

INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

May 2025

CS2 - Risk Modelling and Survival Analysis (Paper B)

Time allowed: 1 Hours 45 Minutes

Total Marks: 100

- Q. 1)** A Cricket League of India (CLI) attracts huge crowds and the league consists of 4 teams, namely,
- Team Delhi (D)
 - Team Mumbai (M)
 - Team Kolkata (K)
 - Team Bengaluru (B)

A broadcasting company Cricket Media of India (CMI) of the CLI analyses the fan-following for each team and based on historical data, has made the following observations:

- D has been highly instrumental in scouting young talented players and providing opportunities and is gaining fans at the annual rates of 0.012 from M, i.e., for every 1000 fans of Mumbai, 12 fans would shift to support Delhi every year. Similarly, it is gaining fans at the annual rates of 0.032 and 0.026 from K and B respectively.
- M has built a strong team and has been performing consistently. Hence, it is gaining fans at the annual rates of 0.02, 0.029 and 0.014 from D, K and B respectively.
- K is the defending champions and also has been a home to star players. K is gaining fans at the annual rates of 0.022, 0.002 and 0.025 from D, M and B respectively.
- B also houses many strong players and has many records to their name. B is gaining fans at the annual rates of 0.009, 0.041 and 0.036 from D, M and K respectively.

You are an Analyst at the CMI and are asked to perform the following tasks:

- i.** Write R code for the generator matrix A and show the output. (3)
- ii.** Derive the matrices of transition probabilities ('P12') over a month and over a day ('P365'). State any suitable approximation used if any. (4)
- iii.** Calculate the matrices of transition probabilities – 'P18y1' and 'P18y2' over 18 years using 'P12' and 'P365', respectively. Compare the results and comment. (5)
- iv.** Calculate the following using 'P12':
 - a. Probability that a Delhi fan would turn into Bangalore fan in 10 years
 - b. Probability that a Mumbai fan would turn into Kolkata fan in 15 years
 - c. Probability that a Bangalore fan will be Bangalore fan in 4.5 years
 - d. Probability that a Kolkata fan will remain Kolkata fan for 4.5 years
 - e. If each team starts with same number of fans, which team would be most popular after 10 years? (10)
- v.** Assume 1000 fans for each of the teams at time 0, using 'P12', simulate the movement of fans between the teams for a period of 30 years and store in a dataframe (named 'Sim1'). (3)

A Senior Analyst mentioned that there is an error in one of the annual rates and has provided you the following fan movement matrix capturing average monthly movement.

	D	M	K	B
D	99575	167	183	75
M	100	99542	17	342
K	181	240	98282	297
B	217	117	208	99458

- vi.** Perform the following to figure out error(s):
Using the above information, simulate the fan movement for 30 years and store it in a dataframe (named 'Sim2'). Add a column – 'Sim' (indicating 'Sim1' as "red" and 'Sim2'

as “green”) to both data frames, and print the first six rows of the combined data frames. For each of the teams, plot a graph comparing the number of fans each team has over the 30 years’ period in both simulations and comment which of the rates could be the wrong one. (6)

- vii. Approximate the generator matrix B from transition probability matrix obtained in (vi) and comment. (2)

[33 Marks]

Q. 2) You are an Actuary at a life insurance company. You are analysing one of the products of your company, which has regular premiums payable annually. You requested the data containing the following columns:

1. **Policy Number:** An identifier variable
2. **Gender:** indicating whether the covered life is “M” or “F”
3. **Date of birth:** birth date of the policyholder
4. **Policy start date:** The date on which coverage starts
5. **Date of death:** death date if the claim exists, else NA
6. **Policy end date:** The date on which coverage ends (due to lapse, surrender or any other reason)

The policy information is available in the file “*Policy_data.csv*”.

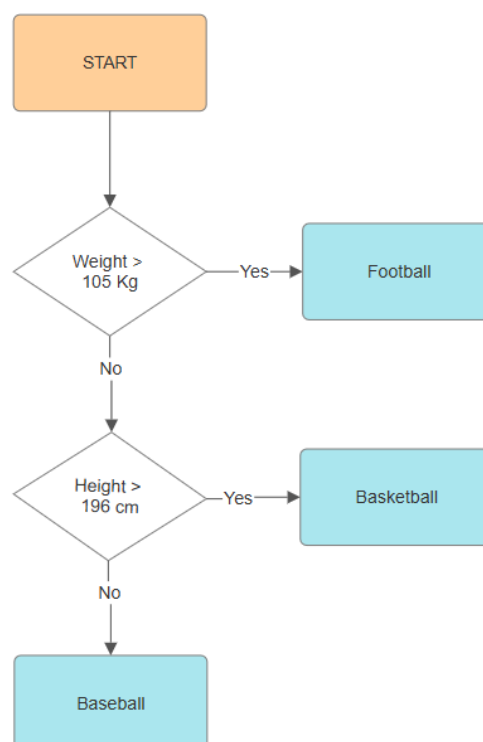
- i. Load the data into R, convert the date columns into date format and print the first 6 rows. (2)
- ii. Carry out suitable checks for this dataset. (3)
- iii. It is observed that few of the policies start date is incorrectly filled with their renewal dates, make the necessary correction. (2)
- iv. Given you would like to analyse the mortality during the period between 01/01/2008 and 31/12/2023, define the exposure start date, exposure end date, age at exposure start, age at exposure end and age at death for each policy. Print the first 6 rows.
(Consider Age last birthday and that each year has 365.25 days) (6)
- v. Construct the mortality table between ages 0 to 70 years inclusive using the above data, stating any assumptions made. Print the summary of the table. (7)
- vi. Create a life table for ages 20 to 50 (inclusive) from the above mortality table which contains only 3 columns:
 - a. Age
 - b. Gender
 - c. Mortality_rate
 Print the first 10 rows. (2)
- vii. Plot a scatter plot to visualise mortality rates for both genders. Add a smoothened trend line for each gender and comment. (3)
- viii. You have decided to fit these crude rates by using Gompertz’ Law.

- a. Divide life table into two datasets based on gender
 - b. Estimate the parameters of Gompertz' law for the datasets in (b) above.
 - c. Calculate graduated rates in a new column (named '**Graduated_rate**', round it to 10 decimals)
 - d. Combine both datasets into single data frame. (Print 6 rows of the final table) (6)
- ix. Plot a scatter plot to visualise mortality rates for both genders. Add the graduated rates as a trend line for each gender and comment. (3)
- x. Perform Signs test to check for overall bias and comment. (3)

[37 Marks]

Q. 3) You are given the following information:

- a. The height of Basketball players usually ranges between 190 cm to 204 cm.
 - b. The weight of Basketball players usually ranges between 90 Kg and 105 Kg.
 - c. The height of Baseball players usually ranges between 178 cm to 196 cm.
 - d. The weight of Baseball players usually ranges between 84 Kg and 98 Kg.
 - e. The height of Football players usually ranges between 174 cm to 191 cm.
 - f. The weight of Football players usually ranges between 98 Kg and 124 Kg.
- i. Based on the above data, create a dataset containing 300 records (100 for each sport) with the following columns: 1. Height, 2. Weight and 3. Sport. Use a seed value of 100 and assume Discrete Uniform distribution of athletes in the above mentioned ranges. Print first five records for each sport. (Hint: Use Sample function for random number generation). (5)
- ii. Comment on how height and weight can be used to predict the sport of an player. (3)



- iii. You have decided to use the above decision tree to predict the sport of a player using the height and weight of the player:

- a. Write a function to predict the sport of the player given height and weight following the decision tree above.
- b. Test your function to predict the sport of a player whose height is 190 cm and weight is 100 Kg.
- c. Apply this function to the dataset and store the prediction in a column named 'Decision_tree'.
- d. To assess the model, calculate and print the confusion matrix.
- e. Calculate Gini index for each of the sports.
- f. Comment on what Gini index suggests.

(9)

iv. You have now decided to use the naïve Bayes approach instead:

- a. Define the likelihood of obtaining a value k from a Uniform discrete distribution ranging from a to b . The expression should be valid for all integer values of k .
- b. Write a function for each sport, that calculates the likelihood of a player playing that sport given the player's height and weight.
- c. Given prior probabilities of Basketball, Baseball and Football as 0.3, 0.4 and 0.3 respectively. Write a function to classify a player into one of the sports, given his height and weight. (Hint: this function should call the above defined functions in (b)).
- d. Test your function to predict the sport of a player whose height is 190 cm and weight is 100 Kg.
- e. Use this function to predict sports of the players in the data and store the output in a column named 'naive_Bayes'.
- f. Print the confusion matrix.
- g. Calculate Gini index for each of the sports.
- h. Comment on what Gini index suggests in the above context.

(13)

[30 Marks]
