ROLL NO. 51
LOKIT NAGI

Date
Page

Probability and Statistics 2
Assignment 2

1.

	x	y	xy	x^2
1	40	56	2240	1600
2	10	62	620	100
3	100	195	19,500	10,000
4	110	240	26,400	12,100
5	120	170	20,400	14,400
6	150	~~6~~270	40,500	22,500
7	20	48	960	400
8	90	196	17640	8,100
9	80	214	17,120	6,400
10	130	286	37,180	16,900

a) $\Sigma x = 850$, $\Sigma y = 1,737$, $\Sigma xy = 182560$, $\Sigma x^2 = 92,500$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 182,560 - \frac{(850)(1737)}{10} = 34915$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 92,500 - \frac{(850)^2}{10} = 20250$$

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = 1.72 \quad , \quad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{\Sigma y}{n} - \hat{b}\frac{\Sigma x}{n}$$

$$= \frac{1737}{10} - 1.72\left[\frac{850}{10}\right] = 27.14$$

$\boxed{\hat{b} = 1.72}$ $\boxed{\hat{a} = 27.14}$

$y = 27.14 + 1.72x$

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

Date
Page

Q2.

i) Fitted Linear Regression Equation

The relevant summary statistics to fit the equation are

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

$S_{xx} = \Sigma x^2 - (\Sigma x)^2/n = 12666.58 - (385.2)^2/12 = 301.66$

$S_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n = 38191.41 - (385.2)(1162.5)/12 = 875.16$

$S_{yy} = \Sigma y^2 - (\Sigma y)^2/n = 119026.9 - (1162.5)^2/12 = 6409.71$

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.9$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \left(\frac{1162.5}{12}\right) - 2.9\left(\frac{385.2}{12}\right) = 3.78$$

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

ii) Confidence Interval for β

Assuming normal errors with a constant variance:

ROLL NO.51
LOKIT NAGI

Date
Page

95% confidence interval for β:

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

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

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

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

95% CI For β: $2.90 \pm 2.228 * 1.13$
$= (0.38, 5.42)$

iii) 95% Confidence Intervals for the mean IBM share price

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

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

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

Thus, 95% Confidence interval:

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

$= 119.78 \pm 2.228 * 10.5989$

$\Rightarrow (96.17, 143.39)$

ROLL NO. 51
LOKIT NAGI

Date
Page

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 non peak 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:

H_0: The mean of differences is same for each city ; against
H_1: 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

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

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

$SS_R = SS_T - SS_B = 600$

The Anova Table is :

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600	25	
Total	27	1,116.11		

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

The 5% critical point is 3.009, so we have sufficient evidence to reject Ho at the 5% level. Therefore it is reasonable to conclude that there are underlying differences between the cities.

iv) Analysis of 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 :

ROLL NO. 51
LOKIT NAGI

Date
Page

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

$= 5.52$

Now we can examine the difference between each of the pairs of means. It has 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 :

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

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 difference between means 2 and 4, so we cannot combine the last two groups.

Therefore the diagram remains as before.

ROLL NO. 51
LOKIT NAGI

Date
Page

Q.5. 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.

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.

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 check can be done on the ratio of maximum to minimum sample variance among the four rates data. A smaller ratio and

ROLL NO. 51
LOKIT NAGI

Date
Page

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.

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 $\Sigma \tau_i = 0$.
- e_{ij} are the independent error terms which follows Normal distribution with mean 0 and common unknown variance σ^2.

ROLL NO. 51
LOKIT NAGI

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
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, $y_i^2 = (\sum_{j=1}^{3} y_{ij})^2 = (3 * \text{mean}_i)^2$
Now,
$SS(\text{Rate}) = \sum_{i=1}^{4} \frac{y_i^2}{3} - \frac{y^2}{12} = \frac{973.373}{3} - \frac{(3*20.110)^2}{12}$
$= 21.153$
$SS(\text{Residuals}) = \sum_{i=1}^{4} \{ \sum_{j=1}^{3} (y_{ij} - \bar{y}_i)^2 \}$
$= \sum_{i=1}^{4} \{ 2 * \text{variance}_i \} = 2 * 0.618$
$= 1.236$

ROLL NO. 51
LOKIT NAGI

Date
Page

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		

The 1% critical value for $F(3,8)$ distⁿ. 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 H_0. Thus it can be concluded that the underlying means are not equal.

v) The 95% CI for the i^{th} rate mean is given by

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

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

The common error variance σ^2 is estimated as:

$$\hat{\sigma}^2 = \frac{\text{Residual sum of Squares}}{8} = 0.155 \quad \{\text{from ANOVA table earlier}\}$$

The 95% CI are tabulated as follows:

ROLL NO. 51
LOKIT NAGI

Rate	Observed Mean	95% CI (Transformed Scale) LCI	UCI	95% CI (Original Scale) LCI	UCI
1	2.992	2.469	3.516	11.810	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

6. i)

$S_{xx} = \Sigma x^2 - n\bar{x}^2 = 207 - 10\left(\frac{39}{10}\right)^2 = 54.9$

$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 2853 - 10\left(\frac{39}{10}\right)\left(\frac{563}{10}\right) = 661.2$

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

Correlation Coefficient $r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$

$= \frac{661.20}{\sqrt{54.90}\sqrt{8923.6}} = \underline{0.945}$

ii) Fitted linear regression Equation

The coefficients of the regression eq^n are ÷

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.2}{54.9} \Rightarrow 12.04$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \left(\frac{562}{10}\right) - 12.04\left(\frac{39}{10}\right) \Rightarrow 9.23$

$Y = \hat{\alpha} + \hat{\beta}X = 9.23 + 12.04\,X$

ROLL NO. 51
LOKIT NAGI

Date
Page

iii) Relation

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

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

$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 8923.60 - \frac{(661.2)^2}{54.96} = 960.30$

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

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

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

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

$S_{xy} = 2176.84$

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

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 17.895 - 0.4545 * 15.83 = \underline{10.79}$

The fitted regression equation is

$\hat{y} = 10.79 + 0.4545x$

ROLL NO.51
LOKIT NAGI

Date
Page

ii) $S_{xx} = 4789.42$ (part i)

$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 1189.21$

$S_{xy} = 2176.84$ (part i)

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right) = \frac{1}{9} * \left(1189.21 - \frac{2176.84^2}{4789.42}\right)$$

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

$$S.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{22.20}{4789.42}} = 0.0681$$

To test $H_0: \beta = 0$
$H_1: \beta \neq 0$

Test Statistic,

$$\frac{\hat{\beta} - 0}{s.e.(\hat{\beta})} = \frac{0.4545}{0.0681} = 6.674$$

Under the assumption that the errors of the regression are iid $N(0, \sigma^2)$ R.V.s, beta has a t dist.^{n} with $n-2$ degrees of freedom.

Critical value for t distribution with 9 degrees of freedom $t_{9, 0.025} = 2.68$

Since the critical value at 95% LOS is less than test statistic, we reject H_0 to reject H_0.

ROLL NO. 51
LOKIT NAGI

Date
Page

iii) Pearson's correlation coefficient is computed as $\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

$S_{yy} = 1189.21$

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{2176.84}{\sqrt{4789.42 * 1189.21}} = 0.91$

iv) The estimated value of y corresponding to $x^3 = 25$ is $10.79 + 0.4545 * 25 = 22.15$

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

The variance of the estimator of mean response ÷

$\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{11} + \frac{84.0889}{4789.42}\right] * 22.20 = 2.41$

The variance of the estimator of individual response ÷

$\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = [1 + 0.1085] * 22.20 = 24.61$

Using t_9 distⁿ, 95% CI for mean and individual responses are ÷

$22.20 \pm 2.262 * \sqrt{2.41}$; $22.20 \pm 2.262 * \sqrt{24.61}$

$(18.7, 25.7)$; $(11.0, 33.4)$

v) The residual plot shows a definite pattern. Although the correlation coefficient is high, the model does not seem to be appropriate. Using the model leads to underestimation of premium rates at low and high mortality ratings.

ROLL NO. 51
LOKIT NAGI

Date
Page

8. i) Factor Analysis / Principal Component Analysis is A method 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 (P(i))	P(i)/(Sum(P(i)) over 1 to 5)
PC_1	0.456	65%
PC_2	0.137	19.5%
PC_3	0.08	11.4%
PC_4	0.0165	2.4%
PC_5	0.012	1.7%
	0.7015	100.0%

iii) As 1st 3 Principal Components explain over 95% of Total variance, dimensionality 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.

ROLL NO. 51
LOKIT NAGE

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 resignation $= (27 + 36 + 30)/3 = 31$

$SS_B = 20((27-31)^2 + (36-31)^2 + (30-31)^2)$
$= 840$

$F_{2,57} = (840/2) / (3591/57) = 6.667$

The 1% point from $F_{2,60}$ is 4.977 and since the test statistic is higher, we reject H_0.