Statistical Inference - Project

Gareth Coffey

26/05/2021

Overview

This is a paper that consists of two parts: 1. A simulation exercise investigating the exponential distribution in R and comparing it with the Central Limit Theorem (CLT). 2. Inferential data analysis of the ToothGrowth data in the R datasets pacakge.

Part 1 - Simulations investigating the exponential distribution in R

Here I perform the simulations of the exponentials and then calculate the means of those exponentials:

First, I replicate 1000 times the simulation of the exponential distribution using the expression rexp(n, lambda)

```
simulation_of_exponentials <- replicate(nosim, rexp(n, lambda))
means_of_exponentials <- apply(simulation_of_exponentials, 2, mean)
```

Some basic exploration of the simulation:

```
# Number of rows and columns of data:
dim(simulation_of_exponentials)
```

```
## [1]   40 1000
```

```
# Value range of the data:
range(simulation_of_exponentials)
```

```
## [1] 8.195983e-06 6.104031e+01
```

```
# First 10 values in the data:
simulation_of_exponentials[0:10]
```

```
##  [1]  0.3752038  1.8535227  6.7953219  0.7189461  0.3205374  4.5755572
##  [7]  4.6738252  4.4648434  0.3227221 13.8080176
```

Sample Mean versus Theoretical Mean

```
sample_mean <- mean(means_of_exponentials)
theoretical_mean <- 1/lambda
```

Question 1 - Show the sample mean and compare it to the theoretical mean of the distribution. The sample mean is:

```
## [1] 4.976277
```

The theoretical mean is:

```
## [1] 5
```

The sample mean is very close to the theoretical mean.

Sample Variance versus Theoretical Variance

```
sd_of_distribution <- sd(means_of_exponentials)
variance_distribution <- sd_of_distribution^2

sd_theoretical <- (1/lambda)/sqrt(n)
variance_theoretical <- ((1/lambda)*(1/sqrt(n)))^2
```

Question 2 - Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution The standard deviation of the sample distribution is:

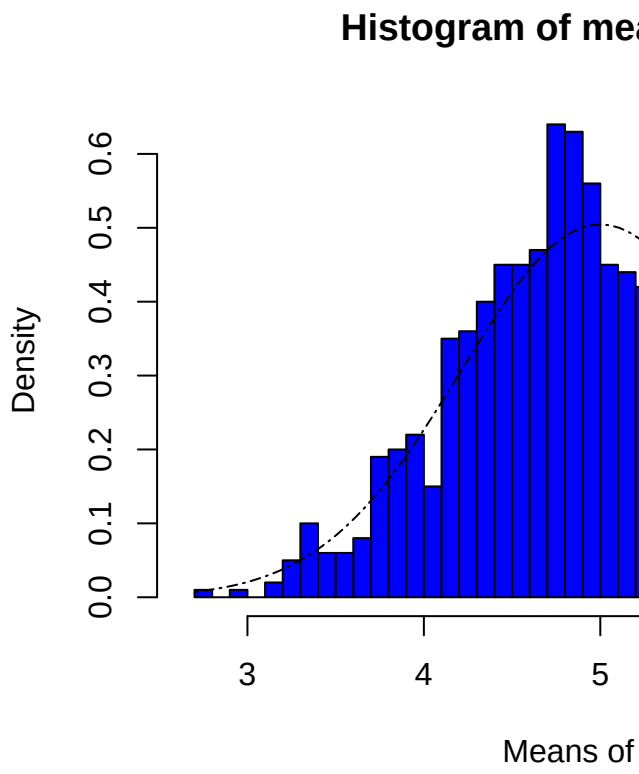
```
## [1] 0.6027554
```

The standard deviation of the central limit theorem is:

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

Distribution

```
x_axis <- seq(min(means_of_exponentials), max(means_of_exponentials), length=100)
y_axis <- dnorm(x_axis, mean=1/lambda, sd=(1/lambda/sqrt(n)))
hist(means_of_exponentials, breaks=n, prob=T, col="blue", xlab="Means of exponentials", main="Histogram
lines(x_axis, y_axis, pch=22, col="black", lty=6)
```



Question 3 - Show the distribution is approximately normal

The distribution of the values is approximately normal as it has the familiar bell-curve.

Part 2 - Inferential Data Analysis of ToothGrowth data

```
# Loading the dataset 'ToothGrowth'
data("ToothGrowth")
```

A basic summary of the data:

```
# A summary of the dataset
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

```
# The first few rows of data
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
# The dimensions of the data
dim(ToothGrowth)
```

```
## [1] 60  3
```

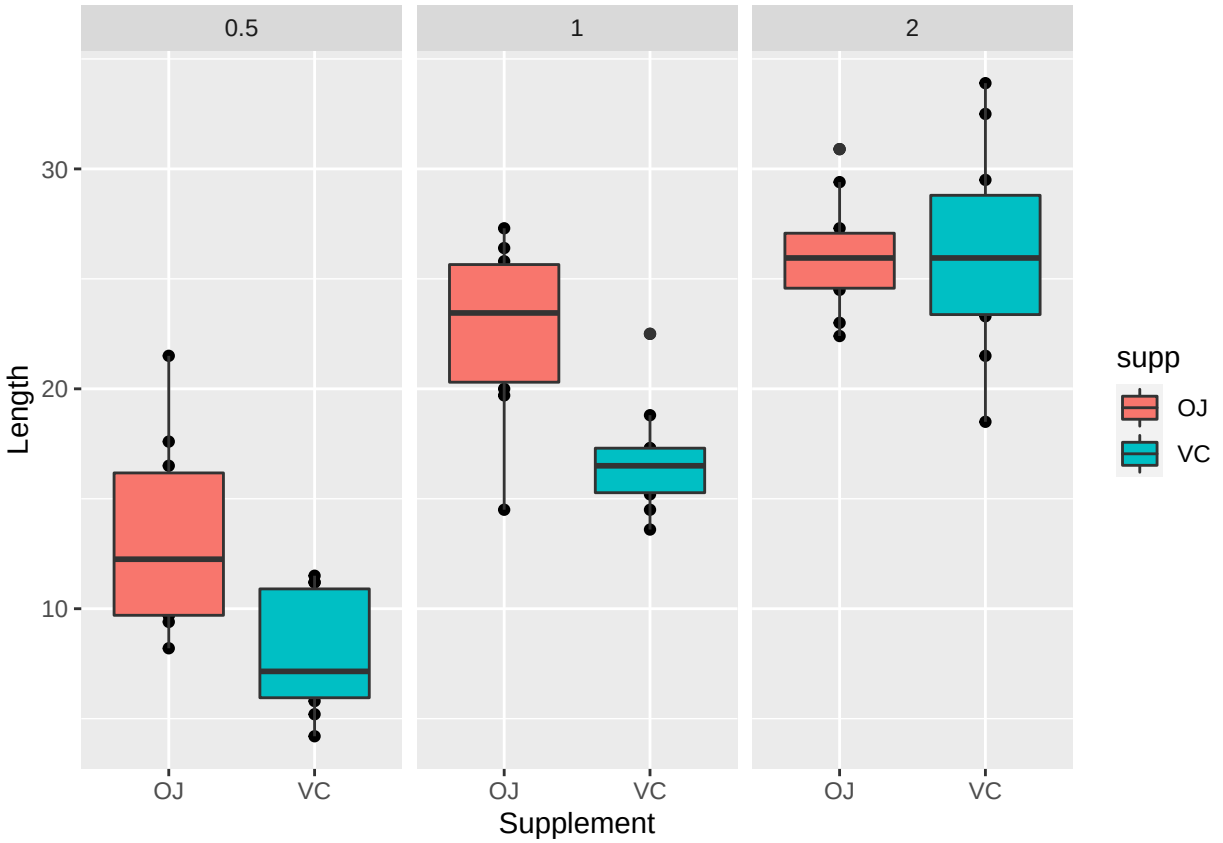
```
# The structure of the data
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

The data set has 60 observations, each with 3 variables: 1. len (number) 2. supp (a factor - supplement type VC or OJ) 3. dose (number)

Comparing the tooth growth in by supp and dose:

```
qplot(supp, len, data=ToothGrowth, facets=~dose,xlab="Supplement", ylab="Length") + geom_boxplot(aes(fi
```



It appears that as the dose increases, the tooth growth increases - this holds true for both supplements.

It seems from the plot above that supplement OJ induces higher relative growth between 0.5 and 1 mg – increasing the dosage to 2mg has little effect.

The growth increases fairly linearly for supplement VC.

Overall, it appears that supplement OJ causes more tooth growth except at the 2mg dosage level, where supplement VC causes higher growth.

Hypothesis Testing - Assumptions

There are a few assumptions that need to be made about the data before I conduct hypothesis testing:

1. The variables are IID
2. A normal distribution for tooth growth

Hypothesis Test - Supplement Type

Null Hypothesis (H0) Neither supplement causes any tooth growth.

Alternative Hypothesis (Ha) Supplement OJ induces more tooth growth than supplement VC.

Testing First, we need to split the dataset by the type of supplement:

```
oj_data = ToothGrowth$len[ToothGrowth$supp == 'OJ']
vc_data = ToothGrowth$len[ToothGrowth$supp == 'VC']
```

Conduct a t confidence interval test on the two sets of data:

```
t.test(oj_data, vc_data, alternative="greater", paired=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj_data and vc_data
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

A P-value of 0.03032 is lower than 0.05 (the 5% tolerance for error - i.e. our confidence level is 95% or 0.95). Therefore, the null hypothesis is rejected - the supplements cause tooth growth - we can say this with 97% confidence.

Hypothesis Test - Dosage

Null Hypothesis (H0) Dosage does not affect the level of tooth growth.

Alternative Hypothesis (Ha) As dosage increases so does the level of tooth growth.

Testing First, we need to split the dataset by the level of dosage:

```
doseHalfMg = ToothGrowth$len[ToothGrowth$dose == 0.5]
doseOneMg = ToothGrowth$len[ToothGrowth$dose == 1]
doseTwoMg = ToothGrowth$len[ToothGrowth$dose == 2]
```

Conduct a t confidence interval test on the 0.5mg vs 1mg

```
t.test(doseHalfMg, doseOneMg, alternative="less", paired=FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  doseHalfMg and doseOneMg
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Here, the P_value of 6.342e-08 is less than 0.05. Therefore, there is a negligle chance of getting a value that disagrees with our alternative hypothesis for doses of 0.5 and 1mg.

Now, let's test 1mg and 2mg:

```
t.test(doseOneMg, doseTwoMg, alternative="less", paired=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  doseOneMg and doseTwoMg
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Here the P-value is slightly higher at 9.532e-06, but is still far less than 0.05 so we can safely reject the null hypothesis. The slightly higher P-value may well represent the lower amount of growth we saw in the bar plot above, for the OJ supplement.