Avani . Paresh . Khatri
Roll no- 40
Sec - A .

PAGE No
DATE / /

PROBABILITY & STASTISTICS - 2.

ASSIGNMENT - 2.

Q.1.
a)

	x	y	xy	x^2
1	40	56	2240	1600
2	10	62	620	100
3	100	195	19,500	10000
4	110	240	26,400	12,100
5	120	170	20,400	14,400
6	150	270	40,500	22,500
7	20	48	960	400
8	90	196	17,640	8100
9	80	214	17,120	6400
10	130	286	37,180	16,900
Total	850	1737	182,560	92,500

$$\therefore S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$= 182560 - \frac{850 \times 1737}{10}$$

$$= 34915$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$= 92500 - \frac{850^2}{10}$$

$$= 20250$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = 1.72$$

$$\hat{a} = \bar{y} - b\bar{x}$$
$$= \left(\frac{1737}{10}\right) - 1.72 \times \frac{850}{10}$$
$$= 27.14.$$

$$\underline{y = 27.14 + 1.72x.}$$

b) Gradient represents the amount of hours per rupee spent.

Q.2. i) Sol^n:

Fitted Linear Regression Equation:

$\Sigma x = 385.2$, $\Sigma x^2 = 12{,}666.58$
$\Sigma y = 1162.5$, $\Sigma y^2 = 119{,}026.9$
$\Sigma xy = 38{,}191.41$, $n = 12.$

$$S_{xx} = \Sigma x^2 - n\bar{x}^2$$
$$= 12666.58 - 12 \times \left(\frac{385.2}{12}\right)^2$$
$$= 301.66$$

$$S_{xy} = \Sigma xy - n\bar{x}\bar{y}$$
$$= 38191.41 - 12 \times \left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right)$$
$$= 875.16$$

$$S_{yy} = \Sigma y^2 - n\bar{y}^2$$
$$= 119026.90 - 12 \times \left(\frac{1162.5}{12}\right)^2$$
$$= 6409.71$$

The coefficients of regression equation are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.90$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

$$= \left(\frac{1162.5}{12}\right) - 2.90 \times \left(\frac{385.2}{12}\right)$$

$$= 3.78$$

$\therefore$ The fitted regression line is

$$y = \hat{\alpha} + \hat{\beta} x = 3.78 + 2.90x.$$

ii) Confidence interval for θ

Assuming normal errors with a constant variance:

95% confidence interval for

$$\beta : \hat{\beta} \pm t_{n-2}(2.50\%) \times s.e(\hat{\beta}).$$

Here,

$$s.e.(\hat{\beta}) = \sqrt{\frac{\sigma^2}{S_{xx}}}$$

$$\sigma^2 = \frac{1}{n-2}\left[S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right]$$

$$= 387.07.$$

$$\therefore s.e.(\hat{\beta}) = \sqrt{\frac{387.07}{301.66}} = 1.13$$

95% confidence intervals for β

$$: 2.90 \pm 2.228 \times 1.13$$

$$= (0.38, 5.42)$$

PAGE No.
DATE / /

iii) 95% confidence intervals for the mean IBM share price:

$$\hat{y}_{x_0} \pm t_{n-2}(2.50\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

The Dell share price is US \$40 (x_0)

$$\hat{y}_{x_0} = 3.78 + 2.90 \times 40$$
$$= 119.78$$

Thus, 95% confidence interval:

$$= 119.78 \pm 2.228 \times \sqrt{387.07 \times \left[\frac{1}{12} + \frac{(40-32.1)^2}{301.66}\right]}$$

$$= 119.78 \pm 2.28 \times 10.5989$$

$$= (96.17, 143.39)$$

Q.3. i) Comments on the plot.

→ The centers of the distribution differ for all the far cities. Thus there is a prima facie case for suggesting that the underlying means are different

→ The difference between the mean time taken to commute to office in peak hours and non-peak hours are in the order City A (highest), City D (low)

→ The variation in the data for City C is lowest compared to city D which appears to be highest. However, with only 7 observations for each city, we cannot be sure that there is a real underlying difference in variance.

PAGE No
DATE / /

ii) Following are the assumptions underlying anaylsis of variance:

→ The populations must be normal.

→ The populations have a common variance

→ The observations are independent.

iii) We are carrying out the following test:

H_0 : The mean of differences is same for each city.

against H_1 : The mean of difference are not the same for all of the cities.

To carry out the ANOVA, we must first compute the sum of squares.

$$SS_T = 1495 - \frac{103^2}{28} = 1116.11$$

$$SS_B = \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right)$$

$$= 516.11.$$

$$SS_R = SS_T - SS_B$$

$$= 600.00.$$

The ANOVA table is:

Sorce	df	SS	MS	F
Treatments	3	516.11	172.04	6.88.
Residuals	24	600.00	25.00	
Total	27	1116.11		

Under Ho, $F = 172.04/25 = 6.88$, using $F_{3,24}$ distribution

The 5% critical point is 3.009, so we have sufficient evidence to reject H_0 at the 5% level.

PAGE No.
DATE / /

Therefore, it is reasonable to conclude that there are underlying differences between the cities.

iv) <u>Analysis of the mean differences</u>

Since, $\bar{y}_1 = 9.14$, $\bar{y}_2 = 6.29$, $\bar{y}_3 = 1.14$, $\bar{y}_4 = 1.86$
we can write

$$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$$

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = 25.$$

The least significant difference between any pair of means is :

$$t_{24,0.025} \times \hat{\sigma}\sqrt{\left(\frac{1}{7}+\frac{1}{7}\right)}$$

$$= 2.064 \times \sqrt{25} \times \sqrt{\left(\frac{1}{7}+\frac{1}{7}\right)}$$

$$= 5.52$$

Now, we can examine the difference between each of the pairs of means. If the difference is less than the least significant difference then there is no significant difference between the means.
We have,

$$\bar{y}_1 - \bar{y}_2 = 2.85$$
$$\bar{y}_2 - \bar{y}_3 = 5.15$$
$$\bar{y}_3 - \bar{y}_4 = 3.00$$

Observing that all these 3 differences are less than 5.52, we underline these pairs to show that they have no significant difference:

PAGE No.
DATE / /

$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$

Examining to see if the first two groups can be combined:

$\bar{y}_1 - \bar{y}_3 = 8.00$

There is significant between means 1 and 3, so we cannot combine the first two groups.

Examining to see if the ~~firs~~ last two groups can be combined

$\bar{y}_2 - \bar{y}_4 = 8.15$.

There is significant between means 2 and 4, we cannot ~~beti~~ combine the last two groups.

$\therefore$ The diagram ~~rea~~ remains as before.

Q.4. a) The mathematical model of one-way ANOVA is given by,

$$y_{ij} = \mu + \tau_i + e_{ij} \quad ; i = 1, 2, \ldots, k.$$

where, $\quad j = 1, 2, \ldots n_i$.

k = number of treatments.

n_i = number of responses from i^{th} treatment

y_{ij} is the j^{th} response from i^{th} treatment.

τ_i is the i^{th} treatment.

μ is the over mean response.

e_{ij} is the error term.

Assumption : e_{ij} is i.i.d $N(0, \sigma^2)$

b) H_0 : Mean claim amounts of five companies are equal.

H_1 : Mean claim amounts of five companies are not equal.

We have, $n_1 = 7$, $n_2 = 8$, $n_3 = 6$, $n_4 = 7$, $n_5 = 5$, $n = 33$

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_{ij}^2 = 174,316$$

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_{ij} = 1856$$

$$C \cdot F = \left[\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_{ij} \right]^2 \Big/ n.$$

$$= 104,385.94$$

$$SS_T = 174316 - C \cdot F$$

$$= 69,930.06$$

$$SS_B = \sum_{i=1}^{5} \left[\sum_{j=1}^{n_i} y_{ij} \right]^2 \Big/ n_i - C \cdot F$$

$$= \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{384^2}{5} - 104,38.94$$

$$= 22,325.69$$

$$SS_R = SS_T - SS_B$$

$$= \cancel{44}\ 47,604.37$$

Sources of variation	d.f	SS	MSS	F
Companies	4	22,326	5581.42	~~33~~ 3.283
Residual	28	47,604	1700.16	
Total	32	69,930		

$F_{observed} = 3.283$

$F_{(4,\ 0.05)} = 2.714$

<u>Reject H_0</u>

c) H_0 : Salaries are independent of number of actuarial papers cleared.
H_1 : Salaries are dependent on number of actuarial papers cleared.

Observed Values (O_i)

Papers cleared	Salary per annum (in Rs. lacs) 3-5	5-8	8-10	10-12	Total.
0-3	45	20	6	5	76
4-6	7	20	9	6	42
7-9	5	8	15	12	40
Total.	57	48	30	23	158.

Under H_0, Expected values (E_i).

Papers cleared	Salary per annum (in Rs. lacs) 3-5	5-8	8-10	10-12	Total
0-3	27.42	23.09	14.43	11.06	76.00
4-6	15.15	12.76	7.97	6.11	42.00
7-9	14.43	12.15	7.59	5.82	40.00
Total	57.00	48.00	30.00	23.00	158.00

$$\chi^2 = \sum_{i=1}^{12} \frac{(O_i - E_i)^2}{E_i} = \frac{(27.42-45)^2}{27.42} + \dots + \frac{(5.82-12)^2}{5.82}$$

$$= 49.916$$

$\chi^2_{observed} = 49.91$; $\chi^2_{(6, 0.05)} = 12.59$, where $df = (3-1)(4-1) = 6$

Reject H_0.

PAGE No
DATE / /

Q.5. i) The assumptions required for one-way analysis of variance (ANOVA) are:

→ The population must be normal.
→ The population have a common variance.
→ The observations are independent.

The sample variance observed for the four rates appear very different from each other. Thus, we can clearly see that the assumptions that the underlying populations have a common variance assumption will not hold for the data as they are.

ii) For the transformation $x \to \sqrt{x}$, the value of sample mean for rate 1 will be:

$$\frac{1}{3}\left(\sqrt{29} + \sqrt{13} + \sqrt{21}\right) = 4.52$$

For the transformation $x \to \log_e x$, the value of sample variance for rate 2 will be

$$\frac{1}{2}\left[(\log_e 180 - 4.827)^2 + (\log_e 90 - 4.827)^2 + (\log_e 120 - 4.827)^2\right] = \underline{0.1213}.$$

iii) The scientist was correct in asserting that the $\log_e$ transformation must be done before carrying out a ~~+~~ one-way (ANOVA) as for this transformation it can be claimed that the assumption of common variance for the underlying population holds.

To justify this, a quick access check can be done on the ratio of maximum to minimum sample variance among the four rates data. A smaller ratio and close to 1 would indicate that the variances are close enough which in turn applies that the assumptions of common variance for the underlying position holds

Variance	x	$\sqrt{x}$	$\log_e(x)$	$1/x$
Min	64	0.79	0.1213	0.0000004
Max	63,300	23.96	0.1861	0.0004721
Ratio	98906	30.17	1.53	1,268.14.

Clearly, the transformation $\log_e x$ produces minimum ratio of maximum to minimum observed sample variance and that too close to 1.

iv) We will perform an ANOVA on the $\log_e x$ data. We would assume the following model:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \quad i = 1, 2, 3, 4 \; ; \; j = 1, 2, 3.$$

Here,

→ Y_{ij} is the $\log_e$ transformed value of the j^{th} observation of the number of germinations per square foot observed when the i^{th} rate was applied.

→ μ is the overall population mean.

→ τ_i is the deviation of the i^{th} rate mean such that $\sum \tau_i = 0$

→ e_{ij} are the independent error terms which follows Normal distribution with mean 0 and common unknown variance σ^2.

We have already argued that we can assume equal underlying variances for this transformation. So, all requisite assumptions hold here.

For ANOVA, the null hypothesis being tested here is:

$H_0 : \tau_i = 0, \; i = 1, 2, 3, 4$ against

$H_1 : \tau_i \neq 0$ for at least one i.

PAGE No
DATE / /

Rate	$\log_e(x)$ Mean	Variance	Y_L^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.716	0.186	294.053
4	6.575	0.148	389.060
	20.110	0.618	973.373

Here, $Y_L^2 = \left(\Sigma_{j=1}^{3} Y_{ij}\right)^2 = (3 \times \text{Mean}_i)^2$

Now,

$$SS_{(Rate)} = \Sigma_{i=1}^{4} \frac{Y_L^2}{3} - \frac{Y^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.100)^2}{12}$$

$$= 21.153.$$

$$SS_{(Residuals)} = \Sigma_{i=1}^{4} \left[\Sigma_{j=1}^{3} (Y_{ij} - \bar{Y}_L)^2 \right]$$

$$= \Sigma_{i=1}^{4} [2 \times \text{Variance}_i]$$

$$= 2 \times 0.618$$

$$= 1.236.$$

The ANOVA table is as follows:

Sources of Variation	d.f	Sum of Squares	Mean Squares	F
Rate	3	21.153	7.051	45.638
Residuals	8	1.236	0.155	
Total	11	22.389		

The 1% critical value for F (3,8) distribution is 7.951
Given that observed F statistic value is much larger than this, we can state the p-value for this test is almost near to 0 or in other words there is

overwhelming evidence against the null Hypothesis H_0. Thus it can be concluded that the underlying means are not equal.

v) A 95% confidence interval for the i^{th} rate mean (in the chosen transformed scale) is given by:

$$\bar{Y}_i \pm t_{8;0.975} \times \frac{\hat{\sigma}}{\sqrt{3}} \quad \text{for } i = 1, 2, 3, 4.$$

Here: $t_{8,0.975}$ is the 97.5% critical point of a t-distribution with 8 degrees of freedom and equals 2.306.

The common error variance σ^2 is estimated as:

$$\hat{\sigma}^2 = \frac{\text{Residual Sum of Squares}}{8} = 0.155.$$

The 95% confidence intervals (in transformed and original scale after applying the transformation $CI \rightarrow e^{CI}$) are tabulated in the following table.

Rate	Observed Mean	95% Confidence Interval (Transformed Scale)		95% Confidence Interval (Original Scale)	
		LCI	UCI	LCI	UCI
1	2.992	2.469	3.516	11810	33.635
2	4.827	4.303	5.350	73.954	210.623
3	5.716	5.193	6.239	179.950	512.505
4	6.575	6.052	7.098	424.770	1209.763

PAGE No.	
DATE	/ /

Q.6. i) Sol^n:

The relevant summary to compute correlation coefficient are:

$$S_{xx} = \sum x^2 - n\bar{x}^2$$

$$= 207 - 10 \times \left(\frac{39}{10}\right)^2$$

$$= 54.90.$$

$$S_{xy} = \sum xy - n\bar{x}\bar{y}$$

$$= 2853 - 10\left(\frac{39}{10}\right)\left(\frac{562}{10}\right).$$

$$= 661.20.$$

$$S_{yy} = \sum y^2 - n\bar{y}^2$$

$$= 40508 - 10 \times \left(\frac{562}{10}\right)^2.$$

$$= 8923.60$$

Correlation coefficient, $r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$

$$r = \frac{661.20}{\sqrt{54.90}\sqrt{8923.60}}$$

$$\underline{r = 0.945}.$$

ii) Fitted linear Regression Equation:

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} - \bar{x} = \left(\frac{562}{10}\right) - 12.04 \times \left(\frac{39}{10}\right).$$

$$= 9.23$$

$\therefore$ The fitted regression line is $y = \hat{\alpha} + \hat{\beta}x$

$$\underline{= 9.23 + 12.04x}$$

PAGE No.	
DATE	/ /

iii) Relation : $SS_{TOT} = SS_{REG} + SS_{RES}$

$$SS_{TOT} = S_{yy} = 8923.60$$

$$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$= 8923.60 - \frac{(661.20)^2}{54.90}$$

$$= 960.30.$$

$$SS_{REG} = SS_{TOT} - SS_{RES}$$

$$= 8923.60 - 960.30$$

$$= 7963.30.$$

iv) Coefficient of Determination :

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SS_{REG}}{SS_{TOT}} = \frac{7963.30}{8923.60} = 0.8924$$

For the simple linear regression model, the value of the coefficient of determination is the square of the correlation coefficient for the data, since,

$$0.945 = r = \frac{S_{xy}}{(S_{xx} \times S_{yy})^{0.5}} = \sqrt{R^2}$$

$$= \sqrt{0.8924}.$$

Q.7 i)

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 4789.42$$

$$S_{xy} = 2176.84$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2176.84}{4789.42} = 0.4545$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 17.895 - 0.4545 \times 15.83$$

$$= 10.79$$

$\therefore$ The fitted regression equation is $\hat{y} = 10.79 + 0.4545x$

iv) The estimated value of y corresponding to $x^3 = 25$ is $10.79 + 0.4545 \cdot 25$

$$= 22.15$$

$$\hat{\sigma}^2 = 22.20$$

The variance of the estimator of the mean response is given by

$$\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{11} + \frac{84.0889}{4789.42}\right] \times 22.20$$

$$= 2.41.$$

The variance of the estimator of the individual response is given by

$$\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[1 + 0.1085\right] \times 22.20$$

$$= 24.61.$$

Using t_9 distribution, the 95% confidence intervals for mean and individual responses are:

$22.20 \pm 2.262 \times \text{sqrt}(2.41)$ and

$22.20 \pm 2.262 \times \text{sqrt}(24.61)$

$= (18.7, 25.7)$ & $(11.0, 33.4)$

v) The residual plot shows a definite pattern. Although the correlation coefficient is high, the model does not seem to be appropriate.

Using this model leads to underestimation of premium rates at low and high mortality ratings.

Q.8 i) Factor Analysis / Principal Component Analysis is -
A method ~~of~~ for reducing the dimensionality of data
It seeks to identify key components necessary to model and understand data.
Original variables may be
→ correlated with each other.
While newly identified principal components are chosen to be
→ uncorrelated
→ linear combinations of the original variables of the data.
→ which maximise the variance.

ii)

Principal Component	Diagonal entry (PC_i)	PC_i / (Sum(PC_i) over 1 to 5)	
PC1	0.456	65.0 %	of total variance explain by PC 1
PC2	0.137	19.5 %	of total variance explained by PC2
PC3	0.08	11.4 %	of total variance explained by PC3
PC4	0.0165	2.4 %	of total variance explained by PC4
PC5	0.012	1.7 %	of total variance explained by PC5
	0.?015	100.0 %	

iii) As 1^{st} 3 Principal Components explain over 95% of total variance, dimensionally can be reduced to 3 for this dataset.

The 1st 3 Principal components can then be used for building further classification or regression modelling purpose.

Q.9. i) The scatter plot suggests an inverse relation between marks obtained and hours spent on social media per day.

ii) $S_{xx} = 277.5 - \frac{45^2}{10} = 75.$

$S_{yy} = 43,956 - \frac{644^2}{10} = 2\cancel{8}240.5.$

$S_{xy} = 2602 - \frac{644 \times 45}{10} = -296.$

$r = \frac{S_{xy}}{\sqrt{(S_{xx} \cdot S_{yy})}} = -0.686.$

−69% correlation coefficient also implied a moderate negative linear relation between the two variables as visible from the scatterplot.

iii) Null hypothesis $H_0 : \rho = 0$ against $H1 : \rho < 0$
Need to assume that data come from a bivariate normal distribution.
from Table Book, $r = 0.5 \times \ln\left(\frac{1-0.686}{1.686}\right) = -0.8404$

And under H_0, this should be a value from a ~~bivariate~~ $N(0, 1/7)$ distribution.
Fisher's standardized statistic $= \frac{(-0.8404 - 0)}{\sqrt{1/7}} = -2.22$

This gives the p-value = $P(z < -2.22)$ $= 0.013$ which is quite small and hence shows a strong evidence to reject the null hypothesis with 95% confidence. We can conclude that marks obtained and hours spent on social media are negatively correlated.

iv) $\text{Beta} = \frac{S_{xy}}{S_{xx}} = \frac{-2.96}{75} = -3.9467.$

$\text{Alpha} = \text{mean of } y - \text{beta} \times \text{mean of } x$

$= \frac{644}{10} + 3.9467 \times \frac{45}{10}$

$= 82.16.$

Fitted line is $y = 82.16 - 3.9476x$

v) $R^2 = -0.686^2 = 0.4706.$

This gives the proportion of total variation explained by the model.

vi) For every additional hour spent on social media per day, the total marks reduce by 3.95 (~4 marks) basis the fitted equation.

Q.10. H_0 : There is no difference among industries.

H_1 : At least one industry differs significantly from the overall mean.

$SS_R = 19(5^2 + 10^2 + 8^2) = 3591.$

Mean of regression $= \frac{(27+36+30)}{3} = 31.$

$$SS_B = 20\left[(27-31)^2 + (36-31)^2 + (30-31)^2\right]$$
$$= 840.$$

$$F_{2,57} = \left(\frac{840}{2}\right) \Big/ \left(\frac{3591}{57}\right)$$
$$= 6.667.$$

The 1% point from $F_{2,60}$ is 4.977 and since the test statistic is higher than this, the null hypothesis is rejected. We conclude that registration rate is different across different industries.