Showing That the Means of 40 Exponentials are Approximately Normally Distributed

Araks Stepanyan

7/20/2017

Overview

This document is created for the final project of the Statistical Inference course offered by Johns Hopkins University through Coursera. This is the first part of the project where we investigate the exponential distribution in R and show that the means of 40 exponentials are approximately normally distributed.

Means of Exponentials

Simulations

To start our analysis let's simulate (thousand times) 40 exponentials and take their means. We are going to use n = 40 and $\lambda = 0.2$ for all simulated exponentials.

```
simulations = 1000
n = 40
lambda = 0.2
exponent.means <- sapply(1:simulations, function(i) mean(rexp(n, lambda)))
length(exponent.means)
```

```
## [1] 1000
```

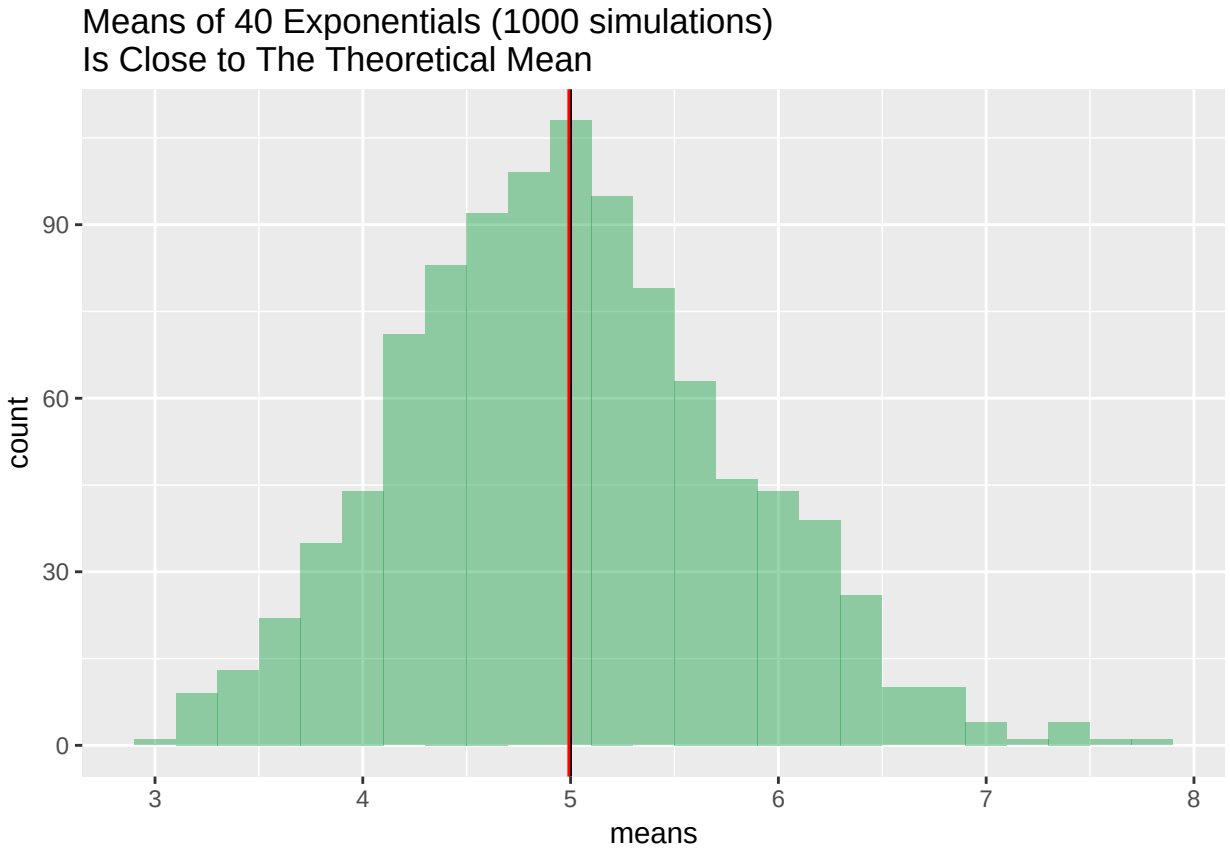
So now we have 1000 means of 40 exponentials. Let's look at the mean of these means and compare it to the theoretical mean. The theoretical mean of exponential distribution is $1/\lambda$.

Sample Mean versus Theoretical Mean

```
library(ggplot2)
```

```
theoretical.mean <- 1/lambda
sample.mean <- mean(exponent.means)
```

```
g <- ggplot(data.frame(means = exponent.means), aes(x = exponent.means))
g <- g + geom_histogram(fill = rgb(0,0.6,0.2,0.4), binwidth = 0.2)
g <- g + geom_vline(xintercept = theoretical.mean, color = "black")
g <- g + geom_vline(xintercept = sample.mean, color = "red")
g <- g + labs(x = "means")
g <- g + labs(title = "Means of 40 Exponentials (1000 simulations)\nIs Close to The Theoretical Mean")
g
```




```
theoretical.mean
```

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

```
round(sample.mean,2)
```

```
## [1] 4.99
```

The mean of sampling distribution of 40 exponentials (the red line on the plot above) is 4.99 and it's pretty close to the theoretical mean, which is 5 (the black line). We will talk more about it later.

What about variability of the sampling distribution? Let's look at the variance of the means and compare it to the theoretical variance. The theoretical standard deviation of exponential distribution is $1/\lambda$, hence the variance is $(1/\lambda)^2$.

Sample Variance versus Theoretical Variance

```
theoretical.variance <- (1/lambda)^2
sample.variance <- var(exponent.means)
theoretical.variance
```

```
## [1] 25
```

```
round(sample.variance, 2)
```

```
## [1] 0.63
```

The variance of sampling distribution is quite different from the theoretical variance. It is smaller, which means that the distribution of averages is less spread out than the original distribution. And it makes

sense, as it is less likely for sample means to be far away from the population mean than it is for individual observations.

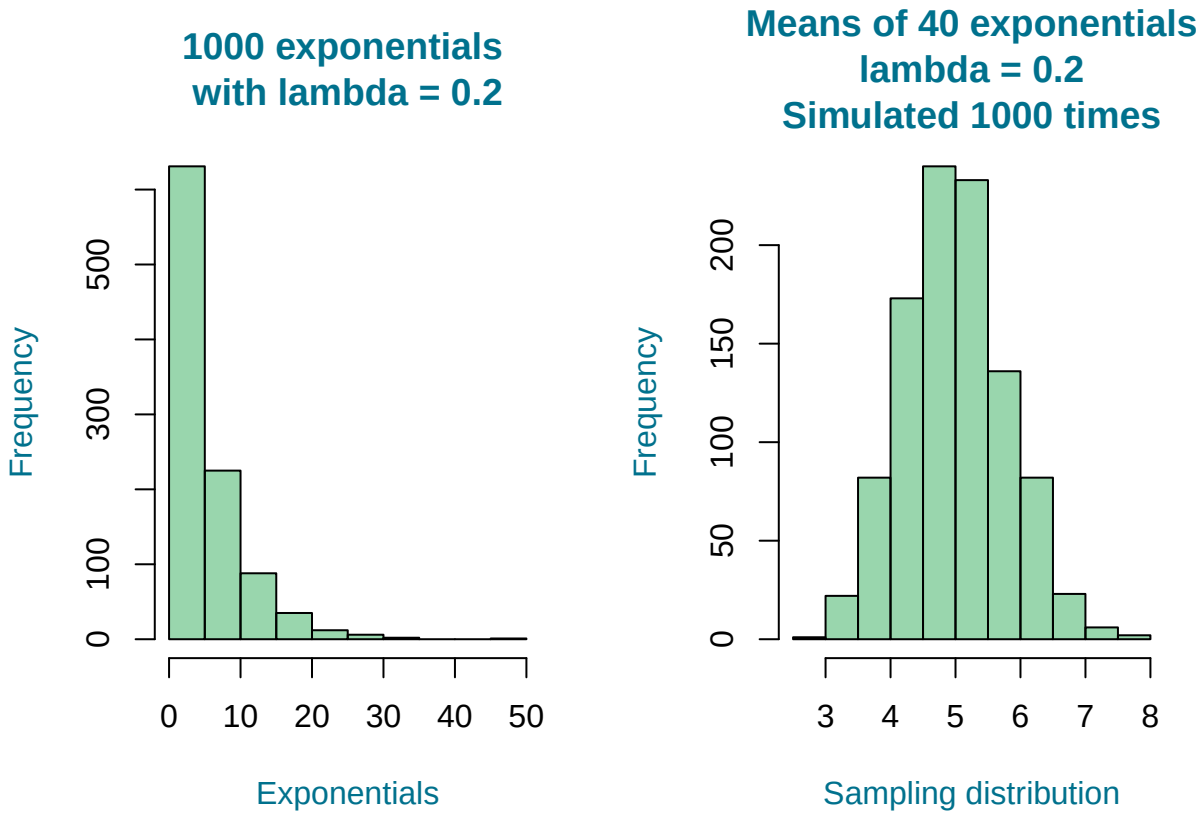
Now let's move on to showing that the distribution of means is approximately normal. For that we will look at the histograms of 1000 random exponentials and the histogram of our simulated averages of 40 exponentials.

Distribution

```
# library(RColorBrewer)
par(mfrow = c(1, 2))

hist(rexp(1000, lambda), col = rgb(0,0.6,0.2,0.4),
     xlab = "Exponentials", main = "1000 exponentials \nwith lambda = 0.2",
     col.lab = colorRampPalette(c("red", "green", "blue"))(10)[8],
     col.main = colorRampPalette(c("red", "green", "blue"))(10)[8])

hist(exponent.means, col = rgb(0,0.6,0.2,0.4),
     xlab = "Sampling distribution",
     main = "Means of 40 exponentials\nlambda = 0.2\nSimulated 1000 times",
     col.lab = colorRampPalette(c("red", "green", "blue"))(10)[8],
     col.main = colorRampPalette(c("red", "green", "blue"))(10)[8])
```



From the two histograms we see that sampling distribution looks far more Gaussian than the original exponential distribution. And that is due to the Central Limit Theorem (CLT).

CLT states that **the distribution of averages of independent, iedntically distributed (iid) variables (properly normalized) becomes standard normal as the sample size increases. In other words, sample average is approximately N(μ, σ^2/n), where μ is population mean and σ^2 is population**

variance.

Let's check conditions for our example.

- we simulated random exponentials, so the **iid condition** is met
- n = 40 **sample size is large enough**

Thus, the distribution of sample mean is approximately $\mathbf{N}(\mu, \sigma^2/n) = \mathbf{N}(1/\lambda, (1/\lambda)^2/n) = \mathbf{N(5, 0.62)}$. Quite true, we have already seen that the mean of 1000 means is 4.99 which is very close to 5. And the variance of 1000 means is 0.63 which is indeed almost equal to 0.62.