Applying the Central Limit Theorem to Averages of Random Exponentials

Calvin Seto

September 7, 2015

Overview

In this exercise, we will use the Central Limit Theorem to show that the distribution of averages of independent and identically distributed exponential variables becomes that of a standard normal as the sample size increases. Using R, we can simulate random exponentials of different sample sizes and compute 1,000 averages from each sample size. Lastly, we can compute the Central Limit Theorem approximation of the distribution of exponential averages and discuss its attributes.

The Law of Large Numbers says that the sample mean or average limits to what it's estimating, the population mean. It assumes the data have to be independent and identically distributed, or iid (independent and all drawn from the same population).

Listing 1 shows R code that computes 1,000 averages from 1,000 samples of random exponentials and plots the cumulative means in Figure 1. Notice how the averages are random at first but converge to 5, the mean of the exponential distribution with $\lambda = 0.2$.

An estimator is consistent if it converges to what you want to estimate. The sample mean of iid samples is consistent with the population mean. The sample variance and standard deviation is consistent as well.

The Central Limit Theorem says that the distribution of averages of iid random variables, properly normalized, becomes that of a standard normal as the sample size increases.

The result is that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

has a distribution like that of a standard normal for large n.

The useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$

Simulations

We'll use the R function rexp(n, λ) to create random samples of exponentials with sample sizes of 10,000, 20,000 30,000, and 40,000. We'll set $\lambda = 0.2$ for all simulations and we know the mean and standard deviation $= 1/\lambda$, or $1/0.2 = 5$, and the variance is $\sigma^2 = 25$. We'll call these samples of exponentials s1, s2, s3, and s4.

In order to compute our averages of random exponentials, we'll reorganize the samples s1, s2, s3, and s4 into matrices m1, m2, m3, and m4 of sizes 1000x10, 1000x20, 1000x30, and 1000x40. Then the R function apply is used with a custom function to normalize the averages with the CLT approximation of subtracting the population mean from the sample mean and dividing by the standard error.

We set up default variables:

nosim - number of simulations - 1,000

lambda - exponential distribution rate parameter - 0.2

mu - mean of the exponential distribution - 1/lambda = 5

s - standard deviation of the exponential distribution - 1/lambda = 5

Listing 2 shows the R code to set up the variables we need for the CLT approximation. We'll set the seed of R's random number generator to 568 so we can reproduce the results.

Sample Mean versus Theoretical Mean

Table 1 shows a comparison of sample means to the theoretical means for each sample size. The sample mean for size of 10,000 exponentials is about 6.96, 1.96 more than 5. The sample mean for size of 40,000 exponentials is about 4.57, 0.43 less than 5. This supports the property of the Law of Large Numbers, as the sample mean approaches 5 as the sample size increases.

Sample Variance versus Theoretical Variance

Table 2 shows a comparison of sample variances to the theoretical variances for each sample size. The sample variance for size of 10,000 exponentials is about 55.93, 30.93 more than 25. The sample variance for size of 40,000 exponentials is about 23.13, 1.87 less than 25. It appears the Law of Large Numbers can be applied to the sample variances, too as the sample variance approaches 25.

Distribution

The function cfunc(x, n) in Listing 2 will normalize a vector x, with sample size n using the CLT result as described before.

The results of the CLT approximations of the distribution of averages of exponentials are stored in the variables x1, x2, x3, x4; four vectors of 1,000 numerics.

Listing 3 shows the R code that will package our CLT approximations into a data frame called dat for ggplot to display. The dat variable has size 4,000x2 to hold our CLT approximations and sample size factors of 10, 20, 30, and 40 used to create four histograms. Figure 2 shows the distributions of the averages of random exponentials for sample sizes 10, 20, 30, and 40.

The properties of the distributions of averages of exponentials is

-it is bell shaped and standard normal

-with a mean = 0 and variance = 1.

Normal distributions are drawn over the 4 histograms. Notice how each histogram is centered close to the red, dashed vertical line at 0.

Table 3 shows a comparison of averages means to the theoretical means for each sample size. The averages mean for size of 10 exponentials is about 1.24, 1.24 more than 0. The averages mean for size of 40 exponentials is about -0.55, 0.55 less than 0. The CLT says the mean of the distribution of averages should be standard normal with a mean = 0. Notice, the averages mean for sample size 30 has a smaller difference of 0.19 compared to sample size 40. The CLT does not guarantee a standard normal distribution as the sample size increases.

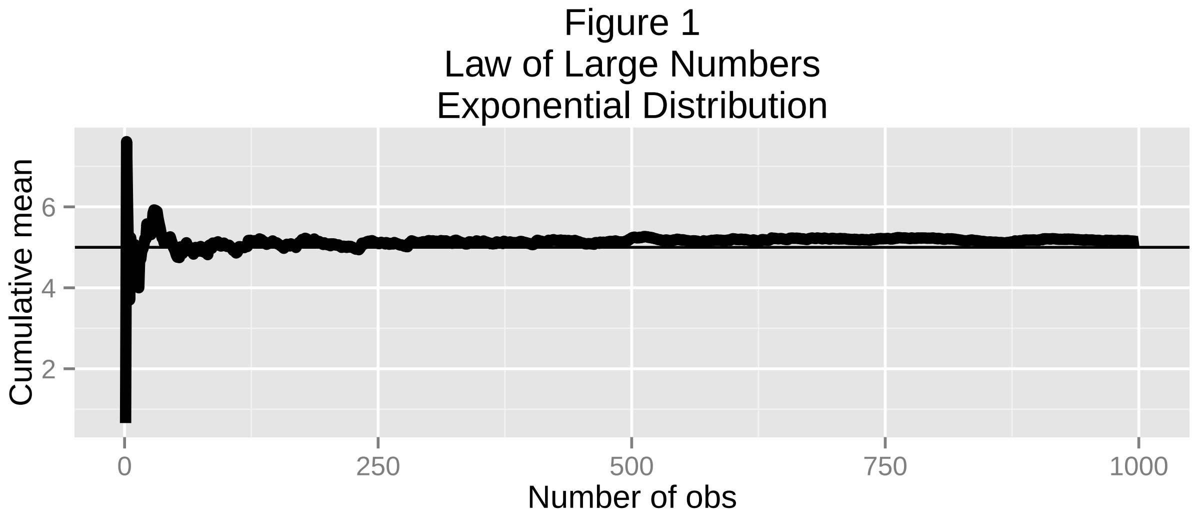
Table 4 shows a comparison of averages variances to the theoretical variances for each sample size. The averages variance for size of 10 exponentials is about 2.30, 1.30 more than 1. The averages variance for size of 40 exponentials is about 0.89, 0.11 less than 1. The CLT says the variance of the distribution of averages should be standard normal with a variance = 1. This supports the property of the Central Limit Theorem, also.

Listing 4 shows the R code which generates a histogram of the sample of random exponentials used to calculate the averages. Figure 3 shows the histgram for random exponentials of sample size 40,000. Notice the shape of this histogram is NOT bell shaped, whereas the histograms of the averages of exponentials of sample size 10, 20, 30, and 40 ARE bell shaped.

Appendix

Listing 1

```
set.seed(568)
library(ggplot2)
n <- 1000
lambda <- 0.2
means <- cumsum(sample(rexp(n, lambda), n , replace = TRUE)) / (1  : n)
g <- ggplot(data.frame(x = 1 : n, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 5) + geom_line(size = 2)
g <- g + labs(x = "Number of obs", y = "Cumulative mean")
g <- g + ggtitle("Figure 1\nLaw of Large Numbers\nExponential Distribution")
g
```



Listing 2

```
set.seed(568)
library(ggplot2)
nosim <- 1000
lambda <- 0.2
mu <- 1/lambda
s <- 1/lambda
cfunc <- function(x, n) sqrt(n) * (mean(x) - mu) / s

s1 <- sample(rexp(10,lambda), nosim * 10, replace = TRUE)
s2 <- sample(rexp(20,lambda), nosim * 20, replace = TRUE)
s3 <- sample(rexp(30,lambda), nosim * 30, replace = TRUE)
s4 <- sample(rexp(40,lambda), nosim * 40, replace = TRUE)
m1 <- matrix(s1, nosim)
m2 <- matrix(s2, nosim)
m3 <- matrix(s3, nosim)
m4 <- matrix(s4, nosim)
x1 <- apply(m1, 1, cfunc, 10)
x2 <- apply(m2, 1, cfunc, 20)
x3 <- apply(m3, 1, cfunc, 30)
x4 <- apply(m4, 1, cfunc, 40)
```

Table 1

	Sample Mean	Theoretical Mean	Difference
10,000	6.962282	5	1.9622820
20,000	6.265946	5	1.2659461
30,000	5.177740	5	0.1777404
40,000	4.567163	5	0.4328372

Table 2

	Sample Variance	Theoretical Variance	Difference
10,000	55.92664	25	30.92664
20,000	58.51260	25	33.51260
30,000	38.22658	25	13.22658
40,000	23.12854	25	1.87146

Listing 3

```
dat <- data.frame(
  x = c(x1, x2, x3, x4),
  size = factor(rep(c(10, 20, 30, 40), rep(nosim, 4))))
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3, colour = "black", a
g <- g + geom_vline(aes(xintercept=0), colour="#990000", linetype="dashed")
g <- g + stat_function(fun = dnorm, size = 2)
g <- g + ggtitle("Figure 2\nCentral Limit Theorem Approximation\nAverages of Exponentials")
g + facet_grid(. ~ size)
```

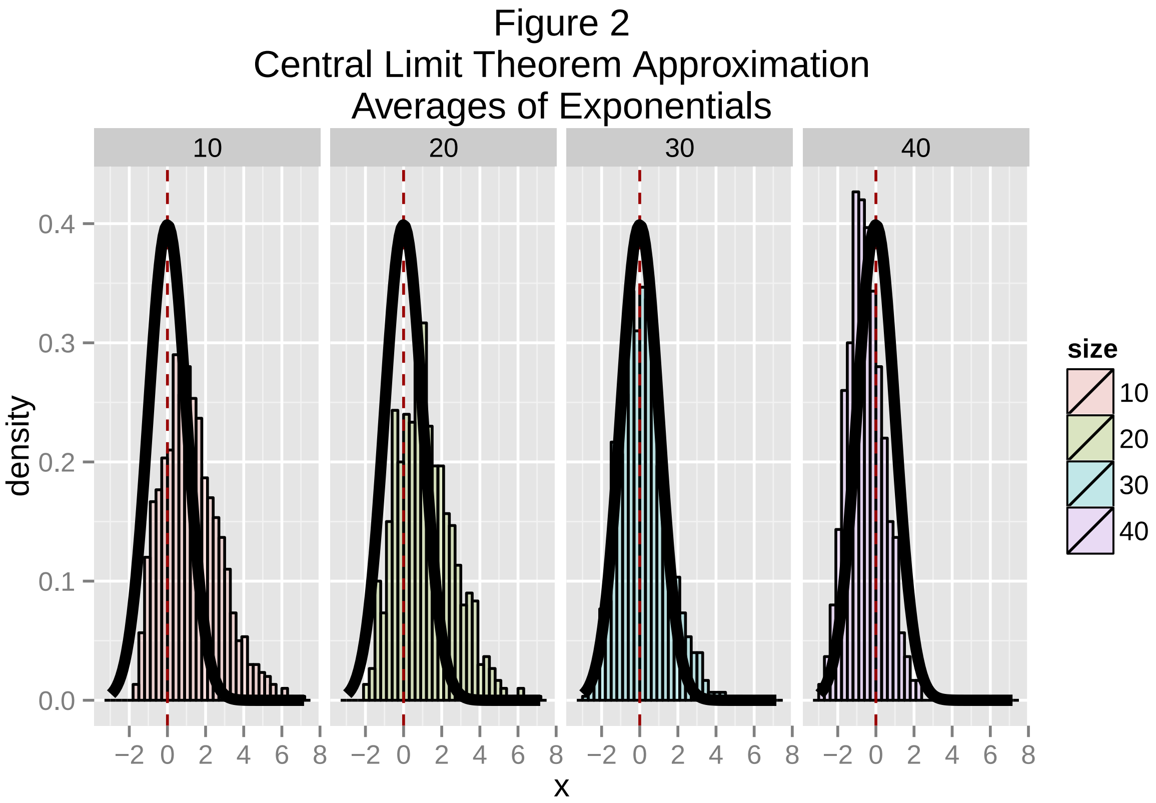


Table 3

	Averages Mean	Theoretical Mean	Difference
10	1.2410561	0	1.2410561
20	1.1322966	0	1.1322966
30	0.1947049	0	0.1947049
40	-0.5475005	0	0.5475005

Table 4

	Averages Variance	Theoretical Variance	Difference
10	2.2961973	1	1.2961973
20	2.2527812	1	1.2527812
30	1.5292529	1	0.5292529
40	0.8912114	1	0.1087886

Listing 4

```
dfs4 <- data.frame(s4)
gs4 <- ggplot(data=dfs4, aes(dfs4$s4)) + geom_histogram() + geom_histogram(alpha = .20, binwidth=.3, col
gs4 <- gs4 + ggtitle("Figure 3 Histogram of Random Exponentials with Sample Size 40,000")
gs4
```

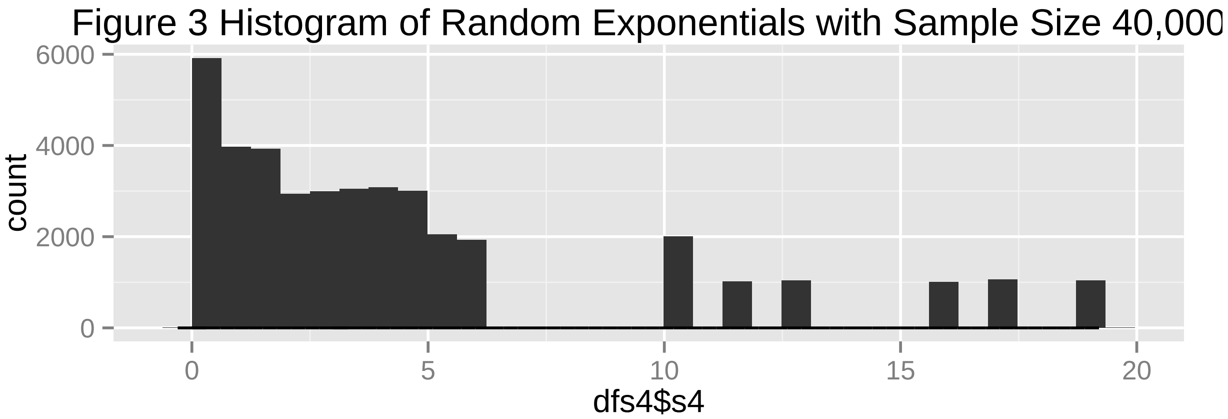


Figure 3 Histogram of Random Exponentials with Sample Size 40,000