Guest Exercise session 2021

Paavo Raittinen

3rd of March 2021

Homework problem (to be solved before the exercise session)

The new minister of education of Never-never land strongly believes that hair color has an effect on ability to learn. The null-hypothesis "Hair color does not have an effect on learning outcomes" was tested at a significance level of 5% in two different schools. The p-values of the tests were both larger than 0.05 and the minister of education was not able to reject the null hypothesis. The minister was very unhappy with the results and decided to test the same null-hypothesis in each of the Never-never land's 3000 schools. The p-value of the test was less than 0.05 in 137 schools. The minister of education is excited about the test results - finally his belief has been proven right! What went wrong, explain?

Classroom exercise problem

In this exercise you are a statistician and your task is to find whether new miracle drug induces changes in lipdomic profile. The study setting is placebo controlled clinical trial with planned 1:1 ratio where placebo arm received daily dose of placebo drug and treatment arm reiceived daily dose of new miracle drug. The dataset is high-dimensional so you will most likely need to use 'apply' function or create for loops. Moreover, due to high-dimensionality, we would like to control for false discoveries, i.e., type I error. Note that, the data is simulated by the author.

You should already know what is t-test, Wilcoxon rank sum test, and Shapiro-Wilk normality test - we will apply those tests. However, you will need some R coding hacks to be efficient. What is new: p-value adjustment.

a) Read the data into R using *read.csv()* function.

b) Describe the data, i.e., calculate statistics such as mean, standard deviation, and maximum. Useful function: *summary()*. What is the placebo / treatment ratio?

c) Separate normally distributed and non-normally distributed covariates using *shapiro.test()*. That is, create two new data frames; both data frames should contain the labels and covariates which are either normally distributed or are non-normal. Use confidence level alpha = 0.05 in the *shapiro.test()*. Note that, we should do the shapiro.test for one of the groups only; let's use Placebo group. This is a three step process:

1. Calculate the p-values and save them into a new vector. Useful function: *apply()*
2. Find which of the p-values are smaller than alpha. Useful function: *which()*

3. Create two new data.frames: normally distributed and not-normal distributed variables.

d) conduct a t.test for normally distributed variables to find statistically signifcant difference between the placebo and the treatment arm using confidence level alpha = 0.05. Which variables are significantly differing?

e) conduct a Wilcoxon Rank Sum test for non-normally distributed variables to find statistically signifcant difference between the placebo and the treatment arm using confidence level alpha = 0.05. Which variables are significantly differing?

f) To control false discovery rate, adjust the p-values using Benjamini-Hochberg p-value adjustement. Which variables are significantly differing now? Useful function: *p.adjust()* and *which()*.

optional: Use random forest classification to classify the data by the studyarm. Study how the so-called variable importance in random forest matches with your findings in part d-f. Use package "randomForest", which implements the original random forest algortihm by Leo Breiman. Note that, this is beyond the scope of the course Statistical Inference; you should self-study random forest before implementing it in any application outside this exercise.