Chapter 7: Multiple Testing

1 Session 1

Objective: ajustement of p-values, correction of Benjamini-Yekutieli, influent covariates.

Exercise 1.1. Import data sets LenzT.rds and LenzI.rds.

1. Import data sets LenzT.rds and LenzI.rds and call them respectively Y and X. Display the dimension of Y. Sample at random 1000 genes from Y.

2. Compute the p-values corresponding to the comparison of the level of expression of genes according to the regime of patients with a `t.test`. How many p-values are below 5% ?

3. Adjust the p-values with the Bonferroni correction. Sort the p-values and display the 10 smallest ones. How many p-values are below 5% ?

4. Repeat with the Benjamini-Hochberg and the Benjamini-Yekutielli corrections. How many p-values are below 5% ?

5. Display the plot of the ecdf of the four vectors of p-values. Comment the difference between the p-values. Order the corrections from the less to the most conservative. Which correction should you prefer ?

6. Are the 10 most influent genes the same depending on the corrections ? Are they order the same way ? What could be the conclusion of this analysis ?

7. Compare the level of expression of genes according to the regime with a variance test and the Benjamini-Yekutieli correction. Compare the most influent genes obtained with those of question 4. Comment.

8. Repeat with the Wilcoxon test (using function `wilcox.test`). What is your conclusion ?

Exercise 1.2. Import data sets LenzT.rds and LenzI.rds.

1. Import data sets LenzT.rds and LenzI.rds and call them respectively Y and X. Display the dimension of Y. Sample at random 1000 genes from Y.

2. Compute the p-values corresponding to the correlation between age and the level of expression of genes with a `t.test`. How many p-values are below 5% ?

3. Adjust the p-values with the Bonferroni correction. How many p-values are below 5% ? Comment the difference. Sort the p-values and display the 10 smallest ones.

4. Repeat with the Benjamini-Hochberg and the Benjamini-Yekutielli corrections. Are the most influent genes the same depending on the corrections ? Are they order the same way ? Conclude on the most influent genes.

5. Display the plot of the ecdf of the four vectors of p-values. Comment the difference between the p-values.

6. Compare the level of expression of genes according to the stage of lymphomia. Which test has to be used ?

7. Sort the vector of p-values (before adjustment). How many p-values are below 5% ?

8. Apply the BY correction of the p-values. How many p-values are below 5% ? Comment and conclude.

2 Session 2

Objectives: on your project dataset, link between phenotypes variables and genomics variables.

Steps of the session:

1. Team work: Study the link between each phenotype variable and the genomics variables on your project dataset. Use adjusted p-values to select the most important and influent genes. Test the link between the level of expression and binary variables through t.test, var.test, ks.test, wilcox.test; between level of expression and qualitative variables using oneway.test; between level of expression and quantitative variables using cor.test.

2. Team work: Write the report corresponding to your results

3. Individual work at home: Read the chapter *Regression models*, section *Linear regression*.