Assignment 1

Deadline: 19/5/2023

Instructor: Panagiotis Papastamoulis
papastamoulis@aueb.gr
Department of Statistics
Athens University of Economics and Business

For this assignment you will need to install the Bioconductor packages leukemiaEset [1] and qvalue [2]. At first you should follow the installation instructions, as shown in the following link: https://www.bioconductor.org/install/. Next, install the two packages using:

```
BiocManager::install("leukemiasEset")
BiocManager::install("qvalue")
```

You will also use base R functions such as prcomp() and p.adjust() among others.

Exercise 1 (Leukemia's microarray gene expression data)**.** Open R and obtain the Leukemia dataset from the leukemiasEset package in Bioconductor.

```
library("leukemiasEset")
data(leukemiasEset)
x <- exprs(leukemiasEset)
```

The dataset (x) contains expresion data for 20172 genes from 60 bone marrow samples of patients with one of the four main types of leukemia:

- ALL: Acute Lymphoblastic Leukemia
- AML: Acute Myeloid Leukemia
- CLL: Chronic Lymphocytic Leukemi
- CML: Chronic Myeloid Leukemia
- NoL: non-Leukemia

There are 12 samples per class, which can be retrieved using the command

```
> leukemiasEset$LeukemiaType
```

Let j denotes the last digit of your student identification number. We are interested to test which genes are differentially expressed between the condition c and NoL groups, where

- c = ALL if $j \leqslant 2$
- c = AML if $3 \leqslant j \leqslant 5$
- c = CLL if $6 \leqslant j \leqslant 7$
- c = CML if $8 \leqslant j \leqslant 9$

So your dataset should consists of a matrix with 20172 rows (gene expression measurements) and 24 columns (12 replicates for each one of the two experimental groups).

1. Explore and visualize the data. Focus on the research question and try to visually describe the variability of the average gene expression between the two groups. Produce some meaningful summaries and descriptive statistics for your dataset.

2. Use PCA in order to visualize the dataset (20172×24). Project the data on the first few principal components and explain your findings. Do the same when considering the transposed input data (24×20172). Describe what you see.

3. Use two independent samples t-tests (you may assume that the variance is equal between groups) in order to test the null hypothesis per gene. State the null and alternative hypothesis per gene, as well as the assumptions you use to model the data. Plot a histogram (relative frequencies) of the p-values.

4. Can you give a rough estimate of the proportion of true null hypotheses?

5. Report how many genes are differentially expressed when controlling the FWER, FDR and pFDR at $\alpha = 0.01$.

6. Visualize the results obtained in question 5 according to whether the corresponding hypothesis is rejected or not when controlling the FDR at 0.01:

   (a) Plot a meaningful summary of the data and colour the genes depending on the result of the test (Differentially Expressed or not Differentially Expressed when controlling the FDR at the given level). Try to take into account both the mean difference as well the standard deviation per gene. Be creative.

   (b) using Principal Components projections.

   and explain your findings.

Exercise 2 (Multiple testing simulation study)**.** Simulate a synthetic dataset from a normal linear model with $n = 500$ observations and $p = 100$ explanatory variables, as follows:

1. Simulate the explanatory variables from independent normal distributions:

   ```
   x <- matrix(rnorm(n*p),nrow = n, ncol = p)
   ```

2. Generate the p regression coefficients $\beta_1, \ldots, \beta_p$ as follows:

   ```
   b <- numeric(p)
   if( runif(1) < 0.3){ b[1] <- rnorm(1) }
   ```

   This means that $\beta_i = 0$ for all $i \geqslant 2$, while the first coefficient (β_1) is zero with probability 0.7, while it is different than zero with probability 0.3.

3. Generate the values of the response variable from a typical normal linear model, that is,

   ```
   y <-  x%*%b + rnorm(n)
   ```

Repeat Steps 1, 2, 3 for $m = 10000$ times (so you will generate 10000 regression datasets). For each synthetic dataset we are interested to test the hypothesis that the response variable is not linearly depending on any of the p explanatory variables, that is,

$$H_0(j) : \beta_1 = \ldots = \beta_p = 0 \quad \text{vs} \quad H_1(j) : \beta_i \neq 0 \quad \text{for at least one} \quad i = 1, 2, \ldots, p,$$

for $j = 1, \ldots, m$. Apply the standard F-test for this purpose. Recall that the p-value of the F-test is returned in the `summary()` method of the `lm()` command. You should extract the p-values for each one of the 10000 synthetic datasets. Since you are generating the data, you know which null hypotheses are true or not. Test all 10000 hypotheses and control the type I error rate using all methods (`c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY")`) described in the `p.adjust()` command of `R`, as well as the q-value.

1. Report a confusion matrix per method with respect to the ground-truth, when controlling the relevant type I error at the $\alpha = 0.05$ level. What is the estimated power (proportion of true discoveries with respect to the total number of non-true null hypotheses) for this target value ($\widehat{\text{power}}(0.05)$), per method?

2. Plot the points $(\alpha, \widehat{\text{power}}(\alpha))$ for a sequence of values $\alpha \in (0, 1)$, that is, the estimated power versus the type I error control-value, for each method (see the relevant plots in the slides of Unit 1). Comment on the ranking of methods.

Advice: be gentle to your machine. There is no need to save 10000 simulated datasets. All you need is the vector of 10000 p-values and the ground-truth per tested hypothesis.

References

[1] S. Aibar, C. Fontanillo, J. D. L. R. Bioinformatics, and F. G. G. C. R. C. S. Spain. *leukemiasEset: Leukemia's microarray gene expression data (expressionSet).*, 2020. R package version 1.26.0.

[2] J. D. Storey, A. J. Bass, A. Dabney, and D. Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2021. R package version 2.26.0.