Statistical Inference: Peer Graded Assigment - Part 1

Wayne Heller

April 22, 2017

Overview

Part 1 Central Limit Theorem Simulation: This part of the analysis compares the sample mean and variance for a large number of means of a simulated exponential distribution. The expected result is that the distribution of means will be Normal with a mean approximately equal to the theoretical mean of the exponential distribution and a variance approximately equal to the theoretical variance of the exponential distribution divided by the number of samples in each group.

Simulations

Create an Exponential distribution with 1,000 simulated observations

```
seed = 12345
nosim <- 1000  # number of simulated values
lambda <- 0.2  # lambda of 0.2 was provided
x.dist <- rexp(nosim, lambda)  # simulated exponential distribution
dat <- data.frame(x = x.dist)  # create a dataframe from the values for
plotting
x.mean = mean(dat$x)  # mean of the simulated distribution
x.var = var(dat$x)  # variance of the simulated distribution
```

Next, create a distribution of Means from 1,000 simulations of 40 observations each

```
nsize <- 40  # number of observations in 1 simulation
x.dist <- rexp(nosim * nsize, lambda)  # simulated exponential distribution
x.avg = apply(matrix(x.dist, nosim), 1, mean)  # calculates the mean of each
simulation
dat.avg <- data.frame(x = x.avg)  # putting in a data.frame for plotting
x.avg.mean = mean(dat.avg$x)  # simulated mean
x.avg.var = var(dat.avg$x)  # simulated variance
x.avg.mean.theoretical = 1/lambda  # theoretical mean
x.avg.var.theoretical = (1/lambda^2)/nsize  # theoretical variance
```

Simulated Exponential Distribution With Lambda = 0.2 and 1,000 Observations

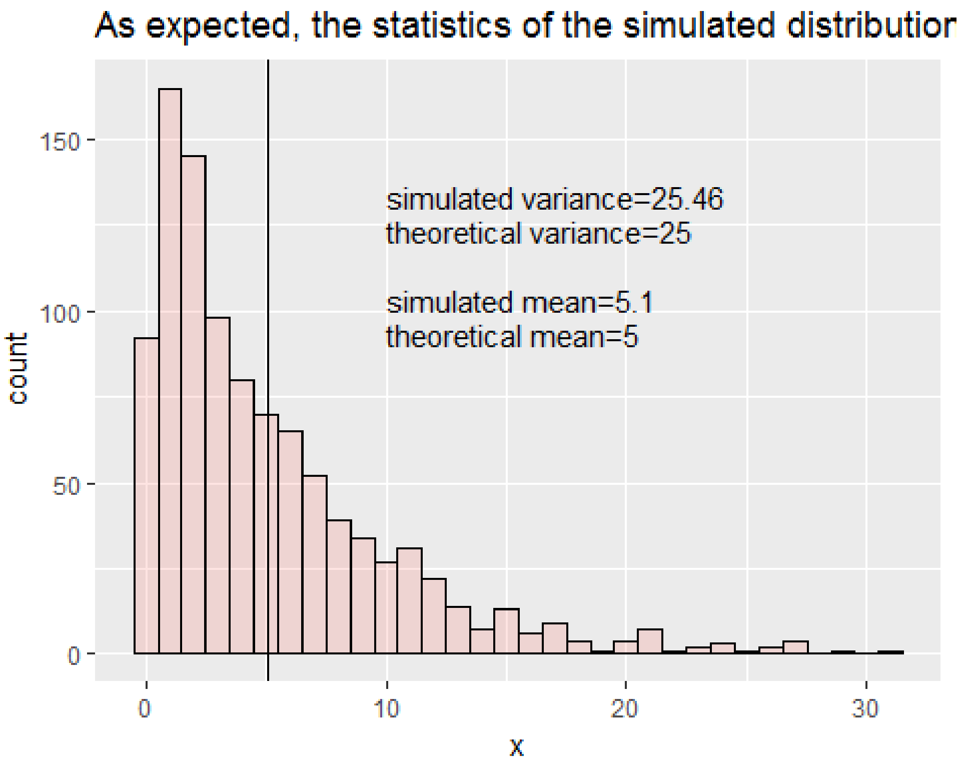

Figure 1: Exponential Distribution

Distribution of Sample Means

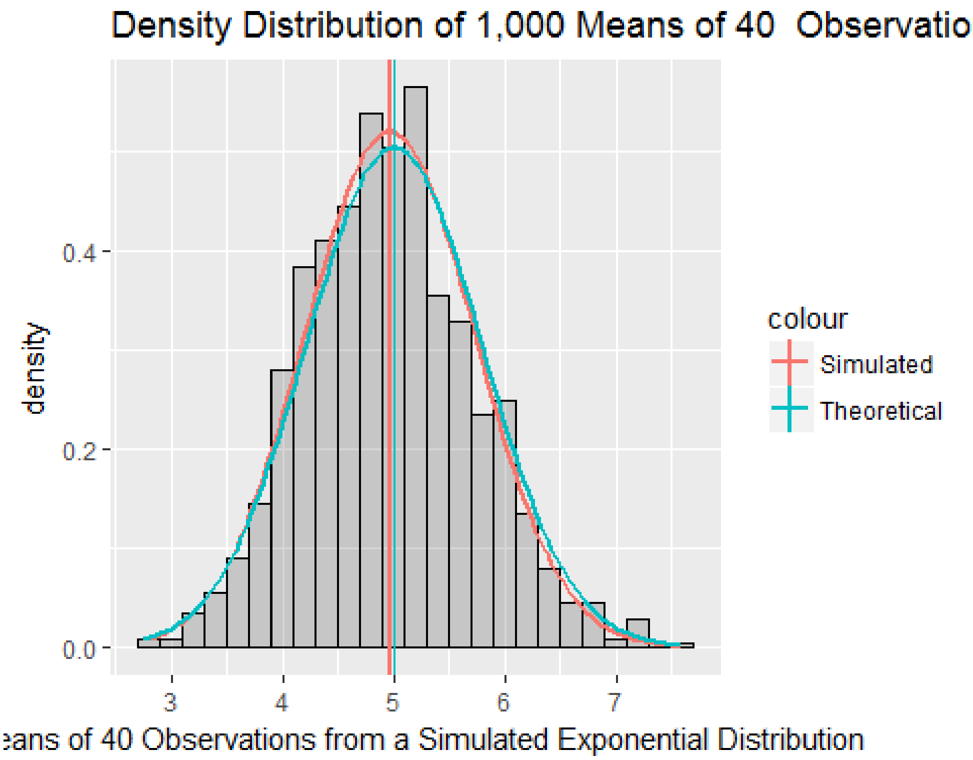

Figure 2: Distribution of Means

Comparison of Simulated and Theoretical Statistics

Statistic	Theoretical Value	Simulated Value
Mean	5.000	4.9638222
Variance	0.625	0.5871752

Summary

Comparing Figure 1 to Figure 2 demonstrates how the distribution of means becomes Gaussian for a large number of simulations. In addition, the distribution of means from the simulation (the green curve and line in Figure 2) is a close approximation to the theoretical normal distribution with mean = 1/lambda and variance = (1/lambda^2) / n, where n is 40 in this simulation (the red curve and line in Figure 2).

APPENDIX

Code for Figure 1

```
createFig1

## function() {
##
```

```
##          g <- ggplot(dat, aes(x = x, fill="red"))
##          g <- g + geom_histogram(alpha = .20, binwidth=1, colour = "black")
##          g <- g + geom_vline(xintercept=x.mean)
##          g <- g + annotate(geom = "text", x=10, y=100,
label=paste0("simulated mean=", as.character(round(x.mean,digits=2))),
size=4, angle=0, vjust=0, hjust=0)
##
##          g <- g + annotate(geom = "text", x=10, y=90,
label=paste0("theoretical mean=", as.character(round(1/lambda,digits=2))),
size=4, angle=0, vjust=0, hjust=0)
##
##          g <- g + annotate(geom = "text", x=10, y=130,
label=paste0("simulated variance=", as.character(round(x.var, digits=2))),
size=4, angle=0, vjust=0, hjust=0)
##
##          g <- g + annotate(geom = "text", x=10, y=120,
label=paste0("theoretical variance=", as.character(round(1/lambda^2,
digits=2))), size=4, angle=0, vjust=0, hjust=0)
##          g <- g + theme(legend.position="none")
##          g <- g + labs(title = "As expected, the statistics of the
simulated distribution are close to the theoretical ")
##          print(g)
## }
```

Code for Figure 2

```
createFig2

## function() {
##          g <- ggplot(dat.avg, aes(x = x))
##
##          g <- g + geom_histogram(alpha = .25, binwidth=.2, color = "black",
aes(y=..density..))
##
##          g <- g + geom_vline(aes(xintercept=x.avg.mean, color="Simulated"),
size = 1.0)
##
##          g <- g + stat_function(fun=dnorm,args=list(mean=x.avg.mean,
sd=sqrt(x.avg.var)), size = 1.0, aes(color="Simulated"))
##
##          g <- g + geom_vline(aes(xintercept=x.avg.mean.theoretical,
color="Theoretical"), size=.5)
##
##          g <- g +
stat_function(fun=dnorm,args=list(mean=x.avg.mean.theoretical,
sd=sqrt(x.avg.var.theoretical)), aes(color = "Theoretical"), size = 1.0)
##
##          g <- g + labs(title = "Density Distribution of 1,000 Means of 40
Observations", x= " Means of 40 Observations from a Simulated Exponential
Distribution")
```

```
##          print(g)
## }
```