data editor

Select descriptive statistics

Opt arithmetic mean

Select grand mean, men deviation, standard error mean etc.

Conducting t test using SPSS

Like normal distribution, student's t distribution is also symmetrical but happens to be flatter than normal distribution. Moreover, there is a different t distribution of every possible sample size. As the sample size gets larger the shape of t distribution loses its flatness and becomes approximately equal to normal distribution.

For applying t test , t value is calculated first of all, and then the calculated t value is compared with the t table value at certain level of significance and degree of freedom.

Enter variables and numeric (Example - sample number, hours, standard deviation

Select statistics viewer or data editor

Select analyze t test

The calculated value will be displayed along with t table value at specified level of significance and degree of freedom.

Correlation analysis using SPSS

Correlation analysis is the statistical tool generally used to describe the degree to which, one variable is related to another. The relationship is usually assumed to be a linear one. This is used frequently in conjunction with regression analysis to measure how well regression line explains variation of the dependent variable.

Statisticians have developed two measures for describing the correlation further i.e., coefficient of determination and the coefficient of correlation.

Enter data in variables and numeric

Select descriptive statistics

Select analyze – bivariate analysis

Select correlation

Opt coefficient of correlation, partial correlation or power of correlation

Conducting ANOVA using SPSS

In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the

differences among various population means. SPSS enables to present all details regarding variances in the form of a table called ANOVA table, showing the source of variation, sum of squares, degrees of freedom and means squares along with the concerned F value. In the case one way classification, there will be one F value and in the case of two way classification, there will be two f values.

Enter data – statistics viewer
Select analyze one way Anova/two way Anova
Give degree of freedom and level of significance
Select variance within sample groups
Select variance between sample groups
Select total variance
Select present Anova table.

Regression analysis

Regression helps to predict unknown values on the basis of known values, based on average relation between variables. SPSS enables regression through scatter diagram method, least square method and linear regression equation.

Enter data in variables and numeric
Select analyze – descriptive statistics
Select dependent and independent variables
Select regression
Find values of X and Y variables in terms of the other variable.

Review questions and exercises

1. What is the business application of MS excel
2. What is MINITAB used for
3. What is SAS
4. Why SPSS is the most widely used statistical software
5. How numeric data is entered in SPSS
6. Can value labels be used fro data entry
7. How is data edited in SPSS
8. What are data restrictions in SPSS
9. How will non- numeric data be entered
10. What is the procedure in creating a variable
11. How is data edited by pull down manner
12. Can output of SPSS be edited
13. Following data is given

    Sample size = 10
    Sample mean = 80
    Population mean = 82stanarad deviation =4

    Conduct student t test, using spss

14. How is Karl Pearson's coefficient of correlation obtained using SPSS

Price	Demand
10	80
11	76
12	80
13	78
14	76
15	70

15. An Agricultural researcher is studying the effect of soil types on the tastes of strawberries. Use Kruskel Wallis H test to examine the hypothesis that soil type affects strawberry taste..

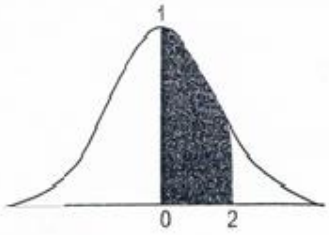
Type 1	Type 2	Type 3
Best 1	Best 1	Sour 4
Tasty 2	Sour 4	Sweet 3
Sweet 3	Tasty 2	Tasty 2
Sour 4	Sweet 3	Best 1

(hint : analyze >non prametric test > k independent sample > define test variable > define groupings > select h test > ok)

16. Obtain the estimating equations by the method of least squares from the following information., using SPSS

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

Table I : Area under normal curve



z	.0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2903	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981

Table II : Values of t distributions

d.f.	Level of significance for two-tailed test					d.f.
	0.20	0.10	0.05	0.02	0.01	
	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.01	0.005	
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.731	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
Infinity	1.282	1.645	1.960	2.326	2.576	Infinity

Table III: Values of Chi-Square

Degrees of freedom	Probability under H_0 that of χ^2 > Chi square						
	.99	.95	.50	.10	.05	.02	.01
1	.000157	.00393	.455	2.706	3.841	5.412	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	.1145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.342	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	72.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	4.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	32.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.652	41.566	44.314
26	12.198	15.379	25.336	35.363	38.885	41.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

Note: For degrees of freedom greater than 30, the quantity $2\chi^2 - \sqrt{2\text{d.f.} - 1}$ may be used as a normal variate with unit variance i.e., $z_\alpha = \sqrt{2\chi^2} - \sqrt{2\text{d.f.} - 1}$.

Table IV: Values of F distribution 5%

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.1	243.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

v_1 = Degrees of freedom for greater variance.

v_2 = Degrees of freedom for smaller variance.

Table V: Values of F distribution 1%

v_2 \ v_1	1									
1	4052	4999.5	5403	5625	5764	5859	5982	6106	6235	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.45
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.87	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	3.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.20	2.96	2.58	2.10
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.45	2.13
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

v_1 = Degrees of freedom for greater variance.

v_2 = Degrees of freedom for smaller variance.

Table VI: Cumulative binomial probabilities

n	r_0	.10	.25	.40	.50
1	0	.9000	.7500	.6000	.5000
	1	1.0000	1.0000	1.0000	1.0000
2	0	.8100	.5625	.3600	.2500
	1	.9900	.9375	.8400	.7500
	2	1.0000	1.0000	1.0000	1.0000
5	0	.5905	.2373	.0778	.0313
	1	.9185	.6328	.3370	.1875
	2	.9914	.8965	.6826	.5000
	3	.9995	.9844	.9130	.8125
	4	.9999	.9990	.9898	.9687
	5	1.0000	1.0000	1.0000	1.0000
10	0	.3487	.0563	.0060	.0010
	1	.7361	.2440	.0463	.0108
	2	.9298	.5256	.1672	.0547
	3	.9872	.7759	.3822	.1719
	4	.9984	.9219	.6330	.3770
	5	.9999	.9803	.8337	.6230
	6	1.0000	.9965	.9452	.8281
	7	1.0000	.9996	.9877	.9453
	8	1.0000	1.0000	.9983	.9892
	9	1.0000	1.0000	.9999	.9990
	10	1.0000	1.0000	1.0000	1.0000
12	0	.2824	.0317	.0022	.0002
	1	.6590	.1584	.0196	.0031
	2	.8891	.3907	.0835	.0192
	3	.9740	.6488	.2254	.0729
	4	.9963	.8424	.4382	.1937
	5	.9999	.9456	.6652	.3871
	6	1.0000	.9857	.8418	.6127
	7	1.0000	.9972	.9427	.8064
	8	1.0000	.9996	.9847	.9269
	9	1.0000	1.0000	.9972	.9806
	10	1.0000	1.0000	.9997	.9977
	11	1.0000	1.0000	1.0000	1.0000
	12	1.0000	1.0000	1.0000	1.0000

Table VII: Values for control charts

	Factors for $\bar{x}$ Charts		Factors for R Charts		
Sample size, n	$d_2 = \frac{R}{\sigma}$	$A_2 = \frac{3}{d_2\sqrt{n}}$	$d_3 = \frac{\sigma_R}{\sigma}$	$D_3 = 1 - \frac{3d_3}{d_2}$	$D_4 = 1 + \frac{3d_3}{d_2}$
2	1.128	1.881	0.853	0	3.269
3	1.693	1.023	0.888	0	2.574
4	2.059	0.729	0.880	0	2.282
5	2.326	0.577	0.864	0	2.114
6	2.534	0.483	0.848	0	2.004
7	2.704	0.419	0.833	0.076	1.924
8	2.847	0.373	0.820	0.136	1.864
9	2.970	0.337	0.808	0.184	1.816
10	3.078	0.308	0.797	0.223	1.777
11	3.173	0.285	0.787	0.256	1.744
12	3.258	0.266	0.779	0.283	1.717
13	3.336	0.249	0.770	0.308	1.692
14	3.407	0.235	0.763	0.328	1.672
15	3.472	0.223	0.756	0.347	1.653
16	3.532	0.212	0.750	0.363	1.637
17	3.588	0.203	0.744	0.378	1.622
18	3.640	0.194	0.739	0.391	1.609
19	3.689	0.187	0.734	0.403	1.597
20	3.735	0.180	0.729	0.414	1.586
21	3.778	0.173	0.724	0.425	1.575
22	3.819	0.167	0.720	0.434	1.566
23	3.858	0.162	0.716	0.443	1.557
24	3.895	0.157	0.712	0.452	1.548
25	3.931	0.153	0.708	0.460	1.540

Note: If $1 - 3d_3/d_2 < 0$, then $D_3 = 0$.