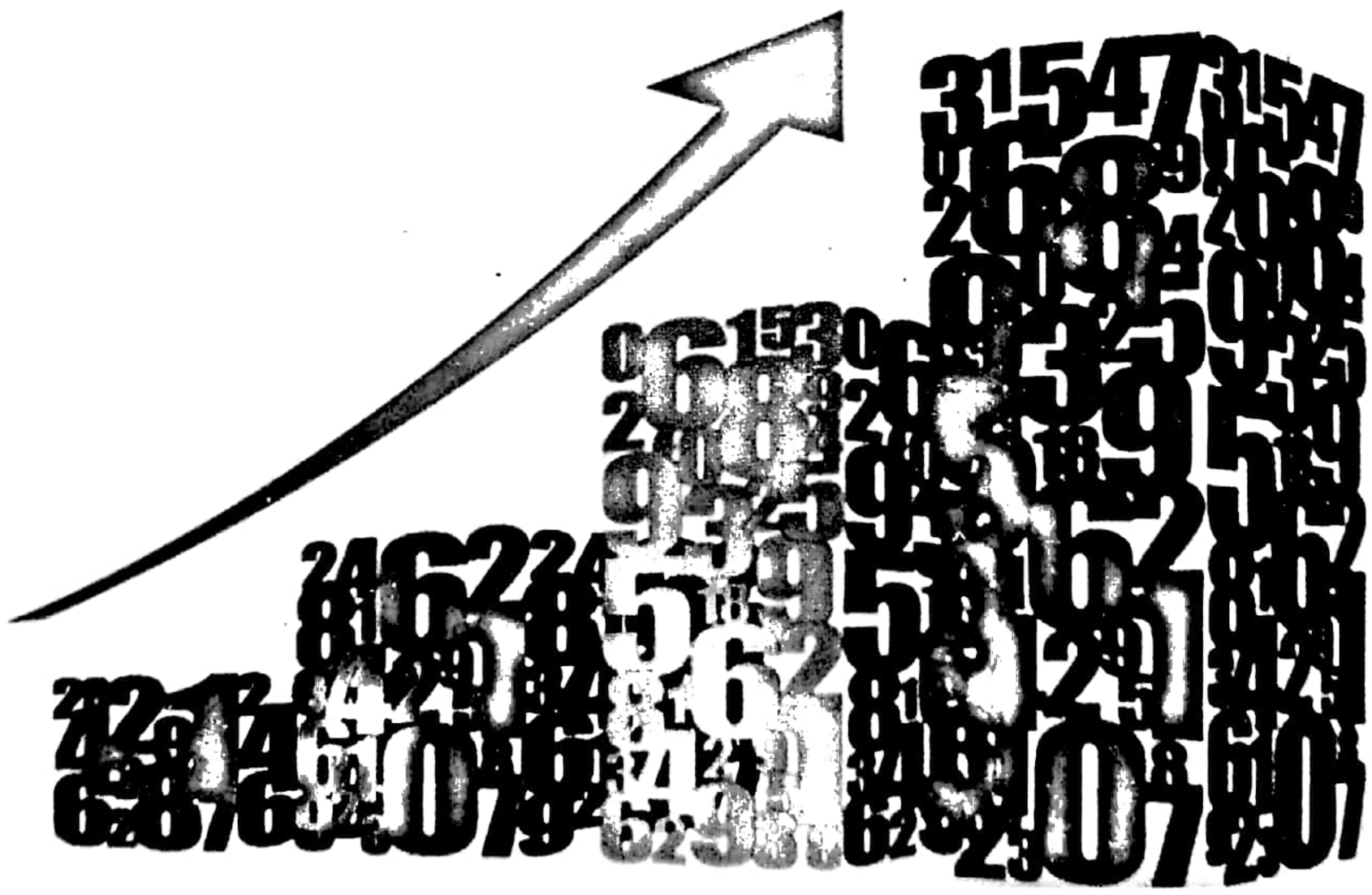
G.L. Book House

a complete manual

PROBABILITY & STATISTICS

Bachelor of Engineering



Govind Kumar Gupta

Arbind Kumar Gupta

Shree Ram Rauniyar

Sujit Singh

CONTENTS

Chapter 1 DESCRIPTIVE STATISTICS AND BASIC PROBABILITY

1.1 Introduction to Statistics and Its Importance in Engineering 2
1.2 Describing Data with Graphs (Bar, Pie, Line Diagram and Box Plot) 2
1.2.1 Bar Graph 2
1.2.2 Pie Graph 3
1.2.3 Line Diagram 3
1.2.4 Box Plot 3
1.3 Describing Data with Numerical Measure (Measuring Center, Meaning Variability) 5
1.3.1 Measure of Central Tangency 5
1.3.2 Measure of Dispersion or Variability 7
1.3.3 Difference between Measure of Central Tendency and Measure of Dispersion 9
1.4 Basic Probability, Additive Law, Multiplicative Law, Bayer's Theorem 9
1.4.1 Mutually Exclusive Events 9
1.4.2 Equally Likely Events 9
1.4.3 Independent Events 9
1.4.4 Dependent Events 10
1.4.5 Axiomatic or Modern Approach of Probability 10
1.4.6 Addition Theorem of Probability or Theorem of Total Probability 10
1.4.7 Multiplication Law of Probability 10
1.4.8 Conditional Probability 11
1.4.9 Baye's Theorem 11
Solved Questions (For BCE/BME/BIE/BAM] 12
Solved Questions (For BCT/BEX/BEL/B. Agri) 24

Chapter 2 DISCRETE PROBABILITY DISTRIBUTION

2.1 Random Variable 36
2.2 Binomial Distribution 36
2.2.1 Characteristics of Binomial Distribution 37
2.3 Negative Binomial Distribution 39
2.3.1 Properties 39
2.3.2 Differences 40
2.4 Poisson Distribution 41
2.4.1 Major Underlying Assumptions 41

2.4.2 Properties 41
2.4.3 Poisson's Distribution Derivation 43
2.4.4 Conditions for Poisson's Approximation to Binomial Distribution 43
2.5 Hyper Geometric Distance 41
2.5.1 Properties 44
Solved Questions (For BCE/BME/BIE/BAM] 46
Solved Questions (For BCT/BEX/BEL/B. Agri) 55

Chapter 3 CONTINUOUS PROBABILITY DISTRIBUTION

3.1 Continuous Random Variable and Probability Densities 68
3.1.1 Continuous Random Variable 68
3.1.2 Probability Density Function (PDF) 68
3.1.3 Cumulative Distribution Function (CDF) 69
3.1.4 Expected Value and Variables of a Continuous Ransom Variable 69
3.2 Normal Distribution 70
3.2.1 Standard Normal Distribution 71
3.3 Gamma Distribution 76
3.4 Chi-squared (χ^2) Distribution 79
Solved Questions (For BCE/BME/BIE/BAM] 80

Chapter 4 SAMPLING DISTRIBUTION

4.1 Population and Sample 96
4.1.1 Parameters and Statistics 96
4.2 Central Limit Theorem 96
4.3 Sampling Distribution of Sample Mean 98
4.4 Sampling Distribution of Sampling Proportion 100
4.5 Some More Information of Sampling Distribution of Sample Mean 100
Solved Questions (For BCE/BME/BIE/BAM] 103
Solved Questions (For BCT/BEX/BEL/B. Agri) 116

Chapter 5 INTERFERENCE CONCERNING MEAN

5.1 Point Estimation and Internal Estimation 126
5.1.1 Confidence Interval and Confidence Level 126
5.2 Test of Hypothesis 131
5.2.1 Procedure for Testing of Hypothesis 132
5.3 Hypothesis Test Concerning One Mean 134
5.3.1 Steps for Test of Significance of Single Mean for Large Sample ($n \geq 30$) 134
5.3.2 Steps for Test of Significance of a Single Mean for Small 138
5.4 Hypothesis Test Concerning Two Mean 139
5.4.1 Steps for Test of Significance of Difference of Two Means for Large Sample ($n > 30$) 139
5.4.2 Steps for Test of Significance of or Equality of Two Means for Large Sample 142
5.4.3 Steps for Test of Significance for Difference of Two Means for Small Sample 142
5.5 One Way Anova 143
5.6 Paired t-Test: (or Matched Pair t-Test) 148
Solved Questions (For BCE/BME/BIE/BAM] 151
Solved Questions (For BCT/BEX/BEL/B. Agri) 177

Chapter 6 INTERFERENCE CONCERNING PROPORTION

6.1 Estimation of Proportion 194
6.2 Hypothesis Concerning One Proportion 195
6.3 Hypothesis Concerning Two Proportions 197
6.3.1 For Difference of Two Proportions 197
6.3.2 For Equality of Two Proportions 198
6.4 Chi Square Test of Independence 199
Solved Questions (For BCE/BME/BIE/BAM] 201
Solved Questions (For BCT/BEX/BEL/B. Agri) 211

Chapter 7 CORRELATION AND REGRESSION

7.1 Least Square Method 226
7.2 Regression 227
7.2.1 Linear Regression with Least Square Method 227
7.2.2 Multiple Regression 227
7.3 Testing Significance of Regression 230
7.4 Correlation 231
7.4.1 Significance of Correlation 231
7.4.2 Types of Correlation 231
7.4.3 Partial Correlation Coefficient 233
7.4.4 Difference between Correlation and Regression 234
Solved Questions (For BCE/BME/BIE/BAM] 236

Chapter 8 APPLICATION OF COMPUTER ON STATISTICAL DATA COMPUTING

Solved Questions (For BCE/BME/BIE/BAM] 255

sample space: set of all possible outcomes from experiment
$S = \{HH, TT, HT, TH\}$

event: outcome or set of outcome of an activity is called event.

Eg: getting odd in die roll $= \{1, 3, 5\}$

exhustive event: if it includes all the possible outcomes.

Priori probability: probab. can be known without performing experiment.

$$P(E) = \frac{m}{n} = \frac{\text{no. of fav. outcomes}}{\text{total no. of outcome}}$$

Posteriori probability:

$$P(E) = \frac{m}{n} = \frac{\text{no. of time E occurs}}{\text{total no of trial}}$$

Chapter 1

DESCRIPTIVE STATISTICS AND BASIC PROBABILITY

1.1 Introduction to Statistics and Its Importance in Engineering 2
1.2 Describing Data with Graphs (Bar, Pie, Line Diagram and Box Plot) ... 2
1.2.1 Bar Graph 2
1.2.2 Pie Graph 3
1.2.3 Line Diagram 3
1.2.4 Box Plot 3
1.3 Describing Data with Numerical Measure (Measuring Center, Meaning Variability) 5
1.3.1 Measure of Central Tangency 5
1.3.2 Measure of Dispersion or Variability 7
1.3.3 Difference between Measure of Central Tendency and Measure of Dispersion 9
1.4 Basic Probability, Additive Law, Multiplicative Law, Bayer's Theorem 9
1.4.1 Mutually Exclusive Events 9
1.4.2 Equally Likely Events 9
1.4.3 Independent Events 9
1.4.4 Dependent Events 10
1.4.5 Axiomatic or Modern Approach of Probability 10
1.4.6 Addition Theorem of Probability or Theorem of Total Probability 10
1.4.7 Multiplication Law of Probability 10
1.4.8 Conditional Probability 11
1.4.9 Baye's Theorem 11

Statistics is a branch of which deals with collection, presentation, analysis and interpretation of numerical data.

1.1 INTRODUCTION TO STATISTICS AND ITS IMPORTANCE IN ENGINEERING

Statistics: A statistics or sample statistics is a single measure of some attribute of a sample. (*For example,* Its arithmetic mean value). It is calculated by applying a function (statistical algorithm) to the values of the items of the sample, which are known together as a set of data.

Its importance in engineering

Statistics has various engineering applications in testing materials, performance test of systems, robotics and automation in general, control of production process and so on. It has roles in different fields of engineering such as:

i) Design of experiments uses statistical techniques to test and contrast models of engineering components and systems.

ii) Quality control and process control use statistics as a tool to manage conformance to specifications of manufacturing process and their products.

iii) Time and methods engineering use statistics to study repetitive operations in manufacturing in order to set standards and find optimum (in some sense) manufacturing procedures.

iv) Reliability engineering uses statistics to measure the ability of system to perform for its intended function (and time) and has tools for improving performance.

v) Probabilistic design uses statistics in the use of probability in product and system design.

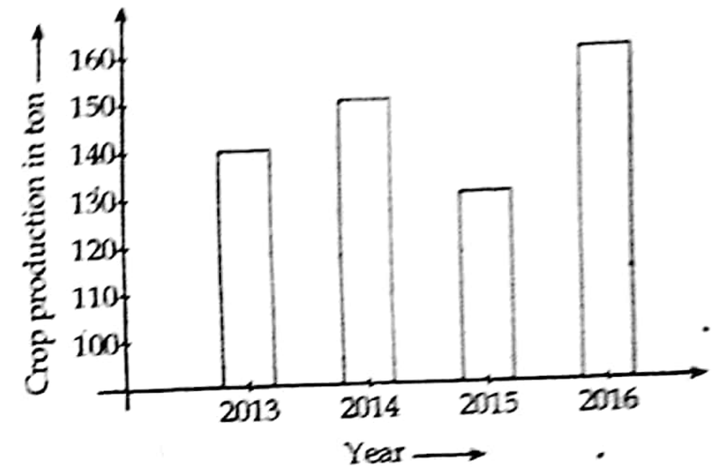
1.2 DESCRIBING DATA WITH GRAPHS (BAR, PIE, LINE DIAGRAM AND BOX PLOT)

The data can be described with the help of graph like bar graph, pie graph, line diagram and box plot daily and the conclusion can be drawn through visualization.

1.2.1 BAR GRAPH

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.

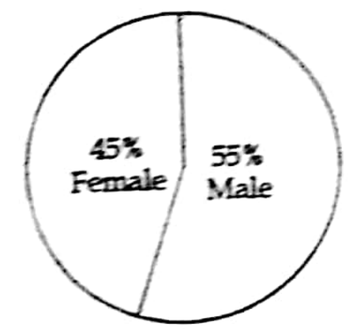
For example;



1.2.2 PIE GRAPH

A pie graph is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie graph, the arc length of each slice is proportional to the quantity it represents.

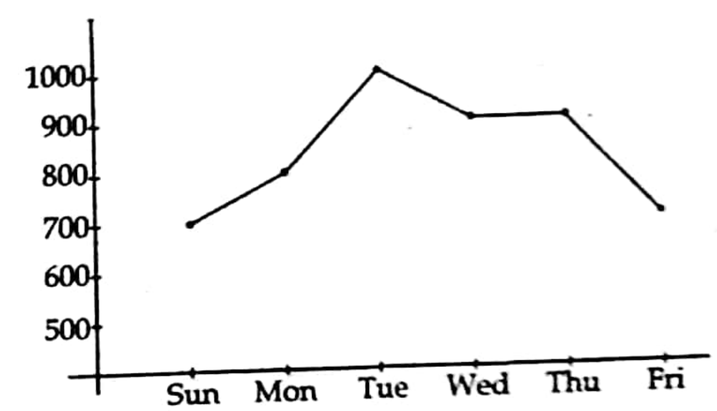
For example; Total population in a city.



1.2.3 LINE DIAGRAM

A line diagram is a type of diagram which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of diagram common in many fields.

For example;

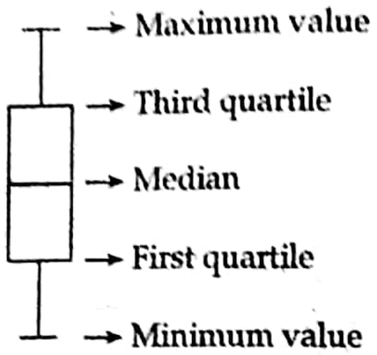


1.2.4 BOX PLOT

A box plot is a graphical rendition of statistics data based on the minimum, first quartile, median, third quartile and maximum (*i.e.*, five number summary) on a rectangular box, aligned either horizontally or vertically.

A box plot is used to summarize a set of data measured on an interval scale and often used in explanatory data analysis. This type of graph is

used to show the shape of line distribution, its central value, and its variability.



In a typical box plot,

- Top of the rectangle indicates the third quartile.
- A horizontal line near the middle of the rectangle indicates the median.
- The bottom of the rectangle indicates the first quartile.
- A vertical line extending from the top of the rectangle indicates maximum value.
- A vertical line extending from the bottom of the rectangle indicates minimum value as shown in the above figure.

Example 1

Contrast a box plot from the following data of marks of students as:

Marks (x)	10-20	20-30	30-40	40-50	50-60
No. of students (f)	2	6	22	13	7

Solution:

Construct the given table with required information as follow:

x	m	f	c.f.
10-20	15	5	5
20-30	25	6	8
30-40	35	22	30
40-50	45	13	43
50-60	55	7	50

Now, position of $Q_1 = \left(\frac{N}{4}\right)^{th} = \frac{50}{4} = 12.5$

c.f. just greater than 12.5 is 30 and corresponding of 30 is 30-40.

$$\therefore \quad Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} \times h = 30 + \frac{12.5 - 8}{22} \times 10 = 32.05$$

Therefore, position of Q_2 or median $= 2\left(\frac{N}{4}\right)^{th} = 25$

of just greater than 25 is 30 and corresponding of 30 is 30-40 class.

$$\therefore \quad \text{Median} = L + \frac{\frac{2N}{4} - c.f.}{f} \times h = 30 + \frac{25 - 8}{22} \times 10 = 37.72$$

$\therefore$ Position of $Q_3 = 3\left(\frac{N}{4}\right)^{th} = 37.5$

The c.f. just greater than 37.5 is 43 and its corresponding class is 40–50.

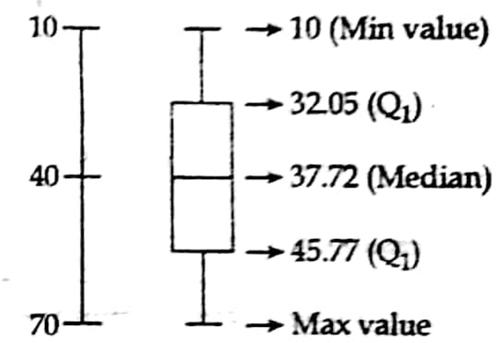
$$\therefore \quad Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} \times h$$

$$= 40 + \frac{37.5 - 30}{13} \times 10$$

$$= 45.77$$

The minimum value is 10 and the maximum value is 59.

The required box plot is given figure.



1.3 DESCRIBING DATA WITH NUMERICAL MEASURE (MEASURING CENTER, MEANING VARIABILITY)

1.3.1 MEASURE OF CENTRAL TANGENCY

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measure of central tendency are sometimes called measure of central location. They are also called as summary statistics. It lies in the middle part of distribution.

Requisites of a good measure of central tendency

i) It should be rigidly defined.
ii) It should be based on all the observations.
iii) It should be suitable for further mathematical treatment.
iv) It should be affected as little as possible by fluctuations of sampling.
v) It should not be affected much by extreme observations.

Different measures of central tendency are

a) Mean
b) Median
c) Mode

a) Mean

The mean is equal to the sum of all the values in the data set dividend by the number of values in the data set.

Mathematically,

$$\bar{x} = \frac{\Sigma x}{N} \qquad \bar{x} = \frac{\Sigma fx}{N}$$

b) Median

The median is the middle score for a set of data that has been arranged in order of magnitude.

Individual series:

$$\text{Median } (M_d) = \text{Values of } \left(\frac{n+1}{2}\right)^{th} \text{ item, when n is odd.}$$

$$\text{Median } (M_d) = \text{Values of } \frac{\left(\frac{n}{2}\right)^{th} \text{ item} + \left(\frac{n+1}{2}\right)^{th} \text{ item}}{2}, \text{ when n is even.}$$

Discrete series:

$$\text{Median } (M_d) = \text{Values of } \left(\frac{N+1}{2}\right)^{th} \text{ item}$$

where, $N = \Sigma f = f_1 + f_2 + f_3 + \cdots\cdots + f_n$.

Continuous series:

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

where, $N = \Sigma f = f_1 + f_2 + f_3 + \cdots\cdots + f_n$

l = Lower limit of the corresponding median class

f = Frequency of the corresponding median class

h = difference of class interval of median class

c.f. = Cumulative frequency of the class preceding median class

c) Mode

The mode is the most frequent score in our data set.

Individual series: Mode is the value of number, whose repetition is maximum.

Discrete series: Mode is the value of variable whose frequency is maximum.

Continuous series: In this series, the class corresponding to the maximum frequency is called the modal class and the mode is defined by the following formula:

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

where, l = Lower limit of modal class

$\Delta_1 = f_1 - f_0$

$\Delta_2 = f_1 - f_2$

f_1 = Maximum frequency

f_0 = Preceding frequency of the modal class

f_2 = Succeeding frequency of the modal class

h = difference of class interval of modal class

mod = 3 median - 2 mean
empirical.

1.3.2 MEASURE OF DISPERSION OR VARIABILITY

Simplest meaning of 'dispersion' is a lack of uniformity in the sizes or quantities of the items of a group or series. According to Reiglemen, "Dispersion is the extent to which the magnitude or quantities of the items differ, the degree of diversity."

Measure of dispersion measure how spread out a set of data is requisites of a good measure of dispersion.

i) It should be simple to understand.
ii) It should be rigidly defined.
iii) It should be based on each individual item of the distribution.
iv) It should be suitable for further mathematical treatment.
v) It should not be affected much by the extreme items.

There are two kinds of measures of dispersion

a) Absolute measures of dispersion
b) Relative measures of dispersion

Objective:
→ determine reliability of an avg.
→ comparative study of two or more series.
→ identify the cause of variability to control it.
→ helps in using other stat. tool

a) Absolute Measures of Dispersion

The measures of dispersion which are expressed in terms of original units of a data are termed as absolute measures. They are expressed in the same units in which the original data are expressed. *For example*, if the series is expressed as marks of the students in particular subject the absolute dispersion will provided the value in marks. The only difficulty is that if two or more series are expressed in different units, the series cannot be compared on the basis of dispersion.

Different absolute measures of dispersion are:

i) Range
ii) Quartile deviation
iii) Mean deviation
iv) Standard deviation

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

b) Relative Measures of Dispersion

Relative or coefficient of dispersion is the ratio or the percentage of a measure of absolute dispersion to an appropriate average. The basic advantage of this measure is that two or more series can be compared with each other despite the fact they are expressed in different units. Theoretically, 'Absolute measure' of dispersion is better. But from a practical point of view, relative or coefficient of dispersion is considered better as it is used to make comparison between series.

Different measure of dispersion are:

i) Coefficient of range
ii) Coefficient of quartile deviation
iii) Coefficient of mean deviation
iv) Coefficient of variation

i) Range = Largest measurement – Smallest measurement
= L – S

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

ii) Quartile deviation (Q.D.):

Inter – Quartile range = $Q_3 - Q_1$

$$Q.D. = \frac{(Q_3 - M_d) + (M_d - Q_1)}{2}$$

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

$$\therefore \quad \text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

iii) Mean Deviation (M.D.):

$$\text{M.D. from Mean} = \frac{\Sigma|X - \mu|}{N}$$

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

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

$$\therefore \quad \text{Coefficient of M.D. from Mean} = \frac{\text{M.D. from Mean}}{\text{Mean}}$$

$$\therefore \quad \text{Coefficient of M.D. from Median} = \frac{\text{M.D. from Median}}{\text{Median}}$$

$$\therefore \quad \text{Coefficient of M.D. from Mode} = \frac{\text{M.D. from Mode}}{\text{Mode}}$$

iv) Standard Deviation (S.D.): (Positive sq. root of variance)

Individual Series

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

Discrete and continuous series

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

Here, in continuous series X is the mid-value of class.

Note: Variance = σ^2

combined S.D → pg 71

Importance of S.D:
1) helps us to determine where the values of the freq. is located relative to mean.
2) great importance in sampling method.
3) ascertain the area under normal curve.
4) testing hypothesis which other measures hardly does.

1.3.3 DIFFERENCE BETWEEN MEASURE OF CENTRAL TENDENCY AND MEASURE OF DISPERSION

Measure of central tendency	Measure of dispersion
i) It is a single value that attempts to describe a set of data by identifying the central position within that set of data.	i) It measures the extent to which the magnitudes or quantities of the items differ, the degree of diversity.
ii) It lies in the middle part of the distribution.	ii) Small dispersion indicates high uniformity of the items.
iii) Mean, mode, median are measures of central tendency.	iii) Range, quartile deviation, mean deviation, standard deviation are measures of dispersion.

1.4 BASIC PROBABILITY, ADDITIVE LAW, MULTIPLICATIVE LAW, BAYER'S THEOREM

1.4.1 MUTUALLY EXCLUSIVE EVENTS

Two events are said to be mutually exclusive events when both cannot occur at the same time. Mutually exclusive events always have a different outcome. Such events are so that when one happens it prevents the second from happening.

For example,

i) If a coin is tossed, either it gives you a "Head" or it gives you a "Tail".
ii) Similarly, in a die role experiment, all the events 1, 2, 3, 4, 5, 6 are mutually exclusive.

1.4.2 EQUALLY LIKELY EVENTS

Two or more events are said to be equally likely events if all the events have the same theoretical probability (or likelihood) of occurring.

For example,

When we toss a win, the probability of getting a head is $\frac{1}{2}$ and the probability of getting a tail is $\frac{1}{2}$. So, it is an equally likely event.

1.4.3 INDEPENDENT EVENTS

Two events are said to be independent event if the result of the second event is not affected by the result of first event and the result of first event is unaffected by the result of second event. *i.e.*, either of the events can occur without being affected by the other.

For example;

If two coins are tossed together, then the occurrence of head or tail on

second coin is not affected and does not affect the occurrence of head or tail on first win.

1.4.4 DEPENDENT EVENTS

Two events are said to be dependent events if the result of first event affects the outcome of the second events so that the probability is changed.

For example;

In a deck of cards, if second cad is drawn without replacement of the first card, then the outcome of the second draw is dependent on the result of first draw.

(v) Extended axiom of prob = $\sum_{i=1}^{\infty} P(A) = \sum_{i=1}^{\infty} P(B)$

1.4.5 AXIOMATIC OR MODERN APPROACH OF PROBABILITY

Axiomatic approach of probability is just another way of describing the probability of an event that includes both *priors* and *posteriori* approaches. In this approach some axioms are predefined before assigning probabilities to quantise the event and hence to ease the calculation of occurrence or non-occurrence of the event.

Probability is a set function P(E) that assigns to every event E a number called the "Probability of E" such that

i) The probability of an event is greater than or equal to zero.

$P(E) \geq 0$ } Prob. of minimum certainty

ii) The probability of sample space is one

$P(\Omega) = 1$. Prob. of certainty.

(iii) Probab. of countable additivity $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

1.4.6 ADDITION THEOREM OF PROBABILITY OR THEOREM OF TOTAL PROBABILITY

Case I: When events are mutually exclusive

If event A and event B are mutually exclusive then probability of either event A or B occurring is given by

$$P(A \text{ or } B)\ P(A \cup B) = P(A) + P(B)$$

Case II: When events are not mutually exclusive

If event A and event B are not mutually exclusive then probability of occurrence of either A or B or both is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

1.4.7 MULTIPLICATION LAW OF PROBABILITY

Case I: When events are independent. If event

If event A and event B are independent then probability of event A and event B occurring together is

$$P(A \text{ and } B) = P(A \cap B) = P(A) + P(B)$$

Case II: When events are dependent

If event A and event B are dependent then probability of event A and event B occurring is

$$P(A \cap B) = P(A)\ P(B/A)$$
$$P(A \cap B) = P(B)\ P(A/B)$$
$$P(A/B),\ P(B/A) = \text{Conditional probability}$$

1.4.8 CONDITIONAL PROBABILITY

The conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. If events A and B are dependent then the conditional probability of event A given that the event B has already occurred denoted by P(A/B) is defined as;

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad \text{where, } P(B) \neq 0.$$

Similarly, if events A and B are dependent then the conditional probability of event B given that the event A has already occurred denoted by P(B/A) is defined as

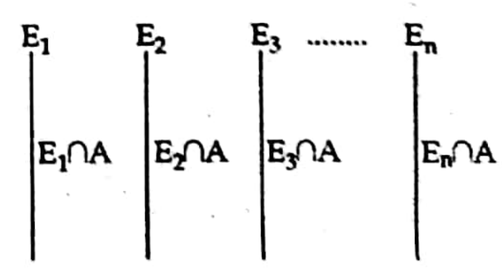
$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \quad \text{where, } P(A) \neq 0.$$

1.4.9 BAYE'S THEOREM

Statement: Let, $E_1, E_2, E_3, \ldots\ldots, E_n$ be n mutually exclusive and exhausted events. Then, they together form a sample space such that $P(E) \neq 0$. Let A be any arbitrary event which is subject $< \bigcup_{i=1}^{n} E_i$ and $P(A) > 0$ then

$$\therefore \quad P(A) = P(E_1 \cap A) + P(E_2 \cap A) + \ldots\ldots + P(E_n)\ P(E_n \cap A)$$
$$= P(E_1) \cdot P(A/E_1) + P(E_2) \cdot P(A \cap E_2) + \ldots\ldots + P(E_n) \cdot P(A/E_n)$$
$$= \sum_{i=1}^{n} P(E_i)\ P(A/E_i)$$

$$\therefore \quad P(E_2/A) = \frac{P(E_2 \cap A)}{P(A)} \qquad [\because \text{By conditional probability}]$$
$$= \frac{P(E_2) \times P(A/E_2)}{\sum_{i=1}^{n} P(E_i) \times P(A/E_i)}$$



SOLVED QUESTIONS (FOR BCE/BME/BIE/BAM)

1. What are difference between measure of central tendency and measure of dispersion? The mean of marks obtained in an exam by a group of 100 students was found to be 49.96. The mean of the marks obtained in the same exam by another group of 200 students was 52.32. Find the mean of the mark obtained by both the group of the students taken together. [2070 Math]

Solution: Please, see the 1.3.3.

Given that;

For 1st group
$n_1 = 100$
$\bar{x}_1 = 49.96$

For 2nd group
$n_2 = 200$
$\bar{x}_2 = 52.32$

Now, mean of both group $(\bar{x}) = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \dfrac{100 \times 49.96 + 200 \times 52.32}{100 + 200}$

$= 51.53$

2. Define prior and posterior probabilities. A company uses thre different assembly lines A_1, A_2 and A_3 to manufacture a particula component. Of these manufactured by line A_1, 5% need rework t remedy a defect, where as 8% of A_2's components need rework an 10% of A_3's need rework. Suppose that 50%, of all components an produced by line A_1, 30% are produced by line A_2 and 20% com from line A_2. If a randomly selected component need rework, wha is the probability that it come from line A_1? [2070 Magh

Solution:

The prior probability distribution that represent your uncertainty over before you have sampled any data and attempted to estimate it. It i usually denoted $\pi(\theta)$.

Whereas, the posterior is a probability distribution representing you uncertainty over θ after you have sampled data-denoted by $\pi(\theta/X)$. It is conditional distribution because it conditions on the observed data.

2nd part

Given that;

$P(A_1) = 50\% = 0.5$
$P(A_2) = 30\% = 0.3$
$P(A_3) = 20\% = 0.2$

Let, $P(D/A_1)$, $P(D/A_2)$ and $P(D/A_3)$ denote the probability of componen that need rework manufactured by A_1, A_2 and A_3 assembly line respectively.

$P(D/A_1) = 5\% = 0.05$
$P(D/A_2) = 8\% = 0.08$
$P(D/A_3) = 10\% = 0.10$

Now, using Baye's Theorem

Therefore, probability that the component selected need rework is from line A_1 is

$$P(A_1/D) = \frac{P(A_1) \times P(D/A_1)}{\sum_{i=1}^{3} P(A_1)\, P(D/A_1)} = \frac{0.5 \times 0.05}{0.5 \times 0.05 + 0.3 \times 0.08 + 0.2 \times 0.10}$$

$= 0.3623$

3. What are the merits and demerits of positional average median? Calculate mean, median and mode from the following data of rainfall. [2070 Bhadra]

Rainfall (in mm)	20-30	30-40	40-50	50-60	60-70
No. of days	15	16	20	24	15

Solution:

Positional average median

Merit

i) It is easy to compute and understand.
ii) It is not affected by extreme valuables and also independent of range or dispersion of the data.
iii) It can be determined graphically.
iv) It is proper average for qualitative data where items are not measured but are scored.

Demerit

i) For computing median data needs to be arranged in ascending or descending order.
ii) It is not based on all the observations of the data.
iii) It cannot be given further algebraic treatment.
iv) It is not accurate when the data is not large.

Given that;

Rain fall (x)	m	No. of dollars (f)	c.f.	m.f.
20-30	25	15	15	375
30-40	35	16	31	560
40-50	45	20	51	900
50-60	55	24	75	1320
60-70	65	15	90	975
		$\Sigma f = 90$		4130

$\therefore$ Mean $(\bar{x}) = \dfrac{\Sigma fm}{\Sigma f} = \dfrac{4130}{90} = 45.89$

Position of median $= \left(\dfrac{N}{2}\right)^{th}$ item $= \dfrac{90}{2} = 45^{th}$ item

c.f. just greater than 45 is 51 and corresponding of 51 is 40-50 class.

$$\therefore \quad \text{Median } (M_d) = L + \frac{\frac{N}{4} - c.f.}{f} \times h = 40 + \frac{45 - 31}{20} \times 10 = 47$$

Mode class is 50-60 since it has highest frequency.

Here, $f_1 = 24$ $\quad f_0 = 20$ $\quad$ and $\quad f_2 = 15$

$\therefore \quad \Delta_1 = f_1 - f_0 = 25 - 20 = 4$

$\therefore \quad \Delta_2 = f_1 - f_2 = 24 - 15 = 9$

$$\therefore \quad \text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = 50 + \frac{4}{4+9} \times 10 = 53.08$$

4. Define addition theorem of probability. In a group of equa number of men and women, 20% of men and 30% of women ar unemployed. if a person selected is random

a) What is the probability that the selected person is an employed?

b) What is the probability that the selected person is me employed? [2070 Bhadra

Solution:

Let x be the total number of men then total number of women will be x.

Now, no. of unemployed men = 20% of x = 0.2x

No. of unemployed women = 30% of x = 0.3x

$\therefore$ No. of employed men = x – 0.2x = 0.8x

$\therefore$ No. of employed women = x – 0.3x = 0.7x

By question

a) Probability that the selected person is an employed

$$= \frac{\text{Total no. of employed person}}{\text{Total no. of person}} = \frac{0.8x + 0.7x}{x + x} = \frac{1.5x}{2x} = 0.75$$

b) Probability that the selected person is men employed

$$= \frac{\text{Total no. of men}}{\text{Total no. of person}} \times \frac{\text{Total no. of employed}}{\text{Total no. of men}} = \frac{x}{2x} \times \frac{0.8x}{x} = 0.4$$

5. What are the chief measures of central tendency? Also describe th techniques to measure the consistency. [2071 Magh

Solution: Please, see the 1.3.1 (mean, median, mode)

Refer 1.3.2. Techniques to measure the consistence (define range, quartil deviation, variance, standard deviation).

6. Five coins are tossed 320 times. If coins are unblased construct th probability distribution table of the result. Also, find mean an variance of the given probability distribution. [2071 Magh

Solution:

No. of coins (n) = 5

No. of times the coin is tossed (N) = 320 times

There are two possible outcome of a win tossed *i.e.*, head or tail with equal probability of $\frac{1}{2}$. Let us consider head is the success. Then, no. of success for 5 coins for once tossed is *i.e.*, r = 0, 1, 2, 3, 4, 5.

Now,

$$P(r) = {}^nC_r\, p^r q^{n-r} = {}^5C_r \left(\frac{1}{2}\right)^r \cdot \left(\frac{1}{2}\right)^{5-r} = {}^5C_r \frac{1}{2^5}$$

$$\therefore \text{Expected frequency} = N \times P(r) = 320 \times {}^5C_r \frac{1}{32} = \frac{320}{32}\, {}^5C_r = 10\, {}^5C_r$$

Probability distribution table

No. of heads	Expected frequency
0	$10 \times {}^5C_0 = 10$
1	$10 \times {}^5C_1 = 50$
2	$10 \times {}^5C_2 = 100$
3	$10 \times {}^5C_3 = 100$
4	$10 \times {}^5C_4 = 50$
5	$10 \times {}^5C_5 = 10$

The expected frequency for no. of tail will be same as no. of head.

$$\therefore \quad \text{Mean} = np = 5 \times \left(\frac{1}{2}\right) = 2.5$$

$$\therefore \quad \text{Variance} = npq = 5 \times \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) = 1.25$$

7. What the differences between measures of central tendency and measures of dispersion? The mean and standard deviation of 20 items is found to be 10 and 2 respectively. At the time of checking it was found that one item 8 was incorrect. Calculate the mean and standard deviation. If a) the wrong item is omitted b) It is replaced by 12. [2071 Bhadra]

Solution: Please, see the 1.3.3.

Given that;

n = 20 items

Mean $(\bar{X}) = 10$

Standard deviation $(\sigma) = 2$

Now,

a) When wrong item is omitted (*i.e.*, wrong item = 8)

n = 19

$$\text{New mean } (\bar{X}) = \frac{10 \times 20 - 8}{19} = 10.10$$

Now standard deviation, $\sigma = \sqrt{\frac{(2^2 + 10^2) \times 20 - 8^2}{19} - (10.10)^2} = 2.023$

$$\left[\because \sigma^2 = \frac{\Sigma X^2}{N} - \bar{X}^2 \text{ or, } \Sigma X^2 = N(\sigma^2 + \bar{X}^2)\right]$$

b) When it is replaced by 12

New mean $= \frac{10 \times 20 - 8 + 12}{20} = 10.2$

$\therefore$ New standard deviation $= \sqrt{\frac{(2^2 + 10^2) \times 20 - 8^2 + 12^2}{20} - (10.2)^2}$

$= \sqrt{108 - 104.04} = 1.99$

8. Difference between equally likely events and mutually exclusive events with suitable example. A consulting firm rents cars form three agencies, 20% from agency D, 20% from agency E and 60% from agency F. It 10% of the cares from D, 12% of the cars from E, and 4% of the cars from F have bad tires, what is the probability that the firm will get a car with bad tires? [2071 Bhadra]

Solution: *Please, see the theory 1.4.1 and 1.4.2.*

Given that;

$P(D) = 20\% = 0.2$

$P(E) = 20\% = 0.2$

$P(F) = 60\% = 0.6$

Let P(B/D), P(B/E) and P(B/F) be the probability that the car has bad tyre from agency D, E and F respectively.

Then,

$P(B/D) = 10\% = 0.1$

$P(B/E) = 12\% = 0.12$

$P(B/F) = 4\% = 0.04$

Now, using Baye's Theorem

Therefore, probability that the film will get a car with bad tires is given as

$\therefore$ $P(B) = P(D)\,P(B/D) + P(E)\,P(B/E) + P(F)\,P(B/F)$

$= 0.2 \times 0.1 + 0.2 \times 0.12 + 0.6 \times 0.04 = 0.068$

9. What are the chef measures of central tendency and measure of dispersion? The mean weight of 100 students in a certain class is 59 kg. The mean weight of boys in the class is 65 kg and that of girls is 50 kg. Find the number of boys and girls in the class. [2072 Magh]

Solution:

The chief measures of central tendency are

i) Mean

ii) Median

iii) Mode

and chief measures of dispersion are

i) Standard deviation

ii) Variance

iii) Mean deviation

iv) Quartile deviation

v) Range

Given, total student (n) = 100

Mean weight $(\bar{X}) = 59$

For boys

Mean weight $(\bar{X}_b) = 65$ kg

Number of boys $(n_b) = ?$

For girls

Mean weight $(\bar{X}_g) = 50$ kg

Number of girls $(n_g) = ?$

We known,

$n_b + n_g = 100$

or, $n_b = 100 - n_g$

Now,

$n \times \bar{X} = n_b \times \bar{X}_b + n_g \times \bar{X}_g$

or, $100 \times 59 = (100 - n_g) \times 65 + n_g \times 50$

$\therefore$ $n_g = 40$

$\therefore$ $n_b = 100 - 400 = 60$

10. A box containing 5000 IC chips of which 1000 are manufactured by company A and the rest by company B. Ten percentage of chips made by company A and five percentage of the chips by company B are defective. It we select a chip at random

i) What is the probability that the chips choosen is defective?

ii) If a randomly choosen chip are found to be defective, what is the probability that it comes from company A? [2072 Magh]

Solution:

Let P(A) and P(B) denotes the probability of chips manufactured by company A and company B respectively. Then,

$$P(A) = \frac{1000}{5000} = \frac{1}{5} = 0.2$$

$$P(B) = \frac{4000}{5000} = \frac{4}{5} = 0.8$$

Let P(E/A) and P(E/B) denotes the probability that the chips are defective and is manufactured by company A and B respectively.

Then,

$P(E/A) = 10\% = 0.1$

$P(E/B) = 5\% = 0.05$

Now,

i) Probability that the chips choosen is defective is

$$P(E) = P(A) \times P(E/A) + P(B) \times P(E/B)$$
$$= 0.2 \times 0.1 + 0.8 \times 0.05$$
$$= 0.06$$

ii) Probability that the choosen chip is defective and come from company A is

$$P(A/E) = \frac{P(A) \times P(E/A)}{P(E)} = \frac{0.2 \times 0.1}{0.06} = 0.333$$

11. Mention and discuss the significance of statistic in engineering. In a moderately asymmetrical distribution the value of mean and median are 20 and 24 respectively. Find the values of mode.

[2072 Ashwin]

Solution: Please, see the 1.1.

Given, in moderately asymmetrical distribution

Mean $(\bar{X}) = 20$

Median $(M_d) = 24$

Now, we known mode $= 3$ median $- 2$ mean $= 3 \times 24 - 2 \times 20 = 32$

12. Define conditional probability. A manufacturing firm produces steel pipes in three plants with daily production volumes of 500, 1000 and 2000 units respectively. According to experience, it is known that the fraction of defective it is known that the fraction of defective output produced by three plants are respectively 0.005, 0.008, 0.010. If a pipe is selected from a day's total production and found to be defective, find out the probability that the pipe might have come from plant II and III.

[2072 Ashwin]

Solution: Please, see the 1.4.8.

Let A, B and C denotes the probability of steel pipes manufactured by plant I, II and III respectively. Then,

$$P(A) = \frac{500}{500 + 1000 + 2000} = \frac{500}{3500} = \frac{1}{7}$$

$$P(B) = \frac{1000}{3500} = \frac{2}{7}$$

$$P(C) = \frac{2000}{3500} = \frac{4}{7}$$

Let E denotes the defective output. Then,

$P(E/A) = 0.005$

$P(E/B) = 0.008$

$P(E/C) = 0.010$

Now, using Baye's theorem: Probability that the pipe is defective and come from plant II and III is given as

$$P(B/E) + P(C/E) = \frac{P(B) \times P(E/B) + P(C) \times P(E/C)}{P(A) \times P(E/A) + P(B) \times P(E/B) + P(C) \times P(E/C)}$$

$$= \frac{\frac{2}{7} \times 0.008 + \frac{4}{7} \times 0.010}{\frac{1}{7} \times 0.005 + \frac{2}{7} \times 0.008 + \frac{4}{7} \times 0.010} = 0.9180$$

13. Write the significance difference between measure of central tendency and measure of dispersion.

In a study conducted by the department of mechanical engineering at Virginia tech, the steel rods supplied by two different companies were compared. Ten sample spring were made out of the steel rods supplied by each company, and a measure of flexibility was recorded for each. The data are as follows

Company A: 9.3, 8.8, 6.8, 8.7, 8.5, 6.7, 8.0, 6.5, 9.2, 7.0

Company B: 11.0, 9.8, 9.9, 10.2, 10.1, 9.7, 11.0, 11.1, 10.2, 9.6

a) Calculate the sample mean and median for the data for the two companies.

b) Calculate the sample standard deviation.

c) Which company is more consistence?

[2073 Magh]

Solution: Please, see the 1.3.3.

Company A: 9.3, 8.8, 6.8, 8.7, 8.5, 6.7, 8.0, 6.5, 9.2, 7.0

Company B: 11.0, 9.8, 9.9, 10.2, 10.1, 9.7, 11.0, 11.1, 10.2, 9.6

Now,

a) Sample mean of company $A(\mu_A) = \frac{\Sigma X}{N} = 7.95$

$$\text{Median } (M_{dA}) = \frac{8.5 + 6.7}{2} = 7.6$$

b) $\mu_B = \frac{\Sigma X}{N} = 10.26$

$$\text{Median } (M_{dB}) = \frac{10.1 + 9.7}{2} \quad 9.9$$

b) $\sigma_A = \sqrt{\frac{\Sigma(x - \mu)^2}{n - 1}} = 1.098$

$\sigma_B = 0.5407$

c) $CV_A = \frac{\sigma_A}{\mu_A} \times 100\% = \frac{1.098}{7.95} \times 100\% = 13.81\%$

$$CV_B = \frac{\sigma_B}{\mu_B} \times 100\% = \frac{0.5407}{10.26} \times 100\% = 5.27\%$$

Hence, company B is more consistence since $CV_B < CV_A$.

14. Define the Bay's rule of probability and its application in engineering? If the probabilities are respectively 0.09, 0.15, 0.21 and 0.23 that a person
   a) Verify that f(x) is a density function
   b) Find $P(0 \le X \le 1)$ [2073 Magh]

Solution: Please, see the 1.4.9.

15. What are graph an diagrams? What are the importance and limitation of diagrammatic and graphical representation of statistical data? [2073 Bhadra]

Solution:

A graph usually means an X–Y plot, using Cartesian co-ordinates.

A diagram means any visual representation that is intended to explain or show a logical idea-categories, relationships, cause and effect, quantities relationships.

Importance

i) Easy to understand and summarize the information.
ii) Easy to compare different information.
iii) Decision making at a glance which is hardly possible through descriptive report.
iv) A complete idea can be easily and quickly generated.

Limitations

i) There is a loss of accuracy of data while representing data through diagram and graph. It is obvious that there will be loss of data as it is the summarization of the whole data.
ii) Sometime the illusionary effect creates a wrong impression on the mind of viewer.
iii) It does not explain the whole information's since it focus on major issues and ideas.
iv) Graphical representation consume more time since it involves graph and figures which need to be plotted precisely.
v) It is costly since it involves images, colours and paints.

16. Define axiomatic approach of probability. In a bolt factory, machine A, B and C manufacture 60%, 25%, and 15% respectively of the total of their output 1%, 2% and 1% are defective bolts. A bolt is drawn at random from the total production and found to be defective. From which machines the defective bolt is expect to have been manufactured? [2073 Bhadra]

Solution: Please, see the 1.4.5.

Given that:

$P(A) = 60\% = 0.6$
$P(B) = 25\% = 0.25$
$P(C) = 15\% = 0.15$

Let E denotes the defective bolt. Then,

$P(E/A) = 1\% = 0.01$
$P(E/B) = 2\% = 0.02$
$P(E/C) = 1\% = 0.01$

Now, using Baye's theorem

$$\therefore \quad P(A/E) = \frac{P(A) \times P(E/A)}{P(A) \times P(E/A) + P(B) \times P(E/B) + P(C) \times P(E/C)}$$

$$= \frac{0.6 \times 0.01}{0.6 \times 0.01 + 0.25 \times 0.02 + 0.15 \times 0.01} = \frac{0.006}{0.0125} = 0.48$$

$$\therefore \quad P(B/E) = \frac{0.25 \times 0.02}{0.0125} = 0.4$$

$$\therefore \quad P(C/E) = \frac{0.15 \times 0.01}{0.0125} = 0.12$$

Since, P(A/E) is highest, so it is more probable that the defective bolt is manufactured by machine A.

17. **We are the merits and demerits of mean? The expenditure of 1000 families is given as below**

Expenditure	40-59	60-79	80-99	100-119	120-139
No. of families	50	–	500	–	50

If the median of frequency distribution is 87, find the missing frequencies. [2074 Bhadra]

Solution: Merits of mean

i) Mean is rigidly defined by algebraic formula.
ii) It is based on all observations and it can be regarded as representative of given data.
iii) It is least affected by the fluctuation of sampling.

Demerits of mean

i) It can neither be determined by inspection or by graphical location.
ii) It cannot be computed for qualitative data like data on intelligence, honestly and smoking habit, etc.
iii) It cannot be computed when class intervals have open ends.

Converting inclusive classes into exclusive classes as given below

Expenditure (x)	No. of families (f)	c.f.
39.5–59.5	50	50
59.5–79.5	y(say)	50 + y
79.5–99.5	500	550 + y
99.5–119.5	400 – y	950
119.5–139.5	50	1000
	$\Sigma f = 1000$	

Now, we know

Median (M_d) = 87

So, the corresponding class is (79.5–99.5)

We have

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

or, $87 = 79.5 + \frac{500 - (y + 50)}{500} \times 20$

or, $y = 262.5$

$\therefore$ $y \approx 263$

and $400 - y = 400 - 263 = 137$ families

Hence, missing frequencies are 263 and 137.

18. Define conditional probability. Two sets of candidates are competing for the positions on the board of directors of a company. The probability that the first and second sets will coin are 0.6 and 0.4 respectively. If the first set wins the probability of introducing a new product is 0.8 of the corresponding probability if the second set wins is 0.3. What is the probability that the product will be introduced? [2074 Bhadra]

Solution: Please, see the 1.4.8.

Given that;

Let A and B be the two sets. Then;

$P(A) = 0.6$

$P(B) = 0.4$

Let E denoted that the new product will be introduced. Then,

$P(E/A) = 0.8$

$P(E/B) = 0.3$

Using Baye's theorem

Probability that product will be introduced

$P(E) = P(A) \times P(E/A) + P(B) \times P(E/B)$

$= 0.6 \times 0.8 + 0.4 \times 0.3$

$= 0.6$

19. We are the merits and demerits of mean? The expenditure of 1000 families is given as below

Expenditure	40-59	60-79	80-99	100-119	120-139
No. of families	50	-	500	-	50

If the median of frequency distribution is 87, find the missing frequencies. [2075 Baishakh]

Solution: Refer Q. No. 17.

20. Define addition theorem of probability. In a group of equal number of men and women, 20% of men and 30% of women are unemployed. if a person selected is random

a) What is the probability that the selected person is an employed?

b) What is the probability that the selected person is men employed? [2075 Baisakh]

Solution: Refer Q.No. 4.

21. What is box plot and what does it measure? Describe the features the a box plot displays. [2075 Bhadra]

Solution: Please, see the 1.2.4.

22. State Baye's theorem. A geologist classified 2015 Nepal earthquake according to intensity as: light, moderate and high. According, to geologist, 90% of earthquakes are of light, 8% are of moderate and 2% of high intensity. An engineer did post earthquake assessment in Sindhupalchowk and found that 1% of damage was done by light, 10% by moderate and 80% by high intensity earthquakes. During assessment an engineer randomly selected a household.

a) What is the probability that the house is damaged?

b) If the house is damaged, what is the probability that it is due to moderate earthquake? [2075 Bhadra]

Solution: Please, see the 1.4.9.

Let L, M, H denote light, moderate and high intensity earthquake respectively. Then,

$P(L) = 90\% = 0.9$

$P(M) = 8\% = 0.08$

$P(H) = 2\% = 0.02$

Let, E denote the house is damaged due to earthquake. Then,

$P(E/L) = 1\% = 0.01$

$P(E/M) = 10\% = 0.1$

$P(E/H) = 80\% = 0.8$

Now, using Baye's theorem

a) Probability that the selected house is damaged

$P(E) = P(L) \times P(E/L) + P(M) \times P(E/M) + P(H) \times P(E/H)$

$= 0.9 \times 0.01 + 0.08 \times 0.1 + 0.02 \times 0.8$

$= 0.033$

b) Probability that the selected house is damaged given that it is due to moderate earthquake

$$P(M/E) = \frac{P(M) \times P(E/M)}{P(E)} = \frac{0.08 \times 0.1}{0.033} = 0.2424$$

SOLVED QUESTIONS (FOR BCT/BEX/BEL/B. Agri)

1. **In statistics paper five candidates obtain the marks as 33, 38, 48, 59 and 72. Calculate the mean and standard deviation of these marks. If 10 marks are added for each student, what will be mean and standard deviation? [2070 Ashad]**

Solution:

Given marks are:

33, 38, 48, 59 and 72.

$\therefore$ Mean $(\mu) = \frac{\Sigma X}{n} = 50$

$\therefore$ Standard deviation $(\sigma) = \sqrt{\frac{\Sigma X^2}{n-1} - \frac{n\mu}{n-1}} = 15.82$

After adding 10 marks to each student. Then, new marks will be

43, 48, 58, 69 and 82

$\therefore$ Mean $(\mu) = \frac{\Sigma X}{n} = 60$

$\therefore$ Standard deviation, $\sigma = \sqrt{\frac{\Sigma X^2}{n-1} - \frac{n\mu}{n-1}} = 15.82$

2. **Distinguish between mutually exclusive and equally likely event with examples. What is the use of Baye's theorem in theory of probability? In a college 45% student belong to civil, 30% electronics and remaining to other faculties. The probability of being top is 5%, 4% and 2% respectively in civil, electronics and others. If this years result is published, what is the probability that the topper is from electronics? [2070 Ashad]**

Solution: *Please see the 1.4.2 and 1.4.3.*

Use of Baye's theorem in theory of probability.

i. It is used to revise previously calculated probabilities based on new informations.

ii. It is an extension of conditional probability.

Let C, E and O denotes the students of civil, electronics and other respectively. Then,

$P(C) = 45\% = 0.45$

$P(E) = 30\% = 0.3$

$P(O) = 25\% = 0.25$

Let, T denotes the student who has topped. Then,

$P(T/C) = 5\% = 0.05$

$P(T/E) = 4\% = 0.04$

$P(T/O) = 2\% = 0.02$

Now, using Baye's theorem

$\therefore$ Probability that the toper is from electrons,

$$P(E/T) = \frac{P(E) \times P(T/E)}{P(C) \times P(T/C) + P(E) \times P(T/E) + P(O) \times P(T/O)}$$

$$= \frac{0.3 \times 0.04}{0.45 \times 0.05 + 0.3 \times 0.04 + 0.25 \times 0.02}$$

$$= 0.3037$$

3. **Calculate the standard deviation from the following data regarding marks obtained by students in a test.**

Marks	1	2	3	4	5	6	7	8	9
No. of students	32	41	57	98	123	83	46	17	3

What will be the value of standard deviation? If the marks obtained by each of the students are increased by one? [2070 Chaitra]

Solution:

Given that;

Marks (X)	f	fX	X^2	fX^2
1	32	32	1	32
2	41	82	4	164
3	57	171	9	513
4	98	392	16	1568
5	123	615	25	3075
6	83	498	36	2988
7	46	322	49	2254
8	17	136	64	1088
9	3	27	81	243
	$\Sigma f = 500$	$\Sigma fX = 2275$	$\Sigma fX^2 = 11925$	

$\therefore$ Standard deviation $(\sigma) = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2}$

$$= \sqrt{\frac{11925}{500} - \left(\frac{2275}{500}\right)^2}$$

$\therefore$ $\sigma = 1.7741$

If the marks obtained by each student is increased by one, then the standard deviation will remain the same.

i.e., $\sigma = 1.7741$

4. **State Baye's theorem. A manufacturer of air conditioning units purchases 70% of its thermostats from company A, 20% from company B and the rest from company C. Past experiences shows that 0.5% of company A's thermostats, 1% of company B's thermostats and 1.5% of company C's thermostat a likely to be**

defective. An air-conditioning unit randomly selected from this manufacture's production line was found to have a defective thermostat. Find the probability that company A supplied the defective thermostat? [2070 Chaitra]

Solution: Please, see the 1.4.9.

Given that;

$P(A) = 70\% = 0.7$

$P(B) = 20\% = 0.2$

$P(C) = 10\% = 0.1$

Let E denotes the defective thermostat's. Then,

$P(E/A) = 0.5\% = 0.005$

$P(E/B) = 1\% = 0.01$

$P(E/C) = 1.5\%\ \ 0.015$

Now, using Baye's theorem

Therefore, probability that the company A provides the defective thermostat's is

$$P(A/E) = \frac{P(A) \times P(E/A)}{P(A) \times P(E/A) + P(B) \times P(E/B) + P(C) \times P(E/C)}$$

$$= \frac{0.7 \times 0.005}{0.7 \times 0.005 + 0.02 \times 0.01 + 0.1 \times 0.015}$$

$$= 0.5$$

5. **Write difference between measure of central tendency and measure of dispersion and their importance. The following table represents the makes of 100 students.**

Marks	0–20	20–40	40–60	60–80	80–100
No. of students	14	?	27	?	15

If the mode value 58, find the missing frequencies and the mean of all 100 students. [2071 Shrawan]

Solution: Please, see the 1.3.3.

Given that;

Marks (X)	No, of students (f)	m	fm
0–20	14	10	140
20–40	y (say)	30	540
40–60	27	50	1350
60–80	44 – y	70	1820
80–100	15	90	1350
			5200

Mode is 58 *i.e.*, it lies in 40–60 class.

$f_1 = 27 \quad f_0 = y \quad f_2 = 44 - y$

$\Delta_1 = f_1 - f_0 = 27 - y$

$\Delta_2 = f_1 - f_2 = 27 - 44 + y = y - 17$

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

or, $58 = 40 + \dfrac{27 - y}{27 - y + y - 17} \times 20$

or, $y = 18$

$\therefore$ $44 - y = 44 - 18 = 26$

Hence, the missing frequencies are 18 and 26.

$\therefore$ $\text{Mean } (\bar{X}) = \dfrac{\Sigma fm}{N} = \dfrac{5200}{100} = 52$

6. **Define multiplication law of probability for dependent and independent events with suitable examples. The independent probabilities that the three sections of a costing department will encounter a computer error 0.2, 0.3 and 0.1 per week respectively. What is the probability that there would be:**

i) At least one computer error per week?

ii) One and only one computer error per week? [2071 Shrawan]

Solution: Please, see the 1.4.7.

Let A, B and C denotes three sections of costing department which encountered computer error of 0.2, 0.3 and 0.1 per week respectively.

i.e., $P(A) = 0.2 \quad \therefore \quad P(\bar{A}) = 1 - 0.2 = 0.8$

$P(B) = 0.3 \quad \therefore \quad P(\bar{B}) = 1 - 0.3 = 0.7$

$P(C) = 0.1 \quad \therefore \quad P(\bar{C}) = 1 - 0.1 = 0.9$

Since the events are independent

i) Probability that at least once computer error per work is

$$P(A \cup B \cup C) = 1 - P(\overline{A \cup B \cup C})$$
$$= 1 - P(\bar{A})\, P(\bar{B})\, P(\bar{C})$$
$$= 1 - 0.8 \times 0.7 \times 0.9$$
$$= 0.496$$

ii) Probability that one and only one computer error per week is

$$P(A \cap \bar{B} \cap \bar{C}) + P(\bar{A} \cap B \cap \bar{C}) + P(\bar{A} \cap \bar{B} \cap C)$$
$$= P(A)\, P(\bar{B})\, P(\bar{C}) + P(\bar{A})\, P(B)\, P(\bar{C}) + P(\bar{A})\, P(\bar{B})\, P(C)$$
$$= 0.2 \times 0.7 \times 0.9 + 0.8 \times 0.3 \times 0.9 + 0.8 \times 0.7 \times 0.1$$
$$= 0.398$$

7. **What is absolute and relative measure of dispersion? Construct a box plot from the following data of marks of students as [2072 Kartik]**

Marks	10–20	20–30	30–40	40–50	50–60
No. of students	2	6	22	13	7

Solution: Please, see the 1.3.2 a) and b)

Please, see the 1.2.4 example 1.

8. **State the law of addition of probability. In a training the 70% of person achieved a rating of satisfactory, 80% had acceptable scores on the personality test. Of these rated as unsatisfactory, 35% has acceptable scores. Find the probability that an applicant would be a satisfactory trainee given the acceptable scores on personality test.** [2072 Kartik]

Solution: Please, see the 1.4.6.

Let S and U denote the rating of satisfactory and unsatisfactory respectively.

Then, P(S) = 0.7

$\therefore$ P(U) = 1 – P(S) = 1 – 0.7 = 0.3

Let A denotes the acceptability

Then, P(A/S) = 0.8

P(A/U) = 0.35

Now, probability that an applicant would be a satisfactory trainee given the acceptable scores on personality test is

$$P(S/A) = \frac{P(S)\,P(A/S)}{P(S)\,P(A/S) + P(U)\,P(A/U)}$$

$$= \frac{0.7 \times 0.8}{0.7 \times 0.8 + 0.3 \times 0.35}$$

$$= \frac{16}{19} = 0.8421$$

9. **Two different of a statistics class take the same quiz and the score are recorded below:**
   a) **Find the range and standard deviation for each section.**
   b) **What do the range values lead you to conclude about the variation in the two sections?**
   c) **Why is the range misleading in this case?**
   d) **What do the standard deviation values lead you to conclude about the variation in two sections?** [2071 Chaitra]

Section 1	1	20	20	20	20	20	20	20	20	20	20
Section 2	2	3	4	5	6	14	15	16	17	18	19

Solution:

Given that;

Section 1	1	20	20	20	20	20	20	20	20	20	20
Section 2	2	3	4	5	6	14	15	16	17	18	19

a) For section 1

Range (R_1) = L – S = 20 – 1 = 19

Standard deviation $(\sigma_1) = \sqrt{\frac{\Sigma X^2}{n-1} - \left(\frac{\Sigma X}{n-1}\right)^2} = 5.728$

For section 2

$\therefore$ Range = L – S = 19 – 2 = 17

$\therefore$ Standard deviation $(\sigma_2) = \sqrt{\frac{\Sigma X^2}{n-1} - \left(\frac{\Sigma X}{n-1}\right)^2} = 6.421$

b) Range in the both sections conclude that the variation in both sections is near about the same. However, verification in first section is less compare to second section.

c) Range is misreading because it's only depend on smallest and target value and does not encounter the intermediate values.

d) Though there is single huge variation in section first and there is a lot variations in section second, the standard deviation is close.

10. **Define multiplication law of probability for dependent and independent events with suitable examples. The independent probabilities that the three sections of a costing department will encounter a computer error 0.2, 0.3 and 0.1 per week respectively. What is the probability that there would be:**
   i) **At least one computer error per week?**
   ii) **One and only one computer error per week?** [2071 Chaitra]

Solution: Please, see the Q. No. 6.

11. **What the differences between measures of central tendency and measures of dispersion? The mean and standard deviation of 20 items is found to be 10 and 2 respectively. At the time of checking it was found that one item 8 was incorrect. Calculate the mean and standard deviation. If a) the wrong item is omitted b) It is replaced by 12.** [2072 Chaitra]

Solution: Please, see the Q. No. 7 of BCE/BME.

12. **Define conditional probability. An assembly plant reserves its voltage regulators from these three different supplies, 60% from supplier A, 30% from supplier B and 10% from supplier C. It is also known that 95% of the voltage regulators from A, 80% of These from B, and 65% these from C perform according to specifications. What is the probability that**
   i) **Anyone voltage regulator received by the plant will perform according to specification.**
   ii) **A voltage regulation that performs according to specifications comes from B.** [2072 Chaitra]

Solution: Please, see the 1.4.8.

Given that;

P(A) = 60% = 0.6

P(B) = 30% = 0.3

P(C) = 10% = 0.1

Let E denotes that the voltage regulations perform according to its specificatio

Then, $P(E/A) = 95\% = 0.95$

$P(E/B) = 80\% = 0.8$

$P(E/C) = 65\% = 0.65$

Now, using Baye's theorem

i) Probability that the voltage regulator perform according to i specification is

$$P(E) = P(A) \times P(E/A) + P(B)\,P(E/B) + P(C)\,P(E/C)$$
$$= 0.6 \times 0.95 + 0.3 \times 0.8 + 0.1 \times 0.65$$
$$= 0.875$$

ii) Probability that the voltage regulator perform according to i specification and come from B is

$$P(B/E) = \frac{P(B) \times P(E/B)}{P(E)} = \frac{0.3 \times 0.8}{0.875} = 0.274$$

13. In two companies A and B engaged in similar type of industry, t average weekly wage and standard deviation are given below:

	Company A	Company B
Average weekly (RS)	460	490
Standard deviation	50	40
No. of wage earners	100	80

i) Which company pays larger amount as weekly wage?

ii) Which company shows grater variability in the distributio weekly wages?

iii) What is the mean and standard deviation of all the workers i two companies taken together? [2073 Shrawa

Solution:

Given that;

	Company A	Company B
Average weekly (RS)	460	490
Standard deviation	50	40
No. of wage earners	100	80

i) For company A

Total wages = 460 × 100

= Rs. 46000

For company B

Total wages = 490 × 80

= Rs. 39200

Hence, company A pays larger amount that company B.

ii) $CV_A = \frac{\sigma_A}{\mu_A} \times 100\%$

$= \frac{50}{460} \times 100\%$

$= 10.87\%$

$CV_B = \frac{\sigma_B}{\mu_B} \times 100\%$

$= \frac{40}{490} \times 100\%$

$= 8.16\%$

Hence, company A shows greater variability since $CV_A > CV_B$ (CV = Coefficient of variation).

iii) Mean together $= \frac{\mu_A n_A + \mu_B n_B}{n_A + n_B} = \frac{460 \times 100 + 490 \times 80}{100 + 80}$

$\therefore \quad \mu_t = 473.363$

Standard deviation together (σ)

$$= \sqrt{\frac{(\sigma_A^2 + \mu_A^2)n_A + (\sigma_B^2 + \mu_B^2)n_B}{n_A + n_B} - (\mu_t)^2}$$

$$= \sqrt{\frac{(50^2 + 460^2)\,100 + (40^2 + 490^2)80}{100 + 80} - (473.33)^2}$$

$= 48.22$

14. State the law of multiplication of probability. An electronics company has an engineering position open. The probability that an applicant is capable is 0.7. Each applicant is given written test and oral examination. A capable applicant passes with probability of 0.9. Find (a) the probability that an applicant passes the test (b) the probability that the applicant is capable given he/she passes the test. [2073 Sharwan]

Solution: Please, see the 1.4.7.

a) Probability that an applicant passes the test is

$= 0.9 \times 0.4 = 0.36$

b) Probability that an applicant is capable given he/she passes the test is

$= 0.7 \times 0.9 = 0.63$

15. Describe the various measures' of central tendency and its application. The following table represents the marks of 100 students.

Marks	0–20	20–40	40–60	60–80	80–100
No. of students	14	18	27	26	15

Find the mean, median and standard deviation of all 100 students. [2073 Chaitra]

Solution: Please, see the 1.3.1.

Marks	No. of students (f)	m	c.f	mf	m^2	fm^2
0–20	14	10	14	140	100	1400
20–40	18	30	32	540	900	16200
40–60	27	50	59	1350	2500	67500
60–80	26	70	85	1820	4900	127400
80–100	15	90	100	1350	8100	121500
	$\Sigma f = 100$			5200		334000

$\therefore$ Mean $= \frac{\Sigma mf}{\Sigma f} = \frac{5200}{100} = 52$

$\therefore$ Standard deviation $(\sigma) = \sqrt{\frac{\Sigma fm^2}{n} - \bar{x}^2} = \sqrt{\frac{334000}{100} - 52^2} = 25.21$

$\therefore$ Position of median $(M_d) = \frac{N^{th}}{2} = \frac{100}{2} = 50^{th}$ item

c.f. just greater than 50 50 is 59 and corresponding of 59 is 40-60.

$\therefore$ $M_d = l + \frac{\frac{N}{4} - c.f.}{f} \times h = 40 + \frac{50 - 32}{27} \times 20 = 53.33$

16. Explain Bay's theorem. A chain of video stores sells three differer brands of DVD players; Of its DVD players sales, 50% are brand (the latest expensive), 30% are brand 2 and 20% are brand 3. Eac manufacturer offers a 1-year warranty on parts and labour. It i known that 25% of brand 1's DVD player's require warranty repa work, whereas the corresponding percentage for brands 2 and 3 ar 20% and 10% respectively.

a) What is the probability that a randomly selected purchase has bought a brand 1 DVD player that will need repair whil under warranty?

b) What is the probability that a randomly selected purchase has a DVD player that will need repair while under warranty

[2073 Chaitr

Solution: Please, see the 1.4.9.

Let A, B and C denotes brand 1, brand 2 and brands respectively.

$P(A) = 50\% = 0.5$

$P(B) = 30\%\ 0.3$

$P(C) = 20\% = 0.2$

Let E denotes 1 year warranty.

Then,

$P(E/A) = 25\% = 0.25$

$P(E/B) = 20\% = 0.2$

$P(E/C) = 10\% = 0.1$

Now, using Baye's theorem

a) $P(A/E) = \frac{P(A) \times P(E/A)}{P(E)} = \frac{0.5 \times 0.25}{0.205} = 0.609$

b) $P(E) = P(A) \times P(E/A) + P(B) \times P(E/B) + P(C) \times P(E/C)$

$= 0.5 \times 0.25 + 0.3 \times 0.2 + 0.2 \times 0.1$

$= 0.205$

17. What is box plot and what does it measure? Explain the meaning of its different parts with diagram. A civil engineering monitors water quality by measuring the amount of suspended solids in a sample of river water. Over 11 week days, he observed 14, 21, 21, 28, 30, 63, 29, 65, 55, 19, 20 suspended solids (parts per million) Find the third quartile and interpret its meaning. [2074 Ashwin]

Solution: Please, see the 1.2.4.

Given that;

14, 21, 21, 28, 30, 63, 29, 65, 55, 19, 20

In ascending order: 14, 19, 20, 21, 21, 28, 29, 30, 55, 63, 65

$\therefore$ Position of $Q_3 = 3\left(\frac{N+1}{4}\right) = 3\left(\frac{11+1}{4}\right) = 3 \times 3 = 9^{th}$ position

$\therefore$ $Q_3 = 55$

This shows that 75% of data have value less than or equal to 55.

18. Write down the difference between the sample space and sample points, dependent and independent events. Urn A contains 2 white, 1 black and 3 red balls. Urn B contain 3 white, 2 black and 4 red balls. Urn C contain 4 white, 3 black and 2 red balls. One urn is choosen at random and 2 balls are drawn. They happen to be red and black. What is the probability that both come from urn B.

[2074 Ashwin]

Solution:

Given that;

Urn A : 2W, 1B, 3R

Urn B : 3W, 2B, 4R

Urn C : 4W, 3B, 2R

where, W → White

B → Black

R → Red

Now, by question;

One run is choosen at random and 2 balls are drawn, they happen to be red and black.

$\therefore$ $P(B) = \frac{1}{3}$

P(R) from urn B $= \frac{4}{3+4+2} = \frac{4}{9}$

P(B) from urn B $= \frac{2}{1+2+3} = \frac{2}{6} = \frac{1}{3}$

Therefore, probability that both red and black balls come from urn B

$= \frac{1}{3} \times \frac{4}{9} \times \frac{1}{3} = \frac{4}{81}$

19. What are measures of central tendency? Write favourable points of each of them. Calculate approximate measures of central tendency from the following data:

Wages in Rs./week	less than 35	35-37	38-40	41-43	over 43
No. of wages earned	14	82	99	18	7

Solution: Refer 1.3.1.

The inclusive table is

Wages in Rs./week	31.5-345	34.5–37.5	37.5-40.5	40.5-43.5	43.5-46.5
No. of wages earned	14	82	99	18	7

Now, find mean, median, mode as in Q. No. 15.

20. From a group of 4 engineers, 3 doctors and 2 statistician a sub-group of 3 has to be made, what is the probability that sub-group consists of

a) One from each profession

b) At least one engineer **[2074 Chaitra]**

Solution:

Let E, D and S denotes engineers, doctors and statistician. Then,

$E = 4, D = 3$ and $S = 2$

$\therefore \quad n = 4 + 3 + 2 = 9$

Now, if a sub-group of 3 members has to be made then

a) Probability that sub-group consists of one from each profession

$$= \frac{{}^4C_1 \times {}^3C_1 \times {}^2C_1}{{}^9C_3} = \frac{2}{7}$$

b) Probability that sub-group consists of atleast one engineer

$$= \frac{{}^4C_1 \times {}^5C_2}{{}^9C_3} + \frac{{}^4C_2 \times {}^5C_1}{{}^9C_3} + \frac{{}^4C_3}{{}^9C_3} = \frac{37}{42} = 0.88$$

21. We are the merits and demerits of mean? The expenditure of 1000 families is given as below

Expenditure	40-59	60-79	80-99	100-119	120-139
No. of families	50	-	500	-	50

If the median of frequency distribution is 87, find the missing frequencies. **[2075 Baisakh]**

Solution: Please, see the Q. No. 17 from BCE/BME.

22. Define addition theorem of probability. In a group of equal number of men and women, 20% of men and 30% of women are unemployed. if a person selected is random

a) What is the probability that the selected person is an employed?

b) What is the probability that the selected person is men employed? **[2075 Baisakh]**

Solution: Please, see the Q. No. 4 from BCE/BME.

Chapter 2

DISCRETE PROBABILITY DISTRIBUTION

2.1 Random Variable 36
2.2 Binomial Distribution 36
2.2.1 Characteristics of Binomial Distribution 37
2.3 Negative Binomial Distribution 39
2.3.1 Properties 39
2.3.2 Differences 40
2.4 Poisson Distribution 41
2.4.1 Major Underlying Assumptions 41
2.4.2 Properties 41
2.4.3 Poisson's Distribution Derivation 43
2.4.4 Conditions for Poisson's Approximation to Bonomial Distribution 43
2.5 Hyper Geometric Distance 41
2.5.1 Properties 44

2.1 RANDOM VARIABLE

A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon. If is a function which maps random process to the actual number

For example;

$$X = \begin{cases} 1 & \text{if a goal is scored} \\ 0 & \text{if no goal} \end{cases}$$

$$X = \begin{cases} 1 & \text{if it is head in coin toss} \\ 0 & \text{if is it tail in coin toss} \end{cases}$$

There are two types of random variable

a) Discrete random variable

b) Continuous random variable

Discrete random variable	Continuous random variable
i) A variable which may take on only a countable number of distinct values.	i) A variable which takes an infinite number of possible values.
ii) Things that you can count is discrete.	ii) Things that you can measure but not by counting.
iii) Example X = {0, 1, 2, 3, ..., 10} X = {Head tail} X = {5, 10, 15} and 10 on	iii) Example X = {amount of rainfall in inches}

Probability Distributions

A probability distribution is a mathematical function that provide the probabilities of occurrence of different possible outcomes in an experiment. For a random variable X we try to find the probabilities of the individual value in the range $R_x = \{x_1, x_2, x_3, ..., x_r\}$ collection of $\{P(X = x_1), P(X = x_2), P(X = x_3), ..., P(X = x_r)\}$ is called probability distribution. The probability of individual value of random variable is expressed as P(x) or P(X = x)

Properties

i) The probability for a particular value or range of value must be between o and 1 *i.e.*, $P(x) \leq = 1$.

ii) The sum of all probabilities for all possible values must be equal to 1. *i.e.*, $\Sigma P(x) = 1$

2.2 BINOMIAL DISTRIBUTION

A random variable X is said to be have binomial distribution if and only if the corresponding probability density function is given by

$$b(x; n) = \binom{n}{x} p^x q^{n-x} \qquad (1)$$

where, x = 0, 1, 2, ..., n

n = No. of trails of experiment

x = No. of success

p = Probability of success

q = 1 - p = Probability of failure

The binomial distribution has only two possible outcomes each with unknown probability. The principle binomial distribution were developed by Jacob Bernoulli. A distribution is a set of Bernouli trails.

The binomial probability distribution defines the probability associated with account of independent binary events. Trails satisfying these assumptions are referred to as Bernouli trials.

i) There are only two possible outcomes for each trail.

ii) The probability of a success is the same for each trail.

iii) There are n trails, where n is a constant

iv) The n trials are independent.

2.2.1 CHARACTERISTICS OF BINOMIAL DISTRIBUTION

i) The mean of binomial distribution is given by

Mean $(\bar{X}) = E(X) = np$

ii) The variance of binomial distribution is given by

Variance (x) $\sigma_x^2 = npq$

Since p and q are numerically less than or equal to 1 so npq < np

i.e., Variance < Mean

iii) It is biparametric distribution as it is characterized by two parameters n and p this means if the value n and p are known then the distribution is known completely.

iv) Since only positive integers values are considered like 0, 1, 2, ..., n so, bionomial distribution is a discrete probability distribution.

v) If follows additive property *i.e.*, let x_1 and x_2 are two independent variables such that $x_1 \sim B(n_1, p)$ and $x_2 \sim B(n_2, p)$ then

$$x_1 + x_2 \sim B(n_1 + n_2, p)$$

Now we are a situation to solve some problems.

Example 1

For a binomial distribution with total no. of trials five and probability of success be 0.4 then find

a) Mean and variance

b) P(X = 0) c) P(X = 3)

d) P(X > 3) e) P(X ≤ 3)

Solution:

Given that;

n = 5, p = 0.4

a) Mean = np = 5 × 0.4 = 2 [∵ q = 1 - p)

Variance = npq = 5 × 0.4 × (1 - 0.4)

= 1.2

b) By binomial distribution

$$P(X) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2, \ldots\ldots, n$$

$$P(X = 0) = \binom{5}{0} (0.4)^0 (0.6)^{5-0} = 1 \times 1 \times 0.077 = 0.077$$

c) $P(X = 3) = \binom{5}{3} (0.4)^3 (0.6)^{5-3}$

$= 10 \times 0.064 \times 0.36$

$= 0.2304$

d) $P(X > 3) = P(4) + P(5)$

$= \binom{5}{4} (0.4)^4 (0.6)^{5-4} + \binom{5}{5} (0.4)^5 (0.6)^{5-5}$

$= 0.0768 + 0.01024$

$= 0.08704$

e) $P(X \le 3) = 1 - [p(4) + p(5))]$

$= 1 - 0.08704$

$= 0.91296$

Example 2

The bank of nay work has recently starts a new credit program for customers. The customers meeting certain requirement can obtain a credit card accepted by participating are a merchant that carry a discount. Pass no. shows that 40% of all applicants for this card are rejected. If 10 applicants are selected, what is the probability that

a) Exactly 7 will be rejected

b) None of them are rejected

c) At least two are rejected

d) Less than 3 are rejected

Solution:

Given that;

Probability of rejected, p = 40%

∴ q = 1 - 40% = 60%

n = 10

Now,

a) $P(X = 7) = \binom{10}{7} p^7 q^{10-7}$

$= \binom{10}{7} (0.4)^7 (0.6)^3$

$= 0.04246$

b) For none of them to be selected x = 0

i.e., $P(X = 0) = \binom{10}{0} (0.4)^0 (0.6)^{10-0} = 0.006$

c) $P(X \ge 2) = 1 - P(X < 2)$

$= 1 - [P(0) + P(1)]$

$= 1 - \left(0.06 + \binom{10}{1} (0.4)^1 (0.6)^{10-1}\right)$

$= 1 - 0.1$

$= 0.9$

d) $P(X < 3) = P(0) + P(1) + P(2)$

$= 0.06 + 0.04 + \binom{10}{2} (0.4)^2 (0.6)^{10-2}$

$= 0.06 + 0.04 + 0.1209$

$= 0.2209$

2.3 NEGATIVE BINOMIAL DISTRIBUTION

A discrete random variable x is said to have negative binomial distribution if and if the corresponding probability density function is given by

$$P(x; r, p) = \binom{x + r - 1}{r - 1} p^r q^x, \; x = 0, 1, 2, \ldots\ldots, \infty$$

where

x = denotes the number of failures preceding the rth success in (x + r) trails, hence x + r = n.

p = Probability of success

r = is the number of success

The distribution basically depends upon the bionomical distribution except the last trial must be success and the number of trial is not fixed in negative binomial distribution. *For example;* Bating score in games in standard deck of cards, replace the card and repeat until two ace are drawn.

2.3.1 PROPERTIES

i) Mean and variance of negative binomial distribution are given as

Mean = $\frac{rq}{p}$ and variance = $\frac{rq}{p^2}$

ii) The number of trials are discrete random variable and number of success is fixed in negative binomial distribution.

iii) If follows additive property for independent random variables.

iv) Negative binomial distribution tends to Poisson's distribution when $r \to \infty$ (infinite) and rp = λ (finite).

2.3.2 DIFFERENCES

Binomial distribution	Negative binomial distribution
i) The number of trials is fixed.	i) The number of success is fixed.
ii) The number of success is random variable	ii) The number of trails is random variables .
iii) *For example;* If a dice is rolled for 10 times, what is the probability that four comes up exactly 4 times?	iii) *For example;* If a dice is repeatedly rolled, what is the probability third time four appear in 10th trials?
iv) If pdf is given as $P(x; n) = \binom{n}{x} p^x q^{n-x}$	iv) Its pdf is given as $P(x; r, p) = \binom{x+r-1}{r-1} p^r q^x$

Now we are in a situation to solve some problems.

Example 3

You are surveying people exerting from a polling booth and asking them if they voted independent. The probability (p) that a person voted independent is 20% what is the probability that 15 people must be asked before you can find 5 people who voted independent?

Solution:

Given that;

Probability that a person voted independent,

$P = 20\% = 0.2$

The number of success, $r = 5$

The total number of people asked, $n = 15$

We know,

$x + r = n$

or, $x + 5 = 15$

$\therefore$ $x = 10$

Now,

$$P(x; r, p) = \binom{x+r-1}{r-1} p^r q^x, \; x = 0, 1, 2, ...$$

$$\therefore \quad P(10; 5, 0.2) = \binom{10+5-1}{5-1} (0.2)^5 (1-0.2)^{10} = 0.034$$

Example 4

The probability of hitting target at any time is 0.2 if a sorter aims at a target, Find the probability that fifth fire is send hit.

Solution:

Given that;

Probability of hitting target (p) = 0.2

Number of success (r) = 2

Number of hit (n) = 5

$x + r = n$

or, $x + 2 = 5$

$\therefore$ $x = 3$

Now,

$$P(x; r, p) = \binom{x+r-1}{r-1} p^r q^x, \; x = 0, 1, 2, 3, 4, 5$$

$$\therefore \quad P(3; 2, 0.2) = \binom{3+2-1}{2-1} (0.2)^2 (1-0.2)^3 = 0.0819$$

2.4 POISSON DISTRIBUTION

A random variable x is said to have Poisson distribution if and only if the corresponding probability density function is given by

$$P(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, ..., \infty$$

where,

$x \rightarrow$ is the number of occurrence of some random event.

$\lambda \rightarrow$ is mean occurrence of random event.

$\lambda = np$, n is the total number of possible cases, p is the probability of success.

Poisson distribution explains the behavior of these discrete random variables whose probability of occurrence of an event is small and the total number of possible cases is very large. Let us consider some example

i) Number of car accident in a town per week.

ii) Number of deaths in a town due to murder.

In case of both examples, we can define them in terms of the discrete random variable *i.e.*, 0, 1, 2, ...

2.4.1 MAJOR UNDERLYING ASSUMPTIONS

i) The occurrence the events are independent *i.e.*, one event occurrence neither diminish or increase the occurrence of another event

ii) The probability of occurrence of a single event within a specified time period is generally proportional to the length of the period.

iii) The probability of occurrence of two or more events is negligible for a very small portion of the time period.

2.4.2 PROPERTIES

i) The mean and variance of Poisson distribution are same *i.e.*, Mean = Variance = x and we know $\lambda = np$.

ii) Poisson distribution is uniparametric distribution as it is by only one parameter x.

iii) Since the variable x take any integer value 0, 1, 2, ..., ∞. So, Poisso distribution is discrete probability distribution.

iv) It follows additive properties but does not follow subtractio property.

Now, we are in a situation to solve some problems.

Example 5

The average number of net work error experienced in a day on LAN distribution with an average 1.4. What is the probability that any give day

a) Zero network errors will occur?

b) Exactly one network error will occur?

c) Atleast two network errors will occur?

Solution:

Given that;

Mean occurrence (λ) = 1.4

Let x be the number of network error occurrence we know,

$$P(x; \lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0, 1, 2, \ldots\ldots$$

Now,

a) $P(0; 1.4) = e^{-1.4}\frac{(1.4)^0}{0!} = 0.2465$

b) $P(1, 1.4) = e^{-1.4}\frac{(1.4)^1}{1!} = 0.3452$

c) $P(X \geq 2; 1.4) = 1 - [P(0; 1.4) + P(1; 1.4)]$
$= 1 - (0.2465 + 0.3452) = 0.4083$

Example 6

If 5% of bulbs manufactured by a company in Kathmandu are defecti find the probability that in a sample of 150 bulbs

i) None is defective

ii) 5 bulbs will be defective.

Solution:

Given that;

Probability that a bulbs defective (p) = 5%

Number of sample (n) = 150 bulbs

Mean occurrence (x) = np = 150 × 0.05 = 7.5

Now, let x be the defective bulb then,

i) $P(X = 0) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-7.5} \times (7.5)^0}{0!} = 0.00055$

ii) $P(X = 5) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-7.5} \times (7.5)^5}{5!} = 0.109$

2.4.3 POISSON'S DISTRIBUTION DERIVATION

From binomial distribution we have

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

Let us assume the following

$n \to \infty$

$p \to 0$

and $np = m$

From equation (i)

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

$$= \frac{n(n-1)(n-2)\ldots\ldots(n-r+1)(n-r)!}{r!\,(n-r)!} p^r q^{n-r}$$

$$= \frac{n^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\left(1-\frac{(r-1)}{n}\right) p^r \left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{(np)^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\left(1-\frac{r-1}{n}\right)\frac{\left(1-\frac{m}{n}\right)^n}{\left(1-\frac{m}{n}\right)^r}$$

Taking limit $n \to \infty$ on both sides.

$$\lim_{n\to\infty} P(x = r) = \frac{m^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\left(1-\frac{r-1}{n}\right)\frac{\left(1-\frac{m}{n}\right)^n}{\left(1-\frac{m}{n}\right)^r}$$

or, $P(x = r) = \frac{m^r}{r!} \cdot 1 \cdot \frac{e^{-m}}{1}$ [∵ r = Finite]

or, $P(x = r) = \frac{e^{-m} m^r}{r!}$ (2)

For r = x and m = λ the equation (2) becomes

∴ $P(X = x) = \frac{e^{-\lambda}\lambda^r}{x!}$ where x = 0, 1, 2, ... Proved.

2.4.4 CONDITIONS FOR POISSON'S APPROXIMATION TO BONOMIAL DISTRIBUTION

From binomial distribution distribution

$$P(x = r) = C(n, r)\, p^r q^{n-r} \quad (1)$$

which works good for fixed number of trails (n) in an experiment and for certain value of P.

But what if 'n' is very large (say $n \to \infty$) and 'P' is sufficiently small (say $P \to 0$) then binomial distribution doesn't work good. To overcome this difficulties, we make following assumptions:

i) n is very large *i.e.*, $n \to \infty$

ii) P is sufficiently small *i.e.*, $P \to 0$

iii) Product np is finite *i.e.*, $np = \lambda$ (say)

Under the above mention assumptions, the probability density functi (pdf) (1) for binomial distribution reduces to

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

where, equation (2) is Poisson's distribution.

2.5 HYPER GEOMETRIC DISTANCE

A discrete random variable x is said to have hyper- geometric distributi if and only if the corresponding probability density function is given by

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, 2, \ldots\ldots,$$

where, x = number of observed success

n = sample size/number of draws

M = total number of items with certain characteristics of populatio

N = population size

Since binomial distribution or Poisson distribution or negative binom distribution can be applied to an experiment where the trials a independent and probability of success remain constant from trial to tri But in an experiment where the probability of success does not rema constant from one trail to another trial then the binomial distributi cannot be applied, in this case we can use hyper geomet distribution (HGD):

2.5.1 PROPERTIES

i) Mean and variance of hyper geometric distribution are given as

$$\text{Mean} = \frac{M}{N} \text{ and variance} = \left(\frac{N-n}{N-1}\right) \times npq$$

ii) It has three parameters n, N and M.

iii) Hypo-geometric distribution tends to follow binomial distributi when $n \to \infty$ and $\frac{M}{N} = P$

iv) The items drawn from sample is not replaced to the sample agai Now, we are in a situation to solve some problems.

Example 7

A deck of cards contours 20 cards: 6 red cards and 14 black cards. 5 cards are drawn randomly without replacement. What is the probability that exactly 4 red cards are drawn?

Solution:

Given that;

Total number of cards (N) = 20

Number of card drawn (n) = 5

Number of red cards drawn (x) = 4

Total number of red cards (M) = 6

Now,

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, 2, \ldots\ldots, n$$

$$\text{Here, } P(4; 5, 6, 20) = \frac{\binom{6}{4}\binom{20-6}{5-4}}{\binom{20}{5}} = \frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}} = 0.0135$$

Example 8

A shipment of 120 burglar alarms contains 5 that are defective. If three of these alarms are randomly selected and shipped to a customer, find the probability that the customer will get one bad unit.

Solution:

Here;

Total number of burglar (N) = 120

Number of burglar drawn (n) = 3

Number of defective burglar drawn (x) = 1

Number of defective burglar (M) = 5

Now,

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, 2, \ldots\ldots, n$$

Here;

$$P(1; 3, 5, 120) = \frac{\binom{5}{1}\binom{120-5}{3-1}}{\binom{120}{3}} = \frac{\binom{5}{1}\binom{115}{2}}{\binom{120}{3}} = 0.1167$$

SOLVED QUESTIONS (FOR BCE/BME/BIE/BAM)

28. What are the differences and similarities between binomial a negative binomial distribution? [2070 Ma

Solution: Please, see the Q. No. 16.

29. Certain mass produced articles which 0.5% are defective, packed in cartoons each containing 100 articles.

i) What proportion of cartons is free from defective articles?

ii) What proportion contains 2 or more defectives?

Solution:

Given that;

Probability that an article is defective (p) = 0.5%

Number of articles in a cartoon (n) = 100

∴ Mean occurrence (λ) = np = 0.005 × 100 = 0.5

Let x be the number of defective article (s). We know,

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{where } x = 0, 1, 2, \ldots\ldots$$

i) $P(x = 0) = \frac{e^{-0.5}(0.5)^0}{0!} = 0.6055$

Proportion of carton's articles are from defect.

ii) $P(x \geq 2) = 1 - P(x < 2) = 1 - \{P(x = 0) + P(x = 1)$

$= 1 - \{0.6065 + 0.3032\} = 0.0903$

30. Write down the difference between a binomial and a negati binomial distribution with an example. [2070 Bhad

Solution: Please, see the 2.3.2.

31. In a certain factory turning out optical lenses, there is a sm change, $\frac{1}{500}$ for any lens to be defective. The lenses are supplied packs of 10 each. What is the probability that a packet will conta

a) No defective lens

b) At least one defective lenses

c) At most two defective lenses. [2070 Ma

Solution: Please, see the Q. No. 21.

32. Define binomial probability distribution with its import characteristics. [2071 Ma

Solution: Please, see the 2.2 and 2.2.1.

33. It has been claimed that in 60% of all solar heat installations utility bill will be reduced by at least one-third. Accordingly w are the probabilities that the utility bill will be reduced by at le one third

a) Four of five installations

b) At least four of five installations?

Solution:

Given that;

p = 60% = 0.6

∴ q = 1 – p = 0.4

n = 5

Now, using Binomial distribution

Let x be the number of installation whose utility bill has reduced by one-third

$$P(x) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2, \ldots\ldots, n$$

a) $P(x = 4) = \binom{5}{4}(0.6)^4(0.4)^1 = 0.2592$

b) $P(x \geq 4) = P(x = 4) + P(x = 5)$

$= 0.2592 + 0.0777 = 0.3369$

34. Define negative binomial distribution. Write the conditions for negative binomial distribution. [2071 Bhadra]

Solution: Please, see the 2.3.

Conditions for negative binomial distribution

i) If the number of success is fixed

ii) If the number of trials is random variables

35. A shipment of 20 digital voice reorders contains 5 that are defective. If 10 of them are randomly choosen for inspection, what is the probability that 2 of the 10 will be defective? Also find the mean and variance of the distribution.

Solution:

Given that;

Total number of recorders (N) = 20

No. of recorder drawn (n) = 10

No. of defective recorder (M) = 5

No. of defective recorder drawn (x) = 2

Now, we know

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad x = 0, 1, 2, \ldots\ldots, n$$

$$\therefore \quad P(x = 2) = \frac{\binom{5}{2}\binom{15}{8}}{\binom{20}{10}} = 0.3482$$

$\therefore$ Mean $= \frac{M}{N} = \frac{5}{20} = \frac{1}{4}$

Variance $= \left(\frac{N-n}{N-1}\right) \times npq = \frac{10}{19} \times 10 \times \frac{1}{4} \times \frac{3}{4} = 0.9868$

36. Define hypergemetric probability distribution with an examp describe the conditions for the binomial approximation hypergeomertic distribution? [2072 Mag

Solution: Please, see the 2.5.

For example; Let us consider a company produce 120 burglar alarms which 5 are defective. If 3 of these alarms are randomly selected, find t probability that 1 is defective.

Here;

$N = 120, n = 3, M = 5, x = 1.$

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, 2, \ldots\ldots, n$$

$$= \frac{\binom{5}{1}\binom{115}{2}}{\binom{120}{3}} = 0.1167$$

Condition for the binomial approximation to hypergeometric distributi are

$$P = \frac{M}{N} \text{ and } n \to \infty$$

In above example;

$$P = \frac{M}{N} = \frac{5}{120} = \frac{1}{24}$$

By binomial distribution

$$P(x = 1) = \binom{n}{x} p^x q^{n-x} = \binom{3}{1}\left(\frac{1}{24}\right)^1\left(\frac{23}{24}\right)^2 = 0.1148 \approx 0.1167$$

37. In a certain factory turning out optical lenses there is a sm chance, $\frac{1}{500}$ for any lens to be defective. The lenses are supplied packets of 10 each. What is the probability that a packet w contain.

i) No defective lens

ii) At least one defective lenses

iii) At most two defective lenses. [2072 Mag

Solution: Please, see the Q. No. 21.

38. What are the characteristics' of hypergeometric probability distribution? How does it differ from binomial probability distribution? [2072 Ashwin]

Solution: Please, see the 2.5.1 and 2.5.

39. It is found that a cricketer hit a century in a game is 25%. He declared that he will retire from cricket after 6th century from now by using negative binomial distribution, find the probability that

i) He will lose none of the game

ii) He will lose only game

iii) He will lose exactly two games

iv) He will lose at least two games. [2072 Ashwin]

Solution:

Given that;

Probability that a cricketer hit a century (p) = 25% = 0.25

Number of success (r) = 6 ; $q = 1 - p = 0.75$

Let x denote the number of failures.

Then, by Poisson's distribution

$$P(x; r, p) = \binom{r+x-1}{r-1} p^r q^r$$

i) $P(x = 0; 6, 0.25) = \binom{0+6-1}{6-1}(0.25)^6(0.75)^0 = 0.0002$

ii) $P(x = 1; 6, 0.25) = \binom{1+6-1}{6-1}(0.25)^6(0.75)^1 = 0.001$

iii) $P(x = 2; 6, 0.25) = \binom{2+6-1}{6-1}(0.25)^6(0.75)^2 = 0.003$

iv) $P(x \geq 2; 6, 0.25) = 1 - P(x < 2; 6, 0.25)$

$= 1\{P(x = 0; 6, 0.25) + P(x = 1; 6, 0.25)\}$

$= 1 - \{0.0002 + 0.001\}$

$= 0.9988$

40. Define the hypergeometric distribution. Discus the condition of hypergeometric distribution. [2073 Magh]

Solution: Please, see the 2.5.

Condition for hypergeometric distribution is when the probability of success does not remain constant from one trial to another trial in an experiment.

41. During a laboratory experiment, the average number of radioactive particles passing through a counter in a millisecond is 4. What is the probability that 6 particles enter the counter in a given millisecond?

OR,

It is conjectured that an impurity exists in 30% of all drinking wells in a certain retail community. In order to gain some insight

in to the true extent of the problem. Its determined that some testing is necessary is is too expensive to test all of the wells in the area. So, to are randomly selected for testing.

a) Using the binomial distribution, what is the probability that exactly 3 wells have the impurity, assuming that the conjecture is correct?

b) What is the probability that more than 3 well impure? [2073 Magh]

Solution:

Given that;

Mean occurrence (λ) = 4

Let x be the number of particle (s) entering in the counter per millisecond.

Then, using Poisson's distribution

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{where } x = 0, 1, 2, \ldots\ldots$$

$$\therefore \quad P(x = 6) = \frac{e^{-4}(4)^6}{6!} = 0.1041$$

OR

Given that;

$p = 30\ \% = 0.3$

$\therefore \quad q = 1 - p = 0.7$

$n = 10$

Let x be the number of well (s) having the impurity.

Then, using binomial distribution

$$P(x = 1) = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2, \ldots\ldots, n$$

a) $P(x = 3) = \binom{10}{3}(0.3)^3(0.7)^7 = 0.2668$

b) $P(x > 3) = 1 - P(x \le 3)$

$= 1 - \{P(0) + P(1) + P(2) + P(3)\}$

$= 1 - \{0.0282 + 0.1210 + 0.2334 + 0.2668\}$

$= 1 - 0.6494$

$= 0.3506$

42. What are the chief characteristics of negative binomial distribution? How does it differ from binomial distribution? [2073 Bhadra]

Solution: Please, see the 2.3 and 2.3.1.

43. Define poisons distribution. A survey report shows that the probability 0.001 that a car will have a flat tyre while driving ever a certain bridge. What is the probability that among 2000 cars driven over this bridge.

i) Not more than one will have a flat tyre.

ii) More than three will have a flat tyre.

iii) Exactly 4 will have a falt tyre.

Solution: Please, see the 2.4.

Here;

$P = 0.001 \qquad n = 2000$

$\therefore$ Mean occurrence (λ) = np = 0.001 × 2000 = 2

Let x be the number of flat tyre.

Then, by Poisson's distribution

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{where } x = 0, 1, 2, \ldots\ldots$$

i) $P(x \le 1) = P(0) + P(1) = \frac{e^{-2} \times (2)^0}{0!} + \frac{e^{-2} \times (2)^1}{1!}$

$= 0.1353 + 0.2706$

$= 0.4059$

ii) $P(x > 3) = 1 - P(x \le 3)$

$= 1 - \{P(0) + P(1) + P(2) + P(3)\}$

$= 1 - \{0.1353 + 0.2706 + 0.2706 + 0.1804\}$

$= 1 - 0.8659$

$= 0.1431$

iii) $P(x = 4) = \frac{e^{-2} \times (2)^4}{4!} = 0.0902$

44. Define hypergemetric probability distribution with an example describe the conditions for the binomial approximation to hypergeomertic distribution? [2074 Bhadra]

Solution: Please, see the Q. No. 36.

45. An office switchboard receives telephone calls at a rate of 3 calls per minute on an average. Find the probability of reviving

i) no calls in one minute interval

ii) at least 3 calls in an one minute interval

iii) at most 2 calls in a five minute interval. [2074 Bhadra]

Solution:

Given that;

Mean call per minute (λ) = 3

Let x be the number of phone calls.

Then, by Poisson's distribution, we have

$$P(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

i) $P(x = 0; 3) = \frac{e^{-2}(3)^0}{0!} = 0.05$

ii) $P(x \geq 3; 3) = 1 - P(x < 3; 3)$
$= 1 - \{P(x = 0; 3) + P(x = 1; 3) + P(x = 2; 3)\}$
$= 1 - \{0.05 + 0.15 + 0.224\}$
$= 0.576$

Now,
$\therefore$ Mean call per 5 minutes $(\lambda') = 3 \times 5 = 15$

iii) $P(x \geq 2; 15) = P(x = 0; 15) + P(x = 1; 15) + (Px = 2; 15)$
$= 0.03 \times 10^{-5} + 0.46 \times 10^{-5} + 3.11 \times 10^{-5}$
$= 3.93 \times 10^{-5}$

46. Define binomial probability distribution with an example how does negative binomial distribution differ from binomial distribution. [2075 Baishakh]

Solution: Please, see the 2.2 and 2.3.2.

47. In a certain factory turning out optical lenses, there is a small chance, $\frac{1}{500}$ for any lens to be defective. The lenses are supplied in packet of 10 each. What is the probability that a packet will contain
i) No defective lens
ii) At least one defective lenses
iii) At most two defective lenses. [2075 Baisakh]

Solution: Please, see the Q. No. 21.

48. What are the difference and similarities between binomial and hyper geometric probability distribution. [2075 Bhadra]

Solution:

Difference between binomial and hyper geometric probability distribution

Binomial distribution	Hyper distribution
i) The probability of success remains constant from trial to trial.	i) The probability of success does not remain constant from trial to trial.
ii) It describe distribution of draws with replacement.	ii) It describes distribution of draws without replacement.
iii) Its probability density function is $b(x; n) = \binom{n}{x} p^x q^{n-x}$	iii) Its probability density function is $P(x; n\ M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
iv) Its mean = np and Variance = npq	iv) Its mean $= \left(\frac{M}{N}\right)$ and Variance $= \left(\frac{N-n}{N-1}\right) npq$

Similarities

i) Both are discrete probability distribution

ii) Hyper-geometric distribution follows binomial distribution when $n \to \infty$ and $\frac{M}{N} = P$.

49. Multi-choice test consists of 10 questions and 4 answer to each question. If each question is answered by shuffling 4 tags tabled 1, 2, 3 and 4 drawing one making the alternative whose number is drawn. Find the probability of i) getting 3 ii) at least one iii) atleast 7 iv) at most three of these questions answered correctly. [2075 Bhadra]

Solution:

Here;
$n = 10$
$p = \frac{1}{4} = 0.25$
$\therefore q = 1 - p = 0.75$

Now, using binomial distribution

Let x be the number of question answered correctly.

$$P(x) = \binom{n}{x} p^x q^{n-x}; x = 0, 1, 2, \ldots\ldots, n$$

i) $P(x = 3) = \binom{10}{3} (0.25)^3 (0.75)^7 = 0.2502$

ii) $P(x \geq 1) = 1 - P(x < 1)$
$= 1 - P(x = 0)$
$= 1 - 0.0563$

iii) $P(x \geq 7) = P(7) + P(8) + P(9) + P(10)$
$= 0.0030 + 0.0003 + 0.000028 + 0.0000009$
$= 0.0033289$

iv) $P(x \leq 3) = P(0) + P(1) + P(2) + P(3)$
$= 0.0563 + 0.1877 + 0.2815 + 0.2502$
$= 0.7757$

50. Define discrete probability distribution with suitable example. Compare negative binomial and binomial probability distribution. [2074 Chaitra]

Solution:

A Discrete probability distribution is defined as the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has countable values, such as a list of non-negative integers. With discrete probability distribution, each possible value of the discrete random variable can be associated with a non-zero probability. Thus, a discrete probability distribution is often presented in tabular from.

For example; For a binomial distribution with total no. of trails five and probability of success be 0.4 then,

P(x = 0)	P(x = 1)	P(x = 2)	P(x = 3)	P(x = 4)	P(x = 5)
0.077	0.2592	0.3456	0.2304	0.0768	0.0102

Solution: Please, see the 2.3.2.

51. A quality control engineer inspects a random sample of 3 batteries from each lot of 24 car batteries that is ready to shipment. If such a lot contain six batteries with slight defects. What is the probabilities that the inspector's sample will contain

a) None of the batteries with defect

b) Only one of the batteries with defect

c) At less two of the batteries with defect **[2074 Chaitra]**

Solution: Please, see the Q. No. 16.

b) Find P(x = 1)

52. Define binomial distribution and explain the condition for binomial distribution. **[2075 Ashwin]**

Solution: Please, see the 2.2.

53. If the probability that an individual suffers a bad reaction from a certain infection is 0.001, among 2000 individual

a) Obtain probability distribution function is 0.001, among 200 individual

b) Determine the probability that

i) Exactly 3 individuals will suffer bad reaction

ii) More than 2 individuals will suffer bad reaction.

[2075 Ashwin]

Solution:

Here;

$P = 0.001 \qquad N = 2000$

$\therefore$ Mean occurrence $(\lambda) = np = 2000 \times 0.001 = 2$

Using Poisson's distribution

Let x be the number of individual suffering bad reaction.

Then, $P(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ where x = 0, 1, 2, ...

a) $P(x) = \dfrac{e^{-2} \times 2^x}{x!}$ is the required pdf.

b) i) $P(x = 3) = \dfrac{e^{-2} \times 2^3}{3!} = 0.1804$

ii) $P(x > 2) = 1 - P(x \le 2)$

$= 1\ \{P(x = 0) + P(x = 1) + P(x = 2)\}$

$= 1 - \{0.1353 + 0.2706 + 0.2706\} = 1 - 0.6765$

SOLVED QUESTIONS (FOR BCT/BEX/BEL/B. Agri)

1. Explain the Poisson process. What are mean and standard deviation of Poisson distribution? Give one example of Poisson distribution. **[5] [2068 Jestha]**

Solution:

The Poisson process is one of the most important random process in probability theory. It is widely used to model random points in time and space, such as the times of radioactive emissions, the arrival time of customers at a service center.

In other word, a Poisson point process is a type of random mathematical object that consists of points randomly located on a mathematical space.

From Poisson's distribution, we have

$$P(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, ..., \infty$$

where, x is the number of occurrence of some random event

λ = Mean occurrence of the random event.

Mean variance $(\lambda) = np$

where, n is the total number of possible cases

p is the probability of success

and standard deviation $(\sigma) = \sqrt{\lambda}$

Please, see the section 2.4.

2. Suppose you are investigating the safety of a dangerous intersection. Past police records indicate a mean of 5 accidents per month at this intersection. Calculate the probability in any month of:

i) No accident occurring

ii) Exactly one accident occurring

iii) At most one accident occurring **[6][2068 Jestha]**

Solution:

Given that;

Mean accident per month $(\lambda) = 5$

From Poisson's distribution, we have

$$P(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Let x denote the number of accident then

i) $P(0; 5) = \dfrac{e^{-5}(5)^0}{0!} = 0.0067$

ii) $P(1; 5) = \dfrac{e^{-5} \times 5^1}{1!} = 0.0336$

iii) $P(X \le 1; 5) = P(0; 5) + P(1; 5) = 0.0068 + 0.0336 = 0.0403$

3. **Define Poisson distribution. Write the limiting case of Poisson distribution as a binomial distribution.** [2+2] [2068 Chaitra]

Solution: Please, see the section 2.4 and 2.4.4.

4. **It is found that 10% of the items produced by a company are defective. Out of 8 items chosen, using binomial distribution, find the probability of (i) no defective item ii) atleast one defective item.** [2+2] [2068 Chaitra]

Solution:-

Given that;

- Probability that an items is defective (p) = 10% = 0.1

  Number of items chosen (n) = 8

Let x denotes the number of defective items

Then by binomial distribution, we have

$$P(x; n, p) = \binom{n}{x} p^x q^{n-x}, \text{ where, } x = 0, 1, 2, \ldots\ldots, n$$

i) $P(0; 8, 0.1) = \binom{8}{0} (0.1)^0 (1 - 0.1)^{8-0}$

$= \binom{8}{0} (0.1)^0 (0.9)^8$

$= 0.4304$

ii) $P(x \geq 1; 9, 0.1) = 1 - P(x < 1; 8, 0.1)$

$= 1 - P(0; 8, 0.1)$

$= 0.5696$

5. **Define hyper geometric distribution. Write down the difference between hyper geometric and binomial distribution.** [2+2]

Solution: Please, see the section 2.5.

Difference between

Binomial Distribution	Hyper geometric Distribution
i) The probability of success of an event remains constant from trail to trail in an experiment.	i) The probability of success of an event doesn't remain constant from one trail to another trail in an experiment.
ii) It has two parameters	ii) It has three parameters.
iii) Its pdf is given as $P(x; n, p) = \binom{n}{x} p^x q^{n-x}$	iii) Its pdf is given as $P(x; n, M, N) = \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

6. **If 6 of 18 new building in a city violate the building code, what is the probability that a building inspector, Who randomly selects 4 of the new building for inspection, will catch**
   a) **None of the buildings that violate the building code.**
   b) **At list 3 of the new buildings that violate the building code.**

OR

A local booth of Nepal telecom on an average 2 calls per minute. Find the probability of (i) no phone call (ii) Exactly 4 calls iii) at lest three calls. [4] [2069 Bhadra]

Solution:

Given that;

Total number of new building (N) = 18

Number of building violating building code (M) = 6

Number of building selected (n) = 4

Let x denote the number of building selected that violate building code.

Then, from hyper-geometric distribution

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

a) $P(0; 4, 6, 18) = \dfrac{\binom{6}{0}\binom{12}{4}}{\binom{18}{4}} = 0.1617$

b) $P(x \geq 3; 4, 6, 18) = P(3; 4, 6, 18) + P(4; 4, 6, 18)$

$$= \frac{\binom{6}{3}\binom{12}{1}}{\binom{18}{4}} + \frac{\binom{6}{4}\binom{12}{0}}{\binom{18}{4}}$$

$$= \frac{4}{51} + \frac{1}{204}$$

$$= \frac{1}{12} = 0.0833$$

OR

Solution:

Given that;

Mean call per minute (λ) = 2

Let x denote the number of phone calls.

Then, by Poisson's distribution, we have

$$P(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

i) $P(0; 2) = \frac{e^{-2}(2)^0}{0!} = 0.1353$

ii) $P(4; 2) = \frac{e^{-2}(2)^4}{4!} = 0.0902$

iii) $P(x \geq 3; 2) = 1 - P(x < 3; 2)$

$= 1\ \{P(0; 2) + P(1; 2) + P(2; 2)$

$= 1 - \{0.1353 + 0.2706 + 0.2706\} = 0.3235$

7. Among the 12 solar collectors on display at a tradeshow, 9 are flat plate collectors and the others are concentrating collectors. If a person visiting the show randomly selects 6 of the solar collectors to check out. What is the probability that

i) Non of them will be flat plate collectors.

ii) At least 3 of them will be flat plate collector.

iii) At most 2 of them will be concentrating collector.

Solution:

Given that;

Total number of solar collectors (N) = 12

Total number of flat collectors (M) = 9

Number of collector selected (n) = 6

Let x be the number of flat collectors to be selected.

Then, from hyper-geometric distribution

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

i) $$P(0; 6, 9, 12) = \frac{\binom{9}{0}\binom{12-9}{6-0}}{\binom{12}{6}} = 0$$

[∴ Since 6 collectors are selected and there are only 3 concentrating collectors that is why there must be 3 flat collectors. so, the probability for 0, 1 or 2 flat collector (s) is zero.]

ii) $P(x \geq 3; 6, 9, 12) = 1 - P(x < 3; 6, 9, 12)$

$= 1 - \{P(x = 0; 6, 9, 12) + P(x = 1; 6, 9, 12)$

$+ P(x = 2; 6, 9, 12)\}$

$$= 1 - 0 + \left\{\frac{\binom{9}{1}\binom{12-9}{6-1}}{\binom{12}{6}} + \frac{\binom{9}{2}\binom{12-9}{6-2}}{\binom{12}{6}}\right\}$$

$= 1 - \{0 + 0 + 0\} = 1$

iii) Out of six at most 2 should be concentrating collectors. *i.e.*, at least 4 should be flat collectors.

So, $P(x \geq 4; 6, 9, 12) = 1 - P(x < 4; 6, 9, 12)$

$= 1 - \{P(x = 0; 6, 9, 12) + P(x = 1; 6, 9, 12)$

$+ P(x = 2; 6, 9, 12) + P(x = 3; 6, 9, 12)\}$

$$= 1 - \left\{0 + 0 + 0 + \frac{\binom{9}{3}\binom{12-9}{6-3}}{\binom{12}{6}}\right\}$$

$$= 1 - \frac{1}{11} = \frac{10}{11} = 0.909$$

8. Write difference between binomial distribution and negative binomial distribution with suitable examples. [2069 Chaitra]

Solution: Please, see the 2.3.2.

9. Write the conditions of using hyper-geometric distribution. An internet-based company that sells discount accessories for cell phones often ships an expensive number of defective products. The company needs better control of quality. Suppose it has 20 identical car chargers on hand but that 5 are defective. If the company decide to randomly select 10 of these items, What is the probability that 2 of the 10 will be defective? Also find mean and variance of the distribution. [1+4] [2069 Poush]

Solution: Please, see the 2.5.

Given that;

Total number of chargers (N) = 20

Total number of defective chargers (M) = 5

Number of charger selected (n) = 10

At x be the defective charger,

Then, from hyper geometric distribution

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\therefore \quad P(2; 10, 5, 20) = \frac{\binom{5}{2}\binom{20-5}{10-2}}{\binom{20}{10}} = 0.3483$$

$$\therefore \quad \text{Mean } (\bar{X}) = \frac{M}{N} = \frac{5}{10} = 0.5$$

$\therefore$ Variance $(\sigma) = \left(\frac{N-n}{N-1}\right) \times npq = \frac{20-10}{20-1} \times 10 \times 0.5 \times (1-0.5) = 1.315$

10. Define poison's probability distribution. Write the conditions fo Poisson approximation to binomial distribution. [2070 Ashad

Solution: Please, see the 2.4 and 2.4.4.

11. A quality control engineers inspects a random sample of batteries from each lot of 24 car batteries that is ready to shipmen If such a lot contain six batteries with slight defects, what is th probability that the inspector's sample will contain

i) None of the batteries with defect?

ii) Only one of the batteries with defect?

iii) At least two of the batteries with defect? [2070 Ashad

Solution:

Given that;

Total number of batteries (N) = 24

Total number of defective batteries (M) = 6

Number of batteries selected (n) = 3

Let x be the number of defective battery selected. Then,

$$P(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

i) $P(x = 0; 3, 6, 24) = \frac{\binom{6}{0}\binom{18}{3}}{\binom{24}{3}} = 0.4031$

ii) $P(x = 1; 3, 6, 24) = \frac{\binom{6}{1}\binom{18}{2}}{\binom{24}{3}} = 0.4535$

iii) $P(x \geq 2; 3, 6, 24) = 1 - P(x < 2; 3, 6, 24)$

$= 1\{P(x = 0; 3, 6, 24) + P(x = 1; 3, 6, 24)\}$

$= 1(0.4031 + 0.4535)$

$= 0.1434$

12. Write the difference and similarities between binomi distribution and negative binomial, distribution. [2070 Chaitr

Solution: Please, see the 2.3.2.

13. The number of accidents in a year attributes to taxi drivers in a city follow poison distribution with mean 3. Out of 1000 taxi driver, find the approximately the number of driver with

i) No accidents' in a year.

ii) More than 3 accident in a year. [2070 Chsitra]

Solution:

Here;

Mean occurrence $(\lambda) = 3$

Let x be the number of driver with accident we know

$$P(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, ...$$

Now,

i) $P(x = 0; 3) = \frac{e^{-3}(3)^0}{0!} = 0.0497$

Therefore, the number of driver with no accident is = 0.0497 × 1000 = 49.7 ≈ 50 drivers.

ii) $P(x > 3; 3) = 1 - P(x \leq 3; 3)$

$= 1 - \{P(0) + P(1) + P(2) + P(3)\}$

$= 1 - \{0.0497 + 0.1493 + 0.2240 + 0.2240\}$

$= 1 - 0.647$

$= 0.353$

Therefore, the number of driver with more than 3 accident in a year

$= 0.353 \times 1000 = 353$

14. Define negative binomial distribution with an example. How does the negative binomial distribution differ from binomial distribution? [2071 Sharawan]

Solution: Please, see the 2.3 and 2.3.2.

15. A heavy machinery manufacture has 3840 large generators in the field that are under warranty. If the probability is $\frac{1}{1200}$ that anyone will fail during the given year, Find the probability.

i) The exactly 3 generators will fail during the given year?

ii) That between 2 and 6 are fail during the given year?

[2071 Shrawan]

Solution:

Given that;

Total number of generators (n) = 3840

Probability of defective generator (p) = $\frac{1}{1200}$

$\therefore$ Mean occurrence $(\lambda) = np = 3.2$

Let x be the number of defective generator.

We know,

$$P(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, ...$$

i) $P(x = 3; 3.2) = \frac{e^{-3.2}(3.2)^3}{3!} = 0.2226$

ii) $P(x = \text{between 2 and 6})$

$= P(x = 3) + P(x = 4) + P(x = 5)$

$= 0.2226 + 0.1780 + 0.1139$

$= 0.5145$

16. Write the difference and similarities between binomial and negative binomial distribution. [2071 Chaitra]

Solution: Please, see the 2.3.2.

Similarities

i) Both binomial and negative binomial distributions describe distribution of draws with replacement.

ii) There are only two possible outcomes for each trial.

iii) Both follow additive property for independent variables.

iv) Both are discrete probability distribution.

17. A quality control engineer inspects a random sample of 4 batteries from each lot of 24 car batteries that is ready to shipment. If such a lot contain six batteries with slight defects. What are the probabilities that the inspectors sample will contain:

i) None of the battery with defects?

ii) At least two of the batteries with defects?

iii) At most three of the batteries with defect?

Solution:

Given that;

Total number of batteries (N) = 24

Total number of defective batteries (M) = 6

Number of battery selected (n) = 4

Let x be the number of defective battery selected then

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

i) $P(x = 0) = \frac{\binom{6}{0}\binom{18}{4}}{\binom{24}{4}} = 0.2879$

ii) $P(x \geq 2) = 1 - P(x < 2)$

$= 1\{P(0) + P(1)\}$

$= 1 - \{0.2879 + 0.4607\} = 0.2514$

iii) $P(x \leq 3) = P(0) + P(1) + P(2) + (P(3)$

$= 0.2879 + 0.4607 + 0.2159 + 0.0338$

$= 0.9983$

18. Define negative binomial distribution with its important characteristics. [2072 Kartik]

Solution: Please, see the 2.3 and 2.3.1.

19. A particularly long traffic light on your morning commute is green 20% of the time that you approach it. Assume that each morning represents as independent trials.

i) Over five morning's what is the probability that the light is green on exactly one day?

ii) Over 20 mornings, what is the probability that the light is green on exactly four days?

Solution:

Given that;

Probability the traffic light is green (p) = 20%

$\therefore$ $q = 100\% - 20\% = 80\%$

Now, using binomial distribution

$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, ..., n$$

i) $n = 5$

$\therefore$ $P(x = 1) = \binom{5}{1}(0.2)^1(0.8)^{5-1} = 0.4096$

ii) $n = 20$

$\therefore$ $P(x = 4) = \binom{20}{4}(0.2)^4(0.8)^{20-4} = 0.2181$

20. Write the differences and similarities between binomial and negative binomial distribution. [2072 Chaitra]

Solution: Please, see the Q. No. 16.

21. In certain factory turning out optical lenses, there is a small change, $\frac{1}{500}$ for any lens to be defective. The lenses are supplied in

packets of 10 each. What is the probability that a packet will contain.

i) No defective lens

ii) At least one defective lenses

iii) At most two defective lenses. **[2072 Chaitra]**

Solution:

Given that;

Probability that a lens is defective (p) = $\frac{1}{500}$

Number of lenses in a packet (n) = 10

$\therefore$ Mean occurrence (λ) = np = $10 \times \frac{1}{500} = \frac{1}{50}$

Now, we know

$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ [where x is defective lense]

i) $P(x = 0) = \frac{e^{-(1/50)}(1/50)^0}{0!} = 0.9801$

ii) $P(x \geq 1) = 1 - P(x < 1)$
$= 1 - P(x = 0)$
$= 1 - 0.9801$
$= 0.0199$

iii) $P(x \leq 2) = P(0) + P(1) + P(2)$
$= 0.9801 + 0.0196 + 0.00019 = 0.9998$

OR

Define mathematical expectation of a discrete random variable. A probability distribution is given.

X = x	0	1	2	3	4	5
P(X) = x	0.26	0.25	0.11	0.02	0.25	0.11

Find a) $P(X \geq 4)$; b) $P(0 < X < 4)$ c) mean and variance of x.

Solution:

Mathematical expectation of a discrete random variable is the summation or integration of a possible values from a random variable.

a) $P(X \geq 4) = 1 - P(X < 4)$
$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$
$= 1\ (0.26 + 0.25 + 0.11 + 0.02$
$= 1 - 0.64 = 0.36$

b) $P(0 < X < 4) = P(1) + P(2) + P(3)$
$= 0.25 + 0.11 + 0.02$
$= 0.38$

c)

X = x	P(X = x)	xp	x^2p
0	0.26	0	0
1	0.25	0.25	0.25
2	0.11	0.22	0.88
5	0.02	0.06	0.54
4	0.25	1	4
5	0.11	0.55	2.75
		2.08	8.42

$\therefore$ Mean (μ) = $\Sigma xp = 2.08$

$\therefore$ Variance (V) = $\Sigma x^2p - \mu^2 = 8.42 - (2.08)^2 = 4.0936$

22. Define negative binomial distribution. If a boy is throwing stone at a target what is the probability that his 10th throw is his 5th hit. If the probability of hitting the target at any trial is 0.6 also find the mean and variance of random variable. **[2073 Shrawan]**

Solution: Please, see the 2.3.

Given that;

Probability of hitting the target (p) = 0.6

Number of success (r) = 5

Number of hit (n) = 10

Now,

$x + r = n$

or, $x + 5 = 10$

$\therefore$ $x = 5$

We know,

$P(x) = \binom{r + x - 1}{r - 1} p^r q^x, x = 0, 1, 2, ...$

or, $P(x = 5) = \binom{9}{4} (0.6)^5 (1 - 0.6)^5 = 0.1$

$\therefore$ Mean $= \frac{rq}{p} = \frac{5 \times 0.4}{0.6} = 3.33$

$\therefore$ Variance $= \frac{rq}{p^2} = \frac{5 \times 0.4}{0.6^2} = 5.56$

23. Define hyper geometric probability distribution. With an example describe the conditions for the binomial approximation to hypergeometric distribution? **[2073 Shrawan]**

Solution: Please, see the 2.5.

24. Define negative binomial distribution with its importa characteristics. [2073 Chait

Solution: Please, see the 2.3 and 2.3.1.

25. If a publisher of nontechnical books takes great pains to ensu that its books are free of typographical errors. So that th probability of any given page containing at least one such error 0.005 and errors are independent from page to page, what is th probability that one of its 400-page novels will contain

a) Exactly one page with errors?

b) At most three pages with errors? [2073 Chait

Solution:

Given that;

Probability of a page with error (P) = 0.005

Probability of a page (n) = 400

∴ Mean occurrence (λ) = np = 2

Now, let x be the number of page with error.

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{where } x = 0, 1, 2, \ldots\ldots$$

a) $P(x = 1) = \frac{e^{-2} \times 2^1}{1!} = 0.1353$

b) $P(x \le 3) = P(0) + P(1) + P(2) + P(3)$

$= 0.1353 + 0.2706 + 0.2706 + .01804$

$= 0.8569$

26. What are the characteristics of binomial distribution and how do it differ from negative binomial distribution. [2074 Ashwi

Solution: Please, see the 2.2.1 and 2.3.2.

27. A quality control engineer inspects a random sample of 4 batteri from each lot of 24 car batteries that is ready to shipment. If such lot contain six batteries with slight defects. What are th probabilities that the inspector's sample will contain.

i) None of the batteries with defect?

ii) At least two of the batteries with defects?

iii) At most three of the batteries with defects? [2074 Ashwin]

Solution: Please, see the Q. No. 16.

Chapter 3

CONTINUOUS PROBABILITY DISTRIBUTION

3.1 Continuous Random Variable and Probability Densities 68

3.1.1 Continuous Random Variable 68

3.1.2 Probability Density Function (PDF) 68

3.1.3 Commulative Distribution Function (CDF) 69

3.1.4 Expected Value and Variables of a Continuous Ransom Variable 69

3.2 Normal Distribution 70

3.2.1 Standard Normal Distribution 71

3.3 Gamma Distribution 76

3.4 Chi-squared (χ^2) Distribution 79

3.1 CONTINUOUS RANDOM VARIABLE AND PROBABILITY DENSITIES

3.1.1 Continuous random variable

a variable which takes an infinite number of possible values is known as continuous random variable. *i.e.*, things that you can measures but not by counting, *For example;*

- x = {amount of rainfall in inches}
- x = {height of student in the class}

3.1.2 PROBABILITY DENSITY FUNCTION (PDF)

A probability density function is a function of a continuous random variable, whose integral across interval gives the probability that the value of the variable lies within the same interval. It is a function $f(x)$ which is associated with a continuous random variable x such that for any two numbers a and b with $a \le b$

Probability $(a < x < b) = \int_a^b f(x)\,dx$

If $f(x)$ satisfies the following property

i) $f(x) \ge 0$ for all x

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Let us consider an example: Let X denotes the amount of rainfall in inches whose graph is shown in figure below.

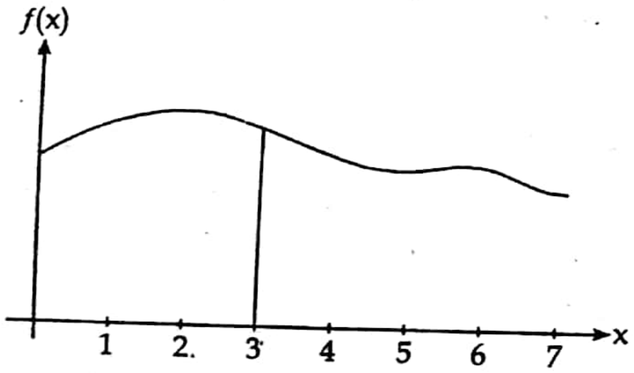


Figure: Rainfall in inches

Now, $P(X = 3) = \int_3^3 f(x)\,dx = 0$

$P(1 < X < 4) = \int_1^4 f(x)\,dx$

$P(X > 2) = \int_2^{\infty} f(x)\,dx = 1 - \int_0^2 f(x)\,dx$

$P(1 \le X < 6) = P(X = 1) + P(1 < X < 6)$

$= 0 + \int_1^6 f(x)\,dx = \int_1^6 f(x)\,dx$

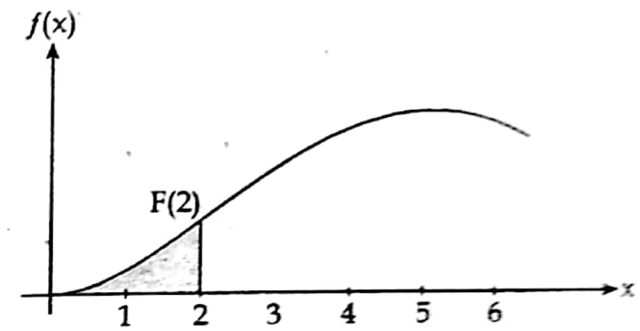
3.1.3 COMMULATIVE DISTRIBUTION FUNCTION (CDF)

Commulative distribution function is a function whose value is the probability that a corresponding continuous random variable has a value less than or equal to the argument of the function.

Mathematically,

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx \qquad \text{for } -\infty < x < \infty.$$

Here, $-\infty$ can be replaced with the lower endpoint of the distribution for practical approach.



Here, $F(2) = P(X \le 2) = \int_{-\infty}^{2} f(x)\,dx$

$= \int_{-\infty}^{0} f(x)\,dx + \int_0^2 f(x)\,dx$

$= 0 + \int_0^2 f(x)\,dx$

$= \int_0^2 f(x)\,dx$

Properties of cdf

i) $F(x) \le 1$

ii) $F(-\infty) = P(X \le -\infty) = \int_{-\infty}^{-\infty} f(x)\,dx = 0$

iii) $F(\infty) = P(X \le \infty) = \int_{-\infty}^{\infty} f(x)\,dx = 1$

iv) $P(a < X < b) = \int_a^b f(x)\,dx = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx$

$\therefore P(a < X < b) = F(b) - F(a)$

v) $f(x) = F'(x)$

3.1.4 EXPECTED VALUE AND VARIABLES OF A CONTINUOUS RANSOM VARIABLE

The average value of a random variable is known as mathematical expectation. Mathematically,

$$\text{Mean} = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$\text{Variance} = V(x) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

where $E(X^2) = \int_{-\infty}^{\infty} x f(x)\,dx$

3.2 NORMAL DISTRIBUTION

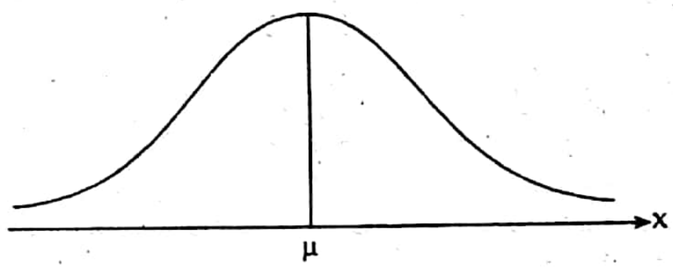
A continuous random variable X is said to have normal distribution with parameter mean (μ_1) and variance (σ^2) if and only if the corresponding probability density function is given by

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$X \sim N(\mu, \sigma)$

Properties of normal distribution

i) The curve of normal distribution is bell shaped and symmetrical about vertical line $x = \mu$.



ii) The x-axis asymptote to the normal curve.

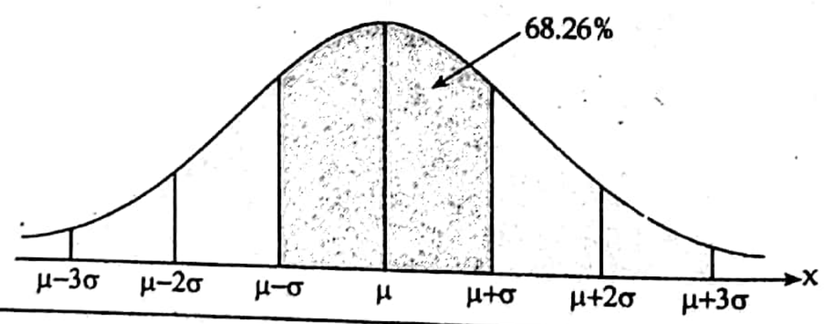
iii) The mean, median and mode of normal distribution lie at same point at centre.

iv) Area under the curve of normal distribution gives probability.

v) Total area of the curve of normal distribution is equal to one.

vi) Standard deviation of distribution control the width of the curve of normal distribution.

Area properties



X	% of area
$(\mu - \sigma)$ to $(\mu + \sigma)$	68.26%
$(\mu - 2\sigma)$ to $(\mu + 2\sigma)$	95.44%
$(\mu - 3\sigma)$ to $(\mu + 3\sigma)$	99.73%

Note: If X is a normal random variable, $E(X) = \mu$

Now, let $Z = \frac{x-\mu}{\sigma}$

Then, $E(Z) = E\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}[E(x) - E(\mu)] = \frac{1}{\sigma}(\mu - \mu) = 0$

$V(Z) = V\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2} \times \sigma^2 = 1$

$\therefore \quad Z \sim N(0, 1)$

3.2.1 STANDARD NORMAL DISTRIBUTION

If a continuous random variable x follows normal distribution with parameter mean (μ_1) and variance (σ^2) then $z = \frac{x-\mu}{\sigma}$ follow standard normal distribution with mean zero and variance one if and only if its probability density function is given by

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

For Z table

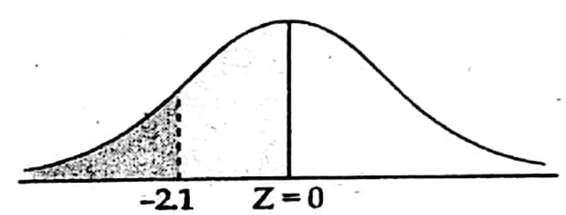


Fig: Negative Z-score

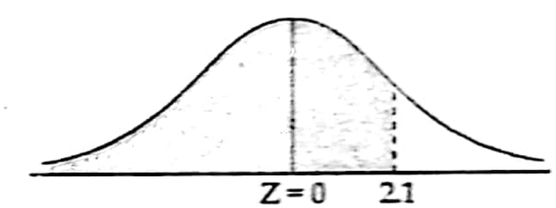
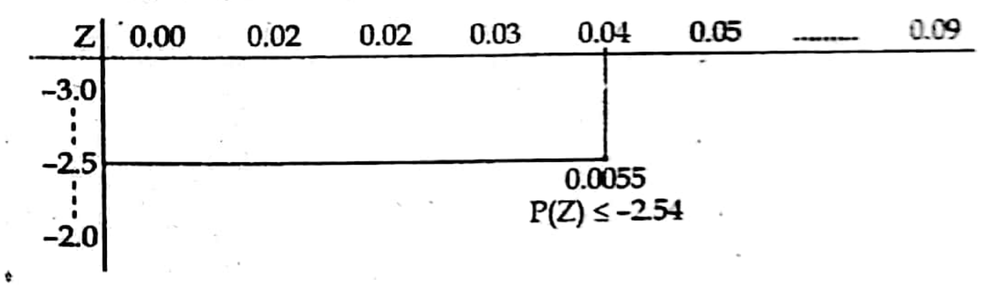
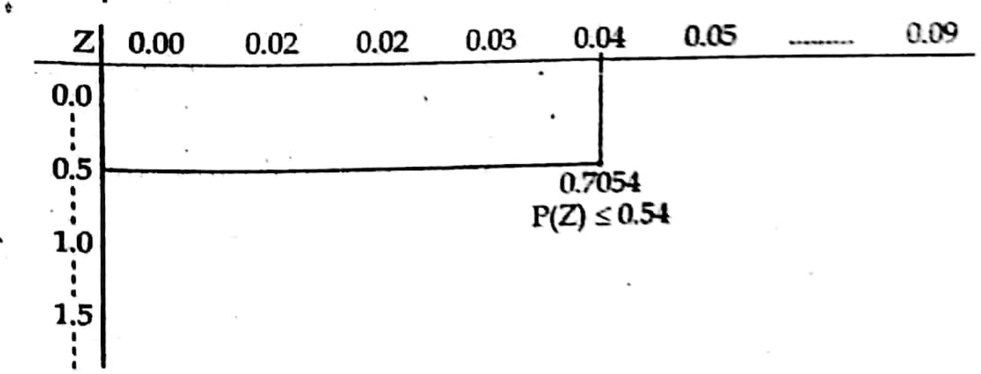


Fig: Positive Z-score





Let us see few examples

Example 1

The burning time of an experimental rocket is a random variable having the normal distribution with mean 4.76 and $\sigma = 0.05$. What is the probability that the rocket will burn

a) less than 4.6 sec
b) more than 4.80 sec
c) anywhere from 4.60 to 4.80 sec

Solution:

Let x be a continuous random variable that represents burning time of experimental rocket.

We have;

Mean (μ) = 4.76

Standard deviation (σ) = 0.05

Now,

a) $P(x < 4.6) = P\left(\frac{x-\mu}{\sigma} < \frac{4.6-4.76}{0.05}\right) = P(Z < -3.2) = 0.0007$

b) $P(x > 4.80) = P\left(\frac{x-\mu}{\sigma} > \frac{4.80-4.76}{0.05}\right)$

$= P(Z > 0.8)$

$= 1 - P(Z \le 0.8)$

$= 1 - 0.7881$

$= 0.2119$

c) $P(4.60 < x < 4.80) = P\left(\frac{4.60-4.76}{0.05} < \frac{x-\mu}{\sigma} < \frac{4.80-4.76}{0.05}\right)$

$= P(-3.2 < Z < 0.8)$

$= P(0.8) - P(-3.2)$

$= 0.7881 - 0.0007$

$= 0.7874$

Example 2

In a school of 1000 students, a test was conducted whose result shows that the average score was 42 and standard deviation was 24 find

a) the number of student exceeding a score 50.
b) the number of student lying between 30 and 60
c) the value of score exceeded by top 50 student

Solution:

Let x be a continuous random variable which represents the marks of the students.

We have;

Mean (μ) = 42

Standard deviation (σ) = 24

Now,

a) $P(x > 50) = P\left(\frac{x-\mu}{\sigma} > \frac{50-42}{24}\right)$

$= P(Z > 0.33)$

$= 1 - P(Z \le 0.33)$

$= 0.3707$

$\therefore$ No. of student $= 0.3707 \times 1000 = 370.7 \approx 371$

b) $P(30 < x < 60) = P\left(\frac{30-42}{24} < \frac{x-\mu}{\sigma} < \frac{60-42}{24}\right)$

$= P(-0.5 < Z < 0.75)$

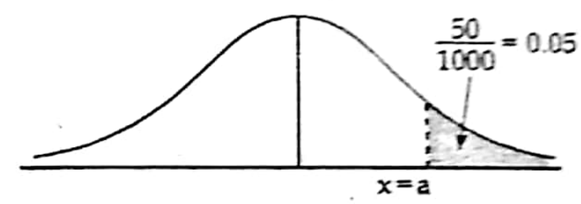
$= P(0.75) - P(-0.05)$

$= 0.7737 - 0.3085$

$= 0.4649$

$\therefore$ No. of students $= 0.4649 \times 1000 = 464.9 \approx 465$ students.

c) Let a be the score exceeded by top 50 students. Then,



$P(X > a) = \frac{50}{1000}$

or, $P\left(\frac{x-\mu}{\sigma} < \frac{a-42}{24}\right) = 0.05$

or, $P\left(Z > \frac{a-42}{24}\right) = 0.05$

or, $1 - P\left(Z \le \frac{a-42}{24}\right) = 0.05$

or, $P\left(Z \le \frac{a-42}{24}\right) = 0.95$

or, $P\left(Z \le \frac{a-42}{24}\right) = P(Z \le 1.645)$

or, $\frac{a-42}{24} = 1.645$

or, $a = 1.645 \times 24 + 42$

$\therefore$ $a = 81.48 \approx 81.5$

Example 3

A random variable has normal distribution with mean (μ) = 62.4. Find its standard deviation if the probability is 0.2 that it will take on value greater than 79.2.

Solution:

Let x be a continuous random variable.

We have;

Mean (μ) = 62.4

Standard deviation (σ) = ?

Now,

$$P(x > 79.2) = 0.2$$

or, $P\left(\frac{x-\mu}{\sigma} > \frac{79.2-62.4}{\sigma}\right) = 0.2$

or, $P\left(Z > \frac{16.8}{\sigma}\right) = 0.2$

or, $1 - P\left(Z \le \frac{16.8}{\sigma}\right) = 0.2$

or, $P\left(Z \le \frac{16.8}{\sigma}\right) = 0.8$

or, $P\left(Z \le \frac{16.8}{\sigma}\right) = P(Z \le 0.84)$

or, $\frac{16.8}{\sigma}$ 0.84

$\therefore$ $\sigma = 20$

Example 4

The distributions of monthly income of 5000 employee of a certain industries was found to be normal distribution with mean of Rs. 2000 and standard deviation of Rs. 200. Estimate

a) The range of income of middle of 60% employees.

b) The lowest income of richest 10% employees.

c) The highest income of poorest 10% employees.

Solution:

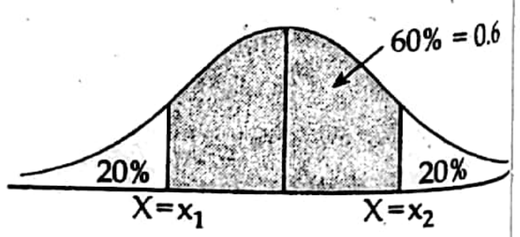
Given that;

Mean (m) = Rs. 2000

Standard deviation (σ) = Rs. 200

a) Let x_1 and x_2 be the lowest and highest income of middle 60% employee. Then,

$$P(x_1 < X < x_2) = 0.6$$



Now,

$$P(X < x_1) = 0.2$$

or, $P\left(\frac{x-\mu}{\sigma} < \frac{x_1-2000}{200}\right) = 0.2$

or, $P\left(Z < \frac{x_1-2000}{200}\right) = 0.2$

or, $P\left(Z < \frac{x_1-2000}{200}\right) = P(Z < -0.84)$

or, $\frac{x_1-2000}{200} = -0.84$

$\therefore$ x_1 = Rs. 1832

Again,

$$P(X > x_2) = 20\% = 0.2$$

or, $P\left(\frac{x-\mu}{\sigma} > \frac{x_2-2000}{200}\right) = 0.2$

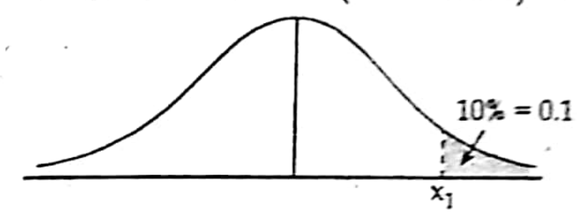
or, $1 - P\left(Z < \frac{x_2-2000}{200}\right) = 0.2$

or, $P\left(Z < \frac{x_2-2000}{200}\right) = 0.8$

or, $\frac{x_2-2000}{200} = 0.84$

$\therefore$ x_2 = Rs. 2168

$\therefore$ Range of income = $x_2 - x_1$ = Rs. (2168 - 1832) = Rs. 336



b) Let x_1 = Lowest income of richest 10% employee

$$P(X > x_1) = 10\% = 0.1$$

or, $P\left(\frac{x-\mu}{\sigma} > \frac{x_1-2000}{200}\right) = 0.1$

or, $P\left(Z > \frac{x_1-2000}{200}\right) = 0.1$

or, $1 - P\left(Z \le \frac{x_1-2000}{200}\right) = 0.1$

or, $P\left(Z \le \frac{x_1-2000}{200}\right) = 0.9$

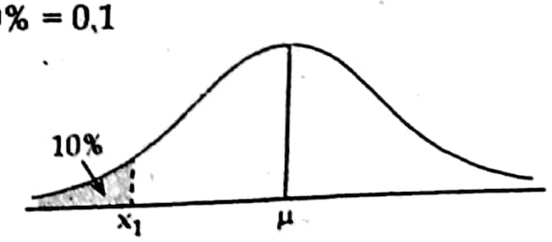
or, $P\left(Z \le \frac{x_1-2000}{200}\right) = P(Z \le 1.28)$

or, $\frac{x_1-2000}{200} = 1.28$

$\therefore$ x_1 = Rs. 2256.

c) Let x_1 be the highest income of poorest 10% employee. Then,

$$P(X < x_1) = 10\% = 0.1$$



or, $\left(\frac{x-\mu}{\sigma} < \frac{x_1-2000}{200}\right) = 0.1$

or, $P\left(Z < \frac{x_1 - 2000}{200}\right) = 0.1$

or, $P\left(Z \le \frac{x_1 - 2000}{200}\right) = 0.1$

or, $P\left(Z < \frac{x_1 - 2000}{200}\right) = P(Z < (-1.28))$

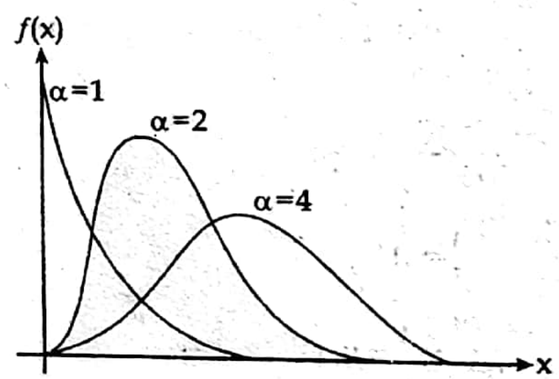
or, $\frac{x_1 - 2000}{200} = -1.28$

$\therefore$ x_1 = Rs. 1744

3.3 GAMMA DISTRIBUTION

A continuous random variable x is said to follow gamma distribution with parameter α and β if and only if it assumes only non-negative value and the probability density function is given by

$$f(x) = f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^2 \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x > 0, \alpha > 0, \beta > 0 \\ 0 & \text{otherwise} \end{cases}$$



Here, α → Shape parameter

β → Scale parameter

Mean, E(X) = αβ

Variance, V(x) = αβ²

Note: $\Gamma n = (n-1)!$

$\Gamma 1 = 1$

$\Gamma(1/2) = \pi$

Uses of gamma distribution

i) Reliability test: Failure time distribution of number of components in electrical system

ii) Quening theory: Waiting to distribution of specific number of events.

iii) To study the distribution of water supply, electricity, petrol supply in a city.

Example 5

Suppose that the time in hours, taken to repair a heat pump, heat pump is random variable X having gamma distribution with parameter α = 2 and $\beta = \frac{1}{2}$. What is the probability that next service call require.

a) Atmost 1 hour to repair the heat pump

b) Atleast 2 hour to repair the heat pump

Solution:

Given that;

$\alpha = 2, \beta = \frac{1}{2}$

From gamma distribution

$$f(x) = \frac{1}{B^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-x/\beta} \quad \text{for } x > 0, \alpha > 0, \beta > 0$$

$$= \frac{1}{(1/2)^2 \Gamma 2} x^{2-1} e^{-x/1/2}$$

$$= \frac{4}{1} xe^{-2x} \quad [\because \Gamma n = (n-1)!]$$

$$= 4\,xe^{-2x}$$

Now,

a) $$P(X \le 1) = \int_0^1 f(x)\,dx$$

$$= \int_0^1 4xe^{-2x}\,dx$$

$$= 4\left[\frac{xe^{-2x}}{-2} - \frac{e^{-2x}}{(-2)^2}\right]_0^1 \quad [\because \int ab = a\int b - \int a'\int b]$$

$$= 4\left[\frac{e^{-2}}{-2} - \frac{e^{-2}}{(-2)^2} - 0 + \frac{1}{(-2)^2}\right]$$

$$= 4\left[\frac{e^{-2}}{-2} - \frac{e^{-2}}{4} + \frac{1}{4}\right] = 0.5939$$

b) $$P(X \ge 2) = \int_2^{\infty} f(x)\,dx$$

$$= \int_2^{\infty} 4xe^{-2x}\,dx$$

$$= 4\left[\frac{xe^{-2x}}{-2} - \frac{e^{-2x}}{(-2)^2}\right]_2^{\infty}$$

$$= 4\left[0 - 0 - \frac{2e^{-4}}{(-2)} + \frac{e^{-4}}{(-2)^2}\right]$$

$$= 4\left[e^{-4} + \frac{e^{-4}}{4}\right]$$

$$= 0.0915$$

Example 6

In a certain city daily consumptions of water (in millions of gallon) follows a gamma distributions with $\alpha = 2$ and $\beta = 3$. If the daily capacity of city is 9 million gallons of water, show that the probability on any given day the water supply is inadequate is $4e^{-3}$.

Solution:

Let x follows the gamma distribution that represents the daily consumption of water in the city with parameter $\alpha = 2$ and $\beta = 3$.

From gamma distribution.

We have;

$$f(x) = \frac{1}{B^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-x/\beta} \qquad \text{for } x > 0, \alpha > 0, \beta > 0$$

$$= \frac{1}{3^2\,\Gamma 2} x^{2-1} e^{-x/3}$$

$$\therefore \quad f(x) = \frac{1}{9} x e^{-x/3} \qquad \text{for } x > 0$$

Now, probability that on any given day the water supply is inadequate is given as;

$$\begin{aligned} P(x > 9) &= 1 - P(x \le 9) \\ &= 1 - \int_0^9 \frac{1}{9} x e^{-x/3}\, dx \\ &= 1 - \frac{1}{9}\left[\frac{xe^{-x/3}}{-(1/3)} - \frac{e^{-x/3}}{(-1/3)^2}\right]_0^9 \\ &= 1 - \frac{1}{9}[-27e^{-3} - 9e^{-3} - 0 + 9] \\ &= 1 - [-3e^{-3} - e^{-3} + 1] \\ &= 3e^{-3} + e^{-3} \\ &= 4e^{-3} \end{aligned}$$

Example 7

The daily consumption of electric power in a certain. City follow a gamma distribution with $\alpha = 2$ and $\beta = 3$. If the power plant of this city has daily capacity of 12 million kilowatt hours, what is the probability that the power supply will be inadequate on any given day? [2071 Chaitra]

Solution:

Let x follow distribution that represents the consumption of electric power in the city with parameter $\alpha = 2$ and $\beta = 3$.

From gamma distribution.

We have;

$$f(x) = \frac{1}{B^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-x/\beta} \qquad \text{for } x > 0, \alpha > 0, \beta > 0$$

$$\text{or,} \quad f(x) = \frac{1}{3^2\,\Gamma 2} x^{2-1} e^{-x/3}$$

$$\therefore \quad f(x) = \frac{1}{9} x e^{-x/3} \qquad \text{for } x > 0$$

Now, probability that the power supply will be inadequate on any given day is

$$\begin{aligned} P(x > 12) &= 1 - P(x \le 12) \\ &= 1 - \int_0^{12} f(x)\, dx \\ &= 1 - \int_0^{12} \frac{1}{9} x e^{-x/3}\, dx \\ &= 1 - \frac{1}{9}\left[\frac{xe^{-x/3}}{-(1/3)} - \frac{e^{-x/3}}{(-1/3)^2}\right]_0^{12} \\ &= 1 - \frac{1}{9}[-36e^{-4} - 9e^{-4} - 0 + 9] \\ &= 1 - [-4e^{-4} - e^{-4} + 1] \\ &= 5e^{-4} \end{aligned}$$

Note: There is inadequate only when either demand has increased with constant supply or the supply has decreased with constant demand.

3.5 CHI-SQUARED (χ^2) DISTRIBUTION

Chi-squared distribution is a special case of gamma distribution where $\alpha = \frac{v}{2}$ and $\beta = 2$. Thus, a continuous random variable is said to have a chi-squared distribution with parameter v degree of freedom if its probability density function is given as

$$f(x) = f(x; v) = \begin{cases} \dfrac{1}{2^{v/2}\,\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Here v denotes degree of freedom.

SOLVED QUESTIONS (FOR BCE/BME/BIE/BAM)

1. **Write the chief characteristics of the normal probability distribution.** [2070 Magh]

Solution: Please, see the 3.2.

2. **A random sample of size 100 is taken from an infinite population having mean 76 and variance 256. Find the probability that $\bar{x}$ will be**
   i) **between 75 and 78**
   ii) **less than 75 and**
   iii) **more than 78**

OR

A continuous random variable X has the density function

$$f(x) = \begin{cases} ce^{-2x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Find i) **The constant c** ii) **$P(1 < x < 3)$**
ii) **E(x)** iv) **V(x)** [2070 Magh]

Solution:

Given that;

Mean (μ) = 76

Variance (σ^2) = 256

$\therefore$ Standard deviation (σ) = $\sqrt{256}$ 16

Now, from normal distribution

Probability that $\bar{x}$ will be

i) Between 75 and 78 is given as

$$P(75 < \bar{x} < 78) = P\left(\frac{75-76}{16} < \frac{\bar{x}-\mu}{\sigma} < \frac{78-76}{16}\right)$$
$$= P(-0.0625 < Z < 0.125)$$
$$= P(0.125) - P(-0.0625)$$
$$= 0.54975 - 0.4751$$
$$= 0.07465$$

ii) Less than 75 is given as

$$P(\bar{x} < 75) = P\left(\frac{\bar{x}-\mu}{\sigma} < \frac{75-76}{16}\right) = P(Z < -0.0625) = 0.4751$$

iii) More than 78 is given as

$$P(\bar{x} > 78) = P\left(\frac{\bar{x}-\mu}{\sigma} > \frac{78-76}{16}\right) = P(Z > 0.125)$$
$$= 1 - P(Z \le 0.125)$$
$$= 1 - 0.54975$$
$$= 0.45025$$

OR

Here;

The density function is

$$f(x) = \begin{cases} ce^{-3x} & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Now,

i) Since it is pdf is it must satisfy

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

or, $\int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} ce^{-3x}\,dx = 1$

or, $0 + c\left[\frac{e^{-3x}}{-3}\right]_0^{\infty} = 1$

or, $c\left[0 - \frac{1}{(-3)}\right] = 1$

or, $c \times \frac{1}{3} = 1$

$\therefore$ $c = 3$

ii) $P(1 < x < 2) = \int_1^2 f(x)\,dx = \int_1^2 3e^{-3x}\,dx$

$$= \left[3\frac{e^{-3x}}{(-3)}\right]_1^2 = [-e^{-3x}]_1^2 = [-e^{-6} + e^{-3}] = e^{-3}[1 - e^{-3}]$$

iii) $E(x) = \int_{-\infty}^{\infty} xf(x)\,dx$

$$= \int_{-\infty}^{0} x \times 0\,dx + \int_0^{\infty} x \times 3e^{-3x}\,dx$$
$$= 0 + 3\left[\frac{xe^{-3x}}{(-3)} - \frac{e^{-3x}}{(-3)^2}\right]_0^{\infty}$$
$$= 3\left[0 - 0 - 0 + \frac{1}{9}\right]$$
$$= \frac{1}{3}$$

iv) $v(x) = E(x^2) - [E(x)]^2$ (1)

$$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^{\infty} x^2 \times 3e^{-3x}\,dx$$
$$= 3\left[\frac{x^2e^{-3x}}{(-3)} - \int \frac{2x\,e^{-3x}}{(-3)}\,dx\right]_0^{\infty}$$
$$= 3\left[\frac{x^2e^{-3x}}{(-3)} + \frac{2}{3}\left\{\frac{xe^{-3x}}{(-3)} - \frac{e^{-3x}}{(-3)^2}\right\}\right]_0^{\infty}$$
$$= 3\left[0 + \frac{2}{3}\{0 - 0\} - 0 - \frac{2}{3}\left\{0 - \frac{1}{9}\right\}\right] = 3 \times \frac{2}{3} \times \frac{1}{9} = \frac{2}{9}$$

From (1)

$$\therefore \quad v(x) = \frac{2}{9} - \left(\frac{1}{3}\right)^2 = \frac{2}{9} - \frac{1}{9} = \frac{1}{9}$$

3. Define the normal distribution. Give the condition for normal approximation of binomial distribution. [2070 Bhadra]

Solution: Please, see the 3.2.

Conditions for normal approximation to binomial distribution.

Let x be a discrete random variable which represents number of success observed from repeated Bernoulli trials. If

i) The number of trails (n) is indefinitely large

i.e., $n \to \infty$

ii) Neither p nor q is very small

Then, $X \sim N(np, npq)$

i.e., $Z = \dfrac{X - np}{\sqrt{npq}}$

4. The mean weight of products Rs. 68.22 grams with variance of 10.8 grams. How many products in a batch of 1000 would you expect be

a) to be over 72 grams

b) between 70 and 72 grams

c) below 65 grams

OR

A collage professor never finishes his lecture before the bell rings to end of the period, and always finishes his lecture within one minute after the bell rings. Let x be the time that elapse between the bell and the end of the lecture, and suppose the probability density function of x is

$$\begin{cases} kx^2 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the value of k

b) What is the probability that the lecture ends within 20 seconds of bell ring?

c) What is the probability that the lecture continue beyond the bell for between 20 to 40 second?

Solution: Similar to Q. No. 2

OR

Here;

Density function is

$$f(x) = \begin{cases} kx^2 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where x is in minute}$$

a) Since it is pdf, it must satisfy

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

or, $\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx = 1$

or, $0 + \int_0^1 kx^2\,dx + 0 = 1$

or, $\left[\dfrac{kx^3}{3}\right]_0^1 = 1$

or, $k \times \dfrac{1}{3} = 1$

$\therefore \quad k = 3$

b) $20 \text{ sec} = \dfrac{20}{60} = \dfrac{1}{3} \text{ min}$

Therefore, probability that the lecture ends within 20 sec. *i.e.,* $\frac{1}{3}$ min of bell ring is

$$P\left(x < \frac{1}{3}\right) = \int_0^{1/3} f(x)\,dx = \int_0^{1/3} 3x^2\,dx = \left[\frac{3x^3}{3}\right]_0^{1/3}$$

$$= \left(\frac{1}{3}\right)^3 - 0 = \frac{1}{27}$$

c) Probability that the lecture continue beyond the bell ring between 20 sec to 40 sec is

$$P\left(\frac{1}{3} < x < \frac{2}{3}\right) = \int_{1/3}^{2/3} 3x^2\,dx = \left[\frac{3x^3}{3}\right]_{1/3}^{2/3} = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

5. Define gamma distribution and write the chief characteristics of it. [2072 Magh]

Solution: Please, see the 3.3.

6. In a photo graphic process, the developing time of prints may be looked upon as a random variable having the normal distribution with a mean of 16.28 second and a standard deviation of 0.12 second. Find the probability that it will take

a) Anywhere from 16.00 to 16.50 second to develop one of the prints.

b) At least 16.20 second to develop one of the prints

c) At most 16.35 seconds to develop one of the prints.

OR

The distribution of amount of the gravel (in ton) sold by a particular contribution supply company in a given week is continuous random variable x with the probability density function.

$$f(x) = \begin{cases} \left(\frac{3}{2}\right)(x^2+1) & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the cumulative distribution function of sales.

b) What is the expected value and variance of sales.

Solution:

Let x be the developing time of prints.

Mean (μ) = 16.28 sec

Standard deviation (σ) = 0.12 sec

Now, using normal distribution

i) $P(16.00 < x < 16.50)$

$$= P\left(\frac{16-16.28}{0.12} < \frac{x-\mu}{\sigma} < \frac{16.50-16.28}{0.12}\right)$$

$= P(-2.33 < Z < 0.018)$

$= P(0.018) - P(-2.33)$

$= 0.5080 - 0.0099$

$= 0.4981$

ii) $P(x \ge 16.20 \text{ sec})$

$$= P\left(\frac{x-\mu}{\sigma} < \frac{16.20-16.28}{0.12}\right)$$

$= P(Z \ge -0.67)$

$= 1 - P(Z < (-0.67))$

$= 1 - 0.254$

$= 0.7486$

iii) $P(x \le 16.35)$

$$= P\left(\frac{x-\mu}{\sigma} < \frac{16.35-16.28}{0.12}\right)$$

$= P(Z \le 0.58)$

$= 0.7190$

OR

Given that;

$$\text{pdf is } f(x) = \begin{cases} \left(\frac{3}{2}\right)(x^2+1) & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Now, we know;

a) Cumulative distribution function is given as

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(x)\,dx$$

When $x < 0$ or $x > 1$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} 0\,dx = 0$$

When $x > 0$ and $x < 1$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} 0\,dx + \int_0^1 \frac{3}{2}(x^2+1)\,dx$$

$$= \left[\frac{3}{2}\left(\frac{x^3}{3} + x\right)\right]_0^x$$

$$= \left[\frac{x^3}{2} + \frac{3x}{2}\right]_0^x$$

$$= \frac{x^3}{2} + \frac{3x}{2} = \frac{1}{2}(x^3+3x)$$

$$\text{Thus, } F(x) = \begin{cases} \frac{1}{2}(x^3+3x) & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

b) $$E(x) = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$= \int_0^1 x \times \frac{3}{2}(x^2+1)\,dx = \int_0^1 \frac{3}{2}(x^3+x)\,dx$$

$$= \frac{3}{2}\left[\frac{x^4}{4} + \frac{x^2}{2}\right]_0^1$$

$$= \frac{3}{2}\left[\frac{1}{4} + \frac{1}{2} - 0 - 0\right]$$

$$= \frac{3}{2} \times \frac{3}{4} = \frac{9}{8}$$

$$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

$$= \int_0^1 x^2 \times \frac{3}{2}(x^2+1)\,dx$$

$$= \frac{3}{2}\int_0^1 (x^4+x^2)\,dx = \frac{3}{2}\left[\frac{x^5}{5} + \frac{x^3}{3}\right]_0^1$$

$$= \frac{3}{2}\left[\frac{1}{5} + \frac{1}{3} - 0 - 0\right] = \frac{3}{10} + \frac{1}{2} = \frac{8}{10} = \frac{4}{5}$$

$$\therefore \text{Variance, } V(x) = E(x^2) - (E(x))^2 = \frac{4}{5} - \left(\frac{9}{8}\right)^2 = -0.4656$$

7. Define standard normal distribution and write the condition for normal approximation to the binomial and Poisson probability distribution.

OR

If the probability density function of a random variable is given by

$$f(x) = \begin{cases} kx^3 & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find i) the value of k

ii) $P\left(X > \frac{2}{3}\right)$

iii) $P\left(\frac{1}{4} < X < \frac{3}{4}\right)$ **and mean of distribution** **[2072 Aswin]**

Solution: Please, see the 3.2.1 and Q. No. 3.

Condition for normal approximation to the Poisson probability distribution is,

If the value of mean of the Poisson's distribution is large.

i.e., λ is large $X \sim P_0(\lambda)$

Then, $X \sim N(\lambda, \lambda)$ approximately

OR

Similarly to Q. No. 4 (OR question)

For (ii) find $P\left(x > \frac{2}{3}\right) = \int_{2/3}^{1} f(x)\,dx$

8. **Income of a group of 10,000 persons were found to be normally distributed with mean Rs. 1520 and standard deviation Rs. 160. Find the number of person whose incomes will be**

   i) Between Rs. 1400 and Rs. 1520

   ii) More than Rs. 1600

   iii) Lowest income of richest 100 persons **[2072 Ashwin]**

Solution:

Let x be the income of a group of 10,000 persons.

Mean (μ) = Rs. 1520

Standard deviation (σ) = Rs. 160

Now,

i) $P(1400 < x < 1520)$

$$= P\left(\frac{1400 - 1520}{160} < \frac{x - \mu}{\sigma} < \frac{1520 - 1520}{160}\right)$$

$$= P(-0.75 < Z < 0)$$

$$= P(0) - P(-0.75)$$

$$= 0.5 - 0.2266$$

$$= 0.2734$$

Therefore, required no. of people having income between Rs. 1400 and Rs. 1520 = 0.2734 × 10,000 = 273.4 persons = 274 person

ii) $P(x > 1600)$

$$= P\left(\frac{x - \mu}{\sigma} < \frac{1600 - 1520}{160}\right)$$

$$= P(Z > 0.5)$$

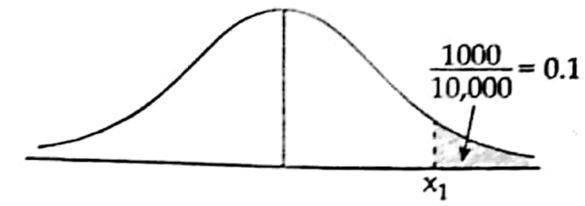
$$= 1 - P(Z \le 0.5)$$

$$= 1 - 0.6915 = 0.3085$$

Therefore, required no. of person having income greater than Rs. 1600 is 0.3085 × 10,000 = 3085 persons.

iii) Let x_1 be the lowest income of richest 1000 persons.

$P(x < x_1) = (1 - 0.1)$



or, $P\left(\frac{x - \mu}{\sigma} < \frac{x_1 - 1520}{160}\right) = 0.9$

or, $P\left(Z < \frac{x_1 - 1520}{160}\right) = 0.9$

The value of probability closet to 0.9 is 0.8997 at Z = 1.28

So, $\frac{x_1 - 1520}{160} = 1.28$

or, x_1 = Rs. 1724.8

9. **Define continuous random variable and probability density function. The probability density function of a random variable is**

$$f(x) = \begin{cases} \frac{3}{4}x & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Find mean and variance of the variable x. **[2071 Magh]**

Solution: Please, see the 3.1.1 and 3.1.2.

The given pdf is $f(x) = \begin{cases} \frac{3}{4}x & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$

We know;

$\therefore$ Mean $E(x) = \int_{-\infty}^{\infty} x\,f(x)\,dx$

$$= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{2} f(x)\,dx + \int_{2}^{\infty} f(x)\,dx$$

$$= 0 + \int_{0}^{2} \frac{3}{4}x\,dx + 0$$

$$= \frac{3}{4}\left[\frac{x^2}{2}\right]_0^2 = \frac{3}{4}\left[\frac{4}{2} - 0\right]$$

$$= 1.5$$

$\therefore$ $E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx$

$$= \int_{0}^{2} x^2 \times \frac{3}{4}x\,dx = \frac{3}{4}\int_{0}^{2} x^3\,dx$$

$$= \frac{3}{4}\left[\frac{x^4}{4}\right]_0^2 = \frac{3}{4}\left[\frac{16}{4} - 0\right]$$

$$= 3$$

$\therefore$ Variance $V(x) = E(x^2) - (E(x))^2 = 3 - (1.5)^2 = 0.75$

10. The time for a super glue to set can be treated as a random variable having a normal distribution with mean 30 seconds. Find its standard deviation if the probability is 0.20 that it will take on a value greater than 39.2 second.

OR

In a certain city daily consumptions of water (in millions of gallon) follows a gamma distributions with $\alpha = 2$ and $\beta = 3$. If the daily capacity of city is 9 million gallons of water, show that the probability on any given day the water supply is inadequate is $4e^{-3}$.

Solution:

Let x be the normal variable that denote the time to set for a super glue.

- Mean (μ) = 30 seconds

  Standard deviation (σ) = ?

We have;

$$P(x > 39.2) = 0.20$$

or, $$P\left(\frac{x-\mu}{\sigma} > \frac{39.2-30}{\sigma}\right) = 0.20$$

or, $$P\left(Z > \frac{9.2}{\sigma}\right) = 0.20$$

or, $$1 - P\left(Z \le \frac{9.2}{\sigma}\right) = 0.8$$

The value of probability closet to 0.8 is 0.7995 at Z = 0.84.

So, $$\frac{9.2}{\sigma} = 0.84$$

or, $$\sigma = 10.95$$

OR

Let x follows the gamma distribution that represents the daily consumption of water in the city with parameter $\alpha = 2$ and $\beta = 3$.

From gamma distribution.

We have;

$$f(x) = \frac{1}{B^{\alpha}\Gamma\alpha} x^{\alpha-1} e^{-x/\beta} \qquad \text{for } x > 0, \alpha > 0, \beta > 0$$

$$= \frac{1}{3^2\,\Gamma 2} x^{2-1} e^{-x/3}$$

$$\therefore \quad f(x) = \frac{1}{9} x e^{-x/3} \qquad \text{for } x > 0$$

Now, probability that on any given day the water supply is inadequate is given as

$$P(x > 9) = 1 - P(x \le 9)$$

$$= 1 - \int_0^9 \frac{1}{9} x e^{-x/3}\, dx$$

$$= 1 - \frac{1}{9}\left[\frac{xe^{-x/3}}{-(1/3)} - \frac{e^{-x/3}}{(-1/3)^2}\right]_0^9$$

$$= 1 - \frac{1}{9}[-27e^{-3} - 9e^{-3} - 0 + 9]$$

$$= 1 - [-3e^{-3} - e^{-3} + 1]$$

$$= 3e^{-3} + e^{-3}$$

$$= 4e^{-3}$$

11. Define standard normal distribution. Write the normal approximation to the

a) Binomial distribution

b) Poisson distribution

OR

A college professor never finishes his lecture before the bell rings to end the period, and always finishes his lectures within one minute after the bell rings. Let x be the time which elapse between the bell and the end of the lecture. Suppose that the pdf of x is

$$f(x) = \begin{cases} kx^2 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the value of k

b) What is the probability that the lecture ends within $\frac{1}{2}$ min the bell ringing?

c) What is the probability that the lecture continuous beyond the bell ring for between 15 and 30 seconds?

d) What is the probability that the lecture continuous for at least 40 seconds beyond the bell?

Solution: Please, see the Q. No. 7.

OR

Similar to Q. No. 4.

b) $\int_0^{1/2} f(x)\, dx$

c) $\int_{1/4}^{1/2} f(x)\, dx$

d) $\int_{2/3}^{1} f(x)\, dx$

12. Suppose that the pH of soil samples taken from a certain geographic region is normally distributed with mean pH 6.00 and standard deviation 0.10. If the pH of a randomly selected soil sample from this region is determined.

a) What is the probability that the resulting pH is between 5.90 and 6.15?

b) What is the probability that the resulting pH exceeds 6.10?

c) **What is the probability that the resulting pH is at most 5.95?** **[2071 Bhadra]**

Solution:

Let x follow the normal distribution that denote the pH of soil samples

Mean (μ) = 6

Standard deviation (σ) = 0.10

Now, we know;

a) $P(5.90 < x < 6.15)$

$$= P\left(\frac{5.90-6}{0.1} < \frac{x-\mu}{\sigma} < \frac{6.15-6}{0.1}\right)$$

$= P(-1 < Z < 1.5)$

$= P(1.5) - P(-1)$

$= 0.9332 - 0.1587$

$= 0.7745$

b) $P(x > 6.10)$

$$= \left(\frac{x-\mu}{\sigma} < \frac{6.10-6}{0.1}\right)$$

$= P(Z > 1)$

$= 1 - P(Z \le 1)$

$= 1 - 0.8413$

$= 0.1587$

c) $P(x \le 5.95)$

$$= \left(\frac{x-\mu}{\sigma} \le \frac{5.95-6}{0.1}\right)$$

$= P(Z \le -0.5)$

$= 0.3085$

13. **Write three important properties of normal distribution. A certain type of storage battery lasts, on average, 3.0 years with a standard deviation of 0.5 years. Assuming that battery life is normally distributed, find the probability that a given battery will last less than 2.3 years.** **[2073 Magh]**

Solution: Please, see the 3.2.

Let x follow normal distribution that denote the battery life in years.

Mean (μ) = 3.0 years

Standard deviation (σ) = 0.5 years

Now, we know;

$P(x < 2.3)$

$$= P\left(\frac{x-\mu}{\sigma} < \frac{2.3-3}{0.5}\right)$$

$= P(Z < -1.4)$

$= 0.0808$

14. **A random variable x has the probability density function $f(x)$ as**

$$f(x) = \begin{cases} kxe^{-x^2/2} & \text{for } 0 \le x < \infty \\ 0 & \text{otherwise} \end{cases}$$

i) **Find the value of constant k $f(x)$ is probability density function**

ii) **Find the distribution function F(x)**

iii) **Also find the mean of random variable x.** **[2073 Bhadra]**

Solution:

Given that;

The pdf is

$$f(x) = \begin{cases} kxe^{-x^2/2} & \text{for } 0 \le x < \infty \\ 0 & \text{otherwise} \end{cases}$$

i) Since it is pdf so it must satisfy

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

or, $\int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx = 1$

or, $0 + \int_{0}^{\infty} kxe^{-x^2/2}\,dx = 1$

Let $\frac{x^2}{2} = y$ then $\frac{2x\,dx}{2} = dy$

or, $x\,dx = dy$ when $x \to 0, y \to 0$

when $x \to \infty, y \to \infty$

Then,

$$\int_{0}^{\infty} ke^{-y}\,dy = 1$$

or, $k[-e^{-y}]_0^{\infty} = 1$

or, $k \times [-0 + 1] = 1$

$\therefore$ $k = 1$

ii) $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{x} f(x)\,dx,$ where $x \ge 0$

$= 0 + \int_{0}^{x} xe^{-x^2/2}\,dx$

Let $\frac{x^2}{2} = y$ then $\frac{2x\,dx}{2} = dy$

or, $x\,dx = dy$

Then,

$$\int_{0}^{\sqrt{2y}} e^{-y}\,dy$$

$= -e^{\sqrt{2y}} + 1 = -1 - e^{-x^2/2}$

$$\therefore \quad F(x) = \begin{cases} 1 - e^{-x^2/2} & \text{for } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

iii) Mean $E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

$= \int_{-\infty}^{0} x\, f(x)\, dx + \int_{0}^{\infty} x\, f(x)\, dx$

$= 0 + \int_{0}^{\infty} x \times e^{-x^2/2}\, dx$

$= \int_{0}^{\infty} x^2 e^{-x^2/2}\, dx = 0$ [Solve by using calculator]

15. The mean elongation of steel bar under a particular tensile load has been established to be normally distribution with parameter $\mu = 0.06$ and $\sigma = 0.008$. Assuming the same distribution applies to new bar, find the probability that the mean elongation fails

a) above 0.08

b) between 0.05 and 0.07

c) either below 0.045 or above 0.065

OR

The distribution of amount of the gravel (in ton) sold by a particular contribution supply company in a given week is continuous random variable x with the probability density function.

$$f(x) = \begin{cases} \left(\frac{3}{2}\right)(x^2+1) & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find the cumulative distribution function of sales.

b) What is the expected value and variance of sales.[2074 Bhadra]

Solution:

Given that;

Mean (μ) = 0.06

Standard deviation (σ) = 0.008

Now,

a) $P(x > 0.08)$

$= P\left(\frac{x-\mu}{\sigma} > \frac{0.08-0.06}{0.008}\right)$

$= P(Z > 2.5)$

$= 1 - P(Z \le 2.5)$

$= 1 - 0.9938$

$= 0.0062$

b) $P(0.05 < x < 0.07)$

$= P\left(\frac{0.05-0.06}{0.008} < \frac{x-\mu}{\sigma} < \frac{0.07-0.06}{0.008}\right)$

$= P(-1.25 < x < 1.25)$

$= P(1.25) - P(-1.25)$

$= 0.8944 - 0.1056$

$= 0.7888$

c) $P(x < 0.045) + P(x > 0.065)$

$= P\left(\frac{x-\mu}{\sigma} < \frac{0.045-0.06}{0.008}\right) + P\left(\frac{x-\mu}{\sigma} > \frac{0.065-0.06}{0.008}\right)$

$= P(Z < (-1.875)) + P(Z > 0.625)$

$= P(Z < (-1.875)) + 1 - P(Z \le 0.625)$

$= 0.0304 + 1 - 0.73405$

$= 0.29635$

OR

Please, see the Q. No. 6 (OR question).

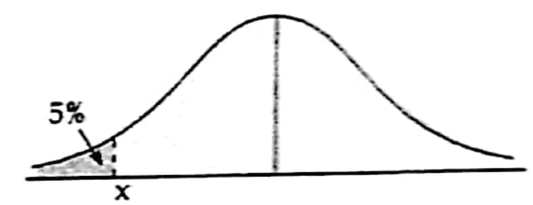
16. The most amount of distilled water dispensed by a certain machine is normally distributed with mean value 64 OZ and standard deviation 0.78 OZ. What container size will ensure that overflow occurs only 0.5% of the time? [2074 Bhadra]

Solution:

Let the volume of container size be x OZ

Mean (μ) = 64 OZ

Standard deviation (σ) = 0.78 OZ



Then, from normal distribution

$P(X < x) = 0.005$

or, $P\left(\frac{x-\mu}{\sigma} < \frac{x-64}{0.78}\right) = 0.005$

or, $\frac{x-64}{0.78} = (-2.575)$ $\left[\because \frac{(-2.57)+(-2.58)}{2} = -2.575\right]$

or, $x = 61.99$ OZ

17. Define the normal distribution. Given the condition for normal approximation of binomial distribution. [2075 Baisakh]

Solution: Please, see the Q. No. 3.

18. The mean elongation of steel bar under a particular tensile load has been established to be normally distribution with parameter $\mu = 0.06$ and $\sigma = 0.008$. Assuming the same distribution applies to new bar, find the probability that the mean elongation fails

a) above 0.08

b) between 0.05 and 0.07

c) either below 0.045 or above 0.065 [2075 Baisakh]

Solution: Please, see the Q. No. 15.

19. Score of 9000 candidates of IOE entrance examination is normally distributed with average score 50 and standard deviation of 20.

a) **Find the value of score exceeded by top 900 candidates.**
b) **Find the number of candidate who secured at least 90.**

[2075 Bhadra]

Solution:

Let x follow normal distribution that denotes the marks. We have,

Mean (μ) = 50

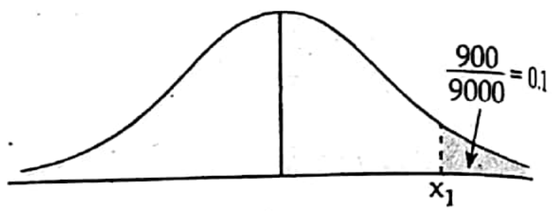
Standard deviation (σ) = 20

Now,

a) Let x_1 be the score exceeded by top 900 students.

$P(X > x_1) = 0.1$

or, $P\left(\frac{X-\mu}{\sigma} < \frac{x_1 - 50}{20}\right) = 0.1$



or, $P\left(Z > \frac{x_1 - 50}{20}\right) = 0.1$

or, $1 - P\left(Z \le \frac{x_1 - 50}{20}\right) = 0.1$

or, $P\left(Z \le \frac{x_1 - 50}{20}\right) = 0.9$

The closet value of probability 0.9 is 08997 at Z = 1.28.

So, $\frac{x_1 - 50}{20} = 1.28$

$\therefore \quad x_1 = 75.6$

b) $P(X \ge 90)$

$= P\left(\frac{X-\mu}{\sigma} < \frac{90-50}{20}\right)$

$= P(Z \ge 2)$

$= 1 - P(Z < 2)$

$= 1 - 0.9772$

$= 0.0228$

Therefore, the number of candidates who secured at least 90 is

$= 0.0228 \times 9000 = 228$ students

20. The mean elongation of steel bar under a particular tensile load has been established to be normally distribution with parameter $\mu = 0.06$ and $\sigma = 0.008$. Assuming the same distribution applies to new bar, find the probability that the mean elongation fails

a) **above 0.08**
b) **between 0.05 and 0.07**
c) **either below 0.045 or above 0.065**

[2075 Baisakh]

Solution: Please, see the Q. No. 15.

Chapter 4

SAMPLING DISTRIBUTION

4.1 Population and Sample 96
4.1.1 Parameters and Statistics 96
4.2 Central Limit Theorem 96
4.3 Sampling Distribution of Sample Mean 98
4.4 Sampling Distribution of Sampling Proportion 100
4.5 Some More Information of Sampling Distribution of Sample Mean 100

4.1 POPULATION AND SAMPLE

Population

A large group of individuals and objects whose characteristics are being studies is called as population.

Sample

A small part (or fraction) of the population which is selected in random to evaluate the approximate characteristics of the population is called a sample and the process of obtaining samples is called of sampling.

4.1.1 PARAMETERS AND STATISTICS

Parameters

The statistical measures computed from the population is called as parameters. They describe certain properties of population. In other words parameter is the function of population.

i.e., parameter = F (population)

Statistics

The statistical measures computed from the samples is called as statistics. They describe certain properties of samples.

In other words, statistics is the function of samples.

i.e., statistics = F(samples)

List of parameters and statistics

S.N.	Parameters	Statistics
1.	Population size = N	Sample size n
2.	Population mean = $\mu = \frac{\Sigma X}{N}$	Sample mean = $\bar{X} = \frac{\Sigma X}{n}$
3.	Population S.D. = $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	Sample S.D. = $S = \sqrt{\frac{\Sigma(x-\bar{X})^2}{n-1}}$
4.	Population variance = σ^2	Sample variance = S^2
5.	Population proportion = $P = \frac{X}{N}$	Sample proportion = $\hat{p} = \frac{x}{n}$

4.2 CENTRAL LIMIT THEOREM

Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample of size 'n' taken from any population with mean from any population with mean (μ) and variance (σ^2). If sample size is sufficiently large; then sampling distribution follows normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

i.e., $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ as $n \to \infty$.

Then, $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

Alternatives definition:

Let $X_1, X_2, \ldots\ldots, X_n$ be a sequence of independent and identically distributed random variable each having mean μ and variance σ^2. Then for n large the distribution of sum

$$S_n = X_1 + X_2 + \ldots\ldots + X_n$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

It follows from central limit theorem (CLT) that

$$Z = \frac{X_1 + X_2 + \ldots\ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

$$= \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

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

Example 1

Arbind bought some gas cylinder to cook food. Using those gas cylinders he found that each gas cylinder has a mean life of 3 months with a standard derivation of 15 days.

a) Find the probability that 5 cylinders will last for at least 1 year.

b) How many gas cylinder should he buy to be 95% sure that, they will last for at least 18 months.

Solution:

Here;

Let X_i denote the usage time of gas cylinder and S_5 denote the usage time of 5-cylinders.

Then, $S_5 = X_1 + X_2 + \ldots\ldots + X_5$

$E(S_5) = n\mu = 5 \times 3 = 15$ [μ = 3 months]

$V(S_5) = n\sigma^2 = 5 \times \left(\frac{1}{2}\right)^2 = 1.25$ $\left[\sigma = 15 \text{ days} = \frac{1}{2} \text{ month}\right]$

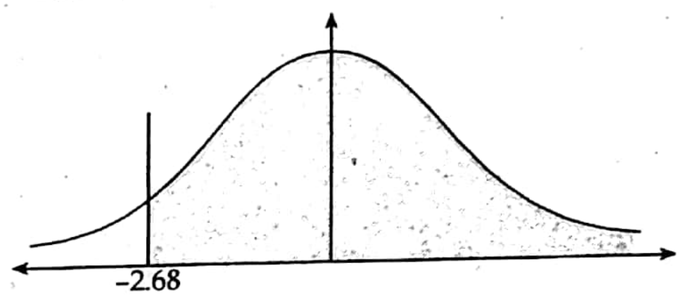
Using formula;

$$Z = \frac{S_5 - n\mu}{\sigma\sqrt{n}}$$

Using CLT to find probability that 5 cylinders will last for at least 1 year (=12 months)

$$P(S_5 \geq 12) = P\left(\frac{S_5 - n\mu}{\sigma\sqrt{n}} \geq \frac{12 - 5 \times 3}{\frac{1}{2} \times \sqrt{5}}\right)$$

$$= P(Z \geq -2.68)$$
$$= 1 - P(Z < -2.68)$$
$$= 1 - 0.0037$$
$$= 0.9963$$



Note: If we were to find probability of 5 cylinders will last for at most 1 year (=12 months) then

$P(S_5 \leq 12)$

b) Let 'n' be the numbers of cylinders required to be 95% sure that it will last for at least 18 months.

$$P(S_n \geq 18) = P\left(\frac{S_5 - n\mu}{\sigma\sqrt{n}} \geq \frac{18 - n \times 3}{\frac{1}{2} \times \sqrt{n}}\right)$$

or, $95\% = P\left(Z \geq \frac{18 - 3n}{0.5\sqrt{n}}\right)$

or, $1 - 0.95 = P\left(Z \leq \frac{18 - 3n}{0.5\sqrt{n}}\right)$.

or, $0.05 = P\left(Z \leq \frac{18 - 3n}{0.5\sqrt{n}}\right)$

Using probability table of normal distribution, the value of probability closer to 0.05 is 0.0505 at Z = –1.645. So,

$$\frac{18 - 3n}{0.5\sqrt{n}} = -1.645$$

Solving using calculator

$n = 6.71 \simeq 7$

Hence, he can buy 7 cylinders.

4.3 SAMPLING DISTRIBUTION OF SAMPLE MEAN

The frequency distribution or probability distribution of all possible values of sample means calculated from different samples drawn from sample population is called sampling distribution of same mean.

a) Sampling from an infinite population

Let $x_1, x_2, \ldots\ldots, x_n$ be a sample of values from this population. Let μ and σ^2 be population mean and population variance.

The sample mean is defined by

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

Since, the value of the sample mean $\bar{x}$ is determined by the values of the random variables in the sample, it follows that $\bar{x}$ is also a random variable. It's expected value and variance are obtained as follows:

$$E[\bar{x}] = E\left(\frac{x_1 + x_2 + \ldots\ldots + x_n}{n}\right)$$

$$= \frac{1}{n}(E[x_1] + E[x_2] + \ldots\ldots + E[x_n])$$

$$= \mu$$

and $$\text{var}(\bar{x}) = \text{var}\left(\frac{x_1 + x_2 + \ldots\ldots + x_n}{n}\right)$$

$$= \frac{1}{n^2}(\text{var}[x_1] + \text{var}[x_2] + \ldots\ldots + \text{var}[x_n]) \quad \text{(By independence)}$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Results

i) Mean $(\mu_x) = E(\bar{x}) = \mu$

ii) Variance $(\sigma_{\bar{x}}^2) = \text{var}(\bar{x}) = \frac{\sigma^2}{n}$

iii) Standard deviation (S.D.) of sample mean

$$(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\text{var}(\bar{x})}$$

Note: S.D. is also called standard error (S.E.)

Hence, S.E. of sample mean $= \frac{\sigma}{\sqrt{n}}$

b) Sampling from a finite population

i) Mean $(\mu_x) = E(\bar{x}) = \mu$

ii) variance $(\sigma_{\bar{x}}^2) = \text{var}(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N - n}{N - 1}\right)$

where, $\left(\frac{N - n}{N - 1}\right)$ is called finite population correction factor.

iii) S.D. $= \sqrt{\text{var}(\bar{x})} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}}$

Since, S.D. is also called standard error (S.E.)

$$\because \quad \text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$$

c) Sampling from two normal populations with mean $\bar{x}_1$ and $\bar{x}_2$, sizes n_1 and n_2 and standard deviations σ_1 and σ_2 then S.E. of difference of two means is

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If population has same S.D. *i.e.*, $(\sigma_1 = \sigma_2 = \sigma)$ then,

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

If population S.D. (σ) is unknown, it is estimated by sample S.D.(S)

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where, $S = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

In spite of calculating S_1 and S_2 separately using above formula, we can also use this formula

$$\text{S.E.}(\bar{x}_1 - \bar{x}_2) = S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where, $S = \sqrt{\dfrac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})}{n_1 + n_2 - 2}}$

4.4 SAMPLING DISTRIBUTION OF SAMPLING PROPORTION

a) **When population size is infinite**

Standard error (S.E.) $= \sqrt{\dfrac{pq}{n}} = \sqrt{\dfrac{p \cdot (1-p)}{n}}$ $\quad [\because p + q = 1]$

b) **When population size is finite**

Standard error (S.E.) $= \sqrt{\dfrac{pq}{n}\left(\dfrac{N-n}{N-1}\right)}$

4.5 SOME MORE INFORMATION OF SAMPLING DISTRIBUTION OF SAMPLE MEAN

a) Distribution without replacement

Formula to determine number of samples $= {}^{N}C_n$

where, N = Population size

n = Sample size

Example 2

Produce samples of size '2' from following population {3, 7, 11, 15}

Solution:

Number of samples $= {}^{N}C_n = {}^{4}C_2 = 6$

Samples	Means
(3, 7)	5
(3, 11)	7
(3, 15)	9
(7, 11)	9
(7, 15)	11
(11, 15)	13

Frequency distribution table

Mean	Frequency (F)	Probability
5	1	1/6
7	1	1/6
9	2	2/6
11	1	1/6
13	1	1/6

Population mean $(\mu) = \dfrac{3+7+11+15}{4} = 9$

Mean of sample means $E(\bar{x}) = \dfrac{5+7+9+9+11+13}{6} = 9$

i.e., $\mu = E(\bar{x})$ [verified]

Population variance $(\sigma^2) = \dfrac{\Sigma(x-\mu)^2}{N}$

$$= \frac{(3-9)^2 + (7-9)^2 + (11-9)^2 + (15-9)^2}{4} = 20$$

Sample variance $(\sigma_{\bar{x}}^2) = \dfrac{\Sigma F(x-\mu)^2}{n}$

$$= \frac{1 \cdot (5-9)^2 + 1 \cdot (7-9)^2 + 2 \cdot (9-9)^2 + 1 \cdot (11-9)^2 + 1 \cdot (13-9)^2}{6}$$

$= 6.667$

Now, $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right) = \dfrac{20}{2} \times \left(\dfrac{4-2}{4-1}\right) = 6.667$ (Verified)

b) With replacement

Number of samples $= N^n$

Example 3

Write samples of size 2 for population {3, 7, 11, 15}.

Solution:

Number of samples = $N^n = 4^2 = 16$

Sample	Mean ($\bar{x}$)	Sample	Mean ($\bar{x}$)
(3, 3)	3	(11, 3)	7
(3, 7)	5	(11, 7)	9
(3, 11)	7	(11, 11)	11
(3, 15)	9	(11, 15)	13
(7, 3)	5	(15, 3)	9
(7, 7)	7	(15, 7)	11
(7, 11)	9	(15, 11)	13
(7, 15)	11	(15, 15)	15

Frequency distribution table

Sample mean ($\bar{x}$)	Frequency (F)	Probability
3	1	1/16
5	2	2/16
7	3	3/16
9	4	4/16
11	3	3/16
13	2	2/16
15	1	1/16

Population mean $(\mu) = \dfrac{3+7+11+15}{4} = 9$

Mean of sample means $E(\bar{x}) = \dfrac{\Sigma F \cdot \bar{x}}{N}$

$$= \frac{(3\times1)+(5\times2)+(7\times3)+(9\times4)+(13\times2)+(11\times3)+(15\times1)}{16} = 9$$

i.e., $\mu = E(\bar{x})$ [verified]

Population variance $(\sigma^2) = \dfrac{\Sigma F \cdot (x-\mu)^2}{n}$

$$= \frac{1\cdot(3-9)^2 + (7-9)^2\cdot1 + 1\cdot(11-9)^2 + 1\cdot(15-9)^2\cdot1}{4} = 20$$

Variance of sample mean $V(\bar{x}) = \sigma_{\bar{x}}^2 = \dfrac{\Sigma F\cdot(\bar{x}-\mu)^2}{\Sigma F}$

$$= \frac{1(-6)^2 + 2\cdot(-4)^2 + 3\cdot(-2)^2 + 4\cdot0^2 + 3\times2^2 + 2\times4^2\times1\times6^2}{16} = 10$$

Also, $V(\bar{x}) = \sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n} = \dfrac{20}{2} = 10$ [Verified]

SOLVED QUESTIONS [FOR BCE/BME/BIE/BAM]

1. **What is sampling distribution? Construct frequency distribution table of sample mean in population 2, 4, 6, 8, 10 with sample size two. Also prove that sample mean is unbiased estimate of population mean.** **[4] [2071 Bhadra, BCE]**

Solution:

The frequency distribution or probability distribution of all possible values of the samples (like: mean mode, variance, etc). Calculated from different samples drawn from same population is called sampling distribution.

Numerical

Population size (N) = 5

Sample size (n) = 2

∴ number of samples without repetition

$= {}^NC_n = {}^5C_2 = 10$

Sample	Mean ($\bar{x}$)	Sample	Mean ($\bar{x}$)
(2, 4)	3	(4, 8)	6
(2, 6)	4	(4, 10)	7
(2, 8)	5	(6, 8)	7
(2, 10)	6	(6, 10)	8
(4, 6)	5	(8, 10)	9

Frequency distribution table

Sample mean ($\bar{x}$)	Frequency (F)	Probability
3	1	1/10
4	1	1/10
5	2	2/10
6	2	2/10
7	2	2/10
8	1	1/10
9	1	1/10

Now, Population mean $(\mu) = \dfrac{\Sigma x}{N} = \dfrac{2+4+6+8+10}{5} = 6$

Mean of sample means $E(\bar{x}) = \dfrac{\Sigma\bar{x}}{N} = \dfrac{3+4+5+5+(6\times2)+(7\times2)+8+9}{10}$

$= 6$

i.e., $\mu = E(\bar{x})$ (verified)

Hence, the sample mean is unbaised estimate of population mean.

2. Define central limit theorem. An auditor from a large credit card company knows that, on average the monthly balance of any given customer deviation is \$56. If the auditor audits 50 randomly selected accounts, sample average monthly balance is

a) below \$100

b) between \$100 and \$130 **[2+4] [2071 Bharda, BCE]**

Solution: Theory

For theory part refers 4.2 (to get full marks it is better to write both the definitions.)

Numerical

Here;

- Population size (N) = 50 > 30. (Hence it is large sample.)

  Population mean (μ) = \$112

  Population S.D. (σ) = \$56

a) $\bar{x} < \$100$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \frac{100 - 112}{56/\sqrt{50}} = -1.515$$

$$P(Z < -1.515) = \frac{0.0655 + 0.0643}{2} = 0.0649$$

b) $\$100 < \bar{x} < \$130$

$= P(\$100 < \bar{x} < \$130)$

$$= P\left(\frac{100 - \mu}{\sigma/\sqrt{n}} < Z < \frac{130 - \mu}{\sigma/\sqrt{n}}\right)$$

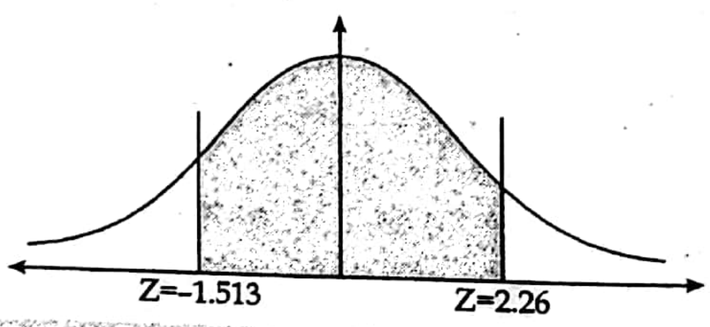
$$= \left(\frac{100 - 112}{56/\sqrt{50}} < Z < \frac{130 - 112}{56/\sqrt{50}}\right)$$

$= P(-1.513 < Z < 2.26)$

$= P(Z < 2.26) - P(Z < -1.513)$

$= 0.9881 - 0.0655$

$= 0.9226$



Note:

1. If sample size is small *i.e.*, n < 30 then even though the sample is normally distributed we use t-table.

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

but if sample is large *i.e.*, n > 30, use z-table.

i.e., $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

2. When working with each of individual value. We use

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

But in same question, if we deal with a group of people (samples) simultaneously. We use

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

3. Define sampling distribution. What is the difference between population and sample. A population consists of 3, 7, 11, 15. Consider all possible samples of size replacement from this population. Find population mean and population S.D.

[4] [2071 Magh]

Solution: Sampling distribution: Please, see the Q. No. (1)

Difference between population and sample: Please, see the definition part 4.1.

Numerical

Population size (N) = 4

Sample size (n) = 2

∴ Number of sample without repetition

$= {}^{N}C_{n} = {}^{4}C_{2} = 6$

Samples	Mean
(3, 7)	5
(3, 11)	7
(3, 15)	9
(7, 11)	9
(7, 15)	11
(11, 15)	13

a) Population mean (μ) $= \dfrac{\Sigma x}{N} = \dfrac{3 + 7 + 11 + 15}{4} = 9$

b) Population S.D. (σ) $= \sqrt{\dfrac{\Sigma(x - \mu)^2}{N}}$

$$= \sqrt{\frac{(3-9)^2 + (7-9)^2 + (11-9)^2 + (15-9)^2}{4}}$$

$= 20$

4. State central limit theorem. If a 1-gallon of paint covers on average 513.3 sq. ft. with a S.D. of 31.5 sq. ft., what is the probability that the sample mean area covered by a sample of 40 of these 1-galloons can be anywhere from 510 to 520 sq. ft.? [6] [2071 Magh]

Solution: Theory: For question see the 4.2.

Numerical

Population mean (μ) = 513.3 sq. ft.

Population S.D. (σ) = 31.5 sq. ft.

Population size (N) = 40 > 30

So these sample is large and we use Z-distribution.

a) $P(510 < \bar{x} < 520)$

$$= P\left(\frac{510-\mu}{\sigma/\sqrt{N}} < Z < \frac{520-\mu}{\sigma/\sqrt{N}}\right)$$

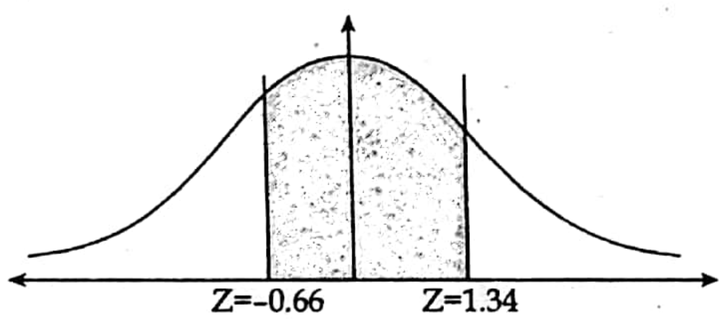
$$= \left(\frac{510-513.3}{31.5/\sqrt{40}} < Z < \frac{520-513.3}{31.5/\sqrt{40}}\right)$$

$= P(-0.66 < Z < 1.34)$

$= P(Z = 1.34) - P(Z = -0.66)$

$= 0.9099 - 0.2546$

$= 0.6553$



Hence, it is 65.53% sure that 40 of these 1-galloons can be between 510 to 520 sq. ft.

5. What do you mean by sampling distribution of a statistical and standard error? [5] [2072 Ashwin]

Solution:

A sampling distribution is a probability distribution of a statistic samples drawn from a specific population. On other words, the sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for statistic of a population.

a) Sampling distribution of a statistical error:

The sampling distribution of statistical error for a given population is the distribution of frequencies of a range of different error (*i.e.* difference between sample and population results). For different samples of a population.

Mathematically,

Statistical error = $\bar{x} - \mu$

In other words statistical error can also be defined as sampling error.

For example: for a population {1, 11, 21, 41}

We choose a sample of size 2 without repetition. Then we obtain samples as: {(1, 11), (1, 21), (1, 41), (11, 21), (11, 41) (21, 41)}.

$$\text{Population mean } (\mu) = \frac{1+11+21+41}{4} = 18.5$$

But we choose sample as (21, 41) then

$$\text{Sample mean } (\bar{x}) = \frac{21+41}{2} = 31$$

i.e., sampling error or statistical error

$= 31 - 18.5 = 12.5$

Hence, we obtain different statistical error for each samples. The set of all those error arranged in tabular from, with their frequency is sampling distribution of statistical error.

b) Sampling distribution of a standard error

The sampling distribution of a standard error for a given population is the distribution of frequencies of different standard deviation of sample mean of different sizes of a population.

The standard deviation of a sample mean ($\bar{x}$) is called as standard error of sample mean.

a) S.E. of sample mean $= \frac{\sigma}{\sqrt{n}}$ (if population is very large)

$$= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

N = Population size, n = Sample size

In above formulas, for a given population N is fixed but n changes as sample size changes and hence standard error changes.

Hence, for different sample size (from 1 to N) S.E. of samples are calculated and arranged in tabular from along with their frequencies to get sampling distribution of standard error.

6. From a population of 5 members 3, 6, 9, 12, 15 draw all possible random sample of size 3 without replacement. Obtain the sampling distribution of sample mean and calculate expectation of sample mean. [5] [2072 Ashwin]

Solution:

Here;

Population size (N) = 5

Sample size (n) = 3

∴ Number of samples without replacement

$= {}^{N}C_n = {}^{5}C_3 = 10$

Samples	Mean
(3, 6, 9)	6
(3, 6, 12)	7
(3, 6, 15)	8
(3, 9, 12)	8
(3, 9, 15)	9
(3, 12, 15)	10
(6, 9, 12)	9
(6, 12, 15)	11
(6, 9, 15)	10
(9, 12, 15)	12

Frequency distribution table

Sample mean ($\bar{x}$)	Frequency (F)	Probability
6	1	1/10
7	1	1/10
8	2	2/10
9	2	2/10
10	2	2/10
11	1	1/10
12	1	1/10

a) Population mean $(\mu) = \frac{\Sigma X}{N} = \frac{3+6+9+12+15}{5} = 9$

b) Expectation of sample means $E(\bar{x}) = \frac{\Sigma F\bar{x}}{\Sigma F}$

$= \frac{(6\times1)+(7\times1)+(8\times2)+(9\times2)+(10\times2)+(11\times1)+(12\times1)}{10}$

$= 9$

7. Define sampling distribution of proportion with suitable example. **[5] [2072 Magh]**

Solution:

The sampling distribution of proportion of a particular population is the probability distribution of all possible sample proportions of same sample of size 'n'.

Population proportion $(p) = \frac{x}{N}$

where, x = number of elements in population, n = population size

Sample proportion $(\hat{p}) = \frac{y}{n}$

where, y = number of elements in sample, n = sample size

For large sample, $\hat{p} \approx p$ and $\hat{q} \simeq q$.

$p + q = 1$

a) For large sample (*i.e.*, $N \to \infty$) [with repetition]

$$S.E.(\hat{p}) = \sqrt{V(\hat{p})} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

If population proportion is not known, we can use

$$S.E.(\hat{p}) = \sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$$

b) For small sample. [without repetition]

$$S.E.(\hat{p}) = \sqrt{\frac{pq}{n}\cdot\left(\frac{N-n}{N-1}\right)} = \sqrt{\frac{p(1-p)}{n}\cdot\left(\frac{N-n}{N-1}\right)}$$

c) Mean of sample proportion $E(\hat{p}) = p$

Hence, sample proportion $(\hat{p})$ is unbiased estimator of the population proportion (p).

For example; **A random of 100 CFL bulbs were taken from 500 bulbs out of which 15 bulbs were found to be defective. Determine S.E. of proportion of fused CFL bulbs.**

Solution:

Here;

Population size (N) = 500

Sample size (n) = 100

Defective bulbs in sample (x) = 15

Proportion of defective bulbs in the sample $(\hat{p}) = \frac{x}{n} = \frac{15}{100} = 0.15$

So, $\hat{q} = 1 - \hat{p} = \frac{85}{100}$ 0.85

For large sample, $p \approx \hat{p}$ and $q \simeq \hat{q}$

So, $S.E.(\hat{p}) = \sqrt{\frac{pq}{n}\cdot\left(\frac{N-n}{N-1}\right)} = \sqrt{\frac{0.15\times0.85}{100}\times\left(\frac{500-100}{500-1}\right)} = 0.032$

8. A population consists of the four numbers 2, 3, 4, 5

i) Write down all possible sample size of two without replacement.

ii) Verify that population mean is equal to the mean of sample mean.

iii) Calculate the standard error of the sampling distribution of the sample mean. [5] (2072 Magh)

Solution:

Here;

i) Population size (N) = 4

Sample size (n) = 2

Therefore, number of samples without replacement

$= {}^{N}C_n = {}^{4}C_2 = 6$

Samples	Mean
(2, 3)	2.5
(2, 4)	3
(2, 5)	3.5
(3, 4)	3.5
(3, 5)	4
(4, 5)	4.5

ii) Frequency distribution table

Sample mean ($\bar{x}$)	Frequency (F)	Probability
2.5	1	1/6
3	1	1/6
3.5	2	2/6
4	1	1/6
4.5	1	1/6

$$\text{Population mean } (\mu) = \frac{\Sigma X}{N} = \frac{2+3+4+5}{4} = 3.5$$

$$\text{Means of sample means } E(\bar{x}) = \frac{\Sigma Fx}{\Sigma F}$$

$$= \frac{(1 \times 2.5) + (1 \times 3) + (2 \times 3.5) + (1 \times 4) + (1 \times 4.5)}{6} = 3.5$$

Hence, $\mu = E(\bar{x})$ (verified)

iii) Standard error of sampling distribution of sample mean

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$\text{Population S.D. } (\sigma) = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$$= \sqrt{\frac{(2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2}{4}}$$

$\therefore \quad \sigma = 1.118$

$$\text{Hence, } S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{1.118}{\sqrt{2}} \times \sqrt{\frac{4-2}{4-1}} = 0.645$$

9. **Define standard error of mean. Write the expression for standard error of mean when**

   i) Sample is drawn from large population

   ii) Sampling is done without replacement from a finite population.

   [5] [2073 Bhadra]

Solution:

The standard deviation of sample mean ($\bar{x}$) is called as standard error of mean.

i) Expression for standard error of mean when sampling is done from large population

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

where, σ = S.D. of population $= \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

n = sample size

ii) Expression for S.E. of mean when sampling is done without replacement from a finite population.

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

where, σ = S.D. of population $= \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

N = population, n = sample size

$\sqrt{\frac{N-n}{N-1}}$ = finite population correction factor as $N \to \infty$, $\frac{N-n}{N-1} \to 1$.

10. **A population consists of four members 12, 13, 14, 15**

    i) Write all possible sample of size two without replacement

    ii) Verify that sample mean is unbiased estimator for population mean

    iii) Calculate the S.E. of mean [5] [2073 Bhadra]

Solution:

Here;

i) Population size (N) = 4

Sample size (n) = 12

∴ Number of samples without replacement

$= {}^{N}C_n = {}^{4}C_2 = 6$

Samples	Mean
(12, 13)	12.5
(12, 14)	13
(12, 15)	13.5
(13, 14)	13.5
(13, 15)	14
(14, 15)	14.5

ii) Frequency distribution table

Sample mean ($\bar{x}$)	Frequency (F)	Probability
12.5	1	1/6
13	1	1/6
13.5	2	2/6
14	1	1/6
14.5	1	1/6

$$\text{Population mean } (\mu) = \frac{\Sigma X}{N} = \frac{12+13+14+15}{4} = 13.5$$

$$\text{Means of sample means } E(\bar{x}) = \frac{\Sigma F\bar{x}}{\Sigma F}$$

$$= \frac{(1\times 12.5)+(1\times 13)+(2\times 13.5)+(1\times 14)+(1\times 14.5)}{6} = 13.5$$

$\mu = E(\bar{x})$ (verified)

Hence, sample mean is unbiased estimator for population mean.

iii) $$\text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

$$\text{Population S.D. } (\sigma) = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$$= \sqrt{\frac{(12-13.5)^2+(13-13.5)^2+(14-13.5)^2+(15-13.5)^2}{4}}$$

$$= \sqrt{\frac{1.5^2+0.5^2+0.5^2+1.5^2}{4}}$$

$\therefore\ \sigma = 1.12$

$$\text{Hence, S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

$$= \frac{1.12}{\sqrt{2}}\times\sqrt{\frac{4-2}{4-1}} = 0.6466$$

11. Write the difference between sample and population write three importance of central limit theorem. [5][2073 Magh]

Solution:

For difference between sample and population refer theory 4.1.

Importance of central limit theorem are:

i) The central limit theorem states that if you have a population with mean (μ) and standard deviation (σ) and take sufficiently large random samples from the population with replacement, then the distribution of sample means will be approximately normally distributed. This will hold true regardless the population is normal, or skewed or unevenly distributed; provided sample size is sufficiently large (*i.e.*, $n > 30$). If sample size is small but normally distributed and min $(np) \geq 5$; then also CLT holds very accurate.

$$\mu_{\bar{x}} = \mu \Rightarrow E(\bar{x}) = \mu$$

$$\text{and, } \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

ii) For large value of 'n' the distribution of standardized sample mean is approximately normal distribution. Hence, we can use test statistics as

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

iii) For large 'n' the distribution of sum $s_n = x_1 + x_2 + x_3 + \cdots\cdots + x_n$ is approximately normal with mean $n\mu$ and S.D. $= \sigma\sqrt{n}$.

$$Z = \frac{s_n - \text{mean}}{\text{S.D.}} = \frac{x_1+x_2+x_3+\cdots\cdots+x_n - n\mu}{\sigma\sqrt{n}}$$

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

12. Define parameter and statistic with examples. [4][2074 Bhadra]

Solution: Please, see the theory 4.1.1.

13. A population consists of the four members 5, 6, 9, 12

a) Write down all possible sample size of two without replacement.

b) Verify that the population mean is equal to the mean of the sample mean.

c) Calculate the standard error of the sampling distribution of the sample mean. [6][2074 Bhadra]

Solution:

Here;

a) Population size (N) = 4

Sample size (n) = 2

∴ Number of sample of size 2 without replacement

$= {}^{N}C_{n} = {}^{4}C_{2} = 6$

Samples	Mean
(5, 6)	5.5
(5, 9)	7
(5, 12)	8.5
(6, 9)	7.5
(6, 12)	9
(9, 12)	10.5

ii) Frequency distribution table

Sample mean ($\bar{x}$)	Frequency (F)	Probability
5.5	1	1/6
7	1	1/6
7.5	1	1/6
8.5	1	1/6
9	1	1/6
10.5	1	1/6

Population mean $(\mu) = \frac{\Sigma X}{N} = \frac{5+6+9+12}{4} = 8$

Means of sample means $E(\bar{x}) = \frac{\Sigma F\bar{x}}{\Sigma F}$

$$= \frac{(1\times 5.5) + (1\times 7) + (1\times 7.5) + (1\times 8.5) + (1\times 9) + (1\times 10.5)}{6} = 8$$

Here, $\mu = E(\bar{x})$

Therefore, population mean is equal to the mean of the sample mean.

c) Population S.D. $(\sigma) = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

$$= \sqrt{\frac{(5-8)^2 + (6-8)^2 + (9-8)^2 + (12-8)^2}{4}}$$

s $= 2.74$

S.E. of mean S.E. $(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

$$= \frac{2.74}{\sqrt{2}} \times \sqrt{\frac{4-2}{2-1}}$$

$= 2.74$

14. Define sampling distribution of proportion with example? **[4] [2075 Baishakh]**

Solution: Please, see the Q. No. 7.

15. A population consists of the four members 12, 13, 14, 15

i) Write down all possible sample size of two without replacement.

ii) Verify that the population mean is equal to the mean of the sample mean.

iii) Calculate the standard error of the sampling distribution of the sample mean. **[6] [2075 Baishakh]**

16. Define the following terms:

i) Parameter

ii) Statistics

iii) Sampling distribution of statistics

iv) Standard error of statistics

Solution: Please, see the theory section.

17. A population consists of four numbers 5, 6, 9, 12.

i) Write down all possible sample size of two without replacement.

ii) Verify that the population mean is equal to the mean of sample mean.

iii) Calculate the standard error of the sampling distribution of sample mean. **[6] [2075 Bhadra]**

Solution: Please, see the previous solution.

SOLVED QUESTIONS (FOR BCT/BEX/BEL/B. Agri)

1. **Define sampling distribution of proportion with example.** [4] [2071 Shrawan]

Solution: Please, see the Q. No. (7) of BCE.

2. **State central limit theorem. An electrical firm manufactures light bulbs that have a length of life that is actually normally distributed with mean deviation of 4 hours. Find the probability that a random sample of 16 bulbs will have an average life less than 12775 hours.** [5] [2071 Chaitra]

Solution: For theory question refer 4.2.

Numerical

Population mean (μ) = 800

Population S.D. (σ) = 4

Sample size (n) = 16

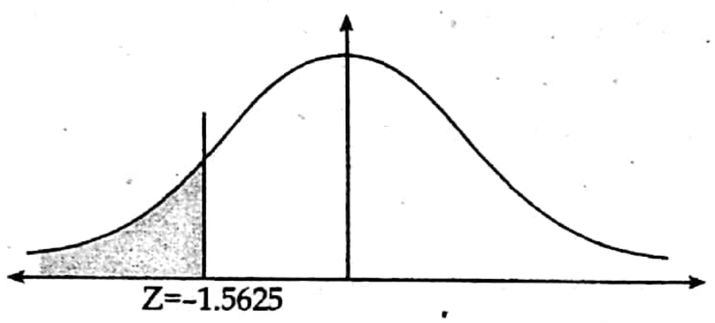
Let S_{16} represent the usage time of then, S_{16} = 12775 hours

Using alternative definition of CLT, we get

$$P(S_{16} < 12775) = P\left(\frac{S_{16} - n\mu}{\sigma\sqrt{n}} < \frac{12775 - 16 \times 800}{4 \times \sqrt{16}}\right)$$

$$= P(Z < -1.5625)$$

$$= 0.0594 \text{ [By Z-table]}$$



Hence, it is only 5.94% certain that 16 bulbs will have an average life less than 12775 hours.

Note: For more practice solve the solved example in section 4.2.

3. **What do you mean by sampling distribution of sample mean and its standard error? What would be the variance of sampling distribution of mean if sample is taken from finite population.** [3+1] [2071 Chaitra]

Solution: For 1st part see the 4.3 (a).

For 2nd part see the 4.3 (b).

4. **Define central limit theorem. An electronics company manufactures resistors that has mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of 25 resistors will have an average resistance less than 95 ohms.** [5] [2072 Kartik]

Solution: For theory see the 4.2.

Numerical

Population mean (μ) = 100

Population S.D. (σ) = 10

Sample size (n) = 25

Sample mean ($\bar{x}$) = 95

Using definition of CLT, we get

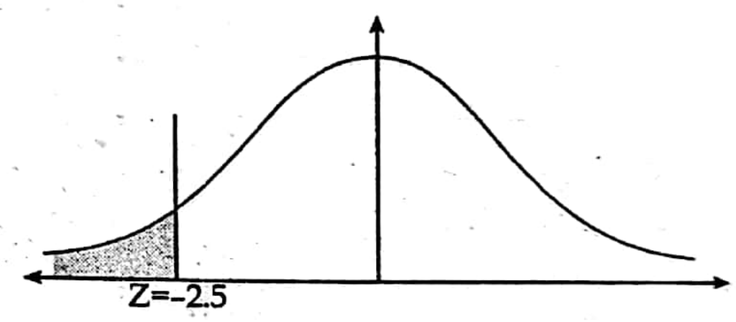
$$P(\bar{x} < 95)$$

$$= P\left(\frac{95 - \mu}{\sigma\sqrt{n}} > Z\right) \qquad \left[\because Z = \frac{\bar{x} - \mu}{\sigma\sqrt{n}}\right]$$

$$= P\left(Z < \frac{95 - 100}{10\sqrt{25}}\right)$$

$$= P(Z < -2.5)$$

$$= 0.0054 \text{ (using Z-table)}$$



Hence, it is only 0.54% sure that a random of 25 samples of resistors selected, will have an average resistance less than 95 ohms.

5. **Define standard error of mean. A population consists of four members 12, 19, 13, 16**

   i) **Write down all possible sample size of two without replacement.**

   ii) **Calculate S.E. ($\bar{x}$).** [5] [2072 Kartik]

Solution:

The standard deviation of sample mean is called standard error of mean. Mathematically, it is given by

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}} \qquad \text{[for finite population]}$$

N = population size

n = sample size

σ = population S.D.

As $N \to \infty$ [for infinite population]

$$\lim_{N\to\infty} \frac{N-n}{N-1}$$

$$= \lim_{N\to\infty}\left(\frac{1-\frac{n}{N}}{1-\frac{1}{N}}\right) = \frac{1-0}{1-0} = 1$$

Hence, above formula becomes

$$\text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Numerical

Here, Population size (N) = 4

Sample size (n) = 2

∴ Number of sample of size 2 without replacement

$= {}^{N}C_n = {}^{4}C_2 = 6$

Samples	Mean
(12, 19)	15.5
(12, 13)	12.5
(12, 16)	14
(19, 13)	16
(19, 16)	17.5
(13, 16)	14.5

ii) Population mean $(\mu) = \frac{\Sigma X}{N} = \frac{12+19+13+16}{4} = 15$

$$\text{Population S.D. } (\sigma) = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$$= \sqrt{\frac{(12-15)^2+(19-15)^2+(13-15)^2+(16-15)^2}{4}}$$

$$= 2.74$$

$$\text{S.E. of mean S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$

$$= \frac{2.74}{\sqrt{2}} \times \sqrt{\frac{4-2}{4-1}}$$

$$= 1.58$$

6. What do you mean by sampling distribution of a sample mean and its standard error? Explain with example. What would be the variance of sampling distribution of mean, if sample is taken from finite population. **[5] [2072 Chaitra]**

Solution:

The frequency distribution or probability distribution of all possible values of sample means calculated from different samples drawn from same population is called as sampling distribution of sample mean.

The standard deviation of sample mean is called standard error of mean. Mathematically, it is given by

$$\text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}} \quad \text{[for finite population]}$$

where, N = population size

n = sample size

σ = population S.D.

As $N \to \infty$ [for finite population]

$$\lim_{N\to\infty}\sqrt{\frac{N-n}{N-1}}$$

$$= \lim_{N\to\infty}\sqrt{\frac{1-\frac{n}{N}}{1-\frac{1}{N}}} = \sqrt{\frac{1-0}{1-0}} = 1$$

Hence, for large population, above formula becomes.

$$\text{S.E. } (\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

For explaining with example see the 4.5 (a).

For 2nd question see the 4.4.

7. Define the central limit theorem? A sample of 100 mobile battery cells tasted to find the length of life produced the following results as mean 13 months and S.D. of 3 months. Assuming the data to be normally distributed by using central limit theorem what percentage of battery cells expected to have battery life?

i) More than 15 months

ii) Less than 9 month **[5] [2072 Chaitra]**

Solution: For theory: Please, see the section 4.2.

Numerical

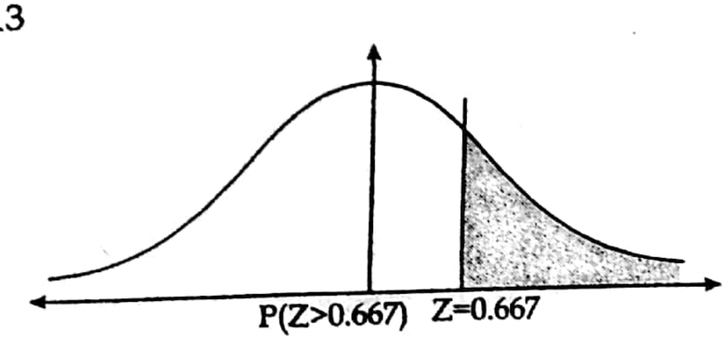
Here, Population mean (μ) = 13

Population S.D. (σ) = 3

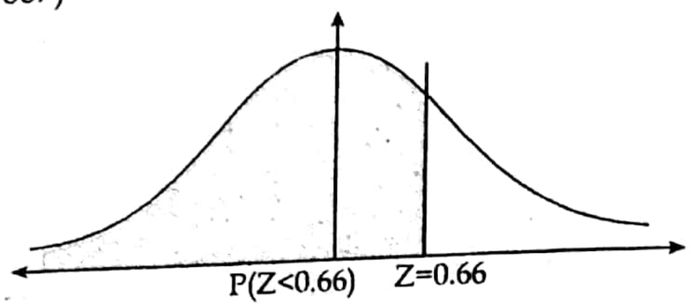
i) $P(\bar{x} > 15)$

$$P\left(Z > \frac{\bar{x}-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{15-13}{3}\right)$$



$= P(Z > 0.667)$



$= 1 - P(Z < 0.667)$

$= 1 - 0.7454$ [using Z table, $P(Z = 0.66) = 0.7454$]

$= 0.2546$

It is 25.46% certain that each bulbs has life more than 15 months

ii) $P(\bar{x} < 9)$

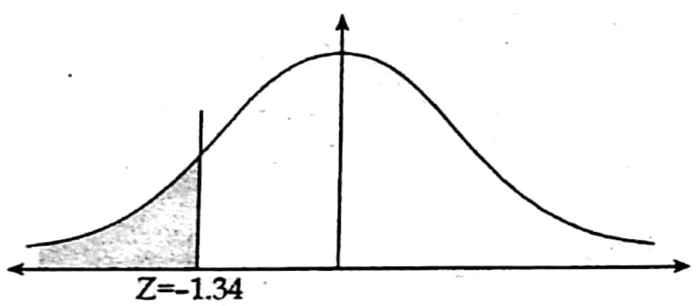
$= P\left(Z < \frac{\bar{x} - \mu}{\sigma}\right)$

$= P\left(Z < \frac{9 - 13}{3}\right)$

$= P\left(Z < -\frac{4}{3}\right)$

$= P(Z < -1.34)$

$= 0.0901$



Therefore, it is only 9% certain that each bulbs has life less than 9 months.

8. **State central limit theorem with an example. Explain why is it important on engineering.** **[5] [2073 Shrawan]**

Solution: For 1st question

Please, see the 4.2 with example.

For 2nd question

Please, see the Q. No.: 11 (2073 Magh) of BCE.

9. **A population consists of the four numbers 2, 8, 14, 20.**
   i) **Write down all possible sample size of two without replacement**
   ii) **Verify that the population mean is equal to the mean of the sample mean.**
   iii) **Calculate the standard error of the sample distribution of the sample mean.** **[5] [2073 Shrawan]**

Solution: Please, see the Q. No. 8, 10, 13 of BCE.

10. **What do you mean by central limit theorem and discuss its applications.** **[4] [2073 Chaitra]**

Solution: Please, see the Q. No. 11 of BCE.

11. **An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equal to 800 hours and standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of**
   a) **less than 850 hours**
   b) **between 750 and 900 hours** **[6] [2073 Chaitra]**

Solution:

Here;

Population mean (μ) = 800 hrs

Population S.D. (σ) = 40 hrs

Sample size (n) = 16

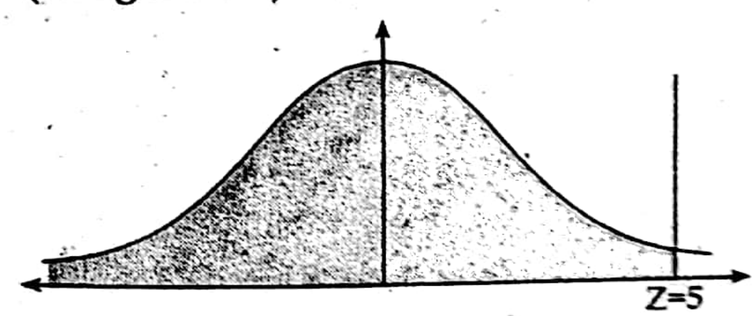
a) $P(\bar{x} < 850)$

$= P\left(Z < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)$

$= P\left(Z < \frac{850 - 800}{40/\sqrt{16}}\right)$

$= P(Z < 5)$

$= 0.999997$ (using Z-table)



Therefore, it is 99.9997% sure that random sample of 16 bulbs will have an average life less than 850 hours.

b) $P(750 < \bar{x} < 900)$

$= \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)$

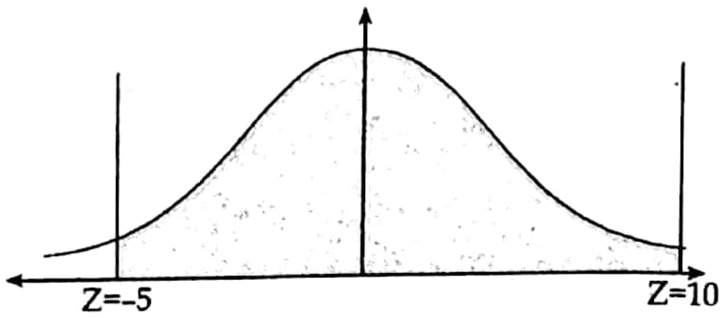
$= P\left(\frac{750 - 800}{40/\sqrt{16}} < Z < \frac{900 - 800}{40/\sqrt{16}}\right)$

$= P(-5 < Z < 10)$

$= P(Z < 10) - P(Z < -5)$

$= 0.999999 - 0.0000003$

$\simeq 0.999999$



Therefore, it is 99.9999% sure that a random sample of 16 bulbs will have an average life between 750 to 900.

12. Define central limit theorem. The amount of impurity in a batch of certain chemical product is a random variable with mean value 4 g and standard deviation 1.5 g. If 50 batches were independently prepared, what is the probability of impurity is between 3.5 and 3.8 g? [5] [2074 Ashwin]

Solution: Theory, Please, see the theory portion 4.2.

Numerical

Population mean (μ) = 4 g

Population S.D. (σ) = 1.5 g

Sample size (n) = 50

$P(3.5 < \bar{x} < 3.8)$

$$= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \bar{x} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$$

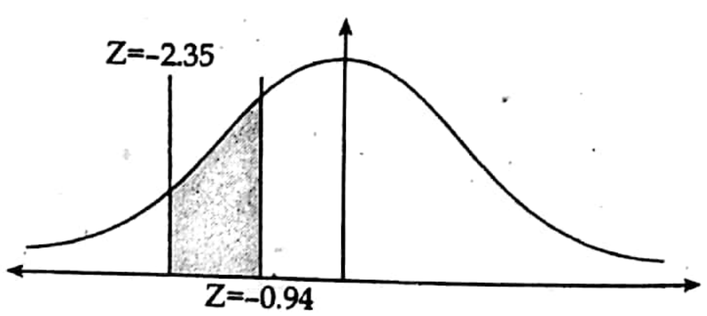
$$= P\left(\frac{3.5-4}{1.5/\sqrt{50}} < Z < \frac{3.8-4}{1.5/\sqrt{50}}\right)$$

$= P(-2.35 < Z < -0.94)$

$= P(Z < -0.94) - P(Z < -2.35)$

$= 0.1736 - 0.0094$

$= 0.1642$



Therefore, it is 16.42% sure that if 50 batches are independently prepared, they will have average amount of impurity between 3.5 g and 3.8 g.

13. Define population, sample, parameter and statistics with suitable examples. A population consists of 3, 7, 11, 15. Consider all possible sample of size two which can be drawn without replacement from this population. Find population mean and standard error of mean. [6] [2074 Ashwin]

Solution: Theory, Please, see the theory in 4.1.

Numerical

Population size (N) = 4

Sample size (n) = 2

$\therefore$ Number of samples formed without replacement

$= {}^{N}C_n = {}^{4}C_2 = 6$

Samples	Mean
(3, 7)	5
(3, 11)	7
(3, 15)	9
(7, 11)	9
(7, 15)	11
(11, 15)	13

ii) Population mean (μ) $= \frac{\Sigma X}{N} = \frac{3+7+11+15}{4} = 9$

Population S.D. (σ) $= \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

$$= \sqrt{\frac{(3-9)^2+(7-9)^2+(11-9)^2+(15-9)^2}{4}}$$

$= 4.47$

S.E. of mean S.E. ($\bar{x}$) $= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

$$= \frac{4.47}{\sqrt{2}} \times \sqrt{\frac{4-2}{4-1}}$$

$= 2.58$

14. A population consists of four numbers 2, 4, 6 and 8.

a) Enumerate all possible sample of size two without replacement.

b) Show that the mean of the sampling distribution of mean is equal to the population mean. [4] [2074 Chaitra]

Solution: Please, see the Q. No. 5.

15. State central limit theorem. A random sample of size 100 is taken from an infinite population with mean 75 and variance 256. Assert the chance of sample mean between 67 and 83. [6] [2074 Chaitra]

Solution: Theory, Please, see the theory portion 4.2.

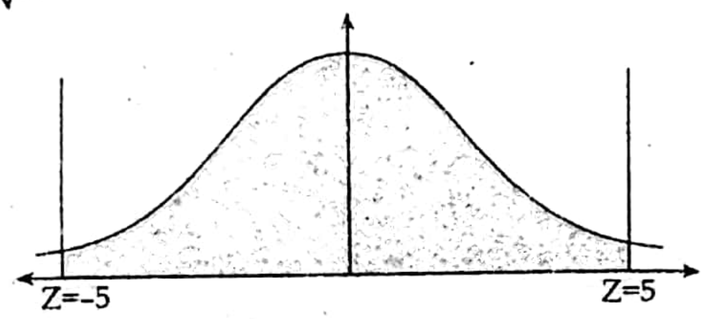
Numerical

Sample size (n) = 100

Population mean (μ) = 75

Population variance (σ^2) = 256

$\therefore \quad \sigma = \sqrt{256} = 16$



$P(67 < \bar{x} < 83)$

$= P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < Z < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$

$= P\left(\frac{67-75}{16/\sqrt{100}} < Z < \frac{83-75}{16/\sqrt{100}}\right)$

$= P(-5 < Z < 5)$

$= P(Z < 5) - P(Z < -5)$

$= 0.999997 - 0.0000003$

$= 0.99999$

Hence, it is 99.9999% sure that sample of size 100 will have mean between 67 and 83.

16. A population consists of four numbers 2, 8, 14, 20

a) Write down all possible sample size of two without replacement

b) Verify that the population mean is equal to mean of sample mean. **[6] [2075 Ashwin]**

Solution: Please, see the Q. No. 5.

Chapter 5

INTERFERENCE CONCERNING MEAN

5.1 Point Estimation and Internal Estimation 126

5.1.1 Confidence Interval and Confidence Level 126

5.2 Test of Hypothesis 131

5.2.1 Procedure for Testing of Hypothesis 132

5.3 Hypothesis Test Concerning One Mean 134

5.3.1 Steps for Test of Significance of Single Mean for Large Sample (n ≥ 30) 134

5.3.2 Steps for Test of Significance of a Single Mean for Small 138

5.4 Hypothesis Test Concerning Two Mean 139

5.4.1 Steps for Test of Significance of Difference of Two Means for Large Sample (n > 30) 139

5.4.2 Steps for Test of Significance of or Equality of Two Means for Large Sample 142

5.4.3 Steps for Test of Significance for Difference of Two Means for Small Sample 142

5.5 One Way Anova 143

5.6 Paired t-Test: (or Matched Pair t-Test) 148

5.1 POINT ESTIMATION AND INTERNAL ESTIMATION

The process of estimating population parameters by using sample data is called estimation.

i) Point estimation

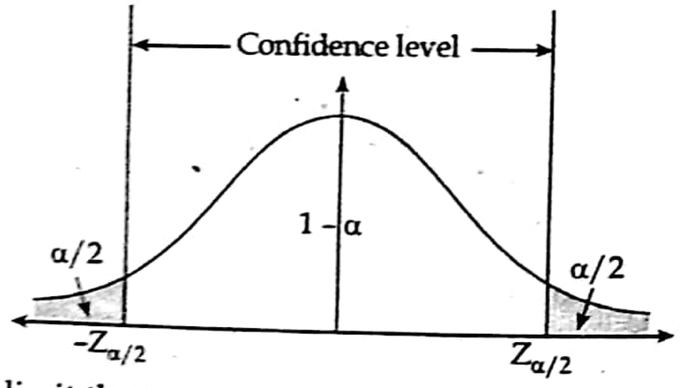
ii) Interval estimation

i) **Point estimate:** If a single value calculated from sample data is used to estimate population parameter 'θ', then the process is called point estimation. In other words, a point estimate is a single value (or point) used to approximate a population parameter.

ii) **Interval estimation:** If an interval or range of values calculated from sample data is used to estimate population parameter 'θ', then process is called interval estimation.

5.1.1 CONFIDENCE INTERVAL AND CONFIDENCE LEVEL

- **Confidence interval:** Confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A population parameter is sometimes abbreviated as CI.
- **Confidence level:** The confidence level is the probability 1–α (often expressed as the equivalent percentage value) that is the proportion of times that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. The confidence level is also called the degree of confidence or confidence co-efficient.
- **Confidence interval for population mean 'μ' for large sample ($n \geq 30$) and having confidence level $(1-\alpha) \times 100\%$.**



By central limit theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Probability $(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1-\alpha$

or, $$P\left(-Z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1-\alpha$$

Multiplying by $\frac{\sigma}{\sqrt{n}}$ on all sides, we get,

$$P\left(-Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Multiplying by (–) on all sides, we get (equality size reverses)

$$P\left(Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} > -\bar{X}+\mu > -Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

adding $\bar{X}$ on all sides, we get

$$P\left(\bar{X}+Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} > \mu > \bar{X}-Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

or, $$P\left(\bar{X}-Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\therefore \quad P(C_1 < \mu < C_2) = 1-\alpha$

Hence, confidence interval for population mean

$$\mu = \left(\bar{X}-Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \bar{X}+Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

$$= \bar{X} \pm Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} = \bar{X} \pm Z_{\alpha/2}\cdot S.E.(\bar{X})$$

Maximum error (E) $= |\bar{X}-\mu| = Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$

Width of interval (w) $= 2\left(Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$

- **Calculation of confidence interval for**

a) Large samples ($n > 30$)

i) $$\mu = \bar{X} \pm Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$$ (Infinite population)

$$= \bar{X} \pm Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$$ (Finite population)

ii) $$P = \hat{p} \pm Z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}\hat{q}}{n}}$$ (Infinite population)

$$= \hat{p} \pm Z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$$ (Finite population)

iii) $$\mu_1 - \mu_2 = (\bar{x}-\bar{y}) \pm Z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

iv) $$P_1 - P_2 = (\hat{p}_1-\hat{p}_2) \pm Z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

b) For small samples ($n < 30$)

i) $\mu = \bar{x} \pm t_{\alpha/2,\, n-1}\left(\frac{S}{\sqrt{n}}\right)$

ii) $\mu_1 - \mu_2 = (\bar{x} - \bar{y}) \pm t_{\alpha/2,\, n_1+n_2-2} \cdot S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where, $S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2}{n_1 + n_2 - 2}}$

OR,

$$\mu_1 - \mu_2 = \bar{x} - \bar{y} \pm t_{\alpha/2,\, n_1+n_2-2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where, $S_1^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n_1 - 1}$ and $S_2^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n_2 - 1}$

Example 1

A sample of 450 items is taken from population where $\sigma = 20$, $\bar{x} = 30$. Calculate 95% confidence interval.

Solution:

Here;

Sample size (n) = 450 (> 30, big sample)

Sample mean ($\bar{X}$) = 30

Population S.D. (σ) = 20

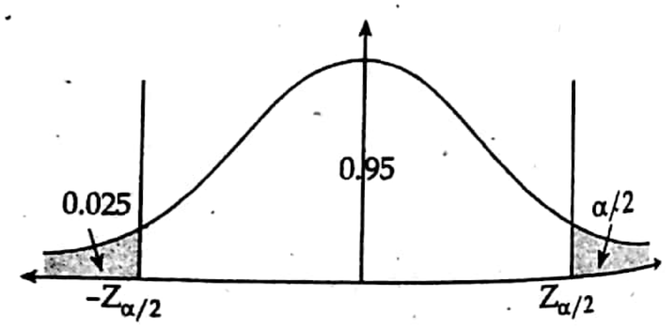
Confidence interval = 95%

or, $(1 - \alpha) \times 100\%$ 95%

or, $1 - \alpha = 0.95$

$\because$ $\alpha = 0.05$

$\therefore$ $\frac{\alpha}{2} = 0.025$



Now,

$$\mu = \bar{X} \pm Z_{\alpha/2} \cdot \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 30 \pm Z_{0.025} \cdot \left(\frac{20}{\sqrt{450}}\right)$$

$$= 30 \pm 1.96 \times \frac{20}{\sqrt{450}} \quad \text{(from Z-table)}$$

$$= 30 \pm 1.848$$

$$= 31.848 \text{ and } 28.152$$

$\therefore$ C.I. is (28.152, 31.848).

Note: P (Z score) = probability (or area)

Here, $Z_{\alpha/2} = Z_{0.025}$ is probability (or area) and we want Z-score. So, we follow this

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
⋮								
-2								
-1.9							0.025	
-1.8								
⋮								

So, Z-score corresponding to area 0.025 is -1.96.

Example 2

In a random sample of 100 of 500 experimental batteries produced by a form lifetime has a mean of 148.2 hrs with σ = 24.9 hrs. Find 80% confidence interval for mean life of 500 batteries.

Solution:

Here;

Population size (N) = 500

Sample size (n) = 100 (> 30, large sample)

Sample mean ($\bar{X}$) = 148.2 hrs

Population S.D. (σ) = 24.9 hrs

Confidence level = 80%

or, $(1 - \alpha) \times 100\%$ 80%

$\therefore$ $\alpha = 0.2 \Rightarrow \frac{\alpha}{2} = 0.1$

Now,

$$\mu = \bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

or, $\mu = 148.2 \pm Z_{0.1} \times \frac{24.9}{\sqrt{100}} \times \sqrt{\frac{500-100}{500-1}}$

$$= 148.2 \pm 1.28 \times \frac{24.9}{\sqrt{100}} \times \sqrt{\frac{400}{499}}$$

$$= 148.2 \pm 2.85$$

$$= 151.05 \text{ and } 145.35$$

$\therefore$ C.I. is (145.35, 151.05).

[$Z_{0.1} \simeq Z_{0.1003} = -1.28$, from Z-table similar to previous question, we find Z-value corresponding to given area = 0.1003]

Example 3

A random sample of 10 boys has following IQ's 70, 120, 110, 101, 83, 88, 95, 98, 107, 100. Find a reasonable range in which most of mean IQ value of 10 boys lie. (Use α = 5%)

Solution:

Here;

Sample population (n) = 10 (<30, small sample)

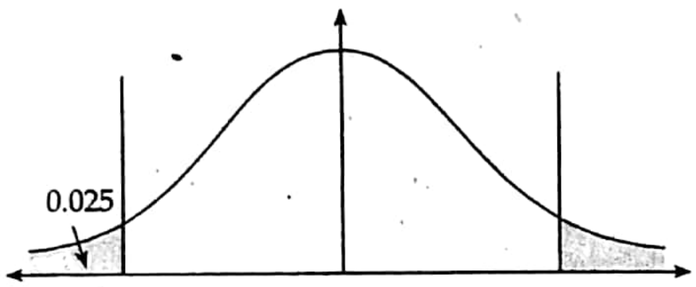
$\Sigma X = 972$

$\Sigma X^2 = 96312$

$$\text{Sample mean } (\bar{X}) = \frac{\Sigma X}{n} = 97.2$$

$$\text{Sample S.D. } (S) = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{\Sigma X^2 - n\bar{X}^2}{n-1}} = \sqrt{\frac{96312 - 10 \times 97.2^2}{9}}$$

$= 14.27$



$$\mu = \bar{X} \pm t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$$

$$= 97.2 \pm t_{0.025, 9} \times \frac{14.27}{\sqrt{10}}$$

$$= 97.2 \pm 2.262 \times \frac{14.27}{\sqrt{10}}$$

$= 97.2 \pm 10.2$

$= 87$ and 107.4

$\therefore$ C.I. is (87, 107.4)

Note: $t_{0.025, 9} \Rightarrow$ See data in area in one tail. Because area of 0.025 is present at each tail. If you want to use area in two tail you can use $t_{0.05,9}$ which give same result area in one tail.

Area in one tail

Degree of freedom v \ Area	0.005	0.01	0.025	0.05	0.1
1					
2					
3					
4					
⋮					
9			2.262		
10					
⋮					

$t_{0.025, 9}$

Also, no need to find ΣX, ΣX^2, manually use calculator to find this. Also check value of sample S.D. by calculator.

Example 4

The following are the Brinell values obtained for samples of 2 magnetism alloys:

Alloy 1:	107	148	123	165	102	119
Alloy 2:	134	115	112	151	133	129

Find 95% confidence interval for $\mu_1 - \mu_2$.

Solution:

Here;

Let Alloy 1 and Alloy 2 be represented by x and y respectively.

$\Sigma x = 764$ $\qquad$ $\Sigma y = 774$

$\Sigma x^2 = 100272$ $\qquad$ $\Sigma y^2 = 100856$

$\bar{x} = \frac{\Sigma x}{n_1}$ 127.34 $\qquad$ $\bar{y} = \frac{\Sigma y}{n_2} = 129$

$$\therefore \quad S = \sqrt{\frac{\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\Sigma x^2 - n\bar{x}^2 + \Sigma y^2 - n\bar{y}^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{100272 - 6 \times 127.34^2 + 100856 - 6 \times 129^2}{6 + 6 - 2}}$$

$\therefore \quad S = 20.011$

$$\alpha = 5\% \Rightarrow \frac{\alpha}{2} = 0.025$$

Now,

$$\mu_1 - \mu_2 = (\bar{x} - \bar{y}) \pm t_{\alpha/2, n_1 + n_2 - 2} \cdot S \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (127.34 - 129) \pm t_{0.025, 10} \times 20.01 \times \sqrt{\frac{1}{6} + \frac{1}{6}}$$

$$= -1.66 \pm 2.228 \times 20.01 \times \sqrt{\frac{1}{3}}$$

$= -1.66 \pm 25.74$

$= (-27.41, 24.07)$

$\therefore$ C.I. is (−27.41, 24.07)

5.2 TEST OF HYPOTHESIS

In statistics, a hypothesis is a claim or statement about a property of a population.

A hypothesis test (or test of significance) is a standard producer for testing a claim about a property of a population.

5.2.1 PROCEDURE FOR TESTING OF HYPOTHESIS

Step 1: State the type of Hypothesis

a) **Null Hypothesis (H_0):** It is a statement that the value of a population parameter (such as: Mean, proportion) is equal to some claimed value.

For example; $H_0 : p = 0.5$ $\quad H_0 : \mu = 98.6$

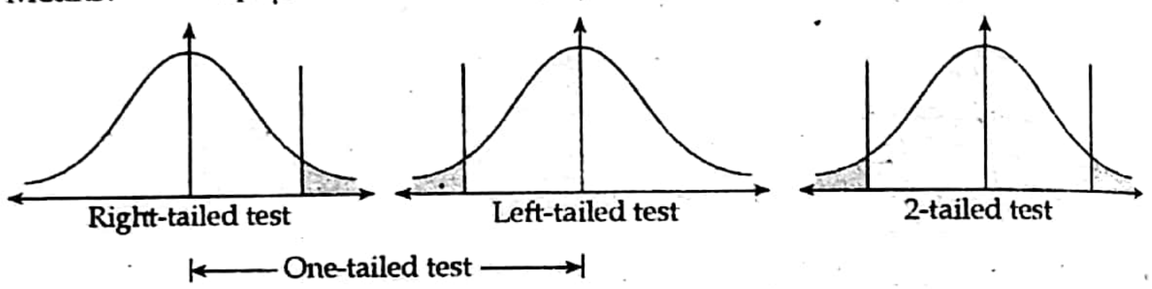
We test the null hypothesis directly in the sense that we assume it is true and reach a conclusion to either reject H_0 or to fail to reject H_0.

b) **Alternative Hypothesis (H_1):** It is the statement that the parameter has a value that somehow differs from the null hypothesis.

For example;

Proportions: $H_1 : p > 0.5$ $\quad H_1 : p < 0.5$ $\quad H_1 : p \neq 0.5$

Means: $H_1 : \mu > 98.6$ $\quad H_1 : \mu < 98.6$ $\quad H_1 : \mu \neq 98.6$



Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true. There can be two types of errors in testing of hypothesis.

a) **Type I Error:** Rejecting null hypothesis when it is true. It is generally represented by α.

b) **Type II Error:** Accepting null hypothesis when it is false. It is generally represented by β.

Actual state	Accept H	Reject H_0
H_0 true	Correct	Type I Error
H_0 false	Type II Error	Correct

Step 3: Select Appropriate Test-statistics Value

It is a value computed from sample data which is used to make decision about accepting or rejecting null hypothesis.

a) If sample is large ($n > 30$)

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

b) If sample is small ($n < 30$)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

c) For proportion

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

The test statistics value (z and t) depends on:

a) Sample size 'n'

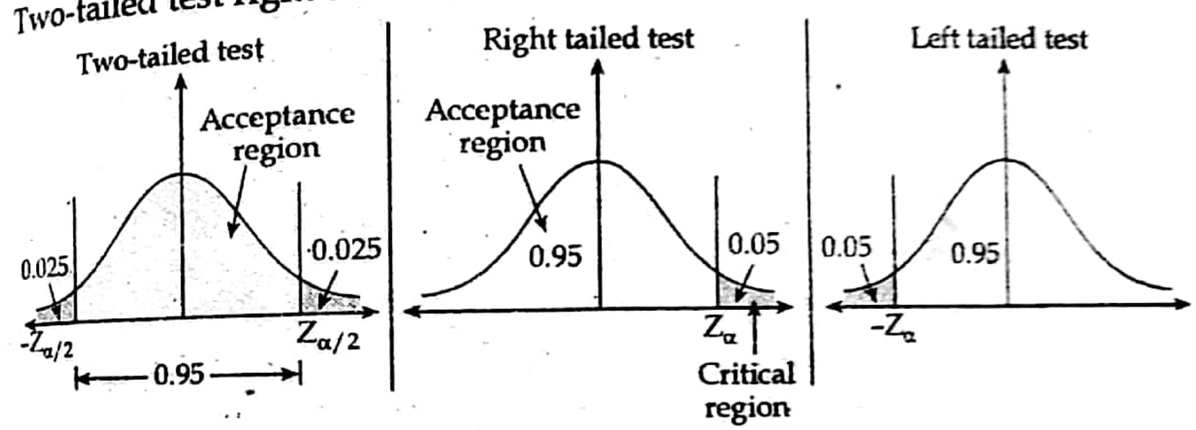
b) Type of test

c) Probability distribution

Step 4 Formulate Design Rule based on Step 1, 2, 3

$$\left(\text{Let } \alpha = 5\% = 0.05, \frac{\alpha}{2} = 0.025\right)$$

Two-tailed test right tailed test



Critical value depends on:

a) Significance level (α)

b) Type of hypothesis

c) Sample size

d) Probability distribution

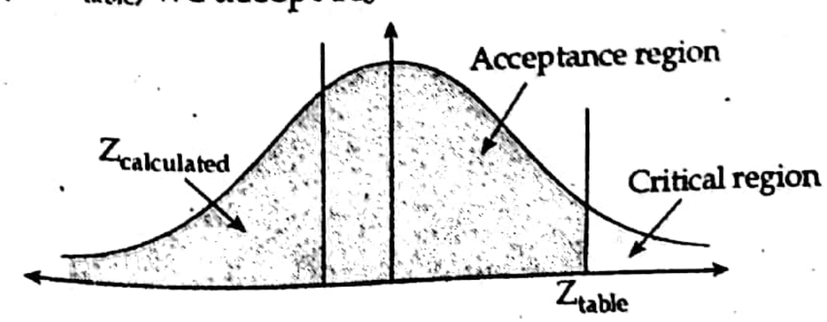
Step 5: Make a decision about null hypothesis using testing statistics value and critical values

a) Using formula from step 3, we calculate value of test statistics = $Z_{calculated}$

b) From table, we find critical values of $Z = Z_{table}$

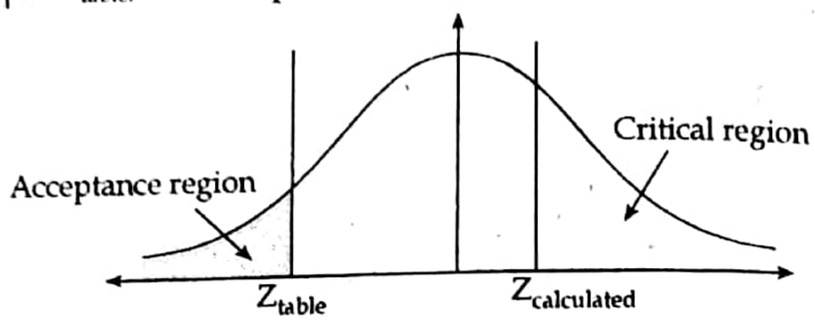
Case I

If $|Z_{calculated}| \leq Z_{table}$, we accept H_0



Case II

If $|Z_{calculated}| > Z_{table}$, we accept H_1 and reject H_0.



5.3 HYPOTHESIS TEST CONCERNING ONE MEAN

5.3.1 STEPS FOR TEST OF SIGNIFICANCE OF SINGLE MEAN FOR LARGE SAMPLE ($n \geq 30$)

Step 1: State the type of Hypothesis

a) **Null Hypothesis:** $H_0 : \mu = \mu_0$

i.e., There is no significant difference between population and sample mean.

b) **Alternative Hypothesis:**

$H_1 : \mu > \mu_0$ (one tailed test)

or, $H_1 : \mu < \mu_0$ (one tailed test)

or, $H_1 : \mu \neq \mu_0$ (two tailed test)

i.e., There is significant difference between population and sample mean.

Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true.

Choose appropriate value of significance level (α). If not given in question, use α = 5%.

Step 3: Test Statistics

Test of a single mean of large sample ($n \geq 30$). Let 'n' be the sample size and μ_0 be claimed mean

$H_0 : \mu = \mu_0$

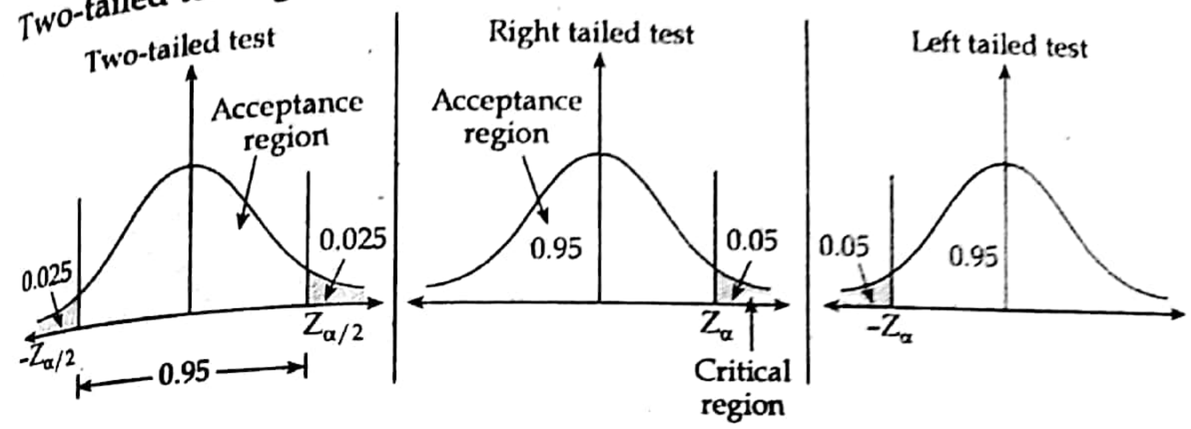
$$\text{Test statistics (T.S.)} = Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

If the value of σ is unknown, sample standard deviation (s) can be used as it is a close approximation to population standard deviation (σ) as large sample is considered.

Step 4 Formulate Design Rule based on Step 1, 2, 3

$\left(\text{Let } \alpha = 5\% = 0.05, \frac{\alpha}{2} = 0.025\right)$

Two-tailed test right tailed test



Critical value depends on:

a) Significance level (α)
b) Type of hypothesis
c) Sample size
d) Probability distribution

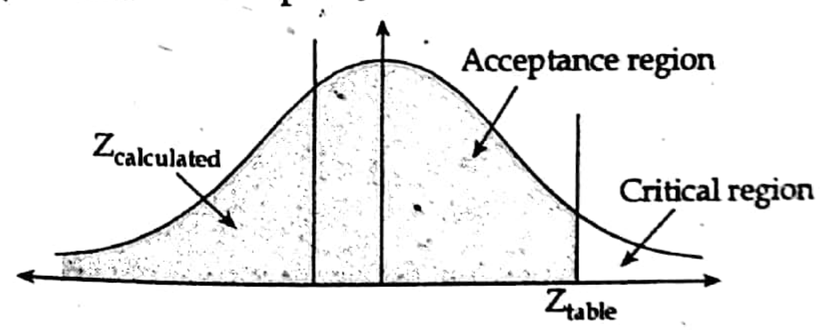
Step 5: Make a decision about null hypothesis using testing statistics value and critical values

a) Using formula from step 3, we calculate value of test statistics = $Z_{calculated}$

b) From table, we find critical values of $Z = Z_{table}$

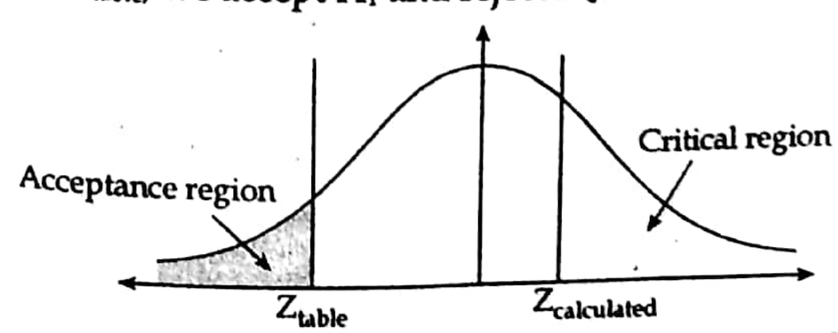
Case I

If $|Z_{calculated}| \leq Z_{table}$, we accept H_0



Case II

If $|Z_{calculated}| > Z_{table}$, we accept H_1 and reject H_0.



H_1	Rejection Region (R.R.)
$\mu < \mu_0$	$Z < -Z_\alpha$
$\mu > \mu_0$	$Z > Z_\alpha$
$\mu \neq \mu_0$	$\lvert Z \rvert > Z_{\alpha/2}$

Example 5

The mean weight of a brick supplied by a manufacturer is 1800 with S.D. of 100. By a new technique in manufacturing, it is claimed that weight of bricks have increased. In other to test his claim, a sample of 50 bricks were tested and found that mean weight is 1850 g. Can we support the level of claim at 0.01 level of significance.

Solution:

Here;

Sample size (n) = 50

Sample mean $(\bar{x})$ = 1850

Hypothesized population mean (μ) = 1800

Population S.D. (σ) = 100

Step 1: State the type of Hypothesis

Null hypothesis $H_0 : \mu = \mu_0 = 1800$. *i.e.*, there is no significant difference between population and sample.

Alternative Hypothesis $H_1 : \mu > 1800$

Step 2: Significance level $\alpha = 1\% = 0.01$

Step 3: Test statistics

n = 50 (> 30), so, large sample. Hence Z test is used.

$$\text{Test statistics } (Z) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.53$$

Step 4: Here in alternative hypothesis, $\mu > 1800$ *i.e.*, right tailed test (one tailed test)

So, $Z_\alpha = Z_{1-\alpha}$ [$\because$ Right tailed test]

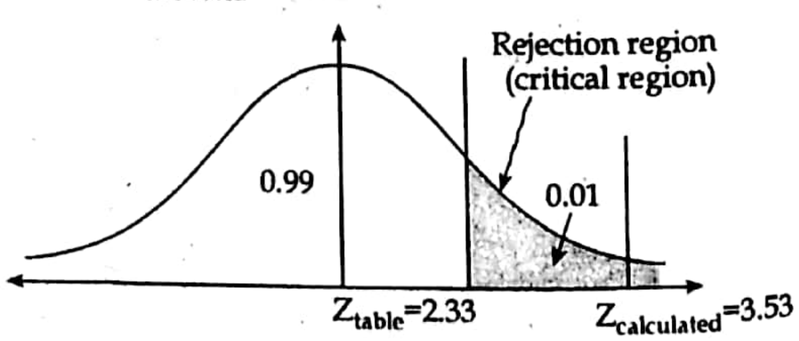
$= Z_{0.99}$

$= 2.33$ [From Z_{table}, $Z_{0.9901} = 2.33$]

i.e., $Z_{table} = 2.33$

Step 5: Decision Making

$Z_{table} = 2.33$ $Z_{calculated} = 3.53$



The test statistics value lies in critical region. So, we reject null hypothesis and accept alternative hypothesis. So, weight of brick has increased by using new techniques.

Example 6

The mean weight of brick supplied by a manufactures is 1800 g with S.D. 100 g. But a person claims to test this claim a sample of 50 bricks were tested and found that mean weight is 1850 g. Can we support the level of claim at 0.01 level of significance.

Solution:

Here;

Sample size (n) = 50

Sample mean $(\bar{x})$ = 1850

Hypothesized population mean (μ) = 1800

Population S.D. (σ) = 100

Step 1: State the type of hypothesis

Null Hypothesis $H_0 : \mu = \mu_0 = 1800$ *i.e.*, there is no significant difference between population and sample mean.

Alternative Hypothesis $H_1 : \mu \neq 1800$ *i.e.*, there is significant difference between population and sample mean.

Step 2: Significance level (α) = 1 % = 0.01

Step 3: Test statistics

n = 50 (> 30), so sample is large and Z-test is used

$$\text{Test statistics } (Z) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.53$$

Step 4: Here in alternative hypothesis $H_1 : \mu \neq 1800$. So, two tailed test is used.

$$Z_{\alpha/2} = Z_{0.005} = \frac{-2.57 - 2.58}{2} \qquad \left[\because Z_{0.005} = \frac{Z_{0.0051} + Z_{0.0049}}{2}\right]$$

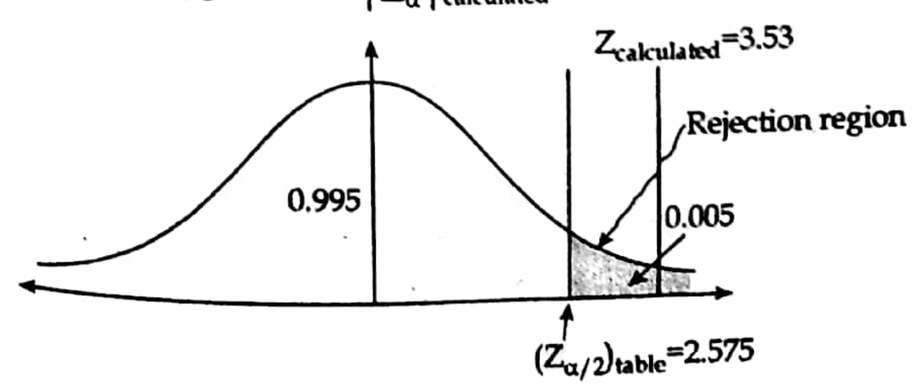
$= -2.575$ (at left)

$Z_{1-\alpha/2} = Z_{0.995} = 2.575$ (at right)

$|Z_{\alpha/2}|_{table} = 2.575$

Step 5: Decision making

$|Z_{\alpha/2}|_{table} = 2.575$ $|Z_\alpha|_{calculated} = 3.53$



The test statistics value lies in rejection region. Hence, we reject null hypothesis and accept alternative hypothesis.

Note: $Z_{0.005}$ can be calculated by calculator FX-991EX

Mode + (7) [Distribution] ↲

(3) [inverse Normal] ↲

Area = 0.005 ↲

$\sigma = 1$ ↲

$\mu = 0$ ↲

−2.575 (calculated by calculator)

5.3.2 STEPS FOR TEST OF SIGNIFICANCE OF A SINGLE MEAN FOR SMALL

Let $x_1, x_2, x_3, \ldots\ldots, x_n$ be a random sample size n (<30) drawn from a normal population with mean m and unknown variance σ^2.

Step 1: State the type of Hypothesis

a) **Null Hypothesis $H_0 : \mu = \mu_0$.**

i.e., there is no significant difference between population and sample mean.

b) **Alternative Hypothesis:**

$H_1 : \mu > \mu_0$ (one tailed test)

or, $H_1 : \mu < \mu_0$ (one tailed test)

or, $H_1 : \mu \neq \mu_0$ (two tailed test)

i.e., There is significant difference between population and sample mean.

Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true.

Choose appropriate value of significance level (α). If not given in question, use α = 5%.

Step 3: Test Statistics

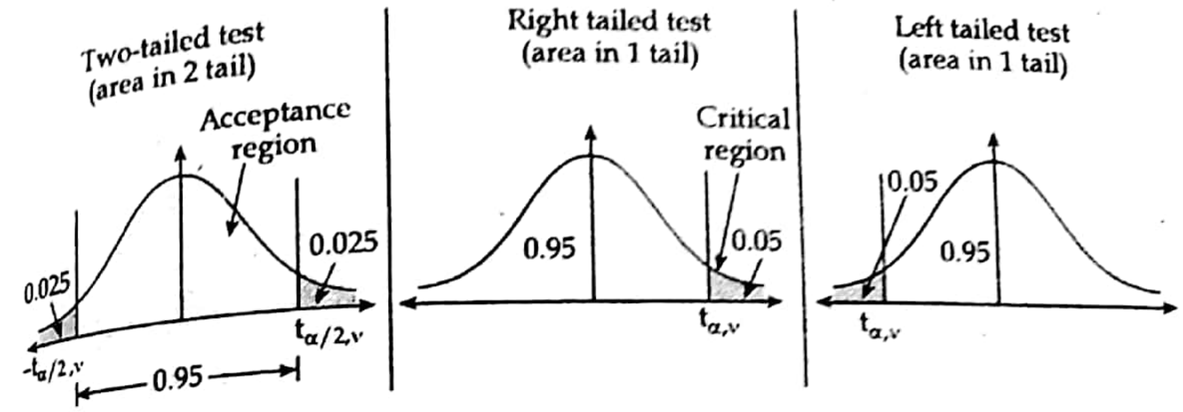
The test statistics used here is 't' and is compared with the student t-distribution for rejection criteria. Let n be the sample size and μ_0 be the claimed mean

Test statistics $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Degree of freedom (df): $\nu = n - 1$

Step 4: Critical Value

Let $\alpha = 5\%, \dfrac{\alpha}{2} = 2.5\% = 0.025$



Critical values depend upon

i) Significance level (α)

ii) Type of hypothesis

iii) Sample size

iv) Probability distribution

Step 5: Decision Making Step

Test statistics value calculated from formula (t) = $t_{calculated}$

Test statistics value from table,

$t_{table} = t_{\alpha,\nu}$ or, $t_{\alpha/2,\nu}$ depending upon type of hypothesis

Alternative Hypothesis (H_1)	Rejection Region
$\mu < \mu_0$	$t < -t_{\alpha,\nu}$
$\mu > \mu_0$	$t > t_{\alpha,\nu}$
$\mu \neq \mu_0$	$\lvert t \rvert > t_{\alpha/2,\nu}$

5.4 HYPOTHESIS TEST CONCERNING TWO MEAN

5.4.1 STEPS FOR TEST OF SIGNIFICANCE OF DIFFERENCE OF TWO MEANS FOR LARGE SAMPLE (n > 30)

Let there be two treatments of sample sizes n_1 and n_2 ($n_1, n_2 > 30$), sample S.D. of σ_1 and σ_2 respectively. Let δ_0 is the claimed difference of their population mean.

Step 1: State the type of hypothesis

a) **Null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$**

b) **Alternative Hypothesis:**

$H_1 : \mu_1 - \mu_2 > \delta_0$ (right tailed test)

or, $H_1 : \mu_1 - \mu_2 < \delta_0$ (left tailed test)

or, $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (two tailed test)

Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true.

Choose appropriate value of significance level (α). If not given in question, use α = 5%.

Step 3: Test Statistics (T.S.)

a) $$\text{TS: } Z = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$ [If σ_1 and σ_2 are known and unequal]

b) $$Z = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sigma\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)}$$ [If σ_1 and σ_2 are known and equal. *i.e.*, $\sigma_1 = \sigma_2 = \sigma$]

c) $$Z = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$ [If σ_1 and σ_2 are unknown and unequal]

For large samples

$\sigma_1 \simeq S_1$ and $\sigma_2 \simeq S_2$

$$\text{Sample S.D.}_1\ (S_1) = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n_1-1}}$$

$$\text{Sample S.D.}_2\ (S_2) = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n_2-1}}$$

d) $$Z = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{S_p\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)}$$ [If σ_1 and σ_2 are unknown but equal]

$$S_p(\text{pooled estimate of S.D.}) = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

where, S_1 and S_2 are calculated as above.

Note: In all test statistics, Z we can replace

$\mu_1 - \mu_2 = \delta_0$

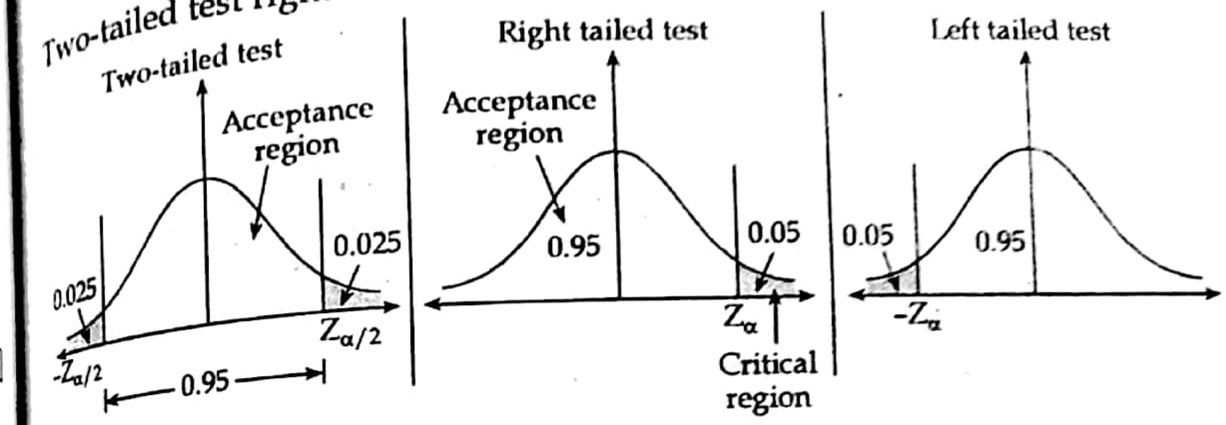
For example; Step (3), (a) can be written as,

$$\text{T.S.: } Z = \frac{\bar{x}-\bar{y}-\delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

Step 4 Formulate Design Rule based on Step 1, 2, 3

$$\left(\text{Let } \alpha = 5\% = 0.05, \frac{\alpha}{2} = 0.025\right)$$

Two-tailed test right tailed test



Critical value depends on:

a) Significance level (α)
b) Type of hypothesis
c) Sample size
d) Probability distribution

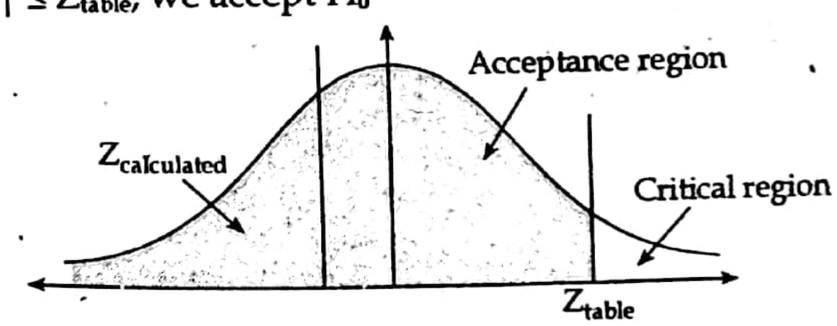
Step 5: Make a decision about null hypothesis using testing statistics value and critical values

a) Using formula from step 3, we calculate value of test statistics = $Z_{calculated}$

b) From table, we find critical values of $Z = Z_{table}$

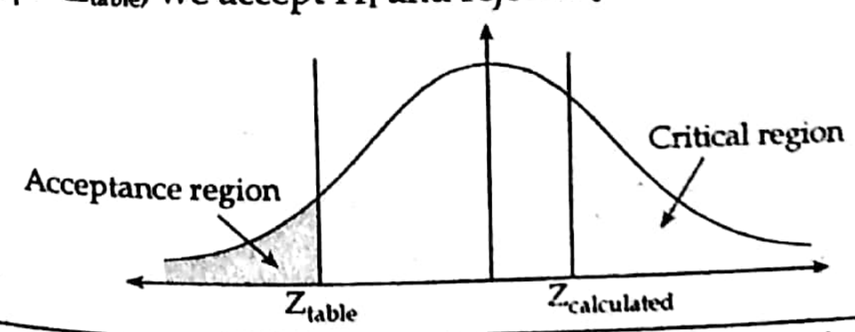
Case I

If $|Z_{calculated}| \le Z_{table}$, we accept H_0



Case II

If $|Z_{calculated}| > Z_{table}$, we accept H_1 and reject H_0.



Alternative Hypothesis (H_1)	Rejection Region
$\mu_1 - \mu_2 < \delta_0$	$Z < -Z_\alpha$
$\mu_1 - \mu_2 > \delta_0$	$Z > Z_\alpha$
$\mu_1 - \mu_2 \ne \delta_0$	$\lvert Z \rvert > Z_{\alpha/2}$

5.4.2 STEPS FOR TEST OF SIGNIFICANCE OF OR EQUALITY OF TWO MEANS FOR LARGE SAMPLE

Follow all the steps of 5.4.1 only replace δ_0 by 0 from step 1 and remove $(\mu_1 - \mu_2)$ in step 3 from Z formula. *i.e.*, $(\mu_1 - \mu_2) = 0$

5.4.3 STEPS FOR TEST OF SIGNIFICANCE FOR DIFFERENCE OF TWO MEANS FOR SMALL SAMPLE

Let there be two treatments of sample sizes n_1 and n_2 (where any sample size (n_1 or n_2) or both are less than 30). Sample means of μ_1 and μ_2, sample S.D. of S_1 and S_2 respectively. Let δ_0 is the claimed difference of their population mean.

Step 1: State the type of Hypothesis

a) **Null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$**

b) **Alternative Hypothesis:**

$H_1 : \mu_1 - \mu_2 > \delta_0$ (right tailed test)

or, $H_1 : \mu_1 - \mu_2 < \delta_0$ (left tailed test)

or, $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (two tailed test)

Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true.

Choose appropriate value of significance level (α). If not given in question, use α = 5%.

Step 3: Test Statistics (T.S.)

a) When the population variance is unknown and unequal

$$\text{T.S.: } t' = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\text{where, Sample S.D.}_1 (S_1) = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n_1 - 1}}$$

$$\text{Sample S.D.}_2 (S_2) = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n_2 - 1}}$$

$$\text{Degree of freedom (d.f.): } v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Note: Round down v.

This test is based on t' approximation to the t-test is called the smith - satterthwaite test.

b) When the population variance is unknown and equal.

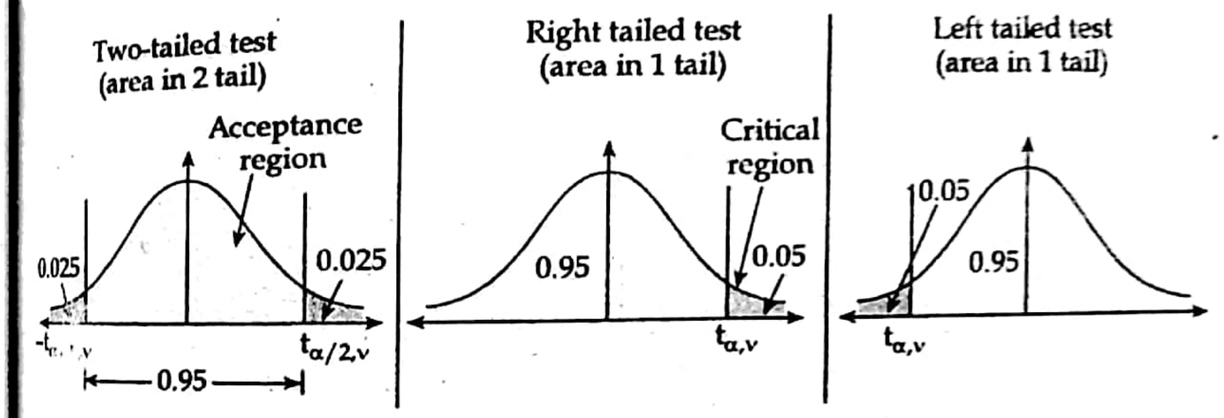
$$\text{T.S.: } t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_p\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}$$

$$S_p = \sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}}$$

Degree of freedom (d.f.): $v = n_1 + n_2 - 2$

Step 4: Critical Value

Let $\alpha = 5\%$, $\frac{\alpha}{2} = 2.5\% = 0.025$



Critical values depend upon

i) Significance level (α)

ii) Type of hypothesis

iii) Sample size

iv) Probability distribution

Step 5: Decision Making Step

Test statistics value calculated from formula (t) = $t_{calculated}$

Test statistics value from table,

$t_{table} = t_{\alpha,v}$ or, $t_{\alpha/2,v}$ depending upon type of hypothesis

Alternative Hypothesis (H_1)	Rejection Region
$\mu_1 - \mu_2 < \delta$	$t < -t_{\alpha,v}$
$\mu_1 - \mu_2 > \delta$	$t > t_{\alpha,v}$
$\mu_1 - \mu_2 \neq \delta$	$\lvert t \rvert > t_{\alpha/2,v}$

5.5 ONE WAY ANOVA

It is a method of testing equality of three or more than 3 means by analysing sample variances. In ANOVA, we use F-test (Fisher's test)

Assumptions

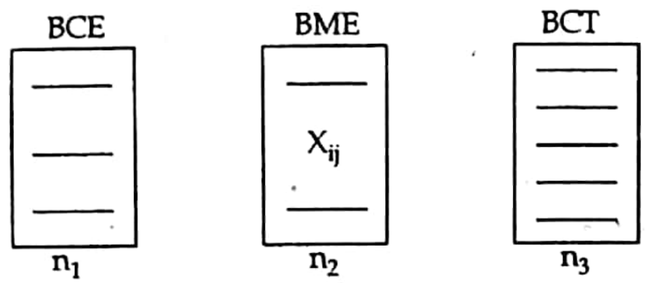
i) The population from which samples are drawn should follow normal distribution.

ii) Samples are drawn randomly and independently.

iii) Various treatments and environmental effects are additive.

Step 3: Test Statistics

Let us assume 3-departments of IOE



$i = 1, 2, 3, \ldots\ldots\ldots, K$

$j = 1, 2, 3, \ldots\ldots\ldots, n$

X_{ij} = j^{th} observation in i^{th} group

$N = n_1 + n_2 + n_3 \ldots\ldots\ldots$

$$\text{Grand mean } (\bar{\bar{X}}) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n} X_{ij}}{N} = \frac{\text{Sum of all elements}}{N}$$

$$\text{Total variation (T.S.S)} = \sum_{i=1}^{K}\sum_{j=1}^{n} (X_{ij} - \bar{\bar{X}})^2$$

$$\text{Variation between treatment } (SST_r) = \sum_{i=1}^{K} n_i (\bar{X}_i - \bar{\bar{X}})^2$$

Variation within sample

$$(SSE) = \sum_{i=1}^{K}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$$

$$TSS = SST_r + SSE$$

Alternative method

$$\text{Correction factor (C.F)} = \frac{\left[\sum_{i=1}^{K}\sum_{j=1}^{n} X_{ij}\right]^2}{N} = \frac{(\text{Sum of all numbers})^2}{N}$$

$$\text{Total variation (T.S.S)} = \left[\sum_{i=1}^{K}\sum_{j=1}^{n} X_{ij}^2\right] - \text{C.F.}$$

Variation between treatment

$$(SST_r) = \left[\sum \frac{T_i^2}{n_i}\right] - \text{C.F.}$$

Variation within sample (SSE) = TSS – SST_r

ANOVA Table

Source of variation	Sum of square	Degree of freedom	Mean sum of square	F-ratio
Between treatment	SST_r	K – 1	$MST_r = \frac{SST_r}{K-1}$	$F = \frac{MST_r}{MSE}$
Within sample	SEE	N – K	$MSE = \frac{SSE}{N-K}$	
Total	TSS	N – 1		

Step 4 For $\alpha = 0.05$

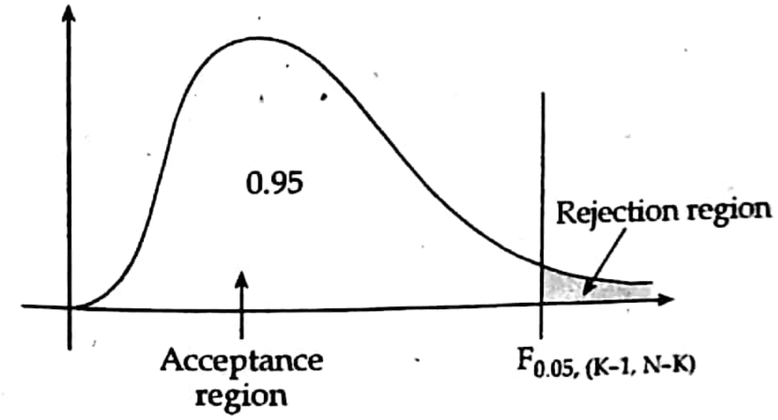
$F_{0.05,}$ (K – 1, N – K) is critical value.

Select $\alpha = 0.005$ (F-table)

Numerator degree of freedom (K – 1) →

Denomenator D.O.F. (N – K) ↓

	1	2	3	4	5 ...
1					
2					
3			*		
⋮					



Example 7

The daily sale of biscuit A, B and C is given in table below:

Product	Sale											
A	14	13	9	15	11	13	14	11				
B	10	12	9	7	11	8	12	9	10	13	9	10
C	11	5	9	10	6	8	8	7				

Construct ANOVA table and test equality of mean using sale

Solution:

Here;

Step 1: Set up Hypothesis

Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$

Alternative hypothesis $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

Step 2: Level of significance

Let $\alpha = 1\%$

Step 3:

Let $A \to 1, B \to 2, C \to 3$

Grand number (N) = 28

$n_1 = 8$

$n_2 = 12$

$n_3 = 8$

$$\text{Grand mean } (\bar{\bar{X}}) = \frac{\sum_{i=1}^{3}\sum_{j=1}^{n} X_{ij}}{N} = \frac{284}{28} = 10.14 \quad \left[\sum_{i=1}^{3}\sum_{j=1}^{n} X_{ij} = \text{Sum of all numbers}\right]$$

X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$
14	2.25	10	0	11	9
13	0.25	12	4	5	9
9	12.25	9	1	9	1
15	6.25	7	9	10	4
11	2.25	11	1	6	4
13	0.25	8	4	8	0
14	2.25	12	4	8	0
11	2.25	9	1	7	1
		10	0		
		13	9		
		9	1		
		10	0		
$\Sigma = 100$	$\Sigma = 28$	$\Sigma = 120$	$\Sigma = 34$	$\Sigma = 64$	$\Sigma = 28$

$$\bar{X}_1 = \frac{100}{8} = 12.5$$

$$\bar{X}_2 = \frac{120}{10} = 10$$

$$\bar{X}_3 = \frac{64}{8} = 8$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{X}_i - \bar{\bar{X}})^2$$

$$= n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + n_3(\bar{X}_3 - \bar{\bar{X}})^2$$

$$= 8 \cdot (12.5 - 10.14)^2 + 12 \cdot (10 - 10.14)^2 + 8 \cdot (8 - 10.14)^2$$

$$= 81.43$$

Variation within sample

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$$

$$= 28 + 34 + 28 \qquad \text{[From table]}$$

$$= 90$$

$\therefore$ Total variation (TSS) = SSE + SST_r = 171.43

ANOVA Table

Source of variation	Sum of square	Degree of freedom	Mean sum of square	F-ratio
Between treatment	$SST_r = 81.43$	3 – 1 = 2	$MST_r = \frac{SST_r}{K-1} = 40.715$	$F = \frac{MST_r}{MSE} = \frac{40.715}{3.6} = 11.31$
Within sample	SSE = 90	28 – 3 = 25	$MSE = \frac{SSE}{N-K} = 3.6$	
Total	TSS = 171.43	28 – 1 = 27		

Step 4 From table

$F_{0.01(2,25)} = 5.57$

Step 5: Decision making

$\because F_{calculated}$ (11.31) > F_{table} (5.57) $\Rightarrow$ in critical zone. So, we reject null hypothesis and accept alternative hypothesis.

Alternative method

(To calculate TSS, SST_r, SSE)

$$C.F. = \frac{\left[\sum_{i=1}^{3}\sum_{j=1}^{n} X_{ij}\right]^2}{N} = \frac{284^2}{28} = 2880.57 \qquad [\because 14 + 13 + \cdots\cdots + 8 + 7 = 284]$$

$$TSS = \sum_{i=1}^{3}\sum_{j=1}^{n} X_{ij}^2 - C.F. = 3052 - 2880.57 = 171.42$$

$$[\because 14^2 + 13^2 + \cdots\cdots + 8^2 + 7^2 = 3052]$$

$$SST_r = \left[\sum \frac{T_i^2}{n_i}\right] - C.F.$$

$$= \left(\frac{100^2}{8} + \frac{120^2}{12} + \frac{64^2}{8}\right) - 2880.57 \quad \left[\begin{array}{l} T_1 = 14 + 13 + \cdots\cdots + 11 = 100 \\ T_2 = 10 + 12 + \cdots\cdots + 10 = 120 \\ T_3 = 11 + 5 + \cdots\cdots + 7 = 64 \end{array}\right]$$

$$= 81.43$$

$$SSE = TSS - SST_r = 171.42 - 81.43 = 90$$

Now, fill the ANOVA table, as above.

5.6 PAIRED t-TEST: (OR MATCHED PAIR t-TEST)

Let x and y be two sets of samples taken in pairs assuming similar conditions of data collection for each pair. It is obvious that $n_1 = n_2 = n$. Let the claimed mean difference between the sets be $\mu_{d,0}$. Compute another set of samples called the difference set D where

$$d_i = x_i - y_i$$

Assumption

i) Small samples $n < 30$

ii) Equal sample size

iii) Every pair of data (x_i, y_i) is dependent.

Let $x_1, x_2, \ldots\ldots, x_n$ be before data and $y_1, y_2, \ldots\ldots, y_n$ be after data.

x_i (Before)	y_i (After)	Difference $(x_i - y_i)$
x_1	y_1	$d_1 = x_1 - y_1$
x_2	y_2	$d_2 = x_2 - y_2$
⋮	⋮	⋮
x_n	y_n	$d_n = x_n - y_n$
Total		Σd_i

$$\bar{d} \text{ (Mean of difference)} = \frac{\Sigma d_i}{n}$$

$$\text{Sample S.D. of } \bar{d} \ (S_d) = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{\Sigma d_i^2 - n\bar{d}^2}{(n-1)}}$$

Step 1: State the type of hypothesis

Null hypothesis $H_0 : \bar{d} = 0$

Alternative hypothesis H_1 : $\bar{d} > 0$ (Right tailed test)

$\bar{d} < 0$ (Left tailed test)

$\bar{d} \neq 0$ (two tailed test)

Step 2: Level of significance (α)

It is the maximum value of probability of rejecting null hypothesis, when it is true. Select appropriate value of level of significance (α). Let $\alpha = 5\%$

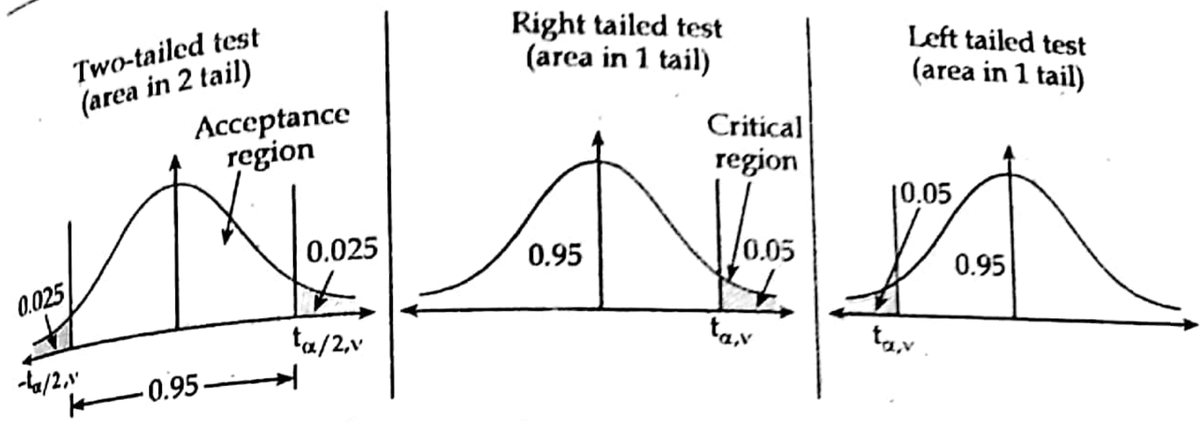
Step 3: Test statistics (TS)

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

Degree of freedom (df): $v = n - 1$

Step 4: Critical value

Let $\alpha = 5\%, \frac{\alpha}{2} = 2.5\% = 0.025$



Critical values depend upon

i) Significance level (α)

ii) Type of hypothesis

iii) Sample size

iv) Probability distribution

Step 5: Decision Making Step

Alternative Hypothesis (H_1)	Rejection Region
$\mu < \mu_{d,0}$	$t < -t_{\alpha,v}$
$\mu > \mu_{d,0}$	$t > t_{\alpha,v}$
$\mu \neq \mu_{d,0}$	$\lvert t \rvert > t_{\alpha/2,v}$

Example 8

The following are the weekly losses of workers hours due to accidents in 10 industrial plants before and after a certain safety.

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

Use $\alpha = 5\%$

Solution:

Here;

Before (x_i)	After (y_i)	Difference (di) = xi – yi
45	36	9
73	60	13
46	44	2
124	119	5
33	35	–2
57	51	6
83	77	6
34	29	5
26	24	2
17	11	6
Total		$\Sigma d_i = 52$ and $\Sigma d_i^2 = 420$

$$\bar{d} = \frac{\Sigma d_i}{n} = \frac{52}{10} = 52$$

$$S = \sqrt{\frac{\Sigma d_i^2 - n\bar{d}^2}{(n-1)}}$$

$$= \sqrt{\frac{420 - 10 \times 2.2^2}{9}} = 4.077$$

Step 1: State the type of hypothesis

Null hypothesis $H_0 : \bar{d} = 0$ *i.e.*, there is no significant improvement in weekly losses.

Alternative hypothesis $H_1 : \bar{d} > 0$ *i.e.*, there is significant improvement in weekly losses.

Step 2: Level of significance (α)

$\alpha = 5\%$ (given)

Step 3: Test statistics (TS)

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{5.2}{4.077/\sqrt{10}} = 4.033$$

df: $\nu = n - 1 = 10 - 1 = 9$

Step 4: Critical value

$t_{\alpha, n-1} = t_{0.05, 9}$

(Use area in 1 tail as it is one tailed test *i.e.*, right tailed test.)

Step 5: Decision making

Here;

$t_{\alpha, n-1} = t_{0.05, 9} = t_{table} = 1.833$

$t_{calculated} = t = 4.033$

$\therefore$ $t_{table} < t_{calculated}$

So, it lies in critical zone.

Hence, null hypothesis is rejected and alternative hypothesis is accepted.

SOLVED QUESTIONS (FOR BCE/BME/BIE/BAM)

1. **A company claims that its light bulbs are superior to those of its competitor. If a study showed that a sample of $n_1 = 40$ of its bulbs has a mean lifetime 1647 hours of continuous use with a standard deviation of 27 hours, while a sample of $n_2 = 40$ bulbs made by its competitor has a mean lifetime deviation of 31 hours, does this substantiate to claim at 0.05 level of significance? [6][2069 Poush]**

Solution:

Given that;

$n_1 = 40, \ \bar{X}_1 = 1647, \ S_1 = 27$

$n_2 = 40, \ \bar{X}_2 = 1638, \ S_2 = 31$

Setting up null and alternative hypothesis:

1. Null hypothesis: The two bulbs are equally effective.
   Symbolically, $H_0: \mu_1 = \mu_2$
2. Alternative hypothesis: The first type of bulb is superior to the second type of bulb.
   Symbolically, $H_1: \mu_1 \neq \mu_2$
3. Since the sample sizes are large *i.e.*, > 30, we assume $\sigma \approx S$ and use Z test. The test statistic is;
   We have,
   $$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(1647 - 1638)}{\sqrt{\frac{27^2}{40} + \frac{31^2}{40}}}$$
4. At $\alpha = 0.05$, from table. $Z_\alpha = Z_{0.05} = 1.645$
5. Since $Z(= 1.385) < Z_{0.05}$ (=1.645), we accept H_0. Thus, the company's claim that it's light bulbs are superior to those of its competitor is incorrect.

2. **Explain the concept of (i) Point estimation (ii) Interval estimation with illustration. [4][2069 Poush]**

Solution: See theory section 5.1 and example 1 of 5.1.1.

3. **Suppose 3 drying formulas for curing a glue are studied and the following times observed.**

Formula	Time					
A	13	10	8	11	8	-
B	13	11	14	14	-	-
C	4	1	3	4	2	4

Construct ANOVA table and test for equality of the mean curing times. [Use $F_{0.05, (2, 12)} = 3.89$] [6][2069 Poush]

Solution:

1. Null hypothesis: There is no significant difference in the mean curing times of 3 formulas.

   Symbolically, $H_0: \mu_1 = \mu_2 = \mu_3$

2. Alternative hypothesis: There is significant difference of at least one curing time mean from others.

   Symbolically, $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

3. Constructing ANOVA table

A (y_1)	$(y_1 - \bar{y}_1)^2$	B (y_2)	$(y_2 - \bar{y}_2)^2$	C (y_3)	$(y_3 - \bar{y}_3)^2$
13	9	13	0	4	1
10	0	11	4	1	4
8	4	14	1	3	0
11	1	14	1	4	1
8	4	–	–	2	1
–	–	–	–	4	1
$\Sigma = 50$	$\Sigma = 18$	$\Sigma = 52$	$\Sigma = 6$	$\Sigma = 18$	$\Sigma = 8$

$n_1 = 5, n_2 = 4, n_3 = 6$

$N = n_1 + n_2 + n_3 = 5 + 4 + 6 = 15$

$T = 50 + 52 + 18 = 120$

$\bar{y}_1 = \frac{50}{5} = 10; \quad \bar{y}_2 = \frac{52}{4} = 13; \quad \bar{y}_3 = \frac{18}{6} = 3$

$K = 3$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{120}{15} = 8$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= n_1(\bar{y}_1 - \bar{\bar{y}})^2 + n_2(\bar{y}_2 - \bar{\bar{y}})^2 + n_3(\bar{y}_3 - \bar{\bar{y}})^2$$

$$= 5(10 - 8)^2 + 4(13 - 8)^2 + 6(3 - 8)^2$$

$$= 270$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 18 + 6 + 8 = 32$$

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 3 - 1 = 2$	$SST_r = 270$	$MST_r = \frac{270}{2} = 135$	$\frac{MSTr}{SSE} = \frac{135}{2.67} = 50.625$
Error	$N - K = 15 - 3 = 12$	$SSE = 32$	$MSE = \frac{32}{12} = 2.67$	
Total	14	302		

4. For $\alpha = 0.05$ and d.f. = (2, 12). Given; $F_{0.05, (2, 12)} = 3.89$
5. Since $F (= 50.625) > F_{0.05, (2, 12)} (= 3.89)$, we reject H_0.
   Hence, there is no significant difference between mean curing time of at least one formula with others.

4. **The following are the number of mistakes made in 5 successive days for 4 technicians working for a photographic laboratory.**

Technician I	Technician II	Technician III	Technician IV
6	14	10	9
14	9	12	12
10	12	7	8
8	10	15	10
11	14	11	11

Test at the level of significance $\alpha = 0.01$ where the difference among the 4 sample means can be attributed to chance.

[6][2069 Poush]

Solution:

1. Null hypothesis: There is no significant difference between mistakes made by four technicians.

   Symbolically, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

2. Alternative hypothesis: There is significant difference between mistake made by at least one technician with others.

   Symbolically, $H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

3. Constructing ANOVA table

$n_1 = n_2 = n_3 = n_4 = 5$

$N = \Sigma n_i = 5 \times 4 = 20$

$K = 4$

Tech 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Tech 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Tech 3 (y_3)	$(y_3 - \bar{y}_3)^2$	Tech 4 (y_4)	$(y_4 - \bar{y}_4)^2$
6	14.44	14	4.84	10	1	9	1
14	17.64	9	7.84	12	1	12	4
10	0.04	12	0.04	7	16	8	4
8	3.24	10	3.24	15	16	10	0
11	1.44	14	4.84	11	0	11	1
$\Sigma = 49$	$\Sigma = 36.8$	$\Sigma = 59$	$\Sigma = 20.8$	$\Sigma = 55$	$\Sigma = 34$	$\Sigma = 50$	$\Sigma = 10$

$\bar{y}_1 = \frac{49}{5} = 9.8; \quad \bar{y}_2 = \frac{59}{5} = 11.8; \quad \bar{y}_3 = \frac{55}{5} = 11; \quad \bar{y}_4 = \frac{50}{5} = 10$

$T = 49 + 59 + 55 + 50 = 213$

$\bar{\bar{y}} = \frac{T}{N} = \frac{213}{20} = 10.65$

$$SST_r = \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{y})^2$$

$$= 5(9.8 - 10.65)^2 + 5(11.8 - 10.65)^2 + 5(11 - 10.65)^2 + 5(10 - 10.65)^2$$

$$= 12.95$$

$$SSE = \sum_{i=1}^{4} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 36.8 + 20.8 + 34 + 10 = 101.6$$

Now,

Variation	d.f.	SS	MS	F-ratio
Treatment	K – 1 = 4 – 1 = 3	$SST_r = 12.95$	$MST_r = \frac{12.95}{3} = 4.317$	$\frac{MST_r}{SSE} = \frac{4.317}{6.35} = 0.68$
Error	N – K = 20 – 4 = 16	SSE = 101.6	$MSE = \frac{101.6}{16} = 6.35$	
Total	19	114.55		

4. For $\alpha = 0.01$ and d.f. = (3, 16), $F_{0.01, (3, 16)} = 5.29$
5. Since F (= 0.68) < $F_{0.01, (3, 16)}$ (= 5.29), we accept H_0.
   Thus, there is no significant difference among the number of mistakes by four different technicians. Thus, the difference between the means, if seen be attributed to chance.

5. To investigate the average time it takes to assemble a certain computer component, the industrial engineer at an electronics firm timed 40 technicians in the performance of this task, getting a mean of 12.73 minutes and a standard deviation of 2.06 minutes. Construction a 98% confidence interval for the true average time it takes to assemble the component? [4][2070 Magh]

Solution:

Given that;

$n = 40$, $\bar{X} = 12.73$ minutes, $S = 2.06$ minutes

$\alpha = 100 - 98 = 2\% = 0.02$

Now, since $n > 30$, $S \approx \sigma$. So,

$$\text{Confidence limits} = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$= 12.73 \pm Z_{\frac{0.02}{2}} \times \frac{2.06}{\sqrt{40}}$$

$$= 12.73 \pm Z_{0.01} \times \frac{2.06}{\sqrt{40}}$$

$$= 12.73 \pm 2.33 \times \frac{2.06}{\sqrt{40}} \quad [\text{From table } Z_{0.01} = 2.33]$$

$$= (11.971, 13.489) \text{ minutes}$$

6. The table given below shows the lifetime in hours of samples from three different types of television tubes manufactured by a company. Set up an ANOVA table and test at 0.05 level of significance whether there is a difference in the three types.

Samples	Observations			
A	407	411	409	
B	404	406	408	405
C	410	408	406	

[6][2070 Magh]

Solution:

Setting up hypothesis:

1. Null hypothesis: There is no significant difference between the lifetime hours of different types of television tubes.
   Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3$
2. Alternative hypothesis: There is a significant difference of lifetime hours of at least one type of tube from others.
   Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3$
3. Setting up ANOVA table

A (y_1)	$(y_1 - \bar{y}_1)^2$	B (y_2)	$(y_2 - \bar{y}_2)^2$	C (y_3)	$(y_3 - \bar{y}_3)^2$
407	4	404	3.0625	410	4
411	4	406	0.0625	408	0
409	0	408	5.0625	406	4
		405	0.5625		
Σ = 1227	Σ = 8	Σ = 1623	Σ = 8.75	Σ = 1224	Σ = 8

$n_1 = 3$, $n_2 = 4$, $n_3 = 3$

$N = n_1 + n_2 + n_3 = 3 + 4 + 3 = 10$

$K = 3$

$$\bar{y}_1 = \frac{1227}{3} = 409; \quad \bar{y}_2 = \frac{1623}{4} = 405.75; \quad \bar{y}_3 = \frac{1224}{3} = 408$$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{1227 + 1623 + 1224}{10} = 407.4$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 3(409 - 407.4)^2 + 3(405.75 - 407.4)^2 + 3(408 - 407.4)^2$$

$$= 19.65$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 8 + 8.75 + 8 = 24.75$$

Variation	d.f.	SS	MS	F-ratio
Treatment	K - 1 = 3 - 1 = 2	$SST_r = 19.65$	$MST_r = \frac{19.65}{2} = 9.825$	$\frac{MSTr}{SSE} = \frac{9.825}{3.54} = 2.775$
Error	N - K = 10 - 3 = 7	SSE = 24.75	$MSE = \frac{24.75}{7} = 3.54$	
Total	9			

4. From F value table for $\alpha = 0.05$ and d.f. = (2, 7), $F_{0.05,(2,7)} = 4.7374$
5. $F (= 2.775) < F_{0.05,(2,7)} (= 4.7374)$. So, we accept H_0.
   Thus, there is no significant difference between the lifetime hours of three different types of television tubes.

7. **10 samples of standard cement had shown an average weight percent calcium with mean 90 and variance 25. 15 samples of the lead-dropped cement had an average weight percent of 87 with variance 16. If the weight percent in population is normally distributed with same variance, test the hypothesis for equality of mean at $\alpha = 0.01$ level of significance.** [5][2070 Magh]

Solution:

Given that;

$n_1 = 10, \ \bar{X}_1 = 90, \ S_1^2 = 25$

$n_2 = 15, \ \bar{X}_2 = 87, \ S_2^2 = 16$

Given, $\sigma_1^2 = \sigma_2^2$

Since, 'σ_1' and 'σ_2' are unknown and the sample sizes are small, we use 't' test.

Setting up hypothesis:

1. Null hypothesis: The mean of the two cement types, calcium percentage is same.
   Symbolically, $H_0: \mu_1 = \mu_2$
2. Alternative hypothesis: The mean of the two cement types' calcium percentage is not same.
   Symbolically, $H_1: \mu_1 \neq \mu_2$
3. Finding 't' value *i.e.*, test statistic.
   We have,

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(90 - 87)}{\sqrt{19.52\left(\frac{1}{10} + \frac{1}{15}\right)}} = 1.6632$$

$$\text{where, } S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(10 - 1) \times 25 + (15 - 1) \times 16}{10 + 15 - 5} = 19.52$$

4. Also, critical value of t at $\alpha = 0.01$
   and d.f. $= n_1 + n_2 - 2 = 10 + 15 - 2 = 23$
   $t_{\frac{\alpha}{2}, d.f.} = t_{\frac{0.01}{2}, 23} = t_{0.005, 23} = 2.807$
5. Since $t(= 1.6632) < t_{0.005, 23} (= 2.807)$, we accept H_0. So, the mean of the two cement types calcium percentage is same.

8. **A company produces automobile tyres, the manager of the company wants to estimate the limits in which expected trend life of his tyres will probably lie. A test sample of 64 tyres was taken and a test run showed the average trend life of 50000 miles. Find the 95% and 99% confidence limits for population mean. Given that population standard deviation is 3000 miles.** [4][2070 Bhadra]

Solution:

$\bar{X} = 50000$ miles, $n = 64$, $\sigma = 3000$ miles

We know;

$$\text{Confidence limits} = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

From table;

At $\alpha = 5\%$ (*i.e.*, 95% confidence)

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

$$Z_{\frac{\alpha}{2}} = 1.96$$

At $\alpha = 1\%$ (*i.e.*, 95% confidence)

$$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$$

$$Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.575$$

So,

i) 95% confidence limits

$$\mu = 50000 \pm 1.96 \times \frac{3000}{\sqrt{64}} = (49265, 50735) \text{ miles}$$

ii) 99% confidence limits

$$\mu = 50000 \pm 2.575 \times \frac{3000}{\sqrt{64}} = (49034.375, 50965.625) \text{ miles}$$

9. **The following data represents the units of production per day turned out by 3 different brands of machines used by 3 machines.**

Machines	Production			
1	15	14	19	18
2	17	12	20	16
3	16	18	16	17

Using ANOVA test whether the difference in performance of the three brands of machines are significant. Use $\alpha = 5\%$.

[6][2070 Bhadra]

Solution:

1. Null hypothesis: The performance of the three machines are same.
   Symbolically, $H_0: \mu_1 = \mu_2 = \mu_3$
2. Alternative hypothesis: The performance of at least one machine is different from others.
   Symbolically, $H_1: \mu_1 \neq \mu_2 \neq \mu_3$
3. Constructing ANOVA table

Machine 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Machine 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Machine 3 (y_3)	$(y_3 - \bar{y}_3)^2$
15	2.25	17	0.5625	16	0.5625
14	6.25	12	18.0625	18	1.5625
19	6.25	20	17.0625	16	0.5625
18	2.25	16	0.0625	17	0.0625
$\Sigma = 66$	$\Sigma = 17$	$\Sigma = 65$	$\Sigma = 32.75$	$\Sigma = 67$	$\Sigma = 2.75$

$n_1 = n_2 = n_3 = 4$

$N = \Sigma n_i = 4 \times 3 = 12$

$K = 3$

$$\bar{y}_1 = \frac{66}{4} = 16.5; \quad \bar{y}_2 = \frac{65}{4} = 16.25; \quad \bar{y}_3 = \frac{67}{4} = 16.75$$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{66 + 65 + 67}{12} = 16.5$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 4(16.5 - 16.5)^2 + 4(16.25 - 16.5)^2 + 4(16.75 - 16.5)^2$$

$$= 0.5$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2$$

$$= 17 + 32.75 + 2.75 = 52.5$$

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 3 - 1 = 2$	$SST_r = 0.5$	$MST_r = 0.25$	$\frac{MSTr}{SSE} = 0.043$
Error	$N - K = 12 - 3 = 9$	$SSE = 52.5$	$MSE = 5.833$	
Total	11	53		

4. From table $F_{0.05\,(2,\,9)} = 4.265$

5. Since $F (= 0.043) < F_{0.05\,(2,\,9)} (= 4.265)$. We, accept H_0.
   Thus, there is no significant difference between the performance of the three machines.

10. **The average hourly wage of sample of 150 workers in a plant 'A' was Rs. 2.56 with a standard deviation of Rs. 1.08. The average of a sample of 200 workers in plant 'B' was Rs 2.87 with a standard deviation of Rs. 1.28. Can an applicant safely assume that the hourly wages paid by plant 'B' are higher than those paid by plant 'A'? Use $\alpha = 0.05$.** **[6][2070 Bhadra]**

Solution:

Setting up null and alternative hypothesis.

1. Null hypothesis: The average wage paid by plant A is equal to that paid by plant B.
   Symbolically, $H_0: \mu_1 = \mu_2$
2. Alternative hypothesis: The average wage paid by plant A is less than that paid by plant B.
   Symbolically, $H_1: \mu_1 \neq \mu_2$
   This is a left tailed test. Since the samples are large even though 'σ_1' and 'σ_2' are unknown, we use 'Z' test.
3. Now, we calculate 'Z' value:
   Given: $\bar{X}_1 = 2.56$; $S_1 = 1.08$; $n_1 = 150$
   $\bar{X}_2 = 2.87$; $S_1 = 1.28$; $n_1 = 200$
   Assuming $\sigma_1 \neq \sigma_2$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(2.56 - 2.87)}{\sqrt{\frac{1.08^2}{150} + \frac{1.28^2}{200}}} = -2.4532$$

4. At $\alpha = 0.05$, for single tailed test
   $Z_\alpha = Z_{0.05} = -1.645$
5. $|Z| = |-2.4532| = 2.4532$
   $|Z_{0.05}| = |-1.645| = 1.645$
   Since $|Z| > |Z_{0.05}|$, we reject null hypothesis.
   Thus, the applicant can safely assume that the hourly wages paid by plant 'B' are higher.

11. **An examination was given to 50 students at college A and 60 students at college B. At A, mean grade was 75 with standard deviation of 9. At B, mean grade was 79 with standard deviation of 7. Is there significant difference between the performance of students at A and those at B, given that $\alpha = 0.05$?** **[6][2071 Magh]**

Solution:

Given that;

At A; $n_1 = 50$; $\bar{X}_1 = 75$, $S_1 = 9$

At B; $n_2 = 60$; $\bar{X}_2 = 79$; $S_2 = 7$

Even though 'σ_1' and 'σ_2' are not given, since the sample sizes are large, we can approximate $S_1 = \sigma_1$ and $S_2 = \sigma_2$. So, we follow Z-test here.

1. Null hypothesis: There is no significant difference in the performance of students at A and B.

   i.e., $H_0: \mu_1 = \mu_2$

2. Alternative hypothesis: There is significant difference in the performance of students at A and B.

   i.e., $H_1: \mu_1 \neq \mu_2$

3. Now, the Z value is given by

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(75 - 79)}{\sqrt{\frac{9^2}{50} + \frac{7^2}{60}}} = -1.561$$

4. From table at $\alpha = 5\%$ *i.e.*, 0.05, we find

$$Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$$

5. Since $|Z|$ (= 1.561) < $Z_{0.025}$ (= 1.96), we accept H_0. Thus, there is no significant difference in the performance of students at A and B.

12. Write down the steps for testing hypothesis on difference between two population means for large sample? [6][2071 Magh]

Solution: See theory section 5.4.1.

13. Three randomly selected groups of chickens are fed on three different diets. Each group consists of five chickens. Their weight gains during a specified period of time are as follows:

Diet I	4	4	7	7	8
Diet II	3	4	5	6	7
Diet III	6	7	7	7	8

Test the hypothesis that mean gains of weight due to the three diets are equal. [6][2071 Magh]

Solution:

We first hypothesis for testing

1. Null hypothesis: The weight gains due to 3-diets are equal.

   i.e., $H_0: \mu_1 = \mu_2 = \mu_3$

2. Alternative hypothesis: The mean weight gain of at least one diet is different.

   i.e., $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

3. Now, we prepare ANOVA table.

Diet I (y_1)	$(y_1 - \bar{y}_1)^2$	Diet II (y_2)	$(y_2 - \bar{y}_2)^2$	Diet III (y_3)	$(y_3 - \bar{y}_3)^2$
4	4	3	4	6	1
4	4	4	1	7	0
7	1	5	0	7	0
7	1	6	1	7	0
8	4	7	4	8	1
$\Sigma = 30$	$\Sigma = 14$	$\Sigma = 25$	$\Sigma = 10$	$\Sigma = 35$	$\Sigma = 2$

$n_1 = n_2 = n_3 = 5$

$N = \Sigma n_i = 15$

$K = 3$

$$\bar{y}_1 = \frac{30}{5} = 6; \quad \bar{y}_2 = \frac{25}{5} = 5; \quad \bar{y}_3 = \frac{35}{5} = 7$$

Now,

$T = 30 + 25 + 35 = 90$

$$\bar{\bar{y}} = \frac{T}{N} = \frac{90}{15} = 6$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 5(6-6)^2 + 5(5-6)^2 + (7-6)^2 = 10$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 14 + 10 + 2 = 26$$

Variation	d.f.	SS	MS	F-ratio
Treatment	K – 1 = 3 – 1 = 2	$SST_r = 10$	$MST_r = \frac{10}{2} = 5$	$\frac{MST_r}{SSE} = \frac{5}{2.17} = 2.304$
Error	N – K = 15 – 3 = 12	SSE = 26	$MSE = \frac{26}{12} = 2.17$	
Total	14	36		

4. At $\alpha = 5\%$ (assume) and d.f. = (2, 12). From table $F_{0.05, (2, 12)} = 3.8853$

5. Since F (= 2.304) < $F_{0.05, (2, 12)}$ (=3.8853), we accept H_0. Thus, the mean weight gains due to 3 diets are equal.

14. As part of the investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three different positions on the roof. The forces required to shear each of these bolts (coded values) are as follows:

Position 1	90	82	79	98	83	91	-
Position 2	105	89	93	104	89	95	86
Position 3	83	89	80	94	-	-	-

Performa an analysis of variance to test at the 0.05 level of significance whether the differences among the sample means at the three positions are significant. [6][2071 Bhadra]

Solution:

1. Null hypothesis: The sample means of shear resistance bolts at three positions are equal.
   Symbolically, $H_0: \mu_1 = \mu_2 = \mu_3$
2. Alternative hypothesis: The sample means of shear resistance of bolts at three positions are significantly different.
   Symbolically, $H_1: \mu_1 \neq \mu_2 \neq \mu_3$
3. Formation of ANOVA table.
   Here,
   $n_1 = 6, n_2 = 7, n_3 = 4$
   $N = n_1 + n_2 + n_3 = 6 + 7 + 4 = 17$
   $K = 3$

Position 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Position 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Position 3 (y_3)	$(y_3 - \bar{y}_3)^2$
90	8.03	105	111.77	83	12.25
82	26.69	89	29.463	89	6.25
79	66.7	93	2.04	80	42.25
98	117.35	104	91.623	94	56.25
83	17.36	89	29.463		
91	14.7	95	0.327		
		86	71.031		
$\Sigma = 523$	$\Sigma = 250.83$	$\Sigma = 661$	$\Sigma = 335.717$	$\Sigma = 346$	$\Sigma = 117$

$$\bar{y}_1 = \frac{523}{6} = 87.167; \quad \bar{y}_2 = \frac{661}{7} = 94.428; \quad \bar{y}_3 = \frac{346}{4} = 86.5$$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{523 + 661 + 346}{17} = 90$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 6(87.167 - 90)^2 + 7(94.428 - 90)^2 + 4(86.5 - 90)^2$$

$$= 234.4056$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2$$

$$= 250.83 + 335.717 + 117 = 703.547$$

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 3 - 1 = 2$	$SST_r = 234.4056$	$MST_r = \frac{234.4056}{2} = 117.2$	$\frac{MSTr}{SSE} = \frac{117.2}{50.25} = 2.3323$
Error	$N - K = 17 - 3 = 14$	$SSE = 703.547$	$MSE = \frac{703.547}{14} = 50.25$	
Total	16			

4. At $\alpha = 5\%$, and d.f. = (2, 14). $F_{0.05, (2, 14)} = 3.7389$
5. Since F_{cal} (= 2.3323) < F_{tab} (=3.7389). Thus, accept H_0.
   Hence, there is no significant difference in the bolt at those three positions.

15. **As part of an industrial training program, some trainees are instructed by method A, which is straight teaching machine instruction, and some are instructed by method B, which also involved the personal attention of an instructor. If random sample of size 10 are taken from large group of trainees instructed by each of the two methods and the scores which they obtained in an appropriate achievement test are:**

Method A	71	75	65	69	73	66	68	71	74	68
Method B	72	77	84	78	69	70	77	73	65	75

Use 0.05 level of significance to test the claim that method B is more effective. [5][2072 Magh]

Solution:

Since 'σ' for both A and B are unknown and the sample size is small, we use 't' test.

From calculator;

$\bar{X}_A = \bar{X}_1 = 70;$ $\bar{X}_B = \bar{X}_2 = 74$

$S_A = S_1 = 3.3664;$ $S_B = S_2 = 5.3955$

$n_1 = 10;$ $n_2 = 10$

First, we set up null and alternative hypothesis.

1. Null hypothesis: Method B and Method A produce equal mean scores. Symbolically, $H_0: \mu_1 = \mu_2$.
2. Alternative hypothesis: Method B produces greater mean score than Method A. Symbolically, $H_1: \mu_1 > \mu_2$ or, $H_1: \mu_1 < \mu_2$.
   Clearly, this is a left tailed test.
3. Now, we calculate test statistic 't' assume $\sigma_1 = \sigma_2$
   For unknown and equal σ_1 and σ_2

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(70 - 74)}{\sqrt{20.22237\left(\frac{1}{10} + \frac{1}{10}\right)}} = -1.989$$

where, $S_p^2 = \dfrac{(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2}{n_1 + n_2 - 2}$

$$= \frac{(10-1)\,(3.3665)^2 + (10-1)\,(5.3955)^2}{10 + 10 - 2} = 20.22237$$

4. At $\alpha = 0.05$ for d.f. $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$, we have
   $t_{0.05,\,18} = -1.734$ (from table)
5. Since $|t|$ (=1.989) > $|t_{0.05,\,18}|$ (= +1.734), we reject null hypothesis
   We accept H_1.
   Thus, the claim that method B is more effective is true.

16. Shyam and Co. produces three varieties of products deluxe, fine and ordinary. A recent market survey is conducted for preference of products. The preference was found as follow:

Product	Production			
Deluxe	15	14	19	18
Fine	17	12	20	16
Ordinary	16	18	16	17

Is there a significant difference in the preference of products using ANOVA test. Use $\alpha = 5\%$. **[5][2072 Magh]**

Solution:

Setting up hypothesis:

1. Null hypothesis: The preference for all 3 products are same.
   Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3$
2. Alternative hypothesis: The preference of at least one product is higher than others.
   Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3$
3. Construction of ANOVA table

Deluxe (y_1)	$(y_1 - \bar{y}_1)^2$	Fine (y_2)	$(y_2 - \bar{y}_2)^2$	Ordinary (y_3)	$(y_3 - \bar{y}_3)^2$
15	2.25	17	0.5625	16	0.5625
14	6.25	12	18.0625	18	1.5625
19	6.25	20	14.0625	16	0.5625
18	2.25	16	0.0625	17	0.0625
$\Sigma = 66$	$\Sigma = 17$	$\Sigma = 65$	$\Sigma = 32.75$	$\Sigma = 67$	$\Sigma = 2.75$

$n_1 = n_2 = n_3 = 4$

$\bar{y}_1 = \dfrac{66}{4} = 16.5;\ \bar{y}_2 = \dfrac{65}{4} = 16.25;\ \bar{y}_3 = \dfrac{67}{4} = 16.75$

Now,

$N = n_1 + n_2 + n_3 = 12$

$K = 3$

$$\bar{\bar{y}} = \frac{T}{N} = \frac{66 + 65 + 67}{12} = 16.5$$

$$SST_r = \sum_{i=1}^{3} n_i\,(\bar{y}_i - \bar{\bar{y}})^2$$

$$= 4(16.5 - 16.5)^2 + 4(16.25 - 16.5)^2 + 4(16.75 - 16.5)^2$$

$$= 0.5$$

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2$$

$$= \Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2 + \Sigma(y_3 - \bar{y}_3)^2$$

$$= 17 + 32.75 + 2.75$$

$$= 52.5$$

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 3 - 1 = 2$	$SST_r = 0.5$	$MST_r = \dfrac{0.5}{2} = 0.25$	$\dfrac{MSTr}{SSE}$
Error	$N - K = 12 - 3 = 9$	$SSE = 52.5$	$MSE = \dfrac{52.5}{9} = 5.83$	$= \dfrac{0.25}{5.83}$
Total	11	$SST = 53$		$= 0.043$

4. At $\alpha = 5\%$, and d.f. = (2, 9). $F_{0.05,\,(2,\,9)} = 4.2565$
5. Since $F(=0.043) < F_{0.05,\,(2,\,9)}$ (= 4.2565), we accept H_0.
   Thus, there is no significant difference in the preference of products.

17. Four salesmen were posted in different areas by a company. The number of units of computer hard disk sold by them are as follows:

Salesmen	Sales		
A	20	23	28
B	25	32	30
C	23	28	35
D	15	21	19

Is there a significant difference in the preference of salesmen? **[5] [2072 Magh]**

Solution:

We, first set up null and alternative hypothesis.

1. Null hypothesis: There is no significant difference in the performance of salesmen.
   Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3$
2. Alternative hypothesis: There is significant difference in sales of at least one salesman.
   Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

3. We now construct ANOVA table

A (y_1)	$(y_1 - \bar{y}_1)^2$	B (y_2)	$(y_2 - \bar{y}_2)^2$	C (y_3)	$(y_3 - \bar{y}_3)^2$	D (y_4)	$(y_4 - \bar{y}_4)^2$
20	13.44	25	16	23	32.11	15	11.11
23	0.44	32	9	28	0.44	21	7.11
28	18.78	30	1	35	40.11	19	0.44
$\Sigma = 71$	$\Sigma = 32.66$	$\Sigma = 87$	$\Sigma = 26$	$\Sigma = 86$	$\Sigma = 72.66$	$\Sigma = 55$	$\Sigma = 18.66$

$n_1 = n_2 = n_3 = n_4 = 3$

$N = \Sigma n_i = 3 \times 4 = 12$

$K = 4$

$\bar{y}_1 = \frac{71}{3} = 23.67;$ $\bar{y}_2 = \frac{87}{3} = 29$

$\bar{y}_3 = \frac{86}{3} = 28.67;$ $\bar{y}_4 = \frac{55}{3} = 18.33$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{71 + 87 + 86 + 55}{12} = 24.9167$$

$$SST_r = \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 3(23.67 - 24.9167)^2 + 3(29 - 24.9167)^2 + 3(28.67 - 24.9167)^2 + 3(18.33 - 24.9167)^2$$

$$= 227.0984$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (\bar{y}_{ij} - \bar{\bar{y}}_i)^2$$

$$= 32.66 + 26 + 72.66 + 18.66$$

$$= 149.98$$

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 4 - 1 = 3$	$SST_r = 227.098$	$MST_r = \frac{227.098}{3} = 75.699$	$\frac{MSTr}{SSE} = \frac{75.699}{18.75} = 4.0378$
Error	$N - K = 12 - 4 = 8$	$SSE = 149.98$	$MSE = \frac{149.98}{8} = 18.7475$	
Total	11	377.078		

4. For $\alpha = 5\%$ and d.f. (3, 8). $F_{0.05, (3, 8)} = 4.0662$
5. F-ratio (= 4.0378) < $F_{0.05, (3, 8)}$ (= 4.0662). So, we accept H_0. Hence, there is no significant difference in the performance of salesmen.

18. The following table represents the sales of three salesmen in your different districts:

Districts	Sales figure		
A	14	20	16
B	12	23	15
C	10	20	10
D	8	18	12

Perform an analysis of variance to test whether there is any significant difference in the sales of different district at 0.05 level of significance. **[5][2072 Ashwin]**

Solution:

We first setup null and alternative hypothesis

1. Null hypothesis: There is no significant difference in the sales of different districts. Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
2. Alternative hypothesis: There is significant difference of least one district sales from the others. Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.
3. Now, we formulate ANOVA table.

Let district A, B, C, D be represented by 1, 2, 3 and 4 respectively.

A (y_1)	$(y_1 - \bar{y}_1)^2$	B (y_2)	$(y_2 - \bar{y}_2)^2$	C (y_3)	$(y_3 - \bar{y}_3)^2$	D (y_4)	$(y_4 - \bar{y}_4)^2$
14	7.111	12	21.778	10	11.111	8	21.778
20	11.111	23	40.111	20	44.444	18	28.444
16	0.444	15	2.778	10	11.111	12	0.444
$\Sigma = 50$	$\Sigma = 18.666$	$\Sigma = 50$	$\Sigma = 64.667$	$\Sigma = 40$	$\Sigma = 66.666$	$\Sigma = 38$	$\Sigma = 50.666$

$\bar{y}_1 = \frac{50}{3} = 16.667;$ $\bar{y}_2 = \frac{50}{3} = 16.667$

$\bar{y}_3 = \frac{40}{3} = 13.33;$ $\bar{y}_4 = \frac{38}{3} = 12.667$

$N = 3 + 3 + 3 + 3 = 12$

$K = 4$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{50 + 50 + 40 + 38}{12} = 14.8333$$

$$SST_r = \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 3(16.667 - 14.8333)^2 + 3(16.667 - 14.8333)^2 + 3(13.33 - 14.8333)^2 + 3(12.667 - 14.8333)^2$$

$$= 41.03$$

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_i)^2$$
$$= \Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2 + \Sigma(y_3 - \bar{y}_3)^2 + \Sigma(y_4 - \bar{y}_4)^2$$
$$= 18.666 + 64.667 + 66.666 + 50.666$$
$$= 200.665$$

Representing the above data in tabular form:

Variation	d.f.	SS	MS	F-ratio
Treatment	K - 1 = 4 - 1 = 3	$SST_r = 41.03$	$MST_r = \frac{41.03}{3} = 13.68$	$\frac{MSTr}{SSE} = \frac{13.68}{25.08} = 0.545$
Error	N - K = 12 - 4 = 8	SSE = 200.665	$MSE = \frac{200.665}{8} = 25.08$	
Total	11	SST = 241.695		

4. For $\alpha = 0.05$ at d.f. (3, 8). $F_{0.5,(3,8)} = 4.0662$
5. F-ratio (= 0.545) < $F_{0.5,(3,8)}$ (= 4.0662). So, we accept H_0. Thus, there is no significant difference between sales of different districts.

19. Differentiate between (a) Confidence level and significance level (b) Type I and Type II error of hypothesis testing. [5][2072 Ashwin]

Solution:

a) See theory section 5.1.1. b) See 5.2 step (2).

20. A manufacturer of sports equipment has developed a new synthesis fishing line that the company claims has a mean breaking strength of 7.8 kilograms with a s.d. of 0.5 kg. Test the hypothesis that $\mu = 8$ kg against the alternative that $\mu \neq 8$ kg if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kg. Use a 0.01 level of significance. [5] [2073 Magh]

Solution:

1. Null hypothesis: The mean breaking strength is 8 kg. Symbolically, $H_0: \mu = 8$ kg.
2. Alternative hypothesis: The mean breaking strength is not 8 kg. Symbolically, $H_1: \mu \neq 8$ kg.
   From 1 and 2. We can see that it's a two tailed test.
3. Since n = 50 (*i.e.*, large sample size > 30), we use Z-distribution:
   $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
   Here, $\mu = 8$ kg, $\bar{X} = 7.8$ kg, $\sigma = 0.5$ kg, n = 50
   $$\text{So, } Z = \frac{(7.8 - 8)}{0.5/\sqrt{50}} = -2.828$$

4. For $\alpha = 0.01$, from Z-table, we get
   $$Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005} = \pm 2.575$$
5. Since $|Z| = 2.8258 > Z_{0.005}$ (= 2.575), it falls in the rejection zone. Thus, reject H_0. Hence, the mean breaking strength not equal to 8 kg.

21. A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean of 81. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean 76. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Use 5% level of significance. [5] [2073 Magh]

Solution:

Given that;

Sample 1: $n_1 = 25$, $\sigma_1 = 5.2$, $\bar{x}_1 = 81$

Sample 2: $n_2 = 36$, $\sigma_2 = 3.4$, $\bar{x}_2 = 76$

1. Null hypothesis: Mean of 2 populations are equal.
   Symbolically, $H_0: \mu_1 = \mu_2$
2. Alternative hypothesis: The mean of 2 populations are unequal.
   Symbolically, $H_1: \mu_1 \neq \mu_2$.
3. The test statistic is:
   $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ (Since } \sigma_1 \text{ and } \sigma_2 \text{ are known and unequal)}$$
   $$i.e., \quad Z = \frac{(81 - 76) - (0)}{\sqrt{\frac{(5.2)^2}{25} + \frac{(3.4)^2}{36}}} = 4.2217$$
4. For a = 0.05; (from table)
   $$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = \pm 1.96$$
5. Since $|Z| = 4.2217 > 1.96$. So, reject null hypothesis. Thus, the mean of the two populations are not equal.

22. Three testing machines are used to determine the breaking load in tension of wire which is believed to be uniform. Nine pieces of wire are cut off, one after another. They are numbered consecutively, and three are assigned to each machine using random numbers. Random numbers are used also to determine the order in which specimens are tested on each machine. The breaking loads (in Newton's) found on each machine are shown in the table below.

Machine 1	1570	1750	1680
Machine 2	1890	1860	2390
Machine 3	1640	1760	2020

Do the data indicate (at α = 5%) that one or two of the machines give higher readings than others? (Value of $F_{2,6}$ = 5.1433)

[5] [2073 Magh]

Solution:

This is a question related to ANOVA test:

1. Null hypothesis: All three machines give equal readings.

   Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3$

2. Alternative hypothesis: The reading in one or two machine is higher than others.

   Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

3. Construction of ANOVA table

Method 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Method 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Method 3 (y_3)	$(y_3 - \bar{y}_3)^2$
1570	9344.44	1890	24544.44	1640	27777.78
1750	6944.44	1860	34844.44	1760	2177.78
1680	177.78	2390	117877.78	2020	45511.11
Σ = 5000	Σ = 16466.66	Σ = 6140	Σ = 177266.6	Σ = 5420	Σ = 75466.67

$$\bar{y}_1 = \frac{5000}{3} = 1666.67$$

$$\bar{y}_2 = \frac{6140}{3} = 2046.67$$

$$\bar{y}_3 = \frac{5420}{3} = 1806.67$$

$N = 3 + 3 + 3 = 9$

$K = 3$

Now,

$$\bar{y} = \frac{T}{N} = \frac{5000 + 6140 + 5420}{9} = 1840$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{y})^2$$

$$= 3(1666.67 - 1840)^2 + 3(2046.67 - 1840)^2 + 3(1806.67 - 1840)^2$$

$$= 221600.0001$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2$$

$$= \Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2 + \Sigma(y_3 - \bar{y}_3)^2$$

$$= 16466.66 + 177266.66 + 75466.67$$

$$= 269200.01$$

Representing the above data in tabular form:

Variation	d.f.	SS	MS	F-ratio
Treatment	K − 1 = 3 − 1 = 2	SST_r = 221600.001	$MST_r = \frac{221600}{2}$ = 110800	$\frac{MSTr}{MSE} = \frac{110800}{44866.67}$ = 2.4695
Error	N − K = 9 − 3 = 6	SSE = 269200.01	$MSE = \frac{269200}{6}$ = 44866.67	
Total	8	SST = 490800		

4. F = 2.4695 < $F_{2,6}$ (= 5.1433). So, accept null hypothesis. Hence, all three machines given equal readings.

23. **Distinguish between Z-test and t-test of hypothesis for testing simple population mean?** **[5][2073 Magh]**

Solution: See theory section 5.3.1 and 5.3.2; step 3, 4 and 5.

24. **Discuss errors of testing of hypothesis? Write the procedure of testing of small population?** **[5][2073 Magh]**

Solution:

a) See theory section 5.2 step (2). b) See theory section 5.3.2.

25. **At what condition pair test should be used. Write the procedure of testing of pair test.** **[5][2073 Magh]**

Solution: See theory section 5.6.

26. **Describe the procedure of the test of significance for paired t-test.** **[5][2073 Bhadra]**

Solution: See theory section 5.6.

27. **Three training methods were compared to see if they led to greater productivity after training. The following are productivity measures for individuals trained by each method.**

Method 1	45	40	50	39	52	44
Method 2	59	43	47	51	39	49
Method 3	41	37	43	40	52	39

Perform an analysis of variance to test at the 0.05 level of significance whether the three training methods lead to different levels of productivity? **[5][2073 Bhadra]**

Solution:

1. We first set null hypothesis (H_0) and alternate hypothesis (H_1) as:
   H_0: The three methods lead to same levels of productivity symbolic representation as; $H_0 : \mu_1 = \mu_2 = \mu_3$.
   H_1: They led to different level of productivity.
   i.e., $H_1: \mu_1 \neq \mu_2 \neq \mu_3$.
2. Assume level of significance (α) = 0.05 *i.e.*, 5%.
3. Formulation of ANOVA table:

Method 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Method 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Method 3 (y_3)	$(y_3 - \bar{y}_3)^2$
45	0	59	121	41	1
40	25	43	25	37	25
50	25	47	1	43	1
39	36	51	9	40	4
52	49	39	81	52	100
44	1	49	1	39	9
$\Sigma y_1 = 270$	$\Sigma = 136$	$\Sigma y_2 = 288$	$\Sigma = 238$	$\Sigma y_3 = 252$	$\Sigma = 140$

$$\bar{y}_1 = \frac{\Sigma y_1}{n_1} = \frac{270}{6} = 45$$

$$\bar{y}_2 = \frac{\Sigma y_2}{n_2} = \frac{288}{6} = 48$$

$$\bar{y}_3 = \frac{\Sigma y_3}{n_3} = \frac{252}{6} = 42$$

$N = n_1 + n_2 + n_3 = 6 + 6 + 6 = 18$

$K = 3$

Now,

$$\bar{\bar{y}} = \frac{T}{N} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3}{N} = \frac{270 + 288 + 252}{18} = 45$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 6(45 - 45)^2 + 6(48 - 45)^2 + 6(42 - 45)^2$$

$$= 108$$

$$SSE = \sum_{i=1}^{3}\sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2$$

$$= \Sigma(y_1 - \bar{y}_1)^2 + \Sigma(y_2 - \bar{y}_2)^2 + \Sigma(y_3 - \bar{y}_3)^2$$

$$= 136 + 238 + 140$$

$$= 514$$

Representing the above data in tabular form:

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 3 - 1 = 2$	$SST_r = 108$	$MST_r = \frac{108}{2} = 54$	$\frac{MSTr}{MSE} = 1.576$
Error	$N - K = 18 - 3 = 15$	$SSE = 514$	$MSE = \frac{514}{15} = 34.27$	
Total	$15 + 2 = 17$	$SST = 622$		

4. From F-table, for α = 5% at d.f. (2, 15); $F_{0.05, (2, 15)} = 3.6823$.
5. F-ratio = 1.576 < 3.6823. We accept H_0. Thus, the three methods lead to same level of productivity.

28. Shyam and co. produces there varieties of products: Deluxe, fine and ordinary. A recent market survey is conducted for preference of products. The preference was found as follows:

Product	Production			
Deluxe	15	14	19	18
Fine	17	12	20	16
Ordanery	16	18	16	17

Is there a significant difference in the preference of products using ANOVA test. [Use α = 5%] **[2074 Bhadra]**

Solution: Refer previous question.

29. A company produces automobile tyres, the manager of the company wants to estimate the limits in which expected trend life of his tyres will probably lie. A test sample of 64 tyres was taken and a test run showed average trend life of 50000 miles. Find the 95% and 99% confidence limits for population mean. Given that population standard is 3000 miles. **[2075 Baisakh]**

Solution: Refer previous question.

30. The following data represent the units of production per day turned out by these different brands of machines used by three mechanists:

Machines	Production			
1	15	14	19	18
2	17	12	20	16
3	16	18	16	17

Using ANOVA test whether the differences in performances of the three brands of machines are significant. [Use α = 5%] **[2075 Baisakh]**

Solution: Refer previous question.

31. A sample of 15 beams of a grand complex are given below in feet 3.

12.8	9.8	10.2	10.0	12.0
10.5	8.9	12.2	10.8	9.0
11.2	12.1	10.1	8.8	10.6

A civil engineer claimed that volume of beams are greater than 10 feet³ test whether his claim is right or wrong. [Use α = 1%]

[2075 Baisakh]

Solution:

1. Null hypothesis. The volumes of beams are equal to (or less than) 10 feet³. symbolically, $H_0 : \mu = 10\ ft^3$ (or $H_0: \mu \le 10\ ft^3$).
2. Alternative hypothesis: The volumes of beams are greater than 10 feet³. Symbolically, $H_1 = \mu > 10\ ft^3$.
3. Now, we find out test statistic:
   From the given data (using calculator)
   $n = 15$
   $\bar{x} = 10.6$
   $s = 1.257$
4. The number of samples is less than 30 and population standard deviation is not known. So, we use t-statistic.
   $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{10.6 - 10}{1.257/\sqrt{15}} = 1.8487$$
5. For α = 0.01 and d.f. = n – 1 = 15 – 1 = 14, since it's a right tailed test, the critical value is:
   $t_{\alpha, n-1} = t_{0.01, 14} = 2.624$
6. Since t(= 1.8487) < $t_{0.01, 14}$ (= 2.624), we accept H_0. Thus, the civil engineer's claim that the volume is greater than 10 feet³ is wrong.

32. In 39 soil samples, tasted for trace elements, the average amount amount of copper was found to be 22 milligrams, with a standard deviation of 4 milligrams. Find a 90% confidence interval for the true mean copper content in the soils from which these samples were taken.

[2075 Bhadra]

Solution:

$n = 39, \bar{x} = 22$ mg, $s = 4$ mg

$\alpha = (100 - 90)\% = 10\% = 0.1$

Now, the confidence interval is

$$= \bar{x} \pm Z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

$$= \bar{x} \pm Z_{\frac{0.1}{2}} \times \frac{s}{\sqrt{n}}$$

$$= \bar{x} \pm z_{0.05} \times \frac{s}{\sqrt{n}}$$

$$= 22 \pm 1.645 \times \frac{4}{\sqrt{39}}$$

$= (20.946, 23.054)$ milligrams.

33. An experiment was conducted to compare three methods of packing a certain frozen good. The criterion was the ascorbic acid content (mg/100 gm) after a specific period of time. The following data were obtained.

Packaging methods							
A	14.29	19.10	19.09	16.25	15.09	16.61	19.63
B	20.06	20.64	18.00	19.56	19.47	19.07	18.38
C	20.04	26.23	22.74	24.04	23.37	25.02	23.27

Do these data provide sufficient evidence at the 0.01 level of significance to indicate a difference in packaging method?

[2075 Bhadra]

Solution:

1. Null hypothesis: There is no significant difference in the three packaging methods. Symbolically, $H_0: \mu_1 = \mu_2 = \mu_3$.
2. Alternative hypothesis: There is significant difference in at least one. Packaging methods from the others. Symbolically, $H_1: \mu_1 \ne \mu_2 \ne \mu_3$.
3. The test statistic under ANOVA test is calculated as follows:
   $n_1 = n_2 = n_3 = 7$
   $N = \Sigma n_i = 21$
   $K = 3$

Variety $A(y_1)$	$(y_1 - \bar{y}_1)^2$	Variety $B(y_2)$	$(y_2 - \bar{y}_2)^2$	Variety $C(y_3)$	$(y_3 - \bar{y}_3)^2$
14.29	8.18	20.06	0.5625	20.04	12.1801
19.10	3.80	20.64	1.7689	26.23	7.29
19.09	3.76	18.00	1.7161	22.74	0.6241
16.25	0.81	19.56	0.0625	24.04	0.2601
15.09	4.24	19.47	0.0256	23.37	0.0256
16.61	0.2916	19.07	0.0576	25.02	2.2201
19.63	6.15	18.38	0.8649	23.27	0.0676
Σ = 120.06	Σ = 27.2316	Σ = 135.18	Σ = 5.0581	Σ = 164.71	Σ = 22.6676

$$\bar{y}_1 = \frac{\Sigma y_1}{n_1} = \frac{120.06}{7} = 17.15$$

$$\bar{y}_2 = \frac{\Sigma y_2}{n_2} = \frac{135.18}{7} = 19.31$$

$$\bar{y}_3 = \frac{\Sigma y_3}{n_3} = \frac{164.71}{7} = 23.53$$

$$\bar{\bar{y}} = \frac{\Sigma y_1 + \Sigma y_2 + \Sigma y_3}{N}$$

$$= \frac{120.06 + 135.18 + 164.71}{21} = 20$$

$$SSTr = \sum_{i=1}^{3} (\bar{y}_i - \bar{\bar{y}})^2 n_i$$

$$= n_1(\bar{y}_1 - \bar{\bar{y}})^2 + n_2(\bar{y}_2 - \bar{\bar{y}})^2 + n_3(\bar{y}_3 - \bar{\bar{y}})^2$$

$$= 7(17.15 - 20)^2 + 7(19.31 - 20)^2 + 7(23.53 - 20)^2$$

$$= 147.4165$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$= 27.2316 + 5.0581 + 22.6676$$

$$= 54.9573$$

Variation	d.f.	SS	MS	F-ratio
Treatment	K – 1 = 3 – 1 = 2	SSTr = 147.4165	$MSTr = \frac{147.4}{2} = 73.7$	$\frac{MSTr}{MSE} = \frac{73.7}{3.05} = 24.164$
Error	N – K = 21 – 3 = 18	SSE = 54.9573	$MSE = \frac{54.96}{18} = 3.05$	
Total	N – 1 = 20	SST = 202.374		

4. For $\alpha = 0.01$ and d.f. = (2, 18). $F_{0.01}$, (2, 18) = 6.01.
5. Since $F_{calc.}$ (24.164) > F_{tab} (= 6.01), we reject H_0. Thus, the data provide sufficient evidence to indicate a difference in packaging methods.

SOLVED QUESTIONS (FOR BCT/BEX/BEL/B. Agri)

1. **Set up an ANOVA table for the following are production of data for 3 varieties of wheat, each growth on 4 plots and state if the variety differences are significant. Use $\alpha = 0.05$.** **[2070 Ashad]**

Plots of land	Variety of wheat		
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution:

1. Null hypothesis: The variety differences on the four plants of land are insignificant. Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
2. Alternative hypothesis: the variety difference of at least one plot of land from the other is significant. Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.
3. Now, we set up ANOVA table:

Plot 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Plot 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Plot 3 (y_3)	$(y_3 - \bar{y}_3)^2$	Plot 4 (y_4)	$(y_4 - \bar{y}_4)^2$
6	0.44	7	2.779	3	0	8	2.778
5	0.111	5	0.111	3	0	7	0.444
5	0.111	4	1.777	3	0	4	5.444
Σ = 16	Σ = 0.662	Σ = 16	Σ = 4.667	Σ = 9	Σ = 0	Σ = 19	Σ = 8.666

$n_1 = n_2 = n_3 = n_4 = 3$

$N = \Sigma n_i = 3 \times 4 = 12$

$K = 4$

$\bar{y}_1 = \frac{16}{3} = 5.333;$ $\bar{y}_2 = \frac{16}{3} = 5.333;$

$\bar{y}_3 = \frac{9}{3} = 3;$ $\bar{y}_4 = \frac{19}{3} = 6.333$

$$\bar{\bar{y}} = \frac{\Sigma y}{N} = \frac{16 + 16 + 9 + 19}{12} = 5$$

$$SST_r = \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 3(5.333 - 5)^2 + 5(5.333 - 5)^2 + 3(3 - 5)^2 + 3(6.333 - 5)^2$$

$$= 18$$

$$SSE = \sum_{i=1}^{4} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2$$

$$= 0.662 + 4.667 + 0 + 0.666$$

$$= 14$$

Now, Expressing these data in ANOVA table:

Variation	d.f.	SS	MS	F-ratio
Treatment	K – 1 = 4 – 1 = 3	$SST_r = 18$	$MST_r = \frac{18}{3} = 6$	$\frac{MST_r}{SSE} = \frac{6}{1.75} = 3.43$
Error	N – K = 12 – 4 = 8	SSE = 14	$MSE = \frac{14}{8} = 1.75$	
Total	11	32		

4. For $\alpha = 0.05$ and d.f. = (3, 8), $F_{0.05,(3,8)} = 4.0662$
5. Since F (= 3.43) < $F_{0.05,(3,8)}$ (= 4.0662), we accept H_0.
   Thus, the variety differences between the four plots of land are insignificant.

2. **The following random samples are measurement of the heat producing capacity (in millions of calories per ton) of specimens of coal from two mines.**

Mine 1	8260	8130	8350	8070	8340	-
Mine 2	7950	7890	7900	8140	7920	7840

Use the 0.01 level of significance to test whether the difference between means of these two samples is significant. [2070 Ashad]

Solution:

The sample sizes are small and, the population standard deviation 'σ' are unknown. So, we use 't' test for difference of two means.

1. Null hypothesis: The difference between means of heat producing capacity of two mines are insignificant. Symbolically, $H_0: \mu_1 = \mu_2$.
2. Alternative hypothesis: The difference between means of heat producing capacity of two mines are significant.
   Symbolically, $H_1: \mu_1 \neq \mu_2$.
3. The test statistic for the problem is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

From the data given (using calculator)

$n_1 = 5$, $n_2 = 6$

$\bar{X}_1 = 8230$, $\bar{X}_2 = 7940$

$S_1 = 125.5$, $S_2 = 104.5$

Now,

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(5-1) \times (125.5)^2 + (6-1) \times (104.5)^2}{5 + 6 - 2}$$

$$= 13066.9167$$

$$\text{So, } t = \frac{(8230 - 7940) - 0}{\sqrt{13066.9167 \times \left(\frac{1}{5} + \frac{1}{6}\right)}} = 4.18963 \approx 4.2$$

4. For a 2 tailed test and $\alpha = 0.01$

$$t_{\frac{\alpha}{2}, d.f} = t_{\frac{0.01}{2}, n_1 + n_2 - 2} = t_{0.005, 5+6-2} = t_{0.005, 9}$$

$$\therefore \quad t_{0.005, 9} = 3.250$$

5. Since, t(= 4.2) > $t_{0.005, 9}$ (= 3.250), we reject H_0. Thus, the difference between means of these two samples is significant.

3. **An analysis for PH (acidity) in an random sample of water from 40 rainfalls showed that mean is 6.7 and s.d. is 0.5. Find a 99% confidence interval for the mean PH in rainfalls.** [2070 Chaitra]

Solution:

$n = 40$, $\bar{X} = 6.7$, $S = 0.5$, $\alpha = 1\% = 0.01$

Given sample is large. So, we approximate $\sigma \approx S$ and use 'Z' parameter.

$$\therefore \text{ 99\% confidence interval } = \bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{S}{\sqrt{n}}$$

$$= 6.7 \pm Z_{\frac{0.01}{2}} \times \frac{0.5}{\sqrt{40}}$$

$$= 6.7 \pm Z_{0.005} \times \frac{0.5}{\sqrt{40}}$$

$$= 6.7 \pm 2.575 \times \frac{0.5}{\sqrt{40}}$$

$$= (6.496, 6.904)$$

4. **As a part of investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three different positions on the roof. The forces required to shear each of these bolts (coded) values are as follows:**

Position 1	**90, 82, 79, 98, 83, 91**
Position 2	**105, 89, 93, 104, 89, 95, 86**
Position 3	**83, 89, 80, 94**

Perform an ANOVA to test 0.05 level of significance whether the difference among the sample means at the three positions are significant. [2070 Chaitra]

Solution: See solution to [2071 Bhadra].

5. **The following are the average weekly losses of worker hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation:**

45 and 36, 73 and 60, 46 and 44, 124 and 119, 33 and 35, 57 and 51, 83 and 77, 34 and 29, 26 and 24, 17 and 11. Use the 0.05 level of significance to test whether the safety program is effective.

[2070 Chaitra]

Solution: See the solution to [2073 Shrawan].

6. The mean weight loss of n = 16 grinding balls after a certain length of time in mill slurry is 3.42 grams with a standard deviation of 0.68 gram. Construct a 99% confidence interval for the true mean weight loss of such grinding balls under the stated conditions.

[2071 Shrawan]

Solution:

$n = 16, \quad \bar{X} = 3.42 \text{ gm}, \quad S = 0.68 \text{ gram}$

Now, $\alpha = 100 - 99 = 1\% = 0.01$

$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$; d.f. = n – 1 = 16 – 1 = 15

For small sample size, we use 't' parameter.

$\therefore$ 99% confidence interval $= \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \times \frac{S}{\sqrt{n}}$

$= 3.42 \pm t_{0.005, 15} \times \frac{0.68}{\sqrt{16}}$

$= (2.919, 3.92)$ grams

7. Four trained operators works on production of new product. The productivity of the operators are recorded as below:

Operators	Production			
1	10	12	14	16
2	12	11	13	16
3	14	15	12	11
4	16	10	17	17

Using ANOVA, test whether the difference in average productivity due to the difference in operators are significant. Use $\alpha = 5\%$.

[2071 Shrawan]

Solution:

It's very similar to the solution of [2071 Chaitra]. Just one extra operator is to be added to the solution. The rest of the process is same.

8. The following are the average weekly losses of worker hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation.

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

Use the 0.05 level of significance to test whether the safety program is effective.

[2071 Shrawan]

Solution: See the solution to [2073 Shrawan].

9. A company claims that its light bulbs are superior to those of its main competitor. If a study showed that a sample of 40 of its bulbs has mean lifetime of 647 hours of continuous use with standard deviation of 27 hour. While a sample of lifetime of 638 hours off continuous use with standard deviation of 31 hours. Does this substantiate claim at 1% level of significance? [2072 Kartik]

Solution: See solution to [2069 Poush].

10. In the investigation of a citizens committee complaint about the availability of fire protection within the country, the distance in miles to the nearest fire station was measured for each of the five randomly selected residences in each of the four areas.
Do these data provide sufficient evidence to indicate a difference in mean distance for the four areas at the 0.05 level of significance?

Area 1	7	5	5	6	8
Area 2	1	4	3	4	5
Area 3	7	9	8	7	8
Area 4	4	6	3	7	5

[2072 Kartik]

Solution:

We use ANOVA test here since more than two means are to be compared.

1. Null hypothesis: There is no significant difference in the mean distance to the nearest fire station in the four areas. Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
2. Alternative hypothesis: There is significant difference in the mean distance to the nearest fire station of at least one area from the others. Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.
3. Now, we prepare ANOVA table

Area 1 (y_1)	$(y_1 - \bar{y}_1)^2$	Area 2 (y_2)	$(y_2 - \bar{y}_2)^2$	Area 3 (y_3)	$(y_3 - \bar{y}_3)^2$	Area 4 (y_4)	$(y_4 - \bar{y}_4)^2$
7	0.64	1	5.76	7	0.64	4	1
5	1.44	4	0.36	9	1.44	6	1
5	1.44	3	0.16	8	0.04	3	4
6	0.04	4	0.36	7	0.64	7	4
8	3.24	5	2.56	8	0.04	5	0
$\Sigma = 31$	$\Sigma = 6.8$	$\Sigma = 17$	$\Sigma = 9.2$	$\Sigma = 39$	$\Sigma = 2.8$	$\Sigma = 25$	$\Sigma = 10$

$n_1 = n_2 = n_3 = n_4 = 5$

$N = \Sigma n_i = 5 \times 4 = 20$

$K = 4$

$$\bar{y}_1 = \frac{31}{5} = 6.2; \quad \bar{y}_2 = \frac{17}{5} = 3.4; \quad \bar{y}_3 = \frac{39}{5} = 7.8; \quad \bar{y}_4 = \frac{25}{5} = 5$$

$$\bar{\bar{y}} = \frac{\Sigma y}{N} = \frac{31 + 17 + 39 + 25}{20} = 5.6$$

$$SST_r = \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 5(6.2 - 5.6)^2 + 5(3.4 - 5.6)^2 + 5(7.8 - 5.6)^2 + 5(5 - 5.6)^2$$

$$= 52$$

$$SSE = \sum_{i=1}^{4} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 6.8 + 9.2 + 2.8 + 10 = 28.8$$

Now, constructing ANOVA table

Variation	d.f.	SS	MS	F-ratio
Treatment	K – 1 = 4 – 1 = 3	$SST_r = 52$	$MST_r = \frac{52}{3} = 17.33$	$\frac{MST_r}{SSE}$
Error	N – K = 20 – 4 = 16	SSE = 28.8	$MSE = \frac{28.8}{16} = 1.8$	$= \frac{17.33}{1.8}$
Total	19	80.8		= 9.63

4. For $\alpha = 0.05$ and d.f. = (3, 16), $F_{0.05, (3, 16)} = 3.2389$
5. Since F (= 9.63) > $F_{0.05, (3, 16)}$ (=3.2389), we reject H_0.
   Thus, there is sufficient evidence to show significant difference between mean distance to the nearest fire station in the four areas.

11. The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected, and the sample means and sample variances are $\bar{X}_1 = 8.73$, $S_1^2 = 0.35$, $\bar{X}_2 = 8.68$ and $S_2^2 = 0.40$, respectively. Assume that $\sigma_1^2 = \sigma_2^2$ and that the data are drawn from a normal distribution. Is there evidence to support the claim that the two machines produce rods with different mean diameters. Use $\alpha = 0.05$ in arriving at this conclusion. [2072 Kartik]

Solution: Given,

$n_1 = 15, \quad \bar{X}_1 = 8.73, \quad S_1^2 = 0.35$

$n_2 = 17, \quad \bar{X}_2 = 8.68, \quad S_2^2 = 0.40$

1. Null hypothesis: The two machines produce rods with same mean diameters. Symbolically, $H_0: \mu_1 = \mu_2$.
2. Alternative hypothesis: The two machines produce rods with different mean diameters. Symbolically, $H_1: \mu_1 \neq \mu_2$.
3. Since the sample sizes are small with unknown 'σ' values, we use 't' test here (Note: 'σ' values denotes standard deviation of the

population which may be different from the standard deviation of the sample 'S_1' and 'S_2' which are given here).

The test statistic is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{where, } S_P^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(15 - 1) \times 0.35 + (17 - 1) \times 0.40}{15 + 17 - 2} = 0.37667$$

$$\text{Thus, } t = \frac{(8.73 - 8.68) - 0}{\sqrt{0.37667\left(\frac{1}{15} + \frac{1}{17}\right)}} = 0.22998$$

4. Now, for 2 tailed test, the critical value is:
   $t_{\frac{\alpha}{2}, n_1 + n_2 - 2} = t_{\frac{0.05}{2}, 15 + 17 - 2} = t_{0.025, 30} = 2.042$
5. Since, t (= 0.22998) < $t_{0.025, 30}$ (= 2.042), we accept H_0. Thus, there is no evidence to support the claim that the two machines produce rods with different mean diameters.

12. In a certain factory, there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 gram with a standard deviation of 12 gram, while the corresponding figures in a sample of 400 items from the other process are 124 and 14 respectively. Test whether the two mean weights differ significantly or not at 5 percent level of significance. [2071 Chaitra]

Solution:

Given that;

$n_1 = 250, \quad \bar{X}_1 = 120, \quad S_1 = 12$

$n_2 = 400, \quad \bar{X}_2 = 124, \quad S_2 = 14$

1. Null hypothesis: There is no significant difference in the mean weights of two different processes. Symbolically, $H_0: \mu_1 = \mu_2$.
2. Alternative hypothesis: There is significant difference in the mean weights of two different processes. Symbolically, $H_1: \mu_1 \neq \mu_2$. (2 tailed test)
3. The given samples are of large size. So, even though 'σ' values are not known, we approximate $\sigma \approx S$ and, use 'Z' test.
   The test statistic is:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad \text{(Assuming } \sigma_1 \neq \sigma_2\text{)}$$

$$= \frac{(120 - 124) - 0}{\sqrt{\frac{12^2}{250} + \frac{14^2}{400}}} = -3.8742$$

4. For a = 5%, since it's a 2 tailed test, we use

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

$$Z_{0.025} = \pm 1.96$$

5. Since, $|Z|$ (= 3.8742) > $Z_{0.025}$ (= 1.96), we reject H_0. Thus, the two mean weights differ significantly.

13. Three trained operators work on production of new product. The productivity of the operators are recorded as below:

Operators	Production			
1	10	12	14	16
2	12	11	13	16
3	14	15	12	11

Using ANOVA test whether the difference in average productivity due to the difference in operators are significant. Use α = 5%.

[2071 Chaitra]

Solution:

1. Null hypothesis: There is no significant difference in the productivity of the three operators. Symbolically, H_0: $\mu_1 = \mu_2 = \mu_3$.
2. Alternative hypothesis: there is significant difference in average productivity of at least one operator from the others. Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3$.
3. Now, we construct ANOVA table:

0 – 1 (y_1)	$(y_1 - \bar{y}_1)^2$	0 – 2 (y_2)	$(y_2 - \bar{y}_2)^2$	0 – 3 (y_3)	$(y_3 - \bar{y}_3)^2$
10	9	12	1	14	1
12	1	11	4	15	4
14	1	13	0	12	1
16	9	16	9	11	4
Σ = 52	Σ = 20	Σ = 52	Σ = 14	Σ = 52	Σ = 10

$n_1 = n_2 = n_3 = 4$

$N = \Sigma n_i = 4 \times 3 = 12$

$K = 3$

$$\bar{y}_1 = \frac{\Sigma y_1}{n_1} = \frac{52}{4} = 13; \quad \bar{y}_2 = \frac{\Sigma y_2}{n_2} = \frac{52}{4} = 13; \quad \bar{y}_3 = \frac{\Sigma y_3}{n_3} = \frac{52}{4} = 13$$

Now,

$$\bar{\bar{y}} = \frac{\Sigma y_1 + \Sigma y_2 + \Sigma y_3}{N} = \frac{52 + 52 + 52}{12} = 13$$

$$SST_r = \sum_{i=1}^{3} n_i (\bar{y}_i - \bar{y})^2$$

$$= 4(13 - 13)^2 + 4(13 - 13)^2 + 4(13 - 13)^2$$

$$= 0$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 20 + 14 + 10 = 44$$

Variation	d.f.	SS	MS	F-ratio
Treatment	K – 1 = 3 – 1 = 2	$SST_r = 0$	$MST_r = \frac{0}{2} = 0$	$\frac{MSTr}{SSE} = 0$
Error	N – K = 12 – 3 = 9	SSE = 44	$MSE = \frac{44}{9} = 4.89$	
Total	11	44		

4. For α = 5% and d.f. = (2, 9). $F_{0.05, (2, 9)} = 4.2565$.
5. Since F (= 0) < $F_{0.05, (2, 9)}$ (= 4.2565), we accept H_0.

Thus, the difference in average productivity due to the difference in operators are not significant.

14. There was a research on voltage supply by two companies. Both company claims that they supply same voltages. But researcher suspects there is significant difference between mean voltages between two companies. To test this, she selected independent samples from both company and in lab test, the result were follows:

	n	Mean	Sample standard deviation
Company A	13	3.59 V	0.3 V
Company B	10	3.15 V	0.4 V

Test the researcher's suspect was correct at 5% level of significance.

[2073 Shrawan]

Solution:

1. Null hypothesis: The mean voltages supplied by both companies is equal. Symbolically, H_0: $\mu_1 = \mu_2$.
2. Alternative hypothesis: There is significant difference between mean voltage supplied by two companies.

   Symbolically, H_1: $\mu_1 \neq \mu_2$.

   This is a 2 tailed test of small sample size with unknown 'σ' (population standard deviation). So, we use 't' test.
3. The test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, $S_p^2 = \dfrac{(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2}{n_1 + n_2 - 2}$

$$= \frac{(13-1)\times(0.3)^2 + (10-1)\times(0.4)^2}{13+10-2} = 0.12$$

Thus, $t = \dfrac{(3.59 - 3.15) - 0}{\sqrt{0.12\left(\frac{1}{13}+\frac{1}{10}\right)}} = 3.01974 \approx 3.02$

4. Degree of freedom (d.f.) = $n_1 + n_2 - 2 = 13 + 10 - 2 = 21$

For $\alpha = 0.05$ and d.f. = 21

$t_{\frac{\alpha}{2}, d.f.} = t_{\frac{0.05}{2}, 21} = t_{0.025, 21} = 2.080$

5. Since $t (= 3.02) > t_{0.025, 21} (= 2.080)$, we reject H_0. Thus, the researcher's suspect that the two voltages were significantly different was correct.

15. Shyam and Co. produces three varieties of certain product: deluxe, fine and ordinary. A recent market survey is conducted for preference of products. The preference was found as follow:

Product	Production			
Deluxe	15	14	19	18
Fine	17	12	20	16
Ordinary	16	18	16	17

Is there a significant difference in the preference of products? Test it using ANOVA test. Use $\alpha = 5\%$. [2073 Shrawan]

Solution: See solution to [2072 Magh].

16. The following are the average weekly losses of worker hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation.

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

Use the 0.05 level of significance to test whether the safety program is effective. [2073 Shrawan]

Solution:

Here, the same parameter is tested on same samples before and after the safety program. So, the data are obviously paired. Hence, we use paired 't' test here.

1. Null hypothesis: There is no difference between the average weekly losses before and after the safety program. Symbolically, H_0: $D_i = 0$ (It's actually H_0: $D_i \le 0$). Here, D_i represents the difference between before and after values of the population. Let the sample differences $x_i - y_i$ be represented by d_i.

2. Alternative hypothesis: There is significant difference between average weekly losses before and after the safety program/OR/The difference between before and after values is greater than zero. Symbolically, H_1: $D_i > 0$.
3. This is a right tailed test with unknown population standard deviation (σ) and small sample. The test statistic for paired 't' test is:

$$t = \frac{\bar{d}_i - D_i}{s_d/\sqrt{n}}$$

Here;

Before (x_i)	After (y_i)	$d_i = x_i - y_i$
45	36	9
73	60	13
46	44	2
124	119	5
33	35	-2
57	51	6
83	77	6
34	29	5
26	24	2
17	11	6

The mean and standard deviation of 'd_i' are: (From calculator)

$\bar{d}_i = 5.2$

$s_d = 4.077$

$n = 10$

So, $t = \dfrac{5.2 - 0}{4.077/\sqrt{10}} = 4.033$

4. Degree of freedom (d.f.) = $n - 1 = 10 - 1 = 9$

$t_{\alpha, n-1} = t_{0.05, 9} = 1.83$

5. Since, $t (= 4.033) > t_{0.025, 9}$, we reject H_0. Thus, the difference between before and after values is greater than zero. In other words, the losses in worker hours before the safety program is greater than after the safety program.

Hence, the safety program is effective.

17. Nine objects were choosen at random from a large population and their weights were found to be in grams.

63, 63, 64, 65, 66, 69, 65, 66.1, 64.5

In the light of above data discuss the suggestion that the mean weight in the population is 65 gm. Use $\alpha = 0.05$. [2072 Chaitra]

Solution:

From the data,

$n = 9$; $\bar{X} = 65.067$; $S = 1.8513$

1. Null hypothesis: The mean weight of the population is 65 gram. Symbolically, $H_0: \mu = 65$.
2. Alternative hypothesis: The mean weight of the population is not 65 gram. Symbolically, $H_1: \mu \neq 65$.
3. This is a 2 tailed test and the test statistic is (for a small sample with 'σ' unknown, we use 't' test.

$$t = \frac{\bar{X} - \mu}{S\sqrt{n}} = \frac{65.067 - 65}{1.8513\sqrt{9}} = 0.1086$$

4. d.f. = n – 1 = 9 – 1 = 8. For a = 0.05, and d.f. = 8; $t_{\frac{0.05}{2}, 8}$ is $t_{0.025, 8} = 2.306$.
5. Since, t (= 0.1086) < $t_{0.025, 8}$ (= 2.306), we accept H_0. Thus, the mean weight of the population is indeed 65 gm.

18. Find 99% confidence interval for true hardness of magnesium alloy. 66.3, 63.5, 64.9, 61.9, 64.3, 64.7, 65.1, 65.1, 68.4, 63.2 [2074 Ashwin]

Solution:

From the data

$\bar{X} = 64.38$ $\quad$ n = 10

S = 2.037

Since it's a small sample, we use 't' value

d.f. = n – 1 = 10 – 1 = 9

$t_{\frac{\alpha}{2}, d.f.} = t_{\frac{0.01}{2}, 9} = t_{0.005, 9} = 3.250$

Thus, 99% confidence limits $= \bar{X} \pm t_{\frac{\alpha}{2}, n-1}$ S.E. $(\bar{X})$

$$= \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \times \frac{S}{\sqrt{n}}$$

$$= 64.38 \pm 3.25 \times \frac{2.037}{\sqrt{10}}$$

$$= (62.286, 66.474)$$

19. An Examination was given to 50 students at college A and 60 students at college B. At A, mean grade was 75 with standard deviation of 9. At B, mean grade was 79 with standard deviation of 7. Is these significant difference between the performance of students at A and those at B, given that α = 0.05? [2074 Ashwin]

Solution: See solution to [2071 Magh].

20. three randomly selected groups of chickens are fed on three different diets. Each group consists off give chickens. Their weight gains during a specified period of time are as follows:

Diet I	4	4	7	7	8
Diet II	3	4	5	6	7
Diet III	6	7	7	7	8

Test the hypothesis that mean gains of weight due to the three diets are equal. [2074 Ashwin]

Solution: See solution to [2071 Magh].

21. Six samples of each of four types of cereal grains grown in a certain region were analyzed to determine thiamin content, resulting in the following data (mg/g).

Wheat	5.2	4.5	6.0	6.1	6.7	5.8
Barley	6.5	8.0	6.1	7.5	5.9	5.6
Maize	5.8	4.7	6.4	4.9	6.0	5.2
Oats	8.3	6.1	7.8	7.0	5.5	7.2

Does this data suggest that at least one of the grains differ with respect to true average thiamin content? Use α = 0.05. [2073 Chaitra]

Solution:

1. Null hypothesis: There is no significant difference between average thiamin content of the four grains.
   Symbolically, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
2. Alternative hypothesis: At least one cereal's thiamin content is significantly different from the others.
   Symbolically, $H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$
3. Constructing ANOVA table:
   Here, $n_1 = n_2 = n_3 = n_4 = 6$
   N = 6 × 4 = 24
   K = 4

Wheat (y_1)	$(y_1 - \bar{y}_1)^2$	Barley (y_2)	$(y_2 - \bar{y}_2)^2$	Maize (y_3)	$(y_3 - \bar{y}_3)^2$	Oats (y_4)	$(y_4 - \bar{y}_4)^2$
5.2	0.267	6.5	0.01	5.8	0.09	8.3	1.734
4.5	1.48	8.0	1.96	4.7	0.64	6.1	0.78
6.0	0.08	6.1	0.25	6.4	0.81	7.8	0.667
6.1	0.147	7.5	0.81	4.9	0.36	7.0	0
6.7	0.967	5.9	0.49	6.0	0.25	5.5	2.2
5.8	0.007	5.6	1	5.2	0.09	7.2	0.047
Σ=34.3	Σ=2.948	Σ=39.6	Σ=4.52	Σ=33	Σ=2.24	Σ=41.9	Σ=5.43

$$\bar{y}_1 = \frac{34.3}{6} = 5.7167; \quad \bar{y}_2 = \frac{39.6}{6} = 6.6; \quad \bar{y}_3 = \frac{33}{6} = 5.5; \quad \bar{y}_4 = \frac{41.9}{6} = 6.9833$$

$$\text{Now, } \bar{\bar{y}} = \frac{T}{N} = \frac{34.3 + 39.6 + 33 + 41.9}{24} = 6.2$$

$$SST_r = \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 6(5.7167 - 6.2)^2 + 6(6.6 - 6.2)^2 + 6(5.5 - 6.2)^2 + 6(6.9833 - 6.2)^2$$

$$= 8.983$$

$$SSE = \sum_{i=1}^{4}\sum_{j=1}^{n_j}(\bar{y}_{ij} - \bar{y}_i)^2 = 2.948 + 4.52 + 2.24 + 5.43 = 15.138$$

Variation	d.f.	SS	MS	F-ratio
Treatment	K - 1 = 4 - 1 = 3	$SST_r = 8.983$	$MST_r = \frac{8.983}{3} = 2.994$	$\frac{MSTr}{SSE} = \frac{2.994}{0.76} = 3.94$
Error	N - K = 24 - 4 = 20	SSE = 15.138	$MSE = \frac{15.138}{20} = 0.76$	
Total	23	24.121		

4. At $\alpha = 0.05$ and d.f. = (3, 20), $F_{0.05\,(3,\,20)} = 3.0984$
5. Since F (= 3.94) > $F_{0.05\,(3,\,20)}$ (= 3.094), we reject H_0.
   Thus, at least one of the grains differ with respect to true average thiamin content.

22. A liquid dietary product implies in its advertising that use of the product for one month results in an average weight loss of at least 3 pounds. Eight subjects use the product for one month, and the resulting weight loss data are reported below. Do the data support the claim of the producer of the dietary product with the probability of a type I error set to 0.05? **[2073 Chaitra]**

Subjects	1	2	3	4	5	6	7	8
Weight (*l*b)	165	201	195	198	155	143	150	187
Weight (*l*b)	161	195	192	193	150	141	146	183

Solution:

Since the same subject's weights are taken before and after the use of the dietary product, the data are linked and "Paired t-test" is to be used in this case.

Subjects	Before (*l*b) (x_i)	After (*l*b) (y_i)	$d_i = x_i - y_i$
1	165	161	4
2	201	195	6
3	195	192	3
4	198	193	5
5	155	150	5
6	143	141	2
7	150	146	4
8	187	183	4

Now, we apply the test for parameter 'D_i'.

1. Null hypothesis: The difference/reduction in weight is at least equal to 3 pounds. *i.e.*, H_0: $D_i \geq 3$ *i.e.*, H_0: $D_i = 3$.
2. Alternative hypothesis: The reduction in weight is less than 3 pounds. *i.e.*, H_1: $D_i < 3$
   This is a left tailed test.

3. The test statistic is given as:

$$t = \frac{\bar{d}_i - D_i}{s_d/\sqrt{n}}; \text{ with d.f.} = n - 1 = 8 - 1 = 7$$

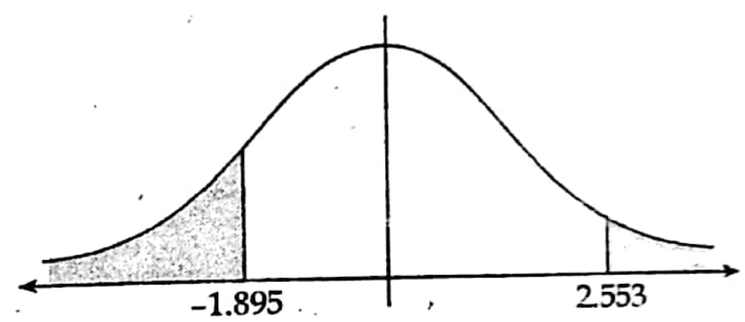
From the above data, using calculator,

$$\bar{d}_i = \frac{\Sigma d_i}{N} = \frac{33}{8} = 4.125$$

$$s_d = 1.2464$$

So, $$t = \frac{4.125 - 3}{\frac{1.2464}{\sqrt{8}}} = 2.553$$

4. For d.f. = 7 and $\alpha = 0.05$. $t_{0.05,\,7} = 1.895$
5.



For a left tailed test, $t_{0.05,\,7}$ is negative *i.e.*, -1.895. But the value t is positive *i.e.*, 2.553. So, it falls in the acceptance region. Hence, H_0 is accepted. Thus, the manufacturer's claim is correct that it results in a weight loss of at least 3 pounds.

23. Three varieties of coal were analyzed by four chemists and the ash-content in the varieties were found as follows:

	Chemists			
Varieties	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Test whether the varieties differ significantly in their ash-content? Test at 5% level of significance.

F(2, 9) = 19.4, F(3, 9) 8.81 **[2075 Ashwin]**

Solution:

1. Null hypothesis: The varieties do not differ significantly in their ash content. Symbolically, $H_0 : \mu_1 = \mu_2 = \mu_3$.
2. Alternative hypothesis: At least one variety differs significantly from the others in it's ash content. Symbolically, H_1: $\mu_1 \neq \mu_2 \neq \mu_3$.

3. The test statistic under ANOVA test is calculated now:

Variety $A(y_1)$	$(y_1 - \bar{y}_1)^2$	Variety $B(y_2)$	$(y_2 - \bar{y}_2)^2$	Variety $C(y_3)$	$(y_3 - \bar{y}_3)^2$
8	3.0625	7	3.0625	3	2.25
5	1.5625	6	0.5625	6	2.25
5	1.5625	4	1.5625	5	0.25
7	0.5625	4	1.5625	4	0.25
$\Sigma = 25$	$\Sigma = 6.75$	$\Sigma = 21$	$\Sigma = 6.75$	$\Sigma = 18$	$\Sigma = 5$

Given,

$n_1 = 4, n_2 = 4, n_3 = 4$

$N = \Sigma n_i = n_1 + n_2 + n_3 = 4 + 4 + 4 = 12$

$k = 3$

$$\bar{y}_1 = \frac{\Sigma y_1}{n_1} = \frac{25}{4} = 6.25; \quad \bar{y}_2 = \frac{\Sigma y_2}{n_2} = \frac{21}{4} = 5.25; \quad \bar{y}_3 = \frac{\Sigma y_3}{n_3} = \frac{18}{4} = 4.5$$

Now,

$$\bar{\bar{y}} = \frac{\Sigma y_1 + \Sigma y_2 + \Sigma y_3}{N} = \frac{25 + 21 + 18}{12} = 5.333$$

$$SSTr = \sum_{i=1}^{3} (\bar{y}_i - \bar{\bar{y}})^2 n_i$$

$$= n_1(\bar{y}_1 - \bar{\bar{y}})^2 + n_2(\bar{y}_2 - \bar{\bar{y}})^2 + n_3(\bar{y}_3 - \bar{\bar{y}})^2$$

$$= 4(6.25 - 5.333)^2 + 4(5.25 - 5.333)^2 + 4(4.5 - 5.333)^2$$

$$= 6.1667$$

$$SSE = \sum_{i=1}^{3} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = 6.75 + 6.75 + 5 = 18.5$$

Constructing ANOVA table

Variation	d.f.	SS	MS	F-ratio
Treatment	$K - 1 = 3 - 1 = 2$	$SSTr = 6.1667$	$MSTr = \frac{6.1667}{2} = 3.083$	$\frac{MSTr}{MSE} = \frac{3.083}{2.055} = 1.5$
Error	$N - K = 12 - 3 = 9$	$SSE = 18.5$	$MSE = \frac{18.5}{9} = 2.055$	
Total	$N - 1 = 11$	$SST = 24.6667$		

4. For d.f. (2, 9) given

$F_{(2, 9)} = 19.4$

5. Since $F_{calc.}$ (= 1.5) < F_{tab} (=19.4), we accept H_0. Thus, there is no significant difference in the ash-content of three varieties.

Chapter 6

INTERFERENCE CONCERNING PROPORTION

6.1 Estimation of Proportion 194
6.2 Hypothesis Concerning One Proportion 195
6.3 Hypothesis Concerning Two Proportions 197
6.3.1 For Difference of Two Proportions 197
6.3.2 For Equality of Two Proportions 198
6.4 Chi Square Test of Independence 199

6.1 ESTIMATION OF PROPORTION

Let $\hat{p} = \frac{x}{n}$, be the sample proportion of success and $\hat{q} = 1 - \hat{p}$ be sample proportion of failure.

From central limit theorem, we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

It follows approximate standard normal distribution.

Hence, confidence interval can be expressed in the probability form as

$$P(C_1 < \theta < C_2) = 1 - \alpha$$

or, $P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$

or, $P\left(-Z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} < Z_{\alpha/2}\right) = 1 - \alpha$

Multiplying by $-\sqrt{\frac{pq}{n}}$ on all sides, we get

$$P\left(Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} > -\hat{p} + p > -Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

Adding $\hat{p}$ on all sides

$$P\left(\hat{p} + Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} > p > \hat{p} - Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

or, $P\left(\hat{p} - Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} < p < \hat{p} + Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}\right) = 1 - \alpha$

Hence, the confidence interval for population proportion 'p' is

$$\left(\hat{p} - Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}, \hat{p} + Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}\right)$$

In summary

$$P = \hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} \quad \text{[If population is large]}$$

$$P = \hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}} \quad \text{[If population is finite]}$$

Note:

a) Upper confidence interval $= \left(\hat{p} - Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}, \infty\right)$

b) Lower confidence interval $= \left(-\infty, \hat{p} + Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}\right)$

Example 1

Find the 90% confidence interval for the percentage defection in a large lot of bulbs, when 18 bulbs were found to be defective in a lot of sample of 50 bulbs.

Solution:

Here;

Sample size (n) = 50 (> 30, large sample) population size is very large.

$(1 - \alpha) \times 100\% = 90\%$

or, $\alpha = 10\% = 0.1$

$\therefore \frac{\alpha}{2} = 0.05$

Sample proportion of defective bulbs $(\hat{p}) = \frac{18}{50} = 0.36$

Sample proportion of good bulbs $(\hat{q}) = 1 - \hat{p}$ 0.64

Using formula

$$P = \hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}}$$

$$= 0.36 \pm Z_{0.05} \sqrt{\frac{0.36 \times 0.64}{50}} \quad \text{[for large sample } \hat{p} \simeq p, \hat{q} \simeq q\text{]}$$

$$= 0.36 \pm 1.645 \times \sqrt{\frac{0.36 \times 0.64}{50}} \quad \text{[form Z-table, } Z_{0.0505} = -1.64$$

$$= 0.36 \pm 0.111 \quad \text{and } Z_{0.0495} = -1.65 \therefore Z_{0.05} = \frac{Z_{0.0505} + Z_{0.0495}}{2}\text{]}$$

$$= 0.249, 0.471$$

$\therefore$ 90% confidence interval = (0.249, 0.471)

Note: $P_1 - P_2 = (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

6.2 HYPOTHESIS CONCERNING ONE PROPORTION

Let n be the sample size out of which x instances are associated with an event. Let the claimed proportion be P_0.

Step 1: State the type of Hypothesis

Null Hypothesis $H_0 : p = p_0$

i.e., there is no significant difference between population proportion and claimed value of proportion.

Alternative Hypothesis (H_1)

$H_1 : p > p_0$ (right tailed test)

or, $H_1 : p < p_0$ (left tailed test)

or, $H_1 : p \neq p_0$ (two tailed test)

i.e., There is significant difference between population and sample mean.

Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true.

Assume appropriate value of significance level (α). Let α = 5%.

Step 3: Test Statistics

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \qquad [p = p_0, \hat{p} = \frac{x}{n}, \hat{q} = 1 - \hat{p}]$$

$$= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \text{ or } \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}}$$

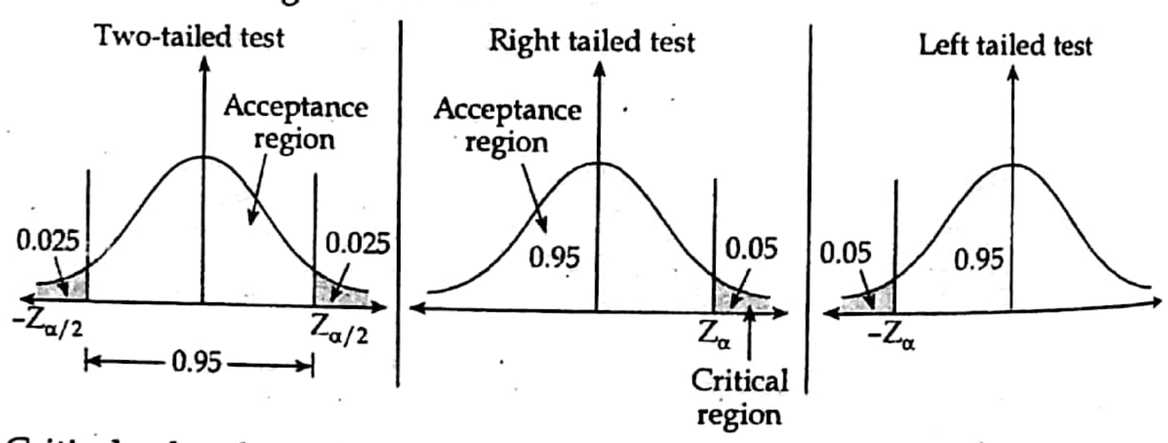
If sampling is done from finite population of size N, then

$$Z = \frac{\hat{p} - p}{\sqrt{\left(\frac{N-n}{n-1}\right)\frac{pq}{n}}}$$

Step 4 Formulate Design Rule based on Step 1, 2, 3

$$\left(\text{Let } \alpha = 5\% = 0.05, \frac{\alpha}{2} = 0.025\right)$$

Two-tailed test right tailed test



Critical value depends on:

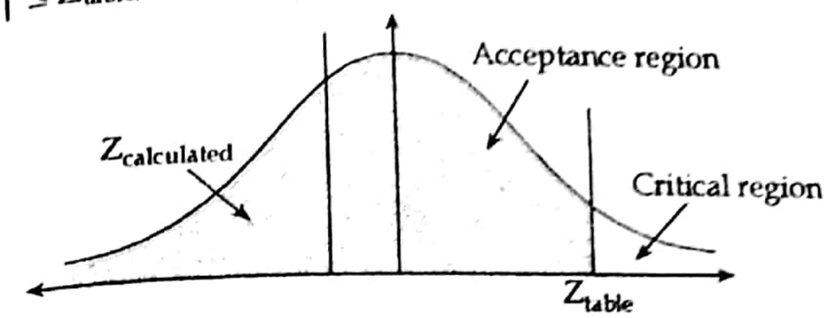
a) Significance level (α)
b) Type of hypothesis
c) Sample size
d) Probability distribution

Step 5: Make a decision about null hypothesis using testing statistics value and critical values

a) Using formula from step 3, we calculate value of test statistics = $Z_{calculated}$
b) From table, we find critical values of $Z = Z_{table}$

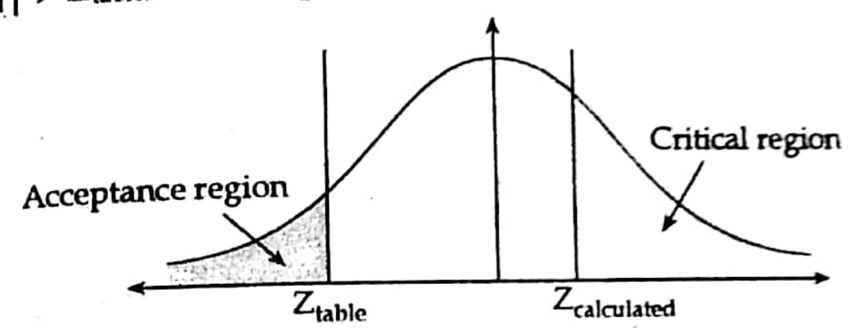
Case I

If $|Z_{calculated}| \le Z_{table}$, we accept H_0



Case II

If $|Z_{calculated}| > Z_{table}$, we accept H_1 and reject H_0.



Alternative Hypothesis (H_1)	Rejection Region
$p < p_0$	$Z < -Z_\alpha$
$p > p_0$	$Z > Z_\alpha$
$p \ne p_0$	$\lvert Z \rvert > Z_{\alpha/2}$

6.3 HYPOTHESIS CONCERNING TWO PROPORTIONS

6.3.1 FOR DIFFERENCE OF TWO PROPORTIONS

Let there be two treatments with population proportion p_1 and p_2 and sample size of n_1 and n_2 respectively. Samples associated with a certain event are x_1 and x_2 with the respective proportion.

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2}$$

Step 1: State the type of Hypothesis

Null Hypothesis $H_0 : p_1 - p_2 = \delta_0$

i.e., there is no significant difference between population proportions.

Alternative Hypothesis (H_1)

$H_1 : p_1 - p_2 > \delta_0$ (right tailed test)

or, $H_1 : p_1 - p_2 < \delta_0$ (left tailed test)

or, $H_1 : p_1 - p_2 \ne \delta_0$ (two tailed test)

Step 2: Select Significance Level 'α'

It is the maximum value of probability of rejecting null hypothesis when it is true.

Assume suitable value of α. Let α = 5%.

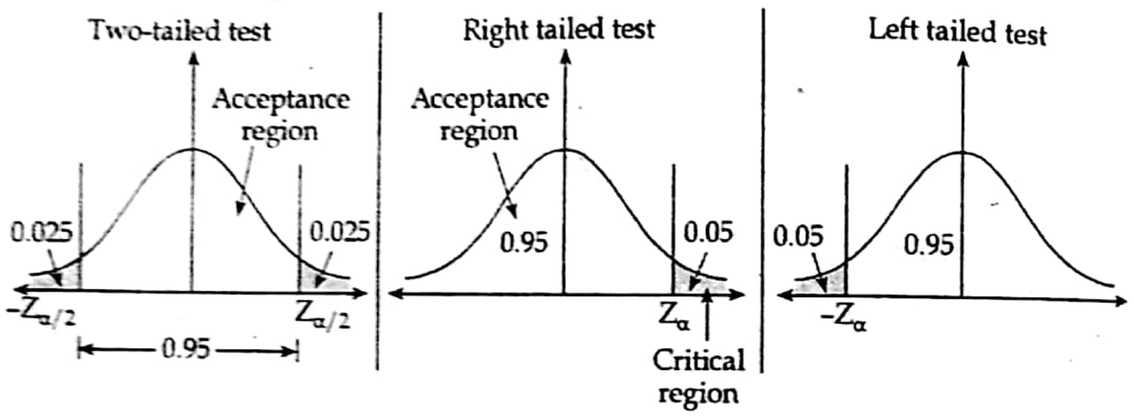
Step 3: Test Statistics

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \quad \text{where, } \delta_0 = p_1 - p_2$$

Step 4 Formulate Design Rule based on Step 1, 2, 3

$$\left(\text{Let } \alpha = 5\% = 0.05, \frac{\alpha}{2} = 0.025\right)$$

Two-tailed test right tailed test



Critical value depends on:

a) Significance level (α)
b) Type of hypothesis
c) Sample size
d) Probability distribution

Step 5: Make a decision

Alternative Hypothesis (H_1)	Rejection Region
$p_1 - p_2 < \delta_0$	$Z < -Z_\alpha$
$p_1 - p_2 > \delta_0$	$Z > Z_\alpha$
$p_1 - p_2 \neq \delta_0$	$\lvert Z \rvert > Z_{\alpha/2}$

6.3.2 FOR EQUALITY OF TWO PROPORTIONS

In all steps of 6.3.1 replace $\delta_0 = 0$ and in

Step 3: Test Statistics

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{where, } \hat{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \qquad \hat{Q} = 1 - \hat{p}$$

All steps are same.

6.4 CHI SQUARE TEST OF INDEPENDENCE

The Chi square distribution denoted as χ^2 is related to the standard normal distribution such as, if the independent normal variable let's say Z assumes the standard normal distribution, then the square of this normal variable Z^2 has the Chi square distribution with 'ν' degree of freedom. Here, ν is the sum of the independent squared normal variables.

Mathematically,

If $Z = \left(\frac{\bar{x} - \mu}{\sigma}\right) \sim N(0, 1)$ then, x defined by

$$X = \left(\frac{\bar{x}_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{\bar{x}_2 - \mu_2}{\sigma_2}\right)^2 + \dots\dots\dots + \left(\frac{\bar{x}_n - \mu_n}{\sigma_n}\right)^2$$

$$= \sum_{i=1}^{n}\left(\frac{\bar{x}_i - \mu_i}{\sigma_i}\right)^2$$

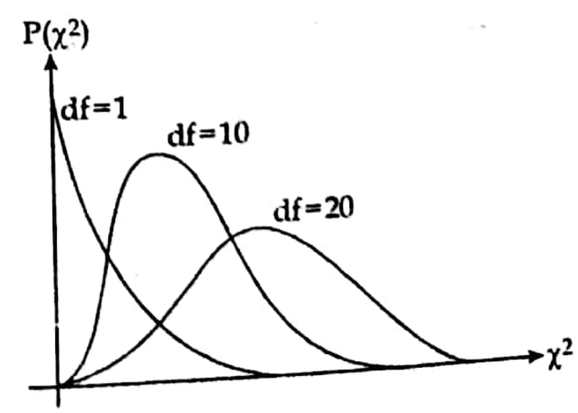
Follows χ^2 distribution with ν degree of freedom.

It's probability density function is given by

$$F(x, \nu) = \begin{cases} \left(\dfrac{1}{2^{\frac{\nu}{2}} \cdot \Gamma\left(\frac{\nu}{2}\right)}\right) x^{\frac{\nu-2}{2}} \cdot e^{-\frac{x}{2}} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Properties of Chi square distribution

1. The chi-square distribution is a continuous probability distribution with the values ranging from 0 to ∞ (infinity) in the positive direction. The χ^2 can never assume negative value.
2. The shape of chi square distribution depends on the number of degree of freedom (ν) when ν is small, the shape of curve tend to be skewed to right and as 'ν' gets larger the shape becomes more symmetrical and can be approximated by the normal distribution.



3. For χ^2 distribution
   Mean = ν
   Variance = 2ν
   Mode = $\nu - 2$
4. The total area under each curve of χ^2 distribution is 1 or 100%.
5. The sum of independent χ^2 is itself a χ^2 variate. Suppose X_1^2 is a χ^2 variate with degree of freedom ν_1 and X_2^2 is another χ^2 variate with degree of freedom ν_2, then their sum $(X_1^2 + X_2^2)$ will be a χ^2 variate with $n_1 + n_2$ degree of freedom. This property is called additive property of χ^2.

Steps for testing X^2 test

Step 1: State the type of Hypothesis

a) **Null Hypothesis (H_0):** There is no significant difference between observed values and expected values (*i.e.*, experimental and theoretical values).

b) **Alternative Hypothesis (H_1):** There is no significant difference between observed values and expected values.

Step 2: Level of Significance

Let $\alpha = 5\%$ (If not provided)

Step 3: Test statistics

$$X^2 = \frac{\sum_{i=1}^{K} (o_i - e_i)^2}{e_i}$$

where, o_i = i^{th} observed frequency
e_i = i^{th} expected frequency

Degree of freedom d.f.: $\nu = n - 1 - K_1 - K_2$

where, K_1 = Number of unknown parameter of specified distribution. *For example;* in Poisson and binomial distribution, there is only one parameter (λ and p respectively) which should be specified. Hence $K_1 = 0$ for both case. But in normal distribution there are two parameter (μ, σ). If only one is given then $K_1 = 1$ for one unknown parameter. If both are given $K_1 = 0$.

K_2 = It is the number of expected frequencies less than 5.

Step 4: Critical Values

For $\alpha = 5\%$, find the table value of χ^2.

Step 5: Decision Making

$(\chi^2)_{calculated} < (\chi^2_{\alpha\nu})_{table}$

We accept the null hypothesis.

SOLVED QUESTIONS [FOR BCE/BME/BIE/BAM]

1. **One thousand articles from factory A are examined and found 97% of good quality. Fifteen hundred similar articles from factory B are found only 2% of bad quality. Would you conclude that the products of factory B are superior to those of factory A? Use $\alpha = 0.05$.** **[5][2070 Magh]**

Solution:

Given that;

$n_1 = 1000$, $\hat{p}_1 = 0.97$

$n_2 = 1500$, $\hat{p}_2 = (1 - 0.02) = 0.98$

1. Null hypothesis: The projects of both factories are equally of good quality. Symbolically, $P_1 = P_2$.
2. Alternative hypothesis: The products of factory A are inferior to that of factory B. Symbolically, H_1: $P_1 < P_2$. This is a left tailed test.
3. Now, the test statistic is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{Here, } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{1000 \times 0.97 + 1500 \times 0.98}{1000 + 1500} = 0.976$$

$$\hat{q} = (1 - \hat{p}) = (1 - 0.976) = 0.024$$

$$\text{Now, } Z = \frac{(0.97 - 0.98) - 0}{\sqrt{(0.976 \times 0.024) \times \left(\frac{1}{1000} + \frac{1}{1500}\right)}} = -1.60$$

4. For $\alpha = 0.05$, $Z_{0.05} = -1.645$
5. Now, $|Z| = (1.60) < |Z_{0.05}|$ (= 1.645). So, we accept H_0 and reject H_1: Thus, the products of factory B is superior of products of factory A.

2. **Discuss the application of χ^2-test of significance.** **[5][2070 Magh]**

Solution:

The chi-square (χ^2) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. The purpose of this test is to evaluate how likely the observations that are made would be true assuming null hypothesis.

(See theory section 6.4 and explain it with example)

Other applications are:

i) In cryptanalysis, chi-square test is used to compare the distribution of plain text and decrypted test. The lower value of test means that the decryption was successful with high probability.

ii) In bio-informatics, chi-square test is used to compare the distribution of certain properties of genes. (*For example;* mutation rate, genomic content, etc) belong to different categories (*For example;* disease genes, genes on certain chromosome, etc.)

3. **A manufacturer claimed that at least 95% of the pumps supplied to the ABC company confirmed to specifications. However, the production manager at ABC company wasn't satisfied with the claim of the manufacturer. Hence, to test the claim, the manager examined a sample of 250 pumps supplied last month and found that 228 pumps as per specifications. Can you conclude that that production manager is right to doubt on the claim of the manufacturer? [5][2070 Bhadra]**

Solution: See the solution of [2072 Magh].

4. **A sample of 500 workers of a factory according to gender and nature of work is as follows:**

Nature of work	Gender	
	Male	Female
Technical	200	100
Non-technical	50	150

Test at 5% level of significance whether there exists any relationship between gender and nature of work. [5][2070 Bhadra]

Solution:

1. Null hypothesis: There exists no relationship between gender and nature of work.
2. Alternative hypothesis. there exists significant relationship between gender and nature of work.
3. Here,

r = No. of rows = 3; c = No. of column = 2

∴ Degree of freedom = $\nu = (r-1)(c-1) = (2-1)(2-1) = 1$

			Row Total (RT)
	200	100	300
	50	150	200
Column Total (CT)	250	250	500

∴ Grand total = N = 500

We know,

$$E = \frac{RT \times CT}{N}$$

Thus, $E(200) = \frac{300 \times 250}{500} = 150$ $\quad E(100) = \frac{300 \times 250}{500} = 150$

$E(50) = \frac{200 \times 250}{500} = 100$ $\quad E(150) = \frac{200 \times 250}{500} = 100$

So,

O	E	O – E	$(O-E)^2/E$
200	150	50	16.67
100	150	-50	16.67
50	100	-50	25
150	100	50	25

$$\therefore \Sigma\left[\frac{(O-E)^2}{E}\right] = 83.34$$

4. For $\alpha = 0.05$ and d.f. = 1, $\chi^2_{0.05,1} = 3.841$
5. Since, $\chi^2 (= 83.34) > \chi^2_{0.05,1} (= 3.841)$, we reject H_0.

Thus, there exists significant relationship between gender and nature of work.

5. **In random sample of 400 industrial accidents, it was found that 231 were due to at least partially to unsafe working conditions. Construct 99% and 95% confidence intervals for corresponding true proportions. [5] [2071 Magh]**

Solution:

$$n = 400,\ X = 231,\ \hat{p} = \frac{231}{400} = 0.5775$$

Let's assume the true proportion is similar to the sample properties.

i.e., $P \approx \hat{p} = 05775$

Now, $Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

$Z_{\frac{0.01}{2}} = Z_{0.005} = 2.575$

Thus, 99% confidence interval

$$= \hat{p} \pm Z_{0.005} \times \sqrt{\frac{pq}{n}}$$

$$= 0.5775 \pm 2.575 \times \sqrt{\frac{0.5775 \times (1-0.5775)}{400}}$$

$$= (0.5139, 0.6411)$$

Again, for 95% confidence interval

$$= \hat{p} \pm Z_{0.025} \times \sqrt{\frac{pq}{n}}$$

$$= 0.5775 \pm 1.96 \times \sqrt{\frac{0.5775 \times (1-0.5775)}{400}}$$

$$= (0.5291, 0.6259)$$

6. **In 40 tosses of a coin, 24 heads were obtained. Find 95% and 99% confidence limit for proportion of heads. [5][2071 Bhadra]**

Solution:

We know, for coin toss: $p = 0.5, q = 0.5$

$$n = 40, X = 24, \hat{p} = \frac{24}{40} = 0.6$$

$$Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$$

$$Z_{\frac{0.01}{2}} = Z_{0.005} = 2.575$$

Now,

$\therefore$ 95% confidence limits $= \hat{p} \pm Z_{\frac{\alpha}{2}} S.E(\hat{p})$

$$= \hat{p} \pm Z_{0.025} \times \sqrt{\frac{pq}{n}}$$

$$= 0.06 \pm 1.96 \times \sqrt{\frac{0.5 \times 0.5}{40}}$$

$$= (0.445, 0.755)$$

99% confidence limit $= \hat{p} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{pq}{n}}$

$$= 0.6 + Z_{0.005} \times \sqrt{\frac{0.5 \times 0.5}{40}}$$

$$= 0.6 \pm 2.575 \times \sqrt{\frac{0.5 \times 0.5}{40}} = (0.3964, 0.8036)$$

7. **A manufacturer claimed at least 95% of the pumps supplied to the ABC company confirmed to specifications. however, the production manager at ABC company wasn't satisfied with the claim of the manufacturer. Hence, to test the claim, the manager examined a sample of 250 pumps. Supplied last month and found that 228 pumps as per the specification. Can you conclude that the production manager is right to doubt on the claim of the manufacturer?** **[2072 Magh]**

Solution:

1. Null hypothesis: The percentage of pups as per specifications is equal to 95%. Symbolically, H_0: $P = 0.95$.
2. Alternative hypothesis: The percentage of pumps as per specifications is less than 95%, Symbolically, H_1: $P < 0.95$.
3. We know that (acc^n to question)

$$n = 250, X = 228, \hat{p} = \frac{X}{n} = \frac{228}{250} = 0.912$$

Now, the test statistic is given as:

$$Z = \frac{\hat{p} - P}{\sqrt{\frac{pq}{n}}} = \frac{0.912 - 0.95}{\sqrt{\frac{0.95 \times (1 - 0.95)}{250}}} = -2.757$$

4. At $\alpha = 5\%$ (say), $Z_{0.05} = -1.65$
5. Since $|Z|$ (= 2.757) > $|Z0.05|$ (= 1.65), we reject H_0. Thus, the percentage of pumps as per specifications is less than 95%. The production manager is right to do not on the claim of the manufacturer.

8. **A study shows that 16 of 200 tractors produced at one assembly line required extensive adjustment before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line. At the 0.05 level of significance, does this support the claim that the second production line does superior work?** **[2072 Ashwin]**

Solution: Given,

$$n_1 = 200, \quad X_1 = 16, \quad \hat{p}_1 = \frac{16}{200} = 0.08, \quad \hat{q}_1 = 1 - 0.08 = 0.92$$

$$n_2 = 400, \quad X_1 = 14, \quad \hat{p}_2 = \frac{14}{400} = 0.035, \quad \hat{q}_2 = 1 - 0.035 = 0.965$$

1. Null hypothesis: The number of defective units produced by both assembly lines are same.
   Symbolically, H_0: $P_1 = P_2$
2. Alternative hypothesis: The number of defective units produced by first assembly line is higher than the second.
   Symbolically, H_1: $P_1 > P_2$
   Clearly, this is a right tailed test.
3. Now, we calculate the test statistic 'Z'. We know,

$$Z = \frac{(\hat{p}_1 - \hat{p}_1) - (P_1 - P_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, $\hat{p}$ and $\hat{q}$ are pooled values calculated as

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{200 \times 0.8 + 400 \times 0.035}{200 + 400} = 0.05$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.05 = 0.95$$

So, $$Z = \frac{(0.08 - 0.035) - 0}{\sqrt{(0.05 \times 0.95) \times \left(\frac{1}{200} + \frac{1}{400}\right)}} = 2.3841$$

4. At $\alpha = 5\% = 0.05$, $Z_{0.05} = 1.645$
5. Since $|Z|$ (= 2.3841) > $Z_{0.05}$ (= 1.645), it falls in the rejection region. Thus, we reject H_0. Hence, the number of defective units produced by first assembly line is higher than the second. Thus, the claim is true.

9. **To determine whether there is really a relationship between an employer's performance in the training program and his ultimate success in the job, it takes a sample of 400 cases from its very extensive files and obtains the result shown in the following table.**

		Performance in training program		
		Below average	Average	Above average
Success in job	Poor	23	60	29
	Average	28	79	60
	Very good	9	49	63

Use the 0.05 level if significance to test whether performance in training program and success in the job are independent.

[2072 Ashwin]

Solution:

This is χ^2 test for association.

1. Null hypothesis: There is no dependence between performance in training program and success in job.
2. Alternative hypothesis: There is significant dependence between performance in training program and success in job.
3. Here,

r = No. of freedom (d.f.) = v = (r – 1) (c – 1) = (3 – 1) (3 – 1) = 4

				Row total (RT)
	23	60	29	112
	28	79	60	167
	9	49	63	121
Column Total (CT)	60	188	152	400 (Grand total)

N = Grand total = 400

$$E = \frac{RT \times CT}{N}$$

$$E(23) = \frac{112 \times 60}{400} = 16.8 \qquad E(60) = \frac{112 \times 188}{400} = 52.64$$

$$E(29) = \frac{112 \times 152}{400} = 42.56 \qquad E(28) = \frac{167 \times 60}{400} = 25.05$$

$$E(79) = \frac{167 \times 188}{400} = 78.49 \qquad E(60) = \frac{167 \times 152}{400} = 63.46$$

$$E(9) = \frac{121 \times 60}{400} = 18.15 \qquad E(49) = \frac{121 \times 188}{400} = 56.87$$

$$E(63) = \frac{121 \times 152}{400} = 45.98$$

O	E	O – E	$(O - E)^2/E$
23	16.8	6.2	2.288
28	25.05	2.95	0.3474
9	18.15	–9.15	4.613
60	52.64	7.36	1.03
79	78.49	0.51	0.003
49	56.87	–7.87	1.089
29	42.56	–13.56	4.32
60	63.46	–3.46	0.1886
63	45.98	17.02	6.3o

$$\chi^2 = \Sigma\left[\frac{(O - E)^2}{E}\right] = 20.179$$

4. For $\alpha = 0.05$ and d.f. = 4, $\chi^2_{0.05, 4} = 9.488$
5. Since χ^2 (= 20.179) > $\chi^2_{0.05, 4}$ (= 9.488), we reject H_0. Thus, there is significant association between performance in training program and success in job.

10. **The results of a survey regarding "regarding listeners" preference for different types of music are given in the following table, is there any association between listeners classified by age group and preference of type of music influenced by age?**

Types of music	Age group		
	19-25	26-35	Above 36
Folk music	80	60	9
Modern music	210	325	44
Indifferent	16	45	32

[2073 Bhadra]

Solution:

Since it is asked to test whether the parameter "type of music" and "age group" are dependent, this is a "χ^2 test for independence of attributes χ^2 test for association".

1. Null hypothesis: The parameters "types of music" and "age group" are independent to each other. *i.e.*, there is no association.
2. Alternative hypothesis: The parameters "types of music" and "age group" are dependent. *i.e.*, there is significant association between the two.
3. Doing χ^2 calculation to calculate χ^2 value.

Here, No. of rows (r) 3, No. of columns (c) = 3

So, degrees of freedom (d.f.) = (r – 1) (c – 1) = (3 – 1) (3 – 1) = 2 × 2 = 4

				Row total (RT)
	80	60	9	149
	210	325	44	579
	16	45	32	93
Column Total (CT)	306	430	85	Grand total = 149 + 579 + 93 = 821

We know,

$$E = \frac{RT \times CT}{N} \text{ Here,}$$

N = Total radio listeners surveyed = Grand total = 821

So, $E(80) = \frac{149 \times 306}{821} = 55.53$; $E(60) = \frac{149 \times 430}{821} = 78.04$

$E(9) = \frac{149 \times 85}{821} = 15.43$; $E(210) = \frac{579 \times 306}{821} = 215.80$

$E(325) = \frac{579 \times 430}{821} = 303.25$; $E(44) = \frac{579 \times 85}{821} = 59.95$

$E(16) = \frac{93 \times 306}{821} = 34.66$; $E(45) = \frac{93 \times 430}{821} = 48.71$

$E(32) = \frac{93 \times 85}{821} = 9.63$

Now,

O	E	O – E	$(O - E)^2$	$(O - E)^2/E$
80	55.53	24.47	598.78	10.783
60	78.04	–18.04	325.44	4.17
9	15.43	–6.43	41.345	2.68
210	215.8	–5.8	33.64	0.156
325	303.25	21.75	473.0625	1.56
44	59.95	–15.95	254.4025	4.24
16	34.66	–18.66	348.1956	10.05
45	48.71	–3.71	13.7641	0.2826
32	9.63	22.37	500.4169	51.964
				85.8856

4. From table, for $\alpha = 5\%$ and d.f. = 4, we have,
$\chi^2_{0.05,4} = 9.488$

5. Since χ^2 (= 85.8856) > $\chi^2_{0.05,4}$ (= 9.488), it lies in rejection region. Thus, we accept alternative hypothesis. Hence, there is significant association between the two parameters.

11. From a log of units produced by machine A, a sample of 500 is drawn and tested for a quality characteristic. It's found that 16 units are not meeting the signification. Another sample of size 100 is drawn from the lot of similar units produced by machine B and tested. In this case, only 3 units are found to be not meeting the specification. Obtain a 99% confidence intervals for the difference of the proportions of defective units produced by two machines.

[2073 Bhadra]

Solution:

Here; $n_1 = 500$, $X_1 = 16$, $\hat{p}_1 = \frac{16}{500} = 0.032$, $\hat{q}_1 = 1 - 0.032 = 0.968$

$n_2 = 100$, $X_2 = 3$, $\hat{p}_2 = \frac{3}{100} = 0.03$, $\hat{q}_2 = 1 - 0.03 = 0.97$

Now, $\alpha = 100 - 99 = 1\% = 0.01$

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005} = 2.575$

Now, Confidence interval

$$= (\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$= (0.032 - 0.03) \pm 2.575 \sqrt{\frac{0.032 \times 0.968}{500} + \frac{0.03 \times 0.97}{100}}$$

$$= 0.002 \pm 2.575 \times 0.018787 = (-0.0464, 0.0504)$$

∴ The required confidence interval is
(–0.0464, 0.0504)

12. A real estate agent claims that 60% of all private residences being built today are 3 bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic state the null and alternative hypothesis to be used in this test and determine the location of the critical region.

[2073 Bhadra]

Solution:

Null hypothesis: The percentage of new homes with 3-bedroom is 60%.

Symbolically, H_1: P = 0.6

Alternative hypothesis: The percentage of new homes with 3-bedrooms is not 60%. Symbolically, H_1: $P \neq 0.6$

This is a two tailed test

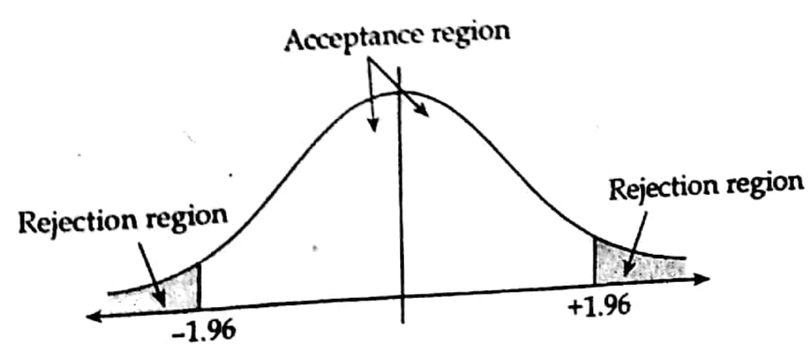
Assume $\alpha = 5\% = 0.05$

Critical point

$= Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}}$

$= Z_{0.025} = \pm 1.96$

Thus, any value exceeding +1.96 or less than –1.96 falls in critical region (rejection region).



13. A manufacturer claimed that at least 95% of the plumps supplied to the ABC Company confirmed to specifications. However, the production manager at ABC Company wasn't satisfied with the claim of the manufacturer. Hence, to test the claim, the manager examined a sample of 250 pumps supplied last month and found that 228 pumps as per specifications Can you conclude that the production manager is right to doubt on the claim of the manufacturer? [$\alpha = 0.01$] [2074 Bhadra]

Solution: See previous question.

14. From the following data can you conclude that there is association between the purchase of brand and geographical region. [Use $\alpha = 5\%$] [2074 Bhadra]

Solution: The table is missing in this question.

15. A manufacturer claimed that at least 95% of the pumps supplied to ABC company confirmed to specifications. However the production manager at ABC company wasn't satisfied with the claim of the manufacturer. Hence to test the claim, the manager examined a sample of 250 pumps supplied last month and found that 228 pumps as per specifications. Can you conclude that the production manager is right to doubt on the claim of the manufacturer. [Use $\alpha = 0.01$] [2075 Baisakh]

Solution: See previous question.

16. A sample of 500 workers of a factory according to gender and nature of work is as follow.

	Gender	
Nature of work	male	female
technical	200	100
non- technical	50	150

Test at 5% level of significance whether there exists any relationship between gender and nature of work. [2075 Baisakh]

Solution: See previous question.

17. From the following data, can you conclude that there is association between purchase of a brand and geographical region? (Use 5% level of significance). [Bhadra 2075]

	Region		
	Central	Eastern	Western
Purchase brand	40	55	45
Do not purchase brand	60	45	55

Solution: See previous question.

SOLVED QUESTIONS (FOR BCT/BEX/BEL/B. Agri)

1. The following test gives the information for the engineering students interest with ability in computer. Is there any significant relationship between interest in engineering and ability in computer? [2070 Ashad]

Ability in computer	Interest	Low	Average	High
	Low	6	12	32
	Average	33	61	18

Solution:

1. Null hypothesis: There is no significant relationship between interest in engineering and ability in computer.
2. Alternative hypothesis: There is significant relationship between interest in engineering and ability in computer.
3. The test statistic is

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

where, $E = \dfrac{RT \times CT}{N}$

Now,

				Row Total (RT)
	6	12	32	50
	33	61	18	112
Column Total (CT)	39	73	50	Grand Total (GT) = N = 162

$E(6) = \dfrac{50 \times 39}{162} = 12.04;$ $\quad E(12) = \dfrac{50 \times 39}{162} = 22.53$

$E(6) = \dfrac{50 \times 50}{162} = 15.432;$ $\quad E(33) = \dfrac{112 \times 39}{162} = 26.963$

$E(61) = \dfrac{112 \times 73}{162} = 50.47;$ $\quad E(18) = \dfrac{112 \times 50}{162} = 34.568$

Now,

O	E	$(O-E)^2$	$(O-E)^2/E$
6	12.04	36.4816	3.03
32	15.432	274.498	17.79
61	50.47	110.881	2.197
12	22.53	110.881	4.92
33	26.963	36.44	1.35
18	374.498	247.498	7.94

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$= 3.03 + 17.79 + 2.197 + 4.92 + 1.35 + 7.94$

$= 37.227$

4. For $\alpha = 0.05$ and d.f. $= (r-1)(c-1) = (2-1)(3-1) = 2$

   $\chi^2_{0.05,2} = 5.991$

5. Since, χ^2 (= 37.227) < $\chi^2_{0.05,2}$ (= 5.991), we reject H_0.

   Thus, there is significant relationship between interest in engineering and ability in computer.

2. **Two different types of injection molding machines are used to form plastic parts. A part is considered defective if it has excessive shrinkage or is discoloured. Two random samples, each of size 300 m are selected and 15 defective parts are found in the sample from machine 1 while 8 defective parts are found in the sample from machine 2. Is it reasonable to conclude that both machines produce the same fraction of defective parts, using $\alpha = 0.05$? [2071 Ashad]**

Solution:

Given that;

$$n_1 = 300, \quad X_1 = 15, \quad \hat{p}_1 = \frac{15}{300} = 0.05$$

$$n_2 = 300, \quad X_2 = 8, \quad \hat{p}_2 = \frac{8}{300} = 0.02667$$

1. Null hypothesis: The two machines produce the same fraction of defective parts. Symbolically, H_0: $p_1 = p_2$.
2. Alternative hypothesis: The two machines produce different fractions of defective parts. Symbolically, H_1: $p_1 \neq p_2$ (2 tailed test).
3. The test statistic for large sample test of two proportion is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)(p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, $\hat{p} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{15 + 8}{300 + 300} = 0.03833$

So, $Z = \frac{(0.05 - 0.02667) - 0}{\sqrt{0.03833\,(1 - 0.03833) \times \left(\frac{1}{300} + \frac{1}{300}\right)}} = 1.48827$

4. For $\alpha = 0.05$, since it's a 2 tailed test;

   $Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

5. Since, Z(= 1.48827) < $Z_{0.025}$ (= 1.96), we accept H_0. Thus, we can conclude that the two machines produce the same fraction of defective parts.

3. **The results of polls conducted two weeks and four weeks before an election are shown in the following table**

	Two weeks before election	Four weeks before election
For republican	79	91
For democrat	84	66
Undecided	37	43

Use 0.05 level of significance to test whether there has been change in opinion during the two weeks between the polls. [2070 Chaitra]

Solution:

The fact whether or not there has been change in opinion can be checked by finding out if there is an association between time and peoples decision. So, we apply χ^2 test for association.

1. Null hypothesis: There is no association between time and people's decision regarding who to vote.
2. Alternative hypothesis: There is significant association between time and people's decision regarding who to vote.
3. The test statistic is:

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

where, $E = \frac{RT \times CT}{N}$

			Row Total (RT)
	79	91	170
	84	66	150
	37	43	80
Column Total (CT)	200	200	Grand Total (GT) = N = 400

$E(79) = \frac{170 \times 200}{400} = 85;$ $\quad E(91) = \frac{170 \times 200}{400} = 85$

$E(84) = \frac{150 \times 200}{400} = 75;$ $\quad E(91) = \frac{150 \times 200}{400} = 75$

$E(37) = \frac{80 \times 200}{400} = 40;$ $\quad E(43) = \frac{80 \times 200}{400} = 40$

Also, r = No. of rows = 3

c = No. of columns = 2

Degree of freedom (d.f.) = (r – 1) (c – 1) = (3 – 1) (2 – 1) = 2

Now,

O	E	$(O - E)^2$	$(O - E)^2/E$
79	85	36	0.423
91	85	36	0.423
84	75	81	1.08
66	75	81	1.08
37	40	9	0.225
43	40	9	0.225

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$= 0.423 + 0.423 + 1.08 + 1.08 + 0.225 + 0.225$$

$$= 3.456$$

4. For $\alpha = 0.05$ and d.f. = 2, $\chi^2_{0.05,2} = 5.991$
5. Since, χ^2 (= 3.456) < $\chi^2_{0.05,2}$ (= 5.991), we accept H_0.

   Thus, there is no association between time and people's decision regarding who to vote. Thus, there hasn't been change in opinion between the polls.

4. **A manufacturer of submersible pumps claims that at most 30% of the pumps require within the first the first 5 years of operation. If a random sample of 120 of these pumps includes 47 which required repairs within the first five years, test the null hypothesis p = 0.30 against the alternative hypothesis p > 0.30 at 0.05 level of significance. [2070 Chaitra]**

Solution:

1. Null hypothesis: The proportion of pumps requiring repairs within the first five years is at most 0.30. Symbolically, H_0: $p \le 0.30$ (OR) H_0: $p = 0.30$.
2. Alternative hypothesis: The proportion of pumps requiring repairs within the first five years is greater than 0.30. Symbolically, H_1: $p > 0.30$. This is a right tailed test:
3. Given,

$$n = 120, \quad X = 47, \quad \hat{p} = \frac{X}{n} = \frac{47}{120} = 0.39167$$

The test statistic for large sample proportion is

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.39167 - 0.30}{\sqrt{\frac{0.30\,(1 - 0.30)}{120}}} = 2.19133$$

4. For $\alpha = 0.05$, since it's a one tailed test;

   $Z_\alpha = Z_{0.05} = 1.645$
5. Since, Z(= 2.19133) > $Z_{0.05}$ (= 1.645), we reject H_0. Thus, the proportion of pumps requiring repairs within the first five years than 0.30.

5. **A manufacturer claimed that at least 95% of the cables supplied to the ABC company confirmed to specifications. However, the production manager at ABC company wasn't satisfied with the claim of the manufacturer. Hence, to test the claim, the manager examined a sample of 250 cables supplied last month and found that 228 cables as per the specifications. Can you conclude that the production manager is right to doubt on the claim of the manufacturer? ($\alpha = 0.01$) [2071 Shrawan]**

Solution: See the solution to [2072 Magh].

6. **A book containing 500 pages was thoroughly checked. The distribution of number of error page was given below as:**

Number of errors	0	1	2	3	4	5
Number of pages	275	138	75	7	4	1

Use Chi-square test of goodness of fit, verity whether the data follow a Poisson distribution at 5% level of significance. [2071 Shrawan]

Solution:

Let 'f' represent number of pages and 'x' represents number of errors.

Total number of pages = N = Σf = 275 + 138 + 75 + 7 + 4 + 1 = 500

$$\lambda = \bar{X} = \frac{\Sigma fx}{N} \text{ (for a discrete data)}$$

$$= \frac{0 \times 275 + 1 \times 138 + 2 \times 75 + 3 \times 7 + 4 \times 4 + 5 \times 1}{500}$$

$$\therefore \quad \lambda = \bar{X} = 0.66$$

Now, we setup null and alternate hypothesis.

1. Null hypothesis: The given data follows a Poisson distribution.
2. Alternative hypothesis: The given data does not follows a Poisson distribution.
3. The test statistic for χ^2 test is

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

where, $E = N \times P(X = r) = N \times \frac{e^{-\lambda}\lambda^r}{r!}$ where, r = 0, 1, 2, 3, 4

or, $E = 500 \times \frac{e^{-0.66}(0.66)^r}{r!}$

r	O		E		$(O-E)^2$	$(O-E)^2/E$
0	275		258.426		274.697	1.063
1	138		170.561		1060.219	6.216
2	75		56.285		350.25	6.223
3	7	= 12	12.383	= 14.7	7.29	0.496
4	4		2.043			
5	1		0.2697			

Since 'E' values at r = 4 at r = 4 and r = 5 are less than 5, we pool 'E' and 'O' of r = 3, 5 and 5 together.

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 1.063 + 6.216 + 6.223 + 0.796 = 13.998$$

4. Degree of freedom = $n_1' - 1 - K_1 - K_2 = 6 - 1 - 1 - 2 = 2$

where, n_1' = Total no. of values = 6

K_1 = d.f. lost due to calculation

(Since 'λ' is calculated at the beginning of the question, we lose 1 d.f. because of it)

K_2 = d.f. lost due to pooling (The last two values are pooled with the third-last data. So, we have to reduce those d.f.)

For $\alpha = 0.05$ and d.f. = 2, $\chi^2_{0.05,2} = 5.991$

5. Since, χ^2 (= 13.998) > $\chi^2_{0.05,2}$ (= 5.991), we reject H_0. Thus, the given data doesn't follow a Poisson distribution.

7. **A random sample of 500 adult residents of Maricopa county found that 385 were in favour of increasing the highway speed limit to 75 mph, while another sample of 400 adult residents of pima county found that 267 were in favour of the increased speed limit. Construct 95% confidence interval on the difference in the two proportions.** **[2072 Kartik]**

Solution:

Given that;

$n_1 = 500, \quad X_1 = 385$

$n_2 = 400, \quad X_2 = 267$

$\hat{p}_1 = \frac{385}{500} = 0.77$

$\hat{p}_2 = \frac{267}{400} = 0.6675$

We know, the 95% confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Note: In case of confidence interval, it is not necessary to pool the sample proportions.

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

∴ 95% confidence limits

$$= 0.77 - 0.6675) \pm 1.96\sqrt{\frac{0.77(1-0.77)}{500} + \frac{0.6675(1-0.6675)}{400}}$$

$= (0.0434, 0.1616)$

8. **From the following data, can you concluded that there is association between the purchase of brand and geographical region?**

	Region		
	Central	Eastern	Western
Purchase brand	40	55	45
Do not purchase brand	60	45	55

[2072 Kartik]

Solution: See the solution to [2073 Chaitra].

9. **1072 students were classified according to their intelligence and economic conditions. Test whether there is any association between intelligence and economic condition.** **[2071 Chaitra]**

Economic condition	Intelligence			
	Excellent	Good	Mediocre	Dull
Good	48	199	181	82
Not good	81	185	190	106

Solution:

1. Null hypothesis: There is no signification association between economic condition and intelligence.
2. Alternative hypothesis: There is significant association between economic condition and intelligence.
3. We know the test statistic for χ^2 test of association is

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

where, $E = \frac{RT \times CT}{N}$

					Row total (RT)
	48	199	181	82	510
	81	185	190	106	562
Column Total (CT)	129	384	371	188	Row Total (N) = 1072

$E(48) = \frac{510 \times 129}{1072} = 61.37;$ $\quad E(199) = \frac{510 \times 384}{1072} = 182.69$

$E(181) = \frac{510 \times 371}{1072} = 176.50;$ $\quad E(82) = \frac{510 \times 188}{1072} = 89.44$

$E(81) = \frac{562 \times 129}{1072} = 67.63;$ $\quad E(185) = \frac{562 \times 384}{1072} = 201.313$

$E(190) = \frac{562 \times 371}{1072} = 194.498;$ $\quad E(106) = \frac{562 \times 188}{1072} = 98.56$

Now,

O	E	$(O - E)^2$	$(O - E)^2/E$
48	61.37	178.76	2.913
181	176.50	20.25	0.1147
81	67.63	178.76	2.643
190	194.498	20.232	0.104
199	182.69	266.0161	1.456
82	89.44	55.354	0.6189
185	201.313	266.114	1.322
106	98.56	55.3536	0.5616

$$\chi^2 = \Sigma\left[\frac{(O - E)^2}{E}\right]$$

$$= 2.913 + 0.1147 + 2.643 + 0.104 + 1.456 + 0.6189 + 1.322 + 0.5616$$

$$= 9.7332$$

4. For $\alpha = 0.05$ (say) and d.f. = $(r - 1) = (c - 1) = (2 - 1)(4 - 1) = 3$

where, r = No. of rows = 2

c = No. of columns = 4

and $\chi^2_{0.05, 3} = 7.815$

5. Since, χ^2 (= 9.7332) > $\chi^2_{0.05, 3}$ (= 7.815), we reject H_0.

Thus, there is significant association between economic condition and intelligence.

10. A manufacturer claimed that at least 95% of the water pumps supplied to the ABC company confirmed to specifications. However, the product manager at ABC company wasn't satisfied with the claim of the manufacturer. Hence, to test the claim, the manager examined a sample of 250 pumps supplied last month

and found that 228 water pumps as per the specification. Can you conclude that the production manager is right to doubt on the claim of the manufacturer? ($\alpha = 0.01$). [2073 Shrawan]

Solution: See the solution to [2072 Magh].

11. From the following data, can you conclude that there is association between the purchase of brand and geographical region? Use $\alpha = 0.05$. [2072 Chaitra]

	Region		
	Central	Eastern	Western
Purchase brand	40	55	45
Do not purchase brand	60	45	55

Solution: See the solution to [2073 Chaitra].

12. In a postal survey of 500 households, 330 said that they thought they were being overcharged for the public services within their area:

i) Calculate an approximate 99% confidence interval for the population proportion, P, of households who thought they were being overcharged.

ii) Estimate the size of sample required to estimate the value of p to be within 99% confidence limits off ±0.025. [2072 Chaitra]

Solution:

Given that;

$$n = 500, \quad X = 330, \quad \hat{p} = \frac{X}{n} = \frac{330}{500} = 0.66$$

$$\alpha = 1\% = 0.01$$

i) The sample size is large and we know that for a large sample size, the confidence interval for proportion is

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= \hat{p} \pm Z_{\frac{0.01}{2}} \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= 0.66 \pm 2.575 \times \sqrt{\frac{0.66 \times (1 - 0.66)}{500}}$$

$$= (0.60545, 0.7146)$$

ii) For 'P' to be within 99% confidence limits of ± 0.025, (considering constant $\hat{p}$)

$$Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.025$$

or, $$Z_{0.005} \times \sqrt{\frac{0.66 \times (1 - 0.66)}{n}} = 0.025 \qquad [\because Z_{0.05} = 2.575]$$

Solving, we get

$$n = \left[\left(\frac{0.025}{2.575}\right)^2 \times \frac{1}{0.66 \times (1 - 0.66)}\right]^{-1} = 2380.66 \approx 2381$$

Thus, the sample size must be at least 2381.

13. **A random sample of smokers was obtained and each individual was classified both with respect to gender and with respect to age at which he/she first started smoking.**
Carry out a test of hypothesis to decide whether there might be an association between gender and the age at which an individual first smokes? **[2074 Ashwin]**

	Gender	
Age	**Male**	**Female**
< 16	25	10
16 – 17	24	32
18 – 20	28	17
> 20	19	34

Solution:

Since, it is asked to test association between two characteristics, we use χ^2 test for association.

1. Null hypothesis: There is no association between gender and age at which a person first smokes.
2. Alternative hypothesis: There is significant association between gender and age at which a person first smokes.
3. The test statistic for χ^2 test is given is:

$$\chi^2 = \Sigma\left[\frac{(O - E)^2}{E}\right]$$

where, E = expected frequency $= \dfrac{RT \times CT}{N}$

O = Observed frequency

Here, r = No. of rows = 4

c = No. of column = 2

Degrees of freedom = (r – 1) (c – 1) = (4 – 1) (2 – 1) = 3

Now,

			Row Total (RT)
	25	10	35
	24	32	56
	28	17	45
	19	34	53
Column Total (CT)	96	93	Grand Total (GT) = N = 189

Now,

$$E(25) = \frac{35 \times 96}{189} = 17.78; \qquad E(10) = \frac{35 \times 93}{189} = 17.22$$

$$E(24) = \frac{56 \times 96}{189} = 28.44; \qquad E(32) = \frac{56 \times 93}{189} = 27.55$$

$$E(28) = \frac{45 \times 96}{189} = 22.86; \qquad E(17) = \frac{45 \times 93}{189} = 22.143$$

$$E(19) = \frac{53 \times 96}{189} = 26.92; \qquad E(17) = \frac{53 \times 93}{189} = 26.08$$

Now,

O	E	$(O - E)^2$	$(O - E)^2/E$
25	17.78	52.1284	2.932
24	28.44	19.7136	0.6932
28	22.86	26.4196	1.1557
19	26.92	62.7264	2.33
10	17.22	52.1284	3.0272
32	27.55	19.8025	0.7188
17	22.143	26.45	1.1945
34	26.08	62.7264	2.405

$$\chi^2 = \Sigma\left[\frac{(O - E)^2}{E}\right]$$

$$= 2.932 + 0.6932 + 1.1557 + 2.33 + 3.0272 + 0.7188 + 1.1945 + 2.045$$

$$= 14.4564$$

4. For $\alpha = 0.05$ (assume) and d.f. = 3

$\chi^2_{0.05, 3} = 7.815$

5. Since, χ^2 (= 14.4564) > $\chi^2_{0.05, 3}$ (= 7.815), we reject H_0.
Thus, there is significant association between gender and age at which people first smoke.

14. **In a random sample of 400 industrial accident, it was found that 231 were due to at least confidence intervals for corresponding true proportions.** **[2074 Ashwin]**

Solution: See the solution to [2071 Magh].

15. **From the following data can you concluded that there is association between the purchase of brand and geographical region? Use 5% level of significance.** **[2073 Chaitra]**

	Region		
	Central	**Eastern**	**Western**
Purchase brand	40	55	45
Do not purchase brand	60	45	55

Solution:

Since this is an association test question, we use χ^2 test of association.

1. Null hypothesis: There is no association between purchase of brand and geographical region.
2. Alternative hypothesis: There is significant association between purchase of brand and geographical region.
3. Now, we know that;

$$E = \frac{RT \times CT}{N}$$

where, RT = Row total

CT = Column total

N = Total number of surveyed people

Here;

				Row total (RT)
	40	55	45	140
	60	45	55	160
Column Total (CT)	100	100	100	Grand Total (GT) = 100

Now, N = GT = 300

So, $E(40) = \frac{RT \times CT}{N} = \frac{140 \times 100}{300} = 46.67$

$E(40) = \frac{140 \times 100}{300} = 46.67$ (55 from first row)

$E(55) = \frac{140 \times 100}{300} = 46.67$ (45 from first row)

$E(60) = \frac{160 \times 100}{300} = 53.33$

$E(45) = \frac{160 \times 100}{300} = 53.33$ (45 from second row)

$E(55) = \frac{160 \times 100}{300} = 53.33$ (55 from second row)

So,

O	E	$(O - E)^2$	$(O - E)^2/E$
40	46.67	44.4889	0.9533
55	46.67	69.3889	1.4868
45	46.67	2.7889	0.0596
60	53.33	44.4889	0.8342
45	53.33	69.3889	1.3011
55	53.33	2.7889	0.0523

Now, $\chi^2 = \Sigma\left[\frac{(O - E)^2}{E}\right]$

$= 0.9533 + 1.4868 + 0.0596 + 0.8342 + 1.3011 + 0.0523$

$= 4.6873$

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

where, r = no. of rows = 2

c = no. of columns = 3

From table, $\chi^2_{0.05, 2} = 5.991$

5. Since, χ^2 (= 4.6873) < $\chi^2_{0.05, 2}$ (= 5.991), we accept H_0.

Thus, there is no association between purchase of brand and geographical region.

16. Two different areas of a city are being considered as sites for day care centers. Of 200 households surveys in one section, the proportion in which 40% of 150 households surveyed has mother at full time jobs. At 0.05 level of significance, is there a significant difference in the proportion of working mothers in the two areas of the city? **[2073 Chaitra]**

Solution:

Given that;

$n_1 = 200$, $\hat{p}_1 = 0.52$

$n_2 = 150$, $\hat{p}_2 = 0.40$

$\alpha = 0.05$

Now,

1. Null hypothesis: There is no significant difference between working mothers at two sites. Symbolically, H_0: $P_1 = P_2$.
2. Alternative hypothesis: There is significant difference between proportion of working moms in the two areas.

   Symbolically, H_1: $P_1 \neq P_2$.

   This is a large sample two tailed test.
3. The test statistic under the given circumstances is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{200 \times 0.52 + 150 \times 0.40}{200 + 150} = 0.4686$

$\hat{q} = (1 - \hat{p}) = (1 - 0.4686) = 0.5314$

So, $Z = \frac{(0.52 - 0.40) - 0}{\sqrt{0.4686 \times 0.5314 \times \left(\frac{1}{200} + \frac{1}{150}\right)}} = 2.2264$

4. The critical value is

$\alpha = 0.05$

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = \pm 1.96$

5. Since, $|Z|$ (= 2.2264) > $|Z_{0.025}|$ (= 1.96), we reject H_0. Thus, there is significant difference between proportion of working mothers in the two areas.

17. **A book containing 500 pages, was thoroughly checked. The distribution of number of error page was given below as;**

Number of errors	0	1	2	3	4	5
Number of pages	275	138	75	7	4	1

Using chi-square test of goodness of fit, verify whether the arrivals follow a Poisson distribution at 5% level of significance.

[2074 Chaitra]

Solution: See previous question.

18. **A manufacturer claimed that at least 95% of the water pumps supplied to the ABC company confirmed to specification. However, the product manager at ABC company wasn't satisfied with the claim of the manufacturer. Hence, to test the claim, the manager examined a sample of 250 water pumps supplied last month and found that 228 water pumps as per specification. Can you conclude that the production manager is right to doubt on the claim of the manufacturer? ($\alpha = 0.01$)** **[2075 Ashwin]**

Solution: See previous question.

Chapter 7

CORRELATION AND REGRESSION

7.1 Least Square Method 226
7.2 Regression 227
7.2.1 Linear Regression with Least Square Method 227
7.2.2 Multiple Regression 227
7.3 Testing Significance of Regression 230
7.4 Correlation 231
7.4.1 Significance of Correlation 231
7.4.2 Types of Correlation 231
7.4.3 Partial Correlation Coefficient 233
7.4.4 Difference between Correlation and Regression 234

7.1 LEAST SQUARE METHOD

The least squares method is a form of mathematical regression analysis that finds the line of best fit of a dataset providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship between a known independent variable and an unknown dependent variable.

The principle of least squares is one of the popular methods for finding a curve fitting a given data.

Least square fit of a straight line

Suppose we have a dataset $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ of n observations from an experiment. Now, we want to fit these data on a straight line whose equation is

$$y = a + bx \qquad (1)$$

where y = is y-axis value

x = is x-axis value

a = is y-intercept

b = is slope

Then, it's normal equation are

$$\Sigma y = na + b\Sigma x \qquad (2)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad (3)$$

Solving equation (2) and (3), we get

Value of a and b and by substituting the value of a and b in equation (1), we get the required straight line equation.

Least square fit for a parabola

Suppose we have a data set $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ of n observations of an experiment. Now, we want to fit these data on a parabola whose equation is

$$y = a + bx + cx^2 \qquad (4)$$

Then, it's normal equations are given as

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 \qquad (5)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \qquad (6)$$

$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \qquad (7)$$

Solving equation (5), (6) and (7), we get

Value of a, b and c and substituting these values in equation (4), we obtain the required equation of parabola of best fit.

7.2 REGRESSION

Linear regression is a statistical method that allows us to summarize and study relationship between two continuous variables.

The general equation for straight line

$$y = a + bx \qquad (1)$$

is called regression equation of y on x.

where, y = y-axis value

x = x-axis value

a = y-intercept

b = slope of regression (coefficient of regression)

The slope of the line 'b' measures the amount by which y changes for each unit change in x.

Regression line always passes through means (mean of x and y) so,

We can write

$$(\bar{y}) = a + b\bar{x} \qquad (2)$$

From (2), we have

$$a = \bar{y} - b\bar{x}$$

and the coefficient of regression can be obtained by

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \qquad \text{[with using mean]}$$

$$b_{yx} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \qquad \text{[without using mean]}$$

7.2.1 LINEAR REGRESSION WITH LEAST SQUARE METHOD

The given straight line equation is

$$y = a + bx \qquad (1)$$

It's normal equations are

$$\Sigma y = na + b\Sigma x \qquad (2)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad (3)$$

Solving (2) and (3) equations, we obtain the value of a and b and substituting these values in equation (1), we get the predictor line (estimator line) equation and is given as

$$\hat{y} = a + b\hat{x}$$

7.2.2 MULTIPLE REGRESSION

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variable. The variable we want to predict is called the dependent variable and the variable(s) we are using to predict the value of the dependent variable are called independent variables.

Multiple regression also allows you to determine overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained.

Suppose we have 3 variables x_1, x_2 and x_3 and x_1 is dependent variable and remaining x_2 and x_3 are independent variables. The multiple regression equation of dependent variable x_1 on x_2 and x_3 is given by

$$x_1 = \alpha + \beta_1 x_2 + \beta_2 x_3 + e_i \quad (1)$$

where i = 1, 2, 3, ..., n

α = y-intercept

β_1 and β_2 are partial regression coefficient

e_i = error terms and F(e) = 0

Using least square method to obtained the value α and partial equations coefficients.

The normal equation of equation (1) are as follow

$$\Sigma x_1 = na + \beta_1 \Sigma x_2 + \beta_2 \Sigma x_3 \quad (2)$$

$$\Sigma x_1 x_2 = \alpha \Sigma x_2 + \beta_1 \Sigma x_2^2 + \beta_2 \Sigma x_2 x_3 \quad (3)$$

$$\Sigma x_1 x_3 = \alpha \Sigma x_3 + \beta_1 \Sigma x_2 x_3 + \beta_2 x_3^2 \quad (4)$$

Solving equation (2), (3) and (4), we obtains the value of α, β_1 and β_2 and substituting these values in equation (1), we get the fitted regression equation which is given as

$$\hat{x}_1 = \alpha + \beta_1 \hat{x}_2 + \beta_2 \hat{x}_3$$

Let us solving some examples

Example 1

Observation on the yield of a chemical reaction taken at various temperature was recorded as follows:

X(°C)	150	150	200	250	250	300	150
Y%	75.4	81.2	85.5	89	90.5	96.7	75.4

Fit a simple linear regression and estimate value of yield at 200°C.

Solution: Given,

Y(%)	X(°C)	XY	X^2
75.4	150	11310	22500
81.2	150	12180	22500
75.4	150	11310	22500
85.5	200	17100	40000
89	250	22250	62500
90.5	250	22625	62500
96.7	300	29010	90000
593.7	1450	125785	322500

The linear regression equation is y = a + bx

Its normal equations are

$$\Sigma y = na = b\Sigma x \quad (1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad (2)$$

From equation (2), $593.7 = 7a + 1450b$ (3)

From equation (3), $125785 = 1450a + 322500b$ (4)

Solving (4) and (5), we get (5)

a = 58.58

b = 0.1266

Note: The value of Σx, Σy, Σxy and Σx^2 can be found directly using calculator.

∴ The required equation is $\hat{y} = 58.58 + 0.1266\,\hat{x}$

∴ The value at 200°C is

$y = 58.58 + 0.1266 \times 200$

∴ y = 83.9%

Example 2

The following table shows the corresponding values of three variables x_1, x_2 and x_3.

a) Find the linear multiple regression equation x_3 on x_1 and x_2.

b) Estimate the value of x_3 when $x_1 = 10$ and $x_2 = 12$.

x_1	3	5	6	9	12	15
x_2	17	11	8	5	3	2
x_3	90	73	56	45	30	13

Solution:

The linear regression equation of x_3 on x_1 and x_2 is given by

$$x_3 = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad (1)$$

From above data, using calculator we obtain

$\Sigma x = 50$, $\Sigma x_1^2 = 520$

$\Sigma x_2 = 46$, $\Sigma x_2^2 = 512$, $\Sigma x_1 x_2 = 265$

$\Sigma x_3 = 307$, $\Sigma x_3 x_2 = 3122$, $\Sigma x_1 x_3 = 1931$

The normal equation of equation (1) are

$$\Sigma x_3 = n\alpha + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2$$

i.e., $307 = 6\alpha + 50\beta_1 + 46\beta_2$ (2)

$$\Sigma x_3 x_1 = \alpha \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_2 x_1$$

i.e., $1931 = 50\alpha + 520\beta_1 + 265\beta_2$ (3)

$\Sigma x_3x_2 = \alpha x_2 + \beta_1 \Sigma x_1x_2 + \beta_2 \Sigma x_2^2$

i.e., $3122 = 46\alpha + 265\beta_1 + 512\beta_2$ (4)

Solving equation (2), (3) and (4), we obtain

$\alpha = 65.68$

$\beta_1 = -3.67$

$\beta_2 = 2.096$

Hence, the required equation is

$\therefore \quad \hat{x}_3 = 65.68 - 3.67\hat{x}_1 + 2.096\hat{x}_2$

b) Value of x_3 when $x_1 = 10$ and $x_2 = 12$

$x_3 = 65.68 - 3.67 \times 10 + 2.096 \times 12$

$\therefore \quad x_3 = 54.132$

7.3 TESTING SIGNIFICANCE OF REGRESSION

After obtaining the regression equation it must be evaluated to determine whether it adequately described the relationship between two variables and whether it can be used effectively for prediction and estimation purposes.

Analysis of variance techniques

Step 1: Setting of hypothesis

H_0: B = 0 (there is no linear relationship between x and y)

H_1: B ≠ 0 (there is linear relationship between x and y)

Step 2: Test statistics *i.e.*, hypothesis testing is tested by F-test using ANNOVA technique.

ANNOVA Table

S.V.	df	Sum of sq. (SS)	Mean sum of sq (MSS)	F-ratio
Linear regression	$\alpha - 1 = 1$	SSR	$\frac{SSR}{df}$	$\frac{MSSR}{MSSE}$
Residual	n – 2	SSE	$\frac{SSE}{df}$	

SST (total sum of square)

$$\Sigma(y - \bar{y})^2 = \Sigma y^2 - \frac{\Sigma y^2}{n}$$

SRR (Explained sum of square/regression sum of square)

$$\Sigma(\hat{y} - \bar{y})^2 = b^2\left[\Sigma x^2 - \frac{\Sigma x^2}{n}\right]$$

Unexplained sum of square (Error sum of square)

SSE = SST – SSR

Step 3: Level of significance (a) = 1%, 5% or 10%

Step 4: Decision

$F_{cal} < F_{table}$ (we accept H_0)

$F_{cal} \geq F_{table}$ (we accept H_1)

Step 5: Conclusion

7.4 CORRELATION

Correlation is a statistical technique that can show whether and how strongly pairs of variable are related. It is a bivariate measure of the relationship between quantitative variables.

It shows the degree of strength of relationship of two fingers. Sample correlation is denoted by r and population correlation is denoted by ρ.

Note: Correlation always lies between –1 to 1.

7.4.1 SIGNIFICANCE OF CORRELATION

i) Help to know the nature, direction and degree of relationship between two or more variables.

ii) Help in understanding the behavior of certain event under specific condition.

iii) Helps in making prediction and decision.

iv) Helps in making regression analysis.

7.4.2 TYPES OF CORRELATION

i) Positive and negative correlation

ii) Linear and non-linear correlation

iii) Simple, partial and multiple correlation

1. Positive and negative correlation

A correlation in the same direction is called positive correlation *i.e.*, if one variable increases the other also increases and vice-versa.

A correlation in the opposite direction is called negatives correlation *i.e.*, if one variable increases, the other variable.

2. Linear and non-linear correlation

Correlation is said to be linear if the ratio of change is constant. In other words, when all points on the scatter diagram tend to lie near a line which looks like a straight line, the correlation is said to be linear.

Correlation is said to be non-linear if the ratio of change is not constant. In other words, when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is said to be non-linear.

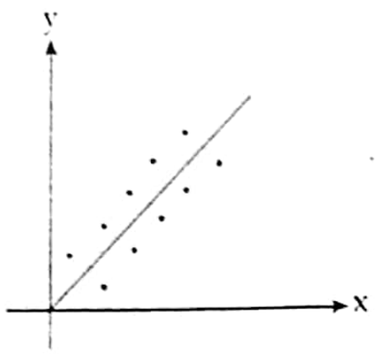


Fig.(i): Linear correlation

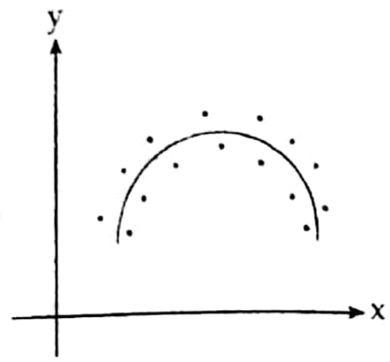


Fig.(ii): Non-linear correlation

3. Simple, partial and multiple correlation

Simple correlation coefficient is defined as the correlation between two variables. Its general measure is Karl person coefficient of correlation. The correlation coefficient between the variables x and y is denoted by

$$r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} \times \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{y})^2}}$$

Also, $r = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}}$

When we have one dependent variable and two more independent variables, then the only means of measuring the association between these variables is multiple correlation. *i.e.*, multiple correlation measures the association between one dependent variable and two or more independent variables jointly.

Let x_1, x_2 and x_3 be three variables then

$R_{1.23}$: Measures the multiple correlation between dependent variable x_1 and joint effect of independent variables x_2 and x_3 on x_1.

Mathematically,

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Similarly,

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

Coefficient of determination

Coefficient of determination is defined as the square of multiple correlation coefficient. It measures at what extent the total variation in dependent variable is explained by independent variables.

For example, $R_{1.23} = 0.60$ then $R_{1.23}^2 = 0.36$ which indicates that the 36% of variation in dependent variable is explained by independent variable and remaining 64% is unexplained.

Properties

1. Multiple correlation coefficient always lies between 0 and 1.
2. Multiple correlation coefficient is not less than simple correlation coefficient.

   i.e., $R_{1.23} \geq r_{12}$, $R_{1.23} \geq r_{23}$, $R_{1.23} \geq r_{13}$
3. The position of subscript to the right side of dot does not make any difference in the meaning

   i.e., $R_{1.23} = R_{1.32}$

7.4.3 PARTIAL CORRELATION COEFFICIENT

A partial correlation coefficient is a measure of the liner dependency of a pair of random variables from a collection of random variables in the case where the influence of the remaining variables is eliminated.

Usually partial correction coefficient is written as $r_{12.3,4,5,\ldots,n}$. Here, r is the correlation studies between x_1 and x_2 by holding all other variables ($x_3, x_4, x_5, \ldots\ldots, x_n$) constant.

Let as consider three variables x_1, x_2 and x_3 then partial correction coefficient between x_1 and x_2 holding x_3 as constant is given as

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

Similarly,

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

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

Coefficient of determination

Coefficient of determination is defined as the square of partial correlation coefficient.

For example, $r_{12.3} = 0.9$ then $r_{12.3}^2 = 0.81$ indicates that the 81% of variation in the x_1 is dependent or associated with x_2 but not associated with x_3 variable.

Properties

1. Partial Correlation coefficient always lies between

   -1 to 1 *i.e.*, $-1 \le r_{12.3} \le 1$

2. The position of subscript on left hand side of dot does not make any difference in the meaning.

   i.e., $r_{12.3} = r_{21.3}$

7.4.4 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation	Regression
1. Correlation is a statistical measure which determines co-relationship or association of two variable.	1. Regression describes how an independent variable is numerically related to the dependent variable.
2. To represent linear relationship between two variables.	2. To fit a best line and estimate one variable on the basis of another variable.
3. Correlation coefficient indicates the extent to which two variables move together. $r_{xy} = r_{yx}$ is symmetric	3. Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y) $b_{xy} \neq b_{yx}$
4. To find a numerical value expressing the relationship between variables.	4. To estimate value of random variable on the basis of the values of fixed variable.

Example 3

To know the effect of certain fertilizer and rainfall of certain crop *i.e.*, experimental observation where collected below:

Yield (kg/acre)	Fertilizers (kg/acre)	Rainfall (in cm)
x_1	x_2	x_3
40	10	10
50	20	20
50	30	10
70	40	30
65	50	20
65	60	20
80	70	30

Calculate partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{12.3}$ and interpret it.

Solution:

We know;

$$\text{Correlation } (r) = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} \times \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{y})^2}}$$

From above data, using calculator we get

$r_{12} = 0.9195$

$r_{13} = 0.8846$

$r_{23} = 0.6614$

Now, partial correlation coefficient

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

$$= \frac{0.9195 - 0.8846 \times 0.6614}{\sqrt{1 - (0.8846)^2}\sqrt{1 - (0.6614)^2}}$$

$$= 0.9561$$

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.9195^2 + 0.8846^2 - 2 \times 0.9195 \times 0.8846 \times 0.6616}{1 - 0.6614^2}}$$

$$= 0.9906$$

SOLVED QUESTIONS (FOR BCE/BME/BIE/BAM)

1. State the differences between correlation and regression. [2070 Magh]

Solution: See the 7.4.4.

2. The performance evolution of five employees on the basis of their aptitude test score and prior work experience in years are as follow:

Performance evaluation (X_1)	9	12	10	7	17
Aptitude test score (X_2)	2	5	4	3	6
Prior experience (year) X_3	4	5	5	3	8

a) Fit a multiple linear regressions model test describing these data.

b) If an employee scored 8 on the aptitude test and has a prior experience of 7 years, what performance evaluation would be expected?

OR

The following show the improvement gain in the reading speed of eight students in a speed reading program, and the number of weeks they have been in program:

No. of weeks	3	5	2	8	6	9	3	4
Speed gain (wood minute)	86	118	49	193	164	232	73	109

Estimate the parameters of a simple linear regression model with no. of weeks as independent variable. [2070 Magh]

Solution:

The multiple linear regression best describing these data is given by

$$x_1 = \alpha + \beta_1 x_2 + \beta_2 x_3 \quad (1)$$

where regression of x_1 on x_2 and x_3 is defined.

Now, from above data, using calculator, we get

$\Sigma x_1 = 55$

$\Sigma x_2 = 20, \Sigma x_2^2 = 90, \Sigma x_1x_2 = 241$

$\Sigma x_3 = 25, \Sigma x_3^2 = 139, \Sigma x_2x_2 = 110$

$\Sigma x_1x_3 = 303$

The normal equations of equation (1) are

$\Sigma x_1 = n\alpha + \beta_1\Sigma x_2 + \beta_2\Sigma x_3$

i.e., $55 = 5\alpha + 20\beta_1 + 25\beta_2$ (2)

$\Sigma x_1x_2 = \alpha\Sigma x_2 + \beta_1\Sigma x_2^2 + \beta_2\Sigma x_2x_3$

i.e., $241 = 20\alpha + 90\beta_1 + 110\beta_2$ (3)

$\Sigma x_1x_3 = \alpha\Sigma x_3 + \beta_1\Sigma x_2x_3 + \beta_2\Sigma x_3^2$

i.e., $303 = 25\alpha + 110\beta_1 + 139\beta_2$ (4)

Solving equation (2), (3) and (4); we get,

$\therefore \alpha = 0.85, \beta_1 = 0.35$ and $\beta_2 = 1.75$

Substituting these values in equation (1); we get,

a) $x_1 = 0.85 + 0.35x_2 + 1.75x_3$

b) When $x_2 = 8$ and $x_3 = 7$

$\therefore x_1 = 0.85 + 0.35 \times 8 + 1.75 \times 7$

$\therefore x_1 = 15.9$

OR

Given that;

Speed gain (word/minute) (y)	No. of weeks (x)	xy	x^2
86	3	258	9
118	5	590	25
49	2	98	4
193	8	1544	64
164	6	984	36
232	9	2088	81
73	3	219	9
109	4	426	16
1024	**40**	**6217**	**244**

The linear regression equation is $y = a + bx$ (1)

It's normal equations are

$\Sigma y = na + b\Sigma x$

i.e., $1024 = 8a + 40b$ (2)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

i.e., $6217 = 40a + 244b$ (3)

Solving (2) and (3); we get,

$a = 3.34$

$b = 24.93$

Substituting these values in equation (1), we get

$y = 3.34 + 24.93x$

3. Write down the properties of regression coefficient and correlation coefficient. [2070 Bhadra]

Solution: Properties of regression coefficient

Definition: The constant 'b' in the regression equation ($Y_e = a + bX$) is called as the regression coefficient. It determines the slope of the line, *i.e.*,

the change in the value of Y corresponding to the unit change in X and therefore, it is also called as a "slope coefficient".

Properties of regression coefficient

i) The correlation coefficient is the geometric mean of two regression coefficients. Symbolically,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

ii) The value of the coefficient of correlation cannot exceed unity *i.e.*, 1. Therefore, if one of the regression coefficients is greater than unity, the other must be less than unity.

iii) The sign of both the regression coefficients will be same, *i.e.*, they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.

iv) The coefficient of correlation will have the same sign as that of the regression coefficients, such as of the regression coefficients have a positive sign then 'r' will be positive and vice-versa.

v) The average value of the two regression coefficients will be greater than the value of the correlation. Symbolically, it can be represented as

$$\frac{b_{xy} + b_{yx}}{2} > r$$

vi) The regression coefficient are independent of the change of origin, but not of the scale. By origin, we mean that there will be no effect on the regression coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or dividend by some constant, then the regression coefficients will also change.

Thus, all these properties should be kept in mind while solving for the regression coefficients.

Properties of coefficient of correlation

- The value of the coefficient of correlation (r) always lies between ± 1, Such as;

  r = +1, perfect positive correlation

  r = –1, perfect negative correlation

  r = 0, no correlation

- The coefficient of correlation is independent of the origin and scale. By origin, it means subtracting any non-zero constant from the given value of X and Y the value of 'r' remains unchanged. By scale it means, there is no effect on the value of 'r' if the value of C and Y divided or multiplied by any constant.

- The coefficient of correlation is a geometric mean of two regression coefficient. Symbolically, it is represented as;

  $$r = \sqrt{b_{xy} \times b_{yx}}$$

- The coefficient of correlation is 'zero; when the variable X and Y are independent. But, however, the converse is not true.

4. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Experience (X)	16	12	18	4	3	10	5	12
Performance (Y)	87	88	89	68	78	80	75	83

a) Find the regression equation of performance rating on experience and estimate the probable performance if an operator has 8 years experience.

b) Calculate coefficient of determination and interpret it.

OR

A Sample of 10 values of three variable x_1, x_2 and x_3 were obtained as

$\Sigma X_1 = 10$	$\Sigma X_2 = 20$	ΣX_3 30
$\Sigma X_1^2 = 20$	$\Sigma X_1^2 = 68$	$\Sigma X_3^2 = 170$
$\Sigma X_1X_2 = 10$	$\Sigma X_1X_3 = 15$	$\Sigma X_2X_3 = 64$

Find,

a) Partial correlation between x_1 and x_3 eliminating the effect of x_2.

b) Multiple correlation between x_1, x_2 and x_3 assuming x_1 as dependent. **[2070 Bhadra]**

Solution:

Given that:

Experience (X)	Performance (Y)	XY	X^2	Y^2
16	87	1392	256	7569
12	88	1056	144	7744
18	89	1602	324	7921
4	68	272	16	4624
3	78	234	9	6084
10	80	800	100	6400
5	75	375	25	5625
12	82	996	144	6889
80		6727	1018	52856

The regression equation of y on x is given by

$$y = a + bx \quad (1)$$

It's normal equations are

$$\Sigma y = na + b\Sigma x$$

i.e., $648 = 8a + 80b$ (2)

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

i.e., $6727 = 80a + 1018b$ (3)

Solving equation (3) and (4), we get;

$a = 69.66$

$b = 1.13$

∴ Hence, the required equation is $y = 69.66 + 1.13x$

When experience (x) = 8 year then

Performance (y) $69.66 + 1.13 \times 8 = 80$

b) Here, $b_{yx} = 1.13$

and $b_{xy} = \dfrac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \dfrac{8 \times 6727 - 80 \times 648}{8 \times 52856 - (648)^2} = 0.6711$

∴ Coefficient of determination $(r^2) = bxy \times byx$

$= 0.6711 \times 1.13 = 0.7583$

This means that one variable explain 75.83% of total variation in other variable.

OR

a) Partial correlation between x_1 and x_3 eliminating the effect of x_2 is

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

We know,

$$r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} \times \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{y})^2}}$$

∴ $r_{12} = -0.5976$

$r_{13} = -0.53$

$r_{23} = 0.0845$

Substituting these values in equation (1), we get

$$r_{13.2} = \frac{(-0.53) - (-0.5976) \times 0.0845}{\sqrt{1 - (-0.5976)^2} \times \sqrt{1 - (0.0845)^2}} = -0.6$$

b) Multiple correlation between x_1, x_2 and x_3 assuming x_1 as dependent is

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(-0.5976)^2 + (-0.53)^2 - 2 \times (-0.5976) \times (0.53) \times (0.0845)}{1 - 0.0845^2}}$$

$= 0.7672$

5. Define person's correlation coefficient. Write down two differences between correlation and regression coefficient. [2071 Magh]

Solution: Karl Pearson's coefficient of correlation

Definition: Karl Pearson's coefficient of correlation is widely used mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Person's method, popularly known as a Personian coefficient of correlation is the most extensively used quantitative methods in practice. The coefficient of correlation is denoted by 'r'.

If the relationship between two variables X and Y is to be ascertained, then the following formula is used;

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2}\sqrt{\Sigma(y - \bar{y})^2}}$$

where, $\bar{x}$ = Mean of x variable

$\bar{y}$ = Mean of y variable

6. The following are the measurements of the air velocity and evaporations coefficient of burning fuel droplets in an impulse engine. [2071 Magh]

Air velocity (x) cm/sec	20	60	100	140	180	220	260	300	340	380
Evaporation coefficient mm²/sec (y)	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1.65

Fit the straight line by method of least square and find the value of evaporation coefficient when air velocity is 200 cm/sec.

Solution:

Here;

We know that, equation of line is

$Y = A + BX$

where, X is independent variable

Y is dependent variable

A and B are constant parameter.

Now, using least square method

$\Sigma Y = nA + B\Sigma X$ (1)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ (2)

Now, let

$X = \frac{x - 200}{10}, \quad Y = y$

x is air velocity in cm/sec.

y is evaporation coefficient in mm^2/sec.

S.N.	x	y	X	Y	XY	X^2
1	20	0.18	–18	0.18	–3.24	324
2	60	0.37	–14	0.37	–5.18	169
3	100	0.35	–10	0.35	–3.5	100
4	140	0.78	–6	0.78	–4.68	36
5	180	0.56	–2	0.56	–1.12	4
6	220	0.75	2	0.75	1.50	4
7	260	1.18	6	1.18	7.085	36
8	300	1.36	10	1.36	13.6	100
9	340	1.17	14	1.17	16.38	196
10	280	1.65	18	1.65	29.70	324
Sum			0	8.35	50.54	1320

Hence,

$8.35 = 10A + B \times O$

or, $A = 0.835$

$50.54 = A \times 0 + B \times 1320$

or, $B = 0.0383$

So, $Y = 0.835 + 0.038\,X$

or, $y = 0.835 + 0.0383\,\frac{(x - 200)}{10}$

or, $\boxed{y = 0.069 + 0.00383x}$

This is linear regression equation.

So, for x = 200 cm/sec.

$y = 0.069 + 0.00383 \times 200$

$\boxed{y = 0.835}$

7. Define the correlation coefficient and mention its important properties. What does the coefficient of determination measure? **[2071 Bhadra]**

Solution: Refer to 7.4.2 (iii).

8. An article in the Tappi Journal (March, 1986) presented data on green liquor Na_2S concentration (in grams per liter) and paper machine production (in tons per day). The data (read from a graph) are shown as follows: **[2071 Bhadra]**

y	40	42	49	46	44	48	46
x	825	830	890	895	890	910	915
y	43	53	52	54	57	58	
x	960	990	1010	1030	1030	1050	

a) Fit a simple line regression model with y = green liquor Na_2S concentration and x = production.

b) Find the fitted value of y corresponding to x = 910 tons.

Solution: Here,

$Y = A + BX$

where X and Y are independent and dependent variable respectively.

A and B are constant parameter.

We know that to find out A and B, we can solve below equations;

$\Sigma Y = An + B\Sigma X$ (1)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ (2)

And, let,

$Y = y - 49$

$X \frac{(x - 915)}{5}$

S.N.	Y	X	Y	X	XY	X^2
1	40	825	–9	–18	162	324
2	42	830	–7	–17	119	289
3	49	890	0	–5	0	25
4	46	895	–3	–4	12	16
5	44	890	–5	–5	25	25
6	48	910	–1	–1	1	1
7	46	915	–3	0	0	0
8	43	960	–6	9	–54	81
9	53	990	4	15	60	225
10	52	1010	3	19	57	361
11	54	1030	5	23	115	529
12	57	1030	8	23	184	529
13	58	1050	9	27	243	729
Sum			–5	66	924	3134

And,

$-5 = 13A + 66B$

$924 = 66A + 3134B$

Solving, we get;

$A = -2.11, B = 0.34$

So, equation becomes,

$Y = -2.11 + 0.34X$

or, $y - 49 = -2.11 + 0.34\left(\frac{x - 915}{5}\right)$

or, $y = 49 - 2.11 + \frac{0.34x}{5} - 62.22$

or, $\boxed{y = -15.33 + \frac{x}{14.71}}$

And, for x = 910 tons

$y = -15.33 + \frac{910}{14.71} = 46.53$ gm per liter.

OR

The following show the important (gain in reading speed) of eight students in a speed reading program, and the number of weeks they have in program;

No. of weeks	3	5	2	8	6	9	3	4
Speed again word/min	86	118	49	193	164	232	73	109

Estimate the parameters of a simple linear regression model with no. of weeks as independent variable.

Solution:

Here, we know

$Y = A + BX$

where,

X and y is independent and dependent variable, respectively.

A and B are constant parameter.

And, to find A and B, we use these two equation.

$\Sigma Y = nA + B\Sigma X$

$\Sigma XY = A\Sigma X + B\Sigma X^2$

Then, let

$X = x - 5$

$Y = y - 164$

where, x is no. of weeks, y is speed gain.

S.N.	x	y	X	Y	XY	X^2
1	3	86	-2	-78	156	4
2	5	118	0	-46	0	0
3	2	49	-3	-115	345	9
4	8	193	3	29	87	8
5	6	164	1	0	0	1
6	9	232	4	68	272	16
7	3	73	-2	-91	182	4
8	4	109	-1	-53	53	1
Sum			0	-288	1097	44

$n = 8$

$\Sigma X = 0$

$\Sigma Y = -288$

$\Sigma XY = 1097$

$\Sigma X^2 = 44$

So, $-288 = 8A + 0B$

or, $A = -36$

$1097 = 0A + 44B$

or, $B = 24.93$

Hence,

$Y = -36 + 24.93X$

or, $y - 164 = -36 + 24.93(x - 5)$

or, $\boxed{y = 3.35 + 24.93x}$

This is linear regression of given data as no. of weeks (x) as independent variable.

9. **Mention the properties of correlation coefficient. What is coefficient of determination and what does its signify? [2072 Magh]**

Solution: See the 7.4.2 (iii).

12. **The following table shows has many weeks a sample of six person have worked at an automobile inspection station and the number of cars each one inspected between noon and 2 pm on given day: [2072 Magh]**

No. of week employed (x)	2	7	9	1	5	12
No. of car inspected	13	21	23	14	15	21

a) **Estimate the parameters of a simple linear regression model with no. of weeks employed as independent variable.**

b) **Use the result of part (a) to estimate how many cars someone who has been working at the inspection station for 8 week can be expected during the given 2 hours period.**

Solution:

Here;

We know that, Simple linear line equation,

$Y = A + BX$

where, X and Y are independent and dependent variable respectively.

A and B are constant parameters.

Let, $X = x - 7$

$Y = y - 21$

where, x is no. of week employed

y is no. of car inspected

X	Y	X	Y	XY	X^2
2	13	–5	–8	40	25
7	21	0	0	0	0
9	23	2	2	4	4
1	14	–6	–7	42	36
5	15	–2	–6	12	4
123	21	5	0	0	25
Sum		–6	–19	98	94

$\Sigma X = -6$

$\Sigma Y = -19$

$\Sigma XY = 98$

$\Sigma X^2 = 94$

$n = 6$

Now, we know that;

$\Sigma Y = nA + B\Sigma X$ (1)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ (2)

So, $-19 = 6A - 6B$

$98 = -6A + 94B$

By solving, we get;

$A = -2.27$ $B = 0.90$

Hence,

$Y = -2.27 + 0.90X$

or, $(y - 21) = -22.7 + 0.90(x - 7)$

or, $y = 18.73 + 0.90x - 6.3$

or, $\boxed{y = 12.43 + 0.90x}$

Hence, the parameter of a simple linear regression is 12.43 and 0.90.

For $x = 8$,

$y = 12.43 + 0.90 \times 8 = 19.63 \approx 20$

Hence, 20 cars are inspected in 8 week of investigation.

11. **Define partial and multiple correlations with examples. Write down the properties of partial and multiple correlation.** **[2072 Ashwin]**

Solution: Refer to 7.4.2 (iii).

12. **The following data gives the experiences of machine operators in year and their performance as given by the number of good parts. Turned out per 100 pieces.** **[2072 Ashwin]**

Experience (x)	16	12	18	4	3	10	5	12
Performance (Y)	87	88	89	68	78	80	75	83

a) **Fit the regression equation of performance rating on experience and estimate the probable performance if an operator has 8 year experiences.**

b) **Calculate 95% confidence interval for regression coefficient (*i.e.*, slope)**

Solution:

Here;

Linear equation is

$Y = A + BX$

where, X is independent variable

Y is dependent variable

A and B are constant parameter

Now, using least square method,

$\Sigma Y = nA + B\Sigma X$ (1)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ (2)

Let, $X = x - 12$

$Y = y - 80$

where, x is experience in years

y is performance in out of 100 pieces.

And,

S.N.	x	y	X	Y	XY	X^2
1	16	87	4	7	28	18
2	12	88	0	8	0	0
3	18	89	6	9	45	36
4	4	68	–8	–12	96	64
5	3	78	–9	–2	18	81
6	10	80	–2	0	0	4
7	5	75	–7	–5	35	49
8	12	83	0	3	0	0
Sum			–16	8	231	252

$\Sigma X = -16$

$\Sigma Y = 8$

$\Sigma XY = 231$

$\Sigma X^2 = 252$

So, $8 = 8A - 16B$

$231 = -16A + 252B$

And, solving we get'

$A = 3.25$

$B = 1.12$

Hence,

$Y = 3.25 + 1.12X$

or, $(y - 80) = 3.25 + 1.12(x - 12)$

or, $y = 69.81 + 1.12x$

And,

i	x_i	y_i	$\hat{y}_i = 69.81 + 1.12x$	$(y_i - \hat{y}_i)^2$	$(x_i - \bar{x}_i)^2$
1	16	87	87.73	0.5329	36
2	12	88	83.25	22.5625	4
3	18	89	89.97	0.9409	64
4	4	68	74.29	39.5641	36
5	3	78	73.17	23.3289	49
6	10	80	81.01	1.0201	0
7	5	75	75.41	0.1981	25
8	12	83	83.25	0.0625	4
				88.18	218

So, standard error

$$S.E. = \sqrt{\frac{\Sigma \frac{(y_i - \hat{y}_i)^2}{n-2}}{\Sigma(x_i - \bar{x})^2}} = \sqrt{\frac{\frac{88.18}{(8-2)}}{218}} = 0.260$$

And, for confidence 95% and df = n – 2 = 8 – 2 = 6

From t-distribution table,

$t = 2.447$

Hence, interval will be for slope will be

Probable error = t × S.E = 2.447 × 0.260 = 0.636

So,

[1.12 – 0.636, 1.12 + 0.636]

= [0.484, 1.756]

OR

A household survey on monthly expenditure on food yield following data:

Monthly expenditure (100 Rs)	10	15	20	25	30	35	40
Monthly income (1000 Rs)	2	4	5	7	6	6	5
Size of the family	4	5	7	10	8	11	4

i) Obtain the multiple correlation coefficients.

ii) Find coefficient of multiple determination and interpret it.

[2072 Ashwin]

Solution: Similar to Q. No. 1 of 7.4.4.

13. Distinguish between correlation and regression. Write importance of coefficient of determination. [2073 Magh]

Solution: See the 7.4.4 and 7.4.2 (iii).

14. The trust of an engine (y) is a function of exhaust temperature (x) in when other important variable are held constant. Consider the following data: [2073 Magh]

Trust	4300	4650	3200	3150	4950	4010	3810	4500	3008
Temp.	1760	1652	1445	1390	1820	1665	1550	1700	1270

Solution:

Here;

For a line regression line,

i.e., $Y = A \times BX$

where, Y and X are dependent and independent variable respectively.

And, constant A and B are calculated by solving following equation

$\Sigma Y = nA + B\Sigma X$ (1)

$\Sigma XY = A\Sigma X + B\Sigma X^2$ (2)

Let, Y = (Thrust – 4010)

X (Temp. – 1665)

S.N.	Thrust	Temp.	Y	X	XY	X^2
1	4300	1760	290	125	36250	15625
2	4650	1652	640	–13	–8320	169
3	3200	1442	–810	–220	178200	48400
4	3150	1390	–860	–275	236500	75625
5	4950	1820	940	155	145700	24025
6	4010	1665	0	0	0	0
7	3810	1550	–200	–115	23000	13225
8	4500	1700	490	35	17150	1225
9	3008	1270	–1002	–395	395790	156025
Sum			–512	–703	1024270	334319

So, equation (1) and (2) becomes,

$-512 = 9A - 703B$

$1024270 = -703A + 334319B$

Solving; we get,

$A = 218.28$

$B = 3.53$

So, $Y = 218.28 + 3.53X$

or, (Thrust – 4010) = 218.28 + 3.53 (Temp – 1665)

or, Thrust = 4228.28 + 3.53 (Temp – 1665)

Now,

Temperature	Trust from data	Thrust from regression line
1760	4300	4563.63
1652	4650	4182.39
1445	3200	3451.68
1390	3150	3257.53
1820	4950	4775.43
1665	4010	4228.28
1550	3810	3822.33
1700	4500	4351.83
1270	3008	2833.93

15. Define the correlation coefficient and mention its important properties. What does the coefficient of determination measure?

[2073 Bhadra]

Solution: See the 7.4.2 (iii).

16. Past experience shows the following result of productivity per hectare with the respective used of chemical fertilizers and seeds. Fit the multiple linear regression equation of y and x_1 and x_2 from the given data:

Fertilizer kgs (x_1)	45	30	70	75	65	80
Seeds kgs (x_2)	2	18	3	2.5	2	3
Productivity kgs (y)	2000	2100	1800	1900	2400	2500

OR

A simply supported beam carries a concentrated loan x(kg) at its mid-point. The following table gives maximum deflection y(cm) corresponding to various value of x.

x	100	120	140	160	180	200
y	0.45	0.55	0.60	0.70	0.80	0.85

Calculate the two regression equations and estimate the value of maximum defection when x = 170 kg.

[2073 Bhadra]

Solution:

The multiple regression equation of y on x_1 and x_2 is given by

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad (1)$$

From above data, using calculator

$\Sigma x_1 = 365$, $\Sigma x_1^2 = 24075$

$\Sigma x_2 = 14.3$, $\Sigma x_2^2 = 35.49$, $\Sigma x_1 x_2 = 911.5$

$\Sigma y = 12700$, $\Sigma y x_2 = 30230$, $\Sigma y x_1 = 777500$

The normal equations are

$$\Sigma y = n\alpha + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2$$

i.e., $12700 = 6\alpha + 365\beta_1 + 14.3\beta_2 \quad (2)$

$$\Sigma x_1 y = \alpha \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2$$

i.e., $777500 = 365\alpha + 24075\beta_1 + 911.5\beta_2 \quad (3)$

$$\Sigma x_2 y = \alpha \Sigma x_2 = \beta_1 \Sigma x_1 x_2 + \beta_2 \Sigma x_2^2$$

i.e., $30230 = 14.3\alpha + 911.5\beta_1 + 35.49\beta_2 \quad (4)$

Solving equation (2), (3) and (4), we get

$\alpha = 2271.26$

$\beta_1 = 9.40$

$\beta_2 = -304.95$

Substituting these values in equation (1); we get,

$$y = 2271.26 + 9.40x_1 - 304.95x_2$$

OR

Given that;

x	y	xy	x^2	y^2
100	0.45	45	10000	0.2025
120	0.55	66	14400	0.3025
140	0.60	84	19600	0.36
160	0.70	112	25600	0.49
180	0.80	144	32400	0.67
200	0.85	170	40000	0.7225
900	3.95	621	142000	2.7175

The linear regression equation of y on x is

$$y = a + b_{yx}x$$

It's normal equations are

$\Sigma y = na + b_{yx}\Sigma x$

i.e., $3.95 = 6a + 900b_{yx}$ (1)

$\Sigma xy = a\Sigma x + b_{yx}\Sigma x^2$

i.e., $621 = 900a + 142000b_{yx}$ (2)

Solving (1) and (2); we get,

$a = \frac{1}{21}$ $b_{yx} = 0.004$

$\therefore$ $a = 0.047$

$\therefore$ The required equation is $y = 0.047 + 0.004x$

Again, the linear regression equation of x on y is

$x = a + b_{xy}y$

i.e., $900 = 6a + 3.95b_{xy}$ (3)

$\Sigma xy = a\Sigma y + b_{xy}\Sigma y^2$

i.e., $621 = 3.95a + 2.7175\, b_{xy}$ (4)

Solving (3) and (4); we get,

$a = -10.25$

$b_{xy} = 243.41$

$\therefore$ Therefore, the required equation is

$x = 243.41y - 10.25$

Now, maximum deflection when $x = 170$ kg is

$y = 0.047 + 0.004 \times 170$

$\therefore$ $y = 0.727$

17. Define correlation and regression with examples. Write down the properties of correlation coefficient and regression coefficient.

[2074 Bhadra]

Solution: See the 7.4 and 7.2.

Refer to Q. No. 3.

18. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Experience (X)	16	12	18	4	3	10	5	12
Performance (Y)	87	88	89	68	78	80	75	83

a) Find the regression equation of performance rating on experience and estimate the probable performance if an operator has 8 years experience.

b) Calculate coefficient of determination and interpret it.

[2074 Bhadra]

Solution: See the Q. No. 4.

19. Write down the properties of correlation coefficient and regression coefficient.

[2075 Baisakh]

Solution: See the Q. No. 3.

20. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Experience (X)	16	12	18	4	3	10	5	12
Performance (Y)	87	88	89	68	78	80	75	83

a) Find the regression equation of performance rating on experience and estimate the probable performance if an operator has 8 years experience.

b) Calculate coefficient of determination and interpret it.

[2075 Baisakh]

Solution: See the Q. No. 4.

21. What does a regression coefficient measure? Mention its important properties.

[2075 Bhadra]

Solution: See the 5.4.

22. Following data give experience (nearest year), performance rating and salary of eight engineers.

Experience (year)	16	12	18	4	3	10	5	12
Performance rating	87	88	89	68	78	80	75	83
Salary (000)	100	125	120	100	120	110	90	125

a) Find Possible multiple correlation coefficient and interpret the result.

[2075 Bhadra]

Solution:

We know,

$$\text{Correlation } (r) = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} \times \sqrt{\frac{1}{n}\Sigma y^2 - (\bar{y})^2}}$$

From above data, using calculator

$r_{12} = 0.8720$

$r_{13} = 0.3273$

$r_{23} = 0.5556$

Now, possible multiple correlation coefficient are

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{0.8720^2 + 0.3273^2 - 2 \times 0.8720 \times 0.3273 \times 0.5556}{1 - 0.5556^2}}$$

$$= 0.8922$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{0.8720^2 + 0.5556^2 - 2 \times 0.8720 \times 0.3273 \times 0.5556}{1 - 0.3273^2}}$$

$$= 0.9176$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{0.3273^2 + 0.5556^2 - 2 \times 0.8720 \times 0.3273 \times 0.5556}{1 - 0.8720^2}}$$

$$= 0.6417$$

Here, $R_{2.13} > R_{1.23} > R_{3.12}$.

Chapter 8

APPLICATION OF COMPUTER ON STATISTICAL DATA COMPUTING

SOLVED QUESTIONS [FOR BCE/BME/BIE/BAM]

1. **The following table shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic.**

7	10	12	4	8	7	3	8	5
12	11	3	8	1	1	13	10	4
4	5	5	8	7	7	3	2	3
8	13	1	7	17	3	4	5	5
3	1	17	10	4	7	7	11	8

a) **Find sample mean, sample variance and sample standard deviation.**

b) **Compute a value that measures the amount of variability relative to the value of mean.** [8] [2071 Bhadra]

Solution:

I am using calculator "CASIO ⇒ FX - 991 EX" to solve the question.

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Select (0) to (9)

(2) ↲ (I choose)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

then; (1) ↲

Step 4: Data entry

Enter data in the box

(7) ↲

(10) ↲

(12) ↲

⋮

(8) ↲

After entering all data

AC ↲ (to save and exit)

Step 5: Result calculation

Option ↲

(2) ↲

All result appear on screen

Select desired results

a) i) Sample mean $(\bar{X}) = \mu = 6.71$

ii) Sample variance $(S_x^2) = 16.21$

iii) Sample S.D. $(S_x) = 4.03$

b) Coefficient of variation (C.V.) $= \frac{S_x}{\bar{X}} \times 100\%$

$$= \frac{4.03}{6.71} \times 100\% = 60.06\%$$

List of formulas

Mean $(\bar{X}) = \frac{\Sigma X}{N}$

Sample variance $(S_x^2) = \frac{\Sigma(X - \bar{X})^2}{N - 1}$

Sample S.D. $(S_x) = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$

2. **An article in computer and industrial engineering describes the time-failure data (in hours) for jet engines some of the datas are presented below.**

Engine	Failure time	Engine	Failure time
1	150	13	213
2	291	14	271
3	93	15	197
4	53	16	200
5	2	17	262
6	65	18	255
7	183	19	286
8	144	20	206
9	223	21	179
10	197	22	232
11	187	23	165
12	197	24	155
		25	203

a) **Find the sample mean, sample variance and sample standard deviation.**

b) **Exclude the lowest 6 data and compute (a) again.**

c) **Analyze the result (a) and (b). Draw your conclusion.**

[8] [2071 Magh]

Solution:

I am using calculator "CASIO FX-991 EX" to solve this question.

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

Finally

(1) ↲

Step 4: Data entry

Enter data in the box

1	150 ↲
2	291 ↲
⋮	⋮
24	155 ↲
25	203 ↲

After entering all data

AC ↲ (to save and exit)

Step 5: Result calculation

Option ↲

(2) ↲

All result appear on screen. Select desired results

a) Sample mean $(\bar{X}) = 184.36$

Sample variance $(S_X^2) = 5147.407$

Sample S.D. $(S_X) = 71.74$

Coefficient of variation $(C.V.)_1 = \frac{S_X}{\bar{X}} \times 100\%$

$$= \frac{71.74}{184.36} \times 100\% = 38.91\%$$

b) The datas are saved in calculator. So, follow the procedure

Option ↲

(1) ↲ (Select type)

(1) ↲ (1-variable)

All datas appear. So erase last 6 datas by using cursor button and DEL

To save and exit,

AC ↲

To get result again

Option ↲

(2) ↲ (1-variable calculation)

Sample mean $(\bar{X}) = 182.58 = \frac{\Sigma X}{N}$

Sample variance $(S_X^2) = 6618.14 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$

Sample S.D. $(S_X) = 81.35 = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$

Coefficient of variation $(C.V.)_2 = \frac{S_X}{\bar{X}} \times 100\% = 44.55\%$

c) Hence out of (a) and (b), set (A) datas are more uniform.

3. **A study was done on a diesel powered light duty pickup truck to see if humidity influence emission of nitrous oxide. Emission measurements were taken at different times, with varying experimental conditions. The datas are as follows:**

Nitrous oxide	Humidity
0.90	72.4
0.91	41.6
0.96	34.3
0.89	35.1
1.00	10.7
1.10	12.9
1.15	8.3
1.03	20.1
0.77	72.2
1.07	24
1.07	23.2
0.94	47.4
1.10	31.5
1.10	10.6
1.10	11.2
0.91	73.3
0.87	75.4
0.78	96.6
0.82	107.4
0.95	54.9

a) Find mean and variance of given data.

b) Calculate degree of relationship between them.

c) Calculate coefficient of determination and interpret the given data. **[8] [2072 Ashwin]**

Solution:

I am using calculator "CASIO FX-991 EX" to solve this question.

Let Nitrous oxide → X

Humidity → Y

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have two variables X and Y in this question. So, open statistics and choose y = a + bx.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

(2) ↲ (y = a + bx)

Step 4: Data entry

Enter data in the box

	X	Y
1	0.90 ↲	72.4 ↲
2	0.91 ↲	41.6 ↲
⋮	⋮	⋮
19	0.82 ↲	107.4 ↲
20	0.95 ↲	54.9 ↲

After entering all data

AC ↲ (to save and exit)

Step 5: Result calculation

Option ↲

(2) ↲

All result appear on screen. Select desired results

a) $(\bar{X}) = 0.971$ [X ⇒ Nitrous oxide, Y ⇒ Humidity]

$(S_x^2) = 0.0133$

$(\sigma_x^2) = 0.0126$

$\bar{y} = 43.155$

$S_y^2 = 927.88$

$\sigma_y^2 = 881.490$

b) To calculate degree of relationship:

Option ↲

(3) ↲ (regression calculation)

r = –0.877

a = 268.04 and b = –231.61

c) Coefficient of determination = $r^2 = 0.769 = 76.9\%$

Hence 76.9% of the total variation in the independent variable Y(Humidity) has been explained by independent variable X(Nitrous oxide). The remaining 23.1% is due to effect of other factors.

4. Following data reveals the 50 samples of speed (km/hrs) of vehicles travelling in an intersection of a busy road.

54	65	58	59	53	67	68	59	64	68
69	72	75	48	44	42	53	52	51	49
48	46	54	52	50	51	64	45	42	49
54	71	72	58	59	52	55	55	56	54
54	68	66	65	67	67	65	65	64	64

i) Find average speed of the vehicles.

ii) Find the sample unbiased standard deviation and standard error of speed of the vehicles.

iii) Also test consistency by applying suitable formula.

[8] [2072 Magh]

Solution:

I am using calculator "CASIO FX-991 EX" to solve this problem.

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

Finally

(1) ↲

Step 4: Data entry

Enter data in the box

	X
1	54 ↲
2	65 ↲
⋮	⋮
49	64 ↲
50	64 ↲

After entering all data

AC ↲ (to save and exit)

Step 5: Result calculation

Option ↲

(2) ↲

i) Average speed $(\bar{X}) = 58.04$

ii) Simple S.D. $(S_X) = 8.720$

Population S.D. $(\sigma_x) = 8.632$

$$S.E.\ (\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{8.632}{\sqrt{50}} = 1.22$$

iii) Coefficient of variation (C.V.) $= \frac{\sigma_x}{\bar{X}} \times 100\% = \frac{8.632}{58.04} \times 100\% = 14.87\%$

Formulas

$$(\bar{X}) = \frac{\Sigma X}{n}$$

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

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

5. Following data reveals the sample of 27 pairs of observation (X, Y) drawn from large population.

X	46	55	49	57	56	46	66	47	51
Y	27	37	24	43	43	50	49	48	29
X	61	56	68	68	48	58	68	45	50
Y	46	43	40	18	39	32	48	18	45
X	59	45	66	62	57	57	57	69	47
Y	32	26	27	29	47	37	37	27	30

Find the

i) Sample mean for each variable X and Y?

ii) Which series is more consistent and why?

iii) Standard error of the difference of mean in the population.

iv) Find the coefficient of Karl Pearson correlation.

[8] [2073 Baishakh]

Solution:

I am using calculator "CASIO FX-991 EX" to solve this question.

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have two variables X and Y in this question. So, open statistics and choose y = a + bx.

Menu (Menu box appear)

Choose Statistics option using arrow keys.

(→) (↓)

(2) ↲ (y = a + bx)

Step 4: Data entry

Enter data in the box

	X	Y
1	46 ↲	27 ↲
2	55 ↲	37 ↲
⋮	⋮	⋮
26	69 ↲	27 ↲
27	47 ↲	30 ↲

Step 5: Result calculation

Option ↲

(2) ↲

i) Mean of X $(\bar{X}) = 56.074$

Mean of Y $(\bar{Y}) = 35.96$

S.D. of X $(\sigma_X) = 7.836$

S.D. of Y $(\sigma_Y) = 9.488$

ii) $(C.V.)_X = \frac{\sigma_X}{\bar{X}} \times 100\% = \frac{7.836}{56.074} \times 100\% = 13.97\%$

$(C.V.)_Y = \frac{\sigma_Y}{\bar{Y}} \times 100\% = \frac{9.488}{35.96} \times 100\% = 26.38\%$

$\therefore (C.V.)_X < (C.V.)_Y$. So, X is more consistent.

iii) Standard error of the difference in mean

$$S.E.(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}} = \sqrt{\frac{7.836^2}{27} + \frac{9.488^2}{27}} = 2.368$$

iv) To find coefficient of Karl Pearson correlation

Option ↲

(3) ↲ (regression calculation)

a = 30.85

b = 0.091

r = 0.075

6. A container of car antifreeze is supposed to hold 3785 ml of liquid realizing that fluctuations are inevitable, the quality control manager wants to be quite sure that the S.D. is less than 30 ml. Otherwise, some containers would overflow while others will not have enough of the coolants. She selects a simple random sample, with the results given here use these sample results to construct 99% confidence interval for true value of S. Using your calculator

a) Calculate the sample mean

b) Calculate the sample standard deviation

c) Does this confidence interval suggest that the fluctuations are at an acceptable level?

d) Observing your result draw your conclusion.

3761	3861	3769	3772	3675	3861	3888
3819	3788	3800	3720	3748	3787	3753
3821	3811	3740	3740	3839	3445	3351
3332	323	3423	3312	3452		

[8] [2073 Magh]

Solution: Here,

$(1 - \alpha) = 99\%$

or, $\alpha = 1\%$

or, $\alpha = 0.01$

$\therefore \frac{\alpha}{2} = 0.005$

Population mean (μ) = 3785 ml

Population S.D. (σ) = 30 ml

I am using calculator "CASIO FX-991 EX" to solve this question.

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

Finally

(1) ↲

Step 4: Data entry

Enter data in the box

1	3761 ↲
2	3861 ↲
⋮	⋮
25	3312 ↲
26	3452 ↲

Step 5: Result calculation

Option ↲

(2) ↲

a) Sample mean ($\bar{X}$) = 3561.19

b) Sample S.D. (S) = 684.4

c) $\mu = \bar{X} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}$

$= 3561.19 \pm Z_{0.005} \cdot \sqrt{\frac{30^2}{50}}$

$= 3561.19 \pm 2.576 \cdot \sqrt{\frac{900}{50}}$

$= 3561.19 \pm 10.929$

$= 3557.24, 3579.1$

$\therefore$ C.I. is (3557.24, 3579.1)

The population mean (μ) = 3785 does not lies in the confidence interval. So, this 1% level of significance does not give acceptable value.

d) The data 323 might have been mid-read because of which we got vast difference between population and sample data results.

Note: Value of $Z_{0.005}$ can be directly found by calculator.

Mode + (7) ↲ [Distribution]

(3) ↲ [Inverse normal]

Fill area = 0.005
$\sigma = 1$
$\mu = 0$ } ↲

$\therefore$ $Z_{0.005} = -2.576$

7. As a part of study monitoring acid rain, measurements of sulphates deposits (kg/hectare) are recorded for different locations on eastern terai of Nepal. The results are listed in following table for 15 recent and consecutive years.

Acid rain: Sulphate deposited (kg/hectare)

Year	Location 1 (X)	Location 2 (Y)	Location 3 (Z)
1	11.94	13.09	7.96
2	11.28	10.88	12.84
3	10.38	12.19	7.38
4	8.00	10.75	7.26
5	12.12	17.21	10.12
6	10.27	10.26	8.89
7	14.80	15.49	11.60
8	13.52	11.61	9.02
9	10.55	10.53	7.78
10	9.81	12.50	8.70
11	11.27	9.94	10.50
12	12.12	11.21	9.95
13	11.68	9.71	15.59
14	11.77	9.37	10.54
15	17.29	13.87	13.64

a) Find sample mean, sample standard deviation and coefficient of variation for sulphate deposits of each location.

b) Give your conclusion about variability and uniformity from the analysis.

[8] [2074 Bhadra]

Solution:

I am using calculator "CASIO FX-991 EX" to solve this question.

(A) for location 1 (X)

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

Finally

(1) ↲

Step 4: Data entry

Enter data in the box

	X
1	11.94 ↲
2	11.28 ↲
⋮	⋮
14	11.77 ↲
15	17.29 ↲

Step 5: Result calculation

Option ↲

(2) ↲

a) i) Sample mean ($\bar{X}$) = 11.786

Sample S.D. (S_X) = 2.191

Population S.D. (σ_X) = 2.116

Coefficient of variation $(C.V.)_X = \frac{\sigma_X}{\bar{X}} \times 100\% = \frac{2.116}{11.786} \times 100\%$

$= 17.85\%$

(B) For location 2 (Y)

Follow same steps as for location 1(X) and write results.

ii) Sample mean $(\bar{Y}) = 11.907$

Sample S.D. $(S_Y) = 2.225$

Population S.D. $(\sigma_Y) = 2.15$

Coefficient of variation $(C.V.)_Y = \frac{\sigma_Y}{\bar{Y}} \times 100\%$

$$= \frac{2.15}{11.907} \times 100\% = 18.06\%$$

(C) for location 3(Z)

Following similar steps, we find results as

iii) Sample mean $(\bar{Z}) = 10.118$

Sample S.D. $(S_Z) = 2.432$

Population S.D. $(\sigma_Z) = 2.350$

Coefficient of variation $(C.V.)_Z = \frac{\sigma_Z}{\bar{Z}} \times 100\%$

$$= \frac{2.350}{10.118} \times 100\% = 23.226\%$$

b) From calculation in (a)

$(C.V.)_X < (C.V.)_Y < (C.V.)_Z$

Hence data of location 1(X) is very uniform. While data Z is very variable.

8. **The height of male and female students are given below:**

Height	Sex of person	
	Male	Female
145 – 150	0	3
150-155	2	8
155-160	6	15
160-165	17	12
165-170	21	4
170-175	10	0
175-180	2	0
180-185	2	0

a) **Calculate mean heights for male and female students.**

b) **Calculate sample standard deviation and sample variance of given data.**

c) **Which data for height is consistent.**

Solution:

Here;

Height (C.I.)	$X = \left(\frac{U+L}{2}\right)$
145 – 150	147.5
150-155	152.5
155-160	157.5
160-165	162.5
165-170	167.5
170-175	172.5
175-180	177.5
180-185	182.5

I am using calculator "CASIO FX-991 EX" to solve this question.

The question is in continuous series and we have found mid-point as x.

Now, turn ON frequency in calculator.

Steps

Shift + Mode

(↓) (Down cursor)

(3) (Statistics)

Now option appears

Frequency?

1 : ON

2 : Off

(1) ↲ (to turn on frequency)

(A) For male

Now follow the steps

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

=

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

Finally

(1) ↲

Step 4: Data entry

Enter data in the box

	X	F
1	147.5 ↲	0 ↲
2	152.5 ↲	2 ↲
⋮	⋮	⋮
74	177.5 ↲	2 ↲
78	182.5 ↲	2 ↲

Step 5: Result calculation

Option ↲

(2) ↲

(a-i) Mean height for male ($\bar{X}$) = 166.25

(b-i) Sample S.D. (S_X) = 6.218

Sample variance (S_X^2) = 38.66

Population S.D. (σ_X) = 6.166

(c-i) $C.V.)_{male} = \frac{\sigma_X}{\bar{X}} \times 100\% = \frac{6.166}{166.25} \times 100\% = 3.7\%$

(B) for female

Repeat same steps as done for male students

(a-ii) Mean height for female ($\bar{Y}$) = 158.214

(b-ii) Sample S.D. (S_Y) = 5.36

Sample variance (S_Y^2) = 28.746

Population S.D. (σ_Y) = 5.297

(c-ii) $C.V.)_{female} = \frac{\sigma_Y}{\bar{Y}} \times 100\% = \frac{5.297}{158.214} \times 100\% = 3.348\%$

Here, $(C.V.)_{male} > (C.V.)_{female}$

So, data for female is consistent.

9. The alternating current (AC) breakdown voltage of an insulating liquid indicates its dielectric strength. The article "Testing practices for all AC breakdown voltage testing of insulation liquids" (IEEE: electrical insulation magazine, 1995 : 21–26) gave the accompanying samples observation on breakdown voltage (KV) of a particular circuit under certain conditions.

62	64	46	57
50	62	55	55
53	50	53	53
57	68	54	59
41	54	52	53
53	55	47	52
55	57	47	50
61	50	55	55
59	55	57	60
64	50	48	50
50	56	63	56
53	55	57	58

a) Constant a box plot of the data and comment on features.

b) Calculate and interpret 95% CI for true average breakdown voltage μ. Does it appear that μ has been precisely estimated? Explain.

Solution:

I am using calculator "CASIO FX-991 EX" to solve this problem.

Step 1: Reset calculator to remove temporarily saved datas

Shift + (9)

(1)

Step 2: Fix decimal

Shift + Menu

(3) ↲

(1) ↲

Option select (0) to (9) appears

(2) ↲ (My choice)

Step 3: We have only one variable x in this question. So, open statistics and choose 1-variable.

Menu (Menu box appear)

Choose Statistics option using arrow keys

(→) (↓)

or

(6) ↲

Finally

(1) ↲

Step 4: Data entry

Enter data in the box

	x
1	62 ↲
2	64 ↲
⋮	⋮
47	57 ↲
48	58 ↲

Step 5: Result calculation

Option ↲

(2) ↲

Mean ($\bar{X}$) = 54.708

Sample S.D. (S_X) = 5.23 ≃ σ_X (for large sample)

Maximum value (L) = 68

Minimum value (S) = 41

Q_1 = 51

Median = 55

Q_3 = 57

Inter quartile range (IQR) = $Q_3 - Q_1$ = 6

a)

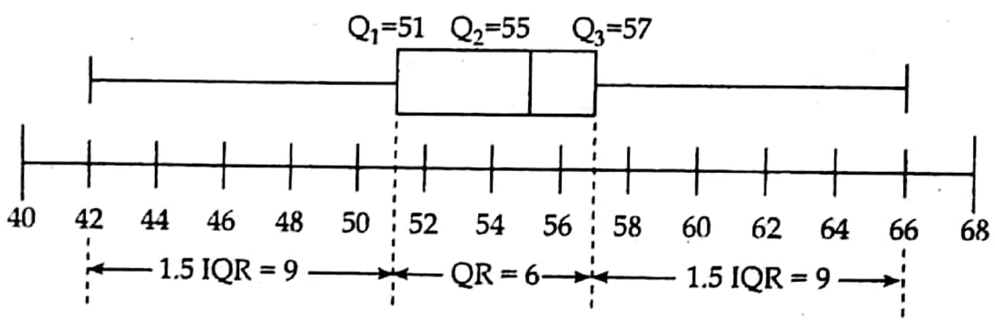


Fig: Box plot

b) $$\mu = \bar{X} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{N}} \qquad [(1-\alpha) = 95\% = 0.95, \alpha = 0.05]$$

$$= 54.708 \pm Z_{0.025} \times \sqrt{\frac{5.23^2}{48}} = 54.708 \pm 1.96 \times \sqrt{\frac{5.23^2}{48}}1$$

$$= 54.708 \pm 1.479 = (53.229, 56.187)$$

∵ The C.I. is (53.229, 56.187)

Here, population mean (μ) is not given in question. Hence, we can't comment on the result of μ, we found.

Note: $Z_{\alpha/2}$ can be directly found by calculator (FX – 991 EX)

Mode + (7) ↲ [Distribution]

(3) ↲ [Inverse normal]

$$\left.\begin{array}{l}\text{Fill area} = 0.025 \\ \sigma = 1 \\ \mu = 0\end{array}\right\} ↲$$

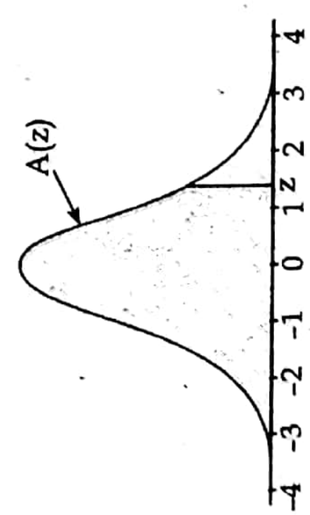
∴ $Z_{0.025} = -1.96$

TABLE A.1

Cumulative Standardized Normal Distribution

A(z) is the integral of the standardized normal distribution from −∞ to z (in other words, the area under the curve to the left of z). It gives the probability of a normal random variable not being more than z standard deviations above its mean. Values of z of particular importance:

z	A(z)	
1.645	0.9500	Lower limit of right 5% tail
1.960	0.9750	Lower limit of right 2.5% tail
2.326	0.9900	Lower limit of right 1% tail
2.576	0.9950	Lower limit of right 0.5% tail
3.090	0.9990	Lower limit of right 0.1% tail
3.291	0.9995	Lower limit of right 0.05% tail



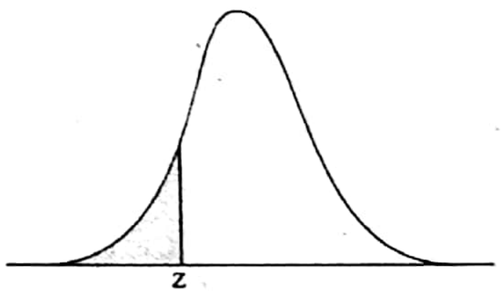
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

TABLE A.2

Standard Normal Cumulative Probability Table

Cumulative probabilities for NEGATIVE z-values are shown in the following table:



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183

-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

TABLE A.3
t Distribution: Critical Values of t

Degrees of freedom	Two-tailed test: / One-tailed test:	*Significance level* 10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850

Degrees of freedom	Two-tailed test: One-tailed test:	10% 5%	5% 2.5%	2% 1%	1% 0.5%	0.2% 0.1%	0.1% 0.05%
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496

Degrees of freedom	Two-tailed test: One-tailed test:	10% 5%	5% 2.5%	2% 1%	1% 0.5%	0.2% 0.1%	0.1% 0.05%
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

TABLE A.4

F Distribution: Critical Values of F (5% significance level)

v_2 \ v_1	2	1	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

v_2 \ v_1	2	1	3	4	5	6	7	8	9	10	12	14	16	18	20
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.4 (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

v_2 \ v_1	25	30	35	40	50	60	75	100	150	200
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.4

F Distribution: Critical Values of F (1% significance level)

v_2 \ v_1	2	1	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

v_2 \ v_1	2	1	3	4	5	6	7	8	9	10	12	14	16	18	20
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

v_2 \ v_1	2	1	3	4	5	6	7	8	9	10	12	14	16	18	20
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.4 (continued)

F Distribution: Critical Values of F (1% significance level)

v_2 \ v_1	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

v_2 \ v_1	25	30	35	40	50	60	75	100	150	200
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

v_2 \ v_1	25	30	35	40	50	60	75	100	150	200
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.5

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