Validating the Central Limit Theorem (CLT) using simulations

BK

Monday, October 19, 2015

Objective

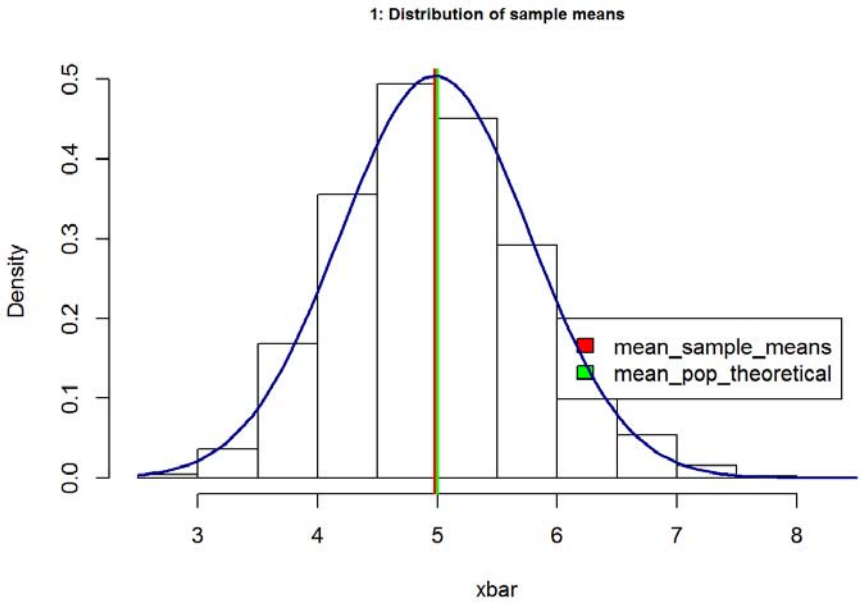
This assignment requires us to validate the following learnings from the study of iid samples and the CLT:
* The mean of sample means converges to the population mean
* The variance of sample means estimates the variance of the population
* The sample means have a standard normal distribution even when the population has random exponential distribution

Simulations

The code does the following in steps,
Step 1: Setting up a seed, sample size and number of simulations per spec given.
Step 2: Generating a population of random exponentials in the form of a “n” column and “nosim” row matrix, and calculating statistical characteristics of each row, across rows and the whole matrix (each row being one random sample set).
Step 3: Plotting charts

The mean of sample means converges to the population mean

The mean of sample means, 4.98, and the populations’ theoretical mean (1/lambda), 5 are nearly identical. The sample means follow a normal distribution. Chart 1 reiterates this.

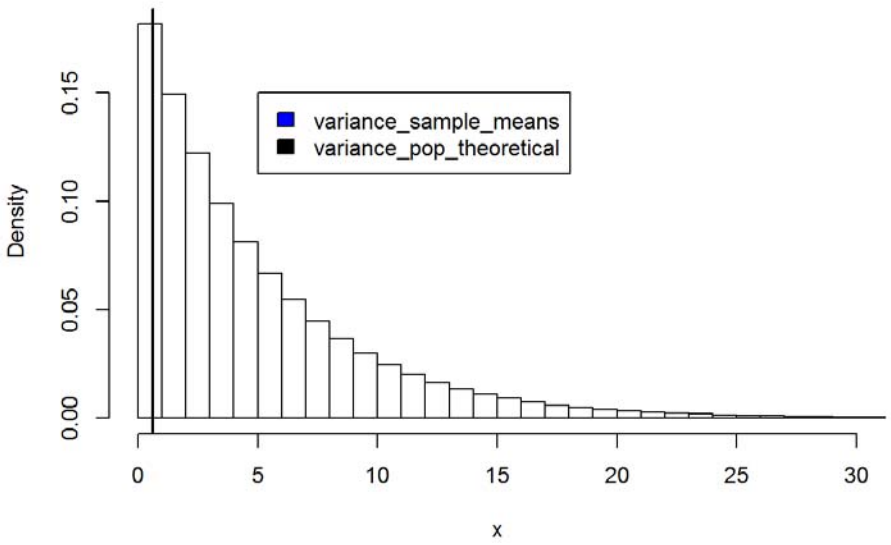


The variance of sample means estimates the variance of the population

According to the CLT, the variance of sample means = The variance of the population/ sample size. This is successfully validated. The variance of sample means is 0.63, and the variance of the population/ sample size is 0.62.

The values of variances included in the above sentence was generated using the following embedded R commands:
* The variance of sample means is: Value = 0.63, Command = r round(var_xbar,2)
* The variance of the population divided by sample size is: Value = 0.62, Command = r round((pop_var/n),2)

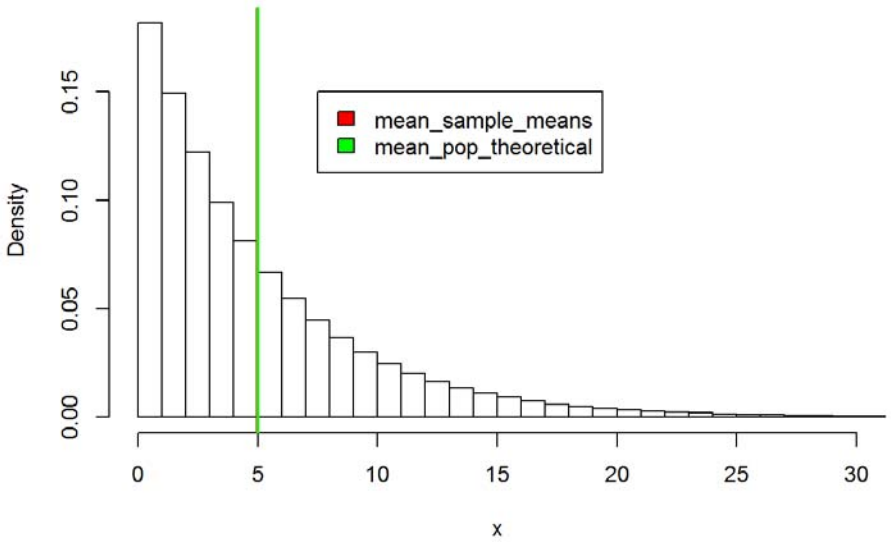
Chart 2: Exponential distribution of random data (comparing population variance to sample variance)



The sample means have a standard normal distribution even when the population has random exponential distribution

The chart below shows the distribution characteristics of the population of random exponentials, from which the samples were drawn. Clearly, this is nowhere near being normally distributed. Whereas, Chart 1 above showed that the sample means are normally distributed.

Chart 3: Exponential distribution of the population



Appendix

All the code used for this assignment is provided below. Explanatory comments are included.

```
#setting up per spec given
set.seed(1000)
lambda <- 0.2
nosim <- 10000
mu <- 1/lambda
sd <- 1/lambda
#selecting one of the 1000 samples (simulations) at random
random_sample <- sample (1:nosim,1)

#run 1 with given sample size
n <- 40

#generating n X nosim random exponentials (n = sample size, nosim = number of simulations)
#each sample is organized as one row of the matrix
x <- matrix(rexp(nosim * n, lambda), nosim)

# calculating the mean of each sample, mean of all sample means, variance of across all sample means,
# variance of each sample, mean of all sample variances, population mean and variance
xbar <- rowMeans(x)
mean_xbar <- mean(xbar)
var_xbar <- var(xbar)
xvar <- apply(x,1,var)
mean_xvar <- mean(xvar)
pop_mean <- mean(as.vector(x))
pop_var <- var(as.vector(x))

hist(xbar, cex.main=0.75, main = "1: Distribution of sample means and its close fit to the normal curve", freq=F)
abline(v = mean_xbar, col = "red", lwd = "2")
abline(v = mu, col = "green", lwd = "2")
legend(x=6, y=0.2,legend=c("mean_sample_means","mean_pop_theoretical"),fill=c("red","green"))
curve(dnorm(x, mean=mean(xbar), sd=sd(xbar)), col="darkblue", lwd=2, add=TRUE, yaxt="n")

hist(x, prob = T, breaks=50, xlim=c(0,30), cex.main=0.65, main = "Chart 2: Exponential distribution of random data (comparing p
opulation to sample variance)")
abline(v = var(xbar), col = "blue", lwd = "2")
abline(v= (1/lambda)* (1/lambda) * (1/n), col = "black", lwd = "2")
legend(x=5,y=.15,legend=c("variance_sample_means","variance_pop_theoretical"),fill=c("blue", "black"))

hist(x, prob = T, breaks=50, xlim=c(0,30), cex.main=0.65, main = "Chart 3: Exponential distribution of the population")
abline(v = mean(xbar), col = "red", lwd = "2")
abline(v= 1/lambda, col = "green", lwd = "2")
legend(x=7.5,y=.15,legend=c("mean_sample_means","mean_pop_theoretical"),fill=c("red","green"))
```