Multiple hypothesis testing: the p-value problem

Tutor: Guillaume CHENNETIER, guillaume.chennetier@polytechnique.edu

Keywords. Multiple hypothesis testing, p-value, mathematical statistics, Bonferroni, Benjamini-Hochberg, FWER, FDR.

Reminder on statistical tests. The theory of statistical tests, founded at the beginning of the 20th century, provides many tools to confront a scientific hypothesis with the data. A test is a binary function of the data that allows one to decide between a null hypothesis H_0 (which in experimental sciences assumes that the effect one wants to highlight does not exist, and that the variations in the data are only due to chance) and an alternative hypothesis H_1 (presence of the effect) that one would ideally like to prove. A test can commit two types of error: type I error by rejecting H_0 when it was true (false positive) and type II error by keeping H_0 when it was false (false negative). We note α the probability of type I error also called significance level and β the probability of type II error. $1-\alpha$ corresponds to the confidence level of the test and $1-\beta$ to its power. It is not mathematically possible to control both α and β as well as we would like. In practice, priority is given to the significance of the test for which a maximum threshold $\overline{\alpha}$ not to be exceeded is fixed and we look for the most powerful test among those respecting this threshold.

The p-value problem. The p-value associated with an observed sample $\boldsymbol{X}$ is the probability p of observing a sample as extreme as $\boldsymbol{X}$ in the case where H_0 is true. If $p < \overline{\alpha}$, then the result is statistically significant and H_0 can be rejected with a confidence level higher than $1-\overline{\alpha}$. The usual threshold is $\overline{\alpha} = 0.05$. The p-value (and its comparison with the 5% threshold) has become in many cases the necessary and sufficient criterion to publish a study in experimental science. In addition to the fact that the interpretation of the p-value is delicate and that it is often made to say what it does not mean, using this criterion induces a survivor bias. Suppose that a large number of hypothesis tests are performed for which H_0 is true (or a large number of times the same test). By construction, there is an absolutely non-negligible probability of obtaining erroneously a p-value below the threshold $\overline{\alpha}$ for at least one of these tests. This undermines the trust that can be placed in a very large number of publications for several reasons:

- No detection of "soft frauds" that repeat the experiments until they fall below the threshold at least once per chance.
- Other studies on the subject that could not reject H_0 are not taken into account because they have not been published.
- In big data situation, it is tempting to test a large number of hypotheses at the same time. Applying this criterion independently to each test is likely to lead to several false discoveries.

The purpose of this work is to study the functioning of some statistical procedures allowing in the case of multiple hypothesis tests to control:

1. the family-wise error rate (FWER) which is the probability of obtaining at least one false positive among all tests,
2. the false discovery rate (FDR) which is the ratio between the number of false positives and the number of true positives among all the tests.

Key points:

- Master the basics of hypothesis testing in classical situations and be able to give a clear interpretation of associated results.
- Formalize the p-value problem in the case of multiple hypotheses and explain the interest of the Bonferroni and Benjamini-Hochberg procedures.
- Implement a test case to highlight the problem and compare different control procedures (using generated data).

Bonus objectives:

- Working with a real data set.
- Investigate other more advanced control procedures.
- Deepen the results for one or more specific test types.

References

[1] A. Durmus, S. Gadat, E. Moulines, Asymptotic statistics. *Lectures notes MAA 304.*

[2] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57.1 (1995): p. 289-300.*

[3] E. L. Lehmann, J. P. Romano, Generalizations of the Familywise Error Rate. *Selected Works of EL Lehmann. Springer, Boston, MA, (2012): p. 719-735.*