Assignment -2

Probability & Statistics -2

1)

Withdrawal	1	2	3	4	5	6	7	8	9	10
x	40	10	100	110	120	150	20	90	80	130
y	56	62	195	240	170	270	48	96	214	286

$\Sigma x = 850 \quad \Sigma y = 1737 \quad \Sigma x^2 = 92500, \quad \Sigma xy = 173560$

a) $\beta = \dfrac{S_{xy}}{S_{xx}}$

$S_{xy} = \Sigma(x_i)(y_i) - \dfrac{(\Sigma x_i)(\Sigma y_i)}{n} = 173560 - 147645 = 25915$

$S_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n} = 92500 - 72250 = 20250$

$\beta = 1.279$

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

$\Rightarrow 173.7 - 85 \times 1.279 = 64.985$

$\bar{y} = 64.985 + \bar{x}(1.279)$

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

NOVEMBER

DECEMBER

2. i) $\Sigma x = 385.2$, $\Sigma y = 1162.5$, $\Sigma x^2 = 12666.58$, $\Sigma y^2 = 119026.9$
$\Sigma xy = 38191.41$

$$S_{xx} = \Sigma x^2 - n\bar{X}^2 = 12668.58 - 12\left(\frac{385.2}{12}\right)^2 = 301.66$$

$$S_{yy} = \Sigma y^2 - n\bar{Y}^2 = 119026.90 - 12\left(\frac{1162.5}{12}\right)^2 = 8409.71$$

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

$$\hat{\beta} = 2.9$$

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

$$\bar{Y} = 3.78 + 2.90\bar{X}$$

ii) 95% CI for $\beta \pm t_{n-2}(2.50\%) . s.e(\hat{\beta})$

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

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

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

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

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

$x_0 = 40 \Rightarrow \hat{y}x_0 = 3.78 + 2.90(40) = 119.78$

$$\therefore 95\ CI = 119.78 \pm 2.228 \times \sqrt{387.07\left[\frac{1}{12} + \frac{(40-32.1)^2}{301.66}\right]}$$

$$\Rightarrow 119.78 \pm 2.228 \times 10.5989$$

$$\Rightarrow (96.17, 143.89)$$

3 i) The centers of the distributions differ for all the four cities.
The difference between the mean time taken to commute to office in peak hours & non-peak hours are in the order city A to city D.

ii)
- The populations must be normal.
- The observations are ~~independent~~ independent.
- The populations have a common variance.

iii) H_0: The mean of differences is same for each city
~~H_1:~~ v/s
H_1: The mean of differences are not same for each city.

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

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

ANOVA Table:

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600	25	
Total	27	1116.11	197.04	

Under H_0, $F = \frac{172.04}{25} = 6.88$ $(\because F_{3,24}$

~~Un~~ The 5% critical point is 3.009, so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is reasonable to conclude that
8 there are underlying differences between the cities.

9

10

11

Noon

01

02

03

04

05

06

i) The following assumptions are required for one-way analyses of variance (ANOVA):
- The populations must be normal.
- The populations have a common variance.
- The observations are independent.

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:

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

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

iv) We will perform an ANOVA on $\log_e x$ 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 $\log_e$ transformed value of the j^{th} observation of the no. of germinators per square foot observed when the i^{th} rate was applie

The 95% confidence intervals are tabulated in the following table.

Rate	Observed Mean	95% C.I (Transformed Scale) LCI	95% C.I (Transformed Scale) UCI	95% C.I (original Scale) LCI	95% C.I (original Scale) 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	512.509
4	6.575	6.052	7.098	424.770	1209.763

i]

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

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

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

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

$= \frac{661.20}{\sqrt{54.90}\sqrt{89236}} = 0.945$

ii] Fitted linear regression Equation

The coefficients of the regression eqⁿ are:

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

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

$\boxed{\therefore\ y = \hat{\alpha} + \hat{\beta}x = 9.28 + 12.04x}$

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) distribution is 7.951.

Considering the F statistic value is much larger than this, we can observe that the p-value for this test is almost near to zero. This is an evidence against the null hypothesis Ho. Thus it can be concluded that the underlying means are not equal.

A 95% confidence interval for the i^{th} rate mean 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:

$\hat{\sigma}^2$ = Residual Sum of Squares = 0.155

• 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.

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.

Sum of Squares table: $(\log_e(x))$.

Rate	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: $Y_i^2 = (\Sigma_{j=1}^{3} Y_{ij})^2 = (3 * \text{Mean}_i)^2$

Now:

$SS(\text{Rate}) = \Sigma_{i=1}^{4} \frac{Y_i^2}{3} - \frac{Y_{..}^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.110)^2}{12}$

$= 21.163$

$SS(\text{Residuals}) = \Sigma_{i=1}^{4} \{ \Sigma_{j=1}^{3} (Y_{ij} - \bar{Y}_i)^2 \}$
$= \Sigma_{i=1}^{4} \{ 2 \times \text{variance}_i \} = 2 \times 0.618 = 1.236$

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:

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

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

$S_{yy} = 43956 - \frac{664^2}{10} = 2482.4$

$S_{xy} = 2602 - 664\left(\frac{45}{10}\right) = -296$

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

iii] Null hypothesis $H0 : \rho = 0$ against $H1 : \rho < 0$
We will assume that the data comes from a bivariate normal distribution.

$\therefore r = 0.5\left[\ln\left(\frac{1-0.686}{1.686}\right)\right] = -0.8404$

Fisher's standardized statistic $= \frac{(-0.8404 - 0)}{\left(\sqrt{1/7}\right)}$

$= -2.22$

From above we get p-value as,
p-value $= P(z < -2.22) = 0.013$
The p-value obtained is quite small and hence shows a strong evidence to reject the null hypothesis with 95% C.I.

Hence, the marks obtained and hours spent on social media are negatively correlated.

iv) Beta $= \frac{Sxy}{Sxx} = \frac{-296}{75} = -3.9467$

Alpha = mean of y − beta × mean of x

$= \frac{644}{10} + 3.9467\left(\frac{45}{10}\right) = 82.16$

Therefore, the fitted line is $\boxed{y = 82.16 - 3.9476x}$

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

vi) From the fitted line equation we can conclude that, for ~~ee~~ every additional hour spent on social media per day, the total marks reduce by 3.95.