

Lecture 3: Multiple testing

May 7, 2019

Lecturer: Aldo Solari

In hypothesis testing the probability of making a type I error is bounded by α , conventionally set at 0.05. Problems arise, however, when researchers do not perform a single hypothesis test but many of them.

There are many ways of dealing with type I errors. We will focus on three types of multiple testing methods:

- those that control the *familywise error* (FWER)
- those that control the *false discovery rate* (FDR)
- those that estimate the *false discovery proportion* (FDP) or make confidence intervals for it.

For a broader discussion, see [1].

1 Error rates

Suppose we have a collection $\mathcal{H} = (H_1, \dots, H_m)$ of m hypotheses of interest:

- an unknown number m_0 of these hypotheses is true, whereas the other $m_1 = m - m_0$ is false
- we denote the proportion of true hypotheses $\pi_0 = m_0/m$

We call the collection of true hypotheses $\mathcal{T} \subseteq \mathcal{H}$ and the remaining collection of false hypotheses $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$. The goal of a multiple testing procedure is to choose a collection $\mathcal{R} \subseteq \mathcal{H}$ of hypotheses to reject.

If we have p -values p_1, \dots, p_m for each of the hypotheses H_1, \dots, H_m , an obvious choice is the collection

$$\mathcal{R} = \{H_i : p_i \leq T\}$$

rejecting all hypotheses with a p -value below a threshold T . In this situation, the multiple testing problem reduces to the choice of T . In some situations, however, rejected sets of other forms may be of interest.

Ideally, the set of rejected hypotheses \mathcal{R} should coincide with the set \mathcal{F} of false hypotheses as much as possible. Two types of error can be made:

- false positives, or type I errors, are the rejected hypotheses that are not false, i.e. $\mathcal{R} \cap \mathcal{T}$
- false negatives or type II errors are the false hypotheses that we failed to reject, i.e. $\mathcal{F} \setminus \mathcal{R}$

Rejected hypotheses are sometimes called *discoveries*, hence the terms *true discovery* and *false discovery* are sometimes used for correct and incorrect rejections.

We can summarize the numbers of errors occurring in a hypothesis testing procedure in a contingency table:

	true	false	total
rejected	V	U	R
not rejected	$m_0 - V$	$m_1 - U$	$m - R$
total	m_0	m_1	m

We can observe m and $R = \#\mathcal{R}$, but all quantities in the first two columns of the table are unobservable.

Multiple testing methods try to reject as many hypotheses as possible while keeping some measure of type I errors in check. This measure is usually either the number V of type I errors or the false discovery proportion (FDP) Q , defined as

$$Q = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

the proportion of false rejections among all rejections, defined as 0 if no rejections are made.

Different types of multiple testing methods focus on different summaries of the distribution of V and Q .

The most popular ones are the *Family-wise Error Rate* (FWER), given by

$$\text{FWER} = P(V > 0) = P(Q > 0)$$

and the *False Discovery Rate* (FDR), given by

$$\text{FDR} = E(Q).$$

The FWER focuses on the probability that the rejected set contains any error, whereas FDR looks at the expected proportion of errors among the rejections. Either FWER or FDR is *controlled* at level α , which means that the set \mathcal{R} (i.e. the threshold T) is chosen in such a way that the corresponding aspect of the distribution of Q is guaranteed to be at most α .

1.1 FWER and FDR

The two error rates FDR and FWER are related. Because $0 \leq Q \leq 1$, we have $Q \leq \mathbb{1}\{Q > 0\}$ and

$$E(Q) \leq P(Q > 0)$$

which means that FWER-controlling methods are automatically also FDR-controlling methods.

Because FDR is smaller than FWER, it is easier to keep the FDR below a level α than to keep the FWER below the same level, and we can generally expect FDR-based method to have more power than FWER-based ones.

In practice, FDR-controlling methods are especially more powerful than FWER-controlling methods if there are many false hypotheses. Conversely, if all hypotheses are true, FDR and FWER are identical; because $R = V$ in this case, Q is a Bernoulli variable, and

$$E(Q) = P(Q > 0)$$

. Both FDR and FWER are proper generalizations of the concept of type I error to multiple hypotheses. If there is only one hypothesis ($m = 1$), the two error rates are identical and equal to the regular type I error.

2 Assumptions of multiple testing methods

All methods we will consider start from a collection of test statistics S_1, \dots, S_m , one for each hypothesis tested, with corresponding p -values

$$p_1, \dots, p_m$$

We call these p -values *raw* as they have not been corrected for multiple testing yet.

Assumptions on the p -values often involve only the p -values of true hypotheses. We denote these by

$$q_1, \dots, q_{m_0}$$

By the definition of a p -value, if their corresponding hypotheses are true, these p -values are either uniformly distributed between 0 and 1, or they can be stochastically greater than uniform if data are discrete: we have, for $i = 1, \dots, m_0$

$$P(q_i \leq u) \leq u$$

2.1 No assumptions

Methods that make no assumptions on the dependence structure of p -values are always based on some probability inequality. The Bonferroni inequality says

$$P\left(\bigcap_{i=1}^{m_0} \left\{q_i > \frac{\alpha}{m_0}\right\}\right) = P\left(q_{(1)} > \frac{\alpha}{m_0}\right) \geq 1 - \alpha$$

and the Hommel inequality says

$$P\left(\bigcap_{i=1}^{m_0} \left\{q_{(i)} > \frac{i\alpha}{m_0 \sum_{j=1}^{m_0} j^{-1}}\right\}\right) \geq 1 - \alpha$$

where $q_{(1)}, \dots, q_{(m_0)}$ are the m_0 ordered p -values of the true hypotheses.

2.2 Assumption of positive dependence

The famous Benjamini & Hochberg procedure requires the assumption of *positive regression dependence on a subset* (PDS). To formally define this assumption, let $T = \{i : H_i \in \mathcal{T}\}$ denote the set of nulls and call $D \subset [0, 1]^m$ an increasing set if $x \in D$ and $x \leq y \leq 1$ (in the coordinate-wise sense) together imply $y \in D$.

Definition 2.1. A set of p -values (p_1, \dots, p_m) is said to satisfy the PDS property if for any increasing set $D \subset [0, 1]^m$ and each null index $i \in T$, the probability

$$P((p_1, \dots, p_m) \in D | p_i \leq t)$$

is non-decreasing in $t \in (0, 1]$.

Examples of cases under which the PDS condition holds include one-sided test statistics that are jointly normally distributed, if all correlations between test statistics are positive.

The PDS assumption is a sufficient condition for a probability inequality due to Simes

$$P\left(\bigcap_{i=1}^{m_0} \left\{q_{(i)} > \frac{i\alpha}{m_0}\right\}\right) \geq 1 - \alpha$$

Equality holds if q_1, \dots, q_{m_0} are i.i.d. $U(0, 1)$.

Simes' inequality strictly improves upon both Hommel's and Bonferroni's inequalities. The critical values of Simes' inequality are larger than those of Hommel's inequality by a factor $\sum_{j=1}^{m_0} 1/j$, which converges to $\log(m_0) + \gamma$ as $m_0 \rightarrow \infty$, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

2.3 Assumption of independence

Much work has been performed under the assumption of independent p -values, but this assumption is often not realistic.

References

- [1] J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.