QUANTITATIVE TECHNIQUES

COURSE CODE: M21CM07DC

Postgraduate Programme in Commerce
Discipline Core Course

SELF LEARNING MATERIAL

SREENARAYANAGURU OPEN UNIVERSITY
The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU OPEN UNIVERSITY

Vision

To increase access of potential learners of all categories to higher education, research and training, and ensure equity through delivery of high quality processes and outcomes fostering inclusive educational empowerment for social advancement.

Mission

To be benchmarked as a model for conservation and dissemination of knowledge and skill on blended and virtual mode in education, training and research for normal, continuing, and adult learners.

Pathway

Access and Quality define Equity.

Quantitative Techniques

Course Code: M21CM07DC

Semester - II

Discipline Core Course
Master of Commerce
Self Learning Material
(With Model Question Paper Sets)

SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

Quantitative Techniques
Course Code: M21CM07DC
Discipline Core Course
Semester - II
Master of Commerce

SREENARAYANAGURU
OPEN UNIVERSITY

www.sgou.ac.in

ISBN 978-81-970547-1-6

9 788197 054716

DOCUMENTATION

Production

July 2024

YouTube

MESSAGE FROM VICE CHANCELLOR

Dear learner,

I extend my heartfelt greetings and profound enthusiasm as I warmly welcome you to Sreenarayanaguru Open University. Established in September 2020 as a state-led endeavour to promote higher education through open and distance learning modes, our institution was shaped by the guiding principle that access and quality are the cornerstones of equity. We have firmly resolved to uphold the highest standards of education, setting the benchmark and charting the course.

The courses offered by the Sreenarayanaguru Open University aim to strike a quality balance, ensuring students are equipped for both personal growth and professional excellence. The University embraces the widely acclaimed "blended format," a practical framework that harmoniously integrates Self-Learning Materials, Classroom Counseling, and Virtual modes, fostering a dynamic and enriching experience for both learners and instructors.

The university aims to offer you an engaging and thought-provoking educational journey. This learning material on Quantitative Techniques for the MCom programme builds on the mathematics skills you learned during your bachelor's degree. It takes basic statistical ideas and makes them more advanced. The course mixes theory with real-world business problems to give you a full understanding. These skills are really important for managing companies well and making good decisions based on data.The Self-Learning Material has been meticulously crafted, incorporating relevant examples to facilitate better comprehension.

Rest assured, the university's student support services will be at your disposal throughout your academic journey, readily available to address any concerns or grievances you may encounter. We encourage you to reach out to us freely regarding any matter about your academic programme. It is our sincere wish that you achieve the utmost success.

Warm regards.
Dr. Jagathy Raj V. P.

01-07-2024

Contents

01 BLOCK

PROBABILITY DISTRIBUTION

Block Content

Unit 1

Introduction to Quantitative Techniques

Learning Outcomes

After completing this unit, the learner will be able to:

- acquaint with the concept of quantitative techniques.
- develop insights on the application of quantitative techniques in the field of business, industry, and management.
- gain knowledge on probability distribution.

Background

Quantitative techniques are mathematical and statistical tools used in business and management to analyze data and make informed decisions. Central to these techniques is the concept of probability distribution, which helps businesses assess uncertainties and risks. By quantifying the likelihood of different outcomes, organizations can forecast trends, plan resources, and strategize effectively. In business, quantitative techniques find diverse applications, from market analysis and demand forecasting to financial modeling and performance evaluation. For instance, in marketing, these methods aid in segmenting markets, targeting specific consumer groups, and measuring the effectiveness of marketing campaigns. In finance, they facilitate portfolio management, risk assessment, and pricing strategies. In management, quantitative techniques support strategic planning, resource allocation, and performance measurement. Through tools like regression analysis, managers optimize resource utilization, simulate scenarios, and evaluate the impact of strategic decisions.

Quantitative techniques empower businesses to make data-driven decisions, enhance operational efficiency, and maintain competitiveness in dynamic markets. By harnessing mathematical and statistical methods, organizations gain deeper insights into their operations, enabling them to adapt and thrive in an ever-changing business landscape.

Keywords

Probability, Probability distribution, Addition theorem, Multiplication theorem, Frequency distribution

Discussion

1.1.1 Introduction to Quantitative Techniques

- Offer solutions to business problems

Quantitative technique is a discipline which has a huge scope in the field of business, economics, finance and agriculture. Businesses make use of quantitative techniques to achieve its goals such as increasing production and revenue, controlling cost etc. Quantitative techniques, thus, helps businesses by offering solutions to their problems and thereby achieve success.

Meaning

- Facilitate decision making

Quantitative techniques are those statistical and operations research or programming techniques which help in the decision making process, especially concerning business and industry. Through the use of numbers, symbols, and other mathematical expressions, these strategies introduce the concept of quantities. These techniques evaluate planning factors and alternatives as and when they arise rather than prescribe courses of action. Accordingly, quantitative procedures can be characterised as those that offer the decision-maker a methodical and potent way to analyse and support the exploration of strategies for accomplishing predetermined goals based on quantitative data. These methods are especially applicable to issues facing intricate corporate structures.

1.1.2 Application of Quantitative Techniques in Business, Industry, and Management

The deployment of Quantitative techniques facilitates decision making in business, industry, and management. The following are the applications of Quantitative techniques:

- Operations and production

a. Quantitative techniques play an important role in production and operations of a business. It facilitates inventory management, schedule production activities efficiently, and ensures timely deliveries. Various quality control techniques also help to prevent defects and improve product quality.

b. Decision making in management benefits from quantitative models and thorough data analysis, offering evidence-based insights that mitigate risks and enhance overall outcomes. In project management, techniques such as PERT, CPM, and others facilitate comprehensive planning, scheduling, and monitoring, ensuring projects are completed promptly and within budget constraints. Strategic planning processes leverage data analysis and modeling to guide decisions re-

- Management decision making

garding market expansion, product innovation, and resource allocation, fostering informed strategies for long-term success. These cohesive quantitative methodologies empower organisations to make sound decisions, manage projects effectively, and pursue strategic initiatives with confidence.

- Market research and segmentation

c. Market research relies on surveys, sampling methods, and statistical analysis to gather valuable customer insights, comprehend market trends, and guide strategic marketing decisions. Through market segmentation techniques like clustering and discriminant analysis, customers are grouped based on shared attributes, facilitating the customization of marketing strategies to target specific segments effectively. Sales forecasting utilizes time series analysis and regression models to predict future sales, optimizing resource allocation, production scheduling, and inventory management processes. These quantitative approaches collectively empower businesses to understand their market, tailor their marketing efforts, and efficiently manage their operations.

- Management of funds

d. Quantitative techniques such as statistical analysis, time series analysis, and risk modeling are utilized within financial markets to optimize portfolios, evaluate risks, assess asset values, and forecast financial trends. These methods enable businesses to make well-informed investment decisions and effectively manage financial risks. Ratio analysis and similar techniques are employed to gauge a company's financial stability, profitability, and liquidity, aiding in investment evaluations. Discounted cash flow (DCF) analysis and comparable methods are employed to assess potential investment profitability, facilitating the selection of optimal investment opportunities. Models like Value at Risk (VaR) quantify financial risks, empowering businesses to devise strategic risk mitigation plans.

- Job recruitment, selection, evaluation

e. Recruitment and selection processes rely on statistical tests and assessment tools to pinpoint suitable candidates for job vacancies, enhancing the quality of hiring decisions. In performance management, the utilization of performance metrics alongside regression analysis enables the evaluation of employee effectiveness, guiding decisions related to training, development, and compensation. Workforce planning benefits from forecasting models, which anticipate future staffing requirements and contribute to strategic decisions regarding recruitment, training, and resource allocation. These integrated quantitative approaches empower organizations to optimize their talent acquisition, performance evaluation, and workforce management practices.

1.1.3 Probability

In daily conversation, the words possible chance, likelihood etc. are commonly used for representing the meaning of probability. You all have a rough idea of what is meant by these words. For example, we may come across the statements like: the rain may come today, the chance of winning the cricket match etc. It means there is uncertainty about the happening of the event(s). We live in a world where we are unable to forecast the future with complete certainty. Our need to cope with uncertainty leads us to the study the use of probability. Initially applied to gambling, probability theory has evolved into a crucial tool for addressing uncertainties in various socio-economic scenarios. Gamblers have used the probability concept to make bets. The probability theory was first applied to gambling and later to other socio economic problems. Lately, the quantitative analysis has become the backbone of statistical application in business decision making and research. Decision making in various situations is facilitated through formal and precise expressions for the uncertainties involved. Probability theory provides us with the means to arrive at preciseexpressions for taking care of uncertainties involved in different situations. If the conditions of certainty only were to prevail, life would have been much simpler. There are numerous real life situations in which conditions of uncertainty and risk prevail. Consequently, we have to rely on the theory of chance or probability in order to have a better idea about the possible outcomes.

- Decision-making under uncertainty and risk

Probability is the branch of Mathematics concerning numerical descriptions of how likely an event is to occur, or likely it is that a proposition is true. The probability of an event is a number between 0 and 1, where '0' indicates impossibility of an event and '1' indicate certainty. A probability distribution is essentially an extension of the theory of probability. All probability distributions have immensely useful applications and explain a wide variety of business situations which call for computation of desired probabilities. The basic objective of calculating probabilities is to facilitate us in decision making. By the theory of probability,

$$P(H_1) + P(H_2) + P(H_n) = 1$$

This means that the unity probability of a certain event is distributed over a set of disjoint events making up a complete group. In general, a tabular recording of the probabilities of all the possible outcomes that could result if random (chance) experiment is done is called "Probability Distribution". It is also termed as theoretical frequency distribution.

- Likeliness of an event to occur

1.1.4 Definition of terms

a. Random experiment

- All the possible outcomes are known

An experiment that has two or more outcomes which vary in an unpredictable manner from trial to trial when conducted under uniform conditions is called a random experiment. In a random experiment all the possible outcomes are known in advance but none of the outcomes can be predicted with certainty. The tossing of a coin, for example is a random experiment since it has two specified outcomes-Head and Tail. But we are uncertain whether head or tail will turn when the coin is tossed.

Throwing a die is another example of a random experiment. The essential features of a random experiment are

- It has more than one result
- The results are unpredictable
- The experiment is repeatable

b. Sample point

- Element in set serving as sample space

Every outcome of a random experiment is called a sample point of that random experiment. For example, when a coin is tossed getting head is a sample point. Similarly, when two dice are thrown getting (2, 3) is a sample point.

c. Sample space

- Complete set of outcomes

A sample space of a random experiment is the set of possible outcome of that random experiment. The following are the examples:

- When a coin is tossed the sample space is (Head, Tail)
- When two coins are tossed sample space is (HH, HT, TH, TT)

d. Event

- Outcome or combination of outcomes

One or more possible outcome of an experiment are said to form an event. So an event is a sub set of a sample space.

For example, when two coins are tossed getting two heads is an event. An event may be simple or compound.

- Simple and compound event

An event is said to be simple if it corresponds to a single possible outcome of an experiment. The following are the examples:

- When two coins are tossed, getting two heads is a simple event

- When a coin is tossed twice, getting Head in both tosses is a compound event.

The classification of events are as follows:

i. Certain, Impossible and Uncertain events

- Event will certainly occur

An event whose occurrence is inevitable is called certain event. For example, getting a white ball from a bag containing all white balls is certain.

An event which can never occur is called impossible event. For example, drawing a white ball from a bag containing all black balls is impossible.

- Event may or may not occur

An event which may or may not occur is known as uncertain event. For example, getting a white ball from a bag containing both white and black balls is an uncertain event.

ii. Equally likely events

Two or more events are said to be equally likely if any one of them cannot be expected to occur in preference to the others. The following are the examples:

- Outcomes are equally probable

- Getting Head or getting Tail when a coin is tossed
- Getting 1, 2, 3, 4, 5, or 6 when a die is thrown

iii. Mutually Exclusive events

- Two events do not happen in one occasion

Two events are said to be mutually exclusive if the occurrence of one of them excludes (or prevents) the possibility of the occurrence of the other. In other words, two mutually exclusive events cannot occur simultaneously in the same trial. For example, when a coin is tossed, we can't get tail if head comes. Similarly, if tail comes, we can't get head. Thus, it is said to be mutually exclusive.

iv. Exhaustive cases

- Total number of possible outcomes

A group of events is said to be exhaustive when it includes all possible outcomes of the random experiment under consideration.

For instance, when a die is thrown, outcomes 1, 2, 3, 4, 5, and 6 together will form exhaustive cases.

v. Complementary events

The event "A" and the event "A does not occur" are called complementary events."A does not occur" is denoted by A^c.

$P(A^c) = 1 - P(A)$

The following are the examples:

- In tossing a coin getting head and getting tail are complementary events
- Getting at least one head and getting no head, when two coins are tossed, are complementary events.

- $P(A^c) = 1 - P(A)$

The outcomes of A^c will be those outcomes of the sample space which are not in A.

Union of two events

The union of two events A and B denoted by AUB is the set of sample point in A or in B

E.g. A= {1,2,3,4}

B= {3,4,5,6}

Then AUB = {1,2,3,4,5,6}

(The outcomes common to A and B will also be present in AUB)

Intersection of two events

The intersection of two events A and B denoted by A∩B is the set of sample points common to both A and B. In the above case A∩B = {3,4}

vi. Independent and Dependent event

Independent Events

Two or more events are said to be independent if the occurrence of one of them in noway affects the occurrence of the other or the others.

Two events are independent if the following are true:

- P(A/B) = P(A)
- P(B/A) = P(B)
- P (A and B) = P(A)P(B)

- Happening of one event is not affected by another

E.g. In the tossing of a coin twice, the result of the second tossing is not affected by the result of the first toss.

Dependent events

Two or more events are said to be dependent if the happening of one of them affects the happening of the others. That is, the chance of one event depends on the happening of the other events. E.g. From a pack of 52 cards, if one card is drawn then 51 cards are left. If another card is drawn, without replacing the first the chance of the second draw is affected by the first draw.

- Happening of one event is affected by another

1.1.5 Theorems of Probability

a. Addition theorem

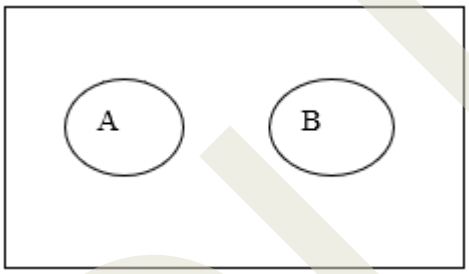
If A and B are two events then the probability for A or B to take place, denoted by(AUB)

= P(A)+P(B) if A and B are mutually exclusive and

= P(A) + P(B) - P (A∩B) if A and B are not mutually exclu sive events

Proof

If A and B are mutually exclusive then A and B are disjoint so that A∩B =ǿ (null set)

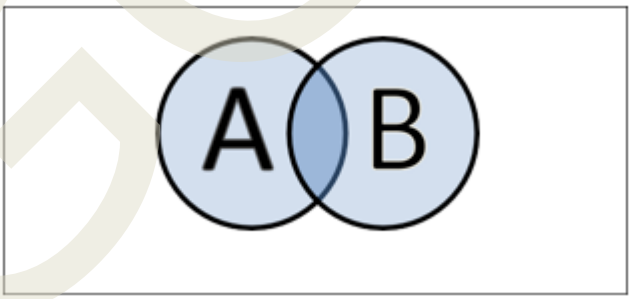


If A and B are not mutually exclusive, then

n(AUB) = n(A) + n(B) - n(A∩B)

Therefore P (AUB) = P(A) + P(B) - P(A∩B)

Where A∩B is the outcome common to A and B



b. Multiplication theorem

If A and B are two events then probability for both A and B to take place denoted by P(A∩B)

=P(A) P(B/A) if A and B are not independent

=P(A)P(B) if A and B are independent

Probability under statistical dependence

- Occurrence of one is affected by occurrence of other

Two or more events are said to be statistically dependent, if the occurrence of any one event affects the probability of occurrence of the other event. There are three types of probability under statistical dependence case.

They are:

i. Conditional probability

ii. Joint probability

iii. Marginal probability

Conditional probability

Probability of an event A, given that B has happened is called the conditional probability of A given B and is denoted by P (A/B)

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

Let us consider the following situation

Consider a family with two children. Then the different outcomes are (Boy, Boy),(Boy, Girl), (Girl, Boy) (Girl, Girl)

Total number of outcomes = 4

If it is known that the first is a boy, the outcomes are (B, B), (B,G). So total number of cases = 2

$$P(\text{both boys/first is a boy}) = \frac{1}{2} \ldots\ldots (i)$$

This is called the conditional probability since the condition is that first is a boy.

Baye's Theorem

The theory is stated by Thomas Bayes, who is a British mathematician. It is used to understand how probability is affected by new piece of information.

Let A be the event that an individual tests positive for some disease and C be the event that the person actually has the disease. We can perform clinical trials to estimate the probability that a randomly chosen individual tests positive given that they have the disease.

- How probability is affected by new piece of information

P (tests positive/has the disease) = P(A/C), by taking individuals with the disease and applying the test. However, we would like to use the test as a method of diagnosis of the disease. Thus, we would like to be able to give the test and assert the chance that the person has the disease. That is, we want to know the probability with the reverse conditioning

P (has the disease/tests positive) =P(C/A)

1.1.6 Frequency Distribution and Probability Distribution

- Based on observations and experimentations

Frequency distributions are based on observations and experimentations. For instance, we may study the profits (during a particular period) of the firm in an industry and classify the data into two columns with class intervals for profits in the first col-

umn and corresponding classes frequencies (No. of firms) in the second column.

The probability distribution is also a two column presentation with the values of the random variable in the first column, and the corresponding probabilities in the second column. These distributions are obtained by expectations on the basis of theoretical or past experience considerations. Thus, probability distributions are related to theoretical or expected frequency distributions. In the frequency distribution, the class frequencies add up to the total number ofobservations (N), whereas in the case of probability distribution the possible outcomes (probabilities) add up to 'one'. Like the former, a probability distribution is also described by a curve and has its own mean, dispersion, and Skewness.

- Based on expected frequencies

Let us consider an example of probability distribution. Suppose we toss a fair coin twice; the possible outcomes are shown in Table below:

Table 1.1.1 Possible Outcomes from Two –toss Experiment of a Fair coin

No. of possible outcomes	1st toss	2nd toss	No of heads on two tosses	Probability of the possible outcomes
1	Head	Head	2	0.5 x 0.5 = 0.25
2	Head	Tail	1	0.5 x 0.5 = 0.25
3	Tail	Head	1	0.5 x 0.5 = 0.25
4	Tail	Tail	0	0.5 x 0.5 = 0.25
				Total=1.00

We would begin by recording any result that did not contain a head, i.e., only the fourth outcome in Table 1.1.1. Next, those outcomes containing only one head, i.e., second and third outcomes in Table 1.1.1, and finally, we would record that the first outcome containing two heads in Table 1.1.1. We recorded the same in Table 1.1.2 to highlight the number of heads contained in each outcome.

Table 1.1.2 Probability Distribution of the Possible No. of Heads from Two toss Experiment of a Fair coin

No. of Heads(H)	Tosses	Probabilityof outcomes P (H)
0	(TT)	$\frac{1}{4} = 0.25$

1	(HT)+(TH)	$\frac{2}{4} = 0.50$
2	(HH)	$\frac{1}{4} = 0.25$

We must note that the above tables are not the real outcome of tossing a fair coin twice. But it is a theoretical outcome, i.e., it represents the way in which we expect our two-tossexperiment of an unbiased coin to behave over time.

Types of Probability Distribution

Probability distributions are broadly classified under two heads:

i. Discrete Probability Distribution

ii. Continuous Probability Distribution

- Can assume only integer values

Discrete Probability Distribution: The discrete probability is allowed to take on only alimited number of values. Consider for example that the probability of having your birthday in a given month is a discrete one. As one can have only 12 possible outcomes representing 12 months of a year.

- Can assume integer and fractional values

Continuous Probability Distribution: In a continuous probability distribution, the variable of interest may take on any values within a given range.

Random variable

- Outcome of random experiment

A random variable is a variable (numerical quantity) that can take different values as a result of the outcomes of a random experiment. When a random experiment is carried out, the totality of outcomes of the experiment forms a set which is known as sample space of the experiment. Similar to the probability distribution function, a random variable may be discrete or continuous.

Discrete Probability Distribution

A representation of all possible values of a discrete random variable together with their probabilities of occurrence is called a discrete probability distribution. There are two kinds of distributions in the discrete probability distribution:

i. Binomial Distribution

ii. Poisson Distribution

These two types of discrete probability distribution will be explained in detail in the forthcoming units.

Summarised Overview

Using statistical and mathematical concepts to analyse data and guide decision-making, quantitative approaches are essential tools in business and management. These methods, which use probability distribution at their core, let companies evaluate risks and model uncertainties. The application of theorems such as addition, multiplication, conditional probability, and Bayes' theorem, which offer frameworks for calculating probabilities in diverse contexts, contributes to the improvement of decision-making.

Probability is the numerical description of how likely an event is to occur. The probability of an event is a number between 0 and 1, where '0' indicates impossibility of an event and '1' indicate certainty. An experiment that has two or more outcomes which vary in an unpredictable manner from trial to trial when conducted under uniform conditions is called a random experiment. Conditional probability is the likelihood that an event will occur given that another event has already occurred. In real-world scenarios, these methods are widely employed in several sectors. Organisations can use quantitative methods to make data-driven decisions that promote growth and profitability in a variety of contexts, including market analysis, demand forecasting, financial modelling, and performance evaluation. These strategies are used by companies in marketing to target particular demographics, segment markets, and assess the success of campaigns. They support pricing strategies, risk assessment, and portfolio management in the financial industry.

Self-Assessment Question

1. What is probability?
2. What do you mean by random experiment?
3. What is conditional probability?
4. What does mutually exclusive events mean?
5. What is a complementary event?
6. What are independent events?
7. What do you mean by dependent events?
8. What is a sample space?
9. What do you mean by certain events?
10. What do you mean by an impossible event?

Assignments

1. A card is drawn at random from an ordinary pack of 52 cards, find theprobability that the card drawn is either spade or the diamond. (Answer- 1/2)
2. A bag contains 5 white, 2 black, 3 yellow and 3 red balls. What is the probabilityof getting a white or a red ball at random in a single draw of one? (Answer- 8/13)
3. There are 6 candies in a bowl, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow candie? (Answer- 1/3)

Reference

1. Anand Sharma. (2017). *Quantitative Techniques for decision making*. Himalaya Publishing House.
2. P.N. Arora, Sumeet Arora, S. Arora. (2010). *Comprehensive Statistical Methods*, S. Chand and Company Private Limited, New Delhi
3. D.V.D. Vohra. (2021). *Quantitative Techniques for Management.* McGraw Hill.
4. G.C. Beri. (2017). *Business Statistics*. Tata McGraw, Hill New Delhi

Suggested Reading

1. Gupta & Khanna. (2011). *Quantitative Techniques for Decision Making*. Prentice Hall of India.
2. Gupta SP. (2021). *Statistical Methods*, S. Chand & Sons.
3. Barry Render. (2022). *Quantitative Analysis for Management*. Prentice Hall of India
4. Levin & Rubin. (1986). *Quantitative Approaches for Management*, Pearson.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 2

Binomial Distribution

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the concept of the binomial distribution and identify its key properties.
- calculate the mean and standard deviation of a binomial distribution
- recognize the utility of the binomial distribution in decision-making, quality control, sampling plans, and assessing randomness.

Background

Imagine you are flipping a fair coin repeatedly, recording how many times it lands on heads. Each flip represents a trial, with only two possible outcomes: heads or tails. This scenario highlights the essence of the binomial distribution, a cornerstone of probability theory. The binomial distribution deals with situations where we repeat a fixed number of trials, each with two potential outcomes. Think of it as a mathematical framework for predicting the likelihood of observing a certain number of successes in a series of experiments.

From quality control in manufacturing to predicting the outcome of election results, the binomial distribution permeates diverse fields. By understanding its mean and standard deviation, we gain insights into the expected outcome and variability, crucial for informed decision-making. So, whether you are tossing coins, conducting product inspections, or analyzing survey data, the binomial distribution offers a powerful toolset for understanding and predicting outcomes in the face of uncertainty.

Keywords

Probability of success, Probability of failure, Mean, Standard deviation

Discussion

1.2.1 Binomial Distribution

- Segregate outcomes into two categories- success or failure

Binomial distribution summarizes the number of trials, or observations when each trial has the same probability of attaining one particular value. The binomial distribution determines the probability of observing a specified number of successful outcomes in a specified number of trials. It is the basic and the most common probability distribution. It has been used to describe a wide variety of processes in business. For example, a quality control manager wants to know the probability of obtaining defective products ina random sample of 10 products. If 10 per cent of the products are defective, he/she can quickly obtain the answer, from table of the binomial probability distributions. The binomial distribution describes discrete, not continuous data resulting from an experiment known a Bernoulli Process. Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives, i.e., success or failure. Hence is also known as Bernoulli distribution, as it was originated by Swiss Mathematician James Bernoulli (1654-1705). Bernoulli distribution is a discrete probability distribution where the Bernoulli random variable can have only '0' or '1' as the outcome, p is the probability of success and 1-p is the probability of failure. As this distribution is very easy to understand, it is used as a basis for deriving more complex distributions.

As per this distribution, the probability of getting 0, 1, 2, …… n heads (or tails) in n tosses of an unbiased coin will be given by the successive terms of the expansion of $(q + p)^n$, where p is the probability of success (heads) and q is the probability of failure (i.e., 1-p).

Binomial law of probability distribution is applicable only when:

a. A trial results in either success or failure of an event.

b. The probability of success 'p' remains constant in each trial.

c. The trials are mutually independent i.e., the outcome of any trial is neither affected by others nor affect others.

Assumptions

i. Each trial has only two possible outcomes either Yes or No, Success or failure etc.

ii. Regardless of how many times the experiment is performed, the probability of the outcome, each time, remains the same.

iii. The trials are statistically independent.

iv. The number of trials is known and is 1, 2, 3, 4, 5 etc.

Binomial probability formula

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

Where,

P (r) = Probability of r success in n trials;

p = Probability of success

q = Probability of failure=1-p

r = No. of success desired

n =No. of trials undertaken

The determining equation for n_{Cr} can easily be written as

$$^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

n! can be simplified as follows:

n! = $n(n-1)!$ = $(n-1)$ $(n-2)!$ = $(n-1)$ $(n-2)$ $(n-3)!$ and so on.

Hence the following form of the equation for carrying out computations of the binomial probability is perhaps more convenient:

$$P(r) = \frac{n!}{r!\,(n-r)!}\; p^{r} q^{n-r}$$

The symbol ! means 'factorial', which is computed as follows:

5! means 5x4x3x2x1=120

Mathematicians define 0! as 1

If n is large in number, say $^{50}c_{3}$, then we write (with the help of the aboveexplanation)

$$^{50}c_{3} = \frac{50!}{3!(50-3)!}$$

$$= \frac{(50)(49)(48)(47)!}{3!(47)!}$$

$$= \frac{50 \times 49 \times 48}{3 \times 2 \times 1} = 19600$$

Similarly,

$$^{75}c_{5} = \frac{75!}{5!(75-5)!}$$

$$= \frac{(75)(74)(73)(72)(71)(70)!}{5!(70)!}$$

$$= \frac{75 \times 74 \times 73 \times 72 \times 71 \times 70}{5 \times 4 \times 3 \times 2 \times 1}$$ and so on

Illustration 1.2.1

A fair coin is tossed six times. What is the probability of obtaining four or more heads?

Solution:

When a fair coin is tossed, the probabilities of head and tail in case of an unbiased coinare equal, i.e.,

$P = q = \frac{1}{2}$ or 0.5

The probabilities of obtaining 4 heads is:

$$P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$$P(4) = {}^6C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{6-4}$$

$$P(4) = \frac{6!}{4!(6-4)!} (0.5)^4 (0.5)^2$$

$$P(4) = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(2 \times 1)} (0.0625)\ (0.25)$$

$$= \frac{720}{(24)(2)} (0.0625)\ (0.25)$$

$$= 15 \times 0.0625 \times 0.25 = 0.234$$

The probability of obtaining 5 heads is

$$P(5) = {}^6C_5 \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^{6-5}$$

$$P(5) = \frac{6!}{5!\,(6-5)!} (0.5)^5 (0.5)^1$$

$$P(5) = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(5 \times 4 \times 3 \times 2 \times 1)(1 \times 1)} (0.03125)\ (0.5)$$

$$= 6 \times 0.03125 \times 0.5 = 0.094$$

The probability of obtaining 6 heads is:

$$P(6) = {}^6C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^{6-6}$$

$$P(6) = \frac{6!}{6!(6-6)!} (0.5)^6 (0.5)^0$$

$$P(6) = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(6 \times 5 \times 4 \times 3 \times 2 \times 1)(1)} (0.015625)\ (1)$$

$= 1 \times 0.015625 \times 1$

$= 0.016$

The probability of obtaining 4 or more heads is:

$0.234 + 0.094 + 0.016 = 0.344$

Illustration 1.2.2

The incidence of a certain disease is such that on an average 20% of workers sufferfrom it. If 10 workers are selected at random, find the probability that:

i. Exactly 2 workers suffer from the disease

ii. Not more than 2 workers suffer from the disease

iii. At least 2 workers suffer from the disease

Solution:

Probability that a worker suffering from the disease $= \frac{20}{100} = \frac{1}{5}$

i.e., $p = \frac{1}{5}$ and

The probability of a worker not suffering from the disease i.e.,

$q = (1- \frac{1}{5}) = \frac{4}{5}$

By binomial probability law, the probability that out of 10 workers, 'x' workers sufferfrom a disease is given by:

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

$= {}^{10}C_{r} (\frac{1}{5})^{r} (\frac{4}{5})^{10-r}$; r = 0,1,2,3,…….10

The required probability that exactly 2 workers will suffer from the disease is given by:

$P(2) = {}^{10}C_{2}(\frac{1}{5})^{2} (\frac{4}{5})^{10-2}$

$P(2) = \frac{10!}{2!(10-2)!} (0.2)^{2}(0.8)^{8}$

$= \frac{10 \times 9 \times 8!}{(2 \times 1)\ (8)!} (0.04)\ (0.1677)$

$= 45\ (0.04)\ (0.1677)$

$= 0.302$

The required probability that not more than 2 workers will suffer from the disease is given by:

$P(0) + P(1) + P(2)$

$P(0) = {}^{10}C_0 (\frac{1}{5})^0 (\frac{4}{5})^{10-0} = 0.107$

$P(1) = {}^{10}C_1 (\frac{1}{5})^1 (\frac{4}{5})^{10-1} = 0.269$

$P(2) = {}^{10}C_2 (\frac{1}{5})^2 (\frac{4}{5})^{10-2} = 0.302$

Probability of not more than 2 workers suffering from the disease

$= 0.107 + 0.269 + 0.302 = 0.678$

We have to find out $P(r \geq 2)$ i.e., $P(r \geq 2) = 1 - P(0) - P(1)$

$= 1 - 0.107 - 0.269 = 0.624$

Thus, the probability at least two workers suffering from the disease is 0.624

1.2.2.1 Properties of Binomial Distribution

i. Binomial distribution is a discrete probability distribution.

ii. The shape and location of Binomial distribution changes as 'p' changes for a given 'n'.

iii. The mode of the binomial distribution is at the value of x which has the maximum probability. Binomial distribution has one or two modes.

iv. Mean of the binomial distribution increases as 'n' increases with 'p' remaining constant.

v. If 'n' is large and if neither 'p' nor 'q' is too close to zero binomial distribution may be approximately normal distribution.

vi. Binomial distribution has mean = np and S.D. = $\sqrt{npq}$

vii. If two independent random variables follow Binomial distribution, their sum also follows Binomial distribution.

1.2.2.2 Mean and Standard deviation of Binomial Distribution

The Binomial distribution has expected values of mean and a standard deviation (σ).

We can express the mean of the binomial distribution as:

Mean = np

Where n= number of trials; p=probability of success and, we can calculate the standard deviation by:

$$\sigma = \sqrt{npq}$$

Where n = Number of trials; p = probability of success and

q = probability of failure =1- p

Illustration 1.2.3

If the probability of defective bolts is 0.1, find the mean and standard deviation for thedistribution of defective bolts in a total of 300.

Solution:

P=0.1, n=300

Hence mean = np = 300 x 0.1 =30

Thus we can expect 30 bolts to be defective

Standard Deviation = $\sqrt{npq}$

n =300, p = 0.1, q = 1 – p = 1 – 0.1 = 0.9

$\sigma = \sqrt{npq} = \sqrt{300 \times 0.1 \times 0.9} = 5.196$

1.2.2.3 Fitting a Binomial Distribution

When a binomial distribution is to be fitted to observed data, the following procedure is adopted:

i. Determine the value of 'p' and 'q'. If one of these values is known, the other canbe found out by the simple relationship p =1-q and q =1-p. If p and q are equal, we say, the distribution is symmetrical. On the other hand, if 'p' and 'q' are not equal, the distribution is skewed. The distribution is positively skewed, in case 'p' is less than 0.5, otherwise it is negatively skewed.

ii. Expand the binomial (p = q) n. The power 'n' is equal to one less than the number of terms in the expanded binomial. For example, if 3 coins are tossed (n = 3) there will be four terms, when 5 coins are tossed (n = 5) there will be 6 terms and so on.

iii. Multiply each term of the expanded binomial by N (the total of frequency), in order to obtain the expected frequency in each category.

1.2.2.4 Utilities of Binomial distribution

- Useful in decision making situations in business.

- It serves as the basis for hypothesis testing and confidence interval estimation in situation involving binary outcomes, such as success/failure, yes/no, or presence/absence.
- In finance, binomial models are employed to price options, where the underlying assets can either increase or decrease in value at each time step, akin to the binomial outcomes of success or failure.
- Applied in quality control.
- Used in sampling plans for carrying out inspection on the article drawn in the sample.

Summarised Overview

The binomial distribution, also known as the Bernoulli distribution, is a fundamental tool in probability theory widely applied across various fields. It models the probability of observing a certain number of successes in a fixed number of trials, each with only two possible outcomes. Its simplicity makes it valuable in decision-making scenarios, quality control processes, and sampling plans. By assuming independence between trials and a constant probability of success, it accurately describes real-life events and aids in assessing the fairness of random processes like coin flips or dice rolls. Overall, the binomial distribution serves as a foundational concept in understanding and analyzing probabilistic phenomena in business and beyond.

Self-Assessment Question

1. What do you mean by Discrete Probability Distribution?
2. Explain the procedure involved when a binomial distribution is to be fitted to observed data
3. What is Binomial distribution?
4. How is the expected value and standard deviation computed in a binomial distribution?
5. What is a random variable?
6. State the assumptions of binomial distribution.
7. What are the utilities of binomial distribution?
8. List out the properties of binomial distribution.

Assignments

1. A survey of the residents of a particular city showed that 20% preferred a white telephone over any other colour available. What is the probability that more than one half of the next 20 telephones installed in this city will be white. (Answer- 0.0006)
2. The probability that a patient recovers from a delicate heart operation is 0.9. What is the probability that exactly 5 of the next 7 patients having this operation survive? (Answer- 0.124)
3. A multiple choice quiz has 15 questions, each with 4 possible answers of which only one is the correct answer. What is the probability that sheer guess work is:
    a. Exactly 2 correct answers (Answer- 0.15)
    b. 5-10 correct answers (Answer- 0.31)
    c. No correct answers (Answer- 0.0.134)
4. Eight unbiassed coins were tossed simultaneously. Find the probability of getting
    a. Exactly 4 heads (Ans. 0.273)
    b. No head at all (Ans 0.0039)
    c. 6 or more heads (Ans 0.144)
    d. Number of heads varying from 3 to 5 (Ans 0.710)
5. The number of tosses of coin that are needed so that the probability pf getting at least one head being 0.875 is
    a. 2
    b. 3
    c. 4
    d. 5

(Ans -ii is the correct answer)

Reference

1. Chawla, D; and Sondhi, N., (2016). *Research Methodology: Concepts and Cases*. Vikas Publishing House Pvt. Ltd.
2. G.C. Beri. (2017). *Business Statistics*. Tata McGraw, Hill New Delhi
3. Barry Render. (2022). *Quantitative Analysis for Management*. Prentice Hall of India
4. Gupta & Khanna. (2011). *Quantitative Techniques for Decision Making*. Prentice Hall of India.

5. Anand Sharma. (2017). *Quantitative Techniques for decision making*. Himalaya Publishing House.
6. D.V.D. Vohra. (2021). *Quantitative Techniques for Management.* McGraw Hill.

Suggested Reading

1. Krishnaswami, O.R., Ranganathan, M., & Harikumar, P.N. (2016). *Research Methodology*. Himalaya Publishing House.
2. Gupta SP. (2021). *Statistical Methods*, S. Chand & Sons.
3. Levin & Rubin. (1986). *Quantitative Approaches for Management*, Pearson.
4. P.N. Arora, Sumeet Arora, S. Arora. (2010). *Comprehensive Statistical Methods*, S. Chand and Company Private Limited, New Delhi

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 3

Poisson Distribution

Learning Outcomes

After completing this unit, the learner will be able to:

- gain insights about the Poisson distribution and understand its properties.
- calculate the mean and variance of a Poisson distribution
- understand the utility of the Poisson distribution.

Background

The Poisson distribution finds application in various domains. Whether it is analyzing the number of customer arrivals at a service center, occurrences of defects in a manufacturing process, or even the number of earthquakes in a region over a period, the Poisson distribution provides a versatile tool for understanding and predicting random events.

So, whether you are managing a call center, optimizing inventory levels, or assessing risks in insurance, the Poisson distribution offers a powerful lens through which to understand the randomness of the world around us. The present unit deals with the concept of poisson distribution and its characteristics.

Keywords

Poisson distribution, Mean, Standard deviation

Discussion

1.3.1 Poisson Distribution

The Poisson distribution is a discrete distribution that measures the probability of a given number of events happening

- Discrete probability distribution

in a specified time period. It can also be used for the number of events in other specified interval types such as distance, area or volume. Poisson distribution is developed by a French mathematician Simeon Poisson. So, it is known by his name and deals with counting the number of occurrences of a particular event in a specific time interval or region of space. It is used in practice where there are infrequently occurring events with respect to time, volume (similar units), area, etc. For instance, the number of deaths or accidents occurring in a specific time, the number of defects in production, the number of workers absent per day etc.

The binomial distribution is determined by two parameters 'p' and 'n'. In a number of cases 'p' (the probability of success) may happen to be very small (even less than 0.01) and the 'n' (the number of trials) is large enough (like more than 50) so that their product 'np' remains a constant. This situation is termed as "Poisson Distribution", and it gives an approximation for binomial probability distribution formula i.e.,

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

The Poisson distribution process corresponds to a Bernoulli process with a very large number of trials (n) and a very low probability of success. Thus, it is also called the Law of improbable events. This would comparatively be simpler in dealing with and is given by the Poisson distribution formula as follows:

- Law of improbable events

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

Where, p (r) = Probability of success desired

r = 0, 1,2,3,4,……∞ (any positive integer)

e = a constant with value; 2.7183 (the base of natural algorithms)

m = mean of the Poisson distribution, i.e., np or the average number ofoccurrences of an event.

- Relation between Binomial and Poisson distribution

When n is large while the probabilities of occurrence of an event is close to 0, so that q = (1-p), the Binomial distribution is very closely approximated by the Poisson distribution with m= np.

1.3.1.1 Characteristics of the Poisson Distribution

a. It is a discrete probability distribution and it is the limiting form of the binomial distribution.

b. The range of the random variable is $0 \leq r < \infty$.

c. It consists of a single parameter "m" only. So, the entire distribution can be obtained by knowing this value only.

d. It is a positively skewed distribution. The skewness, therefore, decreases when'm' increases.

Measures of Central Tendency

In Poisson distribution, the mean (m) and the variance (s^2) represent the same value,i.e.,

Mean = variance = np = m

S.D (σ) = $\sqrt{variance} = \sqrt{np}$

Illustration 1.3.1

2% of the electronic toys produced in a certain manufacturing process turnout to be defective. What is the probability that a shipment of 200 toys will contain exactly 5% defectives? Also find the mean and standard deviation.

Solution:

n = 200

Probability of a defective toy (p) = $\frac{2}{100}$

= 0.02

Since, n is large and p is small, the Poisson distribution is applicable. Apply the formula as follows:

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

The probability of 5 defective pieces in 200 toys is given by:

$$P(5) = \frac{m^5\ e^{-m}}{5!}$$

Where,

m = np = 200 x 0.02 = 4

e = 2.7183 (constant)

$$P(5) = \frac{4^5\ 2.7183^{-4}}{5!}$$

$$= \frac{(1024)(0.0183)}{5 \times 4 \times 3 \times 2 \times 1}$$

$$= \frac{(1024)(0.0183)}{120}$$

$$= \frac{18.7392}{120}$$

$$= 0.156$$

Mean = np = 200 x 0.02 =4

S.D (σ) = $\sqrt{np} = \sqrt{4} = 2$

Illustration 1.3.2

On an average, a certain intersection results in 3 traffic accidents per month, what is the probability that in any given month that through this intersection no accidents will occur.

Solution

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

$$= \frac{3^0 \text{ x } 2.71818^{-3}}{0!}$$

$$= 0.0496$$

1.3.1.2 Fitting of a Poisson Distribution

To fit a Poisson distribution to a given observed data (frequency distribution), the procedure is as follows:

a. We must obtain the value of its mean i.e., m = np

b. The probabilities of various values of the random variables (r) are to be computed:

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

c. Each probability so obtained in step 2 is then multiplied by N (the total frequency) to get expected frequencies.

Summarised Overview

The Poisson distribution, named after Simeon Poisson, is a discrete probability distribution used to model the likelihood of a specific number of events occurring in a given time, area, or volume. It finds practical application in scenarios where events are infrequent but occur randomly, such as accidents, defects in production, or absenteeism. Derived from the binomial distribution under conditions of low success probability and a large number of trials, it simplifies complex calculations. With a single parameter, the mean (m), it offers a compact representation of the distribution. The mean and variance are equal, simplifying analysis. Its fitting process involves calculating the mean and then computing probabilities for different event counts, aiding in analyzing observed data. Overall, the Poisson distribution serves as a vital tool in understanding and predicting rare but significant events, providing a powerful tool for practical statistical analysis and decision-making.

Self-Assessment Question

1. What is a Poisson Distribution?
2. What kind of a probability distribution is Poisson distribution?
3. List out the utilities of Poisson distribution.
4. State the features of Poisson distribution.
5. How is a Poisson distribution different from Binomial distribution?
6. Write a short note about fitting of a Poisson distribution.
7. What are the conditions to be followed to perform Poisson distribution?
8. Why is Poisson distribution known as the Law of improbable events?

Assignments

1. The probability that a person dies from a certain respiratory infection is 0.002. Find the probability that fewer than 5 of the next 2000 persons so infected will die. (Answer- 0.6288)
2. Suppose on an average, one person in every 1000 is an alcoholic. Find the probability that a random sample of 8000 people will yield 6 to 8 alcoholics (Answer- 0.4013)
3. State the relationship between Binomial distribution and Poisson distribution.
4. If 3% of electric bulbs manufactured by a company are defective, find the probability that in a sample of 100 bulbs, exactly 5 bulbs are defective.

   (Answer: 0.1006)

Reference

1. Singh A.K. (2017). Tests, *Measurements and Research Methods in Behavioural Sciences*. Bharti Bhawan
2. Levin & Rubin. (1986). *Quantitative Approaches for Management*, Pearson.
3. Chawla, D; and Sondhi, N., (2016). *Research Methodology: Concepts and Cases*. Vikas Publishing House Pvt. Ltd.

4. G.C. Beri. (2017). *Business Statistics*. Tata McGraw, Hill New Delhi
5. Kothari, C. R., (2004). Research Methodology: Methods and Techniques. New AgeInternational (P) Limited.
6. Anand Sharma. (2017). *Quantitative Techniques for decision making*. Himalaya Publishing House.

Suggested Reading

1. Barry Render. (2022). *Quantitative Analysis for Management*. Prentice Hall of India
2. Krishnaswami, O.R., Ranganathan, M., & Harikumar, P.N. (2016). *Research Methodology*. Himalaya Publishing House.
3. Gupta SP. (2021). *Statistical Methods*, S. Chand & Sons.
4. P.N. Arora, Sumeet Arora, S. Arora. (2010). *Comprehensive Statistical Methods*, S. Chand and Company Private Limited, New Delhi
5. D.V.D. Vohra. (2021). *Quantitative Techniques for Management.* McGraw Hill.
6. Gupta & Khanna. (2011). *Quantitative Techniques for Decision Making*. Prentice Hall of India.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 4

Normal Distribution

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the concept of the normal distribution and its representation by the bell-shaped normal curve.
- identify the key properties of the normal distribution.
- recognize the utility of the normal distribution in statistical inference, characterizing uncertainties, and approximating other probability distributions.
- describe the normal approximation to the binomial distribution and normal approximation to the Poisson distribution, understanding the circumstances in which it is applicable.

Background

In our exploration of random variables, we have encountered discrete distributions, where outcomes are distinct and countable. However, when we consider the continuous distributions, the possibilities expand infinitely across a range of real numbers. Think of variables like heights of individuals, lengths of machine-cut pieces, or yearly rainfall data they span a continuum rather than discrete points.

When we examine the likelihood of different outcomes, we often see a pattern emerge that looks like a bell when graphed. This bell-shaped curve, known as the normal distribution, is incredibly versatile and applicable to many real-world situations. It is like a universal template that fits numerous problems snugly. That is why it is so crucial in statistics, especially when we are trying to understand the relationship between a sample we have collected and the larger population it represents. So, whenever you see that bell-shaped curve, you are likely dealing with the normal distribution, a key player in the world of statistics.

Keywords

Normal distribution, Bell shaped curve, Symmetrical, Continuous distribution

Discussion

1.4.1 Normal Probability Distribution

- Gaussian distribution

As Binomial and Poisson distributions are the most useful theoretical distributions of discrete random variables. Normal distribution is the most useful theoretical distribution for continuous variables. The normal distribution is useful in statistical inferences, in characterizing uncertainties in many real life situations, and in approximating other probability distributions. Normal distribution is also known as Gaussian distribution, as it is a probability distribution that is symmetric about the mean, showing that the data near the mean are more frequent in occurrence than data far from the mean.

Normal distribution was first discovered by De Moivre in 1733 as the limiting form of the binomial distribution. Many statistical data concerning business problems are displayed in the form of normal distribution. Height, weight, and dimensions of a product are some of the continuous random variables which are found to be normally distributed. This knowledge helps us in calculating the probability of different events in varied situations, which in turn is useful for decision making.

- Continuous probability distribution

The normal distributions have key characteristics that are easy to spot in graphs. To define a particular normal probability distribution, we need only two parameters i.e.,the mean and standard deviation (σ).

1.4.2 Normal distribution curve

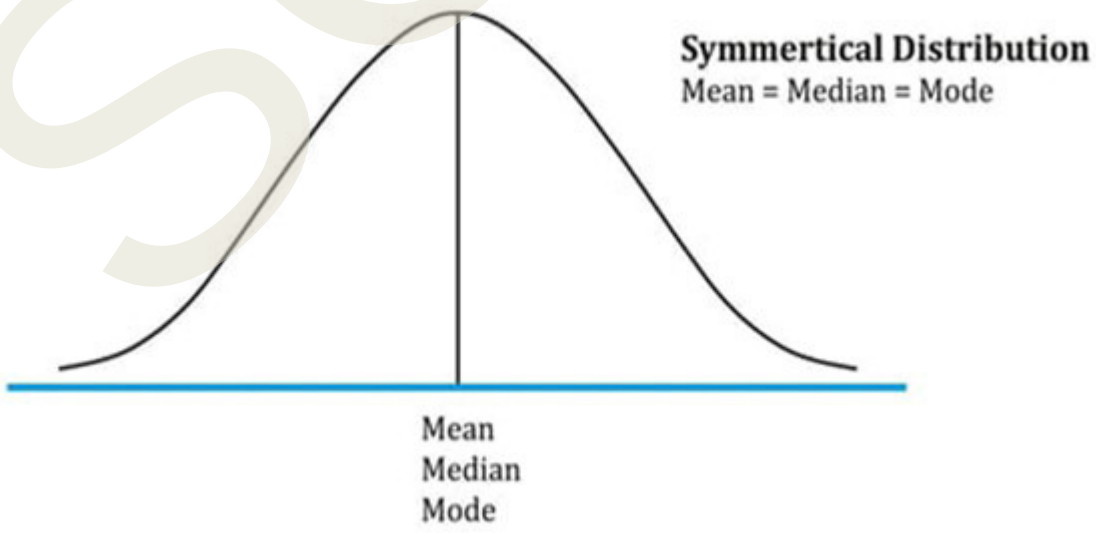


Figure 1.4.1 Normal probability distribution

Characteristics of normal distribution

i. The normal curve is a continuous curve.

ii. It is bell shaped.

iii. It is symmetrical about the mean.

iv. Mean, Median and Mode are equal for a normal curve.

v. The height of normal curve is at its maximum at the mean.

vi. There is only one maximum point, which occurs at the mean.

vii. The ordinate at the mean divides the whole area into two equal parts. The area to the right of the ordinate and that to the left are both 0.5.

viii. The curve is asymptotic to the base on either side: i.e., the curve approaches nearer and nearer to the base but it never touches the base.

ix. The co-efficient of skewness is 0 (i.e., $\beta = 0$).

x. The normal curve is unimodel i.e., it has only one mode.

xi. The point of intersection occurs at $\mu \pm \sigma$. At the point of inflexion, the curve changes from concavity to convexity.

xii. Q1 and Q3 are equidistant from median.

xiii. Mean deviation for normal curve is $\frac{4}{5}\sigma$ and QD is $\frac{2}{3}\sigma$.

xiv. Normal curve is mesokurtic. That is measure of kurtosis is β_2=3.

xv. No portion of the curve lies below the X axis.

xvi. Theoretically the range of the normal curve is $-\infty$ *to* $+\infty$. But practically the range is -3σ to $+3\sigma$.

xvii. A linear function of a number of normal variates is also a normal variate. This is known as additive property of normal distribution.

xviii. Area under the normal curve is distributed as follows:

Mean $\pm \sigma$ covers 68.27 % area

Mean $\pm 2\sigma$ covers 95.45 % area

Mean $\pm 3\sigma$ covers 98.73 % area

This is area property of the normal distribution

Utility of Normal Distribution

The study of the normal distribution is of central importance in statistical analysis because of the following reasons:

i. Most of the discrete probability distributions (Binomial distributions, Poisson distributions) tend to approximate normal distribution as 'n' becomes large.

ii. Almost all sampling distributions such as student's t- distribution, F-distributions, z-distributions, chi-square distributions etc. conform to normal distribution for large values of n.

iii. The various tests of significance like t-test, F –test etc. are based on the assumption that the parent population from which the sample have been drawn follows Normal distribution.

iv. It is extensively used in large sampling theory to find estimates of parameters from statistics, confidence limits etc..

v. Normal distribution has the remarkable property stated in the central limit theorem. As per the theorem, when the sample size is increased the sample means will tend to be normally distributed. Central limit theorem gives the normal distribution its central place in the theory of sampling, since many important problems can be solved by this.

vi. In theoretical statistics as well as applied works many problems can be solved only under the assumptions of a normal population.

vii. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate. The normal curve is reasonably close to many distributions.

viii. It finds many applications in statistical quality control and industrial experiments. Many distributions in social and economic data are approximately normal e.g. Birth, death, etc. are normally distributed. In psychological and educational data many distributions are of normal type.

Area under the Normal Curve

- Probability of getting a value between two numbers

The area under the normal curve gives us the proportion of the cases falling between two numbers or the probability of getting a value between two numbers. Irrespective of the value of the mean (μ) and standard deviation (σ), for a normal distribution, the total area under the curve is 1.00.

The area under the normal curve is approximately distributed by its standard deviation as follows:

$\mu \pm 1\sigma$ covers 68.27 % area i.e., 34.135% area will lie on either side of μ

$\mu \pm 2\sigma$ covers 95.45 % area i.e., 47.725% will lie on either side of μ

$\mu \pm 3\sigma$ covers 98.73 % area i.e., 49.365 % will lie on either side of μ

1.4.3 Normal Approximation to binomial distribution

The continuous normal distribution can sometimes be used to approximate the discrete binomial distribution. It is often used in statistical inference. The binomial distribution is perfectly symmetric if p = 0.5 and has some skewness when $p \neq 0.5$. The normal approximation works best when p is close to 0.5 and when n is large. For a binomial random variable X

$\mu = np$ and $\sigma^2 = np\ (1\text{-}p)$

$z = \frac{x - \mu}{\sigma}$ has (approximately) the standard normal distribution

$z \sim N\ (0, 1)$

- Relation between p, q and n

If p and q are nearly equal (i.e., p is nearly 1/2), then the normal approximation is surprisingly good even for small values of n. However, when p and q are not equal, i.e., when p and q is small, even then the Binomial distribution tends to normal distribution but in this case the convergence is slow. By this, we mean that if p and q are not equal then for Binomial distribution to tend to Normal distribution we need relatively larger value of n as compared to the value of n required in case when p and q are nearly equal. Thus, the normal approximation to the Binomial distribution is better for increasing values of n and is exact in the limiting cases as $n \to \infty$.

1.4.4 Normal approximation to Poisson distribution

The Poisson distribution helps you understand the likelihood of getting a specific number of rare events in a fixed amount of time or space. Sometimes, when you are dealing with a lot of events and they are quite rare, it becomes challenging to use the Poisson distribution directly. So, statisticians use a trick to make the calculations easier. If the average number of events (mean) is relatively large (like more than 10), they can pretend that the Poisson distribution looks like a normal distribution (bell curve).

Conditions for Approximation:

- Large Mean: If, on average, you get a lot of events, the normal approximation tends to work well.
- Continuity Correction: When making approximations for specific values, a small adjustment is made to the boundaries to make the calculations smoother.

- Suitable for large and independent events

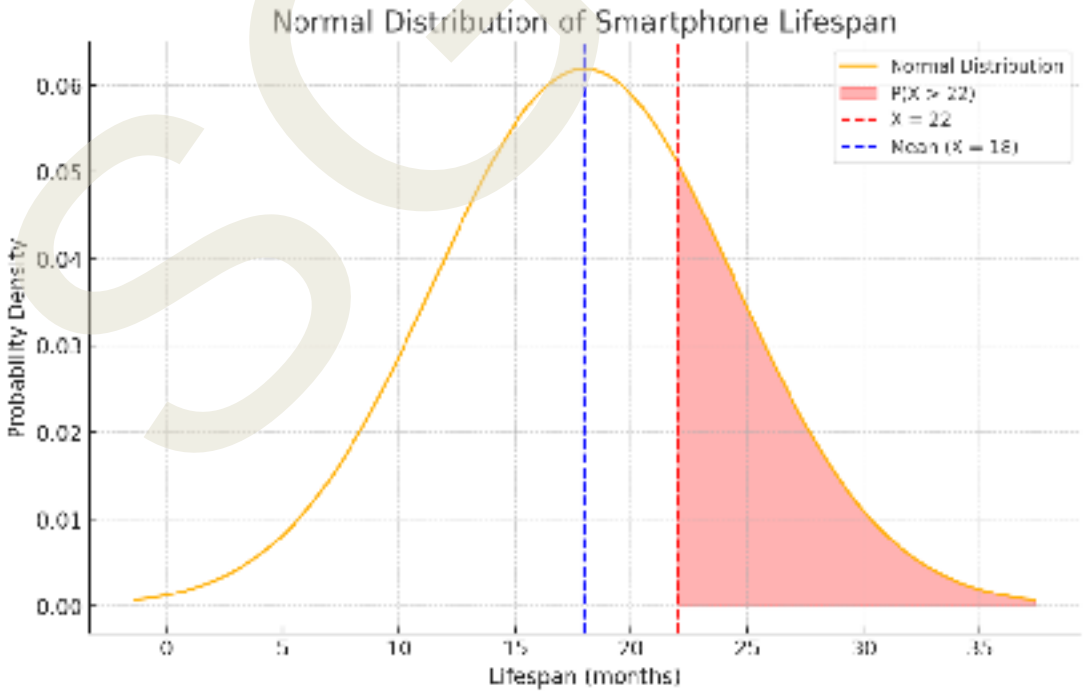
The normal distribution is mathematically convenient and has well-known properties, making calculations easier. It is like using a good approximation when the actual math gets too complicated. The approximation is better suited for large average event counts and events that are rare and independent. If the events are not rare or the mean is small, it's usually better to stick with the original Poisson distribution.

Illustration 1.4.1

An economist asserts that the average lifespan of a particular model of smartphone is 18 months, with a standard deviation of 6.45 months. What is the probability that a smartphone purchased by a customer will still be functional after 22 months.

Solution

For finding the required probability we are interested in the area of the problem of the normal curve as shaded and shown below:



$$Z = \frac{\bar{x} - \mu}{\sigma}$$

$$Z = \frac{22 - 18}{6.45}$$

$$= 0.62$$

The value from the Standard Normal Distribution Table (P no: 243) showing the area under the normal curve for Z = 0.62 is 0.2324. This means that the areas of the curve between μ = 18 and x = 22 is 0.2324. Hence the area of the shaded portion of the curve is (0.5-0.2324) = 0.2676, since the area of the entire right-hand portion of the curve always happen to be 0.5. Thus, the probability that the smartphone will still be functional after 22 months is 0.2676.

Illustration 1.4.2

Suppose the waist measurement (w) of 800 girls are normally distributed with mean 66 cms, and standard deviation 5 cms. Find the number of girls with waists-

i. Between 65cms and 70 cms;

ii. Greater than or equal to 72 cms.

Solution

M = 66

σ = 5cms

W (In cms)	$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{\bar{x} - 66}{5}$
65	$\frac{65-66}{5} = 0.2$
70	$\frac{70-66}{5} = 0.8$
72	$\frac{72-66}{5} = 1.2$

i. The probability that a girl has a waist between 65cms and 70 cms is given by:

$$P(65 \leq W \leq 70) = P(-0.2 \leq Z \leq 0.8)$$

$$= P(-0.2 \leq Z \leq 0) + P(0 \leq Z \leq 0.8)$$

$= P(0 \leq Z \leq 0.2) + P(0 \leq Z \leq 0.8)$

$= 0.0793 + 0.2881$

$= 0.3674$

Hence the group of 800 girls, the expected number of girls with waists between 65 cms and 70 cms is

800 x 0.3674 = 293.92 =294

ii. The probability that a girl has waist greater than or equal to 72 cms is given by

$P(W \geq 72) = P(Z \geq 1.2) = 0.5 - P(0 \leq Z \leq 1.2)$

$= 0.5 - 0.3849$

$= 0.1151$

Hence in a group of 800 girls, the expected number of girls with waist greater than or equal to 72 cms is

800 x 0.1151 = 92.08 = 92

Illustration 1.4.3

Assume the mean heights of NCC volunteers to be 68.22 inches with a standard deviation of 3.286 inches. How many NCC volunteers in a school of 1000 would you expect to be

i. Over 6 feet tall

ii. Below 5.5 feet

Assume heights to be normally distributed.

Solution

Let the variable X denote the height (in inches) of the NCC volunteers. Then we are given:

Mean = μ = 68.22 and standard deviation = σ = 3.286

An NCC volunteer will be over 6 feet tall if X is greater than 12 x 6 = 72 (Because x is height in inches)

When X =72, $Z = \frac{\bar{x}-\mu}{\sigma} = \frac{72-68.22}{3.286} = \frac{3.78}{3.286} = 1.15$

The probability that an NCC volunteer is over 6 feet = 72" tall is given by

$P(X > 72) = (Z > 1.15) = 0.5 - P(0 \leq Z \leq 1.15)$

$= 0.5 - 0.3749 = 0.1251$

Here, in a school of 1000 NCC volunteers, the number of volunteers over 6 feet tall is:

$= 1000 \times 0.1251$

=125.1

=125

The probability that a soldier is below 5.5 = 66 is given by:

$$P(X< 66) = P\left(Z < \frac{66 - 68.22}{3.286}\right)$$

$$= P\left(Z < \frac{-2.22}{3.286}\right)$$

$$= P(Z < -0.6756) = P(Z > 0.6756)$$

$$= 0.5\text{-}P(0<Z<0.6756)$$

$$= 0.5\text{-}0.2501$$

$$= 0.2499$$

Hence, the number of soldiers over 5.5 feet in a regiment of 1000 soldiers is

1000 x 0.2499 = 249.9 = 250

Illustration 1.4.4

The hourly wages of 1000 workmen are normally distributed around a mean of ₹70 and with a standard deviation of ₹5. Estimate the number of workers who hourly wages will be:

i. Between ₹69 and ₹72

ii. More than ₹75

iii. Less than ₹63

iv. Also estimate the lowest hourly wages of the 100 highest paid workers

Solution

Let the random variable X denote the hourly wages in rupees. Then X is a normal variable with $\mu = 70$ and $\sigma = 5$. The standard normal variable corresponding to X is

X	$Z = \frac{\bar{x}-\mu}{\sigma}$
63	-1.4
69	-0.2
72	0.4
75	1

i. $P(69 < X < 72)$

$= P(-0.2 < Z < 0.4)$

$= P(-2 < 2 < Z < 0) + P(0 < Z < 0.4)$

$= P(0 < Z < 0.2) + P(0 < Z < 0.4)$

$= 0.0793 + 0.1554$

$= 0.2347$

Hence the required number of workers is 1000 x 0.2347 = 234.7 = 235

ii. We want $P(X > 75)$

$P(X > 75) = P(Z > 1)$

$= 0.5 - P(0 < Z < 1)$

$= 0.5 - 0.3413$

$= 0.1587$

Thus, the number of workers with hourly wages more than 75 is 1000 x 0.1587 = 158.7 = 159

iii. $P(X < 63)$

$= P(Z < -1.4)$

$= P(Z > 1.4)$

$= 0.5 - P(0 < Z < 1.4)$

$= 0.5 - 0.4192$

$= 0.0808$

Hence, the number of workers with hourly wages less than ₹ 63 is $1000 \times 0.808 = 80.8 = 81$

Proportion of the 100 highest paid workers is $\frac{100}{1000} = \frac{1}{10} = 0.10$

We want to determine $X = x_1$, Such that $(P\ X > x_1) = 0.10$

When $X = x_1$, $Z = \frac{\bar{x} - 70}{5} = z_{1\ (Say)}$

Then $P(Z > z_1) = 0.10 \rightarrow P(0 < Z < z_1) = 0.5 - 0.1 = 0.40$

From the normal probability table, we get

$Z = \frac{\bar{x} - 70}{5} = 1.28 \rightarrow x_1 = 70 + 5 \times 1.28 = 70 + 6.48 = 76.40$

Hence, the lowest hourly wages of the 100 highest paid workers are ₹ 76.40

Summarised Overview

The normal distribution, also known as the Gaussian distribution, is a pivotal concept in statistics, characterizing uncertainties in various real-life scenarios. Originating from De Moivre's work, it approximates many other probability distributions and is fundamental in statistical analysis. With its symmetrical bell-shaped curve, the normal distribution is described by mean and standard deviation parameters, simplifying complex data representations. It is utilized extensively in business and scientific fields, with applications ranging from quality control to psychological research. The area under the curve provides insights into probabilities and proportions, aiding decision-making. Furthermore, the normal distribution serves as a basis for approximating other distributions, such as the binomial and Poisson distributions, particularly when parameters meet certain criteria. Through these approximations, the normal distribution offers a bridge between discrete and continuous distributions, enhancing the understanding and application of statistical principles in diverse contexts. Overall, grasping the principles and characteristics of the normal distribution equips students with a versatile toolset for analyzing and interpreting data across various disciplines.

Self-Assessment Question

1. What is normal distribution?
2. Describe the key characteristics of the normal distribution.
3. How does the normal distribution approximate the binomial distribution?
4. Under what conditions does the normal approximation to the binomial distribution work best?
5. What is the central limit theorem, and why is it important in the study of the normal distribution?
6. How does the area under the normal curve provide insights into probability and proportion?
7. When does the normal distribution provide a good approximation to the Poisson distribution?
8. What are the utilities of the normal distribution in statistical analysis and decision-making?

Assignments

1. Give a normal distribution with Mean-300 and Standard deviation =50, what is the probability that x assumes a value greater than 362? (Answer: 0.1075)

2. Given a normal distribution with μ= 40 and σ= 6, find the value of x that has
    a. 5% of area above the curve (Answer: 49.87)
    b. 38% of area below the curve (Answer: 38.17)
3. Given a normal distribution with μ= 50 and σ= 10, find the probability that x assumes a value:
    a. 50-65 (Answer: 0.4332)
    b. 50-55(Answer: 0.1915)
4. In a normal distribution, 31% of the items are under 45 and 8 % are over 64. Find the mean and standard deviation of the distribution. (Answer: Mean 50, SD 10)
5. A normal distribution has 77 as mean. Find its standard deviation if 20% of the area under the curve lies to the right of 90. (Answer: SD 15.48)
6. Marks obtained by a number of students are assumed to be normally distributed with mean 50 and variance 36. If 4 students are taken at random, what is the probability that exactly two of them will have marks over 62? (Answer: 0.00298)

Reference

1. Anand Sharma. (2017). *Quantitative Techniques for decision making*. Himalaya Publishing House.
2. P.N. Arora, Sumeet Arora, S. Arora. (2010). *Comprehensive Statistical Methods*, S. Chand and Company Private Limited, New Delhi
3. D.V.D. Vohra. (2021). *Quantitative Techniques for Management.* McGraw Hill.
4. G.C. Beri. (2017). *Business Statistics*. Tata McGraw, Hill New Delhi

Suggested Reading

1. Gupta & Khanna. (2011). *Quantitative Techniques for Decision Making*. Prentice Hall of India.
2. Gupta SP. (2021). *Statistical Methods*, S. Chand & Sons.
3. Barry Render. (2022). *Quantitative Analysis for Management*. Prentice Hall of India
4. Levin & Rubin. (1986). *Quantitative Approaches for Management*, Pearson.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

02 BLOCK

PARAMETRIC AND NON PARAMETRIC TESTS

Block Content

Unit 1

Basic Concepts

Learning Outcomes

After completing this unit, the learners will be able:

- to identify the difference between a sample statistic and population parameter.
- to know about the basic concepts of testing of hypothesis.
- to gain insights on Type I and Type II errors.

Background

The impracticality to study all the elements in a population resulted in sampling. As you already know, the process of selecting the sample elements from the population to represent the characteristics of the population is called sampling. After studying the selected samples, we draw inference about the population which is called statistical inference. But an understanding of the various concepts related to statistical inference is essential to make conclusion.

The present unit deals with such basic concepts such as sample statistic, population parameter, hypothesis, process of testing hypothesis, errors in hypothesis testing and so forth.

Keywords

Statistics, Parameter, Hypothesis, Type I error, Type II error, Standard error

Discussion

2.1.1 Statistic and Parameter

A parameter is a function of the population values. It is

- Derived from population

- Derived from samples

a statistical measure derived from the population. A parameter is usually denoted using greek letters. For example, a population mean is a parameter which is represented using the symbol μ.

A function of the sample values is a sample statistic. It is a statistical measure derived from the sample. A sample statistic is usually denoted using english letters. For example, a sample mean is a statistic which is represented using the letter $\bar{x}$.

Thus measures obtained from population are population parameters while measures obtained from sample are sample statistics.

2.1.2 Statistical Inference

- Drawing conclusions

Statistical inference refers to the process of selecting and using a sample statistic to draw conclusions about the population parameter. Statistical inference deals with two types of problems.

They are:-

a. Testing of Hypothesis

b. Estimation

2.1.3 Hypothesis

- Tentative statements

Hypothesis is a statement subject to verification. More precisely, it is a quantitative statement about a population, the validity of which remains to be tested. In other words, hypothesis is an assumption made about a population parameter.

Testing of Hypothesis:

- Testing the validity of formulated hypothesis

Testing of hypothesis is a process of examining whether the hypothesis formulated by the researcher is valid or not. The main objective of hypothesis testing is whether to accept or reject the hypothesis.

Procedure for Testing of Hypothesis:

The various steps in testing of hypothesis involves the following: -

i. Set up a hypothesis:

- Assumption about population parameter

The first step in testing of hypothesis is to set up a hypothesis about population parameter. Normally, the researcher has to fix two types of hypothesis. They are null hypothesis and alternate hypothesis.

Null Hypothesis:-

Null hypothesis is the original hypothesis. It states that there

is no significant difference between the sample and population regarding a particular matter under consideration. The word "null" means 'invalid' or 'void' or 'amounting to nothing'. Null hypothesis is denoted by H_o. For example, suppose we want to test whether a medicine is effective in curing cancer. Hence, the null hypothesis will be stated as follows:-

- No significant difference

H_0: The medicine is not effective in curing cancer (i.e., there is no significant difference between the given medicine and other medicines in curing cancer disease.)

Alternate Hypothesis:-

Any hypothesis other than null hypothesis is called alternate hypothesis. When a null hypothesis is rejected, we accept the other hypothesis, known as alternate hypothesis. Alternate hypothesis is denoted by H_1. In the above example, the alternative hypothesis may be stated as follows:

- Hypothesis other than null hypothesis

H_1: The medicine is effective in curing cancer. (i.e., there is significant difference between the given medicine and other medicines in curing cancer disease.)

ii. Set up a suitable level of significance:

After setting up the hypothesis, the researcher has to set up a suitable level of significance. The level of significance is the probability with which we may reject a null hypothesis when it is true. For example, if level of significance is 5%, it means that in the long run, the researcher is rejecting true null hypothesis 5 times out of every 100 times. Level of significance is denoted by α (alpha).

- Probability of rejecting true H_0

α = Probability of rejecting H_0 when it is true.

Generally, the level of significance is fixed at 1% or 5%.

iii. Decide a test criterion:

The third step in testing of hypothesis is to select an appropriate test criterion. Commonly used tests are z-test, t-test, χ^2 test, F-test, etc.

iv. Calculation of test statistic:

The decision to accept or to reject a null hypothesis is made on the basis of a statistic computed from the sample. Such a statistic is called the test statistic. There are different types of test statistics. All these test statistics can be classified into two groups-Parametric Tests and Non-Parametric Tests. The next step is to calculate the value of the test statistic using appropriate

formula. The formula for computing the test statistic depends on the specific statistical test being used. For example, in the case of z-test, the formula is-

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where,

$\bar{x}$= sample mean

μ= population mean

σ = standard deviation

v. Making Decision:

- Accept or reject null hypothesis

Finally, we may draw conclusions and take decisions. The decision may be either to accept or reject the null hypothesis. If the calculated value is more than the table value, we reject the null hypothesis and accept the alternate hypothesis. If the calculated value is less than the table value, we accept the null hypothesis.

2.1.4 Errors in Testing of Hypotheses

In any test of hypothesis, decision has to be taken whether to accept or to reject a null hypothesis. The decision is based on the information supplied by the sample data. The four possibilities of the decision are:

a. Accepting a null hypothesis when it is true
b. Rejecting a null hypothesis when it is false
c. Rejecting a null hypothesis when it is true
d. Accepting a null hypothesis when it is false

It is clear that the possibilities (a) and (b) are correct decisions. But the possibilities (c) and (d) are errors.

- Rejecting a true H_0

Type I Error:

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

Type II Error:

- Accepting a false H_0

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

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

error.

- 1 – β

Power of a Test

Probability for rejecting the null hypothesis when the alternate hypothesis is true iscalled power of a test.

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

- 1 – Level of significance

Level of Confidence

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

Level of Confidence = 1 – Level of significance.

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

- 1 – Level of acceptance

Level of Significance

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

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

- 100% - Rejection Region

Acceptance Region

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

Acceptance Region = 100% - Rejection Region

- 100% - Acceptance region

Rejection Region (Critical Region)

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

Rejection Region = 100% - Acceptance region

- Total No. observations – No. of constraints

Degree of Freedom

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

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

Two-tailed Test

A two tailed test is one in which we reject the null hypothesis if the computed value of the test statistic is significantly greater than or lower than the critical value (table value) of the test statistic.Thus

- Critical regions is on both tails

in two tailed tests the critical region is represented by both tails. If we test the hypothesis at 10% level of significance, the size of the acceptance region is 90% and the size of the rejection region is 10% on both sides together.

One-tailed Test

- Critical region is either on left tail or right tail

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

2.1.5 Sampling Distribution

- Samples drawn from the same population

The distribution of all possible values which can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population is called Sampling distribution of that statistic.

2.1.6 Standard Error (S.E)

Standard Error is the standard deviation of the sampling distribution of a statistic. Standard error plays a very important role in the large sample theory. The following are the important uses of standard errors:-

a. Standard Error is used for testing a given hypothesis.

b. S.E. gives an idea about the reliability of a sample, because the reciprocal of S.E. is a measure of reliability of the sample.

c. S.E. can be used to determine the confidence limits within which the population parameters are expected to lie.

Summarised Overview

The unit dealt with the basic concepts related to statistical analysis required for drawing informed inferences. The differences between sample statistic, which is a statistical measure derived from the sample, and population parameter, which is a statistical measure derived from the population was discussed in the unit. From the sample statistic, we draw inferences about the population which is termed as statistical inferences. It includes hypothesis testing and estimation.

Hypotheses are tentative statements subject to verification. A step by step procedure must be followed for testing the validity of the hypothesis. The main two types of hypothesis are null hypothesis and the alternate hypothesis. There are chances for errors to creep in while testing hypothesis. Type I and Type II errors are the errors that can occur while testing the hypothesis. Type I error is the error which is committed by rejecting the null hypothesis even when it is true. Type II error is the error which is committed by accepting the null hypothesis even when it is wrong. The distribution of all possible values which can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population is called sampling distribution of that statistic. The standard deviation of the sampling distribution of a statistic is called the standard error.

Self-Assessment Question

1. Differentiate between sample statistic and population parameter.
2. What do you mean by statistical inference?
3. Explain the meaning and types of hypotheses.
4. What is sampling distribution?
5. Elucidate the process involved in testing hypotheses.
6. How does Type I error differ from Type II error?
7. What is a standard error?
8. State the symbols that are used to represent Type I and Type II errors.

Assignments

1. Identify and list the different formulas for calculating the test statistic.
2. Explain the Type I and Type II errors.

3. What is the concept of power of a test?
4. Develop Null hypothesis and Alternate hypothesis for a study that aims to test whether increase in income leads to increase in savings.

Reference

1. Tandon, B.C; (1979). *Research Methodology in Social Sciences*. Chaitanya Publishing House.
2. Chawla, D; and Sondhi, N., (2016). *Research Methodology: Concepts and Cases*. Vikas Publishing House Pvt. Ltd.
3. Kevin, S., (2021). *Research Methodology for Social Sciences*. Ane Books Pvt. Ltd.,New Delhi.
4. Kothari, C. R., (2004). *Research Methodology: Methods and Techniques*. New AgeInternational (P) Limited.
5. Krishnaswami, O.R., Ranganathan, M., & Harikumar, P.N. (2016). *Research Methodology*. Himalaya Publishing House.
6. Panneerselvam, R. (2014) *Research Methodology*. PHI Learning

Suggested Reading

1. Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education (8th ed.)*. Routledge.
2. Gupta, S.K. & Praneet Rangi. (2018). *Business Research Methodology*. Kalyani Publishers
3. Sharma R, N & Sharma R.K. (2019). *Research Methods in Social Science*. Atlantic Publishers.
4. Singh A.K. (2017). *Tests, Measurements and Research Methods in Behavioural Sciences*. Bharti Bhawan

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 2

Parametric Tests

Learning Outcomes

After completing this unit, the learners will be able to:

- explain the principles of hypothesis testing and the role of significance tests such as the Z-test, t-test, and F-test.
- identify scenarios where each type of test (Z-test, t-test, F-test) is applicable and appropriate.
- describe the test for proportion and its application in analyzing categorical data.
- define analysis of variance (ANOVA) and differentiate between one-way ANOVA and two-way ANOVA.

Background

Imagine you are a researcher investigating the effectiveness of a new teaching method on student performance. To draw meaningful conclusions, you need to analyze the data collected from a sample of students. However, you are faced with the challenge of dealing with variability and uncertainty inherent in the data.

Parametric tests come into play in such situations, offering a structured approach to analyze the data and draw inferences about the larger population. These tests rely on assumptions about the distribution of the data, such as normality and homogeneity of variances, to make accurate conclusions. Parametric tests play a crucial role in statistical analysis by providing powerful tools for hypothesis testing, comparison of means, and analysis of relationships between variables. Understanding their underlying principles, assumptions, advantages, and disadvantages is essential for conducting meaningful and valid statistical analyses.

Keywords

Parametric test, t-test, z-test, ANOVA

Discussion

2.2.1 Parametric Tests

- Assumptions about population

Parametric tests assume that the data follows a specific distribution, usually the normal distribution. These tests make assumptions about the population parameters, such as mean and variance. Parametric tests are powerful when these assumptions are met, but they may not be accurate when the assumptions are violated.

- Normally distributed

The statistical tests based on the assumption that population or population parameter is normally distributed are called parametric tests. The important parametric tests are:-

a. z-test

b. t-test

c. F-test

z-test:

z-test is applied when the test statistic follows normal distribution. It was developed by Prof. R.A. Fisher. The following are the important uses of z-test:-

i. To test the population mean when the sample is large or when the population standard deviation is known.

ii. To test the equality of two sample means when the samples are large or when thepopulation standard deviation is known.

iii. To test the population proportion.

iv. To test the equality of two sample proportions.

v. To test the population standard deviation when the sample is large.

vi. To test the equality of two sample standard deviations when the samples are large or when population standard deviations are known.

vii. To test the equality of correlation coefficients.

- Used when variance is known and sample size is large

z-test is used in testing of hypothesis on the basis of some assumptions. The important assumptions in z-test are:-

i. Sampling distribution of test statistic is normal.

ii. For finding standard error, sample statistics are used in place where population parameters are to be used.

t-test

t- distribution was originated by W.S. Gosset in the early 1900. t-test is applied when the test statistic follows t-distribution.

Uses of t-test are: -

i. To test the population mean when the sample is small and the population S.D. is unknown.

ii. To test the equality of two sample means when the samples are small and population

iii. S.D. is unknown.

iv. To compare the means of two groups.

v. To test the significance of correlation coefficients.

The following are the important assumptions in t- test:-

- Used when variance is unknown and sample size is small

i. The population from which the sample drawn is normal.

ii. The sample observations are independent.

iii. The population S.D. is unknown.

iv. have a similar amount of variance within each group being compared

F-test:

F-test is used to determine whether two independent estimates of population variance significantly differ or to establish both have come from the same population. For carrying out the test of significance, we calculate a ratio, called F-ratio. F-test is named in honour of the great statistician R.A. Fisher. It is also called Variance Ratio Test.

F-ratio is defined as follows:-

$$F = \frac{S_1^2}{S_2^2}$$

- Variance ratio test

$$S_1^2 = \frac{\Sigma(x-\bar{x}1)^2}{n_1 - 1}$$

$$S_2^2 = \frac{\Sigma(x-\bar{x}2)^2}{n_2 - 1}$$

While calculating F-ratio, the numerator is the greater variance and denominator is the smaller variance. So,

$$F = \frac{Greater\ variance}{Smaller\ variance}$$

2.2.2 Testing of Given Population Mean

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

Procedure:

1. **Set up H_0 and H_1**

H_0: There is no significant difference between sample mean and population mean (i.e., $\mu = \mu_0$)

H_1: There is significant difference between sample mean and population mean (i.e., $\mu \neq \mu_0$)

2. **Decide the test statistic:**

The test statistic applicable here is z-test or t-test. If population S.D (i.e., σ) is known, apply z-test. If population S.D. (i.e., σ) is unknown but sample is large, apply z-test. If population S.D. (i.e., σ) is unknown but sample is small, apply t-test.

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

$\frac{z}{t}$ =Difference/Standard Error

Difference = Difference between sample mean and the given population mean

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

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

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

Where

σ = population SD.

s = sample S.D.

n = sample size

4. **Specify the level of significance**

If nothing is mentioned about the level of significance, take 5%.

5 **Fix the degree of freedom**

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

6. Locate the table value (critical value)

Table value of the of the test statistic at specified level of significance and fixed degree of freedom.

7. Compare the calculated value of test statistic with the table value and decide whether to accept or reject the null hypothesis.

- Test whether population mean and sample mean are same

If calculated value of the test statistic is numerically less than the table value, the null hypothesis is accepted. If calculated value of the test statistic is numerically more than the table value, the null hypothesis is rejected.

Illustration 2.2.1

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

Solution

H_0: There is no significant difference between sample mean and population mean (i.e., $\mu = 15200$)

H_1: There is significant difference between sample mean and population mean (i.e., $\mu \neq 15200$)

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

$$z = \frac{D}{SE}$$

$$D = \bar{x} - \mu$$

$$= 15269 - 15200$$

$$= 69$$

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{1248}{\sqrt{100}} = 124.8$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z = \frac{69}{124.8}$$

$$= 0.553$$

Level of significance = 5%

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

Table value (Critical value) at 5 % level of significance and infinity degree of freedom is 1.96. Since calculated value of z is less than the critical value, H_0 is accepted. That is, there is no significant difference between sample mean and population mean μ=15200. So, we may conclude that the claim of the company is valid.

Illustration 2.2.2

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

Solution

H_0: There is no significant difference between sample mean and population mean (i.e.; μ = 100)

H_1: There is significant difference between sample mean and population mean (i.e., μ ≠ 100)

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

$$z = \frac{D}{SE}$$

$D = \bar{x} - \mu$

$100 - 99 = 1$

$$S\,E = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{8}{\sqrt{400}} = 0.4$$

$$z = \frac{1}{0.4} = 2.5$$

Level of significance = 5%

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

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

Since calculated value of z is more than the critical value, H_0 is rejected. H_1 is accepted. That is, there is significant difference between sample mean and population mean. So, we may conclude that μ ≠ 100

Illustration 2.2.3

A random sample of 200 bottles of talcum powder gave an

average weight of 49.5 gram with a S.D. of 2.1 gram. Do we accept the hypothesis of weight per bottle is 50 gram at 1% level of significance?

Solution

H_0: There is no significant difference between sample mean and population mean (i.e., $\mu = 50$)

H_1: There is significant difference between sample mean and population mean (i.e., $\mu \neq 50$)

Since sample is large, the test statistic applicable here is z-test.

$$z = \frac{D}{SE}$$

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

$$SE = \frac{s}{\sqrt{n}}$$

$$= \frac{2.1}{\sqrt{200}}$$

$$= 0.148$$

$$z = \frac{-0.5}{0.148} \text{ (Calculated value)} = -3.38$$

Level of significance = 1%

Degree of freedom= infinity (population is large)

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

Since calculated value of z is more than the critical value, H_0 is rejected. H_1 is accepted. That is,there is significant difference between sample mean and population mean. So, we may conclude that $\mu \neq 50$ gram.

2.2.3 Testing of Significance of the Difference between Two Sample Means

- Test whether two sample means are same

1. Setting of hypotheses

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

Procedure

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

H_1: There is significant difference between two sample means

(i.e., $\mu_1 \neq \mu_2$)

2. Decide the test statistic:

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

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

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

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

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

$\frac{z}{t}$ = Difference/Standard Error

Difference = Difference between two sample means

Standard error = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}$ (If population S.D. are known)

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

Standard error = $\sqrt{\frac{(n_1 s_1^2 + n_2 s_2^2)}{n_1 + n_2 - 2} \text{ x } \frac{1}{n_1} + \frac{1}{n_2}}$ (If population S.D. are unknown, and samples are small)

Where

σ_1 = population S.D. of sample 1

s_1 = sample S.D. of sample 1

n_1 = sample size of sample 1

σ_2 =population S.D. of sample 2

s_2 = sample S.D. of sample 2

n_2 = sample size of sample 2

4. Specify the level of significance.

If nothing is mentioned about the level of significance, take 5%.

5. Fix the degree of freedom

For z-test, d.f = infinity;

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

6. Locate the table value

Critical value of the test statistic at specified level of significance and fixed degree of freedom.

7. Compare the calculated value of test statistic with the table value

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

Illustration 2.2.4

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

Solution

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

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

H_1: There is significant difference between two sample means (i.e., $\mu_1 \neq \mu_2$)

Since population S.D. are given, the test statistic applicable here is z-test.

$$z = \frac{\text{Difference}}{\text{SE}}$$

Difference = $\bar{x}_1 - \bar{x}_2 = 210 - 200 = 10$

SE= $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}$ (Population S.D. are known. For the entire State S.D. is 11)

$$= \sqrt{\frac{11^2}{100} + \frac{11^2}{150}}$$

$$= \sqrt{1.21 + 0.81}$$

$= \sqrt{2.02} = 1.42$

$Z = \frac{10}{1.42} = 7.04$

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

Since the calculated value of Z is more than the table value, H_0 is rejected. We accept H_1. So we may conclude that there is significant difference in the mean yields of crops in two districts.

Illustration 2.2.5

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

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

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

Solution

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

H_1: There is significant difference between two sample means (i.e., $\mu_1 \neq \mu_2$)

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

$$z = \frac{\text{Difference}}{\text{SE}}$$

Difference $= \bar{x}_1 - \bar{x}_2$

$= 1300 - 1248$

$= 52$

$$\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}} \text{ (If population S.D.s are unknown)}$$

$$= \sqrt{\frac{82^2}{100} + \frac{93^2}{100}}$$

$$= \sqrt{67.24 + 86.49}$$

$$= \sqrt{153.73}$$

= 12.4

$z = \frac{52}{12.4} = 4.19$

Level of significance = 5%

Degree of freedom = infinity

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

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

Illustration 2.2.6

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

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

Solution

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

H_1 : There is significant difference between two sample means (i.e., $\mu_1 \neq \mu_2$)

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

$$t = \frac{\text{Difference}}{\text{SE}}$$

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

$$\text{Standard error} = \sqrt{\frac{(n_1 s_1^2 + n_2 s_2^2)}{n_1 + n_2 - 2} \times \frac{1}{n_1} + \frac{1}{n_2}}$$

$$= \sqrt{\frac{(10\,x12^2 + 8\,x\,14^2)}{10+8-2} \times \frac{1}{10} + \frac{1}{8}}$$

$$= \sqrt{\frac{3008}{16} \times \frac{1}{10} + \frac{1}{8}}$$

$$= \sqrt{188 \times \frac{1}{10} + \frac{1}{8}}$$

$= \sqrt{42.3} = 6.5$

$t = \frac{70}{6.5} = 10.77$ (Calculated value)

Level of significance = 5%

Degree of freedom = 10 + 8 - 2 = 16

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

Since the calculated value of t is more than the table value, H0 is rejected. We accept H_1. (i.e; $\mu 1 \neq \mu 2$).

So, we may conclude that the lives of products produced in two batches are not same.

Illustration 2.2.7

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

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

Test whether the group means are equal.

Solution

Here we have to find the Means and S.D.s of the two samples.

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

Mean of Group I ($\bar{x}$) = $\frac{333}{9}$ = 37

Mean of Group II($\bar{x}$) = $\frac{238}{7}$ =34

S.D of Group I= $\sqrt{\frac{\sum(x-\bar{x}_1)^2}{n}}$

$= \sqrt{\frac{1134}{9}}$

$= \sqrt{126}$

S_1=11.22

S.D of Group II= $\sqrt{\frac{\sum(x-\bar{x}_2)^2}{n}}$

$= \sqrt{\frac{386}{7}}$

$= \sqrt{55.14}$

S_2 = 7.42

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

H_1: is significant difference between two sample means (i.e., $\mu_1 \neq \mu_2$)

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

$$t = \frac{\text{Difference}}{\text{SE}}$$

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

$$\text{SE} = \sqrt{\frac{(n_1 s_1^2 + n_2 s_2^2)}{n_1 + n_2 - 2} \text{ x } \frac{1}{n_1} + \frac{1}{n_2}}$$

$$= \sqrt{\frac{9 \text{ x } 126 + 7 \text{ x } 55.14)}{9+7-2} \text{ x } \frac{1}{9} + \frac{1}{7}}$$

$$= \sqrt{\frac{1520}{14} \text{x } 0.254}$$

$= \sqrt{27.58}$

$= 5.25$

$t = \frac{3}{5.25} = 0.571$ (Calculated Value)

Level of significance = 5%

Degree of freedom= 9 + 7 - 2 = 14

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

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

2.2.4 Testing of Significance of the Difference in case of Dependent Samples (Paired observations)

Here the observations in one sample are some way related to the observations in the other.Therefore they are called paired observations. The test statistic applicable here is t-test.

Procedure:

a. Set up H_0 and H_1

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

b. Decide test statistic:

Since the paired data are comparatively less, the test statistic applicable here is always t-test.

c. Apply the appropriate formula for computing the value of the test statistic.

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

Where,

d is Arithmetic mean of the difference between the values

SE is $\frac{S}{\sqrt{N-1}}$ [s = standard deviation of the difference]

d. Specify the level of significance.

Take 5%, if nothing is mentioned in the question.

e. Fix the degree of freedom.

d.f = n – 1 , where n= Number of pairs of observations.

f. Locate the critical value of the test statistic (t-test) at specified level of significance and fixed degree of freedom.

g. Compare the calculated value of test statistic with the table value

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

Illustration 2.2.8

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

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

Test whether there is any significant difference in their performance.

Solution

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

Test statistic applicable here is t-test.

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

Computation of men and standard deviation of the difference between the values

Score (Before)	Score (After)	Difference (d)	d^2
67	70	3	9
24	38	14	196
57	58	1	1
55	58	3	9
63	56	-7	49
54	67	13	169
56	68	12	144
68	72	4	16
33	42	9	81
43	38	-5	25
		$\sum d = 47$	**$\sum d^2 = 699$**

Arithmetic mean of d values = $\frac{47}{10} = 4.7$

SD of values = $\sqrt{\frac{\sum d^2}{n} - (\frac{\sum d}{n})^2}$

$= \sqrt{\frac{699}{10} - (\frac{47}{10})^2}$

$= \sqrt{69.9 - 22.09}$

$= \sqrt{47.81}$

$= 6.91$

SE = $\frac{6.91}{\sqrt{10-1}} = 2.3$

t = $\frac{4.7}{2.3} = 2.04$

Level of significance = 5%

Degree of freedom = 10-1 = 9

Table value (critical value) of t at 5% level of significance and 9 degree of freedom is 2.262. Since the calculated value of t is less than the critical value, the null hypothesis is accepted. So, we may conclude that there is no significant difference in the performance of the students.

2.2.5 Testing of Given Population Proportion

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

Procedure:

a. Set up H_0 and H_1:

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

(i.e., H_0: P = p)

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

(i.e., H_0: P ≠ p)

b. Decide the test statistic:

The test statistic applicable here is Z-test

c. Apply appropriate formulae for computing the value of Z (i.e., calculated value):

$Z = \frac{D}{SE}$

i.e., $Z = \frac{p - P}{SE}$

Where p = sample proportion,

P =Population proportion

$SE = \sqrt{\frac{PQ}{n}}$ Where, Q = 1- P

d. Decide the level of significance

Take 5%, if nothing is mentioned in the question.

e. Fix the degree of freedom

(Infinity d.f.)

f. Locate the table value of Z

Locate the value at specified level of significance and fixed degree of freedom.

g. Compare the calculated value of Z with the table value and decide whether to accept or reject the null hypothesis.

If calculated value of Z is numerically less than the table value, the null hypothesis is accepted. If calculated value of Z is numerically more than the table value, the null hypothesis is rejected.

Illustration 2.2.9

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

Solution

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

H_1: There is significant difference between sample proportion and population proportion (i.e., H_0:P≠ 0.02)

$Z = \frac{p - P}{SE}$

P = 0.02,

$p = \frac{30}{500} = 0.06$,

Q = 1 – P

=1 - 0.02

= 0.98,

n =500

$$SE = \sqrt{\frac{PQ}{n}}$$

$$= \sqrt{\frac{0.02 \times 0.98}{500}}$$

$$= \sqrt{\frac{0.0196}{500}}$$

$$= \sqrt{0.0000392}$$

$$= 0.0063$$

$$Z = \frac{0.06 - 0.02}{0.0063} = \frac{0.04}{0.0063} = 6.349$$

Level of significance = 5%

Degree of freedom = infinity

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

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

2.2.6 Testing of the Significance of the Difference between Two Sample Proportions

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

Procedure:

a. Set up H_0 and H_1

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

H_1: There is significant difference between two sample proportions (i.e., $p_1 \neq p_2$)

b. Decide the test statistic

The test statistic applicable here is Z-test.

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

Z = Difference/Standard Error i e; $Z = \frac{p_1 - p_2}{SE}$

Where p_1 and p_2 are the proportions of two samples

$$SE = \sqrt{p_0 q_0 (\frac{1}{n_1} + \frac{1}{n_2})}$$

Where,

$$p_0 = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$q_0 = 1 - p_0$

$n_1 = number\ of\ events\ in\ sample\ number\ one$

$n_2 = number\ of\ events\ in\ sample\ number\ two$

d. Specify the level of significance.

If nothing is mentioned about the level of significance, take 5%.

e. Fix the degree of freedom (d.f. = infinity)

f. Locate the table value (critical value)

Locate the table value (critical value) of the test statistic at specified level of significance and fixed degree of freedom.

g. Compare the calculated value of test statistic with the table value

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

Illustration 2.2.10

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

Solution

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

H_1: There is significant difference between two districts regarding the coffee drinking habits of people (i.e., $p_1 \neq p_2$)

The test statistic applicable here is Z-test.

Z = Difference/Standard Error i e; $Z = \frac{p_1 - p_2}{SE}$

$P_1 = \frac{400}{800} = 0.5$

$P_2 = \frac{450}{1000} = 0.45$

$SE = \sqrt{p_0 q_0 (\frac{1}{n_1} + \frac{1}{n_2})}$

$n_1 = 800, n_2 = 1000$

$p_0 = \frac{(800 \times 0.5 + 1000 \times 0.45)}{(800 + 1000)}$

$= \frac{850}{1800} = 0.472$

$q_0 = 1 - 0.472 = 0.528$

$\therefore SE = \sqrt{0.472 \times 0.528 (\frac{1}{800} + \frac{1}{1000})}$

$= \sqrt{0.249 (0.00125 + 0.001)}$

$= \sqrt{0.249 \times 0.00225}$

$= \sqrt{0.00056}$

$= 0.0237$

$\therefore Z = \frac{(0.5 - 0.45)}{0.0237}$

$= \frac{0.05}{0.0237}$

$= 2.11$

Level of significance = 5%.

Fix the degree of freedom = infinity

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

Since the calculated value of Z is more than the table value, null hypothesis is rejected. Alternate hypothesis is accepted. $p_1 \neq p_2$. So, we may conclude that there is significant difference between the two districts regarding the coffee drinking habits of people.

2.2.7 Analysis of Variance

- Tests three or more sample groups

The testing of hypotheses so far discussed consists of different sample groups which do not exceed two. If there are three or more sample groups, the testing of equality of them cannot be done in anyof the methods which have already been discussed. The testing of significance of the difference among three or more samples is generally done by using the technique of analysis of variance. In case of analysis of variance, as part of testing procedure, we have to prepare a separate statement called Analysis of Variance Table or ANOVA Table. Therefore, this type of testing of hypothesis is also called analysis of variance. The test statistic used for Analysis of Variance is F-test. F-test is a parametric test.

Types of Analysis of Variance

There are two types of Analysis of variance.They are:

i. One-way classification of data (One way analysis of variance)

ii. Two-way classification of data (Two way analysis of variance)

2.2.7.1 One-way classification of data (One way analysis of variance)

- Grouped based on single criteria

In one way classification, observations are classified into different groups on the basis of a single criterion. Suppose we want to study about the yield of a particular crop. You know there a number of factors which influence the productivity of crops. If we undertake this study to know the effect of quality of seed on the yield of crop, it is called one- way analysis of variance. Here yield of crops based on different seed must be given in columns. In other words, in case of one-way analysis of variance, the samples must always be in columns.

Types of variances in One-way ANOVA

a. Variance between samples (Columns): This is the net result of the variation of different sample means from grand mean. Grand mean is the mean of all the observations coming under all sample groups.

b. Variance within the sample: This is the net result of variations of different items of the sample from the respective sample means.

c. Variance about the sample: This is the sum of the variance between samples and the variance within the sample.

Proforma of One-way ANOVA Table

One-way ANOVA Table				
Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Sum of Squares**	**F-Ratio**
Between samples	SSC	C – 1	MSC = $\frac{SSC}{C-1}$	F = $\frac{MSC}{MSE}$ Or
Within Sample	SSE	N – C	MSE = $\frac{SSE}{N-C}$	F = $\frac{MSE}{MSC}$
Total	**SST=**	**N –1**		

SSC = Sum of Squares between Columns (Samples)

SSE = Sum of Square within Column (Sample)

SST = Sum of Square Total

MSC = Mean Sum of Squares between Columns (Samples)

MSE = Mean Sum of Squares within Column (Sample)

C = Number of Columns (Samples)

Procedure for carrying out One-way Analysis of variance

a. Set up H_0 and H_1.

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

b. Decide the test statistic

Test statistic applicable here is F-test

c. Apply the appropriate formula for computing the value of f-test.

$$F = \frac{\text{Larger Variance}}{\text{Smaller Variance}}$$

$$\frac{MSC}{MSE} \text{ or } \frac{MSE}{MSC}$$

i. Find SST.

SST = Sum of square of all items – $\frac{T^2}{N}$

Where

T = Total of all observations,

N = Total Number of observations

$\frac{T^2}{N}$ is generally called correction factor

ii. Find SSC.

$$SSC = \left[\frac{(\varepsilon x1)2}{n1}\right] + \left[\frac{(\varepsilon x1)2}{n1}\right] + \left[\frac{(\varepsilon x1)2}{n1}\right] + \text{.....-} \frac{T^2}{N}$$

Where

$\sum x_1$= sum of items in the first column

$\sum x_2$= sum of items in the second column

n_1= number of items in the first column

n_2 = number of items in the second column

iii. Draw one-way ANOVA Table and enter the values of SST and SSC

iv. Find the value of SSE.

SSE = SST – SSC

v. Find the degree of freedom in the third column as indicated in the proforma.

vi. Find MSC.

$$MSC = \frac{SSC}{C-1}$$

vii. Find MSE.

$$MSE = \frac{SSE}{N-C}$$

viii. Find F–Ratio.

$$F = \frac{\text{Larger Variance}}{\text{Smaller Variance}}$$

$\frac{MSC}{MSE}$ or $\frac{MSE}{MSC}$

d) Specify the level of significance.

Take 5%if nothing is mentioned.

e) Fix the degrees of freedom.

Here we have to fix a pair of d.f. If 'F' is obtained by using $F=\frac{MSC}{MSE}$, then pair of d.f. is (d.f. of MSC, d f of MSE)

If 'F' is obtained by using $F=\frac{MSE}{MSC}$, then pair of d.f. is (d.f. of MSE, d f of MSC)

f) Obtain table value of F at specified level significance and fixed degree of freedom.

g) Compare the Calculated value of F with the Table value

Compare the Calculated value of F with the Table value and

decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, H_0 is accepted. If calculated value is more than the table value, H_0 is rejected.

Illustration 2.2.11

Four varieties of a crop were grown on 3 plots, and the following yield was obtained. You are required to test whether there is significant difference in the productivity of seeds:

Plot	Variety of Seeds			
	P	Q	R	S
I	10	7	8	5
II	9	7	5	4
III	8	6	4	4

Solution

H_0: There is no significant difference in the productivity of seeds.

H_1: There is significant difference in the productivity of seeds.

Test Statistic applicable here is F-test

$$F = \frac{\text{Larger Variance}}{\text{Smaller Variance}}$$

$$\frac{MSC}{MSE} \text{ or } \frac{MSE}{MSC}$$

T= 10+7+8+5+9+7+5+4+8+6+4+4 =77

$$SST = 10^2 + 7^2 + 8^2 + 5^2 + 9^2 + 7^2 + 5^2 + 4^2 + 8^2 + 6^2 + 4^2 + 4^2 - \frac{T^2}{N}$$

$$= 100+49+64+25+81+49+25+16+64+36+16+16 - \frac{77^2}{12}$$

$$= 541 - (5929/12) = 541 - 494 = 47$$

$$SSC = \left[\frac{(\sum x1)^2}{n1}\right] + \left[\frac{(\sum x2)^2}{n2}\right] + \left[\frac{(\sum x3)^2}{n3}\right] + \ldots - \frac{T^2}{N}$$

$$= \left[\frac{(10+9+8)^2}{3}\right] + \left[\frac{(7+7+6)^2}{3}\right] + \left[\frac{(8+5+4)^2}{3}\right] + \left[\frac{(5+4+4)^2}{3}\right] - \frac{T^2}{N}$$

$$= \frac{729}{3} + \frac{400}{3} + \frac{289}{3} + \frac{169}{3} - \frac{77^2}{12}$$

$$= \frac{1587}{3} - 494$$

$$= 529 - 494$$

= 35

SSE = SST- SSC

= 47-35 = 12

Level of significance = 5%

One-way ANOVA Table				
Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Sum of Squares**	**F-Ratio**
Between samples	SSC = 35	C – 1= 4 -1 = 3	MSC = $\frac{SSC}{C-1}$ =11.67	F = $\frac{MSC}{MSE}$ $\frac{11.67}{1.5}$ = 7.78
Within Sample	SSE = 12	N–C =12- 4 = 8	MSE = $\frac{SSE}{N-C}$ =1.5	
Total	SST = 47	N –1=12-1 =11		

Degree of freedom = (3, 8)

Table value of F at 5% level of significance and (3, 8) degrees of freedom is 4.07. Since calculated value of F is more than the table value, H_0 is rejected. We accept alternate hypothesis. So, we may conclude that there is significant difference in the productivity of three varieties of seeds.

Illustration 2.2.12

The following table shows the yield of 3 varieties. Perform analysis of variance andtest whether there is significant difference between varieties:

Varieties	**Plots**				
	A	**B**	**C**	**D**	**E**
I	30	27	42		
II	51	47	37	48	42
III	44	35	41	36	

Solution

Here, we are asked to test whether there is significant difference between varieties.But varieties are given in rows, not in columns. In one way ANOVA, the samples must be in columns. Therefore, we have to rearrange the given data so as to bring the

samples in columns as shown below:

Plots	Varieties		
	I	II	III
A	30	51	44
B	27	47	35
C	42	37	41
D		48	36
E		42	

H_0: There is no significant difference in the productivity of varieties.

H_1: There is significant difference in the productivity of varieties.

Test Statistic applicable here is F-test

F = Larger Variance ÷Smaller Variance, $\frac{MSC}{MSE}$ or $\frac{MSE}{MSC}$

SST = 30^2+ 51^2 + 44^2 + 27^2 + 47^2 + 35^2 +42^2 + 37^2 + 41^2 + 48^2 + 36^2 + 42^2 – $\frac{T^2}{N}$

=900+2601+1936+729+2209+1225+1764+1369+1681+2304+1296+1764 - $\frac{480^2}{12}$

= 19778 – 19200 = 578

SSC = $\left[\frac{(\varepsilon x1)2}{n1}\right] + \left[\frac{(\varepsilon x1)2}{n1}\right] + \left[\frac{(\varepsilon x1)2}{n1}\right]$ +..... $\frac{T^2}{N}$

$\left[\frac{(30+27+42)^2}{3}\right] + \left[\frac{(51+47+37+48+42)^2}{5}\right] + \left[\frac{44+35+41+36^2}{4}\right] - \frac{480^2}{12}$

$= \frac{9801}{3} + \frac{50625}{6} + \frac{24336}{4} - 19200$

= 3267 + 10125 + 6084 – 19200

= 19476 –19200

= 276

SSE = SST- SSC

= 578-276

= 302

Level of significance = 5% Degree of freedom = (2, 9)

One-way ANOVA Table				
Source of variation	**Sum of Squares**	**Degree of freedom**	**Mean Sum of Squares**	**F-Ratio**
Between samples	SSC = 276	3-1= 2	MSC = $\frac{276}{2}$ =138	F= $\frac{138}{33.56}$ = 4.112
Within sample	SSE = 302	12-3 = 9	MSE = $\frac{302}{9}$ = 33.56	
Total	SST = 578	12-1=1		

Table value of F at 5% level of significance and (2, 9) degrees of freedom is 4.26. Since calculated value of F is less than the table value, H_0 is accepted. So we may conclude that there is no significant difference in the productivity of three varieties.

2.2.7.2 Two-way classification of data (Two way analysis of variance)

In two-way classification, observations are classified into different groups on the basis of two criteria. Consider the example mentioned in one- way classification. If we study the effect of both the quality of seeds and the type of fertilizers on the productivity of crop, the data are to be classified on the basis of two criteria, namely type of seed and type of fertilizer. This is called two- way analysis of variance. In case of two-way analysis of variance, we need not make any kind of rearrangement in the given data. Since two criteria are considered, here, there will be two sets of hypotheses.

- Grouped based on two criteria

Types of variances in Two-way ANOVA

a. **Variance between samples (Columns):** This is the net result of the variation of different sample means (in respect of columns) from grand mean. Grand mean is the mean of all the observations coming under all sample groups.

b. **Variance between rows:** This is the net result of the variation of different sample means (in respect of rows) from grand mean.

c. **Variance within the sample (Residual):** This is the net result of variations of different items of the sample from the respective sample means.

d. **Variance about the sample:** This is the sum of the variance between columns, variance between rows and the variance

within the sample(residual)

Proforma of Two-way ANOVA Table

SSC = Sum of Squares between Columns

Two-way ANOVA Table				
Source of variation	Sum of Squares	Degree of Freedom	Mean Sum of Squares	F-Ratio
Between Columns	SSC	c – 1	MSC = $\frac{SSC}{C-1}$	$F_C = \frac{MSC}{MSE}$ or $= \frac{MSE}{MSC}$
Between Rows	SSR	r – 1	MSR= $\frac{SSR}{R-1}$	$F_R = \frac{MSR}{MSE}$ or $= \frac{MSE}{MSR}$
Within Sample	SSE	(c – 1) x (r – 1)	MSE = $\frac{SSE}{(C-1)x\,(r-1)}$	
Total	SST	N –1		

SSR = Sum of Squares between Rows

SSE = Sum of Square within Samples

SST = Sum of Square Total

MSC = Mean Sum of Squares between Columns

MSR = Mean Sum of Squares between Rows

MSE = Mean Sum of Squares within Samples

c= Number of Columns

r = Number of Rows

Procedure for carrying out Two-way Analysis of variance

a. Set up H_0 and H_1

H_0: There is no significant difference between samples (in respect of columns)

H_1: There is significant difference between samples (in respect

of columns)

H_0: There is no significant difference between samples (in respect of rows)

H_1: There is significant difference between samples (in respect of rows)

b. Decide the test statistic

Test statistic applicable here is F-test

c. Apply the appropriate formula for computing the values of F ratios.

$$F_C = \frac{Larger\ Variance}{Smaller\ Variance}$$

. i.e., $\frac{MSC}{MSE} or \frac{MSE}{MSC}$

$$F_R = \frac{LargerVariance}{SmallerVariance}.$$

i.e., $\frac{MSR}{MSE} or \frac{MSE}{MSR}$

i. Find SST.

SST = Sum of square of all items $- \frac{T^2}{N}$

Where

T = Total of all observations,

N = Total Number of observations

$\frac{T^2}{N}$ is generally called correction factor

ii. Find SSC.

$$SSC = \left[\frac{(\sum x1)^2}{n1}\right] + \left[\frac{(\sum x2)^2}{n2}\right] + \left[\frac{(\sum x3)^2}{n3}\right] + \ldots\ldots\ldots - \frac{T^2}{N}$$

Where

$\sum x_1$=sum of items in the first column

$\sum x_2$= sum of items in the second column

n_1 = number of items in the first column

n_2 = number of items in the second column

iii. Find SSR.

$$SSR = \left[\frac{(\sum x1)^2}{n1}\right] + \left[\frac{(\sum x2)^2}{n2}\right] + \left[\frac{(\sum x3)^2}{n3}\right] + \ldots\ldots\ldots - \frac{T^2}{N}$$

Where

$\sum x_1$ = sum of items in the first row

$\sum x_2$ = sum of items in the second row

n_1 = number of items in the first row

n_2 = number of items in the second row

d. Draw ANOVA Table and enter the values of SST, SSC and SSR.

e. Find the value of SSE.

SSE = SST – (SSC + SSR)

f. Find the degree of freedom in the third column as indicated in the proforma.

g. Find MSC.

$$MSC = \frac{SSC}{C\text{-}1}$$

h. Find MSR

$$MSR = \frac{SSR}{r\text{-}1}$$

i. Find MSE.

$$MSE = \frac{SSE}{[(c-1) \times (r-1)]}$$

j. Find F–Ratios (i.e., F_C and F_R)

$$F_C = \frac{\text{Larger variance}}{\text{Smaller variance}}$$

$$[\text{i.e., } F = \frac{MSC}{MSE} \text{ or } \frac{MSE}{MSC}$$

$$F_R = \frac{\text{Larger variance}}{\text{Smaller variance}}; [\text{i.e., } \frac{MSR}{MSE} \text{ or } \frac{MSE}{MSR}]$$

k. Specify the level of significance.

Take 5% if nothing is mentioned.

l. Fix the degrees of freedom.

Fix a pair of d.f. in respect of F_C and F_R.

m. Obtain table value of F_C and F_R at specified level significance and fixed degree of freedom.

n. Compare the calculated value of Fc with the Table value

Compare the calculated value of Fc with the Table value,and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, H_0 is accepted. If calculat-

ed value is more than the table value, H_0 is rejected.

o. Compare the calculated value of F_R with the Table value

Compare the calculated value of F_R with the Table value, and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, H_0 is accepted. If calculated value is more than the table value, H_0 is rejected.

Illustration 2.2.13

Following table shows the yield of crops using 3 varieties of seeds:

Plots	Varieties of Seeds		
	P	Q	R
I	6	7	8
II	4	6	5
III	8	6	10
IV	6	9	9

Test whether there is significant difference in the productivity of varieties of seeds. Also test the significance of the difference between plots.

Solution

H_0: There is no significant difference in the productivity of varieties of seeds.

H_1: There is significant difference in the productivity of varieties of seeds.

H_0: There is no significant difference in the productivity from plots.

H_1: There is significant difference in the productivity from plots.

Test statistic applicable here is F-test

$$F_C = \frac{\text{Larger variance}}{\text{Smaller variance}}$$

$$[\text{i.e., } F = ., \frac{MSC}{MSE} \text{ or } \frac{MSE}{MSC}$$

$$F_R = \frac{\text{Larger variance}}{\text{Smaller variance}}; [\text{i.e., } \frac{MSR}{MSE} \text{ or } \frac{MSE}{T^2 MSR}]$$

$$SST = \text{Sum of square of all items} - \frac{T^2}{N}$$

$= 6^2 + 4^2 + 8^2 + 6^2 + 7^2 + 6^2 + 6^2 + 9^2 + 8^2 + 5^2 + 10^2 + 9^2 - \frac{84^2}{12}$

$=36+16+64+36+49+36+36+81+64+25+100+81 - \frac{7056}{12}$

$= 624 - 588$

$= 36$

$SSC = \left[\frac{(\sum x_1)^2}{n_1}\right] + \left[\frac{(\sum x_2)^2}{n_2}\right] + \left[\frac{(\sum x_3)^2}{n_3}\right] + \ldots\ldots\ldots \frac{T^2}{N}$

$= \left[\frac{(6+4+8+6)^2}{4}\right] + \left[\frac{(7+6+6+9)^2}{4}\right] + \left[\frac{(8+5+10+9)^2}{4}\right] + - \frac{84^2}{12}$

$= \frac{576}{4} + \frac{784}{4} + \frac{1024}{4} - 588$

$= 144 + 196 + 256 - 588 = 8$

$SSR = \left[\frac{(\sum x_1)^2}{n_1}\right] + \left[\frac{(\sum x_2)^2}{n_2}\right] + \left[\frac{(\sum x_3)^2}{n_3}\right] + \ldots\ldots\ldots \frac{T^2}{N}$

$= \left[\frac{(6+7+8)^2}{3}\right] + \left[\frac{(4+6+5)^2}{3}\right] + \left[\frac{(8+6+10)^2}{3}\right] + \left[\frac{(6+9+9)^2}{3}\right] .. - \frac{84^2}{12}$

$= \frac{441}{3} + \frac{225}{3} + \frac{576}{3} + \frac{576}{3} - 588$

$=147+75+192+192-588$

$= 18$

$SSE = SST - (SSC+SSR)$

$= 36 - (8+18) = 10$

Between Columns:

Two-way ANOVA Table				
Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Sum of Squares**	**F-Ratio**
Between Columns	SSC=338.8	4 – 1 = 3	MSC = $\frac{338.8}{3}$ = 112.93	$F_C = \frac{MSC}{MSE}$ $\frac{112.93}{6.142} = 18.387$
Between Rows	SSR=161.5	5 – 1 = 4	MSR = $\frac{161.5}{4}$ = 40.375	

Within Sample	SSE=73.7	(4 – 1) x (5 –1) = 12	MSE = $\frac{73.7}{12}$ =6.142	$F_R = \frac{MSR}{MSE}$ $\frac{40.375}{6.142}$ =6.574
Total	**SST= 574**	**20 –1 = 19**		

Calculated value of F_C = 2.395 Level of Significance = 5% Degrees of freedom = (2, 6)

Table value of F_C at 5% level of significance and (2, 6) degrees of freedom = 5.14

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference in the productivity of three varieties of seeds.

Between Rows:

Calculated value of F_R = 3.593 Level of Significance = 5% Degrees of freedom = (3, 6) Table value of F_R at 5% level of significance and (3, 6) degrees of freedom = 4.76

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference in the productivity from plots.

Coding Method

- To simplify the computation

In analysis of variance, while preparing ANOVA table (both one-way and two-way), at first, we have to find the values of SST, SSC, SSR, etc. But, if the individual observations of the given data are of large values, the computation of SST, SSC, SSR, etc. becomes a tedious task. So, as to avoid this complication, we may apply coding method. Coding method refers to the addition, subtraction, multiplication and division of individual observations of the given data by a constant. The addition, subtraction, multiplication or division of all the individual items by a constant will not affect the value of F.

Illustration 2.2.14

The following table shows the number of units of a product produced by 5 workers using 4 different types of machines:

You are required to test:

i. Whether there is significant difference in the mean productivityof machines.

ii. Whether there is significant difference in the mean pro-

Workers	Machines			
	P	Q	R	S
I	44	38	47	36
II	46	40	52	43
III	34	36	44	32
IV	43	38	46	33
V	38	42	49	39

ductivity of workers.

Solution

Let us apply coding method by subtracting 45 from each observation of the given data.Then we get;

Workers	Machines			
	P	Q	R	S
I	-1	-7	2	-9
II	1	-5	7	-2
III	-11	-9	-1	-13
IV	-2	-7	1	-12
V	-7	-3	4	-6

H_0: There is no significant difference in the productivity of machines.

H_1: There is significant difference in the productivity of machines.

H_0: There is no significant difference in the productivity of workers.

H_1: There is significant difference in the productivity of workers.

Test statistic applicable here is F-test

$$F_C = \frac{\text{Larger variance}}{\text{Smaller variance}}$$

[i.e., F= ., $\frac{MSC}{MSE}$ or $\frac{MSE}{MSC}$

$F_R = \frac{\text{Larger variance}}{\text{Smaller variance}}$; [i.e., $\frac{MSR}{MSE}$ or $\frac{MSE}{MSR}$]

T = -1+-7+2+-9+1+-5+7+-2+-11+-9+-1+-13+-2+-7+1+
-12+-7+-3+4+-6 =80

SST = Sum of square of all items – $\frac{T^2}{N}$

$= -1^2 + 1^2 + -11^2 + -2^2 + -7^2 + -7^2 + -5^2 + -9^2 + -7^2 + -3^2$

$+ 2^2 + 7^2 + -1^2 + 1^2 + 4^2 + 9^2 + -2^2 + -13^2 + -12^2 + -6^2 - \frac{80^2}{20}$

=1+1+121+4+49+49+25+81+49+9+4+49+1+1+16

+81+ 4 +169+144+36 – $\frac{6400}{20}$

= 894 – 320 = 574

SSC = $\left[\frac{(\sum x_1)^2}{n_1}\right] + \left[\frac{(\sum x_2)^2}{n_2}\right] + \left[\frac{(\sum x_3)^2}{n_3}\right] + \ldots\ldots\ldots \frac{T^2}{N}$

$= \left[\frac{(-1+1-11-2-7)^2}{5}\right] + \left[\frac{(-7-5-9-7-3)^2}{5}\right] + \left[\frac{(2+7-1+1+4)^2}{5}\right]$

$+ \left[\frac{(-9-2-13-12-6)^2}{5}\right] - \frac{80^2}{20}$

$= \frac{400}{5} + \frac{961}{5} + \frac{169}{5} + \frac{1764}{5} - 320$

= 80+192.2+33.8+352.8 = 658.8

= 658.8 – 320 = 338.8

SSR = $\left[\frac{(\sum x_1)^2}{n_1}\right] + \left[\frac{(\sum x_2)^2}{n_2}\right] + \left[\frac{(\sum x_3)^2}{n_3}\right] + \ldots\ldots\ldots \frac{T^2}{N}$

$= \left[\frac{(-1-7+2-9)^2}{4}\right] + \left[\frac{(1-5+7-2)^2}{4}\right] + \left[\frac{(-11-9-1-13)^2}{4}\right] +$

$\left[\frac{(-2-7+1-12)^2}{4}\right] + \left[\frac{(-7-3+4-6)^2}{4}\right] - \frac{80^2}{20}$

$= \frac{225}{4} + \frac{1}{4} + \frac{1156}{4} + \frac{400}{4} + \frac{144}{4} - 320$

= 481.5 – 320

= 161.5

SSE = SST – (SSC+SSR)

= 574- (338.8+161.5)

= 73.7

Two-way ANOVA Table				
Source of variation	**Sum of Squares**	**Degree of Freedom**	**Mean Sum of Squares**	**F- Ratio**
Between Columns	SSC = 338.8	4 – 1= 3	MSC = $\frac{338.8}{3}$ =112.93	$F_C = \frac{MSC}{MSE}$ = $\frac{112.93}{6.142}$ =18.387
Between Rows	SSR = 161.5	5 – 1 = 4	MSR = $\frac{161.5}{4}$ = 40.375	$F_R = \frac{MSR}{MSE}$ = $\frac{40.375}{6.142}$ =6.574
Within Sample	SSE = 73.7	(4 –1) * (5-1) = 12	MSE = $\frac{73.7}{12}$ = 6.142	
Total	SST = 574	20 –1 =19		

Between Columns:

Calculated value of F_C = 18.387

Level of Significance = 5%

Degrees of freedom = (3, 12)

Table value of F_C at 5% level of significance and (3, 12) degrees of freedom = 3.49

Since calculated value is more than the table value, null hypothesis is rejected. Alternate hypothesis is accepted. So, we may conclude that there is significant difference in the mean productivity of machines.

Between Rows:

Calculated value of F_R = 6.574

Level of Significance = 5%

Degrees of freedom = (4, 12)

Table value of F_R at 5% level of significance and (4, 12) degrees of freedom = 3.26

Since calculated value is more than the table value, null hypothesis is rejected. Alternative hypothesis is accepted. So, we may conclude that there is significant difference in the mean productivity of workers.

Summarised Overview

Parametric tests, such as the z-test, t-test, and F-test, assume specific distributions, usually normal, to make inferences about population parameters. The z-test is employed for large sample sizes with known population standard deviation, while the t-test is used for smaller samples with unknown population standard deviation. F-test determines differences in population variances. These tests are powerful when assumptions hold but may produce inaccurate results otherwise. Testing of given population mean is used to test whether the given population mean is true or not. In other words, it is used to test whether there is significant difference between sample mean and population mean. Testing of significance of the difference between two sample mean is used to test whether the difference between two sample means are significant or not. Testing of given population proportion is used to test whether there is any significant difference between the sample proportion and the given population proportion. Testing of the significance of the difference between two sample proportions is used to test whether the difference between two sample proportions are significant or not.

Analysis of Variance (ANOVA) assesses differences among three or more sample groups, with one-way ANOVA comparing groups based on a single criterion and two-way ANOVA based on two criteria. Parametric tests offer structured approaches for hypothesis testing, enabling researchers to draw meaningful conclusions from data. However, understanding of assumptions and limitations is crucial for valid interpretation of results. Thus, parametric tests play a vital role in statistical analysis, facilitating hypothesis testing and comparison of population parameters across various fields.

Self-Assessment Question

1. What do you mean by parametric tests?
2. What is type I error?
3. What is Type II error?
4. What do you mean by power of a test?
5. What is meant by critical region and acceptance region?
6. What is one tailed test?
7. What is two-tailed test?
8. What do you mean by analysis of variance?
9. Explain the two types of analysis of variance.
10. What are the different types of variances in case of one-way analysis of variance?
11. What are the different types of variances in case of two-way analysis of variance?
12. Draw the proforma of one-way ANOVA table.

13. Draw the proforma of two-way ANOVA table.
14. Explain the hypothesis testing procedure in case of one-way ANOVA.
15. Explain the hypothesis testing procedure in case of two-way ANOVA.
16. What do you mean by coding method in analysis of variance?

Assignments

1. The average annual income of people in a particular locality is assumed to be ₹15,00,000. The Income Tax Authorities feel that there is gross understatement of income. In order to check this, the tax authorities selected a sample of 100 citizens and found their average annual income to be ₹18,00,000 with a standard deviation of ₹1,50,000. Verify the claim of tax authorities at 5% level of significance. (Answer- Reject null hypothesis)
2. It is claimed that an automobile is driven on the average less than 20000 kilometres per year. To test this claim, a random sample of 100 automobile owners is asked to keep the record of the kilometres they travel. Would you agree with his claim if the random sample showed an average of 23500 kilometres and a standard deviation of 3900 kilometres. Use 0.01 level of significance. (Answer- Reject null hypothesis)
3. In a certain college, it is estimated that fewer than 25% of students have cars on campus. Does this seem to be a valid estimate, if, in a random sample of 90 college students, 28 are found to have cars. Use 0.05 level of significance. (Answer- Accept null hypothesis)
4. The average length of time of students who registered for a certain course in certain college has been 50 minutes. A new registration procedure using modern computing machine is being tried. If a random sample of 12 students had an average registration time of 42 minutes with a S.D. of 11.9 minutes, under the new system, test the hypothesis that the new system is effective using a level of significance of 5 % (Ans: Accept the null hypothesis, t-test)

Reference

1. Barry Render. (2022). *Quantitative Analysis for Management*. Prentice Hall of India
2. Chawla, D; and Sondhi, N., (2016). *Research Methodology: Concepts and Cases*. Vikas Publishing House Pvt. Ltd.
3. Anand Sharma. (2017). *Quantitative Techniques for decision making*. Himalaya Publishing House.

4. Krishnaswami, O.R., Ranganathan, M., & Harikumar, P.N. (2016). *Research Methodology*. Himalaya Publishing House.
5. Panneerselvam, R. (2014) *Research Methodology*. PHI Learning
6. Sharma R, N & Sharma R.K. (2019). *Research Methods in Social Science*. Atlantic Publishers.

Suggested Reading

1. Tandon, B.C; (1979). *Research Methodology in Social Sciences*. Chaitanya Publishing House.
2. Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education (8th ed.)*. Routledge.
3. Kothari, C. R., (2004). *Research Methodology: Methods and Techniques*. New Age International (P) Limited.
4. Gupta,S.K. & Praneet Rangi. (2018). *Business Research Methodology*. Kalyani Publishers
5. Singh A.K. (2017). *Tests, Measurements and Research Methods in Behavioural Sciences*. Bharti Bhawan
6. Kevin, S., (2021). *Research Methodology for Social Sciences*. Ane Books Pvt. Ltd. New Delhi.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 3

Non-Parametric Test

Learning Outcomes

After completing this unit, the learners will be able to:

- understand the concept of non-parametric tests.
- gain insights on the various non-parametric tests.
- interpret the results of each test accurately and draw meaningful conclusions.

Background

In statistical analysis, the reliance on non-parametric tests over parametric counterparts is often necessitated by various factors. These tests, not bound by assumptions of normality or specific population distributions, offer robustness in scenarios where data deviates from the expected patterns. In situations involving small sample sizes, skewed distributions, or the presence of outliers, non-parametric tests shine by providing reliable insights without the stringent requirements of parametric methods. Their flexibility extends to handling diverse data types, including ordinal and categorical variables, rendering them indispensable tools across disciplines like social sciences and ecology. Non-parametric tests offer simplicity and ease of interpretation, requiring fewer assumptions and calculations compared to their parametric counterparts. This makes them suitable for analyzing data in situations where assumptions about the population distribution are uncertain or difficult to verify. This unit focuses into the various non-parametric tests.

Keywords

Chi-square test, Sign test, signed rank test, rank sum test, Run test

Discussion

2.3.1 Non-Parametric Test

- No assumptions about population

Non-parametric test are those tests which can be used for ordinal and nominal data. It does not make assumptions about population such as normality in distribution and randomness like parametric tests. They are used when the data doesn't meet the assumptions of parametric tests, such as when the data is not normally distributed or when it includes outliers. The most common non-parametric tests are Mann-Whitney U test, Kruskal Wallis H test, Friedman test, Wilcoxon signed-rank test etc.

2.3.2 Chi-Square Test

- Test difference between observed and expected frequencies

Chi-square test is a statistical test used to test the significance of the difference between observed frequencies and the corresponding theoretical frequencies (expected frequencies) of a distribution, without any assumption about the nature of distribution of the population. This is the most widely used non-parametric test. It was developed by Prof. Karl Pearson. It is depicted using the symbol χ^2.

Uses of Chi-Square Test (Applications of Chi- Square Test)

Chi-Square test is mainly used for the following purposes:

a. **Used to test goodness of fit:** As a test for goodness of fit, χ^2 test can be used to test how far the theoretical frequencies fit to the observed frequencies.

b. **Used to test independence:** As a test of independence, χ^2 test is used to test whether the attributes of a sample are associated or not.

c. **Used to test homogeneity:** As a test of homogeneity, χ^2 test is used to test whether different samples are homogeneous as far as a particular attribute is concerned.

d. **Used to test population variance:** It used to test whether there is any significant difference between sample variance and population variance. Here, the test statistic value (Chi-square value) is obtained by using the following formulae $(\frac{ns^2}{\sigma^2})$.

Conditions for applying Chi-square Test

i. The total frequencies (N) must be at least 50.

ii. Expected frequencies of less than 5 must be pooled with the preceding or succeeding frequency so that the expected frequency is 5 or more.

iii. The distributions should be of original units.

iv. They should not be of proportions or percentages.

v. All the members or items in the sample must be independent

2.3.2.1 Testing of Goodness of Fit

Procedure:

i. Set up H_0 and H_1

H_0: There is goodness of fit between observed frequencies and expected frequencies.

H_1: There is no goodness of fit between observed frequencies and expected frequencies.

ii. Decide the test statistic

Here, the test statistic is Chi-Square test.

iii. Apply the appropriate formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where,

O_i = Observed frequencies

E_i = Expected frequencies

iv. Specify the level of significance.

If nothing is mentioned, take 5% level of significance.

v. Fix the degree of freedom.

Degree of freedom = n – r –1

Where

n = number of pairs of observations.

r = number of parameters computed from the given data to find the expected frequencies.

vi. Obtain the table value

Obtain the table value of Chi-square at specified level of significance and fixed degree of freedom.

vii. Compare the actual value of Chi-Square with the table value

Compare the actual value of Chi-Square with the table val-

ue and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Illustration 2.3.1

The numbers of road accidents per week in a certain city were as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that the accidents occurred were the same during the 10-week period?

Solution

H_0: There is goodness of fit between observed frequencies and expected frequencies.

H_1: There is no goodness of fit between observed frequencies and expected frequencies.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here the Observed values (Actual values) are 12, 8, 20, 2, 14, 10, 15, 6, 9 and 4.

i.e., O_i = 12, 8, 20, 2, 14, 10, 15, 6, 9, 4

If accidents occurred are same, then the number of accidents per week which we may expect is 10 (i.e., the average of the given values).

$$E_i = \frac{100}{10} 10$$

Now we can find the value of Chi-square as follows:

Computation of Chi-square Value

Observed Values(O_i)	**Expected Values(E_i)**	$\frac{(Oi - Ei)^2}{Ei}$
12	10	0.4
8	10	0.4
20	10	10.0
2	10	6.4
14	10	1.6
10	10	0.0

15	10	2.5
6	10	1.6
9	10	0.1
4	10	3.6
		$\chi^2 = 26.6$

Calculated Value of $\chi^2 = 26.6$

Level of significance = 5%

Degree of Freedom = n - 1 = 10 – 0 -1 = 9

Table value of χ^2 at 5% level of significance and d.f of 9 is 16.919.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternate hypothesis. So we may conclude that the given figures do not agree with the belief that accident occurred were same during the 10 weeks period.

Illustration 2.3.2

The principal of a college made a sample analysis of an examination result of 200 students. It was found that 24 students had got first class, 62 second class, 68 third class and the rest were failed. Are these figures commensurate with the general examination result which is in the ratio of 2:3:3:2 for various categories respectively?

Solution

H_0: There is goodness of fit between the given figures and the figures expected in general examination

H_1: There is no goodness of fit between the given figures and the figures expected in general examination

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here the Observed values (Actual values) for first, second, third and failed categories of students are 24, 62, 68 and 46 respectively.

If results are in the ratio of 2:3:3:2, then the number of students for above categories may be expected as follows:

First Class	200 x $\frac{2}{10}$	40

Second Class	200 x $\frac{3}{10}$	60
Third Class	200 x $\frac{3}{10}$	60
Failed	200 x $\frac{2}{10}$	40
Total		200

So the E_i Values are 40, 60, 60 and 40.

Now we can find the value of Chi-square as follows:

Computation of Chi-square Value			
Observed Values (O_i)	**Expected Values(E_i)**	$(Oi - Ei)^2$	$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
24	40	256	6.400
62	60	4	0.067
68	60	64	1.067
46	40	36	0.900
			χ^2 = 8.434

Calculated Value of χ^2 = 8.434

Level of significance = 5%

Degree of Freedom = n -1 = 4 -1 = 3

Table value of χ^2 at 5% level of significance and d.f. of 3 is 7.815.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternate hypothesis. So, we may conclude that the given figures do not commensurate with the general examination result which is in the ratio of 2:3:3:2.

2.3.2.2 Testing of Independence

Procedure:

i. Set up H_0 and H_1

H_0: There is independence between observed frequencies and expected frequencies.

H_1: There is no independence between observed and expected frequencies.

ii. Decide the test statistic.

Here, the test statistic is Chi-Square test.

iii. Apply the appropriate formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where,

O_i = Observed frequencies

E_i = Expected frequencies

Here E_i values are obtained by using the following formula:

$$E_i \text{ Value} = \frac{\text{(Row Total x Column Total)}}{\text{Grand Total}}$$

E_i Values are computed by preparing a table called Contingency Table.

iv. Specify the level of significance.

If nothing is mentioned, take 5% level of significance.

v. Fix the degree of freedom.

Degree of freedom = (r– 1) x (c – 1)

Where r = number of rows; c = number of columns.

vi. Obtain the table value of Chi-square at specified level of significance and fixed degree of freedom.

vii. Compare the actual value of Chi-Square with the table value

Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Illustration 2.3.3

From the following data, can you say that there is relation between the habit of smoking and literacy?

	Smokers	**Non- Smokers**
Literates	83	57
Illiterates	45	68

Solution

H_0: There is independence between smoking habit and literacy.

H_1: There is no independence between smoking habit and literacy.

Here the Observed values (Actual values) are 83, 57, 45 and 68.

2 x 2 Contingency Table			
	Smokers	**Non- smokers**	**Total**
Literates	$\frac{[(83+57) \times (83+45)]}{253} = 71$	$\frac{(140 \times 125)}{253} = 69$	140
Illiterates	$\frac{(113 \times 128)}{253} = 57$	$\frac{(113 \times 125)}{253} = 56$	113
Total	128	125	253

The E_i Values corresponding to the above 'O_i' values can be

Observed Values(O_i)	**Expected Values(E_i)**	$\frac{(O_i - E_i)^2}{E_i}$
83	71	2.03
57	69	2.09
45	57	2.53
68	56	2.57
		$\chi^2 = 9.22$

found out by preparing a 2 x 2 contingency table:

So, the E values are 71, 69, 57 and 56.

Calculated Value of $\chi 2 = 9.22$

Level of significance = 5%

Degree of Freedom = (2 – 1) x (2 – 1) = 1 x 1 = 1

Table value of $\chi 2$ at 5% level of significance and 1 d.f is 3.841.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternate hypothesis. So we may conclude that there is no independence between smoking habit

and literacy. In other words, smoking habit and literacy are related.

Illustration 2.3.4

In a sample study about the tea drinking habit in a town, following data are observed in a sample of size 200. 46% were male, 26% were tea drinkers and 17% were male tea drinkers. Is there any association between gender and tea habits?

Solution:

H_0: There is independence between gender and tea drinking habits.

H_1: There is no independence between gender and tea drinking habits.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here all the Observed values (Actual values) are not directly given in the question. So, we have to find the missing figures with the help of a 2 x 2 contingency table:

2 x 2 Contingency Table ('O_i' values)			
	Tea drinkers	Non- teadrinkers	Total
Male	$\frac{(200 \times 17)}{100} = 34$	58	$\frac{(200 \times 46)}{100} = 92$
Female	18	90	108
Total	$\frac{(200 \times 26)}{100} = 52$	148	200

"O_i" values are 34, 58, 18 and 90

The E_i Values corresponding to the above 'O_i' values can be found out by preparing a 2 x 2 contingency table:

2 x 2 Contingency Table ('E_i' values)			
	Tea drinkers	Non-tea drinkers	Total
Male	$\frac{(92 \times 52)}{200} = 24$	$\frac{(92 \times 148)}{200} = 68$	92
Female	$\frac{(108 \times 52)}{200} = 28$	$\frac{(108 \times 148)}{200} = 80$	108
Total	52	148	200

So the 'E_i' values are 24, 68, 28 and 80.

Computation of Chi-square Value			
ObservedValues (O_i)	**ExpectedValues (E_i)**	**$(O_i-E_i)^2$**	**$\frac{(O_i - E_i)^2}{E_i}$**
34	24	100	4.17
58	68	100	1.47
18	28	100	3.57
90	80	100	1.25
			$\chi^2 = 10.46$

Calculated Value of $\chi^2 = 10.46$

Level of significance = 5%

Degree of Freedom = (2 – 1) x (2 – 1) = 1 x 1 = 1

Table value of $\chi 2$ at 5% level of significance and 1 d.f. is 3.841.

Since calculated value is more than the table value, null hypothesis is rejected. We accept alternatehypothesis. So, we may conclude that there is no independence between gender and smoking habit. In other words, gender and smoking habit are closely associated.

2.3.2.3 Test of Homogeneity

Procedure:

i. Set up H_0 and H_1

H_0: There is homogeneity between the samples on the basis of the attribute.

H_1: There is no homogeneity between the samples on the basis of the attribute.

ii. Decide the test statistic.

Here, the test statistic is Chi-Square test.

iii. Apply the appropriate formula

$$\chi 2 = \frac{(O_i - E_i)^2}{E_i}$$

where O_i = Observed frequencies

E_i = Expected frequencies

Here 'E_i' values are obtained by using the following formula:

'E_i' Value = $\frac{\text{(Row Total x Column Total)}}{\text{Grand Total}}$

'E_i' Values are computed by preparing a table called Contingency Table.

iv. Specify the level of significance.

If nothing is mentioned, take 5% level of significance.

v. Fix the degree of freedom.

Degree of freedom= (r– 1) x (c – 1)

Where r = number of rows; c = number of columns

vi. Obtain the table value of Chi-square

Obtain the table value of Chi-square at specified level of significance and fixed degree of freedom.

vii. Compare the actual value of Chi-Square with the table value

Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Illustration 2.3.5

In a diet survey the following results were obtained:

	Male	**Female**
No. of families drinking tea	124	16
No. of families not drinking tea	56	10

Is there any difference between the gender in the matter of tea drinking?

Solution

H_0: There is homogeneity between gender in the matter of tea drinking.

H_1: There is no homogeneity between gender in the matter of tea drinking.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here the Observed values (Actual values) are 124, 16, 56, and 10

The 'E_i' values corresponding to the above 'O_i' values can be found out by preparing a 2 x 2 contingency table:

2 x 2 Contingency Table			
	Male	**Female**	**Total**
No. of families drinking tea	$\frac{(140 \times 180)}{206} = 122$	$\frac{(140 \times 26)}{206} = 18$	140
No. of families not drinking	$\frac{(66 \times 180)}{206} = 58$	$\frac{(66 \times 26)}{206} = 8$	66
Total	**180**	**26**	**206**

So, the 'E' values are 122, 18, 58 and 8.

Observed Values (O_i)	**Expected Values (E_i)**	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
124	122	4	0.033
16	18	4	0.222
56	58	4	0.069
10	8	4	0.500
			$\chi^2 = 0.824$

Calculated Value of $\chi 2 = 0.824$

Level of significance =5%

Degree of Freedom = (2 – 1) x (2 – 1) = 1 x 1 = 1

Table value of $\chi 2$ at 5% level of significance and 1 degree of freedom is 3.841.

Since calculated value is less than the table value, null hypothesis is accepted. So, we may conclude that there is homogeneity between gender in the matter of tea drinking.

2.3.2.4 Test of Variance

Procedure:

i. Set up H_0 and H_1

$H_{0:}$ There is no significant difference between sample variance and population variance.

$H_{1:}$ There is significant difference between sample variance and population variance.

ii. Decide the test statistic.

Here, the test applicable is Chi-square test.

iii. Apply the appropriate formula for computing the value of test statistic.

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

where

n = sample size,

s^2 = sample variance,

σ^2 = population variance.

iv. Specify the level of significance.

Take 5%, unless specified otherwise.

v. Fix the degree of freedom.

d.f = n –1.

vi. Locate the table value of Chi-square at specified level of significance and fixed degree of freedom.

vii. Compare the actual value of Chi-Square with the table value

Compare the actual value of Chi-Square with the table value and decide whether to accept or reject the null hypothesis. If calculated value is less than the table value, null hypothesis is accepted and otherwise it is rejected.

Illustration 2.3.6

A sample is drawn from a population which follows normal distribution. The size of sample and S.D. are respectively 10 and 5. Test whether this is consistent with the hypothesis that the S. D. of the population is 5.3.

Solution

H_0**:** There is no significant difference between sample S.D. and population S.D. (i.e., H_0:S.D of population = 5.3)

H_1**:** There is significant difference between sample S.D. and population S.D. (i.e., H_1 : S.D. of population ≠ 5.3)

The test applicable is Chi-square test.

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

Where

n = 10, $s^2 = 5^2$, $\sigma^2 = 5.3^2$

$$= \frac{10 \times 5^2}{5.3^2}$$

$$= \frac{250}{28.09} = 8.899$$

Specify the level of significance. Take 5%, unless specified otherwise.

Fix the degree of freedom. d.f = 10 –1 = 9

Table value of Chi-square at 5% level of significance and 9 degree of freedom is 16.919.

Since calculated value is less than the table value, null hypothesis is accepted. So we may conclude that there is no significant difference between sample S.D. and population S.D.

Illustration 2.3.7

A sample group of 10 students are selected randomly from a class. Their weights (in K.g) are 49, 40, 53, 38, 52, 47, 48, 45, 55, and 43. Can we say that the population variance is 20 Kg?

Solution

$H_{0:}$ There is no significant difference between sample variance and population variance.(i.e., H_0: Variance of population = 20)

$H_{1:}$ There is significant difference between sample variance and population variance.(i.e., H_1: Variance of population ≠ 20)

The test applicable is Chi-square test.

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

Here, n = 10, σ^2 = 20, Sample variance is to be computed from the given data.

Computation of sample variance (s^2)		
Weight (X)	**$(x - \bar{x})$**	**$(X - \bar{x})^2$**
49	2	4
40	-7	49
53	6	36
38	-9	81
52	5	25
47	0	0
48	1	1
45	-2	4
55	8	64
43	-4	16
X = 470		**$(X - \bar{x})^2 = 280$**

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

$$= \frac{470}{10} = 47$$

$$\text{Sample variance } (S^2) = \frac{\sum (x - \bar{x})^2}{n}$$

$$= \frac{280}{10} = 28$$

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$\frac{10 \times 28}{20} = \frac{280}{20} = 14$$

Level of significance =5%

Degree of freedom (d.f) = 10 –1 = 9

Table value of Chi-square at 5% level of significance and 9 degree of freedom is 16.919. Since, calculated value is less than the table value, null hypothesis is accepted. So, we may conclude that there is no significant difference between sample variance and population variance.

∴The population variance = 20 Kg.

2.3.3 Sign Tests

- Based on signs of deviation

t-test is generally used when sample is small and there is an assumption that the population is normal. Therefore, when sample is small but it is not possible to make an assumption about the nature of population distribution, t-test cannot be applied. In such a case sign test is used. In sign test, to find the value of test statistic, we use the proportion of signs (+ve or -ve signs), not the numerical magnitude. That is why, the test is known as sign test.

There are two types of sign tests. They are (a) One sample sign test and (b) Two sample sign test.

a. One sample sign test

It is also called single sample sign test and is used to test the hypothesis concerning the median for one population. One sample sign test is used to test whether the sample belongs to a particular population.

Procedure:

i. Set up null and alternate hypotheses:

H_0: There is no significant difference between sample median and

population median (i.e., $\mu = \mu_{0)}$

H_1: There is significant difference between sample median and population median (i.e., $\mu \neq \mu_0$)

ii. Decide the test statistic.

The test statistic applicable is one sample sign test.

iii. Use the appropriate formula for computing the value of test statistic

$$\text{Test statistic} = \frac{(p-P)}{SE}$$

where P = ½

p = proportion of + signs, (+ or – signs for each observation is determined by subtracting median value from each of them)

S E = Standard deviation

iv. Specify the level of significance.

Take 5%, if not mentioned.

v. Degree of freedom

d.f =infinity

vi. Locate the table value of t-test.

vii. Compare the calculated value with table value

Compare the calculated value with table value and decide whether to accept or reject the hypothesis.If calculated value is less than the table value, null hull hypothesis is accepted and otherwise, it is rejected.

b. Two sample sign test (Paired sample sign test)

- Test whether two population are identical

Two sample sign test is used to test whether two population have the same median. It tests whether two populations are identical. In case of two sample signtest, each pair is replaced by + ve or – ve sign. If first value in a pair is larger, assign + ve sign to thatpair, and otherwise assign – ve sign. The procedure is same as in the case of one sign test.

2.3.4 Signed Rank Test (Wilcoxon Matched Pairs Test)

- Test whether matched paired samples are identical or not

Signed Rank Test is another important non- parametric test used to test whether matched paired samples are identical or not. Here we use the signed ranks for testing. Wilcoxon Matched Pairs Test is used differently depending upon following two situations:

i. When the number of matched pairs ≤ 25, and

ii. When the number of matched pairs >25.

a. Signed Rank Test (When the number of matched pairs ≤ 25)

Here, we find the difference of matched pairs and assign them ranks. Then ranks are classified into two categories based on their respective signs. Then take the sum of two categories of ranks. The minimum of the two is considered as the value of test statistic.

Procedure:

i. Set up null and alternate hypotheses:

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

ii. Decide the test statistic.

The test statistic applicable here is Wilcoxon matched pairs test (i.e., Wilcoxon's T test)

iii. Use the appropriate formula for computing the value of test statistic (Wilcoxon's T test)

T = Sum of Positive Ranks or Sum of Negative Ranks, whichever is less.

iv. Specify the level of significance

Take 5%, if not mentioned.

v. Degree of freedom

d.f = n-1

vi. Locate the table value of Wilcoxon's T test.

vii. Compare the calculated value with table value

Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is accepted andotherwise, it is rejected.

b. Signed Rank Test (When the number of matched pairs > 25)

Here, we find the difference of matched pairs and assign them ranks. Then ranks are classified into two categories based on their respective signs. Then take the total of two categories of ranks. The test statistic is Z test.

Procedure:

i. Set up null and alternate hypotheses:

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

ii. Decide the test statistic.

The test statistic applicable here is Wilcoxon matched pairs test

(i.e., Z test)

iii. Use the appropriate formula for computing the value of test statistic

$$(Z\ test)\ Z = \frac{(T-\mu)}{\sigma}$$

Where T= Sum of Positive Ranks or Sum of Negative Ranks, whichever is less

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

$$\sigma = \sqrt{\frac{[n(n+1)(2n+1)]}{24}}$$

iv. Specify the level of significance.

Take 5%, if not mentioned.

v. Degree of freedom

d.f = infinity

vi. Locate the table value of Z

Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is accepted and otherwise, it is rejected.

2.3.5 Rank Sum Tests

- Assignment of ranks

Rank sum tests are another type of tests used for testing whether the populations are identical. Here, various samples are taken together and then ranks are assigned. There are two important types of rank sum tests. They are (a) Wilcoxon-Mann- Whitney Test (U-test), and (b) Kruskal – Wallis Test (H- test).

2.3.5.1 Wilcoxon-Mann-Whitney Test (Mann-Whitney U Test)

- alternative to the t-test for two independent samples

This non-parametric test for two independent samples was described by Wilcoxon and studied by Mann and Whitney. It is the most widely used test as an alternative to the t- test for two independent samples when we do not make the t- test assumptions about the parent population. The only assumption is that the random variables are continuous. This method is used when there are two groups of samples.

The testing procedure is:

i. Set up null and alternate hypotheses:

H_0: There is no significant difference between two samples

H_1: There is significant difference between two samples

ii. Decide the test statistic.

The test statistic applicable here is Wilcoxon Mann Whitney test (i.e., U-test)

iii. Use the appropriate formula for computing the value of test statistic

(Z test)Test Statistic = $\frac{[(\mu - U)]}{SE}$

Where

$\mu = \frac{n_1 \times n_2}{2}$

U = U_1 or U_2 whichever is less.

$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$

$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$

R_1 = Rank sum of Sample I

R_2 = Rank sum of Sample II

n_1 = Number of observations in Sample I

n_2 = Number of observations in Sample II

$SE = \sqrt{\frac{n_1 \times n_2 (n_1 + n_2 + 1)}{12}}$

iv. Specify the level of significance.

Take 5% unless specified otherwise.

v. Fix the degrees of freedom.

df =infinity

vi. Locate the table value of test statistic

Locate the table value of test statistic (i.e., Z test) at specified level of significance and fixed degreesof freedom.

vii. Compare the calculated value with table value

Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is accepted and otherwise, it is rejected.

2.3.5.2 Kruskal – Wallis Test (H- test):

- Non-parametric analogue of the ANOVA

The Kruskal- Wallis test is the non-parametric analogue of the ANOVA test for the one way classified data. This test is used to test the null hypothesis if k independent samples have been drawn from populations which have identical distributions and does not require the condition of normality of the populations. This test is an extension of Mann-Whitney test to the situations where more than two populations are considered and is based on the ranks of the sample observations. Here, we tests whether three or more independent sample groups come from the population having the same mean.

The testing procedure is:

i. Set up null and alternate hypotheses

H_0: There is no significant difference between samples

H_1: There is significant difference between samples

ii. Decide the test statistic

The test statistic applicable here is Kruskal – Wallis test (i.e., H-test)

iii. Use the appropriate formula for computing the value of test statistic

$$\text{Test statistic} = \frac{12}{n\,(n+1)}\left[\frac{R1^2}{n_1} + \frac{R2^2}{n_2} + \cdots + \frac{Rk^2}{n_k}\right] - 3(n+1)$$

R_1 = Rank sum of Sample I

R_2 = Rank sum of Sample II

n_1 = Number of observations in Sample I

n_2 = Number of observationsin Sample II

iv. Specify the level of significance.

Take 5% unless specified otherwise.

v. Fix the degrees of freedom

d.f. = c–1

vi. Locate the table value of test statistic

Locate the table value of test statistic (i.e; Chi- square test) at specified level of significance andfixed degrees of freedom.

vii. Compare the calculated value with table value

Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is accepted and otherwise, it is rejected.

2.3.6 One Sample Runs Test (Wald-Wolfowitz' Runs Test)

- Checking randomness of observations

Runs test is used to test the randomness of a sample on the basis of the order in which the observations are taken. A 'run' is a succession of identical items. This test was designed by Wald Wolfowitz. The procedure first classifies each value of the variable as falling above or below a cut-point and then tests if there is no specific order to the resulting sequence. The cut-point is either a measure of central tendency or a custom value.

The testing procedure is:

i. Set up null and alternate hypotheses

H_0: There is randomness

H_1: There is no randomness

ii. Decide the test statistic

The test statistic applicable here is Z-test.

iii. Use the appropriate formula for computing the value of test statistic.

$$Z = \frac{(r - \mu)}{\sigma}$$

Where

r =Number of runs

$$\mu = \frac{2n_1 n_2}{n_1 n_2} + 1$$

$$\sigma = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

n_1 = Number of first item in all the runs together

n_2 = Number of second item in all the runs together

iv. Specify the level of significance

Take 5% unless specified otherwise.

v. Fix the degrees of freedom

df = infinity

vi. Locate the table value of Z at specified level of significance

Infinity degrees of freedom.

vii. Compare the calculated value with table value

Compare the calculated value with table value and decide whether to accept or reject the hypothesis. If calculated value is less than the table value, null hull hypothesis is accepted and otherwise, it is rejected.

Summarised Overview

Non-parametric test are those tests which does not make assumptions about population such as normality in distribution and randomness like parametric tests. It can be used for ordinal and nominal data. The most common non-parametric tests explained in the unit were Chi-square test, Mann-Whitney U test, Kruskal Wallis H test, Wilcoxon signed-rank test, and Wald-Wolfowitz' Runs Test.

Chi-square test evaluates the discrepancy between observed and expected frequencies in a single categorical variable, assessing goodness of fit to a specified distribution or testing independence between two variables using a contingency table. The sign test examines differences between pairs of matched samples, suitable for non-parametric data. Wilcoxon Signed Rank sum test assesses differences between paired samples with ordinal or non-normally distributed data. Mann-Whitney U test compares independent samples to evaluate differences in central tendency, regardless of distribution shape. Kruskal Wallis H test extends this to multiple independent groups. Wald Wolfowitz Runs test determines randomness or patterns in a sequence of binary outcomes, assessing independence. Each test addresses specific research questions and data characteristics, providing valuable insights into distributions, relationships, and randomness without stringent assumptions about data distribution or sample size.

Self-Assessment Question

1. What do you mean by non-parametric tests?
2. What are the situations under which non-parametric tests are applied?
3. What are the important assumptions of non- parametric tests?
4. What are the important merits of non-parametric tests?
5. What are the important drawbacks of non- parametric tests?
6. What are the different types of non-parametric tests?
7. Distinguish between parametric tests and non-parametric tests.
8. What do you mean by one sample sign test?
9. Explain the hypothesis testing procedure under one sample sign test.
10. Explain the hypothesis testing procedure under two sample sign test.
11. What do you mean by Wilcoxon matched pair test?
12. Explain the hypothesis testing procedure of Wilcoxon matched pair test.
13. What is meant by Wilcoxon Mann Whitney U-test?
14. Explain the hypothesis testing procedure of Wilcoxon Mann Whitney U-test.
15. What is meant by Kruskal-Wallis H-test?

16. Explain the hypothesis testing procedure of H-test.
17. What do you mean by Chi-square value?
18. What are the important uses of Chi-square test?
19. What do you mean by goodness of fit?
20. Explain the procedure for testing goodness of fit.
21. What do you mean by contingency table?

Assignments

1. A sample of size 50 has S.D. of 10.5. Can you contradict the hypothesis that the population S.D. is12?

   (Answer- $\chi^2 = 7.65$, Table value 9.488)

2. A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck. In a test, you counted 1600 cards, and observed the following:

Spades	404
Hearts	420
Diamonds	400
Clubs	376

Could it be that the suits are equally likely? Or are these discrepancies too much to be random? (Answer- $\chi^2 = 2.480$)

3. A die is thrown 120 times and the frequencies of various faces are as follows:

Face No:	1	2	3	4	5	6
Frequency:	10	15	25	25	18	27

Test whether the die was fair.

(Answer: $\chi^2 = 11.40$)

4. Explain the procedure to compute Kruskal Wallis test and Mann-Whitney U test. Compare and contrast the two tests.

Reference

1. Tandon, B.C; (1979). *Research Methodology in Social Sciences*. Chaitanya Publishing House.
2. Chawla, D; and Sondhi, N., (2016). *Research Methodology: Concepts and Cases*. Vikas Publishing House Pvt. Ltd.
3. Kevin, S., (2021). *Research Methodology for Social Sciences*. Ane Books Pvt. Ltd., New Delhi.
4. Kothari, C. R., (2004). *Research Methodology: Methods and Techniques*. New Age International (P) Limited.
5. Krishnaswami, O.R., Ranganathan, M., & Harikumar, P.N. (2016). *Research Methodology*. Himalaya Publishing House.
6. Panneerselvam, R. (2014) *Research Methodology*. PHI Learning

Suggested Reading

1. Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education (8th ed.)*. Routledge.
2. Gupta, S.K. & Praneet Rangi. (2018). *Business Research Methodology*. Kalyani Publishers
3. Sharma R, N & Sharma R.K. (2019). *Research Methods in Social Science*. Atlantic Publishers.
4. Singh A.K. (2017). *Tests, Measurements and Research Methods in Behavioural Sciences*. Bharti Bhawan

03 BLOCK

CORRELATION AND REGRESSION ANALYSIS

Block Content

Unit 1

Correlation

Learning Outcomes

After completing this unit, the learner will be able to:

- be familiar with the term 'correlation'
- identify the various uses of correlation
- gain insight into the various methods used to calculate correlation
- apply correlation to find out the association between variables.

Background

Imagine you're studying the performance of students in a school. You have a vast amount of data grades, attendance records, hours spent studying, and even data on extra-curricular activities. Your goal is to find a way to understand if there's any relationship between studying habits and grades.

You start by looking at each variable independently: Does the number of hours studied impact the grades? Does attendance correlate with academic performance? Does participation in extracurricular activities influences grades?

As you investigate into the data, you find that students who spend more time studying tend to have higher grades. However, it's not a direct cause; some students who study less might excel naturally, while others who study more might struggle.

Here's where correlation comes into play. By using correlation analysis, you can quantify the relationship between study hours and grades. It helps you understand the strength and direction of this relationship whether they move together positively (more study hours, higher grades), negatively (more study hours, lower grades), or have no apparent relationship at all.

Through correlation analysis, you can determine if there's a meaningful association between the variables. This statistical tool helps you identify patterns, trends, and dependencies between different factors without implying causation. It's a powerful way to grasp the connections between variables and gain insights into how they might relate to each other in real-life scenarios. Correlation is the key to unveiling these relationships and understanding the dynamics between different elements in your data.

Keywords

Correlation, Karl Pearson's correlation Coefficient, Spearman Rank Correlation, Probable error, Standard error

Discussion

3.1.1 Correlation

- Associated Variables

In our day-to-day lives, we often encounter two or more variables closely linked to each other. When one variable changes, there is a corresponding change in another related variable. These variables are termed as correlated variables. For instance, there is a relationship between the price of a commodity and the quantity demanded, fluctuations in rainfall and production of agricultural crops, and advertising expenditure and sales.

- Degree of relationship between variables

The statistical tool used to determine the relationship between two or more variables is called correlation. The measure of correlation is known as the correlation coefficient, which encapsulates both the direction and degree of correlation in a single figure. Therefore, correlation analysis encompasses the techniques employed to measure the proximity or relationship between variables.

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

-Croxton and Cowden

"Correlation is an analysis of the covariation between two or more variables."

-A.M. Tuttle

"Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilising forces may become effective."

-W.A. Neiswanger

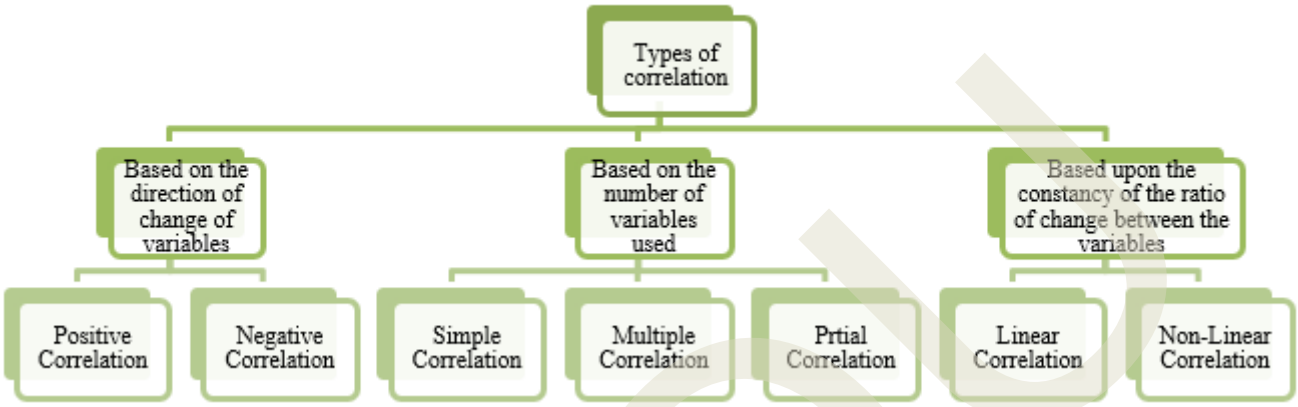
"The effect of correlation is to reduce the range of uncertainty of our prediction."

-Tippett

3.1.1.1 Types of Correlation

Correlation is classified into the following types.

- Positive and negative Correlation
- Linear and non-linear Correlation
- Simple, Partial and multiple Correlation



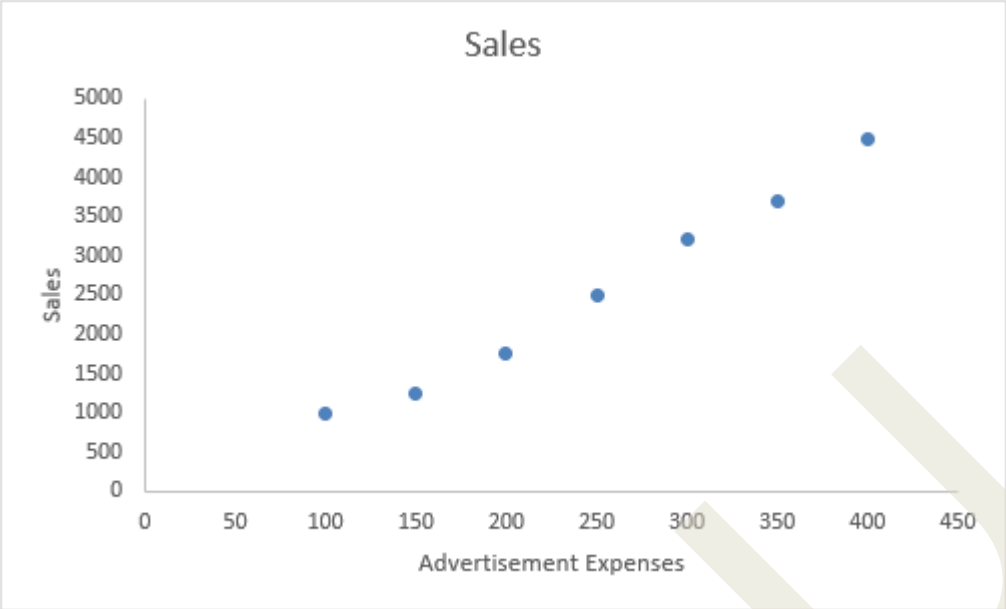
Positive and negative correlation

- Variables moving in same direction

The direction of correlation, whether it is positive (indicating a direct relationship) or negative (indicating an inverse relationship), hinges on how the variables change. When both variables move in the same direction, meaning that as one increases, the other, on average, also increases, or as one decreases, the other, on average, decreases, we classify this as a positive correlation. For instance, the more time you spend running on a treadmill, the more calories you burn, illustrating a positive correlation. The relationship between income and expenditure is another example of a positive correlation.

An example of positive correlation is given below:

Advertisement Expense	Sales
100	1000
150	1250
200	1750
250	2500
300	3200
350	3700
400	4500



- Variable moving in opposite direction

Conversely, when the variables vary in opposite directions, meaning that as one increases, the other decreases, or vice versa, we refer to this as a negative correlation. For example, the correlation between the price of a commodity and its demand is a negative correlation. Similarly, the relationship between volume and pressure in a gas follows a negative correlation pattern.

An example of negative correlation is given below:

Price	Demand
10	150
12	130
15	110
17	95
19	70
21	65
28	50

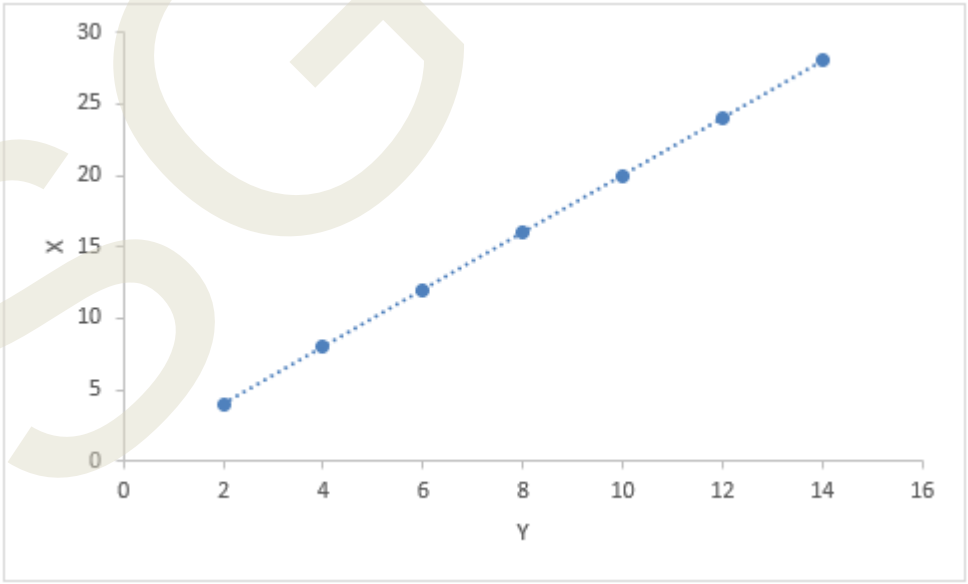
Linear and Non-linear correlation

Consider the values of two variables.

X:	2	4	6	8	10	12	14
Y:	4	8	12	16	20	24	28

- Equal proportion of change

It is observed that ratio of change between the two variables remains constant. If the ratio of change between the two variables is a constant value, then the correlation between the variables is called linear correlation. If we plot these points on a graph, we get a straight line.

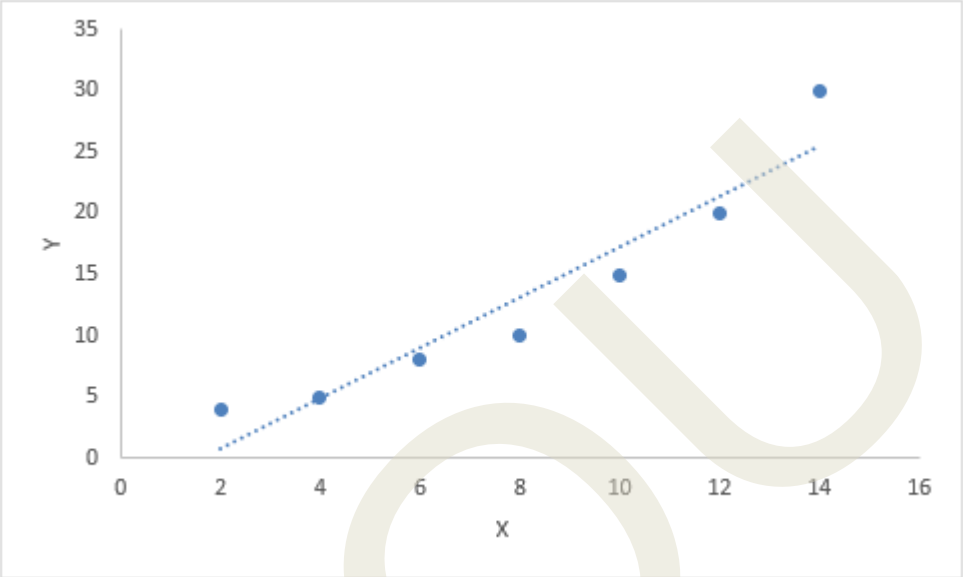


- Unequal proportion of change

However, if the amount of change in one variable does not bear a constant ratio with the amount of change in the other variable, the relation is called non-linear correlation and the resultant graph will be a curve. In a non-linear correlation, the points on the graph will not lie on a straight line.

Consider the following scenario

X:	2	4	6	8	10	12	14
Y:	4	5	8	10	15	20	30



This is an example of Non-linear correlation

Simple, Partial and Multiple correlation

Simple Correlation

- Two Variables

This distinction is based on the number of variables under study. When examining only two variables, it is referred to as simple correlation. For instance, if we study the relationship between the yield of paddy and the fertilizer used, it is a simple correlation. More details about the simple correlation are given as follows:

- ***Number of variables:*** Two.
- ***Concept:*** Assesses the strength and direction of the linear relationship between two variables without controlling for the influence of any other variables.
- ***Calculation:*** Uses Pearson's correlation coefficient (r), ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value of 0 indicates no linear relationship.
- ***Example:*** Examining the correlation between ice cream sales and temperature on a daily basis.

Multiple Correlation

- More than two variables

When three or more variables are studied, it is either multiple or partial correlation. In multiple correlation, three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre, with the amount of rainfall, and the quantity of fertilizers used, it is an example of multiple correlation. More details about the multiple correlation are given as follows:

- ***Number of variables***: Three or more.
- ***Concept:*** Assesses the strength and direction of the relationship between one dependent variable and multiple independent variables taken together. It essentially measures how well the set of independent variables predicts the dependent variable.
- ***Calculation*****:** Uses the multiple correlation coefficient (R), ranging from 0 (no linear relationship) to 1 (perfect linear relationship).
- ***Example*****:** Determining how well income, education level, and work experience affects social status of an individual

Partial Correlation

- One variable is kept constant

On the other hand, in partial correlation, we study more than two variables, but consider only two variables that influence each other, while keeping the effect of other influencing variables constant. In the study of production of rice, if we consider yield and rainfall only, maintaining the daily temperature as constant, it illustrates an example of partial correlation.

- ***Number of variables*****:** Three or more.
- ***Concept*****:** Measures the correlation between two specific variables while controlling for the influence of one or more additional variables. This helps isolate the unique relationship between the two variables of interest.
- ***Calculation*****:** Uses more complex formulas depending on the number of control variables.
- ***Example*****:** Studying the correlation between student test scores and hours of study while controlling for age and socioeconomic status.

Distinctions between Simple, Partial, and Multiple correlation

Feature	Simple Correlation	Partial Correlation	Multiple Correlation
Number of variables	2	3+	3+
Controls other variables	No	Yes	Yes
Measures	Overall linear relationship between two variables	Unique linear relationship between two variables controlling for others	Strength of relationship between one dependent variable and multiple independent variables
Uses	Initial exploration, preliminary analysis	Refining specific relationships, isolating unique effects	Determining predictive power of multiple variables

- Not related to each other

No correlation

If the variables do not have a relationship with each other, then there is no correlation.

3.1.1.2 Uses/Applications of correlation

i. **Financial:** Correlation is used in finance to analyse the relationship between different financial assets.

ii. **Economics:** Correlation analysis helps in analysing the economic behaviour. For example, wages and the level of inflation, savings and the rate of tax collected etc.. By understanding these connections, economists can propose strategies to enhance the prevailing situation.

iii. **Social sciences:** The importance of analysing and understanding the relationship between two or more variables is increasing day by day with the developments in the field of social sciences. For example, to study the relationship between smoking and lung cancer, number of accidents and the sale of vehicles etc.

iv. **Climate science:** Climate scientists use correlation to analyse relationship between different climate variables such as temperature and atmospheric levels to understand how they affect each other.

v. **Education:** Correlation can be used in educational research to study relationship between variables like teaching methods and academic performance.

3.1.1.3 Importance of correlation

The following are the importance of correlation:

i. Understanding relationships

Correlation serves as a tool to gauge relationships between variables. Imagine you are in charge of human resources at a company. You want to assess whether the level of training provided to employees is linked to their job performance. Correlation can help you understand this connection, much like evaluating the relationship between studying for an exam and getting a good grade.

ii. Predicting trends

Correlation empowers commerce professionals to foresee trends. For instance in the stock market, understanding the correlation between two stocks enables traders to make more informed predictions.

iii. Identifying key factors

Correlation helps identify which variables are most crucial. This is analogous to picking the essential ingredients for a recipe. For accountants, it streamlines financial analysis by spotlighting interconnected metrics.

iv. Quality control

Consistency is paramount in commerce. Correlation ensures that changes in one part of a process reflect accurately in another, making it indispensable for quality control in manufacturing and services.

v. Risk management

Correlation aids in managing risks, much like playing cards strategically. Understanding how different variables correlate helps investors diversify portfolios and reduce potential losses.

vi. Optimising resource allocation

In commerce, resources such as advertising budgets must be allocated judiciously. Correlation helps identify which strategies yield the best results, enabling businesses to optimise their spending.

vii. Enhanced decision making

Correlation assists in data-driven decision making. For instance, it reveals whether online or television advertising has a greater impact on sales, facilitating informed choices.

viii. Refining models

In economic modeling, correlation is pivotal for constructing accurate models. It is similar to selecting the right ingredients and proportions for a recipe, ensuring the model mirrors real-world dynamics faithfully.

3.1.1.4 Limitations of correlation

The following are the limitations of correlation:

i. Extreme values can disproportionately affect correlation, leading to inaccurate results.

ii. Correlation captures only strength and direction, not the full nature of relationships.

iii. Small samples might not represent the population accurately, affecting correlation results.

iv. Uncontrolled or unaccounted variables might influence the relationship between the correlated variables, leading to misleading interpretations.

3.1.1.5 Coefficient of correlation

Degree of relationship between two variables is called coefficient of correlation. It is an algebraic method of measuring correlation. Coefficient of correlation is denoted by the symbol 'r' and the value of 'r' lies between −1 and +1.

i.e., $-1 \leq r \leq 1$

3.1.1.6 Properties of Correlation coefficient

The following are the properties of correlation coefficient:

i. Coefficient of correlation lies between −1 and +1.

ii. When r lies between 0 and 1, the correlation is positive, when r lies between -1 and 0, the correlation is negative. If r = 0 there is no correlation.

iii. If $r = +1$ it is perfect positive correlation. If $r = -1$ it is called perfect negative correlation.

iv. Independent variables are uncorrelated but correlation coefficient of X and Y is same as Correlation coefficient of

Y and X

v. Correlation coefficient does not change with reference to change of origin.

3.1.1.7 Methods of Studying Correlation

The various methods of studying correlation coefficient are

a. **Scatter Diagram method**
b. Karl Pearson's Coefficient of Correlation
c. Spearman Rank Method
d. Concurrent Deviation Method, and

a. Scatter diagram method

- Graphical Representation

This method is the simplest and the easiest method for studying correlation. The two variables X and Y are plotted on a two-dimensional graph. One variable is represented along X axis and the other variable along Y axis. The graph thus plotted will represent the relationship between the variables. Thus, this study of relationship between two variables based on the graphical representation is called scatter diagram method.

The following diagrams of the scattered data depict different forms of correlation.

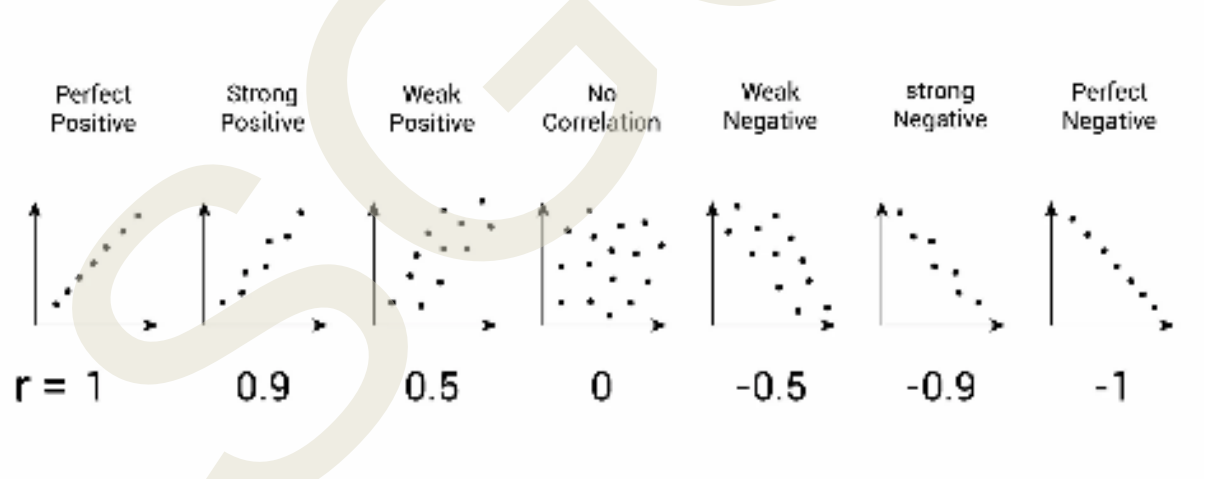
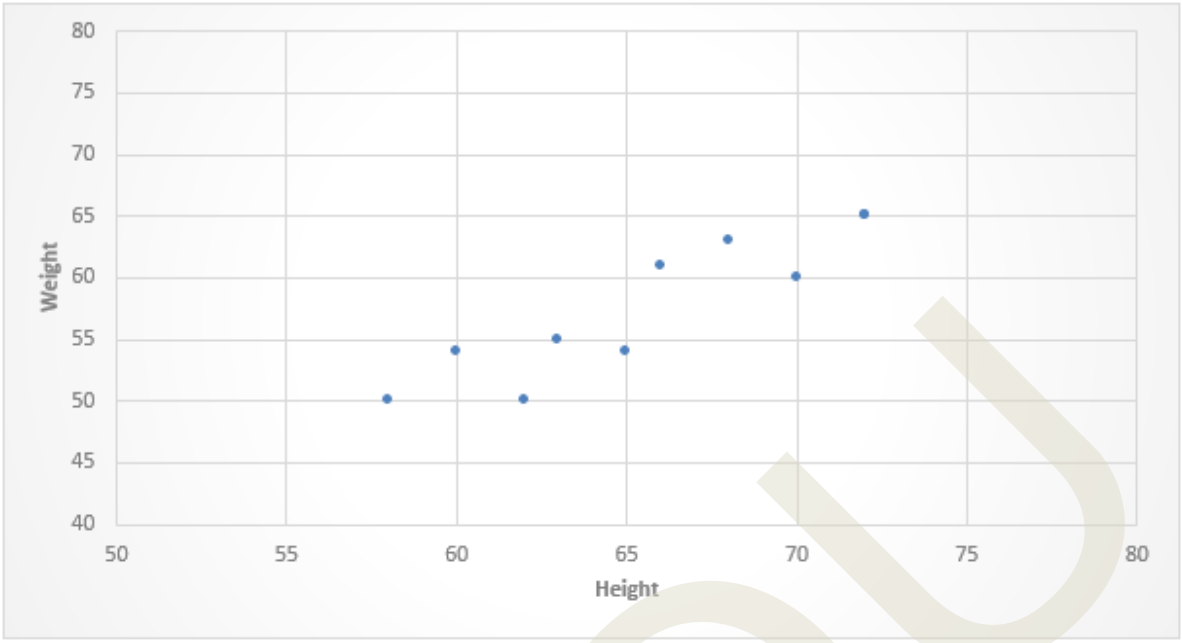


Illustration 3.1.1

Following are the heights and weights of 10 students of a B.Com Class. Calculate correlation using scatter diagram.

Height (in inch):	58	60	62	63	65	66	68	70	72	72
Weights (in kg):	50	54	50	55	54	61	63	60	65	65

Solution:



Since the points are dense i.e., close to each other, we may expect a high degree of correlation between the series of heights and weights. Further. since the points reveal an upward trend starting from left bottom and going up towards the right top. the correlation is positive. Hence, we may expect a fairly high degree of positive correlation between the series of heights and weights in the class of B.Com. students.

b. Karl Pearson's Coefficient of Correlation

This method is the most widely used method for measuring correlation. It is popularly known as Pearson coefficient of correlation. It is denoted by the symbol *"r"*. It is also known as Product Moment Method.

- Denoted as "r"

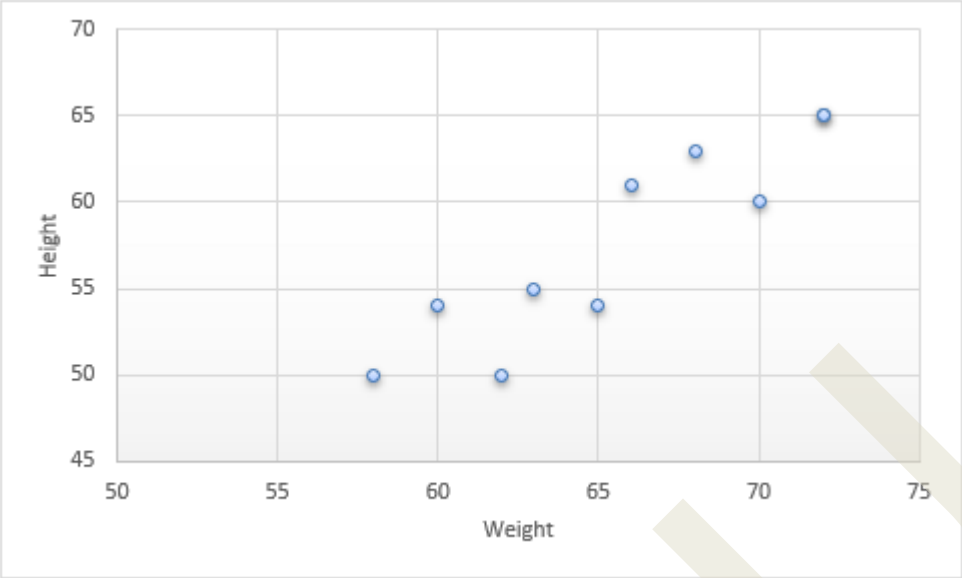
The value of correlation coefficient will always lie between -1 and +1. The positive sign indicates the positive correlation between the variables and the negative sign indicates negative correlation between the variables. The coefficient of correlation describes the magnitude of correlation and its direction.

Assumptions of Karl Pearson Coefficient of Correlation

Karl Pearson's coefficient of correlation is based on three assumptions:

i. **Normality**

The correlated variables are affected by a large number of in-



dependent factors, so that they acquire normality. Variables like age, height, weight, price, supply, demand etc., are affected by such forces so that a normal distribution is formed.

ii. **Causal Relationship**

There is cause and effect relationship between the forces affecting distribution of the items in the two series. So, if there is no such relationship, the correlation is meaningless.

iii. **Linear Nature**

It is assumed that, there is a linear relationship between the variables. In other words, if the pairs of items of both the variables are plotted on a graph paper, the plotted points will form a line.

Properties of Pearson Coefficient of Correlation

The value of the Karl Pearson's Coefficient of Correlation lies between -1 and +1. It cannot be greater than one in any case.

The Pearson Correlation Coefficient is independent of change of origin of X and Y variables

Pearson Coefficient of Correlation is the geometric mean of two regression coefficients, that is $r = \sqrt{b_{xy} \times b_{yx}}$

Merits of Karl Pearson's Coefficient of Correlation

i. ***Widely used and understood*:** The Pearson correlation coefficient is one of the most widely used and understood measures of correlation in statistics. This makes it easy to compare results across different studies and disciplines.

ii. ***Quantitative measure***: The Pearson correlation coefficient is a quantitative measure, which means that it provides a numerical value between -1 and 1 that indicates the strength and direction of the relationship between two variables. This makes it easier to interpret the results than some other measures of correlation, such as scatter plots.

iii. ***Linear relationships***: The Pearson correlation coefficient is specifically designed to measure linear relationships between two variables. This means that it is most accurate when the relationship between the two variables is straight line.

iv. ***Software availability***: The Pearson correlation coefficient is readily available in most statistical software packages, making it easy to calculate and interpret.

Limitations of Pearson Correlation Coefficient

i. ***Assumes normality***: The Pearson correlation coefficient assumes that the data is normally distributed. If the data is not normally distributed, the Pearson correlation coefficient may not be accurate.

ii. ***Sensitive to outliers***: The Pearson correlation coefficient is sensitive to outliers, which can skew the results. It is important to check for outliers and remove them if necessary before calculating the Pearson correlation coefficient.

iii. ***Only measures linear relationships***: The Pearson correlation coefficient can only measure linear relationships between two variables. If the relationship between the two variables is not linear, the Pearson correlation coefficient will not be accurate.

iv. ***Does not tell you about causality***: The Pearson correlation coefficient only tells you about the strength and direction of the relationship between two variables. It does not tell you whether one variable causes the other variable.

Computation of correlation coefficient

$$r(x,y) = \frac{Cov\,(x,y)}{\sigma(x)\sigma(y)}$$

Where, $Cov\ (x,y)$ = covariance of (x,y).

Covariance is a statistical measure that quantifies the degree to which two variables change together. It is the sum of the product of the average of the observations from arithmetic mean.

$$Cov\,(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$\sigma(x) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$ is the standard deviation of x.

$\sigma(y) = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$ is the standard deviation of y.

$\bar{x} = \frac{\Sigma x}{n}, \quad \bar{y} = \frac{\Sigma y}{n},$

So, $r(x,y)$

$$= \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}} = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\,\sqrt{\Sigma(y-\bar{y})^2}}$$

The above formula can be expressed in the following form also

$$r(x,y) = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Illustration 3.1.2

Find the Karl Pearson's coefficient between x and y for the following data.

N = 10, $\Sigma x = 35$, $\Sigma x^2 = 203$, $\Sigma y = 28$, $\Sigma y^2 = 140$,

Σxy =168

Solution:

$$r(x,y) = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10 \times 168 - 35 \times 28}{\sqrt{10 \times 203 - 35^2}\sqrt{10 \times 140 - 28^2}}$$

$$= \frac{1680-980}{\sqrt{805}\sqrt{616}}$$

= 0.99

Since the "r" value is 0.99, it indicates a very strong positive linear relationship between the variables. This means that as one variable increases, the other variable also tends to increase, and the relationship between them is highly correlated.

Illustration 3.1.3

Find Karl Pearson's correlation coefficient between x and y for the following data,

$$n = 15, Cov(x,y) = 8.13, \sigma(x) = 3.01, \sigma(y) = 3.03$$

Solution

$$r = \frac{Cov(x,y)}{\sigma(x)\sigma(y)} = \frac{8.13}{3.01 \times 3.03} = 0.89$$

Since the "r" value of 0.89 suggests a very strong positive linear relationship between the two variables. It indicates a high degree of correlation between the variables, where one variable tends to increase as the other increases.

Illustration 3.1.4

Find Karl Pearson's correlation coefficient between x and y for the following data,

$$n = 1000, \ \sigma(x) = 4.5, \ \sigma(y) = 3.6, \ \sum(x - \bar{x}).(y - \bar{y}) = 4800$$

Solution

$$r(x,y) = \frac{Cov\,(x,y)}{\sigma(x)\sigma(y)}$$

$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n}}{\sigma(x)\sigma(y)}$$

$$= \frac{\frac{4800}{1000}}{4.5 \times 3.6}$$

$$= \frac{4.8}{16.2}$$

$$= 0.296$$

Since the "r" value is 0.296, indicates a weak positive linear relationship between the two variables. Unlike the previous scenarios where the "r" values were close to 1, this value suggests a much weaker correlation.

With an "r" value of 0.296, there's still a positive relationship between the variables, implying that as one variable tends to increase, the other variable also tends to increase, but the relationship is not as strong or clear-cut.

In practical terms, a value of 0.296 suggests that while there

might be some tendency for the variables to move in the same direction, this relationship is relatively weak and might not be very reliable for making predictions or drawing strong conclusions compared to stronger correlations.

Pearson coefficient can be written another way

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

Steps for Computation

i. Take the deviations of X series from the mean of X series, and denote these deviations by "x"

ii. Square these deviations and obtain the total $\sum x^2$

iii. Take the deviations of Y series from the mean of Y series and denote these deviations by "y"

iv. Square these deviations and obtain the total $\sum y^2$

v. Multiply the deviations of X and Y series and obtain the total $\sum xy$

vi. Substitute the values of $\sum xy$, $\sum x^2$ and $\sum y^2$ in the above formula

Illustration 3.1.5

Find correlation from the following data

X:	06	08	09	14	17	28	24	31	07
Y:	10	12	15	15	18	25	22	26	28

Solution:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \text{ x } \sum y^2}}$$

Computation is explained in the following table

X	X-16 = x	X^2	Y	Y-19 = y	Y^2	xy
06	- 10	100	10	- 09	81	+90
08	- 08	64	12	- 07	49	+56
09	- 07	49	15	- 04	16	+28

14	- 02	04	15	- 04	16	+08
17	+01	01	18	- 01	01	- 01
28	+12	144	25	+06	36	+72
24	+08	64	22	+03	09	+24
31	+15	225	26	+07	49	+105
07	- 09	81	28	+09	81	- 81
144	**0**	**732**	**171**	**0**	**338**	**301**

$$\bar{X} = \frac{\sum X}{N}$$

$$= \frac{144}{9}$$

$$= 16$$

$$\bar{Y} = \frac{\sum Y}{N}$$

$$= \frac{171}{9}$$

$$= 19$$

$$r = \frac{301}{\sqrt{732 \times 338}}$$

$$= \frac{301}{497.409}$$

$$= 0.605$$

Since the r value is 0.605, indicates positive correlation, that means variables are moving in the same direction. If values of Variable X are increasing, the values of Y on an average is also increasing or vice versa.

Illustration 3.1.6

Calculate the Pearson coefficient of correlation from the following data;

X:	77	60	30	53	14	35	90	25	56	60
Y:	35	38	60	40	50	40	35	56	34	42

Solution:

X	X - 50 = x	x^2	Y	Y - 43 = y	y^2	xy
77	+27	729	35	- 08	64	- 216
60	+10	100	38	- 05	25	- 50
30	- 20	400	60	+17	289	- 340
53	+03	09	40	- 03	09	- 09
14	- 36	1,296	50	+07	49	- 252
35	- 15	225	40	- 03	09	+45
90	+40	1,600	35	- 08	64	- 320
25	- 25	625	56	+13	169	- 325
56	+06	36	34	- 09	81	- 54
60	+10	100	42	- 01	01	- 10
500	**0**	**5,120**	**430**	**0**	**760**	**- 1,531**

$$\bar{X} = \frac{\sum X}{N}$$

$$= \frac{500}{10}$$

$$= 50$$

$$\bar{Y} = \frac{\sum Y}{N}$$

$$= \frac{430}{10}$$

$$= 43$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$= \frac{-1531}{\sqrt{5120 \times 760}}$$

$$= \frac{1531}{1972.61}$$

= - 0.776

Since the "r" value is - 0.776, it indicates that variables are moving in the opposite directions. That means, if X variables are increasing, Y variables on an average is decreasing or vice versa.

Computation of Pearson Correlation Coefficient Taking Deviations from an Assumed Mean

If the actual mean is in fractions, computing Pearson coefficient of correlation based on actual mean will be difficult. Therefore, unless specified in the question we can compute Pearson coefficient of correlation based on assumed mean. Assumed mean can be a value from the series or outside the series. Irrespective of the assumed mean value, the answer will be uniform. Moreover, this will simplify the calculations also.

- Correlation from assumed mean

The formula for computing Pearson coefficient of correlation using assumed mean =

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{N}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{N}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{N}}}$$

Steps in Computation

i. Take the deviations of X series from an assumed mean, denote these deviations by dx and obtain the total, that is ∑dx

ii. Take the deviations of Y series from an assumed mean and denote these deviations by dy and obtain the total that is, ∑dy

iii. Square dx and obtain the total $\sum dx^2$

iv. Square dy and obtain the total $\sum dy^2$

v. Multiply dx with dy and obtain the total ∑dxdy

vi. Substitute the values in the equation, and apply the formula

Illustration 3.1.7

Using assumed mean, calculate Pearson coefficient of correlation for the following X and Y series.

X:	45	55	56	58	60	65	68	70	75	80
Y:	56	50	48	60	62	64	65	70	74	82

Solution:

For X series, deviations are taken from 65 and for Y series, deviations are taken from 66. As already mentioned, assumed mean can be any value within the series or outside the series. Here value 65 is in the X series, but value 66 is not there in Y series. The intention is to simplify the computations. Computation is explained in the following table:

X	X - 65 = dx	$(dx)^2$	Y	Y - 66 = dy	$(dy)^2$	dxdy
45	- 20	400	56	- 10	100	+200
55	- 10	100	50	- 16	256	+160
56	- 09	81	48	- 18	324	+162
58	- 07	49	60	- 06	36	+42
60	- 05	25	62	- 04	16	+20
65	00	00	64	- 02	04	0
68	+03	09	65	- 01	01	- 03
70	+05	25	70	+04	16	+20
75	+10	100	74	+08	64	+80
80	+15	225	82	+16	256	+240
632	**- 18**	**1,014**	**631**	**- 29**	**1,073**	**921**

Correlation coefficient is calculated using the equation

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{N}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{N}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{N}}}$$

Substituting the values in the equation, we get

$$= \frac{921 - \frac{(-18\,x - 29)}{10}}{\sqrt{1014 - \frac{[-18]^2}{10}} \times \sqrt{1073 - \frac{[-29]^2}{10}}}$$

$$= \frac{868.8}{31.330 \text{ x } 31.447}$$

$$= \frac{868.8}{985.24}$$

$$= +0.8818$$

High positive correlation (+0.8818) indicates that, variables are moving in the same direction, that means if values of X variable is increasing, the values of Y variable is also increasing or vice versa.

Illustration 3.1.8

Taking deviations from an assumed mean, find the correlation coefficient between X and Y variables are given below;

X:	20	30	30	20	19	23	35	16	38	40
Y:	18	35	20	18	25	28	33	18	20	40

Solution:

Deviations are taken from 30 for X and Y series. Computation is explained in the following table:

X	X - 30 = dx	$(dx)^2$	Y	Y - 30 = dy	$(dy)^2$	dxdy
20	- 10	100	18	- 12	144	120
30	0	0	35	+05	25	0
30	0	0	20	- 10	100	0
20	- 10	100	18	- 12	144	120
19	- 11	121	25	- 05	25	55
23	- 07	49	28	- 02	04	14
35	05	25	33	+03	09	15
16	- 14	196	18	- 12	144	168
38	08	64	20	- 10	100	- 80
40	10	100	40	+10	100	100
271	**- 29**	**755**	**255**	**- 45**	**795**	**512**

Correlation coefficient is calculated using the equation

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{N}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{N}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{N}}}$$

Substituting the values in the equation, we get

$$r = \frac{512 - \frac{(-29)(-45)}{10}}{\sqrt{755 - \frac{[-29]^2}{10}} \times \sqrt{795 - \frac{[-45]^2}{10}}}$$

$$= \frac{381.5}{\sqrt{670.9 \; x \; 592.5}}$$

$$= \frac{381.5}{25.9017 \; x \; 24.3413}$$

$$= \frac{381.5}{630.4825}$$

$$= +0.60509$$

A moderately high correlation (+0.60509) indicates that variables are moving in the same direction, that is, if values of X variables are increasing, values on an average Y is also increasing or vice versa.

Probable Error

- Reliability

The probable error of the coefficient of correlation helps in interpreting the value. Therefore, with the help of probable error it is possible to determine the reliability of the value of the coefficient in so far as it depends on the condition of random sampling.

The probable error of the coefficient of correlation is obtained as follows;

$$P.E = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

Where,

P. E = the notation for Probable Error

0.6745 = The Standard Error

r = The Pearson coefficient of correlation

N = Number of paired items

The Method of Interpretation

The interpretation of the computed value of Probable error is explained below;

i. If the value of "r" is less than the probable error, there is no evidence of correlation, that is the value of "r" is not at all significant.

ii. If the value of "r" is more than six times the probable error, the existence of correlation is practically certain, that is the value of "r" is significant.

iii. By adding and subtracting the value of probable error from the coefficient of correlation, we get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie, symbolically;

$\rho = r \neq P.E$

ρ is read as (rho) denotes correlation in the population

Conditions for Use of Probable Error

The measure of probable error can be properly used only when the following conditions are satisfied.

i. The data must approximate a normal frequency curve, that is a bell-shaped curve.

ii. The statistical measure for which the probable error is computed must have been calculated from a sample.

iii. The sample must have been selected in an unbiased manner and the individual items must be independent

Illustration 3.1.9

Find Karl Pearson's coefficient of correlation, from the following series of marks secured by ten students in a class test in mathematics and statistics.

Marks in Mathematics:	45	70	65	30	90	40	50	75	85	60
Marks in Statistics:	35	90	70	40	95	40	60	80	80	50

a. Also calculate its probable error. Assume 60 and 65 as working means?

b. Hence, discuss if the value of "r" is significant or not? Also compute the limits within which the population correlation coefficient may be expected to lie?

Solution

Computation of correlation coefficient is explained in the following table;

X	dx = X- 60	$(dx)^2$	Y	dy =Y- 65	$(dy)^2$	dxdy
45	- 15	225	35	- 30	900	450
70	+10	100	90	+25	625	250
65	+05	25	70	+05	25	25
30	- 30	900	40	- 25	625	750
90	+30	900	95	+30	900	900
40	- 20	400	40	- 25	625	500
50	- 10	100	60	- 05	25	50
75	+15	225	80	+15	225	225
85	+25	625	80	+15	225	375
60	0	0	50	- 15	225	00
610	**+10**	**3,500**	**640**	**- 10**	**4,400**	**3,525**

Here deviations are taken from assumed mean, therefore correlation coefficient is calculated using the following equation;

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{N}}{\sqrt{\sum dx^2 - \frac{[\sum dx]^2}{N}} \times \sqrt{\sum dy^2 - \frac{[\sum dy]^2}{N}}}$$

Substituting the values in the equation, we get;

$$r = \frac{3{,}525 - \frac{(10)(-10)}{10}}{\sqrt{3{,}500 - \frac{[10]^2}{10}} \times \sqrt{4{,}400 - \frac{[-10]^2}{10}}}$$

$$= \frac{3535}{3914.2176}$$

$$= + 0.9031$$

A very high positive correlation between the variables tells that variable are moving in the same direction, that is, if marks in mathematics is high, then marks in statistics will also be high, and viceversa.

Probable error of correlation coefficient is calculated using the equation;

$$P.E = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

Substituting the values in the equation, we get

$$P.E = 0.6745 \frac{1 - (0.9031)^2}{\sqrt{10}}$$

$= 0.0393$

Here the value of r = 0.9031 and P. E= 0.0393

Since the value of "r" is more than P.E, we have to check whether the value of "r" is significant or not, that means the "r" value should be more than six times the probable error, to conclude that the value of "r" is significant. We have r = 0.9031

6 x P.E = 6 x 0.0393 = 0.2358

Here r > 6 x P.E

So, we can conclude that the value of "r" is highly significant. This means that higher the marks of a candidate in Mathematics, higher the marks in Statistics also and lower the marks of a candidate in Mathematics, lower is his/her score in Statistics also.

Limits for population correlation coefficient is given by the formula

r $\pm$ P.E

That means we have to add and subtract the value of "r" to the probable error value to find out the limits of population. So, we get 0.9031 $\pm$ 0.0393=0.8638 to 0.9424

This implies that if we take another sample of ten items from the same population, then its correlation coefficient is expected to lie between 0.8638 and 0. 9424

Illustration 3.1.10

If "r" is 0.6 and N = 64 of a distribution, find out the probable error.

Solution

Probable Error is calculated using the equation

$$P.E = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

$$P.E = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

Substituting the values in the equation, we get;

$$P.E = 0.6745 \frac{1 - 0.6^2}{\sqrt{64}}$$

= 0.6745 x 0.08

= 0.05396

"r" is higher than probable error, so we can check whether "r" is significant or not, that is 6 x P.E, for that we have to calculate by substituting the values in the equation

6 x 0.05396=0.32376

"r" > 6 x P.E

So, "r" is significant

Illustration 3.1.11

Calculate probable error and limits of population from the following data;

N = 6 and r = -0.9203

Solution

Probable error is calculated using the equation

$$P.E = 0.6745 \frac{1 - r^2}{\sqrt{N}}$$

Substituting the values in the equation, we get

$$P.E = 0.6745 \frac{1 - (-0.9203)^2}{\sqrt{6}}$$

= 0.6745 x 0.0624815

= 0.04214

"r" > P.E, so we can check whether the value of "r" is significant or not

6 x P.E = 0.04214 x 6 = 0.25284

Again "r" > 6 x P.E

So, we can conclude that the value of "r" is significant.

Limits of population is calculated using the equation

r $\pm$ P.E

Substituting the values in the equation, we get

-0.9203 + 0.04214= - 0.87816

-0.9203 – 0.04214 = -0.96244

So, the limits of population will be between -0.87816 and -0.9624. This implies that if we take another sample of six items

from the same population, the value of Pearson Correlation Coefficient will be between -0.87816 to -0.96244.

Standard error

The standard error is a measure of the variability or precision of a sample statistic, such as the mean or correlation coefficient, in relation to the true population parameter. A smaller standard error indicates a more precise estimate, suggesting that the sample statistic is likely closer to the true population parameter. Conversely, a larger standard error implies greater uncertainty and less precision in the estimate. The standard error is calculated using the standard deviation of the sample and the sample size. For example, in the context of the correlation coefficient (r), the formula is

- Variability

$$SE(r)=\sqrt{\frac{(1-r)^2}{n-2}}$$

c. Rank Correlation Coefficient

This method was developed by the British Psychologist Charles Edward Spearman in 1904. Karl Pearson's Coefficient of Correlation can be used only if the characteristics under study are quantitative, that is when the variables under study are numerically measurable. But Spearman's Rank Correlation Coefficient can be calculated even if the characteristics under study are qualitative. Examples are evaluation of leadership ability or the judgement of a beauty contest etc.

- Qualitative data

Spearman's Rank Correlation Coefficient, usually denoted by ρ (Rho) is given by the following formula

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

or

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where,

ρ = Spearman's Rank Correlation Coefficient

N = Number of observations

D = The difference of ranks between paired items in the two series

The value of this coefficient also ranges between +1 and -1. When 'ρ' is plus, it indicates that there is complete agreement in the order of ranks and the ranks are in the same direction. When 'ρ' is minus, it indicates that there is no complete agreement in

the order of ranks and they are in opposite directions.

Advantages of Spearman Rank Correlation:

i. **Non-parametric Measure:** Spearman rank correlation is a non-parametric measure, which means it does not rely on the assumption of normal distribution in the data. This makes it suitable for use with ordinal or non-normally distributed data.

ii. **Robust to Outliers:** Spearman's correlation is less sensitive to outliers compared to Pearson's correlation. Outliers can significantly impact Pearson's correlation, but Spearman's correlation is based on ranks, making it more robust in the presence of extreme values.

iii. **Simple Interpretation:** The correlation coefficient in Spearman rank correlation is easy to interpret. A positive value indicates a positive monotonic relationship, while a negative value suggests a negative monotonic relationship.

iv. **No Assumption of Linearity:** Unlike Pearson correlation, Spearman correlation does not assume a linear relationship between variables. It captures monotonic relationships, which may not be strictly linear.

Disadvantages of Spearman Rank Correlation:

i. **Loss of Information:** When using ranks, some information about the magnitude of differences between values is lost. This can result in a reduction in statistical power compared to methods that use the actual values.

ii. **Equal Ranks Issue:** In situations where there are tied ranks (equal values in the dataset), Spearman's correlation tends to assign average ranks. This can lead to a potential loss of precision, especially when dealing with small sample sizes.

iii. **Limited Sensitivity:** Spearman's correlation is generally less sensitive to changes in the center of the distribution compared to Pearson's correlation. It may not detect certain types of relationships that Pearson's correlation would identify.

iv. **Not Suitable for All Data Types:** While Spearman's correlation is robust for ordinal data, it may not be the best choice for all types of data. For example, if the relationship between variables is genuinely linear, Pearson's correlation might be more appropriate.

In rank correlation coefficient we have to do two types of

problems;

i. When actual ranks are given, or

ii. When ranks are not given

iii. When the ranks are repeated

i. When actual ranks are given

If the actual ranks are given, the steps required for computing Spearman's CorrelationCoefficient are;

Take the differences of the two ranks, that is (R1- R2) and denote these differences by 'd'

Square these differences and obtain the total '$\sum d^2$'

Apply the formula $\rho = 1 - \frac{6\sum d^2}{n^3 - n}$

Illustration 3.1.12

Two judges in a beauty contest rank the ten entries as follows. Compute theSpearman's Rank Correlation and interpret the data.

X:	01	02	03	04	05	06	07	08	09	10
Y:	08	07	06	04	03	02	05	01	09	10

Solution

Computation is explained in the following table

R_1	R_2	$R_1 - R_2 = d$	d^2
01	08	- 07	49
02	07	- 05	25
03	06	- 03	09
04	04	00	00
05	03	+02	04
06	02	+04	16
07	05	+02	04
08	01	+07	49

09 10	09 10	00 00	00 00
			156

Rank Correlation coefficient is calculated using the equation,

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Here N= 10 and $d^2 = 156$, substituting the values in the equation, we get

$$\rho = 1 - \frac{6 \times 156}{10^3 - 10}$$

$$\rho = 1 - \frac{936}{990}$$

$$= 0.0545$$

Positive rank correlation indicates that ranks are in the same direction, but here a negligible positive correlation value is obtained. It indicates that though the ranks are in the same order, the judges are not in complete agreement in common tastes to beauty.

Illustration 3.1.13

Ten competitors in a music competition were ranked by three judges X, Y and Z in the following order;

Judge X:	01	06	05	10	03	02	04	09	07	08
Judge Y:	03	05	08	04	07	10	02	01	06	09
Judge Z:	06	04	09	08	01	02	03	10	05	07

By using Rank Correlation Coefficient, find out which pair of judges have the nearest approach to the common taste in music?

Solution

Ranks by judge X is denoted as R_1, by Judge Y as R_2 and by Judge Z as R_3. So, we have to calculate the rank correlation coefficient between R_1 and R_2, R_2 and R_3 and R_1 and R_3.

R1	R2	R3	R1- R2	d^2	R2- R3	d^2	R1-R3	d^2
01	03	06	- 02	04	- 03	09	- 05	25
06	05	04	01	01	01	01	02	04
05	08	09	- 03	09	- 01	01	- 04	16
10	04	08	06	36	- 04	16	02	04
03	07	01	- 04	16	06	36	02	04
02	10	02	- 08	64	08	64	00	00
04	02	03	02	04	- 01	01	01	01
09	01	10	08	64	- 09	81	- 01	01
07	06	05	01	01	01	01	02	04
08	09	07	- 01	01	02	04	01	01
				200		**214**		**60**

Rank correlation coefficient between 1st and 2nd Judge can be calculated using the equation;

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Substituting the values in the equation, we get;

$$\rho = 1 - \frac{6 \text{ x } 200}{10^3 - 10}$$

$$= -0.212$$

Rank correlation coefficient between 2nd and 3rd judge is calculated using the equation;

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Substituting the values in the equation, we get;

$$\rho = 1 - \frac{6 \text{ x } 214}{10^3 - 10}$$

$= -0.2969$

Rank correlation coefficient between 1st and 3rd judge is calculated using the equation,

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Substituting the values in the equation, we get;

$$\rho = 1 - \frac{6 \times 60}{10^3 - 10}$$

$= 0.63636$

Here, we get the highest rank correlation coefficient value between 1st and 3rd judge. Thus, it shows that the first and third Judges have the nearest approach to common tastes in beauty.

ii. When ranks are not given

- Need to assign ranks

When we are given the actual data and not the ranks, we have to assign the ranks. Ranks can be assigned by taking either the highest value as the first rank or the lowest value as the first rank. But the same method has to be followed for both the variables.

Illustration 3.1.14

Calculate Spearman's Coefficient of Correlation from the following data;

X:	53	98	95	81	75	61	59	55
Y:	47	25	32	37	30	40	39	45

Solution

We have to assign ranks for the given X and Y values. Here ranks are assigned by taking the highest value as one, therefore from X series 98 got the first rank, 95 second rank and so on. For Y series, 47 got the first rank and 45 the second rank and so on. Now computation is explained in the following table with the assigned ranks.

X	R_1	Y	R_2	$d = R_1 - R_2$	d^2
53	08	47	01	+07	49
98	01	25	08	-07	49
95	02	32	06	-04	16

81	03	37	05	-02	04
75	04	30	07	-03	09
61	05	40	03	+02	04
59	06	39	04	+02	04
55	07	45	02	+05	25
					160

Spearman's Rank Correlation Coefficient is calculated using the equation,

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Here N= 08 and $d^2 = 160$, substituting the values in the equation, we get

$$\rho = 1 - \frac{6 \text{ x } 160}{8^3 - 8}$$

$$= -0.90476$$

High negative correlation indicates that ranks are not in the same order.

Illustration 3.1.15

Find the Spearman's rank correlation coefficient between marks in Accountancy and Statistics.

Marks in Statistics:	48	60	72	62	56	40	39	52	30
Marks in Accountancy:	62	78	65	70	38	54	60	32	31

Solution

Computation is explained in the following table with the assigned ranks. The ranks are assigned as the highest value gets the first rank.

X	**R_1**	**Y**	**R_2**	**$d = R_1 - R_2$**	**d^2**
48	06	62	04	02	04
60	03	78	01	02	04
72	01	65	03	-02	04

62	02	70	02	00	00
56	04	38	07	-03	09
40	07	54	06	01	01
39	08	60	05	03	09
52	05	32	08	-03	09
30	09	31	09	00	00
					40

Spearman's Rank Correlation Coefficient is calculated using the equation,

$$\rho = 1 - \frac{6\sum d^2}{n^3 - n}$$

Here N= 09 and $d^2 = 40$, substituting the values in the equation, we get

$$\rho = 1 - \frac{6 \times 40}{9^3 - 9}$$

$$= 0.6667$$

A moderately high rank correlation coefficient indicates that higher the marks of a student in Accountancy, higher will be the marks in Statistics also, that is, the ranks are in the same order.

iii. In case the ranks are repeated

In certain cases, we may encounter situations where two or more items share the same rank. In such instances, each individual item is assigned an average rank. "If two values are ranked equal at the third place, they are each given the rank $\frac{3+4}{2} = 3.5$. But if there are three values ranked equal at the third place, then the individual ranks will be $\frac{3+4+5}{3}$

- Presence of equal ranks

When equal ranks are given to some entries, some adjustments in the formula are to be made in calculating the Rank Correlation Coefficient. The formula in that case is written as

$$\rho = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots \frac{1}{12}(m_n^3 - m_2)]}{N^3 - N}$$

Where "m" stands for the number of items which have the common rank. In case, there are more than one such group of items with common rank, the value is added as many times as

the number of such groups. The computation is explained with the help of few examples.

Illustration 3.1.16

Calculate the coefficient of rank correlation from the following data,

X:	48	33	40	09	16	16	65	24	16	57
Y:	13	31	31	06	15	04	20	09	06	19

Solution

Ranks are assigned as follows for X series

65= The highest value, so first rank

57= Gets the second rank

48= Gets the third rank

40= Gets the fourth rank

33= Gets the fifth rank

24= Gets the sixth rank

Now the next highest value 16 is repeated thrice, therefore average of the next three ranks will be taken, that is $\frac{7+8+9}{3} = 8^{th}$ rank. So, rank 8 will be assigned to all the values of 16. The last value 9 gets the 10^{th} rank.

Now, let us explain how the ranks for Y series are assigned

The highest value 31 is repeated twice, so the respective ranks will be $\frac{1+2}{2} = 1.5$ each

The next value 20 gets 3^{rd} rank

Next value 19^{th} gets the 4^{th} rank

Next value 15 gets the 5^{th} rank

Next value 13 gets the 6^{th} rank

Next value 9 gets the 7^{th} rank

Next value 6 is repeated twice, so the next two ranks will be averaged and assigned, that is, $\frac{8+9}{2} = 8.5$

The last value 4 gets 10^{th} rank

Computation of rank correlation coefficient is explained in the following table. Ranks assigned for X series are taken as R_1

X	R_1	Y	R_2	$d=R_1-R_2$	d^2
48	03	13	06	- 03	09
33	05	31	1.5	+3.5	12.25
40	04	31	1.5	+2.5	6.25
09	10	06	8.5	+1.5	2.25
16	08	15	05	+03	09
16	08	04	10	- 02	04
65	01	20	03	- 02	04
24	06	09	07	- 01	01
16	08	06	8.5	- 0.5	0.25
57	02	19	04	- 02	04
					52

and that for Y series are denoted as R_2.

Rank Correlation Coefficient is calculated using the equation;

$$= 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots \frac{1}{12}(m_n^3 - m_2)]}{N^3 - N}$$

Here, three correction factors are used, so that three correction factors are added to the equation; The formula becomes

$$\rho = 1 - \frac{6[(52 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{10^3 - 10}$$

$$\rho = 1 - \frac{6(52 + 2 + 0.5 + 0.5)}{990}$$

$$= 1 - \frac{6 \times 55}{990}$$

$$= 0.66667$$

Here in the X series value 16 is repeated thrice, so for that we added the correction factor $\frac{1}{12}(3^3 - 3)$. Remember that 'm' stands for the number of items which have common rank.

Similarly for Y series 31 and 6 are repeated twice. Therefore, correction factors are added, that is $\frac{1}{12}(2^3 - 2)$.

A moderately high rank correlation coefficient indicates that the ranks are in the same direction, that means higher the ranks

for X series, higher the ranks will be for Y series also.

Illustration 3.1.17

Eight students have obtained the following marks in Economics and Accountancy. Calculate the rank correlation coefficient.

Marks in Accountancy :	25	30	38	22	50	70	30	90
Marks in Economics :	50	40	60	40	30	20	40	70

Solution

Computation is explained in the following table.

X	R_1	Y	R_2	$d = R_1 - R_2$	d^2
25	07	50	03	+04	16
30	5.5	40	05	+0.5	0.25
38	04	60	02	+02	04
22	08	40	05	+03	09
50	03	30	07	- 04	16
70	02	20	08	- 06	36
30	5.5	40	05	+0.5	0.25
90	01	70	01	00	00
					81.5

Here two correction factors are to be added to the equation, for X series, 30 is repeated twice, so the correction factor $\frac{(2^3 - 2)}{12}$ is added. Similarly for Y Series value 40 is repeated thrice, so the correction factor $\frac{(3^3 - 3)}{12}$ is added.

The rank correlation coefficient can be calculated using the equation;

$$\rho = 1 - \frac{6[(\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots \frac{1}{12}(m_n^3 - m_2)]}{N^3 - N}$$

Substituting the values in the equation, we get;

$$\rho = 1 - \frac{6(81.5 + 0.5 + 2)}{504}$$

= 1-1.00

= 0

This indicates that there is no relationship, the correlation value is zero between the marks in economics and accountancy among the eight students.

d. Concurrent deviations method

- Changes denoted as 'sign'

The concurrent deviations method of correlation is a simple way to assess the direction and strength of the relationship between two variables based on their changes. It focuses on whether the variables tend to change together in the same direction (positive correlation), opposite directions (negative correlation), or independently (no correlation).

Summarised Overview

Correlation is a statistical measure that quantifies the degree of association between two variables, providing insights into the strength and direction of their relationship. The correlation coefficient, denoted by "r," ranges from -1 to +1, indicating perfect negative and positive correlations, respectively, with 0 representing no correlation. This versatile statistical tool is widely applied in fields such as economics, finance, biology, and social sciences, aiding in the exploration of relationships between variables and making predictions based on observed patterns.

In practical applications, correlation is crucial for understanding and decision-making in various contexts. In finance, analysts use correlation to assess how different assets move relative to each other, facilitating portfolio diversification and risk management. In medical research, correlation helps investigate potential links between lifestyle factors and health outcomes. The ability to comprehend the correlation between variables empowers researchers, analysts, and decision-makers to identify trends, assess dependencies, and make more informed predictions about future outcomes.

There are several methods for calculating correlation, each with its own strengths and limitations. The Pearson correlation coefficient is commonly used to measure linear relationships, while the Spearman rank correlation assesses monotonic relationships, regardless of linearity. Scatter diagram diagrammatically represents correlation. Additionally, the concurrent deviation method is utilized, providing a practical approach to quantify the correlation between variables by considering the joint deviations from their respective means. These diverse methods offer flexibility in capturing different aspects of relationships between variables, making correlation a versatile and indispensable tool in statistical analysis

Self-Assessment Question

1. Explain the uses of correlation in your daily life.
2. Differentiate Karl Pearson coefficient of correlation and Spearman rank correlation.
3. Making use of the data summarized below, calculate the Pearson coefficient of correlation

Case:	A	B	C	D	E	F	G	H
X:	10	6	9	10	12	13	11	9
Y:	9	4	6	9	11	13	8	4

(Answers; r = + 0.8958)

4. Calculate Pearson's correlation of coefficient from the following data using 44 and 26 respectively as the origin of X and Y series respectively

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

(Answer; r = -0.73266)

5. Calculate the Pearson Correlation Coefficient between income and weight from the following data. Also comment on the result.

Income (Rs):	100	200	300	400	500	600
Weight (Lbs.):	120	130	140	150	160	170

(Answer: r = +1, Perfect positive correlation, *This means that the value of "r" is highly significant, that means as the income increases, weight also increases or vice versa*.)

6. Find out Karl Pearson's Coefficient of Correlation from the following data of marks obtained by ten students in a class test

Marks in Economics	Marks in Accountancy
45	35
70	90
65	70
30	40
90	95
40	40
50	60
75	80
85	80
60	50

(Answer: r = + 0.9031)

7. Find the Karl Pearson's Coefficient of Correlation between the following two variables. Comment on the result through the probable error?

X:	06	08	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

(Answer: r = + 0.958668, P.E = 0.0182006, P = +0.940468 to + 0.976868)

8. A group of ten workers in a factory is ranked according to their efficiency by two judges as follows, compute Spearman's rank Correlation coefficient?

Judge X:	04	08	06	07	01	03	02	05	10	09
Judge Y:	03	09	06	05	01	02	04	07	08	10

(Answer: ρ = +0.87878)

9. From the following data, calculate the Spearman's Coefficient of Correlation between X and Y

X	33	56	50	65	44	38	44	50	15	26
Y	50	35	70	25	35	58	75	60	55	26

(Answer: ρ = -0.07575)

10. Calculate the Spearman's rank Correlation Coefficient for the following data

X	60	34	40	50	45	41	22	43	42	66	64	46
Y	75	32	35	40	45	33	12	30	36	72	41	57

(Answer: ρ = + 0.83217)

11. Given $\sum d^2$ = 40 and N =10, find the value of Rank Correlation Coefficient?

(Answer: ρ = + 0.7575758)

Assignments

1. What are the different methods of calculating correlation?
2. Compute Pearson coefficient of correlation from the following data.

X:	02	04	05	06	08	11
Y:	18	12	10	08	07	05

Multiply each X value in the table by 2 and add 6. Multiply each value of Y in the table by 3 and subtract 15. Find out the correlation coefficient between the variables. Explain why you do or do not obtain the same result as the first set of data?

(Answer; r = -0.9203, *We obtain the same correlation coefficient value for both sets of data. The results are identical because the correlation coefficient is independent of changes in origin and scale*)

3. Given the following, calculate the value of N? r = 0.61, P.E = 0.1312

(Answer: N=10.42, *but the value of N cannot be in fractions, so the value of N= 10)*

4. From the following data of six commodities, calculate Karl Pearson's Correlation Coefficient between sales and price per kg sold

Commodities	Sales (Kgs)	Total Price (Rs)
Rice	100	8,000
Wheat	120	10,000
Sugar	150	13,000
Gram	100	10,000
Oil	160	19,200
Ghee	170	30,940

(Answer: r = +0.84858)

5. From the following data, calculate Karl Pearson's Correlation coefficient. Arithmetic means of X and Y series are 6 and 8 respectively.

X:	06	02	10	---	08
Y:	09	11	---	08	07

(Answers: Missing 'X' Value = 4, Missing Y value = 5, r = - 0.9192)

6. From the following data, compute the Karl Pearson coefficient of Correlation between X and Y

Particulars	X Series	Y Series
Arithmetic Mean	15	28
Sum of Squares of Deviation from Arithmetic Means	144	225

Summation of products of deviations of X and Y series from their respective means = 20
Number of pairs= 10

(Answer: r = + 0.111)

7. The following data relates to the percentage of failures in the Higher Secondary examination. Find out whether there is any correlation between age and failure in the examination and interpret the result by finding the Probable Error.

Age of Candidates (in Years):	13	14	15	16	17	18	19	20	21
Percentage of Failure:	39	40	43	34	36	39	48	47	52

(Answer: r = +0.6741, P.E= 0.12265555, r > P.E, so we have to check whether r > 6 x P.E is 0.7359333, here r < 6 x P.E, hence the value of r is not significant.)

8. Obtain the Rank Correlation Coefficient for the following data showing the ranks of ten girls in a painting competition, given by two judges. Comment on the value obtained.

Judge X:	01	02	03	04	05	06	07	08	09	10
Judge Y:	03	01	05	02	07	04	09	06	10	08

(Answer: ρ = + 0.7939)

9. Calculate the coefficient of correlation from the following data by the Spearman's Rank Differences method.

Price of Tea (Rs);	75	88	95	90	60	80	81	50
Price of Coffee (Rs):	120	134	150	115	110	140	142	100

(Answer: ρ = +0.690476)

10. A psychologist wanted to compare two methods- A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs, so that the students in a pair have approximately equal scores on an intelligent test. In each pair one student was taught by method A and other by method B and examined after the course. The marks obtained by them are tabulated as below;

Pair:	1	2	3	4	5	6	7	8	9	10	11
A:	24	29	19	14	30	19	27	30	20	28	11
B:	37	35	16	26	23	27	19	20	16	11	21

- Find the Karl Pearson's correlation coefficient between the two sets of scores.
- Find the Spearman's rank correlation coefficient.

(Answer: r = +0.3517, ρ = -0.02954)

Reference

1. Ganguli, B. N. (2010). *Statistics and Statistical Methods: In Two Volumes*. New Delhi: New Age International Publishers.
2. Gupta, S. C. (2014). *Fundamentals of Statistics*. Meerut: Rastogi Publications.
3. Kulkarni, P. D., & Deshpande, J. V. (2010). *Statistics:* Essentials and Applications. Pune: Sujata Prakashan.
4. Mehta, P., & Ramachandran, V. (2010). *Statistics for Management*. New Delhi: Vikas Publishing House.
5. Pandey, J. N., & Gupta, T. C. (2017). *Advanced Statistics*. Delhi: Wiley India Pvt. Ltd.
6. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Statistical power analysis for the behavioral sciences* (3rd ed.). Academic Press.
7. Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Sage Publications Ltd.
8. Yadav, V. K. (2016). *An Introduction to Probability and Statistics*. New Delhi: New Age International Publishers.
9. Hayes, A. F. (2023). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (5th ed.). Guilford Publications.
10. Howell, D. C. (2010). *Statistical methods for psychology (7th ed.)*. Wadsworth Cengage Learning.

Suggested Reading

1. Trivedi, A. K. (2012). *Practical Statistics with R.* Noida: McGraw-Hill Education.
2. Urmila, P. (2016). *Fundamentals of Biostatistics & Research Methodology*. Noida: McGraw-Hill Education.
3. Kline, R. B. (2015). *Principles and practice of structural equation modeling (4th ed.).* Guilford Publications.
4. Spearman, C. (1987). *The proof and measurement of association between two things by correlation coefficient*. British journal of psychology, 8(1), 1-28.
5. Kendall, M. G. (1975). *Rank correlation methods (4th ed.)*. Charles Griffin & Company.
6. Cohen, S. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associates.
7. DC Sancheti and VK Kapoor (1979). *Statistics (Theory methods and application)*. Sultan Chand & Sons, New Delhi

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 2

Regression

Learning Outcomes

After completing this unit, the learner will be able to:

- predict one variable based on another.
- gain a strong grasp of how different factors influence each other, enabling the analysis of data to identify patterns or trends.
- learn to construct mathematical models that estimate the relationship between a dependent variable (the variable you're trying to predict) and one or more independent variables (the factors that might influence it).
- encourage a more analytical and data-driven approach to thinking, helping you see the world through a lens of relationships and patterns.

Background

In the previous unit, you studied about correlation. That is the relationship between two variables in a bivariate distribution. In this way, you could understand the relationship between two variables, whether it is positively or negatively correlated or not correlated. However, using correlation, you cannot simply predict one variable using the other variables; you can only determine the direction of the relationship. For example, you can find the relationship between advertisement and sales of a product, but you cannot predict the sales for a given advertisement expense by using correlation. This is where regression is used. By using regression, you can simply predict one variable by using the other variables. This unit is developed to gain insight into the use and application of the statistical technique called regression.

Keywords

Regression, Dependent variable, Independent variable, Regression coefficients, Regression equations, Regression line, Standard error of estimate

Discussion

3.2.1 Regression

- Predict unknown from the known variable

If it is established that two variables are closely related, we can estimate the value of one variable given the value of another variable. For example, if we know that advertisement and sales are related, we can find out the expected amount of sales, for a given advertisement expenditure or the required amount of advertisement expenditure for a given amount of sales. The statistical tool with the help of which we are in a position to estimate or predict the unknown values of one variable from known values of another variable is called regression. Thus, with the help of regression analysis we are in a position to find out the average probable change in one variable given a certain amount of change in another.

Francis Galton

The term Regression was first used by Professor Francis Galton towards the end of the nineteenth century

3.2.1.1 Regression Lines

- X on Y and Y on X

If we take the case of two variables, X and Y, we shall have two regression lines, as the regression line of X on Y and the regression line of Y on X

Regression line of Y on X gives the most probable values of Y for the given values of X and the regression line of X on Y gives the most probable values of X for the given values of Y. Thus we have two regression equations;

3.2.1.2 Regression Equations

- Algebraic expression of the regression lines

Regression equations are algebraic expression of the regression lines. Since there are two regression lines, there are two regression equations.

The regression equation of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression Equation Y on X

The regression equation Y on X is expressed as follows;

$y = a + bx$

In this equation, "a" and "b" are unknown constants. These constants are called the parameters of the line. The values of "a" and "b" can be obtained by solving the following equations simultaneously;

- Predict Y using X

$\sum y = na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

These equations are usually called the normal equations

Regression Equation of X on Y

The regression equation of X on Y is expressed as follows;

$x = a + by$

To determine the values of "a" and "b", the following two normal equations are to be solved simultaneously;

- Predict X using Y

$\sum x = na + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

Computation of the regression equations are explained with the help of the following examples

Illustration 3.2.1

From the following data, obtain the two regression equations

X:	10	06	10	06	08
Y:	06	02	10	04	08

Solution

Computation is explained in the following table

X	Y	XY	X^2	Y^2
10	06	60	100	36
06	02	12	36	04
10	10	100	100	100
06	04	24	36	16
08	08	64	64	64
40	**30**	**260**	**336**	**220**

Regression equation Y on X is given by;

$y = a + bx$

To determine the value of constants "a" and "b", the following two normal equations are to be solved;

$\sum y = na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

Substituting the values in the equation, we get;

30 = 5a + 40b ----------Equation 1

260 = 40a + 336b---------Equation 2

Multiplying equation 1 by 8 we get;

240 = 40a +320b ---------Equation 3

260 = 40a +336b---------Equation 4

Subtracting equation (4) from (3), we get;

- 20 = -16 b (Cancelling the minus signs on both sides, we get)

$b = \frac{20}{16} = 1.25$ Now we have the value of 'b" as 1.25. This value of "b" can be substituted in equation (1), we get the value of "a". That is;

30 = 5a + 40 x 1.25

30 = 5a + 50

30 - 50 = 5a

$\frac{-20}{5} = a = -4$

Substituting the values of "a" and "b" in the regression equation, we get the regression line of Y on X, that is;

Y= – 4 + 1.25x

Now we can calculate the regression equation of X on Y, that is given by the equation;

x = a + by, and the two normal equations are;

$\sum x = na + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

Substituting the values in the equation, we get;

40 = 5a + 30b------- Equation 1

260 = 30a + 220b-----Equation 2

Multiplying equation (1) by 6, we get

240 = 30a + 180b-------Equation 3

260 = 30a + 220b------Equation 4

Subtracting Equation (4) from (3), we get

-20 = - 40b (Cancelling minus signs on both sides, we get)

$b = \frac{20}{40} = +0.5$

Substituting the value of "b" in equation (1), we get;

$40 = 5a + 30 \times 0.5$

$40 = 5a + 15$

$25 = 5a$

$a = \frac{25}{5} = 5$

Now we have to substitute the values of "a" and "b" in the regression equation to obtain the regression line of x on y, that is;

$X = 5 + 0.5y$

Deviations Taken from Arithmetic Means of X and Y Series

The calculations can be simplified if we take deviations from actual means of X and Y series, instead of actual values of X and Y

In such a case, the equation Y on X is written as;

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

Where $\bar{Y} = \frac{\sum Y}{N}$

$\bar{X} = \frac{\sum X}{N}$

The value of "b_{yx}" can be obtained as follows;

$b_{yx} = \frac{\sum xy}{\sum x^2}$

Similarly, regression equation X on Y can be obtained as follows;

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

$b_{xy} = \frac{\sum xy}{\sum y^2}$

Illustration 3.2.2

From the following data, obtain the regression equation of X and Y and Y on X

X:	10	06	10	06	08
Y:	06	02	10	04	08

Solution

Computation is explained in the following table

X	X - 8 = x	x^2	Y	Y - 6 = y	Y^2	xy
10	+02	04	06	00	00	00
06	- 02	04	02	- 04	16	08
10	+02	04	10	+04	16	08
06	- 02	04	04	- 02	04	04
08	00	00	08	+02	04	00
40	**00**	**16**	**30**	**00**	**40**	**20**

Here, deviations are taken from the actual mean

The arithmetic mean of X series is given by; $\bar{X} = \frac{\sum X}{N}$

Substituting the values in the equation, we get;

$$\bar{X} = \frac{40}{5} = 8$$

Similarly, arithmetic mean of Y series is given by the equation;

$$\bar{Y} = \frac{\sum Y}{N}$$

Substituting the values in the equation, we get;

$$\bar{Y} = \frac{30}{5} = 6$$

Regression equation X on Y is given by the equation

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

So, as the first step, we have to calculate the value of "b_{xy}" that is given by the equation;

$$b_{xy} = \frac{\sum xy}{\sum y2}$$

Substituting the values in the equation, we get

$$b_{xy} = \frac{20}{40} = 0.5$$

Now we can substitute the value of "bxy" in the equation, we get;

$(X - 8) = 0.5 (Y - 6)$

$X - 8 = 0.5Y - 3$

$X = 5 + 0.5Y$

Similarly, regression equation Y on X is given by the formula

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

The value of "b_{yx}" can be obtained by the formula;

$$b_{yx} = \frac{\sum xy}{\sum x^2}$$

Substituting the values in the equation, we get

$$b_{yx} = \frac{20}{16} = 1.25$$

Now, we can substitute the value of "b_{yx}" in the equation;

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

$Y - 6 = 1.25 (X - 8)$

$Y - 6 = 1.25X - 10$

$Y = 1.25X - 4$

Deviations Taken from Assumed Means of X and Y Series

If the actual means of X and Y variables are in fractions, then the calculations will be difficult. Therefore, in such cases calculations can be simplified by taking deviations from actual means of X and Y series.

Regression equation X on Y is given by the formula;

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

"b_{xy}" is calculated using the equation;

$$b_{xy} = \frac{N \sum dxdy - (\sum dx)(\sum dy)}{N \sum dy^2 - (\sum dy)^2}$$

Similarly, the regression equation of Y on X is calculated using the equation;

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

b_{yx} is calculated using the equation;

$$b_{yx} = \frac{N \sum dxdy - (\sum dx)(\sum dy)}{N \sum dx^2 - (\sum dx)^2}$$

Illustration 3.2.3

From the following data, obtain the two regression equations taking deviations from 7 in the case of X series and 5 in the case of Y.

X:	10	06	10	06	08
Y:	06	02	10	04	08

Solution

Computation is explained in the following table

X	X -7 = dx	$(dx)^2$	Y	Y - 5 = dy	$(dy)^2$	dxdy
10	+03	09	06	+01	01	03
06	- 01	01	02	- 03	09	03
10	+03	09	10	+05	25	15
06	- 01	01	04	- 01	01	01
08	+01	01	08	+03	09	03
40	**+05**	**21**	**30**	**+05**	**45**	**25**

Regression equation X on Y is given by the equation;

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

Important Point

Here we have to remember that in this equation, $\bar{X}$ and $\bar{Y}$ denotes the actual mean of the series and not assumed means from which deviations are taken.

Constant "b_{xy}" is calculated using the equation;

$$b_{xy} = \frac{N \sum dxdy - (\sum dx)(\sum dy)}{N \sum dy^2 - (\sum dy)^2}$$

Substituting the vales in the equation, we get;

$$b_{xy} = \frac{5x25 - 5x5}{5x45 - 5^2}$$

$$= \frac{100}{200} = 0.5$$

Actual mean of X series $= \frac{\sum X}{N} = \frac{40}{5} = 8$

Actual mean of Y series $= \frac{\sum Y}{N} = \frac{30}{5} = 06$

The regression equation X on Y is calculated using the equation;

$(X - 08) = 0.5 (Y - 06)$

$X - 8 = 0.5Y - 3$

$X = 0.5Y + 5$

Let us now calculate the regression equation Y on X

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

'b_{yx}" is calculated using the equation;

$$b_{yx} = \frac{N\sum dxdy - (\sum dx)(\sum dy)}{N\sum dx^2 - (\sum dx)^2}$$

Substituting the values in the equation, we get;

$$b_{yx} = \frac{5 \times 25 - 5 \times 5}{5x\ 21 - (5)2}$$

$$= \frac{100}{80} = 1.25$$

Substituting the vales in the equation, we get;

Y-6 = 1.25(X-8)

Y-6 =1.25X-10

Y = 1.25X -4

3.2.1.3 Regression Coefficients

- Constant

The constant "b" in the regression equation is called regression coefficient. Since there are two regression equations, there are two regression coefficients. The two regression coefficients are regression coefficient of X on Y and regression coefficient of Y on X.

Regression Coefficient of X on Y

The regression coefficient of X on Y is represented by the symbol, b_{xy} or b_1. This regression coefficient measures the change in X, corresponding to a unit change in Y. The regression coefficient of X on Y is given by;

$$b_{xy} = r\frac{\sigma x}{\sigma y}$$

Where, r = Karl Pearson's Correlation Coefficient

σx = Standard deviation of X series

σy = Standard deviation of Y series

Regression Coefficient of Y on X

The regression coefficient of Y on X is represented by b_{yx} or b_2. This regression coefficient measures the change in Y variable corresponding to unit change in X variable. The value of b_{yx} is given by;

$$b_{yx} = r\frac{\sigma y}{\sigma x}$$

Where 'r" = Karl Pearson's correlation Coefficient

σy = Standard deviation of Y series

σx = Standard deviation of X series

Calculating Correlation Coefficients from Regression Coefficients

We know that; $b_{xy} = r \frac{\sigma x}{\sigma y}$

$$b_{yx} = r \frac{\sigma y}{\sigma x}$$

Therefore $(b_{xy})(b_{yx}) = r \frac{\sigma x}{\sigma y} \times r \frac{\sigma y}{\sigma x}$

(Cancelling the common items, we get)

$(b_{xy}) \times (b_{yx}) = r^2$

Thus, correlation coefficient can be calculated using the equation;

$r = \sqrt{b_1 \times b_2}$ or $\sqrt{b_{xy} \times b_{yx}}$

- The product of the coefficient cannot be greater than one

Since the value of the correlation coefficient cannot exceed one, one of the regression coefficients must be less than one. In other words, both the regression coefficients cannot be greater than one. Similarly, both the regression coefficients will have the same sign, that is they will be either positive or negative.

Illustration 3.2.4

Find the regression coefficients of X on Y and Y on X from the following details;

X	03	02	-01	06	04	-02	05
Y	05	13	12	-01	02	20	00

Solution

Here 2 is the assumed mean taken for X series and 7 is the assumed mean taken for Y series.

Computation is explained in the given table:

X	X - 2 = dx	$(dx)^2$	Y	Y - 7 = dy	$(dy)^2$	dxdy
03	01	01	05	-02	04	-02
02	00	00	13	06	36	00
-01	-03	09	12	05	25	-15
06	04	16	-01	-08	64	-32
04	02	04	02	-05	25	-10
-02	-04	16	20	13	169	-52
05	03	09	00	-07	49	-21
17	**03**	**55**	**51**	**02**	**372**	**-132**

b_{xy} is calculated using the equation;

$$b_{xy} = \frac{N \sum dxdy - (\sum dx)(\sum dy)}{N\sum dy^2 - (\sum dy)^2}$$

Substituting the values in the equation, we get;

$$b_{xy} = \frac{07 \text{ x } -132 - (03 \text{ x } 02)}{07 \text{ x } 372 - (02)^2}$$

$$= \frac{-924 - 06}{2{,}600}$$

$$= \frac{-930}{2600}$$

$$= -0.35769$$

b_{yx} is calculated using the equation;

$$b_{yx} = \frac{N \sum dxdy - (\sum dx)(\sum dy)}{N\sum dx^2 - (\sum dx)^2}$$

Substituting the values in the equation, we get;

$$b_{yx} = \frac{7 \text{ x } - 132 - (03 \text{ x } 02)}{07\text{x } 55 - 3^2}$$

$$= \frac{-930}{376}$$

$$= -2.473$$

Illustration 3.2.5

Given that the means of X and Y are 65 and 67, their standard deviations are 2.5 and 3.5 respectively and the coefficient of correlation between them is 0.8.

Write down the two regression lines.

Obtain the best estimate of X, when Y = 70

Solution

a. Regression equation Y on X is given by the equation;

$$(Y - \bar{Y}) = r \frac{\sigma y}{\sigma x} (X - \bar{X})$$

Substituting the values in the equation, we get;

$$(Y - 67) = 0.8 \frac{3.5}{2.5} (X - 65)$$

(Y-67) = 1.12 (X – 65)

(Y-67) =1.12 X – 72.8

Y = 1.12 X -5.8

Similarly, regression equation x on y is given by the equation;

$$(X - \bar{X}) = r \frac{\sigma x}{\sigma y} (Y - \bar{Y})$$

Substituting the values in the equation, we get;

$$(X - 65) = 0.8 \times \frac{2.5}{3.5} (Y - 67)$$

(X-65) = 0.571 (Y-67)

(X-65) = 0.571Y- 38.257

X= 0.571Y+26.743

b) The best estimate of X, when Y =70 can be obtained from regression equation X on Y, that is;

X =0.571Y+26.743

X= 0.571 x 70 + 26.743

= 66.713

Illustration 3.2.6

Following data are available in respect of sales and advertisement expenditure.

Particulars	**Sales (₹)**	**Advertisement Expenditure (₹)**
Mean	70,000	15,000
Standard Deviation	15,000	3,000

Coefficient of Correlation = +0.8

Find out;

a) Two regression equations

b) If the company desires to achieve the target sales of ₹1,00,000, then how much should be the advertisement budget?

Solution

To simplify the calculations, let us omit the zeros(that is, 000's). Let sales be denoted as X and advertisement expenditure as Y variables.

We are given $\bar{X}$ = 70, $\bar{Y}$ = 15, σx =15, σy = 03, r = +0.8

Regression equation X on Y is calculated using the equation;

$$(X - \bar{X}) = r \frac{\sigma x}{\sigma y} (Y - \bar{Y})$$

Substituting the values in the equation, we get;

$$(X-70) = 0.8 \times \frac{15}{03} (Y - 15)$$

$$(X - 70) = 4 (Y - 15)$$

$$X = 4Y - 60 + 70$$

$$X = 4Y + 10$$

Similarly, regression equation Y on X is calculated using the equation;

$$(Y - \bar{Y}) = r \frac{\sigma y}{\sigma x} (X - \bar{X})$$

Substituting the values in the equation, we get

$$(Y-15) = 0.8 \times \frac{03}{15} (X-70)$$

$$(Y-15) = 0.16 (X-70)$$

$$(Y-15) = 0.16X - 11.2$$

$$Y = 0.16X + 3.8$$

From the regression equation of Y on X, we can find out the likely advertisement budget for a sales target of ₹100.

The equation is

$$Y = 0.16 X + 3.8$$

$$Y = 0.16 \times 100 + 3.8 = ₹19.8$$

Thus, for attaining a sales figure of ₹1,00,000, the company should have an advertisement budget of ₹19,800.

Illustration 3.2.7

Following are the marks in Kannada and English in an annual examination

Particulars	**Kannada (X)**	**English (Y)**
Mean	40	50
Standard Deviation	10	16

Coefficient of correlation is +0.3. estimate the score of English, when score in Kannada is 50 and score in Kannada, when score in English is 30.

Solution

Here, we have to form both the equations, because we have to predict both the lines.

a) Regression equation X on Y is given by the formula;

$$(X - \bar{X}) = r \frac{\sigma x}{\sigma y} (Y - \bar{Y})$$

Substituting the values in the equation, we get;

$$(X - 40) = 0.3 \times \frac{10}{16} (Y - 50)$$

$$X - 40 = 0.1875 (Y - 50)$$

$$X - 40 = 0.1875Y - 9.375$$

$$X = 0.1875Y + 30.625$$

Now, we can predict the score in Kannada, when the score in English is 30 using the above regression equation;

$$X = 0.1875 \times 30 + 30.625$$

$$= 36.25 \text{ marks}$$

Therefore, the score in Kannada will be 36.25, when the score in English is 30.

b) Similarly, regression equation Y on X can be calculated using the formula;

$$(Y - \bar{Y}) = r \frac{\sigma y}{\sigma x} (X - \bar{X})$$

Substituting the values in the equation, we get

$$(Y - 50) = 0.3 \times \frac{16}{10} (X - 40)$$

$$(Y - 50) = 0.48X - 19.2$$

$$Y = 0.48 X + 30.8$$

Now, we can predict the score in English, when score in Kannada is 50, using the above equation;

$$Y = 0.48 \times 50 + 30.8$$

$$= 54.8 \text{ Marks}$$

Therefore, the score in English will be 54.8 marks when the score in Kannada is 50 marks.

3.2.2 Distinction Between Correlation and Regression

Points of Distinction	Correlation	Regression
Inter dependence between variables	Explains the inter-dependence between the variables	Does not explain the interdependence between the variables
Dependent and independent variables	There is no difference between dependent and independent variables	There is difference between dependent and independent variables
Computation objective	To establish in one value the degree and direction of relationship between the variables	To compute the value of one variable given the value of another variable which are established as related variables
Purpose of computation	To establish the relationship between the two variables	To estimate the value of one variable given the value of another variable
Mutually Dependent Variables	Yes	No. one variable is dependent and another independent
Nature of coefficient	Independent of change of origin and scale	Dependent on change of scale but independent on origin

3.2.3 Standard error of the estimate

The standard error of the estimate is a way to measure the accuracy of the predictions made by a regression model.

Estimated standard deviation often denoted σ_{est}, it is calculated as:

$$\sigma_{est} = \frac{\sqrt{\Sigma(y - \hat{y})2}}{n}$$

where:

y: The observed value

ŷ: The predicted value

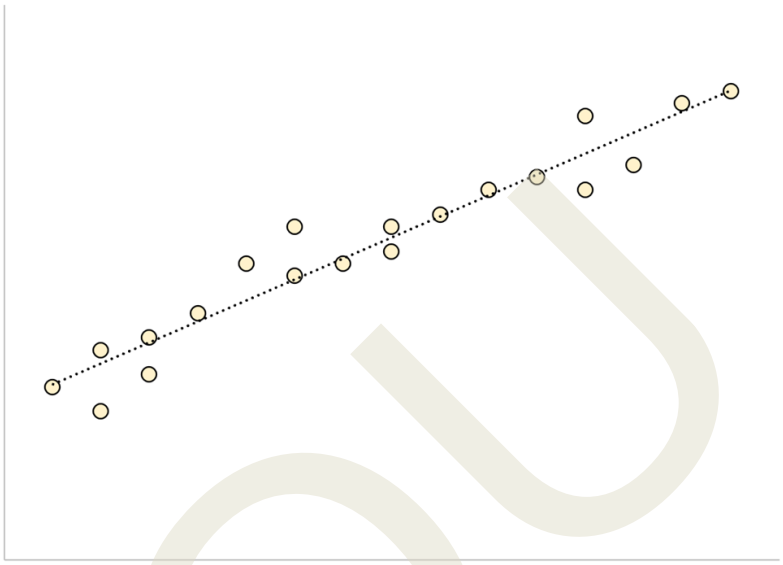
n: The total number of observations

- Measure the accuracy of the predictions

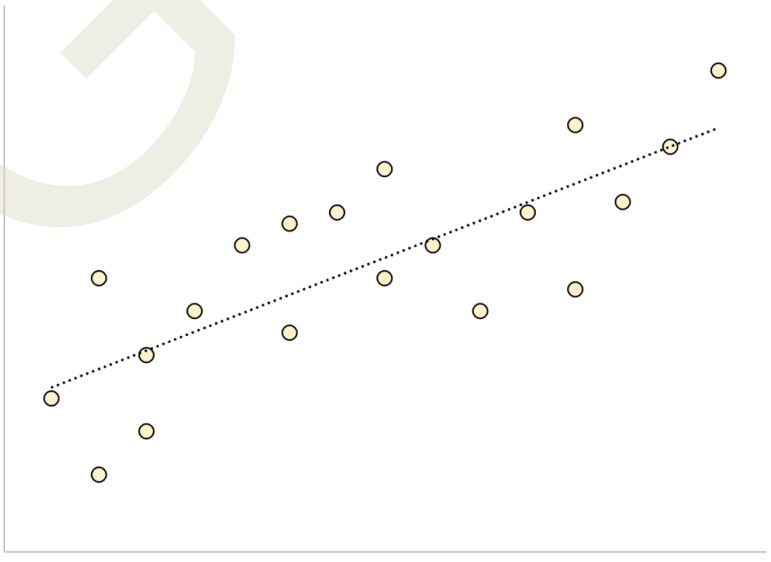
The standard error of the estimate gives us an idea of how well a regression model fits a dataset. In particular:

- The smaller the value, the better the fit.
- The larger the value, the worse the fit.

For a regression model that has a small standard error of the estimate, the data points will be closely packed around the estimated regression line:



Conversely, for a regression model that has a large standard error of the estimate, the data points will be more loosely scattered around the regression line:



Summarised Overview

Regression analysis is a statistical method employed to establish a functional relationship between two variables under study. It aims to discern the average association between a dependent variable and one or more independent variables, providing insights into the nature of their connection. The process involves determining regression coefficients, such as b_{yx} and b_{xy}, which serve as absolute measures representing the change in the value of the dependent variable for a unit change in the independent variable. Unlike correlation, regression analysis goes beyond measuring the direction and degree of linear relationships; it also examines the cause-and-effect dynamics, designating the variable corresponding to the cause as the independent variable and the effect as the dependent variable. This method facilitates the prediction or estimation of dependent variable values based on known independent variable values, offering practical applications in various fields by accommodating both linear and non-linear relationships between variables.

Self-Assessment Question

1. Given the bivariate data, use the actual mean to take deviations from the items and fit the two regression equations;

X:	01	05	03	02	01	02	07	03
Y:	06	01	00	00	01	02	01	05

(*Answers:* $b_{yx} = -0.33,\ Y = 2.99 - 0.33X,\ b_{xy} = -0.277,\ X = 3.554 - 0.277Y$)

2. Given the bivariate data;

X:	02	06	04	03	02	03	08	04
Y:	07	02	01	01	02	03	02	06

Obtain regression equations, taking deviations from 5 in the case of X and 4 in the case of Y series?

(*Answers:* $b_{xy} = -0.2777,\ X = 4.831 - 0.277Y,\ b_{yx} = -0.333,\ Y = 4.332 - 0.333X$)

3. Obtain the value of the correlation coefficient through the method of regression analysis from the data given below, first by taking deviations from the actual means of X and Y series and secondly from assumed means 20 and 18 for X and Y series respectively.

X:	10	20	30	40	50
Y:	10	20	15	25	30

{Answers: (Actual Mean Method = b_{xy} *= 1.8, X = 1.8Y – 6,* b_{yx} *= 0.45, Y =0.45+6.5 Correlation coefficient =+0.9) Computation of Regression Coefficients Taking Deviations from Assumed Means =* b_{xy} *=1.8,* b_{yx} *= 0.45, r = +0.9)}*

4. Given the bivariate data

X:	01	05	03	02	01	01	07	03
Y:	06	01	00	00	01	02	01	05

- Fit the regression line of Y on X and hence predict Y, if X =5
- Fit the regression line of X on Y and hence predict X, if Y = 2.5
- Calculate Karl Pearson's Correlation Coefficient?

(Answer: b_{xy} *= - 0.2777, X =3.431-0.278Y, If Y =2.5, X = 2.736,* b_{yx} *= - 0.304Y = 2.874 - 0.304X , If X =5, y =1.354, r = -0.2907)*

5. You are given the following information about advertisement and sales;

Particulars	**Advertisement Expenses (X) is in Crores of ₹**	**Sales (Y) is in Crores of ₹**
Mean	20	120
Standard Deviation	05	25

Coefficient of correlation is +0.8.

a. Calculate the two regression equations

b. Find out the likely sales, when advertisement expenditure is ₹25 crores

c. What should be the advertisement budget, if the company wants to attain sales target of ₹ 150 crores

(Answers: X = 0.16Y+0.8 X=24.8 Crores to attain a sales target of ₹150 crores, Y= 4X+40 Sales will be ₹140 crores, for an advertisement budget of ₹25 crores.)

6. Karl Pearson's coefficient of correlation between the age of brothers and sisters in a community was found to be 0.8. Average age of the brothers was 25 and that of sisters is 22 years. Their standard deviations were 4 and 5 respectively. Find;

a)The expected age of brother when sister's age is 12 years

b)The expected age of sister when brother's age is 33 years.

(Answers: Brother's age is X variable and sister's age is Y variable, X= 0.64Y+10.92, *If the age of sister is 12 years, brother's age will be 18.6 years, Y= X-3, The age of sister will be 30 years, when the brother's age is 33 years.)*

7. You are given the following information about different media of advertisement and sales;

Particulars	**Different Media of Advertisement (Rs in Lakhs)**	**Sales (Rs. In Lakhs)**
Mean	10	90
Standard Deviation	03	12

Correlation coefficient = -0.8

a) Calculate two regression lines?

b) Find the likely turnover, when advertisement expenditure is ₹20 lakhs.?

c) What should be the advertisement expenditure, if the company wants to attain a sales target of ₹140 lakhs?

(Answers: Advertisement expenditure is the X variable and sales is the Y variable

X= -0.2Y+28, Advertisement expenditure = 0 for a sales target of ₹140 Lakhs

Y= -3.2X+122, Y =₹58 lakhs, that is sales will be ₹58 lakhs for an advertisement budget of ₹20 lakhs.)

8. You are given the following information about advertisement and sales of accompany

Particulars	**Advertisement Expenses (in ₹)**	**Sales (in ₹)**
Mean	10	90
Variance	09	16

Correlation coefficient = +0.8

a) Calculate the two regression coefficients

b) Find the likely sales, when advertising expenditure is ₹15 lakhs, and

c) What should be the advertisement expenses, if the company wants to attain a sales target of ₹190 Lakhs?

(Answer: σx = 3, σy =4, X=0.6Y-44, The advertisement expenses for a sales target of ₹190 lakhs will be ₹70 lakhs, Y = 1.0667X+79.333, The likely sales, when advertisement expenses is ₹15 lakhs is; ₹95.3335 Lakhs)

9. Given the following estimate;

Particulars	**X**	**Y**
Mean	18	100
Standard Deviation	14	20

a) The value of Y, when X = 70

b) The value of X, when Y= 90

c) If two regression coefficients are 0.8 and 0.6, what would be the value of the coefficient of correlation?

(Answers = X= 0.56Y-38, The value of X, when the value of Y=90=12.4, Y= 1.1429X+79.4286. The value of Y, when X= 70 is 159.4316, Regression coefficient "r" = $\sqrt{0.8x\ 0.6}$ *= +0.6928)*

10. The following data relate to the age of husbands and wives. Obtain the two regression equations and determine the most likely age of husband when the age of wife is 25 years.

Age of Husband:	25	28	30	32	35	36	38	39	42	55
Age of Wives:	20	26	29	30	25	18	26	35	35	46

(Answers = Age of husband's is taken as the X and age of wives as y variables respectively. b_{yx} *=0.7623, Y=0.7623X+1.5572, Regression equation X on Y,* b_{xy} *= 0.8261, X= 0.8261Y+12.0431, Age of husband, when the age of wife is 25 years = 33 years)*

11. Obtain the two regression equations by taking deviations from actual means of X and Y series.

X:	02	04	06	08	10
Y:	05	07	09	08	11

(Answer: b_{xy}=1.3, X= 1.3Y-4.4, b_{yx} = 0.65, Y= 0.65X+4.1)

12. The following table shows the number of motor registrations in a certain territory, for a term of five years and the sale of motor tyres by a firm in that territory for the same period.

Year	Motor Registration	Number of Tyres Sold
01	600	1,250
02	630	1,100
03	720	1,300
04	750	1,350
05	800	1,500

Find the regression equation by the method of least squares to estimate the sale of tyres, when motor registration is known. Estimate the sale of tyres, when registration is 850.

(Answers: *Motor registration is denoted as the X variable and sales is denoted as the Y variable, Y= 255.04+1.4928X, When X= 850, Y= 1,523.92*)

13. In the following table, S is the weight of Potassium bromide which will dissolve in 100 gms of water at T^0 C. Fit an equation Y on X using the method of least squares and use this equation to predict S when T= 50^0 C.

T:	00	20	40	60	80
S:	54	65	75	85	96

(*Answers: Let "T" is represented as X and "S" represent Y, Y= 0.52X+54.2, The value of Y, when X= 50 is 80.2*)

14. On the basis of figures recorded below, for "supply" and "price" for nine years, build a regression of price on supply by the method of least squares. Calculate from the equation established, the most likely price when supply is 90.

Years	Supply	Price
2015	80	145
2016	82	140
2017	86	120
2018	91	124
2019	83	133
2020	85	127
2021	89	120
2022	96	110
2023	93	116

(*Answers: Let X denote supply and Y denote the price, Y = 296.072- 1.9486X, Y = ₹120.6978*)

15. Obtain both the regression equations from the following data, where the deviations are taken from mean

N = 20, $\bar{X}$ = 04, $\bar{Y}$ = 02, $\sum x^2$ = 1,680, $\sum y^2$ = 320, $\sum xy$ = 480

(*Answers, X = 1.5Y+1, Y= 0.2857142 X + 0.8571432*)

16. From the following data, obtain the two regression equations

Sales:	91	97	108	121	67	124	51	73	111	57
Purchases:	71	75	69	97	70	91	39	61	80	47

Also find correlation coefficient between sales and purchases?

(*Answers: X= 1.3598326Y-5.188282, Y = 0.6132075X+14.811325, r = +0.913159*)

17. Find the regression equation from the following and estimate the yield for rainfall of 10 inches.

Rain Fall (in inches):	01	02	03	04	05	06	07	08	09
Yield (Tons):	01	03	02	05	05	07	06	09	08

(*Answers: Deviations are taken from assumed means, that is from 5 for X and Y series respectively, X =0.9509433Y+0.1406798, Y= 9.33333X+0.443335, Y=9.77668 tons, when X = 10 inches*)

Assignments

1. To study the relationship between the expenditure on accommodation "X" and expenditure on food and entertainment "Y" an enquiry into 50 families gave the following results;

   $\sum X=8,500$, $\sum Y=9,600$, $\sigma x =60$, $\sigma y =20$ and $r = +0.6$

   Estimate the expenditure on food and entertainment, when expenditure on accommodation is ₹200?

   (Answers: We have to use regression equation Y on X, Y= 0.2X+158, $\bar{X}$ = ₹170, $\bar{Y}$ = ₹192. *The value of Y, when X =200 is ₹198)*

2. From the following data;

   a) Calculate two regression equations

   b) Estimate the value of X, when Y =20

   c) Determine the value of correlation coefficient through regression coefficients?

X	10	12	13	17	18
Y	05	06	07	09	13

(Answers = b_{xy} = 01, X = Y+06, The value of X , when Y =20 is 26, Regression line of Y on X, b_{yx} = 0.86957, Y =0.86957X-4.17398, r = +0.9325)

3. Obtain the regression equations from the following;

Particulars	X series	Y series
Mean	20	25
Variance	04	09

Correlation coefficient is +0.75

(*Answers* = $\sigma x = \sqrt{04} = 02$, $\sigma y = \sqrt{09} = 03$, *X = 0.5Y+7.05, Y =1.125X+2.05*)

4. The following data gives the demand (X) and supply (Y) in a market for 11 days.

   a. Obtain the regression equations of X on Y and Yon X

   b. Calculate the coefficient of correlation from the regression coefficients?

X	61	72	73	73	63	84	80	66	76	74	72
Y	40	52	59	53	61	58	56	42	58	50	50

(*Answers: Deviations are taken from assumed means, that is 72 and 52 for X and Y series respectively.* b_{xy} *=0.51812, X =0.51812Y+44.9062,* b_{yx} *=0.5169, Y=0.5169X+15.33, r = +0.5175)*

5. Given the following information;

   $\bar{X}$ = 65

   $\bar{Y}$= 67

   σx =25

   Variance of Y = 12.25

   r = +0.8

   Obtain

   The two regression equations?

   Predict the value of X, when Y =70 and Y when X= 58?

   (*Answers: X= 5.7143Y- 02, When Y=70, the value of X=398.001, Y= 0.112X+59.72, When X=58, the value of Y= 66.216*)

6. You are given the following data;

Particulars	**X**	**Y**
Mean	47	96
Variance	64	81

   Correlation coefficient between X and Y is 0.36

   a. Calculate Y, when X=50 and X when Y=88

   (*Answers: X =0.32Y+16.28, When Y=88, X =44.44, Y =0.405X+ 76.965, When X = 50, Y = 97.215*)

7. The following results for height and weight of 100 men were calculated as;

   Weight = Mean 150 lbs and Standard Deviation = 20lbs

Height = Mean = 68", Standard deviation= 2.5"

Coefficient of correlation= +0.6

Find an estimate of;

a) Weight of a man whose height is 5 feet, and

b) The height of a man whose weight is 200lbs.

(Answers: Let weight be the X variable and height the Y variable, Regression equation X on Y, X= 4.8Y-176.4, When Y=60 cms, (that is 05x12=60), X=111.6 lbs, Regression equation Y on X , Y=0.075X+56.75, When X = 200, Y= 71.75")

8. A study of wheat price at Mysore and Bangalore yields the following data;

Particulars	Mysore	Bangalore
Average price	2.463	2.797
Standard Deviation	0.326	0.207

r = +0.774

From the above data, estimate the most likely price of wheat

a. At Mysore corresponding to the price of ₹2.354 per kg at Bangalore, and

b. At Bangalore corresponding to the price of ₹3.05 per kg at Mysore?

(*Answers: Let the wheat price at Mysore be represented as X variable and at Bangalore as Y variable, X = 1.219Y-0.94642, Value of X, when Y=2.354 is ₹1.923106, Y= 0.4915X+1.5865 The value of Y, when X= ₹3.05 is ₹3.09625*)

9. The following table shows the exports of raw cotton and the imports of manufactured goods into India for seven years;

Exports (in Crores of ₹) :	42	44	58	55	89	98	60
Imports (in Crores of ₹) :	56	49	53	58	67	76	58

Obtain the two regression equations and estimate the imports when exports in particular year were to the value of ₹70 crores?

(Answers: X = 2.1983Y-67.2409, Y=0.3911X+34.6525, The value of Y, when X =70, X= ₹62.0295 crores)

10. Following data relate to years of service in a factory of seven persons in a specialized field and their monthly income;

Years of Service	11	07	09	05	08	06	10
Income(₹one hundred)	07	05	03	02	06	04	08

Find the two regression equations and also estimate the income of a person with twelve years of service?

(*Answers: Years of service is represented as X variable and income as Y variable. X= 0.75Y+4.25, Y= 0.75X-01, The value of Y, when X =12 is ₹800*)

11. Find the regression equations for the following data and also predict the average value of Y, when X is 09?

X	03	06	05	04	07	02	08	01
Y	03	02	03	05	03	06	06	04

(Answers: X=-0.42857Y+6.2148, Y= -0.07143X+4.3214, The value of Y, when X= 09=3.67853)

12. The following table gives the results of capital employed and profits earned by a firm in 10 successive years.

Particulars	Mean	Standard Deviation
Capital Employed (in Thousands)	₹55	₹28.7
Profit Earned (in Thousands)	₹13	₹8.5

Coefficient of correlation +0.96

a. Obtain the two regression equations

b. Estimate the amount of profit to be earned, if capital employed is ₹50,000?

c. Estimate the amount of capital to be employed, if profit earned is ₹20,000?

(*Answers: X= 3.2414Y+12.8616, When Y = ₹20,000, X= ₹77.6896, Y = 0.284X-2.6376, When X = ₹50,000, Y= ₹11.5774*)

13. A manufacturer of optical lenses has the following data on the cost per unit (in ₹) of a certain type of lenses and the number of units made in each order;

Order Number	**Number of Units(X)**	**Cost per Unit(Y)**
01	01	58

02	03	52
03	05	46
04	07	40
05	10	37
06	12	22

a. Determine the two regression equations?

b. Find the correlation coefficient?

(*Answers: X= -0.3214Y+19.9895, or X = 19.9895- 0.3214Y, Y= -2.9427X+61.1273, or Y= 61.1273-2.9427, R = -0.9725*)

14. A panel of two judges P and Q graded seven dramatic performances by independently awarding marks as follows;

Performances:	01	02	03	04	05	06	07
Marks by P:	46	42	44	40	43	41	45
Marks by Q:	40	38	36	35	39	37	41

The eighth performance which judge Q could not attend was awarded 37 by Judge P. If judge Q had also been present, how many marks would be expected to have been awarded by him for the eighth performance?

(Answers: X= 0.75Y+14.5, Y = 0.75X+5.75, The value of Y, when X=37 is 33.5 marks)

15. The following data is about the sales and advertisement expenditure of a firm; Coefficient of correlation (r) =+0.9. Estimate the likely sales for a proposed advertisement expenditure of ₹10 crores

Particulars	**Sales (in crores of ₹)**	**Advertisement Expenditure (in crores of ₹)**
Mean	40	06
Standard Deviation	10	1.5

(Answers: Let sales be denoted as X variable and advertisement expenditure as Y variable, X= 6Y+04, The value of X, when Y=10 is 64 crores)

Reference

1. Ganguli, B. N. (2010). *Statistics and Statistical Methods: In Two Volumes*. New Delhi: New Age International Publishers.
2. Gupta, S. C. (2014). *Fundamentals of Statistics*. Meerut: Rastogi Publications.
3. Kulkarni, P. D., & Deshpande, J. V. (2010). *Statistics:* Essentials and Applications. Pune: Sujata Prakashan.
4. Mehta, P., & Ramachandran, V. (2010). *Statistics for Management.* New Delhi: Vikas Publishing House.
5. Pandey, J. N., & Gupta, T. C. (2017). *Advanced Statistics*. Delhi: Wiley India Pvt. Ltd.
6. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Statistical power analysis for the behavioral sciences* (3rd ed.). Academic Press.
7. Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Sage Publications Ltd.
8. Yadav, V. K. (2016). *An Introduction to Probability and Statistics*. New Delhi: New Age International Publishers.
9. Hayes, A. F. (2023). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (5th ed.). Guilford Publications.
10. Howell, D. C. (2010). *Statistical methods for psychology (7th ed.)*. Wadsworth Cengage Learning.

Suggested Reading

1. Trivedi, A. K. (2012). *Practical Statistics with R.* Noida: McGraw-Hill Education.
2. Urmila, P. (2016). *Fundamentals of Biostatistics & Research Methodology*. Noida: McGraw-Hill Education.
3. Kline, R. B. (2015). *Principles and practice of structural equation modeling (4th ed.).* Guilford Publications.
4. Spearman, C. (1987). *The proof and measurement of association between two things by correlation coefficient*. British journal of psychology, 8(1), 1-28.
5. Kendall, M. G. (1975). *Rank correlation methods (4th ed.)*. Charles Griffin & Company.
6. Cohen, S. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associates.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

04 BLOCK

STATISTICAL QUALITY CONTROL

Block Content

Unit 1

Statistical Quality Control

Learning Outcomes

After completing this unit, the learner will be able to:

- identify variations and abnormalities in the production process.
- detect and address defects early in the production process.
- learn how to minimize the number of defective products, reducing waste and ultimately improving efficiency and cost-effectiveness in a manufacturing unit
- know how to reduce the costs associated with defects, rework, and waste.

Background

A food processing company that produces canned goods aims to ensure the highest level of product quality for its consumers. Recently, the company has faced challenges with variations in the sealing process of its canned products, leading to occasional defects such as leaks and compromised food safety. To address this issue, the company decides to implement Statistical Quality Control (SQC) methodologies.

By employing SQC techniques, the company begins systematically collecting and analyzing data on the sealing process. Statistical tools such as control charts are utilized to monitor the sealing parameters, identifying patterns and trends over time. This proactive approach allows the company to promptly detect deviations from the desired standards. As a result, the quality control team can take corrective actions before a significant number of defective cans are produced. Through the implementation of SQC, the food processing company not only improves the consistency and reliability of the sealing process but also minimizes the risk of producing substandard products that could harm its reputation and jeopardize consumer safety. SQC provides the company with a structured and data-driven methodology for identifying, analyzing, and addressing variations in the manufacturing process, ensuring the continued delivery of high-quality canned goods to the market.

In this unit, we are going to discuss Statistical Quality Control in detail how it works, why it is used, and what are the advantages and disadvantages of SQC.

Keywords

SQC, Random variations, Assignable variations

Discussion

4.1.1 Introduction

- Quality Assurance

Statistical Quality Control (S.Q.C.) is a crucial application in the industrial field, ensuring the quality of manufactured products. In modern industry, where repetitive work produces seemingly identical products, variations exist due to the inherent imprecision of manufacturing processes. To address this, assembly plants establish quality standards, defining acceptable tolerances within which products must fall to be considered satisfactory.

- Inherent Variations

Traditionally, a 100% inspection system was employed to assess each product's quality individually. However, this method has limitations, as human nature introduces the possibility of overlooking defects, being costly and time consuming. A more effective approach involves statistical quality control, which comprises two main aspects: (i) sampling inspection at various production stages and (ii) statistical inference using tools like charts to analyse quality variability. This system, developed by Walter A. Shewhart and expanded during World War II, aims to detect defects at their origin and replace 100% inspection with continuous sampling.

- Process and lot

Statistical quality control has two key facets: (i) process control, evaluating individual processes' performance to predict future quality variability, and (ii) product or lot control, ensuring that market-released lots do not contain an excessive number of unsatisfactory units. SQC relies on statistical techniques grounded in probability and sampling theory. Its application extends across diverse industries, including aircraft, armaments, automobiles, textiles, plastics, rubber, petroleum, electrical equipment, telephones, transportation, chemicals, and medicine.

- Industry-wide use

The fundamental challenge in any production process lies not in the quantity but in the quality of the product. Producers strive to meet prescribed standards and customer expectations throughout the production chain, from design and specifications to raw materials, machinery, manpower, and final product inspection. Statistical Quality Control plays a pivotal role in maintaining

control over these steps, ensuring that the end product aligns with customer standards.

- Customer Standards

Government bodies, such as the Indian Standards Institute (I.S.I.) in New Delhi, enforce quality standards by granting labels like ISI or Agmark to products meeting specified criteria. As customers become more quality-conscious, these labels signify adherence to stringent quality measures, emphasizing the importance of statistical quality control in today's industrial landscape.

4.1.2 Objectives of Statistical Quality Control (SQC)

Statistical Quality Control (SQC) utilizes statistics to monitor, analyse, and improve the quality of products and processes. Here are the key objectives:

i. Minimize Defects and Rework

Reduce the number of products that do not meet specifications, preventing the need for repairs or replacements. Example: A car manufacturer uses control charts to track paint defects on cars coming off the production line. When a chart shows an upward trend, indicating an increase in defects, they investigate and fix the issue (e.g., faulty equipment) before more cars are affected.

ii. Reduce Scrap and Waste

Minimize the amount of materials or resources wasted due to poor quality. Example: A food processing plant uses statistical sampling to inspect incoming ingredients for quality. This helps them avoid processing unusable materials, minimizing waste and saving costs

iii. Ensure Product Consistency

Guarantee that all products within a batch or group meet defined specifications and customer expectations.

Example: A clothing manufacturer uses control charts to monitor the size and colour consistency of garments. This ensures that all clothes within a batch meet customer specifications and expectations.

iv. Improve Process Efficiency

Optimize processes to increase productivity and reduce wasted time or resources. Example: A call center uses SQC tools to analyze call handling times. This helps them identify bottlenecks and optimize agent schedules to improve efficiency and reduce wait time.

v. Enhance Customer Satisfaction

Deliver products and services that consistently meet or exceed customer expectations.

Example: An online retailer uses data analytics to track customer complaints and product returns. This helps them identify and address common issues, leading to higher customer satisfaction.

vi. Make Data-Driven Decisions

Base decisions on objective data analysis instead of subjective opinions or hunches.

Example: A hospital uses SQC to analyze patient outcomes after different surgical procedures. This helps them identify the most effective procedures and improve patient care quality.

vii. Identify Root Causes of Problems

Find the underlying reasons behind quality issues instead of treating only the symptoms.

Example: A software company uses control charts to track bug reports in new releases. This helps them identify specific areas causing issues and prioritize fixes for efficient problem resolution

viii. Continuously Improve Processes

Regularly seek and implement changes to improve the quality, efficiency, and effectiveness of processes.

Example: A manufacturing plant uses statistical experiments to test different production methods. This helps them identify the most efficient and cost-effective methods for continuous process improvement.

ix. Reduce Costs Associated with Poor Quality

Minimize financial losses resulting from defects, rework, scrap, and other quality-related issues.

4.1.3 Types of variations

In Statistical Quality Control (SQC), understanding and managing variation is crucial for ensuring consistent product quality. This variation can be attributed to two main types of causes: chance causes and assignable causes.

i. Chance causes

These are inherent to any process and occur randomly, even when everything is operating under control. They are small, un-

predictable variations that cannot be readily identified or eliminated individually.

Examples:

- Minor differences in raw materials
- Slight variations in machine settings
- Random fluctuations in temperature or humidity
- Human error within normal bounds

Although individual chance causes are small, their combined effect can lead to noticeable variation in the overall quality of the product.

- Cannot be eliminated

Chance causes cannot be eliminated entirely, but their impact can be minimized by using statistical methods like control charts to establish control limits and identify significant deviations.

ii. Assignable Causes

These are specific, identifiable events or factors that cause abnormal variations in the quality of a product. They are not inherent to the process and can be traced back to a specific root cause.

Examples:

- Can be eliminated

- Equipment malfunction or miscalibration
- Incorrect material used
- Human error beyond normal bounds
- Change in supplier or process parameter

Assignable causes can lead to large and sustained shifts in product quality, potentially causing defects and customer dissatisfaction.

Understanding both types of variation allows SQC practitioners to:

- Differentiate between "normal" random variation and abnormal trends caused by assignable factors.
- Allocate resources effectively, focusing on eliminating assignable causes for significant quality improvement.
- Analyze control charts effectively to detect and react to process shifts or trends.
- Continuously optimize processes by eliminating assignable causes and ensuring the process operates within predictable control limits.
- Identifying and eliminating assignable causes is crucial

for improving process stability and product quality. This requires proactive investigation and corrective action based on control charts and other quality data.

Summarised Overview

Statistical Quality Control (SQC) is a set of statistical techniques and methods employed to monitor, control, and improve the quality of processes and products in business. Its primary objective is to ensure consistency and conformity to predefined standards, ultimately enhancing overall product or service quality. SQC utilizes statistical tools to analyze data collected during the production process, allowing businesses to make informed decisions and identify areas for improvement. By employing SQC, organizations can proactively manage and maintain quality standards, leading to reduced defects, increased efficiency, and higher customer satisfaction.

Self-Assessment Question

1. Define Statistical Quality Control (SQC) and explain its significance in the context of business quality management.
2. Explain the examples of industries or processes where SQC is commonly applied and discuss the benefits it brings to these settings.
3. Explore specific scenarios where SQC objectives can lead to improved decision-making and operational efficiency.
4. Enumerate and elaborate on the primary objectives of Statistical Quality Control. How do these objectives contribute to the overall improvement of processes and products within an organization?
5. Identify and discuss common causes of variation in a manufacturing or service process. How do these inherent variations impact the quality of output?
6. Explore the concept of assignable causes of variation. Provide examples of situations or events that can lead to assignable causes in a business process.
7. Distinguish between chance causes of variation and assignable causes of variation in the context of SQC. Provide real-world examples for each type and explain how recognizing these causes can aid in quality improvement efforts.
8. Discuss the challenges faced by organizations in implementing SQC and propose strategies for overcoming these challenges for effective quality control.

Assignments

1. Outline the primary objectives of Statistical Quality Control (SQC). Discuss how achieving these objectives can benefit an organization in terms of cost, customer satisfaction, and overall efficiency.
2. Explain 10 real-world examples illustrating the successful achievement of SQC objectives
3. Discuss specific examples to illustrate chance causes of variations and explain how businesses can distinguish and manage them using SQC techniques.
4. Discuss the implications of assignable causes on product or service quality and describe how SQC tools and methods can be employed to detect and eliminate them.
5. Examine the chances causes of variation in a manufacturing process. Discuss why it is essential for organizations to identify and understand these inherent variations and propose strategies to manage and control them effectively using SQC techniques.
6. Explore assignable causes of variation and their impact on product or service quality. Discuss how businesses can use SQC to detect, analyze, and eliminate these specific causes to enhance overall process efficiency and meet quality objectives.
7. Imagine you are a quality control manager in a manufacturing company. Develop a step-by-step plan for implementing Statistical Quality Control in your organization. Include key considerations, tools, and strategies to ensure the successful integration of SQC principles to enhance overall quality and customer satisfaction.

Reference

1. Ganguli, B. N. (2010). *Statistics and Statistical Methods: In Two Volumes*. New Delhi: New Age International Publishers.
2. Gupta, S. C. (2014). *Fundamentals of Statistics*. Meerut: Rastogi Publications.
3. Kulkarni, P. D., & Deshpande, J. V. (2010). *Statistics:* Essentials and Applications. Pune: Sujata Prakashan.
4. Mehta, P., & Ramachandran, V. (2010). *Statistics for Management*. New Delhi: Vikas Publishing House.
5. Pandey, J. N., & Gupta, T. C. (2017). *Advanced Statistics*. Delhi: Wiley India Pvt. Ltd.
6. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Statistical power analysis for the behavioral sciences* (3rd ed.). Academic Press.
7. Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Sage Publications Ltd.

8. Yadav, V. K. (2016). *An Introduction to Probability and Statistics*. New Delhi: New Age International Publishers.
9. Howell, D. C. (2010). *Statistical methods for psychology (7th ed.)*. Wadsworth Cengage Learning.

Suggested Reading

1. Trivedi, A. K. (2012). *Practical Statistics with R.* Noida: McGraw-Hill Education.
2. Urmila, P. (2016). *Fundamentals of Biostatistics & Research Methodology*. Noida: McGraw-Hill Education.
3. Kline, R. B. (2015). *Principles and practice of structural equation modeling (4th ed.).* Guilford Publications.
4. Spearman, C. (1987). *The proof and measurement of association between two things by correlation coefficient*. British journal of psychology, 8(1), 1-28.
5. Cohen, S. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associates.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

Unit 2

Quality Control Techniques

Learning Outcomes

After completing this unit, the learner will be able to:

- Familiarize with the concept and function of Statistical Process Control and its tools.
- Understand the concept of Product Control and its main tool of Acceptance Sampling.
- Identify the different types of control charts and their purposes.
- Interpret data using control charts and analyse the results.

Background

In the preceding unit, we explored variations in production, distinguishing between chance causes and assignable causes. These types of causes have the potential to adversely impact overall business operations. While fluctuations in the production process are inherent and cannot be entirely eliminated, it is crucial to distinguish between routine changes and variations that extend beyond the norm. To effectively manage and control such exceptional variations, we rely on control charts. In this unit, we will focus into a comprehensive discussion on control charts and their applications

Keywords

Process Control, Product Control, Acceptance Sampling, AQL, LQL, Control Chart, UCL, LCL, Centre Line, Central Limit Theorem, x-bar chart, R-bar chart and S.D. chart.

Discussion

4.2.1 Statistical Process Control

- Analysing variations in the manufacturing process

A process is a series of operations or actions that transforms input to output. It is said to be stable or repeatable if the resulting output product is of the given specifications or standard quality. Statistical Process Control (SPC) or simply process control (PC) is one of the three categories of Statistical Quality Control (SQC) aside from descriptive statistics and acceptance sampling. SPC is a statistical method of analysing variations in the manufacturing process in order to make it better and more effective. Only a certain number of samples are needed to determine whether the products are acceptable. It works by gathering important data from a specific sample size of a product being manufactured and utilizing statistics to determine the outcome of the process. Any data acquired from this can be used to develop and enhance the process.

- Early detection and correction

SPC is more production and manufacturer oriented. The aim of process control is to analyze the process study, time study, costing and ways of cost minimization. Assignable or special variations are noted and corrective actions may be taken. An advantage of SPC over other methods of quality control, such as "inspection", is that it emphasizes early detection and prevention of problems, rather than the correction of problems after they have occurred.

There are seven major tools used for process control. They are:

i. Histogram
ii. Check sheet
iii. Pareto chart
iv. Cause and effect diagram
v. Process flow diagram
vi. Scatter diagram
vii. Control chart

The application of SPC involves three phases:

i. Understanding the process and the control limits (UCL and LCL).
ii. Eliminating assignable (special) sources of variation.
iii. Monitoring the ongoing production process, using control

charts, to detect significant variation.

The tools of SPC are the normal distribution curve used in conjunction with 3 important parameters: -

- The mean or average of the values measured
- The range - difference between highest and lowest readings measured
- The standard deviation, which is derived by formula.

4.2.2 Product Control

- Detect defects and control the quality of a product

Product control means to control the products in such a way that these are free from defects and conform to their specifications. Product control mainly uses the technique of acceptance sampling to detect defects and control the quality of a product.

- Consumer oriented

This is applicable both in the cases of sourcing components from outside for manufacturing purposes and also in the case of final products. Its purpose is to ensure that the final product meets its specifications and various lots of the product do not contain an excessively large number of defective items. It is consumer oriented and does not concern itself with the production process or time.

- Kinds of inspection

Earlier, product control was done by 100% inspection, that is, each and every unit produced or received from the outside suppliers was inspected. It had the advantage of providing complete assurance of quality but had the problem of being costly and time consuming. It is also not possible in the cases of products such as light bulbs, crackers etc. which are destroyed under inspection. So, sampling inspection or acceptance sampling was developed as an alternative to this. Acceptance sampling is an inspection procedure used to determine whether to accept or reject a specific quantity of material.

- Harold Dodge

The concept and methodology of acceptance sampling were developed by Harold Dodge, a veteran of the Bell Laboratories quality assurance department, who was acting as a consultant to the Secretary of War for the United States in the Second World War. It was originally used by the U.S. military to the testing of bullets.

There are two major classifications of acceptance plans: by attributes and by variables.

This is the basic procedure for acceptance sampling:

i. A random sample is taken from a large quantity of items and tested or measured relative to the quality characteristic of interest.

ii. If the sample passes the test, the entire quantity of items is accepted.

iii. If the sample fails the test, either (a) the entire quantity of items is subjected to 100 percent inspection and all defective items repaired or replaced or (b) the entire quantity is returned to the supplier.

Uses of Acceptance Sampling

i. To determine the quality and acceptability of incoming raw materials, component parts, products etc.

ii. To decide the acceptability of semi-finished products for further processing.

iii. To determine the quality of outgoing products.

iv. For improving maintaining and controlling the quality of the products manufactured.

Producer's Risk and Consumer's Risk

Acceptance sampling involves both the producer (or supplier) of materials and the consumer (or buyer). Both parties face unique risks in accepting or rejecting lots as part of acceptance sampling.

- Type I error

Producer's Risk is the probability of wrongly rejecting a good lot, with level of non- conformance at or below the acceptance quality limit. Generally, the quality of an acceptable lot is expressed as the Acceptance Quality Limit. Producers risk corresponds to Type I error (probability of rejecting good lot) and is usually taken as 0.05 or 5%.

- Type II error

Consumers' Risk is the probability of wrongly accepting a lot that is not of acceptable quality. The corresponding lot quality is called the Limiting Quality (LQ or LQL).

Consumer's risk is the probability of accepting a lot with fraction defective p_t. and is denoted by β. i.e., β = P [accepting a lot of defective quality p_t]. It corresponds to Type II error (probability of accepting bad lot) and is usually taken as 0.10 or 10%.

4.2.3 Difference between Process Control and Product Control

Process control	Product control
Controlling of process sequence or steps to produce desired quality product is called Process control.	The control which is used to decrease defective items within different lots of produced good is known as Product control.

Process control is applied during production process.	Product control is applied after production process.
Process control used to reduce waste, minimize costs and reduce the environmental impact.	In Product control, the efforts are made to identify, sort, and segregate the defective product.
Control charts are the main tool used for process Control	Acceptance Sampling is the main tool used for product control

4.2.4 Control Charts

- Tool for process control

The control chart is the most important tool for process control. It was developed by Walter A. Shewhart of Bell Laboratories in 1924. It is used to detect the assignable causes of variation in order to take corrective action to eliminate them.

- On-line process-monitoring technique

A control chart is a graphic presentation of data over time. When the process is 'in control,' most of the production will fall between the lower and upper control limits (LCL and UCL). When the process is out of control, a larger proportion of the process lies outside of these limits. The control chart is an on-line process-monitoring technique. They may also be used to estimate the parameters of a process, and, hence to determine process capability. The main objective of statistical process control is the elimination of variability in the process, completely, or as much as possible. These charts help distinguish between natural variations and assignable variations.

Control charts are generally of two types:

a) Control charts for variables (Control charts for measurable characteristics) and

b) Control charts for attributes (Control charts for non-measurable characteristics).

- Controlling measurable and non-measurable characteristics

If the quality characteristic to be controlled is measurable such as weight, length, diameter, etc., we use the control chart for variables. If a characteristic to be controlled is non- measurable, e.g., color, surface roughness, etc., we use the control chart for attributes.

The technique for drawing a control chart is the same for both types. These are the main steps are as follows:

i. Select the quality characteristic

A single product/item/unit usually has several quality characteristics such as weight, length,width, strength, thickness, etc. It is difficult to construct a control chart for each characteristic.

So it is necessary to make a careful selection of quality characteristic. While selecting a quality characteristic, higher priority is given to the one that causes more non-conforming (defective) items and increases cost.

ii. Select the type of control chart

The choice of control chart depends upon the measurement, quality characteristic and the cost involved. If the characteristic to be controlled is measurable such as weight, length, diameter, etc., we use the control chart for variables. If a characteristic to be controlled is non- measurable, e.g., color, surface roughness, etc., we use the control chart for attributes.

iii. Selection of rational subgroups

Walter A. Shewhart developed and introduced the concept of rational subgroups for control charts. He suggested that the differences between rational subgroups are an indication of process changes (assignable causes) while differences within rational subgroup are an indication of inherent variability (natural causes). A rational subgroup is a small set of items/units that are produced under similar conditions within a relatively short time; which means, the variation within the subgroup is only due to chance causes. The subgroups are selected in such a way that we are able to differentiate between assignable causes and chance causes. If assignable causes are present in the process, they will present themselves as differences between the subgroups rather than differences within a subgroup.

iv. Size of subgroup (sample size)

To provide maximum homogeneity within a subgroup/sample, the size of the sample should be as small as possible. A sample size of four or five units is commonly used. We know from sampling distribution theory, the distribution of the sample mean $\bar{X}$ is nearly normal for samples of four or more, even though the samples are taken from a non-normal population.

When a sample of size 5 is used, there is ease in the computation of the average. When we have to make the control chart more sensitive, samples of size 10 or 20 are used.

This is because the standard error of a statistic is inversely proportion to sample size, i.e., as the sample size increases, the standard error decreases. Therefore, 3σ limits (upper and lower control limits) will lie closer to the center line. However, if the items produced are destroyed under inspection or are expensive, a small sample of size 2 or 3 is used.

v. Frequency of subgroups (number of samples)

To decide the number of samples, cost considerations must be balanced with the data obtained. There are two ways of taking samples: a) Taking larger samples at less frequent intervals, or b) Taking smaller samples at more frequent intervals. It is better to take samples more frequently in the beginning of the process and as the processes are brought into control, frequency of sampling can be reduced. If the process continuously remains under statistical control for a few weeks, the frequency may be reduced. If problems arise in keeping the process under statistical control, samples canbe taken more frequently.

vi. Design the forms for data collection

After taking the decision about the quality characteristic, subgroup (sample) size and number of subgroups, we design the data recording form. The form for data recordingshould be designed in accordance with the control chart to be used. A data recording form has two parts. The top part of the form contains information about part name, operation machine, gauge used, unit of measurement and specifications. The remaining segment of the form contains information about subgroup (sample) number,the date and time when the sample was selected and raw values of the observations. A column for comment is used to incorporate remark about the process.

vii. Determination of trial centre line and control limits

The quality characteristic usually follows a normal distribution or can be approximated by a normal distribution. The probability of a normally distributed random variable (X) that lies between $\mu - 3\sigma$ and $\mu + 3\sigma$ is 0.9973 where μ and σ are the mean and standard deviation of random variable X.

The probability that the random variable X lies outside the limits $\mu \pm 3\sigma$ is $1 - 0.9973 = 0.0027$, which is very small. It means that if we consider 100 samples, then most probably 0.27 items fall outside $\mu \pm 3\sigma$ limits. Therefore, if an observation falls outside 3σ limits, we can infer that something might have gone wrong. For this reason, the control limits are set up using 3σ limits. Suppose M is a sample statistic (e.g., mean, range, proportion of defectives, etc.) that measures some quality characteristic of interest. Further suppose that μM and σM are the mean and standard error (standard deviation) of the sample statistic M, respectively. Then the centre line and control limits for controlling the quality characteristic are given by:

Center Line (CL) = μM

Upper Control Limit (UCL) = $\mu M + 3\sigma M$

Lower control limit (LCL) = μM - 3σM

viii. Constructing a Control Chart

After calculating center line and control limits, we construct the control chart. In a control chart, the statistic (e.g., mean, range, number of defects, etc.) is taken along the Y-axis and the sample number or time is taken along the X-axis. We represent the center line by a solid line and the control limits by dotted lines. We plot the value of statistic for each sample against the sample number and consecutive sample points are joined by line segments.

ix. Drawing preliminary conclusions from the Control Chart

It is checked whether the plotted sample points lie on or in between the upper and lower control limits or some of them lie outside these control limits. If one or more points lie outside the control limits, we indicate each such point by drawing a dotted circle around it.

If all sample points lie on or in between the upper and lower control limits and there is no unnatural patterns of variation, the chart indicates that the process is under statistical control. When a control chart indicates that the process is under statistical control, periodic samples are taken from the process to determine whether it remains under statistical control continuously. When more data accumulates, the limits may be revised from time to time or whenever necessary. To ensure that the process remains under statistical control, periodic samples are taken once a week, once a month or once every 25, 50 or 100 items. If the control chart indicates that the process is not under statistical control it means some assignable causes are present in the process. We investigate the reasons of assignable causes and take corrective action to eliminate them from the process. Rectification is made by deleting the out-of-control points and calculating the revised centre line and control limits for the chart. These revised limits are known as the revised control limits. This procedure is continued till the process is being under statistical control.

4.2.5 Uses of control chart

Control charts are a very popular SQC tool because of its numerous benefits:

a. Control charts are a proven technique for improving productivity.

A control chart program will reduce scrap and rework, as a re-

sult productivity increases, cost decreases, and production capacity increases.

b. Control charts are effective in defect prevention.

The control chart helps keep the process in control, which is consistent with the "Do it right the first time" philosophy. It is costlier to sort out defective units from non-defective units later on than it is to build it right initially.

c. Control charts prevent unnecessary process adjustment

A control chart can distinguish between background noise and abnormal variation. A human operator is not as effective in making this distinction and may overreact to background noise and go on to make unnecessary adjustments. Such unnecessary adjustments can actually result in a deterioration of process performance.

d. Control charts provide diagnostic information.

The pattern of points on the control chart will contain information of diagnostic value to an experienced operator or engineer. This information allows the implementation of a change in the process that improves its performance.

e. Control charts provide information about process capability.

The control chart provides information about the value of important process parameters and their stability over time. This allows an estimate of process capability to be made which is of tremendous use to product and process designers.

4.2.6 Control Charts for Variables

- Common cause and special cause variation

Control charts for variables are a fundamental statistical tool used in quality management. These charts track continuous data measurements, such as dimensions, weight, or temperature, to monitor a process over time. They help identify variations within the process, separating common cause variation (inherent to the process) from special cause variation (caused by external factors).

- Mean chart, R chart and SD chart

The most common variable charts include X-bar charts (tracking process average), R charts (tracking process range), and S charts (tracking process standard deviation). These charts have a center line representing the process average, and upper and lower control limits calculated based on process data. By plotting sample measurements on the chart, analysts can detect shifts, trends, or unusual patterns, indicating process instability. Control charts are incredibly helpful tools to improve process predictability, reduce defects, and optimize overall quality.

Central Limit Theorem

The statistical foundation for $\bar{x}$ -charts is the central limit theorem. The central limit theorem says that the distribution of sample means will be approximately normally distributed.

In general terms, this theorem states that regardless of the distribution of the population of all parts or services, the distribution of x's (each of which is a mean of a sample drawn from the population) will tend to follow a normal curve as the sample size grows large. The theorem also states that (1) the mean of the distribution of the x's will equal the mean of the overall population and (2) the standard deviation of the sampling distribution, will be the population deviation, divided by the square root of the sample size, n.

- As sample size increases data tend to normal

4.2.6.1 Control Chart for Mean ($\bar{x}$ -chart)

Mean charts, also known as X-bar charts, are used to track the process average. They work by plotting the average of small samples taken from a process over time. A center line represents the overall process average, and upper and lower control limits are calculated. If data points fall outside these limits or patterns emerge, it signals that the process average may have shifted, indicating a potential issue.

- Tracks process average

Procedure for the construction

1. Compute the mean of each sample say $\bar{x}_1, \bar{x}_2 \ldots\ldots \bar{x}_k$ where k denotes the number of samples
2. Compute $\bar{\bar{X}} = \dfrac{\bar{x}_1 + \bar{x}_2 + \cdots \ldots + \bar{x}_k}{k}$
3. Choose the central line as $\bar{\bar{X}}$, CL = $\bar{\bar{X}}$
4. Fix the UCL and LCL using any of the appropriate formula

$$UCL_{\bar{x}1} = \bar{\bar{X}} + A_2 \bar{R}$$
$$LCL_{\bar{x}1} = \bar{\bar{X}} - A_2 \bar{R}$$

This is used when standards are not given. A_2 can be obtained from statistical table for control chart for different values of n from 2 to 25

5. After fixing LCL, UCL the sample means are plotted on the chart.

4.2.6.2 Range chart or R-chart

R charts, or Range charts, focus on process variability. They track the range of each sample, which is the difference between

- Monitors process variability

the highest and lowest values within that sample. Like the Mean chart, it has a center line and control limits. Changes in the range plotted on the R chart indicate fluctuations in the process's consistency, making them useful for spotting instability. R-chart is also drawn for measurable characteristics such as hardness, breadth, thickness etc. It shows variability within the process.

The R -chart is constructed using the following steps

1. Calculate the range R for each sample
2. Compute $\bar{R} = \frac{R_1 + R_2 + \cdots \ldots + R_k}{k}$, k being the sample number.
3. Set the control limits as follows

a. When standards are not given

$CL = \bar{R}$

$UCL_R = D_4 \bar{R}$

$LCL_R = D_3 \bar{R}$

The values D_3, D_4 are obtained from the statistical table.

4.2.6.3 Control chart for standard deviation (S.D Chart)

- Measure of variability

SD charts, or Standard Deviation charts, provide a more refined measure of process variability, particularly for larger sample sizes. They track the standard deviation of each sample. Standard deviation offers a more statistically sound representation of how spread out the data within each sample is. Significant changes in standard deviation signal potential problems with process consistency.

When the standard deviation σ is specified, then chart is constructed as follows:

Central Line = $d_2\sigma$

Upper Control Limit = $D_2\sigma$

Lower Control Limit = $D_1\sigma$

Where σ is the value of standard deviation D_1, D_2, d_2 values are obtained from tables depending on the sample size n. When σ is not known, it can be replaced by the appropriate estimate ($\hat{\sigma} = \frac{\bar{R}}{d_2}$)

Illustration 4.2.1

The following are the values of mean $\bar{X}$ and range R for 20

subgroups of 5 readings each taken from an inspection.

$\bar{X}$	R
1.85	.28
1.81	.14
1.75	.23
1.76	.35
1.83	.26
1.76	.25
1.71	.21
1.80	.08
1.77	.19
1.79	.39
1.82	.36
1.68	.10
1.69	.23
1.81	.29
1.78	.22
1.57	.05
1.72	.23
1.74	.23
1.55	.32
1.57	.47

Draw the $\bar{X}$ and R -charts with the warning and action limits and explain.

Solution

Mean of means $= \bar{\bar{x}} = \frac{\sum \bar{X}}{n}$

= [(1.85+1.81+1.75+1.76+1.83+1.76+1.71+1.8+1.77+1.79+ 1.82+1.68+1.69+1.81+1.78+1.57+1.72+1.74+1.55+1.57)] / 20

$$\bar{\bar{X}} = \frac{34.76}{20}$$

$$= 1.738$$

$$\bar{R} = \frac{\sum R}{20}$$

$$= \frac{4.88}{20}$$

$$= 0.244$$

From the control chart tables for a sample of size 5, we get the values of $A_2 = 0.577$; $D_3 = 0$ and $D_4 = 2.115$.

Control limit for $\bar{X}$-Chart

$$UCL = \bar{\bar{X}} + A_2 \bar{R}$$

$$= 1.738 + 0.577\ (0.244)$$

$$= 1.8788$$

$$= 1.88$$

$$LCL = \bar{\bar{X}} - A_2 \bar{R}$$

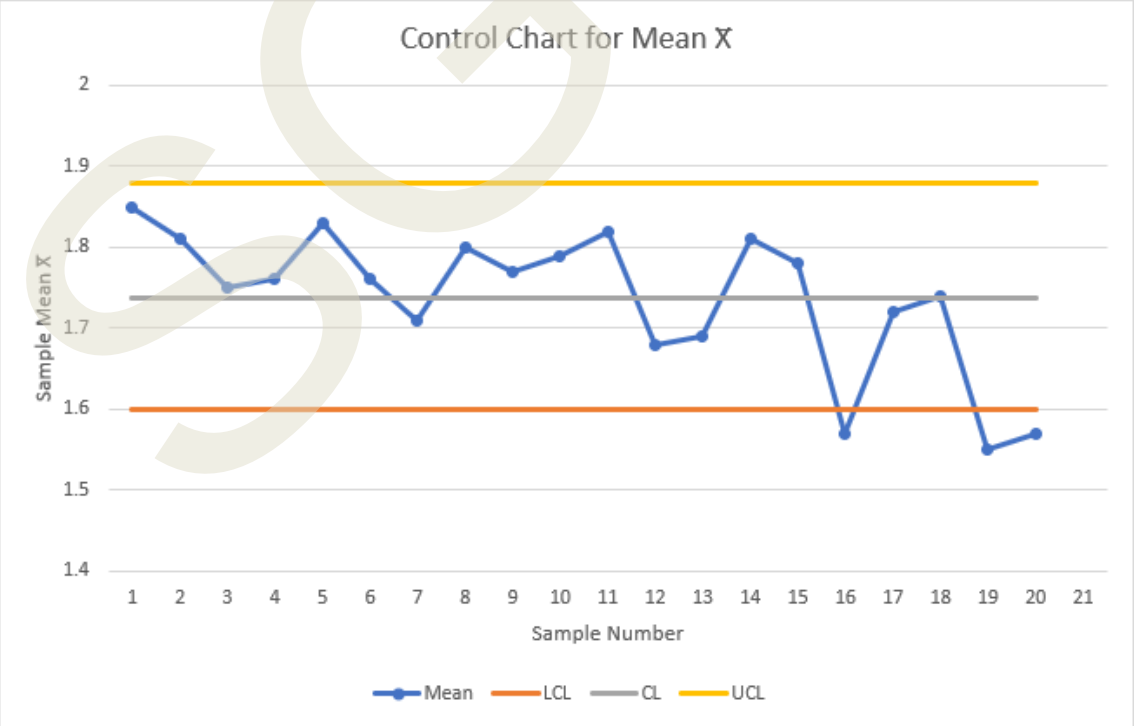
$$= 1.738 - 0.577\ (0.244)$$

$$= 1.5972$$

$$= 1.60$$

$$CL = \bar{\bar{X}} = 1.738$$

From the table given we observe that all the mean values except three values lie inside the control limits. These three values which lie outside the lower limits suggests that the process is out of control.



Control limits for R- Chart

$UCL = D_4 \bar{R}$

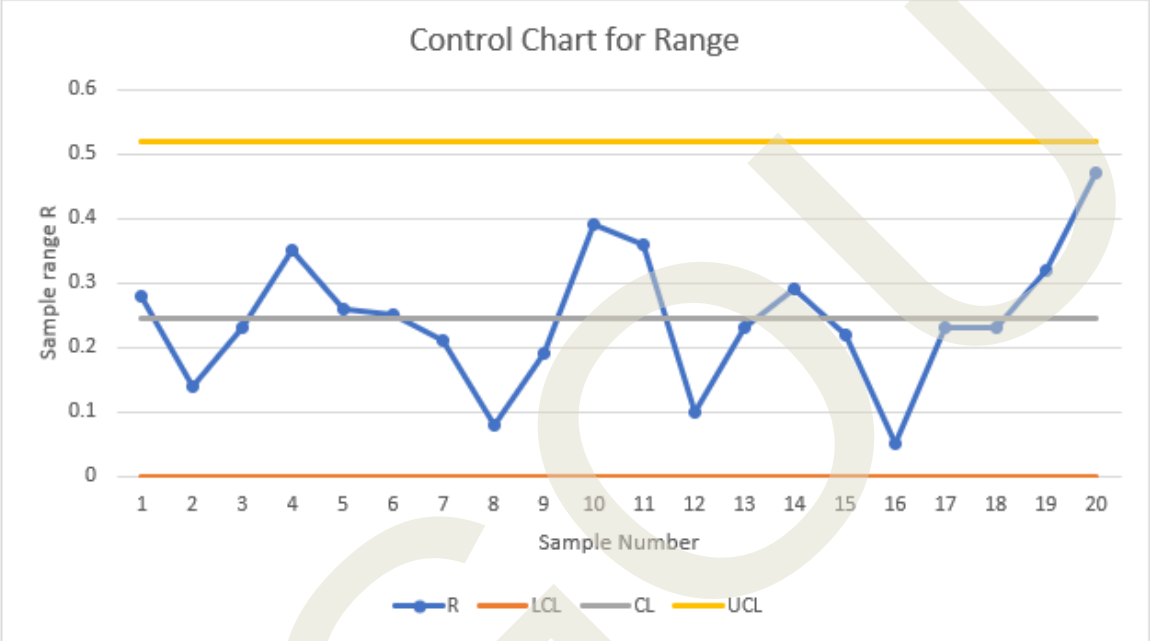
$= 2.114 \times 0.244$

$= 0.52$

$LCL = D_3 \bar{R}$

$= 0 \times 0.244$

$= 0$



From the figure, we find that all the values lie between the upper and lower control limits.

Hence the range (variation) is under control.

Illustration 4.2.2

Based on the following data draw mean chart and R-Chart. Interpret whether it is in control or out of control.

Sample No.	Sample Values				
	X_1	X_2	X_3	X_4	X_5
1	107	60	130	49	78
2	122	124	37	35	117
3	84	109	93	104	91

4	80	78	36	116	122
5	70	116	72	112	63
6	90	66	107	86	26
7	116	108	70	92	118
8	109	119	115	93	23
9	26	120	23	23	120
10	28	68	61	55	85

Solution

Sample No.	Sample Values					Mean	Range
	X_1	X_2	X_3	X_4	X_5	X	R
	107	60	130	49	78	84.8	81
	122	124	37	35	117	87	89
	84	109	93	104	91	96.2	25
	80	78	36	116	122	86.4	86
	70	116	72	112	63	86.6	53
	90	66	107	86	26	75	81
	116	108	70	92	118	100.8	48
	109	119	115	93	23	91.8	96
	26	120	23	23	120	62.4	97
	28	68	61	55	85	59.4	57
	Total					**830.4**	**713**

In case of sample no 1, $X = \frac{107 + 60 + 130 + 49 + 78}{5}$

= 84.8 and

R = Largest value -Smallest value = (L-S)

= 130 - 49

=81

$$\text{Mean} = \bar{\bar{X}} = \frac{830.4}{10} = 83.04$$

$$= \bar{R} = \frac{713}{10} = 71.3$$

From table, for n =5, $A_2 = 0.577$, $D_3 = 0$, $D_4 = 2.115$

Hence for $\bar{X}$ chart,

CL: $\bar{\bar{X}} = 83.04$

UCL $= \bar{\bar{X}} + A_2\bar{R}$

$= 83.04 + 0.577 \times 71.3$

$= 83.04 + 41.14$

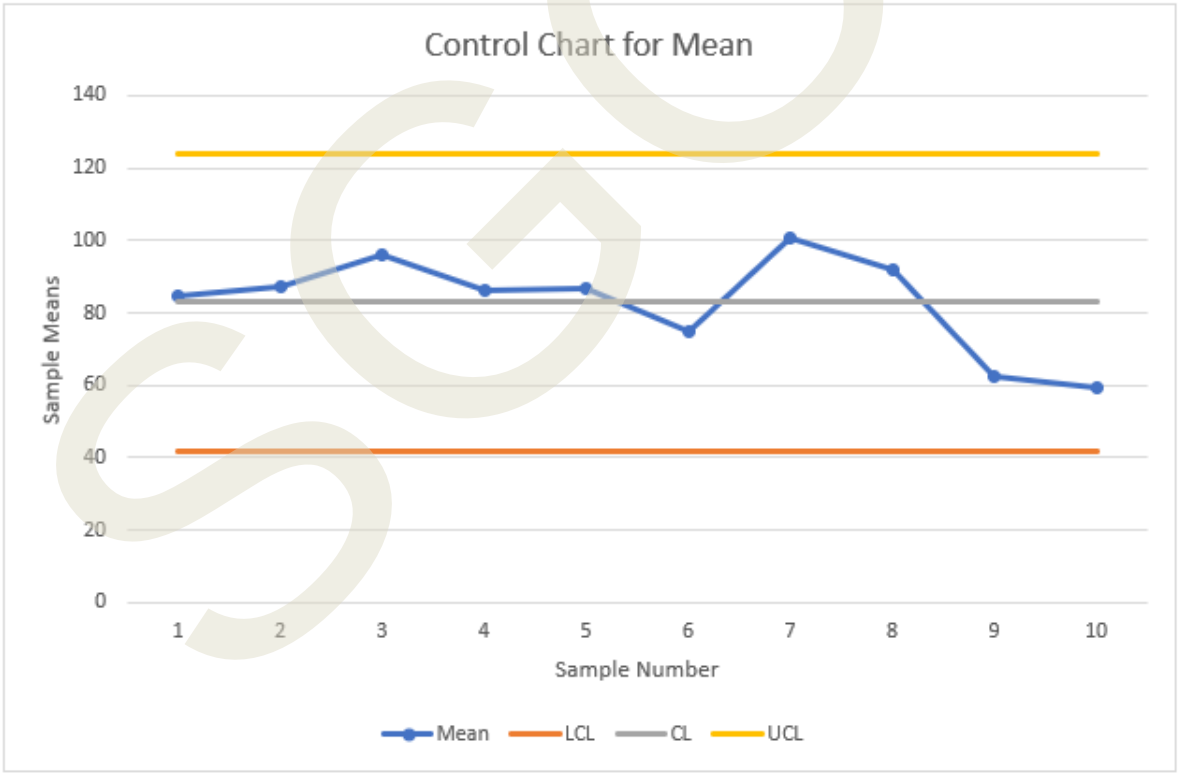
$= 124.18$

LCL $= \bar{\bar{X}} - A_2\bar{R}$

$= 83.04 - 0.577 \times 71.3$

$= 83.04 - 41.14$

$= 41.9$



All the values are inside of the control limits. It indicates that the production process is in control.

For R chart

$$CL = \bar{R} = 71.3$$

$$UCL = D_4\bar{R} = 2.114 \times 71.3$$

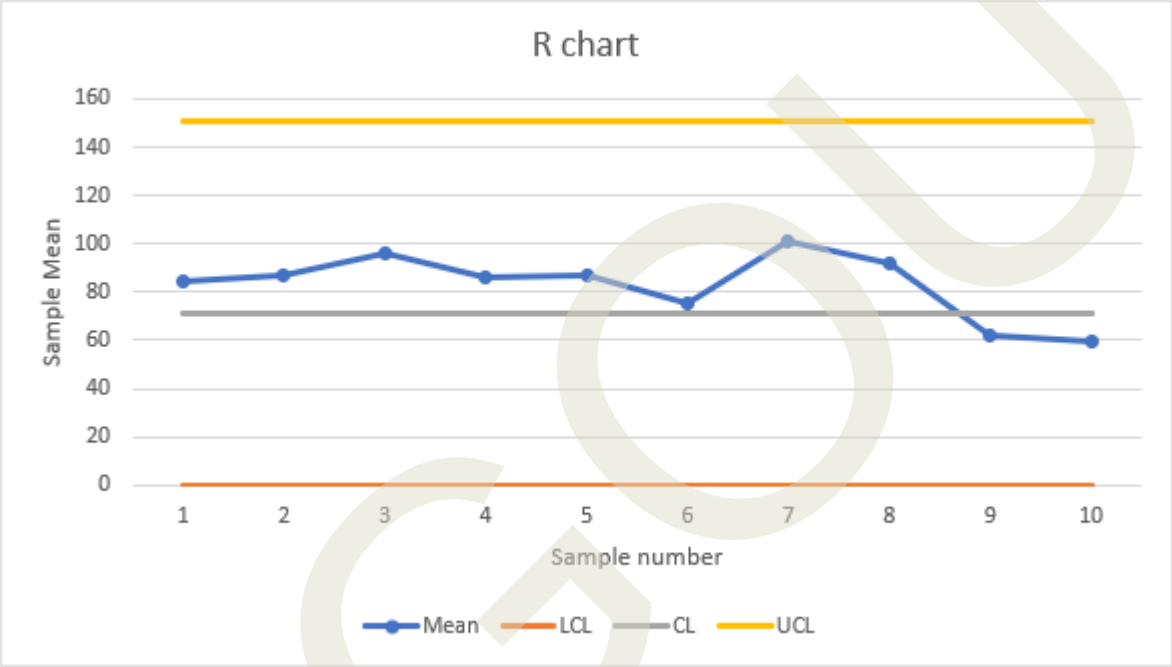
$$= 150.73$$

$$= 150.73$$

$$LCL = D_3\bar{R}$$

$$= 0 \times 71.3$$

$$= 0$$



From the figure, we find that all the values lie between the upper and lower control limits.

Hence the range (variation) is under control.

4.2.7 Control Charts for Attributes

- Categorical

Control charts for attributes are statistical process control (SPC) tools used to monitor processes where the quality characteristic of interest is categorical or "countable" rather than measurable. They help assess whether a process is in a state of statistical control and detect shifts or changes in quality levels over time.

These charts focus on defects or non-conformities. A defect is any characteristic of a product or service that fails to meet specifications or customer requirements. Examples include:

- Surface flaws on a product
- Missing components
- Incorrect paperwork
- Scratches on a component

4.2.7.1 Types of control chart for attributes

i. c-chart (Number of Defects Chart)

- Tracks defects per unit.

A c-chart is used to monitor the number of defects found within a single unit of a product or service. This means that a single item can have multiple flaws or problems. A c-chart assumes that the number of defects follows a Poisson distribution. For example, a c-chart could track the number of scratches on a car door, the number of typos in a document, or the number of missing items in a shipment.

Procedure for the construction of c-chart

Let c denotes the number of defects counted in one unit of cloth of paper or any material. Find the mean $\bar{c} = \frac{(c1 + c2 + \cdots \ldots cn)}{n}$

Where $c_1, c_2 \ldots\ldots\ldots, c_n$ are defects counted in several such units. The expected standard or central line is $\bar{c}$. in a poisson distribution, the variance is equal to mean ie., $\sigma^2 = \bar{c}$ or $\sigma = \sqrt{\bar{c}}$. Based on this 3 σ limits for the upper and lower control limits are obtained.

$$UCL = \bar{C} + 3\sigma = \bar{C} + 3\sqrt{\bar{c}}$$

$$LCL = \bar{C} - 3\sigma = \bar{C} - 3\sqrt{\bar{c}}$$

The c-chart is deemed valid and suitable when the potential for defects in each production unit is infinite, yet the probability of defects at any given point is consistently low. When employing the c-chart, uniform sample sizes or unit sizes are utilized.

ii. p-chart (Proportion Defective Chart)

A p-chart monitors the proportion of defective units within a sample. It focuses on whether the entire unit is considered defective or non-defective (pass or fail). P-charts are based on the binomial distribution. An example would be tracking the percentage of lightbulbs in a batch that are broken, or the percentage of online orders that contain errors.

Procedure for the construction of p-chart

a. Compute the fraction defective for each sample

$$\bar{P} = \frac{\text{Number of defective units in the sample}}{\text{Sample size or Total number of units inspected in the sample}}$$

b. Obtain the average fraction defective p from all the given samples

$$\bar{P} = \frac{\text{Sum of fraction of defectives in all the samples combined}}{\text{Total number of sample size inspected}}$$

c. The expected standard or central line CL = $\bar{p}$

d. The upper and lower control limits are given by

$$UCL_p = \bar{p} + 3\sigma_p$$

$$= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$LCL_p = \bar{p} - 3\sigma_p$$

$$= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

After setting the control limits and marking on the graph sheet we can plot the sample points. if all the points lie within the sample limits the process is under control, otherwise it is not.

iii. np-chart (Number of Defectives Chart)

- Tracks number of defectives in a sample

An np-chart is very similar to a p-chart, but instead of tracking proportions, it focuses on the actual number of defective units within a sample of fixed size. Like the p-chart, it also relies on the binomial distribution. An np-chart might be used to track the number of defective computer chips in a lot, or the number of patients admitted to a hospital with incorrect paperwork.

CL = $n\bar{p}$, Where n = sample size

$$\bar{p} = \frac{Total\ number\ of\ defectives\ of\ all\ samples}{Number\ of\ samples\ inspected}$$

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

We have to plot the values of the number of defectives d = nμ in the chart and find the nature of the process. Here d will never be negative. Hence whenever LCL is less than zero it is taken zero.

Illustration 4.2.3

In a textile manufacturing process, ten pieces of cloth from

various rolls, all with equal length, were inspected for defects. The recorded defect counts were 1, 3, 5, 0, 6, 0, 9, 4, 4, and 3. Create a control chart based on the defect counts, and assess whether the process is currently under statistical control.

Solution

1, 3, 5, 0, 6, 0, 9, 4, 4, 3 are denoted by C_i

Their mean ($\bar{c}$) = $\frac{35}{10}$

= 3.5

CL = $\bar{c}$ = 3.50

UCL = $\bar{C} + 3\sqrt{\bar{c}}$

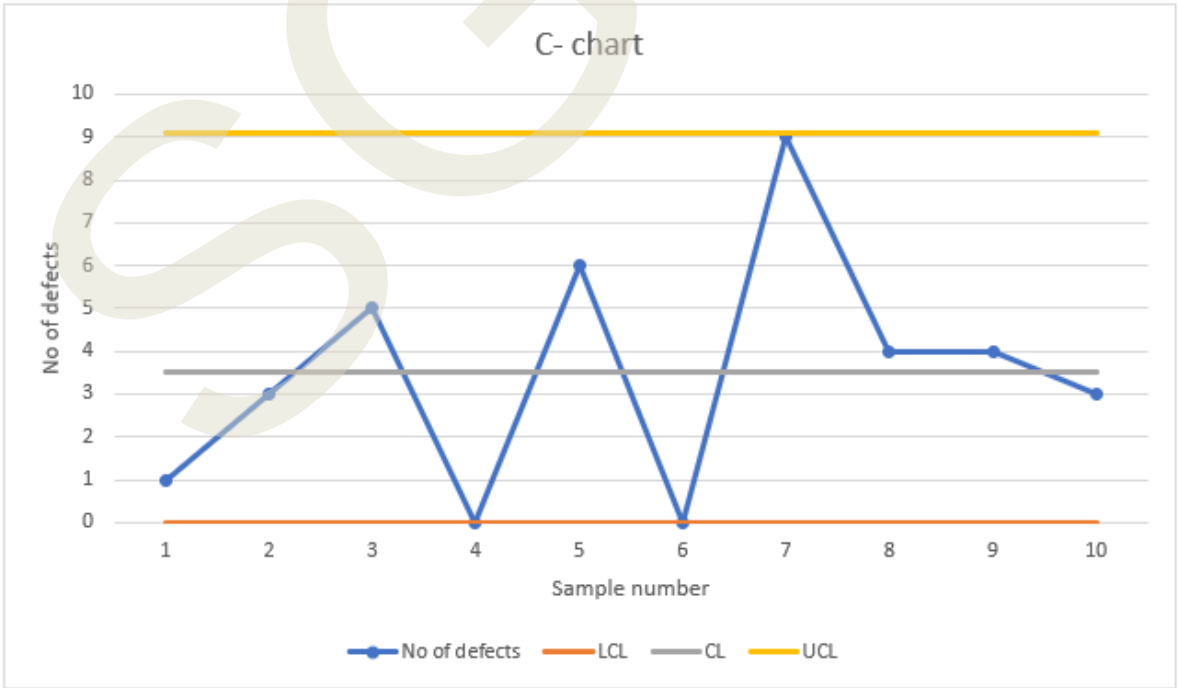
= $3.50 + 3\sqrt{3.50}$

= 9.11

LCL = $\bar{C} - 3\sqrt{\bar{c}}$

= $3.50 - 3\sqrt{3.50}$

= -2.11, is less than 0, it is taken as 0



All the points lie within the control limits in the above control chart for the number of defects. Hence, the process is in statistical control.

Illustration 4.2.4

Twelve sets of 200 bulbs each were inspected during a two-week production period. The number of defective bulbs in each set is recorded below.

Sample No:	1	2	3	4	5	6	7	8	9	10	11	12
No. of defective:	25	30	40	33	48	27	28	24	10	12	9	10

i. Draw the control chart for fraction defective.

ii. What do you find out from the chart?

Solution

$$\text{Fraction defective} = \frac{No\ of\ defective}{No\ of\ sample\ inspeced\ per\ set}$$

In the case of sample, no 1,

$$\text{Fraction defective} = \frac{25}{200} = 0.125$$

Sample No	No. of defective	Fraction defective
1	25	0.125
2	30	0.15
3	40	0.2
4	33	0.165
5	48	0.24
6	27	0.135
7	28	0.14
8	24	0.12
9	10	0.05
10	12	0.06
11	9	0.045

12	10	0.05
		1.48

$$CL = \bar{p} = \frac{\sum pi}{12}$$

$$= \frac{1.48}{12} = 0.123$$

$$UCL_p = \bar{p} + 3\sigma_p$$

$$= 0.123 + 3\sqrt{\frac{0.123 \; x \; 0.877}{200}}$$

$$= 0.123 + 3 \text{ x } 0.0232$$

$$= 0.123 + 0.0696$$

$$= 0.1926$$

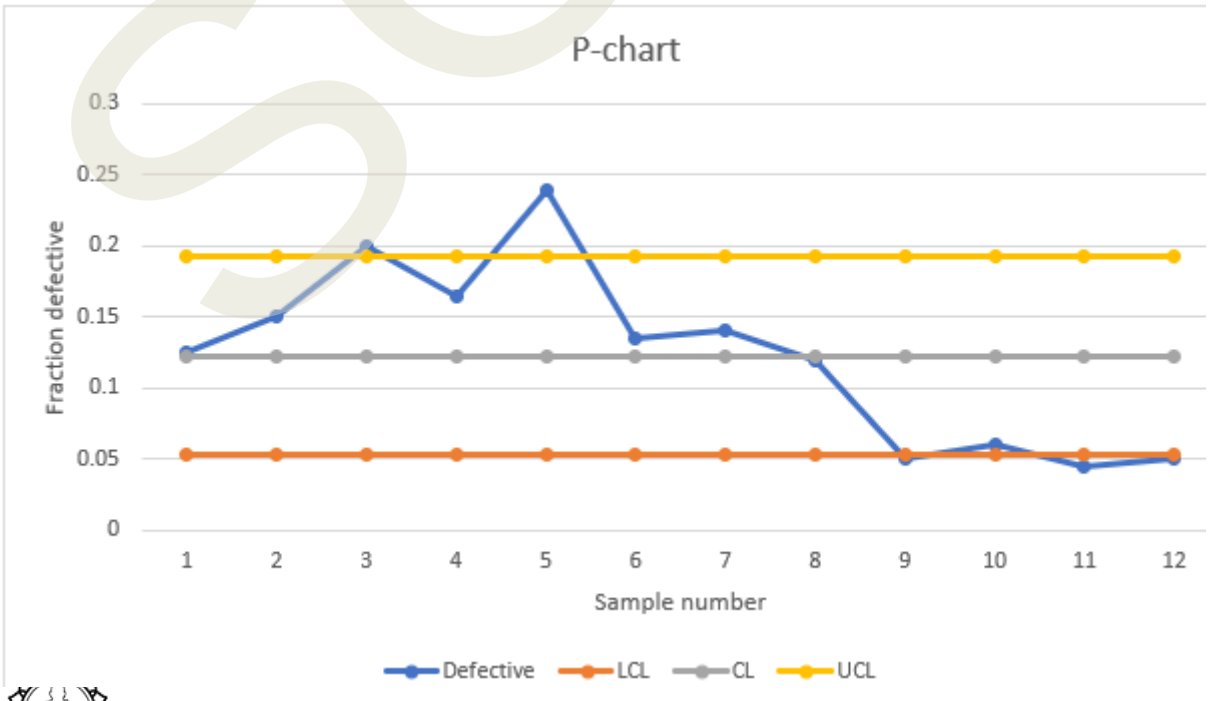
$$LCL_p = \bar{p} - 3\sigma_p$$

$$= 0.123 - 3\sqrt{\frac{0.123 \; x \; 0.877}{200}}$$

$$= 0.123 - 3 \text{ x } 0.0232$$

$$= 0.123 - 0.0696$$

$$= 0.0534$$



There are 5 sample sets found outside the control limits (samples 3, 5, 9, 11, and 12). Therefore, the production process of the bulb is not under statistical control.

Illustration 4.2.5

It was found that when a manufacturing process is under control, the average number of defectives per sample batch of 10 is 1.3. What limits would you set in a quality control chart based on the proportion of defectives in sample batches of 10?

Solution

Proportion of defective (p) = $\frac{1.3}{10} = 0.13$

Standard error of number of defectives = $\sqrt{n\bar{p}(1-\bar{p})}$

$= \sqrt{10x0.13(1-0.13)}$

$= \sqrt{10 \times 0.13 \times 0.87}$

$= 1.063$

Upper Control Limit = 1.3 + (3 x 1.063) = 4.489

Lower Control Limit = 1.3 - (3 x 1.063) = -1.889 that is 0 (number of defectives cannot be negative)

Illustration 4.2.6

The past records of a factory using quality control methods show that on the average 4 articles produced are defective out of a batch of 100. What is the maximum number of defective articles likely to be encountered in the batch of 100, when the production process is in a state of control.

Solution

N = Sample size =400

$\bar{P}$ = Process fraction defective = $\frac{4}{100} = 0.04$

1-p = 0.96

Let d be the number of defectives in a sample of size n. then d is a binomial variate with parameters n and p.

The 3 – σ Control limits for d (number of defectives) are given by

UCL = $n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$

$= 400 \times 0.04 + 3\sqrt{400 \times 0.04 \times 0.96}$

$= 16 + 3\sqrt{15.36}$

$= 16 + 3 \times 3.9192$

$= 27.75$

$LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$

$= 400 \times 0.04 - 3\sqrt{400 \times 0.04 \times 0.96}$

$= 16 - 3\sqrt{15.36}$

$= 16 - 3 \times 3.9192$

$= 4.2424$

Hence, if the production process in state of statistical control, the number of defective items to be encountered. In a batch of 400 should lie within the 3 – σ limits, Viz., (4.2424, 27.7576) ie (4, 28). Hence the maximum number of defectives items likely to be in this batch is 28)

Summarised Overview

Control charts are graphical tools used in statistical process control to monitor and maintain the stability and consistency of a process over time. These charts display the variation in a particular process by plotting data points against predetermined control limits. The central line on the chart represents the process mean, while the upper and lower control limits indicate the acceptable range of variation. Control charts are valuable for identifying trends, shifts, or anomalies in a process, helping organizations detect and address issues promptly. They are widely employed in industries such as manufacturing, healthcare, and service sectors to ensure quality control and optimize processes for efficiency and reliability.

The uses of control charts are diverse and extend across various industries. They play a crucial role in quality management, helping organizations detect deviations from the desired process performance and take corrective actions. By providing a visual representation of process variation, control charts enable businesses to distinguish between common cause and special cause variations, facilitating informed decision-making. There are different types of control charts designed for specific data characteristics. Commonly used types include X-bar charts for monitoring the central tendency of a process, R charts for monitoring the range of variation, and p charts or c charts for attribute data such as defects or non-conformities. Each type of control chart is tailored to address specific aspects of a process, allowing organizations to choose the most appropriate chart based on the nature of the data they are monitoring.

Self-Assessment Question

1. Explain the concept of a control chart. Provide a real-world example of a situation where a control chart could be effectively used.
2. Describe how the Upper Control Limit (UCL) and Lower Control Limit (LCL) are calculated in an X-bar chart. Include the significance of these limits in statistical process control.
3. Compare and contrast the application of p chart and c chart. Provide examples of situations where each chart type would be most appropriate for monitoring processes.
4. Discuss the indicators of a stable process when examining control charts. How can one distinguish between chance cause and Assignable cause variations using the charts?
5. Given the sample mean and range values for ten samples, each consisting of five data points, create both mean and range charts. Evaluate the charts and give comments on the current state of process control.

   You may use the following control chart constants.

For n =5: A_2 =0.58: D_3 =0: D_4 =2.115

Sample No:	1	2	3	4	5	6	7	8	9	10
Mean:	43	49	37	44	45	37	51	46	43	47
Range:	5	6	5	7	7	4	8	6	4	6

Ans: $\bar{X}$-chart (CL=44.20, UCL=47.56, LCL =40.83), $\bar{R}$-chart (CL =5.80, UCL =12.267, LCL =0)

6. Construct a control chart for the mean and range for the data in which sample of 5 are taken

x1	107	117	108	150	121	57	104	123	145	112	81	129	84	142	140	143
x2	82	123	51	60	61	95	133	103	95	132	64	86	85	116	56	98
x3	100	63	78	125	54	72	147	125	97	145	123	70	121	141	135	108
x4	150	140	68	73	121	111	110	126	58	60	83	66	143	124	118	75
x5	104	107	90	147	107	102	130	106	110	113	99	140	72	97	106	95

Ans: $\bar{X}$-chart (CL=104.475, UCL=142.4128, LCL =66.53), $\bar{R}$-chart (CL =65.75, UCL =139.06, LCL =0)

7. The following figures give the number of defectives in 20 samples, each sample containing 2000 items

Sample No	**No of defectives**
1	126
2	409
3	193
4	326
5	280
6	389
7	280
8	306

9	337
10	305
11	356
12	402
13	216
14	264
15	425
16	430
17	216
18	341
19	225
20	322

Prepare p-chart and comment if the process can be regarded in control or not?

Ans: CL=0.1537, UCL=0.178, LCL =0.130

Assignments

1. Explain the relationship between sample size and the np chart. How does a change in sample size impact the sensitivity and effectiveness of the chart in detecting variations?
2. Outline the step-by-step process of constructing an X-bar chart and an R chart. Include the key elements such as sample selection, calculations, and plotting data points.
3. Explain how control charts contribute to the concept of continuous improvement in a manufacturing environment. How can organizations use control charts as a tool for ongoing process optimization?
4. Construct a control chart for the mean and range for the data in which sample of 5 are taken

x_1	x_2	x_3	x_4	x_5
93	103	131	58	131
123	150	83	90	107
119	82	69	75	137
146	112	60	117	58
57	113	123	85	88
149	100	58	84	57
69	115	75	87	94
94	129	126	95	91
63	102	67	66	119
56	80	104	64	58
89	113	60	108	140
81	118	121	144	87
125	137	114	124	99
72	79	145	56	80
101	78	99	85	77
122	114	71	60	106

Comment whether the production seems to be under control

Ans: $\bar{X}$-chart (CL=96.46, UCL=132.45, LCL =60.47), $\bar{R}$-chart (CL =62.375, UCL =131.92, LCL =0)

5. The number of defects in ten pieces of cloth from different rolls of equal length were observed as follows: 2, 1, 3, 7, 3, 0, 2, 8, 4, 1. Construct c chart based on this data and determine whether the production process is currently in a state of statistical control.

Ans: CL=3.10, UCL=8.38, LCL=0

6. In a quality inspection of equal lengths of electronic components, the number of defects observed were 6, 7, 4, 2, 3, 4, 0, 5, 3, 2. Create a control chart based on the

defect data and provide comments on whether the manufacturing process is currently under control.

Ans: CL=3.60, UCL=9.29, LCL=0

7. The following is the number of defective items observed in 15 consecutive sample of size 50 each.

12	9	15	14	10	8	6	12	9	5	12	10	11	9	10

Draw the control chart for fraction defectives and comment upon the state of control of the manufacturing process.

Ans: CL=0.203, UCL=0.374, LCL=0.032

Reference

1. Ganguli, B. N. (2010). *Statistics and Statistical Methods: In Two Volumes*. New Delhi: New Age International Publishers.
2. Gupta, S. C. (2014). *Fundamentals of Statistics*. Meerut: Rastogi Publications.
3. Kulkarni, P. D., & Deshpande, J. V. (2010). *Statistics:* Essentials and Applications. Pune: Sujata Prakashan.
4. Mehta, P., & Ramachandran, V. (2010). *Statistics for Management*. New Delhi: Vikas Publishing House.
5. Pandey, J. N., & Gupta, T. C. (2017). *Advanced Statistics*. Delhi: Wiley India Pvt. Ltd.
6. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Statistical power analysis for the behavioral sciences* (3rd ed.). Academic Press.
7. Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). Sage Publications Ltd.
8. Yadav, V. K. (2016). *An Introduction to Probability and Statistics*. New Delhi: New Age International Publishers.
9. Howell, D. C. (2010). *Statistical methods for psychology (7th ed.)*. Wadsworth Cengage Learning.

Suggested Reading

1. Trivedi, A. K. (2012). *Practical Statistics with R.* Noida: McGraw-Hill Education.
2. Urmila, P. (2016). *Fundamentals of Biostatistics & Research Methodology*. Noida: McGraw-Hill Education.
3. Kline, R. B. (2015). *Principles and practice of structural equation modeling (4th ed.).* Guilford Publications.
4. Spearman, C. (1987). *The proof and measurement of association between two things by correlation coefficient*. British journal of psychology, 8(1), 1-28.
5. Cohen, S. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associates.

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

STATISTICAL TABLES

Cumulative normal distribution

Critical values of the *t* distribution

Critical values of the *F* distribution

Critical values of the chi-squared distribution

LOGARITHAM

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5566	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6345	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7769	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9843	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

ANTILOGARITHAM

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	2	2	2	3	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
.20	1585	1289	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	4	5	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

ANTILOGARITHAM

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6715	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	12	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Standard Normal Distribution Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

TABLE A.3

F Distribution: Critical Values of *F* (5% significance level)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62
250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59
600	3.86	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.90	1.85	1.77	1.71	1.66	1.62	1.59
750	3.85	3.01	2.62	2.38	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.70	1.66	1.62	1.58
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (5% significance level)

v_2 \ v_1	25	30	35	40	50	60	75	100	150	200
1	249.26	250.10	250.69	251.14	251.77	252.20	252.62	253.04	253.46	253.68
2	19.46	19.46	19.47	19.47	19.48	19.48	19.48	19.49	19.49	19.49
3	8.63	8.62	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.54
4	5.77	5.75	5.73	5.72	5.70	5.69	5.68	5.66	5.65	5.65
5	4.52	4.50	4.48	4.46	4.44	4.43	4.42	4.41	4.39	4.39
6	3.83	3.81	3.79	3.77	3.75	3.74	3.73	3.71	3.70	3.69
7	3.40	3.38	3.36	3.34	3.32	3.30	3.29	3.27	3.26	3.25
8	3.11	3.08	3.06	3.04	3.02	3.01	2.99	2.97	2.96	2.95
9	2.89	2.86	2.84	2.83	2.80	2.79	2.77	2.76	2.74	2.73
10	2.73	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.57	2.56
11	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.46	2.44	2.43
12	2.50	2.47	2.44	2.43	2.40	2.38	2.37	2.35	2.33	2.32
13	2.41	2.38	2.36	2.34	2.31	2.30	2.28	2.26	2.24	2.23
14	2.34	2.31	2.28	2.27	2.24	2.22	2.21	2.19	2.17	2.16
15	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.10	2.10
16	2.23	2.19	2.17	2.15	2.12	2.11	2.09	2.07	2.05	2.04
17	2.18	2.15	2.12	2.10	2.08	2.06	2.04	2.02	2.00	1.99
18	2.14	2.11	2.08	2.06	2.04	2.02	2.00	1.98	1.96	1.95
19	2.11	2.07	2.05	2.03	2.00	1.98	1.96	1.94	1.92	1.91
20	2.07	2.04	2.01	1.99	1.97	1.95	1.93	1.91	1.89	1.88
21	2.05	2.01	1.98	1.96	1.94	1.92	1.90	1.88	1.86	1.84
22	2.02	1.98	1.96	1.94	1.91	1.89	1.87	1.85	1.83	1.82
23	2.00	1.96	1.93	1.91	1.88	1.86	1.84	1.82	1.80	1.79
24	1.97	1.94	1.91	1.89	1.86	1.84	1.82	1.80	1.78	1.77
25	1.96	1.92	1.89	1.87	1.84	1.82	1.80	1.78	1.76	1.75
26	1.94	1.90	1.87	1.85	1.82	1.80	1.78	1.76	1.74	1.73
27	1.92	1.88	1.86	1.84	1.81	1.79	1.76	1.74	1.72	1.71
28	1.91	1.87	1.84	1.82	1.79	1.77	1.75	1.73	1.70	1.69
29	1.89	1.85	1.83	1.81	1.77	1.75	1.73	1.71	1.69	1.67
30	1.88	1.84	1.81	1.79	1.76	1.74	1.72	1.70	1.67	1.66
35	1.82	1.79	1.76	1.74	1.70	1.68	1.66	1.63	1.61	1.60
40	1.78	1.74	1.72	1.69	1.66	1.64	1.61	1.59	1.56	1.55
50	1.73	1.69	1.66	1.63	1.60	1.58	1.55	1.52	1.50	1.48
60	1.69	1.65	1.62	1.59	1.56	1.53	1.51	1.48	1.45	1.44
70	1.66	1.62	1.59	1.57	1.53	1.50	1.48	1.45	1.42	1.40
80	1.64	1.60	1.57	1.54	1.51	1.48	1.45	1.43	1.39	1.38
90	1.63	1.59	1.55	1.53	1.49	1.46	1.44	1.41	1.38	1.36
100	1.62	1.57	1.54	1.52	1.48	1.45	1.42	1.39	1.36	1.34
120	1.60	1.55	1.52	1.50	1.46	1.43	1.40	1.37	1.33	1.32
150	1.58	1.54	1.50	1.48	1.44	1.41	1.38	1.34	1.31	1.29
200	1.56	1.52	1.48	1.46	1.41	1.39	1.35	1.32	1.28	1.26
250	1.55	1.50	1.47	1.44	1.40	1.37	1.34	1.31	1.27	1.25
300	1.54	1.50	1.46	1.43	1.39	1.36	1.33	1.30	1.26	1.23
400	1.53	1.49	1.45	1.42	1.38	1.35	1.32	1.28	1.24	1.22
500	1.53	1.48	1.45	1.42	1.38	1.35	1.31	1.28	1.23	1.21
600	1.52	1.48	1.44	1.41	1.37	1.34	1.31	1.27	1.23	1.20
750	1.52	1.47	1.44	1.41	1.37	1.34	1.30	1.26	1.22	1.20
1000	1.52	1.47	1.43	1.41	1.36	1.33	1.30	1.26	1.22	1.19

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (1% significance level)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6142.67	6170.10	6191.53	6208.73
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.44	99.45
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.75	26.69
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.08	14.02
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.77	9.68	9.61	9.55
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.72	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.64	2.56	2.50	2.44
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.46	2.38	2.32	2.27
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.35	2.27	2.20	2.15
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.29	2.21	2.14	2.09
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.27	2.19	2.12	2.07
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.23	2.15	2.09	2.03
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97
250	6.74	4.69	3.86	3.40	3.09	2.87	2.71	2.58	2.48	2.39	2.26	2.15	2.07	2.01	1.95
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.23	2.13	2.05	1.98	1.92
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92
600	6.68	4.64	3.81	3.35	3.05	2.83	2.67	2.54	2.44	2.35	2.21	2.11	2.03	1.96	1.91
750	6.67	4.63	3.81	3.34	3.04	2.83	2.66	2.53	2.43	2.34	2.21	2.11	2.02	1.96	1.90
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (1% significance level)

v_1 / v_2	25	30	35	40	50	60	75	100	150	200
1	6239.83	6260.65	6275.57	6286.78	6302.52	6313.03	6323.56	6334.11	6344.68	6349.97
2	99.46	99.47	99.47	99.47	99.48	99.48	99.49	99.49	99.49	99.49
3	26.58	26.50	26.45	26.41	26.35	26.32	26.28	26.24	26.20	26.18
4	13.91	13.84	13.79	13.75	13.69	13.65	13.61	13.58	13.54	13.52
5	9.45	9.38	9.33	9.29	9.24	9.20	9.17	9.13	9.09	9.08
6	7.30	7.23	7.18	7.14	7.09	7.06	7.02	6.99	6.95	6.93
7	6.06	5.99	5.94	5.91	5.86	5.82	5.79	5.75	5.72	5.70
8	5.26	5.20	5.15	5.12	5.07	5.03	5.00	4.96	4.93	4.91
9	4.71	4.65	4.60	4.57	4.52	4.48	4.45	4.41	4.38	4.36
10	4.31	4.25	4.20	4.17	4.12	4.08	4.05	4.01	3.98	3.96
11	4.01	3.94	3.89	3.86	3.81	3.78	3.74	3.71	3.67	3.66
12	3.76	3.70	3.65	3.62	3.57	3.54	3.50	3.47	3.43	3.41
13	3.57	3.51	3.46	3.43	3.38	3.34	3.31	3.27	3.24	3.22
14	3.41	3.35	3.30	3.27	3.22	3.18	3.15	3.11	3.08	3.06
15	3.28	3.21	3.17	3.13	3.08	3.05	3.01	2.98	2.94	2.92
16	3.16	3.10	3.05	3.02	2.97	2.93	2.90	2.86	2.83	2.81
17	3.07	3.00	2.96	2.92	2.87	2.83	2.80	2.76	2.73	2.71
18	2.98	2.92	2.87	2.84	2.78	2.75	2.71	2.68	2.64	2.62
19	2.91	2.84	2.80	2.76	2.71	2.67	2.64	2.60	2.57	2.55
20	2.84	2.78	2.73	2.69	2.64	2.61	2.57	2.54	2.50	2.48
21	2.79	2.72	2.67	2.64	2.58	2.55	2.51	2.48	2.44	2.42
22	2.73	2.67	2.62	2.58	2.53	2.50	2.46	2.42	2.38	2.36
23	2.69	2.62	2.57	2.54	2.48	2.45	2.41	2.37	2.34	2.32
24	2.64	2.58	2.53	2.49	2.44	2.40	2.37	2.33	2.29	2.27
25	2.60	2.54	2.49	2.45	2.40	2.36	2.33	2.29	2.25	2.23
26	2.57	2.50	2.45	2.42	2.36	2.33	2.29	2.25	2.21	2.19
27	2.54	2.47	2.42	2.38	2.33	2.29	2.26	2.22	2.18	2.16
28	2.51	2.44	2.39	2.35	2.30	2.26	2.23	2.19	2.15	2.13
29	2.48	2.41	2.36	2.33	2.27	2.23	2.20	2.16	2.12	2.10
30	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.13	2.09	2.07
35	2.35	2.28	2.23	2.19	2.14	2.10	2.06	2.02	1.98	1.96
40	2.27	2.20	2.15	2.11	2.06	2.02	1.98	1.94	1.90	1.87
50	2.17	2.10	2.05	2.01	1.95	1.91	1.87	1.82	1.78	1.76
60	2.10	2.03	1.98	1.94	1.88	1.84	1.79	1.75	1.70	1.68
70	2.05	1.98	1.93	1.89	1.83	1.78	1.74	1.70	1.65	1.62
80	2.01	1.94	1.89	1.85	1.79	1.75	1.70	1.65	1.61	1.58
90	1.99	1.92	1.86	1.82	1.76	1.72	1.67	1.62	1.57	1.55
100	1.97	1.89	1.84	1.80	1.74	1.69	1.65	1.60	1.55	1.52
120	1.93	1.86	1.81	1.76	1.70	1.66	1.61	1.56	1.51	1.48
150	1.90	1.83	1.77	1.73	1.66	1.62	1.57	1.52	1.46	1.43
200	1.87	1.79	1.74	1.69	1.63	1.58	1.53	1.48	1.42	1.39
250	1.85	1.77	1.72	1.67	1.61	1.56	1.51	1.46	1.40	1.36
300	1.84	1.76	1.70	1.66	1.59	1.55	1.50	1.44	1.38	1.35
400	1.82	1.75	1.69	1.64	1.58	1.53	1.48	1.42	1.36	1.32
500	1.81	1.74	1.68	1.63	1.57	1.52	1.47	1.41	1.34	1.31
600	1.80	1.73	1.67	1.63	1.56	1.51	1.46	1.40	1.34	1.30
750	1.80	1.72	1.66	1.62	1.55	1.50	1.45	1.39	1.33	1.29
1000	1.79	1.72	1.66	1.61	1.54	1.50	1.44	1.38	1.32	1.28

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (0.1% significance level)

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20
1	4.05e05	5.00e05	5.40e05	5.62e05	5.76e05	5.86e05	5.93e05	5.98e05	6.02e05	6.06e05	6.11e05	6.14e05	6.17e05	6.19e05	6.21e05
2	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.42	999.43	999.44	999.44	999.45
3	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	128.32	127.64	127.14	126.74	126.42
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.95	46.60	46.32	46.10
5	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	26.42	26.06	25.78	25.57	25.39
6	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.99	17.68	17.45	17.27	17.12
7	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.43	13.23	13.06	12.93
8	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	11.19	10.94	10.75	10.60	10.48
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.33	9.15	9.01	8.90
10	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.45	8.22	8.05	7.91	7.80
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.41	7.24	7.11	7.01
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.79	6.63	6.51	6.40
13	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.31	6.16	6.03	5.93
14	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.93	5.78	5.66	5.56
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.62	5.46	5.35	5.25
16	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.55	5.35	5.20	5.09	4.99
17	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.32	5.13	4.99	4.87	4.78
18	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.94	4.80	4.68	4.59
19	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.78	4.64	4.52	4.43
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.64	4.49	4.38	4.29
21	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.51	4.37	4.26	4.17
22	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.40	4.26	4.15	4.06
23	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.30	4.16	4.05	3.96
24	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.21	4.07	3.96	3.87
25	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.31	4.13	3.99	3.88	3.79
26	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	4.06	3.92	3.81	3.72
27	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.99	3.86	3.75	3.66
28	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.11	3.93	3.80	3.69	3.60
29	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.88	3.74	3.63	3.54
30	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.82	3.69	3.58	3.49
35	12.90	8.47	6.79	5.88	5.30	4.89	4.59	4.36	4.18	4.03	3.79	3.62	3.48	3.38	3.29
40	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.47	3.34	3.23	3.14
50	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.44	3.27	3.41	3.04	2.95
60	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.32	3.15	3.02	2.91	2.83
70	11.80	7.64	6.06	5.20	4.66	4.28	3.99	3.77	3.60	3.45	3.23	3.06	2.93	2.83	2.74
80	11.67	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.53	3.39	3.16	3.00	2.87	2.76	2.68
90	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34	3.11	2.95	2.82	2.71	2.63
100	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.91	2.78	2.68	2.59
120	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.85	2.72	2.62	2.53
150	11.27	7.24	5.71	4.88	4.35	3.98	3.71	3.49	3.32	3.18	2.96	2.80	2.67	2.56	2.48
200	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12	2.90	2.74	2.61	2.51	2.42
250	11.09	7.10	5.59	4.77	4.25	3.88	3.61	3.40	3.23	3.09	2.87	2.71	2.58	2.48	2.39
300	11.04	7.07	5.56	4.75	4.22	3.86	3.59	3.38	3.21	3.07	2.85	2.69	2.56	2.46	2.37
400	10.99	7.03	5.53	4.71	4.19	3.83	3.56	3.35	3.18	3.04	2.82	2.66	2.53	2.43	2.34
500	10.96	7.00	5.51	4.69	4.18	3.81	3.54	3.33	3.16	3.02	2.81	2.64	2.52	2.41	2.33
600	10.94	6.99	5.49	4.68	4.16	3.80	3.53	3.32	3.15	3.01	2.80	2.63	2.51	2.40	2.32
750	10.91	6.97	5.48	4.67	4.15	3.79	3.52	3.31	3.14	3.00	2.78	2.62	2.49	2.39	2.31
1000	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13	2.99	2.77	2.61	2.48	2.38	2.30

TABLE A.3 (continued)

F Distribution: Critical Values of *F* (0.1% significance level)

v_1 / v_2	25	30	35	40	50	60	75	100	150	200
1	6.24e05	6.26e05	6.28e05	6.29e05	6.30e05	6.31e05	6.32e05	6.33e05	6.35e05	6.35e05
2	999.46	999.47	999.47	999.47	999.48	999.48	999.49	999.49	999.49	999.49
3	125.84	125.45	125.17	124.96	124.66	124.47	124.27	124.07	123.87	123.77
4	45.70	45.43	45.23	45.09	44.88	44.75	44.61	44.47	44.33	44.26
5	25.08	24.87	24.72	24.60	24.44	24.33	24.22	24.12	24.01	23.95
6	16.85	16.67	16.54	16.44	16.31	16.21	16.12	16.03	15.93	15.89
7	12.69	12.53	12.41	12.33	12.20	12.12	12.04	11.95	11.87	11.82
8	10.26	10.11	10.00	9.92	9.80	9.73	9.65	9.57	9.49	9.45
9	8.69	8.55	8.46	8.37	8.26	8.19	8.11	8.04	7.96	7.93
10	7.60	7.47	7.37	7.30	7.19	7.12	7.05	6.98	6.91	6.87
11	6.81	6.68	6.59	6.52	6.42	6.35	6.28	6.21	6.14	6.10
12	6.22	6.09	6.00	5.93	5.83	5.76	5.70	5.63	5.56	5.52
13	5.75	5.63	5.54	5.47	5.37	5.30	5.24	5.17	5.10	5.07
14	5.38	5.25	5.17	5.10	5.00	4.94	4.87	4.81	4.74	4.71
15	5.07	4.95	4.86	4.80	4.70	4.64	4.57	4.51	4.44	4.41
16	4.82	4.70	4.61	4.54	4.45	4.39	4.32	4.26	4.19	4.16
17	4.60	4.48	4.40	4.33	4.24	4.18	4.11	4.05	3.98	3.95
18	4.42	4.30	4.22	4.15	4.06	4.00	3.93	3.87	3.80	3.77
19	4.26	4.14	4.06	3.99	3.90	3.84	3.78	3.71	3.65	3.61
20	4.12	4.00	3.92	3.86	3.77	3.70	3.64	3.58	3.51	3.48
21	4.00	3.88	3.80	3.74	3.64	3.58	3.52	3.46	3.39	3.36
22	3.89	3.78	3.70	3.63	3.54	3.48	3.41	3.35	3.28	3.25
23	3.79	3.68	3.60	3.53	3.44	3.38	3.32	3.25	3.19	3.16
24	3.71	3.59	3.51	3.45	3.36	3.29	3.23	3.17	3.10	3.07
25	3.63	3.52	3.43	3.37	3.28	3.22	3.15	3.09	3.03	2.99
26	3.56	3.44	3.36	3.30	3.21	3.15	3.08	3.02	2.95	2.92
27	3.49	3.38	3.30	3.23	3.14	3.08	3.02	2.96	2.89	2.86
28	3.43	3.32	3.24	3.18	3.09	3.02	2.96	2.90	2.83	2.80
29	3.38	3.27	3.18	3.12	3.03	2.97	2.91	2.84	2.78	2.74
30	3.33	3.22	3.13	3.07	2.98	2.92	2.86	2.79	2.73	2.69
35	3.13	3.02	2.93	2.87	2.78	2.72	2.66	2.59	2.52	2.49
40	2.98	2.87	2.79	2.73	2.64	2.57	2.51	2.44	2.38	2.34
50	2.79	2.68	2.60	2.53	2.44	2.38	2.31	2.25	2.18	2.14
60	2.67	2.55	2.47	2.41	2.32	2.25	2.19	2.12	2.05	2.01
70	2.58	2.47	2.39	2.32	2.23	2.16	2.10	2.03	1.95	1.92
80	2.52	2.41	2.32	2.26	2.16	2.10	2.03	1.96	1.89	1.85
90	2.47	2.36	2.27	2.21	2.11	2.05	1.98	1.91	1.83	1.79
100	2.43	2.32	2.24	2.17	2.08	2.01	1.94	1.87	1.79	1.75
120	2.37	2.26	2.18	2.11	2.02	1.95	1.88	1.81	1.73	1.68
150	2.32	2.21	2.12	2.06	1.96	1.89	1.82	1.74	1.66	1.62
200	2.26	2.15	2.07	2.00	1.90	1.83	1.76	1.68	1.60	1.55
250	2.23	2.12	2.03	1.97	1.87	1.80	1.72	1.65	1.56	1.51
300	2.21	2.10	2.01	1.94	1.85	1.78	1.70	1.62	1.53	1.48
400	2.18	2.07	1.98	1.92	1.82	1.75	1.67	1.59	1.50	1.45
500	2.17	2.05	1.97	1.90	1.80	1.73	1.65	1.57	1.48	1.43
600	2.16	2.04	1.96	1.89	1.79	1.72	1.64	1.56	1.46	1.41
750	2.15	2.03	1.95	1.88	1.78	1.71	1.63	1.55	1.45	1.40
1000	2.14	2.02	1.94	1.87	1.77	1.69	1.62	1.53	1.44	1.38

TABLE A.4

χ^2 (Chi-Squared) Distribution: Critical Values of χ^2

	Significance level		
Degrees of freedom	5%	1%	0.1%
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515
6	12.592	16.812	22.458
7	14.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

Table of Control Chart Constants

	X-bar Chart Constants	for sigma estimate	R Chart Constants		S Chart Constants	

Sample Size = m	A_2	A_3	d_2	D_3	D_4	B_3	B_4
2	1.880	2.659	1.128	0	3.267	0	3.267
3	1.023	1.954	1.693	0	2.574	0	2.568
4	0.729	1.628	2.059	0	2.282	0	2.266
5	0.577	1.427	2.326	0	2.114	0	2.089
6	0.483	1.287	2.534	0	2.004	0.030	1.970
7	0.419	1.182	2.704	0.076	1.924	0.118	1.882
8	0.373	1.099	2.847	0.136	1.864	0.185	1.815
9	0.337	1.032	2.970	0.184	1.816	0.239	1.761
10	0.308	0.975	3.078	0.223	1.777	0.284	1.716
11	0.285	0.927	3.173	0.256	1.744	0.321	1.679
12	0.266	0.886	3.258	0.283	1.717	0.354	1.646
13	0.249	0.850	3.336	0.307	1.693	0.382	1.618
14	0.235	0.817	3.407	0.328	1.672	0.406	1.594
15	0.223	0.789	3.472	0.347	1.653	0.428	1.572
16	0.212	0.763	3.532	0.363	1.637	0.448	1.552
17	0.203	0.739	3.588	0.378	1.622	0.466	1.534
18	0.194	0.718	3.640	0.391	1.608	0.482	1.518
19	0.187	0.698	3.689	0.403	1.597	0.497	1.503
20	0.180	0.680	3.735	0.415	1.585	0.510	1.490
21	0.173	0.663	3.778	0.425	1.575	0.523	1.477
22	0.167	0.647	3.819	0.434	1.566	0.534	1.466
23	0.162	0.633	3.858	0.443	1.557	0.545	1.455
24	0.157	0.619	3.895	0.451	1.548	0.555	1.445
25	0.153	0.606	3.931	0.459	1.541	0.565	1.435

Control chart constants for X-bar, R, S, Individuals (called "X" or "I" charts), and MR (Moving Range) Charts.

NOTES: To construct the "X" and "MR" charts (these are companions) we compute the Moving Ranges as:

R_2 = range of 1st and 2nd observations, R_3 = range of 2nd and 3rd observations, R_4 = range of 3rd and 4th observations, etc. with the "average" moving range or "MR-bar" being the average of these ranges with the "sample size" for each of these ranges being n = 2 since each is based on consecutive observations ... this should provide an estimated standard deviation (needed for the "I" chart) of

$\sigma = (\text{MR-bar})/d_2$ where the value of d_2 is based on, as just stated, m = 2.

Similarly, the UCL and LCL for the MR chart will be: $UCL = D_4(\text{MR-bar})$ and $LCL = D_3(\text{MR-bar})$

but, since $D_3 = 0$ when n = 0 (or, more accurately, is "not applicable") there will be no LCL for the MR chart, just a UCL.

Model Question Paper Sets

SREENARAYANAGURU OPEN UNIVERSITY

QP CODE: ……

Reg. No : .....

Name : …….....

Model Question Paper- set-I

Second Semester MCom Degree Examination

(Core Course VII)
M21CM07DC: QUANTITATIVE TECHNIQUES

2023-24 - Admission Onwards

Time: 3 Hours Max Marks: 70

Section A

Answer any five of the following questions in one or two sentences each. Each question carries 2 marks. **(5x2=10 marks)**

1. What are regression coefficients?
2. Distinguish between Type I error and Type II error.
3. What is an assignable cause of variation? What are the common types?
4. When is an event considered to be mutually exclusive? Explain with an example.
5. What is a binomial distribution?
6. What are the applications of multiple correlation?
7. Why do organisations use control charts?
8. State the assumptions of parametric tests.

Section B

Answer any six of the following questions in one page each. Each question carries 5 marks. **(6x5=30 marks)**

9. Given the data below, what relationship can you inferred between exercise frequency and dietary preference?

	Regular Exercise	Infrequent Exercise
Plant-based Diet	87	52
Omnivorous Diet	41	68

10. A coin is tossed five times. What is the probability of obtaining three or more heads?

11. The rate of receiving a certain increment is such that on an average 30% of workers receive it. If 10 workers are selected at random, find the probability that:

    i) Exactly 2 workers receive the increment.

    ii) Not more than 2 workers receive the increment.

    iii) At least 2 workers receive the increment.

12. State one null hypothesis? Explain the procedures for testing it.

13. What is a control chart for attributes? What are the various types of control charts for attributes?

14. Below is the data regarding the performance of four fertilizers across three test fields with the specified crop yields. Find whether there is a significant difference in the yield effectiveness of the fertilizers used.

Test Field	Fertilizer A	Fertilizer B	Fertilizer C	Fertilizer D
I	11	7	9	9
II	9	5	5	4
III	10	6	8	6

15. What are the different methods used to study correlation?

16. The academic performance data in two subjects, Mathematics (M) and Science (S), from an annual examination is given below:

Particulars	Mathematics (M)	Science (S)
Mean	41	55
Standard Deviation	10	15

With a coefficient of correlation of + 0.3, estimate the science score for a mathematics score of 45, and the mathematics score for a science score of 38.

17. Explain the role of Statistical Quality Control (SQC), in data -driven decision making.

18. 'Quantitative Techniques play a crucial role in business decision making'. Elucidate.

Section C

Answer any two of the following questions in four pages each. Each question carries 15 marks. **(2x15=30 marks)**

19. Consider the data given below:

Advertisement (₹):	10	50	32	31	54	45	40	35
Sales (₹):	40	120	80	75	130	110	95	92

i. Predict sales, if advertisement expenditure is 60.
ii. Predict the advertisement expenditure, if the sales are 87.
iii. Calculate Karl Pearson's Correlation Coefficient?

20. Explain Chi-square test and its applications. Calculate whether a population variance of 15 kg can be inferred from a randomly selected sample group of 10 students, with weights in kilograms: 50, 48, 53, 38, 57, 47, 48, 45, 55, and 45.

21. During a two-week production phase, twelve batches of 200 smartphones each underwent inspection. The data below represents the number of defective smartphones identified in each batch

Sample No:	1	2	3	4	5	6	7	8	9	10	11	12
No. of defective:	28	35	37	33	48	29	28	24	18	12	8	12

i. Draw the control chart for fraction defective.
ii. What do you find out from the chart

22. What is Poisson distribution, and what are its characteristics? Describe the relationship between the Binomial distribution and the Poisson distribution.

SREENARAYANAGURU OPEN UNIVERSITY

QP CODE: ……

Reg. No : .....

Name : …….....

Model Question Paper- set-II

Second Semester MCom Degree Examination

(Core Course VII)
M21CM07DC: QUANTITATIVE TECHNIQUES

2023-24 - Admission Onwards

Time: 3 Hours Max Marks: 70

Section A

Answer any five questions in one or two sentences each. Each question carries 2 marks.

(5x2=10 marks)

1. When do we use Kruskal Wallis H test?
2. Differentiate process control and product control.
3. Explain the properties of Pearson Coefficient of Correlation
4. Under what circumstances is a normal distribution considered as a suitable approximation for the Poisson distribution?
5. If the probability of defective bolts is 0.3, find the mean and standard deviation for the distribution of defective bolts in a total of 500.
6. Explain central limit theorem
7. Distinguish between one tailed and two tailed tests.
8. Find Karl Pearson's correlation coefficient between 'X' and 'Y' for the following data,

$N = 750$, $\sigma(x) = 4.4$, $\sigma(y) = 3.8$, $\sum(x - \bar{x}).(y - \bar{y}) = 3500$

Section B

Answer any six questions in one page each. Each question carries 5 marks.

(6x5=30 marks)

9. What are the properties of correlation? Explain its various types.
10. In a manufacturing process, 3% of the bicycles produced are found to be defective. What is the probability that a shipment of 150 bicycles will contain exactly 8% defectives? Also find the mean and standard deviation.
11. Differentiate between correlation and regression analysis.
12. In a series of coin flips, where each flip has a probability of 0.6 of landing heads, what is the probability of obtaining exactly 3 heads in 5 flips? Calculate the mean and standard deviation of the number of heads obtained in 5 flips.
13. Define Statistical Quality Control (SQC) and enumerate its types of variation.
14. Upon discovering that a manufacturing process is operating within control parameters, where the average number of defectives per sample batch of 15 is 1.5, determine the appropriate control limits for a quality control chart based on the proportion of defectives in these sample batches of 15.
15. Mobile phones manufactured by X Corporation and Y Corporation provided the following data:

Particulars	**X Corporation**	**Y Corporation**
Number of phones tested	200	200
Mean Battery Life in Hours	36	40
Standard Deviation	3	4

Determine whether there is a significant difference in the battery life between the two manufacturers' phones.

16. Consider the following data on the performance of three different types of fertilizers across four different fields:

Plots	**Fertilizer Types**		
	A	**B**	**C**
I	15	18	20
II	12	17	16
III	18	17	22
IV	16	20	21

Examine whether there is a notable difference in the effectiveness of fertilizers and assess the significance of the variance between the fields.

17. Calculate the Pearson correlation coefficient for the marks obtained by ten students in a class test. The marks are as follows:

Marks in Economics:	48	70	65	37	90	40	50	75	85	60
Marks in Statistics:	39	40	70	40	95	60	40	80	80	50

i. Compute its probable error, considering working means as 61 and 66
ii. Analyse the significance of the correlation coefficient "r". Determine the range within which the population correlation coefficient is expected to fall

18. What is a mean chart? How do you calculate it?

Section C

Answer any two questions in four pages each. Each question carries 15 marks.
(2x15=30 marks)

19. In a cooking competition, two chefs rank the ten dishes as follows. Calculate the Spearman's Rank Correlation and provide an interpretation of the results.

Chef A's Rankings:	5	4	7	8	2	1	6	3	10	9
Chef B's Rankings:	6	5	4	8	1	2	7	3	9	10

20. What is a normal distribution? What are its characteristics and uses?
21. Develop a control chart for the mean and range based on samples of size 5 for the following data representing the daily production rates (in units) of 6 different assembly lines:

Assembly Line 1	104	123	145	112	81	129
Assembly Line 2	133	103	95	132	64	86
Assembly Line 3	147	125	97	145	123	70
Assembly Line 4	110	126	58	60	83	66
Assembly Line 5	130	106	110	113	99	140

22. The sales figures (in thousands) of 10 employees before and after attending a sales training program are provided in the table below:

Before:	28	35	42	39	45	33	40	29	36	41
After:	32	40	45	43	48	36	42	33	38	39

Conduct a test to determine if there's a significant difference in their sales performance before and after the training program.

സർവ്വകലാശാലാഗീതം

വിദ്യയാൽ സ്വതന്ത്രരാകണം
വിശ്വപൗരരായി മാറണം
ഗ്രഹപ്രസാദമായ് വിളങ്ങണം
ഗുരുപ്രകാശമേ നയിക്കണേ

കൂരിരുട്ടിൽ നിന്നു ഞങ്ങളെ
സൂര്യവീഥിയിൽ തെളിക്കണം
സ്നേഹദീപ്തിയായ് വിളങ്ങണം
നീതിവൈജയന്തി പാറണം

ശാസ്ത്രവ്യാപ്തിയെന്നുമേകണം
ജാതിഭേദമാകെ മാറണം
ബോധരശ്മിയിൽ തിളങ്ങുവാൻ
ജ്ഞാനകേന്ദ്രമേ ജ്വലിക്കണേ

കുരീപ്പുഴ ശ്രീകുമാർ

SREENARAYANAGURU OPEN UNIVERSITY

Regional Centres

Kozhikode
Govt. Arts and Science College
Meenchantha, Kozhikode,
Kerala, Pin: 673002
Ph: 04952920228
email: rckdirector@sgou.ac.in

Thalassery
Govt. Brennen College
Dharmadam, Thalassery,
Kannur, Pin: 670106
Ph: 04902990494
email: rctdirector@sgou.ac.in

Tripunithura
Govt. College
Tripunithura, Ernakulam,
Kerala, Pin: 682301
Ph: 04842927436
email: rcedirector@sgou.ac.in

Pattambi
Sree Neelakanta Govt. Sanskrit College
Pattambi, Palakkad,
Kerala, Pin: 679303
Ph: 04662912009
email: rcpdirector@sgou.ac.in

Quantitative Techniques

COURSE CODE: M21CM07DC

YouTube

Sreenarayanaguru Open University
Kollam, Kerala Pin- 691601, email: info@sgou.ac.in, www.sgou.ac.in Ph: +91 474 2966841

ISBN 978-81-970547-1-6