INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

29th May 2024

Subject CS1B – Actuarial Statistics (Paper B)

Time allowed: 1 Hour 45 Minutes (09.30 – 11.15 Hours)

Total Marks: 100

Q. 1) Consider a random variable X which represents the project completion time for a process automation project for a FMCG company (in years). Three scenarios have been contemplated for completion of the project which are as follows:

Scenario		**Probability Distribution**
Scenario 1	:	X has a beta distribution with parameters $\alpha = 5$, $\beta = 1$;
Scenario 2	:	X has a beta distribution with parameters $\alpha = 1$, $\beta = 5$;
Scenario 3	:	X has a beta distribution with parameters $\alpha = 3$, $\beta = 3$.

i) Calculate the following probabilities for X under Scenario 1:

a) Use pbeta function to calculate $P(0.2 < X < 0.8)$; (2)

b) Use qbeta to find x such that $P(X > x) = 0.65$. (2)

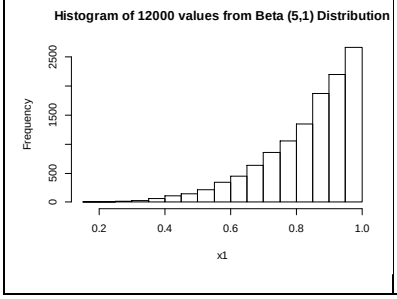
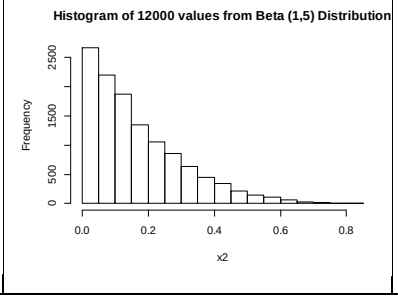
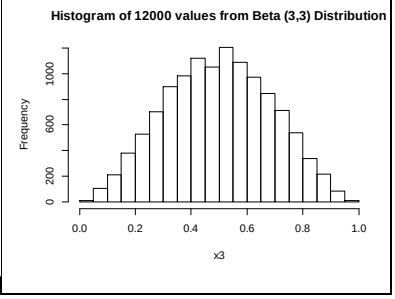
ii) Use the following code to calculate the coefficient of skewness of X under all the three scenarios:

Skew = 2*((β-α)/(α+β+2))* sqrt((α+β+1)/(α*β)) (3)

iii) Under all the three scenarios, write a code to simulate a sample of 12,000 values from a Beta distribution. Use the command set.seed(421967) to initialize the random number generator, before you start the simulation. Simulate this sample by defining vectors "x1", "x2" and "x3" for Scenario1, Scenario2 and Scenario3 and write code to draw histogram for each of the scenarios.

You are NOT required to execute the code or print the result of these vectors or reproduce the histograms. (3)

Histograms for these simulated samples have been given below:

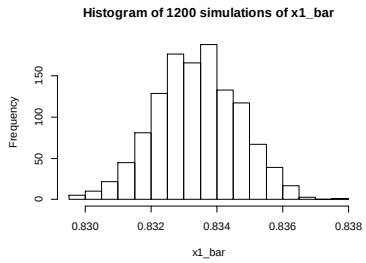
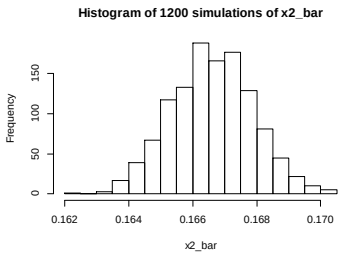
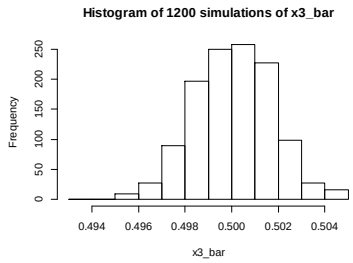
Scenario 1 $X_1 \sim$ **Beta** ($\alpha = 5, \beta = 1$)	**Scenario 2** $X_2 \sim$ **Beta** ($\alpha = 1, \beta = 5$)	**Scenario 3** $X_3 \sim$ **Beta** ($\alpha = 3, \beta = 3$)
		

iv) Comment on the shape of the histograms plotted above by taking into consideration coefficient of skewness calculated in part (ii). (3)

v) Write a code to perform 1,200 repetitions of the simulations in part (iii) for all the three scenarios. You should compute and store the value of the mean of the sample (x1_bar), (x2_bar) and (x3_bar) for each repetition. Use the same command set.seed(421967) to initialize the random number generator, before you start the simulations.

You are NOT required to execute the code or print the output of these simulations or histogram of these simulations. (5)

Histograms of sample means for the three scenarios have been given below:

Scenario 1: x1_bar	**Scenario 2:** x2_bar	**Scenario 3:** x3_bar
		

vi) Comment on the shape of the histograms, by referring to the central limit theorem. Also, compare and contrast with your observations in part (iv). (2)

[20]

Q. 2) As per a recent research, the maximum systolic blood pressure for a person is related to age and can be expressed in terms of the following equation:

maximum systolic blood pressure = 100 + age (in years)

i) If we decide to fit a regression line with maximum systolic blood pressure as the response variable (Y) and age as the explanatory variable (X), what should be the values of the regression coefficients α and β in light of the above equation? (2)

Suppose this is to be empirically proven and 20 people of varying ages are tested for their maximum systolic blood pressure. The following data has been collected:

Age (X)	**Maximum Systolic Blood Pressure (Y)**	**Age (X)**	**Maximum Systolic Blood Pressure (Y)**
28	132	60	149
37	140	55	154
41	155	29	117
52	160	43	146
57	167	36	142
49	148	50	168
38	128	34	144
25	131	40	156
23	118	26	114
48	139	33	133

The data of age in years (X) and corresponding maximum systolic blood pressure (Y) can be entered in R using the below code:

x=c(28, 37, 41, 52, 57, 49, 38, 25, 23, 48, 60, 55, 29, 43, 36, 50, 34, 40, 26, 33)

y=c(132, 140, 155, 160, 167, 148, 128, 131, 118, 139, 149, 154, 117, 146, 142, 168, 144, 156, 114, 133)

Enter the data in R by copying the above code.

ii) Prepare a scatter plot of the data and briefly comment on the same. (3)

iii) Plot the fitted line for regression of Y on X. (4)

iv) Using anova function check whether the slope parameter is significant. (3)

v) Obtain a summary for the linear regression model fitted in part (iv) and clearly state the estimates for the values of the coefficients α and β. (3)

vi) Test the veracity of the above equation as promulgated in the recent research, by referring to your answers in part (i) and part (v). (2)

vii) Calculate the sample Pearson correlation coefficient, sample Kendall correlation coefficient and sample Spearman correlation coefficient. (3)

viii) Comment on each method of calculating correlation coefficient and also comment on the sample correlation coefficients obtained in part (vii). (4)

One of your colleagues says that there exists perfect positive correlation between age and maximum systolic blood pressure.

ix) Perform a hypothesis test to test the above statement at 5% level of significance. You should state the null and alternate hypotheses, report the p-value of the test and arrive at a clear conclusion.(Hint: Use Pearson Method) (6)

[30]

Q. 3) Consider a portfolio of fire insurance policies. The data relating to 192 policies which have claimed at least once till now, is given in the file Firepolicies.csv. Confirm through output that the file contains data regarding four variables:

- Occupancy: The claim has arisen in five different Occupancies:
  TM – for Textile Mills
  DW – for Dwellings
  TG – for Transporters' Godowns
  CS – for Cold Storage Premises
  HG – for Hazardous Goods Storage
- Location: It relates to the loss state:
  M – for Maharashtra
  G – for Gujarat
  T – for Telangana
  K – for Karnataka
  A – for Andhra Pradesh
- Claim.size: It refers to the amount of the claim in INR lakhs.
- Claimed: This is an indicator variable which captures whether there is an incidence of claim in the past one year.
  0 – refers to NO claim in the past one year
  1 – refers to ONE claim in the past one year

i) View the data Firepolicies.csv for 192 entries using read.csv function. Columns Claim.Size, and Claimed have numeric data type, all other columns i.e. Location and Occupancy have character data type. (1)

ii) Using the Firepolicies.csv, calculate the proportion of:

a) Policyholders with no claims in the past one year from the state of Maharahstra (3)

b) Policyholders with no claims in the past one year from the state of Gujarat (1)

iii) Test the hypothesis that the proportion of policyholders with no claims in the past one year is equal in both states (Maharashtra and Gujarat) at 5% level of significance. You should state the null and alternate hypotheses, report the p-value of the test and arrive at a clear conclusion. (5)

iv) Test the hypothesis that there is no significant difference in the average claim size for Textile Mills and Transporters' Godowns at 5% level of significance. You should state the null and alternate hypotheses, report the p-value of the test and arrive at a clear conclusion. Assume that population variances are equal. (6)

v) The number of policies with at least one claim (X) for the fire insurance portfolio is modelled as a random variable with a Binomial distribution X ~ Binomial (n, p).

An Actuary wishes to fit different Generalized Linear Models (GLMs) to the data, assuming that the number of policies with one submitted claim has a Binomial distribution and the link function of the GLM is the logit function.

a) Fit a GLM to the data such that p depends on the Location and Claim Size i.e., Loss State and the Claim size and report the summary. (4)

b) Fit a GLM to the data such that p depends on the Occupancy and Claim Size and report the summary. (2)

vi) Compare the fit of the models in parts (v)(a) and (v)(b) using Akaike's Information Criterion (AIC) and comment on which model is preferable. (3)

vii) Among the four variables as given in the Firepolicies.csv data, which of them are numerical variables and which of them are factor variables? What is the difference between the two in the context of generalised linear models? (5)

[30]

Q.4) The following data represents the average claim size (motor third party - accident injury) under motor insurance policies settled in six different states across a country.

State i	**Year j**				
	2018-2019	**2019-2020**	**2020-2021**	**2021-2022**	**2022-2023**
Assam	219458	240371	289307	264439	279704
Bihar	216594	231311	261915	286211	291934
Chattisgarh	213871	231461	264519	261279	270058
Delhi	389197	400926	393130	391921	373129
Odisha	224879	243361	276718	292564	300135
Kerala	194814	230113	258101	276876	287973

Enter the data in R in the form of a matrix using the following code:

```
Claims<-matrix(c(219458, 216594, 213871, 389197,224879,194814,240371,231311,231461,4
00926,243361,230113, 289307, 261915, 264519, 393130, 276718, 258101, 264439,286211, 2
61279, 391921, 292564, 276876, 279704,291934, 270058, 373129, 300135, 287973), nrow=6
, ncol=5)
```

Or

(You can copy the R code from provided file 'Q4 Reference_RCode.docx)

i) Calculate, using Empirical Bayes Credibility Theory (EBCT) Model 1, the following:

a) E[m(ɵ)] (2)

b) E[s^2(ɵ)] (2)

c) Var [m(ɵ)] (4)

ii) Calculate the credibility factors Z_i and the credibility premiums for Delhi and Kerala. (5)

iii) Comment on the relationship between n and Z_i in case of EBCT Model 1. (2)

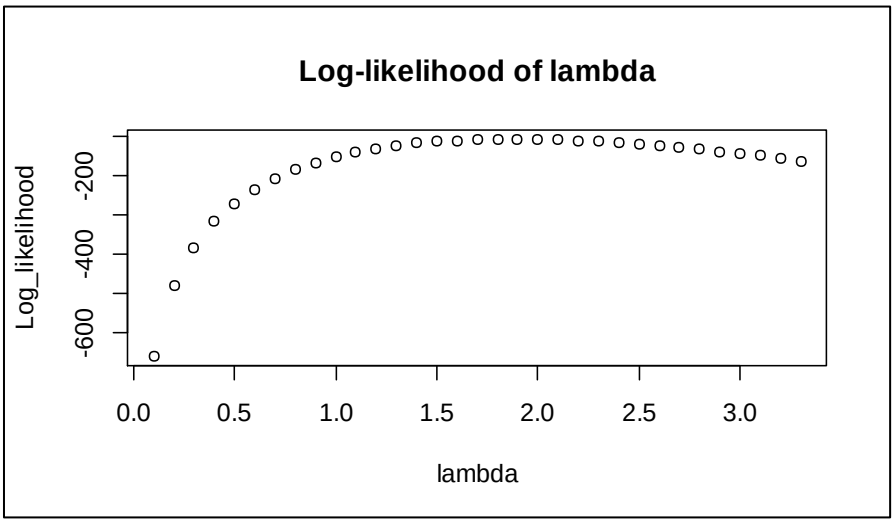
The number of motor third party claims per policy is modelled as a random variable X with a Poisson distribution with unknown parameter ʎ. The log likelihood function for estimating ʎ is given by:

$$L(ʎ) = \log(ʎ) \times \sum x - ʎn$$

where n is the number of observations in the sample data.

There is data for a total of 150 observations and the total number of motor third party claims is 280.

The log likelihood function for the values of ʎ = 0, 0.1, 0.2, …….., 1.8, 1.9, 3.3 has been plotted below:



iv) Determine an approximate maximum likelihood estimate for ʎ using this plot. (2)

v) Determine the exact maximum likelihood estimate for ʎ and compare your answer with the approximate estimate obtained in part (iv). (3)

[20]

****