Analysis of the Exponential Distribution

Charles Wylie - August 25, 2016

Statistical Inference Course Project 1

Overview

In this project we investigate the exponential distribution in R and compare it with the Central Limit Theorem. The Central Limit Theorem states that the means of a large number of iterations of independent random variables of any type of distribution will be approximately normally distributed. We will show that the distribution of means of random exponential samples behaves as predicted by the Central Limit Theorem.

Simulation

The exponential distribution can be simulated in R with rexp(n, lambda), where lambda is the rate parameter. The mean of an exponential distribution is 1/lambda and the standard deviation is also 1/lambda; Lambda is set at 0.2 for this simulation. We will investigate the properties of the distribution of means of one thousand samples of 40 random exponentials.

We will:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

In point 3, we will focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

The following code will generate a data frame "sim" of 1000 samples of 40 random exponential variables, and append a column "means" with the arithmetic mean of each of the 1000 sample observations. Since we are generating random variables, we set a seed so that the experiment can be reproduced:

```
set.seed(3737)
n = 40
B = 1000
lambda = 0.2
sim <- data.frame(matrix(rexp(n*B, lambda), B, n))
sim$means <- apply(sim, 1, mean)
```

Sample Mean vs. Theoretical Mean

We now take the mean of our 1000 sample means. The mean of an exponential distribution is 1/lambda. If the Central Limit Theorem is true then the mean of our simulation should be near the theoretical value of 5.0.

```
round(mean(sim$means), 3)
```

```
## [1] 5.002
```

Sample Variance vs. Theoretical Variance

Similarly, the variance of the 1000 sample means will be near the theoretical value 1/lambda^2/n = 0.625.

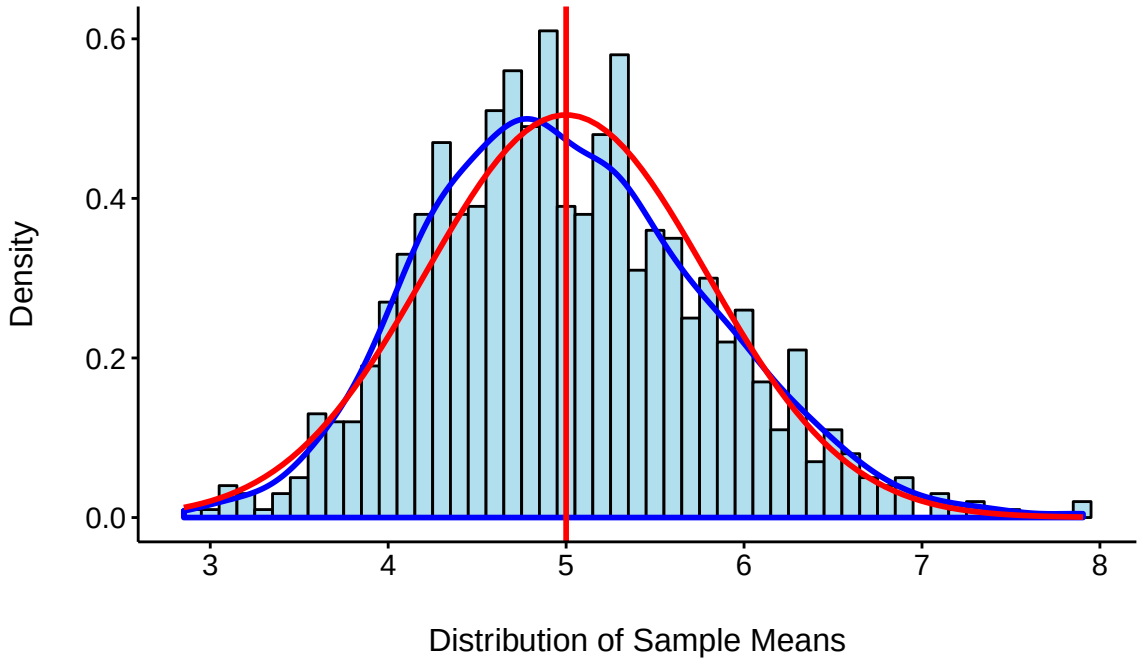
```
round(var(sim$means), 3)
```

```
## [1] 0.625
```

Distribution

Figure 1 shows a histogram of the distribution of means of our exponential samples (blue), with a graph of the normal distribution (the red lines) overlaid. We see that the exponential distribution approximates the normal distribution, as stated by the Central Limit Theorem, though the exponentials appear to be somewhat skewed to the right, that is, the right tail is longer than the left tail.

Figure 1. Exponential Distribution vs. Normal Distribution

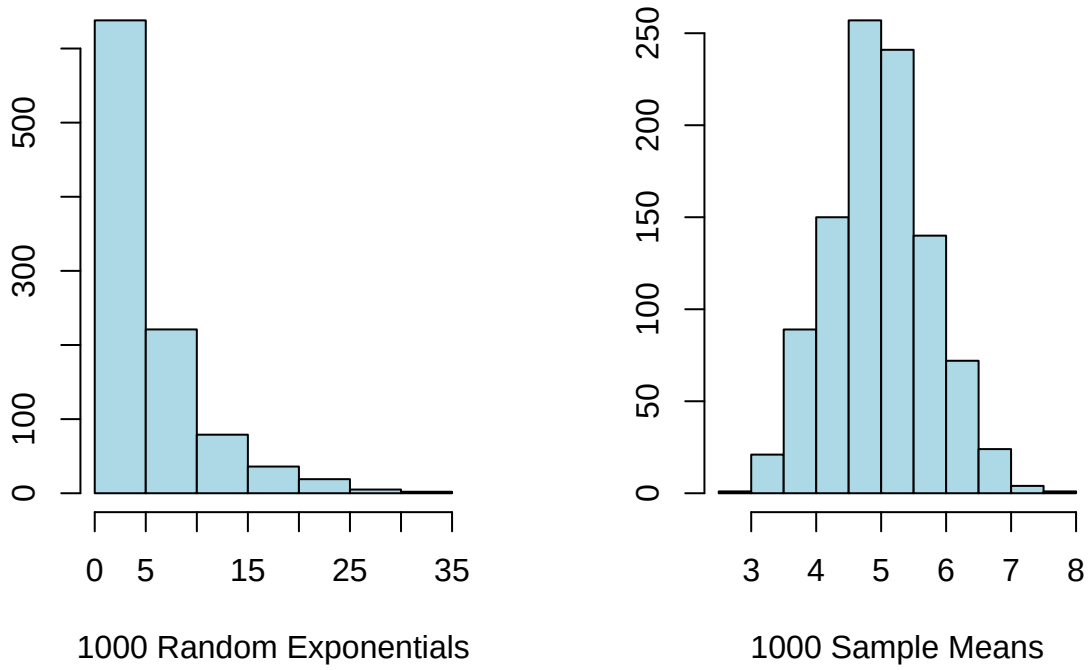


Comparison of Exponential Variables vs. the Means of Samples of Exponential Variables

Finally, we show the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials. We use the rexp() function with lambda = 0.2 as in the first experiment.

Figure 2 shows that the distribution of random exponential variables is, unsurprisingly, exponential, while the distribution of a large number of averages of samples of exponential variables is normal, and thus obeys the Central Limit Theorem. The R code for Figures 1 and 2 is shown in the appendix.

Figure 2. Comparison



Conclusion

The Central Limit Theorem states that the means of a large number of iterations of independent random variables of any type of distribution will be approximately normally distributed. Though the exponential distribution is not itself normal, we have shown that the mean values of a collection of random samples of variables of the exponential distribution are normally distributed and do obey the Central Limit Theoriem.

Appendix

R Code for Figure 1

```
set.seed(3737)
n = 40
B = 1000
lambda = 0.2
sim <- data.frame(matrix(rexp(n*B, lambda), B, n))
sim$means <- apply(sim, 1, mean)

library(ggplot2)

fig1 <- ggplot(sim, aes(means)) +
    geom_histogram(aes(y = ..density..), binwidth = 0.1, fill = 'lightblue2',
                   color = 'black') +
    geom_density(color = "blue", size = 1) +
```

```
    stat_function(fun = dnorm, args = list(mean = 5, sd = sqrt(0.625)), color = 'red',
                  size = 1) +
    geom_vline(xintercept = 5, size = 1, color = "red") +
    labs(x = 'Distribution of Sample Means') +
    labs(y = 'Density') +
    labs(title = 'Figure 1. Exponential Distribution vs. Normal Distribution') +
    theme_bw(base_size = 10) +
    theme(axis.line.x = element_line(colour = "black"),
    axis.line.y = element_line(colour = "black"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank()) +
    theme(plot.margin = unit(c(1, 1, 1, 0.6), 'cm')) + # top, right, bottom, left
    theme(plot.title = element_text(margin = margin(b = 0.6, unit = 'cm'))) +
    theme(axis.title.x = element_text(size = 12, margin = margin(20,0,0,0))) +
    theme(axis.title.y = element_text(size = 12, margin = margin(0,20,0,0))) +
    theme(axis.text.x = element_text(size = 11)) +
    theme(axis.text.y = element_text(size = 11))

print(fig1)
```

R Code for Figure 2

```
exponentials = (rexp(1000, 0.2))
averages = NULL
for (i in 1 : 1000) averages = c(averages, mean(rexp(40, 0.2)))
fig2data <- data.frame(cbind(exponentials, averages))

par(mfrow=c(1,2))
hist(fig2data$exponentials, main=paste('Figure 2. Comparison'), font.main=1, cex.main=1,
     col='lightblue', xlab='1000 Random Exponentials', ylab='')
hist(fig2data$averages, main=paste(''), col='lightblue', xlab='1000 Sample Means',
     ylab='')
```