PAGE No.
DATE

TE Sem-V (AI-DS) Statistics for AI-DS.

Q.1] Attempt any four :

a) Find the standard deviation of the average temperatures recorded over a five-day period last winter: 19, 21, 18, 24, 12 ?

Soln :→

Temp (x)	$x - \bar{x}$	$(x - \bar{x})^2$
19	0.2	0.04
21	2.2	4.84
18	-0.8	0.64
24	5.2	27.04
12	-6.8.	46.24
$\Sigma x = 94$		Σ 78.8

$$\text{Mean } (\bar{x}) = \frac{94}{5} = 18.8$$

$$\text{Variance } \sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

$$= \frac{78.8}{4}$$

$$= 19.7$$

$$\text{Standard Deviation } \sigma = \sqrt{19.7} \quad (\because \sqrt{\text{variance}})$$

$$= 4.4385 \approx 4.44$$

b) X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$ Find:

(i) $P(x < 40)$

(ii) $P(30 < x < 35)$?

Sol^n: → (i) $P(x < 40) = P(z < (40-30)/4)$
$= P(z < 2.5)$
$= 0.5 + P(0 < z < 2.5)$
$= 0.5 + 0.4938$
$= 0.9938.$

NOTE: Table used is area under curve standard normal table (z).

(ii) $P(30 < x < 35) = \cancel{P(30 < x)} -$
$P((30-30)/4) < z < P((35-30)/4)$
$= P(0 < z < 1.25)$
$= 0.3944$

c) Discuss Boot strapping vs re-sampling
Sol^n: → Theory.

d) The school principal wants to test if it is true what teachers say - that high school juniors use the computer an average 3.2 hours a day. What are our null and alternative hypothesis

Sol^n: → Null hypothesis (H_0): Use the computer an average 3.2 hours a day. $\mu = 3.2$

Alternative hypothesis (H_a): Does not use the computer an average 3.2 hours a day. $\mu \neq 30.$

e) What do you mean by correlation & regression? Explain with example
Sol^n: Theory.

Q.2 a]

Age (x)	Glucose level (Y)
43	99
21	65
25	79
42	75
57	87
59	81
$\Sigma x = 247$	$\Sigma y = 486$

$\bar{x} = 41.16$ $\bar{y} = 81$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
43	99	1.84	18	3.38	324	33.12 ~~1095.12~~
21	65	-20.16	-16	406.42	256	322.56
25	79	-16.16	-2	261.14	4	32.32
42	75	0.84	-6	0.70	36	-5.04
57	87	15.84	6	250.90	36	95.04
59	81	17.84	0	318.26	0	0
				$\Sigma(x-\bar{x})^2 =$	$\Sigma(y-\bar{y})^2 =$	$\Sigma(x-\bar{x})(y-\bar{y}) =$
				1240.8	656	478

$$\therefore S_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = \sqrt{\frac{1240.8}{5}} = 15.75$$

$$S_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = \sqrt{\frac{656}{5}} = 11.45$$

$$\therefore r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)S_x S_y}$$

$$= \frac{478}{5 \times 15.75 \times 11.45}$$

$$\boxed{r = 0.53}$$

Q.2 b]

Observation	A	B	C
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

coding data: Subtracting 30 throughout.

A x_1	B x_2	C x_3	x_1^2	x_2^2	x_3^2
-5	1	-6	25	1	36
0	9	0	0	81	0
6	8	-2	36	64	4
8	12	-5	64	144	25
1	5	-2	1	25	4
$\Sigma x_1 = 10$	$\Sigma x_2 = 35$	$\Sigma x_3 = -15$	$\Sigma x_1^2 =$ 126	$\Sigma x_2^2 =$ 315	$\Sigma x_3^2 =$ 69.

Sum of all samples $= \Sigma x_1 + \Sigma x_2 + \Sigma x_3$

$= 10 + 35 - 15$

$= 30$

$\therefore$ Correction factor $= T^2/N = 30^2/15 = 60$

Total sum of squares (SST) $= \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - T^2/N$

$= 126 + 315 + 69 - 60$

$= 450$

Sum of squares between samples (SSC) $= \Sigma x_1^2/N + \Sigma x_2^2/N + \Sigma x_3^2/N - T^2/N$

$= \frac{10^2}{5} + \frac{35^2}{5} + \frac{(-15)^2}{5} - 60$

$= 250$

(SSE)
∴ Sum of squares within sample = Total − SSC
= 450 − 250
= 200.

Source of variation	Sum of square	Degree of freedom	Mean square	F test
Between samples	250	2	125	$\frac{125}{16.67} = 7.49$
Within samples	200	12	16.67	
Total	450	14		

∴ $f = 7.49$

Table value for $v_1 = 2$ & $v_2 = 12$ $\alpha = 0.05$ is 3.89.

∵ calculated value > Table value.

The null hypothesis is rejected.

i.e. The difference in the mean of samples is significant.

Q.3.

a) Theory

b) Kruskal-Wallis Test :

Facility 1	Facility 2	Facility 3	Facility 4
88	77	71	52
82	76	56	65
86	84	64	68
87	59	51	81

Arranging data jointly according to size and assigning ranks, we get:

Values	Ranks	Ranks of Facility 1	Ranks of Facility 2	Ranks of Facility 3	Ranks of Facility 4
51	1	16	10	8	2
52	2	12	9	3	6
56	3	14	13	5	7
59	4	15	4	1	11
64	5				
65	6				
68	7				
71	8				
76	9				
77	10				
81	11				
82	12				
84	13				
86	14				
87	15				
88	16				
	Total →	57	36	17	26

classmate

Date

Page

$N = 16$

By Kruskal - Wallis test,

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{N_i} - 3(N+1)$$

$$H = \frac{12}{16(17)} \left[\frac{57^2}{4} + \frac{36^2}{4} + \frac{17^2}{4} + \frac{26^2}{4} \right] - 3(17)$$

$$= (0.04411)(1377.5) - 51$$

$$= 60.77 - 51$$

$$H = 9.77$$

While for a right tailed chi-square test with 95% confidence level and df=3, critical value χ^2 is 7.81

Null Hypothesis (H_0): The distribution of operator scores are same.

Alternative Hypothesis (H_a): The scores may vary in four facilities.

Therefore, calculated χ^2 value is greater than the critical value of χ^2 for a 0.05 significance level. χ^2 calculated > χ^2 critical hence reject the null hypothesis.

Page No.
Date

Q4.)
a) If the sample mean and expected mean value of the marks obtained by 15 students in a class test is 290 and 300 respectively. What is the t score if the standard deviation of the marks is 50?

Soln: given:

Population mean (μ) = 300
standard deviation (s) = 50
size of the sample (n) = 15
Sample Mean ($\bar{x}$) = 290

To find: t-score

Using the t-distribution formula

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{290 - 300}{\frac{50}{\sqrt{15}}} = \frac{-10}{12.909}$$

$\therefore$ $t = -0.7745$

Ans: T score of the marks is -0.7745

b) Find out what is the relation between the GPA of a class of students and the number of hours of study and the height of the student.

Soln:

For the given dataset,
Let GPA, Height and Study Hours be represented by variables y, x_1 and x_2 respectively.

Hence the relation between them can be given by the linear regression equation

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

The dependent variable in this regression is the GPA, and the independent variables are study hours and the height of the students

Now, we calculate b_0, b_1 and b_2

Step 1:

Calculate X_1^2, X_2^2, $X_1 y$, $X_2 y$ and $X_1 X_2$

	y	x_1	x_2
	2.9	66	7
	3.16	57	7
	3.62	64.5	6
	2	62	7
	3.45	69.5	8
	2.8	65	9
	3.63	63	6
	2.81	68	5
	3.33	59.5	4
	2.75	64	10
	3.86	69	7
Mean	3.12	~~707.5~~ 64.32	~~76~~ 6.91
Sum	34.31	~~64.32~~ 707.5	~~69~~ 76

	x_1^2	x_2^2	$x_1 y$	$x_2 y$	$x_1 x_2$
	4356	49	191.4	20.3	462
	3249	49	180.12	22.12	399
	4160.25	36	233.49	21.72	387
	3844	49	124	14	434
	4830.25	64	239.775	27.6	556
	4225	81	182	25.2	585
	3969	36	228.69	21.78	378
	4624	25	191.08	14.05	340
	3540.25	16	198.135	13.32	238
	4096	100	176	27.5	640
	4761	49	266.34	27.02	483
Sum	45,654.75	554	2211.03	234.7	4902

step 2: Calculate Regression Sums

$$\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n$$
$$= 45{,}654.75 - (707.5)^2 / 11$$
$$= 45{,}654.75 - 45{,}505.11$$
$$= 149.64$$

$$\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n$$
$$= 554 - (76)^2 / 11$$
$$= 554 - 525.10$$
$$= 28.9$$

$$\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n$$
$$= 2211.03 - (707.5 \times 34.31) / 11$$
$$= 2211.03 - 2206.76$$
$$= 4.27$$

$$\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n$$
$$= \overset{234.7}{\cancel{4902}} - (76 \times 34.31) / 11$$
$$= \cancel{4902}\ 234.7 - 237.05$$
$$= -2.35$$

$$\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n$$
$$= 4902 - (707.5 \times 76) / 11$$
$$= 4902 - 4888.18$$
$$= 13.82$$

step 3: Calculate b_0, b_1 and b_2

$$b_1 = \frac{\left[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 \cdot x_2)(\Sigma x_2 \cdot y)\right]}{\left[(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 \cdot x_2)^2\right]}$$

$$= \frac{(28.9)(4.27) - (13.82)(-2.35)}{(149.64)(28.9) - (13.82)^2}$$

$$\therefore b_1 = 0.038$$

$$b_2 = \frac{\left[(\Sigma x_1^2)(\Sigma x_2 \cdot y) - (\Sigma x_1 \cdot x_2)(\Sigma x_1 \cdot y)\right]}{\left[(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 \cdot x_2)^2\right]}$$

$$b_2 = \frac{\left[(149.64)(-2.35) - (13.82)(4.27)\right]}{\left[(149.64)(28.9) - (13.82)^2\right]}$$

$$\therefore b_2 = -0.1$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$= 3.12 - (0.038 \times 64.32) - (-0.1 \times 6.91)$$

$$b_0 = 1.37$$

$\therefore$ The linear equation will be

$$y = 1.37 + 0.038x_1 - 0.1x_2$$

Hence the relation between GPA, Height and the study hours is given by the above linear equation.

Q.5a) $H_0 : \mu_0 \leq 145$

$H_a : \mu_0 > 145.$

Given : $\mu = 145$

$s = 100$

$n = 144.$

$x = 147$

$\therefore$ we have,

$$z = \frac{x - \mu}{s/\sqrt{n}}$$

$$= \frac{147 - 145}{100/\sqrt{144}}$$

$$= \frac{2 \times \sqrt{144}}{100}$$

$$= \frac{24}{100}$$

$$= 0.24$$

The table value for $\alpha = 0.05$ is 1.64.

$\because$ The calculated value < table value.

i.e $0.24 < 1.64$

we accept the null hypothesis.

Q.5]

b) Find the simple linear regression equation that fits the given data and coefficient of determination:

Solⁿ :→

Hour(x)	x^2	$a = x-\bar{x}$	a^2 $(x-\bar{x})^2$	Temp(y)	y^2	$b = y-\bar{y}$	b^2 $(y-\bar{y})^2$	$x \times y$	$a \times b$
2	4	-5	25	21	441	-35.8	1281.64	42	179
4	16	-3	9	27	729	-29.8	888.04	108	89.4
6	36	-1	1	29	841	-27.8	772.84	174	27.8
8	64	1	1	86	7396	29.2	852.64	688	29.2
10	100	3	9	86	7396	29.2	852.64	860	87.6
12	144	5	25	92	8464	35.2	1239.04	1104	176
$\Sigma 42$	$\Sigma 364$		$\Sigma 70$	$\Sigma 341$	$\Sigma 25267$		$\Sigma 5886.84$	$\Sigma 2976$	$\Sigma 589$

$$\text{Mean}(\bar{x}) = \frac{42}{6} = 7 \qquad \text{Mean}(\bar{y}) = \frac{341}{6} = 56.8$$

Y on x

$$y - \bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x - \bar{x})$$

$$y - 56.8 = \frac{589}{70}(x - 7)$$

$$y - 56.8 = 8.414x - 58.898$$

$$y = 8.414x - 2.098$$

$$\therefore y = -2.098 + 8.414x.$$

x on y.

$$x - \bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y - \bar{y})$$

$$x - 7 = \frac{589}{5886.84}(y - 56.8)$$

$$x - 7 = 0.1y - 5.68$$

$$x = 0.1y + 1.32$$

$$\therefore x = 1.32 + 0.1y.$$

Corelation coefficient

$$(r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{6 \times 2976 - (42) \times (341)}{\sqrt{[6 \times 364 - (42)^2][6 \times 25267 - (341)^2]}}$$

$$r = \frac{17856 - 14322}{\sqrt{(2184 - 1764)(151602 - 116281)}}$$

$$r = \frac{3534}{\sqrt{(420)(35321}} = \frac{3534}{\sqrt{14834820}}$$

$$r = \frac{3534}{3851.6} = 0.918$$

$\therefore r^2$ (coefficient of determination) $= (0.918)^2 = \underline{\underline{0.842}}$

classmate

Date
Page

Q 6.(a)

→ The number of people still living has a Binomial $(n=5,\ p=2/3)$.

a) All five people are still living.

~~The~~ $P(X=5) = \binom{5}{5}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^0 = \left(\frac{2}{3}\right)^5 \approx 0.1317$

(b) At least three people are still living.

$P(X\geq 3) = P(X=3) + P(X=4) + P(X=5)$

$= \binom{5}{3}\left(\frac{2}{3}\right)^3(1/3)^2 + \binom{5}{4}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1 + \binom{5}{5}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^0$

$= \frac{5!}{3!\,2!}\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^2 + \frac{5!}{4!\,1!}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1 + \left(\frac{2}{3}\right)^5$

$\approx 0.3292 + 0.3292 + 0.1317$

≈ 0.7901

(c) Exactly two people are still living.

$P(X=2) = \binom{5}{2}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3 = \frac{5!}{2!\,3!}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3 \approx 0.1646$