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 sample size 10, 20, 30, and 40 and compute 1,000 averages for 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 consitent 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 if you subtract the population mean from the sample mean and divide by the standard error, that distribution is standard normal. Replacing the population sd with the sample sd does not change the CLT.

Simulations

We'll use the R function rexp(n, lambda) again to create more 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 = 1/lambda and the standard deviation = 1/lambda. Therefore, the mean and standard deviation = 1/0.2 = 5. We'll call these samples of exponentials s1, s2, s3, and s4.

In order to compute our averages of varying sizes of 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. Now it is easy to use the R function apply 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.

The function cfunc(x, n) will normalize a vector x, with sample size n using the CLT result, subtracting the mean of the estimate from the estimate and dividing by the standard error of the estimate.

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

$\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$

The results of the CLT approximations of the distribution of averages of exponentials are stored in the variables x1, x2, x3, x4 - 4 vectors of 1000 numerics.

The dat variable is a data frame of size 4000x2 to hold our CLT approximations and sample size factors of 10, 20, 30, and 40 used to create histograms with ggplot.

We'll set the seed of R's random number generator to 768 368 468 568 so we can reproduce the results.

The properties of the distribution of averages of 40 exponentials is -it is bell shaped and standard normal -with a mean = 0 and variance = 1.

Sample Mean versus Theoretical Mean

Table 1 shows the means of the 4 samples I used, the theoretical mean, and the differences between the two for sample size 10, 20, 30, and 40. As the sample size increases, the difference between the sample mean and the theoretical mean decreases from 1.96 to 0.43. This supports the CLT property that the sample mean limits to the theoretical mean as the sample size increases. The sample mean is 6.96 for sample size 10, and 4.57 for sample size 40, getting closer to 5.

Sample Variance versus Theoretical Variance

Table 2 shows the variances of the 4 samples I used, the theoretical variance, and the differences between the two. The theoretical variance is (1/lambda)ˆ2 = (1/0.2)ˆ2 = 5ˆ2 = 25.

The variances are different because

Distribution

Listing 3 shows the R code that combines the CLT approximations of averages of random exponentials for sample sizes 10, 20, 30, and 40 and creates histograms of the distributions in Figure 2. A normal distribution is plotted over each histogram for comparison. As the sample size increases from 10 to 40, the distribution of the averages of 40 exponentials becomes more bell shaped.

Notice how each histogram is centered close to 0. Table 3 shows the means of the averages of the 4 samples I used, the theoretical mean, and the differences between the two. The CLT says the distribution of averages is standard normal with a mean = 0 The averages mean changes from 1.24 to -0.55, but the averages mean for sample size of 30 has a smaller difference of 0.19. The CLT does not guarantee a standard normal distribution as the sample size increases.

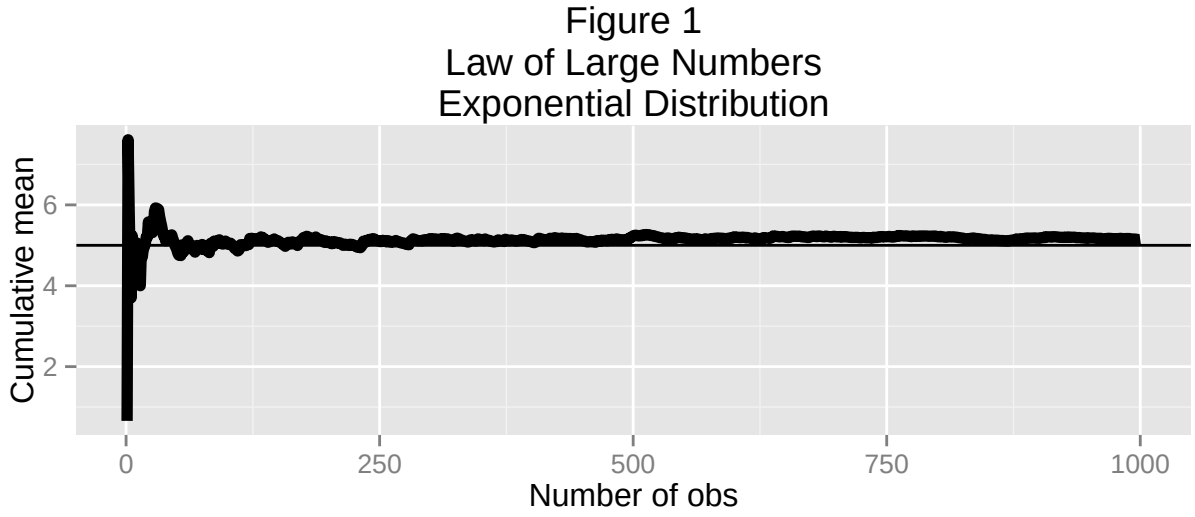
Table 4 shows the variances of the averages of the 4 samples I used, the theoretical variance, and the differences between the two. The CLT say the variance of the distribution of averages should be standard normal with a variance = 1. The variance of the distribution of averages of exponentials for sample size 10 is 2.30 and the variance for sample size 40 is 0.89 supporting the CLT.

Listing 4 shows the R code which generates histograms of the samples of random exponentials used to calculate the averages. Figures 3, 4, 5, and 6 show histgrams for sample sizes of 10,000, 20,000, 30,000, and 40,000. Notice the shapes of all four histograms are NOT bell shaped.

Appendix

Listing 1

```r
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

```r
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	6.962282	5	1.9622820
20	6.265946	5	1.2659461
30	5.177740	5	0.1777404
40	4.567163	5	0.4328372

Table 2

	Sample Variance	Theoretical Variance	Difference
10	55.92664	25	30.92664
20	58.51260	25	33.51260
30	38.22658	25	13.22658
40	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 + stat_function(fun = dnorm, size = 2)
g <- g + ggtitle("Figure 2\nCentral Limit Theorem Approximation\nAverages of Exponentials")
g + facet_grid(. ~ size)
```

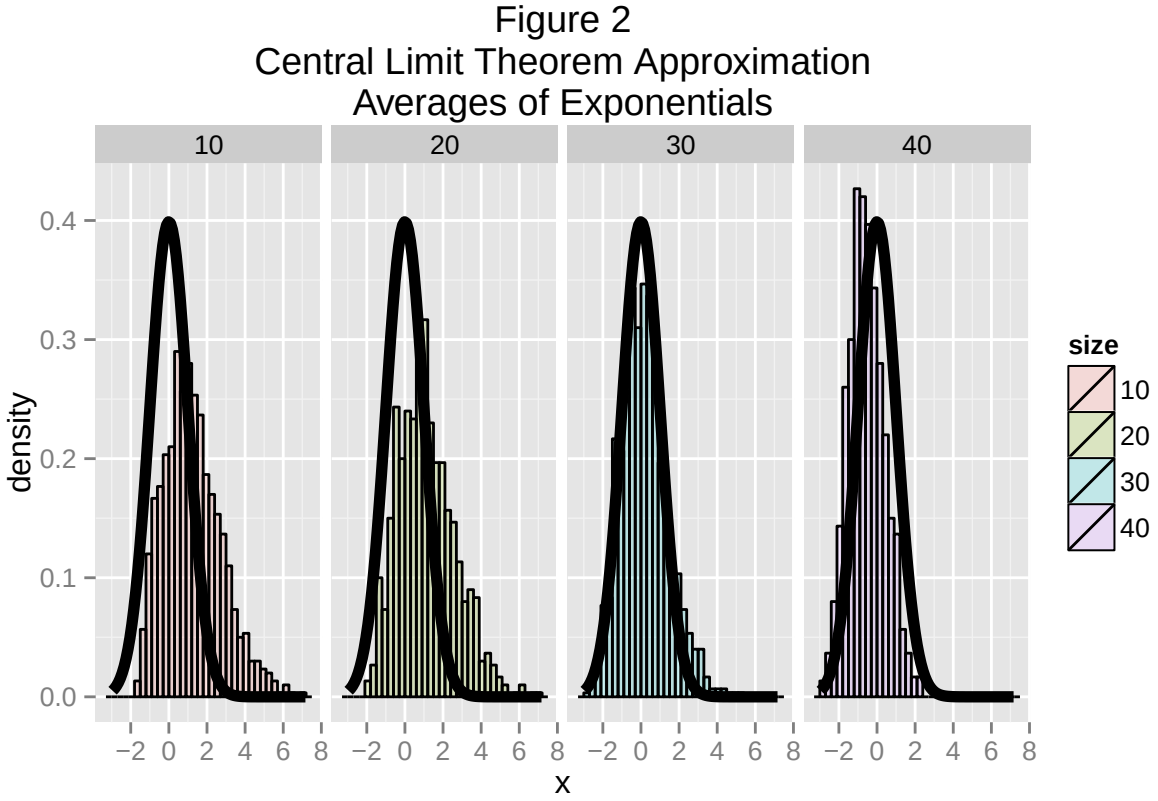


Table 3

	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

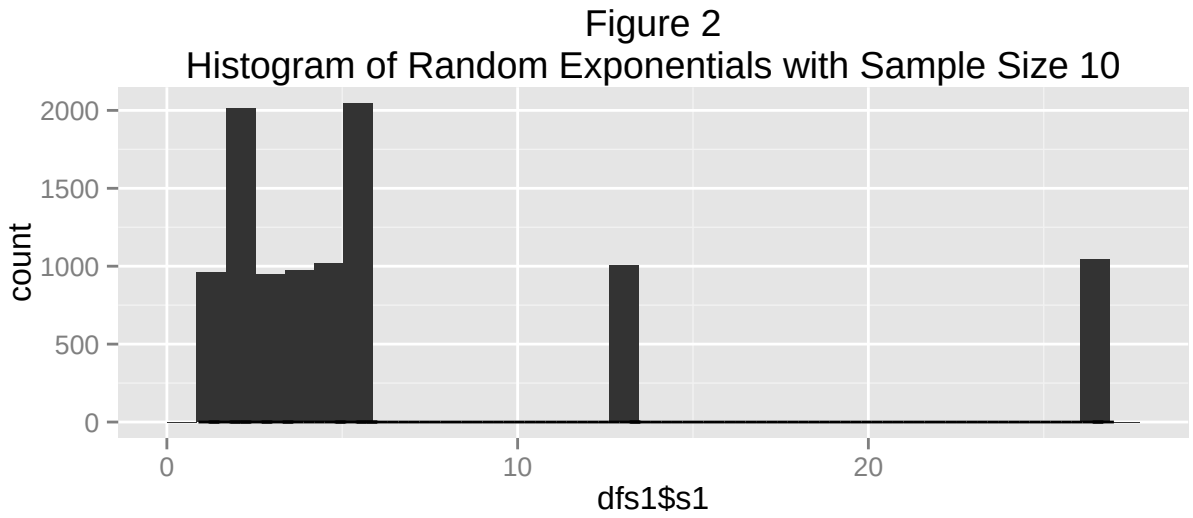
Table 4

	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

```
library(ggplot2)
dfs1 <- data.frame(s1)
gs1 <- ggplot(data=dfs1, aes(dfs1$s1)) + geom_histogram() + geom_histogram(alpha = .20, binwidth=.3, col
gs1 <- gs1 + ggtitle("Figure 2\nHistogram of Random Exponentials with Sample Size 10")
gs1
```

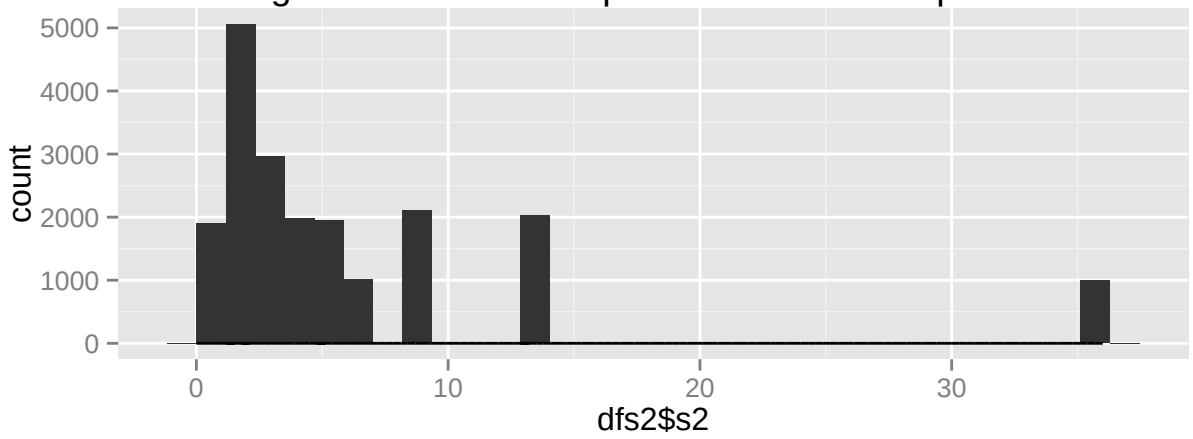
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
dfs2 <- data.frame(s2)
gs2 <- ggplot(data=dfs2, aes(dfs2$s2)) + geom_histogram() + geom_histogram(alpha = .20, binwidth=.3, col
gs2 <- gs2 + ggtitle("Figure 3\nHistogram of Random Exponentials with Sample Size 20")
gs2
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

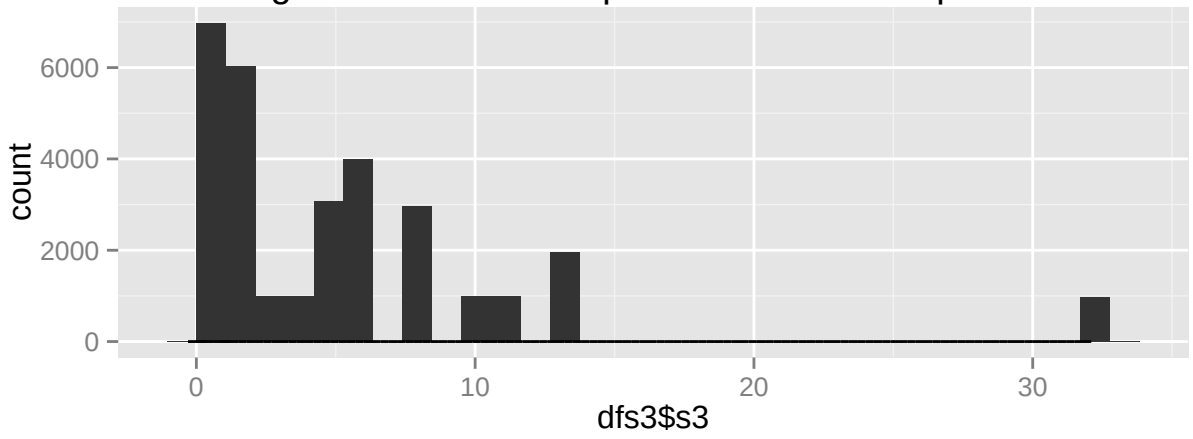
Figure 3
Histogram of Random Exponentials with Sample Size 20




```
dfs3 <- data.frame(s3)
gs3 <- ggplot(data=dfs3, aes(dfs3$s3)) + geom_histogram() + geom_histogram(alpha = .20, binwidth=.3, col
gs3 <- gs3 + ggtitle("Figure 4\nHistogram of Random Exponentials with Sample Size 30")
gs3
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

Figure 4
Histogram of Random Exponentials with Sample Size 30




```
dfs4 <- data.frame(s4)
gs4 <- ggplot(data=dfs4, aes(dfs4$s4)) + geom_histogram() + geom_histogram(alpha = .20, binwidth=.3, col
gs4 <- gs4 + ggtitle("Figure 5\nHistogram of Random Exponentials with Sample Size 40")
gs4
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

Figure 5
Histogram of Random Exponentials with Sample Size 40