Multiple testing

Statistical Learning

Introduction

For many statisticians, *microarrays* provided an introduction to high-dimensional data analysis.

Microarrays were revolutionary biomedical devices that enabled the assessment of individual activity for thousands of genes at once and, in doing so, raised the need to carry out thousands of simultaneous hypothesis tests, done with the prospect of finding only a few interesting genes among a haystack of null cases.

Why multiple testing?

Hypothesis tests are widely used as the gatekeepers of the scientific literature. In many fields, scientific claims are not believed unless corroborated by rejection of some hypothesis.

Hypothesis tests are not free of error, however, and for every hypothesis test there is a risk of falsely rejecting a hypothesis that is true, i.e. a *type I error*, and of failing to reject a hypothesis that is false, i.e. a *type II error*.

In hypothesis testing, type I errors are traditionally considered *more problematic* than type II errors. If a rejected hypothesis allows publication of a scientific finding, a type I error brings a *false discovery*, and the risk of publication of a potentially misleading scientific result. Type II errors, on the other hand, mean missing out on a scientific result. Although unfortunate for the individual researcher, the latter is, in comparison, less harmful to scientific research as a whole.

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

Since each test again has a probability of producing a type I error, performing a large number of hypothesis tests virtually guarantees the presence of type I errors among the findings.

Actually, the expected number of type I errors is

$\mathbb{E}($ number of type I errors $)=$ number of true null hypotheses $\times \alpha$

For example, with $\alpha = 5\%$ and 100 true hypotheses, we obtain on average 5 type I errors.

As the type I errors among the findings are likely to be the most surprising and novel ones, they have a high risk of finding their way into publications.

The key goal of multiple testing methods is to control, or at least to quantify, the flood of type I errors that arise when many hypothesis tests are performed simultaneously.

Different methods do this in different ways, as there are different ways to generalize the concept of type I error to the situation with more than one hypotheses.

It is helpful to see the problem of multiple testing as a problem caused by selection. Although the probability of a type I error in each individual hypothesis remains equal to α regardless of the number of hypotheses that have been tested, the researcher will tend to emphasize only the rejected hypotheses. These rejected hypotheses are a selected subset of the original collection of hypotheses, and type I errors tend to be overrepresented in this selection. The probability of a selected hypothesis to be a type I error is therefore much larger than α. Multiple testing methods correct for this selection process and bring type I error probabilities back to α even for selected hypotheses.

Error rates

The multiple testing problems have a very simple structure. We have a collection $\mathcal{H} = (H_1, \ldots, H_m)$ of null hypotheses, which we would like to reject.

An unknown number m_0 of these hypotheses is true, whereas the other $m_1 = m - m_0$ is false. We call the collection of true hypotheses $\mathcal{T} \subseteq \mathcal{H}$ and the remaining collection of false hypotheses $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$. We denote the proportion of true hypotheses $\pi_0 = m_0/m$.

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 $p_1, \ldots, p_m$ for each of the hypotheses $H_1, \ldots, H_m$, and 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 $\mathcal{F}$ 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 such as Table 1.

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

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

Table 1: Contingency table for multiple hypothesis testing: rejection versus truth or falsehood of hypotheses.

Type I and type II errors are in direct competition with each other, and a trade-off between the two must be made. If we reject more hypotheses, we typically have more type I errors but fewer type II errors.

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* Q, defined as

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

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

Most multiple testing methods choose the threshold T as a function of the data so that the set $\mathcal{R}$ of rejected hypotheses is random, and so both V and Q are random variables.

Different error rates focus on different summaries of the distribution of V and Q.

The most popular methods control either the *familywise error* (FWER), given by

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

or the *false discovery rate* (FDR), given by

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

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 correponding aspect of the distribution of Q is guaranteed to be at most α.

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

$$\mathbb{E}(Q) \leq \Pr(Q > 0)$$

which implies that every FWER-controlling method is automatically also an FDR-controlling method.

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 methods to have more power than FWER-based ones.

Conversely, if all hypotheses are true, FDR and FWER are identical; because $R = V$ in this case, Q is a Bernoulli random variable, and $\mathbb{E}(Q) = \Pr(Q > 0)$.

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.

FDR and FWER generalize type I error in a different way, however. We can say that if FWER of a set of hypotheses $\mathcal{R}$ is below α, then *for every* hypothesis in $H \in \mathcal{R}$ the probability that H is a type I error is below α. FDR control, on the other hand, implies type I error control *on average* over all hypotheses $H \in \mathcal{R}$. This difference has important practical consequences.

In particular, FWER has the *subsetting property* that if a set $\mathcal{R}$ of hypotheses is rejected by an FWER-controlling procedure, then FWER control is also guaranteed for any subset $\mathcal{S} \subset \mathcal{R}$. The corresponding property does not hold for FDR control. In fact, it was shown that a procedure that guarantees FDR control not only for the rejected set itself, but also for all subsets, must be an FWER-controlling procedure. While FWER control is a statement that immediately translates to type I error control of individual hypotheses, FDR control is only a statement on the full set $\mathcal{R}$, and one which does not translate to subsets of $\mathcal{R}$ or individual hypotheses in $\mathcal{R}$.

In methods that control an error rate, such as FDR or FWER, the user chooses an error rate to be controlled, and the multiple testing method finds a rejected set $\mathcal{R}$ for the user according to the criterion. This contrasts with FDP confidence methods that let the user choose the set $\mathcal{R}$ freely, and which subsequently try to make a confidence interval for the FDP of that set. In this type of method, the set $\mathcal{R}$ is not a random quantity, and the number of errors V or FDP Q in this set is fixed but unknown quantities, which can in principle be estimated. In practice, of course, the set $\mathcal{R}$ to be rejected is not determined before data collection, but will often be chosen in some partially data-dependent way. Any estimates and confidence statements for Q or V need to be corrected for bias resulting from such a data-dependent choice.

Assumptions of multiple testing methods

In statistics, stronger assumptions generally allow more powerful statements.

In multiple testing, the most crucial assumptions to be made concern the dependence structure of the p-values.

Much work has been performed under the assumption of independent p-values, but this work is of little practical value in realistic application in which measurements typically exhibit strong but a priori unknown correlations.

Methods with more realistic assumptions come in three major flavours

1. The first kind makes no assumptions at all. They protect against a 'worst case' dependence structure and are conservative for all other dependence structures

2. The second kind gains power by making a certain assumption on the dependence structure of the p-values, known as positive dependence through stochastic ordering (PDS) assumption

3. The third kind uses permutations to adapt to the dependence structure of the p-values.

All three types of assumptions have been used in methods for FWER control, FDR control and FDP estimation.

Table II gives an overview of the methods that use each of the different assumptions to achieve control of each of the error rates. Before we move on to specific methods, we first discuss the assumptions in some detail.

	Error criterion		
Assumptions	FWER control	FDR control	FDP confidence
No assumptions	Holm	Benjamini & Yekutieli	Goeman & Solari
	`p.adjust`	`p.adjust`	`pickSimes` (cherry)
PDS assumption	Hommel	Benjamini & Hochberg	Goeman & Solari
	`p.adjust`	`p.adjust`	`pickSimes` (cherry)
Permutations	Westfall & Young		Meinshausen
	`mt.maxT` (multtest)		`howmay_dependent` (howmany)

Table 2: Overview of error criteria and assumptions considered in thus tutorial, with the main representative method for each combination, if present. The table mentions the name of the method and an R function (with required package) that implements it.

All methods we consider here start from a collection of test statistics $S_1, \ldots, S_m$, one for each hypothesis tested, with corresponding p-values $p_1, \ldots, p_m$. 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 $q_1, \ldots, q_{m_0}$.

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 $i = 1, \ldots, m_0$

$$\Pr(q_i \leq u) \leq u \tag{1}$$

In practice, raw p-values are often only approximate, as they are derived through asymptotic arguments or other approximations. It should always be kept in mind that such asymptotic p-values can be quite inaccurate, especially for small sample sizes, and that their relative accuracy decreases when p-values become smaller.

Methods that make no assumptions on the dependence structure of p-values are always based on some probability inequality. Two such inequalities are relevant for methods described here.

The first is the Bonferroni inequality, which says that simultaneously, with probability at least $1 - \alpha$

$$\bigcap_{i=1}^{m_0} \left\{ q_i > \frac{\alpha}{m_0} \right\}$$

The second inequality is due to Hommel, which states that with probability at least $1 - \alpha$

$$\bigcap_{i=1}^{m_0} \left\{ q_{(i)} > \frac{i\alpha}{m_0 \sum_{i=1}^{m_0} (1/j)} \right\} \tag{2}$$

where $q_{(1)} \leq \ldots \leq q_{(m_0)}$ are the m_0 ordered p-values of the true hypotheses. Hommel's and Bonferroni's inequalities are valid whatever the dependence of p-values, as long as (1) holds.

Probability inequalities have a *worst case* distribution for which the inequality is an equality, but they are strict inequalities for most distributions. Multiple testing methods based on such inequalities are therefore conservative for all p-value distributions except for this 'worst case'. Such 'worst case' distributions are often quite unrealistic, and this is especially true for Hommel's inequality, which can be quite conservative for more realistic distributions.

To avoid having to cater for exotic worst case distributions, assumptions can be made to exclude them. In particular, an assumption of *positive dependence through stochastic ordering* (PDS) can be made, which excludes the 'worst case' distributions of the inequalities of Hommel and Bonferroni. Technically, the PDS condition comes in two forms, The weaker PDS condition says that

$$\mathbb{E}[f(q_1, \ldots, q_{m_0})|q_i = u] \tag{3}$$

is nondecreasing in u for every i and for every coordinate-wise nondecreasing function f. The stronger PDS condition is identical except that (3) is replaced by

$$\mathbb{E}[f(p_1, \ldots, p_m)|q_i = u] \tag{4}$$

so that it involves nondecreasing functions of p-values not only of true hypotheses but also of false ones.

The second PDS condition is slightly more restrictive, but for all practical purposes, we can view the two PDS conditions as identical. In Table II, the FWER-controlling methods and FDP confidence methods require the weaker PDS condition, whereas some FDR-controlling methods, most notably the famous Benjamini & Hochberg procedure, require the stronger one.

The weaker PDS condition is a sufficient condition for a probability inequality due to Simes. This Simes inequality is related to Hommel's inequality, but more powerful. It says that with probability at least $1 - \alpha$

$$\bigcap_{i=1}^{m_0} \left\{ q_{(i)} > \frac{i\alpha}{m_0} \right\} \tag{5}$$

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 $\sum_{i=1}^{m_0}(1/j)$, which, for large m_0, is approximately equal to $\log(m_0) + \gamma$, where $\gamma \approx 0.577$ is the Euler–Mascheroni constant.

As a probability inequality, Simes' inequality is also a strict inequality for some distributions, but the 'worst case', for which the Simes inequality is not conservative, is the case of independent uniform p-values, which is relatively unexotic.

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, or two-sided joint normal test statistics under a different condition that allows some negative correlations; Simes' inequality has also been shown to hold for some important cases for which PDS fails, such as the for certain basic multivariate t -distributions with common denominators.

Even though these conditions are not guaranteed to hold for all distributions relevant for genomics, methods based on the PDS condition turn out to be quite robust in practice. The current consensus among practitioners seems to be that, in genomics data, it is safe to use methods based on the PDS condition.

The third way to deal with the unknown dependence structure of p-values is permutation testing. Permutation tests have two great advantages, both of which translate to permutation-based multiple testing.

1. First, they give exact error control without having to rely on asymptotics for the assumption (1), allowing reliable testing even when asymptotic p-values are unreliable.

2. Second, permutation tests do not use any probability inequality, but generally attain the exact level α regardless of the distribution of the underlying data.

Permutation-based multiple testing is said to *adapt* to the dependence structure of the raw p-values. It is not conservative for any dependence structure of the p-values, and it can be especially powerful in case of strong dependence between p-values.

References

- Goeman and Solari (2014) Multiple Hypothesis Testing in Genomics. Statistics in Medicine 2014 33:1946-78