Central Limit Theorem

Alyssa Goldberg

2015-11-21

Introduction

The Central Limit Theorem (CLT)states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the *sample size* n *increases*. This means that it is possible to get an approximation of mean μ, standard deviation σ, and variance σ^2, for the whole distribution with only one observed average ($\bar{X}_i$) and without knowing the population distribution.

Test the Central Limit Theorem by simulation:

Some Initial Assumptions

The theoretical mean (μ) and variance σ^2 of exponential distribution with parameter λ are respectively $1/\lambda$ and $1/\lambda^2$,

If the CLT is true, then:

- The mean of our simulation $\bar{X}_n$ should approach μ
- The variance of our simulation, Var_n should approach $1/\lambda^2$
- The variance of our sample mean, S^2 should approach σ^2/n
- The standard deviation S of our sample mean should approach $\sqrt{\frac{1}{\lambda^2/n}}$

Let's find out if this is true by using simulation of **random exponentials** with n=40 drawn 1,000 times.

Set Constants

- The rate parameter, λ, as prescribed by the assignment, is 0.2
- The number of random exponentials to means test, n is 40.
- The number of simulations of 40 random exponents is 1000
- The total population, n * 1000 = 40,000
- Theoretical constants:
  - Theoretical means = $\frac{1}{0.2}$ = 5
  - Standard deviation = $\sqrt{\frac{1}{0.2^2/40}}$ =0.7906
  - Variance of sample mean = σ^2/n = 0.625

Create the Population, Draw Samples

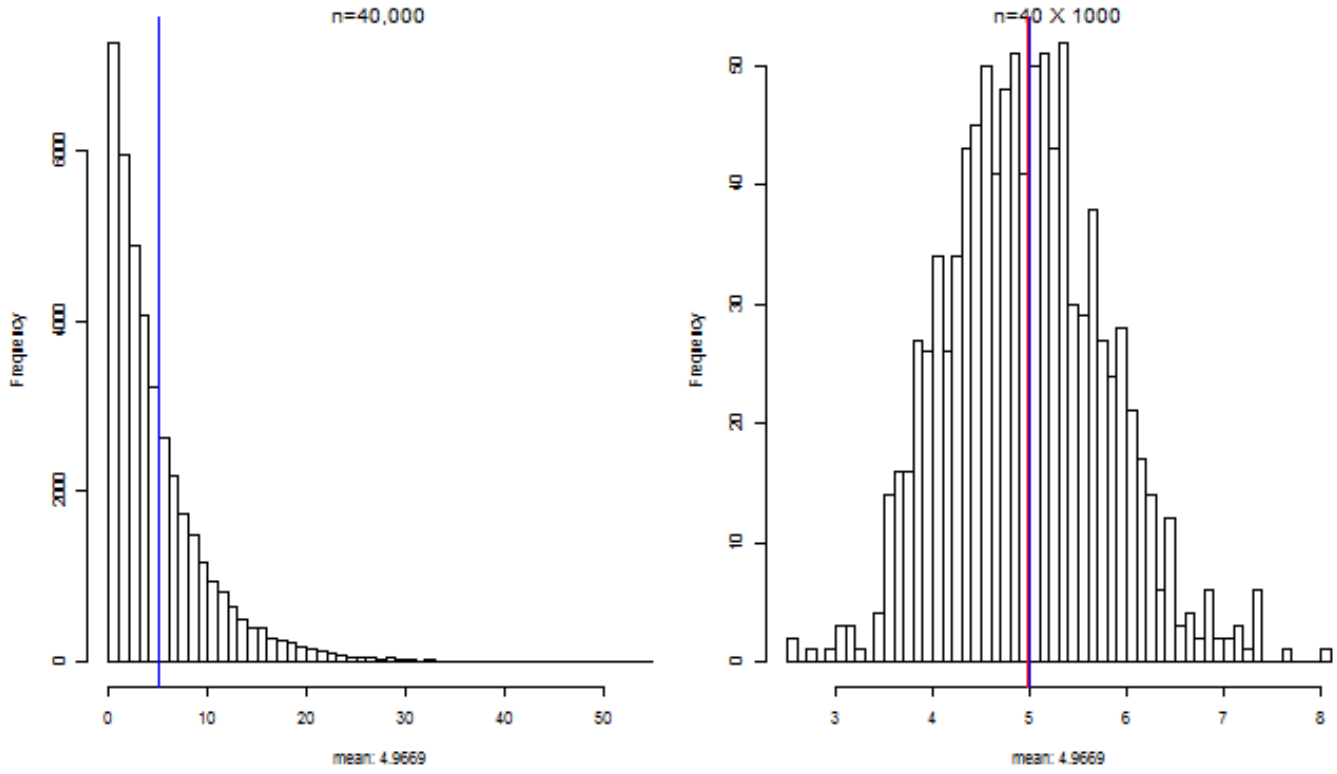
Draw 40 samples of exponentials, 1,000 times, producing a matrix with 1000 rows and 40 columns and create a vector of means of each of those 1000 rows

```
smatrix <- matrix(rexp(ex * sims, .2), nrow = sims, ncol=ex) #produce a 1000 x 40 matrix

smatrixmeans<-data.frame(value=rowMeans(smatrix)) #produce a data frame of the means of t
he 1000 rows in smatrix
```

Compare Means:

The theoretical μ for a population this size= $1/\lambda$ = 5.
The calculated $\bar{X}_n$ = 4.9669



Compare Variance of the Sample Mean:

Let R calculate the sample mean, sample standard deviation and variance of the sample mean:

```
smean<-mean(smatrixmeans$value) #sample mean
ssd<-sd(smatrixmeans$value) #standard deviation of sample means
svar<-var(smatrixmeans$value) #variance of sample means
```

Producing:

Theoretical Values	Sample Values

μ = 5	$\bar{X}_n$ = 4.9669
σ^2 = 0.625	S^2 = 0.6509
σ = 0.7906	S = 0.8068

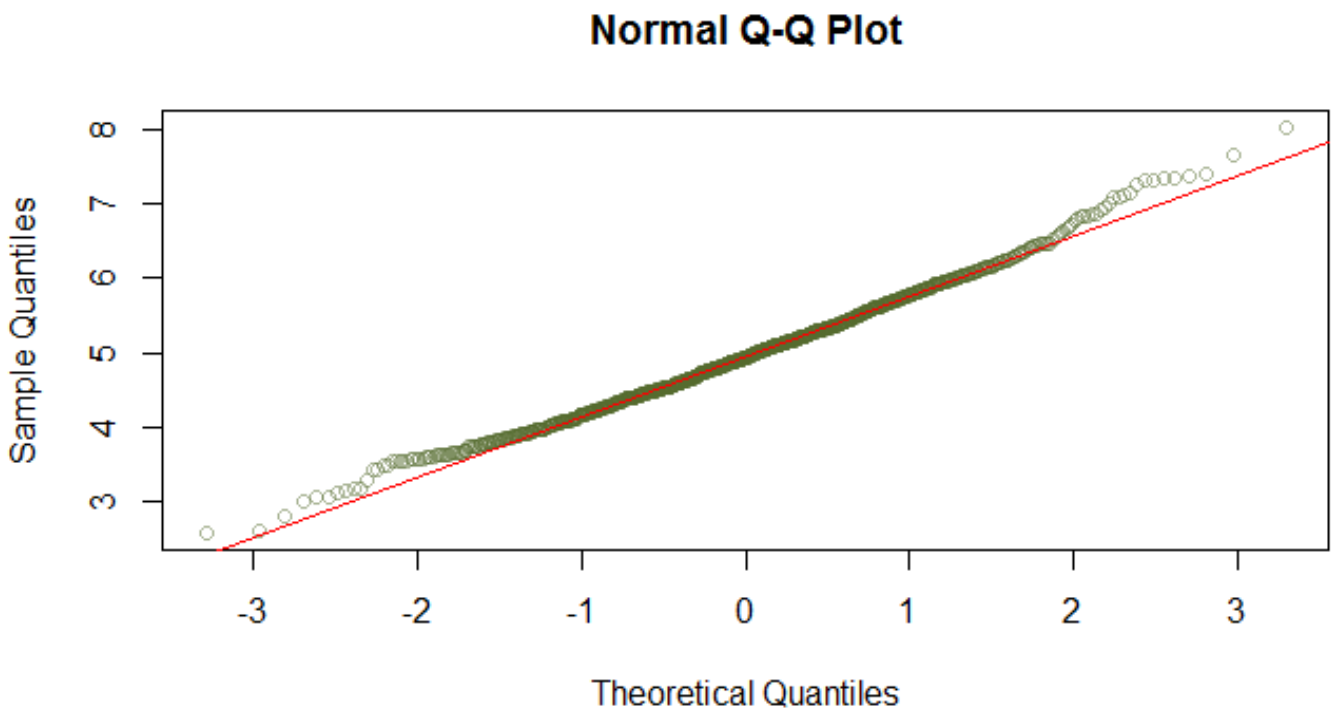
Compare Density Distribution:

As the number of means of sample means (n times the number of simulations) increases, the density distribution should more closely resemble the normal density distribution:

	25%	50%	75%
sample	4.399	4.931	5.492
normal	4.467	5.000	5.533

A Quantile-Quantile plot, displaying both the sorted theoretical normal distribution of means of a large sample (straight line) vs the sorted distribution of calculated means (plot points) shows a fairly tight fit, though it varies a bit more at the tails, yet more evidence that the Central Limit Theorem is useful for working with very large data sets.

```
qqnorm(smatrixmeans$value, col=rgb(.333, 0.42, .18, 0.5))
qqline(smatrixmeans$value, col="red", lwd=1.5)
```



We can see that the behavior of large samples approaches that of the theoretical normal for Mean, Variance and Density, the closer we get to the theoretical Mean, Variance and Density.

APPENDIX

Libraries:

```
library(knitr)
library(dplyr)
library(stats)
library(ggplot2)
```

Density Plot Theoretical vs. Sample

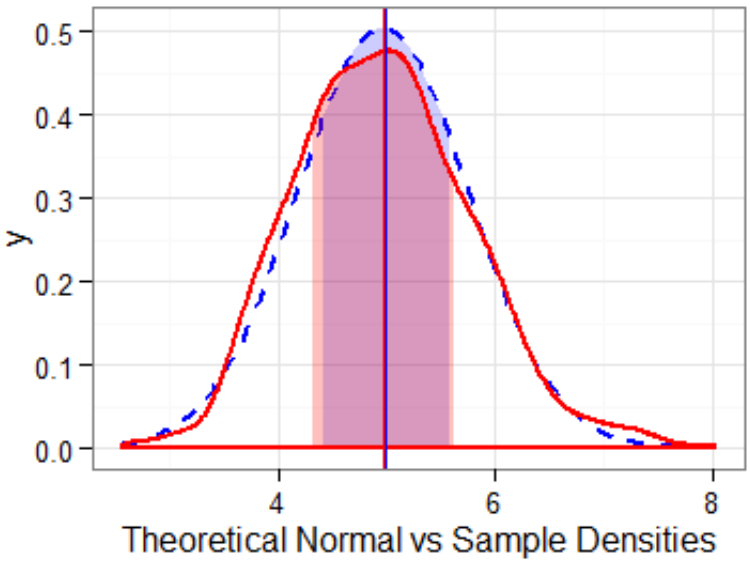
To compare densities, we can overlay a density plot for the distribution of the calculated sample means with a theoretical normal distribution density (the classic $bell$ curve) to test this.

```
# for calculated
sdx <- svar^2  #density standard deviation
x.dens <- density(smatrixmeans$value)  #create list of densities
df.dens <- data.frame(x = x.dens$x, y = x.dens$y)  #create a dataframe with densities
varplot <- df.dens[df.dens$x >= smean - svar & df.dens$x <= smean + svar, ]  #subset the
density data to the area of the variance

dnorm_limit <- function(x) {
    y <- dnorm(x, mean = mean(smatrixmeans$value), sd = sqrt((tmean^2)/ex))
    y[x < tmean - tvar | x > tmean + tvar] <- NA
    return(y)
}

# Normal vs Sample Plot
p <- ggplot(data.frame(x = c(min(smatrixmeans$value), max(smatrixmeans$value))),
    aes(x = x))

p + stat_function(fun = dnorm_limit, geom = "area", fill = "blue", alpha = 0.2) +
    stat_function(fun = dnorm, args = list(mean(smatrixmeans$value), sd = sqrt((tmea
n^2)/ex)),
        col = "blue", lwd = 1, lty = 2) + geom_density(data = smatrixmeans,
    aes(value), alpha = 0.25, col = "red", lwd = 1) + geom_area(data = varplot,
    aes(x = x, y = y), fill = "red", alpha = 0.25) + geom_vline(xintercept = tmean,
    stat = "vline", col = "blue") + geom_vline(xintercept = smean, stat = "vline",
    col = "red") + xlab("Theoretical Normal vs Sample Densities") + theme_bw()
```


0.5
0.4
0.3
0.2
0.1
0.0
y
4
6
8
Theoretical Normal vs Sample Densities