Statistical_Inference_Project

Venkat

8/19/2019

Introduction

The project consists of two parts:

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

You will create a report to answer the questions. Given the nature of the series, ideally you'll use knitr to create the reports and convert to a pdf.

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.

4. 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.

Variables and Data Generation

Based on above requirements, we are defining the variables and also using rexp() generate sample data and using sapply simulate it 1000 times.

```
#set seed for reproducibility
set.seed(99999989)
n <- 40   #tcount of numbers in each simulation
lambda <- 0.2       #set lambda for each simulation
simulations <- 1:1000   # number of simulations
#Generating Data for Exponential Distribution using rexp() of rate 0.2.
avg <- sapply(simulations, function(x) {mean(rexp(n, lambda))})

# now we have averaged out values of the 40 numbers that are simulated 1000 times.
length(avg)
```

```
## [1] 1000
```

1. Compare the distribution with theoretical distribution

- For the **simulated data**, Mean of distribution : `4.988287`
- While for the **theoretical data**, mean of distribution is `5`

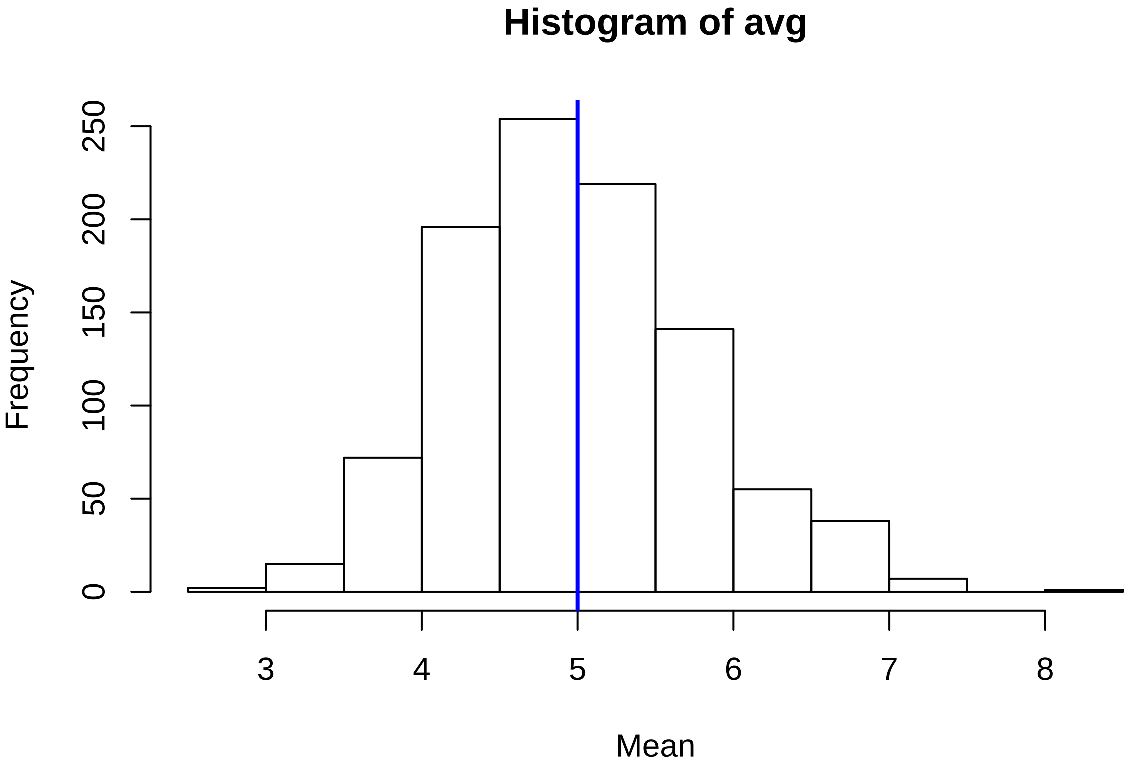
Difference between the 2 distributions: `0.011713`

2. How variable it is and compare it to the theoretical variable of the distribution?

- For the **generated data**, the standard deviation: `0.7886625` & the **variance is :** `0.6219886`
- For **theoretical data**, The standard deviation: `0.7905694` & the **variance is:** `0.625`

3. Perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data?

```r
## Now plot the data as histogram
hist(avg, xlab="Mean", ylab="Frequency")
abline(v=1/lambda , col="blue", lwd=2)
```



Based on output of the histogram, it follows the bell curve, indicating the data is **distributed normally** and most of it is concentrated around the mean.

4. Confidence Interval

Based on the samples, the best estimate would be the sample mean, i.e.here it is 5. But if we have to tell how uncertain we are of this estimate, we need capture the confidence interval.

We can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to the point estimate.

```r
se = sd(avg)/sqrt(n)
low = mean(avg) - 1.96*se
```

```r
high= mean(avg) + 1.96*se
print(c(low, high))
```

```
## [1] 4.743878 5.232696
```

Based on the output we can say **with 95% confidence that the true mean lies between `4.743878, 5.2326961`**