Statistical Inference Project Part 1: Comparing Simulated Data to the Central Limit Theorem

Ryan Hammer

Project Overview

Welcome to Part 1 of the course project in the Statistical Inference course that is part of the Data Science Specialization through Johns Hopkins University in Coursera. For this portion of the project, a simulation exercise will be performed in R to investigate the exponential probability distribution and compare it to the Central Limit Theorem. To do said comparison, we will view the simulated sample mean, variance, and distribution with the theoretical versions of those same three characteristics for a given lambda value.

1. Simulations

To begin, we will need to run 1000 simulations of 40 exponential random variables. The assignment requires using a lambda (λ) value of 0.2 for the randomized data, and then calculating the mean for each sample of size 40. The distribution of those sample means will then be used in the comparison with the Central Limit Theorem. It should be noted that the exponential distribution is related to the geometric distribution, which is a measure of the number of binomial trials required to achieve a single success. The exponential distribution is the continuous version of the discrete geometric distribution. For the simulation in R, I use the rexp() function to generate exponential random variables. I am also employing the set.seed() function to make my results reproducible. Then, I use a for() loop to create a matrix storing each simulation of 40 random exponentials to a single row for a total of 1000 rows. The code for that, and all other R commands, can be found in the Appendix.

Here is a view of the first five rows and columns of the resulting simulated matrix:

0.037520	0.3254638	5.5377835	1.2032695	5.3729986
3.927970	5.5592170	7.8087339	0.1954609	0.5293239
1.443775	10.4110857	2.5677658	1.4573230	2.0893345
5.064839	5.4874516	0.2667949	0.8697219	1.6198950
4.257850	2.6675716	0.0585745	3.1634984	1.0232367

Next, I take the mean and variance of each row of the matrix and store those values in a vector called "simmeans." These represent means for each sample of 40 simulated random variables. The distribution of the means will become the focus of the analysis and comparison.

2. Sample Mean vs. Theoretical Mean

The first question we are exploring is "How does the sample mean compare with the theoretical mean predicted by the Central Limit Theorem?" The theoretical mean for an exponential distribution is equal to the value of $\frac{1}{\lambda}$. As we are using a λ value of 0.2, our theoretical mean, or μ, is equal to 5. The sample mean is calculated by applying the mean() function to the vector "simmeans", which contains the means for each simulation of 40 random exponential variables. That process yeilds a resulting mean of

```
## [1] 4.988198
```

The sample mean of 4.988198 is extremely close to the theoretical mean of $\mu = 5$. This should not be surprising, as the Central Limit Theorem states that for a large number of samples, the distribution of sample means should be approximately normal with center at the population (or in this case, theoretical) mean. The histogram in Figure 1 shows the distribution of sample means. The orange line represents the theoretical

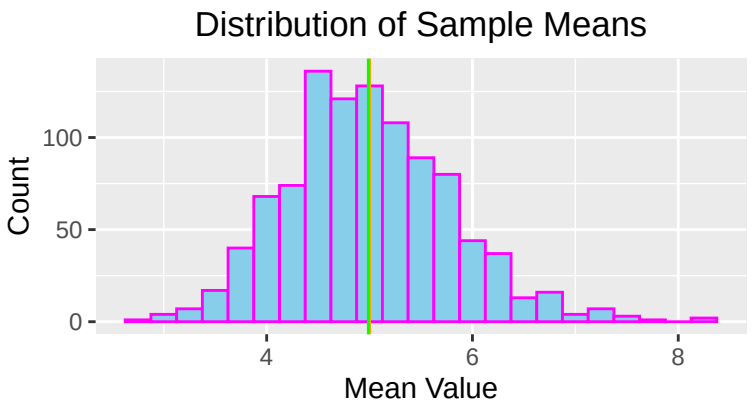

Figure 1: Sample Mean vs. Theoretical Mean

mean and the green line represents the sample mean. The lines are virtually on top of one another, further reinforcing the applicability of the Central Limit Theorem to the distribution of the sample means.

3. Sample Variance vs. Theoretical Variance

We can similarly compare the variance of the sample means with the theoretical variance predicted by the Central Limit Theorem. The theoretical variance of the samples would be calculated using the theoretical variance of the exponential distribution, $\frac{1}{\lambda^2}$, and the size of each sample, n = 40. The theoretical sample variance is equal to $\frac{\frac{1}{\lambda^2}}{n}$. I can find the sample variance using R's var() function to calculate the variance of my sample means. The results of both calcuations are below.

```
## [1] The theoretical variance is 0.625.
## [2] The sample variance is 0.638111306386005.
```

These values are nearly identical! Again, the validity of applying the Central Limit Theorem is confirmed by the results of our simulated data and subsequent calculations. For a visual confirmation of this idea, we can return to the same distribution of sample means from before. Figure 2 has the values of one sample standard deviation from the sample mean and one theoretical standard deviation from the sample mean overlayed to further demonstrate how close these values are to one another. Again, the sample values are in orange, and the theoretical in green. The standard deviation, or the standard error of sampling distributions, is calculated by taking the square root of the variance.

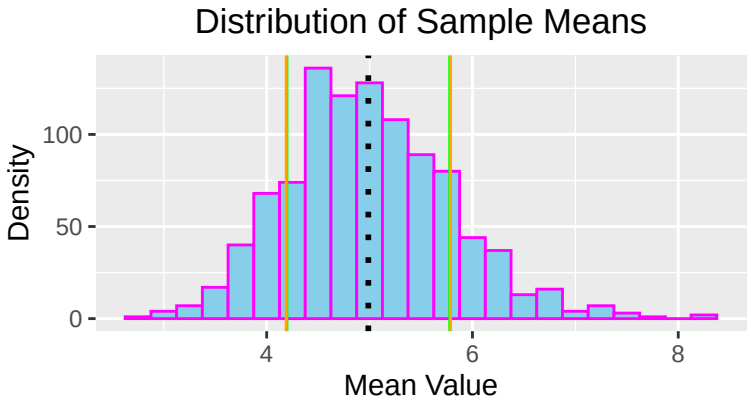

Figure 2: Sample Variance vs. Theoretical Variance

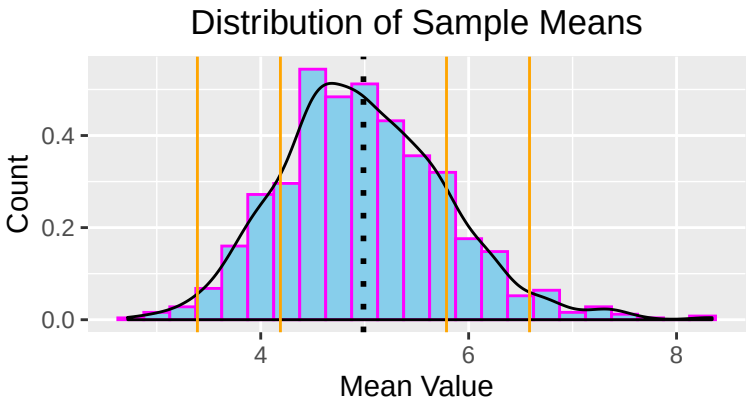

Figure 3: Normality of Distribution

4. Distribution

The final question we would like to answer is "Is the distribution of sample means approximately normal?" To do this we will focus on the location of specific quantiles in the data and the appearance of the distribution. Normal distributions are symmetric about their median, with half the data falling on either side. This leads to the mean being equal to the median. Additionally, approximately 68% of the data should fall within one standard deviation of the mean/median, and about 95% of the data should fall within two standard deviations of the mean. We have already caluclated a sample mean of 4.988. The sample standard deviation, as mentioned above, is equal to the square root of the sample variance, which for the sample data here is $\sqrt{0.6381} = 0.7988187$ Now, we can use the quantile function to view the cutoff points for the median of the data, and the points corresponding to the middle 68% and middle 95% of the data. These values are shown in the first row of the table below. The second row represents the cutoff points within one and two standard deviations of the mean. The mean is the middle value, with the two values before it corresponding to two standard deviations below the mean and the latter two values being those above the mean.

	2.5th Percentil	16th Percentile	Median	84th Percentile	97.5th Percentile
Sample Quantiles	3.5774	4.2324	4.9358	5.7519	6.7085
Standard Dev. Cutoffs	3.3906	4.1894	4.9882	5.7870	6.5858

The values are extremely close to one another, demonstrating a high level of normality in the distribution of the sample means. Again this matches what we would expect from the Central Limit Theorem, which predicts normality of a distribution of sample means for large enough sample sizes. Further evidence of this is seen in Figure 3, which uses a density curve overlaid on the histogram of the sample mean distribution. The vertical axis here has been changed to reflect density of the distribution rather than a simple count; however, this did not change the shape of the distribution. Notice the 'bell-like' shape of the density curve, matching what we would expect to see in normally distributed data.

Conclusion

Though exponentially distributed data are not normally distributed, the Central Limit Theorem states that means of samples taken from data matching an exponential distribution would be approximately normal, assuming large enough sample sizes. Using a simulation of 1000 samples of 40 exponential random variables, we have seen that the mean, variance, and quantile distribution of those samples are extremely close to the corresponding theoretical values predicted by the Central Limit Theorem, seeming to strongly confirm its applicability.

Appendix A: R code used in analysis

```
suppressWarnings(library(ggplot2))
suppressWarnings(library(knitr))
set.seed(35)
simdata <- matrix(data = NA, nrow = 1000, ncol = 40)
for (i in 1:1000) {simdata[i, ] <- rexp(40, 0.2)}
kable(simdata[1:5, 1:5])
simmeans <- apply(simdata, 1, mean)
samplemean <- mean(simmeans)
print(samplemean)
qplot(simmeans, geom = "histogram", binwidth = 0.25, xlab = "Mean Value",
      main = "Distribution of Sample Means", ylab = "Count", fill = I("skyblue"),
      col = I("magenta")) +
    geom_vline(aes(xintercept = 5), col= "orange", size = 0.5) +
    geom_vline(aes(xintercept = mean(simmeans)), col= "green", size = 0.5) +
    theme(plot.title = element_text(hjust = 0.5))
theovar <- (1/0.2^2)/40
samplevar <- var(simmeans)
varsent1 <- paste("The theoretical variance is ", theovar, ".", sep = "")
varsent2 <- paste("The sample variance is ", samplevar, ".", sep = "")
noquote(c(varsent1, varsent2))
theosigma <- sqrt(theovar)
samplesigma <- sqrt(samplevar)
qplot(simmeans, geom = "histogram", binwidth = 0.25, xlab = "Mean Value",
      main = "Distribution of Sample Means", ylab = "Density",
      fill = I("skyblue"), col = I("magenta")) +
    geom_vline(aes(xintercept = samplemean+c(-1, 1)*theosigma),
               col= "green", size = 0.5) +
    geom_vline(aes(xintercept = samplemean+c(-1, 1)*samplesigma),
               col= "orange", size = 0.5) +
    geom_vline(aes(xintercept = mean(simmeans)), col= "black",
               size = 1, linetype = "dotted") +
    theme(plot.title = element_text(hjust = 0.5))
cat(samplesigma)
quant1 <- quantile(simmeans, c(0.025, 0.16, 0.5, 0.84, 0.975))
quant1 <- round(quant1, 4)
quantlabels <- c("2.5th Percentil", "16th Percentile", "Median",
                 "84th Percentile", "97.5th Percentile")
quant2 <- c(-2*samplesigma+samplemean, -1*samplesigma+samplemean,
            samplemean, 1*samplesigma+samplemean,
            2*samplesigma+samplemean)
quant2 <- round(quant2, 4)
quantdf <- as.data.frame(rbind(quant1, quant2))
colnames(quantdf) <- quantlabels
rownames(quantdf) <- c("Sample Quantiles", "Standard Dev. Cutoffs")
kable(quantdf)
sigmavec <- c(-2, -1, 1, 2)*samplesigma
qplot(simmeans, y = ..density.., geom = "histogram", binwidth = 0.25,
      xlab = "Mean Value", main = "Distribution of Sample Means",
      ylab = "Count", fill = I("skyblue"), col = I("magenta")) +
    geom_density(aes(x = simmeans, y=..density..)) +
    geom_vline(aes(xintercept = samplemean+sigmavec), col= "orange",
```

```
            size = 0.5) +
  geom_vline(aes(xintercept = mean(simmeans)), col= "black",
            size = 1, linetype = "dotted") +
  theme(plot.title = element_text(hjust = 0.5))
```