Statistical Inference Project Part 1

J.E. Panzik

5/27/2020

Overview

In this project I investigated the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. I used the distribution of averages of 40 exponentials for 1000 simulations.

The results of the simulated experiments show that the CLT approximates the theoretical mean, standard deviation, and variance. It also shows that the distribution of experiment means is approximately normal.

Instructions

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulated Experiments

This will simulate the Central Limit Theorem (CLT) by creating 1000 simulated experiments of 40 samples.

```
#Set seed for reproducibility
set.seed(64)
#Set lambda to 0.2
lambda <- 0.2
#40 samples
n <- 40
#1000 simulations
nSim <- 1000
#Simulate 1000 experiments of 40 random samples from the exponential distribution
sim <- matrix(rexp(n*nSim,lambda),nSim,n)
```

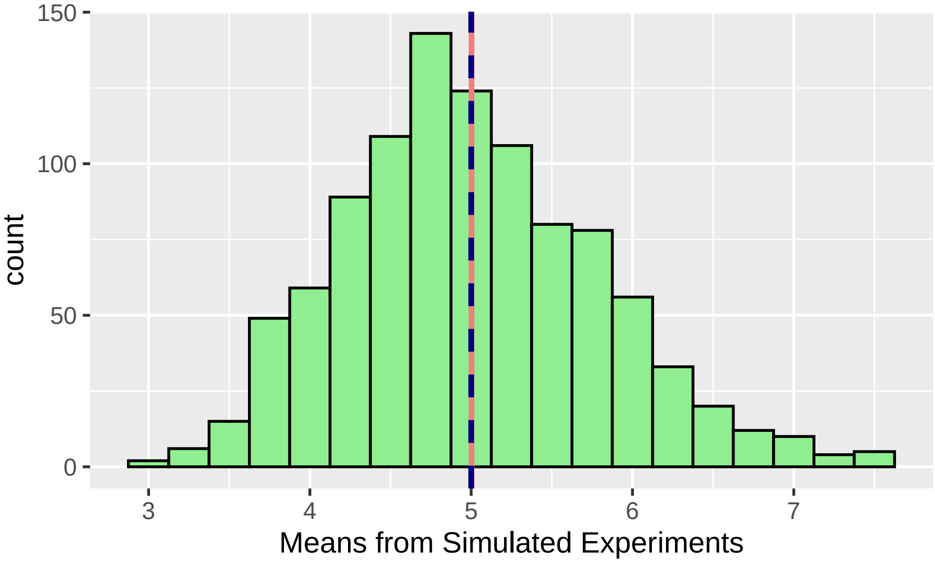
Sample Mean versus Theoretical Mean

I take the simulated experiments and calculate a mean from the 40 samples for each experiment. The mean of the experiment means should approximate the population mean that can be determined theoretically by 1/lambda.

```r
#Calculate the mean of from 40 samples for each simulation
meanSim <- apply(sim,1,mean)
#Calculate the overall mean from each experiment mean
meanTot <- mean(meanSim)
#Calculate the population mean from theory
meanTheory <- 1/lambda
```

The mean calculated using the CLT is **5.0023508**, compared to the expected mean from theory of **5**.

```r
#Histogram plot
library(ggplot2)
ggplot() + aes(meanSim) +
      geom_histogram(color="black", fill="lightgreen", binwidth=0.25) +
      geom_vline(aes(xintercept=meanTot), color="lightcoral", size=1) +
      geom_vline(aes(xintercept=meanTheory), color="navy", linetype="dashed",size=1) +
      xlab("Means from Simulated Experiments")
```



The plot above shows the histogram distribution of each experiment mean. The CLT mean of **5.0023508** is shown in the light red solid line, and the expected mean from theory of **5** is shown in the dark blue dashed line. Notice that the computed mean from the CLT is very close to the expected mean from theory.

Sample Variance versus Theoretical Variance

I take the variance and standard deviation of the distribution of experiments means.

```r
#Standard deviation of experiment means
sdSim <- sd(meanSim)
#Standard deviation from theory
sdTheory <- (1/lambda)/sqrt(n)
#Variance of experiment means
```

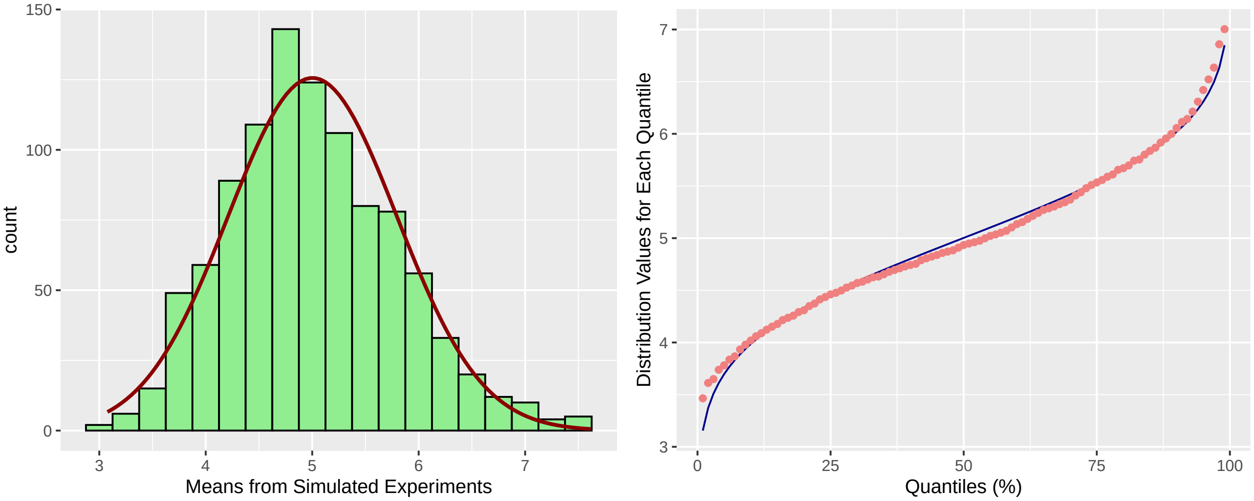
```
varSim <- var(meanSim)
#Variance from theory
varTheory <- ((1/lambda)*(1/sqrt(n)))^2
```

The standard deviation calculated from simulation is **0.7938611** and the expected standard deviation from theory is **0.7905694**. The variance calculated from simulation is **0.6302155** and the expected variance from theory is **0.625**.

Distribution

This will compare the created distribution of simulation means to a normal distribution.

```
#Plots histogram with normal distrobution line
p1 <- ggplot() + aes(meanSim) +
      geom_histogram(color="black", fill="lightgreen", binwidth=0.25) +
      stat_function(fun = function(x)
            dnorm(x, mean = meanTot, sd = sdSim) * length(meanSim) * 0.25,
            color = "darkred", size = 1) +
      xlab("Means from Simulated Experiments")
# Compare simulation quantiles to normal distribution quantiles
x <- seq(0.01,0.99,0.01)
quantSim <- quantile(meanSim, probs=x)
quantNorm <-qnorm(p=x, mean=meanTot, sd=sdSim)
p2 <- ggplot() +
      geom_line(aes(x=x*100,y=quantNorm), color="darkblue") +
      geom_point(aes(x=x*100,y=quantSim), color="lightcoral") + xlab("Quantiles (%)") +
      ylab("Distribution Values for Each Quantile")
#Display both plots
library(gridExtra)
grid.arrange(p1,p2, nrow=1)
```



The red line on the left plot shows the normal distribution for the dataset based on the computed mean and standard deviation from the simulations. A clearer comparison between the two distributions would be to compare the quantiles of the simulation to the expected quantiles of a normal distribution. The right plot shows the values of 1% quantile intervals for the distribution of experiment means (red circles) and a normal distribution (dark blueline). Both plots show that the distribution of experiment means roughly approximates a normal distribution.