TRIBHUVAN UNIVERSITY

Institution of Science and Technology

Bachelor Level/Second Year/Third Semester/Science — Full Marks: 60
Computer Science and Information Technology [STA 210] — Pass Marks: 24

TU QUESTIONS-ANSWERS 2075

Bachelor Level/Second Year/Third Semester/Science — Full Marks: 60
Computer Science and Information Technology (STA 210) — Pass Marks: 24
(Statistics II) — Time: 3 hrs.

All notations have the usual meanings

Group 'A'

Attempt any Two questions **(2 × 10 = 20)**

1. What is multiple Linear Regression (MLR)? From following information of variables X_1, X_2 and Y.

$\Sigma X_1 = 272$, $\Sigma X_2 = 441$, $\Sigma Y = 147$, $\Sigma X_1^2 = 7428$, $\Sigma X_2^2 = 19461$, $\Sigma Y^2 = 2173$, $\Sigma X_1 Y = 4013$, $\Sigma X_1 X_2 = 12005$, $\Sigma X_2 Y = 6485$, $n = 10$. Fit regression equation Y on X_1 and X_2 Interpret the regression coefficient.

Ans: It is a linear function of one dependent variable with two or more independent variables. With the help of two or more independent variables the value of dependent variable is predicted. For example, if we wish to test the hypothesis that whether or not the 'pass grade' of students depends on many causes such as previous test mark, study hours, IQ, ...then we can test a regression of cause (pass grade) with effect variables. This test will give us which causes are really significant in generating effect variable and among the significant cause variables their relative value responsible to generate the effect variable. If we assume more than one causes (called X or independent variable) responsible for one effect (also called Y or dependent variable), it is known as multiple regression. If we assume that the relation between Y and X's is linear it is called multiple linear regression.

Let us consider three variables Y, X_1 and X_2 in which Y is dependent variable, X_1 and X_2 are independent variables, then the mathematical form of the linear relationship of Y with X_1 and X_2 is expressed as

$Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$

Where,

Y = Dependent variable

X_1 and X_2 = Independent variable or explanatory variable or regressors

b_0 = Intercept and is called average value of Y when X_1 and X_2 are zero.

b_1 = Regression coefficient of Y on X_1 keeping X_2 constant. It measures the amount of change in Y per unit change in X_1 holding the X_2 constant.

b_2 = Regression coefficient of Y on X_2 keeping X_1 constant. It measures the amount of change in Y per unit change in X_2 holding the X_1 constant.

ε = Random error.

Random error (ε) is not created from mistake. It is a technical term that denoted the excess of value from real by model estimation. Error is also called Residual.

To fit regression equation $y = b_0 + b_1x_1 + b_2x_2$

$\sum y = nb_0 + b_1\sum x_1 + b_2\sum x_2$

or, $147 = 10b_0 + 272b_1 + 441b_2$ .....(i)

$\sum yx_1 = b_0\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2$

Or, $4013 = 272b_0 + 7428b_1 + 12005b_2$ .....(ii)

$\sum yx_2 = b_0\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2$

Or, $6485 = 441b_0 + 12005b_1 + 19461b_2$ .....(iii)

To find b_0, b_1 and b_2 using Cramer's rule

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constant
10	272	441	147
272	7428	12005	4013
441	12005	19461	6485

Now, $D = \begin{vmatrix} 10 & 272 & 441 \\ 272 & 7428 & 12005 \\ 441 & 12005 & 19461 \end{vmatrix}$

$= 10(144556308 - 144120025) - 272(5293392 - 5294205) + 441(3265360-3275748)$

$= 2858$

$D_1 = \begin{vmatrix} 147 & 272 & 441 \\ 4013 & 7428 & 12005 \\ 6485 & 12005 & 19461 \end{vmatrix}$

$= 147(144556308-144120025) - 272(78096993-77852425) + 441(48176065-48170580)$

$= 29990$

$D_2 = \begin{vmatrix} 10 & 147 & 441 \\ 272 & 4013 & 12005 \\ 441 & 6485 & 19461 \end{vmatrix}$

$= 10(78096993-77852425) - 147(5293392-5294205) + 441(1763920-1769733)$

$= 1658$

$D_3 = \begin{vmatrix} 10 & 272 & 147 \\ 272 & 7428 & 4013 \\ 441 & 12005 & 6485 \end{vmatrix}$

$= 10(48170580-48176065) - 272(1763920-1769733) + 147(3265360-3275748)$

$= -750$

Now $b_0 = \frac{D_1}{D} = 29990/2858 = 10.493$

$b_1 = \frac{D_2}{D} = 1658/2858 = 0.58$

$b_2 = \frac{D_3}{D} = -750/2858 = -0.262$

Hence regression equation is $y = b_0 + b_1x_1 + b_2x_2 = 10.493 + 0.58x_1 - 0.262x_2$

Here $b_1 = 0.58$ it means y changes by 0.58 per unit change x_1 keeping x_2 constant.

$b_2 = -0.262$ it means y changes(decreases) by 0.262 per unit change in x_2 keeping x_1 constant.

2. What do you mean by Latin Square Design? Write down its merit and demerit. Set up the analysis of variance for the following of design.

A (10)	B (15)	C(20)
B (25)	C (10)	A (15)
C (25)	A (20)	B (15)

Ans: When the experimental material is not homogeneous the LSD is better than RBD. In RBD local control is used according to one way grouping i.e. according to blocks but in LSD local control is used according to two way grouping i.e. rows and columns. Hence it is used when two sources of errors are to be controlled simultaneously. In this design number of treatments are equal to the number of replication and the treatments are allocated in such a way that each of the treatment occurs once and only once in each row and column. In this design Latin alphabet are used to denote the treatments, and shape is square due to equal number of treatments and replication so called Latin square design. It is based upon the all principles of design namely replication, randomization and local control.

Let us consider m treatments with m replication each so that there are $N = m^2$experimental unit.

Let us divide the experimental material into m^2 experimental units arranged in square so that each row as well as column contains m units. In this design none of treatments are replicated along row wise or column wise. In this case we study the variation between treatments, the variation between rows and variation between columns. It has only m^2 experimental unit but studies variation of three factors i.e. rows, columns and treatments. Hence it is the case of incomplete three way ANOVA. For complete three way ANOVA we need m^3 experimental unit.

Let us consider t = 4(A, B, C, D) then 4 × 4 LSD is as shown below.

A	D	B	C
B	C	D	A
C	B	A	D
D	A	C	B

Merits of LSD

i) Due to the use of two way grouping of controls more variation than CRD and RBD.

ii) It is incomplete three way layout. It's advantage over complete three way layout is that instead of m^3 experimental units only m^2 units are needed.

iii) The statistical analysis remains simple if some observations are missing.

Demerits of LSD

i) The assumption of factors are independent is not always true.

ii) It is suitable for treatments 5 to 10.

iii) It is not easy in the field layout.

Now,

Problem to test

H_{0R} : Rows are insignificant

H_{1R} : Rows are significant

H_{0C} : Columns are insignificant

H_{1C} : Columns are significant

H_{0T} : Treatments are insignificant

H_{1T} : Treatments are significant

				$T_{i..}$	$T_{i..}^2$
	A 10	B 15	C 20	45	2025
	B 25	C 10	A 15	50	2500
	C 25	A 20	B 15	60	3600
$T_{.j.}$	60	45	50	G =155	$\sum T_{i..}^2$ =8125
$T_{.j.}^2$	3600	2025	2500	$\sum T_{.j.}^2$ = 8125	

$T_{..A} = 10+20+15 = 45$

$T_{..B} = 25+15+15 = 55$

$T_{..C} = 25+10+20 = 55$

$\sum T_{..k}^2 = 45^2 + 55^2 + 55^2 = 8075$

k = A , B , C

m = 3

$N = m^2 = 3^2 = 9$

$CF = \frac{G^2}{N} = \frac{155^2}{9} = 2669.444$

$\sum y_{ijk}^2 = 10^2 + 15^2 + 20^2 + 25^2 + 10^2 + 15^2 + 25^2 + 20^2 + 15^2 = 2925$

$TSS = \sum_{(i,j,k)} y_{ijk}^2 - C.F. = 2925 - 2669.444 = 255.556$

$SSR = \frac{\Sigma_i T_{i..}^2}{m} - C.F. = \frac{1}{3} \times 8125 - 2669.444 = 38.889$

$SSC = \frac{\Sigma_j T_{.j.}^2}{m} - C.F. = \frac{1}{3} \times 8125 - 2669.444 = 38.889$

$SST = \Sigma \frac{T_{..k}^2}{m} - C.F. = \frac{1}{3} \times 8075 - 2669.444 = 22.222$

SSE = TSS -SSR -SSC - SST

= 255.556 - 38.889 - 38.889- 22.222 = 155.556

ANOVA table

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Row	2	38.889	19.444	0.249	$F_{0.05(2,2)} = 19$
Column	2	38.889	19.444	0.249	$F_{0.05(2,2)} = 19$
Treatment	2	22.222	11.111	0.142	$F_{0.05(2,2)} = 19$
Error	2	155.556	77.778		
Total	8	255.556			

Decision

$F_R = 0.249 <> F_{0.05(2,2)} = 19$, Accept H_{0R} at 5% level of significance.

$F_C = 0.249 <> F_{0.05(2,2)} = 19$, Accept H_{0C} at 5% level of significance.

$F_T = 0.142 <> F_{0.05(2,2)} = 19$, Accept H_{0T} at 5% level of significance.

Conclusion

Rows are insignificant.

Columns are insignificant.

Treatments are insignificant.

3. What do you mean by hypothesis? Describe null and alternative hypothesis. A company claims that its light bulbs are superior to those of the competitor on the basis of study which showed that a sample of 40 of its bulbs had an average life time 628 hours of continuous use with a standard deviation of 27 hours. While sample of 30 bulbs made by the

competitor had and average life time 619 hours of continuous use with a standard deviation of 25 hours. Test at 5% level of significance, whether this claim is justified.

Ans: A hypothesis is a tentative theory or supposition provisionally adopted to explain certain facts and to guide in the investigation of others.

A statistical hypothesis which is tentative statement or supposition about the estimated value of one or more parameter of the population is called parametric hypothesis. A statistical hypothesis about attributes is called non-parametric hypothesis.

If a hypothesis completely determines the population, it is called a simple hypothesis; otherwise composite hypothesis.

In testing of hypothesis a statistic is computed from a sample drawn from the parent population and on the basis of the statistic it is observed whether the sample so drawn has come from the population with certain specified characteristic.

Null hypothesis

The supposition about the population parameter is called null hypothesis. It is set up for testing a statistical hypothesis only to decide whether to accept or reject the null hypothesis. According to R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that is true.

It is the hypothesis of no difference between sample statistic and parameter. It is hypothesis of no difference between parameters.

Null hypothesis is denoted by H_0. It is set up as $H_0: \mu = \mu_0$

Suppose we want to test the average score of students in B.Sc. entrance exam is 55 then to start testing the hypothesis we assume the average score is 55. There is no difference between sample average and population average. Then the null hypothesis is $H_0: \mu = 55$

Alternative Hypothesis

A hypothesis which is complementary to the null hypothesis is called an alternative hypothesis.

Any hypothesis which is not null is also called alternative hypothesis. It is hypothesis of difference between sample statistic and parameter. It is hypothesis of difference between parameters.

Alternative hypothesis is denoted by H_1.

Alternative hypothesis are

(i) two tailed

(ii) one tailed right

(iii) one tailed left

Alternative hypothesis is set up as $H_1: \mu \neq \mu_0$ for two tailed or $H_1: \mu$ ▯ μ_0 for one tailed right or $H_1: \mu$ ▯ μ_0 for one tail left.

Let μ_1 *and* μ_2 be mean life of bulbs of company and it's competitor respectively

Sample number of bulb of company(n_1) = 40

Sample mean life of bulb of company ($\bar{X}_1$) = 628,

Sample Sd of life of bulb of company(s_1) = 27

Sample number of bulb of competitor (n_2) = 30,

Sample mean life of bulb of competitor ($\bar{X}_2$) = 619

Sample Sd of life of bulb of competitor (s_2) =25

Let μ_1 = Population mean wage of workers from Pokhara and μ_2 = Population mean wage of workers from Kathmandu.

Problem to test

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Test statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{628-619}{\sqrt{\frac{27^2}{40} + \frac{25^2}{30}}} = \frac{9}{\sqrt{39.058}} = 1.44$$

$Z = 1.44$

Critical value

At $\alpha = 0.05$ be the level of significance then the critical value for one tailed test is

$Z_{tabulated} = Z\,\alpha = 1.645$

Decision

$Z = 1.44 < Z_{tabulated} = 1.645$, accept H_0 at 0.05 level of significance

Conclusion

The claims of company that its light bulbs are superior to those of the competitor is not correct.

Group 'B'

Attempt any Eight questions: **(8 × 5 = 40)**

4. Suppose we are given following information with n = 7, multiple regression mode is $\hat{Y} = 8.15 + 0.56X_1 + 0.54X_2$
Here, Total sum of square = 1493,
Sum of square due to error = 91
Find i) R^2 and interpret it. ii) Test the overall significance of model.

Ans: TSS = 1493

SSE = 91

SSR = TSS – SSE = 1493 – 91 = 1402

MSR = SSR/k = 1402/2 = 701

MSE = SSE/n-k-1 = 91/7-2-1 = 22.75

R^2 = SSR/TSS = 1402/1493 = 0.939 = 93.9%

It means 93.9% variation in y is explained by x_1 *and* x_2

To test overall significance of regression model

Let β_1 *and* β_2 be population regression coefficient of Y on X_1 keeping X_2 constant and population regression coefficient of Y on X_2 keeping X_1 constant

Problem to test

$H_0 : \beta_1 = \beta_2 = 0$

H_1 : At least one β_i is different from zero, i = 1, 2

Test statistic

$$F = \frac{MSR}{MSE} = 701/22.75 = 30.81$$

Critical value

At $\alpha = 0.05$ level of significance, critical value is $F_{\alpha(k,n-k-1)} = 6.944$

Decision

$F = 30.81 > F_{tabulated} = 6.944$, reject H_0 at 5% level of significance.

Conclusion

There is linear relationship of dependent variable y with at least one of the independent variable x's

5. The following data related to the number of children classified according to the type of feed and the nature of teeth.

Type of feed	Nature of Teeth	
	Normal	Defective
Breast	18	12
Bottle	2	13

Do the information provide sufficient evidence to conclude that type of feeding and nature of teeth are dependent? Use chi square test at 5% level of significance

Ans: **Problem to test**

H_0 : Type of feeding and nature of teeth are independent

Against H_1 : Type of feeding and nature of teeth are dependent

Type of feed	Nature of teeth		
	Normal	Defective	Total
Breast	a=18	b=12	a+b=30
Bottle	c=2	d=13	c+d=15
Total	a+c=20	b+d=25	N=45

Test statistic

$$\chi^2 = \frac{N(|ad - bc| - \frac{N}{2})^2}{(a + c)(b + d)(a + b)(c + d)}$$

$$= \frac{45(|18\times13-12\times2|-45/2)^2}{20\times25\times30\times15} = 7.031$$

Critical value

At $\alpha = 0.05$ level of significance critical value is $\chi^2_{\alpha(1)} = 3.84$

Decision

$\chi^2 = 7.031 > \chi^2_{0.01(1)} = 3.84$

Reject H_0 at 0.05 level of significance

Conclusion

Type of feeding and nature of teeth are dependent

6. Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% level of confidence when coefficient of variation is 60%.

Ans: Here, C.V. = 60% = 0.6

$P(|\bar{x} - \mu| \leq 0.1\mu) = 0.95$ .....(i)

Confidence level $(1-\alpha) = 95\% = 0.95$ then $\alpha = 0.05$

Now, $P(|\bar{x} - \mu| \leq \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}) = 1 - \alpha$

$\Rightarrow \quad P(|\bar{x} - \mu| \leq 1.96 \times \frac{\sigma}{\sqrt{n}}) = 0.95 \quad \ldots\ldots(ii)$

From equation (i) and (ii)

$$0.1\mu = 1.96 \times \frac{\sigma}{\sqrt{n}}$$

$\Rightarrow \quad \sqrt{n} = \frac{1.96}{0.1} \times \frac{\sigma}{\mu}$

$\Rightarrow \quad n = (1.96/0.1 \times \sigma/\mu)^2$

$\Rightarrow \quad n = 384.16 \times CV^2$

$\Rightarrow \quad n = 384.16 \times (0.6)^2$

$\Rightarrow \quad n = 138.29 \approx 138$

Hence required sample size is 138.

7. **A manufacture of computer paper has a production process that operates continuously throughout an entire production shift. The paper is expected to have an average length of 11inches and standard deviation is known to be 0.01 inch. Suppose random sample of 100 sheets is selected and the average paper length is found to be 10.68 inches. Set up 95% and 90% confidence interval estimate of the population average paper length.**

Ans: Average length of paper (μ) = 11

Standard deviation of paper (σ) = 0.01

Sample of paper sheet (n) = 100

Sample average length of paper($\bar{x}$) = 10.68

Level of confidence (100 - Θ)% = 95%

Level of significance (Θ) = 5%

Level of confidence (100 - Θ)% = 90%

Level of significance(Θ) = 10%

Confidence interval for population mean = $\bar{x} \pm \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}$

When Θ = 5%

Confidence interval = $10.68 \pm \frac{1.96 \times 0.01}{\sqrt{100}} = 10.68 \pm 0.0019$

Taking (-) sign 10.68 – 0.0019 = 10.678

Taking (+) sign 10.68 + 0.0019 = 10.6819

With 95% confidence interval estimate of population average length lies between 10.678 to 10.6819

When Θ = 10%

Confidence interval = $10.68 \pm \frac{1.645 \times 0.01}{\sqrt{100}} = 10.68 \pm 0.0016$

Taking (-) sign 10.68 – 0.0016 = 10.6784

Taking (+) sign 10.68 + 0.0016 = 10.6816

With 90% confidence interval estimate of population average length lies between 10.6784 to 10.6816

8. **A chemist use three catalyst for distilling alcohol and lay out were tabulated below**

Catalyst	Alcohol (in cc)				
C_1	380	430	410		
C_2	290	350	270	250	270
C_3	400	380	450		

Are there any significant difference between catalyst? Text at 5% level of significance. use KruskalWalli's H test.

Ans: **Problem to test**

H_0: There is no significant difference between catalyst.($Md_1 = Md_2 = Md_3$)

H_1: There is at least one significant difference between catalyst (At least one Mdi is different, i = 1 , 2 , 3)

Catalyst	Alcohol (cc)					R_i	R_i^2/n_i
C_1	380	430	410				
Rank	6.5	10	9			25.5	216.75
C_2	290	350	270	250	270		
Rank	4	5	2.5	1	2.5	15	45
C_3	400	380	450				
Rank	8	6.5	11			25.5	216.75
							$\sum \frac{R_i^2}{n_i} = 478.5$

$n_1 = 3 , n_2 = 5 , n_3 = 3$

$n = n_1 + n_2 + n_3 = 3+5+3 = 11$

$t_1 = 2 , t_2 = 2$

Test statistic

$$H = \frac{\frac{12}{n(n+1)}\Sigma\frac{R_i^2}{n_i} - 3(n+1)}{1 - \Sigma\frac{(t_i^3 - t_i)}{n^3 - n}} = \frac{\frac{12 \times 478.5}{11 \times 12} - 3 \times 12}{1 - \left\{\frac{2^3 - 2}{11^3 - 11} + \frac{2^3 - 2}{11^3 - 11}\right\}} = 7.5/\ 0.9909 = 7.568$$

Level of significance

$\Theta = 5\% = 0.05$

Critical value

From Kruskal Wallis table critical value is p = 0.009

Decision

P = 0.009 α = 0.05, reject H_0 at 0.05 level of significance.

Conclusion

There is at least one significant difference between catalyst.

9. **Consider the partially completed ANOVA table below. Complete the ANOVA table and answer the following:**

Source of Variation	Sum of Square	Degree of freedom	Mean sum of square	F value
Column	72	?	?	2
Rows	?	?	36	?
Treatments	180	3	?	?
Error	?	6	12	
Total	?	?		

i. What design was employed?

ii. How many treatments were compared?

Ans: Design LSD was employed

Here df for treatment = m-1 = 3 , hence m = 4
Hence 4 treatments were compared
df for columns = m-1 = 3
df for rows = m-1 = 3
Total number of observations = m^2 = 16
df for total = $m^2 - 1$ = 16-1 = 15
MSC = SSC/m-1 = 72/3 = 24
MST = SST/m-1 = 180/3 = 60
SSR = (m-1) MSR = 3× 36 = 108
SSE = (m-1)(m-2) MSE = 6× 12 = 72
TSS = SSC + SSR + SST + SSE = 72+108+180+72= 432
F_C = MSC/MSE = 24/12 = 2
F_R = MSR/MSE = 36/12 = 3
F_T = MST/MSE = 60/12 = 5

Source of Variation	Sum of Square	Degree of freedom	Mean sum of square	F value
Column	72	3	24	2
Rows	108	3	36	3
Treatments	180	3	60	5
Error	72	6	12	
Total	432	15		

10. Defined main component of queuing system.

Ans: Components of queuing system are as explained below

Arrival

Job arrives to the queuing system at random times. A counting process A(t) tells the number of arrivals that occurred by time t . In stationary queuing system arrivals occur at arrival rate

λ = Average number of arrivals per unit time = $\frac{EA(t)}{t}$ for any t > 0

Queuing and Routing to Servers

Arrived jobs are processed according to the order of their arrivals, on a first come first serve basis. When new job arrives it may find the system in different states. If one server is available at a time it will certainly take a new job. If several servers are available the job may be randomized to one of them or server may be chosen according to some rule.

Service

Once a server becomes available, it immediately starts processing the next assigned job. In practice service time are random because they depend upon amount of work required by each task. The average service time is μ. It varies from one server to other. The service rate is defined as the average number of jobs processed by a continuously working server during one unit of time. S(t) tells the number of customers served by time t

μ = Average number of customers served per unit time = $\frac{ES(t)}{t}$, t > 0

Departure

When the service is completed, the job leaves the system.

11. Jobs are sent to main farm computer at a rate of 4 jobs per minute. Arrivals are modeled by a binomial process.

i. Choose a frame size that makes the probability of a new received during each frame equal to 0.1.

ii. **using the chosen frame compute the probability of more than 4 jobs received during one minute.**

iii. **Compute mean and variance of inter arrival time?**

Ans: Here, $\lambda = 4$ per minute, $p = 0.1$

(i) $\Delta = p/\lambda = 0.1/4 = 0.025$ min

For $t = 1$, $n = t/\Delta = 1/.025 = 40$ frames

$n = 40$, $p = 0.1$

(ii) $P(X(n) > 4) = 1 - P(X(n) \leq 4) = 1 - \left[\sum_{x=0}^{4} {}^{40}C_x (0.1)^x (0.9)^{40-x}\right]$

$= 1 - [(0.9)^{40} + 40 \times 0.1 \times (0.9)^{39} + 780 \times (0.1)^2 (0.9)^{38} + 9880 \times (0.1)^3 (0.9)^{37} + 91350 \times (0.1)^4 (0.9)^{36}] = 0.37$

(iii) $E(T) = 1/\lambda = 1/4 = 0.25$ min $= 15$ sec

$V(T) = \frac{1-p}{\lambda^2} = \frac{0.9}{4^2} = 0.056$

12. **Write short notes of the following:**

i. Need of non parametric statistical methods.

Ans: Most of the hypothesis testing procedures so far such as Z test, t test, F test are based upon the assumption that the random samples are selected from a normal population. If this is true, these methods can extract all the information that is available in sample and they usually give the best possible precision. Parametric tests depends on parameters, viz., mean or proportion or standard deviation of the population from which sample is taken.

In practice there are many circumstances in which sample are selected from non normal population. In such case, we can have no assumptions about parameters or normality about the population. In such special cases, Parametric are inevitable and are used for testing of hypothesis. Manynon parametric procedures are based upon ranked data or even categorical data. For such data no parametric tests are available.

Non parametric tests are often used in place of their parametric counterparts when certain assumptions about the underlying population are in weak state. However, if the sample size is large enough most non parametric tests can be viewed as (the usual normal theory based procedures applied to ranks- make this statement more clear).

Non parametric test is essential when

- Data size is small and parameter free
- Data is weak scaled
- Data is highly skewed
- Population is distribution free
- Quick decision is required

ii. Efficiency of Randomized Block Design relative to completely Randomized Design.

Ans: The mechanism of the precision of RBD as compared to CRD is called efficiency of RBD relative to CRD.

Let us consider design having t treatments with r replication each.

If we perform RBD then,

The mathematical model of RBD is $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$; $i = 1, 2, 3.....t$, $j = 1, 2, 3..... r$

If we perform CRD then,

The mathematical model of CRD is $y_{ij} = \mu + \tau_i + e_{ij}$; $i = 1, 2, 3.....t$, $j = 1, 2, 3 ... r$

Now the efficiency of RBD relative to CRD is given by

Precision of RBD/ Precision of CRD $= \frac{1}{MSE} / \frac{1}{MSE'} = \frac{MSE'}{MSE} = \frac{\sigma_e'^2}{\sigma_e^2}$

$= \frac{r(t-1)\,MSE + (r-1)\,MSB}{(rt-1)\,MSE}$

If $\frac{\sigma_e'^2}{\sigma_e^2} < 1$ then RBD is less efficient than CRD.

If $\frac{\sigma_e'^2}{\sigma_e^2} > 1$ then RBD is more efficient than CRD.

If $\frac{\sigma_e'^2}{\sigma_e^2} = 1$ then RBD and CRD are equally effective.

❑❑❑

TRIBHUVAN UNIVERSITY

Institution of Science and Technology

Bachelor Level/Second Year/Third Semester/Science — Full Marks: 60
Computer Science and Information Technology [STA 210] — Pass Marks: 24

TU QUESTIONS-ANSWERS 2077

Group A

Attempts any two questions **(2 x 10 = 20)**

1. Describe the concept of sampling distribution of mean with reference to the population data (20, 21, 22 & 23) of size 4. In order to explain this, perform simple random sampling with replacement taking all possible samples with sample size n = 2. While describing the sampling distribution following issues will be covered.

a. population mean & population variance, and its distribution
b. Sample mean & sample variance, and its distribution
c. Comparison of population mean and sample mean; population variance and sample variance; population distribution and sampling distribution based on the given data.
d. Standard error of mean
e. Final comments based on your result

Ans:

Population data(Y) = 20,21,22,23

Population mean $(\bar{Y}) = \frac{\Sigma Y}{n} = \frac{(20 + 21 + 22 + 23)}{4} = 21.5$

Population variance $(\sigma^2) = \frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2$

$= \frac{(20^2 + 21^2 + 22^2 + 23^2)}{4} - (21.5)^2 = 1.25$

Population size(N) = 4
Sample size (n)= 2
Sampling with replacement
Total samples = $N^n = 4^2 = 16$

Sample	Sample mean $(\bar{y}) = \frac{\Sigma y}{n}$	Sample variance $(s^2) = \frac{1}{n-1}\{\Sigma y^2 - n\bar{y}^2\}$
(20,20)	20	0
(20,21)	20.5	0.5
(20,22)	21	2
(20,23)	21.5	4.5
(21,20)	20.5	0.5
(21,21)	21	0
(21,22)	21.5	0.5
(21,23)	22	2
(22,20)	21	2
(22,21)	21.5	0.5
(22,22)	22	0

(22,23)	22.5	0.5
(23,20)	21.5	4.5
(23,21)	22	2
(23,22)	22.5	0.5
(23,23)	23	0
Total	344	20

$$E(\bar{y}) = \frac{344}{16} = 21.5$$

$$E(s^2) = \frac{20}{16} = 1.25$$

Hence, $E(\bar{y}) = \bar{Y}$

$E(s^2) = \sigma^2$

Sampling distribution of mean is normal

Standard error of mean $SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{\sqrt{1.25}}{\sqrt{2}} = 0.79$

Here ,

Sample mean is unbiased estimate of population mean

Sample variance is unbiased estimate of population variance

2. **It was reported somewhere that children whenever plays the game in computer, they used the computer very roughly which may reduce the lifetime of a computer. The random access memory (RAM) of a computer also plays a crucial role on the lifetime of a computer. A researcher wanted to examine how the lifetime of a personal computer which is used by children is affected by the time (in hours) spends by the children per day to play games and the available random access memory (RAM) measured in megabytes (MB) of a used computer. The data is provided in following table.**

Lifetime (years)	5	1	7	2	3	4	6
Play time (hours)/day	2	8	1	5	6	3	2
RAM in MB)	8	2	6	3	2	4	7

Identify which one is dependent variable? Solve this problem using multiple linear regression model and provide problem specific interpretations based on the regression model developed.

Ans: **Dependent variable is life time**

Life time (yrs) y	Play time hrs/day x_1	RAM (MB) x_2	yx_1	yx_2	x_1x_2	x_1^2	x_2^2
5	2	8	10	40	16	4	64
1	8	2	8	2	16	64	4
7	1	6	7	42	6	1	36
2	5	3	10	6	15	25	9
3	6	2	18	6	12	36	4
4	3	4	12	16	12	9	16
6	2	7	12	42	14	4	49
$\Sigma y=28$	$\Sigma x_1 = 27$	$\Sigma x_2 = 32$	$\Sigma yx_1 = 77$	$\Sigma yx_2 = 154$	$\Sigma x_1x_2 = 91$	$\Sigma x_1^2 = 143$	$\Sigma x_2^2 = 182$

To fit regression equation

$y = b_0 + b_1x_1 + b_2x_2$ .....(i)

We have to find b_0, b_1 and b_2

$\Sigma y = nb_0 + b_1\Sigma x_1 + b_2\Sigma x_2$

or $28 = 7b_0 + 27b_1 + 32b_2$ ....(ii)

$\Sigma yx_1 = b_0\Sigma X_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2$

or $77 = 27b_0 + 143b_1 + 91b_2$ ....(iii)

$\Sigma yx_2 = b_0\Sigma x_2 + b_1\Sigma x_1x_2 + b_2\Sigma x_2^2$

or $154 = 32b_0 + 91b_1 + 182b_2$ ....(iv)

Using Cramer's rule

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constant
7	27	32	28
27	143	91	77
32	91	182	154

$$D = \begin{vmatrix} 7 & 27 & 32 \\ 27 & 143 & 91 \\ 32 & 91 & 182 \end{vmatrix}$$

$= 7(143 \times 182 - 91 \times 91) - 27(27 \times 182 - 32 \times 91) + 32(27 \times 91 - 32 \times 143)$

$= 2353$

$$D_1 = \begin{vmatrix} 28 & 27 & 32 \\ 77 & 143 & 91 \\ 154 & 91 & 182 \end{vmatrix}$$

$= 28(143 \times 182 - 91 \times 91) - 27(77 \times 182 - 154 \times 91) + 32(77 \times 91 - 154 \times 143)$

$= 16380$

$$D_2 = \begin{vmatrix} 7 & 28 & 32 \\ 27 & 77 & 91 \\ 32 & 154 & 182 \end{vmatrix}$$

$= 7(77 \times 182 - 154 \times 91) - 28(27 \times 182 - 32 \times 91) + 32(27 \times 154 - 32 \times 77)$

$= -1848$

$$D_3 = \begin{vmatrix} 7 & 27 & 28 \\ 27 & 143 & 77 \\ 32 & 91 & 154 \end{vmatrix}$$

$= 7(143 \times 154 - 77 \times 91) - 27(27 \times 154 - 32 \times 77) + 28(27 \times 91 - 32 \times 143)$

$= 35$

Now,

$$b_0 = \frac{D_1}{D} = 16380/2353 = 6.961$$

$$b_1 = \frac{D_2}{D} = -1848/2353 = -0.785$$

$$b_2 = \frac{D_3}{D} = 35/2353 = 0.014$$

Substitute value in equation (i) we get

$y = 6.961 - 0.785x_1 + 0.014x_2$

Hence

Life time (yrs) = 6.961 – 0.785 play time (hrs/day) + 0.014 RAM (MB)

Here b_0= 6.961 means average life of computer is 6.961 years when play time (hrs/day) is 0 and RAM is 0

b_1= -0.785 means for 1 hour per day increase in play time, life time of computer decrease by 0.785 years keeping RAM constant

b_2 = 0.014 means for 1GB unit increase in RAM life time of computer increase by 0.014 years keeping play time constant

3. **Explain the fundamental concepts of Latin Square Design (LSD) with its necessary conditions. Perform the analysis of variance from the following data and make final comments based on the analysis**

A(5)	B(10)	C(15)
C(20)	A(15)	B(10)
B(20)	C(5)	A(10)

Ans: Latin square design is the design in which all the principles of design are used. It has equal number of rows, columns and treatments. It is square in shape and Latin letters are used to represent treatments hence called Latin square design. In this case treatments are replicated neither along row wise nor along column wise. Local control is used along row wise and column wise. It is incomplete three way ANOVA.

Problem to test

H_{0R}: There is no significant difference between rows

H_{1R}: There is significant difference between rows.

H_{1C}: There is no significant difference between columns

H_{1C}: There is significant difference between columns.

H_{0T}: There is no significant difference between treatments

H_{1T}: there is significant difference between treatments.

				$T_{i..}$
	A(5)	B(10)	C(15)	30
	C(20)	A(15)	B(10)	45
	B(20)	C(5)	A(10)	35
$T_{.j.}$	45	30	35	G= 110

$G = 110, N = 3^2 = 9$

$$C.F. = \frac{G^2}{N} = 110^2/9 = 1344.44$$

$$\sum_{(i,j,k)} y_{ijk}^2 = 5^2 + 10^2 + 15^2 + 20^2 + 15^2 + 10^2 + 20^2 + 5^2 + 10^2 = 1600$$

$$TSS = \sum_{(i,j,k)} y_{ijk}^2 - C.F.$$

$$= 1600 - 1344.44 = 255.56$$

$$SSR = \frac{1}{3}\sum_{i} T_{i..}^2 - C.F.$$

$$= \frac{1}{3}\{(30)^2 + (45)^2 + (35)^2\} - 1344.44$$

$$= 38.89$$

$$SSC = \frac{1}{3}\sum_{j} T_{.j.}^2 - C.F.$$

$$= \frac{1}{3}\{(45)^2 + (30)^2 + (35)^2\} - 1344.44$$

$$= 38.89$$

$T_{..A} = 5+15+10 = 30$

$T_{..B} = 10+10+20 = 40$

$T_{..C} = 15+20+5 = 40$

$$SST = \frac{1}{3}\sum_{k} T_{..k}^2 - C.F.$$

$$= \frac{1}{3}\{(30)^2 + (40)^2 + (40)^2\} - 1344.44$$

$$= 22.22$$

$SSE = TSS - SSR - SSC - SST$

$= 255.56 - 38.89 - 38.89 - 22.22 = 190.56$

S.V.	d.f.	S.S.	M.S.	F_{Cal}	F_{Tab}
Row	2	38.89	19.445	0.204	$F_{0.05(2,2)} = 19$
Column	2	38.89	19.445	0.204	$F_{0.05(2,2)} = 19$
Treatment	2	22.22	11.11	0.116	$F_{0.05(2,2)} = 19$
Error	2	190.56	95.28		
Total	8	255.56			

Decision

$F_R = 0.204 < F_{0.05(2,2)} = 19$, accept H_{0T} at 5% level of significance.

$F_C = 0.204 < F_{0.05(2,2)} = 19$, accept H_{0C} at 5% level of significance.

$F_T = 0.116 < F_{0.05(2,2)} = 19$, accept H_{0T} at 5% level of significance.

Conclusion

There is no significant difference between rows, there is no significant difference between columns and there is no significant difference between treatments.

Group B

Attempts any EIGHT questions **(8 x 5 = 40)**

4. **A dealer of a DELL company located at New Road claimed that the average lifetime of a multimedia projector produced by Dell Company is greater than 60,000 hours with standard deviation of 6000 hours. In order to test his claim, sample of 100 DELL projectors are taken and the average life time was monitored and it was found to be 55,000 hours. Test the claim of the dealer at 5% level of significance.**

Ans: Average life time of multimedia projector (μ) = 60000 hrs
Standard deviation of multimedia projector (σ) = 6000 hrs
Sample size (n) = 100
Sample mean ($\bar{x}$) = 55000 hrs
Level of significance (α) = 5% = 0.05
Problem to test
H_0: Average lifetime of multimedia projector is 60000 ($\mu = 60000$)
H_1: Average lifetime of multimedia projector is more than 60000 ($\mu > 60000$)

Test statistic

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{55000-60000}{6000/\sqrt{100}} = -5000/600 = -12$$

Critical value

At $\alpha = 0.05$ critical value for one tailed test is $Z_{tab} = Z_\alpha = 2.64$

Decision

Here $|Z| = 12 \square Z_{tab} = 2.64$, reject H_0 at 5% level of significance.

Conclusion

The claim of dealer of a DELL company located at New Road that average life of multimedia projector produced by Dell Company is more than 60000 hours is correct.

5. Based on the following information, performed the following:

a. Test whether two mean are significance different (α = 5%) using independent t-test.

b. Compute 95% confidence interval estimation for the difference of mean.

c. Show the linkage between testing of hypothesis and confidence interval estimation in this problem.

	Group A	Group B
Sample mean	10	15
Sample Standard Deviation	3	5
Sample Size	49	64

Ans: $\bar{x}_A = 10, \bar{x}_B = 15, s_A = 3, s_B = 5, n_A = 49, n_B = 64$
Let μ_A and μ_B be population mean of group A and group B respectively
Problem to test
H_0: $\mu_A = \mu_B$
H_1: $\mu_A \neq \mu_B$
Test statistic

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{10 - 15}{\sqrt{\frac{9}{49} + \frac{25}{64}}} = -6.59$$

Critical value

At $\alpha = 0.05$ critical value for two tailed test is

$t_{tab} = t_{\alpha}(n_A + n_B - 2) = t_{0.05\,(111)} = 1.98$

Decision

$|t| = 6.59 > t_{0.05\,(111)} = 1.98$

Reject H_0 at 0.05 level of significance

Conclusion

There is significant difference between two means

Confidence interval estimate for difference of mean

$$(\bar{x}_A - \bar{x}_B) \pm t_{\alpha}\,(n_A + n_B - 2)\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = (10 - 15) = \pm 1.98 \times \sqrt{\frac{9}{49} + \frac{25}{64}}$$

$$= -5 \pm 1.98 \times 0.757 = -5 \pm 1.5$$

Taking (-)

$-5 - 1.5 = -6.5$

Taking (+)

$-5 + 1.5 = -3.5$

Hence 95% confidence interval estimate for difference of mean is -3.5 to -6.5

The linkage between testing of hypothesis and confidence interval estimate is that null hypothesis is accepted if test statistic is between the interval estimates.

6. **A study of 1000 computer engineers conducted by their professional organization reported that 300 stated that their firms' greatest concern was to uplift the professional quality of work. In order to conduct a follow up study to estimate the population proportion of computer engineers to fulfill their greatest concern within ±0.01 with 99% confidence interval, how many computer engineers would be required to be surveyed?**

Ans: Proportion having greatest concern to uplift the professional quality of work

$(p) = \frac{300}{1000} = 0.3$

$q = 1 - p = 1 - 0.3 = 0.7$

$d = 0.01$

Confidence interval = 99%

Level of significance (α) = 1% = 0.01

Sample size (n) = ?

Required sample size $(n) = \frac{Z_{\alpha}^2\,pq}{d^2} = \frac{2.58^2 \times 0.3 \times 0.7}{0.01^2} = 13978.44 \approx 13978$

7. **A survey was conduct to see the association between hacking status of the email and the type of email account. The survey has reported the following cross tabulation.**

Type of e-mail account	Hacking status	
	Yes	No
Yahoo	60	15
Gmail	20	120

Do the information provide sufficient evidence to conclude that the type email account and the hacking status is associated? Use Chi-square test at 1% level of significance.

Ans: Problem to test

H_0 : Type of email account and Hacking status are independent

H_1 : Type of email account and Hacking status are dependent

Type of e-mail account	Hacking status		
	Yes	No	
Yahoo	a=60	b=15	a+b = 75
Gmail	c=20	d=120	c+d = 140
	a+c = 80	b+d = 135	N = 215

Test statistic

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{215\times(60\times120-15\times20)^2}{75\times140\times80\times135} = 90.26$$

Critical value

At 1% level of significance critical value is $\chi^2_{\alpha(1)} = \chi^2_{0.01(1)} = 6.64$

Decision

$\chi^2 = 90.26 > \chi^2_{0.01(1)} = 6.64$

$Reject H_0$ at 1% level of significance

Conclusion

The information provide sufficient evidence to conclude that the type email account and the hacking status is associated

8. **A machine produce metal rods used in an automobile suspension system. A random sample of 6 rods is selected and diameter is measured. The measuring data(in millimeters) are as follows. Assuming that the sample drawn from the normally distributed population.**

8.24	8.26	8.20	8.28	8.21	8.23

Find 95% two sided confidence interval on the mean rod diameter, and interpret the result with reference to the given problem.

Ans:

Diameter of metal rod (x)	x^2
8.24	67.897
8.26	68.227
8.20	67.24
8.28	68.558
8.21	67.404
8.23	67.732
$\sum x = 49.42$	$\sum x^2 = 407.058$

Let population mean of diameter of rod = μ

Sample size (n) = 6

Sample mean $(\bar{x}) = \frac{\sum x}{n} = \frac{49.42}{6} = 8.236$

Sample sd (s) $= \frac{1}{n-1}(\sum x^2 - n\bar{x}^2) = \frac{1}{6-1}(407.058 - 6\times 8.236^2) = 0.013$

Confidence interval = 95%

Level of significance (α) = 5% = 0.05

Confidence interval for mean diameter of rod

$\bar{x} \pm t_{\alpha(n-1)} s/\sqrt{n}$

$= 8.236 \pm \frac{2.57 \times 0.013}{\sqrt{6}} = 8.236 \pm 0.013$

Taking (-)

8.236 – 0.013 = 5.223

Taking (+)

8.236+ 0.013 = 5.249

Hence diameter 5.223 to 5.249 lies within 95% confidence interval.

According to the problem diameters 8.21, 8.26 and 8.28 lies outside the 95% confidence interval

9. **Use Mann-Whitney U test to assess whether the following satisfaction score based on the performance of two different special types of gadgets at 5% level of significance.**

Gadget A	50	40	30	20
Gadget B	40	30	10	40

Ans:

Gadget A	Gadget B	Rank of gadget A	Rank of gadget B
50	40	8	6
40	30	6	3.5
30	10	3.5	1
20	40	2	6
		R_A= 19.5	R_B=16.5

Let Md_A and Md_B be median life of gadget A and gadget B respectively

Problem to test

$H_0 : Md_A = Md_B$

$H_1 : Md_A \neq Md_B$

$U_A = n_A n_B + \frac{n_A(n_A+1)}{2} - R_A$

$= 4 \times 4 + \frac{4\times5}{2} - 19.5 = 6.5$

$U_B = n_A n_B - U_A = 4 \times 4 - 6.5 = 9.5$

$U_0 = \min(U_A, U_B) = 6.5$

Level of significance(α) = 5% = 0.05

Critical value

At 5% level of significance critical value is $U_{\alpha(n_A,n_B)} = 0$

Decision

$U_0 = 6.5 > U_{\alpha(n_A,n_B)} = 0$

Accept H_0 at 5% level of significance

Conclusion

There is no significant difference between gadgets A and B.

10. Define Markov chain and describe its characteristics.

Ans: Let $\{X_n\}$ be a sequence of values describing a mutually exclusive and exhaustive system of events. Let X_n values take only discrete the union I of all possible values of X_n is then a countable set called the state space of the process. Each element i_0 I is called a state. The index n is of time. The number of events may be finite or infinite the values of $\{X_n\}$ is said to be a Markov chain or Markov dependent if for all $i_0, i_1, i_2, \ldots i_{n-1}, i_n \in I$ and for all n.

$P(X_n = i_n (X_0 = i_0, X_1 = i_1 \ldots\ldots\ldots X_{(n-1)} = i_{(n-1)}) = P(X_n = i_n X_{n-1} = i_{n-1})$

The conditional distribution of X_n given the values $X_0, X_1, \ldots\ldots, X_{n-1}$ depends only on X_{n-1} not on the preceding values. If the state space is finite then we have finite Markov Chain.

The probability of transitioning to any particular state is dependent on the current state and time elapsed.

11. Every day is generally considered as either sunny or rainy. A sunny day is followed by another sunny day with probability 0.8 whereas a rainy day is followed by a sunny day with probability 0.4 .Suppose it rains on Monday. Make forecasts for Tuesday and Wednesday.

Ans: Let state 1 = sunny and state 2 = rainy

Transition probabilities are

$P_{11} = 0.8, P_{12} = 0.2, P_{21} = 0.4, P_{22} = 0.6$

One step transition probability matrix

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}$$

For Tuesday rainy chance $(P_{22}) = 0.6 = 60\%$

For Tuesday sunny chance $(P_{21}) = 0.4 = 40\%$

For Wednesday

$$P^{(2)} = PP = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}\begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.64+0.08 & 0.16+0.12 \\ 0.32+0.24 & 0.08+0.36 \end{bmatrix} = \begin{bmatrix} 0.72 & 0.28 \\ 0.56 & 0.44 \end{bmatrix}$$

For Wednesday rainy chance $(P_{22}) = 0.44 = 44\%$

For Wednesday sunny chance $(P_{21}) = 0.56 = 56\%$

12. Write shorts notes on the following:

i. Test of equality of two variances

Ans: Let us consider two independent samples of sizes n_1 and n_2 be selected from two normal populations with unknown variance. Let $x_1, x_2 \ldots\ldots x_{n_1}$ and $y_1, y_2 \ldots\ldots\ldots y_{n_2}$ be samples of sizes n_1 and n_2 respectively. We want to test whether the two random samples have been drawn from the normal population with same variance or not.

Different steps in the test are

Problem to test

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$ (two tailed test) or $H_1 : \sigma_1^2 \sigma_2^2$ (One tailed right)

Test statistic

$F = \dfrac{S_1^2}{S_2^2}$ F distribution with$(n_1\text{-}1 , n_2 - 1)$ degree of freedom.

If $S_1^2 > S_2^2$

Where, $S_1^2 = \frac{1}{n_1-1}\Sigma(x-x)^2$, $S_2^2 = \frac{1}{n_2-1}\Sigma(y-y)^2$

Level of significance:

Let α be the level of significance. Usually we take α = .05 unless we are given.

Critical value:

Critical or tabulated value of F is obtained from table according to the level of significance, degree of freedom and alternative hypothesis.

Decision:

Reject H_0 at α level of significance if $F > F_{\alpha\,(n_1-1\,,n_2-1)}$, accept otherwise.

ii. Adjusted R^2

Ans: It is corrected goodness of fit of regression model. It determines percentage variation in output as explained by inputs. Higher value represents the model fits good.

It is suggested that the adjusted R^2 should be used in place of R^2 in multiple regression model. Adjusted R^2 is simply a R^2adjusted by its degree of freedom and reflects both the number of independent variables and sample size used in the model. Adjusted R^2 is considered as an important measure for the comparison of two or more regression models that predict same dependent variable with different independent variables.

$R^2_{adjusted}$ $(\bar{R}^2) = 1 - \frac{(n-1)}{(n-k-1)}[1 - R^2]$; where n = no of pair of observations, k = no of independent variables.