Statistical Inference Course Project

Andrea Eoli

2022-11-07

Instructions

The project consists of two parts:

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

You will create a PDF report to answer the questions. Each PDF report should be no more than 3 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).

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.

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.

Answer

Generate simulated data

```r
set.seed(1) # Set seed for reproducibility
lambda <- 0.2
n <- 40

# Simulate data
simul <- replicate(n = 1000, expr = rexp(n = n, rate = lambda))

# Calculate mean
mean_sim <- apply(simul, 2, mean)
```

Q1: Sample mean vs Theoretical mean

```r
# Sample mean
sample_m <- mean(mean_sim)
sample_m
```

```
## [1] 4.990025
```

```r
# Theoretical mean
theo_m <- 1/lambda
theo_m
```

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

Both Sample and Theoretical means are very similar, they both approximate to 5.

Q2: Sample variance vs Theoretical variance

```r
# Sample var
sample_v <- sd(mean_sim)^2
sample_v
```

```
## [1] 0.6111165
```

```r
# Theoretical var
theo_v <- (1/lambda/sqrt(n))^2
theo_v
```

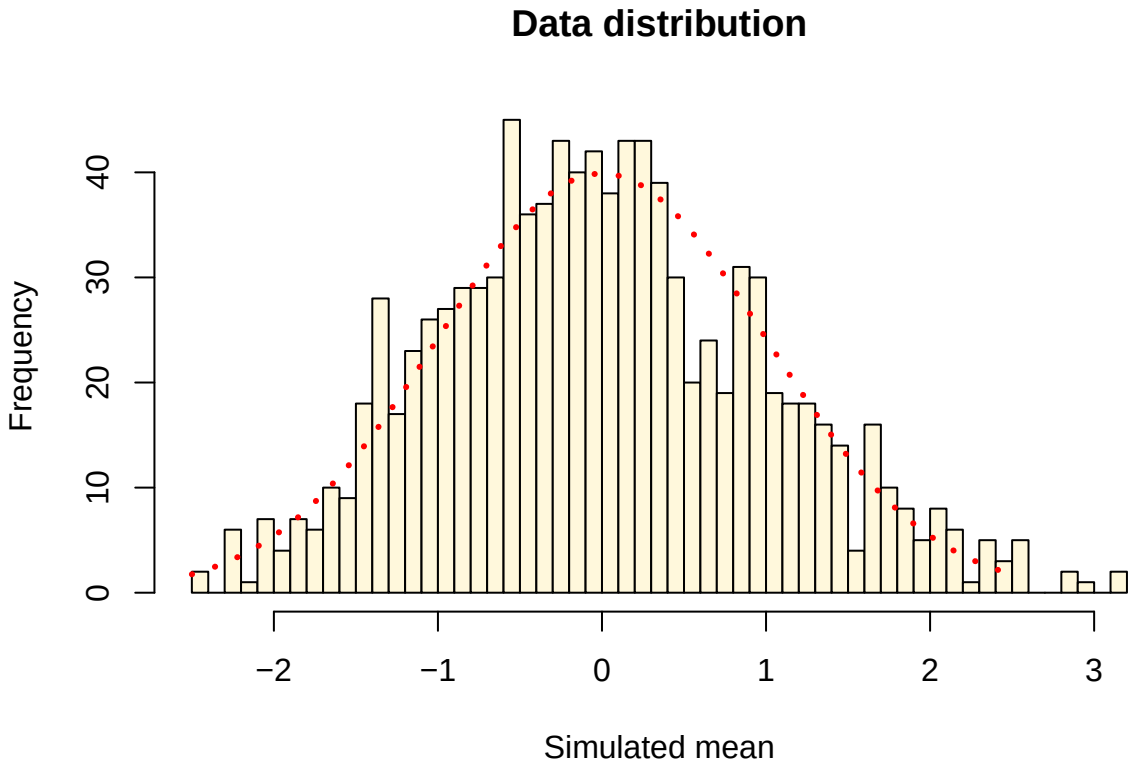
```
## [1] 0.625
```

Both Sample and Theoretical variants are very similar, being the first 0.61 and the second 0.63.

Q3: Check if the distribution is normal

```r
# Histogram
scaled_distr <- scale(mean_sim) # z-transform to make it easier
hist(scaled_distr, main = "Data distribution", breaks = 60, col = "cornsilk", xlab = "Simulated mean") 

xfit <- seq(-2.5, 2.5, length = 100)
yfit <- dnorm(xfit, mean = 0, sd = 1)
lines(xfit, yfit*100, lty = 3, lwd = 3, col = "red") # plot reference line
```



As we can see, because of the Central Limit Theorem, the distribution of the simulated data (here standardized to m = 0 and sd = 1) is approximately normal (red line).

Part 2: Basic Inferential Data Analysis Instructions

Andrea Eoli - Statistical Inference Course Project [2022-11-07]

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

Q1-2: Exploratory analysis and summary

Load the ToothGrowth data and perform some basic exploratory data analyses.

```
library(dplyr)
library(ggplot2)
data("ToothGrowth")
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
```

```
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
```

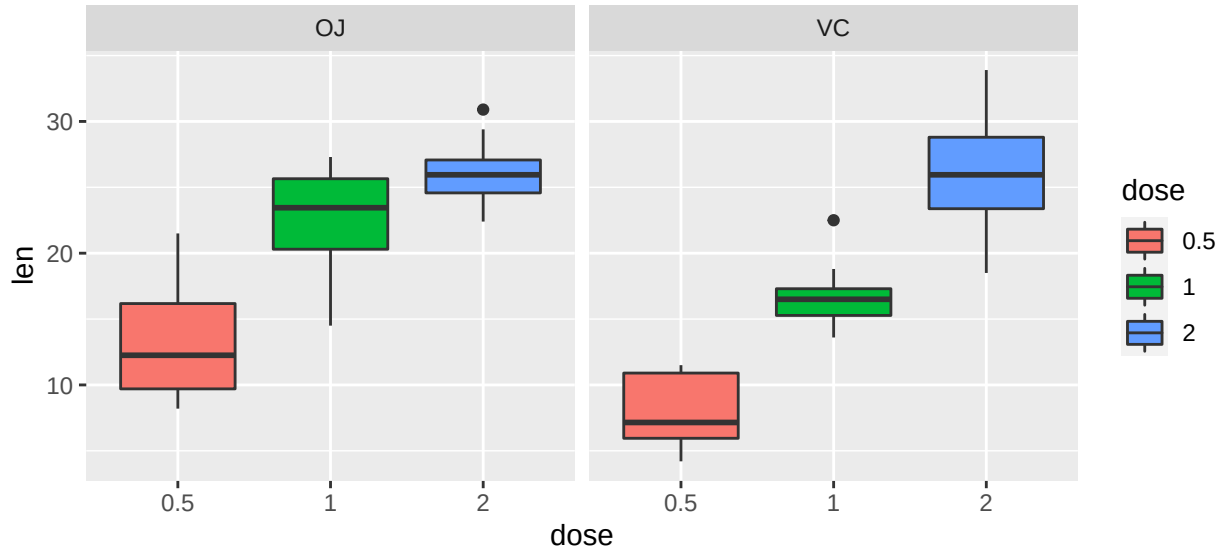
```
unique(ToothGrowth$dose) # Check if dose is really numeric -> only 3 values, convert to factor
```

```
## [1] 0.5 1.0 2.0
```

```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

Let's visualise the data as a boxplot of the tooth growth by dosage and type of supplement.

```
ggplot(aes(x=dose, y=len), data=ToothGrowth) +
  geom_boxplot(aes(fill=dose)) +
  facet_wrap("supp")
```



Q3: Compare tooth growth by supp and dose

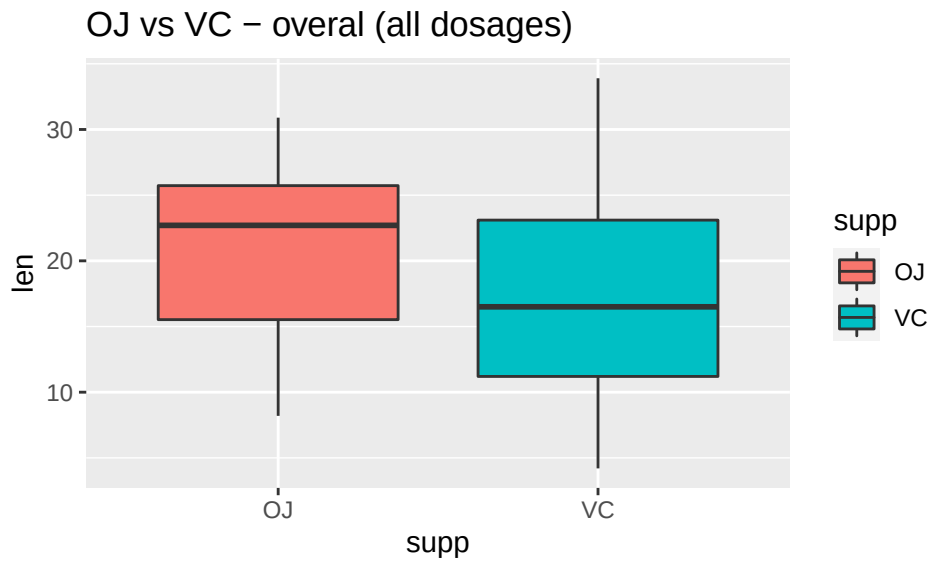
Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) We assume unequal variances between the two groups and test, for each group, growth length between low and high doses.

```
a <- t.test(len ~ dose, data = ToothGrowth[ToothGrowth$dose %in% c(0.5,2) & ToothGrowth$supp == "OJ",])
b <- t.test(len ~ dose, data = ToothGrowth[ToothGrowth$dose %in% c(0.5,2) & ToothGrowth$supp == "VC",])
```

For both supplements the p-value is much smaller than 0.05 (1.3237839×10^{-6} for OJ, 4.6815774×10^{-8} for VC), so we reject the null hypotheses: an higher dose has a significantly higher mean, for both supplements.

Let's now test if one supplement is better than the other.

```
ggplot(aes(x=supp, y=len), data=ToothGrowth) +
  geom_boxplot(aes(fill=supp)) + ggtitle("OJ vs VC - overal (all dosages)")
```

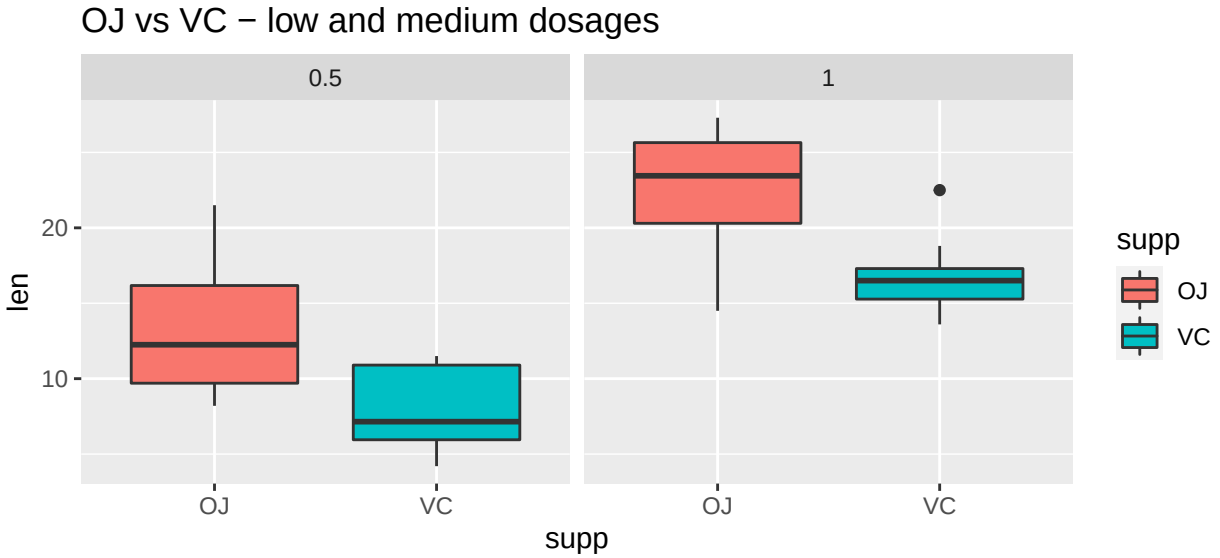


```r
t <- t.test(len ~ supp, data = ToothGrowth)
```

The p-value is 0.0606345, so we fail to reject the null hypothesis, meaning that there is no significant different between the mean of the two supplements.

However, looking at the first boxplot, it seems that OJ has a greater effect on tooth length at a low (0.5) and medium (1) dosage. Let's test it.

```r
ggplot(aes(x=supp, y=len), data=ToothGrowth[ToothGrowth$dose %in% c(0.5,1),]) +
  geom_boxplot(aes(fill=supp)) + facet_wrap("dose") +
  ggtitle("OJ vs VC - low and medium dosages")
```




```r
# "OJ vs VC - low dosage"
a <- t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 0.5,])
# "OJ vs VC - medium dosage"
b <- t.test(len ~ supp, data = ToothGrowth[ToothGrowth$dose == 1,])
```

For both tests, the p-value is smaller than 0.05 (0.0063586 for low dose, 0.0010384 for medium dose) so we reject the null hypotheses and we can confirm that at the same dose the supplement OJ has a greater effect on tooth growth.

Q4: Conclusions

1. Overall, there is no difference in supplement type on tooth growth
2. However, when looking at individual dosages, OJ is better than VJ in stimulating tooth growth at low (0.5) and medium (1) doses.
3. For each supplement type, the high dose is better than the low dose in stimulating tooth growth.

Assumptions
For each t test performed above, variances are assumed to be different.