Coursera Statistical Inference Course

Final Project - Part 1

Darrell Gerber

4/12/2021

Part 1: Simulation Exercise

1.1 Introduction

The first part of the project is a comparison of the exponential distribution and the Central Limit Theorem. Run a series of simulations using R to randomly draw observations with having an exponential distribution. Then, calculate the mean of the observations in each simulation and compare:

- the distribution of means to a normal distribution,
- the average to the population mean, and
- the variance to the theoretical variance.

The exponential distribution is a non-normal distribution. However, by the Central Limit Theorem, the distribution of a statistic of a large number of simulations of the exponential distribution will approach a normal distribution. The formula for the exponential distribution is:

$$f(x) = \lambda e^{-\lambda x}$$

$$\mu_{f(x)} = \sigma_{f(x)} = \frac{1}{\lambda}$$

1.2 Simulations

Simulation parameters:
Number of Observations (n) = 40
lambda = 0.2
Number of Simulations (S) = 1000

```
library(ggplot2)
library(flextable)

n <- 40
lambda <- 0.2
S <- 1:1000
set.seed(1307)
```

Run the simulations and keep the mean of each simulation run. Calculate the average and variance of the distribution of simulated statistics.

```
simulations <- NULL
for(i in S) simulations <- c(simulations, mean(rexp(n, lambda)))
simulations <- data.frame(SimMean = simulations)

mean.sims <- mean(simulations$SimMean)
var.sims <- var(simulations$SimMean)
stats <- data.frame(Statistic=c("Mean", "Variance"))
stats$Simulation <- round(c(mean.sims, var.sims),3)
```

1.3 Compare to theoretical values

Calculate the theoretical mean and variance for the simulation statistics. The Central Value Theorem says that the theoretical mean will be the simulated statistic (the mean) and the theoretical variance is the population variance divided by number of observations.

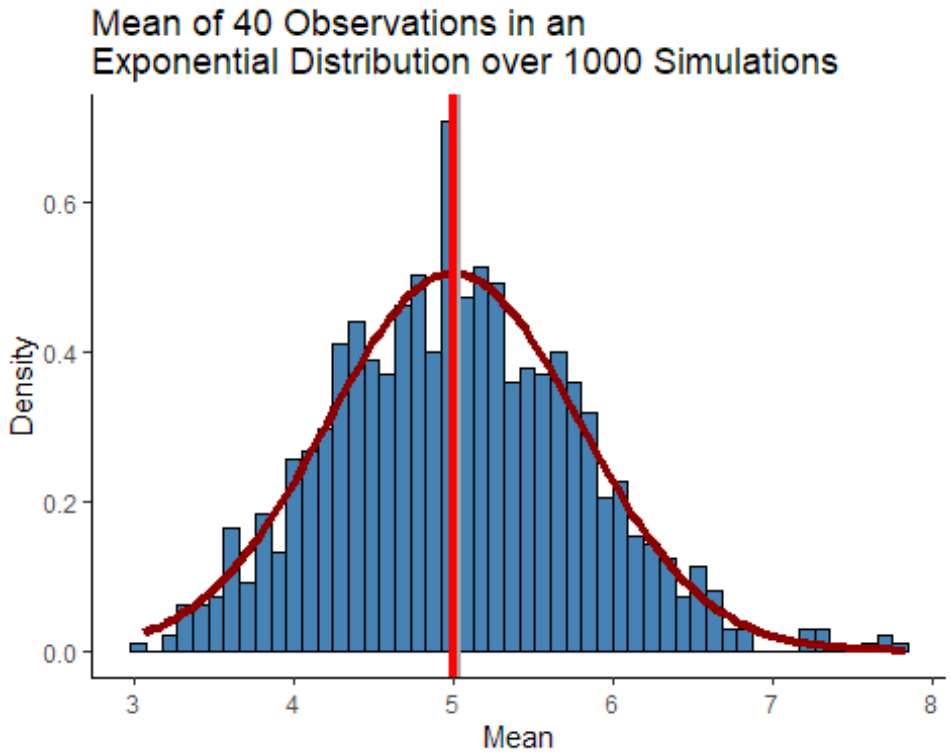
```
mean.theory <- 1/lambda
var.theory <- 1/lambda^2/n
stats$Theoretical <- round(c(mean.theory, var.theory),3)
```

The simulated mean and variance are close to the theoretical values of each:

Statistic	Simulation	Theoretical
Mean	5.027	5.000
Variance	0.611	0.625

1.4 Compare the distribution to the normal distribution

```
SimsHist <- ggplot(simulations, aes(x=SimMean)) + theme_classic() +
  geom_histogram(aes(y=..density..), fill = "steelblue",
                 color="black", bins = 50) +
  geom_function(fun=dnorm, args=list( mean=1/lambda,
                                      sd=((1/lambda))/sqrt(n)),
                lwd=1.5, color = "darkred") +
  geom_vline(xintercept=mean.sims, lwd=1.5, color="darkgray") +
  geom_vline(xintercept=mean.theory, lwd=1.5, color="red") +
  labs(title="Mean of 40 Observations in an \nExponential Distribution over 1000 Simulations",
x="Mean", y="Density")
SimsHist
```



The graph shows the distribution of the averages of 1000 simulations (steelblue) in comparison to the normal distribution of the theoretical mean and variance (dark red). The graph also shows the mean of the simulated statistic (gray) and the theoretical value (red).

After 1000 simulations, the calculated mean of 40 observations in an exponential distribution is nearly normally distributed.