PUSASQFMIN303
Data Analytics Applications of IT - R Basics

Time: 2 hours
Total Marks: 60 marks

Note:

1. **The candidate has option to either attempt question 3A or question 3B. Rest all questions are mandatory.**
2. **Numbers to the right indicate full marks.**
3. **The candidates will be provided with the formula sheet and graph papers (if required) for the examination.**
4. **Use of approved scientific calculator is allowed.**
5. **All answers in R are to be converted to word and the Rscript and Word both are to be submitted.**

Q1 Answer the following **15 Marks**

A. **5 Marks**

The claim amounts (in Lakhs) from a portfolio of insurance policies follow a Lognormal distribution with parameters 5 and 2.

(a) Calculate the probability that a claim amount will be higher than 67.5 Lakhs. (2)

(b) Calculate the 75^{th} percentile of the claim amount using inbuilt R functions. (1)

(c) Generate 100 random values from the given distribution using set.seed(2109) and calculate the empirical value of the probability from (a). (2)

B. **5 Marks**

Use the inbuilt data set iris for this question.

Generate a boxplot of Petal.Width for each Species and comment on the graph. Make sure the plot is well labelled and has separate colored boxes for each species.

C. **5 Marks**

The number of MCQs answered correctly by the students on a test of 10 MCQs are provided below:

Number of MCQs answered correctly	<4	5	6	7	8	>=9
Number of Students	14	15	28	16	18	9

(a) Conduct a Chi-Square goodness of fit test for the Binomial distribution with p = 0.7. (4)

(b) State the p-value for the test above. (1)

Q2 Answer the following **15 Marks**

A. **5 Marks**

Use the Returns_DataPaperA.csv file for this question.

(A) Calculate the mean and standard deviation of the "return" for each value of "rating". (4)

(B) Extract the Fund number and the Rating of the fund with the highest return from the given data. (1)

B. **5 Marks**

The Institute of IQ specializes in education for children with learning disabilities. The students are required to take a standardized test before joining the institute and a similar test is taken after one year of studying at the institute.

The following data is corresponding to 8 different students of the institute:

Student Number	1	2	3	4	5	6	7	8
Before	145	147	123	137	141	142	140	138
After	155	152	146	153	146	160	139	148

(a) Conduct an appropriate test in R to test if the score of the student improves after joining the Institute at 5% level of confidence. (4)

(b) Show the 95% confidence interval for the population mean of the change in marks. (1)

C. **5 Marks**

Use the inbuilt data set iris for this question.

An actuarial student fascinated with the Botany in the above set of plants is currently in the process of understanding the relationship between Sepal.Length and Petal.Length of different flowers.

In particular he wants to fit the following model : $y_i = \alpha + \beta x_i + e_i$

Where y_i is the Sepal.Length and x_i is the Petal.Length of each flower. e_i follows a Normal distribution with mean 0.

(a) Fit the required linear regression model and show the output. (2)

(b) Calculate the confidence interval for β and comment on the significance of Petal.Length in predicting Sepal.Length. (2)

(c) State the R^2 for the fitted model in (a). (1)

Q3 Answer the following **30 Marks**

A. **30 Marks**

The sizes of health insurance claims, Y, for older policyholders are assumed to be independent gamma random variables with $\alpha = \frac{1}{3}$. A sample of claims is taken and the age of the policyholder, gender of the policyholder and the city in which the policyholder lives. (Mumbai, Pune or Nashik).
The results are stored in the file HealthClaims_PaperA.csv.

(a) Construct the null model for the claim size (using the canonical link function) and determine its AIC (3)

(b) Calculate the AIC for the three models for the claim size using one of each of the three covariates as the main effects and comment on the values obtained. (8)

The two covariates that have the most effect on the value of the AIC are retained.

(c) Construct a model for the claim size using both of these covariates as main effects and show that this is the best model so far using the AIC. (4)

(d) Using a Scaled Deviation Test, show that including interaction between the two covariates in part(d) does not improve the model. State the p – value clearly. (4)

(e) Calculate the deviance and Pearson residuals for the model chosen in (c). (3)

(f) Comment on which of these sets of residuals is likely to be better for testing the fit of the model. (1)

(g) Comment on the values obtained for the set of residuals chosen in (f) (3)

(h) Calculate the predicted claim amount for a:

- 75 year old Male who lives in Mumbai
- 90 year old female who lives in Nashik (4)

OR

B. **15 Marks**

The blood pressure of patients are thought to be normally distributed with a mean of 140 and standard deviation 35
Under the Central Limit Theorem, the mean of a large sample of patient's BP will be normally distributed.

(a) State the parameters of the distribution of the sample mean for samples of size 20. (2)

Since the median of a normal distribution is the same as the mean, a student hypothesizes that the median of a sample of patient's BP will also be normally distributed with the same parameters as in part (a)

(b) Perform a simulation of a sample of $y_1, y_2 \ldots y_n$ of patient's BP with sample size 20 and a seed value of 2226. (2)

(c) Calculate the median M, for the sample in part (b) (1)

(d) Perform 1000 repetitions of parts, (b) and (c) to obtain a bootstrap sample of $M_1, M_2, \ldots M_{1000}$ of the median, using the same set seed as before. (7)

(e) Plot a histogram showing the densities of the sample $M_1, \ldots M_{1000}$ from part (d). Label the graph appropriately. (4)

(f) Superimpose the density of the sample mean using your result in part (a) on the histogram from (e) (3)

(g) Compare the distribution of the sample median and the distribution of the sample mean given by the Central Limit Theorem using the graph from part (f). (3)

Despite the differences between the distribution of the sample mode and the sample mean, the student still believes that the sample mode will be normally distributed.

(h) Draw a Q-Q plot of the bootstrap sample from (d) (3)

(i) Comment on whether the diagram in (h) supports the students claim by adding an appropriate red dashed line to the Q-Q plot to show the expected result if the modes were normally distributed. (3)