

PPSAS205
Introduction and Modelling in 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.
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.

Q1 Answer the following

15 Marks

Q1 A

5 Marks

The individual claim amounts from a portfolio of Fire insurance policies follows a Normal distribution with mean 10000 and variance 90000.

- (a) Calculate the probability that a claim amount will be higher than 10600. (1)
- (b) Calculate the 99th percentile of the claim amount using inbuilt R functions only (2)
- (c) Generate 10 random claim amounts following the same Normal distribution and compare the sample variance with the given variance (2)

Q1 B

5 Marks

Use the inbuilt data set iris for this question.

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

Q1 C

5 Marks

The Insurance Company of India has underwritten a large portfolio of motor insurance policies. The actuarial team is currently in the process of checking if the claim count from the portfolio follows a Geometric distribution or not.

The claims received from each of the policies are given below:

X	4	5	6	7	8	9
Frequency	3	9	18	29	25	16

Conduct a Chi-square goodness of fit test for the Geometric distribution along with the p-value for the test.

Q2 Answer the following**15 Marks****Q2 A****5 Marks**

Use the dance.csv provided to you for this question.

The file contains the following fields :

- Judges : The score given by the Judges
- Poll: The score given by the audience
- Final: The final score of the contestant

- (a) Fit a Generalised Linear Regression model for the Final Score and the main effects of other variables store it in fitA under the Gamma family and identity as the link function. (3)
- (b) Construct a plot of Pearson Residuals against the Fitted Values and comment on the relationship between the two. (2)

Q2 B**5 Marks**

Use data_Q2B_PaperB.csv for this question.

A female student actuary is currently helping out her friend in conducting a study on the weight of 25-30 year old women. For this study, the weights of 100 women in the age group were taken. The student actuary is worried that the variance of the data collected is too high and hence wants to conduct a test to make sure that the variance is within reasonable limits. She is

She is conducting the following test:

$$H_0: \sigma^2 = 121 \text{ vs } H_1: \sigma^2 > 121$$

Calculate the test statistic and p - value of the following test in R.

Q2 C**5 Marks**

Use the inbuilt data set mtcars for this question.

- (a) Using the aggregate() function in R , calculate the mean of the “mpg” for each of the different combinations of “am” and “vs”. (2)
- (b) Draw a scatterplot in R of “mpg” and “disp”. The point type/colour should vary with “am”. Make sure the graph is labelled and has a legend (3)

Q3 Answer the following

Q3 A

30 Marks

The heights of 10-year-old boys are thought to be normally distributed with a mean of 1.512m and standard deviation 0.0741m

Under the Central Limit Theorem, the mean of a large sample of boy's heights will be normally distributed.

- (a) State the parameters of the distribution of the sample mean for samples of size 15. (2)

Since the mode of a normal distribution is the same as the mean, a student hypothesizes that the mode of a sample of 10-year-old boy's heights will also be normally distributed with the same parameters as in part (a)

- (b) Perform a simulation of a sample of $y_1, y_2, \dots, y_n$ of 10-year-old boy's heights with sample size 15 and a seed value of 2024. (2)

- (c) Calculate the mode M , for the sample in part (b) (3)

- (d) Perform 1000 repetitions of parts, (b) and (c) to obtain a bootstrap sample of $M_1, M_2, \dots, M_{1000}$ of the mode, using the same set seed as before. (8)

- (e) Plot a histogram showing the densities of the sample $M_1, \dots, M_{1000}$ from part (d), with a y – axis that goes up to 25. (3)

- (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 mode 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)

OR

Q3 B**30 Marks****(I)**

The Royal Medical Institute in Pune claims that they have invented a medicine which allows a Human to breath under water and stay under water longer than before. In order to ascertain the truth value of their claims, an independent study is conducted on 20 people and their time under water was noted, 10 who consumed medicine and 10 who didn't consume the medicine. The data collected was in minutes and is as follows:

T	2.36	4.85	3.65	7.82	5.63	4.89	10.1	8.53	4.12	6.58
C	5.63	5.2	4.78	9.89	8.72	5.98	0.56	4.12	1.85	4.02

T is the group that took the medicine.

- . Test whether there was any increase in average amount of time spent under water due to medicine assuming the variances were equal for both the groups. (3)
- b. Test the assumption that the variances are equal for both the groups. (2)
- c. Carry out a permutation test to test the hypothesis that Control(C) group has a lower time under water than T. Display the *p-value* of the test conducted on every combination using the appropriate R functions. (15)

(II)

1. StarTree Insurance Co. Receives claims daily. The number of claims received for the past 365 days are as follows:

Claims	0	1	3	4	5	8
Freq	28	126	111	31	30	39

- . A new analyst suggests that the claims received daily follow a Poission distribution with λ , the rate parameter equal to the sample mean. Use *chisq.test()* to test the Goodness-of-Fit of the distribution suggested by the new analyst. State the correct *p-value for the test*. (5)
- b. Another suggestion is that, the Claims received daily follow a Binomial Distribution with parameters $n=8$ and $p=0.358$. Test the Goodness-of-Fit for the Binomial Distribution. State the correct *p-value for the test*. (5)