INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

November 2026

CS1 – Actuarial Statistics (Paper B)

Time allowed: 1 Hour 45 Minutes

Total Marks: 100

Q. 1) An insurance company records the total value of claims, in '000s, for three regions of drone insurance policies over five years:

Region	Year 1	Year 2	Year 3	Year 4	Year 5
A	120	150	130	170	140
B	90	110	95	105	100
C	180	200	190	210	195

The data can be entered into R as:

```
claims <- matrix(c(120,150,130,170,140, 90,110,95,105,100,
180,200,190,210,195), nrow = 3, byrow = TRUE )

rownames(claims) <- c("A", "B", "C")

colnames(claims) <- paste0("Year", 1:5)
```

Note: *You can directly copy these code lines from Codes.docx file in your R Script.*

An actuary wishes to estimate the expected claim amount for Region A in the next year.

Assume that, conditional on the risk parameter θ, X|θ~N (θ,σ_1^2) where σ_1^2=200

Before observing the experience of Risk A, the actuary assumes that θ~ N (μ, σ_2^2) with μ = 130 and σ_2^2= 250.

The posterior distribution of θ given X is:

$$\theta \mid \mathbf{x} \sim N\left(\frac{\frac{n\bar{x}}{\sigma_1^2}+\frac{\mu}{\sigma_2^2}}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}, \quad \frac{1}{\frac{n}{\sigma_1^2}+\frac{1}{\sigma_2^2}}\right)$$

i) Calculate the parameters of θ for Region A using the posterior distribution of θ. **[4]**

ii) Calculate the Bayesian estimate of the expected claim amount for Region A in the next year under quadratic loss. **[1]**

iii) Calculate the Bayesian credibility factor and express the Bayesian estimate in credibility form. **[3]**

iv) Using the data for all three regions, estimate the values of the following quantities under the assumptions of Empirical Bayes Credibility Theory Model 1 (EBCT Model 1):

a) E[m(θ)]
b) E[s^2(θ)]
c) Var[m(θ)]. **[6]**

v) Calculate the credibility factor Z and the EBCT credibility premium for Region A. **[3]**

vi) Compare the Bayesian estimate obtained in part (ii) with the EBCT credibility premium obtained in part (v) for Region A. **[3]**

[20 Marks]

Q. 2) Municipal corporation of an ancient religious city has come up with a new policy in order to provide financial support to victims of crowd mismanagement during the annual *Kanwar Yatra* festival celebrations. The corporation has made arrangements to provide compensation for people who suffer from accidents during the *Kanwar Yatra* celebrations.

For a particular day during the festival, let the number of people who suffer from accidents due to crowd mismanagement be modelled as X ~ Poisson (3)

The compensation payable to each accident victim measured in INR '000 is independently distributed as Y ~ Exponential such that the mean compensation payable per victim is INR 50,000.

The corporation wants to budget for the total daily outgoes from this new policy and has hired you as a financial consultant.

You have chosen to model the total outgoes as a random variable S where:

$$S = Y_1 + Y_2 + \ldots\ldots\ldots + Y_X$$

On a day on which no accidents are expected to occur, the corporation wants you to still reserve an amount equal to INR 50,000 to ensure additional prudence.

The corporation wants to use simulation to investigate the relationship between the number of accident victims and the total compensation payable.

An initial attempt was made to simulate data for 10,000 festival days. R code used for the same is reproduced below. However, it has **four** errors.

```
set.seed(9876)

n <- 10000
X <- rpois(n, 3)
S <- numeric(n)

for(i in 1:n){

   Y <- rexp(3, rate = 50)

   S[i] <- mean(Y)
}
```

Note: *You can directly copy these code lines from Codes.docx file in your R Script.*

i) Identify and rectify the errors in the above R code so that it correctly simulates the total compensation (S) payable on each festival day.

Your corrected code should –

- Simulate appropriate number of accident-related compensation amounts per day
- Use the correct parameterisation of the exponential distribution
- Calculate the total compensation payable on each day
- Correctly handle a day on which no accidents occur **[7]**

ii) Use the simulated data to estimate E(S). **[2]**

iii) Using the simulated data, plot –

- a histogram of X
- a histogram of S

[4]

Use `breaks = 50` when plotting histogram of S.

iv) State with reasons an approximate distribution that would be appropriate for total compensation (S). **[2]**

v) Derive a theoretical value for E(S).

Use the formula for unconditional expectation E(S) = E(X) * E(Y).

The theoretical value should take care of the reserve of INR 50,000 where the number of claims X = 0. **[3]**

vi) Compare the theoretical value of E(S) from part (v) with the Monte Carlo estimate of E(S) from part (ii). **[2]**

vii) Using the simulated data, estimate –
- P(S>200,000|X≥3) and
- P(S>200,000)

Comment on whether the number of accident victims (X) and the total compensation payable (S) appear to be independent. **[5]**

[25 Marks]

Q. 3) A financial institution is evaluating whether a new derivative-based hedging strategy reduces the average annual hedging cost associated with its portfolio exposures.

The average annual hedging cost (in INR '000) for two groups of 10 portfolio exposures is given below.

- S = Hedging cost using the standard derivative instrument
- N = Hedging cost using the new derivative instrument

The observed hedging costs are:

S: 52.4, 48.7, 61.3, 55.8, 49.6, 64.2, 57.5, 51.9, 59.1, 53.6

N: 49.8, 46.9, 58.7, 53.1, 48.2, 60.5, 54.8, 50.7, 56.9, 51.4

The following R code is used to enter the data:

```
S <- c(52.4, 48.7, 61.3, 55.8, 49.6, 64.2, 57.5, 51.9, 59.1, 53.6)
N <- c(49.8, 46.9, 58.7, 53.1, 48.2, 60.5, 54.8, 50.7, 56.9, 51.4)
```

Note: *You can directly copy these code lines from Codes.docx file in your R Script.*

Initially, the institution treats the two groups as independent.

i) Use the two vectors and `t.test` function in R to calculate a 90% two-sided confidence interval for the difference in average annual hedging costs between the two groups. Comment on the result. **[4]**

It is subsequently discovered that the two groups actually consisted of the same 10 portfolio exposures, with the hedging costs measured under the standard and new derivative instruments for each exposure.

ii) Recalibrate the 90% two-sided confidence interval determined in part (i) based on this new information and comment on the result. **[3]**

The financial institution wants to determine whether the new derivative instrument significantly reduces the average hedging cost.

iii) Perform an appropriate hypothesis test in R to determine whether the new derivative instrument significantly reduces the average hedging cost at the 5% significance level.

Your answer should clearly indicate the null and alternative hypotheses, the p-value and your conclusion. **[6]**

iv) Comment on the difference between the confidence interval obtained in parts (ii) and (iii). **[2]**

[15 Marks]

Q. 4) A global financial services firm wants to predict the first-year performance of analysts recruited through its graduate hiring program.

The recruitment team has access to following candidate attributes - the candidate's university GPA (Grade Point Average), score on a standardized analytical-aptitude assessment and the quality of recommendations /references.

The firm wants to determine which combination of factors best predicts an analyst's subsequent performance.

The data for 100 newly recruited analysts and four variables are provided where:

Variable	Description
`perf`	First-year analyst performance score, measured on a 0–5 scale
`gpa`	Candidate's university GPA, measured on a 0–5 scale
`aptitude`	Standardized analytical-aptitude assessment score
`recommend`	Quality of professional/academic recommendations, rated from 1–6

The data is available in the file `GLM.csv`.

The recruiter believes that GPA is the most suitable factor in determining the analyst's performance.

i) Load the data into R and draw a scatterplot for the above data with `perf` being the dependant variable and `gpa` being the independent variable. **[3]**

ii) Fit a linear regression model `model_1` with gpa as the explanatory variable. Print the summary of the model. Also state the linear regression equation. Add the regression line onto the scatterplot obtained in part (i). **[5]**

iii) Predict the performance of a student who has scored 3.5 in university GPA. **[2]**

A senior recruiter advises to use all the other factors to predict the analyst's performance.

iv) Produce the correlation matrix and comment on the relationship of the variables. **[2]**

v) Fit a multi-variate linear regression model `model_2` with `perf` as the response variable and rest of the variables as explanatory variables. Print the summary of the model and state the coefficients of the model. **[4]**

vi) Fit a generalized linear model `model_3` to the above data which is equivalent to the multi-variate linear regression model `model_2` determined in part (v). State explicitly the appropriate family and the link function in the arguments. Print the summary of this model. **[4]**

vii) The senior recruiter feels that there might be strong relation amongst factors and hence wants to allow them in the model. Fit a generalized linear model (including all possible interaction terms) `model_4` to the data. Print the summary of this model. Also, comment on the significance of the parameters in `model_4`. **[5]**

viii) Using Akaike's Information Criterion (AIC), compare `model_3` and `model_4`. **[2]**

ix) Use `anova` function to carry out a formal chi-square test to compare `model_3` and `model_4`.

Your answer should clearly indicate the null and alternative hypothesis, the p-value and conclusion at 5% level of significance. **[5]**

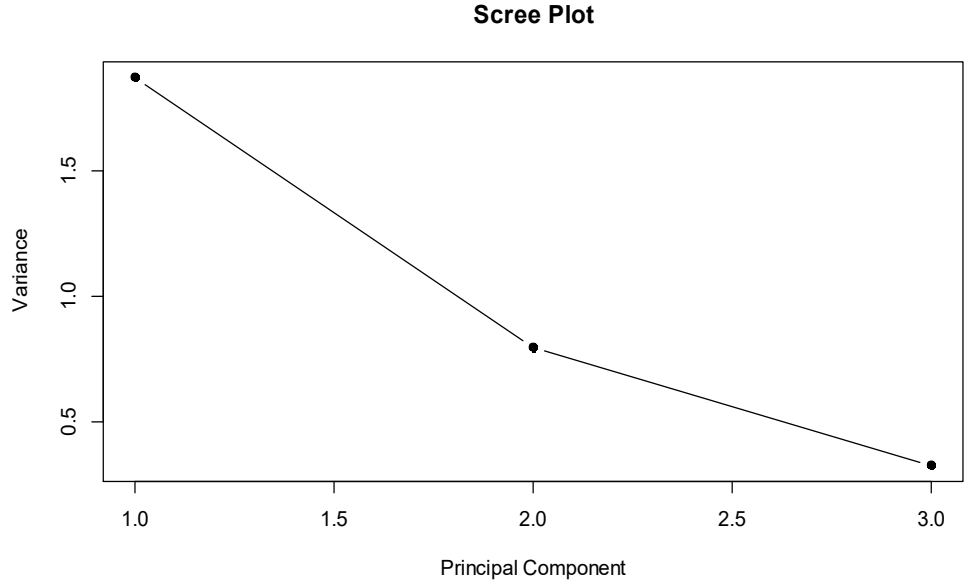
It is decided to perform principal components analysis to reduce the dimensionality of the dataset and following output has been obtained using the `prcomp` function in R:

```
Importance of components:

                          PC1    PC2    PC3
Standard deviation     1.3686 0.8938 0.5727
Proportion of Variance 0.6243 0.2663 0.1093
Cumulative Proportion  0.6243 0.8907 1.0000
```

x) Determine which principal components are to be retained using the Kaiser criterion. **[3]**

Following scree plot of principal components has been constructed using R:



xi) Which principal components out of PC1, PC2 and PC3 will you choose to retain considering the above scree plot? Justify your answer. **[2]**

Based on the principal components retained in part (xi), a new generalised linear model `model_5` is formulated by considering the retained principal components as explanatory variables and `perf` as the response variable. Following is the R summary of this model:

```
Call:
glm(formula = perf ~ PC1 + PC2, data = pc_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.14939  -0.45869  -0.02325   0.40715   1.65747

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.98050    0.05896  33.588  < 2e-16 ***
PC1          0.32424    0.04330   7.488 3.23e-11 ***
PC2          0.17246    0.06630   2.601   0.0107 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.3476807)

    Null deviance: 55.573  on 99  degrees of freedom
Residual deviance: 33.725  on 97  degrees of freedom
AIC: 183.09

Number of Fisher Scoring iterations: 2
```

xii) Based on the above results, comment whether you can conclude that principal components analysis has proved to be useful in improving the fit of the model. **[3]**

[40 Marks]

****