

Statistical Inference Project - Part 1

bzg

11-1-2022

Synopsis

This is a project for the Coursera Statistical Inference Class. The project consists of two parts: - Simulation Exercise to explore inference - Basic inferential analysis using the ToothGrowth data in the R datasets package

Part 1 - Simulation Exercise

Overview

In the first part of the project, we will investigate the behaviour of means for exponential distribution in R and compare it with the expected distribution obtained by Central Limit Theorem. In this part, we will set the rate λ to 0.2. First we draw the histogram of 1000 random exponential values. For the mean distribution, we will consider a mean of $n=40$ exponential random values.

Instructions

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. - Show the sample mean and compare it to the theoretical mean of the distribution. - Show how variable the sample is and compare it to the theoretical variance of the distribution. - Show that the distribution is approximately normal.

Preparing the environment

Load all necessary libraries.

```
# Load packages
library(dplyr, warn.conflicts = F)
library(ggplot2)
library(knitr)
library(data.table)

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

Set variables as defined in the problem.

```
# Exponential function parameters
lambda <- 0.2 # lambda for rexp (limiting factor) (rate)
n <- 40 # number of exponentials (sample size)
num_Sim <- 1000 # number of simulations
quantile <- 1.96 # 95th % quantile to be used in Confidence Interval

# Set the seed
set.seed(1)
```

Create a matrix with 1000 simulations each with 40 samples drawn from the exponential distribution.

```
# Use rexp() and matrix() to generate 40 samples creating a matrix with 1000 rows and 40 columns.
simData <- matrix(rexp(n * num_Sim, rate = lambda), num_Sim)
```

Calculate the averages across the 40 values for each of the 1000 simulations.

```
# Calculate the averages across the 40 values for each of the 1000 simulations.
simMeans <- rowMeans(simData) # Matrix Mean
```

Mean Comparison

Show the sample mean and compare it to the theoretical mean of the distribution.

Mean of Sample Means

Calculate the actual mean of sample data; the average sample mean of 1000 simulations of 40 randomly sampled exponential distributions.

```
sampleMean <- mean(simMeans) # Mean of sample means
sampleMean

## [1] 4.990025
```

Theoretical Mean

Calculate the theoretical mean; the expected mean of the exponential distribution of rate = $1/\lambda$.

```
# Calculate the theoretical mean; the expected mean of the exponential distribution of rate = 1/lambda.
theoMean <- 1 / lambda # Theoretical Mean
theoMean

## [1] 5
```

The distribution of the mean of the sample means is centered at **4.990025** and the theoretical mean is centered at **5**. The mean of the sample means and the theoretical mean (expected mean) are very close.

Variance Comparison

Show how variable the sample is and compare it to the theoretical variance of the distribution.

Sample Variance

Calculate the Actual Variance.

```
# Calculate the Actual Variance.
sampleVar <- var(simMeans)
sampleVar

## [1] 0.6177072
```

Theoretical Variance

Calculate the theoretical variance (expected variance).

```
# Calculate the theoretical variance (expected variance).
theoVar <- (1 / lambda)^2 / (n)
theoVar

## [1] 0.625
```

The variance of the sample means is **0.6177072** and the theoretical variance of the distribution is **0.625**. Both variance values are very close to each other.

Sample Standard of Deviation

Calculate the sample means standard of deviation.

```
# Calculate the sample means standard of deviation.
sampleSD <- sd(simMeans)
sampleSD

## [1] 0.7859435
```

Theoretical Standard of Deviation

Calculate the theoretical standard of deviation.

```
# Calculate the theoretical standard of deviation.
theoSD <- 1/(lambda * sqrt(n))
theoSD

## [1] 0.7905694
```

The sample means standard of deviation is **0.7859435** and the theoretical means of standard deviation is **0.7905694**. Again, the values are close.

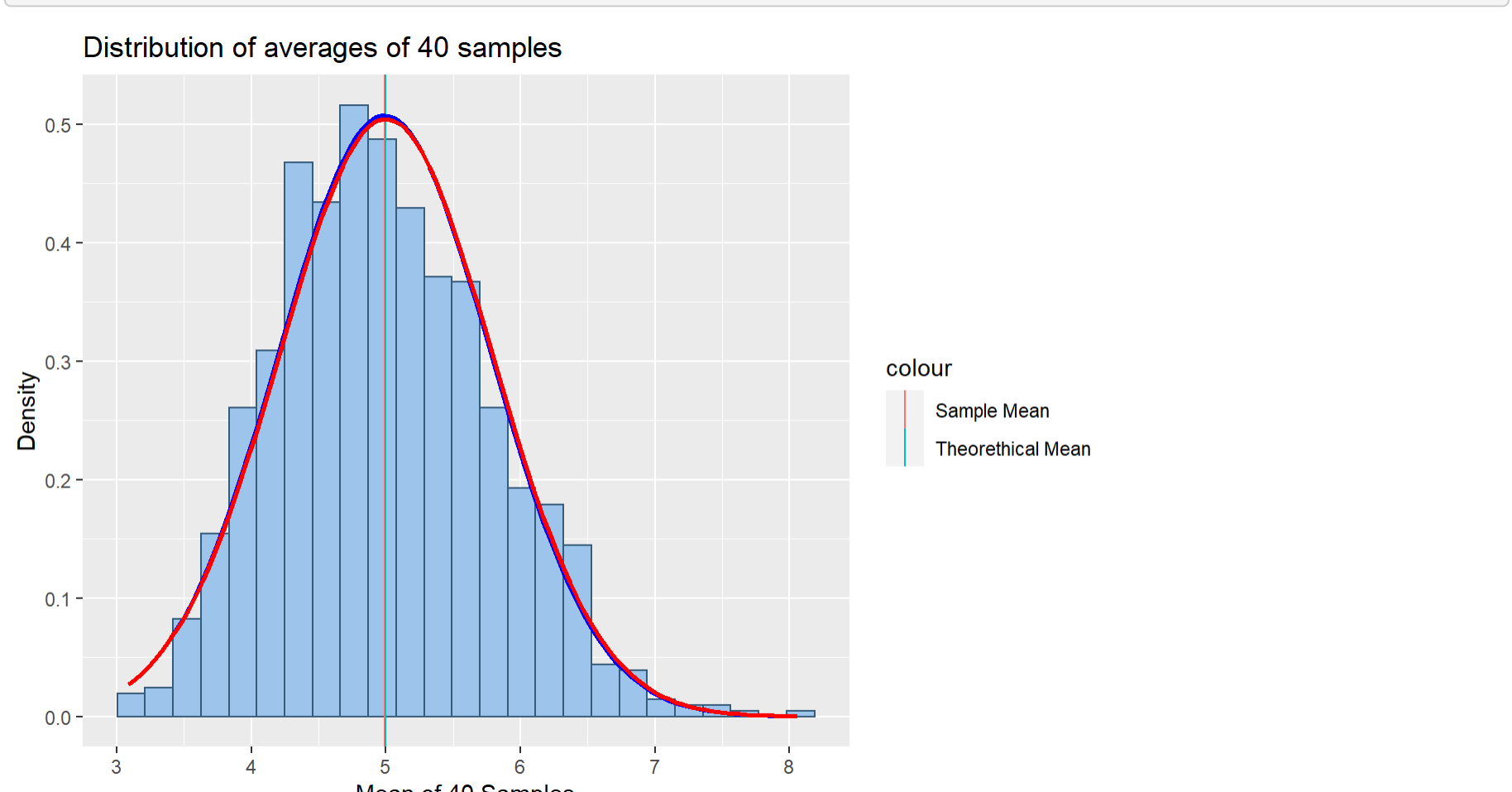
Results

Show that the distribution is approximately normal.

Display the results to visually compare the actual (sample) values versus the theoretical values.

```
# Show that the distribution is approximately normal.
plotdata <- data.frame(simMeans)

ggplot(plotdata, aes(x =simMeans)) +
  geom_histogram(aes(y=..density..), colour="#375B7A",
    fill = "#9DC5EB", bins = 25) +
  labs(title = "Distribution of averages of 40 samples",
    x = "Mean of 40 Samples", y = "Density") +
  geom_vline(aes(xintercept = sampleMean, colour = "Sample Mean")) +
  geom_vline(aes(xintercept = theoMean, colour = "Theoretical Mean")) +
  stat_function(fun = dnorm, args = list(mean = sampleMean,
    sd = sampleSD),
    color = "blue", size = 1.0) +
  stat_function(fun = dnorm, args = list(mean = theoMean,
    sd = theoSD),
    colour = "red", size = 1.0)
```



The density of the actual data is shown by the light blue bars. The theoretical mean and the sample mean are so close to each other that they **overlap**. The "red" line shows the normal curve formed by the theoretical mean and standard deviation. The "royal blue" line shows the curve formed by the sample mean and standard deviation.

As you can see from the graph, the distribution of averages of 40 exponential distributions is close to the normal distribution with the expected theoretical values based on the given lambda.

Confidence Interval Comparison

Check the confidence interval levels to see how they compare.

Sample Confidence Interval

Calculate the sample confidence interval; sampleCI = mean of x plus or minus the .975th normal quantile times the standard error of the mean standard deviation of x divided by the square root of n (the length of the vector x).

```
# Check the confidence interval levels to see how they compare.
sampleConfInterval <- round (mean(simMeans) + c(-1,1)*1.96*sd(simMeans)/sqrt(n),3)
sampleConfInterval

## [1] 4.746 5.234
```

Theoretical Confidence Interval

Calculate the theoretical confidence interval; theoCI = theoMean of x plus or minus the .975th normal quantile times the standard error of the mean standard deviation of x divided by the square root of n (the length of the vector x).

```
# Calculate the theoretical confidence interval
theoConfInterval <- theoMean + c(-1,1) * 1.96 * sqrt(theoVar)/sqrt(n)
theoConfInterval

## [1] 4.755 5.245
```

Conclusion

It is determined that the distribution does indeed demonstrate the Central Limit Theorem; a bell curve. After graphing all the values above and comparing the confidence intervals, the distribution is approximately normal.