Part 1: Exponential Distributions and the Central Limit Theorem

Shayonendra Nath Tagore

25 May, 2020

1 Introduction

This assignment will explore whether the Central Limit Theory (CLT) can be applied to exponential distributions. The CLT states that the distribution of averages for Independent and Identically Distributed (iid) variables will approximate a normal distribution as sample size increases.

This will be explored through three steps:

1. Properties of a single exponential distribution
2. Distribution of Averages for 1000 Exponential Distributions
3. Comparison of Exponential and Normal Distributions

2 Properties of a Single Exponential Distribution

Due to the random selection of values, each sample is only representative of the population. This representation increases in accuracy as the number of values (n) increases, ultimately approaching μ. As such, we will make n = 1000. Exponential Distributions are modeled around λ, with the mean $\mu = \frac{1}{\lambda}$. Lets assume $\lambda = 1/5(0.2)$. This sample will looks as such:

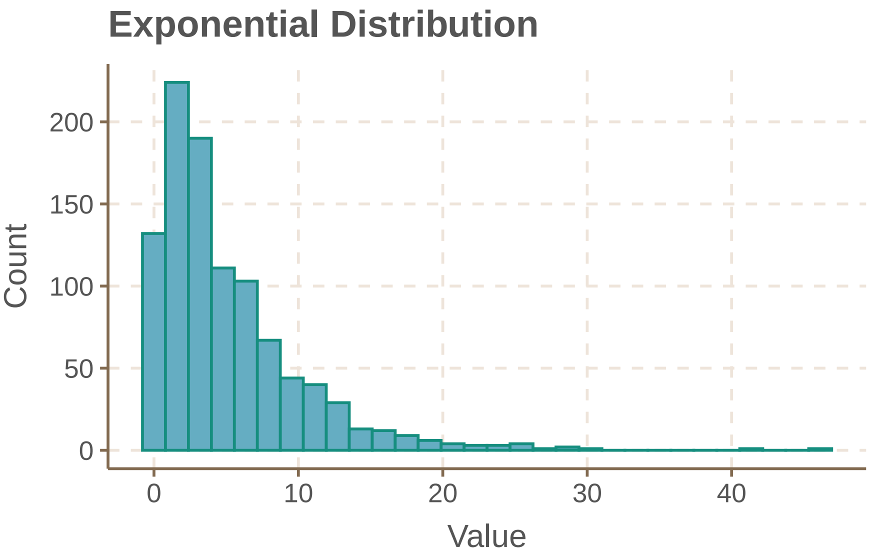

Figure 1: Histogram representing the spread of a single exponential distribution sample with 1000 values.

A necessity for CLT distributions is that it follows the Law of Large Numbers. As n approaches infinity, the sample mean will become asymptotic along μ. This asserts that the each sample results follow a predictable pattern. This can be assessed by observing changes in the average as n_i increases. The Expected Value (EV) for this sample is:

$$E[N] = \frac{1}{\lambda} = \frac{1}{0.2} = 5$$

This single distribution has $E[N] = 5$. This is confirmed through Fig.2 below. As N_i increases, the average approaches 5.

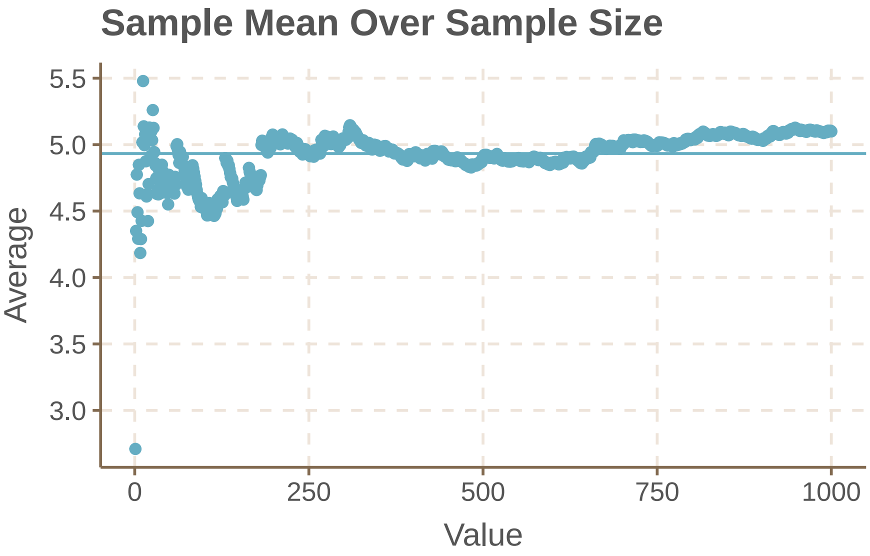

Figure 2: Change in sample mean as sample size increases in an exponential distribution with 1000 values. As $\lambda = 1/5$, the population mean is 5. The blue line is the average of the 1000 values, and as the number of values increases, the sample mean approximates the population mean of 5.

3 Distribution of Averages for 1000 Simulations

To better understand how increasing n leads to approximation of the population, we will examine a large number of samples with 40 values each. Due to random selection, each sample alone will not be an accurate representation of the population. If exponential distributions follow the CLT, the mean of each sample will create a distribution of averages which in turn represent the population.

This section will explore many exponential distributions using the following properties: - $\lambda = 0.2$ - n = 40 per sample - 1000 samples

Plotting the distribution of averages gives this:

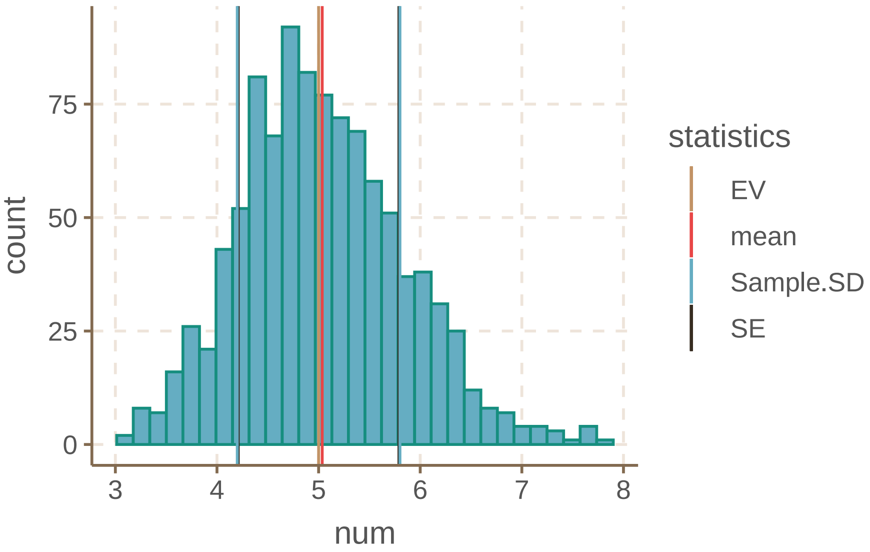

Figure 3: Distribution of averages for 1000 exponential distributions. The sample SD and SE are almost overlapping.

Plotting the distribution of averages gives the figure above. Visually, it is not hard to imagine the shape of normal curve. It is easier than imaging the bell curve on the single distribution represented in Figure 5, an advantage given through use of distribution of averages. However, further statistical analysis is required to confirm that distribution of averages simulates a normal distribution. As with the single expontential distribution above, the $E[N] = 5$ as $\lambda = 0.2$. While the average of the distributions do not equal 5 but 5.0353567, differences can be explained through Standard Error (SE). SE is calculated using:

$$SE = \frac{\sigma}{\sqrt{n}} \approx 0.79$$

As such, the EV of 5 falls in the range of 5.04 ± 0.79. The SD of the distribution of averages is 0.8004739 (population SD is 5, using $\sigma = \sqrt{\frac{1}{\lambda^2}}$). The sample SD appears to be a strong predictor for the population SD. Further confirmation can be found through hypothesis testing.

Since the distribution of means approximates the population mean, the last step is to check if exponential distributions follows the CLT and becomes normally distributed. Figure 4 suggests that the distribution of averages follows the normal distribution. This is done by assessing whether the SD of the distribution of averages approximates the SD of a normal distribution. Using the null hypothesis that distribution of averages represents a normal distribution, with mean μ and standard deviation σ and hypothesis that it does not represent a normal distribution. The calculated P-value is 0.96. With an $\alpha = 0.05$, we fail to reject the null. Thus, we conclude that exponential distributions follow the CLT.

4 Appendix

```r
### loading all packages
pacman::p_load(knitr,
               dplyr,
               ggplot2,
               ggthemr)
ggthemr::ggthemr("fresh")

# for greater stability while exploring where no one
# (except my peers) has gone before
set.seed(1702)

### generating a single exponential distribution
sample_num <- 1000
sample <- rexp(sample_num, rate = 0.2)

### plotting a single exponential distribution
sample %>% data.frame(num = .) %>%
    ggplot(aes(x = num)) +
    geom_histogram(bins = 30, col = ggthemr::swatch()[8]) +
    xlab("Value") + ylab("Count") + ggtitle("Exponential Distribution")
means <- cumsum(sample)/(1:sample_num)

### plotting to see if exponential distributions follow the law of large numbers
means <- cumsum(sample)/(1:sample_num)
asymp_point <- round(mean(means), digits=1)
means %>%
    data.frame(x=1:sample_num, y = .) %>%
    ggplot(aes(x=x, y=y)) +
    geom_point() +
    geom_hline(yintercept = mean(means)) +
    xlab("Value") + ylab("Average") + ggtitle("Sample Mean Over Sample Size")

### Plotting the distribution of averages for 1000 exponential distributions.
sample_averaging_num  <- 40
sim_num <- 1000
mns = NULL
for (i in 1:sim_num) mns = c(mns, mean(rexp(sample_averaging_num, rate = 0.2)))
mns_SE <- (1/0.2)/sqrt(sample_averaging_num) # the Standard Error
mns %>%  data.frame(num = .)  %>%
ggplot(aes(x = num)) +
    geom_histogram(col = ggthemr::swatch()[8], bins = 30) +
    geom_vline(aes(xintercept = mean(mns), color = "mean")) +
    geom_vline(aes(xintercept = 5, color = "EV")) +
    geom_vline(aes(xintercept = 5+mns_SE, color = "SE")) +
    geom_vline(aes(xintercept = 5-mns_SE, color = "SE")) +
    geom_vline(aes(xintercept = sd(mns)+5, color = "Sample.SD")) +
    geom_vline(aes(xintercept = -sd(mns)+5, color = "Sample.SD")) +
    scale_color_manual(name = "statistics", values = c(mean = ggthemr::swatch()[4],
                                                       EV = ggthemr::swatch()[5],
                                                       SE = ggthemr::swatch()[6],
                                                       Sample.SD = ggthemr::swatch()[2])) +
    theme(legend.position = "right")
```

```r
### Hypothesis testing on the distribution of averages for exponential distribution
z_stats <- abs((mean(mns)-5)/mns_SE)
p_val <- 2*pnorm(z_stats, lower.tail = FALSE)
```