PONDICHERRY UNIVERSITY

(A Central University)

DIRECTORATE OF DISTANCE EDUCATION

MASTER OF COMMERCE

M.Com – Second Year

Course Code: 59 **Paper Code: MCOM2004**

STATISTICAL ANALYSIS

DDE – WHERE INNOVATION IS A WAY OF LIFE

PONDICHERRY UNIVERSITY

(A Central University)

DIRECTORATE OF DISTANCE EDUCATION

MASTER OF COMMERCE

M.Com – Second Year

Course Code:59

Paper Code:MCOM2004

STATISTICAL ANALYSIS

STATISTICAL ANALYSIS

Authors:

1. **Unit I & II** **Dr.Kiruthika**
Assistant Professor,
Department of Statistics,
Pondicherry University

2. **Unit III, IV & V** **Dr. R.Vishnu Vardhan**
Assistant Professor,
Department of Statistics,
Pondicherry University

STATISTICAL ANALYSIS

TABLE OF CONTENTS

UNIT	TITLE	PAGE NO.
I	Univariate Analysis	4
II	Probability Theory	32
III	Correlation and Regression Analysis	58
IV	Statistical Estimation and Testing	72
V	Non Parametric Tests	97

M.Com – 2nd Year **PAPER CODE: MCOM 2004**

STATISTICAL ANALYSIS

UNIT I

Univariate Analysis: An overview of Central Tendency, Dispersion and Skewness. Sampling and Data Collection: Sampling and Sampling (Probability and Non-Probability) methods; Sampling and non-sampling errors; Law of Large Number and Central Limit Theorem; Sampling distributions and their characteristics.

UNITII

Probability Theory: Probability – Classical, Relative, and Subjective Probability: Addition and Multiplication Probability Models; Conditional Probability and Baye’s Theorem Probability Distributions: Binomial, Poisson, and Normal Distributions their characteristics and applications.

UNIT III

Correlation and Regression Analysis: Two Variables Case

UNIT IV

Statistical Estimation and Testing: Point and interval estimation of population mean, proportion and variance; Statistical testing – hypotheses and errors; Sample size; Large and small sampling tests –Z tests, T tests, and F tests.

UNIT V

Non Parametric Tests: Chi-square tests; Sign tests Wilcoxon Signed – Rank tests; Wald – Wolfowitz tests; Kruskal – Wallis tests.

Note: Question Paper Shall covers 20% Theory and 80% Problems

REFERENCES:

Arora & Arora, Statistics for Management, Sultan Chand, New Delhi, 2009

Gupta S.P., Statistical Methods, Sultan Chand, New Delhi 2009

Levin, Richard I. and David S Rubin, Statistics for Management, Prentice Hall, Delhi 2009

Sonia Taylor,Business Statistics, Palgrave Macmillan, 2009.

Qaziahmed, Zubuirkhan, Shadabahmedkha, Numerical and Statistical Techniques Ane, 2010

UNIT I

UNIVARIATE ANALYSIS

1.1 – Univariate Analysis

1.2 – Data Collection and Sampling Methods

1.3 – Sampling Distributions

1.1 Univariate Analysis

Statistics is collection, analysis and interpretation of numerical data. Statistics can be broadly classified as descriptive statistics and inferential statistics. Descriptive statistics includes collecting, summarizing, presenting and analyzing a data. The data can be presented as diagrams such as bar chart, pie chart etc. or graphical representations such as histogram, frequency curves etc. The data can also be presented as discrete or continuous tables. More information about data can be obtained by finding the average (mean) and the variation in the data specified by the standard deviation. Measures of central tendency and dispersion are useful in describing the numerical data more precisely. The first section (Lesson 1.1) gives a brief overview of various measures of central tendency, dispersion and skewness.

Measures of Central Tendency

Measures of central tendency give the average value of the data. The average values are also known as measure of central tendency as the values tend to lie centrally within a data set arranged according to magnitude. Measures of central tendency should be rigidly defined, easy to calculate, based on all observations and suitable for further treatment. Measures of central tendency are classified into mathematical averages like Arithmetic Mean, Geometric Mean, Harmonic Mean and positional averages namely, Median and Mode.

Arithmetic mean is most commonly used measure of central tendency. It is rigidly defined and based on all the observations. It is affected by the extreme values. In case of extreme values, arithmetic mean does not represent the distribution properly.

Median divides the distribution into two equal parts. It is the positional average. It is the central value such that the number of observations above and below it is equal. If the number of observations is odd, then the median is the middle value in ordered (ascending order) list of observations. If the number of observations is even then the median is the average of the two middle values. Median is not affected by extreme values and it is located by inspection. Median can also be located graphically by drawing ogives (cumulative frequency curves).

Quartile divides the distribution into four equal parts. The first quartile (Q_1) or lower quartile has 25% of observations of the distribution below it and 75% of items are greater than it. The second quartile (Q_2) is the median and the third quartile (Q_3) has 75% of items of the distribution below it. Deciles divide the series into ten equal parts. There are nine deciles and fifth decile is the median. Percentile divides the distribution into 100 equal parts. There are 99 percentiles (P_1, P_2, ..., P_{99}).

Mode is a value which occur most number of times (highest frequency). The mode may not be unique. Sometimes it may not exist. A distribution having one mode is known as unimodal. The empirical relation between the arithmetic mean, median and mode is given by **Mean – Mode = 3 (Mean – Median)**

Geometric mean is a special mean which gives more weights to small items. It is used to find the rate of population growth and the rate of interest. It is also used in the construction of index numbers.

Harmonic mean is also rigidly defined and not much affected by fluctuations of sampling. It also gives more importance to small items and is useful only when small items are to be given very high weightage.

The formulae for Arithmetic Mean, Geometric Mean and Harmonic Mean for individual series (raw data), discrete frequency distribution and continuous frequency distribution are given in Table 1. Table 2 gives the formulae for calculating Median and Mode for continuous frequency distribution. The steps for calculating mode for discrete series are given below.

1. Prepare a grouping table with 6 columns with 'x' values in the margin.
2. Write the frequencies corresponding to 'x' values in column 1.
3. In column 2, the frequencies are grouped in twos.
4. In column 3, the frequencies are grouped in twos excluding the first frequency.
5. In column 4, the frequencies are grouped in threes.
6. In column 5, the frequencies are grouped in threes excluding the first frequency.
7. In column 6, the frequencies are grouped in twos excluding the first two frequencies.
8. Mark the highest frequency in each column.
9. Mode is the 'x' value having highest frequency, obtained from the analysis table.

Table 1: Formulae for Measures of Central Tendency

Measures of central tendency	Individual Series (Ungrouped data)	Discrete Frequency Distribution	Continuous Frequency Distribution
Arithmetic Mean	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} f_i m_i}{N}$
Geometric Mean	$G.M. = antilog\left[\frac{\sum_{i=1}^{n} log x_i}{n}\right]$	$G.M. = antilog\left[\frac{\sum_{i=1}^{n} f_i log x_i}{N}\right]$	$G.M. = antilog\left[\frac{\sum_{i=1}^{n} f_i log m_i}{N}\right]$
Harmonic Mean	$H.M. = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$	$H.M. = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{x_i}}$	$H.M. = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{m_i}}$

Table 2: Formulae for Median and Mode

Measures of central tendency	Continuous Frequency Distribution
Median	$Median = L + \left(\frac{\frac{N}{2} - c.f}{f}\right) \text{x } i$ where L = Lower limit of median class; f = frequency of median class $c.f$ = cumulative frequency of the class preceding median class i = class interval of median class
Mode	$Mode = L_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \text{x } i$ where L_1 = Lower limit of modal class; f_1 = frequency of modal class f_0 = Frequency of the class preceding the modal class f_2 = Frequency of the class succeeding the modal class i = class interval

1. **Individual Series**: Let $x_1, x_2, \dots, x_n$ be set of n observations.

Example 1: Calculate Arithmetic mean (A.M.), Median, Geometric Mean (G.M.) and Harmonic Mean (H.M.) for the following data.

87, 63, 91, 72, 93, 79, 83, 94, 75

(i) A.M. = (87+63+91+72+93+79+83+94+75+88)/9 = 81.89

(ii) Arrange the values in ascending order : 63, 72, 75, 79, **83**, 87, 91, 93, 94

Median = 83

(iii) G.M. = antilog[(log 87+ ... + log 75)/9] = 81.254

(iv) H.M. = 1/ [1/9 *(1/87 + ... + 1/75)] = 80.5903

Example 2: Find the median and mode for the data 3, 5, 2, 6, 5, 9, 8, 6, 3, 5.

Arrange the values in the ascending order: 2, 3, 3, 5, 5, 5, 6, 6, 8, 9

Median = (5 + 5) / 2 = 5

Mode = number occurring most frequently = 5

2. **Discrete Frequency distribution**: Let $x_1, x_2, \dots, x_n$ occur $f_1, f_2, \dots, f_n$ times respectively and $N = \sum_{i=1}^{n} f_i$.

Example 3: Calculate mean weekly wages of the employees of a company from the following frequency distribution.

Wages (Rs.) (x_i)	2550	2650	2750	2850	2950	3050	3150
Number of employees (f_i)	8	10	16	14	10	5	2
$f_i x_i$	20400	26500	44000	39900	29500	15250	6300

$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$ = 181850/65 = Rs. 2797.69

Example 4: Calculate Q_1, Q_2 and Q_3 for the following series.

Size of shoes	4	4.5	5	5.5	6	6.5	7	7.5	8
Frequency	10	18	22	25	40	15	10	8	7
Cumulative frequency	10	28	50	75	115	130	140	148	155

Steps: 1. Arrange the data in ascending or descending order

2. Find the cumulative frequencies.

3. Median = Size of (N+1)/2 th item

$N = \sum_{i=1}^{n} f_i$ = 155

Q_1 = Size of (N+1)/4 th item = Size of 39th item = 5

Q_2 = Size of (N+1)/2 th item = Size of 78th item = 6

Q_3 = Size of 3*(N+1)/4 th item = Size of 117th item = 6.5

Example 5: Calculate mode for the following frequency distribution.

Value (x_i)	100	110	120	130	140	150	160	170	180
Frequency (f_i)	10	12	15	19	20	8	4	3	2

Grouping Table

Value (x)	**1**	**2**	**3**	**4**	**5**	**6**
100	10	22		37		
110	12		27		**46**	
120	15	**34**				**54**
130	19		**39**	**47**		
140	**20**	28			32	
150	8		12			15
160	4	7		9		
170	3		5			
180	2					

Analysis Table

Column No.	**110**	**120**	**130**	**140**	**150**
1				/	
2		/	/		
3			/	/	
4			/	/	/
5	/	/	/		
6		/	/	/	
Total Frequency	**1**	**3**	**5**	**4**	1

The mode is 130 as this value occurs five times. But through inspection, the mode is 140.

Example 6: Calculate Geometric mean and Harmonic mean from the following data.

Size of items (x_i)	6	7	8	9	10	11
Frequency (f_i)	4	6	9	5	2	8

$G.M. = antilog\left[\frac{\sum_{i=1}^{n} f_i log x_i}{N}\right]$ = antilog (31.41336/34) = antilog (0.923922) = 8.3930

$H.M. = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{x_i}}$ = 34 / 4.131637 = 8.23

3. **Continuous Frequency Distribution**: Let $m_1, m_2, \dots, m_n$ be mid-points of n class intervals with frequencies $f_1, f_2, \dots, f_n$ and $N = \sum_{i=1}^{n} f_i$.

Example 7: Calculate arithmetic mean, median, mode, geometric mean and harmonic mean for the following frequency distribution.

Grade	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of Students	1	3	11	21	43	32	9

Grade	**No. of Students (f_i)**	***c.f.***	***m***			

Grade	**No. of Students (f_i)**	***c.f.***	m_i	$f_i m_i$	$f_i (log m_i)$	f_i / m_i
30-40	1	1	35	35	1.544068	0.02857143
40-50	3	4	45	135	4.959638	0.06666667
50-60	11	15	55	605	19.14399	0.2
60-70	21	36	65	1365	38.07118	0.32307692
70-80	43	79	75	3225	80.62763	0.57333333
80-90	32	111	85	2720	61.74141	0.37647059
90-100	9	120	95	855	17.79951	0.09473684
	120			**8940**	**223.8874**	**1.66285578**

Arithmetic Mean : $\bar{x} = \frac{\sum_{i=1}^{n} f_i m_i}{N}$ = 8940/120 = **74.5**

$\boldsymbol{Median} = L + \left(\frac{\frac{N}{2} - c.f}{f}\right)$ x i = 70 + $\left(\frac{60-36}{43}\right)$ x 10 = 70 + 5.581 = **75.581**

$\boldsymbol{Mode} = L_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)$ x i = 70 + $\left(\frac{43-21}{86-21-32}\right)$ x 10 = 70 + 6.67 = **76.67**

$\boldsymbol{G.M.} = antilog\left[\frac{\sum_{i=1}^{n} f_i log m_i}{N}\right]$ = antilog(223.8874/120) = antilog(1.865728) = **73.405**

$\boldsymbol{H.M.} = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{m_i}}$ = 120 / 1.66285578 = **72.165**

Example 8: Determine the quartiles and the median from the following table.

Income (Rs.)	No. of persons
Below 30	69
Below 40	236
Below 50	443
Below 60	508
Below 70	566
Below 80	593
Below 90	603

Income	f	*c.f*
Below 30	69	69
30-40	167	236
40-50	207	443
50-60	65	508
60-70	58	566
70-80	27	593
80-90	10	603

Median = size of N/2th item = 301.5th item

Median class is 40-50,

$$\boldsymbol{Median} = L + \left(\frac{\frac{N}{2} - c.f}{f}\right) \text{x } i = 40 + \left(\frac{301.5-236}{207}\right) \text{ x } 10 = 43.16$$

Q_1 = size of N/4th item = 150.75th item

First Quartile class is 30-40

$$\boldsymbol{Q_1} = L + \left(\frac{\frac{N}{4} - c.f}{f}\right) \text{x } i = 30 + \left(\frac{150.75-69}{167}\right) \text{ x } 10 = 34.895$$

Q_3 = size of 3N/4th item = 452.25th item

First Quartile class is 50-60

$$\boldsymbol{Q_3} = L + \left(\frac{\frac{3N}{4} - c.f}{f}\right) \text{x } i = 50 + \left(\frac{452.25-443}{65}\right) \text{ x } 10 = 51.423$$

Example 9: A time and motion study of a certain operation shows the following distribution for 100 workers. Calculate mean, median and mode of the distribution.

Time (minutes)	10-15	15-20	20-25	25-30	30-35	35-40	40-45
No. of Workers	8	14	18	25	15	14	6

Time	f_i	*c.f.*	m_i	$d_i=(m_i\text{-}27.5)/5$	f_id_i
10 – 15	8	8	12.5	-3	-24
15 – 20	14	22	17.5	-2	-28
20 - 25	18	40	22.5	-1	-18
25 – 30	25	65	27.5	0	0
30 – 35	15	80	32.5	1	15
35 – 40	14	94	37.5	2	28
40 – 45	6	100	42.5	3	6
	100				**-9**

$$\boldsymbol{Mean} = A + \frac{\Sigma fd}{N} \text{x } i = 27.5 - (9\text{x}5)/100 = 27.05$$

$$\boldsymbol{Median} = L + \left(\frac{\frac{N}{2} - c.f}{f}\right) \text{x } i \ = 25 + \left(\frac{50-40}{25}\right) \text{ x } 5 = 27$$

$$\boldsymbol{Mode} = \ L_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \text{ x } i = 25 + \left(\frac{25-18}{50-18-15}\right) \text{ x } 5 = 27.06$$

Measures of Dispersion

The literal meaning of dispersion is “scatteredness”. Dispersion gives an idea about the homogeneity or heterogeneity of the distribution. Measures of central tendency only tell us about the concentration of the observations about the central part of the distribution. Consider the series (1) 9, 9, 9, 9, 9 (2) 1, 2, 3, 4, 35 (3) 1, 4, 8, 12, 20. In all the cases it is seen that there are 5 observations and the mean value is 9. Here the average value does not tell us about the variation in the data. Measures of Dispersion gives complete information about the variation in the data. According to Spiegel, “The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data”. The average specifies the representative value of a series while the dispersion tells how the values are scattered from the central value. The main purpose of measuring dispersion is to test the reliability of an average, to compare two or more series with respect to variability and to facilitate as a basis for further statistical analysis. Measures of dispersion should be rigidly defined, simple to understand and easy to compute, based on all observations and suitable for further algebraic treatment. The dispersion may be measured either absolutely or relatively. When the dispersion is measured and expressed in terms of the original data, it is called an absolute dispersion. It cannot be used for comparison if expressed in different units. Relative measure of dispersion is computed for comparing the variability. The coefficient of dispersion of each group is calculated for comparing two series. The various measures of dispersion are Range, Quartile deviation, Mean deviation and Standard deviation.

Range is the simplest measure of dispersion. It depends only on the extreme values. It is simple to compute but not reliable measure. It cannot be applied to open end cases. Range is useful in studying the variation in the prices of shares, stocks and other commodities whose price fluctuates from one period to another period. Range is used in construction of control charts.

Range = Largest value – Smallest value = L - S

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

Example 10: Find the range of the series 7, 10, 18, 12, 6, 15, 5, 3.

Range = 18 – 3 = 15

Quartile deviation is defined as half the distance between the third and first quartiles. The quartile deviation gives an idea about the distribution of the middle half of the items around the median. It is simple to understand and not affected by extreme values. Its value is affected by sampling fluctuations.

Quartile Deviation (Q.D.) = $\frac{Q_3-Q_1}{2}$

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

The range and quartile deviations are positional measures of dispersion.

Example 11: Calculate the semi-interquartile range and quartile coefficient from the following data:

Age in years	20	30	40	50	60	70	80
Number of members	3	61	132	153	140	51	3

Age	**Frequency**	**Cumulative Frequency**
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543

Q_1 = size of (N+1)/4 th item = size of 136th item = 40 years

Q_3 = size of 3(N+1)/4 th item = size of 408th item = 60 years

Quartile Deviation (Q.D.) = $\frac{Q_3-Q_1}{2}$ = 10 years

Coefficient of Q.D. = $\frac{Q_3-Q_1}{Q_3+Q_1}$ = 0.2

Example 12: Compute Quartile deviation and coefficient of quartile deviation for the following problem.

Class Interval	Frequency	Cumulative Frequency
250 – 260	8	8
260 – 270	10	18
270 – 280	16	34
280 – 290	14	48
290 – 300	10	58
300 – 310	5	63
310 – 320	2	65

$$Q_1 = L + \left(\frac{\frac{N}{4} - c.f}{f}\right) \text{x } i \; = 268.25$$

$$Q_3 = L + \left(\frac{\frac{3N}{4} - c.f}{f}\right) \text{x } i = 290.75$$

Quartile Deviation (Q.D.) = $\frac{Q_3-Q_1}{2}$ = 11.25

Coefficient of Q.D. = $\frac{Q_3-Q_1}{Q_3+Q_1}$ = 0.04025

Mean deviation is the average of the deviations of a series from either the arithmetic mean or median, ignoring the signs of deviation. Mean deviation is rigidly defined and based on all values of the series. It is used in statistical analysis of economic business and social phenomena. It is useful in forecasting business cycles. The main demerit of mean deviation is that it ignores the algebraic signs of the deviation.

Standard deviation was introduced by Karl Pearson in 1893. Standard Deviation is the most important and widely used measure of dispersion. It is also called Root Mean Square Deviation or Mean Square Error. It is denoted by σ. It is defined as the positive square root of the arithmetic mean of the squares of the deviations of the given value from the arithmetic mean. It is rigidly defined; less affected by fluctuations of sampling and is the basis for measuring the correlation coefficient, sampling and statistical inferences. The relative measure of standard deviation is known as coefficient of variation. According to Karl Pearson, "Coefficient of Variation is the percentage variation in mean, standard deviation being considered as the total variation in the mean". Coefficient of variation is used to compare the variability in two or more series. The series with less coefficient of variation indicates that the series is more consistent or more homogeneous. The series with high coefficient of variation indicates that the series is less consistent or less stable. The variance is a measure of the "average" of the squared deviations from the mean. It is denoted by σ^2.

The formulae for Mean Deviation and Standard Deviation for individual series (raw data), discrete frequency distribution and continuous frequency distribution are given in Table 3.

Table 3 : Formulae for Measures of Dispersion

Measure of Dispersion	Individual Series (Ungrouped data)	Discrete Frequency Distribution	Continuous Frequency Distribution
Mean Deviation (M.D.)	$M.D.(about\ mean) = \frac{\sum\lvert D\rvert}{n}$ where D = $X - \bar{X}$ $M.D.(about\ median) = \frac{\sum\lvert D\rvert}{n}$ where D = $X - Median$	$M.D.(about\ mean) = \frac{\sum f\lvert D\rvert}{N}$ where D = $X - \bar{X}$ $M.D.(about\ median) = \frac{\sum f\lvert D\rvert}{N}$ where D = $X - Median$	$M.D.(about\ mean) = \frac{\sum f\lvert D\rvert}{N}$ where D = $m_i - \bar{X}$ $M.D.(about\ median) = \frac{\sum f\lvert D\rvert}{N}$ where D = m_i - Median
Standard Deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i(m_i - \bar{x})^2}{N}}$
where $N = \sum_{i=1}^{n} f_i$			
Coefficient of M.D.	Coeff. Of M.D. = Mean Deviation / Mean or Median		
Coefficient of Variation (C.V.)	$C.V. = \frac{\sigma}{\bar{x}}$ x 100		

Example 13: Calculate the mean deviation from (1) arithmetic mean (2) mode (3) median with respect to the marks obtained by 9 students: 7, 4, 10, 9, 15, 12, 7, 9, 7.

(1) Mean = 8.9

(2) Mode = 7

(3) Median 9

Marks	$\|D\|$ (Mean)	$\|D\|$ (Median)	$\|D\|$ (Mode)
7	1.9	2	0
4	4.9	5	3
10	1.1	1	3
9	0.1	0	2
15	6.1	6	8
12	3.1	3	5
7	1.9	2	0
9	0.1	0	2
7	1.9	2	0
	21.1	**21**	**23**

MD (about mean) = $\frac{\sum|D|}{n}$ = 2.34

MD (about median) = $\frac{\sum|D|}{n}$ = 2.33

MD (about mode) = $\frac{\sum|D|}{n}$ = 2.56

Example 14: Find the mean deviation about mean and standard deviation for the data given below.

Height (inches)	59.5-62.5	62.5-65.5	65.5-68.5	68.5-71.5	71.5-74.5
No. of male Students	5	18	42	27	8

Height	f_i	m_i	$f_i\, m_i$	$\|D\| = \|m_i - \bar{X}\|$	$f\|D\|$	$f_i\, m_i^2$
59.5-62.5	5	61	305	6.45	32.25	18605
62.5-65.5	18	64	1152	3.45	62.1	73728
65.5-68.5	42	67	2814	0.45	18.9	188538
68.5-71.5	27	70	1890	2.55	68.85	132300
71.5-74.5	8	73	584	5.55	44.4	42632
	100		**6745**		**226.5**	**455803**

$\bar{X} = \frac{\sum_{i=1}^{n} f_i m_i}{N}$ = 6745/100 = 67.45

$M.D.(about\ mean) = \frac{\sum f|D|}{N}$ = 226.5/100 = 2.26 inches

Coeff. of M.D. = Mean Deviation / Mean = 2.26/67.45 = 0.0335

$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (m_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} f_i m_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i m_i}{N}\right)^2}$ = 2.920188

$C.V. = \frac{\sigma}{\bar{x}}$ x 100 = 4.33

Example 15: The price of shares of companies X and y are given below. State which company's share s more stable in value?

Company X	55	54	52	53	56	58	52	50	51	49
Company Y	108	107	105	105	106	107	104	103	104	101

$\bar{X}$ = 53 $\qquad \sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$ =2.645

$C.V.(X) = \frac{\sigma}{\bar{x}}$ x 100 = 4.99

$\bar{y}$ = 105 $\qquad \sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}$ = 2

$C.V.(Y) = \frac{\sigma}{\bar{y}}$ x 100 = 1.90

Company Y shares are more stable in value s compared to Company X shares because the coefficient of variation of Company Y is lower than Company X.

Example 16: Compute the absolute measure of variance and a relative measure of dispersion from the following data.

Income (Rs.)	No. of families
300-399	30
400-499	46
500-599	58
600-699	76
700-799	60
800-899	50
900-999	20

Income (Rs.)	No. of families (f)	*m*	$d = \frac{m-649.5}{100}$	*fd*	*fd*^{2}
299.5-399.5	30	349.5	-3	-90	270
399.5-499.5	46	449.5	-2	-92	184
499.5-599.5	58	549.5	-1	-58	58
599.5-699.5	76	649.5	0	0	0
699.5-799.5	60	749.5	1	60	60
799.5-899.5	50	849.5	2	100	200
899.5-999.5	20	949.5	3	60	180
	N=340			$\sum fd$=**-20**	$\sum fd^2$=**952**

$\bar{X} = A + \frac{\sum fd}{N}$ x i = 643.6

Variance: $\sigma^2 = \left\{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2\right\} \text{x } i^2$ = 27965.39

$\sigma = \sqrt{27965.39}$ = 167.228

Coefficient of Variation: $C.V. = \frac{\sigma}{\bar{x}}$x 100 = 167.23/643.6 x 100 = 25.98%

Example 17: The index number of prices of cotton and coal shares is given below. Which of the two shares is more variable in prices?

Month	Index number of prices of cotton shares	Index number of prices of coal shares
January	188	131
February	178	130
March	173	130
April	164	129
May	172	129
June	183	120
July	184	127
August	185	127
September	211	130
October	217	137
November	232	140
December	240	142

Cotton shares:

$\bar{X}$ = 193.9 $C.V. = \frac{\sigma}{\bar{X}}$x 100 = 12.28%

σ_X = 23.81

Coal shares:

$\bar{Y}$ = 131 $C.V. = \frac{\sigma}{\bar{Y}}$x 100 = 4.42%

σ_Y = 23.81

The cotton shares are more variable in price than the coal shares.

Example 18: A manufacturer of television tubes has two types of tubes, A and B respectively. The tubes have mean lifetime of $\bar{X}_A$ = 1495 hours and $\bar{X}_B$ = 1875 hours, and standard deviation of σ_A = 280 hours and σ_B = 310 hours. Which tube has the greater relative dispersion?

C.V. (A) = σ_A / $\bar{X}_A$ x 100 = 18.7%

C.V. (B) = σ_B / $\bar{X}_B$ x 100 = 16.5%

Tube A has the greater relative dispersion.

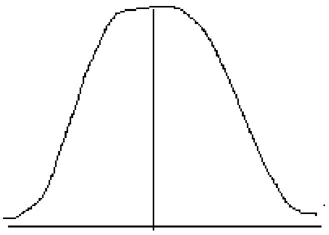
Skewness

The measures of central tendency and dispersion are not adequate to characterize a distribution completely. They do not indicate whether the distribution is symmetric or not. Skewness is a measure of degree of asymmetry of a distribution. Skewness gives an idea about the shape of the curve. According to Croxton and Cowden, "When a series is not symmetrical it is said to be asymmetrical or skewed". The mean, median and mode of a skewed distribution do not coincide. The relative measure of skewness is called coefficient of skewness. The following are the coefficient of skewness:

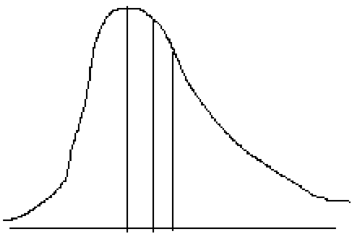
1. **Karl Pearson's coefficient of skewness**: $S_k = \frac{Mean-Mode}{\sigma}$ where σ is the standard deviation of the distribution.
2. **Bowley's Coeficient of skewness**: $S_k = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$

In a symmetrical distribution, the quartiles are equidistant from the median but in a skewed distribution, the quartiles are not equidistant from the median. Bowley's coefficient of skewness lies between ±1.

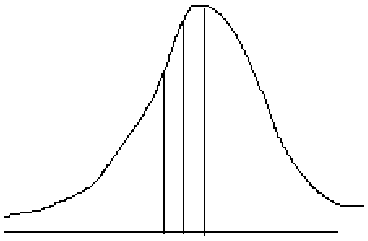
If the curve has longer tail towards the right then it is said to be **positively skewed** (**Mean>Median>Mode**). If the curve has longer tail towards left then it is **negatively skewed** (**Mean < Median < Mode**).



Mean=Median=Mode



Mode<Median<Mean



Mean<Median<Mode

Example 19: Calculate Karl Pearson's coefficient of skewness and Bowley's coefficient of skewness for the information given below.

Mean	Standard Deviation	Median	First Quartile	Third Quartile
26.01	**17.5**	**23.67**	**11.09**	**38.54**

Mode = 3 Median – 2 Mean = 18.99

$S_k = \frac{Mean-Mode}{\sigma}$ = 0.4

$S_k = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}$ = 0.08

Example 20: Compute Karl Pearsons' coefficient of skewness for the following distribution.

Class Interval	35-40	40-45	45-50	50-55	55-60
Frequency	2	14	24	30	15

Class Interval	***f***	***m***	$d = \frac{m-47.5}{5}$	***fd***	***fd***^{2}
35-40	2	37.5	-2	-4	8
40-45	14	42.5	-1	-14	14
45-50	24	47.5	0	0	0
50-55	30	52.5	1	30	30
55-60	15	57.5	2	30	60
	***N*=85**			$\sum fd$= **42**	$\sum fd^2$=**112**

$\bar{X} = A + \frac{\sum fd}{N}$x i = 49.97

$\sigma = \sqrt{\left\{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2\right\}}$x i = 5.18

$Mode = L_1 + \left(\frac{f_1-f_0}{2f_1-f_0-f_2}\right)$ x i = 51.43

$S_k = \frac{Mean-Mode}{\sigma}$ = -0.282

Kurtosis is used to study the peakedness of the distribution curve. A normal curve is symmetrical and bell shaped which is known as mesokurtic curve. If the curve is peaked at the top then it is known as leptokurtic curve. If the curve is more flat than normal curve then it is known as platykurtic curve.

1.2 Data Collection and Sampling Methods

Statistical data may be classified as primary and secondary data. Primary data are collected directly by the investigator to study a particular problem. There are many methods for primary data collection namely, direct method, experiments and surveys. Surveys may be conducted in different ways e.g. personal interview, telephone interview, self-administered questionnaire. Questionnaire should be designed properly to get precise information from the respondents.

Secondary data are the data which are already collected and is available for present study. The various sources of secondary data are international publications, official publications of government, publications of research institutions, financial institutions, journals and newspapers. The investigators should check the suitability, adequacy and reliability of secondary data before using it for analysis.

In a statistical enquiry, the population is set of all possible observations in the study. The population may be finite or infinite. Population can also be classified as hypothetical or real. It may be sometimes difficult to study the entire population, so a sample is drawn from the population to arrive at valid conclusions. Sampling refers to drawing a sample from a population. A sample is expected to represent the population. Sampling theory is a study of relationship, existing between a population and samples drawn from the population. Sampling theory is also useful in determining whether the observed difference between two samples is due to chance variation or whether they are really significant. A representative sample is obtained by random sampling. A sampling plan is a procedure for specifying how a sample is to be drawn from the population. The samples selected must have same characteristics of the population.

If we draw a sample from an urn, there is a choice of replacing or not replacing the sample into the urn before a second drawing. Sampling where each member of the population are chosen more than once is called sampling with replacement, while if each member cannot be chosen more than once is called sampling without replacement. A finite population in which sampling is with replacement is theoretically considered to be infinite. Since any number of samples can be drawn without exhausting the population.

The principal steps in a sample survey are:

1. Objectives of the survey
2. Defining the population to be sampled
3. The frame and Sampling units
4. Data to be collected
5. The questionnaire or schedule
6. Method of collecting information
7. Non-respondents
8. Selection of proper sampling design
9. Organization of field work
10. Summary and analysis of the data

Sampling plan can be classified as probability (random) and non-probability sampling methods. A random sample is one where each observation of the population has an equal chance of being selected.

The various **random sampling methods** are

1. Simple Random Sampling
2. Stratified Random Sampling
3. Systematic Sampling

Simple random sample is a sample selected such that each sample has equal probability of being chosen. Simple random sampling forms the basis for the other random sampling techniques. Let *n* denote the sample size and *N* represent the population size. The probability of selecting any particular item from the population is *1/N*. The samples may be selected with replacement or without replacement. Simple Random Sampling With Replacement (SRSWR) means that after an item is selected, the item is again returned to the frame, where it has the same probability of being selected again. Simple Random Sampling Without Replacement (SRSWOR) means that an item selected cannot be selected again. In otherwords, if an item is selected with probability *1/N* on the first draw then the probability of selecting another item (not previously selected) on second draw is *1/N-1*. This process continues until desired sample is selected. Several methods are used for random selection of samples.

(i) Lottery Method: This is most popular method. In this method, all the items of the population are numbered on separate slips of paper of same size and folded mixed up in a container. A blindfold selection is made.

(ii) Table of random numbers: A table of random numbers consists of a series of digits listed randomly generated sequence. There are several tables of random numbers such as Tippet’s Random Number Tables, Fisher and Yates Table, Kendall and Babington Smith’s Table. Since, the numeric system uses 10 digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9), the chance that any particular digit is randomly generated is equal to the probability of generating any other digit. To select a simple random sample, choose an arbitrary starting point from the table of random numbers.

A simple random sample is more representative of the population and personal bias is completely eliminated as compared to the judgement or purposive sampling. Some of the limitations of simple random sampling are that it requires up-to-date frame and larger sample size.

Stratified random sample is obtained by dividing the population into strata and then drawing simple random sample from each stratum. A stratum is defined by some common characteristic such as gender, age, education level, income, geographical area, economic status and so on. The units within an each group (strata) are homogeneous. A simple random sample is selected within each of the strata. Stratified random sampling is more efficient than simple random sampling or systematic sampling as it ensures the representation of items across the entire population.

In a **systematic sample**, the *N* items in the population is partitioned into *n* groups of *k* items where *k = n/N*. To select a systematic sample, choose the first item at random from the first *k* items in the frame and then select the remaining *n-1* items by taking every *kth* item thereafter from the entire frame.

In a **non-probability sampling**, an item or individual is selected without knowing their probabilities of selection. The various non-probability sampling methods are

1. Judgement or Purposive Sampling
2. Quota Sampling
3. Convenience Sampling

In **judgement sampling**, the choice of the sample items depends on the judgement of the investigator. Judgement sample is obtained according to the discretion of someone who is familiar with relevant characteristic of the population. It is a simple method and is used to obtain a more representative sample. The samples may not be representative due to individual bias.

Quota sampling is similar to stratified sampling. It is stratified cum purposive sampling. The data is collected by dividing the population into quota according to some characteristics.

In **convenient sampling**, convenient elementary units are chosen from a population. The items selected are inexpensive or convenient to sample. It is suitable when the population is not clearly defined and complete source list is not available. The samples are biased.

Sampling and Non-Sampling Errors

The errors involved in the collection, processing and analysis of a data may broadly be classified as sampling errors and non-sampling errors. Sampling error comprises the difference between the sample and the population which is due to improper selection of samples. Sampling bias is a tendency to favour the selection of units that have particular characteristics. Sampling errors are due to the fact that only a part of the population is used to estimate the population parameters and draw inferences about the population. Increase in the sample size usually results in the decrease in sampling error.

Sampling biases are due to the following reasons:

1. Sampling bias may be due to faulty selection of sample. This may be due to the use of defective sampling technique for selection of a sample. The investigator may use purposive or judgement sampling for obtain certain results. This bias can be overcome by using simple random sampling.

2. The investigators may sometimes substitute a convenient member of the population if there is a difficulty in enumerating a particular sampling unit. This leads to bias.
3. Bias due to defective demarcation of sampling units is significant in area surveys such as agricultural experiments in the field.
4. Sampling error may be due to improper choice of the statistic for estimating the population parameters.

Non-sampling error is an error that results from the way in which the observations are obtained. Non-sampling errors primarily arise at the stages of observation, ascertainment and processing of the data. It is present in both complete enumeration survey as well as sample surveys. Non-Sampling errors are more serious in complete enumeration as compared to a sample survey.

Non-sampling errors arise due to the following factors:

1. Faulty planning or definitions: Non-sampling errors may be due to inadequate and inconsistent data specification with respect to the objective of the survey. It may also be due to errors in ill-designed questionnaire, errors in recording the measurements or lack of trained investigators.
2. Response errors: These errors are introduced as a result of the response given by the respondents. The respondent may misunderstand a particular question and give improper information. The respondents may give wrong information due to self interest or prestige. Sometimes investigator may affect the accuracy of the response by the way the question is asked or recorded.
3. Non-response Biases: This bias occurs when full information is not obtained on all the sampling units.
4. Errors in coverage: If the objectives of the survey are not precisely stated then this may result in including the sampling units which are not to be included and excluding the units which are to be included.
5. Compiling errors: Various operations of data processing such as editing and coding of the responses, tabulation and summarizing the observations made in the survey are source of error.
6. Publication errors: The errors committed during presentation and printing of the tabulated results.

1.3 Sampling Distributions

In many applications, inferences about the population are made which are based on statistics calculated from samples. In order to do this, we need to discuss about the sampling distribution. A sampling distribution is a probability distribution consisting of all possible values of a sample statistic.

Consider all possible samples of size *N* that can be drawn from a given population (either with or without replacement). For each sample, a statistic such as mean or standard deviation is computed, which will vary from sample to sample. In this way, a distribution of the statistic is obtained which is known as sampling distribution. Sampling distribution for mean, variances can be derived. The standard deviation of the sampling distribution of a statistic is known as its standard error.

Sampling Distribution of Mean

The sampling distribution of the mean is the probability distribution of all possible values of the sample mean. The sample mean is unbiased because the mean of all possible sample means (of a given sample size *n*) is equal to the population mean. Let us consider a simple example consisting of 4 typists. Each typist is asked to type the same page of the manuscript and the number of typographical errors is recorded.

Typist	Number of Errors
A	3
B	2
C	1
D	4

From the data, $\mu = 2.5$ errors and $\sigma = 1.12$ errors. If we select samples of two typists with replacement from this population, there are 16 possible samples. The average of all 16 sample means is 2.5. Since the mean of the 16 sample means is equal to the population mean, the sample mean is an unbiased estimator of population mean.

Sample	Typists	Sample Observations	Sample Means
1	A, A	3, 3	3
2	A, B	3, 2	2.5
3	A, C	3, 1	2
4	A, D	3, 4	3.5

5	B, A	2, 3	2.5
6	B, B	2, 2	2
7	B, C	2, 1	1.5
8	B, D	2, 4	3
9	C, A	1, 3	2
10	C ,B	1, 2	1.5
11	C, C	1, 1	1
12	C, D	1, 4	2.5
13	D, A	4, 3	3.5
14	D,B	4, 2	3
15	D, C	4, 1	3
16	D, D	4, 4	4
			$\mu_{\bar{X}}$ = 2.5, $\sigma_{\bar{X}}$=0.28

How the sample mean varies from sample to sample can be expressed as the standard deviation of all possible sample means. This is known as the standard error of the mean. The standard error of mean is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.. The standard error is the spread of the values from the mean of the means in a sampling distribution. In this example, although the sample means vary from sample to sample, depending on which two typists are selected, the sample means do not vary much as the individual values in the population.

As the sample size increases, the standard error of the mean goes to zero. The bigger the sample the closer to the true mean we are likely to get and so have more confidence in that estimate. So, we can say that if we are sampling from a normally distributed population with a mean, μ, and standard deviation, σ, then the sampling distribution of the mean will also be normally distributed for any size n with mean μ_{x-} =μ and standard error of the mean, $\sigma_{\bar{X}}$.

Table 3 gives standard errors of sampling distribution for various statistics under the condition of random sampling from an infinite population or of sampling with replacement from a finite population.

Table 3 Standard Errors of Sampling Distribution

Statistic	Standard Error
Mean	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
Proportion	$\sigma_p = \sqrt{\frac{pq}{n}}$
Standard deviation	$\sigma_s = \frac{\sigma}{\sqrt{2n}}$
Median	$\sigma_{med} = \sigma\sqrt{\frac{\pi}{2n}} = \frac{1.2533\sigma}{n}$

Sampling Distribution of Proportion

Consider a categorical variable which has two categories, such as the customer prefers a particular brand of a product or not. Let π be the proportion of product in the entire population and *p* be the proportion of product in the sample with the characteristic of interest.

The sample proportion is used to estimate the population proportion. The sample proportion takes on values between 0 and 1.

p = *X*/*n* = Number of product having the characteristic of interest/ Sample size

The standard error of proportion is given by $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$. The sampling distribution of the proportion follows Binomial distribution.

Central Limit Theorem

There are many instances where sampling is done from non-normally distributed populations. The Central Limit Theorem deals with this situation. The Central Limit Theorem provides the information required for constructing a sampling distribution.

The Central Limit Theorem states that as the sample size increases (gets large enough), the sampling distribution of the mean is approximately normally distributed. For most population distributions, regardless of shape, the sampling distribution of the mean is approximately normally distributed if the samples of at least size 30 are selected. By applying the theorem, the descriptive values for a sampling distribution (usually, the mean and the standard error, which is computed from the sampling variance) is obtained and the probabilities associated with any of the sample means in the sampling distribution are also obtained. The most important reason why we need to construct sampling distributions is to obtain the probabilities associated with varying amounts of sampling error. Once we have this information, it can be used for statistical generalization and for testing the hypotheses. The Central Limit Theorem allows us to directly compute the mean, variance, and standard error of the sampling distribution without taking trouble of drawing all possible samples from the population. The shape of the distribution of sample means is described by using the third principle of the Central Limit Theorem. This principle states that the sampling distribution of means will tend to look like a normal distribution, when the samples are selected from relatively large numbers of observations. The theoretical proof of the central limit theorem requires independent observations in the sample. This condition is met for infinite population and for finite populations where samples are taken with replacement. In general, central limit theorem is applied when the population size is large.

Questions

1. Distinguish between primary and secondary data.
2. Discuss the various methods of collecting primary data.
3. Explain the significance of sampling methods.
4. Describe briefly different types of sampling methods.
5. Explain sampling and non-sampling errors.
6. What are measures of central tendency? Also state the requisites of a good average.
7. State the various measures of central tendency and dispersion.
8. What do you understand by median and quartiles?
9. Briefly explain measures of dispersion. Why standard deviation is called as ideal measure of dispersion.
10. What is the need for sampling as compared to complete enumeration?
11. Differentiate between simple random sampling with replacement and without replacement.
12. Describe the following sampling methods with suitable examples.
    a. Judgement sampling b. Cluster sampling c. Quota sampling
13. Describe different types of random sampling methods.
14. Explain non-probability sampling methods. State its disadvantages.
15. Discuss about sampling distributions and standard error.

EXERCISES

16. The number of ATM transactions per day was recorded at 15 locations in a large city. Find the mean and median number of transactions from the following data

35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32, 60

17. (a) Find the mean, median and mode for the data: 8, 11, 4, 3, 2, 5, 10, 6, 4, 1, 10, 8, 12, 6, 5, 7

(b) Find the geometric mean for the data:

28.5, 73.6, 47.2, 31.5, 64.8

(c) Find the harmonic mean for the data:

3.2, 5.2, 4.8, 6.1, 4.2

18. In automobile mileage and fuel consumption testing, 10 automobiles were road tested for 300 miles in both city and country driving conditions. The following data were recorded for miles-per-gallon performance.

City:	16.2	16.7	15.9	14.4	13.2	15.3	16.8	16.0	16.1	15.3
Country:	19.4	20.6	18.3	18.6	19.2	17.4	17.2	18.6	19.0	21.1

Comment about the difference in performance for city and country driving using mean, median and mode.

19. In an examination of 675 candidates, the examiner supplied the following information:

Marks obtained	**Number of Candidates**
Less than 10%	7
Less than 20%	39
Less than 30%	95
Less than 40%	201
Less than 50%	381
Less than 60%	545
Less than 70%	631
Less than 80%	675

Calculate the arithmetic mean and mode of the percentage marks obtained.

20. The daily wages of the employees of a company is given in the following table. Calculate

i) the quartiles : Q_1, Q_2, Q_3

ii) the deciles : D_1, D_2, D_9

iii) the percentiles : P_{35}, P_{60}

Wages(Rs)	Number of Employees
250.00-260.00	8
260.00-270.00	10
270.00-280.00	16
280.00-290.00	14
290.00-300.00	10
300.00-310.00	5
310.00-320.00	2

21. Calculate arithmetic mean from the following frequency distribution.

Output in units	**Number of workers**
300-309	9
310-319	20
320-329	24
330-339	38
340-349	48
350-359	27
360-369	17
370-379	6

22. Compute the mode of the following distribution.

Size of item	Frequency (*f*)
0 - 5	20
5 – 10	24
10 – 15	32
15 – 20	28
20 - 25	20
25 – 30	16
30 – 35	37
35 – 40	10
40 – 45	8

23. In a small branch of a bank in a rural area, the following is the average deposit balance of current accounts during a month. Calculate the median, seventh decile and 85th percentile.

Deposit balance less than (Rs.)	**No. of deposits**
1000	500
900	498
800	480
700	475
600	440
500	374
400	300
300	125
200	25

24. Compute mean deviation about mean, standard deviation and coefficient of variation for the following distribution.

Profits (Rs. in crores)	**No. of companies**
20-30	4
30-40	8
40-50	18
50-60	30
60-70	15
70-80	10
80-90	8
90-100	7

25. Calculate geometric for the following distribution.

Yield	7.5 – 10.5	10.5-13.5	13.5-16.5	16.5-19.5	19.5-22.5	22.5-25.5	25.5-28.5
No. of farms	5	9	19	23	7	4	1

26. The following table shows the age distribution of persons in a particular region. Find the median age.

Age (years)	Below 10	Below 20	Below 30	Below 40	Below 50	Below 60	Below 70	70 and above
No. of persons (in thousands)	2	5	9	12	14	15	15.5	15.6

27. Compute mean deviation about mean and mean coefficient of dispersion for the following data.

Marks	10	15	20	30	40	50
Frequency	8	12	15	10	3	2

28. Obtain Karl Pearson's coefficient of skewness and Bowley's coefficient of skewness for the following data.

Class Interval	130-134	135-139	140-144	145-149	150-154	155-159	160-164
Frequency	3	12	21	28	19	12	5

29. In a moderately skewed frequency distribution, the mean is Rs.15 and the median is Rs. 14. If the coefficient of variation is 30% then find the Karl Pearsons's coefficient of skewness.

30. An analysis of the monthly wages paid to workers in the firms A and B belonging to the same industry gives the following results. Which firm has greater variability in individual wages?

	Firm A	**Firm B**
No. of workers	500	600
Average Monthly Wage (Rs.)	480	475
Variance of distribution of wages (Rs.)	400	625

UNIT II

PROBABILITY THEORY

CONTENTS:

2.1 – Introduction to Probability theory

2.2 – Conditional Probability

2.3 – Probability Distributions

2.1 Introduction to Probability Theory

A probability is the numeric value representing the chance or possibility a particular event will occur, such as a rainy day, a defective product or price of a stock etc. According to Ya-Lin Chou, "probability is the science of decision making with calculated risks in the face of uncertainty". If the probability of success is based on prior knowledge of the process involved then it is called a priori probability. In the empirical probability approach, the probability is based on the observed data. Subjective probability differs from person to person. The basic terminologies used in probability theory are given below.

1. **Random Experiment**: If in each trial of an experiment conducted under identical conditions the outcome is not unique, but may be any one of the possible outcomes, then such an experiment is called a random experiment.

   e.g : tossing a coin, throwing a die, selecting a card from a pack of playing cards etc.
2. **Outcome**: The result of a random experiment is called an outcome.
3. **Trial and Event**: Any particular performance of a random experiment is called a trial and outcomes or combination of outcomes is called events.

   e.g: if a coin is tossed repeatedly, we may get head or tail.

   Tossing of coin - Trial

   Getting head or tail - Event.

   Throwing a die – Trial, getting outcomes 1,2,3,4,5,6 – Event.
4. **Exhaustive Events:** The total number of possible outcomes of a random experiment is known as exhaustive events.

   e.g: In tossing of a coin, there are two exhaustive cases namely, Head and Tail

   In throwing a die, there are 6 exhaustive cases namely 1, 2, 3, 4, 5, 6
5. **Favourable Events**: The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event.

e.g. In drawing a card from a pack of cards, the number of cases favourable to drawing an Ace is 4.

6. **Mutually Exclusive Events**: Events are said to be mutually exclusive if happening of any one of the them precludes the happening of all the others (i.e.) if no two or more of them can happen simultaneously in the same trial.

   e.g. In tossing a coin, the events head and tail are mutually exclusive.

7. **Equally Likely Events**: Outcomes of trial are said to be equally likely if taking into consideration all the relevant evidences, there is no reason to expect one is preference to the others.

   e.g: In a random toss of an unbiased coin, head and tail are equally likely.

8. **Independent Events:** Events are independent if the happening of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events.

   e.g : In tossing an unbiased coin, the event of getting a head in the first toss is independent of getting a head in the second, third etc.

Mathematical or Classical or apriori Probability:

If a random experiment or a trial results in 'n' exhaustive, mutually exclusive and equally likely outcomes, out of which 'm' are favourable to the occurrence of an event E, then the probability 'p' of occurrence of E, denoted by P(E) is given by

$$p = P(E) = \frac{Number\ of\ favourable\ cases}{Total\ number\ of\ exhaustive\ cases} = \frac{m}{n}$$

Remark:

1. Since $m \geq 0$, $n > 0$ and $m \leq n$, $P(E) \geq 0$ and $P(E) \leq 1$. => $0 \leq p \leq 1$
2. The non-happening of the event E is called the complementary event of E and is denoted by $\bar{E}$ or E^c. The number of favourable cases of $\bar{E}$ is (n-m). Then, the probability q that E will not happen is given by ,

$$q = P(\bar{E}) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p \quad => p + q = 1$$

Statistical or Empirical Probability:

If an experiment is performed repeatedly under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the number of trials, as the number of trial becomes indefinitely large, is called the probability of happening of the event,

$$P(E) = \lim_{N\to\infty} \frac{M}{N}$$

Sample space: The set of all possible outcomes of a given random experiment is called the sample space associated with that experiment. Each possible outcome or element in a sample space (*S*) is called a sample point or an elementary event.

Event: Every non-empty subset of *S*, which is a disjoint union of single element subsets of the sample space S of a random experiment *E* is called an event.

Examples:

1. Tossing a single coin – *S* = {*H, T*} ; *n(S)* = 2 {*n(S)* is total number of sample points in *S*.}
2. Tossing of two coins – *S* = {*HH, HT, TH, TT*} ; *n(S)* = 4.
3. Throwing a die - *S* = { 1, 2, 3, 4, 5, 6 } *n(S)* = 6.
4. Throwing two dice –*S* = $\begin{Bmatrix} (1,1), & (1,2) & . & . & . & (1,6) \\ (2,1), & (2,2), & - & - & - & (2,6) \\ (3,1) & (3,2) & - & - & - & (3,6) \\ (4,1) & - & - & - & - & (4,6) \\ (5,1) & - & - & - & - & (5,6) \\ (6,1) & - & - & - & - & (6,6) \end{Bmatrix}$

E_1 = Sum of points on two dice in 5

E_1 = $\{(1,4), \quad (4,1), \quad (2,3), \quad (3,2)\}$, $n(E_1) = 4$

E_2 = Sum of points on two dice is odd

$$\begin{Bmatrix} (1,2), & (1,4), & (1,6), & (2,1), & (2,3), & (2,5), & (3,2), & (3,4), & (3,6), \\ (4,1), & (4,3), & (4,5), & (5,2), & (5,4), & (5,6), & (6,1), & (6,3), & (6,5) \end{Bmatrix}$$

$n(E_2) = 18$

Example 1: A ball is drawn at random from a box containing 6 red balls, 4 green balls and 5 blue balls. Find the probability that the ball drawn is

i) red *ii)* green *iii)* blue *iv)* red or green

i) P(red) = 6/15 = 2/5
ii) P(green) = 4/15
iii) P(blue) = 5/15= 1/3
iv) P(red or green) = 10/15 = 2/3

Example 2: *a)* Two cards are drawn at random from a pack of 52 cards. Find the probability of drawing two aces.

$$P(A) = \frac{4C_2}{52C_2} = \frac{1}{221}$$

b) Four cards are drawn from a pack of cards. Find the probability that

i) all are diamonds *ii)* two are spades and two are hearts

$$i)P(A) = \frac{13C_4}{52C_4} \qquad ii)P(A) = \frac{13C_2 \times 13C_2}{52C_4}$$

Example 3: Out of 20 employees in a company, 5 are post graduates. Three employees are selected at random. Find the probability that all the three are post graduates.

$$P(E) = \frac{\mathbf{5}C_3}{20C_3}$$

Example 4: A coin is tossed two times. The sample space is S = {HH, HT, TH, TT}, where H and T denote head and tail. What is the probability of getting at least one head?

Let A be an event of getting at least one head. A = {HH, HT, TH}

P(A) = P(HH) + P(HT) + P(TH)

= 1/4 + 1/4 + 1/4

P(A) = 3/4

Example 5: If two television tubes are picked in succession from a shipment of 240 television tubes of which 15 tubes are defective. What is the probability that they will both be defective?

Prob. (both the selected tubes will be defective) = 15/240 x 14/239 = 7/1912

Example 6: From 25 tickets, marked with the first 25 numerals, one is drawn at random. Find the probability that

i) it is a multiple of 5 or of 7,

ii) it is a multiple of 3 or of 7.

i) Numbers (out of the first 25 numerals) which are **multiples of 5** are **5, 10, 15, 20** and **25**, and the numbers which are **multiples of 7** are **7, 14 and 21.**

Pr (a multiple of 5 or 7) = 8 / 25

ii) Numbers (among the first 25 numerals) which are **multiples of 3** are **3, 6, 9, 12, 15, 18, 21, 24** and the numbers which are **multiples of 7** are **7, 14, 21**.

Pr (a multiple of 3 or 7) = 10 / 25 = 2 / 5.

Example 7: Three groups of chicken contain respectively 3 girls and 1 boy, 2 girls and 2 boys, 1 girl and 3 boys. One child is selected at random from each group. Determine the probability that the three selected children consist of 2 boys and 1 girl.

To select 2 boys and 1 girl:

Group No.	I	II	III
(i)	Girl	Boy	Boy
(ii)	Boy	Girl	Boy
(iii)	Boy	Boy	Girl

$$\text{Prob} = P(i) + P(ii) + P(iii)$$

$$= \frac{3}{4}\times\frac{2}{4}\times\frac{3}{4}+\frac{1}{4}\times\frac{2}{4}\times\frac{3}{4}+\frac{1}{4}\times\frac{2}{4}\times\frac{1}{4}$$

$$\text{Prob} = \frac{9}{32}+\frac{3}{32}+\frac{1}{32} = \frac{13}{32}$$

Example 8: If two dice are thrown, what is the probability that the sum is

a) Greater than 8 *b)* neither 7 or 11

a) Let *S* denote the sum of the two dice.

$$P(S > 8) = P(S = 9) + P(S = 10) + P(S = 11) + P(S = 12)$$

$$P(S > 8) = \frac{4}{36}+\frac{3}{36}+\frac{2}{36}+\frac{1}{36}=\frac{10}{36}=\frac{5}{18}$$

b) Let *A* denote the event of getting the sum of 7 and *B* denote the event of getting the sum of 11.

$$P(A) = \frac{1}{6} \quad , \qquad P(B) = \frac{1}{18}$$

$$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) = 1 - P(A) - P(B)$$

$$= 1 - \frac{1}{6} - \frac{1}{18} = \frac{7}{9}$$

Example 9: A factory finds that, average, 20% of the bolts produced by a given machine will be defective for certain specified requirements. If 10 bolts are selected at random from the day's production of this machine. Find the probability

a) that exactly 2 will be defective
b) that 2 or more will be defective
c) that more than will be defective

Let X be number of defective bolts.

a) $P(X = 2) = 10C_2(0.2)^2(0.8)^8 = 45(0.04)(0.1678) = 0.3020$

b) $P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$

$= 1 - 10C_0(0.2)^0(0.8)^{10} - 10C_1(0.2)^1(0.8)^0$

$= 1 - (0.8)^{10} - 10(0.2)(0.8)^9$

$= 1 - 0.1074 - 0.2684 = 0.6242$

c) $P(X > 5) = P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$

$= 10C_6(0.2)^6(0.8)^4 + 10C_7(0.2)^7(0.8)^3 + 10C_8(0.2)^8(0.8)^2 + 10C_9(0.2)^9(0.8)^1 + 10C_0(0.2)^{10}$

$= 0.00637$

Results on probability:

1. $P(\emptyset) = 0.$
2. $P(\bar{A}) = 1 - P(A).$
3. $P(\bar{A} \cap B) = P(B) - P(A \cap B).$
4. $P(A \cap \bar{B}) = P(A) - P(A \cap B).$

Addition Theorem of Probability:

If A and B are any two events (subsets of sample space S) and are not disjoint, then

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

Proof:

$$P(A \cup B) = \frac{n(A \cup B)}{n(s)} = n(A) + n(B) - n(A \cap B)$$

$$= \frac{n(A)}{n(s)} + \frac{n(B)}{n(s)} - \frac{n(A \cap B)}{n(s)} = P(A) + P(B) - P(A \cap B)$$

*Corollary*1: If the events A and B are mutually disjoint then

$$A \cap B = \emptyset \qquad => \quad P(A \cap B) = P(\emptyset) = 0$$

*Corollary*2: For three non-mutually exclusive events to A, B and C

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

Example 10: The chance of winning the race by person A is 1/5 and that of person B is 1/6. What is the probability that the race will be won by A or B?

P($A \cup B$) = P(A) + P(B) = 11/30

Example 11: Find the probability of a 4 turning up at least once in two tosses of a die.

Let A_1 be an event of '4' turning up on the first toss.

Let A_2 be an event of '4' turning up on the second toss.

Since, A_1 and A_2 are not mutually exclusive,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

$$= \frac{1}{6} + \frac{1}{6} - \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)$$

$$P(A_1 \cup A_2) = \frac{11}{36}$$

Example 12: A problem in Mathematics is given to three students *X*, *Y* and *Z* whose chances of solving it are 3/4, 1/2 and 1/4 respectively. What is the probability that the problem will be solved?

$$P(X \cup Y \cup Z) \ = P\ \ (X) + P(Y) + P(Z) - P(X \cap Y) - P(X \cap Z) - P(Y \cap Z) + P(X \cap Y \cap Z)$$

$$= \frac{3}{4} + \frac{1}{2} + \frac{1}{4} - \frac{3}{4} \times \frac{1}{2} - \frac{3}{4} \times \frac{1}{4} - \frac{1}{2} \times \frac{1}{4} + \frac{3}{4} \times \frac{1}{2} \times \frac{1}{4} \ = \ \frac{29}{32}$$

Example 13: In a city (based on a sample survey), the probabilities that a family owns a television set, a washing machine or both television and washing machine are 0.86, 0.35 and 0.29 respectively. What is probability that a family owns either or both?

Let A be an event that a family owns a television set.

Let B be an event that a family owns a washing machine.

P(A) = 0.86

P(B) = 0.35

P(A ∩ B) = 0.29

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

= 0.86 + 0.35 – 0.29

= 0.92

Example 14: *A* can hit a target 3 times in 5 shots, *B* 2 times in 5 shots, and *C* 3 times in 4 shots. Find the probability of the target being hit at all when all of them try.

Let E_1 be the event that *A* hits the target.

$$P(E_1) = \frac{3}{5}\ \ and\ \ P(\overline{E_1}) = \ 1 - \frac{3}{5} = 2/5$$

and E_2 be the event that *B* hits the target,

$$P(E_2) = \frac{2}{5} \ and \ P(\overline{E_2}) = \ 1 - \frac{2}{5} = 3/5$$

and E_3 be the event that *B* hits the target,

$$P(E_3) = \frac{3}{4} \ and \ P(\overline{E_3}) = 1 - \frac{3}{4} = \frac{1}{4}$$

The required probability 'p' that the target is hit when they all try is given by

$$p \ = P[atleast \ one \ of \ the \ three \ hits \ the \ target]$$

$$= 1 - P[none \ hits \ the \ target]$$

$$= 1 - P(\overline{E_1} \cap \overline{E_2} \cap \overline{E_3})$$

$$= 1 - P(\overline{E_1})P(\overline{E_2})P(\overline{E_3}),$$

by compound probability theorem, since E_1, E_2 and E_3 are independent

⇨ $\overline{E_1}, \overline{E_2} \ and \ \overline{E_3}$ are also independent

$$p = 1 - \frac{2}{5}.\frac{3}{5}.\frac{1}{4} = \frac{47}{50}$$

Example 15: The odds against *A* solving a certain problem are 4 to 3 and odds in favour of *B* solving the same problem are 7 to 5. What is the probability that the problem is solved if they both try independently?

Let E_1 denote the event that ' *A*' solves the problem.

Let E_2 denote the event that '*B*' solve the problem.

$$P(E_1) = \frac{3}{4+3} = \frac{3}{7} \qquad P(E_1^c) = 1 - \frac{3}{7} = \frac{4}{7}$$

$$P(E_2) = \frac{7}{12} \qquad P(E_2^c) = 1 - \frac{7}{12} = \frac{5}{12}$$

P(atleast one of *A* and *B* solves the problem) $= 1 - P(E_1^c \cap E_2^c) \ = \ 1 - \left(\frac{4}{7} \times \frac{5}{12}\right)$

$$= 16/21$$

2.2 Conditional Probability

Let *P(A)* represent the likelihood that a random experiment will result in an outcome in the set *A* relative to the sample space *S* of the random experiment. If we have prior information that the outcome of the random experiment must be in a set *B* of *S* then this information must be used to re-appraise the likelihood that the outcome will also be in B. This re-appraised probability is denoted by

P(A|B) = conditional probability of the event A given that event B has already happened

When we know that a particular event *B* has occurred instead of *S*, we concentrate our attention on *B* only and the conditional probability of *A* given *B* will be the ratio of that part of *A* which is included in *B* (ie $A \cap B$) to the probability of *B*.

e.g : Let us consider a random experiment of drawing a card from a pack of cards.

A: drawing a king card ; *n*(*A*) = 4

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

Now suppose that a card is drawn and we are informed that the drawn card is red. How does this information affect the likelihood of the event *A*?

B: the card drawn is red ; *n*(*B*) = 26

Now, the *P(A)* must be computed relative to the new sample space *B* which consists of 26 sample points (red cards only).

Among the 26 red cards, 2(red) kings so $n(A \cap B) = 2$.

Hence the required probability

$$P(A|B) = \frac{n(A\cap B)}{n(B)} = \frac{2}{26} = \frac{1}{13}$$

Multiplication Theorem of Probability:

$$P(A \cap B) = \begin{cases} P(A).P(A|B), P(A) > 0 \\ P(B).P(A|B), P(B) > 0 \end{cases}$$

where *P(B|A)* represents conditional probability of occurrence of *B* when the event *A* has already happened and *P(A|B)* is the conditional probability of happening of A, given that *B* has already happened.

Independent Events: An event *A* is said to be independent of another event *B*, if the conditional probability of *A* given *P(A|B)* is equal to the unconditional probability of *B* (ie) if *P(A|B) = P(A)* and *P(B|A)= P(B).*

Multiplication theorem of probability for independent events:

If A and B are two events with positive probabilities $\{P(A) \neq 0, P(B) \neq 0\}$ then A and B are independent if and only if $P(A \cap B) = P(A).P(B)$

Pairwise independent events: The events $A_1, A_2, \dots \dots . A_n$ are said to be pairwise independent if and only if

$$P(A_i \cap A_j) = P(A_i).P(A_j); \quad i \neq j = 1,2,3, \dots, n$$

Example 16: Given that P(A) = 1/3, P(B) = 1/4, P(A|B) = 1/6. Find P(B|A) and P(B|A^c).

P(B|A) = P(A|B) P(B) / P(A) = 1/8 , P(B|A^c) = 5/16

Example 17: Three marbles are drawn successively from a bag containing 5 blue marbles, 4 green marbles and 6 black marbles. Find the probability that they drawn in the order black, green and blue if each ball is (a) replaced (b) not replaced

Let *A* be an event of drawing black marble on the first draw

Let *B* be an event of drawing green marble on the second draw

Let *C* be an event of drawing blue marble on the third draw

a) If each marble is replaced, then *A*, *B* and *C* are independent events

$$P(ABC) = P(A)P(B)P(C) = \frac{6}{15} \times \frac{4}{15} \times \frac{5}{15} = \frac{8}{225}$$

b) If each marble is not replaced, then *A*, *B* and *C* are dependent events.

$$P(ABC) = P(A)P\left(\frac{B}{A}\right)P\left(\frac{C}{AB}\right)$$

$$= \frac{6}{15} \times \frac{4}{14} \times \frac{5}{13} = \frac{4}{91}$$

Example 18: A consumer research organization has studied the services under warranty provided by the 50 new-car dealers in a certain city. The following table gives the results of the findings of the study.

	Good service under warranty	Poor service under warranty
Dealer in business(5 years or more)	16	4
Dealer in business (less than 5 years)	10	20

(i) What is the probability that a customer (who randomly selects one of the new-car dealers), gets the dealer who provides good service under warranty?

(ii) What is the probability a customer (who randomly selects the dealer who has been in the business for 5 years or more) gets the dealer who provides good service under warranty?

(iii) What is the probability that one of the dealers who has been in business less than 5 years will provide good service under warranty?

Let A denote the selection of a dealer who provides good service under warranty.

Let D denote the selection of a dealer who has been in business for 5 years or more.

(i) P(A) = (16 + 10) / 50 = 0.52

(ii) P(A| D) = 16 / 20 = 0.80

(iii) $P(A \mid D^c) = P(A \cap D^c) / P(D^c) = \frac{\frac{10}{50}}{\frac{30}{50}} = 0.20 / 0.60 = 1/3 = 0.33$

Example 18: Find the probabilities of getting

(a) Three heads in three tosses of a coin

(b) Four sixes and then another number in five throws of a die.

(a) Pr.(Three heads in three tosses of a coin) = 1/2 x 1/2 x 1/2 = 1/8

(b) Pr. (Four sixes and then another number in five throws of a die)

= 1/6 x 1/6 x 1/6 x 1/6 x 5/6

= 5 / 7776

Example 19: A coin is tossed three times and the outcomes are HHH, HHT, HTH, HTT, THH, THT, TTH, TTT. If A is the event that a head occurs on each of the first two tosses, B is the event that a tail occurs on the third toss and C is the event that exactly two tails occur in three tosses. Show that

(a) Events A and B are independent

(b) Events B and C are dependent

A = {HHH, HHT}

P(A) = 1/4

B = {HHT, HTT, THT, TTT}

P(B) = 1/2

C = {HTT, THT, TTH}

P(C) = 3/8

A ∩ B = {HHT}

P(A ∩ B) = 1/8

B ∩ C = {HTT, THT}

P(B ∩ C) = 1/4

(a) Since P(A) . P(B) = 1/4 . 1/2 = 1/8 = P(A ∩ B) ,

Events A and B are independent.

(b) Since P(B) . P(C) = 1/2 . 3/8 = 3/16 ≠ P(B ∩ C) ,

Events A and B are dependent.

Bayes' Theorem:

If $E_1, E_2, \dots \dots . E_n$ are mutually disjoint event with *P (Ei)*≠ 0 (*i=1,2,….,n*), then for any arbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that *P(A)* > 0, we have

$$P(E_i|A) = \frac{P(E_i)P(A|E_i)}{\sum_{i=1}^{n} P(E_i)P(A|E_i)} = \frac{P(E_i)P(A|E_i)}{P(A)}, i = 1,2, \dots, n$$

Remark:

1. The probabilities $P(E_1).P(E_2) \dots \dots P(E_n)$ are termed as 'a prior' probabilities.
2. $P(E_i|A)$ are called 'likelihood'.
3. $P(E_i|A)$ are called 'posterior' probabilities.
4. If the events constitutes a disjoint partition of the sample space *S* and *P (Ei)*≠ 0; *i=1,2,…,n* then for any event *A* in *S*, we have

$$P(A) = \sum_{i=1}^{n} P(E_i)P(A|E_i)$$

Example 20: Basket I contains 1 white, 2 black and 3 red balls. Basket II contains 2 white, 1 black and 1 red ball. Basket III contains 4 white, 5 black and 3 red balls. One basket is chosen at random and two balls are drawn. They happen to be white and red. What is the probability that the balls are drawn from Basket I, II or III?

Let E_1, E_2, E_3 be the event of choosing Basket I, II and III.

Let *A* be an event that the two balls drawn from the selected basket are white and red.

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$$

$$P(A|E_1) = \frac{1 \times 3}{6C_2} = \frac{1}{5}, \quad P(A|E_2) = \frac{2 \times 1}{4C_2} = \frac{1}{3}, \quad P(A|E_3) = \frac{4 \times 3}{12C_2} = \frac{2}{11}$$

$$P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P\left(\frac{A}{E_3}\right)} \quad = \frac{33}{118}$$

$$P(E_2|A) = \frac{P(E_2)P(A|E_2)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P\left(\frac{A}{E_3}\right)} \quad = \frac{55}{118}$$

$$P(E_3|A) = \frac{P(E_3)P(A|E_3)}{P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)} \quad = \frac{30}{118}$$

Example 21: An industrial unit has 3 machines- I, II and III which produce the same item. Machines I and II, each produce 30% of the total output, Machine III produces 40% of the remaining output. 2% of Machines I defective while Machine II and III, each produces 3% defective items. All the items are put into one stockpile, and then one item is chosen at random. Find the probability that

a) the selected item is defective.

b) What is the probability that the selected defective item was produced by Machine I?

Let *E₁*, *E₂*, *E₃* be the events that the item chosen is produced by Machines I, II and III respectively. Let *A* be an event that the item chosen is defective.

a) $P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$

$= (0.3) \times (0.02) + (0.03) \times (0.03) + (0.4) \times (0.03)$

$= 0.027$

b) $P(E_1|A) = \frac{P(E_1)P(A|E_1)}{P(A)} \quad = \quad \frac{(0.02)\times(0.3)}{0.027}$

$= 0.223$

Example 22: The members of a consulting firm rent cars from three rental agencies: 60 percent from agency 1, 30 percent from agency 2 and 10 percent from agency 3. If 9 percent of the cars from agency one need a tune-up , 20 percent of the cars from agency 2 need a tune-up and 6 percent of the cars from agency 3 need a tune-up.

(a) What is the probability that a rental car delivered to the firm will need a tune-up?

(b) If a rental car delivered to the consulting firm needs a tune-up, what is the probability that it came from agency 2?

Let E_1, E_2, E_3 be the events that the car comes from rental agencies 1, 2 and 3 respectively. Let A be an event that the car needs a tune-up.

a) $P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$

$= (0.6) \times (0.09) + (0.3) \times (0.20) + (0.1) \times (0.06)$

$= 0.12$

b) $P(E_2|A) = \frac{P(E_2)P(A|E_2)}{P(A)} = \frac{(0.3)\times(0.2)}{0.12}$

$= 0.5$

Random Variable

A random variable is a function X(ω) with domain S and range (-∞, ∞) such that for every real number a, the event [ω : X(ω) ≤ a] ∈ $\mathcal{B}$. In other words, we are considering a function whose domain is the set of possible outcomes, and whose range is a subset of the set of real.

Consider a random experiment (E) consisting of two tosses of a coin.

Outcomes: *HH* *HT* *TH* *TT*

Let X be a real number associated with the outcome of the experiment. Let X be number of heads.

X : 2 1 1 0

Random variables are denoted by capital letters X, Y, Z,……

Examples: 1. If a coin is tossed $S = \{\omega_1, \omega_2\}$; $\substack{\omega_1 = H \\ \omega_{2=T}}$

$$X(\omega) = \begin{cases} 1 \; if \; \omega = H \\ 0 \; \; if \; \omega = T \end{cases}$$

2. If a pair of fair dice is tossed then $S = \left\{\begin{matrix} (1,1), & (1,2) & . & . & . & (1,6) \\ (2,1), & (2,2), & - & - & - & (2,6) \\ (3,1) & (3,2) & - & - & - & (3,6) \\ (4,1) & - & - & - & - & (4,6) \\ (5,1) & - & - & - & - & (5,6) \\ (6,1) & - & - & - & - & (6,6) \end{matrix}\right\}$

and *n(s)* = 36. Let X be a r.v. with image set $X(S) = \{1\,, 2, 3, 4, 5, 6\}$

$$P(X=1)=P\{1,1\}=\frac{1}{36}$$

$$P(X=2)=P\{(2,1),(2,2),(1,2)\}=\frac{3}{36}$$

Discrete Random Variable: A variable which can assume only a countable number of real values and for which the value which the variable takes depends on chance (a real valued function defined as a discrete sample space is called discrete random variable).

Probability Mass Function: If *X* is a discrete random variable, with distinct values$x_1, x_2, \dots \dots . x_n$,… then the function *p(x)* defined as

$$p_X(x)=\begin{cases} P(X=x_i)= \ p_i\,; \ if \ \ x=x_i \\ 0 \qquad\qquad\quad ; \quad if \ x \neq \ \ x_i \end{cases}$$

Continuous Random Variable: A random variable *X* is said to be continuous if it can take all possible values between certain limits. e**.g**. age, height, weight etc.

Probability density function:

$$f_X(x)= \lim_{\delta x \to 0} \frac{P(x \le X \le x+\delta x)}{\delta x}$$

Distribution Function: Let *X* be a random variable. The function F defined for all real *x* by

$$F(x)=P(X \le x)=P\{\omega : X(\omega) \le x\}, \qquad -\infty < x < \infty$$

is called the distribution function of the random variable *X*.

Example 23: A random variable *X* has the following probability function:

X=*x* :	0	1	2	3	4	5	6	7
p(x) :	0	*k*	*2k*	*2k*	*3k*	k^2	$2k^2$	$7k^2+k$

i) Find *k*

ii) Evaluate *P(X<6)*, *P(X≥6)* and *P(0<X<5)*,

iii) Determine the distribution function of X.

Since $\sum_{x=0}^{7} p(x)=1$, we have

i) $$k+2k+3k+k^2+2k^2+7k^2+k=1$$

$$10k^2+9k-1 \quad =10$$

$$(10k-1)(k+1)=0$$

⇨ $$k = 1/10$$

ii) $P(X < 6) = P(X = 0) + P(X = 1) + \cdots + P(X = 5)$

$$= \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} = \frac{81}{100}$$

$$\mathrm{P}(X \geq 6) = 1 - P(X < 6) = 19/100$$

$$P(0 < x < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 8k = \frac{4}{5}$$

iii)

X=*x*:	0	1	2	3	4	5	6	7
$F_x(x) = P(X \leq x)$:	0	1/10	3/10	5/10	4/5	81/100	83/100	1

2.3 Probability Distributions

Bernoulli Distribution

A random variable *X* is said to have a Bernoulli distribution with parameter *p*, if its probability mass function is given by

$$P(X = x) = \begin{cases} p^x(1-p)^{1-x}\ , & for\ x = 0{,}1 \\ 0 & otherwise \end{cases}$$

A random experiment whose outcomes, are of two types, success *S* and failures *F*, occurring with probabilities p and q respectively, is called a Bernoulli trial. If for this experiment, a random variable *X* is defined such that if takes values 1 for *S* and 0 for F then *X* follows Bernoulli distribution.

Let 'p' be the probability of success and 'q' be the probability of failure.

$$E[X] = \sum_{x=0}^{1} xP(X = x) = 0.P(x = 0) + 1.P(x = 1) \\ = 0.q + 1.p$$

$$E[X] = p.$$

$$E[X^2] = \sum_{x=0}^{1} x^2P(X = x) = 0.P(x = 0) + 1.P(x = 1) \\ = 0.q + 1.p \\ = p$$

$$V[X] = E[X^2] - \{E[X)\}^2 = p - p^2 = p(1-p) = pq$$

The binomial distribution is called by this name because the values of the probabilities are the success terms of binomial experiment $(q+p)^n$

Binomial Distribution

Binomial distribution was introduced by James Bernoulli. Consider n independent Bernoulli trial. Let occurrence of an event be success (S) Let occurrence of an event be failure (F). Let p be the probability of success in any trial (and is constant for each trial). Let $q= 1-p$ be the probability of failure in any trial.

The probability of x successes and consequently (n-x) failures in n independent trails is a specified order (say) *SSFSFFS……FSF*

$$P(SSFSFFS\ldots\ldots FSF) = P(S)P(S)P(F)\ldots\ldots P(F)$$

$$= p.p.q\ldots\ldots q$$

$$= p^x q^{n-x.} \qquad (x \text{ factors and } (n-x) \text{ factors})$$

The x successes in n trials occurs in nc_x ways and the probability for each of these ways is same $p^x q^{n-x}$.

Hence the probability of x successes in n trials in any order is given by addition theorem $nc_x\, p^x q^{n-x}$.

A random variable X is said to follow Binomial distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = \begin{cases} nc_x\, p^x q^{n-x}; & x = 0,1,2,\ldots,\mathrm{n}; \mathrm{q} = 1 - \mathrm{p} \\ 0 & otherwise \end{cases}$$

The mean of Binomial distribution is $E[X] = np$ and the variance is $V[X] = npq$

Example 24: If the probability is 0.70 that a student with very high grades will get into law school, what is the probability that 3 of 5 students with very high grades will get law school?

X=3, n=5, p=0.70

$$P[X = 3] = 5c_3\ (0.70)^3 (0.30)^2 \quad = 0.3087$$

Example 25: If the probability is 0.60 that a person shopping at a certain market will spend at least Rs.100. Find the probability that among five persons shopping at this market 0, 1, 2, 3, 4 or 5 will spend at least Rs.100.

x	$P(X = x) = nC_x p^x q^{n-x}$
0	0.010
1	0.077
2	0.230
3	0.346
4	0.256
5	0.078

Example 26: Find the probability of getting exactly 2 heads in 6 tosses of a fair coin.

Probability of getting head (*p*) = 1/2 , *q* = 1/2 , *n* = 6

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

$$P(X = 2) = \binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{6-2} = \frac{15}{64}$$

Example 27: The mean and standard deviation of a Binomial distribution is 20 and 4 respectively. Find *n*, *p* and q.

Given *np* = 20 $\sqrt{npq}$ = 4

q = 16/20 = 0.8, *p* = 1 – *q* = 0.2, *n* = 20/0.2 = 100

Poisson Distribution

Poisson distribution was discovered by French mathematician and physicist Simeon Denis Poisson. Poisson distribution is a limiting case of Binomial distribution under the condition

1. *n* is large. (*n*→ ∞)
2. *p* (constant probability of success for each trail is very small)
3. *np*=λ is finite.

A random variable *X* is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} ; & x = 0,1,2 \ldots.. \\ 0 & otherwise \end{cases}$$

E[X] = λ *V[X]* = λ

Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials of an experiment but which occur at random points of time and space. The following are some examples where Poisson distribution is used.

(i) Number of deaths due to a disease

(ii) Number of printing mistakes at each page of the book

(iii) Number of defective material in a packing of manufactured item.

(iv) Number of telephone calls received at a particular telephone exchange

(v) Number of cars passing a crossing per minute during busy hours of a day.

Example 28: A very large shipment of books contains 2% with defective bindings. Find the probability that among 400 books taken at random from this shipment only five will have defective bindings.

n=400, x=5, λ= np=8, p=0.02

$$P(X = x) = \begin{cases} \dfrac{e^{-8}8^5}{5!} \end{cases} = 0.093$$

Example 29: Suppose a telephone exchange receives telephone calls at the rate of 3 calls per minutes on an average. Find the probability of receiving at most one call in one minute.

$$P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} \ ; & x = 0,1,2 \ldots\ldots \\ 0 & otherwise \end{cases}$$

$$P(X \ \leq 1) = \frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} = 0.19915$$

Example 30: A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality?

n=100.

Let p= probability of a defective pin = 5% = 0.05

λ = mean number of defective pins in a box of 100

= np =100*0.005 =5

Since ‘p’ is small, we may use Poisson Distribution.

Probability of x defective pins in a box of 100 is

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}\quad=\frac{e^{-5}5^{x}}{x!};\ x=0,1,2,\ldots$$

Probability that a box will fail to meet the guaranteed quality is

$$P(X>10)=1-P(X\leq 10)=1-\sum_{x=0}^{10}\frac{e^{-5}5^{x}}{x!}=1-e^{-5}\sum_{x=0}^{10}\frac{5^{x}}{x!}$$

Example 31: If *X* and *Y* are independent Poisson variates such that

$$P(X=1)=P(X=2)$$

$$P(Y=2)=P(Y=3)$$

Find the variance of *X-2Y*

Let *X~ P(λ) and Y~P(μ).*

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!},\ \ x=0,1,2,3,\ldots;\ \lambda>0$$

$$P(Y=y)=\frac{e^{-\mu}\mu^{y}}{x!},\ \ y=0,1,2,3,\ldots;\mu>0$$

$$\lambda e^{-\lambda}=\frac{\lambda^{2}e^{-\lambda}}{2!}\qquad and\qquad \frac{\mu^{2}e^{-\mu}}{2!}=\frac{\mu^{3}e^{-\mu}}{3!}$$

$$\lambda e^{-\lambda}[\lambda-2]=0\quad and\quad \mu^{2}e^{-\mu}[\,\mu-3]=0$$

$\lambda=2\ and\ \ \mu=3,$ since $\lambda>0,\mu>0.$

Var(*X*)= $\lambda=2, and\ Var(Y)=\ \mu=3$

Covariance term vanishes since X and Y are independent.

Var(*X* - *2Y*) = 2 + 4 *3 = 14

Example 32: The average number of trucks arriving on any one day at a truck depot in a certain city is 12. What is the probability that on a given day fewer than 9 trucks will arrive at the depot?

Let X be the number of trucks arriving on a given day.

P(X < 9) = $\sum_{x=0}^{8}\frac{e^{-12}\lambda^{x}}{x!}$ = 0.1550

Normal Distribution

Normal distribution is a continuous distribution. It is most important probability distribution in statistical analysis. A random variable X is said to have a normal distribution with parameters μ (mean) and σ^2 (variance) if the density function is given by

$$f(x;\ \mu,\sigma^{2})=\frac{1}{\sqrt{2\pi}\ \sigma}\exp\left[-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^{2}\right];\ -\infty<x<\infty,-\infty<\ \mu<\ \infty,\sigma^{2}>0$$

Chief characteristics of Normal distribution:

(i) The normal curve is bell shaped and symmetrical.

(ii) Mean, median and mode of the distribution coincide.

(iii) $\beta_1 = 0$ and $\beta_2 = 3$.

(iv) X-axis is an asymptote to the curve.

(v) Area property :

a. $P(\mu - \sigma < X < \mu + \sigma) = 0.6826$

68.26% of the values of a normal random variable are within plus or minus one standard deviation of its mean.

b. $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544$

95.44% of the values of a normal random variable are within plus or minus two standard deviation of its mean.

c. $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

99.73% of the values of a normal random variable are within plus or minus three standard deviation of its mean.

(vi) The entire family of normal probability distribution is differentiated by its mean μ and its standard deviation σ.

(vii) The highest point on the normal curve is at the mean.

A random variable that has a normal distribution with mean '0' and standard deviation '1' is said to have standard normal probability distribution. The letter 'Z' is used to denote a standard normal random variable. The probability calculation with any normal probability distribution is made by computing the area under the graph of the probability density function. The area under the normal curve of standard normal distribution is computed and is available in tables that can be used in computing probabilities. The formula used to convert any normal random variable X with mean μ and standard deviation σ to the standard normal distribution is $Z = \frac{X-\mu}{\sigma}$.

Example 33: *X* is normally distributed and the mean of *X* is 12 and S.D. is 4

(a) Find out the probability of the following:

i) $X \geq 20$, ii) $X \leq 20$, and iii) $0 \leq X \leq 12$

(b) find x', when $P(X > x') = 0.24$.

(c) Find x_0' $and\ x_1', when\ P(x_0' < X < x_1') = 0.50\ and\ P(X > x_1') = 0.25$

(a) $\mu = 12, \sigma = 4, i.e., X \sim N(12,16)$.

i) When X = 20, $Z = \frac{20-12}{4} = 2$

$P(X \geq 20) = P(Z \geq 2) = 0.5 - P(0 \leq Z \leq 2) = 0.5 - 0.4772 = 0.0228$

ii) $P(X \leq 20) = 1 - P(X \geq 20)$ = 1 − $0.0228 = 0.9772$

iii) $P(0 \leq X \leq 12) = P(-3 \leq Z \leq 0)$ $\left(Z = \frac{X-12}{4}\right)$

$= P(0 \leq Z \leq 0) = 0.49865$

(b) When X > $x^{'}, Z = \frac{x^{'}-12}{4} = z_1(say)$

then, we are given

$$P\left(X > x^{'}\right) = 0.24 \quad => \quad P(Z > z_1) = 0.24 \;\; i.e., P(0 < Z < z_1) = 0.26$$

From tables

$$z_1 = 0.71$$

Hence

$$\frac{x^{'} - 12}{4} = 0.71 \qquad => \quad x_1^{'} = 12 + 4 \times 0.71 = 14.84$$

(c) We are given

$$P\left(x_0^{'} \;\; < X < x_1^{'}\right) = 0.50 \;\; and \; P\left(X > x_1^{'}\right) = 0.25$$

When $X = x_1^{'}$, $\quad Z = \frac{x^{'}-12}{4} \; = z_1$

And when $X = x_0^{'}$, $\quad Z = \frac{x^{'}-12}{4} \; = -z_1$

We have

$$P(Z > z_1) = 0.25 \quad => P(0 < Z < z_1) = 0.25$$

$$z_1 = 0.67$$

Hence $\quad \frac{x^{'}-12}{4} = 0.67 \qquad => \quad x_1^{'} = 12 + 4 \times 0.67 = 14.68$

$\frac{x^{'}-12}{4} = -0.67 \qquad => \quad x_1^{'} = 12 - 4 \times 0.67 = 9.32$

Example 34: X is a normal variate with mean 30 and S.D. 5. Find the probabilities that

$i)\; 26 \leq X \leq 40, \;\; ii) X \geq 45, \qquad and \;\; |X - 30| > 5$

Here $\mu = 30, \sigma = 5.$

i) When $X = 26, \quad Z = \frac{X-\mu}{\sigma} \; = \; \frac{26-30}{5} \; = \; -0.8$

and when $X = 40, \;\; Z = \; \frac{40-30}{5} \; = 2$

$$P(26 \le X \le 40) = P(-0.8 \le Z \le 2)$$
$$= P(-0.8 \le Z \le 0) + P(0 \le Z \le 2)$$
$$= P(-0.8 \le Z \le 0) + 0.4772$$
$$= P(0 \le Z \le 0.8) + 0.4772$$
$$= 0.2881 + 0.4772 \quad = 0.7653$$

ii) $P(X \ge 45) = ?$

$$X = 45, \qquad Z = \frac{X-\mu}{\sigma} = \frac{45-30}{5} = 3$$

$$P(X \ge 45) = P(Z \ge 3) = 0.5 - P(0 \le Z \le 3)$$
$$= 0.5 - 0.49865 = 0.00135$$

iii) $\mathrm{P}(|X-30| \le 5) = P(25 \le X \le 35) = P(-1 \le Z \le 1)$

$$= 2P(0 \le Z \le 1) = 2 \times 0.49865 = 0.00135$$

$$\mathrm{P}(|X-30| > 5) = 1 - P(|X-30| \le 5)$$
$$= 1 - 0.6826 \ = 0.3174$$

Normal distribution is most important of all the distributions in statistical theory. Most of the discrete distributions namely, Binomial, Poisson distributions tend to normal distribution as the sample size (*n*) increases. Almost all sampling distributions chi-square, Student's-t and Snedecor's F tend to normal for large degrees of freedom.

Questions and Exercises

1. Define mutually exclusive events and exhaustive events with example.
2. Define random experiment and sample space with suitable examples.
3. An experiment has five possible outcomes I, II, III, IV, V, that are mutually exclusive. Check whether the following assignments of probabilities are permissible. State reasons.

(a)	P(I) = 0.21	P(II) = 0.26	P(III) = 0.58	P(IV) = 0.01	P(V) = 0.06
(b)	P(I) = 0.18	P(II) = 0.20	P(III) = 0.22	P(IV) = 0.19	P(V) = 0.21
(c)	P(I) = 0.10	P(II) = 0.10	P(III) = 0.30	P(IV) = 0.60	P(V) = -0.10

4. What is the probability of getting a 9 exactly once in 3 throws with a pair of dice?
5. From a bag containing 10 black and 5 white balls, a ball is drawn at random. What is the probability that it is white?
6. A bag contains 10 white, 6 red, 4 black and 7 blue balls. 5 balls are drawn at random. What is the probability that 2 of them are red and one black?
7. An urn contains 3 red and 4 black balls. Two balls are taken out at random. Find the probabilities that the balls are (*i*) different colours, (*ii*) black colour, (*iii*) red colour.

8. If 3% of electric bulbs manufactured by a company are defective, find the probability that in a sample of 100 bulbs (a) 0 (b) 3 (c) 5 bulbs are defective.

9. Two cards are randomly drawn from a deck of 52 playing cards. Find the probability that both cards will be greater than 3 and less than 8.

10. State and prove the addition theorem of probability for two events when they are not disjoint.

11. The odds in favour of certain event are 5:8, and odds against another event are 4:3. What is the chance that at least one of them will happen?

12. The chance of getting an order by three salesmen A, B and C are 1/3, 3/4 and 1/2 respectively. What is the probability that any one of them will get the order?

13. Suppose that two people are randomly selected from a group of 4 women and 6 men

    (a) What is the probability that both are women?

    (b) What is the probability that one is a woman and other a man?

14. In a bolt factory, machines A, B and C manufacture respectively 25%, 35% and 40% of the total. Of their output respectively 5, 4, 2 percents are defective. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines A, B and C?

15. Three newspapers A, B and C are published in a city and a recent survey of readers indicate the following:

    20% read A ; 16% read B ; 14% read C,
    8% read both A and B ; 5% read both A and C ;
    4% read B and C ; 2% read all three.

    For one person chosen at random, compute the probability that

    i) the person reads none of the papers,

    ii) the person reads exactly one of the papers, and

    iii) the person reads at least A and B if it is known that he reads at least one of the papers.

16. Three urns are given each containing red and white chips as indicated.

    Urn 1: 6 red and 4 white
    Urn 2: 2 red and 6 white
    Urn 3: 1 red and 8 white

    i) An urn is chosen at random and a ball is drawn from the urn. The ball is red. Find the probability that the urn chosen was urn 1.

ii) An urn is chosen at random and two balls are drawn without replacement from this urn. If both balls are red, find the probability that urn I was chosen. Under these conditions, what is the probability that urn III was chosen.

17. State Bayes' theorem. The members of a consulting firm rent cars from three rental agencies: 60 percent from agency 1, 30 percent from agency 2, and 10 percent from agency 3. If 9 percent of cars from agency 1 need a tune-up, 20 percent of the cars from agency 2 need a tune-up and 6 percent of cars from agency 3 need a tune-up. What is the probability that a rental car delivered to the firm will need a tune-up?

18. An electronic company manufactures MP3 players at three locations. The company at location I manufactures 50% of the MP3 players and 1% are defective. The company at location II manufactures 30% and 2% are defective. The company at location III manufactures 20% and 3% are defective. If an MP3 player is selected at random, what is the probability that it is defective?

19. From a lot of 10 items containing 3 defectives, a sample of 4 items is drawn at random. Let the random variable *X* denote the number of defective items in the sample. Answer the following, when the sample is drawn without replacement.

i) Find the probability distribution of *X*.

ii) Find $P(X\leq 1)$, $P(X<1)$ and $P(0<X<2)$

20. In a survey of 500 persons were asked questions

(1) Do you own a cell phone?

(2) Do you own an ipod?

(3) Do you have internet connection?

The results of the survey were (no one answered no to all the three questions)

Cell phone	329
ipod	186
internet connection	295
cell phone and ipod	83
cell phone and internet connection	217
ipod and internet connection	63

Find the probability for the following events

(a) Answered yes to all three questions

(b) Had a cell phone but not an internet connection

(c) Had a cell phone but not an ipod or an internet connection

(d) Had an internet connection but not an ipod.

21. A random variable X has the following probability distribution:

X= x:	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
p(x):	*a*	*3a*	*5a*	*7a*	*9a*	*11a*	*13a*	*15a*	*17a*

i) Determine the value of *a*.

ii) Find $P(X<3), P(X>3), P(0<X<5)$.

22. Define Binomial distribution. The mean and variance of a binomial variate X with parameters *n* and *p* are 16 and 8. Find (i) P (X = 0) (ii) P (X ≥ 2)

23. A machine manufacturing screws is known to produce 5% defectives. In a random sample of 15 screws, what is the probability that there are

i) Exactly three defectives?

ii) not more than three defectives?

24. A multiple choice examination has 3 possible answers for each of 5 questions. What is the probability that a student will get 4 or more correct answers just by guessing?

25. An automobile safety engineer claims that 1 in 10 accidents is due to driver fatigue. What is the probability that at least 3 of 5 automobile accidents are due to driver fatigue?

26. The number of complaints that a establishment receives per day is a random variable having Poisson distribution with $\lambda = 3.3$. Find the probability that it will receive only 2 complaints on any given day.

27. Suppose that 2 batteries are randomly chosen from a bin containing 12 batteries, of which 8 are good and 4 are defective. What is the expected number of defective batteries chosen?

28. Define normal distribution and state its various properties.

29. The daily trading volumes (millions of shares) for stocks traded in a leading stock exchange for 12 days are given below.

917, 983, 1046, 944, 723, 783, 813, 1057, 766, 836, 992, 973

Assuming the probability distribution of trading volume to be normal, compute the probability that on a particular trading day, the trading volume will less than (i) 800 million shares and (ii) will be between 900 and 1020 million shares.

30. The time needed to complete a final examination in a particular college course is normally distributed with mean of 80 minutes and a standard deviation of 10 minutes.

a. What is the probability of completing the exam in one hour or less?

b. What is the probability that a student will complete the exam in more than 60 minutes but less than 75 minutes?

UNIT - III

CORRELATION AND REGRESSION ANALYSIS

Contents:

3.1 – Regression Analysis

3.2 – Correlation Analysis

3.1 Regression Analysis

Managerial decisions often are based on the relationship between two or more variables. One such technique is the correlation analysis and modeling these related variables is the Regression Analysis. The main purpose of regression analysis is prediction. Usually, the regression model is used to predict the values of a dependent variable based on the values of one or more independent variables. Here, the discussion is limited to a two variable model with one dependent (Y) and one independent (X) variable.

In general, the scatter diagram will be used to plot and explain the relationship between an X variable on horizontal axis and a Y variable on the vertical axis. The nature of relationship between two variables can take wither of the positive, negative and absence of relationships. The simplest form of the relationship is the straight line or linear relationship. Apart from straight line form, there are many other forms like exponential, power, cubic, parabolic and many more to explain the nature of the data.

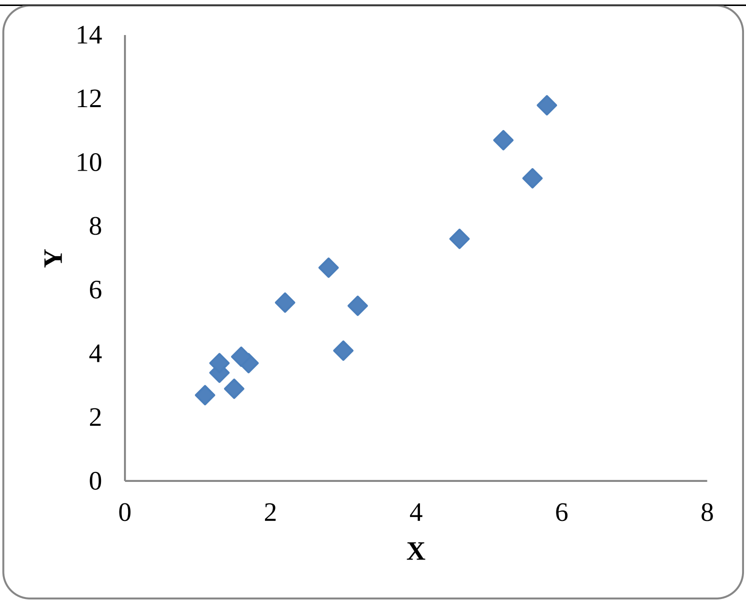


Figure 3.1: Scatter Diagram showing linear relationship between X and Y variables

The straight line (linear) model can be represented as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad ; i = 1,2,\ldots, n \tag{3.1}$$

where

Y_i is the response in the ith observation

X_i is the ith value of the independent variable

$\beta_0$and β_1 are the intercept and slope

ε_i is the random error and

n is the number of observations

The estimated regression equation can be obtained using least squares method.

Least Squares Method

Simple linear regression analysis is concerned with finding the straight line that fits the data best. This can be addressed in two ways:

(a) Finding the straight line for which the difference between the actual values Y_i and the values that would be predicted from the line of regression.

(b) However, because these difference are positive and negative for some observations instead, an approach called, the least squares method that minimizes the sum of the squared differences is used.

i.e., $\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2 = \sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$

here, $\widehat{Y_i} = \beta_0 + \beta_1 X_i$

β_0 and β_1 are the unknown constants which are to be estimated. Using the least square method, one can determine the values of β_0 and β_1 that minimizes the sum of squared differences. The estimated regression equation and the expressions for determining β_0 and β_1 are given below:

$$\widehat{y_i} = b_0 + b_1 x_i \; ; i = 1,2,\ldots,n \tag{3.2}$$

where

y_i is the estimated value of the ith observation

b_0 is the y intercept of the estimated regression line or model

b_1 is the slope of the estimated regression line or model

x_i is the value of the ith observation of independent variable

In other words, the least square criterion can be given as min $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$

where

y_i is the observed value of the dependent variable for the ith observation

$\widehat{y_i}$ is the estimated value of the dependent variable for the ith observation

The expressions for b_0 and b_1 can be derived using normal equations and are given below:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \; or \; b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (3.3)$$

and $b_0 = \bar{y} - b_1\bar{x}$ (3.4)

here

x_i is the value of the independent variable for ith observation

y_i is the value of the dependent variable for ith observation

$\bar{x}$ is the mean value for the independent variable = $\frac{\sum_{i=1}^{n} x_i}{n}$

$\bar{y}$ is the mean value for the dependent variable = $\frac{\sum_{i=1}^{n} y_i}{n}$

n is the total number of observations

In order to have a better understanding of the concept of regression, let us have an numerical illustration using the following data on store size and its annual sales.

Problem:

It is to examine the relationship between the store size (i.e., square footage) and its annual sales, a sample of 14 stores was selected.

Store	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Square footage (in 1000's)	1.7	1.6	2.8	5.6	1.3	2.2	1.3	1.1	3.2	1.5	5.2	4.6	5.8	3.0
Annual sales (in lakhs)	3.7	3.9	6.7	9.5	3.4	5.6	3.7	2.7	5.5	2.9	10.7	7.6	11.8	4.1

For illustration and computational purpose, prepare the following table:

Table 3.1 Computation table

Store	Square footage (in 1000's)	Annual sales (in lakhs)	x^2	y^2	xy
1	1.7	3.7	2.89	13.69	6.29
2	1.6	3.9	2.56	15.21	6.24
3	2.8	6.7	7.84	44.89	18.76
4	5.6	9.5	31.36	90.25	53.2
5	1.3	3.4	1.69	11.56	4.42
6	2.2	5.6	4.84	31.36	12.32
7	1.3	3.7	1.69	13.69	4.81
8	1.1	2.7	1.21	7.29	2.97

9	3.2	5.5	10.24	30.25	17.6
10	1.5	2.9	2.25	8.41	4.35
11	5.2	10.7	27.04	114.49	55.64
12	4.6	7.6	21.16	57.76	34.96
13	5.8	11.8	33.64	139.24	68.44
14	3	4.1	9	16.81	12.3
Total	**40.9**	**81.8**	**157.41**	**594.9**	**302.3**

Here, it clearly indicates that increasing store size has increased annual sales (figure 3.2). So, store size (square foot in 1000's) is independent variable, X and the annual sales purely depends on the square foot of store size, Y.

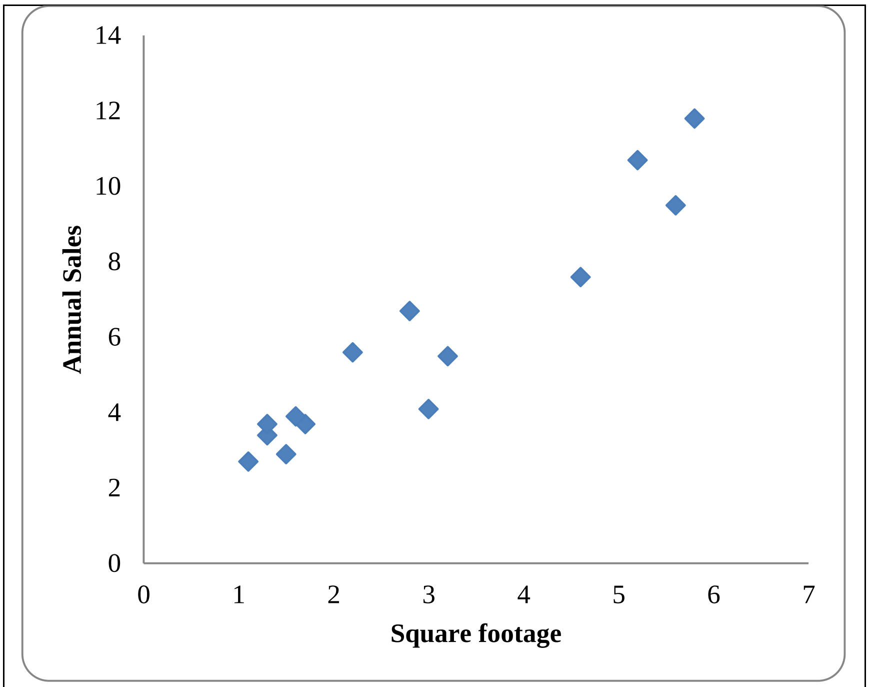


Figure 3.2: Graphical view of linear relationship between square footage and its annual sales

Using (3.3) and (3.4), the values of b₀ and b₁ can be computed as

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

$$b_1 = \frac{302.3 - \frac{(40.9)(81.8)}{14}}{157.41 - \frac{(40.9)^2}{14}}$$

$$b_1 = \frac{302.3 - 238.9728}{157.41 - 119.4864}$$

$$b_1 = \frac{63.32715}{37.92358}$$

$$\Longrightarrow b_1 = 1.66986$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = 5.84285 - (1.66986) * (2.92143)$$

$$\Longrightarrow b_0 = 0.96447$$

Therefore, b_0= 0.96447 and b_1=1.66986

So, the estimated regression equation is

Annual sales = 0.96447 + 1.66986 (square footage)

The slope of the estimated regression (b_1 = 1.66986) is positive, implying that as square foot of store size increases, annual sales increases. In fact, we can conclude that an increase in square footage of store size of 2000 sq. ft is associated with an increase of 3.324 times in annual sales. That is, if we wish to predict the annual sales for a store size of 2000 (=x) square feet, we would have

Annual sales = 0.96447 + (1.66986) (2000)

Annual Sales = 0.96447 + 3339.72

Therefore, Annual Sales = 3340.6844

Hence, we would predict annual sales of $ 3340.6844 for this particular store size.

Once, the regression equation is estimated, the next step is to compute the coefficient of determination, R^2, is given by

$$R^2 = \frac{sum\ of\ squares\ due\ to\ regression}{total\ sum\ of\ squares} = \frac{SSR}{SST}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{b_0\sum_{i=1}^{n} y_i + b_1\sum_{i=1}^{n} xy - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}$$

For the considered problem, R^2 can be computed as follows

$$\text{SSR} = (0.964478) * (81.8) + (1.66986) * (302.3) - \frac{(81.8)^2}{14} = 105.74726$$

SST = 594.9 - $\frac{(81.8)^2}{14}$

SST = 594.9 – 477.94571

SST = 116.9542

$$\therefore\ R^2 = \frac{SSR}{SST} = \frac{105.74726}{116.9542} = 0.9042$$

The ratio SSR/SST, which will take values between zero and one, is used to evaluate the goodness of fit for the estimated regression equation.

R^2 can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation. That is, for the considered problem, the R^2 value is 0.9042, with this values, we can conclude that 90.42 % of the total sum of squares can be explained by using the estimated regression equation, $\hat{y} = 0.964478 + 1.66986\, x$ to predict annual sales. In other words, 90.42% of variability in annual sales can be explained by the size of the store and annual sales.

Let us consider another example to explain the procedure of estimating regression equation in a better manner. Data is collected from a sample of 10 restaurants located near college campuses. The data is presented in the following table:

Restaurant	Student Population (in 1000's)	Quarterly sales (in 1000's)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Here from the given information, student population influences the quarterly sales of the restaurants, hence, student population can be considered as independent variable (x) and quarterly sales as dependent variable (y). Now, let us compute the following terms such as mean of x and y, $(x_i - \bar{x})$, $(y_i - \bar{y})$, $(x_i - \bar{x})(y_i - \bar{y})$ and $(x_i - \bar{x})^2$.

Restaurant	Student Population (in 1000's)	Quarterly sales (in 1000's)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Total	**140**	**1300**			**2840**	**568**
Mean	**14**	**130**				

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

$b_0 = \bar{y} - b_1\bar{x}$

$b_0 = 130 - 5 * 14$

$b_0 = 60$

Thus, the estimated regression equation is $\hat{y} = 60 + 5x$

In similar lines, the interpretation about the parameters of the estimated regression equation can be given as follows:

The slope, $b_1 = 5$ and is positive, which implies that as student population increases, sales increases. In fact, we can conclude that an increase in the student population of 1000 is associated with an increase of \$5000 in expected sales. Suppose if we wish to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute the quarterly sales, $\hat{y}$ as, $\hat{y}$=60+5 (16) = 140. Hence, we would predict quarterly sales of \$140,000 for that particular restaurant. Further, the coefficient of determination can be computed as R^2 = SSR/SST = 14,200/15,730 = 0.9027

We can conclude that 90.27% of the total sum of squares can be explained by using the estimated regression equation. In other words, 90.27% of the variability in quarterly sales can be explained by the linear relationship between the size of the student population and quarterly sales.

3.2 – Correlation Analysis

Correlation is another way of assessing the relationship between variables. To be more precise, it measures the extent of correspondence between the ordering of two random variables. There is a large amount of resemblance between regression and correlation but for their methods of interpretation of the relationship. For example, a *scatter diagram* is of tremendous help when trying to describe the type of relationship existing between two variables.

Measuring correlation

We make use of the *linear product-moment correlation coefficient*, also known as *Pearson's correlation coefficient*, to express the strength of the relationship. This coefficient is generally used when variables are of *quantitative* nature, that is, ratio or interval scale variables.

Pearson’s correlation coefficient is denoted by *r* and is defined by

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2 * \sqrt{n\sum y^2 - (\sum y)^2}}}$$

The value of *r* always lies between –1 and 1 inclusive, that is, $-1 \leq r \leq 1$. If *Y* increases when *X* increases, we say that there is *positive* or *direct* correlation between them. However, if *Y* decreases when *X* increases (or *vice versa*), then we say that they are *negatively* or *inversely* correlated. The reader must have noticed that *direct* and *inverse* are terms that are used in the context of variation or *proportionality*.

Interpretation of the correlation coefficient

The extreme values of *r*, that is, when *r* = ±1, indicate that there is *perfect* (positive or negative) correlation between *X* and *Y*. However, if *r* is 0, we say that there is *no* or *zero* correlation. When *r* = 0, we may *not* assert that there is no correlation *at all* between *X* and *Y*. Pearson’s correlation coefficient is meant to measure *linear* relationship only. It should *not* be used in the case of *non-linear* relationships since it will obviously lead to an erroneous interpretation.

The remaining values, falling in subintervals of [–1, 1], describe the relationship in terms of its *strength*.

We observe that the strength of the relationship between *X* and *Y* is the same whether *r* = 0.85 or – 0.85. The only difference is that there is *direct* correlation in the first case and *inverse* correlation in the second. We should bear in mind that *r* is the *linear* correlation coefficient and that, as mentioned earlier, its value can be wrongly interpreted whenever the relationship between *X* and *Y* is non-linear. That is the reason why we should have a look at a scatter diagram of points (*x*, *y*) and verify whether the relationship is, for example, of quadratic, logarithmic, exponential or trigonometric (briefly, *non-linear*) nature.

If *r* = 0, we should not jump to the conclusion that there is no correlation at all between *X* and *Y*. Consider the case where there is *perfect* (but unsuspected) *non-linear* correlation between the two variables.

With practice and experience, it is even possible to know approximately the value of *r* by inspection of a scatter diagram. The location (amount of scattering) of the points with respect to the least-squares regression line indicates the strength of the relationship between the variables. The *more scattered* the points are, the *weaker* is the relationship and the *closer* is the value of *r* to zero.

Example

Determine the degree of correlation between the prime lending rate and the inflation rate and test whether it is significant at 0.05 level.

Inflation rate	3.3	5.8	6.2	6.5	7.6	9.1	11
Prime lending rate	5.2	6.8	8	6.9	9	7.9	10.8

Solution

We first summarise the data from the above table as follows:

Sample number	Inflation rate	Prime lending rate	X^2	Y^2	XY
1	3.3	5.2	10.89	27.04	17.16
2	5.8	6.8	33.64	46.24	39.44
3	6.2	8	38.44	64	49.6
4	6.5	6.9	42.25	47.61	44.85
5	7.6	9	57.76	81	68.4
6	9.1	7.9	82.81	62.41	71.89
7	11	10.8	121	116.64	118.8
Totals	$\sum x$ =**49.5**	$\sum y =$ **54.6**	$\sum x^2$ =**386.79**	$\sum y^2$ =**444.94**	$\sum xy$ =**410.14**

Pearson's correlation coefficient is calculated as

$$r = \frac{(7)(410.14) - (49.5)(54.6)}{\sqrt{\left[(7)(386.79) - (49.5)^2\right]\left[(7)(444.94) - (54.6)^2\right]}} = 0.9082$$

Hence, there is a *very strong direct* correlation between inflation rate and prime lending rate.

EXERCISES

Problems on Correlation

1. Five observations taken for two variables follow.

x_i	4	6	11	3	16
y_i	50	50	40	60	30

Compute and interpret the sample correlation coefficient. At α = 0.05 test for significant correlation.

2. Nielsen Media Research provides two measures of the television viewing audience: a television program rating, which is the percentage of households with televisions watching a program, and a television program share, which is the percentage of households watching a program among those with televisions in use. The following

data show the Nielsen television ratings and share data for the Major League Baseball World Series over a nine-year period (Associated Press, October 27, 2003).

Rating	19	17	17	14	16	12	15	12	13
Share	32	28	29	24	26	20	24	20	22

Compute the sample correlation coefficient. What does this value tell us about the relationship between rating and share? At α = 0.05 level, does there appear to be a relationship between rating and share?

3. A department of transportation’s study on driving speed and miles per gallon for midsize automobiles resulted in the following data:

Speed (Miles per Hour)	30	50	40	55	30	25	60	25	50	55
Miles per Gallon	28	25	25	23	30	32	21	35	26	25

Compute and interpret the sample correlation coefficient. Test the correlation at α = 0.10.

4. The Dow Jones Industrial Average (DJIA) and the Standard & Poor’s 500 Index (S&P 500) are both used to measure the performance of the stock market. The DJIA is based on the price of stocks for 30 large companies; the S&P 500 is based on the price of stocks for 500 companies. If both the DJIA and S&P 500 measure the performance of the stock market, how are they correlated? The following data show the daily percent increase or daily percent decrease in the DJIA and S&P 500 for a sample of nine days over a three-month period (The Wall Street Journal, January 15 to March 10, 2006).

DJIA	0.2	0.82	-0.99	0.04	-0.24	1.01	0.3	0.55	-0.25
S&P 500	0.24	0.19	-0.91	0.08	-0.33	0.87	0.36	0.83	-0.16

Compute the sample correlation coefficient for these data and test at α = 0.05.

5. The daily high and low temperatures for 14 cities around the world are shown (The Weather Channel, April 22, 2009).

City	High	Low	City	High	Low
Athens	68	50	London	67	45
Beijing	70	49	Moscow	44	29
Berlin	65	44	Paris	69	44
Cairo	96	64	Rio de Janeiro	76	69
Dublin	57	46	Rome	69	51
Geneva	70	45	Tokyo	70	58
Hong Kong	80	73	Toronto	44	39

What is the correlation between the high and low temperatures? Is the observed correlation significant at α = 0.05? Discuss.

6. Education and crime rate ratings for selected US cities are given below. Education rating is an index for public/teacher ratio, academic options in higher education the higher the rating the better and other factors and crime is the crime rate per 100 people

City	Education (x) ordered data	Crime (y)
New York	30	25
Detroit	31	16
Los Angeles	32	20
Boston	35	12
Chicago	35	10
Washington, DC	36	13

Compute and interpret the correlation coefficient and test at α = 0.05.

7. The data below summarized the relationship between number of employees (x) and number of openings (y) at 11 Boston area hospitals.

$\sum x = 56{,}562$; $\sum x^2 = 456{,}525{,}234$; $\sum y = 2611$; $\sum y^2 = 818{,}149$; $\sum xy = 18{,}267{,}023$. Find the correlation coefficient, r and test at α = 0.05.

8. The number of hours of study and the exam scores of 10 students are collected. Find the degree of correlation between the hours invested and scores obtained and test at α = 0.05.

No. of Hrs of Study	4	6	8	4	2	1	5	7	4	6
Exam Score	85	80	92	70	65	60	89	82	81	95

Problems on Regression

9. Given are five observations for two variables, x and y.

xi	1	2	3	4	5
yi	3	7	5	11	14

Construct a simple linear regression model for the data and use the estimated regression equation to predict the value of y when x = 4.

10. The following data were collected on the height (inches) and weight (pounds) of women swimmers.

Height	68	64	62	65	66
Weight	132	108	102	115	128

a. Develop the estimated regression equation by computing the values of regression coefficients.

b. If a swimmer's height is 63 inches, what would her weight be?

11. Elliptical trainers are becoming one of the more popular exercise machines. Their smooth and steady low-impact motion makes them a preferred choice for individuals with knee and ankle problems. But selecting the right trainer can be a difficult process. Price and quality are two important factors in any purchase decision. Are higher prices

generally associated with higher quality elliptical trainers? Consumer Reports conducted extensive tests to develop an overall rating based on ease of use, ergonomics, construction, and exercise range. The following data show the price and rating for eight elliptical trainers tested (Consumer Reports, February 2008).

Brand and Model	Price ($)	Rating
Precor 5.31	3700	87
Keys Fitness CG2	2500	84
Octane Fitness Q37e	2800	82
LifeFitness X1 Basic	1900	74
NordicTrack Audiostrider 990	1000	73
Schwinn 430	800	69
Vision Fitness X6100	1700	68
ProForm XP 520 Razor	600	55

Use the estimated regression equation to predict the rating for an elliptical trainer with a price of $1500.

12. The cost of a previously owned car depends upon factors such as make and model, model year, mileage, condition, and whether the car is purchased from a dealer or from a private seller. To investigate the relationship between the car's mileage and the sales price, data were collected on the mileage and the sale price for 10 private sales of model year 2000 Honda Accords (PriceHub website, October 2008).

Miles	(1000s)	90	59	66	87	90	106	94	57	138	87
Price	($1000s)	7	7.5	6.6	7.2	7	5.4	6.4	7	5.1	7.2

a. Use the least squares method to develop the estimated regression equation.

b. Provide an interpretation for the slope of the estimated regression equation.

c. Predict the sales price for a 2000 Honda Accord with 100,000 miles.

13. A sales manager collected the following data on annual sales and years of experience.

Salesperson	Years of Experience	Annual Sales ($1000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. Develop an estimated regression equation that can be used to predict annual sales given the years of experience.

b. Use the estimated regression equation to predict annual sales for a salesperson with 9 years of experience.

14. Bergans of Norway has been making outdoor gear since 1908. The following data show the temperature rating (F°) and the price ($) for 11 models of sleeping bags produced by Bergans (Backpacker 2006 Gear Guide).

Model	Temperature Rating (F°)	Price($)
Ranger 3-Seasons	12	319
Ranger Spring	24	289
Ranger Winter	3	389
Rondane 3-Seasons	13	239
Rondane Summer	38	149
Rondane Winter	4	289
Senja Ice	5	359
Senja Snow	15	259
Senja Zero	25	229
Super Light	45	129
Tight & Light &	25	199

a. Use the least squares method to develop the estimated regression equation.

b. Predict the price for a sleeping bag with a temperature rating (F°) of 20.

15. To avoid extra checked-bag fees, airline travelers often pack as much as they can into their suitcase. Finding a rolling suitcase that is durable, has good capacity, and is easy to pull can be difficult. The following table shows the results of tests conducted by Consumer Reports for 10 rolling suitcases; higher scores indicate better overall test results (Consumer Reports website, October 2008).

Brand	Price ($)	Score
Briggs & Riley	325	72
Hartman	350	74
Heys	67	54
Kenneth Cole Reaction	120	54
Liz Claiborne	85	64
Samsonite	180	57
Titan	360	66
TravelPro	156	67
Tumi	595	87
Victorinox	400	77

a. Use the least squares method to develop the estimated regression equation.

b. Provide an interpretation for the slope of the estimated regression equation.

c. The Eagle Creek Hovercraft suitcase has a price of $225. Predict the score for this suitcase using the estimated regression equation.

16. A personal watercraft (PWC) is a vessel propelled by water jets, designed to be operated by a person sitting, standing, or kneeling on the vessel. In the early 1970s, Kawasaki Motors Corp. U.S.A. introduced the JET SKI® watercraft, the first commercially success-ful PWC. Today, jet ski is commonly used as a generic term for personal watercraft. The following data show the weight (rounded to the nearest 10 lbs.) and the price (rounded to the nearest $50) for 10 three-seater personal watercraft (Jetski News website,2006).

Make and Model	Weight (lbs.)	Price ($)
Honda AquaTrax F-12	750	9500
Honda AquaTrax F-12X	790	10500
Honda AquaTrax F-12X GPScape	800	11200
Kawasaki STX-12F Jetski	740	8500
Yamaha FX Cruiser Waverunner	830	10000
Yamaha FX High Output Waverunner	770	10000
Yamaha FX Waverunner	830	9300
Yamaha VX110 Deluxe Waverunner	720	7700
Yamaha VX110 Sport Waverunner	720	7000
Yamaha XLT1200 Waverunner	780	8500

a. Use the least squares method to develop the estimated regression equation.

b. Predict the price for a three-seater PWC with a weight of 750 pounds.

UNIT - IV

STATISTICAL ESTIMATION AND TESTING

CONTENTS:

4.1 – Point Estimate of the Population Mean and Variance

4.2 – Point Estimate for the population proportion and variance

4.3 – Statistical Hypotheses Testing

In statistical hypotheses testing, the main objective is to determine an appropriate value of a parameter on the basis of a sample statistic. Usually, we can provide two estimates of the populations namely, point and interval estimates. A *Point estimate* is a sample statistic that is used to estimate an unknown population parameter. An *Interval estimate* is a range of values used to estimate a population parameter. In other words, interval estimate indicates the error by the extent of its range and by the probability of the true population parameter lying within that range. An estimator is said to be a good estimator if it possess the qualities of unbiasedness, consistency and efficiency.

4.1 Point Estimate of the Population Mean and Variance

It is already defined that point estimator helps in drawing inferences about a population parameter by estimating the value of an unknown parameter using a single point. Sample mean is the best estimator of the population mean since it is unbiased, consistent and most efficient estimator. However, to claim that the sample mean is best estimator of population mean, either the sample should be drawn from normal population or the sample should be sufficiently large so that the sampling distribution approximates to normal distribution. Therefore, the point estimate of the population mean is given by $\hat{\mu} = \bar{X}$.

A point to be noted that the point estimator gets closer to the population parameter value as sample size increases. However, the degree of error will not be reflected by point estimator and thus we can make use of the concept of interval estimator. By doing this, the probability value of the population parameter will lie between two values with middle value being represented by point estimator. To obtain such interval, the estimate which can be used for the population standard deviation, σ, is the sample standard deviation s, and is given by

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n-1}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}}{n-1}$$

where s = $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$ is the unbiased estimator of the population standard deviation, $\widehat{\sigma}$.

Further, the interval estimates for the estimator of population mean can be observed under two cases, namely with known and unknown σ. In practice, the data which are collected will have sampling variability. This can be addressed in terms of sampling error. With point estimator, we cannot explain the error involved in the samples collected. Our knowledge of sampling error would indicate that the standard error provides a measure of this error in terms of a probability value that the value of the population mean lie within a specified interval. This interval is termed as interval estimate. Let us assume that the sampling distribution of the mean follows normal distribution, then the 95% of the distribution lies within plus or minus $Z\alpha/2$ standard deviation of the mean, i.e., $\bar{X} \pm Z\alpha/_2 S.E(\bar{X}) \Rightarrow \bar{X} \pm Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$

With 95% level, the confidence interval for the mean is $\bar{X} \pm (1.96)\frac{\sigma}{\sqrt{n}}$ and the lower, upper confidence intervals are given by$\left(\bar{X} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}},\ \bar{X} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}\right)$. In general, it is known that the point estimate will lie between these lower and upper confidence intervals.

Interval Estimation of the Population Mean (σ known and unknown)

Let $x_1, x_2, \ldots, x_n$ be the 'n' random samples drawn from a normal population with mean μ and variance σ^2, i.e., $X \sim N(\mu, \sigma^2)$, then the sampling distribution of the sample mean is $\bar{X} \sim N(\mu, \sigma^2)$. The confidence interval of the population mean is given by $\bar{X} - Z\alpha/_2 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$

But in most cases, the population standard deviation would be unknown value and we would use the sample value to estimate the population value. Then the distribution of sample standard deviation underpins students' t - distribution with (n-1) degrees of freedom. Therefore, the 100(1-α)% confidence interval for the population mean when σ is unknown can be given as

$$\bar{X} - t_{n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1}\frac{s}{\sqrt{n}}$$

The above expression is to be used when there are less than 30 samples. Suppose, the samples drawn are greater than (or) equal to 30, then the expression can be rewritten as

$$\bar{X} - Z\alpha/_2 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + Z\alpha/_2 \frac{s}{\sqrt{n}}$$

4.2 Point Estimate for the population proportion and variance

In some situations, the point estimates can be obtained using proportions calculated using a sample instead of using sample mean, we can use the sample proportions to give point estimates of the population proportions and the standard error of the proportion are as follows:

Estimate of the population proportion, $\hat{P} = p$

Estimate of the standard error, $\hat{\sigma}_p = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$

If more than one sample is taken from a population then the population proportion is obtained by the pooled sampled proportion, i.e., $\hat{P} = \frac{n_1p_1+n_2p_2}{n_1+n_2}$

Interval estimate of a population proportion

If the population is normally distributed or the sample size is large, then the confidence interval for a proportion is given by

$$p - Z * \sqrt{\frac{p(1-p)}{n}} \leq P \leq p + Z * \sqrt{\frac{p(1-p)}{n}}$$

Problems on Means and Proportions

1. A random samples of five values was taken from a population, 8.1,6.5, 4.9, 7.3 and 5.9. Estimate the populations mean, standard deviation and standard error of the estimate for the population mean.

   Solution: Given a random sample of five values and let denote it by X, now compute mean and standard deviation for the given data i.e.,

Sample number	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	8.1	1.56	2.4336
2	6.5	-0.04	0.0016
3	4.9	-1.64	2.6896
4	7.3	0.76	0.5776
5	5.9	-0.64	0.4096
Total	**32.7**		**6.112**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} ; n = 5 \; and \; i = 1,2, \dots, 5$$

$$\bar{X} = \frac{8.1 + 6.5 + 4.9 + 7.3 + 5.9}{5}$$

$$\bar{X} = \frac{32.7}{5} = 6.54$$

$$s = \sqrt{\frac{\sum_{i=1}^{5}(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{6.112}{4}} = \sqrt{1.528} = 1.2361$$

Standard error of mean, $\hat{\sigma}_X = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{1.2361}{\sqrt{5}} = \frac{1.2361}{2.2361} = 0.5528$

Therefore, the values of the unbiased estimates of mean, standard deviation and standard error of mean are 6.54, 1.2361 and 0.5528 respectively.

2. **(case of σ known)** 8 samples measuring the length of cloth are sampled from a population where the length is normally distributed with population standard deviation 0.2. Calculate 95% confidence interval for the population mean based on the 8 samples, 4.9, 4.7, 5.1, 5.4, 4.7, 5.2, 4.8 and 5.1.

Solution: Given that 8 samples are drawn from normally distributed population and relates to the length of cloth.

Population standard deviation = 0.2

In order to obtain the 95% confidence interval for the mean, first we need to compute sample mean. Since the population standard deviation is known, the expression for obtaining confidence interval is given by $\bar{X} \pm Z\alpha/_2 \frac{\sigma}{\sqrt{n}}$

Here, $\bar{X}$ = sample mean which is to computed

$Z\alpha/_2$ = the table value ($Z_{0.05}/_2 = 1.96$)

σ = population standard deviation = 0.2

n = sample size = 8

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}\ ; n = 8\ and\ i = 1,2,\dots,8$$

$$\bar{X} = \frac{4.9 + 4.7 + 5.1 + 5.4 + 4.7 + 5.2 + 4.8 + 5.1}{8}$$

$$\bar{X} = \frac{39.9}{8} = 4.9875$$

The confidence interval for the mean is 4.9875 $\pm(1.96)\ \frac{0.2}{\sqrt{8}}$

$$\Rightarrow 4.9875 \pm (1.96)\frac{0.2}{2.8284}$$

$$\Rightarrow 4.9875 \pm (1.96)(0.0707)$$

$$\Rightarrow 4.9875 \pm 0.1386$$

Therefore, 95% confidence interval for the mean is (4.8489, 5.1231)

3. **(case of unknown σ)** The masses in gram of 13 ball bearings taken at random from a batch are 21.4, 23.1, 25.9, 24.7, 23.4, 24.5, 25.0, 22.5, 26.9, 26.4, 25.8, 23.2 and 21.9. Calculate 95% confidence interval for the mean mass of the population assuming that the samples were drawn from the normal population.

Solution: Given that 13 samples were drawn at random from a normally distributed population. Now, we need to calculate the 95% confidence interval for the mean. In this problem, information about population standard deviation is not known, so, we need to compute the sample standard deviation from the data.

The expressions for obtaining sample mean and standard deviation are

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ } and \text{ } s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

and the 95% confidence interval for mean when σ is unknown is $\left(\bar{X} - t\alpha_{/2}\frac{s}{\sqrt{n}}, \bar{X} + t\alpha_{/2}\frac{s}{\sqrt{n}}\right)$

Sample number	X_i	$(X_i - \bar{X})^2$
1	21.4	7.8832
2	23.1	1.2269
3	25.9	2.8639
4	24.7	0.2424
5	23.4	0.6524
6	24.5	0.0854
7	25.0	0.6277
8	22.5	2.9162
9	26.9	7.2484
10	26.4	4.8062
11	25.8	2.5354
12	23.2	1.0154
13	21.9	5.3255
Total	**314.7**	**37.429**

$$\bar{X} = \frac{314.7}{13} = 24.2077$$

$$s = \sqrt{\frac{\sum_{i=1}^{13}(X_i - \bar{X})^2}{13-1}} = \sqrt{\frac{37.429}{12}} = \sqrt{3.1191} = 1.7661$$

$t_{0.025}$ at 12 degrees of freedom = 2.1788

therefore, 95% confidence interval for mean is

$$\Rightarrow 24.2077 \pm (2.1788)\frac{1.7661}{\sqrt{13}}$$

$$\Rightarrow 4.9875 \pm (1.96)(0.4898)$$

$$\Rightarrow 4.9875 \pm 1.0672$$

The lower and upper 95% confidence interval for the mean of the given samples is (23.1405, 25.2749)

4. A random samples of 20 children in a large school were asked a question and 12 answered correctly. Estimate the proportion of children in the school who answered correctly and the standard error of this estimate.

Solution: Given that a random sample of 20 children in a large school were selected and asked a question, i.e., n =20 of which 12 children answered correctly i.e., X = 12 (say).

Now, the sample proportion, p = $\frac{X}{n} = \frac{12}{20} = 0.6$

The standard error of this sample proportion is $\hat{\sigma}_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$\Rightarrow \hat{\sigma}_p = \sqrt{\frac{0.6 * (1-0.6)}{20}}$$

$$\Rightarrow \hat{\sigma}_p = \sqrt{\frac{0.6 * 0.4}{20}}$$

$$\Rightarrow \hat{\sigma}_p = \sqrt{\frac{0.24}{20}} = 0.1095$$

Therefore, the sample proportion and the standard error values are 0.6 and 0.1095

5. 80 samples were drawn at random from a normally distributed population with a sample proportion of 0.26. Using this information, construct the 95% confidence interval for the sample proportion.

Solution: Basing on the given information, it is understood that n is large i.e., n =80 and p=0.26.

Therefore, 95% confidence interval for the sample proportion can be obtained as

$p \pm Z\alpha/_2 * \sqrt{\frac{p(1-p)}{n}}$

$$\Rightarrow 0.26 \pm (1.96) * \sqrt{\frac{0.26 * (1-0.26)}{80}}$$

$$\Rightarrow 0.26 \pm (1.96) * \sqrt{\frac{0.26 * (0.74)}{80}}$$

$$\Rightarrow 0.26\ \pm (1.96) * (0.0490)$$

$$\Rightarrow 0.26\ \pm (0.09611)$$

Thus, the lower and upper 95% confidence intervals are (0.16389, 0.35611).

Z- test for two Proportions

Let p_1 and p_2 are the two proportions from two independent normally distributed populations with sample sizes n_1 and n_2. Here the statement to be test is that the sample proportions of two populations are one and the same. Thus, the null and alternative hypothesis can be defined as

H_0: the sample proportions, p_1 and p_2 are equal, i.e., $p_1 = p_2$

H_1: the sample proportions, p_1 and p_2 are not equal, i.e., $p_1 \neq p_2$

The test statistic for testing the proportion based hypothesis is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

where $\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$ and $\hat{Q} = 1 - \hat{P}$

If $Z_{cal} \leq Z_{cri}$ at α % level of significance, there is no evidence to reject the null hypothesis, otherwise.

Example:

A Business Week/Harris survey asked senior executives at large corporations their opinions about the economic outlook for the future. One question was, “Do you think that there will be an increase in the number of full-time employees at your company over the next 12 months?” In the current survey, 220 of 400 executives answered yes, while in a previous year survey, 192 of 400 executives had answered yes.

Solution: From the given information, the proportions for two populations can be calculated as follows. Sample proportion for current survey is 220/400 = 0.55 and the sample proportion for the previous survey is 0.48.

and $\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{400*0.55 + 400*0.48}{400 + 400} = 0.515$ and $\hat{Q} = 1 - \hat{P} = 1 - 0.515 = 0.485$ with $n_1 = 400$ and $n_2 = 400$ are

Now on substituting the above computed values, the Z test statistic is given by

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.55 - 0.48}{\sqrt{0.515 * 0.485\left(\frac{1}{400} + \frac{1}{400}\right)}} = 1.99$$

$Z_{0.05}$ = 1.96 and from the results it can be concluded that there is an evidence to reject the null hypothesis. Thus, there will be an increase in the number of full-time employees in the company over the next 12 months.

4.3 STATISTICAL HYPOTHESES TESTING

Hypothesis: Statement about the population (or) parameters of the population. In other words, it is a statement of the perceived value of a variable or relationship between two or more variables that can be measured.

In dealing with a hypothesis test, we have to formulate our initial research into two statements such as Null and Alternative hypothesis.

Null Hypothesis (H_0): It is the hypothesis of no difference and is formulated in anticipation of being rejected as false.

Alternative Hypothesis (H_1): It is the positive proposition which states that a difference exists.

Errors: In statistical hypotheses testing, usually we come across four possible states of decision making, of which two states provides correct decision making and the rest two states have wrong decision making process. The following matrix or table will given an idea about these four possible states.

	H_0 is true	H_0 is false
Accept H_0	Correct	Type II error
Reject H_0	Type I error	correct

These two type I and II errors are to be considered to minimize the error in decision making process.

Type I error:

It is the probability that rejecting H_0 when it is true and is denoted by α and is given as

$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$

The maximum probability of occurring type I error is the level of significance.

Type II error:

It is the probability that acepting H_0 when H_1 is true and is denoted by β and is given as

$\beta = P(\text{accept } H_0 \mid H_1 \text{ is true})$

in testing a particular hypothesis, sample size plays a prominent role. Basing on the number of samples involved in the study, the selection about an appropriate statistical test will be done. In general, the branch of statistical tests is divided into small and large samples. It is the rule of thumb that if number of samples are less than 30, then we need to make use of exact sampling tests like, t, F and chi-square. If the number of samples are greater than 30, then the large sample tests namely Z test should be used. However, if the number of samples are at larger number then the t and F or exact sampling distributions approximates to normal distribution.

Small sample tests

Statistical tests based on t- distribution

The major applications branched from this t-distribution are

- One sample t-test
- Two samples or Independent samples t-test and
- Paired samples t-test

One Sample t-test

The following is the procedure of one sample t-test

1. Let x_1, x_2, …, x_n be the random samples drawn from a normal population with a specified or known population mean, μ_0 (say)
2. Define the null and alternative hypothesis as:

H_0: there is no difference between the estimated sample mean and population mean.

H_1: there is a difference between the estimated sample mean and population mean

3. The test statistic to test the above defined hypothesis is

$$t= \left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right| \sim t_{n-1} \text{degrees of freedom}$$

4. Now, compare t with t_{n-1} degrees of freedom at α% level of significance.
5. If $t \leq t_{n-1}$ then there is no evidence to reject the H_0, otherwise.

Example: The mean outer diameter of a bearing of 25 randomly selected units is observed to 10.6 cm with standard deviation of 1.15cm. Is it significantly different from mean of 11 cm.

Solution: In the given problem, information about the population mean and sample statistics such as sample mean and sample standard deviation are given.

i.e., population mean, $\mu = 11$ cm

sample mean, $\bar{x} = 10.6$ cm

sample standard deviation, $s = 1.15$ cm

Hypothesis: H_0: $\bar{x} = \mu$ against H_1: $\bar{x} \neq \mu$

The test statistic for testing the above defined hypothesis is

$$t= \left|\frac{\bar{x}\text{-}\mu}{s/\sqrt{n}}\right| \sim t_{n-1}\text{degrees of freedom}$$

$$t= \left|\frac{10.6\text{-}11}{1.15/\sqrt{25}}\right| \Rightarrow t = \frac{0.4}{0.23} \Rightarrow t = 1.74$$

$t_{n-1} = t_{25-1} = t_{24}$ at 0.05 level of significance is 2.06.

therefore, $t < t_{n-1}$ i.e., 1.74 < 2.06, hence there is no evidence to reject the null hypothesis and it can be concluded the sample mean is not significantly different from the given population mean.

Example: A random sample of 10 boys had the following I.Q.'s: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.

Solution: Given that, a random sample of 10 boys are selected and their I.Q.'s are measured. Now the hypothesis to be tested is that the mean I.Q. values of samples support the assumption of population mean I.Q. of 100 or not.

First define, null and alternative hypothesis as

H_0: mean I.Q. values of samples do not support the assumption of population mean I.Q. of 100

H_1: mean I.Q. values of samples support the assumption of population mean I.Q. of 100

Now, to claim any one of the hypothesis, we need to compute sample mean and sample standard deviation.

Sample mean,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} ; n = 10 \; and \; i = 1,2, \dots, 10$$

$$\bar{X} = \frac{972}{10} = 97.2$$

Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = 14.27352$$

Therefore, test statistic value is

$$\text{t}=\left|\frac{97.2\text{-}100}{14.27352/\sqrt{10}}\right| \Rightarrow t = \frac{2.8}{4.514} \Rightarrow t = 0.6203$$

$t_{n-1} = t_{10-1} = t_9$ at 0.05 level of significance is 1.8331.

Therefore, $t < t_{n-1}$ i.e., 0.6203 < 1.8331, hence there is no evidence to reject the null hypothesis. Thus, mean I.Q. values of 10 samples support the assumption of population mean I.Q. of 100.

t-test for two independent samples

This test is applicable only when there are two independent samples and the objective of comparison between these two groups. Further, it is assumed that the samples are drawn at random from a normally distributed populations.

- Let there be two groups, G1 and G2 with $\bar{x}_1$ and $\bar{x}_2$, s_1 and s_2 as their respective sample means and sample standard deviations
- The hypothesis, in general can be defined as follows:

H_0: There is no mean difference between two groups

H_1: There is a mean difference between two groups

- To test the defined hypothesis, the test statistic based on t-distribution is given by

$$\text{t}=\left|\frac{\bar{x}_1\text{-}\bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| \sim t_{n_1+n_2\text{-}2} \text{ degrees of freedom}$$

where, $s = \sqrt{\frac{1}{n_1+n_2\text{-}2}\left(\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2+\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2\right)}$ and

$$\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1} \; and \; \bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2}$$

- If $t \leq t_{n_1+n_2\text{-}2}$ degrees of freedom at α% level of significance, then there is no evidence to reject the null hypothesis, otherwise.

Example: Two types of teaching aids A and B, say, were introduced and as a response, students' scores for each teaching aid is collected independently. Analysis showed the following results $n_1 = 6$, $n_2 = 5$, $\bar{x}_1=7.5$ and $\bar{x}_2=7.2$, $s_1 = 0.024$ and $s_2 = 0.032$. It is to be tested that which teaching aid is better for students?

Solution: It is stated that two teaching aids A and B are adopted independently on two groups of students and the responses are collected as scores.

On conducting such experiment, it is given that the sample means and sample standard deviations obtained from two teaching aids are $n_1 = 6$, $n_2 = 5$, $\bar{x}_1 = 7.5$ and $\bar{x}_2 = 7.2$, $s_1 = 0.024$ and $s_2 = 0.032$

Here, the hypothesis to be defined is

H_0: There is no mean difference between two teaching aids

H_1: There is a mean difference between two teaching aids

Then, the test statistic value can be obtained by substituting the given values in the following formula

$$t = \left| \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \sim t_{n_1+n_2-2} \text{ degrees of freedom}$$

$$t = \left| \frac{7.5 - 7.2}{0.01685\sqrt{\frac{1}{6} + \frac{1}{5}}} \right| \Rightarrow t = \left| \frac{0.3}{0.0102} \right| \Rightarrow t = 29.4025$$

t_9 at 0.05 level of significance is 1.8331

Therefore, $t > t_{n_1+n_2-2}$ i.e., 29.4025 > 1.8331, hence there is an evidence to reject the null hypothesis. This means that there is difference between two teaching aids and the teaching aid A is comparatively better than teaching aid B.

Note: If data pertaining to random samples of two populations are given, then the sample means and sample standard deviations are to be computed and then substituted in the test statistic for obtaining outcome.

t-test for paired samples

This test is applied to paired data of observations from one sample only when each individual gives a pair of observations. For example, the comparison of prices of certain commodities estimated before and after raising of shares. The point to be noted is that the number of random samples which are drawn should be same during before and after experimentation. For this test, the hypothesis is defined as follows:

H_0: There is no true discrepancy between before and after values

H_1: There exists true discrepancy between before and after values

The test statistic to test the defined hypothesis is given as

$$\mathrm{t}= \left|\frac{\bar{d}}{{}^{\mathrm{s}}/_{\sqrt{\mathrm{n}}}}\right| \sim \mathrm{t}_{\mathrm{n-1}}\text{degrees of freedom}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$; $d_i = x_i - y_i$; here x_i and y_i are used to denote before and after values and the sample standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}}$$

Now, compare t with t_{n-1} degrees of freedom at α% level of significance. If $t \leq t_{n-1}$ then there is no evidence to reject the H_0, otherwise.

Example: The marks or opinions in form of scores are collected on 12 trainees before commencement of a training program and at the end of the program. Is there any change in the scores indicating the effect of the training program?

Trainee No.	1	2	3	4	5	6	7	8	9	10	11	12
Scores before training	6	8	8	6	5	9	6	7	6	6	4	8
Scores after training	8	8	10	7	6	10	9	8	5	7	4	6

Solution: It is stated that the effectiveness of the training program is observed by collecting data in terms of scores before and after the program from 12 trainees. Now, basing on the given information, the hypothesis can be defined as:

H_0: There is no effectiveness of the training program

H_1: Training program is effective

To test the defined hypothesis the test statistic is given by

$$\mathrm{t}= \left|\frac{\bar{d}}{{}^{\mathrm{s}}/_{\sqrt{\mathrm{n}}}}\right| \sim \mathrm{t}_{\mathrm{n-1}}\text{degrees of freedom}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$; $d_i = x_i - y_i$; here x_i and y_i are used to denote before and after values and the sample standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}} = \sum_{i=1}^{n} d_i^2 - \frac{(\sum_{i=1}^{n} d_i)^2}{n}$$

Here, $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{9}{12} = 0.75$

Trainee No.	Scores before training	Scores after training	d_i	d_i^2
1	6	8	-2	4
2	8	8	0	0
3	8	10	-2	4
4	6	7	-1	1
5	5	6	-1	1
6	9	10	-1	1
7	6	9	-3	9
8	7	8	-1	1
9	6	5	1	1
10	6	7	-1	1
11	4	4	0	0
12	8	6	2	4
				27

$$\Rightarrow s = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}} = \frac{20.25}{11} = 1.84$$

$$\text{t}= \left|\frac{\bar{d}}{\text{s}/\sqrt{\text{n}}}\right| \Rightarrow \text{t}= \left|\frac{0.75}{1.84/\sqrt{12}}\right| \Rightarrow t = 1.92$$

$t_{n-1} = t_{12-1} = t_{11}$ at 0.05 level of significance is 2.20.

Therefore, t < t_{n-1} i.e., 1.92 < 2.20, hence there is no evidence to reject the null hypothesis. This helps to claim that there is no effectiveness of training program.

F - test for equality of two sample variances

- Let us suppose two random samples are drawn from two independent normal populations, x_i and y_j, say
- Here the main objective is to test whether the sample variances obtained from these two independent samples are equal or not.

- To meet this objective, the null and alternative hypothesis can be defined as

H$_0$: The population variances are equal, i.e., $\sigma_x^2 = \sigma_y^2$

H$_1$: The population variances are not equal, $\sigma_x^2 \neq \sigma_y^2$

- The test statistic which is used to test the hypothesis is as follows

$$F = \frac{s_x^2}{s_y^2} \sim F_{(n_1-1,n_2-1)} \text{ degrees of freedom}$$

where $s_x^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$; $s_y^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(y_j - \bar{y})^2$ and

$$\bar{x} = \frac{\sum_{i=1}^{n_1} x_i}{n_1};\ \bar{y} = \frac{\sum_{j=1}^{n_2} y_j}{n_2}$$

Here, $s_y^2\ and\ s_y^2$ are the sample variances and unbiased estimates of $\sigma_x^2\ and\ \sigma_y^2$ respectively

- If F$_{cal}$ ≤ $F_{(n_1-1,n_2-1)}$degrees of freedom at α% level of significance, then there is no evidence to reject the null hypothesis, otherwise.

Example: The sample variances in 25 males and 25 females are 5.0 and 9.0. can we conclude that the variance in prices is same in both males and females?

Solution: Given n$_1$ = 25, n$_2$ = 25, s_1^2= 5.0 and s_2^2 = 9.0

Therefore, $F = \frac{s_x^2}{s_y^2} = \frac{9.0}{5.0} = 1.8$

F$_{(24,\ 24)}$ at 5% level of significance is 1.98

Here F$_{cal}$ < F$_{(24,\ 24)}$, 1.8 < 1.98, hence, we accept and conclude that there is no variability in both the groups with respect to prices.

Example: The College Board provided comparisons of Scholastic Aptitude Test (SAT) scores based on the highest level of education attained by the test taker's parents. A research hypothesis was that students whose parents had attained a higher level of education would on average score higher on the SAT. During 2003, the overall mean SAT verbal score was 507 (The World Almanac, 2004). SAT verbal scores for independent samples of students follow. The first sample shows the SAT verbal test scores for students whose parents are college graduates with a bachelor's degree. The second sample shows the SAT verbal test scores for students whose parents are high school graduates but do not have a college degree.

College Grads	485	534	650	554	550	572	497	592	487	533	526	410	515	578	448	469
High School Grads	442	580	479	486	528	524	492	478	425	485	390	535				

Formulate the hypotheses that can be used to determine whether the sample data support the hypothesis that students show a higher variability in verbal score on the SAT if their parents have not attained a higher level of education. At α = 0.05, what is your conclusion?

Solution:

Basing on the given information, we need to test whether the sample data support the hypothesis that students show a higher variability in verbal score on the SAT if their parents have not attained a higher level of education

Let us denote the SAT verbal test scores for students whose parents are college graduates with a bachelor's degree as Group A and the other as group B

The null and alternative hypothesis for this problem is

H_0: The sample variability between groups A and B are equal

H_1: The sample variability between groups A and B are not equal

Sample number	**College Grads**		**High School Grads**	
	x_i	$\sum_{i=1}^{n_1}(x_i-\bar{x})^2$	**y_j**	$\sum_{j=1}^{n_2}(y_j-\bar{y})^2$
1	485	1600	442	2025
2	534	81	580	8649
3	650	15625	479	64
4	554	841	486	1
5	550	625	528	1681
6	572	2209	524	1369
7	497	784	492	25
8	592	4489	478	81
9	487	1444	425	3844
10	533	64	485	4
11	526	1	390	9409
12	410	13225	535	2304
13	515	100		
14	578	2809		
15	448	5929		
16	469	3136		
Totals	**525**	**3310.125**	**487**	**2454.667**

$$s_A^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2 = \frac{3310.125}{15} = 220.675$$

$$s_B^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_j-\bar{y})^2 = \frac{2454.667}{11} = 223.1515$$

Therefore, $F = \frac{s_A^2}{s_B^2} = \frac{220.675}{223.1515} = 0.9889$

$F_{(15,11)}$ at 0.05 level of significance is 2.72. hence, from the computed results, it is evident that there is no variability between the groups A and B. In other words, the variability in the SAT verbal test scores for students whose parents are college graduates with a bachelor's degree and with SAT verbal test scores for students whose parents attained higher degree is equal.

EXERCISES

Problems on tests for single Proportion

1. A builder claims that heat pumps are installed in 70% of all homes being constructed today in the city of Richmond, Virginia. Would you agree with this claim if a random survey of new homes in this city shows that 8 out of 15 had heat pumps installed? Use a 0.10 level of significance.
2. A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use a 0.05 level of significance.
3. Suppose that, in the past, 40% of all adults favored capital punishment. Do we have reason to believe that the proportion of adults favoring capital punishment today has increased if, in a random sample of 15 adults, 8 favor capital punishment? Use a 0.05 level of significance.
4. It is believed that at least 60% of the residents in a certain area favor an annexation suit by a neighboring city. What conclusion would you draw if only 110 in a sample of 200 voters favor the suit? Use a 0.05 level of significance.
5. A new radar device is being considered for a certain defense missile system. The system is checked by experimenting with actual aircraft in which a kill or a no kill is simulated. If in 300 trials, 250 kills occur, accept or reject, at the 0.04 level of significance, the claim that the probability of a kill with the new system does not exceed the 0.8 probability of the existing device.
6. In a controlled laboratory experiment, scientists at the University of Minnesota discovered that 25% of a certain strain of rats subjected to a 20% coffee bean diet and then force-fed powerful cancer-causing chemical later developed cancerous tumors. Would we have reason to believe that the proportion of rats developing tumors when

subjected to this diet has increased if the experiment were repeated and 16 of 48 rats developed tumors? Use a 0.05 level of significance.

Problems on tests for Proportions between two populations

7. Consider the following results for independent samples taken from two populations.

Sample 1	**Sample 2**
$n_1 = 400$	$n_2 = 300$
$p_1 = 0.48$	$p_2 = 0.36$

a. What is the point estimate of the difference between the two population proportions?

b. Develop a 90% confidence interval for the difference between the two population proportions.

c. Develop a 95% confidence interval for the difference between the two population proportions.

8. Consider the following hypothesis test.

$$H_0: p_1 - p_2 \leq 0$$

$$H_0: p_1 - p_2 > 0$$

The following results are for independent samples taken from the two populations

Sample 1	**Sample 2**
$n_1 = 200$	$n_2 = 300$
$p_1 = 0.22$	$p_2 = 0.16$

With α = .05, what is your hypothesis testing conclusion?

9. The Professional Golf Association (PGA) measured the putting accuracy of professional golfers playing on the PGA Tour and the best amateur golfers playing in the World Amateur Championship (Golf Magazine, January 2007). A sample of 1075 6-foot putts by professional golfers found 688 made puts. A sample of 1200 6-foot putts by amateur golfers found 696 made putts.

1. Estimate the proportion of made 6-foot putts by professional golfers. Estimate the proportion of made 6-foot putts by amateur golfers. Which group had a better putting accuracy?
2. What is the point estimate of the difference between the proportions of the two populations? What does this estimate tell you about the percentage of putts made by the two groups of golfers?
3. What is the 95% confidence interval for the difference between the two population proportions? Interpret his confidence interval in terms of the percentage of putts made by the two groups of golfers.

10. Chicago O’Hare and Atlanta Hartsfield-Jackson are the two busiest airports in the United States. The congestion often leads to delayed flight arrivals as well as delayed flight departures. The Bureau of Transportation tracks the on-time and delayed performance at major airports (Travel & Leisure, November 2006). A flight is considered delayed if it is more than 15 minutes behind schedule. The following sample data show the delayed departures at Chicago O’Hare and Atlanta Hartsfield-Jackson airports.

	Chicago O’Hare	**Atlanta Hartsfield-Jackson**
Flights	900	1200
Delayed Flights	252	312

a. State the hypotheses that can be used to determine whether the population proportions of delayed departures differ at these two airports.

b. What is the point estimate of the proportion of flights that have delayed departures at Chicago O’Hare?

c. What is the point estimate of the proportion of flights that have delayed departures at Atlanta Hartsfield-Jackson?

d. What is your conclusion?

11. In a test of the quality of two television commercials, each commercial was shown in a separate test area six times over a one-week period. The following week a telephone survey was conducted to identify individuals who had seen the commercials. Those individuals were asked to state the primary message in the commercials. The following results were recorded.

	Commercial A	**Commercial B**
Number Who saw Commerical	150	200
Number Who Recalled Message	63	60

a. Use α = .05 and test the hypothesis that there is no difference in the recall proportions for the two commercials.

b. Compute a 95% confidence interval for the difference between the recall proportions for the two populations.

12. During the 2003 Super Bowl, Miller Lite Beer’s commercial referred to as “The Miller Lite Girls” ranked among the top three most effective advertisements aired during the Super Bowl (USA Today,December 29, 2003). The survey of advertising effectiveness, conducted by USA Today’s Ad Track poll, reported separate samples by respondent age group to learn about how the Super Bowl advertisement appealed to different age groups. The following sample data apply to the “The Miller Lite Girls” commercial.

Age Group	Sample size	Liked the Ad a lot
Under 30	100	49
30 to 49	150	54

a. Formulate a hypothesis test that can be used to determine whether the population proportions for the two age groups differ.

b. What is the point estimate of the difference between the two population proportions?

c. Conduct the hypothesis test. At α = .05, what is your conclusion?

d. Discuss the appeal of the advertisements to the younger and the older age groups. Would the Miller Lite organization find the results of the USA TodayAd Track poll encouraging? Explain.

13. A 2003 New York Times/CBSNews poll sampled 523 adults who were planning a vacation during the next six months and found that 141 were expecting to travel by airplane (New York Times News Service, March 2, 2003). A similar survey question in a May 1993 New York Times/CBSNews poll found that of 477 adults who were planning a vacation in the next six months, 81 were expecting to travel by airplane.

a. State the hypotheses that can be used to determine whether a significant change occurred in the population proportion planning to travel by airplane over the 10-year period.

b. What is the sample proportion expecting to travel by airplane in 2003? In 1993?

c. Use α = .01 and test for a significant difference. What is your conclusion?

d. Discuss reasons that might provide an explanation for this conclusion.

Problems on tests for means when variances are known

14. The following results come from two independent random samples taken of two populations. If $n_1 = 50, \bar{x}_1 = 13.6, \sigma_1 = 2.2$ and $n_2 = 35, \bar{x}_2 = 11.6, \sigma_2 = 3.0$.

a. What is the point estimate of the difference between the two population means?

b. Test whether the means are equal or not with α = 0.05.

a. When conducting a hypothesis test with the values given for the standard deviation, sample size, and α, how large must the increase from 2007 to 2008 be for it to be statistically significant?

b. Use the result of part (d) to state whether the increase for J.C. Penney from 2007 to 2008 is statistically significant.

15. A manufacturer claims that the average tensile strength of thread A exceeds the average tensile strength of thread B by at least 12 kilograms. To test his claim, 50 pieces of each

type of thread are tested under similar conditions. Type A thread had an average tensile strength of 86.7 kilograms with known standard deviation of $\sigma_A = 6.28$ kilograms, while type B thread had an average tensile strength of 77.8 kilograms with known standard deviation of $\sigma_B = 5.61$ kilograms. Test the manufacturer's claim at α = 0.05.

Problems on tests for means when variances are unknown

16. The following results are for independent random samples taken from two populations. If $n_1 = 20, \bar{x}_1 = 22.5, s_1 = 2.5$ and $n_2 = 30, \bar{x}_2 = 20.1, s_2 = 4.8$.
    a. What is the point estimate of the difference between the two population means?
    b. What is the degrees of freedom for the t distribution?
    c. At 95% confidence, what is the margin of error?
    d. What is the 95% confidence interval for the difference between the two population means?
17. Consider the following hypothesis test.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_0: \mu_1 - \mu_2 \neq 0$$

The following results are from independent samples taken from two populations.

$n_1 = 35, \bar{x}_1 = 13.6, s_1 = 5.2$ and $n_2 = 40, \bar{x}_2 = 10.1, s_2 = 8.5$.

a. What is the value of test statistic?
b. With α = 0.05, what is your hypothesis testing conclusion?

18. Consider the following data for two independent random samples taken from two normal populations.

Sample 1	10	7	13	7	9	8
Sample 2	8	7	8	4	6	9

a. Compute the two sample means.
b. Compute the two sample standard deviations.
c. What is the point estimate of the difference between the two population means?
d. What is the 90% confidence interval estimate of the difference between the two population means?

19. FedEx and United Parcel Service (UPS) are the world's two leading cargo carriers by volume and revenue (The Wall Street Journal, January 27, 2004). According to the Airports Council International, the Memphis International Airport (FedEx) and the Louisville International Airport (UPS) are 2 of the 10 largest cargo airports in the world. The following random samples show the tons of cargo per day handled by these airports. Data are in thousands of tons.

Memphis

9.1	15.1	8.8	10	7.5	10.5
8.3	9.1	6	5.8	12.1	9.3

Louisville

4.7	5	4.2	3.3	5.5
2.2	4.1	2.6	3.4	7

a. Compute the sample mean and sample standard deviation for each airport.

b. What is the point estimate of the difference between the two population means? Interpret this value in terms of the higher-volume airport and a comparison of the volume difference between the two airports.

c. Develop a 95% confidence interval of the difference between the daily population means for the two airports.

20. Are nursing salaries in Tampa, Florida, lower than those in Dallas, Texas? Salary data show staff nurses in Tampa earn less than staff nurses in Dallas (The Tampa Tribune, January 15, 2007). Suppose that in a follow-up study of 40 staff nurses in Tampa and 50 staff nurses in Dallas you obtain the following results.

Tampa	**Dallas**
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = \$56{,}100$	$\bar{x}_2 = \$59{,}400$
$s_1 = \$6000$	$s_2 = \$7000$

a. Formulate hypothesis so that, if the null hypothesis is rejected, we can conclude that salaries for staff nurses in Tampa are significantly lower than for those in Dallas. Use α = 0.05.

b. What is the value of the test statistic?

c. What is the p-value?

d. What is your conclusion?

21. Injuries to Major League Baseball players have been increasing in recent years. For the period 1992 to 2001, league expansion caused Major League Baseball rosters to increase 15%. However, the number of players being put on the disabled list due to injury increased 32% over the same period (USA Today, July 8, 2002). A research question addressed whether Major League Baseball players being put on the disabled list are on the list longer in 2001 than players put on the disabled list a decade earlier.

a. Using the population mean number of days a player is on the disabled list, formulate null and alternative hypotheses that can be used to test the research question.

b. Assume that the following data apply:

2001 Season	**1992 Season**
$n_1 = 45$	$n_2 = 38$
$\bar{x}_1 = 60\ days$	$\bar{x}_2 = 51\ days$
$s_1 = 18\ days$	$s_2 = 15\ days$

What is the point estimate of the difference between population mean number of days on the disabled list for 2001 compared to 1992? What is the percentage increase in the number of days on the disabled list?

c. Use α = 0.01. What is your conclusion about the number of days on the disabled list? What is the p-value?

d. Do these data suggest that Major League Baseball should be concerned about the situation?

22. Periodically, Merrill Lynch customers are asked to evaluate Merrill Lynch financial consultants and services. Higher ratings on the client satisfaction survey indicate better service, with 7 the maximum service rating. Independent samples of service ratings for two financial consultants are summarized here. Consultant A has 10 years of experience, whereas consultant B has 1 year of experience. Use α = 0.05 and test to see whether the consultant with more experience has the higher population mean service rating.

Consultant A	**Consultant B**
$n_1 = 16$	$n_2 = 10$
$\bar{x}_1 = 6.82$	$\bar{x}_2 = 6.25$
$s_1 = 0.64$	$s_2 = 0.75$

a. State the null and alternative hypotheses.

b. Compute the value of the test statistic.

c. What is your conclusion?

Problems on F-test for variances

23. A sample of 16 items from population 1 has a sample variance $s_1^2 = 5.8$ and a sample of 21 items from population 2 has a sample variance $s_2^2 = 2.4$ Test the following hypotheses at the .05 level of significance.

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2$$

24. Consider the following hypothesis test.

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_a: \sigma_1^2 \neq \sigma_2^2$$

What is your conclusion if $n_1 = 21$, $s_1^2 = 8.2$, $n_1 = 26$ and $s_2^2 = 4.0$? Use α = 0.05

25. The variance in a production process is an important measure of the quality of the process. A large variance often signals an opportunity for improvement in the process by finding ways to reduce the process variance. Conduct a statistical test to determine whether there is a significant difference between the variances in the bag weights for two machines. Use a .05 level of significance. What is your conclusion? Which machine, if either, provides the greater opportunity for quality improvements?

Machine 1	2.95	3.45	3.5	3.75	3.48	3.26	3.33	3.2
	3.16	3.2	3.22	3.38	3.9	3.36	3.25	3.28
	3.2	3.22	2.98	3.45	3.7	3.34	3.18	3.35
	3.12							
Machine 2	3.22	3.3	3.34	3.28	3.29	3.25	3.3	3.27
	3.38	3.34	3.35	3.19	3.35	3.05	3.36	3.28
	3.3	3.28	3.3	3.2	3.16	3.33		

26. On the basis of data provided by a Romac salary survey, the variance in annual salaries for seniors in public accounting firms is approximately 2.1 and the variance in annual salaries for managers in public accounting firms is approximately 11.1. The salary data were provided in thousands of dollars. Assuming that the salary data were based on samples of 25 seniors and 26 managers, test the hypothesis that the population variances in the salaries are equal. At a .05 level of significance, what is your conclusion?

Problems on t-test for single mean

27. A machinist is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.

28. The mean weekly sales of soap bars in departmental stores were 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?

29. The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% level of significance assuming that for 9 degrees of freedom P (t > 1.83) = 0.05.

30. A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43.5 inches for the population is not reasonable.
31. A manufacturer of gun powder has developed a new powder which is designed to produce a muzzle velocity equal to 3000 ft/sec. Seven shells are loaded with the charge and the muzzle velocities measured. The resulting velocities are as follows: 3005; 2935; 2965; 2995; 3905; 2935 and 2905. Do these data present sufficient evidence to indicate that the average velocity differs from 3000 ft/sec.
32. The average length of time for students to register for summer classes at a certain college has been 50 minutes with a standard deviation of 10 minutes. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with s.d. of 11.9 minutes under the new system, test the hypothesis that the population mean has not changed, using 0.05 level of significance.
33. The average breaking strength of steel rods is specified to be 18.5 thousand pounds. To test this a sample of 14 rods was tested. The mean and standard deviations obtained were 17.85 and 1.955 thousand pounds respectively. is the result of the experiment significant?

UNIT – V

NON PARAMETRIC TESTS

CONTENTS:

5.1 Introduction to Non Parametric Tests

5.2 Sign Test

5.3 Wilcoxon Signed Rank test

5.4 Kruskal-Wallis H test

5.5 Wald-Wolfowitz Run test

5.6 Chi-Square test

5.1 Introduction to Non Parametric Tests

In statistics, most of the methods require an assumption that the data follows a normal distribution. Sometimes, this may not be applicable when data does not follow normal. Further, the normality assumption can be obtained by adapting a suitable transformation method. In some situations, the researchers will be in need of those methods which are free from distributional assumptions that the data follows a particular distribution. Those methods are usually named as Non Parametric methods. Another way of usage of these methods arise when your data is of nominal or ordinal type. So, in general, for a statistical method to be classified as non-parametric, it must satisfy at least one of the following conditions: (Conover, 1998)

1. The method can be used with nominal/ ordinal data.
2. The method can be used with interval/ ratio data when no assumption can be made about population probability distribution.

Here, a point to be noted is that if the data with level of measurement interval/ ratio has the necessary probability distribution then parametric methods result in providing more powerful statistical procedures. It is also to be noted that in many cases, both non-parametric and parametric methods can be applied, in such cases non-parametric method is as good as (or) almost as powerful as parametric method. Finally, it can be understood that whenever there are less restrictions on data measurement and fewer assumptions about population distributions, non-parametric methods are regarded as more generally applicable than parametric methods.

In this chapter, a detailed discussion is made on each of the non-parametric method like Sign, Wilcoxon Signed Rank test, Wald-Wolfowitz runs test, Kruskal-Wallis test and Chi-Square tests with numerical illustrations.

5.2 Sign Test

Sign test is a non-parametric analogue to one sample Z test under small and large sample sizes. The applicability of Sign test will come into role when we need to test the data values against a perceived hypothesis statement with the following inclusions:

- When $n \leq 20$, the small sample case of Sign test is to be used by replacing one sample t-test with underlying distribution as Binomial.
- When $n > 230$, the large sample case of Sign test is to be used by replacing one sample Z test with underlying approximated distribution as normal.

Now, let us see and understand how the small and large sample versions of the Sign test can be used to test the hypothesis and helps in drawing proper conclusions about the defined hypothesis.

Assumptions

1. The samples are drawn at random
2. Continuous distribution and
3. Measurements can be expressed using plus or minus signs

The sign test purely depends on the usage of the measure, median to note the number of counts that fall above and below the median. Generally, for those counts which lie above the median a plus sign is given and a negative sign is assigned for the values which are below the median. Let ‘p’ be the proportion or probability that 50% of all values lie above and 50% of all values lie below the median value. It is expected that the number of counts of distribution is approximated symmetrically around median. Here, the data so collected is represented in terms of plus or minus sign; hence this non-parametric test is called the SIGN TEST. The number of plus signs is the test statistic.

In small sample case under the assumption that null hypotheses is true, the underlying sampling distribution is Binomial distribution with $p = 0.5$. The hypothesis under one tail and two tail test can be defined as : (usually $\alpha = 0.05$ or 0.01)

Table 5.1

One tail test		Two tail test
Upper one tail	Lower one tail	
H_0 : Sample median ≤ Population median	H_0 : Sample median ≥ Population median	H_0 : Sample median = Population median
H_1 : Sample median > Population median	H_1 : Sample median < Population median	H_1 : Sample median ≠ Population median

Procedure: Consider an experiment where the pairs of observations on two things are being compared and different pairs are observed under different conditions. These number of paired observations (n) are the number of trials with probability of success equal to 0.5. In this case, we can handle the situation using a Binomial distribution.

X ~ Bin(n, p) with probability mean function

$$P(X = r) = {}^{n}C_{r}p^{r}q^{n-r} \ ;$$

$$\text{Mean } \mu = np \text{ and variance } \sigma^2 = npq.$$

In general, there will be five steps of conducting any test of hypothesis in both parametric/ non-parametric analysis.

1. Define null hypothesis
2. Define alternative hypothesis
3. Set the level of significance
4. Test statistic and
5. Inference

Let us consider an example to illustrate the procedural format of conducting sign test.

Example: A company has conducted a study to know the preferences of individuals in terms of taste for two brands of a product. The following table contains the list of preferences indicated by 10 individuals

Table 5.2

Individuals	1	2	3	4	5	6	7	8	9	10
Brand A vs Brand B	+	+	+	-	+	+	-	+	-	+

At $\alpha = 0.05$, test for a significant difference in the preferences for the two brands. Here, + indicates preference for Brand A over Brand B.

From the data listed in Table 5.2, seven plus signs are observed indicating that the individuals' preference is towards brand A rather than brand B.

Now, on using the binomial probabilities, it is easy to determine the p-value for the test. The binomial probabilities can be obtained using the table ____ given in appendix with p = 0.5.

For the mentioned problem, hypothesis can be stated as

H_0 : Preference for Brand A = Preference for Brand B

H_1 : Preference for Brand A ≠ Preference for Brand B

Upon using the binomial probabilities, with n = 10; p = 0.5 and number of successes as 7, p-value is observed to be 0.000, indicating that the chance of preferring brand A of a product is very high over brand B.

So, with respect to taste of a product, it is evident to claim that individual's preference differs significantly for the two brands of a product.

In large sample case, under the null hypothesis H_0 : p =0.5 and a sample size of n > 20, the sampling distribution for the number of plus signs can be approximated by a normal distribution.

The mean and standard deviation under normal approximation are $\mu = 0.5n$; $\sigma = \sqrt{0.25n}$.

To have a better understanding about the practical applicability of sign test under large sample case, let us have an example of a national survey which is conducted to see whether the annual income will fulfill the dreams to come true or not. For which, the data was collected from the adults of median annual income. A sample of 225 individuals in a country are randomly selected, of which 122 individuals reported that the amount of income is needed to make their dreams come true is less than \$ 152,000 and the rest 103 individuals reported that the amount needed is more than \$ 152,000.

Now, the problem here is to test whether the median amount of annual income needed to make dreams come true in that country is \$ 152,000 or not.

On using the sign test, we have that n = 225. Now, we compute the mean and standard deviation for the given problem.

$$\mu = 0.5n = 0.5(225) = 112.5$$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(225)} = \sqrt{56.25} = 7.5$$

With 0.05 level of significance and based on the response from the individuals, (i.e., x= 122), we can calculate the Z value as

$$Z = \frac{x-\mu}{\sigma} = \frac{122-112.5}{7.5} = \frac{9.5}{7.5} = 1.2667$$

The standard normal probability table shows that the area in the tail towards left of Z = 1.2667 is 0.8962. Under two tail test, p-value = 2(0.8962) = 1.7924. This depicts that at α = 0.05, we accept H_0 indicating that individual's perception about median amount of annual income $152,000/- to make dreams come true is not true.

Let us consider another example with a different set of information about a study. Taken a situation of competition in the personal computer market intensity, a sample of 500 purchases showed 202 Brand A computers, 158 Brand B computers and 140 other computers. It is necessary to test the hypothesis that brand A and brand B have the same share of the personal computer market. (α = 0.05)

Using sign test, we have n = 500-140 = 360 individuals were able to purchase Brand A and Brand B personal computers. From this, we can compute mean and standard deviation as

$$\mu = 0.5n = 0.5(360) = 180$$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(360)} = 9.486$$

Here n = 360, it can be assumed that the sampling distribution is normally distributed (approximated).

Using 0.05% level of significance, the test statistic can be computed as

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

Here x = 202; μ = 180; σ = 9.486

$$\Rightarrow Z = \frac{202 - 180}{9.486} = \frac{22}{9.486} = 2.32$$

From the standard normal probability table, the area in the tail to the left of Z = 2.32 is 0.9898. under two tail test, the p-value is 2(0.9898) = 1.96

With the p-value, we do not reject the hypothesis

i.e., H_0 : Personal computer market for Brand A = Personal computer market for Brand B

H_1 : Personal computer market for Brand A ≠ Personal computer market for Brand B

Therefore, it can be concluded that both Brand A and Brand B have the same share of the personal computer market.

5.3 Wilcoxon Signed Rank test

When an experiment is conducted to generate a matched pairs of observations with a normality assumption, then, in such situations, the standard test which is derived from exact sampling distribution, namely the t-test for paired sample is generally used. However, if there is a non-normality and particularly for samples the t-test may not be valid. Hence, a non-parametric analogue to paired samples t-test is the Wilcoxon- Signed rank sum test, which is used to test the null hypothesis that the median of a distribution is equal to some value.

The method considers ‘n’ matched pairs as one sample. The following are the assumptions of the test:

- Each paired sample is randomly distributed.
- The differences obtained through matched pairs should be symmetrically distributed.

In the case of large samples of paired data ($n \geq 10$), it is suggested to have a normal approximation. If $n \leq 10$, we shall test the hypothesis using the critical values of the statistic.

A test was conducted of two overnight delivery services. Two samples of identical deliveries were setup so that both delivery services were notified of the need for a delivery at the same time. The hours required to make each delivery is reported in table 5.3. The problem which is to be addressed is that is there any difference in the delivery times for the two services?

Table 5.3											
Delivery	1	2	3	4	5	6	7	8	9	10	11
Service 1	24.5	26	28	21	18	36	25	21	24	26	31
Service 2	28	25.5	32	20	19.5	28	29	22	23.5	29.5	30

In order to verify the given statement of the problem that is there any similarity in the delivery times of two services, first we need to define the hypothesis in the following way.

H_0 : The delivery times of both services are identical

against H_1 : The delivery times of both services are not identical

If the null hypothesis cannot be rejected, indicates that the delivery time for the two services does not differ. However, if H_0 is rejected, we will have an evidence to claim that the delivery time for the two services differ in time.

The following way outlines the procedures of conducting the test.

1. Compute the difference for each matched sample.
2. Consider the absolute difference of the differences between the matched pairs.
3. Rank these absolute differences from lowest to highest. Suppose, if there is any zero difference is present, then ranking is done by discarding that zero difference pair.
4. Once, the ranking is done for the absolute differences, assigning of signs to those ranks on par with the original difference in the data.
5. Calculate the sum of the signed ranks, denoted as T.
6. The sampling distribution of T for identical populations can be approximated to normal distribution when $n \geq 10$.

i.e., Mean = $\mu_T = 0$

Standard Deviation = $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}}$

7. The test statistic, to test the defined hypothesis is

$$Z = \frac{T - \mu_T}{\sigma_T}$$

8. The Z value can be used to compute the p-value. This can be done using the standard normal probability table.
9. If p-value ≤ 0.05 (= α), then it is evident to reject the null hypothesis, otherwise.

Here, for the considered problem, the following table 5.4, gives us the absolute values, ranks and signed ranks.

Table: 5.4

Delivery	**Service 1**	**Service 2**	**Difference**	**Absolute Difference**	**Rank**	**Signed Rank**
1	24.5	28.0	-3.5	3.5	7.5	-7.5
2	26.0	25.5	0.5	0.5	1.5	+1.5
3	28.0	32.0	-4.0	4.0	9.5	-9.5
4	21.0	20.0	1.0	1.0	4.0	+4.0
5	18.0	19.5	-1.5	1.5	6.0	-6.0
6	36.0	28.0	8.0	8.0	11.0	+11.0
7	25.0	29.0	-4.0	4.0	9.5	-9.5
8	21.0	22.0	-1.0	1.0	4.0	-4.0
9	24.0	23.5	0.5	0.5	1.5	+1.5
10	26.0	29.5	-3.5	3.5	7.5	-7.5
11	31.0	30.0	1.0	1.0	4.0	+4.0
						T = -22

The tied observations are assigned the average ranks. The last column of the above table 5.4 takes the sign as that of the original difference. It is noted that the sum of signed ranks is -22, which is significantly from zero (since the sum of signed rank values would be approximately zero, if the populations are identical). Here, the numbers of matched pairs is 11, hence the sampling distribution of T can be approximated by a normal distribution. The mean and standard deviation for the given problem are as follows:

$$\mu_T = 0 \text{ (assuming identical nature of two services)}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{11(12)(23)}{6}} = \sqrt{\frac{3036}{6}} = \sqrt{506} = 22.494$$

Suppose, if there are any matched pairs with zero difference, then σ_T should be calculated by discarding such observations.

Now, the test statistic value is computed as

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{-22 - 0}{22.494} \Longrightarrow Z = -0.9780$$

The two tailed p-value can be obtained using the standard normal probability table, i.e., 2(0.1660) = 0.332 > 0.05. This indicates that there is no evidence to reject the H_0, and it can be concluded that the delivery times of services one and two are identical.

Let us consider another example for a better understanding of the concept.

In a study of bilingual coding, 6 bilingual (French and English) college students an article written in French, and each answers a series of 25 multiple-choice questions covering the content of the article. The questions are written in French in one examination and the other examination is in English. The score (total correct) is:

Table 5.5

Examination in French	11	12	16	22	25	25
Examination in English	10	13	17	19	21	24

Is this evidence at the 5 percent significance level that there is difficulty in transferring information from one language to another?

H_0 : There is no difficulty in transferring information from one language to another.

H_1 : There is a difficulty in transferring information from one language to another.

In order to test the given hypothesis, the test statistic is to be computed. This requires computation of T, μ_T and σ_T. The table 5.6 consists of absolute values, ranks and signed ranks

Table: 5.6

Student	Examination in French	Examination in English	Difference	Absolute Difference	Rank	Signed Rank
1	11	10	1	1	2.5	+2.5
2	12	13	-1	1	2.5	-2.5
3	16	17	-1	1	2.5	-2.5
4	22	19	3	3	5	+5
5	25	21	4	4	6	+6
6	25	24	1	1	2.5	+2.5
						T = 11

The tied observations are assigned the average ranks. Absolute difference 1 is repeated four times thus the ranks 1,2,3 and 4 that are to be assigned to these numbers are replaced by the average of the four ranks i.e., 2.4. Here n = 6, the hypothesis is tested using the critical values of the statistic. The overall rank sum is T = 11. The mean and standard deviation for the given problem are as follows:

$$\mu_T = 0 \text{ (assuming identical nature of two languages)}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = \sqrt{\frac{6(7)(13)}{6}} = \sqrt{\frac{546}{6}} = \sqrt{91} = 9.5394$$

Now, the test statistic value is computed as

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{11 - 0}{9.5394} \Rightarrow Z = 1.1531$$

The Wilcoxon table value at n=6 and $\alpha = 0.05$ is 0 which is less than the Z value, i.e., 1.1531 > 0. This indicates that there is evidence to reject the H_0, and it can be concluded that the there is a difficulty in transforming information from one language to another.

5.4 Kruskal-Wallis H test

In usual hypothesis testing problems, we come across the situations where there will be a need to compare two groups. But, if there are more than two groups which are to be compared then the conventional independent samples t-test (when data is assumed to be normal and interval/ratio scale) or the non-parametric approach of Mann-Whitney U test fails to provide such comparisons. Hence, Kruskal and Wallis framed a statistical methodology that accommodates the problem of ‘k’ populations at a time. So, the test was named after them as KRUSKAL-WALLIS test. Whenever, the data from $k \geq 3$ populations or ordinal (or) whenever the normality assumption is violated then the Kruskal-Wallis test is an alternate statistical procedure for testing whether the populations are identical (or) the populations are with same median.

Assumptions

- The samples in each population (group) are drawn at random.
- The populations (groups) are independent of each other
- The populations have similar distribution shape and variability

This test is based on the sum of the ranks for each sample and using the rank position of the population, it assesses whether the population median values come from population with the same global median. That is, in hypothetical terms

H_0 : All the ‘k’ populations are identical (have same median)

H_1 : All the ‘k’ populations are not identical

The following are the steps which are to be used in computing the Kruskal-Wallis test value under the setup of ‘k’ populations. Let ‘k’ be the number of populations and n_k be the number of samples in each of the k^{th} population.

1. First rank all the n_k samples in descending order. (i.e., the lowest value will be assigned rank 1)
2. If there are any ties in ranking the data samples, assign the ranks by taking the average of these ranks.
3. Now, compute the totals for each of the k populations, R_1,R_2,…R_k (say)
4. The Kruskal-Wallis test statistic is computed using

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Here k = number of populations

n_i = size of the i^{th} population

Ri = rank sum of the i^{th} population

$n = \sum_{i=1}^{k} n_i$ = Total number of observations

5. To determine the critical value, we can now use the χ^2 distribution table (Since the Kruskal-Walls test approximates to χ^2 distribution with k-1 degrees of freedom.)

i.e., $H \sim \chi^2_{k-1}$ d.f

If computed H > critical value at α = 0.5, we will have an evidence to reject the null hypothesis, otherwise accept.

In order to understand the methodology of this test, let us consider an example of best selling candies which are with high calories. Consider data on calorie content which is taken from the samples of M & M’s, Kit Kat and Milky Way II. Here, the problem is to test for significant differences in the calorie content of these candies (α = 0.5)

M & M's	230	210	240	250	230
Kit Kat	225	205	245	235	220
Milky Way II	200	208	202	190	180

From the given information, we have k = 3; $n_1 = n_2 = n_3 = 5$ and n = 5+5+5 = 15. Now, we need to compute R_i^2, for this, first rank all the observations in descending order.

M & M's	Rank	Kit Kat	Rank	Milky Way II	Rank
230	10.5	225	9	200	3
210	7	205	5	208	6
240	13	245	14	202	4
250	15	235	12	190	2
230	10.5	220	8	180	1
R_1 =	56	R_2 =	48	R_3 =	16

Here, the lowest value is 180, observed in Milky Way II is assigned rank 1, the next is 190 and is given rank 2 and this process is continued until all the data observations are given ranks. If we observe, the observation 230 appears twice with ranks 10 and 11, so, the average this turns to be 10.5. from the rank sums, it can be understood that the calorie content is more in M & M's candy and least in Milky Way II. However, there is a need to have evidence about the significant difference in calorie content of these candies, for which, the hypothesis can be defined as

H_0 : The calorie content is same in all the three candies

H_1 : The calorie content is not same in all the three candies

The test statistic, H is given by

$$H = \frac{12}{15(15+1)}\left[\frac{(56)^2}{5} + \frac{(48)^2}{5} + \frac{(16)^2}{5}\right] - 3(15+1)$$

$$= \frac{12}{15(16)}\left[\frac{3136+2304+256}{5}\right] - 3(16)$$

$$= \frac{12}{240}\left[\frac{5696}{5}\right] - 48$$

$$= (0.05)(1139.2) - 48$$

$$= 56.96 - 48$$

$H = 8.96$

For k = 3, the $\chi^2_{k-1} = \chi^2_{3-1} = \chi^2_2$ at 0.05 level is 7.378

$$\therefore H > \chi^2_2 \Rightarrow 8.96 > 7.37$$

This creates an evidence to reject the null hypothesis and we can claim that the calorie content is not same in all the three candies. However, it can be concluded from the rank sums that, the samples of M & M's candy contain high calorie content and next to it is samples of Kit Kat and the least calorie is observed in Milky Way II.

Now consider another example where three admission test preparation programs are being evaluated. The scores obtained by a sample of 20 people who used the programs provided the following data. α = 0.05.

Program		
A	B	C
540	450	600
400	540	630
490	400	580
530	410	490
490	480	590
610	370	620
	550	570

The null and alternative hypothesis for the above problem can be defined as

H_0: There is no difference among the three test preparation programs.
H_1: There is a significant difference among the three test preparation programs.

From the given information, we have k = 3; $n_1 = 6$, $n_2 = n_3 = 7$ and n = 6+7+7 = 20. Now, we need to compute R_i^2, for this, first rank all the observations in descending order.

A	Rank	B	Rank	C	Rank
540	11.5	450	5	600	17
400	2.5	540	11.5	630	20
490	8	400	2.5	580	15
530	10	410	4	490	8
490	8	480	6	590	16
610	18	370	1	620	19
		550	13	570	14
R_1=	58	R_2=	43	R_3=	109

The test statistic, H is given by

$$H = \frac{12}{20(20+1)}\left[\frac{(58)^2}{6} + \frac{(43)^2}{7} + \frac{(109)^2}{7}\right] - 3(20+1)$$

$$= \frac{12}{20(21)}\left[\frac{3364}{6} + \frac{1849}{7} + \frac{11881}{7}\right] - 3(21)$$

$$= (0.02857)(2522.0952) - 63$$

$$= 72.0563 - 63$$

$$H = 9.0563$$

For k = 3, the $\chi^2_{k-1} = \chi^2_{3-1} = \chi^2_2$ at 0.05 level is 7.378

$$\therefore H > \chi^2_2 \Longrightarrow 9.0563 > 7.378$$

This creates evidence to reject the null hypothesis and we can claim that the test preparation programs are different from each other. However, it can be concluded from the rank sums that, the scores obtained from test C are the highest among the three tests, scores from test A are next highest and scores obtained in test B are the least.

5.5 Wald-Wolfowitz Run test

This test is used to examine whether the random samples have come from populations with same distribution. Let 'r' and 's' be the number of samples drawn randomly from two independent populations. Let x_1, x_2,…x_r be the ordered sample from first population, P_1 (say) and y_1, y_2, …,y_s be the ordered sample from second population, P_2 (say). Here, the hypothesis is to test if the samples have been drawn from same population i.e., $P_1 = P_2$.

Let us now define the hypothesis as

H_0 : Two samples come from populations having same distribution

H_1 : Two samples come from populations having different distribution

Let 'U' denote the number of runs. A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of other kind. To obtain r, let us combine two samples by arranging the observations in the order of magnitude, as

$x_1x_2y_1y_2x_3y_3\ldots x_ky_{k+1}\ldots$

$$E(U) = \mu = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$V(U) = \sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

and the test statistic for testing the hypothesis is given by

$$Z = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0, 1) \text{ asymptotically}$$

Here U is the observed number of runs. n_1 and n_2 are the number of observations in first and second population.

To understand the procedure of conducting Wald-Wolfowitz run test, let us consider an example of two methods of teaching. A group of students was randomly divided into two groups. One group was taught to read using a uniform method of teaching under teacher's direction. The second group was taught to read using an individual method of teaching with programmed workbook. The following are the test scores which projects the reading ability of two groups of students.

First Group	227	176	252	149	16	55	234	194	247	92	184	147	88	161	171
Second Group	202	14	165	171	292	271	151	235	147	99	63	284	53	228	271

Test the equality of the distribution of scores of two groups.

As per the procedure, first we need to define the hypothesis then an arrangement of given scores of two groups is to be made as per the order of magnitude.

H_0 : The distribution of scores from first and second groups are same.

H_1 : The distribution of scores from first and second groups are not same.

Arrangement of sequence and defining runs:

Data	14	16	53	55	63	88	92	99	147	147	149	151	161	165	171
Group	2	1	2	1	2	1	1	2	1	2	1	2	1	2	1
Data	171	176	184	194	202	227	228	234	235	247	252	271	271	284	292
Group	2	1	1	1	2	1	2	1	2	1	1	2	2	2	2

Total number of runs U = 23 and $n_1 = n_2 = 15$.

$$E(U) = \mu = \frac{2(15)(15)}{15+15} + 1$$

$$= \frac{2 \times 225}{30} + 1$$

$$= \frac{450}{30} + 1$$

$$= 15 + 1$$

$$E(U) = \mu = 16$$

$$V(U) = \sigma^2 = \frac{2(15 \times 15)(2(15 \times 15) - 15 - 15)}{(15+15)^2(15+15-1)}$$

$$= \frac{2(225)(2(225) - 30)}{(30)^2(29)}$$

$$= \frac{450(450 - 30)}{(900)(29)}$$

$$= \frac{450 \times 420}{900 \times 29} = \frac{189{,}000}{26{,}100}$$

$$\Rightarrow Var(U) = 7.2413$$

$$\therefore Z = \frac{U - E(U)}{\sqrt{V(U)}} \Rightarrow Z = \frac{23 - 16}{\sqrt{7.2413}}$$

$$= \frac{7}{2.6909}$$

$$\Rightarrow Z_{cal} = 2.6013$$

At α = 0.05, the Z_{cri} value is 1.96

$$\therefore Z_{cal} > Z_{cri} \Rightarrow 2.6013 > 1.96$$

This claims that the distribution of test scores is not the same in first and second groups.

Let us consider another example of winning or losing of a local softball team. The following are the outcomes of the last 24 games played by a local softball team. The letter W signifies a win and L a loss. W, L, L, L, W, L, L, W, L, L, W, L, L, W, L, W, L, L, L, L, W, L, W, L. Is this dataset consistent with randomness?

In order to test the given statement about the randomness, first we need to define the hypothetical statements which are given below

H_0 : Wins and losses are random.

H_1 : Wins and losses are not random.

From the given information, we need to identify the pattern of runs which is denoted with a flower bracket as below

$$\underbrace{W}\ \underbrace{LLL}\ \underbrace{W}\ \underbrace{LL}\ \underbrace{W}\ \underbrace{LL}\ \underbrace{W}\ \underbrace{LL}\ \underbrace{W}\ \underbrace{L}\ \underbrace{W}\ \underbrace{LLLL}\ \underbrace{W}\ \underbrace{L}\ \underbrace{W}\ \underbrace{L}$$

Total number of runs = U =16, n_1 = 8 and n_2 = 16

$$E(U) = \mu = \frac{2(8)(16)}{8+16} + 1$$

$$= \frac{2 \times 128}{24} + 1$$

$$= \frac{256}{24} + 1$$

$$= 10.6667 + 1$$

$$E(U) = \mu = 11.6667$$

$$V(U) = \sigma^2 = \frac{2(8 \times 16)(2(8 \times 16) - 8 - 16)}{(8 + 16)^2(8 + 16 - 1)}$$

$$= \frac{2(128)(2(128) - 24)}{(24)^2(23)}$$

$$= \frac{256(256 - 24)}{(576)(23)}$$

$$= \frac{256 \times 232}{576 \times 23} = \frac{59{,}392}{13{,}248}$$

$$\Rightarrow Var(U) = 4.4830$$

$$\therefore Z = \frac{U - E(U)}{\sqrt{V(U)}} \Rightarrow Z = \frac{16 - 11.6667}{\sqrt{4.4830}}$$

$$= \frac{4.3333}{2.1173}$$

$$\Rightarrow Z_{cal} = 2.0466$$

At α = 0.05, the Z_{cri} value is 1.96

$$\therefore Z_{cal} > Z_{cri} \Rightarrow 2.0466 > 1.96$$

This provides evidence to claim that the wins and losses of the last 24 games are not at random.

5.6 Chi-Square test

The Chi-Square test is a distribution free statistic, usually use for testing the goodness of fit of data whether it is consistent with respect to the hypothesized distribution and another use is to test for independence between two attributes or categorical variables. Many statistical tests are approximated to Chi-Square distribution. It describes the magnitude of difference between observed and expected frequencies.

Chi-Square test for Goodness of fit

This test is applied when you have a categorical variable from a single population. It is used to compare observed data with expected data under a specific hypothesis. Means, the interest lies in knowing the “goodness of fit” between observed and expected data. For this test, the null hypothesis can be stated as that "there is no significant difference between the expected and observed result". In simple, the Chi-square test for goodness of fit is used to how much the observed data is fit for a hypothesized distribution.

Let o_i, denotes the number of counts in i^{th} category of a categorical variable and e_i represent the expected count of the i^{th} category of the same categorical variable. Here i= 1, 2,…k; k being the number of categories. The expected value of i^{th} category can be computed as

$e_i = n * p_i$; p_i is the proportion or percentage of i^{th} category and n is total count.

For this particular test, the hypothesis takes the following form

H_0 : The data are consistent with a specified distribution.

H_1 : The data are not consistent with a specified distribution.

Then the Karl Pearson χ^2 test is given by

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(o_i - e_i)^2}{e_i}\right] \sim \chi^2_{k-1}\ d.f.$$

If $\chi^2_{cal} \leq \chi^2_{k-1}$, there will not be any evidence to reject the null hypothesis, otherwise.

The implementation of the procedure of goodness of fit can be illustrated using the following example of Acme toy company prints baseball cards.

The company claims that 30% of the cards are rookies, 60% of veterans and 10% are All-stars. A random sample of 100 cards is taken, of which, 50 rookies, 45 veterans and 5 All-stars are observed. Is this consistent with Acme's claim? Test this at α = 5% level of significance.

For the given problem, first we need to define the null hypothesis as

H_0 : The proportions of rookies, veterans and All-stars is 30%, 60% and 10% respectively.

H_1 : At least one of the proportion in null hypothesis is false.

The expected frequencies can be computed as $e_i = n * p_i$

i.e., $e_1 = 100 * 0.30 = 30$; $e_2 = 100 * 0.60 = 60$; $e_3 = 100 * 0.10 = 10$

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3}$$

$$= \frac{(50 - 30)^2}{30} + \frac{(45 - 60)^2}{60} + \frac{(5 - 10)^2}{10}$$

$$= \frac{(20)^2}{30} + \frac{(-15)^2}{60} + \frac{(-5)^2}{10}$$

$$= \frac{400}{30} + \frac{225}{60} + \frac{25}{10}$$

$$= 13.33 + 3.75 + 2.5$$

$$\chi^2 = 19.58$$

The $\chi^2_{k-1} = \chi^2_{3-1} = \chi^2_2$ at 0.05 is 5.99

$$\therefore \chi^2_{cal} = 19.58 > \chi^2_2 = 5.99$$

This claims to reject the null hypothesis and with evidence we can conclude that the proportions claimed by Acme's are not correct.

Another example is as follows

Over a long period of time, a research team monitored the number of car accidents which occurred in a particular county. The following table summarizes the data relating to the day of the week on which the accident occurred.

Day	M	Tu	W	Th	F	S	Su
No. of Accidents	60	54	48	53	53	75	77

Investigate the hypothesis that these data are a random sample from uniform distribution.
The null and alternative hypothesis can be defined as
H_0 : The data are a random sample from uniform distribution.

H_1 : The data are not a random sample from uniform distribution.

The expected frequencies are based on uniform distribution and each e_i can be computed as

$$\frac{1}{7}(60 + 54 + 48 + 53 + 53 + 75 + 77) = 60$$

The Chi square statistic is computed using the following table

Day	o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2$	$\frac{(o_i - e_i)^2}{e_i}$
M	60	60	0	0	0.000
Tu	54	60	-6	36	0.600
W	48	60	-12	144	2.400
Th	53	60	-7	49	0.817
F	53	60	-7	49	0.817
S	75	60	15	225	3.750
Su	77	60	17	289	4.817
					χ^2=13.200

The $\chi^2_{k-1} = \chi^2_{7-1} = \chi^2_6$ at 0.05 is 12.592

$$\therefore \chi^2_{cal} = 13.2 > \chi^2_6 = 12.592$$

This claims to reject the null hypothesis and with evidence we can conclude that the dta does not follow uniform distribution. i.e., The accidents occurring on different days of the week are not uniform.

Chi-square test for independence of attributes

This test is used to determine whether there is a significant association between two categorical variables. The null and alternative hypothesis can be stated as

H_0 : The two categorical variables are not associated.

H_1 : The two categorical variables are associated.

Let us consider two attributes (categorical variables) A and B in such a way that A and B are divided into r and c subcategories i.e., A_1, A_2, …A_r and B_1, B_2, …B_c respectively. Let (A_i) and (B_j) denote the number of frequency (count) possessing attribute A and B respectively. In other words, O_{ij} represents the observed frequency of i^{th} row and j^{th} column of the categorical variables A and B. In order to test the null hypothesis, first the expected frequencies are to be computed in the following way:

$$E_{ij} = \frac{R_i \times C_j}{N} \text{ (or) } E_{ij} = \frac{Row\ Total * Column\ Total}{Total\ sample\ size}$$

Here R_i is the sum or total of the i^{th} subcategory of attribute A

C_j is the sum or total of j^{th} subcategory of attribute B

N is the total number; i = 1, 2, …r; j = 1, 2, …c

The number of degrees of freedom for the test is (r-1)(c-1) with 'r' number of rows and 'c' number of columns.

Now, we need to compute the χ^2 test statistic as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}\right] \sim \chi^2_{(r-1)(c-1)}\ d.f.$$

If $\chi^2_{cal} > \chi^2_{(r-1)(c-1)}$, then we do not accept the null hypothesis.

Let us consider an example of a public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (republic, democrat or independent). The following table shows the distribution of the voters.

		Voting Preferences			Row Total
		Republican	Democrat	Independent	
Gender	Male	200	150	50	400
	Female	250	300	50	600
Column Total		450	450	100	1000

Do the men's voting preferences significantly differ from the women's preferences? Use a 0.05 level of significance.

In the given information on voting preferences, the statement to be tested is defined in terms of hypothesis as

H_0 : Gender and voting preferences are independent.

H_1 : Gender and voting preferences are not independent.

The expected frequencies are calculated as follows

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{400 \times 450}{1000} = \frac{180000}{1000} = 180$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{400 \times 450}{1000} = \frac{180000}{1000} = 180$$

$$E_{13} = \frac{R_1 \times C_3}{N} = \frac{400 \times 100}{1000} = \frac{40000}{1000} = 40$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{600 \times 450}{1000} = \frac{270000}{1000} = 270$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{600 \times 450}{1000} = \frac{270000}{1000} = 270$$

$$E_{23} = \frac{R_2 \times C_3}{N} = \frac{600 \times 100}{1000} = \frac{60000}{1000} = 60$$

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{13} - E_{13})^2}{E_{13}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} + \frac{(O_{23} - E_{23})^2}{E_{23}}$$

$$= \frac{(200 - 180)^2}{180} + \frac{(150 - 180)^2}{180} + \frac{(50 - 40)^2}{40} + \frac{(250 - 270)^2}{270} + \frac{(300 - 270)^2}{270} + \frac{(50 - 60)^2}{60}$$

$$= \frac{(20)^2}{180} + \frac{(-30)^2}{180} + \frac{(10)^2}{40} + \frac{(-20)^2}{270} + \frac{(30)^2}{270} + \frac{(-10)^2}{60}$$

$$= \frac{400}{180} + \frac{900}{180} + \frac{100}{40} + \frac{400}{270} + \frac{900}{270} + \frac{100}{60}$$

$\chi^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67$

$\Longrightarrow \chi^2 = 16.2$

Degrees of freedom: (r-1)(c-1) = (2-1)(3-1) = 2

χ_2^2 at 0.05 level of significance = 5.99

$$\therefore \chi_{cal}^2 > \chi_2^2 \Rightarrow 16.2 > 5.99$$

With this result, there is evidence to reject the null hypothesis. Thus, we can conclude that the voting and gender preference are not independent. In other words, we can say that there is a relationship between gender and voting preference.

Let us have one more illustration to have a clear picture about the computational aspects and the practical application of this test. A market research firm has distributed samples of a new shampoo to a variety of individuals. The following data summarizes the comments of these individuals about the shampoo as well as provide the age group into which they fall.

Rating	Age Group (in years)		
	15-20	21-30	Over 30
Excellent	18	20	41
Good	25	27	43
Fair	17	15	26
Poor	3	2	8

Do these data prove that different age groups have different opinions about the shampoo at 5 percent level of significance?

From the above explained situation on comments given by individuals about the shampoo, the null and alternative hypothesis can be defined as

H_0 : Age group and Ratings (opinions) are independent.

H_1 : Age group and Ratings (opinions) are not independent.

Before going to compute the expected frequencies, we need to calculate the row and column totals which are given in the following table.

Rating	Age Group (in years)			**Row Total**
	15-20	21-30	Over 30	
Excellent	18	20	41	**79**
Good	25	27	43	**95**
Fair	17	15	26	**58**
Poor	3	2	8	**13**
Column Total	**63**	**64**	**118**	**245**

The expected frequencies can now be calculated as

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{79 \times 63}{245} = \frac{4977}{245} = 20.3142$$

$$E_{12} = \frac{R_1 \times C_2}{N} = \frac{79 \times 64}{245} = \frac{5056}{245} = 20.6367$$

$$E_{13} = \frac{R_1 \times C_3}{N} = \frac{79 \times 118}{245} = \frac{9322}{245} = 38.0489$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{95 \times 63}{245} = \frac{5985}{245} = 24.4286$$

$$E_{22} = \frac{R_2 \times C_2}{N} = \frac{95 \times 64}{245} = \frac{6080}{245} = 24.8163$$

$$E_{23} = \frac{R_2 \times C_3}{N} = \frac{95 \times 118}{245} = \frac{11210}{245} = 45.7551$$

$$E_{31} = \frac{R_3 \times C_1}{N} = \frac{58 \times 63}{245} = \frac{3654}{245} = 14.9143$$

$$E_{32} = \frac{R_3 \times C_2}{N} = \frac{58 \times 64}{245} = \frac{3712}{245} = 15.1510$$

$$E_{33} = \frac{R_3 \times C_3}{N} = \frac{58 \times 118}{245} = \frac{6844}{245} = 27.9347$$

$$E_{41} = \frac{R_4 \times C_1}{N} = \frac{13 \times 63}{245} = \frac{819}{245} = 3.3428$$

$$E_{42} = \frac{R_4 \times C_2}{N} = \frac{13 \times 64}{245} = \frac{832}{245} = 3.3959$$

$$E_{43} = \frac{R_4 \times C_3}{N} = \frac{13 \times 118}{245} = \frac{1534}{245} = 6.2612$$

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{13} - E_{13})^2}{E_{13}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$
$$+ \frac{(O_{23} - E_{23})^2}{E_{23}} + \frac{(O_{31} - E_{31})^2}{E_{31}} + \frac{(O_{32} - E_{32})^2}{E_{32}} + \frac{(O_{33} - E_{33})^2}{E_{33}}$$
$$+ \frac{(O_{41} - E_{41})^2}{E_{41}} + \frac{(O_{42} - E_{42})^2}{E_{42}} + \frac{(O_{43} - E_{43})^2}{E_{43}}$$

$$= \frac{(18-20.3142)^2}{20.3142} + \frac{(20-20.6367)^2}{20.6367} + \frac{(41-38.0489)^2}{38.0489} + \frac{(25-24.4286)^2}{24.4286} + \frac{(27-24.8163)^2}{24.8163} + \frac{(43-45.7551)^2}{45.7551} + \frac{(17-14.9143)^2}{14.9143} + \frac{(15-15.1510)^2}{15.1510} + \frac{(26-27.9347)^2}{27.9347} + \frac{(3-3.3428)^2}{3.3428} + \frac{(2-3.3959)^2}{3.3959} + \frac{(8-6.2612)^2}{6.2612}$$

$$= \frac{(-2.3142)^2}{20.3142} + \frac{(-0.6367)^2}{20.6367} + \frac{(2.9511)^2}{38.0489} + \frac{(0.5714)^2}{24.4286} + \frac{(2.1837)^2}{24.8163} + \frac{(-2.7551)^2}{45.7551} + \frac{(2.0857)^2}{14.9143} + \frac{(-0.1510)^2}{15.1510} + \frac{(-1.9347)^2}{27.9347} + \frac{(-0.3428)^2}{3.3428} + \frac{(-1.3959)^2}{3.3959} + \frac{(1.7388)^2}{6.2612}$$

$$= \frac{4.5548}{20.3142} + \frac{0.4054}{20.6367} + \frac{8.7090}{38.0489} + \frac{0.3265}{24.4286} + \frac{4.7685}{24.8163} + \frac{7.5906}{45.7551} + \frac{4.3501}{14.9143} + \frac{0.0228}{15.1510} + \frac{3.7431}{27.9347} + \frac{0.1175}{3.3428} + \frac{1.9485}{3.3959} + \frac{3.0234}{6.2612}$$

$$\chi^2 = 0.2242 + 0.0196 + 0.2288 + 0.0133 + 0.1921 + 0.1659 + 0.2916 + 0.0015 + 0.1339 + 0.0351 + 0.5737 + 0.4828$$

$$\Rightarrow \chi^2 = 2.3625$$

Degrees of freedom: (r-1)(c-1) = (4-1)(3-1) = 6

χ^2_6 at 0.05 level of significance = 12.59

$$\therefore \chi^2_{cal} < \chi^2_2 \Rightarrow 2.3625 < 12.59$$

With this result, there is evidence to reject the null hypothesis. Thus, we can conclude that the voting and gender preference are not independent. In other words, we can say that there is a relationship between gender and voting preference.

EXERCISES

Problems on Sign Test

17. The following table lists the preferences indicated by 10 individuals in taste tests involving two brands of a product.

Individual	Brand A versus Brand B	Individual	Brand A versus Brand B
1	+	6	+
2	+	7	-
3	+	8	+
4	-	9	-
5	+	10	+

With α = 0.05, test for a significant difference in the preferences for the two brands. A plus indicates a preference for brand A over brand B.

18. The following hypothesis test is to be conducted.

$$H_0 : \text{Median} \leq 150$$

$$H_1 : \text{Median} > 150$$

A sample of size 30 yields 22 cases in which a value greater than 150 is obtained, three cases in which a value of exactly 150 is obtained, and five cases in which a value less than 150 is obtained. Use α = 0.01 and conduct the hypothesis test.

19. Are stock splits beneficial to stockholders? SNL Securities studied stock splits in the banking industry over an 18-month period and found that stock splits tended to increase the value of an individual's stock holding. Assume that of a sample of 20 recent stock splits, 14 led to an increase in value, four led to a decrease in value, and two resulted in no change. Suppose a sign test is to be used to determine whether stock splits continue to be beneficial for holders of bank stocks.
    a. What are the null and alternative hypothesis?
    b. With α = 0.05, what is your conclusion?

20. A poll asked 1253 adults a series of questions about the state of the economy and their children's future. One question was, "Do you expect your children to have a better life than you have had, a worse life, or a life about as good as yours?" The responses were 34% better , 29% worse, 33% about the same, and 4% not sure. Use the sign test and a

0.05 level of significance to determine whether more adults feel their children will have a better future than feel their children will have a worse future. What is your conclusion?

21. Nielsen Media Research identified American Idol and Dancing with the Stars as the two top-rated prime-time television shows (USA Today, April 14, 2008). In a local television preference survey, 750 individuals were asked to indicate their favorite prime-time television show: Three hundred thirty selected American Idol, 270 selected Dancing with the Stars, and 150 selected another television show. Use a .05 level of significance to test the hypothesis that there is no difference in the preference for the American Idol and Dancing with the Stars television shows. What is your conclusion?

22. Competition in the personal computer market is intense. A sample of 450 purchases showed 202 Brand A computers, 175 Brand B computers, and 73 other computers. Use a .05 level of significance to test the null hypothesis that Brand Aand Brand B have the same share of the personal computer market. What is your conclusion?

23. The median annual income of subscribers to Shutterbug magazine is $75,000 (Home Theater website, August 18, 2008). A sample of 300 subscribers to Popular Photography & Imaging magazine found 165 subscribers with an annual income over $75,000 and 135 with an annual income under $75,000. Can you conclude that the median annual income of Popular Photography & Imaging subscribers differs from the median annual income of Shutterbug subscribers? Use $\alpha = 0.05$.

24. The median number of part-time employees at fast-food restaurants in a particular city was known to be 18 last year. City officials think the use of part-time employees may be increasing. A sample of nine fast-food restaurants showed that seven restaurants were employing more than 18 part-time employees, one restaurant was employing exactly 18 part-time employees, and one restaurant was employing fewer than 18 part-time employees. Can it be concluded that the median number of part-time employees has increased? Test using $\alpha = 0.05$.

25. Net assets for the 50 largest stock mutual funds show a median of $15 billion (The Wall Street Journal, March 2, 2009). A sample of 10 of the 50 largest bond mutual funds follows.

Bond Fund	Net Assets	Bond Fund	Net Assets
Fidelity Intl Bond	6.1	T Rowe Price New Income	6.9
Franklin CA TF	11.7	Vanguard GNMA	15
American Funds	22.4	Oppenheimer Intl Bond	6.6
Vanguard Short Term	9.6	Dodge & Cox Income	14.5
PIMCO: Real Return	4.9	iShares: TIPS Bond	9.6

Using the median, can it be concluded that bond mutual funds are smaller and have fewer net assets than stock mutual funds? Use α = 0.05.

a. What are the hypotheses for this test?

b. What is the p-value? What is your conclusion?

26. The median annual income for families living in the United States is $56,200 (The New York Times Almanac, 2008). Annual incomes in thousands of dollars for a sample of 50 families living in Chicago, Illinois, are shown. Use the sample data to see if it can be concluded that the families living in Chicago have a median annual income greater than $56,200. Use α = 0.05. What is your conclusion?

66.3	60.2	49.9	75.4	73.7
65.7	61.1	123.8	57.3	48.5
74	146.3	92.2	43.7	86.9
59.7	64.2	56.2	48.9	109.6
39.8	60.9	79.7	42.3	52.6
60.9	43.5	61.7	54.7	95.2
70.4	43.8	57.8	83.5	56.5
51.3	42.9	87.5	43.6	67.2
48.7	79.1	61.9	53.4	56.2
57	49.6	109.5	42.1	74.6

27. A Pew Research Center survey asked adults if their ideal place to live would have a faster pace of life or a slower pace of life (USA Today, February 13, 2009). A preliminary sample of 16 respondents showed 4 preferred a faster pace of life, 11 preferred a slower place of life, and 1 said it did not matter.

    a. Are these data sufficient to conclude there is a difference between the preferences for a faster pace of life or a slower pace of life? Use α = 0.05. What is your conclusion?

    b. Considering the entire sample of 16 respondents, what is the percentage who would like a faster pace of life? What is the percentage who would like a slower pace of life? What recommendation do you have for the study?

28. An engineering firm is involved in selecting a computer system, and the choice has been narrowed to two manufacturers. The firm submits eight problems to the two computer manufacturers and has each manufacturer measure the number of seconds required to solve the design problem with the manufacturer's software. The times for the eight design problems are given below.

Design problem	1	2	3	4	5	6	7	8
Time with computer A	15	32	17	26	42	29	12	38
Time with computer B	22	29	1	23	46	25	19	47

Determine the p-value of the sign test when testing the hypothesis that there is no difference in the distribution of the time it takes the two types of software to solve problems.

29. To test the hypothesis that the median weight of 16-year-old females from Los Angeles is at least 110 pounds, a random sample of 200 such females was chosen. If 120 females weighed less than 110 pounds, does this discredit the hypothesis? Use the 5 percent level of significance. What is the p-value?

30. In 2004, the national median salary of all U.S. financial accountants was $124,400. A recent random sample of 14 financial accountants showed 2007 incomes of (in units of $1,000)

125.5, 130.3, 133.0, 102.6, 198.0, 232.5, 106.8,
114.5, 122.0, 100.0, 118.8, 108.6, 312.7, 125.5

Use these data to test the hypothesis that the median salary of financial accountants in 2007 was not greater than in 2004. What is the p-value?

31. An experiment was initiated to study the effect of a newly developed gasoline detergent on automobile mileage. The following data, representing mileage per gallon before and after the detergent was added for each of eight cars, resulted.

Car	Mileage without Additive	Mileage with Additive
1	24.2	23.5
2	30.4	29.6
3	32.7	32.3
4	19.8	17.6
5	25.0	25.3
6	24.9	25.4
7	22.2	20.6
8	21.5	20.7

Find the p-value of the test of the hypothesis that mileage is not affected by the additive when

a. the sign test is used; and (b) the signed rank test is used.

32. An engineer claims that painting the exterior of a particular aircraft affects its cruising speed. To check this, the next 10 aircraft off the assembly line were flown to determine cruising speed prior to painting, and were then painted and reflown. The following data resulted.

Aircraft	Cruising Speed (knots)	
	Not Painted	Painted
1	426.1	416.7
2	418.4	403.2
3	424.4	420.1
4	438.5	431
5	440.6	432.6
6	421.8	404.2
7	412.2	398.3
8	409.8	405.4
9	427.5	422.8
10	441.2	444.8

Do the data uphold the engineer's claim?

33. Ten pairs of duplicate spectrochemical determinations for nickel are presented below. The readings in column 2 were taken with one type of measuring instrument and those in column 3 were taken with another type.

Sample	Duplicates		Sample	Duplicates	
1	1.94	2	6	1.96	1.98
2	1.99	2.09	7	1.95	2.03
3	1.98	1.95	8	1.96	2.03
4	2.07	2.03	9	1.92	2.01
5	2.03	2.08	10	2	2.12

Test the hypothesis, at the 5 percent level of significance, that the two measuring instruments give equivalent results.

Problems on Wilcoxon Signed-rank Test

34. Two fuel additives are tested to determine their effect on miles per gallon for passenger cars. Test results for 12 cars follow; each car was tested with both fuel additives. Use α = 0.05 and the Wilcoxon signed-rank test to see whether there is a significant difference between the median miles per gallon for the additives.

Additive			Additive		
Car	1	2	Car	1	2
1	20.12	18.05	7	16.16	17.2
2	23.56	21.77	8	18.55	14.98
3	22.03	22.57	9	21.87	20.03
4	19.15	17.06	10	24.23	21.15
5	21.23	21.22	11	23.21	22.78
6	24.77	23.8	12	25.02	23.7

35. A sample of 10 men was used in a study to test the effects of a relaxant on the time required to fall asleep. Data for 10 subjects showing the number of minutes required to fall asleep with and without the relaxant follow. Use a 0.05 level of significance to determine whether the relaxant reduces the median time required to fall asleep. What is your conclusion?

Relaxant			Relaxant		
Subject	No	Yes	Subject	No	Yes
1	15	10	6	7	5
2	12	10	7	8	10
3	22	12	8	10	7
4	8	11	9	14	11
5	10	9	10	9	6

36. Percents of on-time arrivals for flights in 2006 and 2007 were collected for 11 randomly selected airports. Data for these airports follow (Research and Innovative Technology Administration website, August 29, 2008). Use α = 0.05 to test the hypothesis that there is no difference between the median percent of on-time arrivals for the two years. What is your conclusion?

	Percent on Time	
Airport	2006	2007
Boston Logan	71.78	69.69
Chicago O’Hare	68.23	65.88
Chicago Midway	77.98	78.4
Denver	78.71	75.78
Fort Lauderdale	77.59	73.45
Houston	77.67	78.68
Los Angeles	76.67	76.38
Miami	76.29	70.98
New York (JFK)	69.39	62.84
Orlando	79.91	76.49
Washington (Dulles)	75.55	72.42

37. The PGA Players Championship was held at the Sedgefield Country Club in Greensboro, North Carolina, August 11–17, 2008. Shown here are first-round and second-round scores for a random sample of 11 golfers. Use α = 0.05 to determine whether the first- and second-round median scores for golfers in the Players Championship differed significantly.

Golfer	1st Round	2nd Round
Marvin Laird	63	74
Jimmy Walker	70	73
Kevin Chappell	72	70
Kevin Duke	65	71
Andrew Buckle	70	74
Paul Claxton	69	73
Larry Mize	72	71
Chris Riley	68	70
Bubba Watson	70	68
Carlos Franco	71	71
Richard Johnson	72	69

38. Fifteen cities, of roughly equal size, are chosen for a traffic safety study. Eight of them are randomly chosen, and in these cities a series of newspaper articles dealing with traffic safety is run over a 1 month period. The number of traffic accidents reported in the month following this campaign is as follows:

Treatment group	19	31	39	45	47	66	74	81
Control group	28	36	44	49	52	52	60	72

Determine the exact p-value when testing the hypothesis that the articles have not had any effect.

Problems on Kruskal-Wallis Test

39. A sample of 15 consumers provided the following product ratings for three different products. Five consumers were randomly assigned to test and rate each product. Use the Kruskal-Wallis test and α = 0.05 to determine whether there is a significant difference among the ratings for the products.

Product		
A	B	C
50	80	60
62	95	45
75	98	30
48	87	58
65	90	57

40. Forty-minute workouts of one of the following activities three days a week will lead to a loss of weight. The following sample data show the number of calories burned during 40- minute workouts for three different activities. Do these data indicate differences in the amount of calories burned for the three activities? Use a .05 level of significance. What is your conclusion?

Swimming	Tennis	Cycling
408	415	385
380	485	250
425	450	295
400	420	402
427	530	268

41. Condé Nast Traveler magazine conducts an annual survey of its readers in order to rate the top 80 cruise ships in the world (Condé Nast Traveler, February 2008). With 100 the highest possible rating, the overall ratings for a sample of ships from the Holland America, Princess, and Royal Caribbean cruise lines are shown. Use the Kruskal-Wallis test with α = 0.05 to determine whether the overall ratings among the three cruise lines differ significantly. What is your conclusion?

Holland America		Princess		Royal Caribbean	
Ship	Rating	Ship	Rating	Ship	Rating
Amsterdam	84.5	Coral	85.1	Adventure	84.8
Maasdam	81.4	Dawn	79	Jewel	81.8
Ooterdam	84	Island	83.9	Mariner	84
Volendam	78.5	Princess	81.1	Navigator	85.9
Westerdam	80.9	Star	83.7	Serenade	87.4

42. A large corporation sends many of its first-level managers to an off-site supervisory skills training course. Four different management development centers offer this course. The director of human resources would like to know whether there is a difference among the quality of training provided at the four centers. An independent random sample of five employees was chosen from each training center. The employees were then ranked 1 to 20 in terms of supervisory skills. A rank of 1 was assigned to the employee with the best supervisory skills. The ranks are shown. Use α = 0.05 and test whether there is a significant difference among the quality of training provided by the four programs.

Course			
A	B	C	D
3	2	19	20
14	7	16	4
10	1	9	15
12	5	18	6
13	11	17	8

Problems on Wald-Wolfowitz Run test

43. A production run of 50 items resulted in 11 defectives, with the defectives occurring on the following items (where the items are numbered by their order of production): 8, 12, 13, 14, 31, 32, 37, 38, 40, 41, 42. Can we conclude that the successive items did not constitute a random sample?

44. The win-lose record of a certain basketball team for its last 50 consecutive games was as follows-

W W W W W W L W W W W W W L W L W W W L L W W W W L W W W L L W W W
W W W L L W W L L L W W L W W W

Apply run test to test that sequence of wins and losses is random.

45. A random sample of 8 household is selected from a village A whose daily expense on milk is 11, 15, 17, 19, 25, 27, 31, 33. Another sample of 9 households is selected from village B whose expense on milk is 12, 16, 20, 22, 28, 30, 36, 38, 42. Test whether the households of the two villages are same on spending daily milk expenses.

46. Two makes of cars were sampled randomly, to determine the mileage (in thousands) until the brakes required relining. Calculate the probability value for the null hypothesis that make A and make B are similar.

Make A	Make B
30	22
41	26
48	32
49	39
61	

47. A random sample of 8 men’s heights and an independent random sample of 8 women’s heights were observed as follows

Men’s Height (M)	Women’s Height (M)
65	62
67	63
69	64
70	65
71	66
73	68
76	69
77	71

Use run test to check whether the distribution of heights is same in men and women.

48. 12 people are polled to find out if they use a certain product and the outcomes are recorded according to their sex by the symbols M and F. Use run test to check for the randomness if the observed sequence is M M F F F M F F M M M M.

49. A machine is adjusted to dispense acrylic paint thinner into a container. Would you say that the amount of paint thinner being dispensed by this machine varies randomly if the contents of the next 15 containers are measured and found to be 3.6, 3.9, 4.1, 3.6, 3.8, 3.7, 3.4, 4.0, 3.8, 4.1, 3.9, 4.0, 3.8, 4.2 and 4.1 liters? Use a 0.01 level of significance.

50. A series of 20 coin tosses might produce the following sequence of heads (H) and tails (T). H H T T H T H H H H T H H T T T T T H H. Test for the randomness of occurrence of heads and tails.

51. A college is interested in whether there is any predictable change in the academic performance of international students from the first to the second semester of their first year at a university. A random sample of 20 students is selected and their first and second semester grade point averages are

First Sem	1.53	2.00	1.93	3.90	2.14	1.52	0.91	1.95	3.00	1.67
	2.78	1.21	1.66	1.75	2.96	1.50	2.25	2.66	1.87	3.50
Second Sem	3.67	2.74	3.50	3.27	1.97	1.54	3.42	1.04	2.45	2.09
	2.00	3.00	1.78	2.31	2.25	2.20	0.91	1.52	1.61	2.56

Problems on Chi-Square for Goodness of fit

52. Test the following hypotheses by using the χ^2 goodness of fit test.

$$H_0: p_A = 0.40, p_B = 0.40, \text{ and } p_C = 0.20$$

$$H_a: \text{The population proportions are not}$$

$$p_A = 0.40, p_B = 0.40, \text{ and } p_C = 0.20$$

A sample of size 200 yielded 60 in category A, 120 in category B, and 20 in category C. Use $\alpha = 0.01$ and test to see whether the proportions are as stated in H_0.

a. Use the p-value approach.

b. Repeat the test using the critical value approach.

53. Suppose we have a multinomial population with four categories: A, B, C, and D. The null hypothesis is that the proportion of items is the same in every category. The null hypothesis is

$$H_0: p_A = p_B = p_C = p_D = 0.25$$

A sample of size 300 yielded the following results.

A: 85 B: 95 C: 50 D: 70

Use $\alpha = 0.05$ to determine whether H_0 should be rejected. What is the p-value?

54. During the first 13 weeks of the television season, the Saturday evening 8:00 p.m. to 9:00 p.m. audience proportions were recorded as ABC 29%, CBS 28%, NBC 25%, and independents 18%. A sample of 300 homes two weeks after a Saturday night schedule revision yielded the following viewing audience data: ABC 95 homes, CBS 70 homes, NBC 89 homes, and independents 46 homes. Test with $\alpha = 0.05$ to determine whether the viewing audience proportions changed.

55. M&M/MARS, makers of M&M® chocolate candies, conducted a national poll in which more than 10 million people indicated their preference for a new color. The tally of this poll resulted in the replacement of tan-colored M&Ms with a new blue color. In the brochure “Colors,” made available by M&M/MARS Consumer Affairs, the distribution of colors for the plain candies is as follows:

Brown	Yellow	Red	Orange	Green	Blue
30%	20%	20%	10%	10%	10%

In a follow-up study, samples of 1-pound bags were used to determine whether the reported percentages were indeed valid. The following results were obtained for one sample of 506 plain candies.

Brown	Yellow	Red	Orange	Green	Blue
177	135	79	41	36	38

Use α = 0.05 to determine whether these data support the percentages reported by the company.

56. Where do women most often buy casual clothing? Data from the U.S. Shopper Database provided the following percentages for women shopping at each of the various outlets (The Wall Street Journal, January 28, 2004).

Outlet	**Percentage**	**Outlet**	**Percentage**
Wal-Mart	24	Kohl’s	8
Traditional department Stores	11	Mail order	12
JC Penney	8	Other	37

The other category included outlets such as Target, Kmart, and Sears as well as numerous smaller specialty outlets. No individual outlet in this group accounted for more than 5% of the women shoppers. A recent survey using a sample of 140 women shoppers in Atlanta, Georgia, found 42 Wal-Mart, 20 traditional department store, 8 JC Penney, 10 Kohl’s, 21 mail order, and 39 other outlet shoppers. Does this sample suggest that women shoppers in Atlanta differ from the shopping preferences expressed in the U.S. Shopper Database? What is the p-value? Use α = 0.05. What is your conclusion?

57. The American Bankers Association collects data on the use of credit cards, debit cards, personal checks, and cash when consumers pay for in-store purchases (The Wall Street Journal, December 16, 2003). In 1999, the following usages were reported.

In-Store Purchase	Percentage
Credit card	22
Debit card	21
Personal check	18
Cash	39

A sample taken in 2003 found that for 220 in-stores purchases, 46 used a credit card, 67 used a debit card, 33 used a personal check, and 74 used cash.

a. At α = 0.01, can we conclude that a change occurred in how customers paid for in-store purchases over the four-year period from 1999 to 2003? What is the p-value?

b. Compute the percentage of use for each method of payment using the 2003 sample data. What appears to have been the major change or changes over the four-year period?

c. In 2003, what percentage of payments was made using plastic (credit card or debit card)?

58. The Wall Street Journal’s Shareholder Scoreboard tracks the performance of 1000 major U.S. companies (The Wall Street Journal, March 10, 2003). The performance of each company is rated based on the annual total return, including stock price changes and the reinvestment of dividends. Ratings are assigned by dividing all 1000 companies into five groups from A (top 20%), B (next 20%), to E (bottom 20%). Shown here are the one-year ratings for a sample of 60 of the largest companies. Do the largest companies differ in performance from the performance of the 1000 companies in the Shareholder Scoreboard? Use $\alpha = 0.05$.

A	B	C	D	E
5	8	15	20	12

59. How well do airline companies serve their customers? A study showed the following customer ratings: 3% excellent, 28% good, 45% fair, and 24% poor (BusinessWeek, September 11, 2000). In a follow-up study of service by telephone companies, assume that a sample of 400 adults found the following customer ratings: 24 excellent, 124 good, 172 fair, and 80 poor. Is the distribution of the customer ratings for telephone companies different from the distribution of customer ratings for airline companies? Test with $\alpha = 0.01$. What is your conclusion?

Problems on Chi-Square for Independence of Attributes

60. The following 2 X 3 contingency table contains observed frequencies for a sample of 200. Test for independence of the row and column variables using the χ^2 test with $\alpha = 0.05$.

Row Variable	A	B	C
P	20	44	50
Q	30	26	30

61. Visa Card USA studied how frequently consumers of various age groups use plastic cards (debit and credit cards) when making purchases (Associated Press, January 16, 2006). Sample data for 300 customers shows the use of plastic cards by four age groups.

	Age Group			
Payment	18–24	25–34	35–44	45 and over
Plastic	21	27	27	36
Cash or check	21	36	42	90

a. Test for the independence between method of payment and age group. What is the p-value? Using α = 0.05, what is your conclusion?

b. If method of payment and age group are not independent, what observation can you make about how different age groups use plastic to make purchases?

c. What implications does this study have for companies such as Visa, MasterCard, and Discover?

62. With double-digit annual percentage increases in the cost of health insurance, more and more workers are likely to lack health insurance coverage (USA Today, January 23, 2004). The following sample data provide a comparison of workers with and without health insurance coverage for small, medium, and large companies. For the purposes of this study, small companies are companies that have fewer than 100 employees. Medium companies have 100 to 999 employees, and large companies have 1000 or more employees. Sample data are reported for 50 employees of small companies, 75 employees of medium companies, and 100 employees of large companies.

	Health Insurance		
Size of Company	Yes	No	Total
Small	36	14	50
Medium	65	10	75
Large	88	12	100

a. Conduct a test of independence to determine whether employee health insurance coverage is independent of the size of the company. Use α = 0.05. What is the p-value, and what is your conclusion?

b. The USA Today article indicated employees of small companies are more likely to lack health insurance coverage. Use percentages based on the preceding data to support this conclusion.

63. The National Sleep Foundation used a survey to determine whether hours of sleeping per night are independent of age (Newsweek, January 19, 2004). The following show the hours of sleep on weeknights for a sample of individuals age 49 and younger and for a sample of individuals age 50 and older.

	Hours of sleep				
Age	Fewer than 6	6 to 6.9	7 to 7.9	8 or more	Total
49 or younger	38	60	77	65	240
50 or older	36	57	75	92	260

a. Conduct a test of independence to determine whether the hours of sleep on weeknights are independent of age. Use α = 0.05. What is the p-value, and what is your conclusion?

b. What is your estimate of the percentage of people who sleep fewer than 6 hours, 6 to 6.9 hours, 7 to 7.9 hours, and 8 or more hours on weeknights?

64. Samples taken in three cities, Anchorage, Atlanta, and Minneapolis, were used to learn about the percentage of married couples with both the husband and the wife in the workforce (USA Today, January 15, 2006). Analyze the following data to see whether both the husband and wife being in the workforce is independent of location. Use a .05 level of significance. What is your conclusion? What is the overall estimate of the percentage of married couples with both the husband and the wife in the workforce?

	Location		
In Workforce	Anchorage	Atlanta	Minneapolis
Both	57	70	63
Only one	33	50	90

65. On a syndicated television show the two hosts often create the impression that they strongly disagree about which movies are best. Each movie review is categorized as Pro (“thumbs up”), Con (“thumbs down”), or Mixed. The results of 160 movie ratings by the two hosts are shown here.

	Host B		
Host A	Con	Mixed	Pro
Con	24	8	13
Mixed	8	13	11
Pro	10	9	64

Use the chi-square test of independence with a .01 level of significance to analyze the data. What is your conclusion?

.

References

1. Levin D M, Krehbiel T C, Berenson M L and Viswanathan P K (2011), Business Statistics – A First Course, 5/e, Pearson.
2. Sheldon M Ross (2006), Introductory Statistics, 2/e, Academic Press.
3. Gupta S C and Kapoor V K (2002), Fundamentals of Mathematical Statistics, 11/e, Sultan Chand and Sons.
4. Gupta S C and Kapoor V K (2007), Fundamentals of Applied Statistics, 4/e, Sultan Chand and Sons.
5. Spiegel M R and Stephens L J (2010), Schaum's Outlines – Statistics, 4/e, Tata McGraw Hill Education Private Limited.
6. Pillai R S N and Bagavathi V (2005), Statistics – Theory and Practice, S.Chand and Company Limited.
7. Pillai R S N and Bagavathi V (2003), Practical Statistics, 2/e, S.Chand and Company Limited.
8. Irwin Miller and Marylees Miller (2002), Mathematical Statistics, 6/e, Prentice-Hall of India Pvt. Ltd.
9. Anderson D R, Sweney D J and Williams T A(2005), Statistics for Business and Economics, 8/e, Thomson.

****