Statistical Inference - Course project - Part 1

Christoph Wagner

03/05/2020

1). Introduction

This document contains my solutions for part 1 of the final course project of Coursera course “Statistical Inference” (Specialization “Data Mining”). The document has been created using R Markdown and the knitr framework. Part 1 of the course project deals with the exponential distribution. It analyzes the distribution and shows the applicability of the Central Limit Theorem (CLT) using simulation / random sampling. All experiments below have been executed in R (IDE RStudio). Graphics have been plotted using the ggplot2 package.

2.) The exponential distribution and it’s averages

The exponential distribution is a common probability distribution which is based on a exponential function. It is mainly used to model the duration of random time intervals, e.g. radioactive decay or lifetime of mechanical applications.

The density of the exponential distribution is defined as

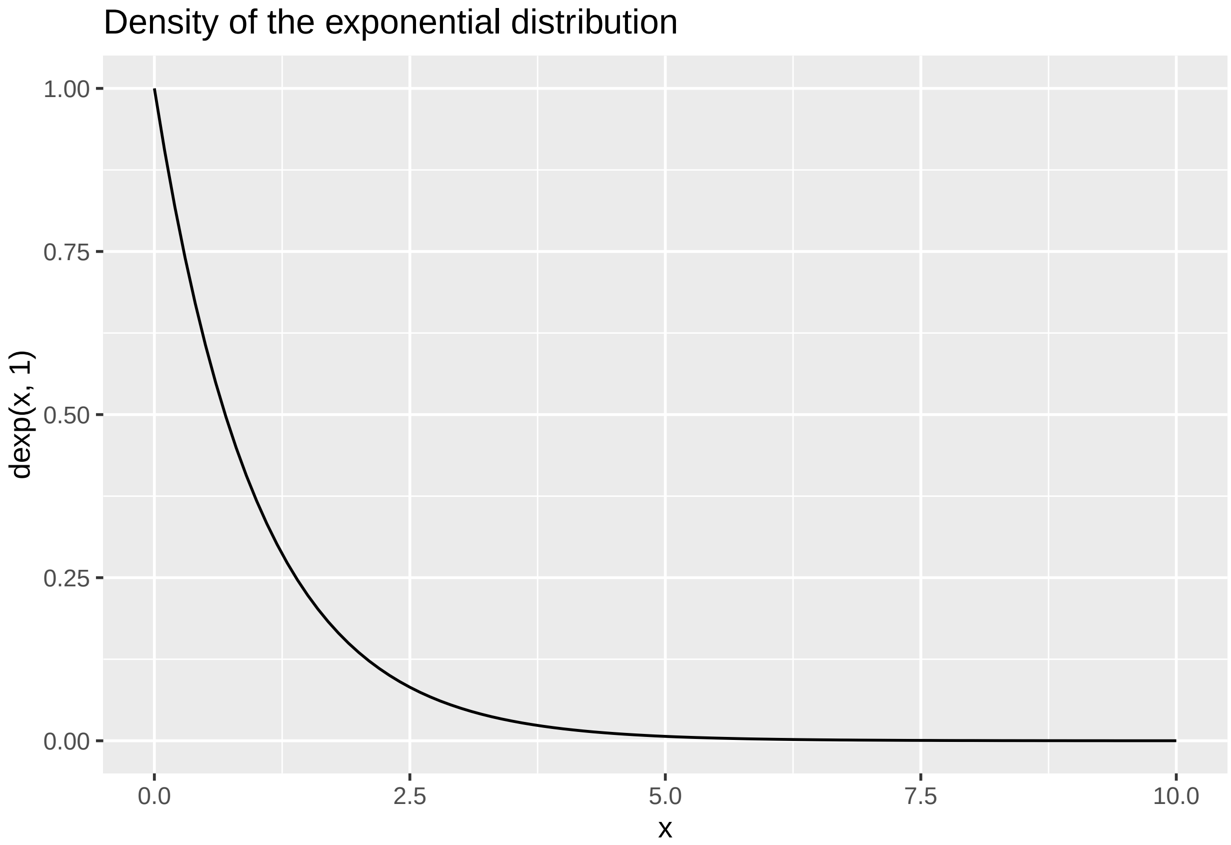
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

In R, the density of the exponential distribution can be calculated using function dexp(). The shape of the distribution can be displayed by applying the function for a value interval, e.g. [0, 10], and drawing the results using the ggplot2 framework.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
x<-seq(0, 10, 0.1)
y<-dexp(x, rate=1)
d<-data.frame(x, y)
ggplot(d, aes(x=x, y=y))+geom_line()+xlab("x")+ylab("dexp(x, 1)")+ggtitle("Density of the exponential di
```



Further relevant R functions related to the exponential distribution:

- pexp: Exponential distribution function
- qexp: Exponential quantile function
- rexp: Function for generating random values of the exponential distribution.

As visible above in the formula, the exponential distribution has one "rate" parameter lambda, which controls important statistics of the distribution:

- The mean of the exponential distribution is equal to $\frac{1}{\lambda}$
- The standard deviation of the exponential is also equal to $\frac{1}{\lambda}$

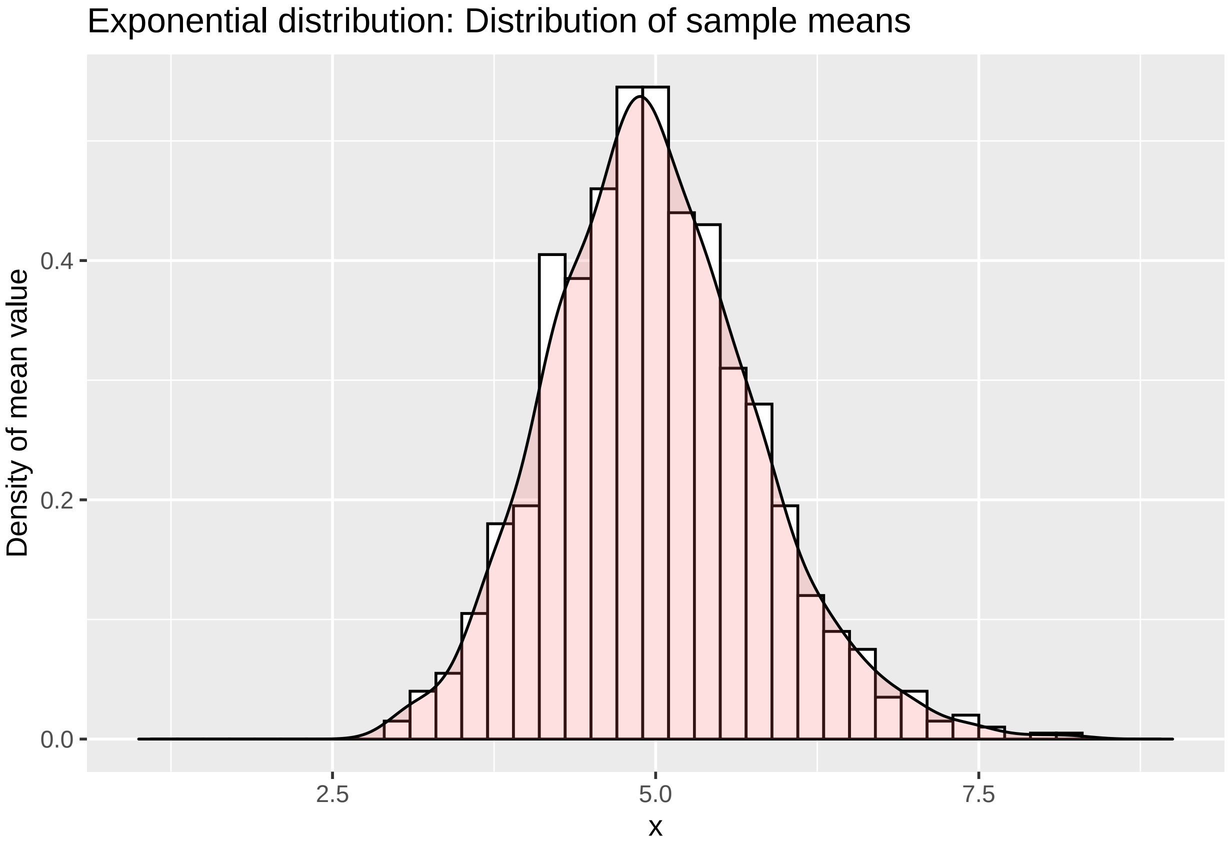
In the following experiments, we will analyze the exponential distribution by simulation. We will calculate the mean of 40 random draws from the exponential distribution (with 1000 repetitions) and verify the CLT for the distribution of the means. For all experiments, the rate value lambda=0.2 is used.

```r
set.seed(42)  #Setting seed for reproducibility
n_sim<-1000   #Number of repetitions of the simulation
n_draws<-40   #Number of draws per repetition
lambda<-0.2   #Used lambda value

#Random draw and storage of the result in dataframe
avgs<-replicate(n_sim, mean(rexp(n_draws, rate=lambda)))
avgs_df<-data.frame(avgs)
```

```
#Display of the random draws in a historgram (interval [1;9]) and a smoothed density graph
ggplot(avgs_df, aes(x=avgs))+
  geom_histogram(aes(y=..density..), binwidth=0.2, colour="black", fill="white")+
  geom_density(alpha=0.2, fill="#FF6666")+xlab("x")+ylab("Density of mean value")+
  xlim(1, 9)+
  ggtitle("Exponential distribution: Distribution of sample means")
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



The resulting distribution of the means doesn't have the shape of the exponential distribution any more but resembles a bit the normal distribution.

3.) The mean of the distribution of averages

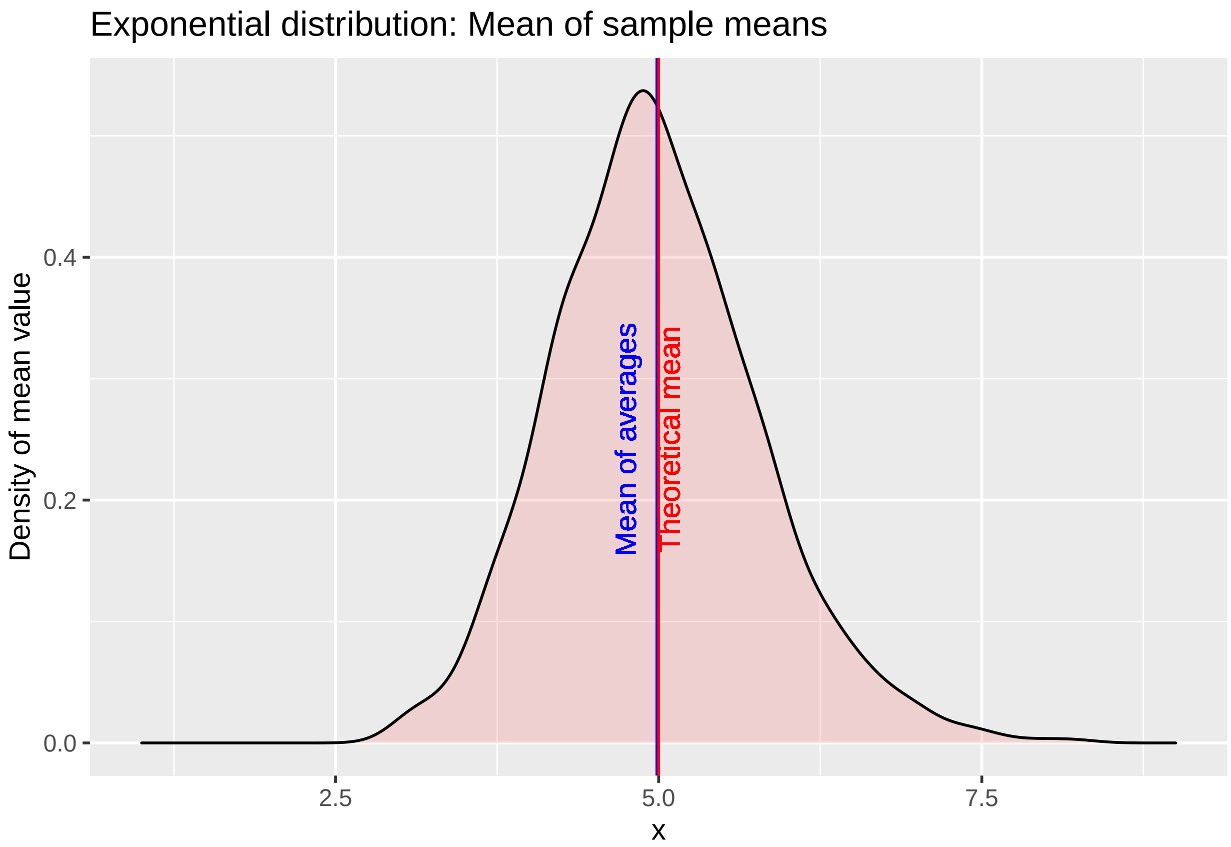
The mean of distribution of the means can be calculated using the R mean function. Theoretically, the sample mean is unbiased and so the mean of it's distribution is equal to what it's estimating, i.e. the mean of the exponential distribution.

The experimental mean of the averages of the simulation above is 4.9865083. The theoretical mean is equal to $\frac{1}{\lambda} = 5$.

The following R code displays the distribution of the mean of 40 random draws from the exponential distribution and marks the theoretical and the experimental mean of the averages:

```r
experimental_mean<-mean(avgs)    #Calculation of the real mean
th_mean<-1/lambda        #Calculation of the sample mean

#Drawing the experimental mean and the theoretical mean
ggplot(avgs_df, aes(x=avgs))+
  geom_density(alpha=0.2, fill="#FF6666")+
  geom_vline(xintercept=experimental_mean, colour="blue")+
  geom_text(aes(x=experimental_mean, y=0.25, label="Mean of averages"), colour="blue",
          angle=90, vjust = -1)+
  geom_vline(xintercept=th_mean, colour="red")+
  geom_text(aes(x=th_mean, y=0.25, label="Theoretical mean"), colour="red", angle=90,
          vjust = +1)+xlab("x")+ylab("Density of mean value")+
  xlim(1, 9)+
  ggtitle("Exponential distribution: Mean of sample means")
```



4.) The variance of the distribution of average

The variance of distribution of the means can be calculated using the R var function. Theoretically, the variance of the sample mean is $Var(X_{average}) = \sigma^2/n$, where σ^2 is the variance of the population where we take our samples from.

The experimental variance of the averages of the simulation above is 0.6344405. The theoretical variance is equal to

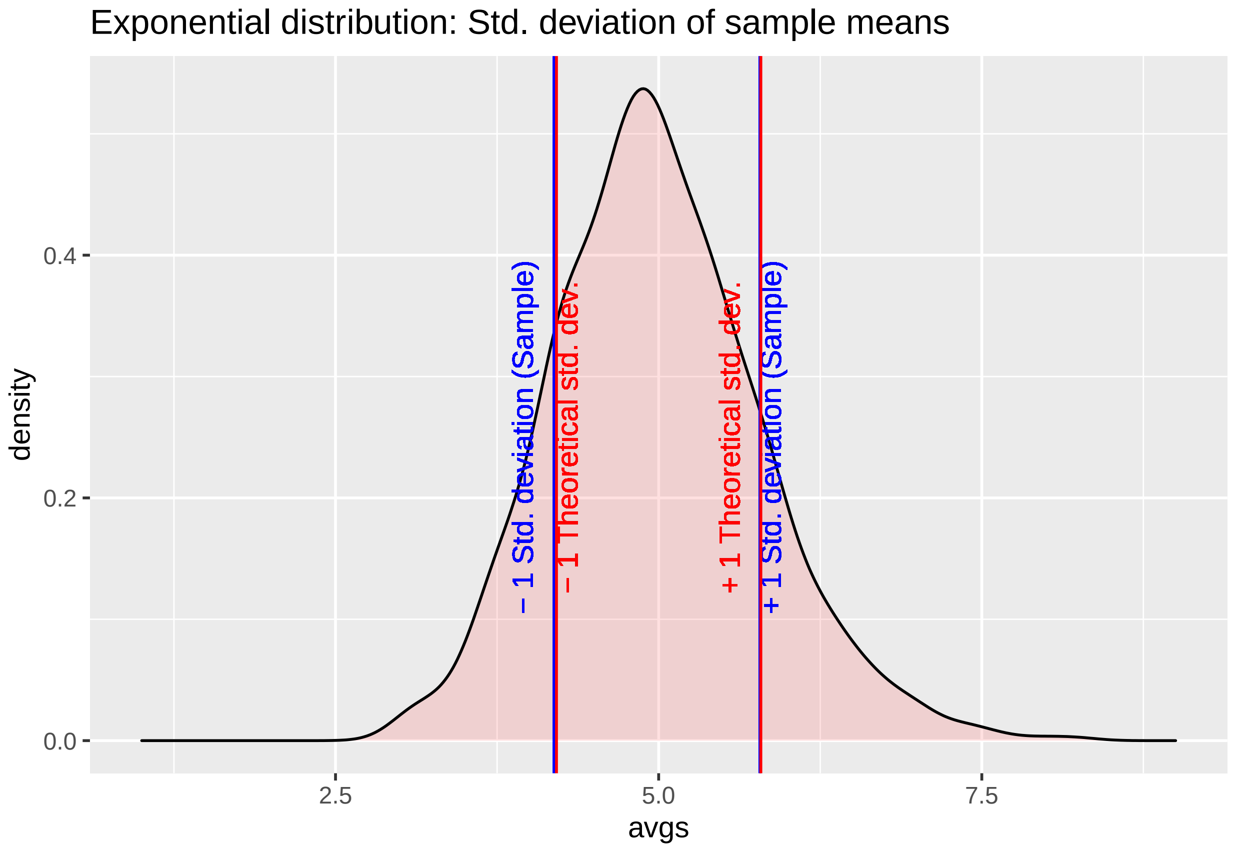
$$\frac{1}{\lambda}^2/n = 0.625$$

.

The following R code displays the distribution of the mean of 40 random draws from the exponential distribution and marks the theoretical and the experimental standard deviations of the averages. I don't display the variances here because they have a different scale than the distribution of the means. Equally as the variance, the standard deviation is a measure for the variation of the values of the distribution.

```
th_std_dv<-(1/lambda)/sqrt(n_draws)

ggplot(avgs_df, aes(x=avgs))+
  geom_density(alpha=0.2, fill="#FF6666")+
  geom_vline(xintercept=mean(avgs)-sd(avgs), colour="blue")+
  geom_text(aes(x=mean(avgs)-sd(avgs), y=0.25, label="- 1 Std. deviation (Sample)"),
            colour="blue", angle=90, vjust = -1)+
  geom_vline(xintercept=mean(avgs)+sd(avgs), colour="blue")+
  geom_text(aes(x=mean(avgs)+sd(avgs), y=0.25, label="+ 1 Std. deviation (Sample)"),
            colour="blue", angle=90, vjust = +1)+
  geom_vline(xintercept=th_mean-th_std_dv, colour="red")+
  geom_text(aes(x=th_mean-th_std_dv, y=0.25, label="- 1 Theoretical std. dev."),
            colour="red", angle=90, vjust = +1)+
  geom_vline(xintercept=th_mean+th_std_dv, colour="red")+
  geom_text(aes(x=th_mean+th_std_dv, y=0.25, label="+ 1 Theoretical std. dev."),
            colour="red", angle=90, vjust = -1)+
  xlim(1, 9)+
  ggtitle("Exponential distribution: Std. deviation of sample means")
```



5.) The distribution of averages and the CLT

The distribution of the mean of random samples of the exponential distribution can be described using the central limit theorem. It states that the distribution of averages of i.i.d. (and normalized) variables becomes that of a standard normal distribution when the sample size increases.

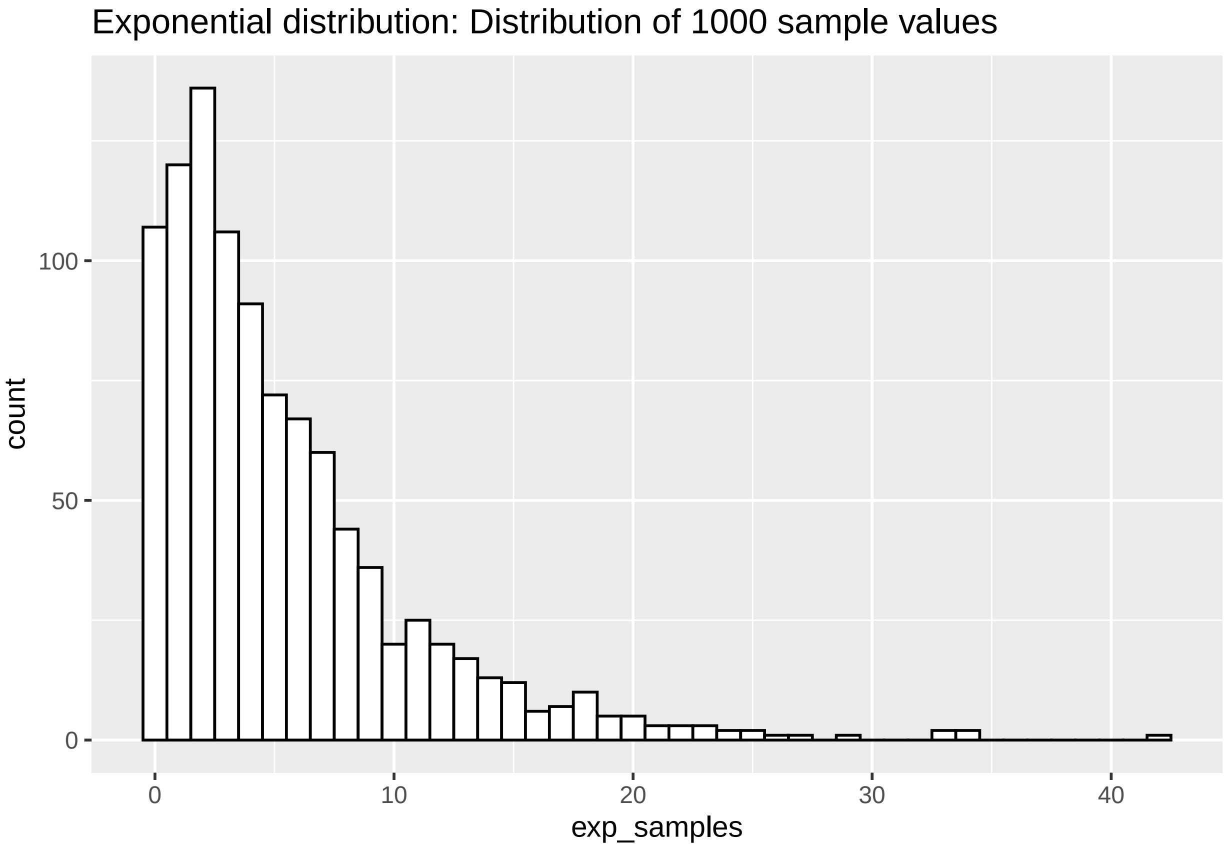
We can experimentally verify this thesis by simulation. First of all, we take 1000 random samples of the exponential distribution and display it, just to show the shape of the exponential distribution for lambda=0.2.

```r
library(ggplot2)

set.seed(42)  #Setting seed for reproducibility
n_sim<-1000   #Number of repetitions of the simulation
lambda<-0.2   #Used lambda value

#Take random sample from the exponential distribution
exp_samples<-rexp(n=n_sim, rate=lambda)
df_values<-data.frame(exp_samples)

#Display random sample in a histogram
ggplot(df_values, aes(x=exp_samples))+
  geom_histogram(binwidth=1, colour="black", fill="white")+
  ggtitle("Exponential distribution: Distribution of 1000 sample values")
```



Then we take 1000 averages of n_draws_1=3 and n_draws_2=40 random samples of the exponential distribution. When the averages are normalized using the formula

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

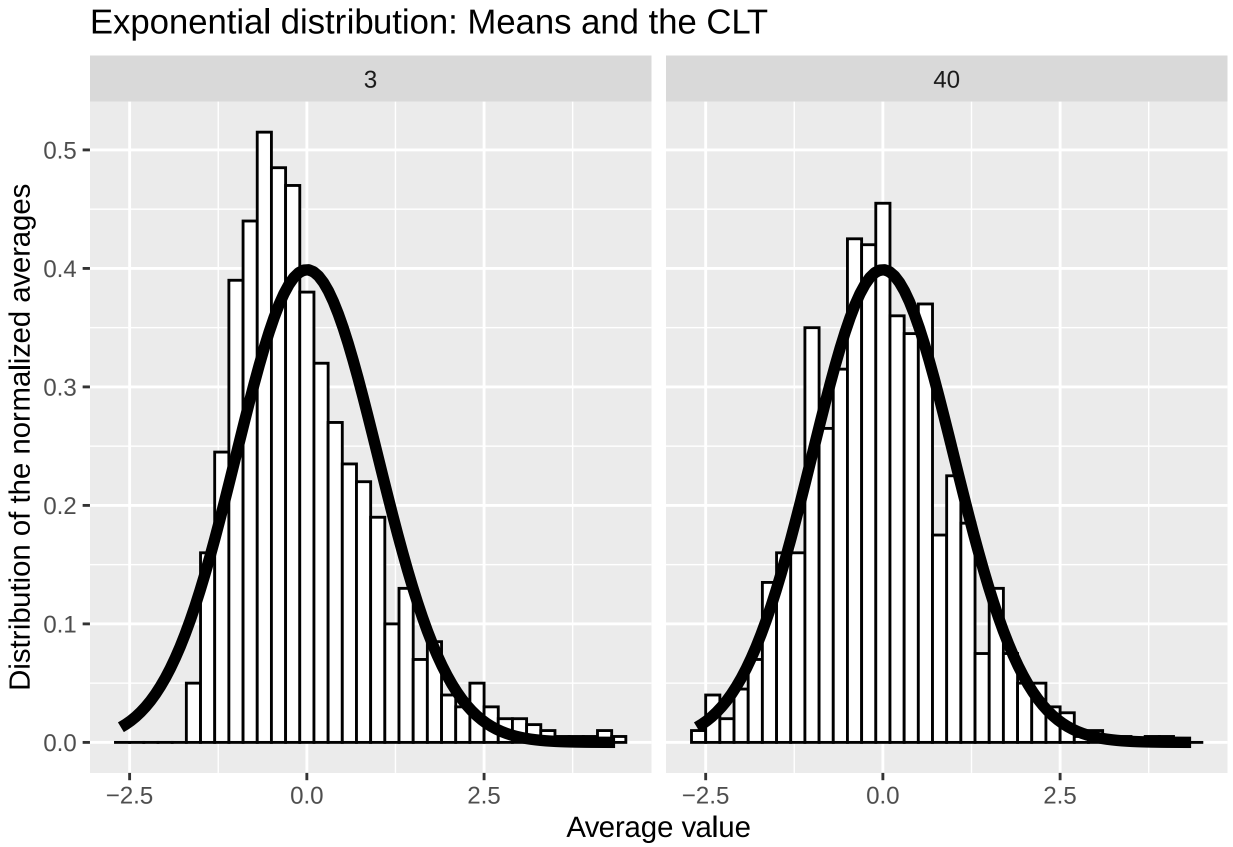
the resulting distibutions resemble a normal distribution.

```
n_draws_1<-3   #Number of draws per repetition
n_draws_2<-40  #Smaller number of draws per repetition (for comparison)

#Random draw and storage of the result in dataframe
avgs_1<-replicate(n_sim, mean(rexp(n_draws_1, rate=lambda)))
avgs_2<-replicate(n_sim, mean(rexp(n_draws_2, rate=lambda)))

#Normalize averages
mean_exp<-1/lambda
sd_exp<-1/lambda
avgs_norm_1<-(avgs_1-mean_exp)/(sd_exp/sqrt(n_draws_1))
avgs_norm_2<-(avgs_2-mean_exp)/(sd_exp/sqrt(n_draws_2))
df_avgs_norm<-data.frame(avgs_norm=c(avgs_norm_1, avgs_norm_2),
                         size=factor(rep(c(n_draws_1, n_draws_2), c(n_sim, n_sim))))

#Display averages and standard normal distribution
ggplot(data=df_avgs_norm, aes(x=avgs_norm))+
  geom_histogram(aes(y=..density..), binwidth=0.2, colour="black", fill="white")+
  stat_function(fun=dnorm, size=2)+
  xlab("Average value")+
  ylab("Distribution of the normalized averages")+
  facet_grid(. ~ size)+
  ggtitle("Exponential distribution: Means and the CLT")
```



Conclusion: The example clearly shows that with increasing n, the shape of the sample distribution becomes more and more similiar to a normal distribution.