Exploring the CLT with the Exponential Distribution

jrfoster

Overview

The Central Limit Theorem (CLT) says that if independent samples of size `n` are repeatedly taken from any population, then when `n` is large the distribution of sample means will approach a normal distribution. In this report, we investigate the CLT using R to generate random independent samples of the Exponential Distribution. A brief investigation of the Exponential Distribution is given in the Appendix. Note that all R code used to produce the calculations and plots is included the Appendix but not displayed in the narrative for the sake of brevity.

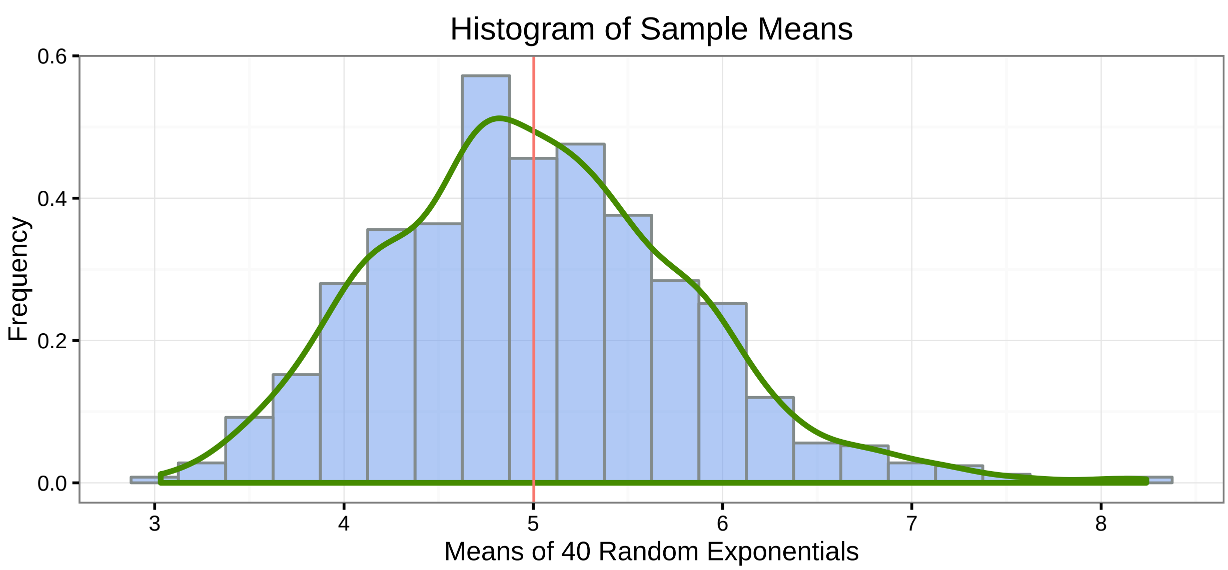
Simulations

For this report we will perform 1000 simulations of 40 random exponentials with $\lambda = .2$. To generate the simulation data, we use the `rexp` function to generate a single random exponential distribution, `replicate` to repeat the generation 1000 times and `t` to transpose the resulting matrix columns into rows. Finally, `data.frame` is used to create an appropriate structure for `ggplot`. We then use `dplyr:mutate()` to add a column for each distribution's mean and the fluctuation around the theoretical mean μ using the formula

$$\sqrt{n}(S_n - \mu)$$

Comparison of Theoretical and Sample Mean

The following plot shows the distribution of sample means along with its density curve. The distribution of the mean of our simulated 40 exponentials should resemble a normal distribution with mean $\frac{1}{\lambda}$. The red vertical line shows the sample distribution's mean.



To see how the sample mean compares with the theoretical mean of our distribution, we calculate both in the table below.

Sample Mean	Theoretical Mean
5.002795	5

As you can see, the sample distribution mean of 5.0028 is very close to the theoretical mean of 5. Note that the 95% confidence interval for the sample mean is [4.953, 5.0526].

Comparison of Theoretical and Sample Variance

Next we compare the variance of the 1000 sample means with the theoretical variance for our sample. The theoretical variance for our sample is given by the formula $\frac{\frac{1}{\lambda^2}}{n}$

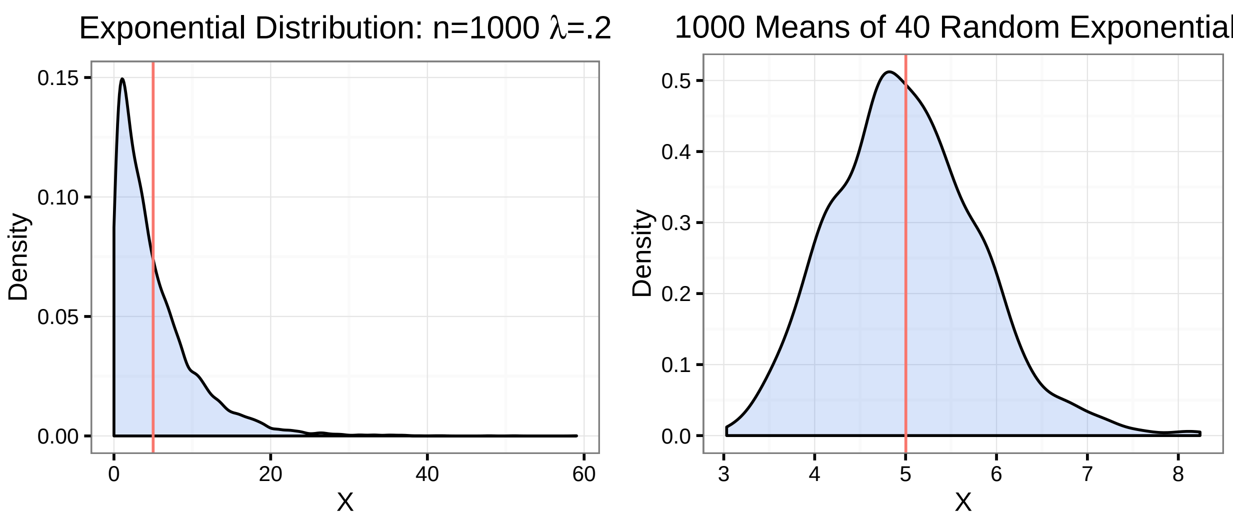
So, lets calculate the variance of our sample means and compare it with the theoretical value

Sample Variance	Theoretical Variance
0.6440784	0.625

As you can see, the sample distribution variance 0.6441 is very close to the theoretical variance of 0.625.

Sample Distribution vs. Normal Distribution

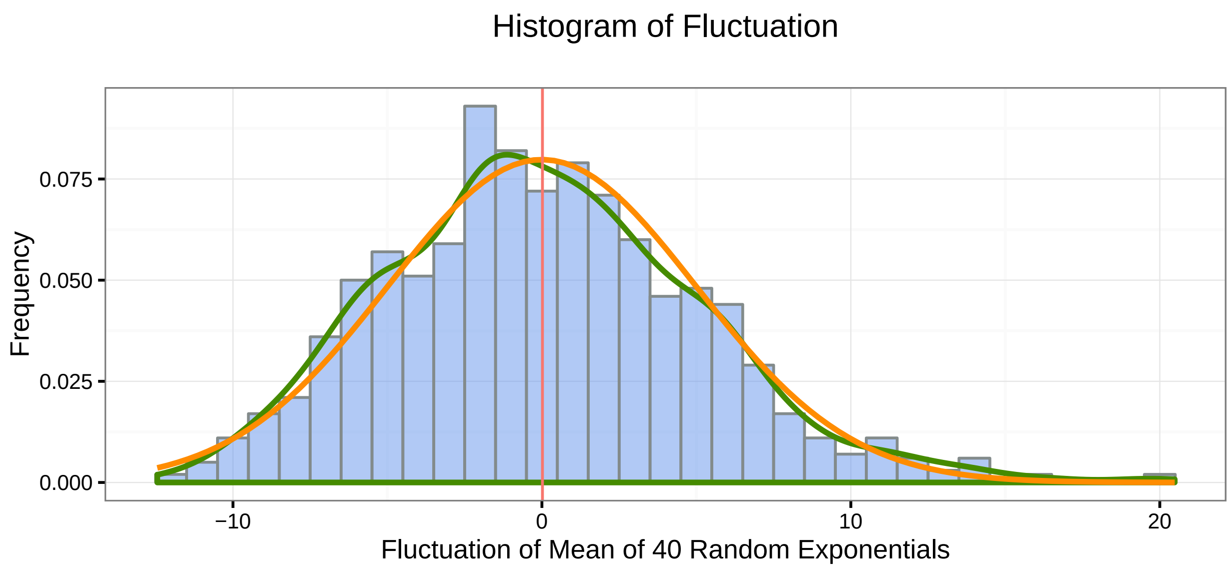
The great importance of the Normal Distribution is that the sum of a large number of independent random variables will be approximately normally distributed almost regardless of their individual distributions (Bulmer 1967). To demonstrate this, let us start by visually contrasting an exponential distribution with $n = 1000, \lambda = .2$ with the distribution sample we have been looking at already.



From this plot we can see the difference in the shape of the density functions, and that the Exponential Distribution doesn't resemble a Normal Distribution. Lets take a look at some further comparisons of this distribution of means as it relates to the normal distribution.

According to (Wikipedia, n.d.), Classical CLT states that as n gets larger, the distribution of the difference between the sample average S_n and its limit μ, when multiplied by the factor $\sqrt{n}$ approximates the normal distribution with mean 0 and variance σ^2.

The simulation data contains this random fluctuation in the variable `flux` so we can examine its distribution with a histogram of these fluctuations, along with its associated density curve. A normal distribution is also overlaid for comparison.

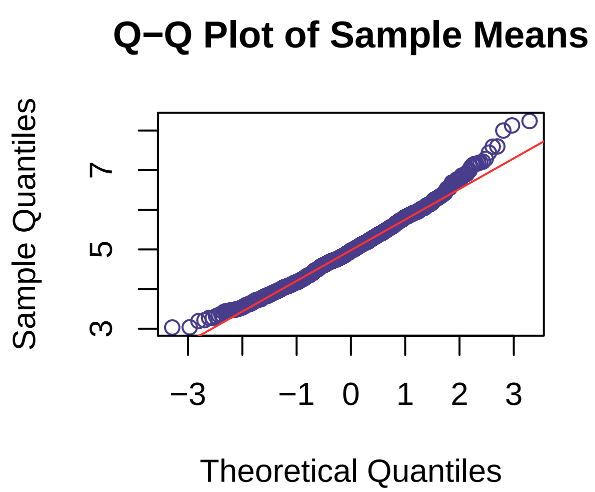


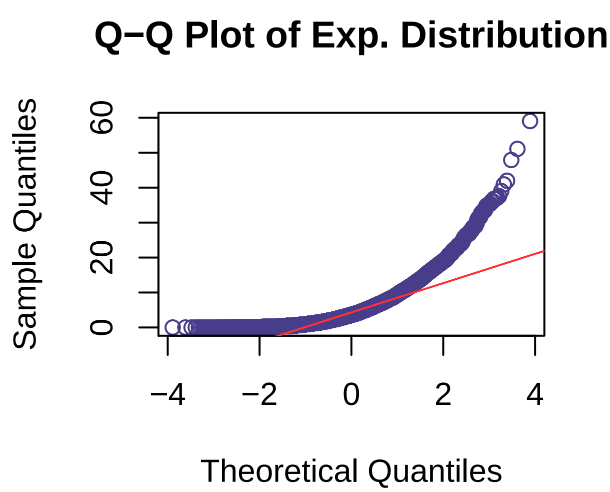
As expected, this histogram resembles a normal distribution. Lets take a look at the mean and variance of this distribution to see if it approaches what the CLT would assert.

Mean	Variance
0.0176757	25.76313

From this we can see that the CLT's assertion about the distribution being normal with mean 0 is true, since 0.0176757 is close to zero.

Another method of comparing this distribution of means with a normal distribution is by using a Quantile-Quantile Plot, or q-q plot. A q-q plot plots the quantiles of one dataset against the quantiles of another dataset and can be useful in determining if two datasets come from a population with a common distribution. In our case, instead of plotting with two different datasets, we can plot our sample distribution against the the normal distribution by using the `qqnorm` and `qqline` functions in R. By way of comparison, we also display a q-q plot of a large random exponential distribution with $n = 1000$ and $\lambda = .2$.





We can see from these q-q plot that the quantiles of our sample distribution follow fairly closely with the theoretical quantiles from the normal distribution, indicating that it approximates the normal distribution. By way of comparison, the q-q plot of the large random sample exponential distribution does not follow nearly as closely.

Appendix

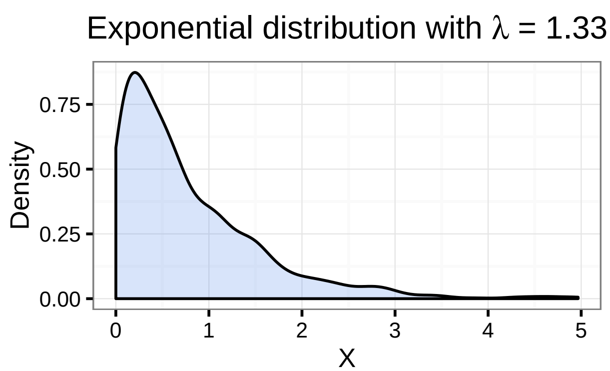
A Note on Reproducibility

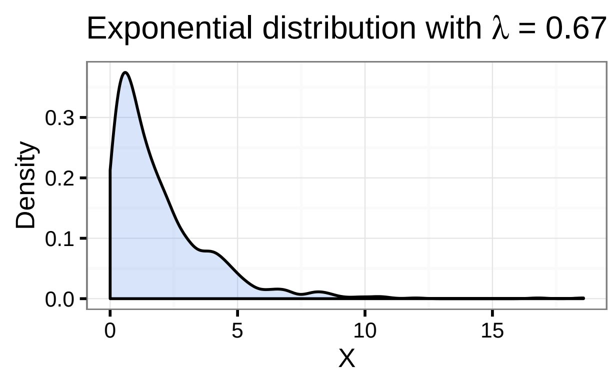
For the reviewer of this analysis: I have chosen to not use `set.seed` because I believe that this analysis could be run repeatedly, with different random samples, any number of times, and my assertions would still be valid.

Exponential Distribution

The exponential distribution is defined as $f(x) = \lambda e^{-\lambda x} \qquad x > 0, \lambda > 0.$

Two exponential distributions are plotted below with $n = 1000$. Note that while they both have the same basic shape, the shape does not resemble a normal distribution. Notice, also that the higher the λ the more likely it is that a random variable X will have a small value, which makes sense, given that the mean and standard deviation of the exponential distribution is $\frac{1}{\lambda}$ (adapted from (Gordon 2013))





Code Used in Report

This section contains all the code used to generate the diagrams.

Code to include libraries

```
suppressMessages(library(ggplot2))
suppressMessages(library(gridExtra))
suppressMessages(library(dplyr))
suppressMessages(library(knitr))
```

This code generates all the simulations used in the analysis:

```
lambda <- .2
n <- 40
numSims <- 1000
mu <- 1 / lambda
simulations <- data.frame(t(replicate(numSims, rexp(n, lambda)))) %>%
  mutate(xBar = rowMeans(.), flux = sqrt(n) * (xBar - mu))
```

This code generates the histogram of sample means:

```
ggplot(simulations, aes(x=xBar)) +
  ggtitle("Histogram of Sample Means") +
  labs(x = "Means of 40 Random Exponentials", y = "Frequency") +
  geom_histogram(aes(y = ..density..), col = "azure4", fill = "cornflowerblue",
                 alpha = .5, binwidth = .25) +
  geom_density(color = "chartreuse4", size = 1) +
```

```
  geom_vline(aes(xintercept=mean(xBar), color="firebrick"), show.legend=FALSE) +
  theme_bw(base_size = 10)
```

This code calculates and displays the theoretical and sample means and the confidence interval

```
sampleMean <- mean(simulations$xBar)
kable(data.frame(sampleMean, mu), col.names = c("Sample Mean","Theoretical Mean"))
confInt <- t.test(simulations$xBar)$conf.int
lowerCI <- confInt[1]
upperCI <- confInt[2]
```

This code calculates and displays the theoretical and sample variance

```
sigmaSq <- (1 / (lambda^2)) / n
sampleVariance <- var(simulations$xBar)
kable(data.frame(sampleVariance, sigmaSq),
      col.names = c("Sample Variance","Theoretical Variance"))
```

This code generates the comparison plot

```
bigSim <- rexp(10000, .2)
ex1 <- data.frame(val = bigSim)
ex1Plot <-  ggplot(ex1, aes(x=val)) +
  geom_density(col = "black", fill = "cornflowerblue", alpha = .25) +
  ggtitle(expression(paste("Exponential Distribution: n=1000 ", lambda, "=.2"))) +
  labs(x = "X", y = "Density") +
  geom_vline(aes(xintercept = 1/.2, color = "firebrick"), show.legend = FALSE) +
  theme_bw(base_size = 10)
sd1Plot <- ggplot(simulations, aes(x=xBar)) +
  geom_density(col = "black", fill = "cornflowerblue", alpha = .25) +
  ggtitle("1000 Means of 40 Random Exponentials") +
  labs(x = "X", y = "Density") +
  geom_vline(aes(xintercept=mean(xBar), color="firebrick"), show.legend=FALSE) +
  theme_bw(base_size = 10)
grid.arrange(ex1Plot, sd1Plot, nrow=1, ncol=2)
```

This code generates the histogram of fluctuations around the mean

```
ggplot(simulations, aes(x=flux)) +
  ggtitle("Histogram of Fluctuation\n") +
  labs(x = "Fluctuation of Mean of 40 Random Exponentials", y = "Frequency") +
  geom_histogram(aes(y = ..density..), col = "azure4", fill = "cornflowerblue",
               alpha = .5, binwidth = 1) +
  geom_density(color = "chartreuse4", size = 1) +
  stat_function(fun = dnorm, color = "darkorange", size = 1,
              args = list(mean = 0, sd = 5)) +
  geom_vline(aes(xintercept=mean(flux), color="firebrick"), show.legend=FALSE) +
  theme_bw(base_size = 10)
```

This code calculates the mean and standard deviation of the fluctuation around the mean

```
fluxMean <- mean(simulations$flux)
fluxVar <- var(simulations$flux)
kable(data.frame(fluxMean, fluxVar), col.names=c("Mean", "Variance"))
```

This code creates a large exponential distribution and displays the q-q plots comparing it with the sample

```
par(mfrow=c(1,2))
qqnorm(y = simulations$xBar, col = "darkslateblue", main="Q-Q Plot of Sample Means")
qqline(y = simulations$xBar, col = "firebrick1")
qqnorm(y = bigSim, col = "darkslateblue", main="Q-Q Plot of Exp. Distribution")
```

```
qqline(y = bigSim, col = "firebrick1")
```

Code to produce the two exponential distributions used in the appendix

```
ex1 <- data.frame(val = rexp(1000, 1.33))
ex1Plot <-  ggplot(ex1, aes(x=val)) +
  geom_density(col = "black", fill = "cornflowerblue", alpha = .25) +
  ggtitle(expression(paste("Exponential distribution with ", lambda, " = 1.33"))) +
  labs(x = "X", y = "Density") +
  theme_bw(base_size = 10)
ex2 <- data.frame(val = rexp(1000, .5))
ex2Plot <-  ggplot(ex2, aes(x=val)) +
  geom_density(col = "black", fill = "cornflowerblue", alpha = .25) +
  ggtitle(expression(paste("Exponential distribution with ", lambda, " = 0.67"))) +
  labs(x = "X", y = "Density") +
  theme_bw(base_size = 10)
grid.arrange(ex1Plot, ex2Plot, nrow=1, ncol=2)
```

References

Bulmer, M.G. 1967. *Principles of Statistics*. New York: Dover Publications, Inc.

Gordon, Ian. 2013. “Exponential and Normal Distributions.” http://amsi.org.au/ESA_Senior_Years/PDF/ExpoNormDist4f.pdf.

Wikipedia. n.d. “Central Limit Theorem.” https://en.wikipedia.org/wiki/Central_limit_theorem‘.