Statistical Inference Final Project - Part 1

Jerry Lakin

9/22/2020

Exponential Distribution Simulation Exercise

Central Limit Theorem

The purpose of this exercise is to explore the Central Limit Theorem (CLT). The CLT establishes that the distribution of averages of independent, individually distributed variables becomes that of a standard normal as the sample size increases. In order to examine this theorem, we will simulate a large number of averages of random variables drawn from the exponential distribution, and compare the characteristics this distribution of means to those predicted by the theorem.

The Exponential Distribution

The exponential distribution is the probability distribution of the time between events in a Poisson point process. An important parameter of the distribution is lambda, also known as the rate parameter. In this exercise, we will set lambda equal to 0.2. Some important facts about the exponential distribution are that the mean is equal to 1/lambda and the variance is equal to 1/lambda^2. We can simulate random variables from the distribution in R using the function rexp().

Simulation

We will run 1000 simulations taking the average of 40 random variables drawn from the exponential distribution. This snippet of R code initializes an empty vector expMeans of length 1000. Then it runs a loop 1000 times generating 40 random variables from the exponential distribution and adding their mean to expMeans.

```
n = 40
lambda = 0.2
expMeans <- rep(NA,1000)
for(i in 1:1000){
    expMeans[i] = mean(rexp(n, lambda))
}
```

Now we want to compare this distribution to the predictions of the CLT. The CLT predicts that the distribution of averages should be centered at the mean of the original distribution. For the exponential distribution, the mean equals 1/lambda

```
muPred <- 1/lambda
muSim <- mean(expMeans)
muPred
```

```
## [1] 5
```

```r
muSim
```

```
## [1] 5.024265
```

```r
muSim - muPred
```

```
## [1] 0.02426546
```

As we can see, the simulated mean is very close to the predicted mean, deviating by only about 1%. Because the expected values of the CLT depend on the sample size approaching infinity, this result is well within our expectations.

Next we will look at the variance. The CLT predicts that the variance of the distribution of means should be equal to the variance of the original distribution divided by the sample size.

```r
varPred <- 1/(lambda^2*n)
varSim <- var(expMeans)
varPred
```

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

```r
varSim
```

```
## [1] 0.6162585
```

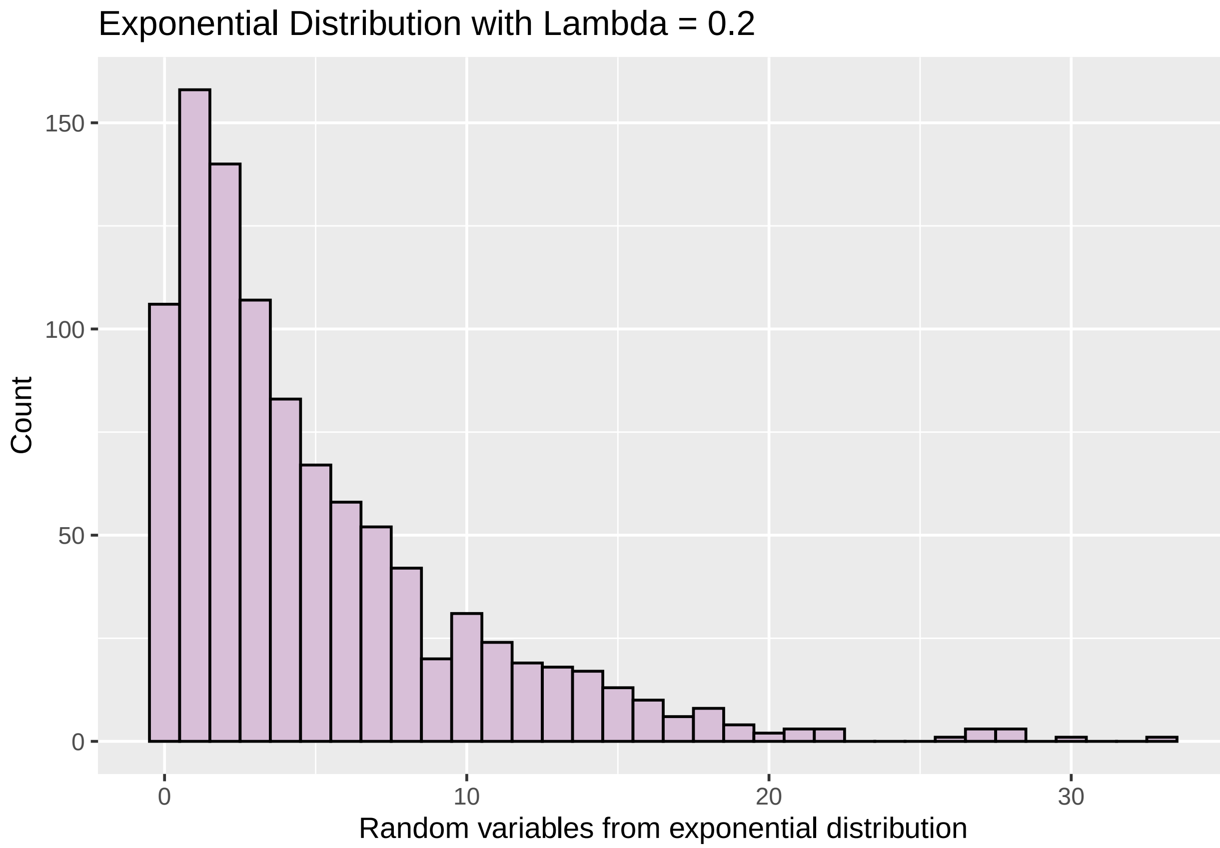
```r
varSim - varPred
```

```
## [1] -0.008741546
```

Once again our predicted value is very close to the observed value, this time deviating by about 6%, well within our expectations.

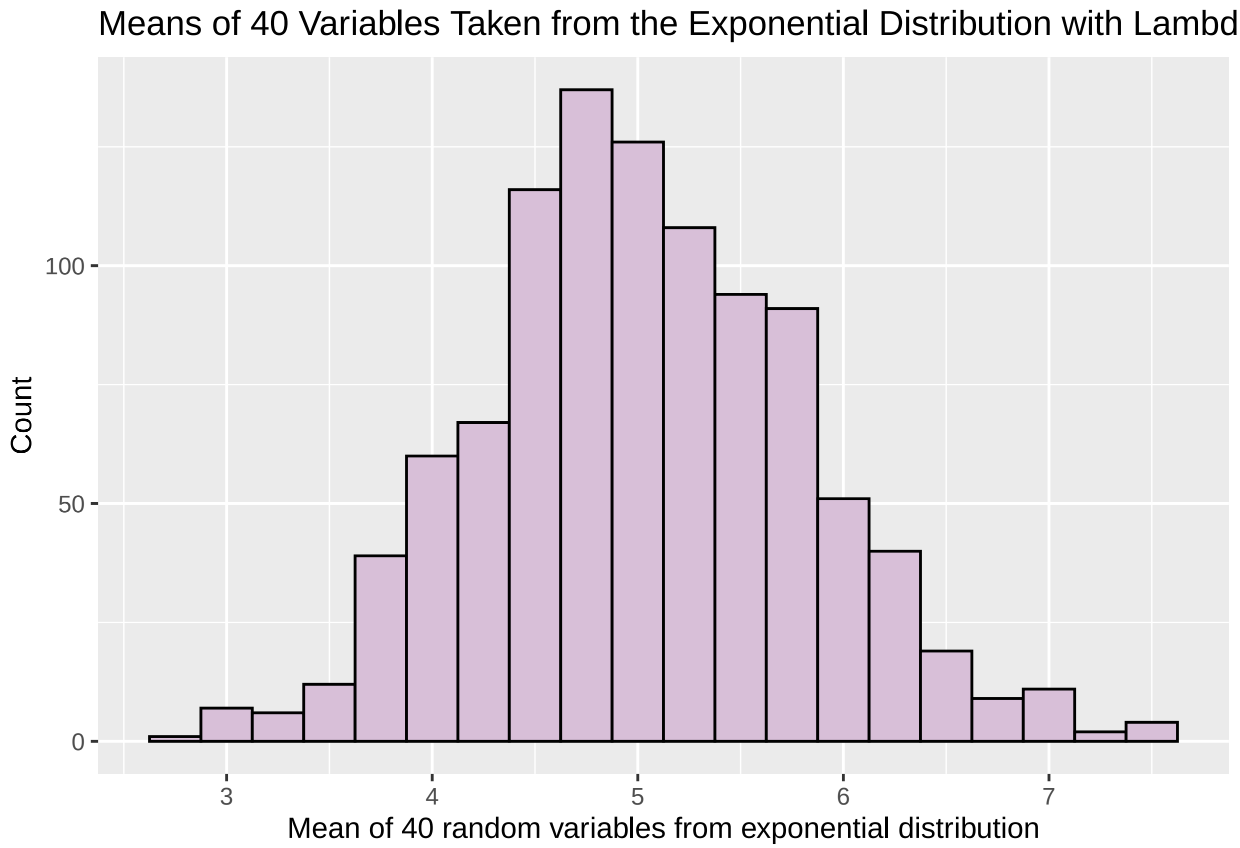
Finally, we would like to examine the distribution of means to prove that it is approximately normal as predicted by the CLT. To do this we will look at a histogram of random variables taken from the exponential distribution compared to a histogram of the simulated means. First, the random variables from the exponential distribution.

```r
exps <- rexp(1000, lambda)
library(ggplot2)
g1 <- ggplot(data.frame(exps = exps), aes(x = exps))
g1 + geom_histogram(color = "black", fill = "thistle", binwidth = 1) +
     ggtitle("Exponential Distribution with Lambda = 0.2") +
     xlab("Random variables from exponential distribution") +
     ylab("Count")
```



Next, the distribution of simulated means

```
g2 <- ggplot(data.frame(expMeans = expMeans), aes(x = expMeans))
g2 + geom_histogram(color = "black", fill = "thistle", binwidth = 0.25) +
     ggtitle("Means of 40 Variables Taken from the Exponential Distribution with Lambda = 0.2") +
     xlab("Mean of 40 random variables from exponential distribution") +
     ylab("Count")
```



As we can see, while the exponential distribution peaks around 1 and decays for higher values, the distribution of simulated means is symmetric and appears to be roughly normal, centered around 5 which is the theoretical mean of the exponential distribution for lambda = 0.2.