1) Suppose a population of 4 computers with their lifetimes 3, 5, 7 & 9 years. Comement on the population distribution. Assume that you sample with replacement, select all possible samples of n=2, and construct sampling distribution of mean and compare the population distribution and sampling distribution of mean. Compare population mean versus mean of all sample means and population variance versus variance of sample means and comment on them with the support of theoritical considerations if any.

Soln

Here, Population size (N) = 4

Sample size (n) = 2

Number of possible samples of n=2, that can be drawn from the population of size N=4 by using replacement is :- $N^n = 4^2 = 16$

∴ possible samples: (3,3), (3,5) (3,7), (3,9), (5,3), (5,5), (5,7), (5,9), (7,3), (7,5), (7,7), (7,9) (9,3), (9,5), (9,7), (9,9)

Now, Calculation of pop^n mean & variance

Y_i	$Y_i - \bar{Y}$	$(Y-\bar{Y})^2$
3	-3	9
5	-1	1
7	1	1
9	3	9
$\Sigma Y = 24$		$\Sigma(Y-\bar{Y})^2 = 20$

Population mean ($\bar{Y}$) $= \frac{\Sigma Y_i}{N} = \frac{26}{4} = 6$

Pop^n variance (σ^2) $= \frac{\Sigma(Y_i - \bar{Y})^2}{N} = \frac{20}{4} = 5$

calculation of sample mean & variation of the sampling distribution of means

S.No	Sample value (y_i)	Sample Means ($\bar{y}_i$)	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	(3,3)	3	-3	9
2	(3,5)	4	-2	4
3	(3,7)	5	-1	1
4	(3,9)	6	0	0
5	(5,3)	4	-2	4
6	(5,5)	5	-1	1
7	(5,7)	6	0	0
8	(5,9)	7	1	1
9	(7,3)	5	-2	4
10	(7,5)	126	0	0
11	(7,7)	7	1	1
12	(7,9)	8	2	4
13	(9,3)	6	0	0
14	(9,5)	7	1	1
15	(9,7)	8	2	4
16	(9,9)	9	3	9
		96		43

Mean of sample mean $(\bar{\bar{y}}) = \frac{\Sigma \bar{y}_i}{\text{No. of samples } (N^n)}$

$= \frac{96}{16}$

$= 6$

Mean of sample mean = 6, mean of popⁿ mean = 6
So we can conclude that the mean of the sampling distribution of the sample means is equal to the popⁿ mean.

Variation of sample means is

$$V(\bar{y}) = \frac{\Sigma(\bar{y}_i - \bar{\bar{y}})^2}{N^n} = \frac{43}{16} = 2.6875$$

Here, popn variation (σ^2) = 5 and sample variance = 2.6875. Which means that sample variation is greater than popn variance.

- A computer manager is keenly interested to know how efficiency of her new computer program depends on the size of incoming data and data structure. Efficiency will be measured by the number of processed request per hour. Data structure may be measured on how many tables were used to arrange each data set. All information was put together as follows.

Data size (GB)	6	7	7	8	10	10	15
No. of tables	4	20	20	10	10	2	1
Processed req.	40	55	50	41	17	26	16

Identify which one is dependent variable? Fit the appropriate multiple regression model and provide problem specific interpretations of the fitted regression coefficients.

Let y = GB, no. to tables $= x_1$ and processed request $= x_2$

y	x_1	x_2	x_1^2	x_2^2	yx_1	yx_2	x_1x_2
6	4	40	16	1600	24	240	160
7	20	55	400	3025	140	385	1100
7	20	50	400	2500	140	350	1000
8	10	41	100	1681	80	328	410
10	10	17	100	289	100	170	170
10	2	26	4	676	20	260	52
15	1	16	1	256	15	240	16
$\Sigma y = 63$	$\Sigma x_1 = 67$	$\Sigma x_2 = 245$	$\Sigma x_1^2 = 1021$	$\Sigma x_2^2 = 10027$	$\Sigma yx_1 = 519$	$\Sigma yx_2 = 1973$	$\Sigma x_1x_2 = 2908$

$$D_2 = \begin{vmatrix} 7 & 63 & 245 \\ 67 & 519 & 2968 \\ 245 & 1973 & 10027 \end{vmatrix} = 1449841\ 60536$$

$$D_3 = \begin{vmatrix} 7 & 67 & 63 \\ 67 & 1021 & 519 \\ 245 & 2968 & 1973 \end{vmatrix} = -285036 - 285612$$

$$b_0 = \frac{D_1}{D} = \frac{24182701}{1640633} = 15.221\ 14.77$$

$$b_2 = \frac{D_2}{D} = \frac{1449841}{1211841} = 0.1196 = \frac{60536}{1640633} = 0.0368$$

$$b_3 = \frac{D_3}{D} = \frac{-285036}{1211841} = 0.210 = \frac{-285612}{1640633} = 0.174086$$

GB & dependent variable,

substitute value in eqn (1),

$y = 15.221 + 0.1196x_1 - 0.21x_2$

$y = 14.7739 + 0.0368x_1 - 0.174086x_2$

To fit $y = b_0 + b_1 x_1 + b_2 x_2$

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

$63 = 7b_0 + 67 b_1 + 245 b_2 \longrightarrow (1)$

$\Sigma y x_1 = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$

$519 = 67 b_0 + 1021 b_1 + 2908 b_2 \rightarrow (11)$

$\Sigma y x_2 = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$

$1973 = 245 b_0 + 2908 b_1 + 10027 b_2 \rightarrow (111)$

Using cramer's rule.

coefficient of b_0	coefficient of b_1	coefficient of b_2	constant
7	67	245	~~63~~ 519 63
67	1021	2908	519
245	2908	10027	1973

Now,

$$D = \begin{vmatrix} 7 & 67 & 245 \\ 67 & 1021 & 2908 \\ 245 & 2908 & 10027 \end{vmatrix}$$

$= 7(1021 \times 10027 - 2908 \times 2908) - 67(67 \times 10027 - 245 \times 2908) + 245(67 \times 2908 - 1021 \times 245)$

$= $ ~~1211841~~ 1640633

$$D_1 = \begin{vmatrix} 63 & 67 & 245 \\ 519 & 1021 & 2908 \\ 1973 & 2908 & 10027 \end{vmatrix} = $$ ~~1844525~~ 2418701

3- State and explain the mathematical model for randomized complete block design. Explain all the steps to be adopted to carry out the analysis and finally prepare the ANOVA table.

When the experimental material is not homogeneous the RBD is better than CRD. The RBD is the design where the treatments are allocated in random manner but randomization is restricted that each treatments ~~are atte~~ must occur one in each row or once in each column. It is row wise or column wise. It is based upon the all principles of design namely replication, randomization and local control.

Mathematical Model

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$$

where,

y_{ij} = j^{th} block receiving i^{th} treatment.

$i = 1, 2, \cdots, t$, $j = 1, 2, \cdots r$

μ = constant effect

τ_i = effect due to i^{th} treatment

β_j = effect due to j^{th} treatment

e_{ij} = error due to chance.

Statistical Analysis

In the model $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$, where μ, τ, β are parameters determined by the

principle of least square by minimizing error sum of square.

$TSS = SST + SSB + SSE$

where

TSS = Total sum of square

SST = sum of square due to treatment

SSB = sum of square due to block

SSE = sum of square due to error.

Degree of freedom (d.f.) for various sum of square:

Degree of freedom for total sum of square $= rt - 1 = N - 1$

d.f. for sum of square due to treatment $= t - 1$

d.f for $SSB = r - 1$

d.f for. $SSE = t(r - 1)$

Mean sum of square (MSS):

The sum of square divided by the corresponding degree of freedom gives the respective mean of square.

Mean sum of sq. due to treatment $(MST) = \frac{SST}{t-1}$

Mean sum of square due to block $(MSB) = \frac{SSB}{r-1}$

Mean sum of square due to error $(MSE) = \frac{SSE}{(t-1)(r-1)}$

ANOVA TABLE

S.V	df	SS	M.S	fcal	ftab
treatment	t-1	SST	MST = SST/(t-1)	f_T = MST/MSE	$f_{\alpha\{(t-1),(t-1)(r-1)\}}$
Block	r-1	SSB	MSSB = SSB/(r-1)	f_B = MSB/MSE	$f_{\alpha\{(t-1)(t-1)(r-1)}$
Error	(r-1)(t-1)	SSE	MSE = SSE/(t-1)(r-1)		
Total	(rt-1)	TSS			

Decision:

Reject H_{0T} at α% level of significance if $f_T > f_{\alpha\{(t-1),(t-1)(r-1)\}}$ accept otherwise.

Reject H_{0c} at α% level of significance if $f_c > f_{\alpha\{(t-1),(t-1)(r-1)\}}$ accept otherwise.

Group-B

In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7 with a sample deviation of 2.2. At the 1% level of significance, do these data provide considerable evidence that the mean number of concurent users is greater than 36? Draw your conclusion based on your result.

Given, sample size(n) = 100

sample mean($\bar{x}$) = 37.7

sample [illegible]

population mean(μ) = 35

Problem to test

H_0: The mean number of concurrent users is 35
H_1: The mean number of concurrent users is greater than 35. (one tail right)

Test statistic

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{37.7 - 35}{9.2/\sqrt{100}} = 2.72$$

Critical value.

At $\alpha = 1\%$. critical value for one tailed test is

$Z_{tab} - Z_{\alpha/2} = 2.32$

Decision:

$Z = 2.72 > Z_{tab} = 2.32$, reject H_0 at 1% level of significance.

Conclusion:

The mean number of concurrent users is greater than 35.

A sample of 250 items from lot A contains 10 defective items and a sample of 300 items from lot B is found to contain 18 defective items. At a significance level $\alpha = 0.05$, is there a significant difference between the quality of the two lots?

Soln:-

Here, lot A

Sample size $(n_1) = 250$

defective items $(x_1) = 10$

$$p_1 = \frac{x_1}{n_1} = \frac{10}{250} = 0.04$$

lot B

sample size $(n_2) = 300$

defective items $(x_2) = 18$

$$p_2 = \frac{x_2}{n_2} = \frac{18}{300} = 0.06$$

Let P_1 and P_2 be pop^n proportion of lot A and lot B.

Problem to test

H_0: There is no significant difference between the quality of the two lots.

H_1: There is significant difference between the quality of the two lots. (Two tailed)

Test statistic

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{250 \times 0.04 + 300 \times 0.06}{250 + 300} = 0.051$$

$$Z_{\alpha} \cdot Q = 1 - P = 1 - 0.051 = 0.95$$

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.04 - 0.06}{\sqrt{0.95 \times 0.051\left(\frac{1}{100} + \frac{1}{300}\right)}}$$

$$= -1.061$$

$$|z| = 1.061$$

critical value

At $\alpha = 0.05$, critical value is $Z_{\alpha/2} = 1.96$

Decision

$z = 1.061 < Z_{tab} = 1.96$, accept H_0 at 5% level of significance.

Conclusion

There is no significant difference between the quality of two lots.

Modern email servers and anti-spam filters attempt to identify spam emails and direct them to a junk folder. There are various ways to detect spam and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to spam depends on whether it contains images or not. The following data were collected on $n = 1000$ random email messages.

Spam status	Image containing status: With images	No images	Total
Spam	160	240	400
No spam	140	460	600
Total	300	700	1000

Access whether being spam and containing images are independent factors at 1% level of significance.

Here,

Spam status	With images	No images	Total
Spam	160 (a)	240 (b)	400
No spam	140 (c)	460 (d)	600
Total	300	700	1000

Problem to test

H_0: Being spam and containing images are independent.
H_1: Being spam and containing images are dependent.

Test stalistic

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

$$= \frac{1000(160\times460 - 240\times140)^2}{300\times700\times400\times600}$$

$$= 31.75$$

Critical value

At $\alpha = 0.01$, critical value for 1 degree of freedom is $X^2_{0.01(1)} = 6.635$

Decision

$X^2 = 31.75 > X^2_{0.01(1)} = 6.635$, reject H_0.

conclusion

Being spam and containing images are dependent.

Two computer maker A and B compete for a certain market. Their users rank the quality of computer on a 4 point scales. will be recommended to others. The following counts were observed.

computer maker	Not satisfied	satisfied	Good	Excellent
A	20	40	70	20
B	10	30	40	20

Is there a significant difference on customer

satisfaction of the computers produced by A and B using Mann-Whitney U test at 5%. l.o.s.

A	Rank	B	Rank
20	3	10	1
40	6.5	30	5
70	8	40	6.5
20	3	20	3
	$R_1 = 20.5$		$R_2 = 15.5$

Sample size of A $(n_1) = 4$
Sample size of B $(n_2) = 4$
Sum of ranks of A $(R_1) = 20.5$
Sum of ranks of B $(R_2) = 15.5$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= 4 \times 4 + \frac{4(4+1)}{2} - 20.5$$

$$= 5.5$$

$$U_2 = n_1 n_2 - U_1 = 4 \times 4 - 5.5 = 10.5$$

$$U_0 = \min\{U_1, U_2\} = \min\{5.5, 10.5\} = 5.5$$

let Md_1 and Md_2 be median of A and B.

Problem to test

H_0: There is no significant difference between computer produced by A and B.

H_1: There is significant difference between computer produced by A and B.

Test statistic

$U_0 = 5.5$

Critical value

Let $\alpha = 0.05$ be level of significance then critical value is $p = 0.2429$

For two tailed,

$2p = 2 \times 0.2429 = 0.4858$

Decision

$2p = 0.4858$ & $\alpha = 0.05$, accept H_0.

conclusion

There is no significant difference between computer produced by A and B.

In some town, each day is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7 whereas a rainy day is followed by a sunny day with probability 0.4. Weather conditions on this problem represent homogeneous markov chain with 2 state : state1 = "sunny" and state2 = "rainy". Transition probability matrix of sunny ~~da~~ and rainy days is given:

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$$

Compute the probability of sunny days and rainy day

using the steady-state equation for the Markov chain

Given,

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Let $\pi = (\pi_1 \ \pi_2)$

Now, $\pi P = \pi$

$$[\pi_1 \ \pi_2] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = [\pi_1 \ \pi_2]$$

$$\text{or, } [0.7\pi_1 + 0.4\pi_2 \quad 0.3\pi_1 + 0.6\pi_2] = [\pi_1 \ \pi_2]$$

So,

$$0.7\pi_1 + 0.4\pi_2 = \pi_1 \rightarrow (1)$$

$$0.3\pi_1 + 0.6\pi_2 = \pi_2 \rightarrow (2)$$

From (1),

$$0.7\pi_1 - \pi_1 = -0.4\pi_2$$

$$\text{or, } -0.3\pi_1 = -0.4\pi_2$$

$$\text{or, } \pi_2 = \frac{0.3}{0.4}\pi_1 = \frac{3}{4}\pi_1 = 0.75\pi_1$$

Since $\pi_1 + \pi_2 = 1$

$$\text{or, } \pi_1 + 0.75\pi_1 = 1$$

$$\text{or, } 1.75\pi_1 = 1$$

$$\text{or, } \pi_1 = 1/1.75$$

$$\therefore \pi_1 = 0.57 = \frac{4}{7}$$

$$\pi_2 = 0.75\pi_1 = 0.75 \times 0.57 = 0.4 = 3/7$$

The probability of sunny day is $4/7$ and rainy day is $3/7$.

10. Consider a completely randomized design with 4 treatments with 7 observation on each. For the NOVA summary table below, fill all the missing result. Also indicate your statistics decision

Source	d.f	S.S	MS.S	fratio
Treatment	?	SSA = ?	70	F = ?
Error	?	SSE = 590	?	
Total	?	SST = ?		

Here,

treatment (t) = 3 , & r = 7

Source	d.f	S.S	MSS	Fratio	Ftab
Treatment	4-1=3	210	70	2.85	f0.05(3, 24)
Error	24	590	24.58		
Total	27	800			

We know,

~~MSST = SST~~

Error = t(r-1)

= 4(7-1)

= 4 × 6 = 24

Treatment = t-1 = 4-1 = 3

Total = n-1 = 28-1 = 27

$$MSST = \frac{SST}{t-1}$$

$$70 = \frac{SST}{3}, \quad \therefore SST = 70 \times 3 = 210$$

$TSS = SST + SSE$

$= 210 + 590$

$= 800$

$$MSSE = \frac{SSE}{t(r-1)} = \frac{590}{24} = 24.58$$

$$F_{ratio} = \frac{MSST}{MSSE} = \frac{70}{24.85} = 2.85$$

Decision

~~F-tab~~

$F_T = 2.85 < f_{0.05}(3,24) = 3.01$, accept H_0. at 5% level of significance.

~~conclusion~~

Following are the scores obtained by 10 university staffs on the computer proficiency skills before and after training. It was assumed that the proficiency of computer skills is expected to be increased after training.

Staffs	Score	
	Before	After
1.	50	55
2.	30	40
3.	15	30
4.	22	30
5.	34	36
6.	45	45

7	40	41
8	10	30
9	26	40

Test 5% level of significance whether the training is effective to improve the computer proficiency skills applying appropriate statistical test. Assume that the given score follows normal distribution.

Here,

sample size (n) = 9

$\alpha = 5\%$.

Staff	Before training (X)	After training (Y)	$d = X - Y$	d^2
1	50	55	-5	25
2	30	40	-10	100
3	15	30	-15	225
4	22	30	-8	64
5	34	36	-2	4
6	45	45	0	0
7	40	41	-1	1
8	10	30	-20	400
9	26	40	-14	196
			$\Sigma d = -75$	$\Sigma d^2 = 1015$

$$d' = \frac{\Sigma d}{n} = \frac{-75}{9} = -8.33$$

$$S_d^2 = \frac{1}{n-1}\{\Sigma d^2 - n d'^2\} = \frac{1}{9-1}\{1015 - 9 \times (-8.33)^2\}$$

$sd^2 = 48.81$

$\therefore sd = 6.99$

~~critical~~

Problem to test

H_0: Training is not effective

H_1: Training is effective. (two tailed)

Test statistic

$$t = \frac{d'}{\frac{sd}{\sqrt{n}}} = \frac{-8.33}{\frac{6.99}{\sqrt{9}}} = -3.58$$

Critical value.

At $\alpha = 0.05$, critical value is $t_{tab} = t_{\alpha/2(n-1)} = 2.306$

Decision

$|t| = 3.58 > t_{tab} = 2.306$, reject H_0.

Conclusion

Training is effective.

Q. Write short notes:

a) Concept of LSD:

When the experimental material is not homogenous 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 error are to be controlled simultaneously. It is based upon the all principles of design namely replication, randomization and local control.

b) Multiple correlation:

The relationship among three or more variables at the same time is called multiple correation. Let us consider three variables X_1, X_2 and X_3 the multiple correlation coefficient of X_1 with X_2 and X_3 is denoted by $R_{1.23}$ and given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}}$$

It lies between 0 and 1. i.e.

(i) $0 \leq R_{1.23} \leq 1$

(ii) $0 \leq R_{2.13} \leq 1$

(iii) $0 \leq R_{3.12} \leq 1$

3 - Define queuing system with suitable examples. Also explain the main components of queuing systems in brief.

Queuing system is facility consisting of one or several servers designed to perform certain tasks or process certain jobs and a queue of jobs waiting to be processed.

Eg:

→ A medical office serving patients.

→ A printer processing job sent to it from different computers etc.

Following are the main components of queuing:

- Arrival:

Job arrives to the queuing system at random times. A counting process A(t) tells the no. of arrivals that occured by time t. In stationary queuing system arrivals occur at arrival rate

λ = average no. 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 the server becomes available, it immediately ~~stop~~ start processing the next assigned job. The average service time is μ. It varies from one to other server. The service rate is defined as the average no. of jobs processed by a continuously working server during one unit of time. $s(t)$ tells the no. of customers served per unit time $= \frac{Es(t)}{t}$, $t>0$.

Departure:

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