GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

Seshadri Rao Knowledge Village, Gudlavalleru – 521 356.

Department of Computer Science and Engineering

HANDOUT

On

STATISTICAL METHODS USING R SOFTWARE

Vision

To provide quality education embedded with knowledge, ethics and advanced skills and preparing students globally competitive to enrich the civil engineering research and practice.

Mission

- To aim at imparting integrated knowledge in basic and applied areas of civil engineering to cater the needs of industry, profession and the society at large.
- To develop faculty and infrastructure making the department a centre of excellence providing knowledge base with ethical values and transforming innovative and extension services to the community and nation.
- To make the department a collaborative hub with leading industries and organizations, promote research and development and combat the challenging problems in civil engineering which leads for sustenance of its excellence.

Program Educational Objectives

PEOI : Exhibit their competence in solving civil engineering problems in practice, be employed in industries and undergo higher studies.

PEOII : Adapt to changing technologies with societal relevance for sustainable development in the field of their profession.

PEO III: Develop multidisciplinary team work with ethical attitude &social responsibility and engage in life - long learning to promote research and development in the profession.

STATISTICAL METHODS USING R SOFTWARE

Class & Sem.: IV B. Tech. – I Sem. Academic Year: 2018 – 2019
Branch : CE / EEE / ME / ECE / CSE / IT Credits: 3

1. Brief History and Scope of the Subject:

What is R Programming Language?

R language is an open source program maintained by the R core-development team – team of volunteer developers from across the globe. R language used for performing statistical operations and is available from the R-Project website www.r-project.org. R is a command line driven program. The user enters commands at the prompt (> by default) and each command is executed one at a time.

Many routines have been written for R analytics by people all over the world and made freely available from the R project Website as packages. However, the basic installation (for Linux, Windows, or Mac) contains a powerful set of tools for most purposes.

R is a consolidated environment for performing statistical operations and generating R data analysis reports in graphical or text formats. R commands entered in the console are evaluated and executed. R cannot handle certain auto-formatting characters such as en-dashes or smart quotes; therefore, you need to be careful while copying and pasting commands into R from other applications. Let us now learn something about the History of R in this Introduction to R Programming.

History of R language

John Chambers and colleagues developed R at Bell Laboratories. R is an implementation of the S programming Language and combines with lexical scoping semantics inspired by Scheme. R was named

partly after the first names of two R authors. The project conceives in 1992, with an initial version released in 1995 and a stable beta version in 2000. Let us also Understand in this Introduction to R Programming Tutorial, that Why should learn R Programming.

R Applications

- Many data analysts and research programmers use R because R is the most prevalent language. Hence, R is used as a fundamental tool for finance.
- Many quantitative analysts use R as their programming tool. Hence, R helps in data importing and cleaning, depending on what manner of strategy you are using on.
- R is best for data Science because it gives a broad variety of statistics. In addition, R provides the environment for statistical computing and design. Rather R considers as an alternate execution of S.

2. Pre-Requisites:

(a) Genuine Interest in statistical programming

(b) Computer ready to run R and R Studio

(c) Basic understanding of statistics and data structure

(d) NO prior knowledge in programming is required

3. Course Objectives:

(a) To understand statistical concepts
(b) To know R software

4. Course Outcomes:

Students should be able to:
CO1: examine the relationship between the variables and forecast
CO2: apply suitable range of statistical tests
CO3: use R for statistical programming, Computation, Graphics, and modeling
CO4: expand their knowledge of R on their own.

5. Program Outcomes: Graduates of the Computer Science and Engineering Program will have

a) an ability to apply knowledge of mathematics, science, and engineering
b) an ability to design and conduct experiments, as well as to analyze and interpret data
c) an ability to design a system, component, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability
d) an ability to function on multidisciplinary teams
e) an ability to identify, formulate, and solve engineering problems
f) an understanding of professional and ethical responsibility
g) an ability to communicate effectively
h) the broad education necessary to understand the impact of engineering solutions in a global, economic, environmental, and societal context
i) a recognition of the need for, and an ability to engage in life-long learning,
j) a knowledge of contemporary issues

6. Mapping of Course Outcomes with Program Outcomes:

	a	b	c	d	e	f	g	h	i	j	k	l
CO1	H	H			M							
CO2	H	H			M							
CO3	M	H										
CO4	M	H										

7. Prescribed Text Books:

1. S.C.Gupta and V.K.kapoor-Fundamentals of Mathematical Statistics-S.chand & co
2. Probability and Statistics, Dr. T. K. V. Iyengar, Dr. B. Krishna Gandhi, S. Ranganatham and Dr. M.V. S. S. N. Prasad, S. Chand & Company Ltd.
3. Peter Dalgaard. Introductory Statistics with R (Paperback) 1st Edition Springer-Verlag New
   York, Inc. ISBN 0-387-95475-9
4. W. N. Venables and B. D. Ripley. 2002. Modern Applied Statistics with S. 4th Edition.
   Springer. ISBN 0-387-95457-0

8. Reference Books:

1. An Introduction to R. Online manual at the R website at http://cran.r-project.org/manuals.html

2. Andreas Krause, Melvin Olson. 2005. The Basics of S-PLUS. 4th edition. Springer-Verlag, New York. ISBN 0-387-26109-5.

9. URLs and Other e – Learnings:

https://onlinecourses.nptel.ac.in/noc17_ma17

https://en.wikipedia.org/wiki/R_(programming_language)

a.

10. Digital Learning Meterial:

11. Lecture Schedule:

Topic	Theory
UNIT –1: Correlation-Regression	
Simple correlation ,types of correlation, correlation co-efficient	1
Problems on correlation coefficient	1
Problems on correlation coefficient	1
rank correlation -problems	2
Simple regression –problems	1
Simple regression –problems	1
UNIT – 2: Testing Of Hypothesis (Large Samples)	
Population, samples, parameter, statistic, random sample, sampling distribution, standard error.	1
Test of hypothesis- simple, composite hypotheses, Null hypothesis and alternative Hypothesis, Test statistic.	1
Type I & Type 2 errors in sampling.	1
L.O.S – one tail and two tail tests, degrees of freedom,	1
procedure of testing of hypothesis	1
UNIT – 3: One Sample Significance Tests	
Large Sample: Test for single mean	1
Problems	2
Single Proportion	1
Problems	2
Small Sample : t – test for single mean	1
Problems	2
UNIT – 4: Two Sample Significance Tests	
Large sample : test for two means,	1
Problems	2
Test for two proportions	1
Problems	2
Small sample: t – test for two means	1
Problems	2
F – test	1
Problems	2
UNIT – 5: Introduction to R Software	
Introduction to R session(Downloading and installing R, Explaining R screen)	2

R as a calculator getting help and loading packages	3
Data entry and exporting data :Vectors, Matrices and Data frames lists	1
Calculating correlation coefficient using R	2
Finding regression lines – interpretations using R	2
UNIT – 6 : One Sample and Two Sample Tests using R	
Large sample: Calculating Z value for single mean - interpretation	2
Calculating Z value for two means - interpretation	1
Calculating Z value for single proportion - interpretations	2
Calculating Z value for two proportions - interpretations	1
Small Sample : Calculating t for single mean - interpretations	1
Calculating t for two means - interpretations	1
Calculating F value - interpretations	2
TOTAL	**54**

12. Seminar Topics

1. Correlation-Regression
2. Testing Of Hypothesis (Large Samples)
3. One Sample Significance Tests
4. Two Sample Significance Tests
5. Introduction to R Software
6. One Sample and Two Sample Tests using R

UNIT – I

CORRELATION - REGRESSION

Objectives:

- To define correlation, regression and analyze the types

Syllabus:

Simple Correlation for ungrouped data, rank correlation and simple regression.

Course Outcomes: After completion of the course the student should be able to

- examine correlation between variables and find the relation between them
- construct the regression lines and forecast

Learning Material

Correlation

Correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

The correlation expresses the relationship or interdependence of two sets of variables upon each other. One variable may be called the independent and the other variable dependent etc...

Correlation: If the change in one variable affects the change in the other variable then the two variables are said to be correlated and the relationship is called correlation

Types of Correlation:

Correlation is classified into many types.

- Positive and negative
- Simple and multiple
- Partial and total
- Linear and non-linear

Positive and negative correlation:

If two variables deviate in the same direction i.e. an increase or decrease in the value of one variable is accompanied by an increase or decrease of the other variable then the correlation is called positive or direct correlation.

If two variables deviate in opposite directions i.e. an increase or decrease in the values of one variable is accompanied by a decrease or increase in the value of the other variable then the correlation is called negative or inverse correlation.

Simple and Multiple Correlations:

If the study is between two variables then it is simple correlation and examples are quantity of money and price level, demand and price etc, But in multiple correlations we study more than two variables simultaneously and examples are the relationship of price, demand, supply of a commodity.

Partial and Total Correlation:

Two variables excluding some other variables is called partial correlation. Example, we study price and demand, eliminating the supply side. In total correlation, all the facts are taken into account.

Linear and non-linear correlation:

If the ratio of change between two variables is uniform, then there will be linear correlation between them.

In a curvilinear or non-linear correlation, the amount of change in one variable does not bear a constant ration of the amount of change in the other variables.

Scatter diagram or scatter gram:

The scatter diagram is pictorial representation by plotting two variables to find out whether there is any relationship between them.

Karl Pearson's correlation coefficient:

Karl Pearson is a British Biometrician and Statistician suggested a mathematical method for measuring the magnitude of linear relationship between two variables. This is known as Pearson's Coefficient of correlation or Product-Moment correlation coefficient. It is denoted by $r_{x,y}$

$$r= \frac{Cov(X,Y)}{\sigma_x\sigma_y} \quad \text{OR} \quad r=\frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{OR} \quad r= \frac{\sum xy}{N\sigma_{x\sigma_y}} \quad \text{OR} \quad r= \frac{(\sum XY*n)-(\sum X*\sum Y)}{\sqrt{(\sum X^2*n-(\sum X)^2)*(\sum Y^2*n-(\sum Y)^2)}}$$

Where n is number of paired observations

Limits of correlation coefficient ($-1 \leq r_{x,y} \leq +1$)

PROBLEM: Calculate coefficient of correlation from the following data.

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Solution:

We have $r= \frac{(\sum XY*n)-(\sum X*\sum Y)}{\sqrt{(\sum X^2*n-(\sum X)^2)*(\sum Y^2*n-(\sum Y)^2)}}$

X	Y	X^2	Y^2	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
70	63	728	651	676

Here n=7

$\therefore r=\frac{(676*7)-(70*63)}{\sqrt{(728*7-70^2)*(651*7-63^2)}} \quad =0.95$

Note: When deviations are taken from an assumed mean the coefficient of correlation is

$$r=\frac{\sum XY-\frac{\sum X\sum Y}{n}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{n})-(\sum Y^2-\frac{(\sum Y)^2}{n})}}$$

Rank correlation coefficient:

The method of finding the coefficient of correlation by ranks. This method is based on ranks and is useful in dealing with qualitative characteristics such as morality, character, intelligence and beauty. Rank correlation is applicable only to the individual observations. The formula for Spearman's rank correlation coefficient is given by

$$\rho = 1\text{-}\frac{6\sum d^2}{n(n^2-1)} \quad \text{(For untied ranks)}$$

Where ρ is rank coefficient of Correlation

d^2 is Sum of the squares of the difference of two ranks

n is Number of paired observations

Properties of rank correlation coefficient:

- **T**he value of ρ lies between 1 and -1
- If ρ=1, there is complete agreement in the order if the ranks and the direction of the rank is same.
- If ρ=-1, then there is complete disagreement in the order of the ranks and they are in opposite directions.

PROBLEM: A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be

Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

Calculate Spearman's rank correlation coefficient.

Solution:

X	Y	Ranks in x	Ranks in y	d_i =x-y	D^2
85	93	2	1	1	1
60	75	4	3	1	1
73	65	3	4	-1	1
40	50	5	5	0	0
90	80	1	2	-1	1
					4

Here N=5 $\sum D^2$=4

Spearman's rank correlation coefficient is

$$\rho = 1\text{-}\frac{6\sum d^2}{n(n^2-1)} \quad =1\text{-}\frac{6*4}{5(5^2-1)} \quad =0.8$$

Equal or Repeated ranks:

If there is more than one item with the same value in the series then the Spearman's formula for calculating the rank correlation coefficient is

$$\rho = 1\text{-}6\left\{\frac{\sum d^2+coreection\ factor\ of\ X\ and\ Y}{n(n2-1)}\right\}$$

Where correction factor(C.F)= m(m²-1)/12

Where m= the number of times the item is repeated

PROBLEM: Obtain the rank correlation coefficient for the following data

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

X	Y	Rank of X(x)	Rank of Y (y)	d=x-y	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				0	72

In X-series, 75 occurs 2 times, so rank = $\frac{2+3}{2}$ =2.5

64 occur 3 times, so rank= $\frac{5+6+7}{3}$ = 6

To $\sum d^2$ we add $\frac{m(m^2-1)}{12}$ for each value repeated, so for 75 m=2, for 64, m=3.

So far X series,C.F is $\frac{2(4-1)}{12}+\frac{3(9-1)}{12}=\frac{5}{2}$

In Y series, 68 occurs twice, so rank = $\frac{3+4}{2}$ = 3.5

68 occurs twice so m=2

So far Y series, C.F is $\frac{2(4-1)}{12}=\frac{1}{2}$ $\therefore\ \rho=\frac{1-6(\sum d^2+\frac{5}{2}+\frac{1}{2)}}{N(N^2-1)}$=0.545

Regression

In regression analysis the nature of actual relationship if it exists, between two (or more variables) is studied by determining the mathematical equation between the variables. It is mainly used to predict or estimate one (the dependent) variable in terms of the other (independent) variable(s).

Definition: Regression is a mathematical measure of the average relationship between two or more variables in terms of original units of the data.

Simple regression: It establishes the relationship between two variables (one dependent and one independent variable)

Linear regression: if the relationship between the two variables is linear and is represented by straight line then it is regression line or the line of average relationship or prediction of equation.

Regression lines are of two types (i) regression line y on x (ii) regression line x on y

The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression.

Uses:

- It is used to estimate the relation between two economic variables like income and expenditure.
- It is highly valuable tool in economic and business.
- It is useful in statistical estimation of demand curves, supply curves, production function, cost function and consumption function etc.

Properties of Regression coefficients:

1. Regression lines pass through the points (x, y)
2. Correlation coefficient is the geometric mean between the regression coefficients
3. If one of the regression coefficients is greater than unity, the other must be less than unity
4. Arithmetic mean of the regression coefficient is greater than the correlation coefficient
5. Regression coefficients are independent of the change of origin but not scale

Deviation taken from arithmetic mean X on Y:

This method is simpler to find the values of a and b. We can find out the deviations of X and Y series from their respective means.

Regression equation X on Y is

$$X-\overline{X} = r\,\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

Regression equation Y on X is

$$(Y-\overline{Y}) = r\,\frac{\sigma_y}{\sigma_x}(X-\overline{X})$$

Where $\overline{X}$ and $\overline{Y}$ be the means of X and Y series

The regression coefficient of X on Y = $r\,\frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2} = b_{xy}$

The regression coefficient of Y on X = $r\,\frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum X^2} = b_{yx}$

PROBLEM:

Find the most likely production corresponding to a rainfall 40 from the following data.

	Rain fall(X)	Production(Y)
Average	30	500kgs
Standard deviation	5	100kgs
Coefficient of correlation	0.8	

We have to calculate the value of Y when X =40

So we have to find the regression equation of Y on X.

Mean of X series, $\overline{X}$ =30; Mean of Y series, $\overline{Y}$ = 500
σ of X series, σ_x= 5 , σ of Y series , σ_y =100
Regression line Y on X

$$(Y-\overline{Y}) = r\,\frac{\sigma_y}{\sigma_x}\,(X-\overline{X}) \quad = (Y-500) = 0.8\,\left(\frac{100}{5}\right)(X-30)$$

When X=40, Y-500=160
Y=660
Hence the expected value of Y is 660kgs.

Deviations taken from the assumed mean:

If the actual mean is fraction this method is used.
In this method we take deviations from the assumed mean instead of A.M

$$X-\overline{X} = r\,\frac{\sigma_x}{\sigma_y}\,(Y-\overline{Y})$$

We can find out the value of r $\frac{\sigma_x}{\sigma_y}$ by applying the following formula

$$r\,\frac{\sigma_x}{\sigma_y} = \frac{\sum dx\,dy - \frac{\sum dx * \sum dy}{n}}{\sum dy^2 - \frac{(\sum dy)^2}{n}} \quad , \quad dx = X-A;\ dy = Y-A$$

Regression equation Y on X is

$$(Y-\overline{Y}) = r\,\frac{\sigma_y}{\sigma_x}\,(X-\overline{X})$$

We can find out the value of r $\frac{\sigma_y}{\sigma_x}$ by applying the following formula

$$r\,\frac{\sigma_y}{\sigma_x} = \frac{\sum dx\,dy - \frac{\sum dx * \sum dy}{n}}{\sum dx^2 - \frac{(\sum dx)^2}{n}}$$

PROBLEM: Price indices of cotton and wool are given below for the 12 months of a year. Obtain the equations of lines of regression between the indices.

X	78	77	85	88	87	82	81	77	76	83	97	93
Y	84	82	82	85	89	90	88	92	83	89	98	99

Calculation of regression equation

X	dx=(X-84)	dx^2	Y	dy=(Y-88)	dy^2	Dxdy
78	-6	36	84	-4	16	24
77	-7	49	82	-6	36	42
85	1	1	82	-6	36	-6
88	4	16	85	-3	9	-12
87	3	9	89	1	1	3
82	-2	4	90	2	4	-4
81	-3	9	88	0	0	0
77	-7	49	92	4	16	-28
76	-8	64	83	-5	25	40
83	-1	1	89	1	1	-1
97	13	169	98	10	100	130
93	9	81	99	11	121	99
1004	-4	488	1061	5	365	287

Regression line X on Y:

$$X-\overline{X} = b_{xy}(Y-\overline{Y})$$

$$b_{xy} = \frac{\sum dx\,dy - \frac{\sum dx * \sum dy}{n}}{\sum dy^2 - \frac{(\sum dy)^2}{n}} = \frac{287-(\frac{-4*5}{12})}{365-\frac{5^2}{12}} = 0.795$$

X-83.7 = 0.795(Y-88.42)

X=0.795Y+13.38

Regression line Y on X:

$$(Y-\overline{Y}) = b_{yx}(X-\overline{X})$$

$$b_{yx} = \frac{\sum dx\,dy - \frac{\sum dx * \sum dy}{n}}{\sum dx^2 - \frac{(\sum dx)^2}{n}} = \frac{287-(\frac{-4*5}{12})}{488-\frac{-4^2}{12}} = 0.59$$

Y-88.42 = 0.59(X-83.67)

Y=0.59X+39.05

PROBLEM:

Determine the equation of a straight line which best fits the data.

X	10	12	13	16	17	20	25
Y	10	22	24	27	29	33	37

Let the required straight line is Y=a+bX

The two normal equations are $\sum Y$ = b$\sum X$ + na

$\sum XY$ = b$\sum X^2$+a$\sum X$

X	X^2	Y	XY
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925
113	1938	182	3186

Substituting the values:

113b+7a =182 ------------------- (1)

1983b+113a=3186---------------- (2)

Then a=0.82, b=1.56

The equation of straight line is Y=0.82 +1.56 X

Assignment-cum-Tutorial Questions
Section A

Objective Questions:

1. The functional relationship of a dependent variable with independent variable is called_________________

2. Scatter diagram of the variate values (X,Y) gives the idea about____
   (a) functional relationship (b) regression model
   (c) distribution of errors (d) none of the above [ ]

3. The range of correlation coefficient is__________________ [ ]
   (a) 0 t o∞ (b) -∞ to ∞ (c) 0 to 1 (d) -1 to 1

4. In calculating r with raw scores, the numerator of r represents___________ [ ]
   (a) the variance of X (b) the variance of Y
   (c) the variance of X multiplied by the variance of Y
   (d) the covariance of X and Y

5. Which of the following values could not represent a correlation Coefficient? [ ]
   (a) r = 0.99 (b) r = 1.09 (c) r = -0.73 (d) r = -1.0

6. The arithmetic mean of two regression coefficients is _________________ to the correlation coefficient.

7. Regression coefficient is independent of____________ [ ]
   (a) Origin (b) scale (c) both (a) & (b) (d) neither (a) nor (b)

8. One regression coefficient is positive then the other regression coefficient is______ [ ]
   (a) Positive (b) negative (c) equal to zero (d) cannot say

9. If the regression equation is equal to Y=23.6−54.2XY=23.6−54.2X, then 23.6 is the _____ while -54.2 is the ____ of the regression line[ ]
   (a) Slope, Regression coefficient (b) Radius, Intercept
   (c) Intercept, Slope (d) Slope, Intercept

10. If the slope of the regression line is calculated to be 2.5 and the intercept 16 then the value of *Y* when *X* is 4 is_______ [ ]
    (a) 16 (b) 66.5 (c) 26 (d) 2.5

11. When two regression lines coincide then r is ______ [ ]
    (a) 0 (b) -1 (c) 1 (d) 0

12. The regression lines cut each other at the point of____ [ ]
    (a) Average of X and Y (b) average of X only
    (c) Average of Y only (d) none

13. Coefficient of correlation is equal to_______________ [ ]
    (a) $b_{xy}{}^* \; b_{yx}$ (b) $\sqrt{b_{xy} * b_{yx}}$ (c) $\sqrt{b_{xy}}$ (d) $\sqrt{b_{yx}}$

14. Which of the following would not allow you to calculate a Correlation ? [ ]
    (a) a negative relationship between X and Y
    (b) a positive relationship between X and Y
    (c) a curvilinear relationship between X and Y

(d) a linear relationship between X and Y

15. Rank the score of 8 in the following set: [ ]

2; 7; 1; 8; 4; 2; 7; 10; 20

(a) 5 (b) 3 (c) 8 (d) 7

16. If $b_{yx} < 0$ and $b_{xy} \leq 0$, then r is [ ]

(a) =0 (b) < 0 (c) > 0 (d) ≠0

Section B

Subjective Questions:

1. Define correlation and types of correlation.
2. Calculate Karl Pearson's correlation coefficient for the following data.

X	380	402	370	365	410	392	385
Y	560	543	564	573	550	554	540

What inference would you draw from estimate?

3. Given n=10,σ_x = 5.4,σ_y= 6.2 and sum of product of deviation from the mean of X and Y is 66 find the correlation coefficient.
4. Find coefficient of correlation between X and Y for the following data.

X	60	62	64	66	68	70	72
Y	61	63	63	63	64	65	67

5. Following are the rank obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in two subjects is related?

Statistics	3	5	8	4	7	10	2	1	6	9
Mathematics	6	4	9	8	1	2	3	10	5	7

6. Compute the rank correlation between Eco marks and Statistics marks as given below:

Eco marks	80	56	50	48	50	62	60
Stats marks	90	75	75	65	65	50	65

7. Price indices of cotton and wool are given below for the 12 months of a year. Obtain the equations of lines of regression between the indices.

X	36	23	27	28	28	29	30	31	33	35
Y	29	18	20	22	27	21	29	27	29	28

Estimate the Y when X is 25.

8. The following calculations have been made for prices of 12 stocks (X) in stock exchange, on a certain day along with the volume of the sales in thousands of shares(Y). From these calculations find the regression equation of prices of stocks, on the volume of the sales of shares.

$\sum X$ =580, $\sum Y$ =370, $\sum XY$ =11499, $\sum X^2$=41658, $\sum Y^2$=17206

9. The equations of two regression lines are 7X-16Y+9=0 and 5Y-4X-3=0. Find the coefficient of correlation and the means of X and Y.
10. Distinguish between correlation and regression.

Section C

GATE/IES/Placement Tests/Other competitive Exam

1. Regression analysis was applied to return rates of sparrow hawk colonies. Regression analysis was used to study the relationship

between return rate (***x***: % of birds that return to the colony in a given year) and immigration rate (***y***: % of new adults that join the colony per year). The following regression equation was obtained.
y= 31.9 – 0.34*x*

Based on the above estimated regression equation, if the return rate were to decrease by 10% the rate of immigration to the colony would:

a. increase by 34%
b. increase by 3.4%
c. decrease by 0.34%
d. decrease by 3.4%

2. Larger values of r^2 (R^2) imply that the observations are more closely grouped about the

a. average value of the independent variables
b. average value of the dependent variable
c. least squares line
d. origin

3. A regression analysis between sales (in $1000) and price (in dollars) resulted in the following equation:
y= 50,000 - 8X

The above equation implies that an

a. increase of $1 in price is associated with a decrease of $8 in sales
b. increase of $8 in price is associated with an increase of $8,000 in sales
c. increase of $1 in price is associated with a decrease of $42,000 in sales
d. increase of $1 in price is associated with a decrease of $8000 in sales

4. If the coefficient of determination is a positive value, then the regression equation

a. must have a positive slope
b. must have a negative slope
c. could have either a positive or a negative slope
d. must have a positive y intercept

5. If the coefficient of determination is equal to 1, then the correlation coefficient

a. must also be equal to 1
b. can be either -1 or +1
c. can be any value between -1 to +1
d. must be -1

6. In regression analysis, if the independent variable is measured in kilograms, the dependent variable

a. must also be in kilograms
b. must be in some unit of weight
c. cannot be in kilograms
d. can be any units

7. If the correlation coefficient is 0.8, the percentage of variation in the response variable explained by the variation in the explanatory variable is
a. 0.80%
b. 80%
c. 0.64%
d. 64%

8. If the correlation coefficient is a positive value, then the slope of the regression line

a. must also be positive
b. can be either negative or positive
c. can be zero
d. can't be zero

9. Regression analysis was applied between $ sales (***y***) and $ advertising (***x***) across all the branches of a major international corporation. The following regression function was obtained.
y= 5000 + 7.25

If the advertising budgets of two branches of the corporation differ by $30,000, then what will be the predicted difference in their sales?

a. $217,500
b. $222,500
c. $5000
d. $7.25

10. Suppose the correlation coefficient between height (as measured in feet) and weight (as measured in pounds) is 0.40. What is the correlation coefficient of height measured in inches versus weight measured in ounces? [12 inches = one foot; 16 ounces = one pound]
a.0.40
b.0.30
c.0.533
d. cannot be determined from information given
e. none of these

Unit-II

Testing Of Hypothesis

Objectives:

- To know Sampling Theory
- Understand how to develop Null and Alternative Hypotheses
- Know the principles of hypothesis testing

Syllabus:

Introduction: Population-Sample-Large sample and Small sample.
Testing of Hypothesis-hypothesis-Null hypothesis-Alternative hypothesis-level of significance-degrees of freedom-One tail and two tailed tests – Procedure of Testing of hypothesis.

Learning Outcomes: The students will be able to

- Understand the concept of sampling
- setup null and alternative hypothesis
- understand the hypothesis testing

Learning Material

Population:

In statistics population does not only refers to people but it may defined as any collection of individuals or objects or units which can be specified numerically.

Population may be mainly classified into two types.

(i) Finite population
(ii) Infinite population

(i) Finite population: The population contains finite number of individuals is called 'finite population'. For example, total number of students in a class.

(ii) Infinite population: The population which contains infinite number of individuals is known as 'infinite population'. For example, the number of stars in the sky.

Parameter: The statistical constants of a population are known as parameter.

For example, mean (μ) and variance (σ^2).

Statistic:

Any function of sample observations is called sample statistic or statistic.

Standard error: The standard deviation of the sampling distribution of a statistic is known as its 'standard error'.

Sample:

A portion of the population which is examined with a view to determining the population characteristics is called a sample. Or A sample is a subset of the population and the number of objects in the sample is called the size of the sample size of the sample is denoted by ‘n’.

Classification of samples: Samples are classified in 2 ways.

(i) **Large sample:** The size of the sample (n) ≥ 30, the sample is said to be large sample.

(ii) **Small sample:** If the size of the sample (n) < 30, the sample is said to be small sample or exact sample.

Statistical Hypothesis:

Hypothesis is a statement or assumption about the population which may or may not be true

Testing of hypothesis:

It is used to testing the hypothesis about the parent population from which the samples are drawn.

Test of Significance:

A very important aspect of the sampling theory is the study of the test of significance, which enables us to decide on the basis of the sample results, if

- ➢ The deviation between the observed sample statistics and the hypothesis parameter value (or)
- ➢ The deviation between two independent sample statistics is significant.

Null Hypothesis: A definite statement about the population parameter. Such hypothesis which is usually a hypothesis of no difference is called ‘Null hypothesis’ and is usually denoted by ‘H_0’.

Alternative Hypothesis: Any hypothesis which is complementary to the null hypothesis is called ‘Alternative hypothesis” and is usually denoted by ‘H_1’.

Eg: If we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e., H_1:$\mu \neq \mu_0$-----------------(i)

H_1:$\mu < \mu_0$------------------(ii)

H_1:$\mu > \mu_0$-------------------(iii)

Then the alternative hypothesis in (i) is known as Two-tailed test and the alternatives in (ii) and (iii)are known as left and right tailed tests respectively.

Critical Region:

The region of rejection of null hypothesis H_0 when H_0 is true is that region of the outcomes at where H_0 is rejected. If the sample points falls in that region is called the 'critical region', size of the critical region is α.

Type-I error:

P (rejecting H_0 /H_0 is true) i.e. when H_0 is true it is to be accepted but it is a rejected. Therefore there is an error

Type-II error:

P (Accepting H_0/H_0 is false) i.e. when H_0 is false it is to be rejected but it is accepted. Therefore there is an error

One tailed and two tailed tests:

If the alternative hypothesis is of the type (< or >) and the entire critical region lies in the normal probability curve on one side then it is said to be one tailed tests (OTT)

Again the one tailed test is two types (i) Right one tailed test (ii) Left one tailed test

If the alternative hypothesis is of the type ($\neq$) and the

Critical region lies in the normal probability curve on both sides then it is said to be two tailed tests (TTT)

Level of significance (LOS): The probability of committing Type-I error is known as the level of significance which is denoted by 'α'. Usually LOS are 10%, 5% or 1%.

Degrees of freedom: It is very clear that in a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of the sample varies since it depends either on the experimenter or on the resources available. Moreover, the test statistic involves the estimated value of the parameter which depends on the number of observations. Hence the sample size plays an important role in testing of hypothesis and is taken care of by degrees of freedom.

Definition: The number of independent observations in a set is called degrees of freedom. It is denoted by ν (read as Nu). In general, the number of degrees of freedom is equal to the total number of observations less than the number of independent constraints imposed on the observations. i.e., in a set of n observations, if k is the number of independent constraints then $\nu = n - k$.

Procedure for testing of hypothesis:

Step (1): Set up Null hypothesis (H_0)

Step (2): Set up Alternative hypothesis (H_1) which enables us to apply one tailed test / Two tailed test.

Step (3): Choose Level of significance (LOS) α

Step (4): Under the null hypothesis H_0, the test statistic Z=$\frac{t-E(t)}{S.E\ of(t)} \sim N(0,1)$ where'tʼ is a statistic

Step (5): Conclusion: If calculated Z < (tabulated) Z_α at α % LOS then accept null hypothesis otherwise reject null hypothesis.

Table: Critical value of Z when n≥30			
Level of significance	1%	5%	10%
Two-Tailed test	2.58	1.96	1.645
Right-Tailed test	2.33	1.645	1.28
Left-Tailed test	-2.33	-1.645	-1.28

UNIT-II

Assignment-Cum-Tutorial Questions

SECTION-A

Objective Questions

1. The statistical constant of the population are called [ ]
   (a) Statistic (b) Parameter (c) Sample Statistic (d) One
2. A sample consists of [ ]
   (a) All units of the population (b) 50% units of the population
   (c) 5% units of the population (d) Any fraction of the population
3. Area of critical region depends on [ ]
   (a) Size of Type-I error (b) size of Type-II error
   (c) Value of the statistic (d) No. of observations
4. A hypothesis is true, but it is rejected, this is an error of type [ ]
   (a) I (b) II (c) I and II (d) None
5. Level of significance is the probability of committing________ [ ]
   (a) Type-I error (b) Type-II error (c) both I and II (d) None
6. Degrees of freedom is related to_______________ [ ]
   (a) No. of observations in a set (b) Hypothesis under test
   (c) No. Of independent observations in a set (d) none of these
7. Whether a test is one-sided or two sided depends on_____________ .
8. A single-tailed test is used when ________________________
9. The sizes of Type-I and Type-II errors are also known as_______________ and ______________
10. A null hypothesis is rejected if the value of a test statistic lies in the _____________
11. A statistical test is a ________________to decide about H_0.
12. The hypothesis which is under test for possible rejection is called _______________.
13. A hypothesis contrary to null hypothesis is known as __________________hypothesis.
14. The number of independent values in a set of values is known as_______________

SECTION-B

SUBJECTIVE QUESTIONS:

1. Define population and sample with one example each.
2. Write a short note on sample theory.
3. Define parameter and statistic
4. Define (a) Critical region (b) Level of significance
5. Define degrees of freedom.
6. Explain about two types of errors in testing of hypothesis.
7. Explain One-tailed and two-tailed tests.
8. Explain about Null hypothesis and Alternative hypothesis.
9. Define simple and composite hypothesis.
10. Explain the procedure for Testing of Hypothesis?

Learning Material

UNIT-III

One Sample Significance Tests

Objectives:

- Compare the difference between a sample mean and proportion with a target value using Z-test for single mean and proportion
- Find the difference between a sample mean and a target value using a one-sample t-test.

Syllabus:

One sample tests: Large sample-Test for single mean, single proportion-
Small sample tests: t-test for single mean

Learning Outcomes:

Students will be able to

- Apply hypothesis test about population mean and proportion
- Analyze single-sample t-test

Prerequisite information: The estimation of population parameters and the testing of hypothesis concerning those parameters are similar techniques, but at the same time there are major differences in interpretation of results arising from each method. When we are concerned with measurement, say, of expenditure on entertainment, the appropriate method would be the process of estimation. When we are involved in decision making such as whether we should raise the price of our product by 5 percent or not, it is the hypothesis testing that would be enable us to take a proper decision. In addition, hypothesis testing is very helpful in examining the validity or otherwise of theories such as wage increase leads to rising prices. It may, however, be noted that sometimes such situations arise that it may be difficult to interpret correctly the results emerging from hypothesis testing.

Basic Terms:

Classification of samples based on size: Samples are classified in 2 ways.

(i) **Large sample:** The size of the sample (n) ≥ 30, the sample is said to be large sample.

(ii) **Small sample:** If the size of the sample (n) < 30, the sample is said to be small sample or exact sample. In this unit one sample tests in large sample and small sample are

Large sample tests: (i) Z-test for single mean (ii) Z-test for single proportion

Small sample tests: t-test for single mean

Significance test for a single mean

Working Rule

Step (1): Null hypothesis:$H_{0:}$μ=μ_0

Step (2): Alternative hypothesis: H_1:μ≠ μ_0 / H_1:μ<μ_0 / H_1:μ>μ_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: We have the following two cases.

Case (1): When the S.D ( ) of population is known. Then the test statistic is

$$Z=\frac{\overline{X}-\mu}{S.E(\overline{X})} \sim N(0,1)$$

Where S.E($\overline{x}$) =$\sigma/\sqrt{n}$

Where σ = standard deviation of population

n=Sample size.

Case (2): When the S.D (σ) of population is unknown. The test statistic is Z=$\frac{\overline{X}-\mu}{S.E(\overline{X})}$

Where S.E ($\overline{x}$) =$s/\sqrt{n}$

Where s = standard deviation of sample

n=Sample size.

Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

(i). A sample of 400 items is taken from a population whose standard deviation is 10. The mean of the sample is 40. Test whether the sample has come from a population with mean 38.Also calculate 95% confidence interval for the population?

Given n=400, $\overline{x}$ =40, μ=38, $\sigma = 10$

Step (1): Null hypothesis:$H_{0:}$μ=38

Step (2): Alternative hypothesis: H_1:μ≠ 38

Step (3): Level of significance: α = 5%

Step (4): Test statistic: When the S.D ( ) of population is known. Then the test

statistic is Z=$\frac{\overline{X}-\mu}{S.E(\overline{X})}$ Where S.E ($\overline{x}$) =$\sigma/\sqrt{n}$

=4

Step (5): Conclusion: Z_{cal} =4, Z_{tab} =1.96

If Z_{cal}> Z_{tab} at 5% LOS. So we reject H_0.

95% confidence interval is ($\overline{x}$ ±1.96 $\sigma/\sqrt{n}$) =(39.02,40.98)

Test of significance of single proportion: Suppose a large random sample of size n has a sample proportion p of members possessing a certain attribute. To test the hypothesis that the proportion P in the population has a specified value p_0.

Step (1): Null hypothesis:$H_{0:}$p=p_0

Step (2): Alternative hypothesis: H_1:p≠ p_0/ H_1:p<p_0 / H_1:p>p_0

Step (3): Level of significance: Choose 5% (or) 1%

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$

Where p=sample proportion=x/n

P=population proportion, Q=1-P

N=sample size

Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.

If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.

Problem:

In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% LOS?

Given n=1000

P=sample proportion of rice eaters=540/1000=0.54

P=population proportion of rice eaters=1/2=0.5 Q=1-P=0.5

Step (1): Null hypothesis:H_0: both rice and wheat are equally popular in the state. i.e P=0.5

Step (2): Alternative hypothesis:

H_1:p≠ 0.5

Step (3): Level of significance: 1% =2.58

Step (4): Test statistic: Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$ =2.532

Step (5): Conclusion: Z_{cal} = 2.532, Z_{tab} =2.58.

If Z_{cal}<Z_{tab} at 1% LOS then we accept H_0.

Small sample tests: t-test for single mean

Degrees of freedom: It is very clear that in a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of the sample varies since it depends either on the experimenter or on the resources available. Moreover, the test statistic involves the estimated value of the parameter which depends on the number of observations. Hence the sample size plays an important role in testing of hypothesis and is taken care of by degrees of freedom.

Definition: The number of independent observations in a set is called degrees of freedom. It is denoted by ν (read as Nu). In general, the number of degrees of freedom is equal to the total number of observations less than the number of independent constraints imposed on the observations. i.e. in a set of n observations, if k is the number of independent constraints then $\nu = n - k$.

Before going to discuss the tests of significance under small samples, we need some knowledge about exact sampling distributions: t- distribution (or Student's t- distribution), F- distribution and χ^2 - distribution (or Chi-Square distribution).

t- distribution: It is discovered by W.S. Gosset in 1908. The statistician Gosset is better known by the pen name (pseudonym) 'student' and hence t- distribution is called student's t- distribution.

In practice, the standard deviation σ is not known and in such a situation the only alternative left is to use S, the sample estimate of standard deviation σ. Thus, the variate $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ is approximately normal provided n is sufficiently large. If n is not sufficiently large (small) the varite $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ is distributed as t and hence, $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$ where $S^2 = \frac{1}{(n-1)}\sum_i (x_i - \bar{x})^2$.

Properties of t- distribution:

(1) The shape of t- distribution is bell-shaped and is symmetrical about mean.
(2) The curve of t- distribution is asymptotic to the horizontal axis.
(3) It is symmetrical about the line t = 0.
(4) The form of the probability curve varies with degrees of freedom.
(5) It is unimodel with mean = median = mode.
(6) The mean of t- distribution is zero and variance depends upon the parameter ν, is called the degrees of freedom.
(7) The t- distribution with ν degrees of freedom approaches standard normal distribution as $\nu \to \infty$, ν being a parameter.

The t- distribution is extensively used in hypothesis about one mean or single mean, or about equality of two means or difference of means when σ is known.

Some applications of t- distribution are:

(1). To test the significance of the difference between two sample means or to compare two samples.
(2). To test the significance of an observed sample correlation coefficient and sample correlation coefficient.
(3). To test the significance of difference between two sample means or to compare two samples.
Assumptions about t- test: t- test is based on the following five assumptions.
(1). The random sample has been drawn from a population.
(2). All the observations in the sample are independent.
(3). The sample size is not large. (One should note that at least five observations are desirable for applying a t- test.)
(4). The assumed value μ_0 of the population mean is the correct value.
(5). The sample values are correctly taken and recorded.
(6).The population standard deviation σ is unknown

In case the above assumptions do not hold good, the reliability of the test decreases.
Confidence limits: For example, 95% confidence limits for the population mean μ are $\bar{x} \pm t_\alpha \frac{S}{\sqrt{n}}$ or $\bar{x} \pm t_\alpha \frac{s}{\sqrt{n-1}}$ where $\alpha = 0.025$ for two-tailed test and $\alpha = 5\%$ for one-tailed test, as mentioned in the above example. i.e. For two-tailed test at α los, the value of is taken for $\alpha / 2$ from statistical tables of t.

Problem: The life expectancy of people in the year 1970 in Brazil is expected to be 50 years. A survey was conducted in 11 regions of Brazil and the data obtained are given below. Do the data confirm the expected view?
Life expectancy (in years): 54.2, 50.4, 44.2, 49.7, 55.4, 57.0, 58.2, 56.6, 61.9, 57.5, 53.4.

Solution:
Null hypothesis, H$_0$: $\mu = 50$.

Alternative hypothesis, H$_1$: $\mu \neq 50$ (Two-tailed test).

Under the null hypothesis H$_0$, the test statistic is

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$, $d_i = x_i - y_i$ and $S^2 = \frac{1}{(n-1)} \sum_i (x_i - \bar{x})^2$.

Calculation of $\bar{x}$ and S^2:

$$\bar{x} = \frac{54.2 + 50.4 + 44.2 + 49.7 + 55.4 + 57.0 + 58.2 + 56.6 + 61.9 + 57.5 + 53.4}{11} = \frac{598.5}{11} = 54.41.$$

and

$$S^2 = \frac{(54.2-54.41)^2+(50.4-54.41)^2+(44.2-54.41)^2+\ldots+(53.4-54.41)^2}{10} = \frac{236.07}{10} = 23.607$$

$\Rightarrow S = 4.853.$

Therefore, the value of test statistic is

$$t = \frac{54.41-50}{4.859/\sqrt{11}} = 3.01.$$

t- critical value or table value at 5% level of significance and 10 degrees of freedom is 2.228 (from t- tables).

Since t- calculated value is greater than t- critical value, we reject the null hypothesis, H_0. i.e. we accept H_1. It means that the life expectance more than 50 years.

Problem: Mean life time of computers manufactured by a company is 1120 hours. (a) Test the hypothesis that mean lifetime of computers has not changed if a sample of 8 computers has a mean lifetime of 1070 hours with a standard deviation of 125 hours. (b) Is there decrease in mean lifetime? Use (i) 0.05 and (ii) 0.01 level of significance.

Solution:

(a) Null hypothesis, H_0: $\mu = 1120$.

Alternative hypothesis, H_1: $\mu \neq 1120$ (Two-tailed test).

Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{(n-1)}$$

Where $\bar{x} = \frac{\sum_i x_i}{n}$ and $s^2 = \frac{1}{n}\sum_i (x_i-\bar{x})^2$, a sample variance.

Calculation: We are given n = 8, $\bar{x} = 1070$, $s = 125$

Therefore, the value of test statistic is

$$t = \frac{1070-1120}{125/\sqrt{7}} = -1.05$$

$or\ |t| = 1.05.$

(i) t- critical value or table value at 5% level of significance with 7 degrees of freedom is 2.365

Since t- calculated value is less than t- critical value, we accept the null hypothesis, H_0.
i.e. the sample has been come from the population whose mean life-time of computers is 1120 hours.

(ii) t- critical value or table value at 1% level of significance with 7 degrees of freedom is 3.499 and we accept null hypothesis H$_0$.

(b) Null hypothesis, H$_0$: $\mu = 1120$.

Alternative hypothesis, H$_1$: $\mu < 1120$ (left-tailed test).

Under the null hypothesis H$_0$, the test statistic is

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n-1}} \sim t_{(n-1)}$$

Where $\overline{x} = \frac{\sum_i x_i}{n}$ and $s^2 = \frac{1}{n} \sum_i (x_i - \overline{x})^2$, a sample variance.

Calculation: We are given n = 8, $\overline{x} = 1070$, $s = 125$

Therefore, the value of test statistic is

$$t = \frac{1070 - 1120}{125 / \sqrt{7}} = -1.05$$
$$or \;\; t = -1.05$$

(i) t- critical value at 5% level of significance with 7 degrees of freedom for left tailed test is -1.895. Since t- calculated value is greater than t- critical value, we accept the null hypothesis, H$_0$.

i.e .the sample has been come from the population whose mean life-time of computers is 1120 hours.

(ii) t- critical value at 1% level of significance with 7 degrees of freedom is -2.998.

Since t- calculated value is greater than t- critical value, we accept the null hypothesis. i.e. it indicates no decrease in mean lifetime at either of the level of significances.

(These are the applications of t- test for single mean.)

Unit-VI

One Sample and Two Sample Significance Tests using R

Objectives:

- Define R code to compare the difference between a sample mean and proportion with a target value using Z-test for single and two means, proportions
- Find the difference between a sample mean and a target value using a one and two-samples t-test with R code.
- Explain how to use an *F*-test to judge whether two population variances are equal with R code.

Syllabus:

Large Sample: Calculating Z value for single and two means-interpretations; Calculating Z value for single and two proportions-interpretations.

Small Sample: Calculating t value for single and two means-interpretations; Calculating F value-interpretations.

Learning Outcomes: The students will be able to

- Apply R code to test the hypothesis about population one and two means, proportions
- Analyzeone and two-samples t-test with R code.
- Apply R code to*F*-test, to judge whether the two population variances are equal.

<u>Learning Material</u>

Calculating Z value for single and two means-interpretations:

- The test of statistic z for single mean is : $z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$
- The test of statistic z for two means is : $z = \frac{\overline{x_1} - \overline{x_2}}{\sigma\left[\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]}$ or

$$z = \frac{\overline{x_1} - \overline{x_2}}{\left[\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]}$$

Problem

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim by the manufacturer?

Solution

The null hypothesis is that μ ≥ 10000. We begin with computing the test statistic.

```
> xbar = 9900
> mu = 10000
> sigma = 120
> n = 30
> z = (xbar−mu)/(sigma/sqrt(n))
> z
[1] −4.5644
```

We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1−alpha)
> −z.alpha
[1] −1.6449
```

Conclusion:

The test of statistic -4.5644 is less than the critical value -1.6449. Hence, at .05 significance level, we reject the claim that mean lifetime of a light bulb is above 10,000 hours.

Calculating Z value for single and two proportions-interpretations:

- The test of statistic z for single proportion is : $z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$
- The test of statistic z for two proportions is : $z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Problem:

In a random sample of 1000 persons from town A, 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between from town A and town B, so far as the proportions of wheat consumers is concerned at 10% significance level.

Solution

The null hypothesis is that $P_1 = P_2$

```
> p1=400/1000

> p2=400/800

> n1=1000

> n2=800

> p=((n1*p1)+(n2*p2))/(n1+n2)

> q=1-p

> z=(p1-p2)/sqrt((p*q)*((1/n1)+(1/n2)))

>z

[1] -4.242641
```

Now, compute the critical value at 0.1 level of significance

```
>alpha=0.1
>zhalf.alpha=qnorm(1-(alpha/2))
>c(-zhalf.alpha,zhalf.alpha)
[1] -1.644854  1.644854
```

Conclusion:

The test of statistic -4.242641 does not lies between critical value -1.644854 and 1.644584. Hence, at 0.1 significance level, we reject the null hypothesis.

Calculating t value for single and two means-interpretations:

- The test of statistic t for single mean is : $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

- The test of statistic t for two means is : $t = \frac{\overline{x_1} - \overline{x_2}}{s\left[\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]}$

Problem:

Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 28 light bulbs, it was found that they only last 9,900 hours on average. Assume the sample standard deviation is 125 hours. At .05 significance level, can we reject the claim by the manufacturer?

Solution

The null hypothesis is that µ ≥ 10000.

```
>mu=10000

>xbar=9900

> n=28

> s=125

> t=(xbar-mu)/(s/sqrt(n))

>t

[1] -4.23320
```

then compute the critical value at .05 significance level.

```
>alpha=0.05

>talpha=qt(1-alpha,df=n-1)

> -talpha

[1] -1.703288
```

Conclusion:

The test of statistic -4.23320is less than the critical value-1.703288. Hence, at .05 significance level, we can reject the claim that mean lifetime of a light bulb is above 10,000 hours.

Calculating F value-interpretations:

- The test of statistic F for significant difference between the variances is: $F = \frac{S_1^2}{S_2^2} \ or \ \frac{S_2^2}{S_1^2}$

Problem:

The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of specimens of coal from two mines:

Mine 1	8260	8130	8350	8070	8340	
Mine 2	7950	7890	7900	8140	7920	7840

Use the 0.05 level of significance to test whether it is reasonable to assume that the variances of the two populations are equal.

Solution:

```
> mine1<-c(8260,8130,8350,8070,8340)

> mine2<-c(7950,7890,7900,8140,7920,7840)

>var.test(mine1,mine2)

        F test to compare two variances

data:  mine1 and mine2

F = 1.4423, numdf = 4, denomdf = 5, p-value = 0.6872

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.195226    13.506448

sample estimates:

ratio of variances

1.442308

compute the critical value at .05 significance level

>alpha=0.05
```

```
>falpha=qf(1-alpha,df1=n1-1,df2=n2-1)

>falpha

[1] 5.192168
```

Conclusion:

The test of statistic 1.4423 is less than the critical value 5.192168. Hence, at .05 significance level, we can accept the claim.

Unit-V

Introduction to R software

Objectives:

- Classify the menus ‘help’ and ‘packages’.
- How to use R as a calculator and define different types of data structures.
- Define the commands for correlation coefficient and regression lines.

Syllabus:

An introductory R session-R as a calculator-Getting help and loading packages-Data entry and exporting data.
Correlation and Regression using R:
Calculating correlation coefficient-calculating rank correlation-finding regression lines-interpretations.

Learning Outcomes: The students will be able to

- Discuss different types of data structures with exporting data.
- Apply the commands to find correlation coefficient and regression lines.

Learning Material

R is a statistical computing language. It is based on S and it was initially developed by Ross Ihaka and Robert Gentleman at the University of Auckland in early 90’s. It is an open source software that means it is reviewed and improved continuously.

It is widely used among statisticians and data miners for developing statistical software and data analysis.

Using R:

There are several ways to work with R:

- With the R console GUI
- With the R studio IDE
- With the Tinn-R editor and R console
- From one of the other IDE such as JGR
- From a command line R interface (CLI)
- From the ESS (Emacs speaks statistics) module of the Emacs editor.

Of these, here mainly we will discuss about R console GUI.

Advantages of using R:

- It is completely free and is freely available over the internet.
- It runs on many operating systems like Windows, Linux and Macintosh etc.,
- It can work on objects of unlimited size and complexity with a consistent, logical expression language.
- Data can be imported in various forms such as CSV, Excel and also output can be exported in different forms, namely, CSV, Excel, Charts and Graphs.
- It is extensively used for manipulating data.
- It provides the best statistical analysis and is excellent for visualization purpose.

Disadvantages of using R:

The main drawbacks of R are: Lack of efficiency, Low speed and Poor memory management.

- *Lack of efficiency*: Some machine learning models doesn't yield the efficient results.
- *Low speed*: the computation speed for complex models in considerably low.
- *Poor memory management*: certain tasks use memories up to a greater extent. Hence, memory management scheme is required to be implemented in certain cases.

Downloading and Installing R:

Downloading R:

The program R is easily downloadable from CRAN i.e., Comprehensive R Archive Network and is the network of websites. The benefit of having this network of websites is improved download speeds.

Now, perform the following steps to download R.

- Open the main R webpage with the link www.r-project.org, then we will get a page with getting started box and with a link to download R. click on that link and then we directed to select a local CRAN mirror sites from which to download R.
- Next, click on any one of those CRAN mirrors and then directed to a CRAN page with some icons on left side. In those icons, click on R binaries under software.
- After selecting R binaries, we directed to a simple directory containing folders for a variety of operating systems.

- Now select the appropriate operating system (windows) on which you want to download R and depending on this operating system, the installation procedure may vary.

Installing R:

After selecting windows, we moved to a page “R for windows” with the link install R for the first time.

- Next, click on that link and then we directed to a page “R-3.x.x for windows” which supports both 32- and 64-bit versions and on this page again we have a link “Download R-3.x.x. for windows”. Here x’s indicates the current version of R and this changes periodically as improvements are made.
- Finally, R was installed by selecting “Download R-3.x.x. for windows” and run the installer with all the default settings.

Exploring the Environment:

The R screen contains main application window and within it there is a console window with a short introductory message (that appears in a little difference on each OS)

The main application window contains a menu bar with six menus and toolbar with eight icons. Let’s explore some of the features of R environment.

Menu bar:

The menu bar in R is similar to that in most of windows based programs. It contains six pull down menus, which are briefly described below.

File: The file menu contains options for opening, saving and printing R documents, as well as the option for exiting the program. The options that begin with “Load” are used to open the previously saved work.

Edit: The edit menu contains standard functionalities of cut, copy, paste and select all. In addition there is an option to ‘Clear Console’ which creates a blank workspace with only a command prompt. Moreover, the ‘Data editor’ option allows you to access the data editor, a spreadsheet like interface for manually editing the data.

*Misc:*The most notable feature of this menu is the first option i.e., “Stop Current Computation” which can also be accessed with the ESC key on your key board. If the coding of R gets stuck then by selecting this option (or Esc) we should get the situation under control and return the console to a new command line.

The other functionality provided by ‘Misc’ is listing and removing the objects.

*Packages:*Packages are collections of R functions, data and compiled code in a well-defined format.

For example, the survival package is used for survival analysis, ggplot2 is used for plotting and sp is used for spatial data etc.,

Some packages are installed with R and automatically loaded at the start of an R session. The standard packages contain the basic functions that allow R to work, the data sets, standard statistical and graphical functions. The directory where packages are stored on your computer is called the library and the command “library()” shows what packages we were saved in the library.

Other packages (nearly there are more than 5500) are available for download and installation. Once installed, they must be loaded into the session in order to use. The command “search()”tells you which packages are loaded and ready to use.

*Installing a package:*To install a package for the first time, use the command “install.packages()” in R console or select “Install packages” from the packages menu (to use this option, your computermust be connected to internet) and the selection opens a dialog box with list of packages. Now select the package of interest and click on OK. Then R will automatically download the package to your computer and put it in library. Moreover, use the option “update packages” to update any package that we have installed.

Some packages are not available directly from the CRAN site. So, download these packages from their source site to an appropriate folder on your computer. Now, install them to select the option “Install packages from local files” on the packages menu and R will again automatically put them in library.

If we want to see the details of our packages then use the command “installed.packages()” in R console. It lists the packages we have, along with their version, dependencies and other information.

Loading Packages: To load an installed package, select the “Load package” option from the packages menu. This produces a dialog box with a list of installed packages. Select the package of interest and click on OK then we are able to use the features of that package.

Windows: The windows menu provides option for cascading and tiling windows. If there is more than one opened window then we use those opened windows list on the bottom of this menu to access these different windows.

*Help:*By selecting 'Help' on RGui (main window), it directs to the following options which are available.

- Console: It gives useful shortcuts.

  Ex: ctrl+L to clear the R console screen, ctrl+X to first copy and then paste etc.,
- FAQ on R: Frequently asked questions concerning general R operation.
- FAQ on R for Windows: Frequently asked questions about R, tailored to the Microsoft Windows operating system.
- Manuals: It contains technical manuals about all features of the R system including installation, the complete language definition and add-on packages.
- R functions (text)... : If we know the exact name of the function then by using this option we can know more information about it.

  Ex: mean, plot etc.,

  Select the option "R functions (text)..." from help menu, it directs a help on window and type 'mean' in that window and again click on OK. Then it directs a new window with more information about mean.

Note: This verification only works if the function of interest is contained in a package which is already loaded into the search path.

- HTML help: This option is used to browse the manuals with point-and-click links. It also has a search engine and keywords option for searching the help page titles with point-and-click links for the search results and is the best possible method for beginners.
- Search help...: If we do not know the exact name of the function of interest or if the function is in a package that has not yet loaded then we can use this option for identifying it.

  Ex: If you enter any word of starting letters say 'plo' then a text window will return by listing all the help files with an alias, concept or title matching 'plo' using regular expression matching.
- Search.r-project.org...: This will search for words in help lists and email archives of the R project. It can be very useful for finding other questions that other users have asked.
- Apropos...: This is used for more sophisticated partial name matching of functions. For more details, give the command as "?apropos" at command line.

The Toolbar:

Toolbar is below the menu bar and it provides quick access icons to the main features of the menu bar. If we scroll over the icons with the mouse slowly then we will get rollover messages about the feature of each icon.

Openscript: It opens an existing file directly.

Load workspace: It loads the previous workspace.

Save workspace: It saves the current workspace.

Copy: It makes a copy of our current workspace.

Paste: It paste the selected line.

Copy and Paste: This option at a time copy and paste the selected line.

Stop current computation: If R will continue to compute most recent code until it is stopped then by clicking on this we can stop that computation.

Print: It prints the current workspace.

R Console:

R console is below the toolbar with short introductory message. In this, the prompt always starts with a symbol '>' on the command line. If the command line is not complete in one line then the continuation prompt is '+'

R as a Calculator:

Being a statistical programming language, certainly R can be used to do basic math.

```
Ex: >1+(2*3)  ;  >log(42/7.3)    ;  > log(6,10)     etc.,
        [1]  7    ;   [1]  1.749795  ;  [1]  0.7781513
```

Moreover, R can work with variables.

Variable: A variable is a name for a value. Letters and digits can be used while naming a variable. Also special characters such as dot and underscore are permissible in the name. The variable name must start with a letter or dot.

The variables can be assigned by using three operators. They are

1. The Leftward operator (<-)
2. The Rightward operator (->)
3. The Equal to or Assignment operator (=)

Of these, the most popular assignment operator is Leftward i.e.,(<-)

Mainly in R, the calculations are done by using basic operators. The basic operators are of three types. They are

1. Arithmetic operator
2. Comparison operator
3. Logical operator

*Arithmetic Operators:*Addition (+), Subtraction (-), Multiplication (*), Division (/) and exponentiation (^ or **)

Comparison Operators: Equal (==), not equal (!=), greater/less than (>/<), greater /less than or equal (>=/<=)

```
Ex: > x=5

> y=10

>x>5          ;x!=y

[1] FALSE   ; [1] TRUE
```

Logical Operators:

AND: The sign '&' gives 'true' if both the comparisons are true.

```
Ex:> x<-1:10

> y<-10:1

> x>y&x>5

 [1] FALSE FALSEFALSEFALSEFALSE  TRUETRUETRUETRUETRUE
```

OR: The sign '|' gives 'true' if at least one comparison is true.

```
Ex:> x<-1:10

> y<-10:1

>x==y|x!=y

 [1] TRUE TRUETRUETRUETRUETRUETRUETRUETRUETRUE
```

NOT: The sign '!' gives negation (opposite) of a logical vector.

```
Ex:> x<-1:10

> y<-10:1

>!x>y

 [1]  TRUETRUETRUETRUETRUE FALSE FALSEFALSEFALSEFALSE
```

Data Types in R:

The data types in R are Numeric, Integer, Logical, Character and Complex.

Numeric: The most commonly used numeric data is numeric. It handles integers and decimals (both +ve,-ve, 0). A numeric value stored in a variable is automatically assumed to be numeric.

Testing whether a variable is numeric or not by the command "is.numeric()" or "class()"

Ex:
```
> x=3          ;          > y=2.8
>class(x)                 > class(y)
[1] "numeric"                       [1] "numeric"
>is.numeric(x)                      >is.numeric(y)
[1] TRUE                            [1] TRUE
```

Integer: In R, integers are denoted by the symbol 'L' (L=1)

```
Ex: > x<-2L      ;             > z=2L*3.8
> x                       > z
[1] 2                     [1] 7.6
>class(x)                 > class(z)
[1] "integer"             [1] "numeric"
```

Logical: Logicals are the way of representing data that can be either 'TRUE (T)' or 'FALSE (F)'. Numerically 'TRUE' is same as '1' and 'FALSE' is same as '0'.

```
Ex: > TRUE*3 or >T*3;   >F*3              >class(T)
[1] 3             [1] 0              [1] "logical"
```

Character: Thecharacter (string) data type is very common in statistical analysis. It is specified by using quotes (both single and double quotes will work). The length of a character is given by the command "nchar()"

```
Ex: > x<-'data'   ;    >nchar(x)          ;     >nchar("studentofGEC")
>x                [1] 4                    [1] 12
[1] "data"        >nchar(125689)          >nchar("student of GEC")
                       [1] 6                    [1] 14
```

Complex:

```
Ex: > x<-3+2i      ;     > Re(x)     ;     > Mod(x)          ; >Arg(x)
>x                 [1] 3             [1] 3.605551     [1] 0.5880026
[1] 3+2i                 >Im(x)            >Conj(x)
                         [1] 2             [1] 3-2i
```

Data Structures:

R has many data structures. These include vectors, matrices, data frames, lists, factors and tables etc.,

Vectors: Vectors are the collection of same data type (numerical, character, complex or logical) components or elements.

The most common way to create or construct a vector is with ‘c’ function, which combines its elements.

```
Ex: > x<-c(12,25,36,65,14,58)        ; > x[c(1,3,5)]
>x                                 ; [1] 12 36 14
[1] 12 25 36 65 14 58
```

Vector operations: The standard arithmetic operators and functions can be applied directly to each vector (the calculation can be done automatically on an element-wise).

```
Ex: > x<-c(12,25,36,65,14,58)        ; > y<-c(5,10,15,20,25,30)
> x                                >nchar(y)
[1] 12 25 36 65 14 58    [1] 1 2 2 2 2 2
>sin(x)
[1] -0.5365729 -0.1323518 -0.9917789  0.8268287  0.9906074  0.9928726
>x/2
[1]  6.0 12.5 18.0 32.5  7.0 29.0
>x/y
[1] 2.400000 2.500000 2.400000 3.250000 0.560000 1.933333
>mean(x)
[1] 35
```

Matrices: Matrix is a two dimensional rectangular array in which each element has the same mode (numeric, character, logic). Matrices are created with the 'matrix()' function.

The basic syntax for creating a matrix in R is

Matrix(data, nrow, ncol, byrow, dimnames)

Where data - the input vector which becomes the data elements of the matrix.

nrow – number of rows to be created

ncol – number of columns to be created

byrow – logical clue. If it is set to be TRUE then only the elements are arranged in row-form (otherwise the elements are arranged in column form).

Dimnames – the names assigned to rows and columns.

```
Ex: > x<-c(3:14)

>x

[1]  3  4  5  6  7  8  9 10 11 12 13 14

> D<-matrix(x,nrow=3,byrow=TRUE); >D<-matrix(x,nrow=3,byrow=FALSE)

> D                                    > D

    [,1] [,2] [,3] [,4]                          [,1] [,2] [,3] [,4]

[1,]    3    4    5    6               [1,]    3    6    9   12

[2,]    7    8    9   10               [2,]    4    7   10   13

[3,]   11   12   13   14                     [3,]    5    8   11   14
```

Data frames: A data frame is a table or a two-dimensional array like structure in which different columns contain different modes of data (numeric, character etc.,)

A data frame can be created by using the function 'data.frame()'

Note: 1) In a data frame, each column contains values of one variable and each row contains one set of values from each column.

2) Each column should contain the same number of data items.

3) A data frame can be expanded by adding rows and columns.

```
Ex: > x<-c("A","C","G","H","T")

> y<-c(23,24,56,95,41)

> z<-c(2.5,3.6,1.8,4.5,9.1)
> d<-data.frame(x,y,z)

>d                                  ; >nrow(d)

x  y   z                            [1] 5

1 A 23 2.5                          >ncol(d)

2 C 24 3.6                            [1] 3

3 G 56 1.8

4 H 95 4.5

5 T 41 9.1
```

Lists: List is an ordered collection of objects. It can store any number of items in any type i.e., A list can store either numerics or characters or mix of two or data frames or matrices.

Lists can be created with the function ‘list()’ in which each argument to the function becomes an element.

```
Ex: >  x<-c(1,2); > y<-c("A","B"); > z<-c(3,"C"); > t<-data.frame(z,x);

> w<-matrix(x,nrow=2,byrow=TRUE); > l<-list(x,y,z,t,w); > l

[[1]]

    [1] 1 2

[[2]]

   [1] "A" "B"

[[3]]

   [1] "3" "C"

[[4]]

 z x

1 3 1

     2 C 2

[[5]]

    [,1]

[1,]    1

[2,]    2
```

Correlation and Regression using R:

Correlation coefficient using R:

- The formula for Karl Pearson's Coefficient of correlation or Product-Moment correlation coefficient is given by $r_{x,y}$

$$r= \frac{Cov(X,Y)}{\sigma_x\sigma_y} \quad \text{OR} \quad r=\frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{OR} \quad r= \frac{\sum xy}{N\sigma_{x\sigma_y}} \quad \text{OR}$$

$$r= \frac{(\sum XY*n)-(\sum X*\sum Y)}{\sqrt{(\sum X^2*n-(\sum X)^2)*(\sum Y^2*n-(\sum Y)^2)}}$$

- The formula for Spearman's rank correlation coefficient is given by

$$\rho = 1-\frac{6\sum d^2}{n(n^2-1)} \text{ (For untied ranks)}$$

$$\text{And} \quad \rho = 1\text{-}6\left\{\frac{\sum d^2+coreectionfactorofXandY}{n(n2-1)}\right\}\text{(For tied ranks)}$$

Problem:

```
> x<-c(5,7,10,12,15,22)

> y<-c(3,8,9,13,19,25)

>cor(x,y)

[1] 0.9836706

>cor(x,y,method="pearson")

[1] 0.9836706

>cor(x,y,method="spearman")

[1] 1
```

Simple Linear Regression using R:

The general mathematical equation for a linear regression is −

y = ax + b where

- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

Note:

- The basic syntax for linear relationship model is "lm(formula)"where 'formula' is a symbol presenting the relation between x and y.
- Find the new value of a variable by using the command "predict(object,newdata)**"** where 'object' is the formula which is

already created by using lm() function and 'newdata' is the vector containing the new value for predictor variable.

Problem:

```
> x<-c(5,7,10,12,15,22)

> y<-c(3,8,9,13,19,25)

>relation<-lm(y~x)          ;     > relation1<-lm(x~y)

>relation                   > relation1

Call:                             Call:

lm(formula = y ~ x)               lm(formula = x ~ y)

Coefficients:                     Coefficients:

(Intercept)        x              (Intercept)        y

    -2.420     1.289                  2.2000     0.7506
```

UNIT – IV
TWO SAMPLE SIGNIFICANCE TESTS

Objectives:

- Define principles of hypothesis testing in case of Large and small samples.
- Find the difference between two sample means using a two-sample t-test.
- Explain how to use an *F*-test to judge whether two population variances are equal.

Syllabus:

Two sample tests: Large sample-test for two means, two proportions, small sample: t-test for two means, F-test

Learning Outcomes: The students will be able to

- Apply a range of statistical tests appropriately.
- Choose independent-samples t-test.
- Apply *F*-test to judge whether two population variances are equal.
- **Test of equality of Two means**: Let $\overline{x}_1, \overline{x}_2$ be the sample means of two independent random samples sizes n_1 and n_2 drawn from two populations having the means μ_1 and μ_2 and standard deviation σ_1 andσ_2. To test whether the two population means are equal .
- Step (1): Null hypothesis: H_0 : μ_1=μ_2
- Step (2): Alternative hypothesis:
- $H_1 : \mu_1 \neq \mu_2 \,/\, H_1 : \mu_1 \leq_2 \,/\, H_1 : \mu_1 \geq \mu_2$
- Step (3): Level of significance: Choose 5% (or) 1%
- Step (4): Test statistic: Z= $\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma^2}_2}{n2}}}$
- Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.
- If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.
-
- **Problem:**
- The mean of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from same population of s.d 2.5 inches?
- Given n_1=1000,n_2=2000 and $\overline{x_1}$ =67.5 $\overline{x_2}$ =68 population S.D σ=2.5
- Step (1): Null hypothesis:$H_{0:}\mu_1$=μ_2
- Step (2): Alternative hypothesis:

- $H_1 \colon \mu_1 \neq \mu_2$
- Step (3): Level of significance: Choose 5% (or) 1%
- Step (4): Test statistic: Z= $\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{\sigma^2{}_2}{n2}}}$ =5.16
- Step (5): Conclusion: Z_{cal}=5.16 Z_{tab} =1.96
- If Z_{cal}> Z_{tab} then we reject our H_0.
- ∴ The samples have not been drawn from same population of S.D 2.5 inches.
- **Test of equality of two proportions:**
- Let p_1 and p_2 be the sample proportions in two large random samples of sizes n_1 and n_2 drawn from two populations having proportions p_1 and p_2.
- To test whether the two samples have been drawn from the same population
- Step (1): Null hypothesis: $H_{0:}$P$_1$= P_2
- Step (2): Alternative hypothesis:
- $H_1 : P_1 \neq P_2 / H_1 : P_1 \leq P_2 / H_1 : P_1 \geq P_2$
- Step (3): Level of significance: Choose 5% (or) 1%
- Step (4): Test statistic: **(a)** when the population proportion P_1 and P_2 are known.
- The test statistic is Z=$\frac{p_{1-} \, p_2}{S.E(p_{1-} \, p_2)}$ $\sim N(0{,}1)$
- Where S.E (p_1-p_2) =$\sqrt{\frac{P_1 Q_1}{n_1}}$ + $\sqrt{\frac{P_2 Q_2}{n_2}}$ Q_1=1-P_1 Q_2=1-P_2
- **(b)** When the population proportion P_1 and P_2 are unknown.
- In this case we have two methods to estimate P_1 and P_2 .
- **(i)Method of substitution:**
- In this method sample proportion p_1 and p_2 are substituted for P_1 and P_2.
- ∴ S.E (p_1 - p_2)= $\sqrt{\frac{p_1 q_1 + p_2 q_2}{n_1 + n_2}}$
- ∴ Test statistic is Z= $\frac{p_1 - p_2}{S.E\ (p_1 - p_2)}$
- **(ii) Method of pooling:**
- In this method, the estimate value for the two population proportions is obtained by pooling the two sample proportions p_1 and p_2 into a single proportion p by the formula is given below.
- Sample proportion of two samples or estimated values is given by
- P= $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$
- ∴ Test statistic is Z= $\frac{p_1 - p_2}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$

- Step (5): Conclusion: Z_{cal} is compare with Z_{tab} value.
- If Z_{cal}<Z_{tab} accept H_0. Otherwise reject H_0.
- **Problem:**
- In two large populations, there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?
- Given n_1=1200 n_2 =900
- p_1= proportion of fair haired people in first population=30/100=0.3
- p_2= proportion of fair haired people in first population=25/100=0.25
- Step (1): Null hypothesis:H_0: The two sample proportions are equal p_1=p_2
- Step (2): Alternative hypothesis:
- $H_1: p_1 \neq p_2$
- Step (3): Level of significance: 5% = 1.96
- Step (4): Test statistic: Z= $\frac{p_{1-}\, p_2}{S.E(p_{1-}\, p_{2)}}$= 2.56
- Step (5): Conclusion: Z_{cal} = 2.56, Z_{tab} = 1.96
- If Z_{cal}> Z_{tab} at 5% LOS then we reject H_0.
- **Problem:**
- (1) Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women in favour of the proposal. Test the hypothesis that proportion of men and women in favour of the proposal at 5% LOS?
- Given n_1=400 n_2 =600
- p_1= population of men =200/400=0.5
- p_2= population of women=250/600=0.541
- Step (1): Null hypothesis:H_0: There is no significance difference between the option of men and women $H_0: p_1$=p_2=p
- Step (2): Alternative hypothesis:

$$H_1: p_1 \neq p_2$$

- Step (3): Level of significance: 5%
- Step (4): Test statistic: Z= P= $\frac{n_1p_1 + n_2p_2}{n_1+n_2}$

∴ Test statistic is Z= $\frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ =1.28

- Step (5): Conclusion: Z_{cal} = 1.28, Z_{tab} = 1.96
- If Z_{cal}< Z_{tab} at 5% LOS then we reject H_0.

t- Test for difference of means:

Assumptions: (i) parent populations, from which the samples have been drawn are normally distributed

(i) The population variances are equal and unknown

(ii) The two samples are random and independent of each other

Suppose we want to test if two independent samples xi (i=1,2..n1) and yj(j=1,2..n2) of sizes n1 and n2 have been drawn from two normal populations with μ_x and μ_y respectively.

Under the null hypothesis H$_0$, $\mu_x = \mu_y$, H$_1$: $\mu_x \neq \mu_y$,

then test statistic

$$t = \frac{\bar{x} - \bar{y}}{S / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{\sum_i x_i}{n}$, $\bar{y} = \frac{\sum_i y_i}{n}$ and $S^2 = \frac{1}{(n_1+n_2-2)}\left[\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2\right]$, a combined sample mean square.

Confidence limits: For example, 95% confidence limits for the population mean μ are $\bar{x} \pm t_\alpha \frac{S}{\sqrt{n}}$ or $\bar{x} \pm t_\alpha \frac{s}{\sqrt{n-1}}$ where $\alpha = 0.025$ for two-tailed test and $\alpha = 5\%$ for one-tailed test, as mentioned in the above example. i.e. For two-tailed test at α los, the value of is taken for $\alpha / 2$ from statistical tables of t.

Problem: A group of 5 patients with medicine A weigh 42, 39, 48, 60, and 41 kilograms. Another group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kilograms. Do you agree with the claim that medicine B increase the weigh significantly?

Solution: Null hypothesis, H$_0$: There is no significant difference between the medicines A and B with reference to their effect on increase in weight. i.e. $\mu_x = \mu_y$.

Alternative hypothesis, H$_1$: $\mu_x < \mu_y$ (left-tailed test).

Under the null hypothesis H$_0$, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{S / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

where $\bar{x} = \frac{\sum_i x_i}{n}$ and $S^2 = \frac{1}{(n_1 + n_2 - 2)}\left[\sum_i (x_i - \bar{x})^2 + \sum_i (y_i - \bar{y})^2\right]$, a combined sample mean square.

Calculation of $\bar{x}$ and S^2 :

We are given $n_1 = 5, \quad n_2 = 7$.

From the given data,

$$\bar{x} = \frac{42+39+48+60+41}{5} = \frac{230}{5} = 46,$$

$$\bar{y} = \frac{38+42+56+64+68+69+62}{7} = \frac{399}{7} = 57$$

and

$$\sum_{i=1}^{5} (x_i - 46)^2 = (42-46)^2 + (39-46)^2 + \ldots + (41-46)^2 = 290$$

$$\sum_{i=1}^{7} (y_i - 57)^2 = (38-57)^2 + (42-57)^2 + \ldots + (62-57)^2 = 926$$

Hence,

$$S^2 = \frac{290+926}{5+7-2} = 121.6$$

$\Rightarrow S = 11.03.$

Therefore, the value of test statistic is

$$t = \frac{46-57}{11.03\sqrt{\frac{1}{5}+\frac{1}{7}}} = -\frac{11}{6.46} = -1.7$$

$\therefore t = -1.7$

t- Critical value at 5% los with 10 degrees of freedom for right tailed test is -1.812.

Clearly, t- calculated value is greater than t- critical value at 5% los, we accept the null hypothesis. i.e. the medicines A and B do not differ significantly with reference to their effect on increase in weight.

(This is an application of Test for difference of means.)

Problem: Memory capacity of 10 students was tested before and after training. State whether the training was effective or not from the following scores:

Before training:	12	14	11	8	7	10	3	0	5	6
After training:	15	16	10	7	5	12	10	2	3	8

Solution:

Null hypothesis, H$_0$: There is no significant effect of the training on memory capacity of the students. i.e. $\mu_1 = \mu_2$.

Alternative hypothesis, H_1: Memory capacity of the students has been increased or improved after training. i.e. $\mu_1 < \mu_2$ (left-tailed test).
Under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

Where $\bar{d} = \frac{\sum_i d_i}{n}$, $d_i = x_i - y_i$ and $S^2 = \frac{1}{(n-1)}\sum_i (d_i - \bar{d})^2$.

Calculation of $\bar{d}$ and S^2:
Let memory capacity before training and after training be x and y respectively.

$$\bar{d} = \frac{(12-15)+(14-16)+(11-10)+(8-7)+(7-5)+(10-12)+(3-10)+(0-2)+(5-3)+(6-8)}{10}$$

$$= -\frac{12}{10} = -1.2$$

and

$$S^2 = \frac{(-3-(-1.2))^2 + (-2-(-1.2))^2 + (1-(-1.2))^2 + ... + (-2-(-1.2))^2}{9} = 7.73$$

$$\Rightarrow S = 2.78$$

Therefore, the value of test statistic is

$$t = \frac{-1.2}{2.78/\sqrt{10}} = -1.365.$$

t- Critical value at 5% los with 9 degrees of freedom for left-tailed test is -1.833.

Since t- calculated value is greater than t- critical value at 5% los with 9 degrees of freedom for left tailed test, we accept the null hypothesis, H_0 i.e. we conclude that there is no change in memory capacity after the training programme (or) there is no use of training programme.

F-distribution :- “The ratio of two sample variances is distributed of F.” F-distribution was worked out by G.W. Snedecor and as a mark of respect for Sir R.A.Fisher (Father of modern statistics). Who was defined a statistics Z which is based upon the ratio of two –sample variances initially and hence it is denoted by F. (The first letter of Fisher).

Let s_1^2 be the sample variance of an independent sample of size n_1 drawn from a normal population N $(\mu_1, \sigma_1{}^2)$. Similarly , let S_2^2 be the sample variance in an independent sample of size n2 drawn from another normal population N $(\mu_2, \sigma_2{}^2)$.Thus s_1^2 and S_2^2 are the variances of two random samples of sizes n_1 and n_2 respectively drawn from two normal populations. In order to determine whether the two samples came from two populations having equal variances of the two independent random samples defined by

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$

Which is an F-distribution with v1=n1-1 and v2=n2-1 degrees of freedom.

Properties of F-distribution:

(1) F-distribution curve extends on abscissa from 0 to ∞.

(2) It is an unimodel curve and its mode lies on the point

F= $\frac{k_2(k_1-2)}{k_1(k_2+2)}$ or $\frac{v_2(v_1-2)}{v_1(v_2+2)}$ which is always less than unity

(3) F-distribution curve is a positive skew curve. Generally, the F-distribution curve is highly positive skewed where v2 is small

(4) The mean and variance are defined when v2 ≥ 3 and v2≥5 respectively.

(5) There exists a very useful relation for interchange of degrees of freedom v1 and v2 i.e

$$F_{1-\alpha}(v_1,v_2)=\frac{1}{F_\alpha(v_2,v_1)}$$

(6) The moment generating function of F-distribution does not exist.

F-test is used to

(1) Test the hypothesis about the equality of two population variances**.**

(2) Test the hypothesis about the equality of two or more population means.

F-test for equality of two population variances: Suppose we want to test whether two independent samples xi (i=1,2..n1) and yj(j=1,2..n2) of sizes n1 and n2 have been drawn from two normal populations with the same variance or not then

Null hypotheses, $H_0:\sigma_x^2=\sigma_y^2$ $,H_1:\sigma_1^2\neq\sigma_2^2$.

Under the null hypothesis, H$_0$, the test statistics is:

$$F=\frac{s_x^2}{s_y^2}\sim F_{(v_1,v_2)} \quad \text{(OR)} \quad F=\frac{s_y^2}{s_x^2}\sim F_{(v_2,v_1)}$$

When $s_x^2>s_y^2$ OR $s_y^2>s_x^2$ respectively

Where $s_x^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(x_i-\bar{x}\right)^2$ with $\bar{x}=\frac{\sum_{i-1}^{n_1}x_i}{n_1}$

And $s_y^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(y_i-\bar{y}\right)^2$ with $\bar{y}=\frac{\sum_{i-1}^{n}y_i}{n_2}$

Besides t-test , we can also apply a F-test for testing equality of two population means.

F-distribution is a very popular and useful distribution because of its utility in testing of hypothesis about the equality of several population means, two population variances and several regression coefficients in multiple regression coefficient etc., As a matter of fact , F-test is the backbone of analysis of variance(ANOVA)

Note: (1) F determines whether the ratio of two sample variances s1 and s2 is too small or too large.

(2) When F is close to 1, the two sample variances s1 and s2 are likely same

(3) F-distribution also known as variance ratio distribution

(4) Dividing S_1^2 and S_2^2 by their corresponding population variances standardizes the sample variance, and hence on the average both numerator and denominator approach. Therefore, its customer, to take the large sample variance as the numerator.

(5) F-distribution depends not only on the two parameters, $V_1$and V_2 but also on the order in which they are slated.

Problem: Life expectancy in 9 regions of Brazil in 1900 and in 11 regions of Brazil in 1970 was as given in the table below:

(Source: The review of income and wealth, June 1983)

Regions Life Expectancy	1	2	3	4	5	6	7	8	9	10	11
1900	42.7	43.7	34.0	39.2	46.1	48.7	49.4	45.9	55.3	-	-
1970	54.2	50.4	44.2	49.7	55.4	57.0	58.2	56.6	61.9	57.5	53.4

It is desired to confirm, whether the variation in life expectancy in various reigns in 1900 and in 1970 in same or not.

Solution: Let the populations in 1900 and in 1970 be considered as $N\left(\mu_1, \sigma_1^2\right) \, and \, N\left(\mu_1, \sigma_1^2\right)$ respectively.

Null hypotheses, H_0 : The variation of life expectancy in various regions in 1900 and in 1970 is same. $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Under the null hypothesis, H_0, the test statistics is :

$$F = \frac{s_1^2}{s_2^2} \sim F_{(v_1,v_2)} \quad \text{(OR)} \quad F = \frac{s_2^2}{s_1^2} \sim F_{(v_2,v_1)}$$

When $s_1^2 > s_2^2$ OR $s_2^2 > s_1^2$ respectively

$$\text{Where } s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(x_i - \overline{x}\right)^2 \text{ with } \overline{x} = \frac{\sum_{i-1}^{n_1} x_i}{n_1}$$

$$\text{And } s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(y_i - \overline{y}\right)^2 \text{ with } \overline{y} = \frac{\sum_{i-1}^{n} y_i}{n_2}$$

Calculation: $\bar{x}=\frac{405}{9}=45,\ \bar{y}=\frac{598.5}{11}=5441$(approximate)

$$\sum_{i=1}^{9}(x_i-45)^2 = 5.29+1.69+121+33.64+1.21+13.69+0.9+106.9$$

$$= 288.51+19.36=302.87$$

$$\Rightarrow s_1^2=\frac{302.87}{8}=37.85$$

$$\sum_{i=1}^{11}(y_i-54.41)^2 = 0.04+16+104.04+22.09+1+6.76+14.44+4.84+56.25+9.61+1 =236.07$$

$$\Rightarrow s_2^2=\frac{236.07}{10}=23.607$$

Since $s_1^2 > s_2^2$, the value of test statistics is:

$$F=\frac{37.85}{23.607}=1.603$$

The table value of F at 5% los with (8, 10) degrees of freedom for two tailed test is 3.85 (From F-tables).

Since F-Calculated value is less than f-tabulated value, we accept H$_0$. i.e. The sample data confirms the equality of variances in 1900 and 1970 in various regions of brazil or $\sigma_1^2=\sigma_2^2$

Problem: The house-hold net expenditure on health care in south and north India, in two samples of households, expressed as percentage of total income is shown the following table:

South:	15.0	8.0	3.8	6.4	27.4	19.0	35.3	13.6	
North:	18.8	23.1	10.3	8.0	18.0	10.2	15.2	190.0	20.2

Test the equality of variances of household's net expenditure on health care in south and north India.

Problem: The time taken by workers in performing a job by method I and Method II is given below.

Method I	20	16	26	27	23	22	-
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution of population from which these samples are drawn do not differ significantly?

Solution:

Null hypothesis, H$_0$: There is no significant difference between the variances of time distribution of populations. i.e. $\sigma_1^2=\sigma_2^2$.

Alternative hypothesis, H$_1$: $\sigma_1^2\neq\sigma_2^2$ (Two-tailed test)

Level of significance : Choose $\alpha=5\%=0.05$

Under the null hypothesis, H_0, the test statistics is :

$$F = \frac{s_1^2}{s_2^2} \sim F_{(v_1,v_2)} \quad \text{(OR)} \quad F = \frac{s_2^2}{s_1^2} \sim F_{(v_2,v_1)}$$

When $s_1^2 > s_2^2$ OR $s_2^2 > s_1^2$ respectively

Calculation: we are given $n_1 = 6$, $n_2 = 7$

$$\bar{x} = \frac{134}{6} = 22.3, \ \bar{y} = \frac{241}{7} = 34.4$$

$$\sum_{i=1}^{6}(x_i - 22.3)^2 = 81.34, \ \sum_{i=1}^{7}(y_i - 34.4)^2 = 133.72$$

$$\therefore s_1^2 = \frac{81.34}{5} = 16.26 \ and \ s_2^2 = \frac{133.72}{6} = 22.29$$

The value of test statistics is

$$F = \frac{22.29}{16.26} = 1.3699 \cong 1.37$$

F-critical value at 5% los with (5, 6) degrees of freedom for two tailed test is 4.39 (From F-tables)

Since F-Calculated value is less than F-tabulated value t 5% los, we accept H_0. i.e. there is no significant deference between the variances of the time distribution by the workers.

Problem: The nicotine contents in milligrams in two samples of tobacco were found to be as follows:

Sample A	24	27	26	21	25	-
Sample B	27	30	28	31	22	36

Can it be said that the two samples have come from the same normal population?

Hint: When testing the significance of the difference of the means of two samples, we assumed that the two samples came from the same population or from populations with same variances. If the variances of the population are not equal, a significant difference in the means may arise. Hence, to test the two samples have come from the same population or not, we need to apply both t-test and F-test. But here we note that first apply F-test, as usual manner.