An Analysis of the Mean of Many Exponential Random Variables

Matthew Halbasch

February 25, 2019

Overview

In this paper we will conduct 1000 simulations of 40 exponential random variables each, and compute their mean in each simulation. We will examine the distribution of the mean, and compare it to the theoretical distribution we expect. In particular, we will show that the mean of these 40 random variables is approximately normally distributed.

Simulations

First, we simulate the exponential random variables using the `rexp` function. These variables follow a probability density function,

$$f(x) = \lambda e^{-\lambda x},$$

with λ the rate parameter. The mean and variance of the distribution are given by

$$\mathrm{E}[X] = \frac{1}{\lambda} \quad , \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

The code for simulating these variables is below. In particular, we simulate 40,000 random exponential variables and organize them into a matrix with 1000 rows and 40 columns. We will use a rate parameter $\lambda = 0.2$.

```
set.seed(1032) # We set the seed so the results are reproducible
M = matrix(rexp(40000, rate=0.2), nrow=1000, ncol = 40)
```

Now we calculate the mean of each row of 40 variables, to give us a set of 1000 means to compare to the theoretical distribution of the mean.

```
means <- apply(M, MARGIN=1, mean)
```

Sample Mean versus Theoretical Mean

The theoretical mean of our distribution of exponentials is given above as $\mu = 1/\lambda$. In our case, $\lambda = 0.2$, so we should expect a theoretical mean of $\mu = 5$. Figure 1 shows the empirical distribution of the mean via a density plot:

```
library(ggplot2)

ggplot(data=as.data.frame(means), aes(x=means)) +
  geom_density(fill = 'deepskyblue3', alpha=0.4) +
  geom_vline(xintercept=5, color = 'tomato', linetype=2) +
  xlab("Mean") +
  ylab("Density") +
  labs(title = "Density of the Empirical Mean") +
  theme_minimal(10)
```

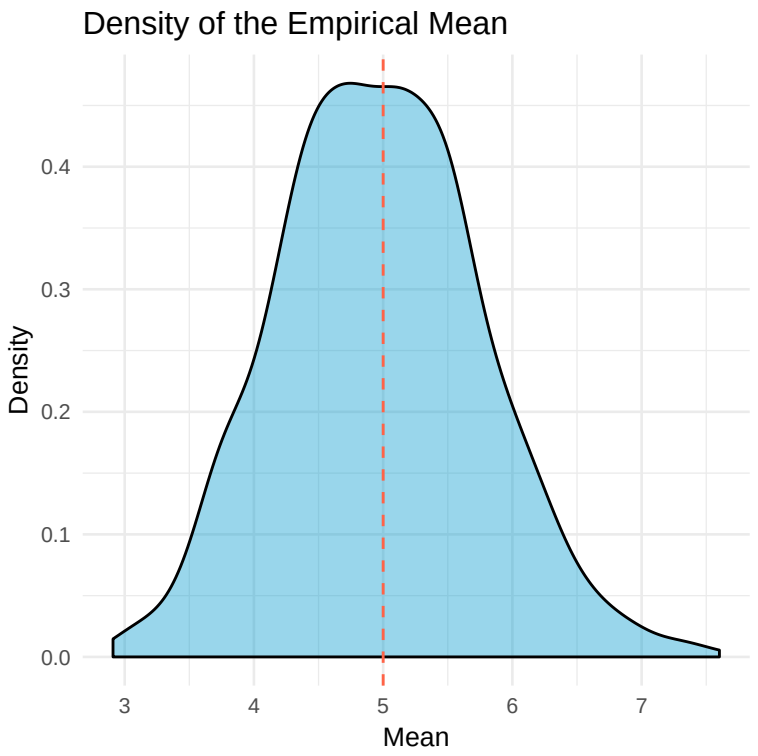

Figure 1: A density plot of the mean of 40 random exponential variables. The dashed line shows the theoretical mean of 5

Figure 1 shows that the empirical mean distribution over these 1000 simulations is concentrated near the theoretical mean of 5, but has a significant spread from this value. We will return to this distribution in the Distribution section.

Sample Variance Versus Theoretical Variance

As explained above, for the exponential distribution we expect a variance of $1/\lambda^2 = 25$ for points drawn from this distribution. We would expect that the sample variance of each group of 40 random variables is near this value of 25. We have plotted the distribution of sample variances in Figure 2.

```
vars = apply(M, MARGIN=1, var)
ggplot(data = as.data.frame(vars), aes(x=vars)) +
  geom_density(fill = 'deepskyblue3', alpha=0.4) +
  geom_vline(xintercept = 25, color='tomato', linetype=2) +
  theme_minimal(10) +
  xlab("Sample Variance") +
  ylab("Density") +
  labs(title = "Density of the Sample Variance")
```

What we find in Figure 2 is that the variance actually peaks before this theoretical value, and has a very long tail to the right hand side. We see that the variance is certainly not normally distributed, as it is not a simple average of random variables. From the plot, it seems that the mean of the distribution is potentially close to the theoretical value of 25, while the mode is significantly smaller.

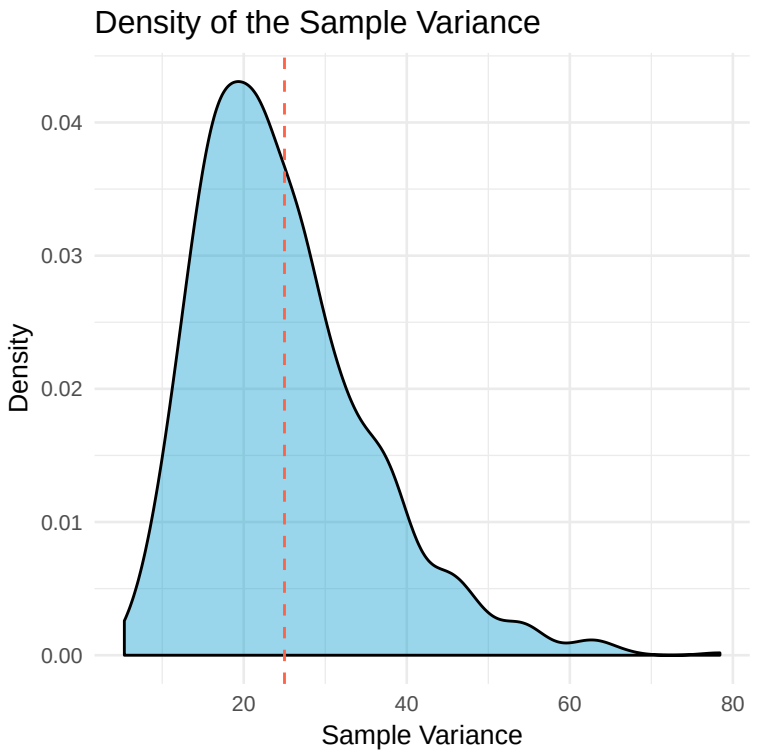

Figure 2: A density plot of the sample variance of each group of 40 random variables. The dashed line shown is the theoretical variance of the underlying exponential distribution. We see that the variance is peaked slightly before this value, but has a very long tail off to the right.

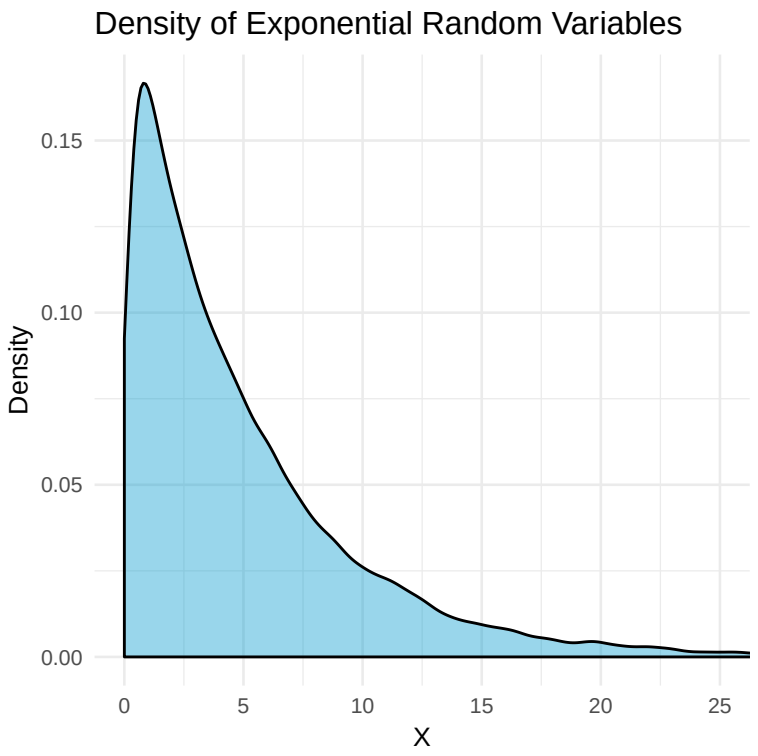

Figure 3: A plot of the distribution of the exponential random variables themselves. We see that these follow an exponential distribution, as we would expect.

Distribution

The distribution of the mean of the 40 exponential random variables seen in Figure 1 looks to be approximately Normal. This is in contrast to the distribution of the random variables themselves, as seen in Figure 3.

```
exp_vars <- as.numeric(M)

ggplot(data=as.data.frame(exp_vars), aes(x=exp_vars)) +
  geom_density(fill='deepskyblue3', alpha=0.4) +
  xlab("X") +
  ylab("Density") +
  labs(title = "Density of Exponential Random Variables") +
  coord_cartesian(xlim=c(0,25)) +
  theme_minimal(10)
```

The sample mean follows a different distribution from the variables themselves because it is an average of many independent random variables. Some intuition for this phenomenon comes from dice: when rolling two dice, the 7 is much more common than the 2 because many different rolls can give a 7, despite every number being equally likely on a single die. Similarly, for the mean of many different random variables, the numbers close to the sample mean become more likely as the fringe values require many unlikely draws to be realized.

The central limit theorem tells us that in the asymptotic case of infinitely many samples, the sample mean will follow a normal distribution according to

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) ,$$

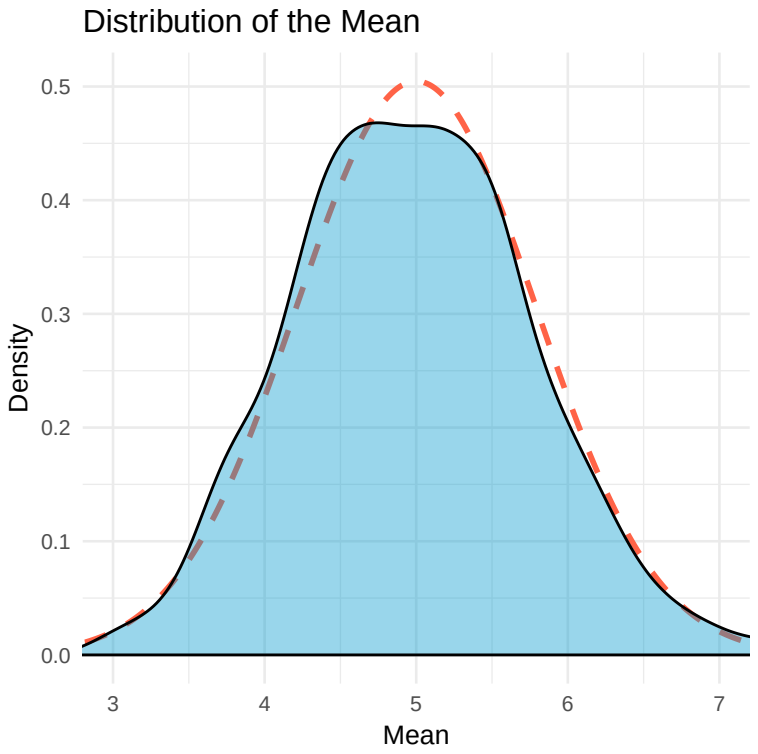

Figure 4: A plot showing the theoretical asymptotic distribution of the mean (dashed red) and the empirical distribution of the sample mean (blue). We see that the empirical distribution approximates the theoretical distribution, but is not exactly the same.

where μ is the population mean of X_i and σ is the population standard deviation. In our case, we know the population mean is $\mu = 5$ and the population standard deviation is $\sigma = 5$ as well. So, using $n = 40$, we can compare this theoretical distribution of the mean to our empirical distribution of the mean we depicted in Figure 1. This comparison is shown in Figure 4.

```
xs = 0:10000/1000
vals <- dnorm(xs, mean=5, sd = 5/sqrt(40))
norms <- as.data.frame(cbind(xs,vals))
ggplot() +
  geom_path(data = norms, aes(x=xs, y=vals),
            col="tomato", linetype=2, size=1) +
  geom_density(data = as.data.frame(means),
               aes(x=means), col="black", fill='deepskyblue3',
               alpha=0.4) +
  coord_cartesian(xlim=c(3,7)) +
  xlab("Mean") +
  ylab("Density") +
  labs(title = "Distribution of the Mean")+
  theme_minimal(10)
```

Figure 4 shows that the emirical and theoretical distributions of the mean are indeed similar, though the empirical distribution is not exactly equal to the asymptotic distribution.

If we were to take more and more samples, these distributions would get closer to each other according to the Central Limit Theorem.