Statistical Inference Course Project

Luis David Bedon Gomez

23 Mai 2017

Introduction

The following report discuss the questions asked for the first part of the Course Project of the Statistical Inference Course of the Data Science Specialization from the Johns Hopkins University an Coursera, concerning the dataset "ToothGrowth" from the R documentation.

1. Simulation of the Exponential Distribution in R and Comparisson with the Central Limit Theorem

In this part of the assignment 1000 averages of 40 randomized points will be generated by the "rexp"-function in R. Its distribution will be compared to the Gauß-distribution, therefore relating the performed simulation to the Central Limit Theorem.

1.1 Simulating the data

The simulation of the exponentially distributed data starts as follows:

First the working directory is set and the libraries are loaded:

```
setwd("~/DataScienceSpecialization/06StatisticalInference")
library(dplyr)
library(ggplot2)
library(extrafont)
```

A reproducible seed is set before the creation of the simulation data.

For the simulation, a function "fsim" is declared, that gives always the mean of 40 random numbers generated after the exponential distribution "rexp" with $\lambda = 0.2$, using the silly variable x.

```
set.seed(1)
fsim<- function(x,n=40,lambda=0.2) {
  mean(rexp(n,lambda))
}
```

At last, the data frame variable "sim" is created with the simulation for 1000 cycles.

```
sim<-data.frame(sim=sapply(1:1000, fsim))
```

1.2 Comparing the Sample Mean and the Theoretical Mean

1.2.1 Theoretical Data

From the exponential distribution ed we expect for the sample data, as explained in the instructions:

- $mean_{\text{theo}} = \frac{1}{\lambda}$
- $var_{\text{theo}} = \frac{1}{\lambda * n}$

In our example, with $\lambda = 0.2$, we get:

- the expected $mean_{\text{theo}} = 1/\lambda = 5$.
- the expected variance is $var_{\text{theo}} = \frac{1}{\lambda^2 * n} = 0.625$ with $n = 40$,
- so is the expected standard deviation $sd_{\text{theo}} = 0.791$.

1.2.2 Simulation Data

From the simulation we get:

```
paste("The sample mean is",round(mean(sim$sim),3))
```

```
## [1] "The sample mean is 4.99"
```

```
paste("The sample variation is",round(var(sim$sim),3))
```

```
## [1] "The sample variation is 0.611"
```

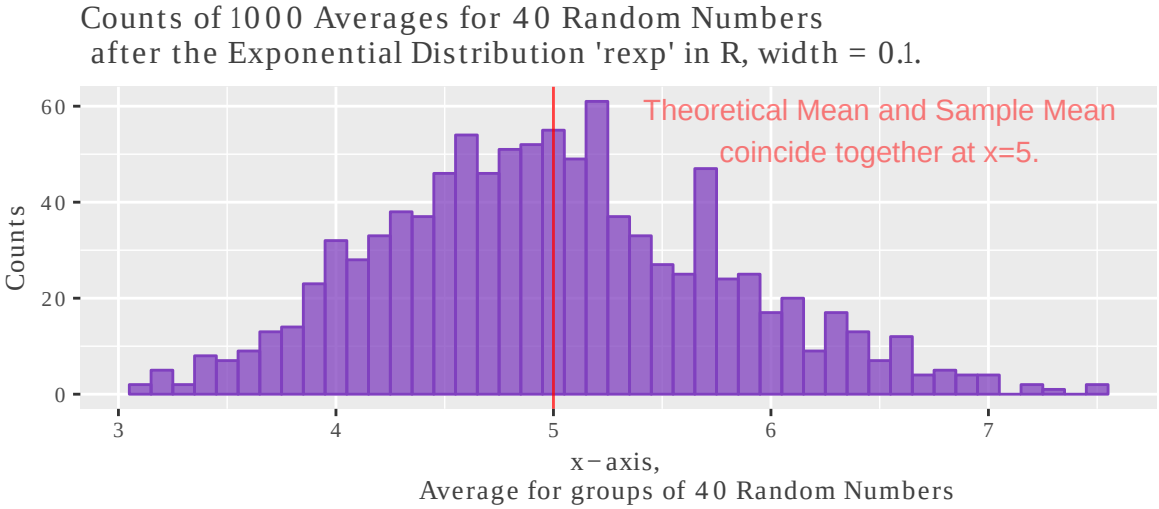
```
paste("The sample standard deviation is",round(sd(sim$sim),3))
```

```
## [1] "The sample standard deviation is 0.782"
```

1.2.3 Theoretical Mean and the Sample Mean

The obtain sample mean of 4.99 is only 0.2% lower than the expected value.

The mean in the obtained distribution can be seen in the following diagram:



1.2.3 Theoretical Variance and Sample Variance

Resuming, we expect from the theoretical variance and standard deviation:

- $\mathbf{var}_{\text{theo}} = 1/(0.2^2 * n) = \mathbf{0.625}$ with $n = 40$.
- $\mathbf{sd}_{\text{theo}} = 1/(0.2 * \sqrt{n}) = \mathbf{0.791}$ with $n = 40$.

We got from the simulation

- $\mathbf{var}_{\text{sim}} = \mathbf{0.611}$ with $n = 40$.
- $\mathbf{sd}_{\text{sim}} = \mathbf{0.782}$ with $n = 40$.

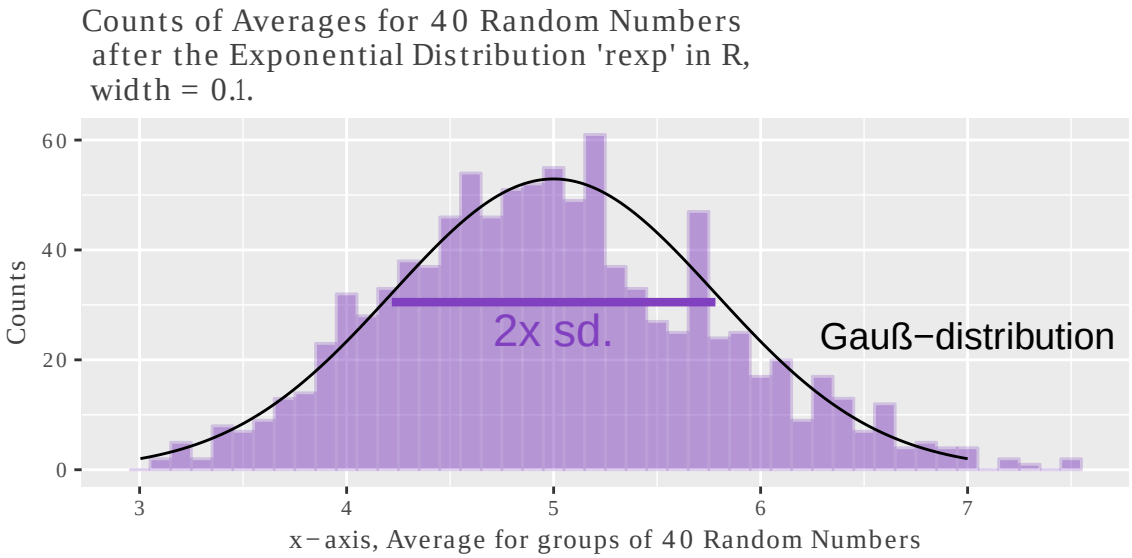
Taking the difference from each other and dividing by the theoretical value $(var_{\text{theo}} - var_{\text{sim}})/var_{\text{theo}}$ we see that the theoretical and sample variation differ by 2.24% from each other.

Regarding the theoretical and the sample standard deviation, they differ by 1.14% from each other.

1.2.3 Sample and Central Limit Theorem

From the Central Limit Theorem we know, that the averages of samples are approximately normal distributed for a large number of samples n.

In this exercise, we have $n = 40$ and the result can be seen in the following diagram:



The black line represents a Gaussian or normal distribution, and even for the 1000 averages of only 40 numbers of the population, this example seems to be related to the Gaussian curve.

To complement the first diagram, in this one you can appreciate the with that corresponds to the mean plus and minus the standard deviation. Together with the mean, this to components are responsible for the shape of the Gaussian distribution.

Appendix

Codes from the plots

Diagram 1

```
ggplot()+
geom_histogram(data=sim,
               aes(x=sim),
               binwidth=0.1,
               colour=rgb(.5,.25,.75,1),
               fill=rgb(.5,.25,.75,.75))+
scale_x_continuous(name="x-axis, Average for groups of 40 Random Numbers",
                   breaks=seq(0,10,1))+#,minor_breaks = seq(1000,24000,2000))+
scale_y_continuous(name="Counts")+#,breaks=c(1,2,3,4,5,6,7,8,9,10))+
geom_vline(aes(xintercept=5),colour="red")+
annotate("text",x=6.5,y=55,
         label="Theoretical Mean and Sample Mean\ncoincide
         together at x=5.",
         colour="red",
         alpha=.7)+
ggtitle("Counts of 1000 Averages for 40 Random Numbers\n
        after the Exponential Distribution 'rexp' in R,\n width = 0.1.")
```

Diagram 2

```
# Plot
g2 <- ggplot() + geom_histogram(data = sim, aes(x = sim), binwidth = 0.1, colour = rgb(0.5,
    0.25, 0.75, 0.25), fill = rgb(0.5, 0.25, 0.75, 0.5)) + geom_line(aes(x = sim$disx,
    y = sim$disy))

# Create a variable 'vg2' with the plotted data to extract the maximum of
# counts
vg2 <- ggplot_build(g2)
# Extract the maximum:
countsmax <- max(vg2$data[[1]]$count)


# Create distribution data
sim$disx <- seq(3.004, 7, 0.004)
sim$disy <- dnorm(seq(3.004, 7, 0.004), 5, sd(sim$sim)) * countsmax * 1.7


# Plot the histogram with the additional information for the variances
g2 + scale_x_continuous(name = "x-axis,
                    Average for groups of 40 Random Numbers",
    breaks = seq(0, 10, 1)) + scale_y_continuous(name = "Counts") + geom_segment(aes(x = 5 -
    sd(sim$sim), xend = 5 + sd(sim$sim), y = countsmax/2, yend = countsmax/2),
    colour = rgb(0.5, 0.25, 0.75, 1), size = 1.5) + annotate("text", x = 5,
    y = countsmax/2 - 5, label = "2x sd.", colour = rgb(0.5, 0.25, 0.75, 1),
    alpha = 1, size = 6) + annotate("text", x = 7, y = countsmax/2.5, label = "Gauß-distribution",
```

```
    colour = "black", alpha = 1, size = 5) + # geom_smooth(aes(x=vg2$data[[1]]$x,y=vg2$data[[1]]$count),
ggtitle("Counts of Averages for 40 Random Numbers\n
        after the Exponential Distribution 'rexp' in R,\n width = 0.1.")
```