

Statistical Inference

Reza Mofidi

15/06/2020

Peer Graded Assignment: Statistical Inference Course (Part-1)

Simulation Exercise

Introduction

This assignment is divided into two parts. Part one is a simulation exercise. Part 2 involves basic inferential data analysis. I will address each of these projects separately.

In this project I investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate Advanced Math and Statistics.

Exponential Distribution

The exponential distribution is a continuous distribution that is commonly used to measure the expected time for an event to occur (1). For example, in physics it is often used to measure radioactive decay, in engineering it can be used to measure the time associated with receiving a defective part on an assembly line, and in finance it is used to measure the likelihood of the next default for a portfolio of financial products (1). The exponential distribution can be simulated in R with `rexp(n, lambda)` where λ 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. I aim to investigate the distribution of averages of 40 exponentials. Note that As per instructions I have performed a thousand simulations.

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.

```
library(dbplyr)
library(ggplot2)
if(!file.exists("~/data")){dir.create("~/data")}
```

The first part of this project is completed in the following steps:

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

```
n<- 40
simulations<- 1000
lambda<- 0.2
sample_mean<- as.numeric()
for(i in 1:simulations){
  sample_mean <- c(sample_mean, mean(rexp(n, lambda)))
}
mean(sample_mean)
## [1] 4.963341
```

As you can see the sample mean is close to the value 5 it varies slightly every time the simulation is run.

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

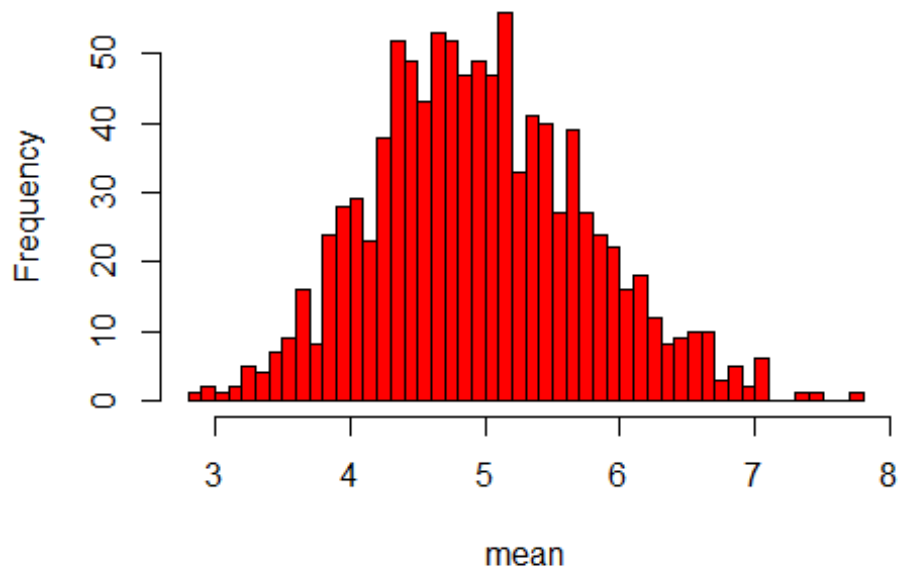
```
variance<- var(sample_mean)
variance
## [1] 0.6256575
```

The variance value is close to 0.6

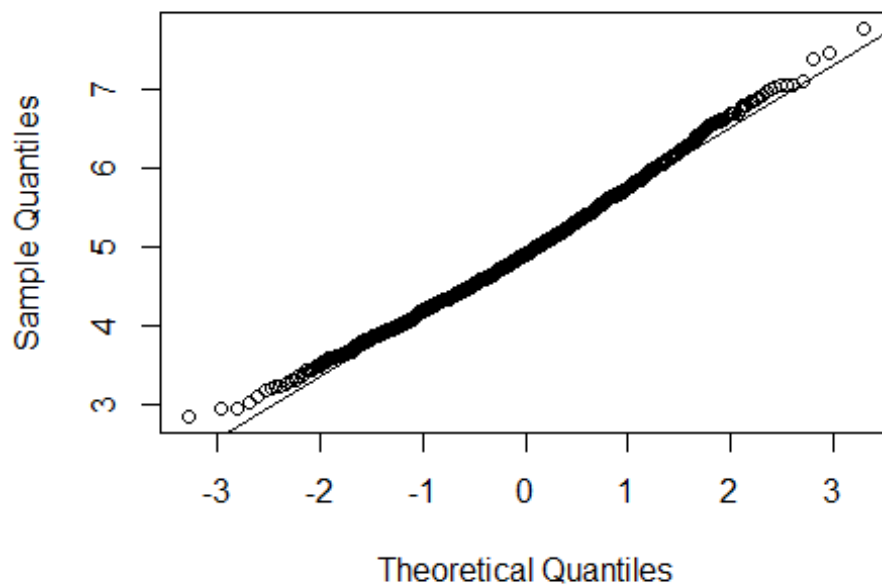
3- Show that the distribution is approximately normal.

A picture (or 2) is worth a thousand words therefore I plan to illustrate this in graphical form. The following plots illustrate how closely the sample means are similar to a normal gaussian distribution:

distribution of sample means



Normal Q-Q Plot



As the histogram above and the plot of sample quantiles versus theoretical quantiles indicate the sample means obtained from a simulated exponential distribution of a relatively small sample size ($n=40$) closely approximate a normal distribution.

Reference

1- Kissell, R. L., & Poserina, J. (2017). Optimal Sports Math, Statistics, and Fantasy. Academic Press.