Statistical Inference Final Project: Demonstrating the Central Limit Theorem and using Confidence Intervals - Part 1

James Wright

October 2016

Overview

This report covers the Statistical Inference course project, which is made up of two parts. The first part of the project is a demonstration of the Central Limit Theorem using the exponential distribution in R - how sampling works with the Central Limit Theorem. The second part of the project was an analysis of the ToothGrowth dataset available in the R datasets package.The aim here is to use confidence intervals to compare tooth growth by supp and dose.

Part 1: Demonstrating the Central Limit Theorem using the Exponential Distribution in R

The Simulation

The exponential distribution function in R, `rexp()`, is a function which will randomly generate a number from the exponential distribution. It takes as input arguments n, the sample size, and a rate.

The objective is to take 1000 samples, each of size 40, from the population of the numbers in the exponential distribution. We'll find the mean of each sample and look at how those means are distributed. According to the Central Limit Theorem, the distribution of sample means should be approximately normally distributed (because our sample size is large enough, i.e. >30).

This was done as follows:

40 random numbers from the exponential distribution were obtained using the `rexp()` function, and the mean of this numeric vector was calculated and stored. This procedure was repeated 1000 times.

The R code for the simulations is listed below:

```
n=40
lambda=0.2
mns=NULL
vars=NULL
for (i in 1:1000){
        samp=rexp(n,lambda)
        mns=c(mns,mean(samp))
        }
```

The sample size, n, was 40 and the number of observations was 1000. The mns variable is a numeric vector containing the 1000 means. Lambda, the rate parameter, was set to 0.2 for the simulations.

Before we look at the sampling distribution of the means, it would be interesting to look at the sort of data we're dealing with to get a better idea of the bigger picture!

Let's generate a random sample, of size 40, from the exponential distribution:

```
exps = rexp(n,lambda)
```

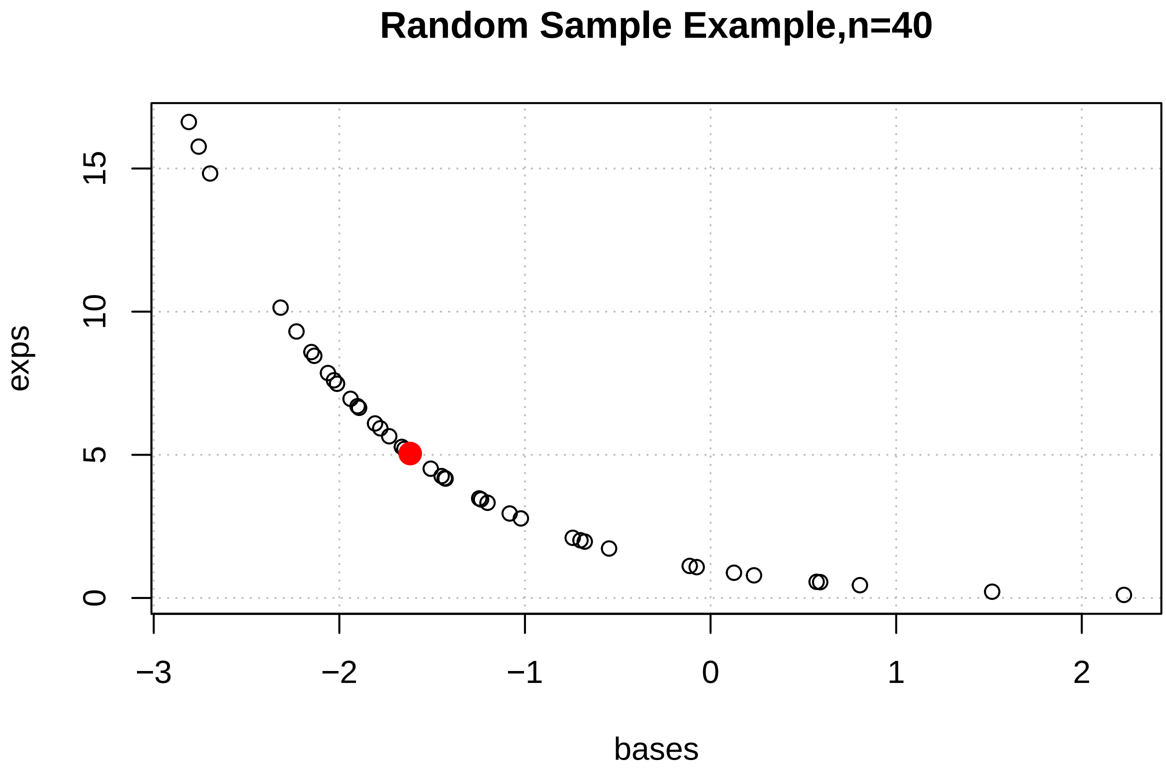
We can find the base of the expression the generated each number by performing the inverse function, i.e. taking the natural log:

```
bases = log(1/exps)
```

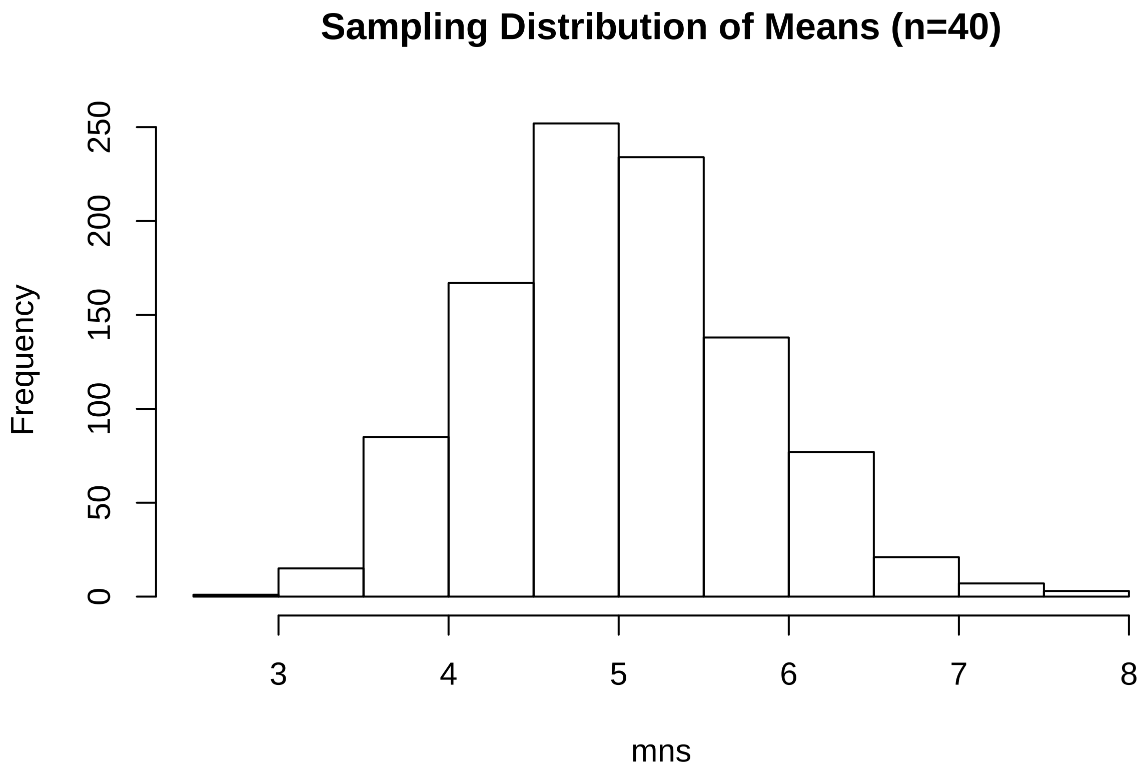
We've taken the log of the inverse exponential because, in actual fact, the exponential distribution used in statistics, and in R, is also known as the negative exponential distribution (see https://en.wikipedia.org/wiki/Exponential_distribution for more detail) - now we can be consistent when looking at the example later on.

When we plot the exponentials against the bases, we can see that the sample is (of course) distributed exponentially. Just for fun, we can also highlight where the mean is too:

```
m = mean(exps)
b = log(1/m)
exps = c(exps,m) # just in case there's no point that
bases = c(bases,b) # corresponds to the mean in our sample!
plot(bases,exps,
     title('Random Sample Example,n=40'),
     col=ifelse(bases==b,'red','black'),
     cex=ifelse(bases==b,1.5,1),
     pch=ifelse(bases==b,19,1),
     panel.first=grid(col='gray'))
```



Ok, back to the problem now. If we now plot a histogram of the means that we got for each sample earlier, we see that the distribution of the sample means is quite different to the distribution of the samples and the original population! More on this a bit later on!



1.Comparing the Sample Mean to the Theoretical Mean of the Exponential Distribution

The theoretical mean of the exponential distribution is given by 1/lambda = 1/0.2 = 5.

The mean of the sampling distribution can be found by:

```
mean(mns)
```

```
## [1] 4.999148
```

We can see that this is approximately equal to the theoretical value of 5. This proves that the mean of the distribution of sample means is approximately equal to the theoretical mean, and we would expect this as our sample size was large enough.

2.Comparing the Sample Variance to the Theoretical Variance of the Exponential Distribution

The variance measures the variability, or spread, of the data, and is calculated as the square of the standard deviation. The standard deviation of the exponential distribution is given by 1/lambda = 5, so the variance is 25. This tells us that the expected squared deviation around the population mean is 25.

The variance of the sampling distribution of the mean is found by the square of the standard error of the distribution of the mean, (sigma/sqrt(n))ˆ2 = sigmaˆ2/n, where sigma is the population standard deviation.

We're lucky to have the population variance, sigmaˆ2 = (1/lambda)ˆ2, in this case, as it's not always available. This allows us to compare the variance of our sample of means to the theoretical variance calculated with the above equation.

```
# Actual variance of sample mean
var(mns)
```

```
## [1] 0.6110345
```

```
# Theoretical variance of sample mean
25/40
```

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

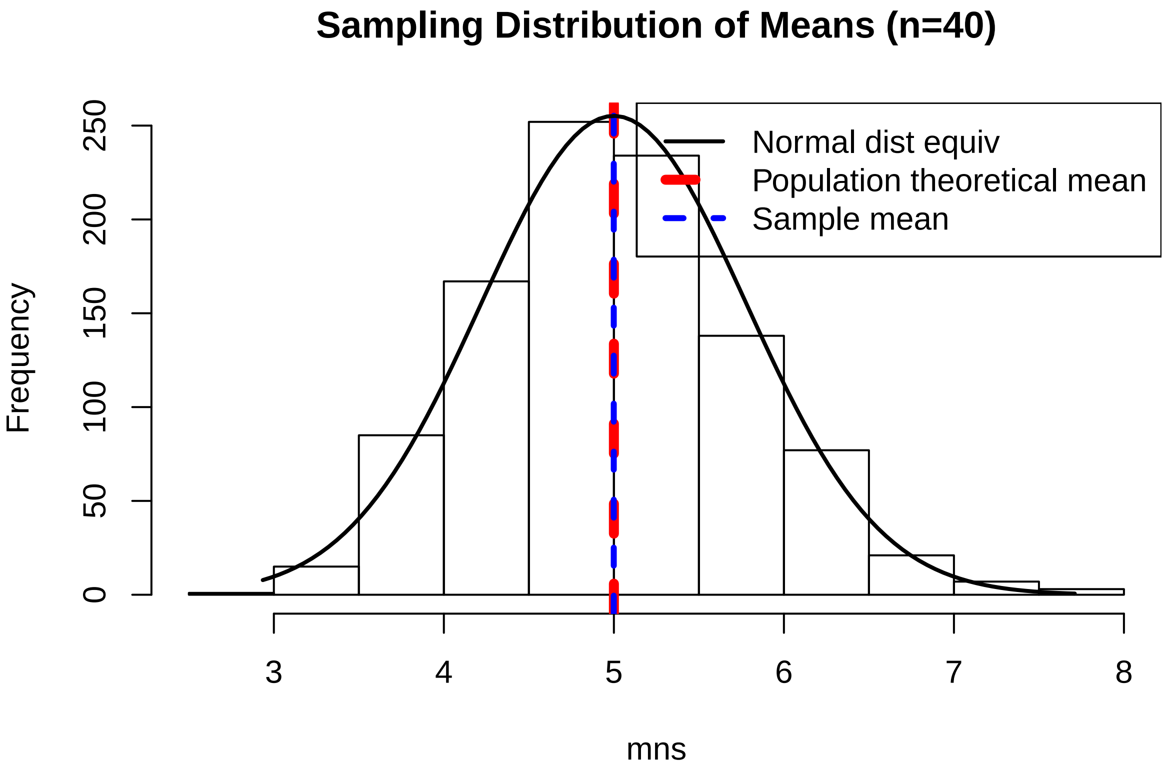
The theoretical variance of the sample mean appears to estimate the variability of the sample mean distribution well.

3.Central Limit Theorem in Action!

To recap, the Central Limit Theorem (CLT) states that the distribution of means of a population sample will approximate a normal distribution, even if the population itself is not normally distributed, if the sample size is large enough.

Applied to our case here, we're saying that the population is every number available from the exponential distribution (which, by its defintion, is not normally distributed!). Each sample is made up of the 40 numbers randomly taken from this exponential distribution. We have 1000 samples of size 40, and when we take the mean of each sample, we have 1000 means. From the CLT, we know that the distribution of these sample means should be approximately normal.

This is clearly seen by looking at the distribution of means in the histogram again. We can plot a line showing a normal distribution centred at the mean of mns, with a standard deviation equal to the standard deviation of the mns vector - we'll call this a "normal distribution equivalent". We can see that the histogram roughly takes this shape, and is centred on the sample mean (dashed blue line). Additionally, if we plot a dashed red line on the at the location of the theoretical mean, we see that the "equivalent" normal distribution that is approximated by the sampling distribution is roughly centred at the theoretical mean, and therefore the mean of the sample means distribution is a good estimator of the population mean.

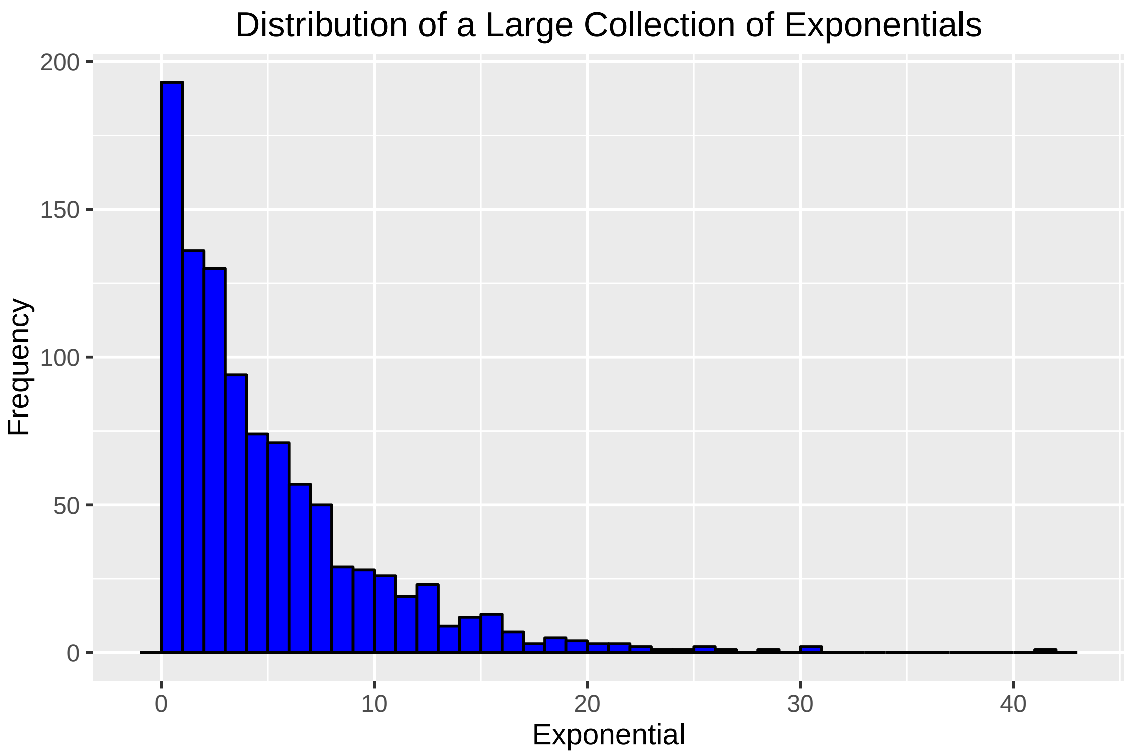


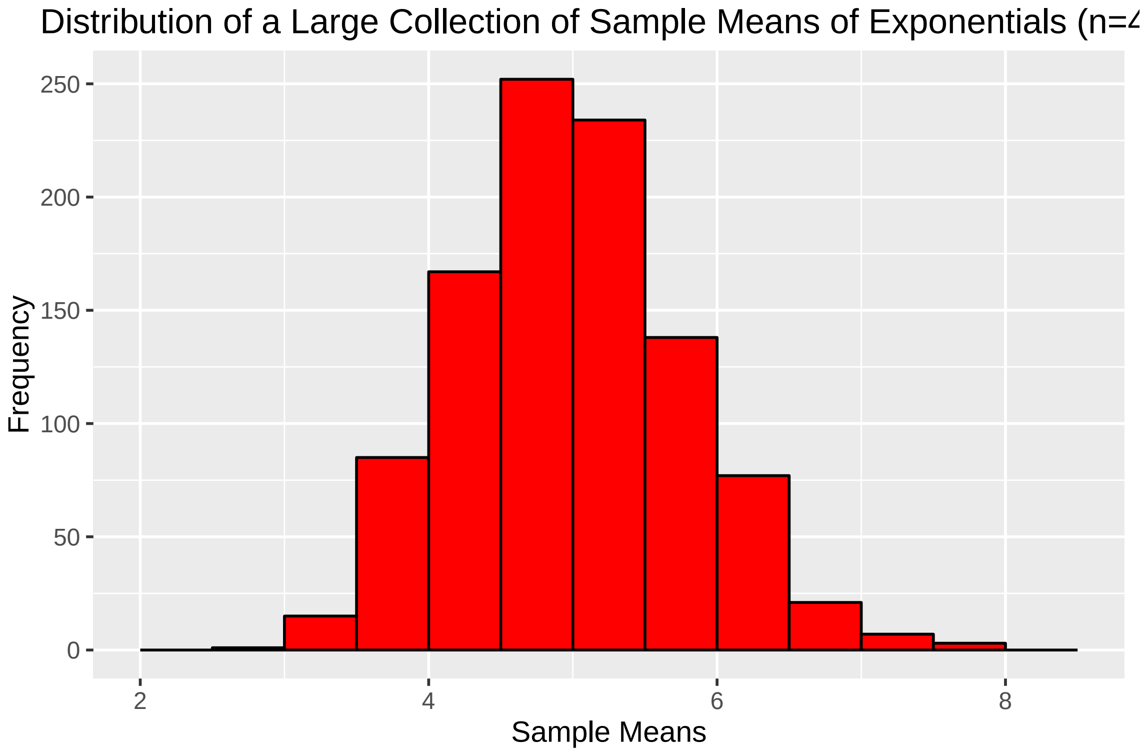
We can see just how different the sample distribution of means is to the sample data itself. Therefore, it is important to realise that the distribution of a large collection of random exponentials is completely different to the distribution of a large collection of averages of samples containing exponentials (making sure that the sample size is greater than 30). In this case, we go from an exponential distribution to an approximately normal distribution. This is seen from the figures in the Appendix.

To conclude on Part 1 of this report:

1. We have seen that a large sample (>30) of random exponentials is itself exponentially distributed - just like the population it was taken from.
2. A large collection of these samples (like the 1000 that we took) will still be exponentially distributed.
3. When we take the mean of each sample, however, and then plot a histogram of the means, this distribution is approximately normal!

Appendix


Distribution of a Large Collection of Exponentials
Frequency
200
150
100
50
0
0
10
20
30
40
Exponential


Distribution of a Large Collection of Sample Means of Exponentials (n=4
Frequency
250
200
150
100
50
0
2
4
6
8
Sample Means