Golas, Ekal

UNIVERSITY OF TEXAS AT DALLAS | 800 W CaMPBELL RD RICHARDSON TX 75080

MINI PROJECT 1

Statistical methods for data science



Table of Contents

[Methodology 2](#_Toc430548572)

[Algorithm 2](#_Toc430548573)

[Answers 3](#_Toc430548574)

[R Code 4](#_Toc430548575)

[References 5](#_Toc430548576)

# Methodology

As discussed in the problem statement (Baron, 2009), theoretical calculation in HW1 Solutions (Choudhary, 2015) was difficult. Hence the following algorithm is proposed as an alternative:-

## Algorithm

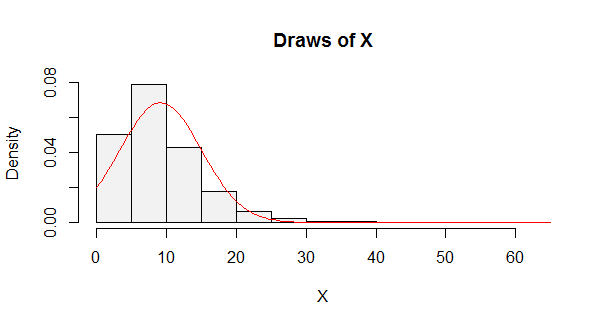
As stated in (Choudhary, Mini Project 1, 2015), a simulation study is done as:-

1. A function is defined for getting N draws from distribution of X:
   1. Use “rexp” function to get 3 random block execution times with rate as 0.2 min-1
   2. Get the maximum of the random numbers generated in step a to simulate X
   3. Repeat the above two steps N times and return X from the function
2. Call the function for N = 10000 and plot a histogram of the resulting variable
3. Superimpose the density function by calling the curve function in R to plot the density curve on the histogram
4. Compare estimated mean with the answer
   1. Define “answer.exact” as 9.167 which is the answer obtained from theoretical calculation
   2. Use the mean function in R to get the mean of the resulting variable obtained in step 2
   3. Simulate again and repeat the above two steps 4 more times
   4. Observe the results
5. Repeat step 4 with n = 1000 and n = 100000

# Answers

The answers to the questions stated in (Choudhary, Mini Project 1, 2015) are as follows:-

1. **Superimpose the density function obtained in Exercise 4.6. Try using the R function `curve’ for drawing the density. Note what you see.**

The plot is stated as follows:

The peak of the plot is around the mean and the distribution is approximately equal on both sides of the mean. We see that the curve is symmetric around the mean (~9) and attains and normal shape.

1. **Use the draws to estimate E(X). Compare your answer with the exact answer obtained in Exercise 4.6. Note what you see.**

We see that the estimated value is considerably close to the exact answer obtained. This is because N is considerably large and as per the previous answer, the plot of the distribution is centered on the exact answer.

1. **Repeat the process of obtaining an estimate of E(X) five times. Compare each estimate with the exact value. Note what you see.**

We see that the estimated values are considerably close to the exact answer obtained and are within the range (-0.04, 0.08).

1. **Comment on how your results would change if you use 1,000 Monte Carlo replications instead of 10,000. What if you use 100,000 replications? Justify your answers.**

For N = 1000, the values are close to exact answer within the range (-0.08, 0.37). The range of estimated answer is less accurate than N = 10000, because as the sample size decreases, the error in estimation increases.

For N = 100000, the values are close to the exact answer within the range (-0.005, 0.007). The range of estimations are more accurate than N = 10000, because as the sample size increases, the error is estimation decreases and the estimated mean becomes the distribution mean.

# R Code

# Get 10000 draws

f = function(n, X = c()){

for(index in 1:n) {

# Simulate the block execution times

# Simulate X by taking max of block execution times

X[index] = max(rexp(3, 0.2))

}

# Return X

X

}

# Plot the histogram

par(mfrow = c(1, 1))

X = f(10000)

hist(X, freq = FALSE, probability = TRUE, main="Draws of X", col = gray(.95))

# Superimpose the density function

curve(dnorm(x, mean = mean(X), sd = sd(X)), col = 2, add = TRUE)

# Estimate E(X) and compare with the exact answer 5 times

answer.exact = 9.167

for (index in 1:5){

# Print the comparison

cat("n = 10000,", index, "- Differnce in E(X) and exact answer:", mean(X) - answer.exact, "\n")

# Simulate again

X = f(10000)

}

# Repeat with n = 1000 and n = 100000

for (n in c(1000, 100000)){

for (index in 1:5){

# Simulate

X = f(n)

# Print the comparison

cat("n =", n, ",", index, "- Differnce in E(X) and exact answer:", mean(X) - answer.exact, "\n")

}

}

# References

Baron, M. (2009). Exercise 4.6. In M. Baron, *Probability and Statistics for Computer Scientists* (p. 97). Richardson, USA: CRC Press.

Choudhary, P. (2015). *HW1 Solutions.* Richardson, USA.

Choudhary, P. (2015). *Mini Project 1.* Richardson, USA.