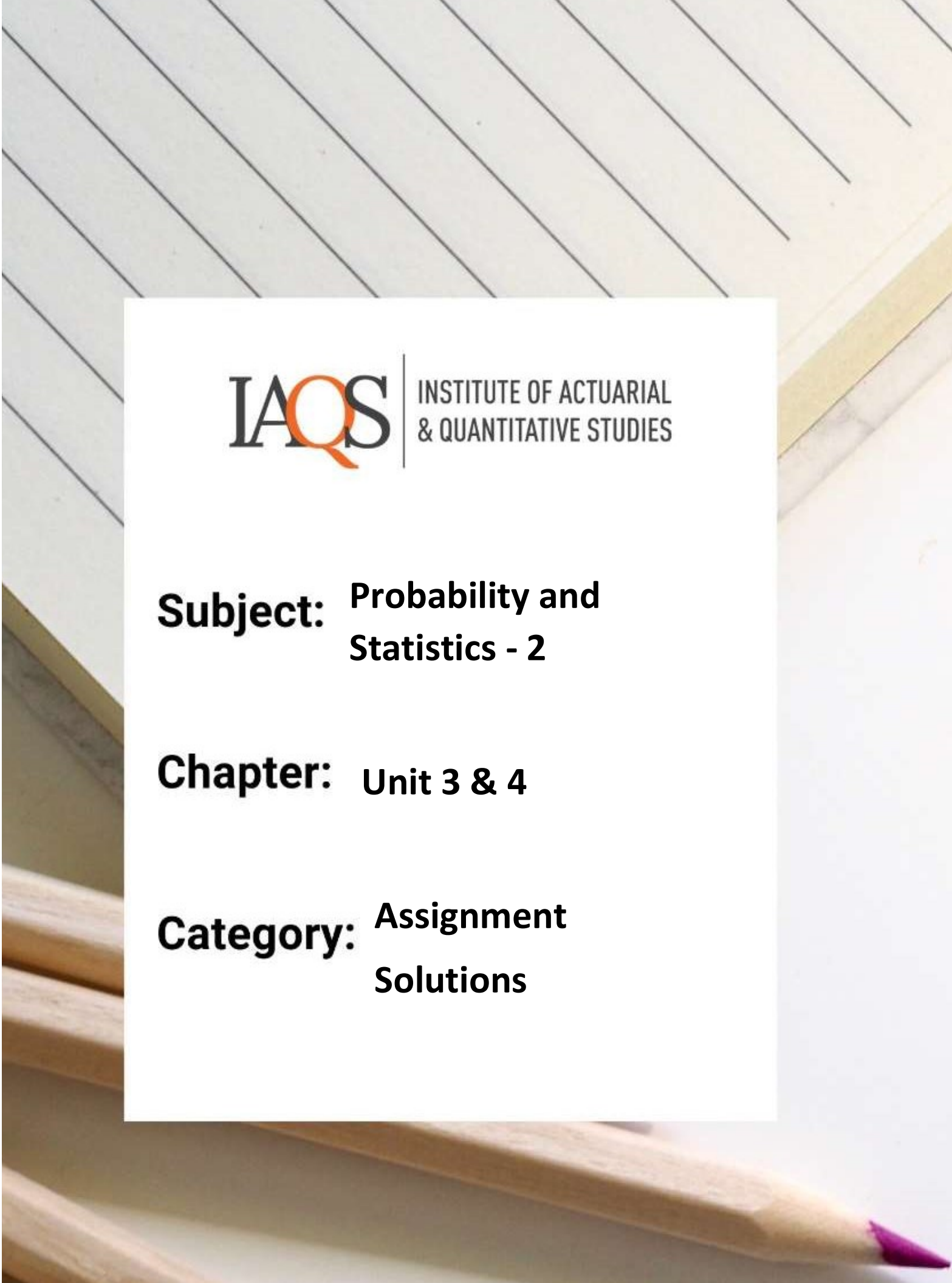


IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability and Statistics - 2

Chapter: Unit 3 & 4

Category: Assignment Solutions

1.

	1	2	3	4	5	6	7	8	9	10	Total
x	40	10	100	110	120	150	20	90	80	130	**850**
y	56	62	195	240	170	270	48	196	214	286	**1,737**
xy	2,240	620	19,500	26,400	20,400	40,500	960	17,640	17,120	37,180	**182,560**
x^2	1,600	100	10,000	12,100	14,400	22,500	400	8,100	6,400	16,900	**92,500**

$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i / n = 182560 - 850 * 1737/10 = 34915$

$S_{xx} = \sum x_i^2 - (\sum x_i)^2 / n = 92500 - 850^2/10 = 20250$

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

$\hat{a} = \bar{y} - b\bar{x} = (1737/10) - 1.72 * (850/10) = 27.14$

$y = 27.14 + 1.72x$

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

2.

i. Fitted Linear Regression Equation

The relevant summary statistics to fit the equation are:

$\sum x$ = 385.2; $\sum x^2$ = 12,666.58;
$\sum y$ = 1,162.5; $\sum y^2$ = 119,026.9;
$\sum xy$ = 38,191.41; n = 12.

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 12666.58 - 12 * \left(\frac{385.2}{12}\right)^2 = 301.66$$
$$S_{xy} = \sum xy - n\overline{xy} = 38191.41 - 12 * \left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right) = 875.16$$
$$S_{yy} = \sum y^2 - n\bar{y}^2 = 119026.90 - 12 * \left(\frac{1162.5}{12}\right)^2 = 6409.71$$

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.90$$
$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{1162.5}{12}\right) - 2.90 * \left(\frac{385.2}{12}\right) = 3.78$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta}x = 3.78 + 2.90\,x}$

ii. Confidence interval for β

Assuming normal errors with a constant variance:

95% confidence interval for β: $\hat{\beta} \pm t_{n-2}(2.50\%) * s.e.(\hat{\beta})$

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

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[S_{yy} - \frac{S_{xy}{}^2}{S_{xx}}\right] = 387.07$$

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

95% confidence interval for β: $2.90 \pm 2.228 * 1.13 = (0.38, 5.42)$

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).

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

Thus, 95% Confidence interval:

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

$$= 119.78 \pm 2.228 * 10.5989$$

$$= (96.17, 143.39)$$

3.
i) Comments on the plot
The centers of the distributions differ for all the four 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 nonpeak hours are in the order City A (highest), City D (lowest).

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.

ii. Following are the assumptions underlying analysis 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:
H0: The mean of differences is same for each city
against
H1: The mean of differences are not the same for all of the cities

To carry out the ANOVA, we must first compute the Sum of Squares

$$SS_T = 1{,}495 - \frac{103^2}{28} = 1{,}116.11$$

$$SS_B = \frac{1}{7}(64^2 + 44^2 + 8^2 + (-13)^2) - \frac{103^2}{28} = 516.11$$

$$SS_R = SS_T - SS_B = 600.00$$

The ANOVA table is:

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600.00	25.00	
Total	27	1,116.11		

Under $H_0, F = \frac{172.04}{25.00} = 6.88$, using the $F_{3,24}$ distribution.

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

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

iv. Analysis of the mean differences

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:

$$t_{24,0.025} * \hat{\sigma} \sqrt{\left(\frac{1}{7} + \frac{1}{7}\right)} = 2.064 * \sqrt{25} * \sqrt{\left(\frac{1}{7} + \frac{1}{7}\right)} = 5.52$$

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:

$$\underline{\bar{y}_{1*} > \bar{y}_{2*}} > \underline{\bar{y}_{3*} > \bar{y}_{4*}}$$

Examining to see if the first two groups can be combined

$$\bar{y}_{1*} - \bar{y}_{3*} = 8.00$$

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

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

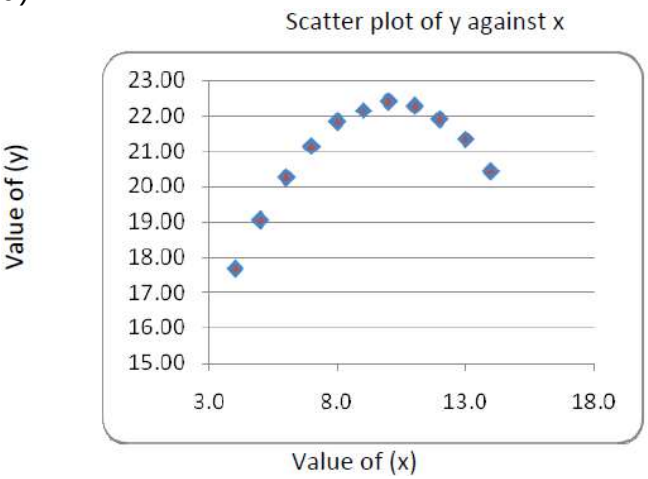
$$\bar{y}_{2*} - \bar{y}_{4*} = 8.15$$

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

Therefore the diagram remains as before.

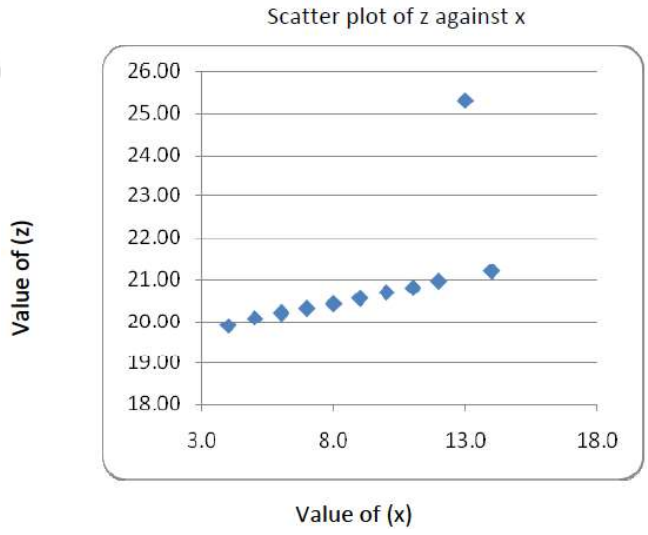
4.

a)



From the scatter plot we see that the line of regression may not be a best fit, but a polynomial regression equation may fit.

b)



From the scatter plot we see that the line of regression may be a best fit. Only a point seems to be an outlier.

c)

$\bar{x} = 9.00$; $\bar{y} = 20.95$; $n = 11$;

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 1{,}001.00 - 11 \times (9.00)^2 = 110.00$$

$$S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = 4{,}850.30 - 11 \times (20.95)^2 = 22.37$$

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\,\bar{y} = 2{,}104.85 - 11 \times (9.00 \times 20.95) = 30.80$$

$$\hat{\beta}_{xy} = \frac{S_{xy}}{S_{xx}} = \frac{30.80}{110.00} = 0.28$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}_{xy}\bar{x} = 20.95 - 0.28 \times 9.00 = 18.43$$

The least squares fitted regression line of y on x = 18.43 + 0.28 *x*

d)

$\sum z = 230.45$; $\bar{z} = 20.95$; $\sum z^2 = 4{,}850.30$; $\sum xz = 2{,}104.85$

$$S_{zz} = \sum(z_i - \bar{z})^2 = \sum z_i^2 - n\bar{z}^2 = 4{,}850.30 - 11 \times (20.95)^2 = 22.37$$

$$S_{xz} = \sum(x_i - \bar{x})(z_i - \bar{z}) = \sum x_i z_i - n\bar{x}\,\bar{z} = 2{,}104.85 - 11 \times (9.00 \times 20.95) = 30.80$$

So the least square fitted regression line of z on x is same as obtained in (c) above = 18.43 + 0.28 *x*

e)

Correlation coefficient of x and y $(\rho_{xy}) = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{30.80}{\sqrt{110.00 \times 22.37}} = 0.621$

Also $(\rho_{xz}) = \frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}} = \frac{30.80}{\sqrt{110.00 \times 22.37}} = 0.621$

Though the relationship of y on x and z on x are different, as found in the above scatter plots, the correlation coefficient are same and positive.

f) i)

After omitting the values for city 3, we calculate,

$\bar{x} = 8.60$; $\bar{z} = 20.51$; $n = 10$; $\sum x^2 = 832.00$; $\sum z^2 = 4,209.71$; $\sum xz = 1,775.62$

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 832.00 - 10 \times (8.60)^2 = 92.40$$

$$S_{zz} = \sum(z_i - \bar{z})^2 = \sum z_i^2 - n\bar{z}^2 = 4,209.71 - 10 \times (20.51)^2 = 1.46$$

$$S_{xz} = \sum(x_i - \bar{x})(z_i - \bar{z}) = \sum x_i z_i - n\bar{x}\bar{z} = 1,775.62 - 10 \times (8.60 \times 20.51) = 11.62$$

$$\hat{\beta}_{xz} = \frac{S_{xz}}{S_{xx}} = \frac{11.62}{92.40} = 0.126$$

$$\hat{\alpha} = \bar{z} - \hat{\beta}_{xz}\bar{x} = 20.51 - 0.126 \times 8.60 = 19.43$$

The regression line of z on x = 19.43 + 0.126 *x*

ii)

$$(\rho_{xz}) = \frac{S_{xz}}{\sqrt{S_{xx} S_{zz}}} = \frac{11.62}{\sqrt{92.40 \times 1.46}} = 0.999$$

g)

After omitting the outlier value for city 3, the correlation coefficient is approximately equal to1, i.e. the minimum temperatures on a certain day is perfectly correlated with the maximum temperatures on weekend on the basis of the data provided.

5.

a)

The mathematical model of one-way ANOVA is given by

$Y_{ij} = \mu + \tau_i + e_{ij}$; i= 1,2,.....,k j = 1,2,.....,n_i

where,

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

$$\sum_{i=1}^{5}\sum_{j=1}^{ni} y_{ij} = 1,856$$

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

$$\text{C.F.} = \left(\sum_{i=1}^{5}\sum_{j=1}^{ni} y_{ij}\right)^2 / n = 104,385.94$$

SS_T = 174316- C.F. = 69,930.06;

$$SS_B = \sum_{i=1}^{5}\left(\sum_{j=1}^{ni} y_{ij}\right)^2 / n_i - C.F. = 354^2/7 + 386^2/8 + 87^2/6 + 645^2/7 + 384^2/5 - 104{,}385.94$$

= 22,325.69

$SS_R = SS_T - SS_B$ = 47,604.37

Sources of Variation	d.f	SS	MSS	F
Companies	4	22,326	5,581.42	3.283
Residual	28	47,604	1,700.16	
Total	32	69,930		

$F_{observed}$ = 3.283 ; $F_{(4, 28, 0.05)}$ = 2.714 ;

Reject H_0

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)				Total
	3 - 5	5 - 8	8 - 10	10 - 12	
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)				Total
	3 - 5	5 - 8	8 - 10	10 - 12	
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

$$\chi^2 = \sum_{i=1}^{12} (O_i - E_i)^2 / E_i = (27.42 - 45)^2 / 27.42 + \ldots\ldots + (5.82 - 12)^2 / 5.82 = 49.919$$

$\chi^2_{Observed} = 49.91$ $\chi^2_{(6,\ 0.05)} = 12.59$,where d.f. 6 = (3-1) × (4-1)

Reject H_0

6.

[i] The assumptions required for one-way analyses of variance (ANOVA) are:

The populations must be **normal**

The populations have a **common variance**

The observations are **independent**. [1]

The sample variance observed for the four rates appear very different from each other.

Thus, we can clearly see that the assumption that the underlying populations have a common variance assumption will not hold for the data as they are. [1]

[ii]

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

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

[1]

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

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

[2]

[iii]

The scientist was correct in asserting that the loge 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. [1]

To justify this, a quick 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 implies that the assumption of common variance for the underlying population 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	**989.06**	**30.17**	**1.53**	**1,268.14**

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

[iv]

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

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

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 [1]

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:

$$H_0: \tau_i = 0, \quad i = 1,2,3,4 \quad against \quad H_1: \tau_i \neq 0 \; for \; at \; least \; one \; i \qquad [1]$$

To carry out the ANOVA, we must first compute the Sum of Squares. We have the following table using the information given in the question:

Rate	$\log_e(x)$		
	Mean	Variance	$Y_{i.}^2$
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: $Y_{i.}^2 = \left(\sum_{j=1}^{3} Y_{ij}\right)^2 = (3 * Mean_i)^2$

Now:

- $SS(\text{Rate}) = \sum_{i=1}^{4} \frac{Y_{i.}^2}{3} - \frac{Y_{..}^2}{12} = \frac{973.373}{3} - \frac{(3*20.110)^2}{12} = 21.153$
- $SS(\text{Residuals}) = \sum_{i=1}^{4} \left\{\sum_{j=1}^{3} \left(Y_{ij} - \bar{Y}_{i.}\right)^2\right\} = \sum_{i=1}^{4} \{2 * Variance_i\} = 2 * 0.618 = 1.236$

[3]

The ANOVA table is as follows:

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

[2]

The 1% critical value for F (3, 8) distribution is 7.951.

Given the observed F statistic value is much larger than this, we can state the p-value for this test is almost near to zero or in other words there is overwhelming evidence against the null hypothesis H0. Thus it can be concluded that the underlying means are not equal. [1]

7.

i)

The relevant summary statistics to compute correlation coefficient are:

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 207 - 10 * \left(\frac{39}{10}\right)^2 = 54.90$$

$$S_{xy} = \sum xy - n\overline{xy} = 2853 - 10 * \left(\frac{39}{10}\right)\left(\frac{562}{10}\right) = 661.20$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 40508 - 10 * \left(\frac{562}{10}\right)^2 = 8923.60$$

$$\text{Correlation Coefficient } r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{661.20}{\sqrt{54.90}\sqrt{8923.60}} = 0.945$$

ii)

<u>Fitted Linear Regression Equation</u>

The coefficients of the regression equation are:

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

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{562}{10}\right) - 12.04 * \left(\frac{39}{10}\right) = 9.23$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta}x = 9.23 + 12.04x}$

iii)

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

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

$$SS_{RES} = S_{yy} - \frac{S_{xy}^{\ 2}}{S_{xx}} = 8923.60 - \frac{(661.20)^2}{54.90} = 960.30$$

$$SS_{REG} = S_{TOT} - S_{RES} = 8923.60 - 960.30 = 7963.30$$

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,

$$0.945 = r = \frac{S_{xy}}{(S_{xx} * S_{yy})^{0.5}} = \sqrt{R^2} = \sqrt{0.8924}$$