Investigating the Exponential Distribution and the Central Limit Theorem

Jason Giacomo

November 20, 2015

Synopsis

This report will investigate the exponential distribution using R. Simulated random samplings from the exponential distribution will be taken. The distribution of the means from these samplings and the sample mean and variance will be compared to the population values within the context of the central limit theorem.

Simulations

Simulated random samples will be taken from the exponential distribution using the R function `rexp()`. The simulation will utilize a function to evaluate 1000 means from 40 samples. The rate of the exponential distribution will be fixed at 0.2.

```
expSim <- function(nSamples, nMeans){
    expMean <- NULL
    for(i in 1:nMeans) expMean <- c(expMean, mean(rexp(nSamples, 0.2)))
    expMean <- data.frame("means" = expMean)  # for ease of use with ggplot
    return(expMean)
}
```

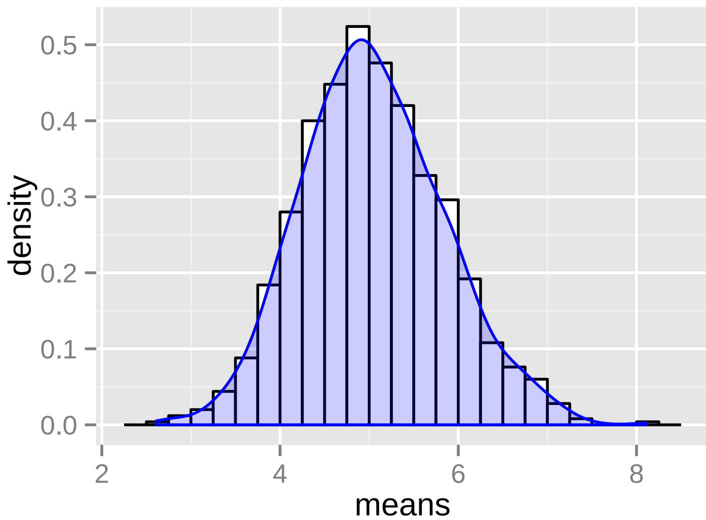
This function is then used to generate 1000 means from 40 random samples. Here we run the function and plot a histogram of the resulting means.

```
library(ggplot2)
set.seed(333)
simMeans <- expSim(40, 1000)
str(simMeans)
```

```
## 'data.frame':    1000 obs. of  1 variable:
##  $ means: num  5.15 5.07 5.63 5.62 3.49 ...
```

```
simHistogram <- ggplot(simMeans, aes(x=means)) +
    geom_histogram(aes(y=..density..),
                   binwidth = 0.25,
                   color="black", fill="white") +
    geom_density(color="blue", fill="blue", alpha=0.2)

print(simHistogram)
```



The density has been overlayed in blue to give a sense of the distribution, more on this later.

Sample Mean vs. Theoretical Mean

To compare our simulation to the population we will look at the sample mean and compare that to the population mean. For an exponential distribution the mean is related to the rate, λ. Recall that our function, `expSim()`, fixed the rate at 0.2, so the population mean is just $1/\lambda$ or 5. Now, we can calculate the mean of our 1000 means and compare it to the population mean.

```
sampleMean <- mean(simMeans$means)
popMean <- 1/0.2
```

Statistic	Value
Population mean	5
Sample mean	5.05

From these results it is clear that the estimator (sample mean) closely approximates the estimand (population mean) as we would expect from the law of large numbers.

Sample Variance vs. Theoretical Variance

Next, we will compare the sample variance to the population variance. The population variance for the exponential distribution is given as λ^{-2}. However, we need to be more careful when comparing this to the variance of our simulated means. Remember that our means are all taken from 40 random samples of the exponential distribution and thus should have less variance than from a single sampling. In fact, the variance of the sample mean is given by:

$$Var(\overline{X}) = \sigma^2/n,$$

where σ^2 is the variance of the population we are sampling from. Now, the variance of the sample mean and the variance of our simulation are calculated as follows:

```r
Var_X = (1 / 0.2)^2 / 40
S2 = sd(simMeans$means)^2
```

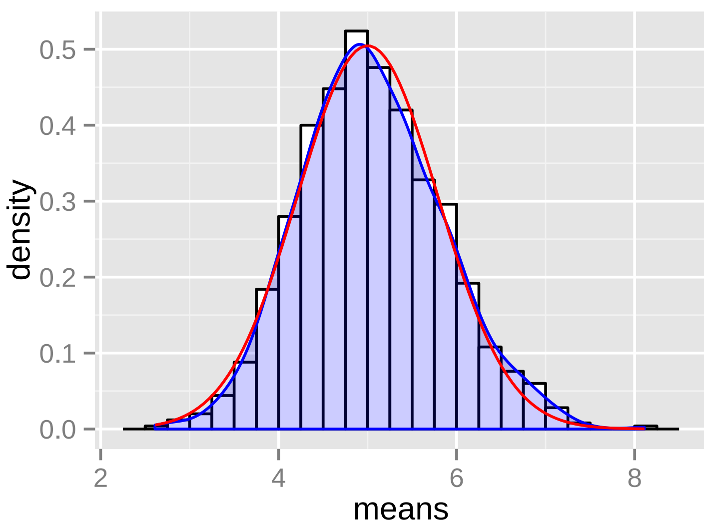
Statistic	Value
Theoretical variance	0.62
Sample variance	0.64

These results again confirm that our simulation is a good estimate of the theoretical variance.

Distribution of Means

Recall in the first plot that I inserted the density to get a feel for what the distribution of the sample means looked like. The central limit theorem states that a sampling of means from any distribution will be distributed normally. This means that our sampling of means from the exponential distribution should look like a normal distribution. We can see a similarity to the bell-shaped curve of the normal distribution in the density in the first plot. Now, let's take it a step further and plot the normal distribution with our simulation means. We can plot the normal distribution by using the `dnorm()` function and supplying it with the theoretical mean and standard deviation ($\sqrt{Var(\overline{X})}$) so it will be consistent with our data.

```r
normHist <- simHistogram + stat_function(fun=dnorm, color="red",
                                         args = list(mean = popMean,
                                                     sd = sqrt(Var_X)))
print(normHist)
```



From the plot, we can easily see that the central limit theorem applies to our simulation. The density from the simulated means (blue) very nearly matches a normal distribution (red) with the same mean and variance.