Submission Deadline: October 15, 2023 EOD

Assignment 1

Instructor: Rajeeva Karandikar, rlk@cmi.ac.in
TAs: {nidhi, sayantika, ujan, varuna, vinodh}@cmi.ac.in

Probability and Statistics with R

Instructions

1. Each question is worth 10 marks.
2. For your submission, please include an R Markdown (Rmd) file containing both code and output, along with a report written in LaTeX or handwritten that explains your approach and answers.

Question 1

Consider a continuous random variable X with probability density function

$$f(x) = \frac{ax^{a-1}}{c^a}; 0 \leq x \leq c$$

(i) Obtain the Cumulative Distribution function for the given pdf.

(ii) Obtain the inverted CDF

(iii) Write a function in R to generate a sample from the given distribution using the inverse-transform method which takes in parameters a and c. You may use this article for reference.

(iv) Using R, plot the inverse CDF when $a = 5$ and $c = 10$ and determine the range of values with the least probability of being generated.

(v) Use the above function to draw a sample of size 1000 from the given distribution with parameter values $a = 5$ and $c = 10$.

(vi) Using R, check if the empirical mean is approximately equal to the theoretical mean.

Question 2

You are conducting a study to understand the distribution of the average heights of people in a large city. You have access to the heights of 10,000 individuals, but collecting data from such a large population is time-consuming and costly. Instead, you decide to use a random sample of 100 individuals to estimate the population mean height.Your task is to:

(i) Simulate the heights of 10,000 individuals in the city. Assume that the heights follow a) Uniform, b) Poisson, c) Normal distribution. Assume the mean and variance for each distribution is `160` cm and `160` cm^2 respectively.

(ii) Randomly select 100 individuals from all three (a, b, c above) simulated population and calculate the sample mean height.

(iii) Repeat step 2 a large number of times (e.g., 1,000 or more) to create a distribution of sample means.

(iv) Calculate the population mean height and the standard error of the sample mean from the simulated data.

(v) Plot a histogram of the distribution of sample means and overlay a normal density plot with mean and standard deviation calculated from previous question.

(vi) (Optional) Discuss your findings and any insights regarding the sample mean distribution compared to the population distribution.

Question 3

In this question, you will determine whether pregnancy smoking is harmful to a child's health. You will use the "Birth Weight" and "Maternal Smoker" columns from the baby dataset.

(i) Plot overlaying histograms of birth weights of babies born to non-smoking mothers and smoking mothers. Write down your observations about the mean of each distribution.

(ii) Find the difference between the average weight of the smoking group and the average weight of the non-smoking group. This is the observed difference.

(iii) If there were no difference between the two distributions in the underlying population, then whether a birth weight has the label True or False with respect to maternal smoking should make no difference to the average. The idea, then, is to randomly shuffle all the labels among the mothers. This process is called random permutation. Using this idea, shuffle the labels and find the difference you found in (ii) for this new labelling. Repeat this step a large number of times (1,000 or more), calculate the mean and variance of the differences, and plot a histogram.

(iv) Plot the observed difference you obtained in (ii) and overlay this point to the histogram mentioned in (iii).

(v) Calculate how many standard deviations the observed difference is from the mean of the simulated variables. What conclusions can you draw from this?