INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

24th November 2023

Subject CS1A – Actuarial Statistics (Paper A)

Time allowed: 3 Hours 15 Minutes (10.15 – 13.30 Hours)

Total Marks: 100

INSTRUCTIONS TO THE CANDIDATES

1. ***Please read the instructions inside the cover page of answer booklet and instructions to examinees sent along with hall ticket carefully and follow without exception.***

2. ***Mark allocations are shown in brackets.***

3. ***Attempt all questions, beginning your answer to each question on a separate sheet.***

4. ***Please check if you have received complete Question Paper and no page is missing. If so, kindly get new set of Question Paper from the Invigilator.***

5. ***While attempting multiple choice questions, you are only required to state the right option number in the answer sheet. You are NOT required to give reasons / show supporting calculations to justify your choice.***

AT THE END OF THE EXAMINATION

Please return your answer book and this question paper to the supervisor separately. You are not allowed to carry the question paper in any form with you.

Q. 1) Let $M_x(t)$ denote the moment generating function (MGF) and $C_x(t)$ denote the cumulant generating function (CGF) of a random variable X.

i) Which of the following is TRUE for the relationship between $M_x(t)$ and $C_x(t)$?

A. $C_x(t) = e^{\,Mx(t)}$

B. $M_x(t) = e^{\,Cx(t)}$

C. $C_x(t) = \log_{10}(M_x(t))$

D. $M_x(t) = \log_{10}(C_x(t))$ (1)

The series expansion formula of moment generating function (MGF) for a random variable X is as follows:

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \frac{t^4}{4!}E(X^4) + \cdots$$

ii) Based on the above series expansion, show that the variance of random variable X is given by the following expression:

$$\text{var}(X) = M''_X(0) - [M'_X(0)]^2$$ (3)

iii) In terms of $C_x(t)$, how will you represent variance of random variable X? Choose the correct option from those given below:

A. $\text{var}(X) = C''_X(0) - [C'_X(0)]^2$

B. $\text{var}(X) = [C'_X(0)]^2$

C. $\text{var}(X) = C''_X(0)$

D. $\text{var}(X) = [C''_X(0)]^2$ (2)

[6]

Q. 2) The Inspector General of Police wants to evaluate the effectiveness of the lie-detector test carried out using a polygraph on those who are charged with the offence of first degree murder at all police station lock-ups in the country.

The police authorities have prepared the following matrix for consideration of the Inspector General based on the past records relating to 1,000 cases which have been subsequently decided by the judicial authorities:

Particulars	Persons found **Innocent** as per the **Lie-Detector Test**	Persons found **Guilty** as per the **Lie-Detector Test**
Persons acquitted by **Judicial Authorities** as **Innocent**	"Area **A**" **(✓–)**	"Area **B**" **(☒ +)**
Persons found **Guilty** of first degree murder by **Judicial Authorities**	"Area **C**" **(☒ –)**	"Area **D**" **(✓ +)**

i) Which of the following is TRUE with reference to Type I Error (False Positives) and Type II Error (False Negatives) based on the given matrix?

A. Type I Error = Area A and Type II Error = Area D

B. Type I Error = Area C and Type II Error = Area B

C. Type I Error = Area B and Type II Error = Area C

D. Type I Error = Area D and Type II Error = Area A

(2)

Following events have been defined in this context:

G : A person charged with the offence is actually guilty of the offence

I : A person charged with the offence is innocent

LG : A person charged with the offence is found to be guilty as per the lie-detector test

LI : A person charged with the offence is found innocent as per the lie-detector test

ii) Using the events as defined above, you are required to show the following:

a) Probability of committing Type I Error = P(I | LG);

b) Probability of committing Type II Error = P(G | LI). (2)

Based on the historical records for 1,000 cases as collected by the police authorities, following data has been obtained:

A = 356; B = 111; C = 105; D = 428

iii) Based on the above data, which of following is the probability that the Lie-Detector Test correctly identifies the perpetration / non-perpetration of the crime?

A. 0.784

B. 0.461

C. 0.533

D. 0.216 (1)

iv) Based on the historical data as given above, calculate the probabilities of Type I Error and Type II Error as shown in part (ii). (3)

[8]

Q. 3) **i)** Briefly explain the three components of a generalised linear model (GLM). (3)

An Actuary is using a GLM to model the remaining time until an actuarial aspirant qualifies as an actuary. The covariates he uses are:

- Age
- Passes: List of papers already passed.
  [This can be expressed as as a series of variables P_i, where i is a number between 1 and 13 (both inclusive), and each P_i is either 1 (the paper has been passed) or 0 (the paper has not been passed).]

- Experience
- Duration: Time elapsed since clearing ACET

The model he fits is given below:

Age + Passes + Experience + Duration + Experience . Duration

ii) The number of parameters for the main effects (i.e., not interactions) under the above GLM is –

A. 15

B. 16

C. 17

D. 18 (2)

iii) The number of parameters relating to interaction terms in the above GLM is –

A. 30

B. 20

C. 10

D. 1 (1)

The Actuary wants to simplify the model and check whether any of the covariates can be removed. He has calculated scaled deviance of the model as 15. However, he has been informed that the Akaike's Information Criterion (AIC) could be a better metric to decide on the model fit.

iv) Calculate the AIC for the model based on the scaled deviance given above and the number of parameters determined in parts (ii) and (iii). You are given that the log-likelihood of the saturated model is 16. (4)

[10]

Q. 4) Space scientists from Planet Actuaria have been making attempts to understand the extent of gravitational pull on various regions of a nearby Planet Numerica. For this purpose, an object weighing 10 pounds on Actuaria has been sent through space vehicles on various regions of Planet Numerica and the weight of this object at each of these regions is being measured.

Let W be the weight (in pounds) of the object on Planet Numerica.

Let X be the ratio of the weight of the object on Planet Numerica to the weight of the object on Planet Actuaria. In other words, X = W / 10.

X is assumed to follow a continuous uniform distribution within the interval $[0, \theta]$.

An attempt is being made to arrive at the estimate of θ (given $\theta > 0$) in order to assess the maximum gravitational pull on Planet Numerica.

A sample of 10 measurements for X has been obtained as given below:

0.70, 0.55, 0.31, 0.40, 0.35, 0.62, 0.34, 0.77, 0.45, 0.64

i) If $\hat{\theta}_{MOM}$ is the method of moments estimator of θ, obtain an estimate for $\hat{\theta}_{MOM}$ based on the given sample.

$\sum x = 5.13$

$\sum x^2 = 2.8761$ (3)

ii) Which of the following is a correct expression for the likelihood function of θ assuming a sample $\underline{x} = x_1, x_2, \ldots\ldots., x_{10}$ given $0 \leq x_1, x_2, \ldots\ldots., x_{10} \leq \theta$?

A. $L(\theta) = \frac{1}{\theta^{10}}$ for all x_i in $\underline{x}$

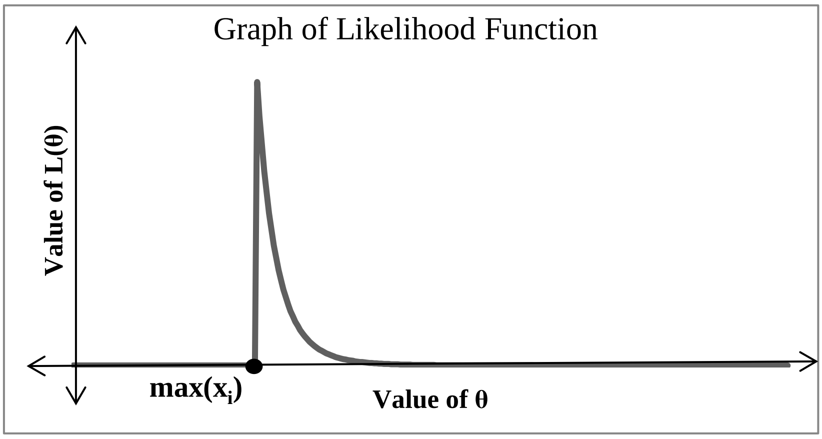
B. $L(\theta) = \frac{1}{\theta^{10}}$ if $\theta > \max(x_i)$;
= 0 otherwise

C. $L(\theta) = \frac{1}{\theta^{10}}$ if $\theta > \min(x_i)$;
= 0 otherwise

D. $L(\theta) = \frac{1}{\theta^{10}}$ if $\theta \leq \max(x_i)$;
= 0 otherwise (1)

iii) Obtain an equation to be solved for finding $\hat{\theta}_{MLE}$ i.e. the maximum likelihood estimator of θ using the likelihood function chosen in part (ii). Also comment on why method of differentiation doesn't work in this case. (3)

Assuming value of θ from 0.01 to 3.00, a graph was constructed plotting the value of θ on x axis and the value of L(θ) on y-axis:



iv) Based on the given graph, show that max (x_i) is $\hat{\theta}_{MLE}$ i.e. the maximum likelihood estimator of θ. (2)

Let us define a random variable Z = max (X_i) where X_is given i = 1,2, ….., 10 are independent and identically distributed continuous uniform variables over interval [0, θ].

The probability density function of Z is given by the following expression:

$$f_Z(Z) = \frac{10\,z^9}{\theta^{10}}. \quad \text{for } 0 \leq z \leq \theta$$

v) Using the probability density function of Z, show that the bias of the estimator $\hat{\theta}_{MLE}$ is $- 11^{-1} * \theta$.

Hint: *Bias* $(\hat{\theta}_{MLE}) = E(\hat{\theta}_{MLE}) - \theta = E(Z) - \theta$. (2)

vi) Which of the following is TRUE about the mean square error of the estimator $\hat{\theta}_{MLE}$?

A. MSE $(\hat{\theta}_{MLE}) = \text{Bias}^2(\hat{\theta}_{MLE})$

B. MSE $(\hat{\theta}_{MLE}) = \text{Variance}\ (\hat{\theta}_{MLE})$

C. MSE $(\hat{\theta}_{MLE}) = \text{Variance}\ (\hat{\theta}_{MLE}) + \text{Bias}^2(\hat{\theta}_{MLE})$

D. MSE $(\hat{\theta}_{MLE}) = \text{Variance}\ (\hat{\theta}_{MLE}) - \text{Bias}^2(\hat{\theta}_{MLE})$ (1)

[12]

Q. 5) Country Money-Land has two leading stock exchanges MSE and LSE. Stock market growth in the country is measured using two important indices:

1. MSE’s MIFTY which is an index comprising of 50 stocks from diverse sectors listed on MSE.
2. LSE’s LENSEX which is an index comprising of 30 large cap stocks listed on LSE.

There are no common stocks in MIFTY and LENSEX and hence annual returns from both these indices are assumed to be independent of each other.

Column A represents various available data sets relating to MSE and NSE and Column B represents the type of data.

Column “A” – Data Sets	Column “B” – Type of Data
1. LSE’s LENSEX returns over the last 10 years.	i. Truncated Data
2. Returns on stocks which were listed on MSE in the middle of the year.	ii. Longitudinal Data
3. Closing value of LSE’s LENSEX on 31st March, 2023.	iii. Cross-sectional Data
4. Data of stocks on MSE during a truncated week (i.e. a week where there are less than 5 working days in the week due to public holidays).	iv. Censored Data

i) Above pairs are incorrectly matched. Which one of the following options from those given below represents correctly matched pairs?

A. 1 – iii, 2 – iv, 3 – i, 4 – ii

B. 1 – ii, 2 – i, 3 – iv, 4 – iii

C. 1 – i, 2 – iv, 3 – iii, 4 – ii

D. 1 – ii, 2 – iv, 3 – iii, 4 – i (2)

A leading firm in the stock market which undertakes technical analysis has fitted a normal distribution on returns from both MSE’s MIFTY as well as LSE’s LENSEX.

Random variables X and Y represents the average annual returns on MSE’s MIFTY and LSE’s LENSEX respectively. $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\alpha, \beta^2)$.

Random samples (X_1, X_2, …….., X_{10}) and (Y_1, Y_2, …….., Y_{10}) comprising of average annual returns on MIFTY and LENSEX respectively for the past 10 years have been collected with sample means $\overline{X}$ and $\overline{Y}$ and sample variance S^2x and S^2y.

ii) Show that P ($S^2_x > \sigma^2$) is 0.41. Use χ^2 result for sample variance. (4)

iii) Is the sample mean $\overline{X}$ independent of sample variance S^2_x? (1)

iv) Determine P ($\overline{X} > \mu \mid S^2_x > \sigma^2$) using the result for sampling distribution of $\overline{X}$. (3)

MSE's MIFTY is considered to be a more reliable index for tracking stock market growth due to its wider and more diverse composition. LSE's LENSEX however, tends to show higher annual returns coupled with higher volatility.

You are given that: μ = 17%, σ^2 = 16%%, α = 19%, β^2 = 24%%.

v) Show that P ($S_x \leq S_y$) is greater than 10%. Use F-result for variance ratios. (4)

vi) Technical analysts in the firm have developed a 30 × 30 (scaled) variance / covariance matrix for average annual returns of the 30 stocks which form a part of LSE's LENSEX using principal component analysis (PCA). What is the maximum number of non-zero values that this matrix can contain?

A. 30

B. 900

C. 0

D. None of the above. (1)

[15]

Q. 6) Seismologists consider that the approximate time (in days) between the occurrence of two earthquakes in a particular seismic zone can be modelled as a random variable X with an exponential distribution, having the density function: $f(x) = \frac{1}{\mu} e^{-x/\mu}$.

It is proposed to use the following prior distribution for μ:

$$f(\mu) = \frac{\theta^{\alpha} e^{-\theta/\mu}}{\mu^{\alpha+1}\Gamma(\alpha)} \qquad \mu>0$$

The mean of this distribution is: θ / (α – 1).

i) Write down the likelihood function of μ, based on observations x_1, ……………., x_n from an exponential distribution. (2)

ii) Determine the posterior probability density function of μ and state its parameters. (3)

iii) Based on the posterior distribution derived in part (ii), show that an expression for the Bayesian estimate of μ under squared error loss is given by:

$$\hat{\mu} = \frac{\theta + \Sigma x}{n+\alpha-1}$$ (2)

iv) This is a case of conjugate priors. Which of the following is TRUE in respect of conjugate priors?

A. If the prior distribution leads to a posterior distribution is exactly identical to the distribution of the prior, only then this prior is called the conjugate prior for this likelihood.

B. Conjugate distributions often make Bayesian calculations simpler.

C. If the distribution of the random variable X is identical with the posterior distribution, then it is considered to be a case of conjugate priors.

D. If the prior distribution of a parameter is uniform, then the posterior distribution of the parameter will always be a uniform distribution. (1)

v) Show that the Bayesian estimate of μ can be written in the form of a credibility estimate giving formula for the credibility factor. (3)

You are given that the parameters of the prior distribution are θ = 40 and α = 1.5.

You are given the following summary statistics from the sample data relating to the past 100 earthquakes in the seismic zone:

$$n = 100, \sum x = 9{,}000, \sum x^2 = 12{,}00{,}000$$

vi) Calculate the prior mean, the sample mean, the Bayesian estimate of μ and the value of the credibility factor. (3)

vii) Based on the calculations in part (vi), what can you infer about the estimation exercise? Choose the right option from those given below:

A. Bayesian estimate of μ is closer to the mean of the sample data.

B. Bayesian estimate of μ is closer to the mean of the prior distribution.

C. Bayesian estimate of μ is an average of the sample mean and prior mean.

D. Bayesian estimate of μ is equal to the mean of the sample data. (1)

[15]

Q. 7) The insurance regulator has announced a move to a new solvency regime to align with international practice. Under the new regime, the capital requirement for each line of business will be determined based on expected losses arising from a line of business.

Let Y be a random variable representing the expected losses arising from a line of business. The capital requirement for that line of business is the value y such that P(Y≤y) = 0.995.

The Appointed Actuary of a general insurance company is reading some examples of capital requirement calculation. All examples are in lakhs of INR.

In the first example, the losses Y arising from a line of business are modelled as Y ~ Gamma(θ, 1/υ).

For a particular line of business, θ = 4 and υ = 25.

i) Express the mean and variance of Y in terms of θ and υ and calculate the mean and standard deviation of the losses from this business. (2)

ii) What is the capital requirement for this line of business?
(The table of percentages for the distribution is at the end of this question.) (1)

In the second example, frequency N and severity X are modelled separately. Here, N is a random variable representing the number of claims, and X is a random variable representing the cost of claim. Thus, Y is modelled as

$$Y = \sum_{i=1}^{N} X_i$$

The X_i's are assumed to be independent and identically distributed. N is assumed to be independent of the X_i's.

iii) In the context of the above example, which of the following is necessary and sufficient for random variables to be "independent and identically distributed"?

A. Both variables have identical distributions and are uncorrelated to each other.

B. Both variables belong to the same family of distributions and are not dependent on each other.

C. Both variables have identical distributions and are not dependent on each other.

D. Both A and C (1)

iv) Which type of random variables will be used to model N and X in the second example? Choose the right option from those given below:

A. Discrete random variables would be used to model both N and X.

B. X would be modelled as a discrete random variable whereas N would be modelled as a continuous random variable.

C. N would be modelled as a discrete random variable whereas X would be modelled as a continuous random variable.

D. Continuous random variables would be used to model both N and X. (1)

The example further states that N ~ Poisson (μ) and X ~ Gamma (α, 1/β), where μ = 10, α = 2/3 and β = 15.

v) Show that $E(Y) = \mu\alpha\beta$ and $var(Y) = \mu\alpha\beta^2 (1+\alpha)$. Using the values of μ, α and β given above, calculate the mean and variance of Y.

Hint: *Use the results E(Y) = E[E(Y|N)] and Var(Y) = E[Var(Y|N)] + var[E(Y|N)]* (7)

In the second example, capital requirement is calculated using Monte Carlo simulation. Simulation is to be done using the following three steps:

Step 1: First, we need to simulate a value for number of claims 'n' given by random variable N where N ~ Poisson (10).

Step 2: Using the simulated value 'n' generated above, 'n' values are simulated for claim amounts 'x' represented by random variable X where X ~ Gamma (2/3, 1/15)

Step 3: The sum of these n claim amounts is the simulated value of Y.

vi) Follow the given steps to simulate a value of Y. Use the random variates and the table of gamma probabilities as given below.

Random variates taken from a U(0,1) distribution: 0.19, 0.95, 0.70, 0.80, 0.20, 0.10, 0.20, 0.50, 0.60, 0.80
[The first random variate can be used to simulate n, and the remaining can be used to simulate x values.] (4)

Table of percentages of Gamma distributions

P(Y≤y)	Gamma (4, 1/25)	Gamma (2/3, 1/15)
0.1%	10.71	0
0.5%	16.80	0.01
1.0%	20.58	0.01
2.5%	27.25	0.05
5.0%	34.16	0.14
10.0%	43.62	0.41
20.0%	57.42	1.21
30.0%	69.09	2.32
40.0%	80.28	3.77
50.0%	91.80	5.65
60.0%	104.38	8.11
70.0%	119.06	11.48
80.0%	137.88	16.46
90.0%	167.02	25.39
95.0%	193.84	34.64
97.5%	219.18	44.10
99.0%	251.13	56.82
99.5%	274.44	66.56
99.9%	326.56	89.43

[16]

Q. 8) A real estate analyst has collected data of the average floor price index (per-square-foot apartment price) and the population density of various metropolitan areas in the country. The data is given below:

Population Density (X) (In unit of 1000 persons)	**Floor price index (Y) (In unit of 1000 INR)**
per sq. km	*per sq. foot*
17	8.50
12	6.90
25	18.00
10	4.00
5	0.50

She decides to fit a linear model to the data, where population density is the explanatory variable X and floor price index is the response variable Y.

i) Calculate $\bar{x}$ and $\bar{y}$. (2)

ii) Calculate S_{xx}, S_{yy} and S_{xy}.

You are given that:

$\sum x^2$ = 1183.00
$\sum y^2$ = 460.11
$\sum xy$ = 719.80 (3)

iii) Determine the fitted regression line: $\hat{y} = \alpha + \beta x$. (3)

iv) Calculate a 99% confidence interval for the slope parameter β. (4)

The residuals $y - \hat{y}$ from the linear regression are as follows:

x	Residual
5	0.42
10	-0.34
12	0.85
17	-1.81
25	0.87

v) Which of the following is TRUE about the distribution of residuals? Choose the right option from those given below:

A. The residual values don't seem to have any relationship with x and seem to be distributed approximately normally around the origin.

B. Residual values seem to be positively correlated with x and hence are not normally distributed.

C. Residual values seem to be negatively correlated with x and hence are not normally distributed.

D. Residual values are not independent of each other and hence are not normally distributed. (1)

The analyst obtains another data point to add to her data, so n = 6.

After adding this data point, S_{xx} = 250, S_{yy} = 190, S_{xy} = 200.

vi) What proportion of variance in the floor price index is now explained by the model? Choose the right option from those given below:

A. 84%

B. 72%

C. 42%

D. 93% (1)

The analyst adds a second explanatory variable to her regression and R^2 becomes 87%.

vii) Calculate the adjusted R^2 for both the one-variable model (considering 6 data points) and the two-variable model, and comment on which is the better model. (4)

[18]

****