Statistical Inference Course Project Part 1

Katya Demidova

24 December 2015

Overview

The main goal of this document is to investigate the exponential distribution in R and compare it with the Central Limit Theorem. We will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials and show that it behaves as predicted by the CLT. The code used in this work can be seen in Appendix.

Simulation

The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter.

The requirements are:

- Set `lambda = 0.2` for all of the simulations.
- Investigate the distribution of averages of 40 exponentials.
- Do a thousand simulations.

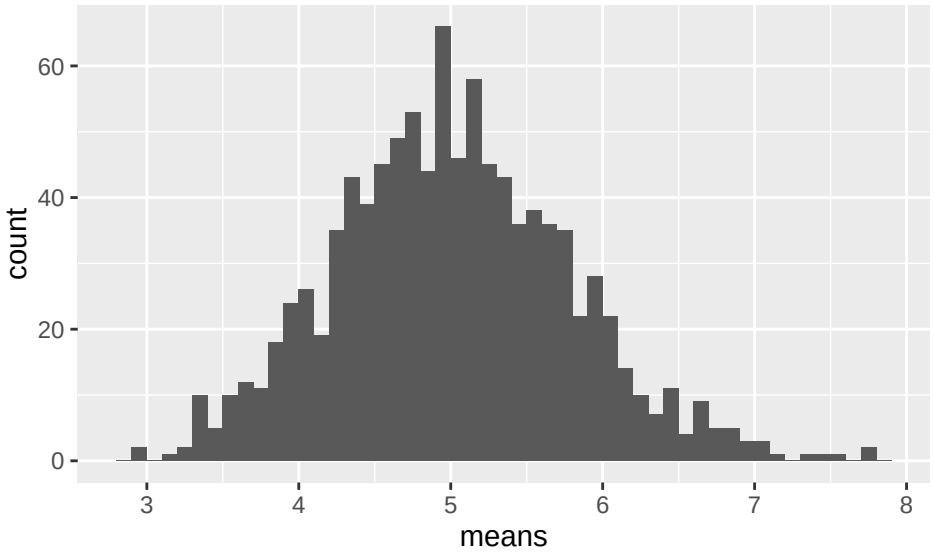

Figure: Distribution of simulated averages of 40 exponentials

Sample Mean versus Theoretical Mean

The mean of exponential distribution is `1 / lambda`. Given the Central Limit Theorem, our expected mean would be:

```
(theor.mean <- 1 / lambda)
```

```
## [1] 5
```

Next, we will evaluate the sample mean:

```
(sample.mean <- mean(means$means))
```

```
## [1] 5.011911
```

Let's add the sample mean and the theoretical mean to the plot we've constructed before:

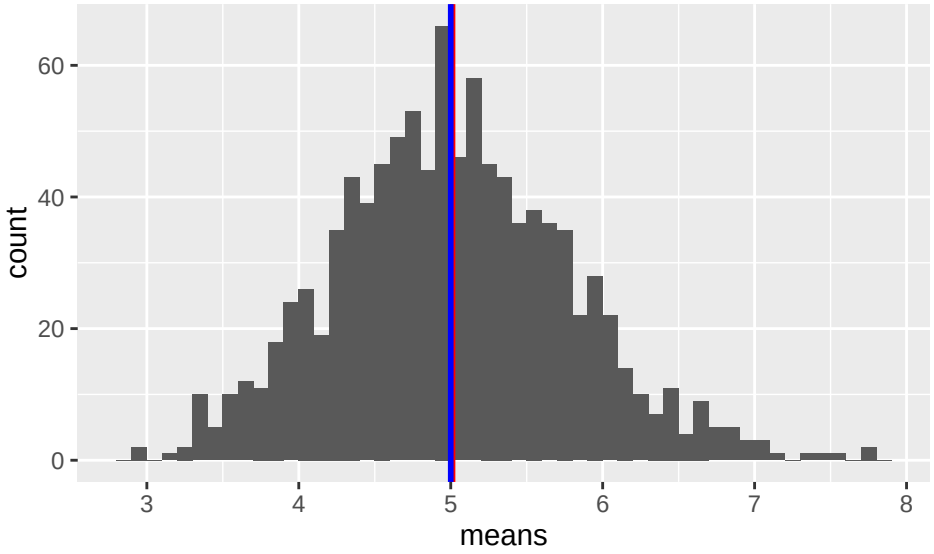

Figure: Sample mean (red line) vs. Theoretical mean (blue)

As predicted by CLT, the center of mean distributions (5.0119113) is very close to the theoretical mean (5).

Sample Variance versus Theoretical Variance

The standard deviation of exponential distribution is `1 / lambda`. Given the Central Limit Theorem, our expected variance equals to `sd ^ 2 / n`, where `n` is the sample size (which is 40).

Theoretical variance and standard deviation of the sample means:

```
(theor.variance <- ((1 / lambda) / sqrt(n)) ^ 2 )
(theor.sd <- sqrt(theor.variance))
```

```
## [1] 0.625
## [1] 0.7905694
```

Sample variance and standard deviation of averages of simulations:

```
(sample.variance <- var(means$means))
(sample.sd <- sd(means$means))
```

```
## [1] 0.6004928
## [1] 0.7749147
```

Sample standard deviation (0.7749147) is close to the theoretical standard deviation (0.7905694). Sample and expected variances (0.6004928 and 0.625) also look similar (since variances are measured in square units, standard deviations are preferred).

Distribution

Given the Central Limit Theorem, the distribution of the means should be approximately normal. Let's look at this figure:

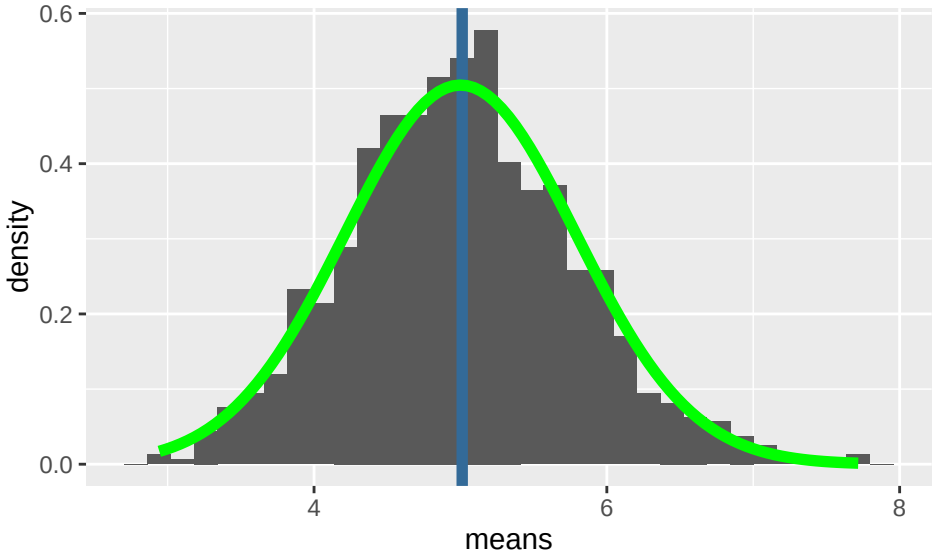

Figure: distribution of sample means. Green line: sample mean. Yellow curve: normal distribution

Indeed, density of calculated means is somehow similar to a normal (bell-shaped) curve (its mean and sd were calculated earlier). We can also build a QQ plot:

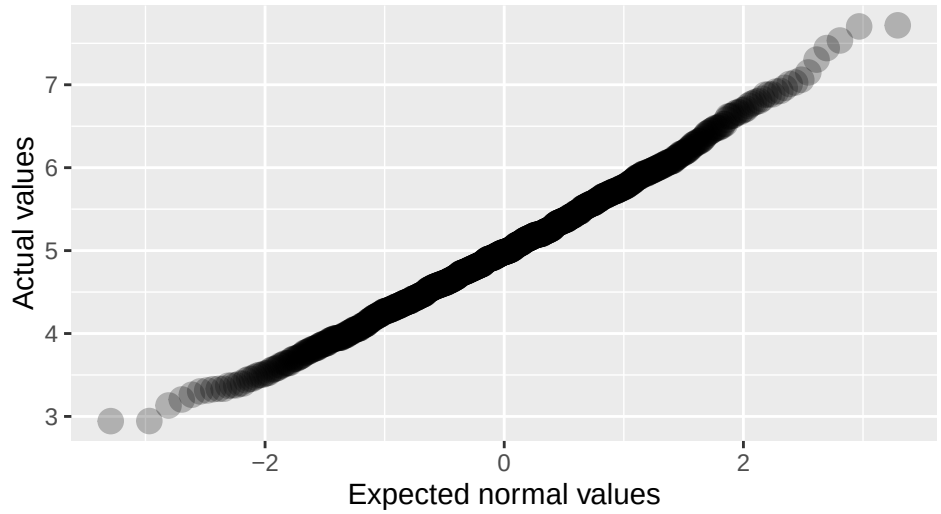

Figure: QQ plot for sample distribution of the means against theoretical distribution

From the look of this QQ plot, we can make the conclusion that the distribution of actual data (sample means) is approximately normal.

We have investigated the exponential distribution in R and compared it with the Central Limit Theorem. We have compared the sample mean to the theoretical mean, the variability of the mean of 40 exponentials to the theoretical variance, and from that we concluded that the distribution of sample means behaved as predicted by the Central Limit Theorem.

Appendix

```
library(ggplot2)

set.seed(123)

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

sims <- data.frame(replicate(nosim, rexp(n, lambda)))

means <- sapply(sims, mean)
means <- data.frame(means)

dist <- ggplot(data = means, aes(x = means)) + theme(legend.position = "none") +
  geom_histogram(binwidth=0.1)

dist
```

```
dist +
  geom_vline(aes(xintercept = sample.mean), colour="red", size = 1) +
  geom_vline(aes(xintercept = theor.mean), colour="blue", size = 1)
```

```
ggplot(data = means, aes(x = means)) +
  geom_histogram(aes(y=..density..)) +
  geom_vline(aes(xintercept = sample.mean, color = sample.mean), size = 2) +
  stat_function(fun = dnorm, args = list(mean = theor.mean, sd = theor.sd), size=2, color = "green") + t
```

```
ggplot(data = means, aes(sample = means)) +
  stat_qq(size=4, alpha=0.25) +
  labs(title = "",
       x = "Expected normal values",
       y = "Actual values")
```