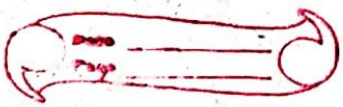
STA 210

Tribhuvan University
Institute of Science and Technology
Model Question



Group 'A'

1. Suppose a population of 4 computers with their lifetimes 3, 5, 7 and 9 years. Comment on the population distribution. Assuming 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.

⇒ Solution:

Population size $(N) = 4$

Population units = 3, 5, 7, 9

Sample size $(n) = 2$

Here,

$$\text{Population mean } (\mu) = \frac{X_1 + X_2 + X_3 + X_4}{N} = \frac{3+5+7+9}{4} = 6$$

Calculation Table:

X	N	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$
3	4	6	-3	9
5			-1	1
7			1	1
9			3	9
				$\Sigma(X - \bar{X})^2 = 20$

$$\therefore \text{Population variance } (\sigma^2) = \frac{\Sigma(X - \bar{X})^2}{N} = \frac{20}{4} = 5$$

All possible samples of $n=2$ drawn from the population with replacement is :

$N^n = 4^2 = 16$.

And the samples are (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) and (9,9).

- Calculation table for sampling distribution of mean ($\bar{\bar{x}}$) and variance ($V(\bar{x})$) :

S.N.	Samples	Sample mean ($\bar{x}_i$)	$\bar{\bar{x}} = \frac{\Sigma \bar{x}_i}{n}$	$\bar{x}_i - \bar{\bar{x}}$	$(\bar{x}_i - \bar{\bar{x}})^2$
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	6	0	0
8.	(5,9)	7		1	1
9.	(7,3)	5		-1	1
10.	(7,5)	6		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
		$\Sigma \bar{x}_i = 96$			$\Sigma(\bar{x}_i - \bar{\bar{x}})^2 = 40$

Here,

Sampling distribution of mean, $\bar{\bar{x}} = \frac{\Sigma \bar{x}_i}{n} = \frac{96}{16} = 6$

Here, Population mean (μ) and mean of all sample means ($\bar{\bar{x}}$) are equal.

$E(\bar{x}) = \mu = 6$

Thus, Sample mean ($\bar{x}$) is an unbiased estimate of the population mean (μ).

Similarly,

Variance of sample means, $V(\bar{x}) = \frac{\Sigma(\bar{x}_i - \bar{\bar{x}})^2}{n} = \frac{40}{16} = 2.5$

Also,

Sample variance $= \frac{\sigma^2}{n} = \frac{5(4-2)}{2(4-1)} = \frac{10}{6} = \frac{5}{2} = 2.5$

S.E$(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{5}}{\sqrt{2}} = \sqrt{5/2} = \sqrt{2.5}$

$= 1.58$

2. 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 requests per hour. Data structures may be measured on how many tables were used to arrange each data set. All the information was put together as follows.

Data size (gigabytes)	6	7	7	8	10	10	15
Number of tables	4	20	20	10	10	2	1
Processed requests	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.

⇒ Solution:

Here, Data size is a dependent variable as it depends on the number of tables and processed requests.

Let, y = data size (gigabytes)

x_1 = no. of tables

x_2 = processed requests

Data size (y)	No. of tables (x_1)	Processed requests (x_2)	x_1^2	x_2^2	x_1x_2	yx_1	yx_2
6	4	40	16	1600	160	24	240
7	20	55	400	3025	1100	140	385
7	20	50	400	2500	1000	140	350
8	10	41	100	1681	410	80	328
10	10	17	100	289	170	100	170
10	2	26	4	676	52	20	260
15	1	16	1	256	16	15	240
$\Sigma y = 63$	$\Sigma x_1 = 67$	$\Sigma x_2 = 245$	$\Sigma x_1^2 = 1021$	$\Sigma x_2^2 = 10027$	$\Sigma x_1x_2 = 2908$	$\Sigma yx_1 = 519$	$\Sigma yx_2 = 1973$

To fit $y = b_0 + b_1x_1 + b_2x_2$:

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

or, $63 = 7b_0 + 67b_1 + 245b_2$ ——— (I)

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

or, $519 = 67b_0 + 1021b_1 + 2908b_2$ ——— (II)

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

or, $1973 = 245b_0 + 2908b_1 + 10027b_2$ ——— (III)

Solving equations (I), (II) and (III) using Cramer's rule :

Coefficient of b_0	Coefficient of b_1	Coefficient of b_2	Constants
7	67	245	63
67	1021	2908	519
245	2908	10027	1973

Here,

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

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

$= 1640633$

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

$= 63(10237567 - 8456464) - 67(5204013 - 5737484) + 245(1509252 - 2014433)$

$= 24182701$

$$D_2 = \begin{vmatrix} 7 & 63 & 245 \\ 67 & 519 & 2908 \\ 245 & 1973 & 10027 \end{vmatrix}$$

$= 7(519 \times 10027 - 2908 \times 1973) - 63(67 \times 10027 - 2908 \times 245) + 245(67 \times 1973 - 519 \times 245)$

$= 60536$

$$D_3 = \begin{vmatrix} 7 & 67 & 63 \\ 67 & 1021 & 519 \\ 245 & 2908 & 1973 \end{vmatrix}$$

$= 7(1021 \times 1973 - 519 \times 2908) - 67(67 \times 1973 - 519 \times 245) + 63(67 \times 2908 - 1021 \times 245)$

$= -285612$

Now,

$$b_0 = \frac{D_1}{D} = \frac{24182701}{1640633} = 14.739$$

$$b_1 = \frac{D_2}{D} = \frac{60536}{1640633} = 0.0368$$

$$b_2 = \frac{D_3}{D} = \frac{-285612}{1640633} = -0.174$$

∴ The appropriate multiple regression model is:

$$y = 14.739 + 0.0368 b_1 - 0.174 b_2$$

Interpretations of regression coefficients:

$b_1 = 0.0368$ means datasize is increased by 0.0368 gigabytes when one table is used to arrange each data set holding the effect of processed requests constant.

$b_2 = -0.174$ mean data size is decreased by 0.174 gigabytes when one request is processed per hour holding the effect of number of tables used constant.

Group B

4. 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 standard deviation of 9.2. At the 1% level of significance, do these data provide considerable evidence that the mean number of concurrent users is greater than 35? Draw your conclusion based on your result.

⇒ Solution:

Sample size (n) = 100

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

Sample s.d. (s) = 9.2

Population mean (μ) = 35

- Problem to test

H_0: The mean number of concurrent users is ~~greater than~~ 35. ($\mu = 35$)

H_1: The mean number of concurrent users is greater than 35. ($\mu > 35$)

(one tail right)

- Test statistic

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{37.7 - 35}{\frac{9.2}{\sqrt{100}}} = 2.934$$

- Critical value

At $\alpha = 1\% = 0.01$ level of significance, the critical value for one tailed test is $Z_{tab} = Z_{\alpha} = 2.32$.

- Decision: $|Z| = 2.934 > Z_{tab} = 2.32$, so reject H_0 at 1% level of significance.
- Conclusion: The mean number of concurrent users is greater than 35.

5. 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 two lots?

⇒ Solution:

Sample size of lot A $(n_1) = 250$

No. of defective items in lot A $(x_1) = 10$

∴ Sample proportion of defective items in lot A is

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

Sample size of lot B $(n_2) = 300$

No. of defective items in lot B $(x_2) = 18$

∴ Sample proportion of defective items in lot B is

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

Let P_1 = population proportion of defective items in lot A

P_2 = population proportion of defective items in lot B

$$\therefore P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{10 + 18}{250 + 300} = \frac{28}{550} = 0.05$$

And $Q = 1 - P = 1 - 0.05 = 0.95$

- Problem to test:

H_0: There is no significant difference between the quality of two lots. $(P_1 = P_2)$

H_1: There is significant difference between the quality of two lots. $(P_1 \neq P_2)$ (both tailed)

- Test statistic

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.04 - 0.06}{\sqrt{0.05 \times 0.95\left(\frac{1}{250} + \frac{1}{300}\right)}}$$

$$= \frac{-0.02}{\sqrt{0.0003483}}$$

$$Z = -1.071$$

$\therefore |Z| = 1.071$

- Critical Value

At $\alpha = 5\% = 0.05$ level of significance, the critical value for ~~both~~ two tailed test is $Z_{tab} = Z_{\alpha/2} = 1.96$.

- Decision

Here, $|Z| = 1.071 < Z_{tab} = 1.96$, accept H_0 at 5% level of significance.

- Conclusion

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

11. Following are the scores obtained by 9 university staffs on the computer proficiency skills before training and after training. It was assumed that the proficiency of computer skills is expected to increased after training.

Staffs	Scores	
	Before training	After training
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 at 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.

⇒ Solution.

Given:

Sample size (n) = 9

Level of significance (α) = 5%.

Staffs	Scores		$d = X - Y$	d^2
	Before training (X)	After training (Y)		
1	50	55	-5	25
2	30	40	-10	100
3	15	30	-15	225
4	22	20	-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
			$\Sigma d = -75$	$\Sigma d^2 = 1015$

Here,

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

$$S_d = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}} = \sqrt{\frac{\Sigma d^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{1015 - 9 \times (-8.33)^2}{9-1}} = 6.986$$

- Problem to test

H_0 : The training is not effective ($\mu_d = 0$)

H_1 : The training is effective ($\mu_d \neq 0$) (Two tailed)

- Test statistic

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} = \frac{-8.33}{\frac{6.986}{\sqrt{9}}} = -3.577$$

- Critical value : At $\alpha = 5\%$ level of significance, the critical value for two tailed test is $t_{tab} = t_{\alpha/2\,(n-1)} = 2.306$.
- Decision : Here $|t| = 3.577 > t_{tab} = 2.306$, accept H_0 at 5% level of significance.
- Conclusion : The training is effective to improve the computer proficiency skills.

6. 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 be spam depends on whether it contains images or not. The following data were collected on n = 1000 random email messages.

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

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

⇒ Solution:

	Image containing status		
Spam status	With images (B)	No images ($\bar{B}$)	Total ($O_{i.}$)
Spam (A)	160 = a	240 = b	400 = a+b
No spam ($\bar{A}$)	140 = c	460 = d	600 = c+d
Total ($O_{.j}$)	300 = a+c	700 = b+d	1000 = N

- Problem to test

H_0: Being spam and containing images are independent.

H_1: Being spam and containing images are dependent.

• Test statistic

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} = \frac{1000(160\times460 - 240\times140)^2}{300\times700\times400\times600}$$

$$= 31.746$$

• Critical value

At $\alpha = 1\%$ level of significance, the critical value for 1 degree of freedom is $\chi^2_{0.01(1)} = 6.635$.

• Decision

$\chi^2 = 31.746 > \chi^2_{0.01(1)} = 6.635$, reject H_0 at 1% level of significance.

• Conclusion

Being spam and containing images are dependent factors.

7. Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as "Not satisfied", "Satisfied", "Good quality", and "Excellent quality", will recommend to others. The following counts were observed:

Computer maker	Not satisfied	Satisfied	Good quality	Excellent quality
A	20	240	70	20
B	10	30	40	20

Is there a significant difference in customer satisfaction of the computers produced by A and by B using Mann-Whitney U test at 5% level of significance.

=> Solution:

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

Here,

Sample size of A (n_1) = 4 } ≤ 10

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

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 4\times4 + \frac{4(4+1)}{2} - 20.5$$

$$= 5.5$$

$U_2 = n_1 n_2 - U_1 = 4\times4 - 5.5 = 10.5$

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

Let Md_1 and Md_2 be the median customer satisfaction for A and B respectively.

- Problem to test

H_0: There is no significant difference in ~~computer~~ customer satisfaction of both the computers. ($Md_1 = Md_2$)

H_1: There is significant difference in customer satisfaction of both the computers. ($Md_1 \neq Md_2$)

- Test statistic

$U_0 = 5.5$

- Critical value

At $\alpha = 5\%$ level of significance, the critical value is

$U_{tabulated} = U_{\alpha(n_1, n_2)} = 0$.

- Decision

$U_0 = 5.5 > U_{\alpha(n_1, n_2)} = 0$, accept H_0 at 5% level of significance.

- Conclusion

There is no significant difference in ~~computer~~ customer satisfaction of the computers produced by A and B.