PUSASQF206
Application of IT – Basics of R

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.**

Q1. **15 Marks**

Q1A. **5 Marks**

A random variable Y is assumed to follow a Gamma Distribution with mean 4 and variance 8

(a) Simulate 100 values from the above distribution , using the seed as the year of your birthdate. Show the empirical mean , median and mode.

$$If\ X \sim Ga(\alpha, \lambda)\ then\ mode\ of\ X\ is \frac{\alpha - 1}{\lambda}$$

(b) Compare your answers in part (a) , which the population values for the measures of central tendency.

Q1B. **5 Marks**

A mentor at an actuarial classes is investigating the claim that number of students in class does not affect Exam Marks. His observations of nine separate class that wrote the same paper were as follows :

Number of students (n)	35	32	27	21	34	30	28	24	7
Average Exam Marks (m)	59	41	24	17	63	53	35	26	16

(a) Create a scatterplot of the above data and comment on the relationship between the two variables. The points should be green colour * shaped. The graph should contain appropriate title and axis labels. (3)

(b) Conduct a test at 5% level of significance to check whether the Pearson's correlation co-efficient is equal to 0.8 against an alternative that it is greater than 0.8 . (2)

Q1C **5 Marks**

From the package 'datasets' , procure the 'Orange' database.

(a) Fit a linear model of Circumference (in mm) over Age (number of days since 31/12/1968) (1.5)
(b) Obtain the standard error for the above model. (0.5)
(c) By creating an appropriate plot, comment whether the residuals follow a normal distribution or not. (3)

Q2 **15 Marks**

Q2A. **5 Marks**

You are interesting in understanding the time people have to wait for a Cab at the Mumbai Airport. You decide to spend a day at the pick – up point and record the waiting time for the passengers on that day. *Run the code below :*

```
set.seed(1729)
n = 100
wait.times = -log(runif(n))/0.1
```

(a) Assuming that the wait times are exponentially distribute use set.seed(2022) to obtain the median of 1000 samples of size 100. (3.5)
(b) Hence calculate the 97% parametric bootstrap confidence interval for the median wait time. (1.5)

Q2B. **5 Marks**

A collegue of yours has recently created a Generalized Linear Model with Weight as the response variable and Time, Time^2 and Diet as the explanatory variables using the following code :

```
glm = glm(weight~Time + I(Time^2) + Diet,
          data = ChickWeight,
          family = gaussian(link = "log"))
```

Unfortunately, he fell ill before he could conclude the next phase and the work has now been assigned to you.

(a) Create a QQ - plot of the deviance residuals and comment if they follow a Normal Distribution. (3)
(b) Calculate the predicted value of Weight when Time = 12 and 3rd Diet is being supplied to the Chick. (2)

Q2C. **5 Marks**

Ram and Karim just completed an investigation which required the application of a two – sampled *t-test* to compare two independent samples each of size 11. Upon discussion , they found that they had missed to conduct an test to check their equal variance assumption required for this test. Their data was stored in vectors `males` and `females` : Run the below code to load them in your R session :

```
Males <- c(21,22,28,27,20,23,26,32,25,21,30)

Females <- c(19,18,38,33,24,39,22,29,28,26,30)
```

(a) Conduct an appropriate test in R. Your output should contain the alternative hypothesis and the p – value for the test. (4)
(b) Comment on the validity of the test conducted by Karim prior to (i) (1)

Q3. Attempt any one from 3A and 3B. **30 Marks**

Q3A. **30 Marks**

The runs scored by Sachin and Lara in 8 instances each is given a**s** below:

Sachin	12	35	65	125	118	48	28	48
Lara	75	08	28	107	137	45	36	60

A cricket enthusiast actuarial student wants to test if the runs scored by both the legends are distributed similarly.

(a) Conduct a paired t – test , as the scores are from the same day and stadium. Comment on the result of the test. (3)

(b) By assuming that the scores are normally distributed and have the same variance, calculate the 99% confidence interval of the variance of the scores. (5)

Now the student wants to test if the scores scored by Sachin and Lara have the same median. I.e. the difference of median scores by Sachin and Lara is 0. As there are no sampling distributions for the median, he decides :

(c) He stores the differences of the two vectors in a separate vector called *D*, and its median in ObsT. (2)

He says: "As the median would be 0, each score would be equally
likely to be positive or negative, and hence if I calculate the median of the differences for each possible permutation of the signs, and compare it with the
observed median , I would be able to conclude if the median is0".

(d) Create an object **sign** that contains -1 and 1 (1)

(e) Use the permutations() on the object **sign** to generate all possible permutations. (3)

(f) Use a for()/while() loop to store the median of the differences in an object called *dif* (8)

(g) Plot a labeled histogram of the median of the differences of all possible permutation. (3)

(h) Use abline() to add vertical lines to show critical value and observed value. Use red for critical value and blue for Observed value. State your conclusion for the test using the Histogram. (5)

OR

Q3B. **30 Marks**

HealthStar is a health insurance company in the country of Actuaria. You are an analyst at the company and are asked to analyse the claims experience of the past 6 months since the inception of the company.
The claims department has provided the data of the claims in a file called *Claims_Experience.csv* along with the details regarding what each column means :

LOCATION : The location of residence of the insured
JOB : Profession of the insured
SEX : Biological gender of the insured
AGE : Age of the Insured
CLAIM : Amount of the claim paid by the insurer.

(a) Import the file in R and print the number of rows it contains to the console. (1)

(b) Plot the empirical density of CLAIM and suggest an appropriate distribution for its modelling (3)

(c) Create a GLM model with inverse link and distribution assumption, with CLAIM being the explanatory variable and all other variables and all their interaction as the explanatory variables. Why are some of the co – efficients NA? (3)

(d) State the effectiveness of SEX as an explanatory variable for estimating claim amount. (1)

(e) Conduct a test at 5% level of significance to test if the intercept parameter is equal to 0.001 (3)

(f) By performing forward selection, obtain the model with the best fit. You should use the criteria that all variables in the model should be significant at 7.5% level of significance. (10)

(g) Plot a Q – Q Plot of the Pearson residuals and state whether they follow a Normal Distribution. (3)

(h) Compare the RMSE (Root Mean Square Error) for models in (c) and (e) (3).

(i) Find the ratio of the estimated claim amount for :
- 34 year old MALE living in Region 1 who owns a Business
- Self – employed 27 year old Female living in Region 8 (3)