Statistical Inference Assigment - Kevin Carhart

PART ONE: Exponentials and the Central Limit Theorem

Overview

The first portion of this report is devoted to an illustration of the Central Limit Theorem. The principle described here is that a set consisting of the mean of a large set of draws from a population will hew to a normal distribution, even when the direct draws from that population do not. In various course materials we have shown this for Poisson and Binomial data, and in this project we will do something similar for the Exponential distribution.

Simulations

The Central Limit Theorem (CLT) states that the distribution of averages of independent and identically distributed (iid) variables becomes that of a standard normal as the sample size increases. In order to illustrate this principle on exponential data, I created an R function to perform 1000 draws of exponential data using rexp. There are 40 values in each simulation, and 1000 simulations altogether. The other parameter to rexp is set to 0.2. Here are the starting constants in R:

```
lambda = 0.2;nosim = 1000;n = 40#lambda is 0.2,nosim is 1000,n is 40
```

Here is the function that will perform the simulations.

```
DrawFromExp <- function (lambda,nosim,n) {
lambda = .2;nosim <- 1000;n <- 40;all_means <- (1:1000) * 0;loop_over <- 1:1000
    sapply(loop_over, function(iterator){
    next_exp_vector <- rexp(n, lambda)
    next_mean <- mean(next_exp_vector)
 all_means[iterator] <- next_mean
})
}
```

This function returns a vector of 1000 means which are suitable for further analysis:

```
all_means <- DrawFromExp(lambda,nosim,n)
```

This is the end of the simulation itself, so now we can use these 1000 means to answer the questions.

A comparison of means

The implication of the CLT is that the mean and other attributes of these draws will approach the attributes of the original population. The potentially surprising thing is that merely taking 1000 draws of exponential data does not do this. First let's show this by looking at the means.

```
# This is the sample mean or mean of means of our 1000 simulation results
mean(all_means)
```

```
## [1] 5.019333
```

```
# And this is the mean of an exponential distribution,
# defined as 1/lambda = 1/.2 = 5
1/lambda
```

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

The sample mean is very close to 5. These values are very close.

A comparison of variances

The second way of confirming the CLT is by comparing what the variance is supposed to be against what it calculates out to from the simulation data. The theoretical variance is…

sigma squared over n

Or

(population standard deviation) ^2 / n

Or

(1/lambda) ^ 2 / n

```
(1/lambda) ^ 2 / n
```

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

The other side of the equation is simply the variance found from the simulation results.

```
var(all_means)
```

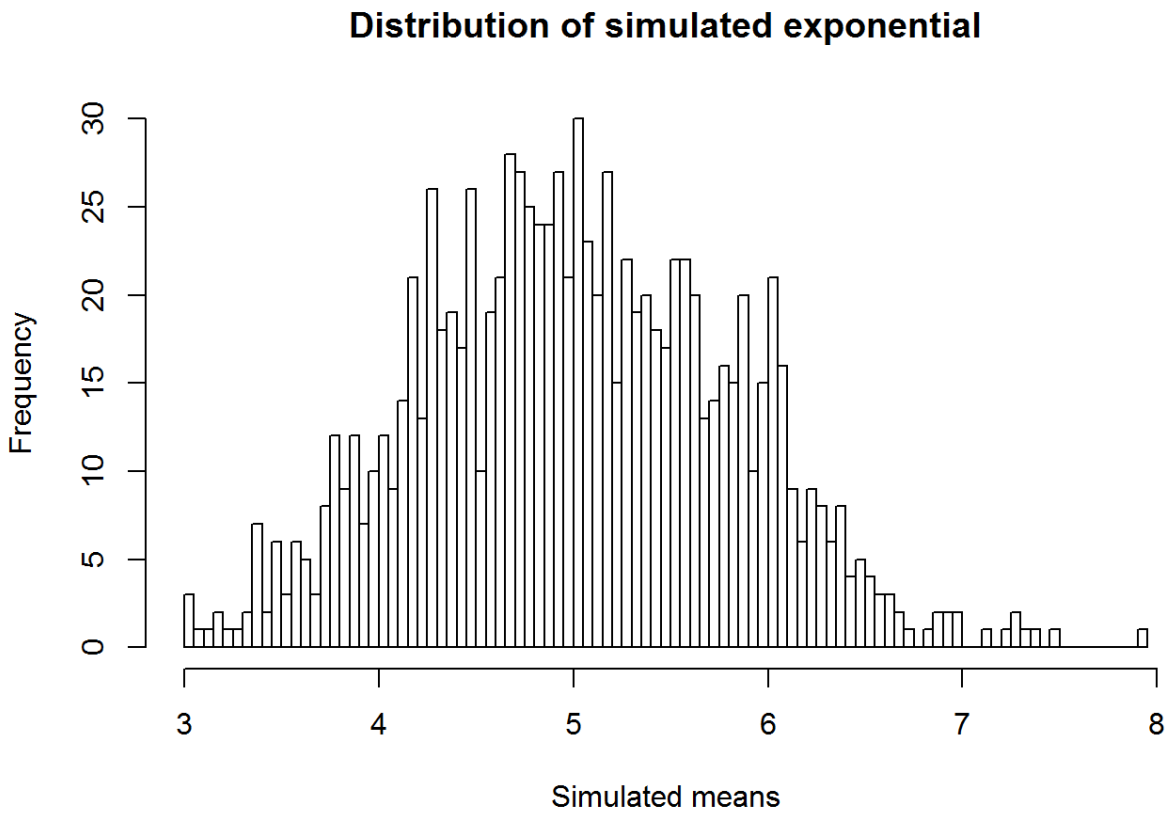
```
## [1] 0.6392729
```

This is also rather close.

Show that the distribution is approximately normal

Simply from eyeballing, when the volume of data is sufficient, the sample approaches a Gaussian appearance as the CLT says it should:

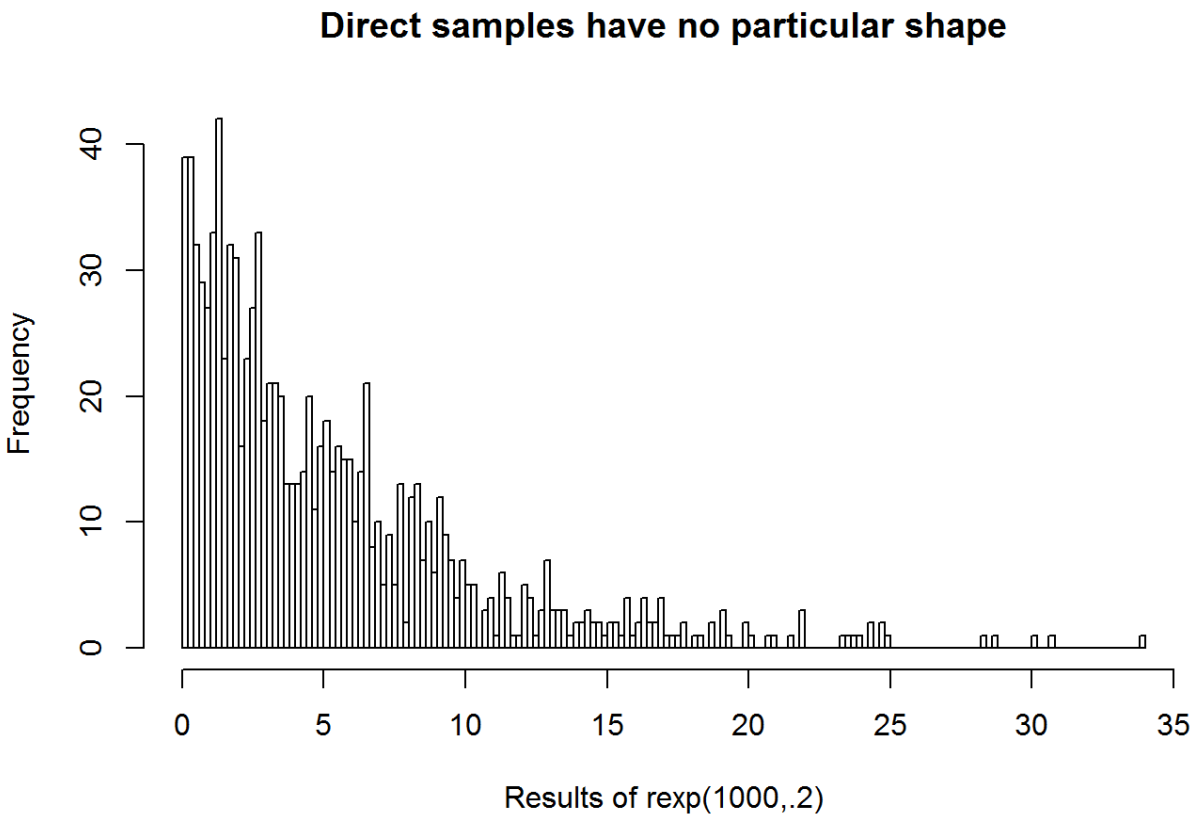
```
hist(all_means,150,main="Distribution of simulated exponential",xlab="Simulated means")
```



So we know that the CLT has held up here because of the similarity of means, the similarity of variances, and the colloquial yet effective means of examining the histogram to see whether it looks like the normal.

In addition, the instructions state “focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.” Compare the above histogram with simply drawing 1000 exponential values once.

```
hist(rexp(1000, lambda),150,main="Direct samples have no particular shape",xlab="Results of rexp(10
00,.2)")
```



The assignment materials make this point for uniforms. I have laid out the histograms in the reverse order, but the point is that the simulated means look “far more Gaussian than the original [exponential] distribution!”