

**INSTITUTE OF ACTUARIES OF INDIA**

**EXAMINATIONS**

**14<sup>th</sup> September 2021**

**Subject CS1B – Actuarial Statistics (Paper B)**

**Time allowed: 2 Hours (14.30 - 16.30 Hours)**

**Total Marks: 100**

- Q. 1)** Assume  $X$  is a random variable which follows Poi (2) where  $X = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$  i.e. values from 1 to 10
- i) Calculate the probability for each of the values of  $X$ . (2)
  - ii) Calculate the cumulative probability for each of the values of  $X$ . (2)
  - iii) Plot a graph of the distribution function of  $X$ . (1)
- [5]**
- Q. 2)** A student is performing study to understand the correlation of temperature among different days of the week. Temperature on weekdays is recorded for 15 weeks in the file Temperature\_data.csv.
- i) Generate,
    - a) Pearson's correlation matrix (3)
    - b) Kendall's rank (3)
    - c) Spearman's rank (3)
  - ii) Comment on the outcomes of the results in a, b and c. (3)
- [12]**
- Q. 3)** A child tosses  $n$  coins and the outcome of heads and tails are recorded in  $n$  samples as  $X_1, X_2, \dots, X_n$ , where,  $X_i$ 's are independent Bernoulli variables with  $p = 0.5$ . The total outcome of  $n$  variables is  $Y = X_1 + \dots + X_n$
- i) Specify the distribution of  $Y$  (2)
  - ii) Simulate a sample of 10 values for  $Y$  (2)
  - iii) Assess the value of  $Y$  for the sample created in (ii) (2)
  - iv) what is the probability of  $Y = 4$  (1)
- [7]**
- Q. 4)** For a sample data  $X = 0.5820, 0.04981, 0.1552, 0.1555, 0.9036, 0.8501, 0.9288, 0.4408, 0.9688, 0.6300$
- i) Find the mean value of  $X$  (2)
  - ii) Calculate the standard deviation of  $X$  (1)
  - iii) Find the median of  $X$  (3)
- [6]**
- Q. 5)** (Students can copy the R code as provided file 'Reference\_RCode')

**Claim payment data (Year 1 to Year 4) in INR Crores**

	Year 1	Year 2	Year 3	Year 4
Insurer A	112	130	178	150
Insurer B	38	50	80	68
Insurer C	89	127	210	150
Insurer D	70	75	77	80

- i) Analyse the data using EBCT Model 1 and calculate the expected total claim payment to be made by each insurer (A, B, C and D) in the Year 5 by replacing question marks‘?’ with appropriate function in the R Code shared in file “Reference\_RCode”. (2)
- ii) You are now required to calculate the expected total claim payment to be made by each insurer (A, B, C, D and E) in year 5, when additional data of insurer E is made available to you (By using EBCT Model 1).

**Additional claim payment data for Insurer E in INR Crores**

	Year 1	Year 2	Year 3	Year 4
Insurer E	73	87	113	112

- a) Provide the new R code so that ‘data’ includes insurer E and calculate the expected claim payment to be paid by each insurer in year 5. (2)
- b) Comment on the values of Z and ‘s/v’ ratio in part (ii) compared to part (i). (3)
- c) Comment on the change in expected claim payment for insurers A, B C and D in year 5 in part (ii) compared to part (i). (3)
- iii) Your manager is not happy with results based on EBCT 1 model. He has now provided you additional information about volume measure for these 5 insurers over Year 1 to Year 5.

**Volume measure (Year 1 to Year 5)**

	Year 1	Year 2	Year 3	Year 4	Year 5
Insurer A	165	186	198	200	210
Insurer B	51	63	78	83	91
Insurer C	119	159	219	188	192
Insurer D	108	112	122	133	144
Insurer E	103	116	126	151	181

Analyse the data using EBCT Model 2 to calculate the expected total claim payment to be made by each insurer (A, B, C, D and E) in Year 5.

- a) Complete the reference R Code (by replacing question marks‘?’ with appropriate function for data and zi, where data is as used in part (ii) to arrive at the ‘claims5i’ which represents the estimated claim payment for each insurer in Year 5. (2)

- b) Comment on change in expected claim payment for insurer A, B, C, D and E in year 5 in part (iii) compared to part (ii).

(3)

**[15]**

- Q. 6)** You are investigating the level of premium charged by two companies for certain group. Random samples of 10 policies from Company1 and Company2 are compared. Below R code provides the premiums charged by Company1 and Company2 in current year for 10 sample policies.

```
Company1<-c(1350,1790,1500,1150,2100,2350,1550,1800,1650,1450)
```

```
Company2<-c(1500,1200,1300,1700,1800,2400,1450,1950,1850,2100)
```

**Students are required to provide hypothesis, R code, Output and conclusion**

- i) Assuming that the premiums are normally distributed, carry out a statistical test to check equal variance assumption so that it is appropriate to apply a two-sample t test to these data.

Use R code – var.test

(4)

- ii) Test whether the level of premiums charged by Company1 and Company2 are same.

Use R code – t.test (use var.equal =TRUE)

(4)

- iii) The average premium charged by Company2 in the previous year was INR 1500. Test whether Company2 appears to have increased its premiums since the previous year.

Use R code -t.test (#one sided)

(4)

- iv) It was found that out of sample of 200 policies of Company1 and 100 policies of Company2 sold at the start of the year, 60 policies of Company1 and 50 policies of Company2 resulted in claim. Carry out a hypothesis test for the difference in proportions.

Use R code – prop.test (use correct=FALSE)

(4)

- v) Company2 wants to study the claim frequency between group1 policies having premium less than INR 1500 and remaining policyholders. There were 65 claims out of 250 policies of group1 and 45 claims out of 110 other policies in a year. Assuming number of claims having Poisson distribution, test at 2.5% level whether the ratio of claim frequency between group1 and other policyholders is less than 1.

Use R code – poisson.test (#one sided)

(4)

**[20]**

- Q.7)** You are working as an Actuarial analyst in a medium size health insurance company in India. Your manager has asked you to analyse the claim amounts paid data of the past six months.

You have received the data set “HealthClaims.csv” from the claims department of your company with the following explanations of the data fields

GEOGRAPHY: The geographical region of residence of the Insured

PROFESSION: Profession of the Insured

GENDER: Gender of the Insured

AGE: Age of the Insured

CLAIM\_AMOUNT: Amount of health claim paid by the Insurer

Refer to the data set “HealthClaims.csv”.

- i) Fit a linear regression model to the data with “CLAIM\_AMOUNT” as the response and other variables as explanatory variables (consider “Age” as numerical variable and others as categorical variables).

Provide your interpretation of the model by explaining R-Squared, Adjusted R-Squared & p-value of the model. Identify the significant variables in the prediction of claim “CLAIM\_AMOUNT”.

(10)

- ii) Determine 95% confidence intervals for the parameters of the regression model.

(3)

- iii) Plot “QQ plot of the residuals” and comment on applicability of linear regression model.

(4)

- iv) Your manager has also suggested you to include the interaction effects between Geography and Profession, Profession and Gender, Gender and Geography as additional explanatory variables to the set of independent variables taken in (i) for the purpose of improvement of the model. Comment on the suitability of inclusion of interaction effects for the purpose of improvement of the model.

(10)

- v) One of your friend working with Actuarial consulting firm told you that natural logarithm{ $\log_e(\text{CLAIM\_AMOUNT})$ } is better fit to normal distribution than “CLAIM\_AMOUNT”. You wanted to validate the same by fitting a linear regression model to the data with  $\log_e(\text{CLAIM\_AMOUNT})$  as the response and other variables as explanatory variables. Identify and comment on the key differences between the models in (i) and (iv).

(8)

[35]

\*\*\*\*\*