Central Limit Theorem

Normand Desmarais

January 27, 2016

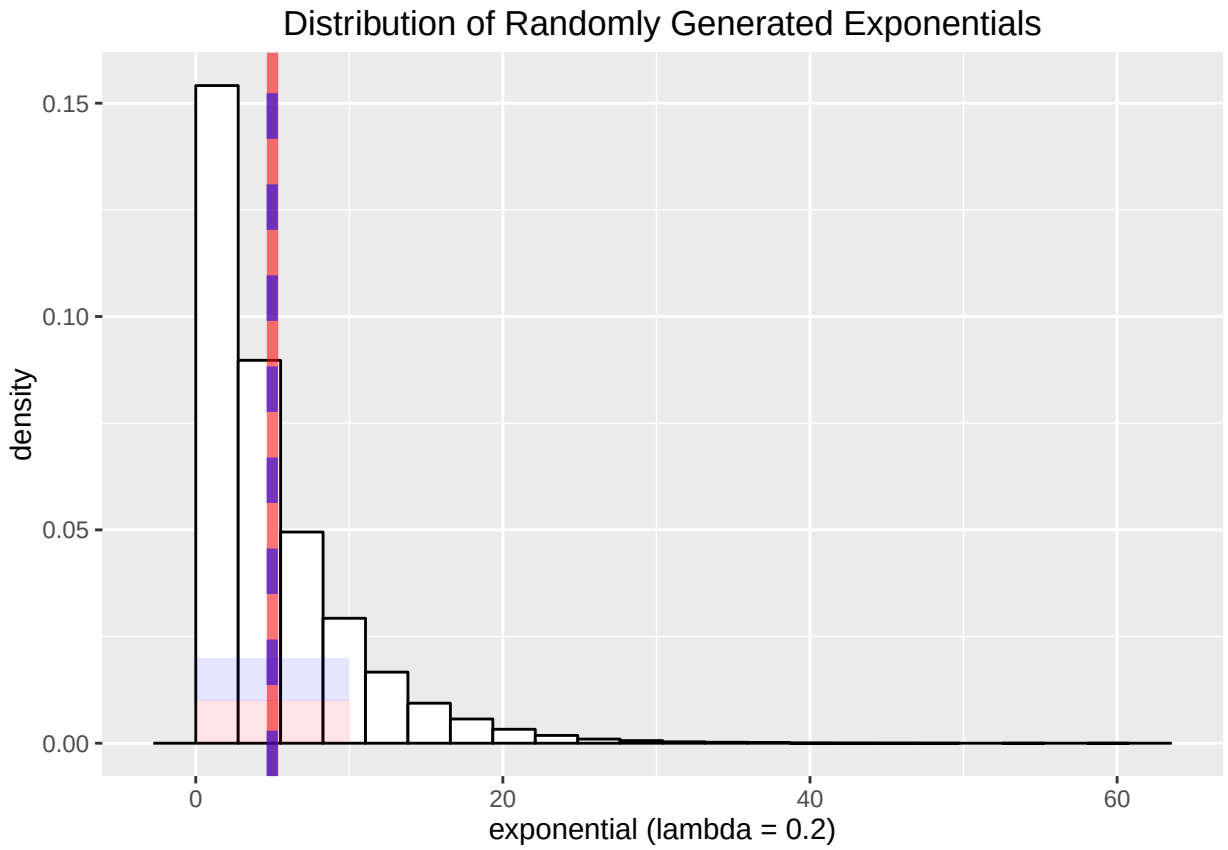
In this document we are going to demonstrate the Central Limit Theorem, namely that the distribution of the averages of randomly generated sets of iid variables follows a normal distribution as the sample size increases. In our study we will focus our attention on the distribution of randomly generated exponentials.

Let's first generate $1000 \times 40 = 40000$ exponentials with the parameter λ set to 0.2. We have plotted the resulting distribution in the following histogram. We'll describe the meaning of the vertical line as well as the shaded rectangles in the following sections.

```
set.seed(5319)
n = 40
m = 1000
lambda = 0.2

sample_data <- rexp(n * m, lambda)  # generate the sample data
sample_mean <- round(mean(sample_data), 3)  # compute the sample mean
sample_sd <- round(sd(sample_data), 3)  # compute the sample standard deviation
sample_var <- round(sd(sample_data)^2, 3)  # compute the sample variance

# need a data frame for ggplot
theDataF <- as.data.frame(sample_data)
names(theDataF) <- "exp"
binWidth <- diff(range(theDataF$exp))/21  # generate 21 bins in histogram
g <- ggplot(theDataF, aes(x = exp))
g <- g + geom_histogram(binwidth = binWidth, fill = "white", colour = "black",
    aes(y = ..density..))
g <- g + annotate("rect", xmin = 0, xmax = 10, ymin = 0, ymax = 0.01, alpha = 0.1,
    fill = "red")
g <- g + annotate("rect", xmin = sample_mean - sample_sd, xmax = sample_mean +
    sample_sd, ymin = 0.01, ymax = 0.02, alpha = 0.1, fill = "blue")
g <- g + geom_vline(xintercept = 1/lambda, colour = "red", size = 2, alpha = 0.55)
g <- g + geom_vline(xintercept = sample_mean, colour = "blue", linetype = 2,
    size = 2, alpha = 0.55)
g <- g + ggtitle("Distribution of Randomly Generated Exponentials")
g <- g + xlab("exponential (lambda = 0.2)")
g
```



1. Sample mean vs Theoretical mean

The dashed red line in the above histogram represents the position of the theoretical average ($\mu = 1/\lambda = 5$). The dashed purple line represents the position of the sample mean $\bar{X} = 4.976$. We can clearly see that they both overlap, confirming that our sample of randomly generated 40000 exponentials has its mean approximately centered at $1/\lambda$ as theoretically expected.

2. Sample variance vs Theoretical variance

Before considering the variance, let's consider first its square root, namely the standard deviation. The shaded red rectangle in the above histogram represents one theoretical standard deviation ($\sigma = 1/\lambda = 5$) from the theoretical mean on both sides of the mean. The height of the rectangle just serves visual purpose and has no meaning. Similarly, the blue shaded rectangle represents one sample standard deviation ($S = 4.975$) from the sample mean. We can clearly see that they both overlap, confirming that our sample of randomly generated 40000 exponentials has its standard deviation spread as $1/\lambda$ (as theoretically expected).

The variance is simply the square of the standard deviation. Hence we can see that $S^2 = 24.753 \approx 25 = 1/\lambda^2 = \sigma^2$. We would get a better approximation by increasing the size of the sample.

3 The distribution of averages

We now want to validate the CLT, which, once again, tell us that the distribution of averages of a large number of samples follow a standard normal distribution. Hence:

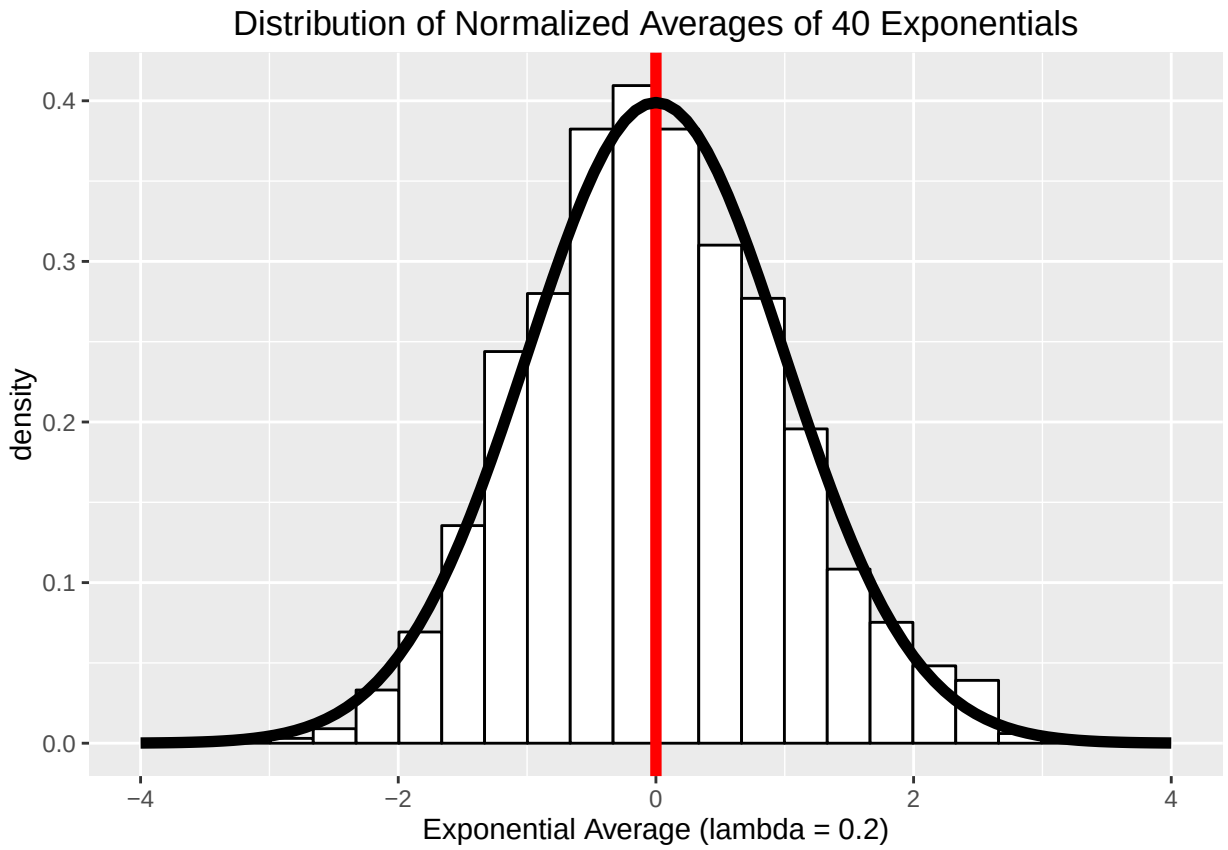
- We will divide our previously generated sample of 40000 exponentials into a thousand samples of 40 exponentials.
- For each new sample of 40 exponentials, we will compute the mean value (that we will call X_{40}).
- This will give us a 1000-long vector of averages (of 40 exponentials).
- We will compute the mean ($\bar{X}_{40}$) of this new vector as well as its standard deviation (S_{40}).
- We will use these two values to renormalize our vector of average: **normalized**$(X_{40,i}) = (X_{40,i} - \bar{X}_{40})$ / S_{40}; $\quad i = 1, 1000$
- Finally we will plot an histogram of the distribution of the normalized averages and compare it to a standard normal distribution.

```
# divide the sample of 40000 into 1000 row of 40
theDataF <- as.data.frame(rowMeans(matrix(sample_data, m, n)))
names(theDataF) <- "X_40"

# compute the mean and sd of this new distribution
mean_X_40 <- mean(theDataF$X_40)
sd_X_40 <- sd(theDataF$X_40)

# normalize the original distribution
theDataF <- mutate(theDataF, X_40 = (X_40 - mean_X_40)/sd_X_40)

# and let's plot the histogram
binWidth <- diff(range(theDataF$X_40))/21
g <- ggplot(theDataF, aes(x = X_40))
g <- g + geom_histogram(binwidth = binWidth, fill = "white", colour = "black",
    aes(y = ..density..))
g <- g + geom_vline(xintercept = 0, colour = "red", size = 2) + xlim(-4, 4)
g <- g + stat_function(fun = dnorm, size = 2)
g <- g + ggtitle("Distribution of Normalized Averages of 40 Exponentials")
g <- g + xlab("Exponential Average (lambda = 0.2)")
g
```



In the above histogram, the thick black curve represents a standard normal distribution. The red vertical line indicates the mean of the standard normal (= 0).

We can clearly see that our normalized distribution of averages (represented by the white bins) closely approximate a standard normal distribution and is centered around 0, just as stated by the Central Limit Theorem. In comparison, the distribution of exponential (as shown in the first graph) doesn't follow a normal distribution (it follows an exponential distribution - what else did you expect). And that's the beauty of the CLT: no matter how the original distribution looks like, as long as it is constituted of iid variables, the normalized distribution of mean of a large number of samples will always approximate a standard normal distribution.