Assignment 2

STAT 3006, 2021/22

Due date: Mar 25, 2022 23:59, Friday

Question 1

(25%) Please use the inverse method to generate 5000 samples from a Poisson distribution($\lambda = 5$) and draw a histogram for the 5000 samples. The probability mass function of a Poisson distribution(λ) is

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

with $k = 0, 1, 2, 3, ...$

Question 2

(25%) Please use the accept-reject (AR) method to generate 5000 samples from a truncated Gamma distribution $X \sim Gamma(\frac{1}{2}, 1)I(x \geq 5)$ with density function

$$f(x) = \frac{x^{-\frac{1}{2}}exp(-x)I(x \geq 5)}{\int_5^{+\infty} y^{-\frac{1}{2}}exp(-y)dy}$$

and show why AR method works. Moreover, please give the theoretical acceptance rate and compare it with the actual acceptance rate calculated by your sampling.

Hint:

- Sample from a shifted exponential distribution $g(x) = exp(-(x-5))I(x \geq 5)$;
- The integration $\int_5^{+\infty} y^{-\frac{1}{2}}exp(-y)dy$ can be calculated by `[1 - pgamma(5, shape = 1/2, rate = 1)] * gamma(1/2)` in R.

Question 3

(25%) Please estimate the integration $\int_5^{+\infty} cos(x)x^{-\frac{1}{2}}exp(-x)dx$:

1. using 5000 samples from $Q2$;
2. using importance sampling (drawing 5000 samples from $g(x) = e^{-(x-5)}I(x \geq 5)$ to estimate the integration).

Question 4

(25%) Salary data set contains all persons' salary in a county. Your goal is to use stratified sampling to estimate this county's salary mean. In the data set, the first column represents salary per month for each person and the second column indicates the age range. "1" denotes the age interval (20, 30], "2" denotes (30, 50], and "3" denotes (50, 70]. There are total $11,000$ persons. $1,500$ persons have age indicator "1", $4,500$ persons have age indicator "2", $5,000$ persons have age indicator "3".

1. Randomly draw 100 samples from the data set, and estimate the standard deviation for each subpopulation;

2. If the total sample number is set as $1,000$, please determine the sample number for each subpopulation;

3. Estimate the population mean salary based on these $1,000$ samples and compare it to $4,307.189$ (underlying truth of the population mean salary).

The data set is stored in `salary_data.txt`.

Requirements

-	in the paper report	in the R code file
Q1	algorithm histogram	R code
Q2	algorithm proof comparison results	R code
Q3	algorithms integration values of the two approaches	R code
Q4	standard deviation sample number population mean salary estimate	R code