The Central Limit of an Exponential PDF

Salma Rodriguez

Thursday, March 12, 2015

Overview

In this analysis, we will investigate the exponential density function in R and compare it with the Central Limit Theorem. The goal of our analysis is to show that the Law of Large Numbers is conserved for repeated sampling of the arithmetic mean of exponentials. We will demonstrate that the Central Limit Theorem holds, and that when the sample size is large, the sample mean is an accurate estimate for the theoretical mean of a standard normal distribution and the sample variance is an accurate estimate for the theoretical variance of a standard normal distribution.

Motivation

In order to build motivation for our experiment, we begin with a conceptual understanding of exponential distributions, which will help us construct an accurate analysis later on.

To make this experiment reproducible, we must set a seed. We use 31415927, but any integer should work.

```
set.seed(31415927)
```

Now, we set the value of λ:

```
lambda = 0.2
```

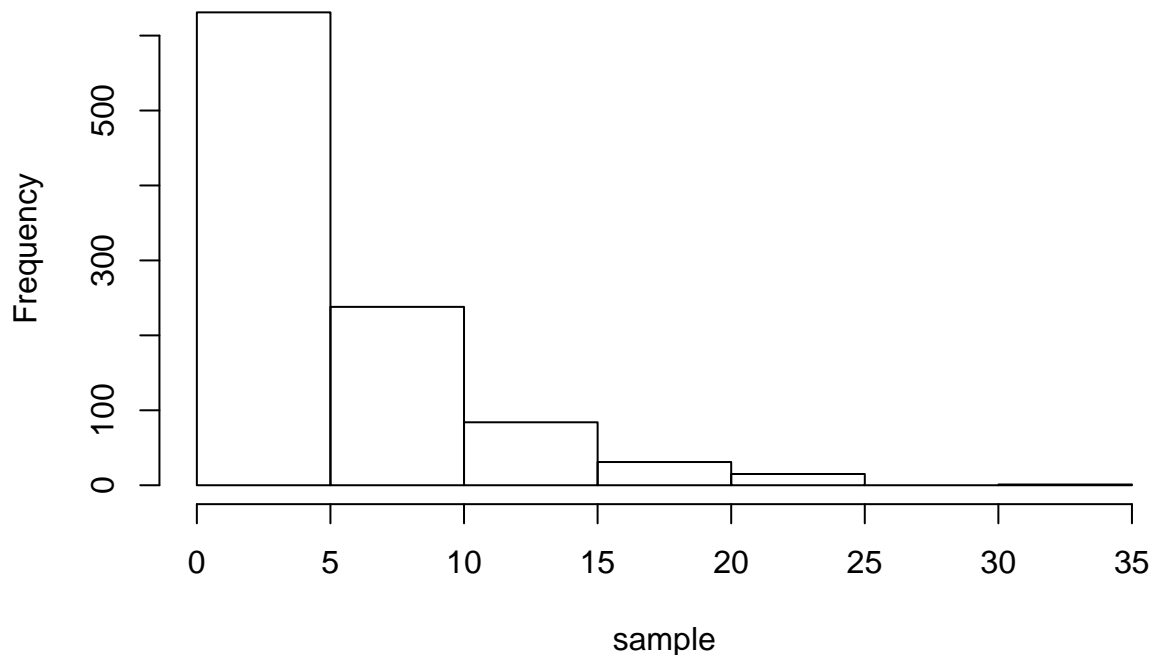We will now collect a sample of 1000 exponentials, with mean λ and variance λ^2.

```
# We would like to save this, since we'll be using it later.

runexp <- function(n = 1000) {
  sample = rexp(n, lambda)
  return (sample)
}

n = 1000
sample <- runexp(n)

hist(sample)
```

Histogram of sample


```
sample_mean = mean(sample)
theoretical_mean = 1 / lambda
```

The distribution looks exponential, as expected, with a center of mass located at 4.8751495. The center of mass of our sample can be compared to the theoretical center of mass, as in equation (1) above.

With a theoretical mean of 5, we may conclude that the sample mean is close to the theoretical mean of an exponential distribution with λ equal to 0.2.

And now for the sample variance, and theoretical variance:

```
sample_variance = sd(sample)^2
theoretical_variance = 1 / lambda^2
```

The variance of our sample is 22.3485675, which is somewhat close to the theoretical variance of 25. This shows that with a large sample of exponentials, the sample mean and variance can be used to approximate the theoretical mean and variance of an exponential distribution with λ equal to 0.2.

Simulation

Before we describe our simulation, we begin with some background details about exponential distributions.

Background

The probability density function of an exponential distribution is piecewise defined as follows:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 1 & \text{if } x < 0 \end{cases}$$

The exponential distribution can be simulated in R with rexp(n, λ), where λ is the rate parameter. The mean of the exponential distribution can be computed as follows:

$$\mu = \frac{1}{\lambda} \qquad (1)$$

and the standard deviation is also as equation (1) above.

The Simulation

To make this experiment reproducible, we must set a seed. We use 31415927, but any integer should work.

```r
set.seed(31415927)
```

Now, we set the value of λ.

```r
lambda = 0.2
```

We will simulate 1000 arithmetic averages of 40 exponentials (i.e., 1000 Monte Carlo repetitions of an experiment with sample size 40). This should enable us to study the expected value of an exponential random variable with theoretical mean λ and theoretical variance $1/\lambda^2$. We will use λ equal to 0.2, as defined above.

We take 1000 random samples of the mean of 40 exponentials:

```r
# We would like to save this, since we'll be using it later.

runsim <- function(n = 40, itrs = 1000) {
  mns = NULL
  for (i in 1 : itrs) mns = c(mns, mean(rexp(n, lambda)))
  return (mns)
}

n = 40 # sample size
means <- runsim(n)
```
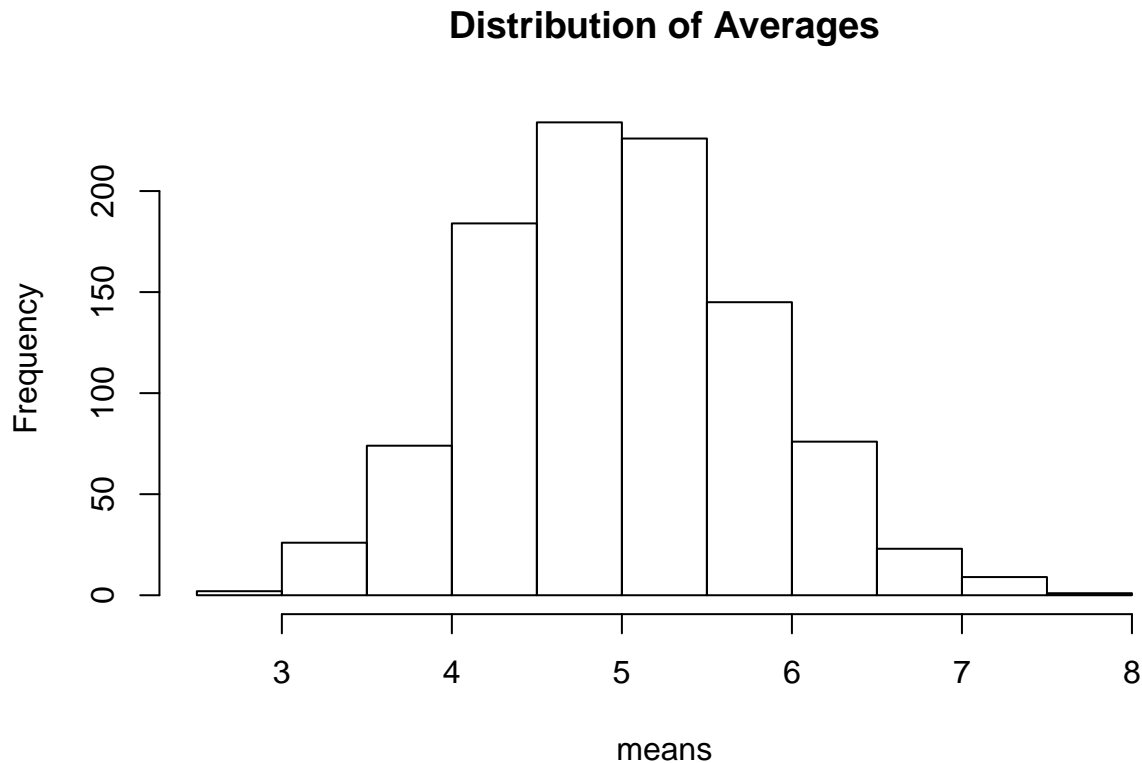
Sample Mean versus Theoretical Mean

Now that the simulation run is complete, we can take a close inspection at the mean of averages for an exponential distribution with λ equal to 0.2.

```r
hist(means, main = "Distribution of Averages")
```

Distribution of Averages

The distribution looks Gaussian, with a center of mass located at 4.9754542. This mean of averages, taken from 1000 averages of 40 mutually independent outcomes of an exponential random variable, is very close to the theoretical mean of an exponential distribution with λ equal to 0.2; namely, $1/\lambda$, or 5.

Sample Variance versus Theoretical Variance

Since the distribution is Gaussian, we can take the standard deviation of our simulation runs. This value is 0.7951854.

We will now normalize the distribution, in order to make the deviation from the mean of our exponential averages comparable to the standard deviation of a standard normal distribution. We can normalize the errors with the following normalization procedure:

$$z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

In R, this is equivalent to:

```
theoretical_sd = sqrt(theoretical_variance)
std_error = theoretical_sd/sqrt(n)
zmns = (means - theoretical_mean)/std_error
```

The standard error, 0.7905694, is close to the standard deviation from the mean of our simulation runs, without normalization. The sample variance of our normalized arithmetic means is the square of the standard deviation; namely, 1.0117116. The standard deviation of 1.0058388 indicates that the distribution of the

4

normalized averages of 40 exponential random variable outcomes is close to a standard normal distribution, with a center of mass at 0, and a standard deviation of 1. We will look further into this in the following section.

Distribution

We can now make a histogram of our 1000 samples of 40 averages, with the distance from the mean of estimates normalized over the standard error of our estimates (namely, the averages of 40 exponentials).

```
hist(zmns, main = "Normalized Distribution of Averages")
```

Normalized Distribution of Averages
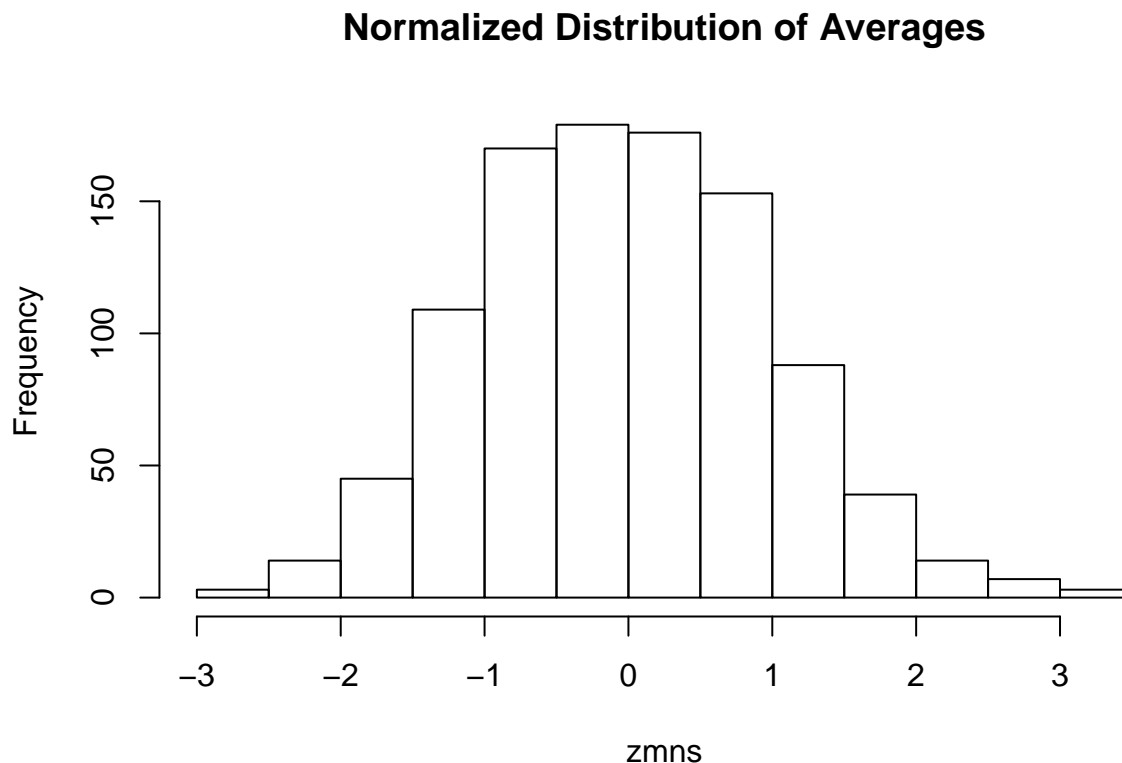
Notice how the mass is approximately centered at 0. Here, zero implies a standard deviation of 0, or zero distance from the expected value of averages. This result is consistent with the Law of Large Numbers for idependent and identically distributed random variables, with the mass centered at -0.031 for 1000 simulation runs, and standard deviation 1.006.

An earlier observation revealed that the standard error of the mean (SEM) is 0.7905694, which is smaller than the standard deviation (SD) computed above. However, note that the SEM is **always** smaller than the SD. With a large sample of averages, the SEM is expected to decrease further, and the SD is expected to converge to 1, or the standard deviation of the standard normal distribution.

Results and Conclusion

In the *Motivation* section, we showed that the sample mean and sample variance of a large number of exponentials is close to the theoretical mean and theoretical variance of an exponential distribution. Continuing

with the theme of motivation, we would like to see how the mean and variance is affected by increasing the sample size.

```r
exvar = NULL
exmns = NULL

for (i in 1:10000) {
  run <- runexp(i)
  exvar = c(exvar, sd(run)^2)
  exmns = c(exmns, mean(run))
}

par(mfrow = c(1, 1))

plot(exvar,
     xlab = "sample size",
     ylab = "value",
     col = "blue",
     type = "l",
     main = "Convergence of Mean and Variance of Exponentials")

lines(exmns,
      col = "red",
      type = "l")

legend("topright",
       col = c("red", "blue"),
       lwd=1,
       bty="n",
       legend=c("mean", "variance"))
```
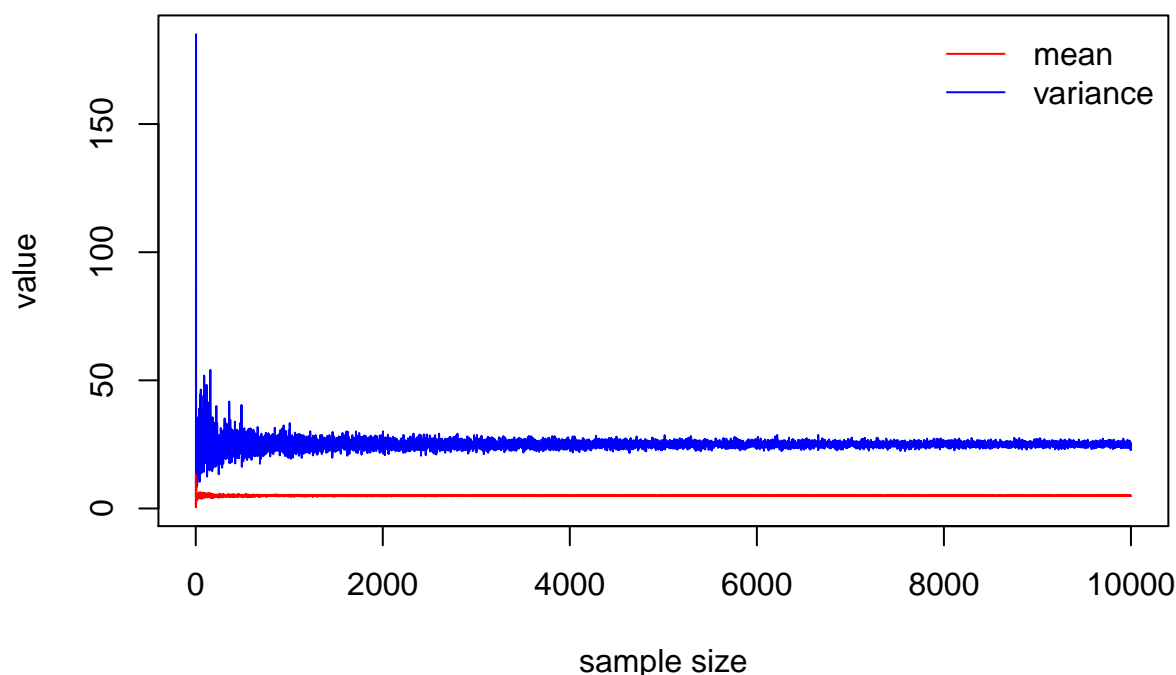
Convergence of Mean and Variance of Exponentials

As expected, the mean and variance of exponentials with λ equal to 0.2 converge to 5 and 25 respectively, which were the theoretical values computed in the *Motivation* section, under *Simulation*.

We will now analyze the results for averages of an exponential random variable. The following plot explores both the center of mass and standard error convergence for increasing number of simulation runs, and the center of mass and standard error convergence for increasing sample size.

```
# for increasing size

xstd = NULL
xmns = NULL

for (i in 1:1000) {
  mns = runsim(n = i, itrs = 1000)
  xstd = c(xstd, theoretical_sd / sqrt(i))
  xmns = c(xmns, mean((mns - theoretical_mean) / xstd[i]))
}

# for increasing runs

zstd = NULL
zmns = NULL

for (i in 1:1000) {
  mns = runsim(n = 40, itrs = i)
  zstd = c(zstd, theoretical_sd / sqrt(40))
  zmns = c(zmns, mean((mns - theoretical_mean) / zstd[i]))
```

```r
}

par(mfrow=c(1, 2))

plot(xmns,
     xlab = "sample size",
     ylab = "value",
     ylim = c(0, 5),
     col = "red",
     type = "l",
     main = "Increasing Size")

lines(xstd,
     col = "blue",
     type = "l")

legend("topright",
       col=c("red", "blue"),
       lwd=1,
       bty="n",
       legend=c("mean", "std. err"))

plot(zmns,
     xlab = "iterations",
     ylab = "value",
     ylim = c(0,5),
     col = "red",
     type = "l",
     main = "Increasing Runs")

lines(zstd,
     col = "blue",
     type = "l")

legend("topright",
       col=c("red", "blue"),
       lwd=1,
       bty = "n",
       legend=c("mean", "std. err"))
```
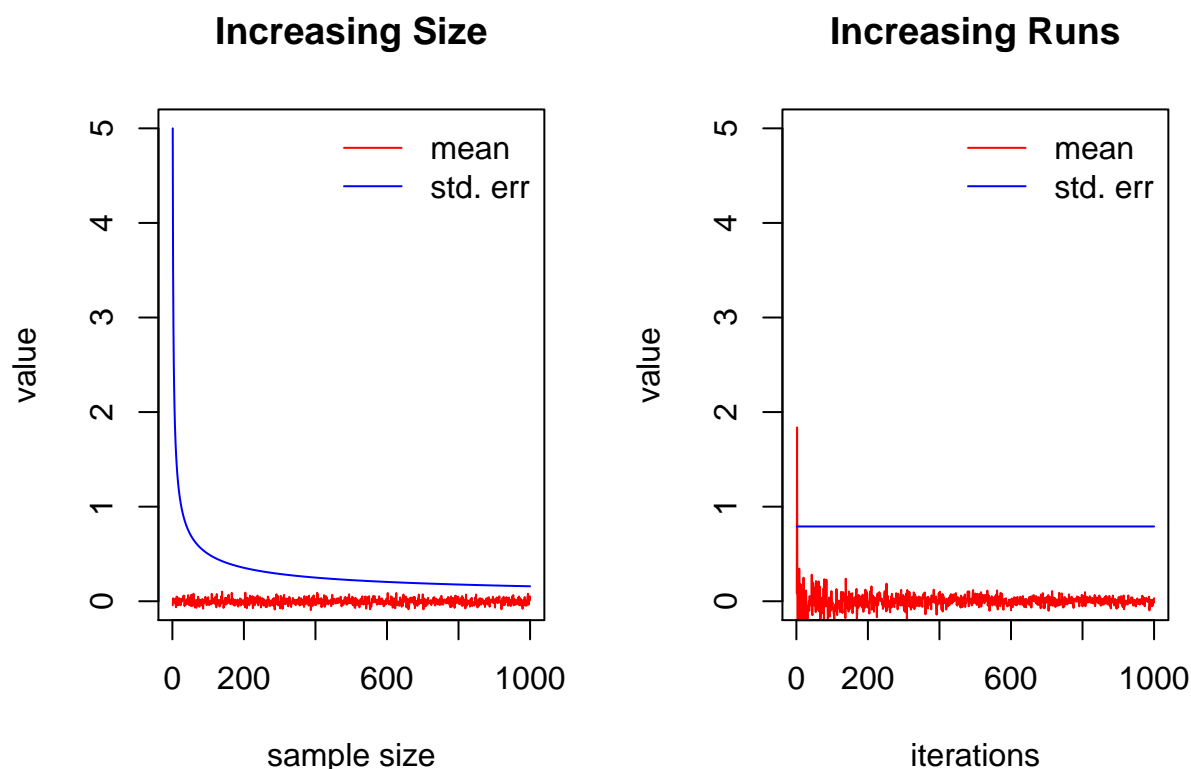
Increasing Size / **Increasing Runs**

The plots above show that the mean and standard error for the averages of an exponential random variable with λ equal to 0.2 converge with increasing size and simulation runs. The plot to the left shows how the normalized sample mean and standard error change as the sample size increases, with the number of simulation runs held constant at 1000. The plot to the right shows how the normalized sample mean and standard error change as the number of simulation runs increase, with the sample size held constant at 40. Notice that the standard error never changes on the right-hand plot. This shows that the sample size dominates the distribution parameters, yet an adequately large number of simulation runs is needed to accurately approximate a standard normal distribution. For increasing sample size, the standard error decreases to zero, as expected, while the normalized mean of the arithmetic means of exponentials fluctuates around 0.

To conclude our analysis, we showed that the Central Limit Theorem holds for a large sample of averages, with the sample parameters limiting to the parameters of a standard normal distribution.