Institute of Actuaries of India

Subject CS1-Actuarial Statistics (Paper A)

December 2022 Examination

INDICATIVE SOLUTION

Introduction

The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiner have given credit for any alternative approach or interpretation which they consider to be reasonable.

Solution 1: $S'^2_n = \frac{\sum_i (X_i - \overline{X})^2}{n} = {(n-1)S_n^2}/{n}$

$\mathrm{E}(S'^2_n)$ = E [${(n-1)S_n^2}/{n}$] = ${(n-1)E[S_n^2]}/{n}$ = ${(n-1)\sigma^2}/{n}$

So for n = 13 , $\mathrm{E}(S'^2_{13}) = {(13-1)\sigma^2}/{13}$ = 3.1591

Thus, bias = $\mathrm{E}(S'^2_{13}) - \sigma^2$= 3.1591 – 3. 4224 = - 0.2633

[3 Marks]

Solution 2: **i) Answer: C**

PCA will not reduce the number of variables. So a 10 variable data will give 10 components. PCA only transforms the data into uncorrelated linear combination of the variables. The components are then selected to maximise variance. The initial PCAs (say PCA 1 and PCA 2) usually try to capture maximum possible information.

[2]

ii) Answer: B

A linear predictor is linear in the parameters. It does not have to be linear in covariates as in case of A) and C)

[2]

[4 Marks]

Solution 3: **i)**

PDF of a standard normal distribution is: $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

$M_x(t)$ = E(e^{tx})

$= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}x^2}\, dx$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx + -\frac{1}{2}x^2}\, dx$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{1}{2}t^2 - \frac{1}{2}t^2 + tx + -\frac{1}{2}x^2}\, dx$ *………… Adding and subtracting ½ t^2 in the exponent*

$= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 + tx + -\frac{1}{2}x^2}\, dx$

$= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t^2 - 2tx + x^2)}\, dx$

$= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{(1)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-t}{1}\right)^2}\, dx$

$= e^{\frac{1}{2}t^2} * 1$

……. As the PDF is that of a normal distribution with mean t and sd 1 and it integrates to 1

$= e^{\frac{1}{2}t^2}$

[3]

ii)

We have to prove that a normal variable X with mean μ and variance δ^2 is symmetrical about its mean.

It means that we have to prove that the coefficient of skew-ness of the normal variable X is equal to 0.

Symbolically, we have to prove – $E[(X - \mu)^3] = 0$ .... (1)

We know that standard normal variable Z is a special case of a normal variable with $\mu = 0$ and $\delta^2 = 1$

We also know the relationship that $Z = (X - \mu) / \delta$

Rearranging, $(X - \mu) = Z * \delta$

Hence we have to prove that-

From (1) $E[(X - \mu)^3] = 0$

i.e. $E(Z * \delta)^3 = 0$

i.e. $\delta^3 * E(Z^3) = 0$ .... (2)

$M_X(t)$ for a standard normal variable $= e^{\frac{1}{2}t^2}$

Using the Taylor expansion, $e^x = \sum_0^\infty \frac{x^r}{r!}$

$$M_X(t) = \sum_0^\infty \frac{(\frac{1}{2}t^2)^r}{r!}$$
$$= 1 + ½ t^2 + (½ t^2)^2 / 2! + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..$$

It is given in the question that the $E(Z^r)$ is the coefficient of the term t^r / $r!$ in the Taylor expansion.

So,

$E(Z) = 0$ …… *since term t/1 is not there in the Taylor expansion*

$E(Z^2) = 1$ …… *coefficient of term t^2 / 2 in the Taylor expansion*

$E(Z^3) = 0$ …….. *there is no term t^3 / 6 in the Taylor expansion*

Thus, $E(Z^3) = 0$ .... (3)

From (2) and (3),

L.H.S.

$= \delta^3 * E(Z^3)$

$= \delta^3 * 0$

$= 0$

R.H.S = 0

Hence we are able to prove that the coefficient of skew-ness for a normal variable X is 0 and hence we infer that it is symmetrical about its mean. .

[3]

[6 Marks]

Solution 4: **i)**

Since only one claim is eligible for each of the ailments, claims from Heart, Cancer and Liver related ailments can be modelled as three Bernoulli Variables (indicator variables). It is given in the question that the three can be assumed to be independent.

H = Claims from Heart related ailments H ~ Bernoulli (0.01)

C = Claims from Cancer related ailments C ~ Bernoulli (0.02)
L = Claims from Liver related ailments L ~ Bernoulli (0.005)

Let X be the claim amount to be paid out in the next year on a single policy

X = 20 × H + 25 × C + 15 × L

We have to find E(X) and s.d.(X)

E(X) = 20 × E(H) + 25 × E(C) + 15 × E(L)
= 20 × (0.01) + 25 × (0.02) + 15 × (0.005) ……. *E(A) = p for A ~ Bernoulli(p)*
= 0.775 lakhs
= INR 77,500

Var(X) = 20^2 × Var(H) + 25^2 × Var(C) + 15^2 × Var(L)
…… *Since H, C and L are independent, no co-variance terms*

= 400 × (0.01)(1-0.01) + 625 × (0.02)(1-0.02) + 225 × (0.005)(1-0.005)
………… *Var(A) = p(1-p) for A ~ Bernouli(p)*
= 17.32938 lakhs

SD(X) = $(17.32938)^{1/2}$
= INR 4.1628 lakhs

[2]

ii)
Exactly one claim has occurred. We don't know whether it is related to H, C or L.

P(exactly 1 claim)
= P(H) × (1-P(C)) × (1-P(L)) + (1-P(H)) × P(C) × (1-P(L)) + (1-P(H)) × (1-P(C)) × P(L)
= (0.01)(1-0.02)(1-0.005) + (1-0.01)(0.02)(1-0.005) + (1-0.01)(1-0.02)(0.005)
= 0.009751 + 0.019701 + 0.004851
= 0.034303

P(H | 1 claim has occurred) = 0.009751 / 0.034303 = 0.284261
P(C | 1 claim has occurred) = 0.019701 / 0.034303 = 0.574323
P(L | 1 claim has occurred) = 0.004851 / 0.034303 = 0.141416

These should total up to 1.

So, we have to find E(X | 1 claim has occurred)

E(X | 1 claim has occurred)
= 20 × P(H | 1 claim has occurred) + 25 × P(C | 1 claim has occurred) + 15 × P(L | 1 claim has occurred)
= 20 × 0.284261 + 25 × 0.574323 + 15 × 0.141416
= INR 22.16453 lakhs

………… *Kindly note that even after taking account the condition that one claim has occurred, H,C and L continue to be Bernoulli variables and hence their mean will be equal to p …… although the value of p has changed now.*

[3]

iii)

There are three independent risks covered under this policy with relatively very small probability of incidence of a claim in the next year.
The probability of no claim during the next one year = (1-0.01) (1-0.02) (1-0.005) = 0.965349

Since in almost 96% of the cases, there will be no claim, the expected pay-out at the inception of the policy is quite low (lower than1 lakh).

However, after one claim has occurred, we have actually experienced something which has a possibility of 3.4% to occur. After its occurrence we are finding out the expected amount since we don't know whether it relates to H, C or L (otherwise there was no need of expectation, we could directly infer it to be 20 lakhs, 25 lakhs or 15 lakhs).

Since something which was only 3.4% probable has actually occurred, there is a significant increase in the expected claim pay-out from (i) to (ii).

[2]

[7 Marks]

Solution 5: **i)**

Most suitable distribution for N is Binomial (85, p) where p is the probability of head.

Estimate of p is 40/85.

Mean of N is np =85*(40/85) = 40.

Variance is np(1-p) = 85*(40/85) * (1-40/85) = 21.1765. [2]

ii)

N approximately follows Normal with mean 85*(1/2) = 42.5 and variance 85*(1/2)*(1-(1/2)) = 21.25.

Using continuity correction,

PBin (N > 40) = PNor(N≥ 40.5)

p = ½	**p = 40/85**
= PNor (Z ≥ (40.5 -42.5)/(21.25^(1/2))) = PNor (Z ≥ -0.43386) = 0.6678	= PNor (Z ≥ (40.5 -40)/(21.1765^(1/2))) = PNor (Z ≥ 0.11) = 0.4562

[2]

iii)

Using the above probability, the P-value of the test is 0.6678 ***(or alternatively 0.4562)***. We do not have sufficient evidence to reject Ho at 5% significance level. [1]

iv)

Null hypothesis to be rejected for P-value < 0.05,

Then PBin (N>=n) = 0.05

p = ½
P[((N-42.5)/21.25^0.5) >=((n-42.5)/21.25^0.5)] = 0.05 P(Z>=1.64485) = 0.05 Thus, n=1.64485*21.25^0.5+42.5 = 50.0824
p = 40/85
P[((N-40)/21.1765^0.5) >=((n-40)/21.1765^0.5)] = 0.05 P(Z>=1.64485) = 0.05 Thus, n=1.64485*21.1765^0.5+40 = 47.56926

[2]

[7 Marks]

Solution 6: **i)**

We know that X and Y are proportions. Hence they need to lie between 0 and 1.

Hence, preliminary bounds are: 0 ≤ x, y ≤ 1

It is also given in the problem that you cannot opt for accidental death benefit rider unless you have opted for the group term insurance policy. This further implies that y ≤ x.

So, the final bounds are –
$0 \leq x \leq 1$ for X
$0 \leq y \leq x$ for Y

[1]

ii)

We have to determine the marginal density function of X.

fx(X) $= \int_{y=0}^{x} f(x,y)\,dy$
$= \int_{y=0}^{x} 2\,(x+y)\,dy$
$= \int_{y=0}^{x} 2x\,dy + \int_{y=0}^{x} 2y\,dy$
$= 2x\,(x - 0) + 2/2\,(x^2 - 0^2)$
$= 2x^2 + x^2$
fx(X) $= 3x^2$

[2]

iii)

We are given that X = 0.10 and we have to calculate P(Y<0.05 | X = 0.10).

P(Y<0.05 | X = 0.10) $= \frac{h(x=0.10, y<0.05)}{f_x(x=0.10)}$

h(x=0.10, y<0.05) $= \int_{y=0}^{0.05} 2(0.10 + y)\,dy$

$= [0.2(0.05 - 0) + 2 * \frac{1}{2}(0.05^2 - 0^2)]$
= (0.01 + 0.0025)
= 0.0125

f_x(X=0.1) $= 3(0.10)^2$
= 0.03

P(Y<0.05 | X = 0.10) = 0.0125/0.03
= 0.4167

Hence the probability that less than 5% of the employees will opt for the accidental death benefit rider, given that 10% of them have opted for the group term insurance policy is 0.4167.

[4]

[7 Marks]

Solution 7: **i)**

P(X = k+1)
Substituting x by K+1
P(X=k+1) = $e^{-\lambda}\lambda^{k+1}/(k+1)!$
$= e^{-\lambda}(\lambda^k * \lambda)/[(k+1) * k!]$
$= e^{-\lambda}\lambda^k/k! * [\lambda/(k+1)]$
$= \frac{\lambda}{k+1}\,P\,(X = k)$ for k=0,1,2,3,…

[2]

ii)

As MLE of $\hat{\lambda}$ = 1.186,
P(X=0)=$e^{-1.186}$ = 0.3054
Also, P(X=8+)=1-$\sum_{i=1}^{7} P(X = i)$

K	0	1	2	3	4	5	6	7	8+

Probability using MLE and $\frac{\lambda}{k+1}\ P\ (X = k)$	0.3054	0.3623	0.2148	0.0849	0.0252	0.0060	0.0012	0.0002	0.0000
Expected No of policies = prob*1000	305.44	362.25	214.82	84.92	25.18	5.97	1.18	0.20	0.03

[3]

iii)

To perform Chi-square goodness of fit
combining 4 categories to obtain >5

$$\chi^2 = \sum_{i=0}^{5\&more} \frac{(f_i - e_i)^2}{e_i}$$

K	0	1	2	3	4	5+
e_i	305.4	362.3	214.8	84.9	25.2	7.4
f_i	300	365	216	70	30	19

$$\chi^2 = \ 0.10 + 0.02 + 0.01 + 2.61 + 0.91 + 18.18 = 21.84$$

Degrees of freedom = 6-1-1 = 4 due to MLE estimate
From tables, $\chi^2_{0.05,4}$ = 9.488
21.84>9.488
Thus, the number of claims does not come from a Poisson (1.186) distribution.

[5]

[10 Marks]

Solution 8: **i)**

Spearman Rank correlation coefficient :

Rank in low to high order and differences are :

Zones	A	B	C	D	E	F	G	H	I
Police	9	7	4	2	8	6	5	3	1
Cases	8	6	3	2	9	7	5	4	1
Differences	1	1	1	0	-1	-1	0	-1	0
diff square	1	1	1	0	1	1	0	1	0

$r_s = 1 - \frac{6\ X\ 6^2}{9\ X\ (9^2-1)}$ = 0.95

Kendall Rank correlation coefficient :

Arranging in order of Policemen rank

Zones	Police	Cases	Concordant Pairs	Disconcordant Pairs
I	1	1	8	0
D	2	2	7	0
H	3	4	5	1
C	4	3	5	0

G	5	5	4	0
F	6	7	2	1
B	7	6	2	0
E	8	9	0	1
A	9	8	0	0
Total			33	3

$$\tau = \frac{33 - 3}{\mathbf{33 + 3}} = 0.83$$

[4]

ii)

Both Spearman and Kendall rank correlation coefficient indicates a strong positive correlation between policemen and cases. In other words, zones with more policemen have more cases.

However, correlation does not necessarily infer causation. Even though more cases are present where more policemen are present, it doesn't indicate cases will go down with reduction of policemen.

It could be other way, i.e., more policemen are deployed where more crime is present. Or It could be more policemen and crime depends upon the size of zones. Bigger zones have more crime and more force. [2]

iii)

Calculation of Sample correlation coefficient using Pearson's Method:

X be Policemen and Y be Cases

$$\sum x = 1212 \ , \sum y = 934 \, , \sum x^2 = 178000 \ \sum y^2 = 120476 \ \sum xy = 140790$$

S_{xx} = 14784.00
S_{yy} = 23547.56
S_{xy} = 15011.33

$$r = \frac{Sxy}{\sqrt{Sxx \quad Syy}} \quad = 0.8045$$

Test whether is no correlation:

$$H_0\colon \rho = 0 \qquad vs \ H_1\colon \rho \neq 0$$

Under H_0:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Observed value of test statistic is

$$\frac{0.8045\sqrt{9-2}}{\sqrt{1-0.8045^2}} = 3.58$$

Upper 0.5 % point of t distribution t_7 distribution = 3.499 < 3.584 (observed value). Thus, we have sufficient evidence to reject H_0 at 1% level. Thereby, it indicates there is strong correlation between policemen and cases. [6]

[12 Marks]

Solution 9: **i)**

We need to find the parameters of the Gamma distribution, say α and λ

Then

$$\frac{E(X)}{Var(X)} = \frac{\alpha/_{\lambda}}{\alpha/_{\lambda^2}} = \lambda = \frac{50}{25} = 2$$

And hence $\alpha = E(X) * \lambda = 50 * 2 = 100$

The posterior distribution is given by:

$$f(\theta_1|x) \propto f(x|\theta_1) * f(\theta_1)$$
$$\propto \left(\prod_{j=1}^{5} e^{-\theta_1}\,\theta_1^{n_{1j}}\right) * \theta_1^{\alpha-1} e^{-\lambda\theta_1}$$
$$\propto e^{-(\lambda+5)\theta_1}\theta_1^{\alpha+\sum_{j=1}^{5} n_{1j}-1}$$

Which is the pdf of a gamma distribution with parameters

$$\alpha + \sum_{j=1}^{5} n_{1j} = 100 + 232 = 332$$

And $\lambda + 5 = 7$

Under quadratic loss the Bayes estimate is the mean of the posterior distribution. So we have an estimate of 332/7 = 47.43 [5]

ii)

We have

$$\bar{n}_1 = \frac{232}{5}$$
$$\bar{n}_2 = \frac{260}{5}$$
$$\bar{n}_3 = \frac{145}{5}$$

This gives $\bar{n} = \frac{46.4+52+29}{3} = 42.4667$

$$\sum_{j=1}^{5}(n_{1j} - \bar{n}_1)^2 = \sum_{j=1}^{5} n_{1j} - 2\sum_{j=1}^{5} n_{1j} * \bar{n}_1 + 5 * \bar{n}_1^2$$

= 11,434 – 2*232 * 46.4 + 5*46.4^2

= 669.2

Similarly,

$$\sum_{j=1}^{5}(n_{2j} - \bar{n}_2)^2 = \sum_{j=1}^{5} n_{2j} - 2\sum_{j=1}^{5} n_{2j} * \bar{n}_2 + 5 * \bar{n}_2^2$$

= 14028 – 2* 260 * 52.0 + 5*52^2

= 508

$$\sum_{j=1}^{5}(n_{3j} - \bar{n}_3)^2 = \sum_{j=1}^{5} n_{3j} - 2\sum_{j=1}^{5} n_{3j} * \bar{n}_3 + 5 * \bar{n}_3^2$$

= 4399 – 2* 145 * 29.0 + 5*29^2

= 194

So

E(s^2(θ)) = $\frac{1}{3} * \frac{1}{4}$ * (669.2 + 508 + 194) = 114.2667

Var(m(θ)) = $\frac{1}{2}$ * ((46.2-42.4667)^2+ (52-42.4667)^2 + (29-42.4667)^2) - $\frac{1}{5} * 114.2667$

= 121

So

$Z = \frac{5}{5+\frac{114.2667}{121}} = 0.8411$

So expected claims for next year are:
Cat 1 0.1589 × 42.4667 + 0.8411 × 46.4 = 45.78
Cat 2 0.1589 × 42.4667 + 0.8411 × 52 = 50.49
Cat 3 0.1589 × 42.4667 + 0.8411 × 29 = 31.14

[6]

iii)

The main differences are that:

- The approach under (i) makes use of prior information about the distribution of θ_1 whereas the approach in (ii) does not.
- The approach under (i) uses only the information from the first category to produce a posterior estimate, whereas the approach under (ii) assumes that information from the other categories can give some information about category 1.
- The approach under (i) makes precise distributional assumptions about the number of claims (i.e. that they are Poisson distributed) whereas the approach under (ii) makes no such assumptions. [2]

iv)

The insurance policies were newly introduced 5 years ago, and it is therefore likely that the volume of policies written has increased (or at least not been constant) over time. The assumption that the number of claims has a Poisson distribution with a fixed mean is therefore unlikely to be accurate, as one would expect the mean number of claims to be proportional to the number of policies. Let P_{ij} be the number of policies in force for risk i in year j.

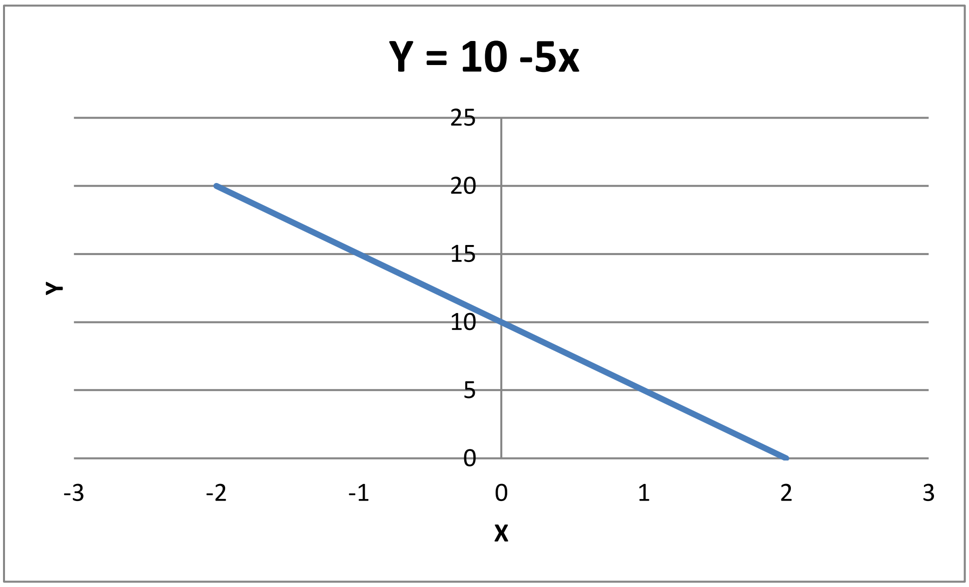
Then the models can be amended as follows: The approach in (i) can be taken assuming that that the mean number of claims in the Poisson distribution is $P_{ij}\theta_i$. The approach in (ii) can be generalised by using EBCT Model 2 which explicitly incorporates an adjustment for the volume of risk

[2]

[15 Marks]

Solution 10: **i)**

X	-2	-1	0	1	2
Y	20	15	10	5	0



[2]

ii)

a)

Assumption of linear model does not seem valid in this case. The histogram shows a positively skewed distribution for the residuals and suggests that the errors ~N(0,б^2) distribution does not fit in this case. [2]

b)

The plot in (i) suggests a negative relationship between Y and X with slope parameter = -5. However, the correlation shows a positive relationship. Thus, it seems an error has been made in either fitting the model or while computing correlation. [2]

iii)

For exponential family, we have write in the form:

$$g(y) = exp\left[\frac{y\theta - b(\theta)}{a(\emptyset)} + c(y, \emptyset)\right]$$

Poisson Distribution is:

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}$$

$$f(y) = exp\left[\frac{ylog\mu - \mu}{1} - \log y!\right]$$

Where:

$$b(\theta) = \mu, \theta = log\mu\,, a(\emptyset) = 1\,, c(y, \emptyset) = -\log y!$$ [2]

iv)

a)

Log Likelihood Function is:

$$Log\ L = \sum y_i log\mu_i - \sum \mu_i - \sum \log y_i!$$

Model 1 $\log \mu_i = \alpha$

$$Log\ L = \alpha \sum_{i=1}^{20} y_i - 20e^{\alpha} - \sum_{i=1}^{20} \log y_i! = 18\alpha - 20e^{\alpha} - \sum_{i=1}^{20} \log y_i! \quad -(*)$$

Differentiating this with respect to α , and setting the result equal to 0, we get

$$18 - 20e^{\hat{\alpha}} = 0 \rightarrow \hat{\alpha} = \log\left(\frac{18}{20}\right) = -0.1054$$

Model 2 : $\log \mu_i = \alpha + \beta\, x_i\ where\ xi = \begin{cases} 1\ for\ City\ I \\ 0\ for\ City\ II \end{cases}$

$$Log\, L = \alpha \sum_{i=1}^{20} y_i + \beta \sum_{i=1}^{20} y_i x_i - 10e^{\alpha} - 10e^{\alpha+\beta} - \sum_{i=1}^{20} \log y_i!$$

[5]

$$= 18\alpha + 6\beta - 10e^{\alpha} - 10e^{\alpha+\beta} - \sum_{i=1}^{20} \log y_i! - (**)$$

Differentiating this in turn with respect to α and β and setting it equal to 0, we get

$$18 - 10\, e^{\hat{\alpha}} - 10\, e^{\hat{\alpha}+\widehat{\beta 1}} = 0 \qquad (*)$$

$$6 - 10\, e^{\hat{\alpha}+\hat{\beta}} = 0 \rightarrow 6 = 10\, e^{\hat{\alpha}+\hat{\beta}} \quad (**)$$

Substituting this in (*), we get

$$18 - 10\, e^{\hat{\alpha}} - 6 = 0 \rightarrow \quad \hat{\alpha} = \log\left(\frac{12}{10}\right) = 0.1823$$

Substituting $\hat{\alpha}$ *in* $(**)$, *we get*

$$6 = 10\, e^{0.1823+\hat{\beta}} \rightarrow \quad \hat{\beta} = \ln\left(\frac{6}{12}\right) = -0.6932$$

b)

City 1 : $\log \mu_{City1} = \alpha + \beta \rightarrow \mu_{City1} = 0.6$

City 2 : $\log \mu_{City2} = \alpha \rightarrow \mu_{City2} = 1.2$

P(cancellation =3) for City 1 = $\frac{e^{-0.6} 0.6^3}{3!}$ = 0.0126

P(cancellation =3) for City 1 = 0.0867 [2]

v)

Scaled Deviance = 2 (log Ls – Log Lm)

Where LogLs is the value of log likelihood function for the saturated model

And Log Lm is the value of log likelihood function for Model M

For Saturated model, $\mu_i = y_{i,}$. This implies

$$Log\, L = \sum y_i log y_i - \sum y_i - \sum \log y_i! = -13.84$$

Using iv.a. (*) and (**),

$$Log\, L(model\ 1) = 18\hat{\alpha} - 20e^{\hat{\alpha}} - \sum_{i=1}^{20} \log y_i!$$

$$= 18 * -0.1054 - 20\, ex(-0.1054) - \sum_{i=1}^{20} \log y_i!$$

$$= -24.055$$

Similarly,

$$Log\, L(Model\ 2) = 18\hat{\alpha} + 6\hat{\beta} - 10e^{\hat{\alpha}} - 10e^{\hat{\alpha}+\hat{\beta}} - \sum_{i=1}^{20} \log y_i!$$

putting $\hat{\alpha} = 0.1823$ *and* $\hat{\beta} = -0.6932$ in above equations

logL (model2) = -23.036

Scaled deviance model 1 = 20.43

Scaled deviance model 2 = 18.39

[5]

vi)

We can compare Model 1 and 2 by using chi-square distribution and scaled deviance difference.

Scaled deviance difference = 2[LogL(model 1) – LogL(model 1)) = 2.04

This follows chi-square distribution with 2-1 =1 degrees of freedom

At 5% level of confidence, value is 3.84.

No significant improvement. Thus, prefer Model 1 [2]

vii)

a)

Model 3: $\log \mu_i = \begin{cases} \delta \; for\ City\ I \\ \gamma \; for\ City\ II \end{cases}$

So Log Likelihood function is:

$$Log\ L = \delta \sum_{i=1}^{10} y_i + \gamma \sum_{i=11}^{20} y_i - 10e^{\gamma} - 10e^{\delta} - \sum_{i=1}^{20} \log y_i!$$
$$= 6\delta + 12\gamma - 10e^{\gamma} - 10e^{\delta} - \sum_{i=1}^{20} \log y_i!$$

Differentiating this , and setting the result equal to 0, we get

$$6 - 10e^{\hat{\delta}} = 0 \rightarrow \hat{\delta} = \log\left(\frac{6}{10}\right) = -0.5108$$
$$12 - 10e^{\hat{\gamma}} = 0 \rightarrow \hat{\gamma} = \log\left(\frac{12}{10}\right) = 0.1823$$

[2]

b)

Scaled Deviance = 18.39

AIC = -2 * LogL (model3) + 2 X number of paramters = 50.72 [1]

viii)

Model 3 and Model 2 are essential the same but represented in different way.

Under Model 2, exp(α) represents mean of City II and exp(α + β) represents City I. In Model 3, α + β is given as .

In other words, exp(β) under Model 2 is expressed as change in mean between City I and City II. In Model 3, mean of the 2 cities are expressed separately.

Thus, it can be observed that Scaled Deviance for Model 2 and 3 are same.

[2]

ix) [2]

Pearson residuals is defined as: **[29 Marks]**

$$\frac{y_i - \widehat{y_\iota}}{\sqrt{\widehat{y_\iota}}}$$

As variance equals mean for poison distribution.

Pearson residual distribution is often skewed for non-normal distributed data. This makes the interpretation of residual plots difficult.

****