Statistical Inference Project: Part 1

Angie Marchany-Rivera

7/22/2021

Overview

This project consists of 2 parts: a simulation excercise and basic inferential data analysis. Part 1 shows a comparison between the exponential distribution and the Central Limit Theorem (CLT). Part 2 shows the results of confidence intervals and hypothesis tests of the ToothGrowth dataset in R.

Part 1: Simulation Excercise

The means of 40 exponentials will be compared to a normal distribution. The following parameters will be used:

```
sim <-1000 # number  of simulations
n <- 40 # number of exponentials
lambda <- 0.2 # limiting factor
set.seed(340) # for reproducibility
theoMean <- 1/lambda # theoretical mean
theoSd <- 1/(lambda*sqrt(n)) # theoretical standard deviation
```

Simulating the means of 40 exponentials

```
expMeans = NULL
for (i in 1 : sim) expMeans = c(expMeans, mean(rexp(n,lambda)))
summary(expMeans)
```

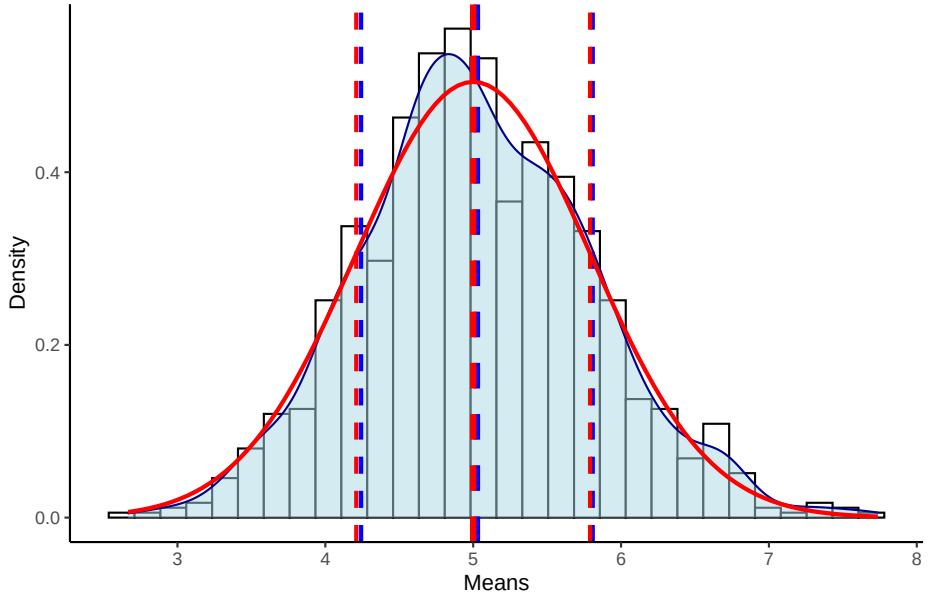
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.666   4.528   4.982   5.025   5.543   7.737
```

Sample Mean versus Theoretical Mean

Figure 1 shows the density distribution of the simulated means of 40 exponentials (solid blue line) compared to the correspoding theoretical normal distribution (solid red line). The centered blue dashed line represents the sample mean, 5.025. The centered red dashed line represents the theoretical mean, 5. It can be concluded that the sample mean is very close to the theoretical mean.

```
library(ggplot2)
df <- data.frame(expMeans)
ggplot(df,aes(expMeans))+
        geom_histogram(aes(y=..density..),colour="black", fill="white")+
        geom_density(color="darkblue", fill="lightblue", alpha=0.5)+
        geom_vline(aes(xintercept=mean(expMeans)),
            color="blue", linetype="dashed", size=1.5)+
        geom_vline(aes(xintercept=mean(expMeans)-sd(expMeans)),
            color="blue", linetype="dashed", size=1)+
        geom_vline(aes(xintercept=mean(expMeans)+sd(expMeans)),
            color="blue", linetype="dashed", size=1)+
        stat_function(fun = dnorm, args = list(mean = theoMean, sd = theoSd),
                      colour = "red", size = 1)+
        geom_vline(aes(xintercept=theoMean),
            color="red", linetype="dashed", size=1.5)+
        geom_vline(aes(xintercept=theoMean-theoSd),
            color="red", linetype="dashed", size=1)+
         geom_vline(aes(xintercept=theoMean+theoSd),
            color="red", linetype="dashed", size=1)+
        labs(title="Figure 1: Means of 40 exponentials vs. the theoretical normal distribution",
        x="Means", y = "Density")+
        theme_classic()
```

Figure 1: Means of 40 exponentials vs. the theoretical normal distribution



Sample Variance versus Theoretical Variance

The variances are calculated below.

```
var(expMeans) # sample variance
```

```
## [1] 0.6134735
```

```
(1/lambda)^2/n # theoretical variance
```

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

The sample variance is very close to the theoretical variance. This can also be verified by comparing the 68% confidence intervals shown on figure 1. The blue dashed lines show the 1-sigma region around the sample mean while the red dashed lines show the 1-sigma region around the theoretical mean. Both regions are very close to each other which indicates a normal distribution.

Comparing distributions

A Q-Q plot is used to compare the sample quantiles to the quantiles of a normal distribution. Figure 2 shows that the relationship between the sample quantiles versus the theoretical quantiles is almost linear. This relationship further confirms that the means of 40 exponentials density distribution approaches a normal distribution with a sample size of 1000.

```
ggplot(df,aes(sample = expMeans))+
        stat_qq()+
        stat_qq_line(alpha = 1, color = "red", linetype = "dashed")+
        labs(title="Figure 2: Means of 40 exponentials Q-Q Plot",
        x="Theoretical Quantiles", y = "Sample Quantiles")
```