Statistical Inference Course Project

Cristobal Morgado

2022-09-07

Introduction

The following report is composed by two parts. The first one generate a distribution of the means and variances of an exponential distribution, compare the estimated parameters with the theoretical ones and plot the resulting distributions.

The second part is a series of significance test of the ToothGrowth database, testing differences between treatment and doses.

Code is include in the rmd file in: https://github.com/CMorgadoM/Statistical_Inference_Project

Part 1: Simulations

1.1 - Generate data

The first step is to generate the simulations data, here I've used the rexp() function to generate 1000 samples of 40 observations, with lambda = 0.2. Then get the mean and variance of every individual sample and store them the "avg" and "var" variables respectively.

1.2 - Plotting the results

Once generated the estimated parameters of the mean and variance of the distribution I've used ggplot to create an histogram and look at the form of the distribution. The resulting plots are presented in the following figures, when we can observe that the distribution of the means resembles a normal distribution, on the other hand, the distribution of the variance resembles more a poisson distribution.

In both cases the sample average of the parameters (mean of 4.9835069 and variance of 24.9535481), showed by the red lines, are close to the theoretical parameters, represented by the green line. **(Results showed in appendix 1)**

Part 2: Statistical inference

2.1 - Understanding the database

The database for the project is the ToothGrowth database, that measure the growth of 60 guinea pigs tooth, treated by two types of supplements, orange juice (OJ) and a form of vitamin C called ascorbic acid (VC). The treatments are supplemented in doses of 0.5, 1 and 2 mg/day. The database contains the following variables:

1.-len(num) = indicates the length of the tooths at the end of the treatment

2.-supp(factor) = indicates the treatment method

3.-dose(num) = indicates the doses in mg/day

2.2 - Summary of the data

```
##       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
```

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
len	60	18.813	7.649	4.200	33.900
dose	60	1.167	0.629	0.500	2.000

(Distribution of results showed in appendix 2)

2.3 - Statistical test

Test 1: Is there a difference between treatment methods?

H0: mean(OJ) = mean(VC)

Conclusion: Confidence interval with 95% confidence is equal to -0.7510928, 8.1510928 and includes zero, therefore we fail to reject H0 and there is no evidence that the treatment affect differently on the toothgrowth for guinea pigs. Also P-value is equal to 0.0606345, and we fail to reject the null hypothesis under 95% confidence.

Test 2: Is there a difference between doses?

H0a: mean(0.5) = mean(1)

H0b: mean(1) = mean(2)

H0b: mean(0.5) = mean(2)

a) dif 0.5 mg/day and 1 mg/day

With a P-value of 1.2683007×10^{-7} we reject the H0 and conclude that there is significant difference between 0.5 and 1 mg per day at any reasonable significance level.

b) dif 1 mg/day and 2 mg/day

With a P-value of 1.9064295×10^{-5} we reject the H0 and conclude that there is significant difference between 1 and 2 mg per day at any reasonable significance level.

c) dif 0.5 mg/day and 2 mg/day

With a P-value of 4.397525×10^{-14} we reject the H0 and conclude that there is significant difference between 0.5 and 2 mg per day at any reasonable significance level.

test 3: What treatment-doses pairs are signiticantly different?

For this test, as it's being tested 9 hypothesis, I've applied a **Bonferroni correction** as the criteria to find the significant differences. This has been done saving the p-values in the pValues variable and applied the p.adjust function to get the p-values adjusted by the Bonferroni correction.

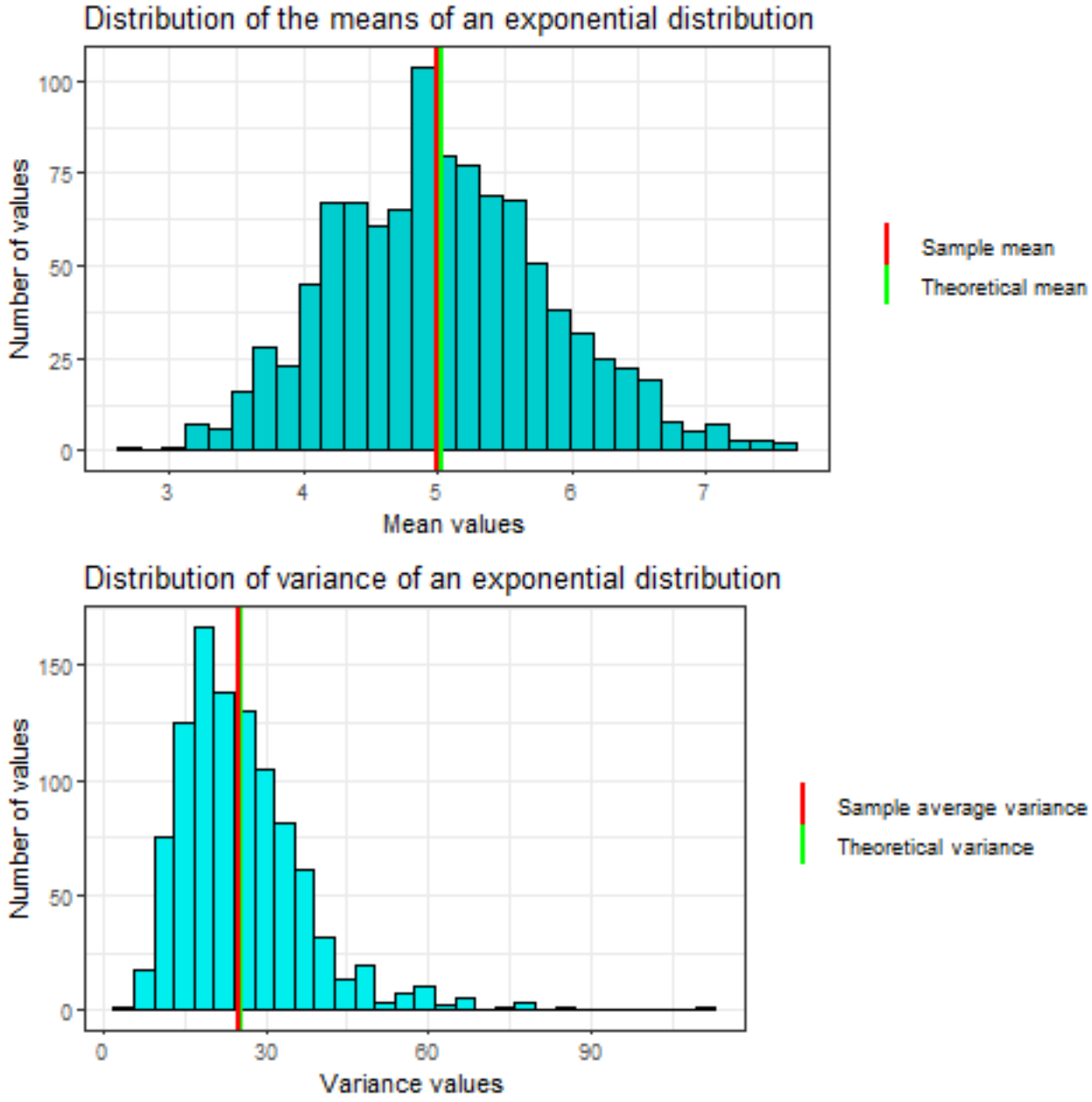
Also, since the samples have only 10 observations in every table, a **bootstrap** method is used to generate 1000 iterations of the means of resamples values, with repetitions, to be able to draw better conclusions from the data.

Conclusions: The following results have a significant difference

Significant test at 5% with bonferroni correction	
Pair_treatment_dose	adj_values
OJ-VC, 0.5mg/day	0.0000000
OJ-VC, 1mg/day	0.0000000
OJ, 0.5-1 mg/day	0.0000000
OJ, 0.5-2 mg/day	0.0000000
OJ,2-1 mg/day	0.0303419
VC, 0.5-1 mg/day	0.0000000
VC, 0.5-2 mg/day	0.0000000
VC, 2-1 mg/day	0.0000000

Appendix

1) Plot 1: Exponential distribution, in red sample mean and in green theoretical mean



2) Plot 2: Distribution of results in ToothGrowt database