

INSTITUTE OF ACTUARIES OF INDIA

CS1A - Actuarial Statistics

May 2026 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 examiners have given credit for any alternative approach or interpretation which they consider to be reasonable.

Solution 1 :

Correct Answer is Option **B**.

$$E(X) = C \cdot X(0) = 2 + 6 \cdot 0 = 2.$$

$$\text{Var}(X) = C \cdot X(0) = 6.$$

[2 marks]

Solution 2:

Correct Answer is Option **E**.

Probability that a player is the final winner

= probability that he wins in all rounds

= (probability of winning round 1) (probability of winning round 2) (probability of winning round $n-2$) (probability of winning round $n-1$)

$$= (n-1)/n \cdot (n-2)/(n-1) \cdot \dots \cdot 2/3 \cdot 1/2$$

$$= 1/n.$$

Alternatively, there a total of n players playing the game and only one player will win. Considering this, the probability that an individual wins is $1/n$.

[2 marks]

Solution 3:

Correct Answer is Option **C**.

Let us consider each statement:

1. UPSC mains score will also be affected by number of hours of study per week and by the intercept parameter. So, this statement is incorrect.
2. UPSC mains score will increase by 3 marks for each additional hour of study per week put by the aspirant holding number of mock tests to be constant. So, this statement is also incorrect.
3. UPSC mains score increases by 2 marks for every additional mock test given by the aspirant, holding the weekly hours of study to be constant. This statement is correct.
4. UPSC mains score increases by 2 marks for every additional mock test given by the aspirant only when the weekly hours of study is held constant. If the number of hours of study per week also increases, then it will also lead to increase in the UPSC mains score by 3 marks for every additional hour of weekly study. Hence, this statement is not correct.

Even if a candidate has not appeared for any mock test, the candidate will score 150 marks plus 3 times the number of hours of weekly study. Hence, this statement is factually incorrect.

[2 marks]

Solution.4

Correct Answer is Option **D**.

We want $E[B|B \geq 1]$

For each b , sum over all $a = 0, 1, 2, 3, 4$.

For $b = 1$:

$$P(B = 1) = \sum_{a=0}^4 \frac{a+3}{75} = \frac{(0+1+2+3+4) + 5 \cdot 3}{75} = \frac{10+15}{75} = \frac{25}{75}$$

For $b = 2$:

$$P(B = 2) = \sum_{a=0}^4 \frac{a+6}{75} = \frac{10+30}{75} = \frac{40}{75}$$

Thus:

$$P(B \geq 1) = \frac{25}{75} + \frac{40}{75} = \frac{65}{75}$$

conditional probabilities:

$$P(B = 1 | B \geq 1) = \frac{25/75}{65/75} = \frac{25}{65}$$

$$P(B = 2 | B \geq 1) = \frac{40/75}{65/75} = \frac{40}{65}$$

$$E[B|B \geq 1] = 1 \cdot \frac{25}{65} + 2 \cdot \frac{40}{65} = \frac{105}{65} \approx 1.62$$

[2 marks]

Solution.5

Correct Answer is Option **B**.

$$\text{Standard deviation} = s = 400,000^{1/2} = 632.45$$

$$\text{Standard error} = s / \sqrt{n} = 632.45 / (36^{1/2}) = 105.41$$

[2 marks]

Solution.6

Correct Answer is Option **D**.

Y follows a lognormal distribution.

$$E[Y] = \exp(\mu + \frac{1}{2} \sigma^2) = \exp(\mu + 0.08) = \exp(\beta_0 + \beta_1 X + 0.08)$$

With the flat at the second floor,

$$E[Y] = \exp(\beta_0 + 2 \cdot 0.5 + 0.08)$$

With the flat at the fourth floor,

$$E[Y] = \exp(\beta_0 + 4 \cdot 0.5 + 0.08)$$

If we calculate the increase,

Increase

$$= \exp(\beta_0 + 2 + 0.08) - \exp(\beta_0 + 1 + 0.08)$$

$$= \exp(\beta_0 + 1.08) * (\exp(1) - 1)$$

% Increase

$$= \exp(\beta_0 + 1.08) * (\exp(1) - 1) / \exp(\beta_0 + 1.08)$$

$$= 2.7182 - 1$$

$$= 1.72 = 172\%$$

[2 marks]

Solution.7

Correct Answer is Option A.

$MU(t)$

$$= E[e^{t(X-Y)}]$$

$$= E[e^{tX-tY}]$$

$$= M_{X,Y}(t, -t) \dots \dots \text{Using } M_{X,Y}(s_1, s_2) = E[\exp(s_1X + s_2Y)]$$

[2 marks]

Solution.8

Correct Answer is Option C.

Let number of patients with rare eye condition be X

$$X \sim \text{Binomial}(n=2000, p=0.02).$$

$$\text{Mean} = 2000 \cdot 0.02 = 40$$

$$\text{Var} = 2000 \cdot 0.02 \cdot 0.98 = 39.2$$

$$\text{SD} = \sqrt{39.2} = 6.26$$

Using Central Limit Theorem,

$$X \sim N(\mu=40, \sigma^2=39.2)$$

$$P(X > 50)$$

$$\begin{aligned}
 &\approx P(X > 50.5), \\
 &= P(Z > (50.5 - 40)/6.26) \\
 &= P(Z > 1.68) \\
 &= 0.046 \quad \dots\dots\dots \text{from standard normal tables.} \\
 &= 0.05.
 \end{aligned}$$

[2 marks]**Solution.9**

Correct Answer is Option **E**.

Let X be the number of rolls needed to get the first '6'.

$$X \sim \text{Geometric}(p = 1/6)$$

$$P(X = 3)$$

$$= p * (1 - p)^{3-1}$$

$$= 1/6 * (5/6)^2$$

$$= 25 / 216.$$

[2 marks]**Solution.10**

Correct Answer is Option **B**.

Let us consider each of the pairs:

1. Poisson distribution – log link function – correctly matched
2. Gamma distribution – inverse link function – correctly matched
3. Binomial distribution – identity link function – incorrect
4. Normal distribution – logit link function – incorrect

Canonical link functions for binomial and normal distributions is logit and identity link functions respectively.

[2 marks]**Solution.11**

Correct Answer is Option **B**.

$$E(S)$$

$$= E(N) * E(X)$$

$$= 2.70 * 7,350$$

$$= 19845.$$

$$\text{Var}(S)$$

$$= E(N) \text{Var}(X) + \text{Var}(N) [E(X)]^2$$

$$= 2.70 \cdot 51202 + 2.70 \cdot 73502$$

$$= 216639630$$

$$\text{SD}(S)$$

$$= 14718.683.$$

Required Premium

$$= 19845 + 30\% * 14718.683$$

$$= 24,261.$$

[2 marks]

Solution.12

Correct Answer is Option C.

$$(\bar{X} - \mu) / (S / \sqrt{n}) \sim t_{n-1}$$

$$(\bar{X} - 0) / (S / \sqrt{9}) \sim t_8$$

$$3\bar{X} / S \sim t_8$$

$$P(\bar{X} > S)$$

$$= P(\bar{X} / S > 1)$$

$$= P(3\bar{X} / S > 3)$$

$$= P(t_8 > 3)$$

$$= 0.89\% \dots\dots \text{Interpolating between } P(t_8 > 2.896) = 1\% \text{ and } P(t_8 > 3.355) = 0.5\%$$

[2 marks]

Solution.13

Correct Answer is Option D.

From the graph it is clear that XS has the lowest variance and XJ has the highest variance, XB lies in between.

$$P(XS > 6) = P(ZS > 1/\sigma_S)$$

$$P(XB > 6) = P(ZB > 1/\sigma_B)$$

$$P(XJ > 6) = P(ZJ > 1/\sigma_J)$$

We know that:

$$1/\sigma_S > 1/\sigma_B > 1/\sigma_J$$

It follows that:

$$P(Z_S > 1/\sigma_S) < P(Z_B > 1/\sigma_B) < P(Z_J > 1/\sigma_J)$$

$$P(X_S > 6) < P(X_B > 6) < P(X_J > 6) \text{ i.e.}$$

$$P(X_J > 6) > P(X_B > 6) > P(X_S > 6)$$

In simple words, the lower the variance, the lower will be the probability of time exceeding 6 hours and higher the variance, higher will be the probability of time exceeding 6 hours.

[2 marks]

Solution.14

Correct Answer is Option A.

Mean of posterior distribution

$$= (\alpha + x) / (\alpha + x + \beta + n - x)$$

$$= (\alpha + x) / (\alpha + \beta + n)$$

$$= (2 + x) / 18$$

Mean of prior distribution

$$= \alpha / (\alpha + \beta)$$

$$= 2 / 6$$

$$\text{Sample Mean} = (x/12)$$

Posterior Mean

$$= Z (x/12) + (1 - Z) \text{ Prior Mean}$$

$$= (12/18) (x/12) + (6/18) (2/6)$$

Credibility Factor

$$= 12/18$$

$$= 0.6667.$$

[2 marks]

Solution.15

Correct Answer is Option E.

$$\begin{aligned}
& f_X(x) \\
&= \int_0^{20} f_{X,P}(x,p) dp \\
&= \int_0^{20} (x+p)/300 dp \\
&= 1/300 \left[\int_0^{20} x dp + \int_0^{20} p dp \right] \\
&= 1/300(20x+200) \\
&= (20x+200)/300 \\
&= (x+10)/15.
\end{aligned}$$

[2 marks]

Solution.16

i) The probability density function of the exponential distribution is:

$$f(t | \lambda) = \lambda e^{-\lambda t}, \quad t > 0$$

Since the observations $T_1, T_2, \dots, T_n$ are independent, the likelihood is the product of the individual densities:

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda t_i} \quad L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n t_i}$$

Thus the likelihood function is

$$L(\lambda) \propto \lambda^n e^{-\lambda \sum t_i}$$

[2]

ii) Likelihood from part (i):

$$L(\lambda | \mathbf{t}) \propto \lambda^n \exp\left(-\lambda \sum_{i=1}^n t_i\right)$$

The prior distribution is

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$

The prior density of λ is:

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0.$$

Ignoring normalising constants that do not depend on λ , the prior can be written as:

$$\pi(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}.$$

Using Bayes' theorem:

Posterior $\propto$ Likelihood $\times$ Prior

Substituting:

$$\begin{aligned}\pi(\lambda \mid \mathbf{t}) &\propto L(\lambda \mid \mathbf{t}) \pi(\lambda) \\ &\propto \lambda^n e^{-\lambda \sum t_i} \times \lambda^{\alpha-1} e^{-\beta\lambda}.\end{aligned}$$

Combine terms:

$$\pi(\lambda \mid \mathbf{t}) \propto \lambda^{\alpha+n-1} \exp \left[-(\beta + \sum_{i=1}^n t_i) \lambda \right].$$

This has the same functional form as the density of a Gamma distribution.

Therefore, the posterior distribution of λ is:

$$\lambda \mid t_1, \dots, t_n \sim \text{Gamma} \left(\alpha + n, \beta + \sum_{i=1}^n t_i \right)$$

[4]

iii) Given:

- Number of observations: $n = 20$
- $\sum_{i=1}^{20} t_i = 40$
- Prior distribution:

$$\lambda \sim \text{Gamma}(\alpha, \beta), \quad \alpha = 2, \beta = 1$$

Under quadratic loss, the Bayesian estimator is the posterior mean.

For a Gamma(shape, rate) distribution:

Mean = shape / rate

Posterior distribution:

Gamma ($\alpha + n$, $\beta + \sum t_i$)

Substituting the given values:

$$\lambda|t \sim \text{Gamma}(2 + 20, 1 + 40) = \text{Gamma}(22, 41)$$

Thus

$$E[\lambda] = \alpha/\beta$$

$$\hat{\lambda} = \frac{22}{41} \approx 0.54$$

[2]

iv)

- Bayesian estimate under all-or-nothing loss is the mode of the posterior distribution. At point A the value of the gamma density function is the highest. Hence, the x-coordinate corresponding to Point A i.e. XA represents the mode of the posterior distribution. Hence, the statement is correct. XA represents the Bayesian estimate under all or nothing loss.
- Bayesian estimate under absolute error loss is the median of the posterior distribution. It is that which exactly divides the density function into two halves. A perpendicular drawn from Point B on X-axis divides the density function into two halves. Hence XB represents the median of the posterior distribution. Hence, the statement is correct. XB represents the Bayesian estimate under absolute error loss.

[2]

[10 marks]

Solution.17

i) Sample Mean

$$\bar{X}$$

$$= ((0 \times 280) + (1 \times 150) + (2 \times 50) + (3 \times 15) + (4 \times 4) + (5 \times 1)) / 500$$

$$= 0.632 \text{ (1)}$$

Sample Variance

$$= S^2$$

$$= ((0 - 0.632)^2 \times 280 + (1 - 0.632)^2 \times 150 + (2 - 0.632)^2 \times 50 + (3 - 0.632)^2 \times 15 + (5 - 0.632)^2 \times 1) / (500 - 1)$$

$$= 0.750 \text{ (1)}$$

[2]

ii)

$$\text{So, } \hat{\lambda}_{\text{MOM}} = \bar{X} = 0.632.$$

Expected frequencies under Poisson($\lambda=0.632$):

- 0 passengers: $500 \times e^{-0.632} = 500 \times 0.531 = 265.5 \approx 266$
- 1 passenger: $500 \times (e^{-0.632} \times 0.632) = 500 \times 0.336 = 168$
- 2 passengers: $500 \times 0.106 = 53$
- 3 passengers: $500 \times 0.022 = 11$

- 4 passengers: $500 \times 0.0035 = 2$
- ≥ 5 passengers: $500 \times 0.000498 = 0.25 \approx 0$

[2]

iii)

(i) Sample mean and sample variance

- Total hours

$$N = 500$$

- Treat " ≥ 5 " as 5

$$\sum fx = 316 \Rightarrow \bar{x} = \frac{316}{500} = 0.632$$

- Sample variance (from data):

$$s^2 = 0.750$$

(ii) Poisson model using method of moments

- For Poisson distribution:

$$\lambda = \bar{x} = 0.632$$

- Expected frequencies using

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Passengers	0	1	2	3	4	≥ 5
Observed	280	150	50	15	4	1
Expected	266	168	53	11	2	0

- Expected frequency for ≥ 5 is less than 1 and hence taken as approximately 0
- Total expected frequency = 500

(iii) Suitability of the Poisson model

- As can be seen from the above table, there are differences between the actual observed no of hours and the fitted number of hours as per the Poisson Model. This indicates that Poisson Model may not be a good fit for the data.
- Poisson distribution assumes that mean = variance = parameter λ . But in case of the observed data, sample variance (0.750) is greater than the sample mean (0.632). Hence, a Poisson model is not an appropriate fit for the data.
- A chi-square goodness-of-fit test is performed to justify this conclusion:

(iii a) Goodness-of-fit assessment (Chi-square test)

To satisfy chi-square test conditions, upper classes are combined.

Grouped data

Passengers	0	1	2	≥ 3	Total
Observed (O)	280	150	50	20	500

Expected frequencies

Using Poisson

$$\begin{aligned}
 E(0) &= 500 \times P(0) = 265.76 \\
 E(1) &= 500 \times P(1) = 167.96 \\
 E(2) &= 500 \times P(2) = 53.08 \\
 (\lambda = 0.632) E(\geq 3) &= 500 \times [1 - (P_0 + P_1 + P_2)] = 13.20
 \end{aligned}$$

Chi-square statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(280-265.76)^2}{265.76} + \frac{(150-167.96)^2}{167.96} + \frac{(50-53.08)^2}{53.08} + \frac{(20-13.20)^2}{13.20} = 6.37$$

Degrees of freedom and decision

$$\text{df} = \text{number of classes} - 1 - \text{parameters estimated} = 4 - 1 - 1 = 2$$

Critical value at 5% level:

$$\chi_{0.05,2}^2 = 5.99$$

Since:

$$6.37 > 5.99$$

the null hypothesis is rejected.

Final Conclusion

- The sample mean (0.632) and variance (0.750) are not equal, violating a key Poisson assumption.
- The chi-square goodness-of-fit test rejects the Poisson model at the 5% significance level.
- Therefore, **the Poisson distribution is not an appropriate model for this dataset.**

[2]

iv) Using the bootstrap samples, the expected number of hours corresponding to each number of passengers would be simply the mean number of hours based on the 5 sample values corresponding to each number of passengers. This is shown in the below table:

No of passengers		0	1	2	3	4	≥ 5	Total
No of hours	Sample B1	278	152	49	14	5	2	500
	Sample B2	285	145	52	13	4	1	500
	Sample B3	276	154	51	15	3	1	500
	Sample B4	282	148	50	16	3	1	500
	Sample B5	279	151	48	15	6	1	500
Expected Number of Hours (Bootstrap)		280	150	50	14.6	4.2	1.2	500

[2]

v)

No of passengers	0	1	2	3	4	≥ 5	Total
No of hours (Actual)	280	150	50	15	4	1	500
No of hours (Poisson Model)	266	168	53	11	2	1	500
No of hours (Bootstrap)	280	150	50	14.6	4.2	1.2	500

If we compare the expected number of hours using the bootstrap samples, they are very close to the actually observed number of hours based on the observed data.

They are not close to the expected number of hours as determined using the Poisson model.

If a parametric approach (with Poisson distribution and parameter = 0.632) would have been used for creation of bootstrap samples, the expected number of hours as per the bootstrap samples would have been very close to the expected number of hours determined using the Poisson model in part (ii).

So, it is safe to conclude that a non-parametric approach was used for generating bootstrap samples.

[2]

[10 marks]

Solution.18

i)

$$S_{xx} = (\sum x^2 - (\sum x)^2 / n) = 2310$$

$$\bar{x} = \frac{770}{10} = 77 \quad \bar{y} = \frac{35}{10} = 3.5 \quad S_{xy} = \sum xy - n \bar{X} \bar{Y}$$

$$S_{xy} = -225$$

$$b = S_{xy}/S_{xx} = -0.09740$$

$$a = \bar{y} - b\bar{x}$$

$$a = 3.5 - (-0.09740)(77) = 10.9998 \approx 11.0$$

So, the fitted regression line is:

$$= 11 - 0.09740x.$$

[2]

ii)

H0: $\beta = 0$ (no linear relationship hypothesis)

H1: $\beta \neq 0$ (1)

$$SS_{REG} = S^2_{xy} / S_{xx} = (-225)^2 / 2310 = 21.92$$

$$SSTOT = S_{yy} = \sum y^2 - (\sum y)^2 / n = 22.50$$

$$SS_{RES} = SSTOT - SS_{REG} = 22.50 - 21.92 = 0.58 \text{ (1)}$$

Following is the ANOVA table: (1)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression	1	21.92	21.92
Residual (Error)	8	0.58	0.073
Total	9	22.50	—

Under H0, we have

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} = \frac{21.92}{0.073} \approx 300$$

on (1, 8) degrees of freedom.

As $300 > 11.26$ (critical value at 1%), so H0 is rejected at 1% level of significance.

Hence, the slope of the regression line is significantly different from zero.

[4]

iii) For the standard model,

$$R^2 = SS_{REG} / SSTOT$$

$$= 21.92 / 22.50$$

$$= 97.40\%$$

- The standard linear regression model explains **97.4%** of the variation in the number of minor survival incidents.
- This R² value is **higher than that of all four transformed models**.
- Therefore, the standard linear regression model provides the **best fit** to the orangutan rehabilitation data among the models considered.

Conclusion:

The standard linear regression model is preferred, as it has the highest coefficient of determination and captures the relationship between training hours and minor incidents most effectively.

[2]

iv) Using the **linear regression model obtained in part (i)**, the fitted (predicted) number of minor incidents for each orangutan is obtained by substituting the value of training hours into the estimated regression equation.

For orangutan E:

- Training hours: $x = 75$
- Observed number of minor incidents: $y = 4$

From the fitted regression model obtained in part (i), the **predicted value** for orangutan E is:

$$\hat{y} = 3.695$$

The **residual** is defined as:

$$\text{Residual} = y - \hat{y}$$

Substituting the values:

$$\text{Residual} = 4 - 3.695 = \mathbf{0.305}$$

[2]

[10 marks]

Solution.19

i) $n = 300$

Mean distance from centre to maximum height = 19 metres

Standard deviation of distance from centre to maximum height = 3 metres

$$\bar{X} = 19 + 22 = 41$$

$S = 3$ (standard deviation is unaffected if a constant i.e. 22 is added to all the values)

95% confidence interval for maximum height

$$= \bar{X} \pm t_{2.5\%, 299} * S / \sqrt{n}$$

$$= 41 \pm 1.96 * 3 / \sqrt{300}$$

$$= (40.66 \text{ metres}, 41.34 \text{ metres})$$

[2]

ii) 95% prediction interval for maximum height

$$\begin{aligned}
 &= \bar{X} \pm t_{2.5\%, 299} * S * \sqrt{1 + 1/n} \\
 &= 41 \pm 1.96 * 3 * \sqrt{1 + 1/300} \\
 &= (35.11 \text{ metres}, 46.89 \text{ metres})
 \end{aligned}$$

95% prediction interval is wider than the 95% confidence interval as the prediction interval is for a single observation and hence it not only accounts for the variability from the mean but also for the variability among the individual data points.

[3]

iii) On the restructured wheel,

$$\bar{X} = \mu = 35$$

$$S = \sigma = 2$$

$$P(X > 36)$$

$$= P(Z > (36-35)/2)$$

$$= P(Z > 0.50)$$

$$= 31\%$$

As $31\% > 5\%$, the restructured wheel does not comply with the new regulatory guideline at the 5% level of significance.

[2]

iv) S^2_p

$$= ((300 - 1) 32 + (300 - 1) 22) / (300 + 300 - 2)$$

$$= 6.50 (1)$$

95% confidence interval for difference in mean maximum height of the original and the restructured giant wheel

$$= (\bar{X}_1 - \bar{X}_2) \pm t_{2.5\%, 598} * S_p * \sqrt{1/n_1 + 1/n_2}$$

$$= (41 - 35) \pm 1.96 \sqrt{6.50} \sqrt{1/300 + 1/300}$$

$$= (5.59 \text{ metres}, 6.41 \text{ metres})$$

[3]

[10 marks]

Solution.20

i) REM, DPM and MDT are numerical variables.

MRT and EMP are factor variables.

[1]

ii) We will have

- 1 parameter for the intercept
- 1 parameter for each numerical variable viz. REM, DPM, MDT

- (4-1) i.e. 3 parameters for MRT

(5-1) i.e. 4 parameters for EMP

Model 1: REM + DPM

No of parameters

$$= 1 + (1 + 1)$$

$$= 3 (1)$$

Model 2: REM + DPM + MDT + MRT

No of parameters

$$= 1 + (1+1+1+3)$$

$$= 7 (1)$$

Model 3: REM + DPM + MDT + MRT + MDT:MRT

No of parameters

$$= 1 + (1+1+1+3) + (1*3)$$

$$= 7 + 3$$

$$= 10 (1)$$

Model 4: REM + DPM + MDT + MRT + EMP + MDT:MRT + MRT:EMP + MDT:EMP

No of parameters

$$= 1 + (1+1+1+3+4) + (1*3) + (3*4) + (1*4)$$

$$= 11 + 3 + 12 + 4$$

$$= 30 (1)$$

Model 5: MRT*EMP*REM

Let us split this into three:

1. Main effects
2. 2-way interactions
3. 3-way interactions

No of parameters for main effects

$$= (1+3+4+1)$$

$$= 9$$

Possible two-way interactions are:

$$\text{MRT:EMP} + \text{MRT:REM} + \text{EMP:REM}$$

No of parameters for 2-way interactions

$$= (3*4) + (3*1) + (4*1)$$

$$= 19$$

Possible three-way interactions are:

$$\text{MRT:EMP:REM}$$

No of parameters for 3-way interactions:

$$= (3*4*1)$$

$$= 12$$

Total number of parameters

$$= 9 + 19 + 12$$

$$= 40 \text{ (3)}$$

[7]

iii) A model (x) is a significant improvement over the previous model (x-1) if the difference in scaled deviance for model (x-1) and model (x) is greater than the chi-square critical value at 5% level of significance.

- Model 1 is a significant improvement over Model 0 (as $8 > 5.99$)
- Model 2 is a significant improvement over Model 1 (as $11 > 9.49$)
- Model 3 is a significant improvement over Model 2 (as $9 > 7.82$)
- Model 4 is not a significant improvement over Model 3 (as $20 < 31.41$)
- Model 5 is not a significant improvement over Model 4 (as $17 < 18.31$)

So, Model 3 is the best fit at the 5% level of significance.

[2]

[10 marks]

Solution.21

i) Type I error (False Positive): Genuine claim flagged as fraudulent i.e. unnecessary investigation costs and customer dissatisfaction.

Type II error (False Negative): Fraudulent claim not flagged i.e. insurer pays out fraud, financial loss.

[1]

ii)

- Fraudulent claims: $10,000 \times 0.08 = 800$
- Genuine claims: $10,000 - 800 = 9,200$

- True positives: $800 \times 0.85 = 680$
- False positives: $9,200 \times 0.10 = 920$

[1]

iii) Under H_0 , $p=0.80$, the expected number of correctly flagged fraudulent claims is:

$$E(\text{fraud flagged}) = 800 \times 0.80 = 640$$

Expected number of missed fraudulent claims:

$$E(\text{fraud missed}) = 800 - 640 = 160$$

So under H_0 :

$$\text{Fraud flagged} = 640$$

$$\text{Fraud missed} = 160$$

Observed values:

$$\text{Fraud flagged} = 680$$

$$\text{Fraud missed} = 120$$

Chi-Square test statistic for observed vs expected counts:

$$\begin{aligned} \chi^2 &= (680 - 640)^2 / 640 + (120 - 160)^2 / 160 \\ &= 12.5 \end{aligned}$$

We have 2 categories (fraud flagged vs fraud missed),

So degrees of freedom = 1.

From chi-square tables: $\chi^2(0.01, 1) = 6.635$ (Tables page 169)

Since $12.5 > 6.635$, we reject H_0 . At the 1% level, the chi-square test shows strong evidence that the fraud detection system's sensitivity is greater than 80%. The system is statistically effective at catching fraud.

[4]

iv) Sample proportion

$$\hat{p} = 680 / 800 = 0.85$$

$$SE = (\hat{p}(1 - \hat{p}) / n)^{1/2} = 0.0126$$

95% confidence interval for sample proportion p

$$= \hat{p} \pm Z_{2.5\%} * SE$$

$$= 0.85 \pm 1.96 * 0.0126$$

$$= (0.825, 0.875)$$

The confidence interval shows that the true sensitivity is likely between 82.5% and 87.5%.

Because the whole interval lies above 80%, this backs up the insurer's claim that the system is performing better than the minimum standard.

In everyday terms, the evidence suggests the fraud detection system is reliably catching more than 80% of fraudulent claims, so it's working at a level the insurer can be confident in.

[4]
[10 marks]

Solution.22

i) Correct Answer is Option C.

A. Longitudinal data refers to repeated measurements over a period of time. The mission data is not longitudinal

B. Censored data has some values which are partially unknown. The mission data is not censored.

C. Cross-sectional data refers to measurements at one point of time relating to different aspects. The mission data can be referred to as cross-sectional data.

D. Truncated data refers to case where only certain observations are included due to selection criteria. The mission data is not truncated.

[2]

ii)

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{xx} = 15.824$$

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

$$S_{yy} = 327.541$$

$$S_{xy} = \sum xy - n \bar{X} \bar{Y}$$

$$S_{xy} = 8.2426$$

Karl Pearson's correlation coefficient:

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

$$= 0.11449237 \approx 0.115.$$

- The correlation coefficient $r \approx 0.115$ indicates a **very weak positive linear relationship** between launch mass and travel time.
- This suggests that **launch mass has little or no influence on travel time** for the Jupiter missions considered.

[2]

iii)

- a) In this context,
- Concordant pair (C) would mean a pair where higher number of instruments leads to higher mission cost or where lower number of instruments lead to lower mission costs
 - Discordant pair (D) would mean a pair where higher number of instruments leads to lower mission cost or where lower number of instruments leads to higher mission costs
- b) As the number of instances where higher number of instruments lead to higher mission costs is lower than the number of instances where higher number of instruments lead to lower mission costs, Kendal's correlation coefficient is likely to be negative.

Kendal's correlation coefficient

$$= (C - D) / (C + D)$$

As $C < D$, Kendal's correlation coefficient is likely to be negative.

- c) Adding space missions with higher number of high-cost state of art instruments would contribute to increasing the number of concordant pairs and hence would lead to improvement in the Kendal's correlation coefficient.

[3]

iv)

1. An outlier is simply a data point that sits noticeably away from the rest of the observations. Because a simple linear regression line is fitted by minimising squared errors, a point that is far from the overall pattern can pull the line towards itself.
2. If such a point is extreme in launch mass or travel time, it can have a strong "leveraging" effect. This means the fitted slope might tilt more steeply or become flatter than it should.
3. In this dataset, mission J10 has an unusually low travel time compared with the others. A point like that would tend to drag the regression line downward, giving a slope that doesn't really represent the general relationship seen in the rest of the missions.

[3]

[10 marks]
