Sampling from the Exponential Distribution: Exploring the Sampling Distribution when n = 40

Brian Waismeyer

Sunday, May 24, 2015

Overview

The Central Limit Theorem suggests that sampling distributions - regardless of the distribution being sampled from - will converge to a Gaussian distribution if the sample size is sufficiently large.

How large a sample size is sufficient before the sampling distribution is approximately Gaussian? This depends on how biased (i.e., non-normal) the distribution being sampled from is.

This report will explore the behavior of a distribution of sample means when the distribution sampled from is an exponental distribution with a rate parameter (aka - lambda) of 0.2. The sample size for each draw from the original distribution will be 40.

Specifically, we will test whether a sample size of 40 is sufficient to observe the following Central Limit Theorem properties:

- Per the Law of Large Numbers, the expected value of the sampling distribution should approximate the expected value of the source distribution (here 1/lamda or 5).
- The variance of the sampling distribution should scale to the variance observed in the source distribution such that sampling distribution variance = source distribution variance / n.
- The sampling distribution should be approximately normal.

Load Supporting Resources

Here we load any supporting R packages used for the report. These will need to be installed and loaded in an R environment if you want to reproduce in full the computations in the report.

```r
library(ggplot2)
library(ggthemes)
```

Simulating Sampling from the Exponential Distribution (n = 40)

The first step in this report is simulating the act of sampling from our target exponential distiribution (lambda = 0.2).

We will sample 40 exponents from the distribution and then take the mean of this sample. This process will repeat 1000 times to build our distribution of sample means.

```r
# establish our simulation parameters
rate <- 0.2
sample_size <- 40
simulated_means <- c()

# set the simulation seed so that our results are consistent across reports
set.seed(1)
```

```
# 1000 times: sample 40 values from an exponential distribution with lambda =
# 0.2, taking the mean value each time
for(index in 1:1000) {
    simulated_means <- c(simulated_means,
                         mean(rexp(sample_size, rate)))
}
```

Comparing the Observed Mean of Random Samples v. the Theoretical Mean

Now that we have a collection of sample means, we can start exploring how that distribution compares to what we expect based on the Central Limit Theorem.

We first check for the first of our expected Central Limit Theorem properties: Per the Law of Large Numbers, the expected value of the sampling distribution should approximate the expected value of the source distribution. The mean of an exponential distribution is 1/lambda.

```
# determine the mean value we expect (the source exponential distribution mean)
expected_mean <- 1/rate

# determine the observed value for the collection of sample means
observed_mean <- mean(simulated_means)

# compare
expected_mean
```

```
## [1] 5
```

```
observed_mean
```

```
## [1] 4.990025
```

```
expected_mean - observed_mean
```

```
## [1] 0.009974799
```

At least for the current `set.seed()`, the observed sampling distribution mean appears to be a close approximation to what we would expect (i.e., the same value as the source distribution mean).

Comparing the Observed Sample Variance v. the Theoretical Variance

Now we check the second of our expected Central Limit Theorem properties: The variance of the sampling distribution should scale to the variance observed in the source distribution such that sampling distribution variance = source distribution variance / n.

For the exponential distribution, it is easier start from the standard deviation first. The standard deviation for this distribution is also 1/lambda.

```
# determine the source standard deviation and then switch it to variance
source_sd <- 1/rate
source_variance <- source_sd^2

# determine the expected variance for the sampling distribution
expected_variance <- source_variance/sample_size

# determine the observed variance for our collection of sample means
observed_sd <- sd(simulated_means)
observed_variance <- observed_sd^2

# compare
expected_variance
```

```
## [1] 0.625
```

```
observed_variance
```

```
## [1] 0.6111165
```

```
expected_variance - observed_variance
```

```
## [1] 0.01388353
```

At least for the current `set.seed()`, the observed sampling distribution variance appears to be a close approximation to what we would expect (i.e., the source distribution variance divided by the sample size).

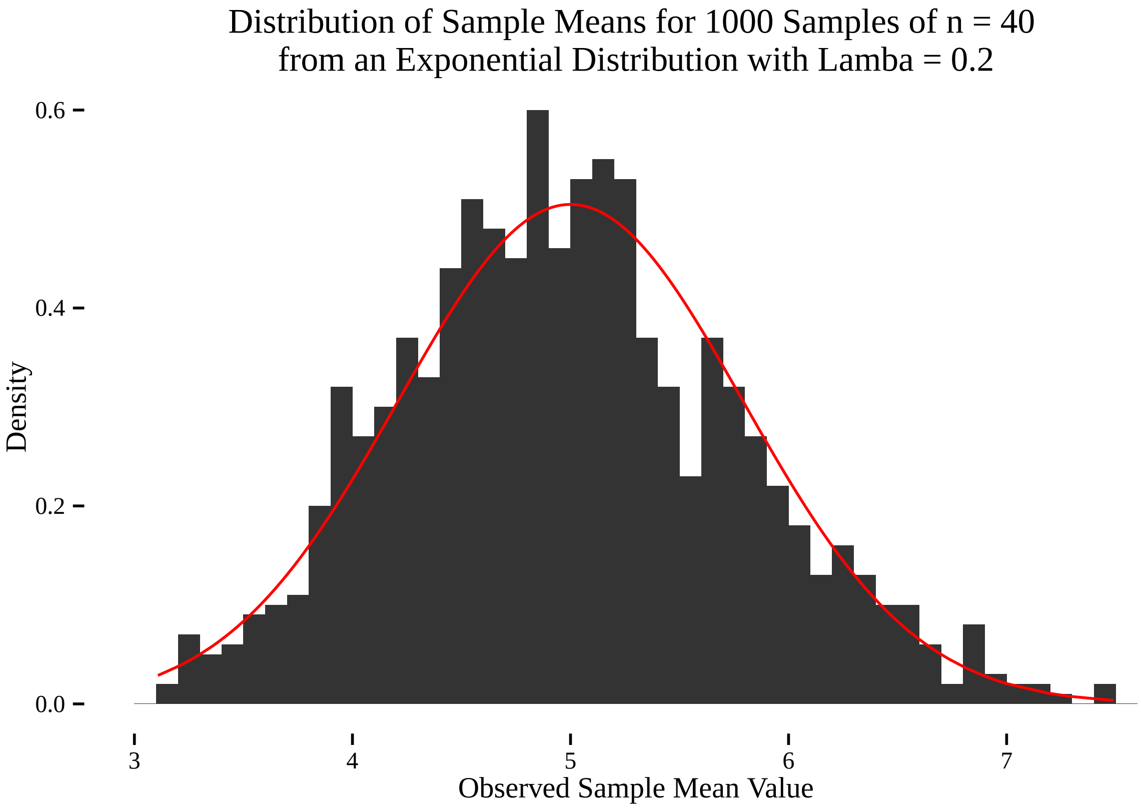
Assessing Normality of the Distribution of Random Samples

We'll conclude the report by assessing the final of our desired Central Limit Theorem properties: The sampling distribution should be approximately normal.

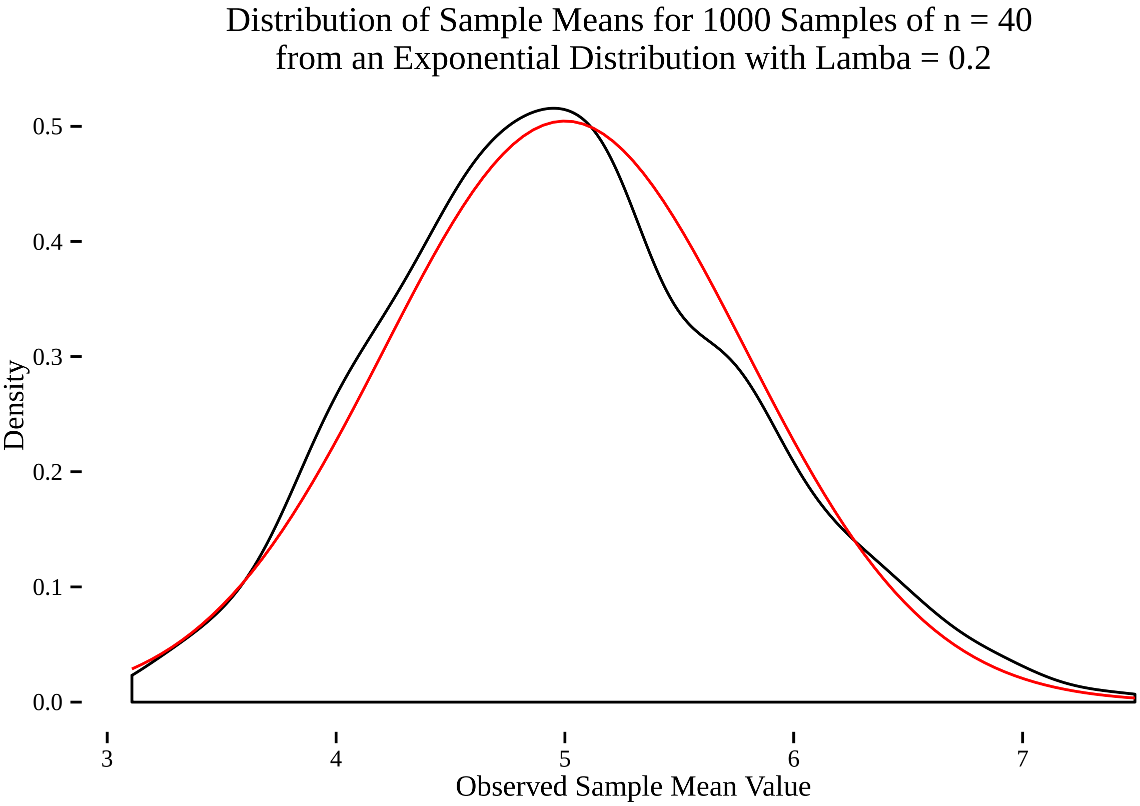
Although there are a variety of statistics for assessing the normality of a distribution (e.g., skew, kurtosis), we will keep our assessment simple. We will do a simple visual contrast between our observed sampling distribution and a normal distribution.

```
# get our expected standard deviation (we already know the expected mean)
expected_sd <- sqrt(expected_variance)

# create a dataframe from our simulated means object so that it plays nicely
# with ggplot
simulated_means_df <- data.frame(simulated_means)

# expand our sample means data-frame to include the probability density we
# would expect to see for each sample mean value if they were part of a
# normal distribution built from our expected_mean and expected_sd
simulated_means_df$expected_density <- dnorm(simulated_means_df$simulated_means,
                                             expected_mean,
                                             expected_sd)

# now we plot the histogram of sample means, setting the y-axis to show us the
```

```r
# density of observed sample means that occur at the various values;
# on top of this, we add a line showing the density we would have expected to
# for our sample mean values
ggplot(simulated_means_df, aes(x = simulated_means, y = ..density..)) +
    geom_histogram(binwidth = 0.1) +
    geom_line(aes(y = expected_density), color = "red") +
    labs(title = paste0("Distribution of Sample Means for 1000 Samples of n = ",
                        "40 \nfrom an Exponential Distribution with Lamba = ",
                        "0.2"),
         y = "Density",
         x = "Observed Sample Mean Value") +
    theme_tufte()
```



```r
# as an alternative, we can also draw a smooth density of our observed sample
# means along with a an appropriate normal density curve
ggplot(simulated_means_df, aes(x = simulated_means)) +
    geom_density(binwidth = 0.1) +
    stat_function(fun = dnorm,
                  arg = list(mean = expected_mean,
                             sd = expected_sd),
                  color = "red") +
    labs(title = paste0("Distribution of Sample Means for 1000 Samples of n = ",
                        "40 \nfrom an Exponential Distribution with Lamba = ",
                        "0.2"),
         y = "Density",
        x = "Observed Sample Mean Value") +
    theme_tufte()
```



At least for the current `set.seed()`, the observed sampling distribution appears to have a slight right-skew (aka - more density than expected to the left). However, this skew seems sufficiently minor to suggest that our sampling distribution is relatively normal.