Exponential Distribution Simulation

Ash Chakraborty

Wednesday, June 17, 2015

PART 1: SIMULATION - Sampling the Exponential Distribution

Synopsis

Part 1 of this project investigates the *Exponential Distribution* and compares it to the *Central Limit Theorem*. The exponential distribution is given by the probability distribution function, $P(x) = \lambda e^{-\lambda x}$. We also know that it has a mean, $\mu = 1/\lambda$, and variance, $\sigma^2 = 1/\lambda^2$. This report will conduct an appropriate number of simulations on this distribution with a sample size of 40 exponentials by generating a sampling distribution of sample means. The consequences of this sampling distribution will be evaluated for adherence to the central limit theorem.

Exponential Distribution

We are to work with a sample size, n = 40 with a rate parameter, $\lambda = 0.2$. We first generate a sample of 40 random exponentials, and look at its central tendencies:

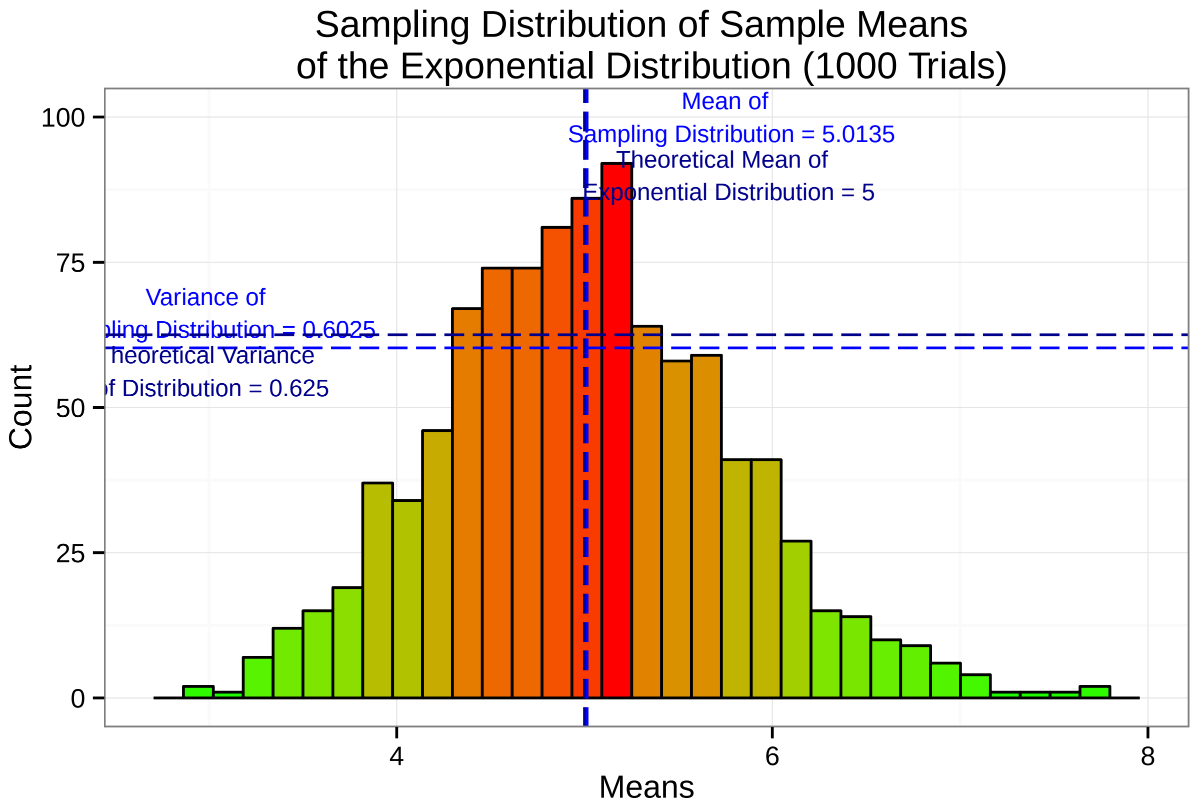
```r
library(PerformanceAnalytics)
library(ggplot2)
library(gridExtra)
set.seed(123)
sample <- rexp(40, rate = 0.2)
```

Simulating 1000 Trials

In order to understand the properties of the distribution of the mean of 40 exponentials, we conduct a 1000 simulations, and extract the mean of each sample:

```r
samp.dist <- NULL
for(i in 1:1000){samp.dist <- c(samp.dist, mean(rexp(40, rate=0.2)))}
```

Now, we plot this sampling distribution:



The distribution of a 1000 means of 40 random exponentials has begun to resemble a Gaussian distribution.

Comparing Theoretical Variance and Sample Mean

We continue to note from the earlier plot that the sampling distribution's mean and variance may be given by:

```
## [1] "Mean of Sampling Distribution: 5.0135"
```

```
## [1] "Variance of Distribution: 0.6025"
```

We noted earlier that the theoretical mean of the exponential distribution is given by $1/\lambda$ ($\lambda = 0.2$), which is equal to 5. When we compare this to sampling distribution's mean $= 5.014$, we see that it is a fairly good approximation.

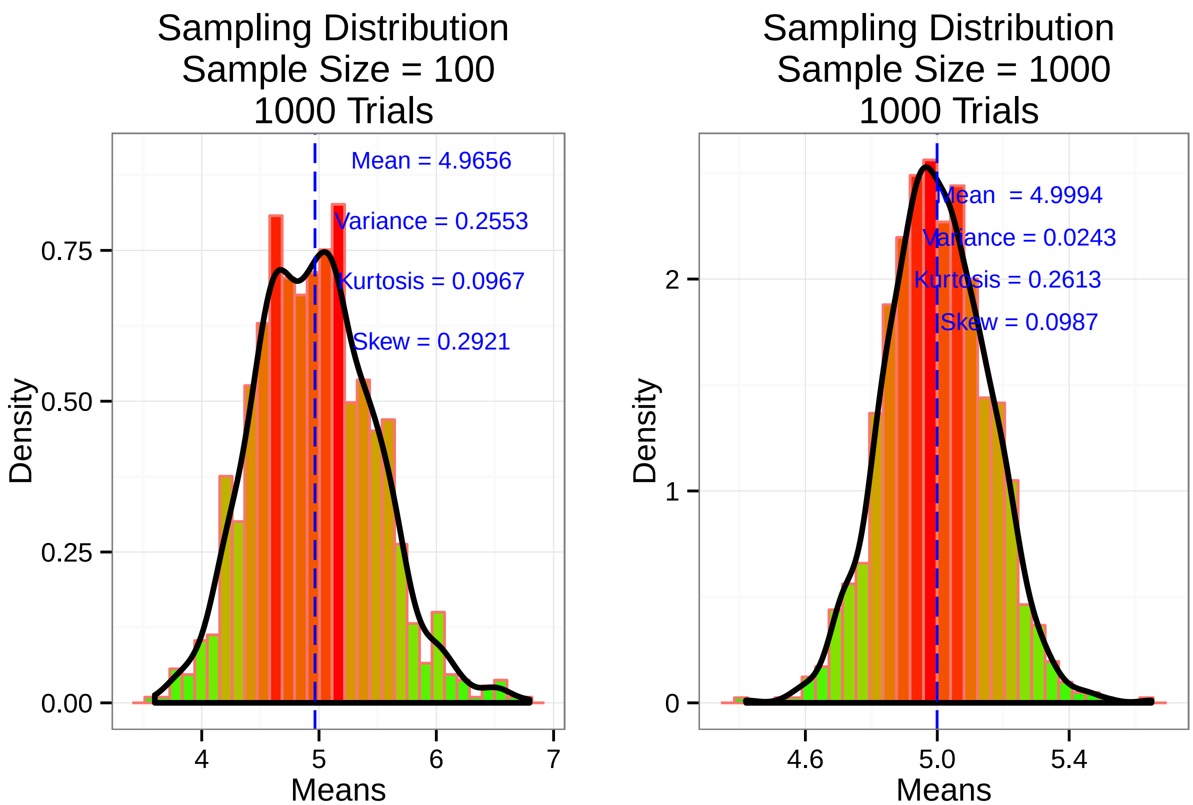
Furthermore, the theoretical variance of the sampling distribution is given by the variance of the population mean/sample size, or $(1/\lambda^2)/n$, which $= 0.625$. Here, it seems that the variance obtained from the sampling distribution of 0.6025, is a fair approximation of the theoretical variance of the sampling distribution obtained from the population mean.

Verifying Normality

In order to confirm that the sampling distribution obtained above is indeed normal, we will test a consequence of the central limit theorem, i.e. *as the sample size approaches infinity, the variation of the sampling* *distribution gets smaller and thus results in a closer approximation of the population mean.* We will conduct two simulations with larger sample sizes, n= 100 and n = 1000. We will conduct a 1000 trials and compute the variance and skew to give us an idea of the density curve.

```
#sim 1, sample size = 100
samp.dist1 <- NULL
for(i in 1:1000){samp.dist1 <- c(samp.dist1, mean(rexp(100, rate=0.2)))}

#sim 2, sample size=1000
samp.dist2 <- NULL
for(i in 1:1000){samp.dist2 <- c(samp.dist2, mean(rexp(1000, rate=0.2)))}
```



Consequently, we observe the following:

- The mean is approximated more closely as the sample size increases
- Variance of sampling distribution significantly decreases
- The skew of the density curve gets smaller

We can thus conclude that the sampling distribution of means of the exponential distribution approaches a normal distribution as the sample sizes increase.

END OF PART 1

APPENDIX 1: GIT CODE

The entire markdown file can be found in this github repo.

APPENDIX 2: PLOT CODE

Comparing Variance and Mean to Theoretical Variance and Mean

```r
#sampling distribution of sample means
g3 <- ggplot()
g3+geom_histogram(aes(x=samp.dist, fill=..count..), col="black")+
        scale_fill_gradient("Count",
                            low = "green",
                            high = "red")+

        #mean of sampling dist.
        geom_vline(aes(xintercept=mean(samp.dist)),
                   linetype="longdash",
                   col="blue")+
        geom_text(aes(x=mean(samp.dist)*1.15,
                      y=100,
                      label=paste0("Mean of \n Sampling Distribution = ",
                                   round(mean(samp.dist), 4))
                      ),
                  col="blue", size=3
                  )+
        #theoretical mean
        geom_vline(aes(xintercept=1/0.2),
                       linetype="longdash",
                       col="darkblue"
                   )+
        geom_text(aes(x=(1/0.2)*1.15,
                      y=90,
                      label=paste0("Theoretical Mean of \n Exponential Distribution = ",
                                   round(1/0.2, 4))
                      ),
                  col="darkblue", size=3
                  )+

        #variance of sampling dist.
        geom_hline(aes(yintercept=var(samp.dist)*100),
                       linetype="longdash",
                       col="blue"
                   )+
        geom_text(aes(y=var(samp.dist)*110,
                      x=3,
                      label=paste0("Variance of \n Sampling Distribution = ",
                                   round(var(samp.dist), 4))
                      ),
                  col="blue", size=3
                  )+
```

```
        #theoretical variance = population variance/n
        geom_hline(aes(yintercept=((1/(0.2^2))/40)*100),
                       linetype="longdash",
                       col="darkblue")+
        geom_text(aes(y=((1/(0.2^2))/40)*90,
                      x=3,
                      label=paste0("Theoretical Variance \n of Distribution = ",
                                   round(((1/(0.2^2))/40), 4))
                      ),
                  col="darkblue", size=3
                  )+
        labs(title="Sampling Distribution of Sample Means \n of the Exponential Distribution (1000 Trial
             x="Means", y="Count")+
        theme_bw()+
        theme(legend.position="none")
```

Verifying Normality

```
# PLOT SIM 1
g4 <- ggplot() + geom_histogram(aes(x=samp.dist1, y=..density.., fill=..count.., col="black"))+
        scale_fill_gradient("Count", low = "green", high = "red")+
        geom_density(aes(samp.dist1), col="black", size=1)+
        #mean of sampling dist.
        geom_vline(aes(xintercept=mean(samp.dist1)),linetype="longdash", col="blue")+
        geom_text(aes(x=mean(samp.dist1)*1.2, y=.9,label=paste0("Mean = ",
                                   round(mean(samp.dist1), 4))), col="blue", size=3)+
        #variance of sampling dist.
        geom_text(aes(y=.8, x=mean(samp.dist1)*1.2,
                      label=paste0("Variance = ", round(var(samp.dist1), 4))), col="blue", size=3)+
        #kurtosis
        geom_text(aes(y=.7,x=mean(samp.dist1)*1.2,
                      label=paste0("Kurtosis = ", round(kurtosis(samp.dist1), 4))),col="blue", size=3)+
        #skew
        geom_text(aes(y=.6,x=mean(samp.dist1)*1.2,
                      label=paste0("Skew = ", round(skewness(samp.dist1), 4))),col="blue", size=3)+
        labs(title="Sampling Distribution \n Sample Size = 100 \n 1000 Trials",x="Means", y="Density")+

# PLOT SIM 2
g5 <- ggplot() + geom_histogram(aes(x=samp.dist2, y=..density.., fill=..count.., col="black"))+scale_fil
        geom_density(aes(samp.dist2), col="black", size=1)+
        #mean of sampling dist.
        geom_vline(aes(xintercept=mean(samp.dist2)),linetype="longdash", col="blue")+
        geom_text(aes(x=mean(samp.dist2)*1.05, y=.9+1.5,
                      label=paste0("Mean  = ",round(mean(samp.dist2), 4))), col="blue", size=3)+
        #variance of sampling dist.
        geom_text(aes(y=.8+1.4, x=mean(samp.dist2)*1.05,
                      label=paste0("Variance = ", round(var(samp.dist2), 4))), col="blue", size=3)+
        #kurtosis
        geom_text(aes(y=.7+1.3,x=mean(samp.dist1)*1.05,
                      label=paste0("Kurtosis = ", round(kurtosis(samp.dist2), 4))),col="blue", size=3)+
        #skew
```

```r
        geom_text(aes(y=.6+1.2,x=mean(samp.dist2)*1.05,
                      label=paste0("Skew = ", round(skewness(samp.dist2), 4))),col="blue", size=3)+
        labs(title="Sampling Distribution \n Sample Size = 1000 \n 1000 Trials",
               x="Means", y="Density")+theme_bw()+theme(legend.position="none")
grid.arrange(g4, g5, ncol=2)
```