Statistical Inference Project - Part 1 Simulation Excercise

ALIU

September 11, 2016

Overview:

In this report, we will analyze the result of 1,000 simulations of exponential distributions and illustrate that as the Central Limit Theorem porangeicts, the sample mean and sample variance produced by the simulations is very similiar to the theoretical mean and variance. Additionally, the distribution of the sample means is much more Gaussian than the original distributions.

Simulations:

Simulate 1,000 exponential distributions with the sample size (n=40) and lambda (rate= 0.2):

- set.seed will set the seed of the simulation so that the results are reproducible
- rexp(40,rate=0.2) gives us 1 simulation of exponetial distribution with the sample size 40 and rate 0.2
- a "for" loop is used to repeat the simulation 1,000 times
- "rbind" is used to combine the result of each simulation into a data frame called "exp", with each row containing the results from one simulation.

```
set.seed(17)
exp <- NULL
for (i in 1:1000) {exp <- rbind(exp, rexp(40,rate=0.2))}
```

Sample Mean versus Theoretical Mean:

The Theorectical Mean of an exponential distribution is given by:

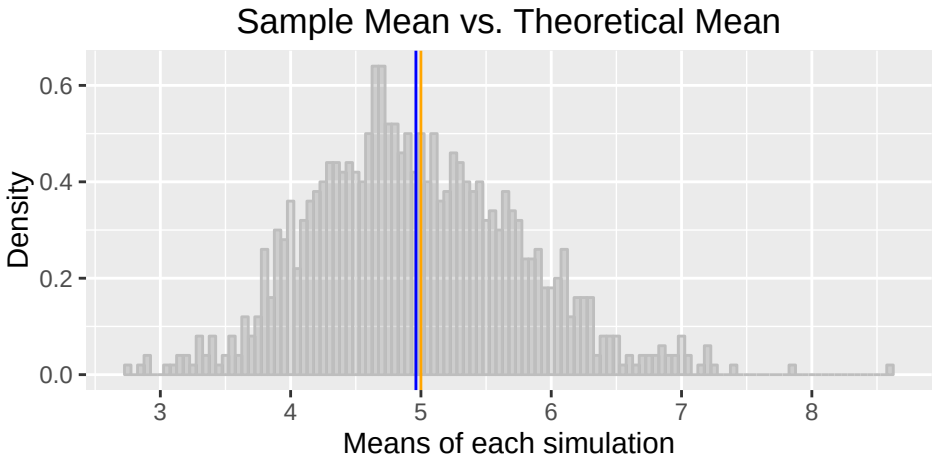
- $1/lambda = 1/0.2 = 5$

Take the mean of results of each sumulation "exp" will give us the 1,000 sample means ("exp_mean"). The average of the sample means is calculated by mean(exp_mean) and the output below compares the values of the two means

```
theo_mean <- 1/0.2
exp_mean <- apply(exp,1, mean)
sample_mean <- mean(exp_mean)
print(cbind(theo_mean,sample_mean))
```

```
##      theo_mean sample_mean
## [1,]         5    4.961683
```

The plot below shows the sample means simulated in this exercise with the Sample Mean (blue vertical line, 4.9616831) and Theoretical Mean (orange vertical line, 5). We can see that the two vertical times almost overlap with each other.



Sample Variance versus Theoretical Variance:

The Theoretical Variance is given by:

- $(1/lambda)^2 = 25$

The sample variance can be obtained by:

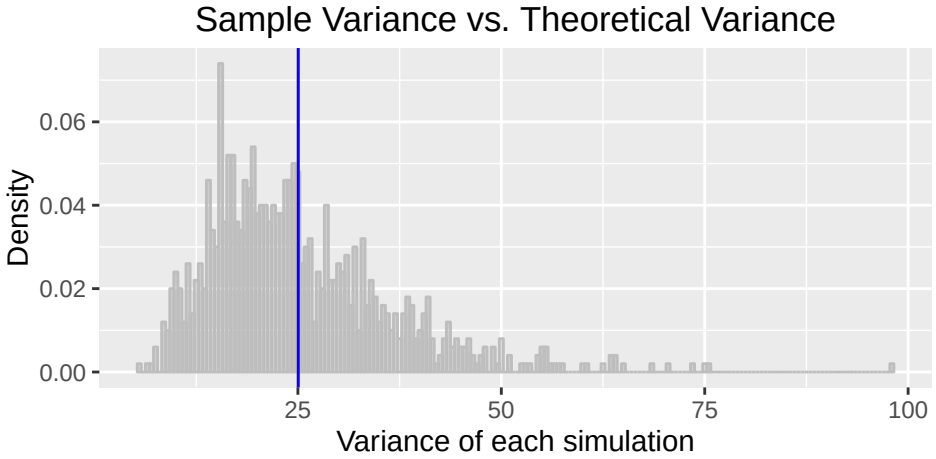
- Calcualte the variance of results of each sumulation "exp"
- Take the mean of variances of each sample will give us the sample variance ("sample_var")

The output below compares the values of the two variances.

```
theo_var <- (1/0.2)^2
exp_var <- apply(exp,1,var)
sample_var <- mean(exp_var)
print(cbind(theo_var,sample_var))
```

```
##      theo_var sample_var
## [1,]       25   25.04348
```

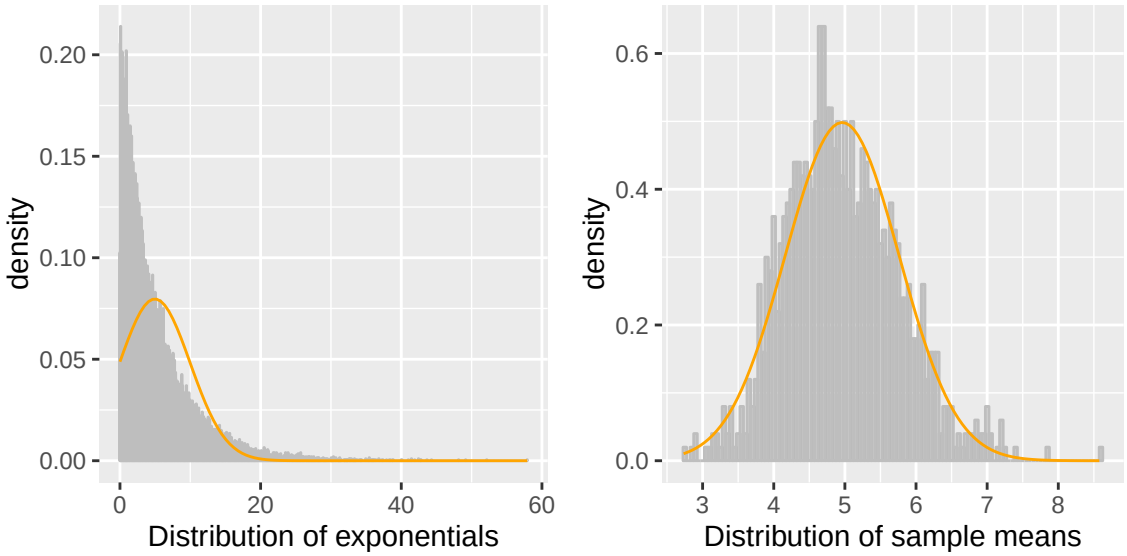
The plot below shows the sample variances simulated in this exercise with the Sample Variance (blue vertical line, 25.0434773) and Theoretical Variance (orange vertical line, 25). We can see that the two vertical times almost overlap with each other.



Distribution:

To verify that the distribution of the sample means is approximately normal, we 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 plot compares the distribution of averages and the distribution of the random exponentials
- We can see that the distribution of the averages of the exponentials is far more Gaussian than the original distributions, with the orange curve showing how a normal distribution of same mean and standard deviation looks like



Appendix

The codes for producing the three plots are as follows:

Plot of Theoretical Mean vs. Sample Mean

```
exp_mean <- as.data.frame(exp_mean)
p1 <- ggplot(data = exp_mean, aes(x = exp_mean)) +
  geom_histogram(alpha = 0.2, binwidth=0.05,
                 color="gray", aes(y = ..density..)) +
  geom_vline(xintercept=theo_mean, color="orange") +
  geom_vline(xintercept=sample_mean, color="blue") +
  labs(x = "Means of each simulation", y = "Density") +
  scale_x_continuous(breaks=c(3,4,5,6,7,8)) +
  ggtitle("Sample Mean vs. Theoretical Mean")
p1
```

Plot of Theoretical Variance vs. Sample Variance

```
g_var <- ggplot() + aes(x = exp_var) +
  geom_histogram(alpha=0.2, binwidth=0.5,
                                color="gray", aes(y=..density..)) +
  geom_vline(xintercept=theo_var, color="orange") +
  geom_vline(xintercept=sample_var, color="blue")+
  labs(x = "Variance of each simulation", y = "Density") +
  scale_x_continuous(breaks=c(0,25,50,75,100)) +
  ggtitle("Sample Variance vs. Theoretical Variance")
g_var
```

Plot of Distribution of Exponentials vs. Distribution of averages

```
g_mns <- ggplot() + aes(x = exp_mean) +
  geom_histogram(alpha = 0.2, binwidth=0.05,
                 color="gray", aes(y = ..density..)) +
  stat_function(fun=dnorm, color = "orange",
                args=list(mean=mean(exp_mean$exp_mean),
                          sd=sd(exp_mean$exp_mean))) +
  labs(x="Distribution of sample means")

g_exp <- ggplot() + aes(x = as.vector(exp)) +
  geom_histogram(alpha = 0.2, binwidth=0.05,
                 color="gray", aes(y = ..density..)) +
  stat_function(fun=dnorm, color = "orange",
                args=list(mean=mean(exp),
                          sd=sd(exp))) +
  labs(x="Distribution of exponentials")

grid.arrange(g_exp, g_mns,nrow=1)
```