Statistical Inference - Course Project Part 1

Pieter Wessel

February 19, 2017

Overview

This report is for the final (and only) course project for the Statistical Inference course. It addresses Part 1, which covers looking at the simulated mean and variance of an exponetial distributions and contrasting them to the theorical values. I also look at the sampling distribution to see that it follows the normal distribution as predicted by the CLT.

Part 1 - Simulation Exercise

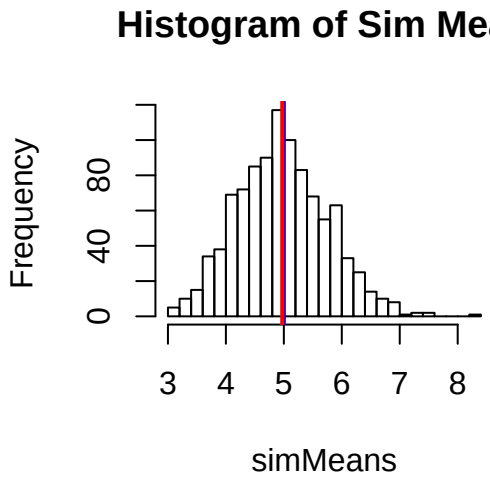
Mean

Start by generating the simulation data. Each simulation samples the exponential 40 times, and this is run 1000 times. The for each simulation, I compute the mean and the mean of the sample means. All code is included in the appendix.

The theoritical mean of the exponential distribution is calculated by

$$\mu = \frac{1}{\lambda} = \frac{1}{0.2} = 5$$

while the sample mean of the simulations is 4.971972. A histogram is plotted below with the true mean in blue and the sample mean of the simulations in red. Given n = inf the sample mean would converge to the true mean of the population.



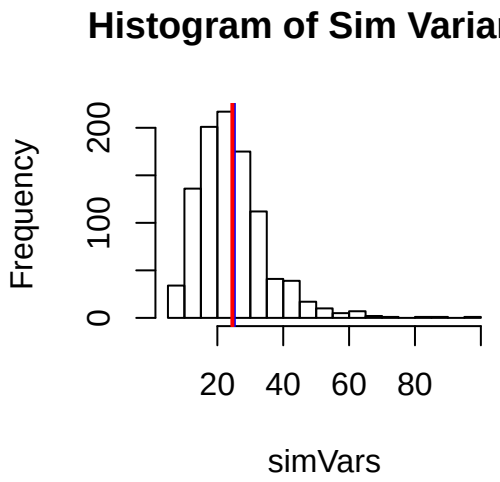
Variance

This was computed similiar to the means, only now I calculate the variance for each simulation and then calculate the mean of the variances.

The theoritical variance of the exponential distribution is calculated by

$$Var(X) = \frac{1}{\lambda^2} = \frac{1}{0.2^2} = 25$$

while the mean of the sample vairances of the simulations is 24.5680618, which is in line with what I would expect. A histogram is plotted blot with the true variance in blue and the sample variance of the simulations in red.

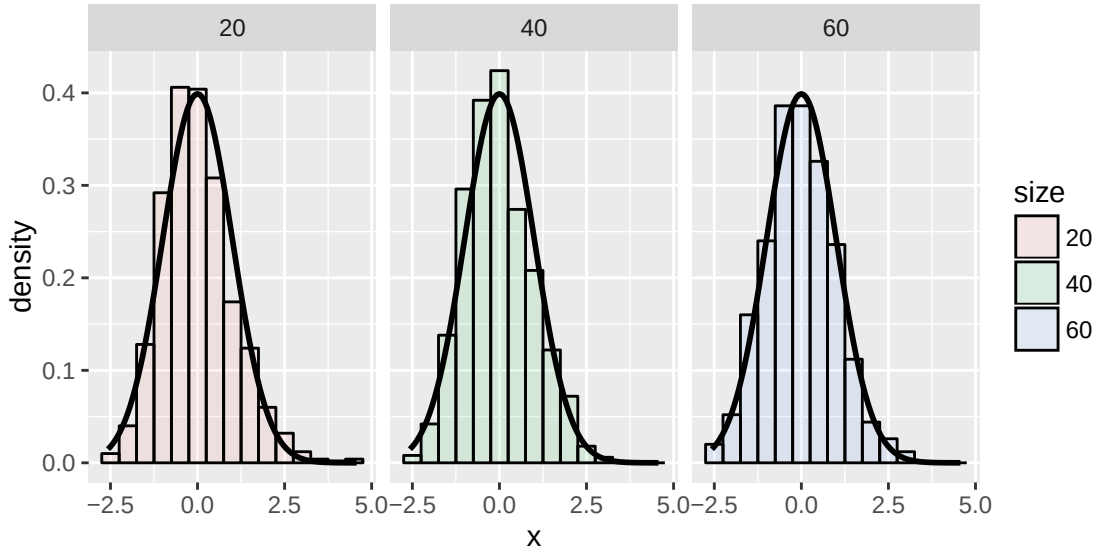


Distribution

Knowing that the $\mu = 5$ and $Var(X) = 25$, then $SE(X) = 5/sqrtn$. So I will repeat the simulations using n = 20,40,60 aplying the CLT

$$fracX_n - \mu SE(X)$$

The code below runs three experiements 1000 times for different simulation sizes, namely 20, 40 and 60 sampled values per simulation. It the applys the CLT formula to show that the gaussian sampling distribution is that of the standard normal distribution.



We see that as n increases the sampling distribution appears quite like the normal distribution as gaurenteed by the Central Limit Theorem. Given a binwidth of sufficient granularity, the historgram would follow the normal distribution (curved line) perfectly.

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
set.seed(12345)
n <- 1000; sims <- 40; lambda <- 0.2
simData <- matrix(rexp(n*sims, lambda),n,sims) #Generate Data
simMeans <- apply(simData, 1, mean) #Calcluate the mean of each simulation
simMean <- mean(simMeans)
hist(simMeans, breaks = 30, main = "Histogram of Sim Mean")
abline( v = 5, col = "blue", lwd = 2)
abline( v = simMean, col = "red", lwd = 2)
simVars <- apply(simData, 1, var)
simVar <- mean(simVars)
hist(simVars, breaks = 20, main = "Histogram of Sim Variance")
abline( v = 25, col = "blue", lwd = 2)
abline( v = simVar, col = "red", lwd = 2)
set.seed(1435)
library(ggplot2)
cfunc <- function(x,no) sqrt(no)*(mean(x) - 5)/5
dat <- data.frame(
        x = c(apply(matrix(rexp(n*20, lambda),n),1,cfunc,20),
              apply(matrix(rexp(n*40, lambda),n),1,cfunc,40),
              apply(matrix(rexp(n*60, lambda),n),1,cfunc,60)
        ),
        size = factor(rep(c(20,40,60),rep(n,3))))
p3 <- ggplot(dat, aes(x=x, fill = size)) + geom_histogram(alpha = 0.1, binwidth = .5,
                colour = "black", aes(y = ..density..))
p3 + facet_grid(.~size) + stat_function(fun = dnorm, size = 1)
```