Probability & Statistics - 2

Q1. a) $\hat{y_i} = \hat{\alpha_i} + \hat{\beta} x_i$

b) $\hat{\beta} \Rightarrow 1.724197$

2. i $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha}_i = \bar{y} + \bar{\beta}\bar{x}$

$S_{xx} \Rightarrow \Sigma x_i - \frac{(\Sigma x_i)^2}{n} \Rightarrow \boxed{301.66}$ $S_{xy} \Rightarrow \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$

$S_{yy} \Rightarrow \Sigma y_i - \frac{(\Sigma y_i)^2}{n} \Rightarrow \boxed{6409.7125}$ $\Rightarrow \boxed{875.18}$

$\hat{\alpha}_j \Rightarrow 3.748181$ $\hat{\beta} = 2.901148.$

$\hat{y}_i \Rightarrow \boxed{\cancel{5.6493329}}\ \hat{\alpha}_i + \hat{\beta} x_i \Rightarrow \boxed{3.748181 + 2.901148 x_i}$

ii $PQ \Rightarrow \frac{\hat{\beta} - B}{se(\hat{\beta})} \sim t_{m-2}$

$se(\hat{\beta}) \Rightarrow \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}};\ \hat{\sigma}^2 \Rightarrow \frac{1}{10}\left[6409.7125 - \frac{(875.16)^2}{301.66}\right] \Rightarrow \boxed{387.0744}$

$\therefore se(\hat{\beta}) \Rightarrow \sqrt{\frac{387.0744}{S_{xx}}} \Rightarrow$

$\Rightarrow \sqrt{\frac{387.0744}{301.66}} \Rightarrow \boxed{1.1327611906}$

$P\left[-2.228 < \frac{2.901 - \beta}{1.133} < 2.228\right] \Rightarrow \boxed{0.95}$

$P\left[-2.228(1.133) - 2.901 < -\beta < 2.228(1.133) - 2.901\right] = 0.95$

$\Rightarrow (0.378904, 5.423096)$ for 95% confidence interval

iii $PQ \Rightarrow \frac{\hat{\mu}_i - \mu_i}{se(\hat{\mu}_i)} \sim t_{m-2}$

$\hat{\mu}_i \Rightarrow 3.748181 + 2.901146(40) \Rightarrow \boxed{119.79}$

$se(\hat{\mu}_i) \Rightarrow \sqrt{\left[\frac{1}{12} + \frac{(7.9)^2}{301.66}\right] 387.707} \Rightarrow \boxed{10.5986}$

$$P\left[-2.228 < \frac{119.79 - \mu_i}{10.5988} < 2.228\right] \Rightarrow 0.95$$

$\Rightarrow (96.17, 143.40)$ for 95% confidence interval

3 i Each city has a different distribution center which indicates that means are also different from each other. The variance of the last 2 cities i.e. C and D are different as in C being the largest and D being the highest.

ii Assumptions are :-

a It must follow a normal distribution.

b Population must have a common variance.

c The observations should be obtained by random sampling i.e. unbiased

iii $H_0 \Rightarrow$ Mean is same for all cities

$H_1 \Rightarrow$ Mean is not same for all cities.

$$SS_{TOT} \Rightarrow S_{yy} \Rightarrow \sum y_i^2 - \frac{(\sum y_i)^2}{n} \Rightarrow 1495 - \frac{(103)^2}{28} \Rightarrow \boxed{1116.11}$$

$$SS_{RES} \Rightarrow SS_{TOT} - SS_{REG}$$

$$SS_{REG} \Rightarrow \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right) \Rightarrow \boxed{516.11}$$

$$\therefore SS_{RES} \Rightarrow \cancel{516.11}\ 1116.11 - 516.11 \Rightarrow \boxed{600}$$

ANOVA Table

	df	Sum of squares	mean of squares
REG	3	516.11	172.04
RES	24	600	25
Total	27	1116.11	

$\frac{MSS_{REG}}{MSS_{RES}} \Rightarrow \frac{172.04}{25} \Rightarrow \boxed{6.88}$

$F_{3,24} @ 5\% \Rightarrow 3.009$ i.e. right tailed.
∴ we reject H_0 ~~as~~ TB

iv $\bar{y}_1 = 9.14$, $\bar{y}_2 = 6.29$, $\bar{y}_3 = 1.14$, $\bar{y}_4 = -1.86$

$\hat{\sigma}^2 = \frac{SS_{RES}}{n-2} \Rightarrow \boxed{25}$

$t_{24, 0.025} \times \hat{\sigma}^2 \sqrt{\frac{1}{7} + \frac{1}{7}} = 2.064 \times \overset{5}{\cancel{25}} \times \sqrt{\frac{2}{7}} \Rightarrow \boxed{5.516272}$

$\bar{y}_1 - \bar{y}_2 \Rightarrow \boxed{2.85}$; $\bar{y}_2 - \bar{y}_3 \Rightarrow \boxed{5.15}$; $\bar{y}_3 - \bar{y}_4 \Rightarrow \boxed{3}$

4a $Y_{ij} \Rightarrow \mu + T_i + e_{ij}$ where $i \Rightarrow 1,2,3 \ldots k$; $j \Rightarrow 1,2, \ldots n_i$
μ = Over mean response
e_{ij} = error term
k = number of treatments.
T_i = treatment and Y_{ij} = the j^{th} response from i^{th} treatment.
There are certain assumptions which states that it should follow a normal distribution and have a common variance.

b. $H_0 \Rightarrow$ Mean claim amounts are equal
$H_1 \Rightarrow$ Mean claim amounts are not equal.
7, 8, 6, 7, 5
$\sum_i^5 \sum_j^{n_i} y_{ij} \Rightarrow \boxed{1856}$ and $\sum_i^5 \sum_j^{n_i} y_{ij}^2 \Rightarrow \boxed{174316}$

$CF \Rightarrow \frac{\left(\sum \sum y_{ij}\right)^2}{n} \Rightarrow \frac{3444736}{33} \Rightarrow \boxed{104385.93939}$

$SS_{REG} \Rightarrow \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{643^2}{7} + \frac{384^2}{3} - CF \Rightarrow \boxed{22325.69}$

$SS_{TOT} \Rightarrow 174316 - CF \Rightarrow \boxed{69930.06}$

$SS_{RES} \Rightarrow SS_{TOT} - SS_{REG} \Rightarrow \cancel{69930.06 - 104385.47} \Rightarrow$

$\Rightarrow 69930.06 - 22325.69 \Rightarrow \boxed{47604.37}$

	df	Sum of squares	Mean sum of squares
REG	4	22326	5581
RES	28	47604	1700
Total	32	69930	

Test statistic $\Rightarrow \dfrac{MSS_{REG}}{MSS_{RES}} \Rightarrow \dfrac{5581}{1700} \Rightarrow \boxed{3.282941}$

$F_{4,20,0.05} \Rightarrow \boxed{2.714}$

$\therefore H_0$ is rejected.

c) O_i

⊙	3-5	5-8	8-10	10-12	Total
0-3	45	20	6	5	~~402~~ 76
4-6	7	20	9	6	42
7-9	5	8	15	12	40
Total	57	48	30	23	158

E_i

	3-5	5-8	8-10	10-12
0-3	27.4170	23.088	14.43038	11.0632
4-6	15.151	12.759	7.974	6.113
7-9	14.430	12.1518	7.594	5.822
Total				

$H_0 \Rightarrow$ Sales are independent

$H_1 \Rightarrow$ Sales are dependent.

Test statistic $\Rightarrow \dfrac{\Sigma (O_i - E_i)^2}{E_i} \Rightarrow \boxed{49.91146}$

degree of freedom $\Rightarrow 2 \times 3 \Rightarrow \boxed{6}$

$$\chi^2_{c,5\%} \Rightarrow \boxed{12.59}$$

$\therefore H_0$ is rejected.

5 i Assumption are population must follow normal distribution, ~~[illegible]~~ observation must be obtained through random sampling i.e. unbiased and it should have a common variance

ii Mean $\Rightarrow \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3} \Rightarrow \boxed{4.52443}$

variance $\Rightarrow \frac{\Sigma(x_i - \bar{x}_i)^2}{n-1} \Rightarrow \boxed{0.12127}$

iii

iv

a) $Y_{ij} = \mu + T_i + e_{ij}$ j ~~&~~ $i = 1,2,3,4$ & $j = 1,2,3$

- $\mu \Rightarrow$ population mean
- $T_i \Rightarrow$ deviation of i^{th} rate mean.
- $E_{ij} \Rightarrow$ independent error term.

b) ANOVA

$H_0 \Rightarrow T_i = 0$; $H_1 \Rightarrow T_i \neq 0$

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	389.063
	20.11	0.618	973.373

DATE

$Y_{i}^{2} \Rightarrow \left(\sum Y_{ij}\right)^{2} = (3 \times \text{mean}_i)^{2}$

$SS_{RATE} \Rightarrow \sum^{4} \frac{Y_i^2}{3} - \frac{Y^2}{12} \Rightarrow \boxed{21.153}$

$SS_{RES} \Rightarrow \sum\left(\sum (Y_{ij} - \bar{Y}_i)^2\right) \Rightarrow \sum_{i=1}^{4} 2 \times \text{variance}$

$\Rightarrow 2 \times 0.618 \Rightarrow \boxed{1.236}$

~~Rate~~	d.f.	Sum of squares	Mean of Sum of squares
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
	11	22.389	

Test Statistics $\Rightarrow \frac{MSS_{Rate}}{MSS_{Res}} \Rightarrow \frac{7.051}{0.155} \Rightarrow \cancel{45.638} \Rightarrow \boxed{45.490322}$

$F_{3,8,1\%} \Rightarrow \boxed{7.951}$

$\Rightarrow$ The p-value of the test is near to zero.

6(i) $S_{xx} \Rightarrow 207 - \frac{(39)^2}{10} \Rightarrow \boxed{54.9}$

$S_{yy} \Rightarrow 40505 - \frac{(562)^2}{10} \Rightarrow \boxed{8923.6}$; $S_{xy} = 2853 - \frac{562 \times 39}{10} \Rightarrow \boxed{661.2}$

$r \Rightarrow \frac{S_{xy}}{\sqrt{S_{xy} \times S_{yy}}} \Rightarrow \frac{661.2}{\sqrt{54.9 \times 8923.6}} \Rightarrow \boxed{0.94466}$

(ii) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} \Rightarrow \frac{661.2}{54.9} \Rightarrow \boxed{12.04371}$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \Rightarrow 56.2 - 12.04371(3.9) \Rightarrow \boxed{9.2295}$

(iii) $SS_{TOT} = SS_y = \boxed{8923.6}$

$SS_{REG} \Rightarrow \frac{S^2_{xy}}{S_{xx}} \Rightarrow \frac{(661.2)^2}{54.9} \Rightarrow \boxed{7963.30492}$

$SS_{RES} \Rightarrow SS_{TOT} - SS_{REG} \Rightarrow \boxed{160.29508}$

iv $R^2 \Rightarrow \frac{S^2_{xy}}{S_{xx} \cdot S_{yy}} \Rightarrow \boxed{89.24\%}$

7.

i $S_{xx} \Rightarrow 7545.9 - \frac{179.13^2}{11} \Rightarrow$

$S_{yy} \Rightarrow 4747.45 - \frac{197.84^2}{11} \Rightarrow$

$S_{xy} \Rightarrow \boxed{2176.86}$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \Rightarrow \boxed{0.45451}$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \boxed{10.78056}$

$\hat{y} = 10.78056 + 0.45451x$

ii $H_0 \Rightarrow B = 0$ and $H_1 \Rightarrow B \neq 0$

Test Statistics $\Rightarrow \frac{\hat{\beta} - B}{S_e(\hat{\beta})} \sim t_{n-2}$

$\hat{\sigma}^2 \Rightarrow \frac{1}{n-2}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right) \Rightarrow \boxed{22.20136}$

$S_e(\hat{\beta}) \Rightarrow \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

Test Statistic $\Rightarrow \frac{0.45451 - 0}{\sqrt{\frac{22.20136}{4789.4271}}} \Rightarrow 6.6755 \sim t_9$

$t_9 =$

$t_{9, 2.5\%} \Rightarrow \boxed{2.262}$

$\therefore$ H_0 is rejected.

Assumption are population must follow a normal distribution, have a common variance and observation should be collected by random sampling i.e. unbiased.

iii $r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \Rightarrow \boxed{0.91213}$

iv $x_0 = 25$

$\mu_0 = \frac{\hat{\mu} - \mu}{S_c(\hat{\mu})} \sim t_{n-2}$

$SE(\hat{\mu}) = \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \hat{\sigma}^2}$ $\quad \left(\hat{\mu} = \hat{\alpha} + \hat{\beta} \times 25\right)$

$SE(\hat{\mu}) = \sqrt{\frac{1}{11} + \frac{(25 - 174.1311)^2}{4789.4221} \times 22.02136} \Rightarrow \boxed{1.55}$

$\hat{\mu} \pm t_{9, 2.5\%} \Rightarrow 22.15331 \pm 3.51018$

$\Rightarrow (18.64313, 25.66349)$

~~[illegible]~~

v The model is not a good fit as the assumption ~~[illegible]~~ taken in consideration might be wrong.

8 (i) ~~The~~ Factor analysis is done to reduce many individuals items into fewer dimensions. In regression models it is ~~[illegible]~~ used to simplify the data. Data can be easily compressed into a short ~~[illegible]~~ form. The things taken into consideration while under factor analysis are that they are ~~[illegible]~~ not correlated and has a linear combination of the original variables of the data.

9.

(i) · Marks scored and hours spent on social media have a negative ~~[illegible]~~ correlation between them.

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

$S_{yy} \Rightarrow 43956 - \frac{(644)^2}{10} \Rightarrow \boxed{2482.4}$

$S_{xy} \Rightarrow 2602 - \frac{644 \times 45}{10} \Rightarrow \boxed{-296}$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} \Rightarrow \boxed{-0.686002}$$

Hence proved.

(iii) $H_0 \Rightarrow \rho = 0$; $H_1 \Rightarrow \rho < 0$.

$$\tanh^{-1}(r) - \tanh^{-1}(\rho) \sim N\left(0, \frac{1}{n-3}\right)$$

$$\text{Test Statistics} \Rightarrow \frac{\tanh^{-1}(r) - \tanh^{-1}(\rho)}{\sqrt{1/n-3}} \sim N(0,1)$$

$$\Rightarrow \boxed{-2.07893}$$

$$Z_{5\%} \Rightarrow \boxed{1.6448}$$

$\therefore$ H_0 is rejected as calculated value is greater than the calculated value.

(iv) $$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \Rightarrow \boxed{-3.9467} \qquad \hat{y} = 82.16 - 3.9467X$$

$$\hat{\alpha} = \bar{Y} - \bar{X}.\hat{\beta} \Rightarrow \boxed{82.16}$$

(v) $$R^2 \Rightarrow \frac{S^2_{xy}}{S_{xx} S_{yy}} \Rightarrow \boxed{47.0598\%}$$

The model is not a good fit as it has only 47% variability.

(vi) $X_0 = 1$

$$\text{Expected change} = \hat{\beta} x_0 \Rightarrow \boxed{-3.9467}$$

10. $$S_{xx} \Rightarrow 0.2925 - \frac{0.93^2}{3} \Rightarrow \boxed{0.0042}$$

$$S_{yy} \Rightarrow 0.0189 - \frac{0.23^2}{3} \Rightarrow \boxed{0.00126}$$

$$S_{xy} \Rightarrow 0.0735 - \frac{0.93 \times 0.23}{3} \Rightarrow \boxed{0.0022}$$

$$SS_{TOT} \Rightarrow S_{yy} \Rightarrow \boxed{0.00126}$$

$SS_{REG} \Rightarrow \frac{S^2_{xy}}{S_{xx}} \Rightarrow \boxed{0.00115}$

$SS_{RES} \Rightarrow 0.00011$

Table

	Df	Sum of Square	Mean of sum of square
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011
Total	2	0.00126	

$H_0 \Rightarrow \beta = 0 \quad ; \quad H_1 \Rightarrow \beta \neq 0$.

Test Statistics $\Rightarrow \frac{0.0018}{0.00011} \sim F_{1,1}$

$10.4545 \sim F_{1,1}$

$F_{1,1,5\%} \Rightarrow 161.4$

$\therefore$ We accept H_0 as tabulated value is greater than the calculated value