Exponenetial Distribution Analysis

Beau Norgeot

May 21, 2015

Introduction

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

```
library(dplyr)
library(ggplot2)
```

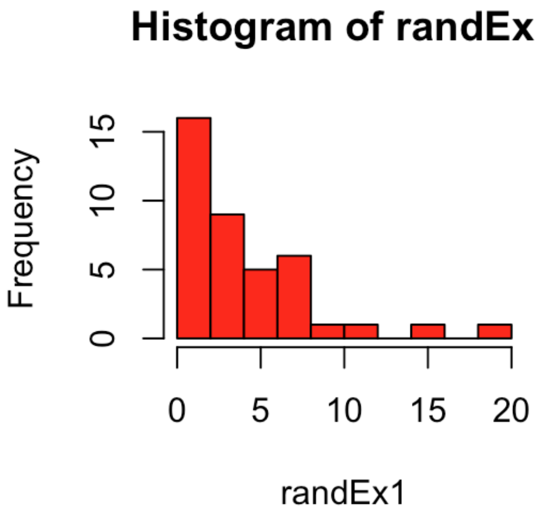
Part 1: Show the sample mean and compare it to the theoretical mean of the distribution

Generate a sample of 40, random, exponential values, with a lambda of .2

```
set.seed(101) # Let's keep this reproducible
randEx1 <- rexp(40, .2); mean(randEx1)
```

```
## [1] 4.034012
```

```
hist(randEx1, col = "red")
```

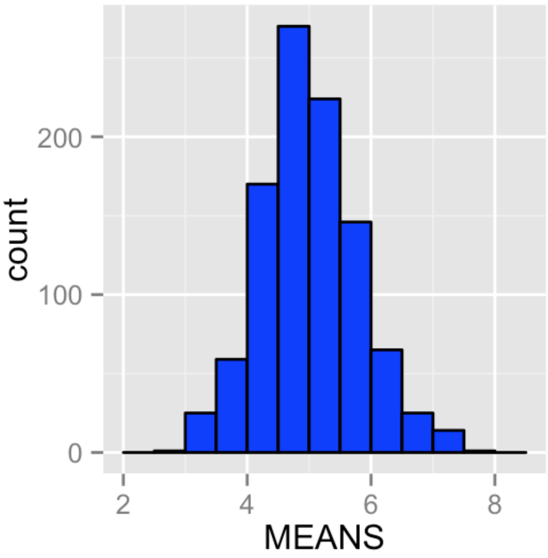


The mean from my first sample was 4.034 The theoretical mean for the distribution would be 1/lambda: (1/.2) = 5 Our sample value is about 20% lower than our expected value, but that's OK, we haven't sampled the entire population of exponential random variables that have a lambda of (1/5); The values we chose in this sample are just a bit smaller than the average of their kind. We'll see the frequency distribution of the sample means normalize as we take more and more samples.

We can see from the histogram that the intial sample isn't even remotely normal.

Generate 1000 random exponenetial samples, with a sample size of 40, and a lambda of .2

```
set.seed(101)
lambda = .2
sampleSize = 40
meansList <- data.frame(x = sapply(1:1000,function(x) {mean(rexp(sampleSize,lambd
a))}))
meansTbl <- tbl_df(meansList)
meansTbl <- rename(meansTbl, MEANS = x)
sampleData <- ggplot(meansTbl, aes(MEANS))
sampleData + geom_histogram(binwidth = .5, colour="black", fill="blue")
```



```
# The mean of all 1000 trials
sampleMean <- mean(meansTbl$MEANS)
# Theoretical mean
TheorectialMean <- 1/lambda
```

After generating 1000 samples of size 40, we can see that frequency distribution looks far more normal. The mean of actual distribution is 5.013, while the theoretical mean is 5.0

Part 2: Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

```
#Sample Data Variance
sampleVariance <- var(meansList$x);sampleVariance
```

```
## [1] 0.5985383
```

```
sampleSD <- sd(meansList$x); sampleSD
```

```
## [1] 0.7736526
```

```
#Theoretical Variance. theorySD == Standard Error
theorySD <- (1/lambda)/sqrt(sampleSize); theorySD
```

```
## [1] 0.7905694
```

```
theoryVariance <- theorySD^2; theoryVariance
```

```
## [1] 0.625
```

The expected variance was .625 while the observed sample variance was .599 The expected Standard Error was .791 while the observed Standard Error was .774 Observed values are within 2.5% of expected values. If we were to increase the sample size and/or the number of trials, we would see values converge even closer.

Part3: Show that the distribution is approximately normal

There are a couple of different ways to go about this. The Shapiro-Wilk Normality Test tests the Null Hypothesis that samples come from a Normal Distribution against the Alternative Hypothesis that the samples do not come from a Normal Distribution. You are testing against the assumption of Normality. This means that if p-value that results from the test is <= .05, we would reject the Null Hypothesis and say the sample data is NOT Normal. R has a simple function for this shapiro.test()

```
shapiro.test(meansList$x)
```

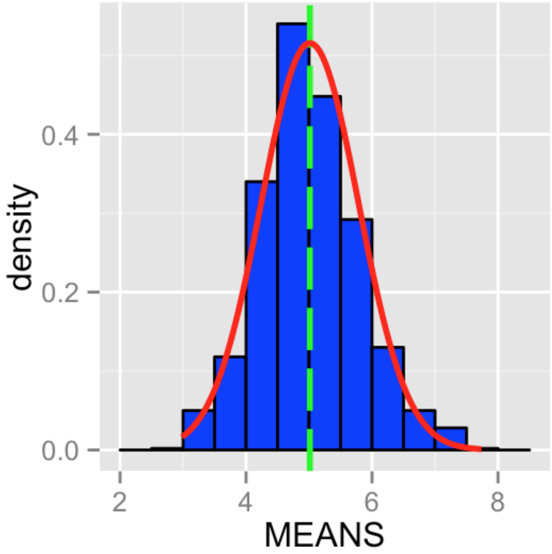
```
## 
##  Shapiro-Wilk normality test
## 
## data:  meansList$x
## W = 0.9934, p-value = 0.0001926
```

We can see that resulting p-value is substantially less than .05 and will not reject the Null Hypothesis. It’s reasonable to say that our sample data comes from a Normal Distribution.

Note that the Shapiro-Wilk test begins to loose it’s practical effectiveness as the size of data (number of trials) becomes larger. This is because when large amounts of data are present, even very small deviations from normality can be detected, leading to rejection of the null hypothesis, despite the fact that for practical purposes the data is effectively normal. The R-based Shapiro.test() protects against this by limiting the size of the input to 5000. However, it’s also a good idea to use plotting functions to check for normality.

If we make the assumption that our sample data is normal, we can plot a normal density function, using the sample mean and SD. We can lay the normal density function ontop of the our sample data and visually determine if the histogram appears to fit neatly under the normal curve.

```
sampleData <- ggplot(meansTbl, aes(MEANS))
sampleData + geom_histogram(binwidth = .5, colour="black", fill="blue",aes(y=..density..)) + geom_vline(xintercept = sampleMean, color="green",size=1,linetype="longdash") + stat_function(fun=dnorm, args = list(mean =sampleMean, sd = sampleSD),color="red",size=1.0)
```


density
MEANS