Estadística avanzada para ciencia de datos

Contrastes de Hipótesis - Ejercicio 1 Tests Estadísticos

Jakub Maciążek

Exercise description

> **Scenario** A medicine (M) has been developed that improves the results in the Computer Engineering Master's exams. The drug (M) or a placebo (P) is given to a group of students and scores are collected on a questionnaire (from 0 to 100) of the students in that group. It is desired to verify if there are differences between the students who took the medication and those who took the placebo (control group) with respect to the results of said questionnaire.

Given above scenario follow instructions

Prepare an RMarkdown file explaining the steps to follow to study the proposed case. Deliver in the CV in a task the .Rmd file along with the output:

- Download the file experimento1.csv from the datasets directory of the CV.
- Check if there is a difference in the results due to taking the medication.
- Use the statistical tests seen in class. Explain the results.

Data analysis

Main goal of the exercise is to determine, whether Medication has an effect on students performance. This can be achieved by comparing average result between the gropus, and comparing how different it was. Also important is to check whether that statistical information is significant, therefore reliable.

Data loading and primary samples analysis

First step to follow, is to load given data and analyze it's format, in order to decide on what tests to choose.

```
experimento1 <- read_delim("S:/0_Universidad_de_Malaga/MI_Ingenieria_y_ciencia_de_datos/Estatistica_ava
## Rows: 48 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (1): Producto
## dbl (1): Nota en un testc- 1:100
## lgl (1): (M: Medicamento , P: Placebo)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

head(experimento1)
```

```
## # A tibble: 6 x 3
##   `Nota en un testc- 1:100` Producto `(M: Medicamento , P: Placebo)`
##                       <dbl> <chr>    <lgl>
## 1                        43 M        NA
## 2                        65 P        NA
## 3                        31 M        NA
## 4                        54 P        NA
## 5                        41 P        NA
## 6                        37 M        NA
```

From above information, we can see, that relevant data is presented in two columns, 1st one indicating the score, and 2nd one indicating whether student took Medicine or Placebo.

Data preparation

In order to conduct the test, data needs to be divided into two vectors, 1st one containing scores, and the 2nd one corresponding to information about taken Medication or Placebo.

```
scores <- experimento1[[1]]

labels <- experimento1[[2]]
```

Running tests

In order to conduct tests, we need to state null hypothesis against which test will be conducted and alternative hypothesis, therefore:

- H_0: Mean exam score does not differ between students that took Medicine or Placebo. (Meaning it has no effect)
- H_A: Mean exam score does differ between students that took Medicine or Placebo. (Medicine has an effect)

To choose the test, we need to think about 3 requirements samples need to pass: independence, normal distribution and equal variances:

- In this case we can assume independence of samples, as results of one student are not related to any other one.

- Regarding distribution, our test statistic is the note from the exam, score between 0 and 100. If this is a count of correctly answered questions, it would be not a continuous therefore binomial distribution. If this is any number from given range, being a continuous distribution, data would follow required normal distribution. As it is not stated, later one was assumed, allowing null hypothesis to predict that samples are normally distributed.

- Variance among samples is unknown, therefore for the test it will not be stated that it is equal.

Considering all above, to test if difference between two means is real, we will use **Two Sample T-test**, and as the variance equality is unknown, **Welch's Test** will be conducted:

```
t.test(scores ~ labels)
## 
##  Welch Two Sample t-test
## 
## data:  scores by labels
## t = -0.087763, df = 45.996, p-value = 0.9304
## alternative hypothesis: true difference in means between group M and group P is not equal to 0
## 95 percent confidence interval:
##  -18.94914  17.36581
## sample estimates:
## mean in group M mean in group P 
##        46.08333        46.87500
```

Test result analysis and conclusions

Given that p-value, which is smallest Type I error rate ( ) that we have to be willing to tolerate if we want to reject the null hypothesis, is very high, and equal to 0.9304, when by convention it should be at most 0.05, we cannot qualify result of the test as significant, and reject N_0.

This is because probability of making Type I Error would be ~93%. **Therefore, based on given data we cannot conclude that there is a statistically significant difference in the means of compared samples, and we can not confirm that the Medicine has any effect.**