Statistical Inference Course Project: Simulation Exploration of the Exponential Distribution

Philip Bulsink

2017-01-28

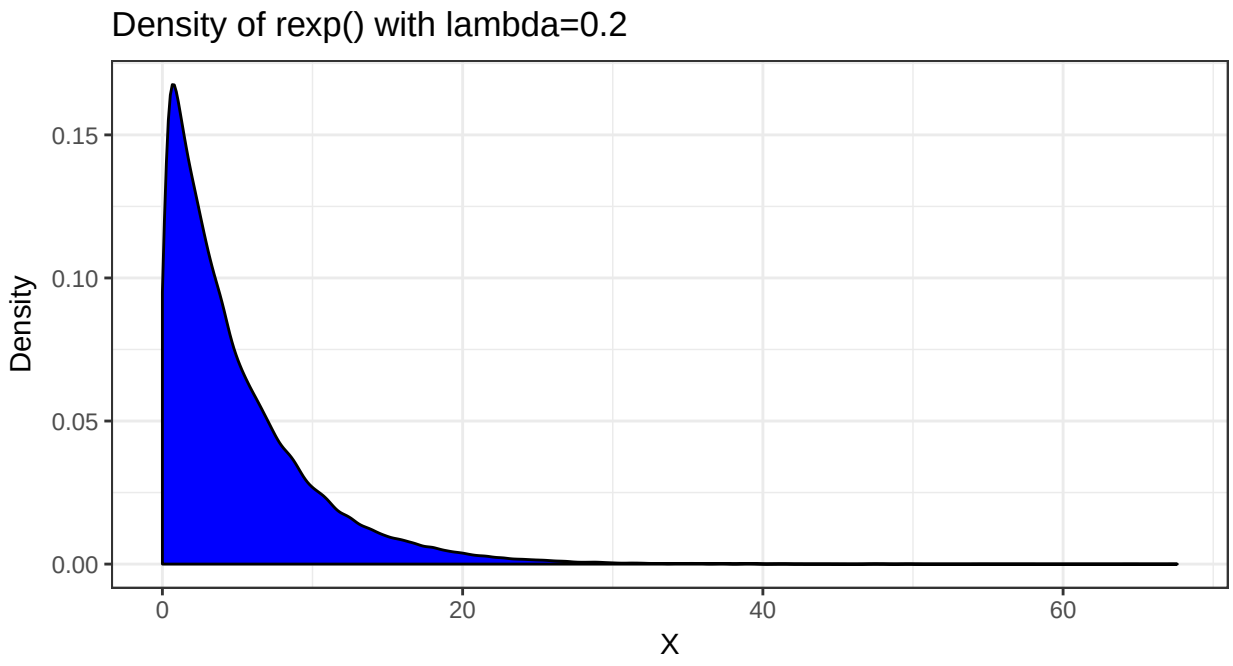
Introduction

There are a range of statistical distributions that serve unique purposes in statistical analysis of data. These distributions include Poisson, binomial, uniform, and normal distributions. This report will investigate the exponential distribution and compare it to the normal distribution.

Simulations

To investigate this distribution, simulations of samples from the population of the exponential distribution will be analyzed. A sample from the population can be drawn with the `R` command `rexp(n, lambda)`, where `n` is the number of samples to draw, and `lambda` provides both the mean and the standard deviation (both equal to `1/lambda`). Lambda is also known as `rate`, according to the coding standard. For this study, we will use a lambda value of 0.2, providing us with a mean of `1/0.2 = 5`, and a standard deviation of `1/0.2 = 5` Thus, when we sample we draw from a population with this approximate density:

```
ggplot(data = data.frame(x = rexp(100000, 0.2)), aes(x = x)) +
    geom_density(fill = 'blue') +
    ggtitle("Density of rexp() with lambda=0.2") +
    xlab("X")+
    ylab("Density")+
    theme_bw()
```



We can 'validate' our distribution by sampling from it `n` times and taking a mean, and comparing that to the expected mean of `1/lambda`. However, if we perform that repeatedly, we can generate a distribution of the means, which is useful for later exploration.

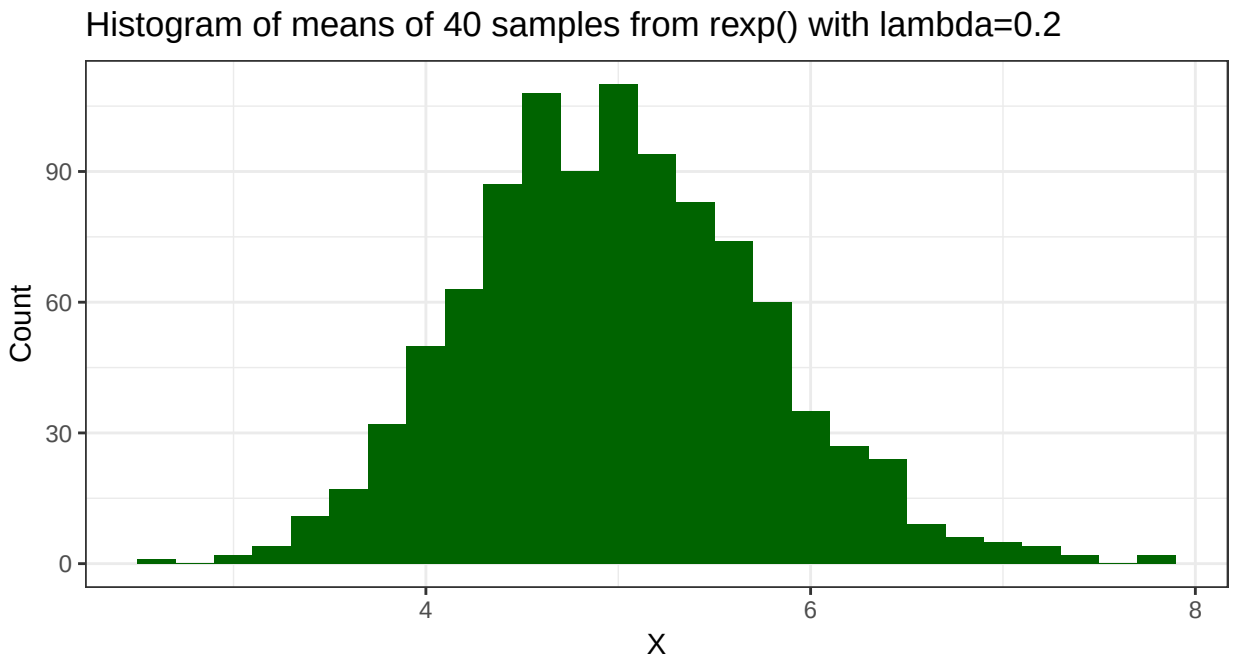
We can do these simulations with the following short `R` code:

```
sample_means<-NULL
for (i in 1 : 1000) sample_means = c(sample_means, mean(rexp(40, 0.2)))
```

Sample and Theoretical Mean

The Central Limit Theorem states that, with a large enough sample size, the distribution of means of a population of independent and identically distributed values will be normally distributed. This extends to mean that, with enough repeats, the mean of the distribution of sample means will be centered around the true mean of the population.

```
ggplot(data = data.frame(SampleMeans = sample_means), aes(x = SampleMeans)) +
    geom_histogram(binwidth = 0.2, fill = 'darkgreen') +
    ggtitle("Histogram of means of 40 samples from rexp() with lambda=0.2") +
    xlab("X")+
    ylab("Count")+
    theme_bw()
```



This has valuable implications. We know that the population has a mean of `1/lambda`, or for our simulations, when `lambda` = 0.2, a mean of 5. With our simulation from above, we can calculate the mean of the distribution as simply as `mean(sample_means)`, which comes out to 4.993. This is within the limits of the simulation, running with 10,000 or 100,000 samples instead of 1,000 may provide us with a more exact result.

Sample and Theoretical Variance

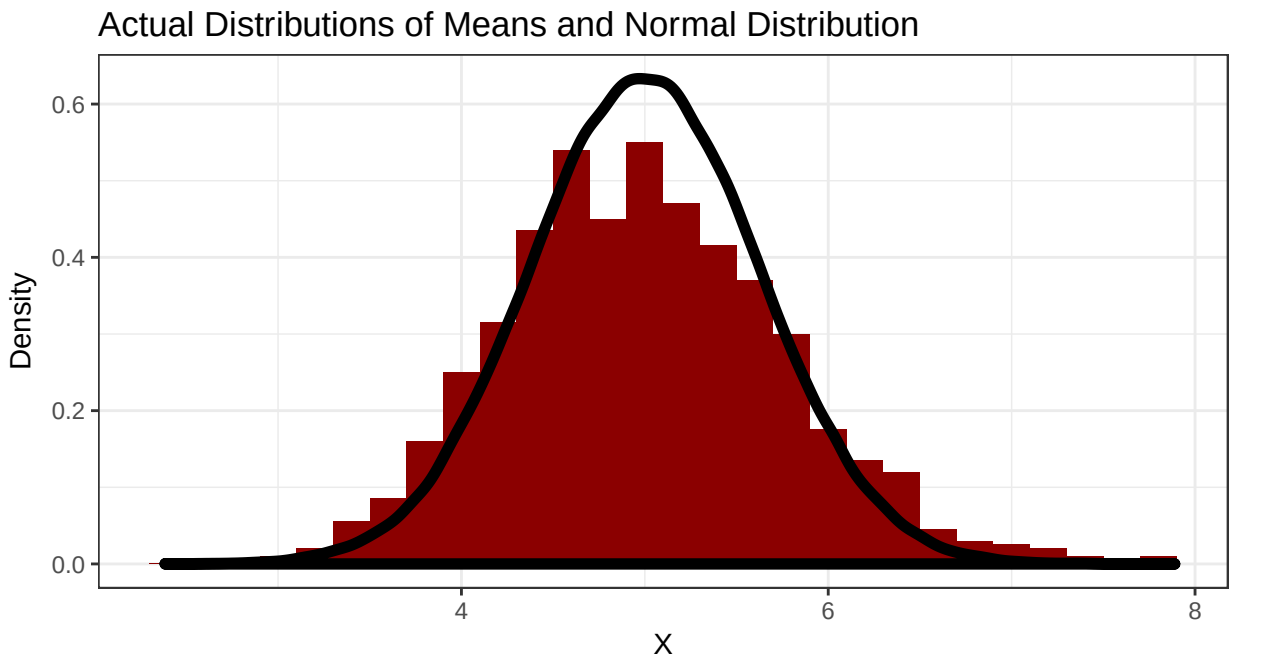
Similarly, we can compare the expected population variance calculated by the means to the actual variance of the population, since the variances of the sample means, and population variance are correllated. The

standard deviation of the population was set to equal 5, thus the variance will be `5^2 = 25`. The population variance can be estimated from the distribution of means using the formula `sd(sample_means)^2 * n`, where `n` is the number of samples. For our `n = 40`, this equals 23.376, which is close to our expected 25.

Comparison to Normal Distribution

The final comparison the normal distribution can be done visually. Recall from above the distribution of the means. This will be presented again, as a density (where for each bar in the histogram the formula `count/total simulations` has been applied) but with an overlayed standard normal density with the predicted mean of 5 and variance of 0.791.

```
ggplot(data = data.frame(SampleMeans = sample_means,
                         NormalDensity=rnorm(n = 100000, mean = 5, sd = (5/sqrt(40))^2))) +
    geom_histogram(aes(x=SampleMeans, y=..density..), fill='darkred', binwidth=0.2) +
    geom_density(aes(x=NormalDensity), colour="black", size=2) +
    ggtitle("Actual Distributions of Means and Normal Distribution") +
    xlab("X")+
    ylab("Density")+
    theme_bw()
```



Thus, from the image above, we can see that the distribution of means somewhat matches the normal distribution. We know (as above) that the results will more closely match theory if we increase the number of simulations.

Conclusion

In conclusion, we showed that the mean of the sample means was 4.993, compared to a theoretical sample mean of 5. Similarly, we showed that the variance of sample was 23.376, close to the theoretical value of 25. Finally, we displayed that the distribution was similar to the normal distribution, centered around the same point, with the same variance.