Statistical Inference Course Project Part 1

Naveen Venkat Raj Jaladi

Overview

As the part of coursera statistical inference course project this is the first part of the project which make calculation and plots and compare confidence intervals, and eventually proves that the distribution is approximately normal.

Tasks

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should - 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.

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.

Simulation

We simulate the exponential distribution using R `rexp()` function and use `ggplot2` package to compare its properties. At first, we should to load required packages, set up seed and other parameters values.

```
set.seed(12031987)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

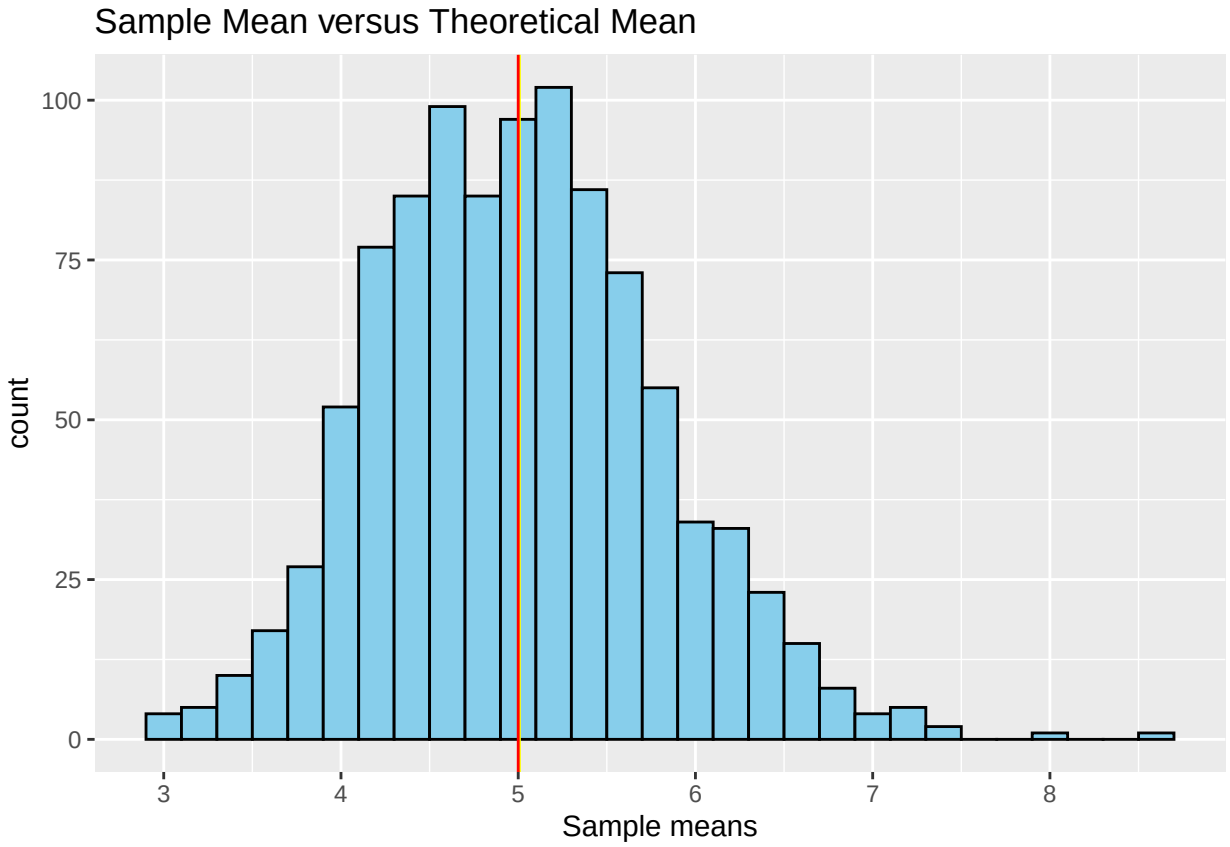
```
lambda <- 0.2
sims.n <- 1000
n <- 40
```

Then we do **1000** simulations of **40** exponentials and calculate their means.

```
sims.exp <- replicate(sims.n, rexp(n, lambda))
sims.exp.means <- colMeans(sims.exp)
```

Q1 Sample Mean Vs Theoretical Mean

The difference between the sample mean **(5.0069177)** and theoretical mean **(5)** is **0.0069177**. It is a very small value cause their values are very close and it was be expected. Lets visualize the difference by two vertical lines per each mean on the histogram.

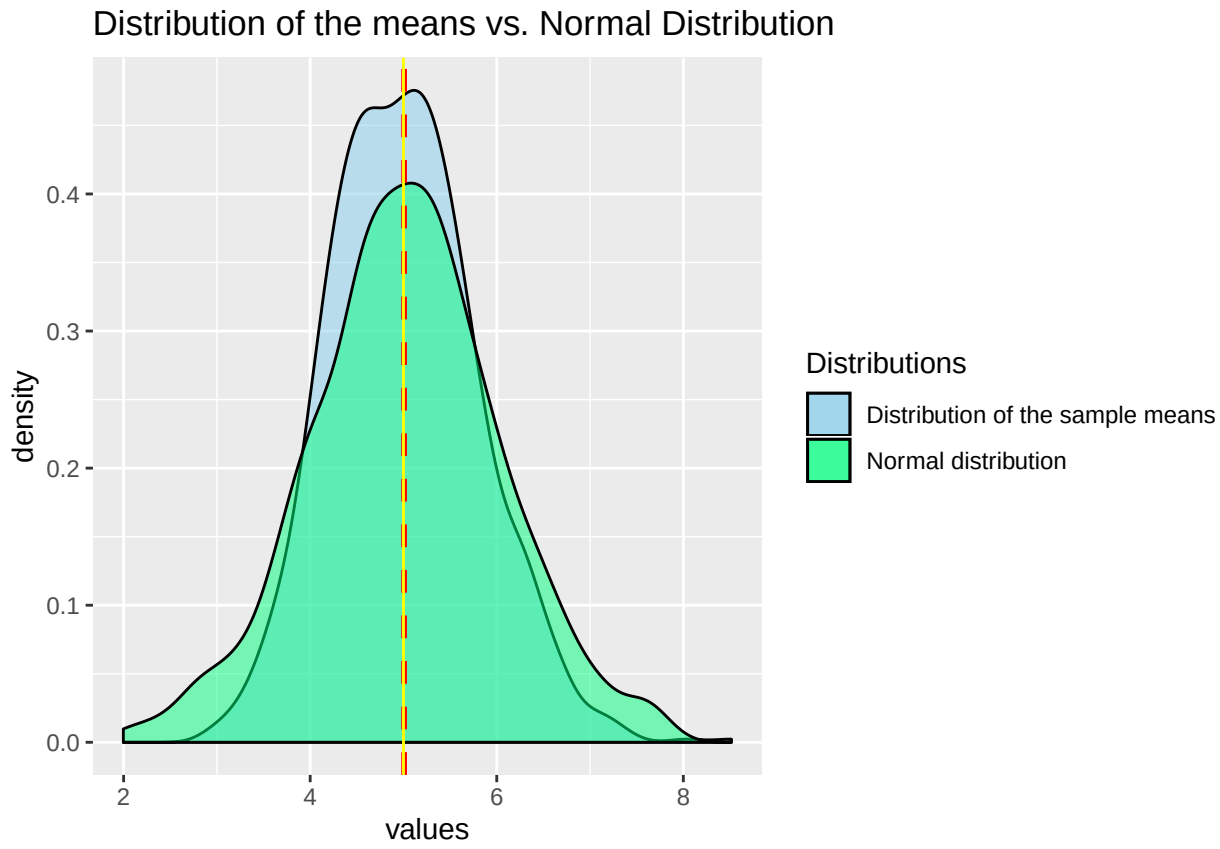


Q2 Sample variance Vs Theoretical Variance

The difference between the sample variance **(0.619896)** and theoretical variance **(0.625)** is **-0.005104**. In this case, the sample variance is **less** than theoretical variance, but the difference between their values are also very small. And according **the Central Limit Theorem** sample variance will be closer to theoretical variance as we increase our sample size.

Q3 Distribution

As we already know from the Central Limit Theorem distribution of the means should looks like more similar to normal distributions. And this process depends on numbers of iteration. And it is clearly seen in the visualization below:



Appendices

Q1 code listing:

```
sample.mean <- mean(sims.exp.means)
theory.mean <- 1/lambda
```

```
ggplot(as.data.frame(sims.exp.means), aes(x=sims.exp.means)) +
     geom_histogram(binwidth = .2, color = "black", fill = "skyblue") +
     geom_vline(xintercept = sample.mean, color = "yellow") +
     geom_vline(xintercept = theory.mean, color = "red") +
     labs(title = "Sample Mean versus Theoretical Mean", x = "Sample means")
```

Q2 code listing:

```
sample.var <- var(sims.exp.means)
theory.var <- ((1/lambda)^2)/n
```

Q3 code listing:

```
sims.norm <- rnorm(sims.n, mean=1/lambda)
```

```r
ggplot(as.data.frame(sims.exp.means), aes(x=sims.exp.means, fill = "skyblue")) +
     geom_density(alpha= .5) +
     geom_density(data = as.data.frame(sims.norm), aes(sims.norm, fill = "springgreen"), alpha= .5) +
     geom_vline(xintercept = sample.mean, linetype = "dashed", color = "red", size = 1) +
     geom_vline(xintercept = 1/lambda, color = "yellow", size = .5) +
     labs(title = "Distribution of the means vs. Normal Distribution", x = "values") +
     scale_fill_identity(name = "Distributions",
                         guide = "legend",
                         labels = c("Distribution of the sample means", "Normal distribution"))
```