Statistical Inference

Peer Assessment Part 1

Sergei Titov

Overview

This part of Statistical Inference project consider with the exponential distribution and the Central Limit Theorem. At this report we will:

- Generate a sample of a thousand simulations of averages of 40 exponentially distributed randoms
- Show the sample mean and compare it to the theoretical mean of the distribution
- Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution
- Show that the distribution is approximately normal

Simulation

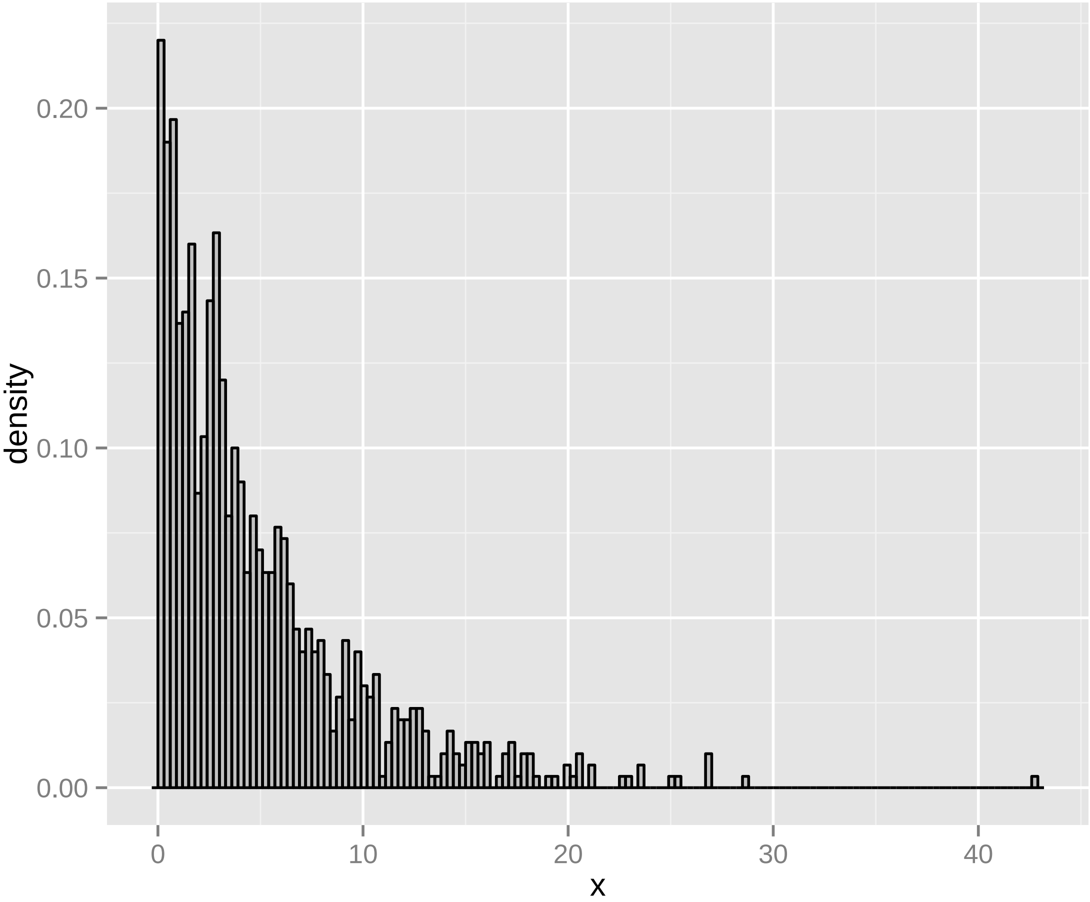
Set the default parameters.

```
lambda <- 0.2
n <- 40
nosim <- 1000
```

Fast look on original exponential distribution.

```
library ("ggplot2")
g <- ggplot (data.frame (x = rexp (nosim, lambda)), aes(x = x))
g <- g + geom_histogram (alpha = .20, binwidth = .3,
                         colour = "black", aes(y = ..density..))
g
```

```
## Warning: position_stack requires constant width: output may be incorrect
```

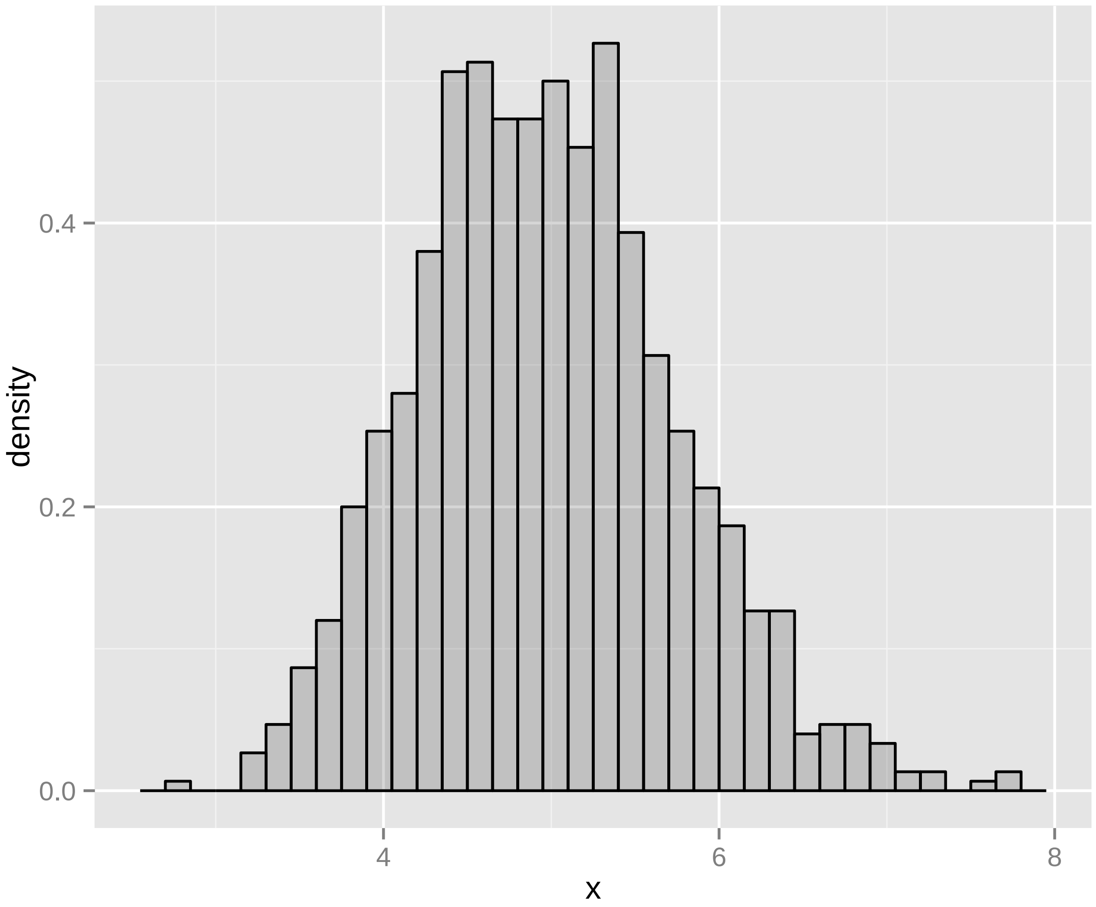


Create a matrix with 1000 rows, 40 observations per row and calculate row means.

```
set.seed (1234)
data <- matrix (rexp (nosim * n, lambda), nrow = nosim)
data <- rowMeans(data)
```

Plotting data.

```
g <- ggplot (data.frame (x = data), aes(x = x))
g <- g + geom_histogram (alpha = .20, binwidth = .15,
                         colour = "black", aes (y = ..density..))
g
```



Sample vs. theoretical parameters

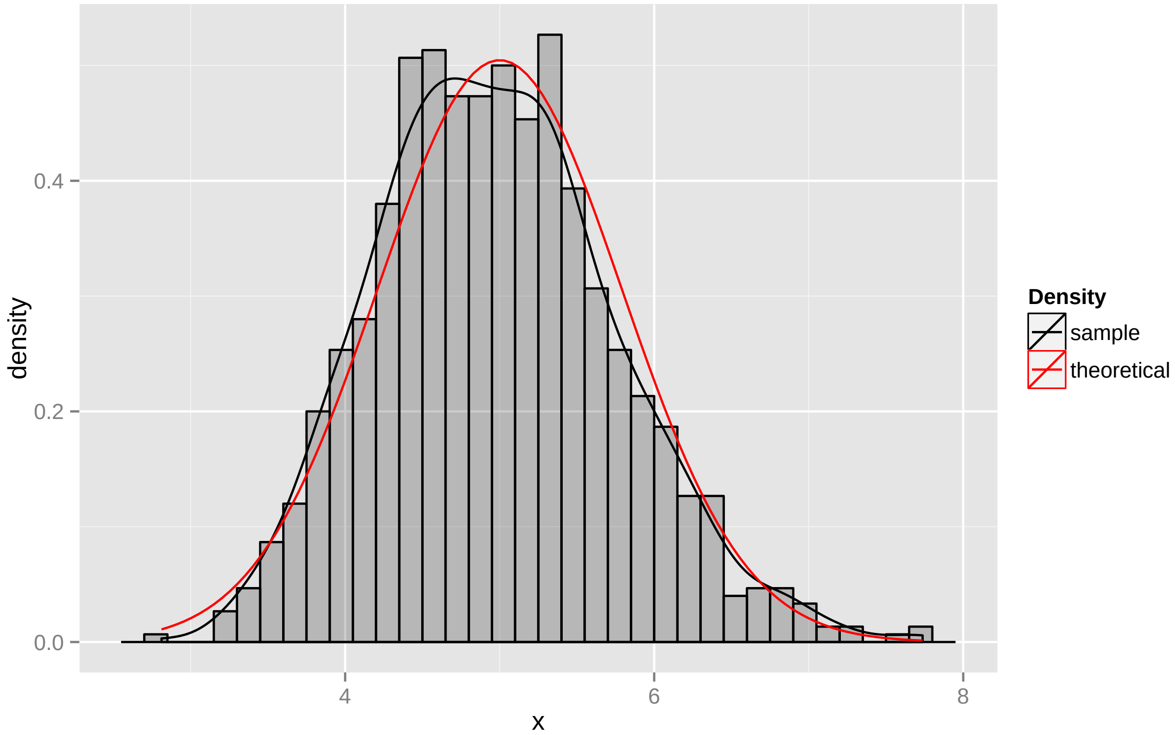
```
sample_mean <- mean (data)
sample_var <- var (data)
sample_sd  <- sd (data)
```

The sample mean equals 4.974 and theoretical mean of the distribution $1/lambda = 1/0.2 = 5$
The variance of sample means is 0.595, where the theoretical variance of the distribution is $(1/lambda)/n = 5/40 = 0.625$
The sample mean equals 0.771 and theoretical standard deviation equals $(1/lambda)/\sqrt{n} = (1/0.2)/\sqrt{40} = 0.791$

Comparison with Normal distribution

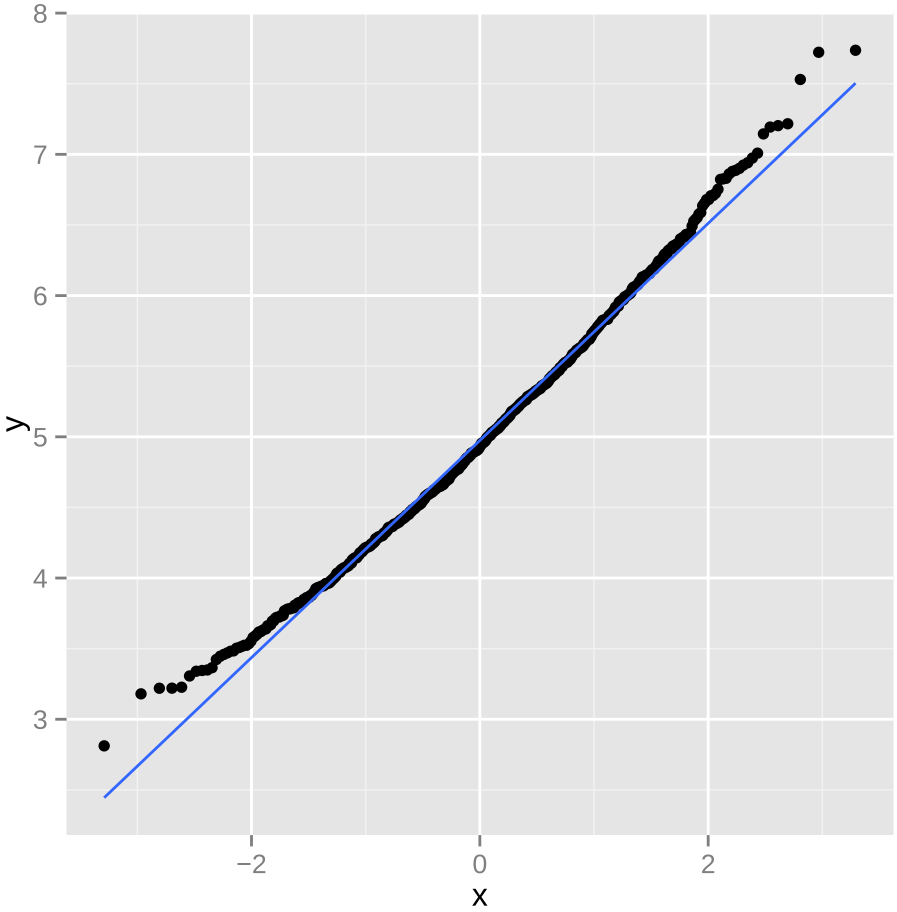
To compare the sample distribution with normal distribution plot density functions.

```
g <- ggplot (data.frame (x = data), aes(x = x)) +
geom_histogram (alpha = .2, fill = "black", binwidth = .15,
                colour = "black", aes (y = ..density..)) +
geom_density (fill = NA, aes (colour = "sample")) +
stat_function (fun = dnorm, args = list (mean = 1/lambda, sd = 5/sqrt (n)),
               aes (colour = "theoretical")) +
scale_colour_manual (name = "Density",
                     values = c ("sample" = "black", "theoretical" = "red"))
g
```



Q-Q Plot.

```
ggplot(data = as.data.frame(qqnorm (data ,plot=F)), mapping = aes(x=x, y=y)) +
    geom_point() + geom_smooth(method="lm", se=FALSE)
```



Due to the central limit theorem (CLT), the distribution of averages of 40 exponentials is very close to a normal distribution.