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 X is assumed to follow an exponential distribution with mean 3.4

(a) Simulate 10 random values of X , using the seed as the year of your birth and show the empirical mean and variance. (2)

(b) Repeat part (a) by , simulate 100 values from the same seed. (1)

(c) Compare your answers in parts above to the population mean and variance. (2)

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) Calculate Pearson's and Kendall's correlation coefficients and comment on the values (2)

(b) By stating appropriate hypothesis and using Spearman's correlation coefficient to test whether or not the data agrees with the claim that class size does not affect exam marks. (3)

Q1C **5 Marks**

From the package ‘datasets’ , procure the ‘Orange’ database

(a) Fit an intercept – free linear model of Circumference (in mm) over Age (number of days since 31/12/1968) (2)

(b) By creating an appropriate diagnostic plot, comment if there exists any relationship between errors and the fitted values (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 normally distributed , calculate the 92.5% confidence interval for the mean wait times. (3)

(b) By creating a histogram of the wait times comment on the assumptions in part (i) (2)

Q2B. **5 Marks**

During an interview process, you are required to assess two different GLM models. The interviewer has provided you with the following code being used to build the model.

```
model1 = glm(Sepal.Length~Petal.Length + Species,
             data = iris)
model2 = glm(Sepal.Length~Petal.Length + Sepal.Width,
             data = iris)
```

(a) Compare both models using Residual deviance as a criteria. (1)

(b) State why residual deviance is not an appropriate criteria for comparing these two models. (2)

(c) Compare the two models using any appropriate criteria. (2)

Q2C. **5 Marks**

CSK Team management is concerned about the fairness of the coins used for tosses in the recent IPL matches. Upon their request , the match officials conducted multiple sets of 4 coin tosses and recorded the number of times **heads** shows up. The figures reported by the officials were as follo ws :

2,1,2,3,1,1,1,1,2,2

(a) Store the above values in a vector called `heads`

(b) Conduct an appropriate test in order to ascertain the fairness of the coin by stating the NULL and ALTERNATE hypothesis and the exact p – value for the NULL hypothesis.

Q3. Attempt any one from 3A and 3B.

Q3A. **30 Marks**

An investigative agency is currently conducting a health survey for its clients. The focus of the investigation is to understand how the heights of young children are distributed in the small city of Galactica.

At the moment , the height of young children are thought to be normally distributed wih a mean height of 132 cm with a standard deviation of 12.32 cm.

(a) State the distribution and the parameter/s of the sample mean for samples of size 20. (2)

Another trainee states that the mode of the normal distribution is the same as the mean, he asserts that mode of a sample of heights will also be normally distributed with the same parameters as (a)

(b) Perform a simulation of a sample $x_1, x_2, x_3 \ldots, x_n$ of the heights for a sample of size 20 and a seed value of 1947. (2)

(c) Calculate the empirical mode for the sample in (b) (2)
[Hint : `density()` outputs `x` and `y` , where `y` is the empirical pdf for each corresponding value of `x`]

(d) Perform 10000 repetitions of parts (b) and (c) to obtain a bootstrapped sample of the mode using the same set seed as before. (9)

(e) Plot a histogram showing the densities of the sample modes from part (d) , with a y – axis that goes up to 0.15. Superimpose the density of the distribution discussed in part (a) on the histogram. (6)

(f) Compare the distribution of the sample mode with that of the distribution of the sample mean given by the Central Limit Theorem, using the graph in part (e). (2)

Despite the differences observed in (e) , the trainee still believes that empirical mode should be normally distributed.

(g) Create a Q - Q Plot of the sample modes from part (d) and by adding a line to the Q – Q plot to show the expected result if the modes were normally distributed, comment whether there is any proof of trainee's assertion. (7)

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) Fit a linear regression model to the data with CLAIM as the response variable and AGE as the explanatory variable.
Provide your interpretation of the model by commenting on the p – value of the F test. (4)

(c) Generate all the diagnostic plots of the model in (b) and comment on whether the underlying assumptions of the linear regression model are fulfilled. (6)
Your manager has asked you to update the methods and instead fit a GLM model by still assuming that the CLAIM are from a Normal distribution.

(d) Fit an equivalent model to the model in (b) (2)

(e) Add the main effects of the other variables to the model in (d) and comment on the effect of each variable to the claim amount. (4)

(f) Another suggestion mentions that the log of claim amounts is a better fit for the normal distribution then the claim amounts directly. Change the link function of the model with all explanatory variables remaining the same. Compare the model with that in (e) (4)

(g) By performing forward selection and AIC as a criteria, determine which is the best possible model. You can use all variables, their two – way and multi- way interaction. (9)