Methods for familywise error control

Statistical Learning

The most classical way of controlling for multiple testing is by FWER control, and consequently, this is the type of multiple testing for which most methods are available. We first discuss the methods of Bonferroni and Holm, which are valid under any dependence of the p-values, then Hochberg's and Hommel's methods, which are valid if the PDS assumption holds, before looking at the permutation-based method of Westfall & Young. It has been argued that FWER control is too conservative and therefore not suitable for genomics research, but this is too simplistic.

Bonferroni

The method of Bonferroni controls FWER at level α by rejecting hypotheses only if they have raw p-value smaller than α/m.

This single-step adjustment of the significance threshold is the simplest, oldest and most well-known multiple testing method, and is attractive because of its simplicity. However, it is also known to be conservative, especially if many hypotheses are false, or if strong positive correlations between p-values occur.

Despite being so well known, or perhaps because of this, there is a lot of misunderstanding about the method of Bonferroni in the literature. We will start this section with a discussion of these misunderstandings.

One of the most widespread of these misunderstandings is that the method would be based on an assumption of independence between p-values. This misunderstanding comes from a frequently used, but deficient motivation for Bonferroni, saying that the probability of making a false rejection if all m_0 p-values of true hypotheses are independent, and we perform each test at level α/m is $1-(1-\alpha/m)^{m_0}$. Expanding this expression, the first term, which is dominant for small α or large m, is $m_0\alpha/m \leq \alpha$. By this reasoning, Bonferroni seems like a method that only provides approximate FWER control and that requires an assumption of independence for its validity. In fact, the method of Bonferroni provides exact FWER control under any dependence structure of the p-values.

To properly motivate the Bonferroni method, we should look at it as a corollary to Boole's inequality, which says that for any collection of events $E_1, \ldots, E_k$, we have

$$\mathrm{P}\big(\bigcup_{i=1}^{k} E_i\big) \leq \sum_{i=1}^{k} \mathrm{P}(E_i).$$

It follows from Boole's inequality that, if $q_1, \ldots, q_{m_0}$ are the p-values of the true null hypotheses, that the probability that there is some i for which $q_i \leq \alpha/m$ is given by

$$\Pr\big(\min_i q_i \leq \alpha/m\big) \leq \mathrm{P}\big(\bigcup_{i=1}^{m_0}\{q_i \leq \alpha/m\}\big) \leq \sum_{i=1}^{m_0} \mathrm{P}(q_i \leq \alpha/m) = m_0\frac{\alpha}{m} \leq \alpha. \tag{1}$$

Because the method of Bonferroni only commits a type I error if $q_i \leq \alpha/m$ for some i, this proves FWER control at level α for the Bonferroni method.

A few things can be learnt from this derivation. First, the FWER control of Bonferroni is not approximate, and it is valid for all dependence structures of the underlying p-values. Second, the three inequalities indicate in which cases the Bonferroni method can be conservative.

- The right-hand one shows that Bonferroni does not control the FWER at level α but actually at the stricter level $\pi_0\alpha$, where $\pi_0 = m_0/m$. If there are many false hypotheses, Bonferroni will be conservative.

- The middle inequality says that Bonferroni is conservative if the raw p-values are
- The left-hand inequality is due to Boole's law. This inequality is a strict one in all situations except the one in which all events $\{q_i \leq \alpha/m\}$ are disjoint. From this, we conclude that Bonferroni is conservative in all situations except in the situation that the rejection events of the true hypotheses are perfectly negatively associated, that is, if the p-values have strong negative correlations.

The conservativeness of Bonferroni in situations in which Boole's inequality is strict deserves more detailed attention.

- With independent p-values, this conservativeness is present but very minor. To see this, we can compare the Bonferroni critical value α/m with the corresponding critical values $1-(1-\alpha)^{1/m}$ that can be calculated for independent p-values. For $m = 5$ and $\alpha = 0.05$ we find a critical value of 0.01021 for Sidak against 0.01 for Bonferroni. As m increases, the ratio between the two increases to a limit of $-\log(1-\alpha)/\alpha$, which evaluates to only 1.026 for $\alpha = 0.05$.
- Much more serious conservativeness can occur if p-values are positively correlated. For example, in the extreme case that all p-values are perfectly positively correlated, FWER control could already have been achieved with the unadjusted level α, rather than α/m. Less extreme positive associations between p-values would also allow less stringent critical values, and Bonferroni can be quite conservative in such situations.

A second, less frequent misunderstanding about Bonferroni is that it would only protect in the situation of the global null hypotheses, i.e. the situation that $m_0 = m$. This type of control is known as *weak* FWER control. On the contrary, Bonferroni controls the FWER for any combination of true and false hypotheses. This is known a *strong* FWER control. In practice, only strong FWER controlling methods are of interest, and methods with only weak control should, in general, be avoided. To see this, consider the method that, if there is at least one p-value below α/m, rejects all m hypotheses, regardless of the p-values they have. This nonsensical method has weak, but not strong FWER control. Related to weak control methods but less overconfident are global testing methods that test the global null hypothesis that $m_0 = m$. If such a test is significant, one can confidently make the limited statement that at least one false hypotheses is present, but not point to which one. In contrast, methods with strong FWER control also allow pinpointing of the precise hypotheses that are false.

The *adjusted* p-value for the Bonferroni procedure is given by $\min(mp_i, 1)$, where p_i is the raw p-value.

Holm

Holm's method is a sequential variant of the Bonferroni method that always rejects at least as much as Bonferroni's method, and often a bit more, but still has valid FWER control under the same assumptions. From this perspective, there is no reason, aside from possibly simplicity, to even use Bonferroni's method in preference to Holm's.

Holm remedies part of the conservativeness in the Bonferroni method arising from the right-hand inequality of (1), which makes Bonferroni control FWER at level $\pi_0\alpha$. It does that by iterating the Bonferroni method in the following way. In the first step, all hypotheses with p-values at most α/h_0 are rejected, with $h_0 = m$ just like in the Bonferroni method. Suppose this leaves h_1 hypotheses unrejected. Then, in the next step, all hypotheses with p-values at most α/h_1 are rejected, which leaves h_2 hypotheses unrejected, which are subsequently tested at level α/h_2. This process is repeated until either all hypotheses are rejected, or until a step fails to result in any additional rejections. Holm gave a very short and elegant proof that this procedure controls the FWER in the strong sense at level α. This proof is based on Boole's inequality just like that of the Bonferroni method, and consequently makes no assumptions whatsoever on the dependency structure of the p-values.

It is immediately obvious that Holm's method rejects at least as much as Bonferroni's and possibly more. The gain in power is greatest in the situation that many of the tested hypotheses are false, and when power for rejecting these hypotheses is good. Rejection of some of these false hypotheses in the first few steps of the procedure may lead to an appreciable increase in the critical values for the remaining hypotheses. Still, unless the proportion of false hypotheses in a testing problem is very large, the actual gain is often quite small.

An alternative way of describing Holm's method is via the ordered p-values $p_{(1)}, \ldots, p_{(m)}$. Comparing each p-value $p_{(i)}$ to its corresponding critical value $\alpha/(m-i+1)$, Holm's method finds the smallest j such that $p_{(j)}$ exceeds $\alpha/(m-j+1)$, and subsequently rejects all $j-1$ hypotheses with a p-value at most $\alpha/(m-j)$. If no such j can be found, all hypotheses are rejected.

Hochberg and Hommel

Bonferroni's and Holm's methods make no assumptions on the dependency structure of the p-values, and protect against the 'worst case' according to Boole's inequality, which is that the rejection regions of the different tests are disjoint. If we are willing to make assumptions on the joint distribution of the p-values, it becomes possible exclude this worst case a priori, and as a result gain some power.

One such assumption could be that the Simes inequality holds for the subset of true hypotheses. This assumption makes the use of Hochberg's method possible, which is very similar to Holm's method but more powerful. Hochberg's method (not to be confused with Benjamini and Hochberg's method) compares each ordered p-value $p_{(i)}$ to a critical value $\alpha/(m-i+1)$, the same as Holm's. It then finds the largest j such that $p_{(j)}$ is smaller than $\alpha/(m-j+1)$, and subsequently rejects all j hypotheses with p-values at most $\alpha/(m-j+1)$.

Comparing to Holm's method, it is clear that Hochberg's method rejects at least as much as Holm's method, and possibly more. If the curves $p_{(1)}, \ldots, p_{(m)}$ and $\alpha/m, \alpha/(m-1), \ldots, \alpha$ never cross or only once, Holm's and Hochberg's methods reject the same number of hypotheses. If the same curves cross multiple times, Holm's method uses the first crossing point as the final critical value, while Hochberg's uses the last crossing point.

In the jargon of multiple testing methods, Holm's method is known as a *step-down* method and Hochberg's as its *step-up* equivalent. This is somewhat confusing as one would say that Holm's method steps up from the smallest p-value to find the first crossing point, while Hochberg's steps down from the largest one to find the last crossing point, but the terminology was originally formulated in terms of test statistics rather than p-values. An illustration of the difference between step-up and step-down methods is given in Figure 1.

Hochberg's method is a special case of the powerful closed testing procedure, and the proof of its FWER control is based on combining that procedure with the Simes inequality. Hommel showed that a more powerful, although more complicated and computationally more demanding procedure can be constructed from the same ingredients. Hommel's resulting procedure rejects at least as much as Hochberg's, but possibly more. On modern computers the additional computational burden of Hommel's procedure is not an impediment anymore.

A gain in power of Hochberg's and Hommel's methods over Holm's method can be expected in case that a large proportion of the hypotheses is false, but the corresponding tests have relatively low power, or if there are positive associations between the p-values of these false hypotheses.

Permutations and Westfall & Young

To be discussed later.

Use of familywise error rate control

Since the advent of genomics data, FWER as a criterion has been heavily criticised for being too conservative for genomics research. For many data sets, application of methods of FWER control result in very few rejected hypotheses or none at all, even when other analyses suggested the presence of some differential

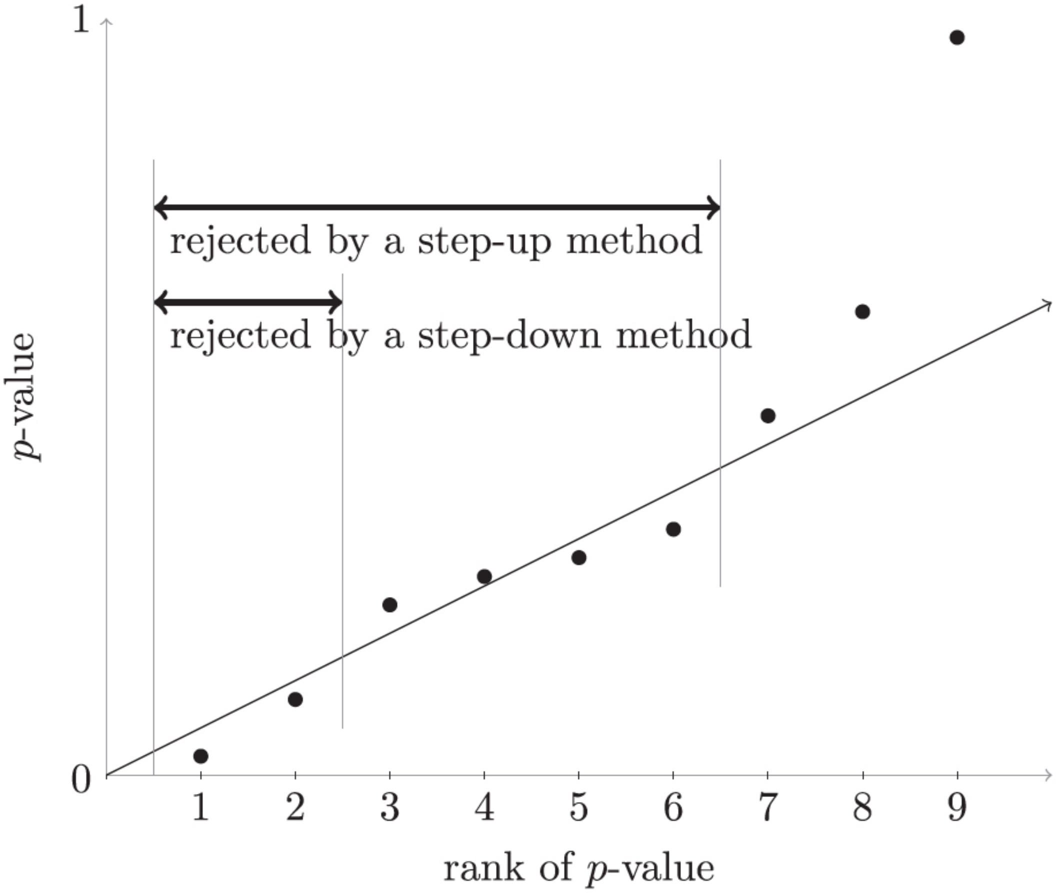

Figure 1: Comparison of rejections by step-up and step-down methods with the same critical values. The dots are observed ranked p-values. The line represents the critical values. Step-down methods reject all hypotheses up to, but not including, the first p-value that is larger than its critical value. Step-up methods reject all hypotheses up to and including the last p-value that is smaller than its critical value.

expression. This criticism of FWER stands at the basis of the development and popularity of the various FDR and FDP-based methods.

Indeed, FWER is a very rigorous and strict, and therefore conservative, criterion. It avoids type I errors at all cost, and as a consequence it introduces a large number of type II errors. The payback for this is that all hypotheses rejected by FWER controlling methods are individually reliable. FWER control implies $1 - \alpha$ confidence that each individual rejected hypothesis is correctly rejected. For many genomics experiments such great confidence is much more than necessary. If the experiment will be followed up by replication or proper validation before publication of the results, confidence that at least a substantial proportion of the findings is real is often sufficient to continue, and FWER-type confidence is not needed. What's more, at this stage the cost of type II errors is non-negligible, as missing out on an important finding can result in an expensive experiment wasted. More lenient criteria than FWER are in order for such experiments.

All this does not mean, however, that FWER has no place in genomics research. For the analysis of any experiments that are end-stage, not followed up by independent validation, such as the validation experiments themselves, merely saying that the proportion of true discoveries in the list is large is hardly sufficient. Such results have to be individually reliable, since they are likely to be taken out of the context of the list they were presented in. This individual reliability of results is precisely what FWER control guarantees. Since the power of validation experiments if typically large, and since the number of hypothesis tests done at this stage is limited, any conservativeness of FWER should negligible at this stage, especially if powerful methods such as Westfall & Young's are used.

References

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