Statistical Inference Course Project - Central Limit Theorm from Means of Exponential Distribution

Coursera Data Science Specialization

S Carroll

April 2018

Overview

The goal of this project is to demonstrate that the means of a distribution tend to be distributed normally. This is a key tenant of the Central Limit Theorem. As an example, histograms of the means of a simulated exponential distribution are shown to approach normal distributions, particularly with larger sample draws. Interestingly, while histograms of **means** are more normally distributed, histograms of **individual** samples are not normally distributed.

A 1000-sample simulation investigating the distribution of averages of 1, 40, and 10,000 sample draws from the exponential distribution was performed. The exponential distribution is described by $f(x) = \lambda\ e^{-\lambda x}$, for $x \geq 0$ and choosing $\lambda = 0.2$. The samples are assumed to be independent and individually distributed. Later, the confidence interval is formed for the 40 and 10,000 sample draw data, assuming the distribution is approximately normal with the mean and standard deviation.

This project satisfies the following objectives: * An exploratory data analysis highlighting basic features of the data with a single table
* Comparison of the mean of the sample means to the theoretical mean $(1/\lambda)$ of the distribution
* Comparison of the variance $(1/\lambda^2)$ of the sample means to the theoretical variance of the distribution
* Show that the distribution is **approximately normal** by focussing on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of **averages**
* Show the two-tailed, 95% confidence interval around the mean and correctly interpret the data to show the proportion of the distribution contained within the interval
* Restate assumptions and summarize conclusions

Simulation Code

```
library(ggplot2)
library(pander)

set.seed(04252018)
lambda = 0.2

# 1, 40, and 10,000 - sample draw from exponential distribution, simulation size = 1000
n=1000
iexp <- rexp(n,lambda)

emns = NULL
for (i in 1:n) emns = c(emns, mean(rexp(40,lambda)))

emns10k = NULL
for (i in 1:n) emns10k = c(emns10k, mean(rexp(10000,lambda)))
```

```r
# tabulate summary statistics and theoretical comparisons compatible with .pdf output
sumall <- round(rbind(summary(iexp),summary(emns),summary(emns10k)),3)
rownames(sumall) <- c('1-sample draw', '40-sample draw', '10k-sample draw')

sumtheo <- round(rbind(c(1/lambda,(1/lambda)^2, (1/lambda)^2), c(mean(iexp), var(iexp),var(iexp)*1),c(m

rownames(sumtheo) <- c('Theoretical', '1-sample draw','40-sample draw', '10k-sample draw')
colnames(sumtheo) <- c('Mean', 'Variance','Variance * # Draws')
```

Results of Data from the Exponential Distribution

Initial Exploration

Summary statistics for the **means** of each set of random draws are shown in the following table. The 1-sample draw is a classic exponential distribution, and is not normally distributed. As the number of drawn samples for averaging increases, however, the median and mean are more similar in value. This is a characteristic of normally-distributed data.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1-sample draw	0.002	1.599	3.563	5.021	6.926	47.8
40-sample draw	2.932	4.448	4.976	5.012	5.519	8.278
10k-sample draw	4.851	4.966	4.999	5	5.035	5.177

Compare Simulation Samples With Theoretical Values

As shown in the following table, as the number of draws averaged increases, the mean approaches the theoretical value. The variance for each dataset decreases as the number of draws averaged increases. However, the variances are similar to theoretical values once the number of draws are accounted for, in the last column.

	Mean	Variance	Variance * # Draws
Theoretical	5	25	25
1-sample draw	5.021	24.29	24.29
40-sample draw	5.012	0.62	24.79
10k-sample draw	5	0.003	25.52

Normality of Distribution

There are several ways to assess normalcy of data. Histograms, q-q plots, and a numeric test of normality are shared below.

Histograms

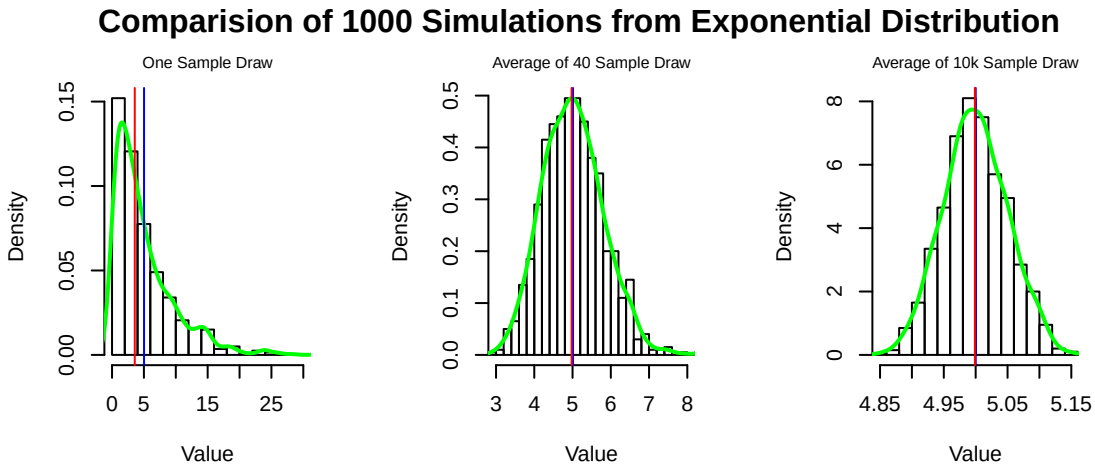
```r
# histograms to compare means with indidual draws
par(mfrow = c(1,3), mar = c(4, 4, 2, 3), oma = c(0, 0, 2, 0))

hplot <- function(df,mtitle,xl){
      hist(df, main = list(mtitle,cex = 0.8, font = 1), xlab = 'Value', density = NULL, xlim = xl, prob
      lines(density(df), col = 'green', lwd = 2)
```

```
      abline(v = mean(df), col = 'blue')
      abline(v = median(df), col = 'red')
}

hplot(iexp,'One Sample Draw',c(0,30))
hplot(emns, 'Average of 40 Sample Draw',c(3,8))
hplot(emns10k, 'Average of 10k Sample Draw',c(4.85,5.15))

mtext("Comparision of 1000 Simulations from Exponential Distribution", font = 2, outer = TRUE)
```



As the number of averaged draws increases, the histogram better approaches a bell-shaped curve. The left-most histogram is not normal. The middle and right-most histograms are more normally distributed; the mean and median are similar in these cases.

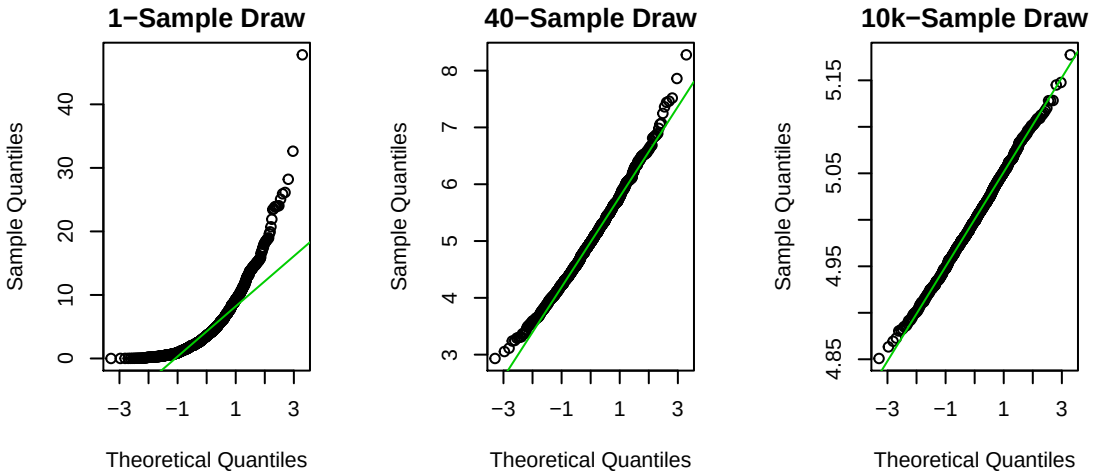
Quantile Plots

```
# QQ plot for distributions
par(mfrow = c(1,3), mar = c(4, 4, 2, 3))

qqplt <- function(df, mt) {
  qqnorm(df, main = mt)
  qqline(df, col = "3")
}

qqplt(iexp, '1-Sample Draw')
qqplt(emns, '40-Sample Draw')
qqplt(emns10k, '10k-Sample Draw')

mtext("Quantile Plot for Increasing Number of Averaged Draws", font = 2, outer = TRUE)
```



The means of each sample set were transformed to a plot that compares quantiles from the sample data with a theoretical normal distribution. Sample data that is normally distributed tends to follow the diagonal line. The left-most graph is from individual data, and is not normally distributed. The middle graph averages 40 samples; it is closer to a normal distribution but has some deviation at the tails. The right-most graph better approaches a normal distribtion as it is closest to the diagonal line.

Shapiro-Wilk Test of Normality

This numerical test claims the null hypothesis that the data are normally distributed. The alpha level is chosen at 0.05. If the p-value is less than alpha, the null hypothesis is rejected.

```
STresult <- signif(as.numeric(c(shapiro.test(iexp)[2],shapiro.test(emns)[2],shapiro.test(emns10k)[2])),1
names(STresult) <- c('1 Draw', '40 Draws', '10k Draws')
STresult
```

```
##    1 Draw  40 Draws 10k Draws
##   1.3e-32   2.3e-04   6.1e-01
```

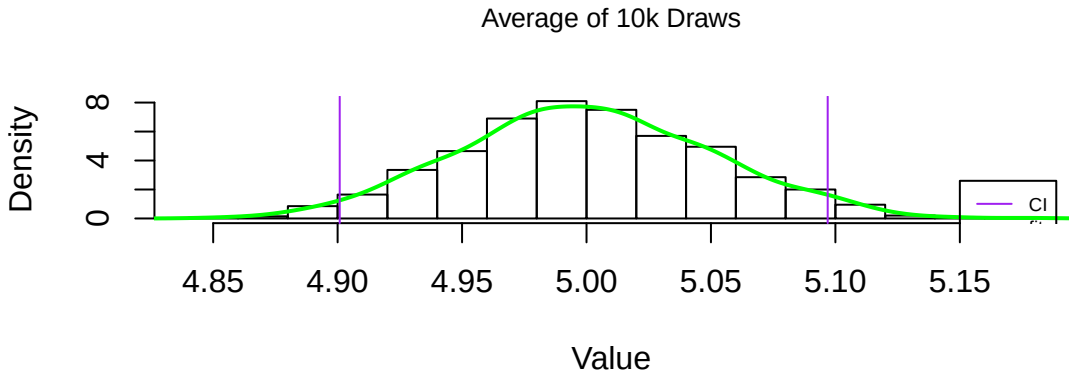
For sample draws of 1 and 40, the high p-value indicates the null hypothesis is rejected, concluding that these datasets are not normally distributed. On the other hand, for sample draws of 10k, the p-value is less than alpha. The null hypothesis is accepted and it is concluded that the data is normally distributed.

Confidence Interval for 10k sample draw

```
# 95% confidence interval
CI <- c(qnorm(0.025, mean = 4.99887, sd = 0.05001534, lower.tail = TRUE), qnorm(0.025, mean = 4.99887, s

hist(emns10k, main = list('Average of 10k Draws', cex = 0.8, font = 1),xlab = 'Value', prob = TRUE,breal
lines(density(emns10k), col = 'green', lwd = 2)
abline(v = CI[1], col = 'purple')
abline(v = CI[2], col = 'purple')
legend(x = 5.15, y = 2.6,col = c('purple','green'), lty = 1, legend = c('CI','fit'), cex = .6)

mtext("Confidence Interval from 1000 Simulations from Means of 10k-Sample Draws from Exponential Distril
```



Confidence interval calculations assume a normal distribution, and the only case that meets this assumption is for the 10k sample draw. The two-tailed, 95% confidence interval around the mean was calculated using R's qnorm to find the quantiles. This is shown on the histogram. The interpretation is that at least 95% of the means sampled should fall within this interval. In this case, the proportion of the distribution contained within the interval is 94%.

Conclusions

- Individual samples from the exponential distribution are not normally distributed.

- Assuming each draw is independent and identically distributed, the means of a set of draws approach normal distribution with increasing number of draws or simulations.
- The 10k draw dataset was normally distributed upon examination of the similarity of the mean/median, bell shaped histogram, conformity of the quantile plot with a theoretical normal quantile plot, and passing the Shapiro-Wilk hypothesis test for normality. The 40-sample draw was more normal than the individual sample draw, but had deviations from normality at the tails and did not pass the Shapiro-Wilk test.
- A confidence interval was constructed and it was demonstrated that at least 95% of the means fell within this interval.