Investigation of Central Limit Theorem

Kevin Tham

April 25, 2018

```
if (!require("pacman"))
  install.packages("pacman", repos = "http://cran.us.r-project.org")
pacman::p_load(knitr, dplyr, ggplot2, tidyr, grid, gridExtra)
```

Overview

In this exercise we investigate the Central Limit Theorem by simulating samples drawn from an exponential distribution. We expect that the means of a large number of samples drawn from identical exponential distributions will themselves approach a normal distribution with known mean and standard deviation.

Theory

The Central Limit Theorem (CLT) considers a sequence of random variables $X_1, X_2, ..., X_n$ drawn from a common distribution F with mean μ and variance σ^2. The statement is that as $n \to \infty$, the mean

$$S_n = \frac{X_1 + X_2 + ... + X_n}{n},$$

which is a random variable itself, converges in distribution to a normal distribution with the same mean μ as the original distribution F:

$$\bar{S_n} = \mu$$

The variance of this normal distribution can be obtained by considering the following:

$$\begin{aligned} \text{Var}(S_n) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(X_i\right) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Simulation Exercise

In this section, we will set out to investigate the CLT via simulating a distribution of averages sampled from exponential distributions. The means and variances of the sampled distribution will be compared with the expected theoretical mean and variance. Finally we will compare the distribution with the appropriate normal distribution it is expected to converge to by the CLT.

To begin with, set the seed to ensure that the following analysis is reproducible:

```
set.seed(123)
```

The probability density function of an exponential distribution is given by the following:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

It is defined by a single parameter, λ, otherwise known as the rate. It has a known mean of λ^{-1} and variance of λ^{-2}.

We prepare a thousand repetitions of $n = 40$ samples drawn from an exponential distribution of rate $\lambda = 0.2$.

```
lambda <- 0.2
n <- 40
rep <- 1000
rawdat <- rexp(40000, rate = lambda)
matdat <- matrix(rawdat, rep, n)
```

Next, we can calculate the theoretical mean and variance of the sample means.

```
th_mean <- 1/lambda
th_var <- (1/lambda)^2/n
th_sd <- sqrt(th_var)
```

From our simulated data we can obtain the sample mean:

```
means <- apply(matdat, 1, mean)
sample_mean <- mean(means)
variance <- var(means)

results <- data.frame('Simulated' = c(sample_mean, variance),
                      'Theoretical' = c(th_mean, th_var),
                      row.names = c('Mean', 'Variance'))

kable(x = signif(results, 3))
```

	Simulated	Theoretical
Mean	5.010	5.000
Variance	0.609	0.625

We can see from the above table that the simulated and theoretical values of the mean and variance of the sample means closely match.

Next, we can compare the distribution of the sample means with an actual normal distribution with the stated theoretical mean and variance by plotting a histogram of the distribution of the sample means:

```
df_means <- data.frame('means'= means)

ggplot(df_means, aes(means)) + geom_histogram(aes(y=..density..), binwidth=0.3) +
  stat_function(fun=function(means) dnorm(means, mean=th_mean, sd=th_sd)) +
  labs(x = 'Means', y = 'Density')
```

We can see from the Fig. 1 that the distribution of sample means closely resembles that of the expected normal distribution. There is deviation from the expected distribution as it can be observed that the distribution of sample means is slightly skewed to the left. Therefore we can conclude that the means of 40 exponentials behave as predicted by the Central Limit Theorem.

Finally, we can compare the distribution of the sample means with the distribution of the samples themselves.

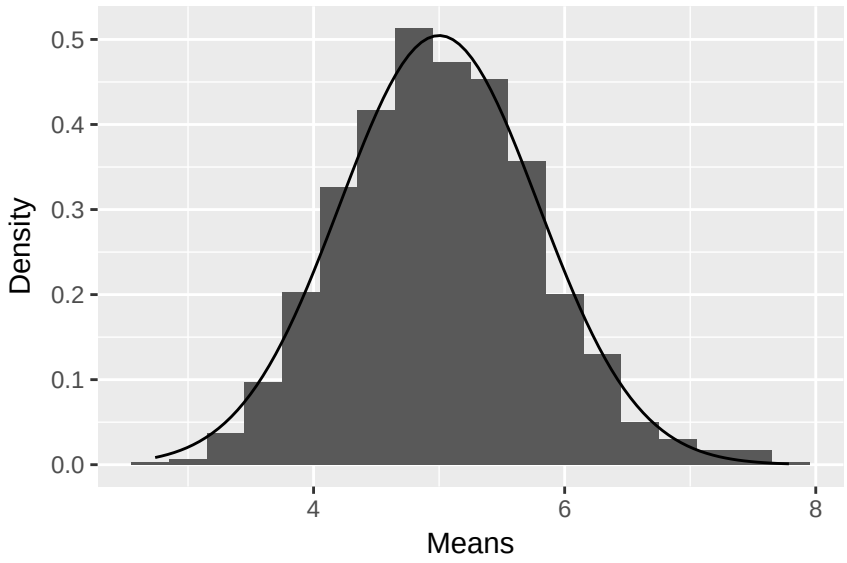

Figure 1: Scaled Histogram of means drawn from exponential distribution. Black curve represents CLT normal distribution for comparison

```
ggplot(data.frame(X = rawdat), aes(X)) + geom_histogram(aes(y=..density..), binwidth=1) +
  stat_function(fun=dexp, args=list(rate=lambda)) + labs(x = 'X', y = 'Density')
```

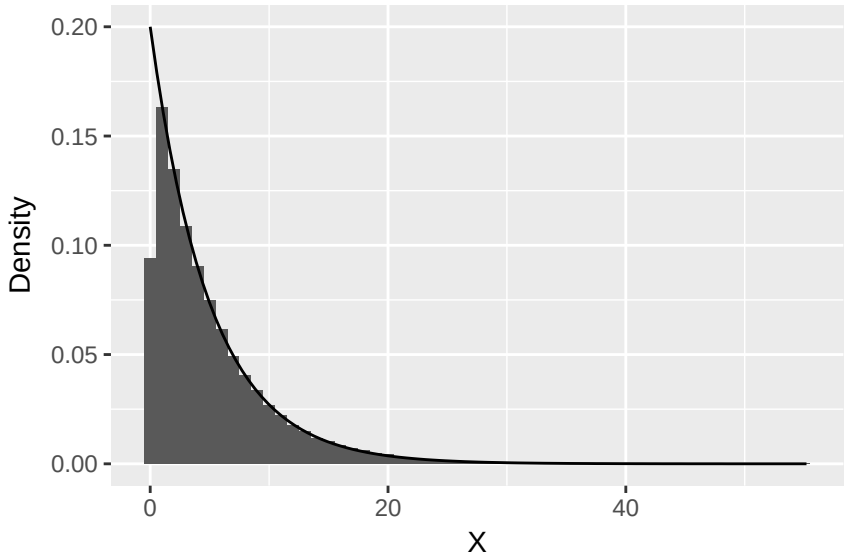

Figure 2: Scaled Histogram of samples drawn from exponential distribution

We can see from Fig. 2 that the original samples, being drawn from an exponential distribution, have a vastly different distribution from the sample means.