Exponential Distribution and Central Limit Theorem

Ivana Lipnerová

December 20, 2015

Overview

The goal of this project is to investigate exponential distribution compared to distribution of averages of 40 exponentials, as these according to Central Limit Theorem should approximate normal distribution. I will test this via 1000 simulations of drawing 40 random exponentials with set parametr lambda ($\lambda = 0.2$). In the instruction there is useful information about the mean of exponential distribution (mean = $1/\lambda$) and the standard deviation of same distribution (sd = $1/\lambda$).

Simulations

Setting the environment

```
#load necessary libraries:
library(ggplot2)
library(knitr)
library(gridExtra)

# Set the lambda to 0.2:
lambda <- 0.2

# Set the number of simulations:
n_simulations <- 1000

# Set the number of drawn samples:
n <- 40

# Set random seed, so that the conclusion will be valid for every run of knitr:
set.seed(12345)
```

Simulation

```
# Random exponential variables:
sim_vars <- matrix(data=rexp(n*n_simulations, rate=lambda), nrow=n_simulations, ncol=n)

# Means:
sim_means <- rowMeans(sim_vars)
```

Sample Mean versus Theoretical Mean

Show the sample mean and compare it to the theoretical mean of the distribution.

The mean of simulated averages can be safely computed by base function mean(). As was stated in overview, the theoretical mean of exponential distribution is $1/\lambda$, with $\lambda = 0.2$.

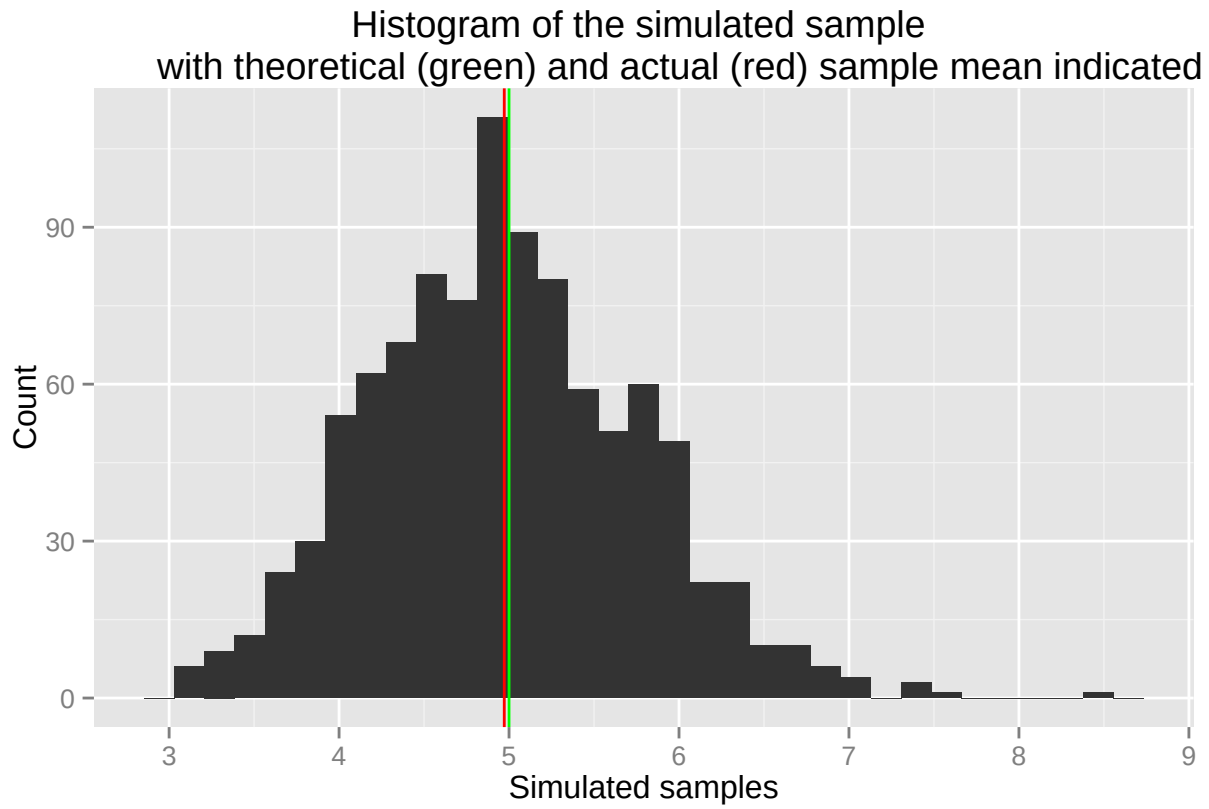
```r
means<-data.frame("Mean"=c(mean(sim_means), 1/lambda),
                  row.names=c("Sample mean", "Theoretical mean"))

kable(x=means, digits=3, align="c", caption="Table of sample and corresponding theoretical mean")
```

Table 1: Table of sample and corresponding theoretical mean

	Mean
Sample mean	4.972
Theoretical mean	5.000

As we can see, the actual sample mean is not so far from the theoreticaly predicted one. Let's see how it stands in histogram of simulated data.

```r
ggplot(as.data.frame(sim_means), aes(sim_means)) +
  geom_histogram() +
  geom_vline(xintercept=mean(sim_means), colour="red") +
  geom_vline(xintercept=1/lambda, colour="green") +
  labs(title="Histogram of the simulated sample
       with theoretical (green) and actual (red) sample mean indicated",
       x="Simulated samples", y="Count")
```



Sample Variance versus Theoretical Variance

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. According to Wikipedia (https://en.wikipedia.org/wiki/Variance#Exponential_distribution), the variance of exponential distributed random variable is equal to its mean to the power of two: $\sigma^2 = \mu^2$.

```
var_sim <- (mean(sim_means))^2
var_theo <- (1/lambda)^2


vars<-data.frame("Variances"=c(var_sim, var_theo),
                 row.names=c("Sample variance", "Theoretical variance"))

kable(x=vars, digits=3, align="c", caption="Table of sample
      and corresponding theoretical variance")
```

Table 2: Table of sample and corresponding theoretical variance

	Variances
Sample variance	24.721
Theoretical variance	25.000

Distribution

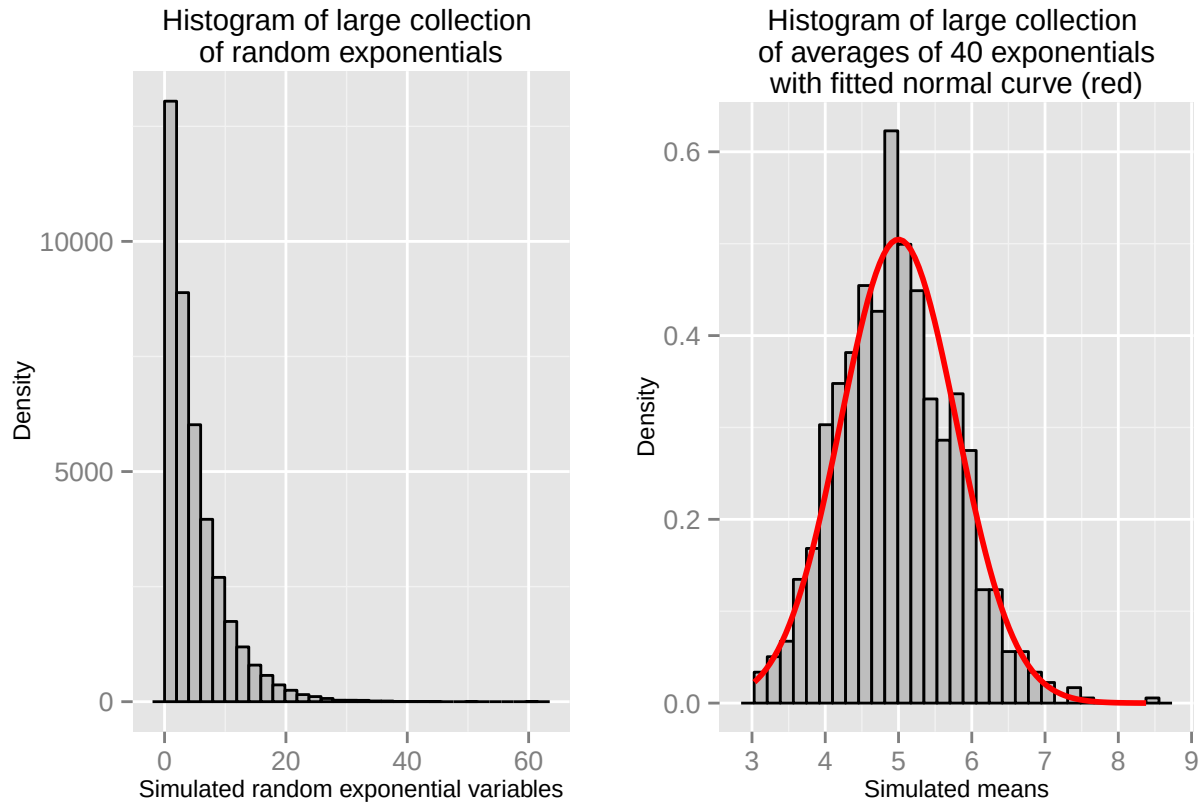
Show that the distribution is approximately normal. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

I decided to look into this question by showing the fit of normal curve to the histogram (see plot below) and by testing the normality of averages by Shapiro-Wilk test of normality (at the bottom of the document).

For the first part, there are two histograms. The left one shows the original simulated data (thus showing them as exponentials), the right one shows the drawn averages of 40 exponentials with fitted normal curve for theoretical normal distribution given set lambda.

```
# Plot of raw simulated data:
p1 <- ggplot(as.data.frame(as.vector(sim_vars)), aes(as.vector(sim_vars))) +
  geom_histogram(aplha=0.5, fill="grey", col="black") +
  ggtitle("Histogram of large collection
of random exponentials") +
  xlab("Simulated random exponential variables") +
  ylab("Density") +
  theme(title = element_text(size = rel(0.75), hjust = 0.5))

# Plot of averages:
p2 <- ggplot(as.data.frame(sim_means), aes(sim_means)) +
  geom_histogram(aes(y=..density..), aplha=0.5, fill="grey", col="black") +
      #that aes() necessary for normal curve displayed correctly
  stat_function(fun=dnorm, colour="red", lwd=1,
                args=list(mean = 1/lambda, sd = sqrt(var_theo/n))) +
      #necessary  manualy set mean and sd for a curve
  labs(title="Histogram of large collection
```

```
of averages of 40 exponentials
with fitted normal curve (red)",
       x="Simulated means",
       y="Density") +
  theme(title = element_text(size = rel(0.75), hjust = 0.5))

# Arrange them:
grid.arrange(p1, p2, ncol=2, nrow=1)
```



The plots are ilustrative, though at least simple test for normality should be done. In Shapiro-Wilk test of normality the p-value lower than 0.01 indicates non-normal distribution of tested data.

```
shapiro.test(sim_means)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sim_means
## W = 0.99438, p-value = 0.0008625
```

Thus I would conclude that the tested collection of averages of 40 exponentials is just approximately normal. Given the lambda and setted random seed, to achieve normality of averages one should use larger dataset (e.g. simulate more data).