Probability assignment II ②

Q1 a)

x	y
40	56
10	~~92~~ 62
100	195
110	240
120	170
150	270
20	48
90	196
80	214
130	286

$\Sigma x^2 = 92500$ $\Sigma y^2 = 372712$

$\Sigma x = 850$ $\Sigma y = 1737$

$\Sigma xy = 182560$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 92500 - \frac{(850)^2}{10}$$

$$= \underline{\underline{20250}}$$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 182560 - \frac{850 \times 1737}{10}$$

$$= \underline{\underline{34915}}$$

$$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 372712 - \frac{(1737)^2}{10}$$

$$= \underline{\underline{70995.1}}$$

$$\beta = \frac{S_{xy}}{S_{xx}} = \frac{34915}{20250} = 1.724197531$$

$$\hat{\alpha} = \bar{y} - \beta\bar{x} = \frac{1737}{10} - (1.724197531) \times 85$$

$$= \underline{\underline{27.14320988}}$$

$$y = \underline{\underline{27.14320988 + 1.724197531\, x}}$$

b) ∴ we can interpret that the money spend by Anand's ~~girf~~ girlfriend is affected by the amount of cash he withdrew at his previous visit to an ATM.

Q2 i $\Sigma x = 385.2$ $\quad \Sigma x^2 = 12666.58$

$\Sigma y = 1162.5$ $\quad \Sigma y^2 = 119026.9$

$\Sigma xy = 38191.41$ $\quad n = 12$

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

$$= \underline{\underline{301.66}}$$

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

$$= \underline{\underline{875.16}}$$

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

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.901$$

$$\hat{a} = \bar{y} - \hat{\beta}\bar{x} = \frac{1162.5}{12} - 2.901\left(\frac{385.2}{12}\right)$$

$$= \cancel{3.7481} \quad \underline{\underline{3.7529}}$$

$\therefore$ Therefore the regression line:-

ii Confidence interval for β

$$95\% \text{ CI} = \hat{\beta} \pm t_{n-2}(2.5\%)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right] = \frac{1}{10}\left[6409.7 - \frac{875.16^2}{301.66}\right] = \underline{\underline{387.07}}$$

$$95\% \text{ CI} = 2.901 \pm 2.228 \times \sqrt{\frac{387.07}{301.66}}$$

$$= 2.901 \pm 2.228 \times 1.1327$$

$$= (0.3764, 5.4237)$$

iii) 95% CI for mean IBM share price

$$\hat{y}_{x_0} \pm t_{n-2}(2.5\%) \sqrt{s^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

$x_0 \rightarrow \$40$

$\hat{y}_{x_0} = 3.7529 + 2.901 \times 40 = 119.7929$

95% CI

$$= 119.7929 \pm 2.228 \times \sqrt{387.07\left[\frac{1}{12} + \frac{(40-32.2)^2}{301.66}\right]}$$

$= 119.7929 \pm 2.228 \times 10.5989$

$= 119.7929 \pm 23.61434$

$\underline{(96.1785, 143.407)}$

Q3 i) <u>Comment</u>:-

- Centers of the distributions differ for all 4 cities. Thus there is a prima facie case for suggesting that the underlying means are different.
- Difference between the mean time taken to commute to office in peak hours and non-peak hours are in order city A (highest), city D (lowest).

ii) <u>Assumption</u> required to carry ANOVA:-

- population must be normal
- The observations are independent
- population have a ~~com~~ comman variance.

Date
Page

Q4 a) Mathematical one way analysis

$y_{ij} = \alpha + \beta x_{ij} + e_{ij}$ $\quad i = 1, 2, 3, \ldots k \quad j = 1, 2, 3 \ldots n_i$

k = no. of treatments

n_i = number of responses from ith treatment

y_{ij} is the jth response from ith treatment

e_{ij} is the error term $e_{ij} \sim N(0, 1)$

b - ?

c) Ho - Salaries are independent v/s Salaries are dependent

Oi

Paper cleared	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

Ei

Paper cleared	3-5	5-8	8-10	10-12	Total
0-3	$\frac{76 \times 57}{158} = 27.42$	23.09	14.43	11.06	76
4-6	15.15	12.76	7.97	6.11	42
7-9	14.43	12.15	7.59	5.82	40
Total	57	48	30	23	158

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{(45 - 27.42)^2}{27.42} + \frac{(20 - 23.09)^2}{23.09} + \frac{(6 - 14.43)^2}{14.43} + \frac{(5 - 11.06)^2}{11.06} +$$

$$\frac{(7 - 15.15)^2}{15.15} + \frac{(20 - 12.76)^2}{12.76} + \frac{(9 - 7.97)^2}{7.97} + \frac{(6 - 6.11)^2}{6.11} +$$

$$\frac{(5 - 14.43)^2}{14.43} + \frac{(8 - 12.15)^2}{12.15} + \frac{(15 - 7.59)^2}{7.59} + \frac{(12 - 5.82)^2}{5.82}$$

$= 49.919$

χ^2_{tab} df $(4-1)(3-1) = 6$ $\quad \chi^2_{tab} < \chi^2_{cal}$

$\chi^2_6 (0.05) = 12.59$ $\quad \therefore$ Reject Ho

$\therefore$ Salaries is dependent on the number of papers cleared.

classmate
Date
Page

Q5 i a) Same as Q3 ii)

ii b) Sample mean for rate 1 will be $\sqrt{\frac{1}{3}(29+13+21)} = \sqrt{21} = 4.5827$
$(\sqrt{x})$

Sample variance for rate 2 will be $\frac{1}{2}\left[(\log 180 - 4.827)^2 + (\log 90 - 4.827)^2 + (\log 120 - 4.827)^2\right]$
$\log x$

$= 0.1213.$

iii) Scientist are correct in asserting that log transformation must be done before performing ANOVA test. Because for the original data the underlying assumption is not true (equal variance) which seems true after log transformation.

iv a) same as 4 a)

b - ?

classmate

Date ______
Page ______

6) i) Correlation coefficient -

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

$S_{xy} = \sum xy - n\bar{x}\bar{y} = 2853 - 10 \times \left(\frac{39}{10}\right)\left(\frac{562}{10}\right) = 661.20$

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

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

ii) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04371585 = 12.04$

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

∴ Regression line

$y = \hat{\alpha} + \hat{\beta}x = 9.23 + 12.04x$

iii Relation $SS_T = SS_{Reg} + SS_{Res}$

$SS_T = S_{yy} = 8923.60$

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

$SS_{Reg} = SS_T - SS_{Res}$

$= 8923.60 - 960.30$

$= 7963.30$

iv Coefficient of Determination

$R^2 = \frac{S_{xy}^2}{S_{xx} \times S_{yy}} = \frac{SS_{Reg}}{SS_{Tot}} = \frac{7963.30}{8923.60} = 0.8924$

$R^2 = (0.945)^2 = 0.89302$ $\quad r = \sqrt{0.8924} = 0.945$

7) i) $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 7545.90 - \frac{(174.13)^2}{11} = 75490 - 2756.4779$

$= \underline{\underline{4789.4221}}$

$S_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) = \underline{\underline{2176.84}}$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2176.84}{4789.4221} = 0.4545$

$\alpha = \bar{y} - \hat{\beta}\bar{x} = \frac{197.84}{11} - 0.4545\left(\frac{174.13}{11}\right)$

$= 17.985454 - 7.194892511$

$= \underline{\underline{10.79056}}$

Linear regression eq:- $\alpha + \beta x$

$= \underline{\underline{10.79056 + 0.4545x}}$

ii) $S_{xx} = 4789.4221$

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

$S_{xy} = 2176.84$

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

$= 22.20$

$s.e.(\hat{\beta}) = \sqrt{\frac{\sigma^2}{S_{xx}}} = \sqrt{\frac{22.20}{4789.42}} = \underline{\underline{0.06808}}$

$H_0 \; \beta = 0 \qquad H_1 \; \beta \neq 0$

$\frac{\hat{\beta} - 0}{\sqrt{\sigma^2 / S_{xx}}} = \frac{0.4545}{0.06808} = 6.675.$

$t_{tab}(0.025) = t_{9,0.025} = 2.68$

$t_{tab} < t_{cal}$ ∴ We reject H_0 ∴ we reject H_0 @ 95% significance

∴ $\beta \neq 0$

iii Pearson's correlation is computated as

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad S_{yy} = 1189.21$$

$$r = \frac{2176.84}{\sqrt{4789.42 \times 1189.21}} = 0.91$$

iv) estimated value of y corresponding to $x = 25$ = $10.79 + 0.4545 \times 25$

$= 22.15$

$\sigma^2 = 22.20$

Variance of estimator of mean response is given by $\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\sigma^2$

$$= \left[\frac{1}{11} + \frac{84.0889}{4789.42}\right] \times 22.20$$

$$= 2.407952$$

Variance of estimator of individal response is given by $\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\sigma^2$

$$= [1 + 0.1085]\, 22.20$$

$$= 24.61$$

Q9

x	y
6	65
5.5	45
0.5	75
3.5	78
7.5	45
4.5	80
4	~~7~~ 49
1.5	91
2	67
10	49

$\Sigma x = 45$ $\Sigma x^2 = 277.5$ $\Sigma y = 644$ $\Sigma y^2 = 43956$

$\Sigma xy = 2602$

i) The scatter plot shows a inverse relationship between marks scored and hours spent per day on social media.

ii) $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 277.5 - \frac{45^2}{10} = 75$

$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 43956 - \frac{644^2}{10} = 2482.4$

$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 2602 - \frac{644 \times 45}{10} = -296$

$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{-296}{\sqrt{75 \times 2482.4}} = -0.686$

iii $H_0\ \rho = 0$ v/s $\rho < 0$

~~$r = \frac{1}{2} \log($~~

~~$r = \frac{1}{2} \ln\left(\frac{1-0.686}{1+0.686}\right)$~~ $\quad z_r \sim N\left(z_r, \frac{1}{n-3}\right]$

$z_P = \tan^{-1} \rho = \tan h(0.686) = -0.8403605763$

$= -0.8404$

$z_r = \tanh^{-1} r = 0$

$\therefore w_{cal} = \frac{z_r - z_r}{\sqrt{1/n-3}} = \frac{-0.8404}{\sqrt{1/7}} = -2.223385097.$

$w_{tab} = \pm 1.96$

$\therefore$ we reject H_0 at 95% confidence interval. we can conlude that the marks obtained and hours spent on social media are negatively correlated.

iv $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-296}{75} = -3.9467$

$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} = \frac{644}{10} - \frac{45}{10} \times -3.9467$

$= 64.4 - 4.5 \times (-3.9467)$

$= 64.4 + 17.76$

$= \underline{\underline{82.16}}$

Fitted line $y = 82.16 - 3.9467x$.

v) $R^2 = \frac{S^2xy}{S_{xx} S_{yy}} = \frac{-296^2}{75 \times 2482.4} = \underline{\underline{0.4706}}$.

vi For every additional houe spend on social media the total marks are reduced by 3.9467 i.e 4 marks (based on the fitted equation.

Q10 H_0: No difference among industries v/s H_1: at least one industry different

$n-1 = 19$; $n = 20$

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

Mean of regr resignation $= \frac{27+36+30}{3} = 31$

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

$= 20(16 + 25 + 1)$

$= 840$

$F_{cal} = \frac{840/\overset{3-1}{2}}{3591/\underset{60-1-2}{57}} = 6.67$

~~$F_{2,57}$ (~~

$F_{2,60}(1\%) = 4.677 < F_{cal}$

$\therefore$ We reject null hypothesis and we can conclude that at least one industry is different