Exploration of Behavior of Sample Means for Exponentially Distributed Data

Alex Hudson

2019-01-23

Setup

```
library(ggplot2)
library(grid)
library(gridExtra)
library(knitr)
library(kableExtra)
```

Overview

For this analysis, we will explore the behavior of sample means when the population follows an exponential distribution. We will use simulation to generate sample means from the exponential distribution, and we will compare the results to those mathematically expected by way of algebra of random variables and the Central Limit Theorem.

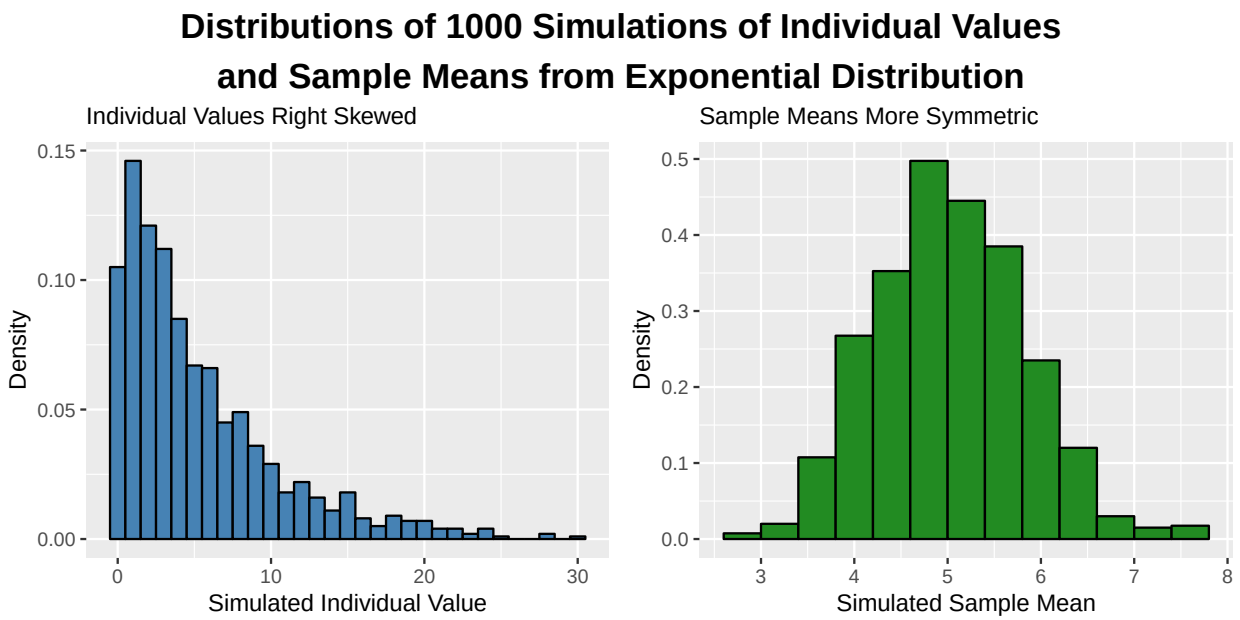
Simulations

The exponential distribution is characterized by a single parameter, λ, which is also called the rate parameter. Both the mean and standard deviation of the distribution are $1/\lambda$. For our purposes, we'll be using a value of 0.2 for λ, we'll be generating 1000 simulations, our sample means will use a sample size of 40, and we'll be setting a random number generation seed so that our results are reproducible. For comparison, we'll first simulate 1000 individual values from the exponential distribution, and then we'll simulate 1000 sample means of size 40 by simulating 40,000 individual values, organizing them into a matrix with 1000 rows and 40 columns, and taking row-wise means using the `apply()` function in R.

```
set.seed(37)
n <- 40
numsims <- 1000
lambda <- 0.2

xe <- rexp(numsims, lambda)
xem <- apply(matrix(rexp(n * numsims, lambda), nrow = numsims, ncol = n),
             1, mean)
```

Before we proceed to check results against known results, let's have a look at the simulated data. Below, we first plot a histogram of the simulated individual values, then we plot a histogram of the simulated sample means. The individual values are heavily right skewed with a range between 0 and 30. Meanwhile, the sample means are much closer to symmetric (though they are slightly right skewed) with a range between 2.5 and 8.



Empirical Mean versus Theoretical Mean

Using algebra of random variables, the mean of the distribution of sample means is expected to be the population mean. As we mentioned earlier, the mean of the exponential distribution with rate parameter λ is $1/\lambda$. Since we set λ to be 0.2 for our simulations, the associated distribution should have a population mean of 5. Thus, we expect the mean of the simulated sample means to be near 5. In our simulation, the mean of the simulated means is about 5.03, which is within 0.03 of the expectation.

Theoretical and Empirical Means for Simulated Sample Means

Theoretical Mean	Empirical Mean	Difference
5	5.03	0.03

Empirical Variance versus Theoretical Variance

Using algebra of random variables, the variance of the distribution of sample means is expected to be the population variance divided by the sample size. As we mentioned earlier, the standard deviation of the exponential distribution with rate parameter λ is $1/\lambda$, and so its variance is $1/\lambda^2$. Since we set λ to be 0.2 for our simulations, the associated distribution should have a population variance of 25. Our simulations used a sample size of 40. Thus, we expect the variance of the simulated sample means to be near $25/40 = 0.625$. In our simulation, the variance of the simulated means is about 0.633, which is within 0.008 of the expectation.

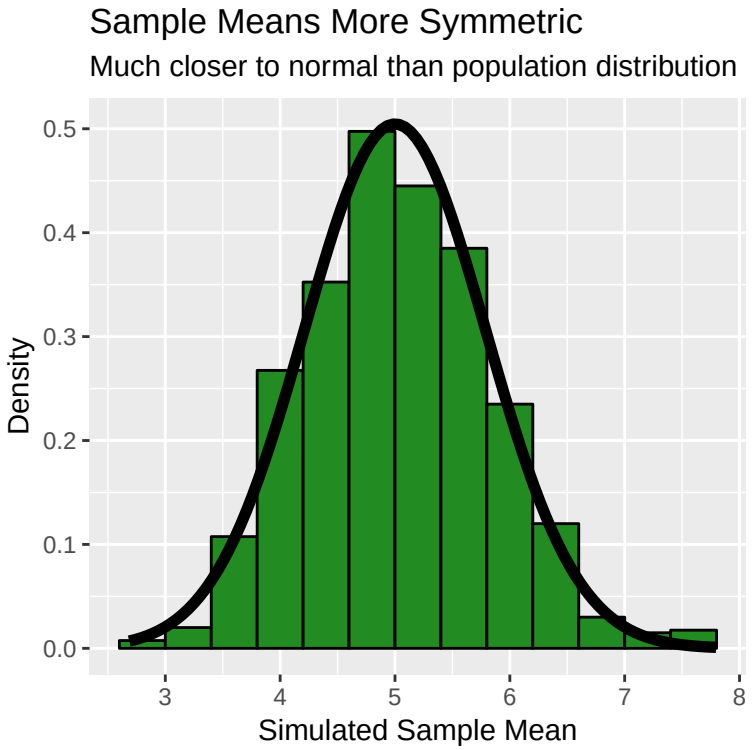
Theoretical and Empirical Variances for Simulated Sample Means

Theoretical Variance	Empirical Variance	Difference
0.625	0.633	0.008

Distribution

According to the Central Limit Theorem, the distribution of the sample mean becomes normal as the sample size increases. While there are no strict guidelines with regard to how large the sample size must be, our simulated sample size of 40 should be enough to at least show that the distribution is beginning to approach normality. In the following plot, we've overlaid a normal density with the theoretical mean and variance

on the histogram of simulated sample means. From this plot, we can see that the histogram is much more symmetric and bell-shaped than the original plot of individual values, though the histogram does retain some slight right skew. While the histogram and the normal density curve do not agree exactly, it does appear that the distribution of sample means is approaching normality as stipulated by the Central Limit Theorem.



Conclusion

Between algebra of random variables and the Central Limit Theorem, the distribution of sample means, for which individual observations are independent and identically distributed, has a mean equal to the population mean, it has a variance equal to the population variance divided by the sample size, and its distribution becomes normal as the sample size increases. Interestingly, this places no requirement on the actual population distribution. We've seen in the previous sections that these properties of the distribution of sample means appear to hold true through simulation methods, even with a population distribution such as the exponential distribution, which is very much unlike the normal distribution. The results appear to hold both quantitatively, as we compared the theoretical and empirical means and variances of the distribution of sample means, and qualitatively, as we compared the histogram of simulated sample means to the theoretical distribution of sample means.

Appendix

What follows is the code required to produce each of the figures seen in the report.

The following code chunk produces the side-by-side histograms of the simulated individual observations and the simulated sample means.

```
ge1 <- ggplot(data = data.frame(x = xe)) +
    geom_histogram(mapping = aes(x = x, y = ..density..),
                   binwidth = 1, color = "black", fill = "steelblue") +
    labs(x = "Simulated Individual Value", y = "Density",
         subtitle = "Individual Values Right Skewed")
ge2 <- ggplot(data = data.frame(x = xem)) +
    geom_histogram(mapping = aes(x = x, y = ..density..),
                   binwidth = 0.4, color = "black", fill = "forestgreen") +
    labs(x = "Simulated Sample Mean", y = "Density",
         subtitle = "Sample Means More Symmetric")
grid.arrange(ge1, ge2, ncol = 2,
             top = textGrob(paste("Distributions of 1000 Simulations of",
                                  " Individual Values\nand Sample Means",
                                  " from Exponential Distribution",
                                  sep = "", collapse = ""),
                            gp = gpar(fontsize = 16, font = 2)))
```

The following code chunk produces the first table, which compares the theoretical and empirical means for the simulated sample means. Note that this requires the `knitr` chunk option `results = "asis"` to be rendered as seen above.

```
kab1 <- kable(data.frame(x = 1/lambda,
                         y = mean(xem),
                         z = mean(xem) - 1/lambda),
              digits = 2,
              col.names = c("Theoretical Mean",
                            "Empirical Mean",
                            "Difference"),
              caption = paste("Theoretical and Empirical Means",
                              " for Simulated Sample Means",
                              sep = "", collapse = ""))
kable_styling(kab1, latex_options = "hold_position")
```

The following code chunk produces the second table, which compares the theoretical and empirical variances for the simulated sample means. Note that this requires the `knitr` chunk option `results = "asis"` to be rendered as seen above.

```
kab2 <- kable(data.frame(x = ((1/lambda)^2)/n,
                         y = var(xem),
                         z = var(xem) - ((1/lambda)^2)/n),
              digits = 3,
              col.names = c("Theoretical Variance",
                            "Empirical Variance",
                            "Difference"),
              caption = paste("Theoretical and Empirical Variances",
                              " for Simulated Sample Means",
                              sep = "", collapse = ""))
kable_styling(kab2, latex_options = "hold_position")
```

The following code chunk produces the histogram of the sample means with the normal distribution overlaid

on it.

```
ggplot(data = data.frame(x = xem)) +
    geom_histogram(mapping = aes(x = x, y = ..density..),
                   binwidth = 0.4, color = "black", fill = "forestgreen") +
    stat_function(fun = dnorm,
                  args = list(mean = 1/lambda, sd = 1/(lambda*sqrt(n))),
                  size = 2) +
    labs(x = "Simulated Sample Mean", y = "Density",
         title = "Sample Means More Symmetric",
         subtitle = "Much closer to normal than population distribution")
```