INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

20th July 2022

Subject CS1A – Actuarial Statistics (Paper A)

Time allowed: 3 Hours 30 Minutes (9.30 - 13.00 Hours)

Total Marks: 100

Q. 1) **i)** List the key steps of a data analysis process. (4)

ii) Random variable X follows a distribution with an unknown parameter μ. A statement "The probability of μ being between 10 and 50 is 95%" is **valid** in… (pick the correct option below)

A. Classical statistics only
B. Bayesian statistics only
C. Both Classical & Bayesian statistics
D. Neither Classical nor Bayesian statistics (1)

iii) Match items A-D with the corresponding items I-IV below.

A. Type I error
B. Type II error
C. Specificity
D. Power

I. True positive
II. True negative
III. False positive
IV. False negative (2)

[7]

Q. 2) An Insurance company is analysing the claim data. Determine the probabilities of the following events.

i) The number of claims reported in a year by 100 policyholders is less than 6.

Assume claims reporting from each policyholder follows Poisson distribution with mean 0.03 per year independently of the other policyholder. (2)

ii) The number of claims examined up to and including the fourth claim that exceeds £50,000 is less than 7.

Assume the above follows negative binomial distribution with probability of a claim exceeding £50,000 as 0.4 independent of any other claim. (2)

iii) The number of deaths in the coming year amongst a group of 1000 policyholders is less than 10.

Assume each policyholder has a 0.015 probability of dying in the coming year independently of any other policyholder. (2)

[6]

Q. 3) A Student Actuary is analysing the time taken between two consecutive claims in a health insurance policy. It is believed that the time period (denoted by random variable X) between two consecutive claims in a health insurance policy follows an exponential distribution with mean μ.

i) Identify the correct expression for moment generating function of X.

A. $E(e^{tX}) = (\frac{1}{\mu} - t) \int_0^\infty e^{-z} \, . \, dz$

B. $E(e^{tX}) = \frac{1}{\mu} (\frac{1}{\mu} - t)^{-1} \int_0^\infty e^{-z} \, . \, dz$

C. $E(e^{tX}) = \frac{1}{\mu}\int_0^{\infty} e^{-z}$. dz

D. $E(e^{tX}) = \frac{1}{\mu}(\frac{1}{\mu} - t)\int_0^{\infty} e^{-z}$. dz

E. None of the above (2)

ii) If random variable Y denotes sum of time periods of two consecutive claims of N policies, determine the moment generating function of Y. (3)

iii) Identify the distribution of Y. (1)

[6]

Q. 4) Claim sizes on a fixed benefit health insurance policy are normally distributed about a mean of INR 900 and with a standard deviation of INR 100. Claim sizes on a indemnity based health insurance policy are normally distributed about a mean INR 1,400 and with a standard deviation of INR 300. All claim sizes are assumed to be independent and in units of INR 100.

To date, there have already been fixed benefit health insurance claims amounting to INR 900 and no indemnity based health insurance claims. Assuming that there will be further 4 fixed benefit health insurance claims and 3 indemnity based health insurance claims in next year, calculate the probability that the total claim amount under the indemnity based health insurance claims exceeds the total claim amount under fixed benefit health insurance claims.

[4]

Q. 5) **i)** Briefly explain what is meant by Independent and Identically Distributed (IID) Variables. (1)

A coin is tossed five times and the outcome is as follows: heads, heads, heads, heads, tails. Assume that the coin tosses are IID and the probability of each coin toss being either heads or tails is θ and (1-θ) respectively.

ii) Under the null hypothesis where θ = 0.5:

a) Calculate the probability of the outcome observed. (1)

b) Calculate the p-value of four or more heads and comment whether the null hypothesis can be rejected at significance level of 5%. (3)

iii) Considering θ as an unknown parameter:

a) Write down a formula for the likelihood of the outcome observed. (1)

b) Which of the following is the Maximum Likelihood Estimate (MLE) of θ?

A. 0.5
B. 0.4
C. 0.8
D. 0.2
E. None of the above (3)

The prior distribution of θ was as follows:

θ	0.1	0.3	0.5	0.7	0.9
f(θ)	0.2	0.2	0.2	0.2	0.2

c) Calculate the prior expected value of θ. (1)

d) Briefly explain what the prior distribution indicates about our knowledge of θ. (1)

e) Based on the outcome observed (4 heads, 1 tail), the posterior expected value of θ is… (pick the correct option).

A. Less than the prior expected value
B. Equal to the prior expected value
C. Greater than the prior expected value
D. None of the above (2)

The coin is tossed 3 more times and the outcome is tails, tails, tails.

f) The posterior distribution is recalculated incorporating the 3 additional coin tosses as well. The revised expected value of θ is also recalculated from this posterior distribution. This expected value is… (pick the correct option)

A. Exactly 0.2
B. Greater than 0.2 but less than 0.5
C. Exactly 0.5
D. Greater than 0.5 but less than 0.8
E. Exactly 0.8 (2)

g) The Bayesian estimator under quadratic loss is… (pick the correct option)

A. Mean of the prior distribution
B. Mean of the posterior distribution
C. Median of the prior distribution
D. Median of the posterior distribution
E. Mode of the posterior distribution (1)

h) The Bayesian estimator under absolute loss is… (pick the correct option)

A. Mean of the prior distribution
B. Mean of the posterior distribution
C. Median of the prior distribution
D. Median of the posterior distribution
E. Mode of the posterior distribution (1)

iv) So far, we have assumed that each coin toss is independent of the previous toss. It is now desired to test this assumption.

a) State, with reason, whether the chi-squared test can be used for this purpose, considering the sample of 8 coin tosses as above. (1)

The coin is tossed 13 more times (21 tosses in total). A contingency table is prepared as below. The columns shows the outcomes of the coin toss, and the rows shows the outcome of the immediately preceding coin toss. (The first toss is omitted.)

	Heads	Tails	Total
Heads	7	3	10
Tails	3	7	10
Total	10	10	20

b) From the above table, test whether each coin toss is independent of the immediately preceding toss and state your inference. (3)

[21]

Q. 6) The joint probability density function of random variables X and Y is:

$$f(x,y) = \begin{cases} 3e^{-(x+3y)} & , x > 0, y > 0 \\ 0 & otherwise \end{cases}$$

i) Determine $f_Y(y)$ the marginal density function of Y. (1)

ii) Determine the conditional density function f(y | Y>4). (2)

iii) Identify which one of the following expressions is equal to the conditional expectation E[Y | Y>4] :

A. $\int_0^\infty 3e^{-3t}dt + \int_0^\infty 12e^{-3t}dt$
B. $\int_0^\infty 3e^{-3t}dt + \int_0^\infty 12te^{-3t}dt$
C. $\int_0^\infty 3te^{-3t}dt + \int_0^\infty 12e^{-3t}dt$
D. $\int_0^\infty 3te^{-3t}dt + \int_0^\infty 12te^{-3t}dt$
E. None of the above (2)

[5]

Q. 7) You are an actuarial analyst working at a life insurance company in India. The company is analysing the force of mortality, μ_X, of a particular group of policyholders. The company believes that μ_X is related to age, X, by the formulae:

$$\mu_X = BC^X$$

You are provided herewith the following summary results for 10 ages

Age,X	30	32	34	36	38	40	42	44	46	48
Force of mortality, $\mu_X(\times 10^{-4})$	5.22	5.64	6.25	6.82	7.46	8.73	10.63	12.23	14.45	16.28
$\ln \mu_X$	-7.56	-7.48	-7.38	-7.29	-7.20	-7.04	-6.85	-6.71	-6.54	-6.42

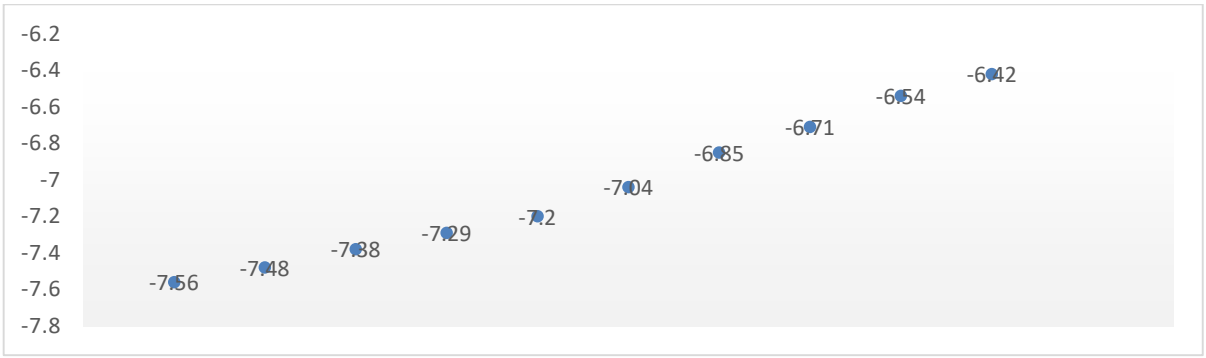
$\sum X_i = 390$, $\sum X_i^2 = 15{,}540$, $\sum \ln \mu_{X_i} = -70.47$, $\sum (\ln \mu_{X_i})^2 = 498.05$, $\sum X_i \ln \mu_{X_i} = -2{,}726.66$

Your reporting manager has asked you to perform a linear regression analysis on this data to identify the relationship between age and force of mortality.

It is decided to analyse the assumptions by using the linear regression model:

$Y_i = \alpha + \beta X_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ where $Y_i = \ln \mu_{X_i}$, $\alpha = \ln B$, $\beta = \ln C$

i) The following is a plot of a graph of ln μ_X against the age of the policyholder, X.



Comment on the suitability of the regression model. (1)

ii) Use the data to calculate least squares estimates of B and C in the original formula. (3)

iii) Write down the value of Pearson's correlation coefficient between the variables ln μ_{X_i} and X_i hence comment on the relationship between the variables after taking into consideration the value of the slope parameter β. (2)

iv) Calculate the coefficient of determination between ln μ_X and X. Hence comment on the fit of the model to the data. (2)

v) Complete the table of residuals and use it to comment on the fit.

Age,X	30	32	34	36	38	40	42	44	46	48
Residual, $\hat{e}_i$	0.079	0.028		-0.045	-0.087	-0.058		0.009	0.048	

(3)

vi) Calculate a 95% confidence interval for the mean predicted response [ln μ_{45}] and hence obtain a 95% confidence interval for the mean predicted value of μ_{45} . (4)

vii) Comment on the width of a 95% confidence interval for the predicted mean response if X = 41, as compared to the width of the interval in part (vi), without calculating the new interval. (2)

[17]

Q. 8) You have been investing in shares of unrelated industries X and Y for diversification. Your friend opines that this does not achieve diversification, because share indices of the two industries are strongly positively correlated with each other. To support this line of argument, he gives you the following index values for two different dates over two different periods in time:

Period 1		Period 2	
X	Y	X	Y
10	10	10	10
30	20	30	15

i) Calculate Pearson's correlation coefficient between the two indices X and Y for each of the two periods. (5)

On digging deeper, it turns out that the bases of the indices differ between periods 1 and 2. To express the period 2 figures in the period 1 base, the X figures need to be multiplied by 2 and the Y figures by 0.4.

ii) Expressing all figures in the period 1 base, calculate the combined correlation of X and Y across both the periods. (4)

iii) Comment on whether your portfolio is diversified in view of your friend's opinion based on the results of (i) and (ii). (2)

[11]

Q. 9) An Indian life insurer has written a large book of policies which provides the Sum Assured on diagnosis of major stage cancer.

The claim frequency per mille (number of claims per 1,000 policies) arising on this book over the past 3 years is as below:

Year	Claim frequency (per mille)
2019-20	16.4
2020-21	17.3
2021-22	16.7

The age composition and other aspects of the book have been relatively unchanged over this period.

The claim frequency per mille is modelled as following a Poisson distribution with an unknown parameter λ.

The Pricing Actuary models λ as following a Gamma (A, B) distribution, with A = 15 and B = 1.

The Bayesian credibility factor is given by $n/(n + B)$, where n is the number of years for which data is available.

i) Calculate the Bayesian credibility estimate for the number of claims per 1,000. (3)

The Pricing Actuary had chosen the prior distribution of Gamma (15, 1) based on inputs from a global reinsurer, who expected the claim frequency to be 15 per mille.

The Appointed Actuary disagrees with the Pricing Actuary's choice of Gamma parameters, as he believes the reinsurer's experience may not be very relevant to the Indian market. He therefore suggests using Gamma (3, 0.2) as the prior distribution.

ii) Briefly explain, by general reasoning, how the suggested parameters reflect the greater uncertainty regarding the relevance of the reinsurer's data. (1)

iii) Calculate the revised Bayesian credibility estimate. (1)

[5]

Q. 10) An insurer has written n personal accident policies, which pay the Sum Assured in case of death of insured due to accident.

The probability of a claim payout is assumed to be q. Each claim is IID.

The actual number of claims paid is x. Policy-wise data is available, showing exactly which policies have turned into claims.

i) Which of the following is the likelihood function for the situation described?

A. ${}^{n}C_{x}q^{x}(1-q)^{n-x}$

B. $q^{x}(1-q)^{n-x}$

C. ${}^{n}C_{x}q^{1-x}(1-q)^{n-x}$

D. $q^{n-x}(1-q)^{x}$

E. ${}^{n}C_{x}q^{n-x}(1-q)^{x}$

(1)

Given that n = 10,000 and x = 3.

ii) Using the normal approximation, calculate the 95% confidence interval for $\hat{q}$ (the sample estimator of q). (4)

iii) Consider q = 0.2 per mille as the null hypothesis.
Based on the above, comment on the validity of the null hypothesis. (1)

[6]

Q. 11) A random variable z has a binomial distribution with parameters n and μ and has the following density function:

f(z) = $\binom{n}{z}\mu^{z}(1-\mu)^{(n-z)}$, where 0 < μ < 1

i) Show that the distribution function of Y = $\frac{z}{n}$ can be written in the standard form of the exponential family of distributions, stating the natural and scale parameters, θ and φ, and the associated functions of these parameters. (4)

ii) Verify the mean and variance of the Binomial Distribution, using the expressions from part (i) together with the properties of the exponential family of distributions. (3)

A researcher is investigating the number of students who pass in a particular examination. The researcher believes that the number of students who pass follows binomial distribution.

He also believes that probability of passing, μ, depends on the followings

- The number of assignment, N, submitted by the student
- The student's mark in the mock exam S
- Whether student attended tutorials or not (Yes/No)

The researcher specifies the following linear predictor, where α_i, β_1 and β_2 are parameters to be estimated

$$\eta(\mu) = \alpha_i + \beta_1 N + \beta_2 \mathrm{S}$$

Where α_i takes one value for those attending tutorials (α_Y) and a different value for those who do not (α_N).

The researcher then runs computer model that fits generalized linear model (using binomial canonical link function) basis of data collected from 30 observation points.

Parameters:	Estimate	Standard Error
Intercept, α_Y	-1.501	0.29190
Intercept, α_N	-3.196	0.13401
β_1, no. of assignment	0.5459	0.08352
β_2, mark in mock exam	0.0251	0.00156

iii) Explain, using the model output shown above, whether the variable "no. of assignment" is significant or not. (2)

iv) Estimate using the fitted model, the probability of passing for a student who attends tutorials, submitted 4 assignments and scored 65 marks in the mock exam. (3)

[12]

****