**Master IMSD - Mines 3A, Université de Lorraine** **2024-2025**
**Modélisation Statistique des données complexes**

**TP1 : Multiple testing**

# 1 Objectives

1. Use R software
2. Use simulations to verify mathematical properties empirically
3. Understand the risk of type I associated with a test
4. Understand the power of a test
5. Understand the problem of multiple testing
6. Use R packages to adjust p-values
7. Compare the methods on simulated data
8. Use the methods on real data
9. Code the different methods to adjust p-values

# 2 Multiple testing and p-values

The purpose of this section is to study the p-values behaviour under hypothesis $H_0$ and/or $H_1$.

**Empirical Type I error**

The purpose of this section is to study the empirical type I error in the case of the Student-test of comparison of the means of two series

1. Simulate a vector of binary variable $Y$ of size $n = 50$
2. Simulate a matrix of quantitatives covariates $\mathbf{X}$ of size $n = 50$ and $p = 100$ that is not linked to $Y$.
3. What is the appropriate test to verify if one covariate is linked to $Y$ ?
4. Perform the test for each covariate and save the p-values in a vector. What is the min and the max of the p-values ? Plot the histogram. Comment.
5. Compute the number of False Positive Significant tests for a level $\alpha$
6. Compute the empirical type I error and compare to $\alpha$
7. Do it again for $p = 100$, $p = 1000$ and $p = 10000$.

**Empirical Power of test**

8. We want now to simulate a matrix $X$ of covariates of size $n = 50$ and $p_1 = 100$ that are linked with $Y$ by the following way.
   (a) Simulate $p_1$ covariates such that with probability 0.7, $X_j \sim \mathcal{N}(3Y, 1)$, and with probability 0.3, $X_j \sim \mathcal{N}(0, 1)$. The notation is

$$X_j \sim 0.7\mathcal{N}(3Y, 1) + 0.3\mathcal{N}(0, 1)$$

   — Perform the test for each covariate and save the p-values in a vector. What is the min and the max of the p-values ? Plot the histogram. Comment.
   — Compute the number of Significant tests for a level $\alpha$

— Compute the empirical power of the test.

(b) Do the same trick for $p_1$ other covariates :

$$X_j \sim 0.7\mathcal{N}(2Y,1) + 0.3\mathcal{N}(0,1)$$

(c) Do the same trick for $p_1$ other covariates :

$$X_j \sim 0.7\mathcal{N}(Y,1) + 0.3\mathcal{N}(0,1)$$

(d) Do the same trick for $p_1$ other covariates :

$$X_j \sim 0.3\mathcal{N}(3Y,1) + 0.7\mathcal{N}(0,1)$$

(e) Do the same trick for $p_1$ other covariates :

$$X_j \sim 0.3\mathcal{N}(2Y,1) + 0.7\mathcal{N}(0,1)$$

(f) Do the same trick for $p_1$ other covariates :

$$X_j \sim 0.3\mathcal{N}(Y,1) + 0.7\mathcal{N}(0,1)$$

(g) Comment the results

(h) Do it again for $n = 100$ and $n = 200$, comment

**Mixture of $H_0$ and $H_1$**

Put together the $6 * p1$ covariates simulated above in a same data frame, and add $6 * p1$ other covariates independent from $Y$.

9. Compute the vector of the p-values of the $6 * p_1$ tests that test the link between the $X_j$ and $Y$ (for $1 \leq j \leq 6 * p_1$). Look and the min, the max and do the histplot, comment.
10. Compute the vector of the p-values of the other $6 * p_1$ tests that test the link between the $X_j$ and $Y$ (for $6 * p_1 + 1 \leq j \leq p$). Look and the min, the max and do the histplot, comment.
11. Aggregate the two vectors of the p-values and look and the min, the max and do the histplot, comment.
12. Choose a risk $\alpha$ for each test and compute each entry of the following table (frequencies associated with the test's decisions). Remember what is fixed and what is dependent of the simulation.

| | Significant | Not significant | Total |
|---|---|---|---|
| Null true | V | U | $p_0$ |
| Alternative true | S | T | $p - p_0$ |
| Total | R | p-R | p |

## 3 Adjusted p-values

1. We have compute the p-values and we will now compute the adjusted p-values according to different methods. For each method, give the frequencies of the table of question 12, compute the FDR. The command is `adj.pv<-p.adjust(pval, method = "method")`, use the help of the command to see the different methods.

   (a) For each of the method, Bonferroni, Sidak, Holms and Bejamini-Hochberg, compute the vector of adjusted p-values, look at the min, the max, plot the histogram and comment.

   (b) Compare the previous methods by plotting the point clouds of each method 2 to 2.

(c) To compute the q-values, download package `qvalue` available on bioconductor thanks to the following code :

```
source("https:\\bioconductor.org\biocLite.R")
biocLite("qvalue")
```

depending on your version of $R$

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")

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

(d) Compute the qvalues by the following code and do the table and the comparison

```
qvalcom<-qvalue(pval2)
qval<-qvalcom $ qvalues
```

2. The command `qvalue` compute also $\hat{\pi}_0 = \dfrac{\#\{p_i > \lambda\}}{m(1-\lambda)}$ for different values of $\lambda$. Compare the result with the real $\pi_0$

## 4 Application to real data

The dataframe "datacancer" gives informations (103 variables) about 143 patients who received a treatment for a cancer. The purpose is to study the link between covariates and the survival at 6 or 12 months after the beginning of the treatment. The first column is the ID of the patients. The followings gives the time of treatment (TRT)and survival (OS) that could be censored (censoros ; censortrt). Interest variable are the survival at 6 and 12 maths (Surv6, Surv12) the following variables are covariates that could explain the survival. The purpose is to use multiple testing methods to select the covariates linked with the survival.

1. Select the qualitative covariates and perform the appropriate tests to see the link with Surv12. Save the p-values in a vector.
2. Select the quantitative covariates and perform the appropriate tests to see the link with Surv12. Save the p-values in a vector.
3. Aggregate the two vectors and adjust the p-values with different methods. Gives the variables that can explain the survival.