An Exploration of the Central Limit Theorem As Applied To The Exponential Distribution

Tim Chopoorian

December 27, 2015

Overview

This project investigates the exponential distribution and compares the characteristics of its distribution of sample means with those of its underlying population in an attempt to confirm what the Central Limit Theorem (CLT) tells us about the sampling distribution of the mean of IID random variables. Specifically, we will compare the mean and variation of the sample means with their corresponding population parameters, as well as apply some tests to confirm that the distribution of sample means is aproximately normal.

Simulations

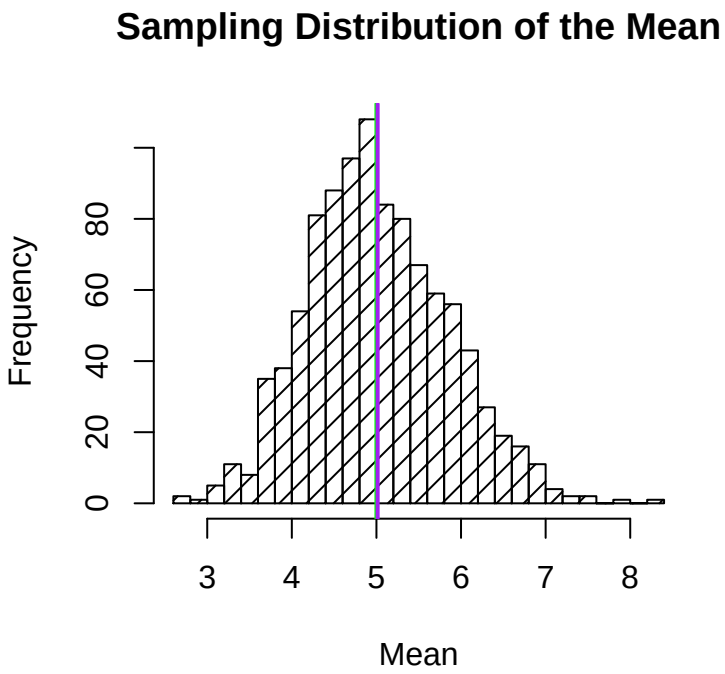
We've been given the theoretical mean and standard deviation of the exponential distribution, both of which are 1/lambda. For all of the simulations that follow, we will be using a value of 0.2 for lambda.

First, we'll require some libraries, set up some variables, and run our simulation.

```
library(ggplot2) # require some libraries
library(pander)
set.seed(23) # set a seed so our results are reproducible
lambda <- 0.2 # set our lambda value
mu <- 1/lambda # calculate the population mean, mu
sigma <- 1/lambda # calculate the population standard deviation, sigma
n <- 40 # sample size
sim_count <- 1000 # number of simulations
simulations <- t(replicate(sim_count, rexp(n, lambda))) # create a matrix of simulations
# calculate mean and variance for each sample
sample_stats <- data.frame(Mean= rowMeans(simulations),
                Variance= apply(simulations,1, var))
```

Appendix 1 shows the exponential distribution. By its shape, we see that it is clearly not a normal distribution. The CLT tells us that the sample means of IID random variables will be normally distributed, even if the underlying population is not normally distributed. Let's see if that's the case.

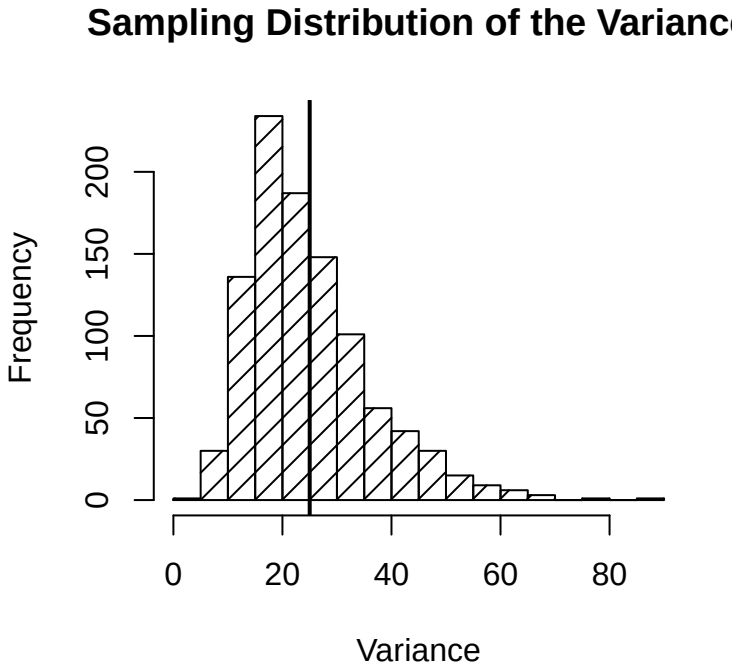
Sample Mean versus Theoretical Mean



From the plot (see R code in Appendix 2), we can see that the distribution of the means of random samples taken from the exponential distribution does seem to be normal in shape. But let's look at the numbers. The theoretical mean of the population is **5** (green line in graph). The average of the distibution of sample means is **5.0142497** (purple line). These values, nearly identical, confirm what we expected per the CLT: that the sample means are distributed around the population mean, making the mean of sample means a reliable approximation of the population mean for sufficiently large sample size.

Sample Variance versus Theoretical Variance

The CLT applies to the sampling distribution of any statistic, not just the mean. So let's try a similar analysis for sample variance: (code for plot in Appendix 3)

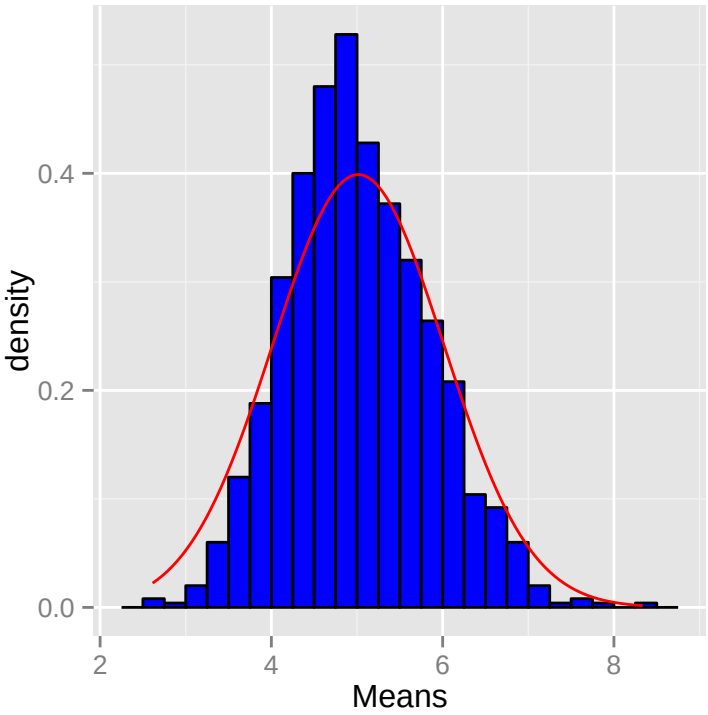


From its shape, the distribution appears to be approximately Gaussian. The theoretical standard deviation of the exponential distribution is **5**, so the theoretical variance is its square, **25** (black vertical line in graph). According to the CLT, the mean of the sample variances should estimate the population variance. Taking the mean of variances from each of our 1000 simulations of size n, we get **24.9289518**. So, as expected, the sample variance is normally distributed around the population variance.

We can also consider the theoretical variance of the sample mean. This is the theoretical variance of the population, **25**, divided by n (40), which gives us **0.625**. The actual variation of our sample means should approximate this, which we find to be the case as follows: **var(sample_stats$Mean)= 0.6862199**.

Distribution

Finally, we'll see if, as predicted by the CLT, the sampling distribution of the mean is a normal distribution. (See code for plot in Appendix 4)



We've superimposed a normal curve on our sampling distribution of the mean. The shape of the sampling distribution tells us that it is most likely a normal distribution. Also, the mean of the distribution, **5.0142497**, and the median, **4.9350236**, are approximately the same, another indication of normality. As another check, we can calculate the percentages of the distribution that lie between one, two, and three standard deviations, and compare them to those expected in a normal distribution. (See R code in Appendix 5.)

	normal	sample
% within 1 std devs	0.68	0.5925
% within 2 std devs	0.95	0.9024
% within 3 std devs	0.99	0.9871

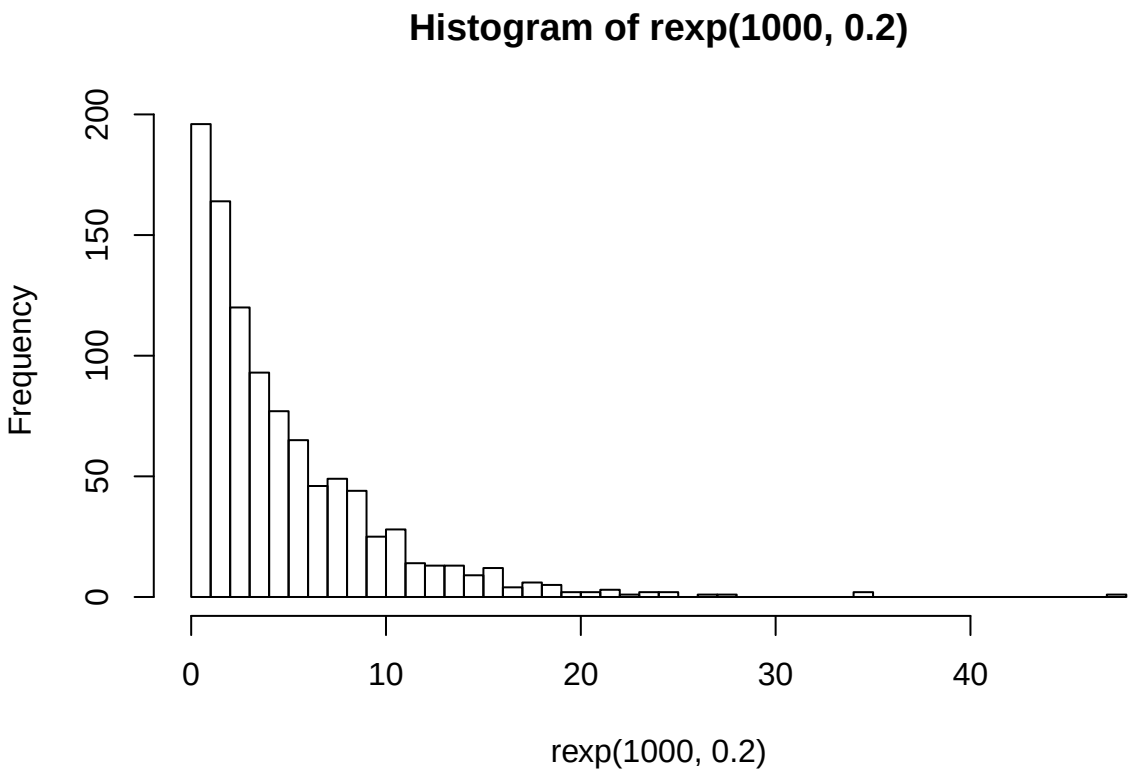
While not exact, they're close enough to be considered approximately normal. Finally, in Appendix 6, the linear shape of the Q-Q plot of our sample versus a normal distribution shows that the quantiles of our sample approximate those of a normal distibrution.

(See Appendix 7 for the url where the RMD file used to compile this report can be found.)

Appendices

Appendix 1: The distribution of 1000 random exponentials using rexp(n, lambda) in R.

```
hist(rexp(1000,0.2), breaks=40)
```



Appendix 2: R code for the plot of the sampling distribution of the mean of the exponential distribution

```
hist(sample_stats$Mean, breaks=40, main="Sampling Distribution of the Mean", density=10, xlab="Mean")
abline(v= mu, col="green", lwd=2)
abline(v= mean(sample_stats$Mean), col="purple", lwd=2)
```

Appendix 3: R code for the plot of the sampling distribution of the variation of the exponential distribution

```
hist(sample_stats$Variance, breaks=20, main="Sampling Distribution of the Variance", density=10, xlab="Varianc
abline(v= sigma^2, col="black", lwd=2)
```

Appendix 4: R code for plot of sampling distribution of mean with normal curve overlayed

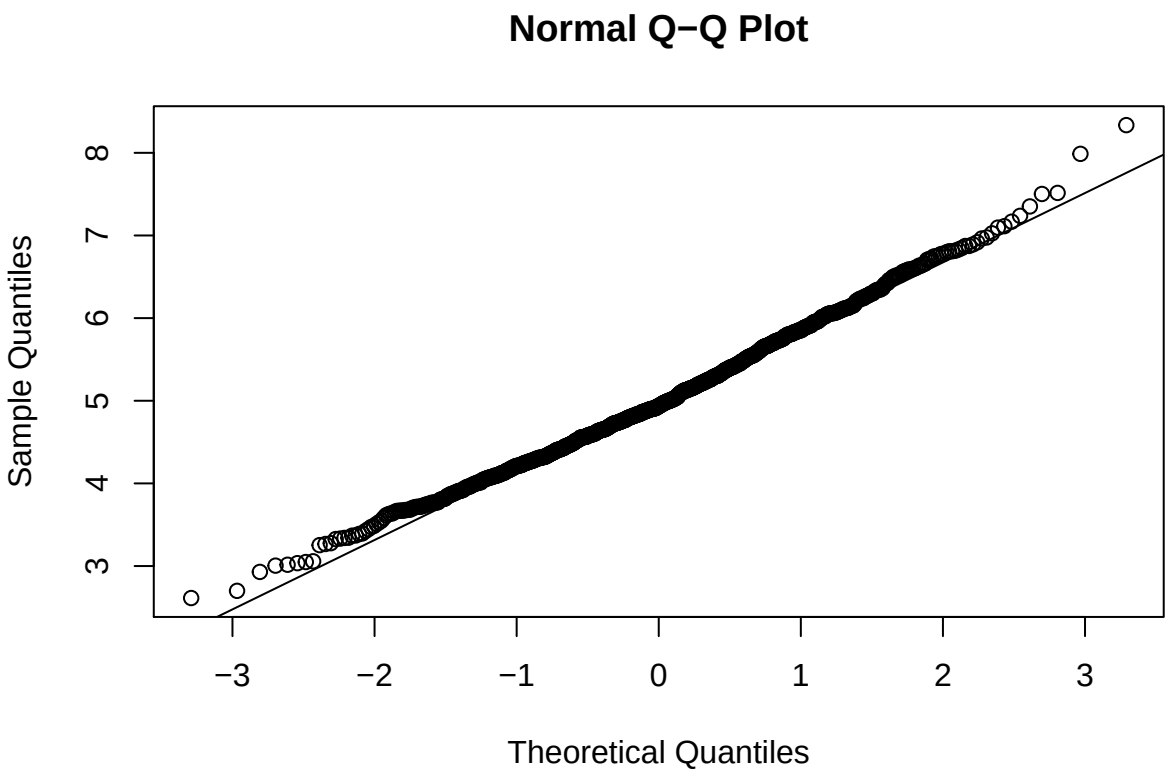
```
x <- sample_stats$Mean
ggplot(data = sample_stats, aes(x = Mean)) + xlab("Means") +
    geom_histogram(binwidth=0.25, fill="blue", color="black", aes(y=..density..)) +
    stat_function(fun=dnorm, color = "red", arg=list(mean=mean(x)) )
```

Appendix 5: R code for table of percentages of distribution between 1, 2, and 3 standard deviations

```
display <- data.frame(normal=c(.68, .95, .99), sample=c(pnorm(sd(x)) - pnorm(-sd(x)),
    pnorm(2*sd(x)) - pnorm(-2*sd(x)), pnorm(3*sd(x)) - pnorm(-3*sd(x)) ),
    row.names=c("% within 1 std devs", "% within 2 std devs", "% within 3 std devs") )
pander(display)
```

Appendix 6: Q-Q Plot of Sampling Distribution

```
qqnorm(sample_stats$Mean)
qqline(sample_stats$Mean)
```



Appendix 7: Complete R Markdown File

The complete R Markdown file used to compile this pdf report can be found here...

https://github.com/tchopoorian/Statistical-Inference