

INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

May 2025

CS1B - Actuarial Statistics (Paper B)

Time allowed: 1 Hours 45 Minutes

Total Marks: 100

- Q. 1)** A sports analyst aims to examine how varying levels of humidity across different stadiums influence a cricket team's run-scoring patterns. By analysing venue-specific data, the analyst seeks to understand whether higher humidity correlates with lower scores, possibly due to increased swing or changes in pitch behaviour.

The data shows the below:

- Humidity: An average relative humidity level (in %) at the location of the match
- Runs: The number of runs scored by the team in that match.

- i) Load the data `runs.csv` in R which contains data related to 30 matches. (1)
- ii) Plot the scatter plot for the data. (2)
- iii) Calculate the Pearson correlation coefficient r and state your conclusion. (2)
- iv) Fit a linear regression model called '`model_runs`' to the data with runs as the dependent variable and humidity as the independent variable, and state the regression equation. (3)
- v) Superimpose the scatter plot with the linear regression line for `model_runs`. (2)
- vi) What is the expected runs scored in the next match where the relative humidity level is predicted to be 48% ? (2)
- vii) Display a residual plot with fitted values on the horizontal axis and comment on it. (3)
- viii) State the value of the coefficient of determination, R^2 for `model_runs` and comment on it. (2)
- ix) Calculate the confidence intervals for the model parameters. (2)
- x) Create an analysis of variance (ANOVA) table and test if there is a linear relationship between humidity and runs. (2)

You should:

- a. State the null and alternative hypothesis
- b. Print the ANOVA table
- c. State the p-value
- d. Reject or accept the test
- e. Make a conclusion

(6)

[25 Marks]

- Q. 2)** A health club conducted a survey of 1000 people in a district to understand the relationship between their exercise habits and the intensity of their workouts. The data representing the number of people in each category is categorized as follows:

	Light	Moderate	High
Sedentary	38	120	42
Moderately Active	232	357	73
Very Active	48	29	4
Extremely Active	15	34	8

- i) Enter and print the data in the form of a 4x3 matrix assigning appropriate row and column names. (3)
- ii) Perform chi-square test to test if there is any dependence between the exercise habits and the intensity of workouts.

You should clearly:

- State the null and alternative hypothesis
- State the p-value
- Reject or accept the test
- Make a conclusion

(5)

The health club rolled out a workout routine to reduce the cholesterol level among individuals. The cholesterol levels (in mmol/L) was measured for 10 members before and after maintaining a consistent workout routine for a month.

Member	1	2	3	4	5	6	7	8	9	10
Before (mmol/L)	9.1	8	7.7	10	9.6	7.9	9	7.1	8.3	9.6
After (mmol/L)	8.2	6.4	6.6	8.5	8	5.8	7.8	7.2	6.7	9.8

You can enter the data in R using:

```
before <- c(9.1, 8.0, 7.7, 10.0, 9.6, 7.9, 9.0, 7.1, 8.3, 9.6)
after <- c(8.2, 6.4, 6.6, 8.5, 8.0, 5.8, 7.8, 7.2, 6.7, 9.8)
```

iii) Calculate the mean cholesterol levels before and after enrolling into the workout routine. (2)

iv) Calculate Pearson's correlation coefficient, Spearman's rank correlation and Kendall's Tau coefficient between **before** and **after** cholesterol levels. (3)

v) Comment on the suitability of the above correlation coefficients. (3)

vi) Comment on your results in part (iii) above. (1)

vii) Perform a suitable test at 5% level of significance to examine if working out regularly changed the mean cholesterol level or not (Assume equal and unknown variance).

You should clearly:

- State the null and alternative hypothesis
- State the p-value
- Reject or accept the test
- Make a conclusion

(5)

The health club claims that their workout routine has helped in reducing the cholesterol level of the members who enrolled for the program.

viii) Carry out a suitable test at 1% level of significance to test if the claim is true or not.

You should clearly:

- State the null and alternative hypothesis
- State the p-value
- Reject or accept the test
- Make a conclusion

(5)

[27 Marks]

Q. 3) An institute has launched a scholarship program and 100 students have applied for financial assistance. Among these 100 applicants, 40% come from low-income families.

Use the command `set.seed(2025)` to initialise the random number generator. When you execute any R code in this question, make sure you run the entire R script including the line `set.seed(2025)` at the start.

- i) Generate a random sample X of size 100 where $X_i = 1$ if student is from low-income family and $X_i = 0$ otherwise. (2)
- ii) Compute the sum of the elements of X to find out how many students are from low-income families in your simulated applications. (1)
- iii) Modify your code to simulate 5000 values of X and save it in vector `applications`. (2)
- iv) Compute the mean and variance of `applications`. (2)
- v) Compute the mean of the Binomial distribution with parameter $n = 100, p = 0.4$. (1)
- vi) Compare your result with mean computed in part (iv). (1)
- vii) Plot a labelled histogram of `applications`. (2)
- viii) State the distribution along with the parameters that can be a good approximation to the distribution of X . (2)
- ix) Use `lines`, `density` and `curve` to superimpose the histogram in (vii) with the
 - a) plot of the density estimate from the simulated data and
 - b) plot the pdf of the distribution stated in part (viii) above (3)
- x) Comment on the plot in part (ix) comparing the simulated values of X with the PDF. (2)

[18 Marks]

Q. 4) You are analysing whether to purchase an electric vehicle (EV) or not. Your sales agent provided you with a data containing the purchase decision (PurchaseEV: 1 = Purchased an EV, 0 = Did not purchase) based on the following variables :

- AvgDailyCommute: 1 = >50km, 2 = 25-50kms, 3 = <25kms
- ChargingAccess: Number of charging stations accessible near home or work
- HouseholdAdults: Number of adults in the household (e.g., spouse, siblings)

- i) Load the data `ev_purchase.csv` in R containing decisions made by 50 customers. (1)
- ii) Fit a Generalised Linear Model (GLM) to the data assuming a binomial distribution using PurchaseEV as the response variable and all others as explanatory variable without an interaction term, calling this `model_a`. Your answer should include the estimated coefficients, p-value and the Akaike's Information Criterion (AIC) of the fitted model. (6)
- iii) Predict if you would buy an EV given the average daily commute is 20kms, with 4 charging stations accessible near home or work and have 2 adults in your household. (3)

You are thinking to refine the model to include the interaction between the variables.

- iv) Fit an appropriate model, `model_b`, to include only **two-way** interactions between the explanatory variables.

Your answer should include the estimated coefficients, p-value and the Akaike's Information Criterion (AIC) of the fitted model. (4)

- v) Comment on whether any of the interactions are significantly associated with the car purchase decision. (2)

- vi)** Justify whether `model_b` is a better fit than `model_a`. (2)

You are concerned about the maintenance costs of the EV. Your sales agent provided you data containing the total maintenance costs (in thousands) per year for five brands of electric vehicles (EVs) over a six-year period ($n=6$).

Brand	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
A	9.8	10.5	11	12.3	15.2	18.5
B	21.5	22	23.3	25	22.5	20.1
C	6	6.5	7.2	8.7	9.8	10.5
D	12.2	13	14.3	16.5	19	22
E	7.9	8.5	9.5	11	12.8	14.5

- vii)** Load and print the data `ev_cost.csv` into R as matrix `cost` assigning proper row names. (2)
- viii)** Estimate the values $m = E[m(\theta)]$, $s = E[s^2(\theta)]$, $v = var[m(\theta)]$ using EBCT Model 1, and hence estimate the credibility factor, Z . (8)
- ix)** Calculate the credibility weighted maintenance cost for each brand. (2)

[30 Marks]
