Statistical Inference Course Project

Andrea Eoli

2022-11-07

Instructions

The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

You will create a PDF report to answer the questions. Each PDF report should be no more than 3 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).

Part 1: Simulation Exercise Instructions

Overview

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

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. 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.

Answer

Generate simulated data

```r
set.seed(1) # Set seed for reproducibility
lambda <- 0.2
n <- 40

# Simulate data
simul <- replicate(n = 1000, expr = rexp(n = n, rate = lambda))

# Calculate mean
mean_sim <- apply(simul, 2, mean)
```

Q1: Sample mean vs Theoretical mean

```r
# Sample mean
sample_m <- mean(mean_sim)
sample_m
```

```
## [1] 4.990025
```

```r
# Theoretical mean
theo_m <- 1/lambda
theo_m
```

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

Both Sample and Theoretical means are very similar, they both approximate to 5.

Q2: Sample variance vs Theoretical variance

```r
# Sample var
sample_v <- sd(mean_sim)^2
sample_v
```

```
## [1] 0.6111165
```

```r
# Theoretical var
theo_v <- (1/lambda/sqrt(n))^2
theo_v
```

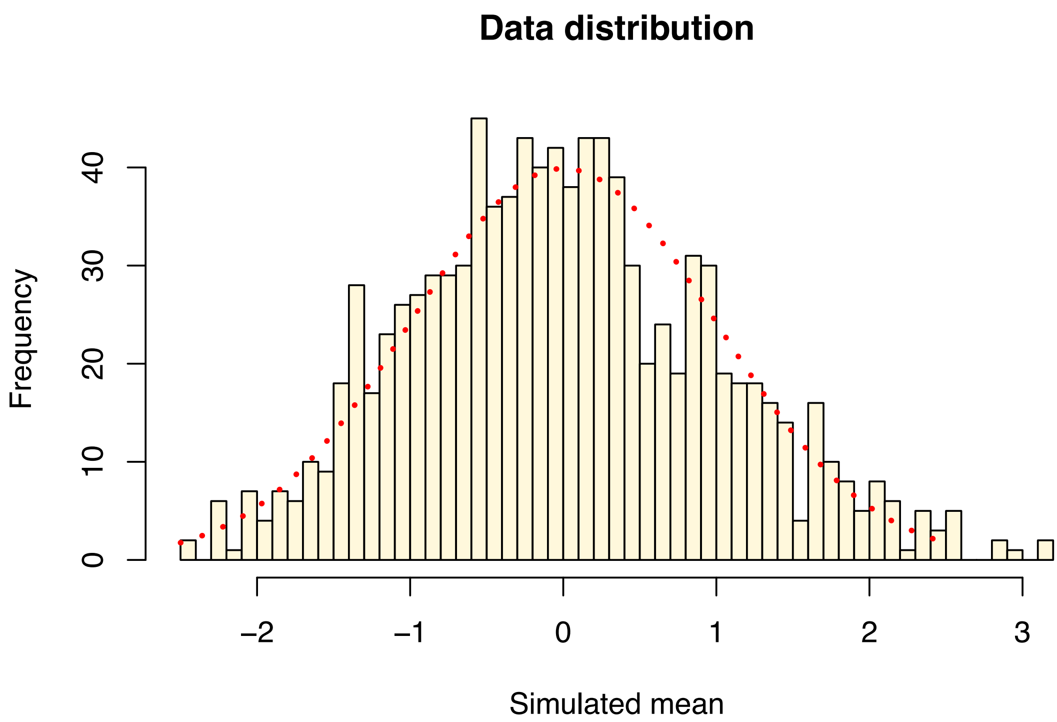
```
## [1] 0.625
```

Both Sample and Theoretical variants are very similar, being the first 0.61 and the second 0.63.

Q3: Check if the distribution is normal

```r
# Histogram
scaled_distr <- scale(mean_sim) # z-transform to make it easier
hist(scaled_distr, main = "Data distribution", breaks = 60, col = "cornsilk", xlab = "Simulated mean")

xfit <- seq(-2.5, 2.5, length = 100)
yfit <- dnorm(xfit, mean = 0, sd = 1)
lines(xfit, yfit*100, lty = 3, lwd = 3, col = "red") # plot reference line
```



As we can see, because of the Central Limit Theorem, the distribution of the simulated data (here standardized to m = 0 and sd = 1) is approximately normal (red line).