Statistics Assignment 1: Central Limit Theorem

JB

September 22, 2017

Overview

This assignment demonstrates the Central Limit Theorem for an exponential probability density function (pdf). The idea is that if 40 samples are drawn 1000 times from the exponential distribution and if the mean is taken across all 40 samples in each draw, then the distribution of the 1000 resulting sample means should be approximately normally distributed. The average value of the sample means should be very close to the theoretical mean of the exponential pdf. Also, the standard deviation of the sample mean distribution should be close to the theoretical standard error, which is the standard deviation of the exponential pdf divided by the square root of the sample size (n=40 in this case). The exponential rate, λ, will be set to 0.2 for this assignment. The theoretical mean, μ, and standard deviation, σ of the exponential pdf is $1/\lambda$. The sample mean distribution should look completely different (normal, or bell-shaped) compared to the original exponentional pdf.

Simulations

This section outlines the code needed to do the analysis.

Load the plotting libraries needed to show the distributions.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.4.1
```

Set λ, μ and standard error, $\sigma/\sqrt{n}$.

```
lambda = 0.2
theoretical_mean <- 1/lambda
standard_error <- (1/lambda)/sqrt(40)
```

For the first plot, we want to show the original theoretical exponential pdf, but we also want to draw 1000 random variables from the exponential pdf to show that *without any averaging*, these raw samples are in fact exponentially distributed. This code chuck draws those 1000 exponential variables. These variables will also be put into a data frame for plotting.

```
exp_values <- rexp(1000,lambda)
exp_frame <- data.frame(exp_values)
```

This next code chunk, generates 1000 draws of 40 samples each from the exponential distribution, and then calculates a sample mean across all 40 samples in each draw. This results in a vector of 1000 sample means. Since an average was taken across all 40 samples in each draw, then according to the Central Limit Theorem, this averaging will result in a vector of sample means that is *normally* (not exponentially) distributed. The resulting histogram of sample means should therefore have a bell-curve shape. The average of all 1000 sample means is calculated, which should agree closely with the theoretical mean of the exponential distribution determined previously. The standard deviation of the sample means should also agree closely with the expected standard error calculated previously, according to the Central Limit theorem. For plotting, the sample means are put into a data frame.

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

average_means <- mean(mns)
sd_means <- sd(mns)
means_frame <- data.frame(mns)
```

Sample vs. Theoretical Comparison: Mean, Standard Deviation, and Distributions

This section compares the average of the sample mean distribution to the theoretical mean of the exponential distribution, and they should be very close according to the Central Limit Theorem. The standard deviation of the sample means distribution should also agree closely to the theoretical standard error, according to this theorem. The following code summarizes these comparisons:

```
cat("Theoretical mean: ", theoretical_mean, "\n")
```

```
## Theoretical mean:  5
```

```
cat("Average of sample means: ", round(average_means,3), "\n")
```

```
## Average of sample means:  5.011
```

```
cat("Theoretical standard error: ", round(standard_error,3), "\n")
```

```
## Theoretical standard error:  0.791
```

```
cat("Standard deviation of sample means: ", round(sd_means,3), "\n")
```

```
## Standard deviation of sample means:  0.806
```

From these results, the agreement between these quantities is good as expected from the central limit theorem. These results are explored more closely in the following plot figure. The plot on the left shows the histogram of the 1000 original random variable samples from the exponential distribution. As expected, they follow this distribution and the theoretical exponential curve is overlaid on the histogram. μ is shown as solid blue vertical line on this plot.

In the plot on the right, the density histrogram of the 1000 sample means is shown, and this is clearly no longer an exponential distribution. The theoretical normal distribution with mean, μ and standard deviation $\sigma/\sqrt{n}$ is overlaid on the histogram and it fits the sample distribution very well. The average of the sample means, $\bar{X}$, is shown as a vertical blue line, and it agrees closely with μ. To show that the theoretical standard error and sample standard deviations agree closely, the theoretical and sample 95% confidence intervals are shown on the plot. The theoretical 95% confidence interval, shown by the solid green lines, is given as $\mu \pm 1.96\sigma/\sqrt{n}$. The sample 95% confidence interval, shown as the dashed red lines, is given as $\bar{X} \pm 1.96s$, where s is the sample standard deviation of the sample means vector. These confidence intervals agree very closely on the plot.

The following code generates a figure and the two plots side by side.

```
exp_plot <- ggplot(exp_frame, aes(x=exp_values)) +
geom_histogram(alpha = .80, binwidth=1, boundary=1, colour = "black",
fill = "limegreen",aes(y = ..density..))+
stat_function(fun = dexp, args = list(rate=lambda), size=2)+
geom_vline(xintercept = theoretical_mean, linetype = "solid",color="blue",
size=2,alpha = 0.8)+
```

```
labs(title="Sample Exponential Density Plot",x = "x",y="P(x)")+
theme(text = element_text(size=16))

sample_means_plot <- ggplot(means_frame, aes(x=mns)) +
geom_histogram(alpha = .80, bins = 15, colour = "black",
fill = "orange",aes(y = ..density..))+
stat_function(fun = dnorm, args = list(mean = theoretical_mean, sd = standard_error)
, size=2)+
geom_vline(xintercept = average_means, linetype = "solid",color="steelblue",size=2,
alpha = 0.7)+
geom_vline(xintercept = c(average_means-1.96*sd_means, average_means+1.96*sd_means),
linetype = "dashed",color="red",size=2,alpha=0.6)+
geom_vline(xintercept = c(theoretical_mean-1.96*standard_error,
theoretical_mean+1.96*standard_error),
linetype = "solid",color="green",size=2,alpha=0.6)+
labs(title="Sample Mean Density Plot",x = "Sample Mean",y="Probability")+
theme(text = element_text(size=16))

plot_CLT <- grid.arrange(exp_plot, sample_means_plot, ncol = 2)
```

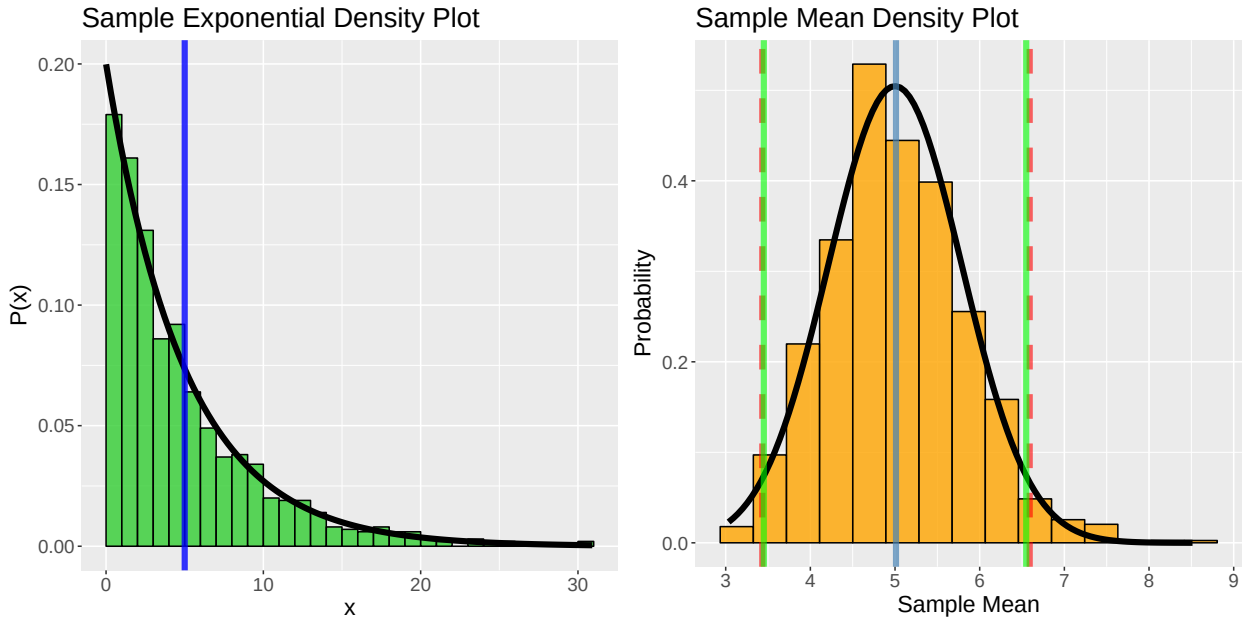

Figure 1: Density histgram plots of the exponential distribution (left) and the sample means normal distribution (right). The theoretical density functions are shown overlaid on the histograms. The means are shown on both plots. The theoretical and sample 95% confidence intervals are also shown on the sample means plot.

Conclusion

This exercise emphasizes the Central Limit theorem effectively, demonstrating that regardless of the shape of the distribution of the original random variables, averaging samples of these variables over many simulations will produce a data set of mean values for the variable that is normally distributed.