Multiple Testing

1 Multiple Testing

In practice, particularly in modern scientific applications, one is often faced with the problem of testing not just one null hypothesis, but many.

Example 1: In brain imaging one might collect brain image data on two groups of individuals (patients and controls) and wish to identify in which areas of the brain activity levels differ between the groups. This is sometimes set up as a hypothesis testing problem, with the null hypothesis for each pixel in an image being "the distribution of the intensity of this pixel is the same in patients as in controls". The aim is then to identify pixels for which the null is false.

Example 2: In genetics, one might want to see which genes differ in their "expression" (activity level) in tumor cells vs normal cells. This is often set up as a hypothesis testing problem, with the null hypothesis for each gene being that "the distribution of the expression of this gene is the same in tumor and normal cells".

In each of the above examples we have thousands of different null hypotheses. In general, we consider m distinct null hypotheses, and let $H_i \in \{0, 1\}$ denote the (unknown) status of the ith hypothesis, with $H_i = 0$ denoting that the ith null is true. Typically one proceeds by computing test statistics $T = (T_1, \ldots, T_m)$, where T_i conveys the evidence in the data regarding the ith null, $H_i = 0$. If we assume that small values of T_i indicate stronger evidence against $H_i = 0$ then a typical procedure would reject those tests i for which $T_i < \gamma$ for some suitably-chosen threshold γ. In a nutshell, the question we want to address is how to choose γ. In what follows the expectations and probabilities are with respect to the random variables T: that is, they are expectations over hypothetical repetitions of the experiment that lead to T. Note that, because the distribution of T depends on H, the probabilities and expectations also depend on H (which, recall, is unknown). In most cases this dependance on H is not made explicit in our notation, with it being made explicit only when it is necessary.

2 Family-wise Error rate and False Discovery Rate

The first, and perhaps most crucial, decision to be made in a multiple testing situation is what type of error rate you want to control. Effectively this is connected with selecting your "loss function". The two most common error rates considered in practice are "family-wise error rate" (FWER), and the "false discovery rate" (FDR), or variations on this.

To define these quantities, consider the possible outcomes of testing multiple hypotheses (Table 1).

The family-wise error rate (FWER) is defined to be the probability that even a single null hypothesis is wrongly rejected, $\Pr(V \geq 1)$. Recall that the randomness here is over the random distribution of the test statistics T, over repeated experiments; that is, this is a frequentist measure. Also note that this probability depends on the (unknown) value of H, so the FWER depends on H. The value of the FWER when all the nulls are true ("$H \equiv 0$") indicates the probability of making at least one wrong rejection when all the nulls are true, and in some cases this may be a particular focus of attention - see below.

The FDR is the expected value of V/R, defined to be 0 if $R = 0$. Thus

$$\text{FDR} = E(V/R|R > 0) \Pr(R > 0). \tag{1}$$

As pointed out by Storey (2001), one problem with FDR is that control of FDR can be achieved by making $\Pr(R > 0)$ small, without necessarily making $E(V/R|R > 0)$ small. And in practice we really want $E(V/R|R > 0)$ to be small (and $\Pr(R > 0)$ large if possible!). So he introduced the "positive FDR" (pFDR), defined to be

$$\text{pFDR} = E(V/R|R > 0). \tag{2}$$

Although in these notes we will distinguish between the FDR and pFDR, in practice we (and others) are often sloppy about this distinction, and simply use FDR even when we mean pFDR. Indeed, many (most?) users of FDR-based methods will probably not be aware of the distinction.

In most multiple testing problems I would argue that pFDR is a more appropriate choice of error measure than FWER for two reasons. First, the FWER seems much harder to calibrate. Or, perhaps more accurately, it is easier to forget that one needs to worry about this. For example, how does one decide what level to control the FWER? This is the same problem as calibrating p values that we covered previously. In practice everyone chooses 0.05 (or 0.01? or some other arbitrary value) without much thought or comment. Second, controlling FWER to be small is often unnecessarily stringent: in practice one would probably be OK with making a few errors ($V > 1$) provided $R = S + V$ is large compared with V. Note also that intuitively S depends on the power of the test, and V depends on the size, so controlling pFDR can be thought of, informally, as balancing power vs size. The fact that this balancing act is explicitly part of the calculation seems, in my opinion, to substantially reduce the potential problems of calibration.

Hypothesis	Accept	Reject	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

Table 1: Outcomes when testing m hypotheses

3 Controlling the FWER

There are two senses of controlling the FWER: strong control and weak control. Weak control means controlling the FWER only under the global null where all the nulls are true, $H \equiv 0$. Strong control means controlling the FWER under all possible values of H.

For example, consider the rule where we reject H_i if $T_i \leq \gamma$. We say this rule provides weak control of FWER at level α if

$$\Pr\left(\bigcup_{i:H_i=0} \{T_i \leq \gamma\} \middle| H \equiv 0\right) \leq \alpha. \tag{3}$$

For strong control at level α the same has to hold for all H. Note here that "controlled" means that the error does not exceed α: it may be considerably less than α. Ideally we might like equality to hold here. If strict inequality holds then the procedure is sometimes said to be "conservative".

The simplest and most widely-used approach to control FWER is the so-called "Bonferroni" method. This is based on the Bonferroni inequality, a truncation of Boole's formula: for any sequence of events $A_1, A_2, \ldots$

$$\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i). \tag{4}$$

Applying this to (3) gives

$$\Pr\left(\bigcup_{i:H_i=0} \{T_i \leq \gamma\} \middle| H \equiv 0\right) \leq \sum_i \Pr(T_i \leq \gamma | H \equiv 0) \tag{5}$$

which is less than α if each of the terms in the sum is less than α/m. Thus, the overall FWER can be controlled at level α by controlling each of the m individual tests at level α/m. If the test statistics T are all p values, then the Bonferroni procedure is to reject H_i if $T_i < \alpha/m$. (Note that if the test statistics are independent given H, with $P(T|H) = \prod_i \Pr(T_i|H_i)$ then this also provides strong control of the FWER.)

Note that the Bonferroni procedure is usually "conservative" in that it actually controls the FWER to be substantially less than α. If the tests of the m hypotheses are independent (which is sometimes approximately the case) and m is large (as is typical) then this conservativeness is not appreciable.

There some potential for confusion in the statement that "the Bonferroni procedure is conservative", because use of the Bonferroni procedure can also be thought of as conservative in another sense: it controls the very stringent criteria FWER, rather than the less stringent (and usually more appropriate) pFDR. It is important to distinguish in your mind these two different types of conservativeness. Further, I suggest you should really reserve the statement that the "Bonferroni is conservative" for the fact that it usually controls FWER to be strictly less than α, rather than for the fact that controlling the FWER is stringent. For example, I suggest it is unhelpful to say things like "The Bonferroni procedure is conservative, so we chose instead to control the pFDR".

4 Controlling the FDR and pFDR

Benjamini and Hochberg (1995) consider controlling the FDR, whereas Storey (2002, 2003) considers the pFDR, and the reader is referred to these papers for more details. What follows is a summary of some of the main ideas. For simplicity we assume the test statistics T are p values (so have a uniform distribution under the null).

4.0.1 Controlling FDR: The BH Procedure

To control the FDR one would specify a rate α and the state a rule that ensures that the FDR does not exceed α. The following Benjamini–Hochberg (BH) procedure provides this type of control of FDR. [1]

Let $P_{(1)} \leq P_{(2)} \cdots \leq P_{(m)}$ be the ordered p values, and let $H_{(i)}$ denote the null hypothesis corresponding to $P_{(i)}$. The BH procedure is to reject $H_{(1)}, \ldots, H_{(k)}$, where k is the largest i for which $P_{(i)} \leq i\alpha/m$.

Note a connection here with the Bonferroni procedure: if the smallest P value, $P_{(1)}$, is less than α/m then $H_{(1)}$ will be rejected.

[1] Note that, as pointed out by Benjamini and Hochberg (1995), controlling the FDR automatically provides weak control of the FWER, because when $H \equiv 0$ the FDR is equal to the FWER. This may appear to contradict the idea that, in general, controlling FDR is "less stringent" than controlling FWER. The resolution of this is that in practice we are not usually in the situation $H \equiv 0$; in rare cases where we do consider it possible that $H \equiv 0$ then arguably controlling FWER, or at least testing $H \equiv 0$, may be the first question of interest.

4.0.2 Storey's procedure

Analogous to the above, to control the pFDR one hope to specify a rate α and provide a rule that ensures that pFDR does not exceed α. However, as pointed out by Benjamini and Hochberg, pFDR cannot be controlled in the same way, because if $H \equiv 0$ then pFDR is identically 1. Indeed, this seems to be one reason why Benjamini and Hochberg chose FDR rather than pFDR in their initial paper.

To get around this, Storey uses a different approach: he fixes the rejection region, and attempts to estimate (conservatively) the pFDR for that region. The approach is easiest to describe in the case where the test statistic for each test is a p value (so has a uniform distribution under the null). Then, for any p value threshold γ,

$$\text{pFDR}(\gamma) = \pi_0 \gamma / \Pr(P \leq \gamma). \tag{6}$$

where π_0 is the proportion of true nulls ($H_i = 0$). Storey uses the test statistics near 1 to estimate π_0. For example, one could assume that all tests with p values between 0.95 and 1.0 correspond to true nulls, and then multiply the number of these tests by 20 to get an estimate of the total number of nulls, and then divide by m to get an estimate of π_0. To estimate $\Pr(P \leq \gamma)$ one can just count up how many tests have $P \leq \gamma$ and divide this by m. This is the essence of Storey's approach. (He has some twists to deal with the case where the denominator is 0, although I'm not sure they make that much sense to me: fortunately that case is not that interesting.)

5 Connection between pFDR and Bayesian quantities

Storey (2003) has the following theorem (his Theorem 1):

Suppose that H_i is Bernoulli($\pi_1 = 1 - \pi_0$), and that $T_i \sim (1 - H_i)F_0 + H_i F_1$, where F_0 and F_1 denote null and alternative distributions for T respectively. For rejection region Γ,

$$\text{pFDR}(\Gamma) = \Pr(H_i = 0 | T_i \in \Gamma) \tag{7}$$

Note that in the quantity on the left the randomness is over T (and H), whereas in the quantity on the right we condition on $T \in \Gamma$, and use Bayes Theorem to compute the posterior probability of $H_i = 0$ (considered to be random, jointly with T_i). This theorem can thus be viewed as connecting the frequentist measure pFDR with a Bayesian quantity.

Note, however, that the quantity on the right is not what a fully Bayesian analysis would compute, because it conditions only on $T \in \Gamma$, whereas a full analysis would condition on the *actual observed value* of T. In some cases this could lead to pFDR overstating the evidence against the null. For example, if you observe $T = \gamma$ then your posterior probability for $H = 0$ should be $\Pr(H = 0 | T = \gamma)$ which will naturally

be greater than $\Pr(H = 0|T \leq \gamma)$ because the latter includes cases where $T < \gamma$ which provide greater evidence against $H = 0$ than does $T = \gamma$.

Note also that this theorem implies that pFDR does not depend on m!!! That is, *if what you care about is the false discovery rate, then the number of tests you did is effectively irrelevant in selecting an appropriate threshold.* The intuition here is that, although doing more tests increases the number of false positive findings (at a given threshold), it will also increase the number of true positives in a similar way, and the two effects balance each other in the pFDR. Thus, if what you care about is the pFDR then there is no "problem" with multiple tests, and no need to "correct" for multiple tests. You will repeatedly come across people (statisticians especially!) who worry about the "problem of multiple testing", or "correcting for multiple tests". Ask them if they are worried about the FWER or pFDR, and if they say the pFDR point out that it does not depend on the number of tests.[2]

Note that this relies on the assumption that all the tests are "equivalent", and in particular are equally powerful, and equally likely to be null (a priori). For example, if you do additional less powerful tests, or additional tests of hypotheses that are clearly null, then this could hurt your overall false discovery rate (compared with not doing them) unless you take these differences into account. In some situations where there seem to be identifiably different types of test then it may be prudent to group tests into sets that seem somewhat equivalent, and analyze each set separately. For example, in a genetic context, one sometimes has an idea that certain genes are good candidates for influencing a particular outcome - it could be useful to consider those genes separately from all the others, particularly if you are right about them being "good candidates"! (It might help to think of it this way. Doing the tests separately allows that π_0 may differ between the groups; doing the tests together assumes that π_0 is the same for both groups. So if π_0 does indeed differ considerably between groups it will be better to do them together, provided you have enough data to accurately estimate the two values for π_0.)

6 q values

In practice one doesn't really want to fix the rejection region in advance. To get around this Storey effectively estimates pFDR for every possible rejection region. He gives each test an estimated q value, which is the estimated pFDR for all tests

[2] In contrast the FDR can, in principle, depend on the number of tests because the additional term $\Pr(R > 0)$ can depend on the number of tests. However, in many cases of practical interest $\Pr(R > 0)$ is near 1, and does not depend "much" on the number of tests, so in practice controlling FDR is also not really "correcting for multiple tests". However, this whole issue is complicated by the fact that the BH procedure *does* depend on the number of tests. The way I think of this is that the BH procedure is conservative.

that are as or more significant than the test in question. He ensures that these estimated q values are decreasing with decreasing p value by starting with large p values, and as he goes down the list he always requires that the q value is no bigger than the last one:

$$\hat{q}(p_{(i)}) = \min[\widehat{\text{pFDR}}(p_i), \hat{q}(p_{(i+1)})]. \tag{8}$$

Note that the q values, if misinterpreted, tend to overstate the evidence against the null for any given test. For example, a q value of 0.05 does not mean that this test has a probability 0.05 of being a false discovery. It means that, on average, that test *and all more significant tests* have an expected (postive) false discovery rate of 0.05.

7 local FDR; connection with Bayesian inference

The "local FDR" at a threshold γ is the probability that a given discovery is a false discovery:

$$\text{lfdr}(\gamma) = \Pr(H = 0 | T = \gamma). \tag{9}$$

It is really a Bayesian quantity. To compute this you have to have some way of estimating the density of T at γ, which is perhaps harder (but perhaps only slightly harder) than estimating the $\Pr(T \leq \gamma)$ which is required for the pFDR.

If you average the local fdrs over all tests that you reject then you get a Bayesian analogue of the False Discovery Rate for those rejected tests.

8 Multiple tests of the same null hypothesis

In these notes we have considered multiple tests of lots of different hypotheses. It is important not to confuse this with another situation that might be considered to involve "multiple testing": that is, multiple tests of the same hypothesis. For example, in genetics, one might test a genetic variant for association using two different tests, one that is aimed at detecting additive effects, and another that is aimed at detecting dominant or recessive effects. Clearly the minimum of these two p values (for example) is not a valid p value for because it is not uniform under the null. One would need to correct for this, for example, by modifying the null distribution. The Bonferroni procedure also works here, but is generally conservative because the tests are highly dependent; usually permutation is used in this kind of situation.

To illustrate in more detail how permutation testing works, suppose that we are attempting to test for association between a variable y and a variable x, each measured on n individuals. So the data are $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$. Let $T_1(\mathbf{x}, \mathbf{y})$ and $T_2(\mathbf{x}, \mathbf{y})$ denote two different test statistics testing for association.

For example, T_1 could be the p value from a linear regression of $\mathbf{y}$ on $\mathbf{x}$, and T_2 could be a p value from a Mann-Whitney test statistic for association between $\mathbf{x}$ and $\mathbf{y}$. Now define

$$T_{\min}(\mathbf{x}, \mathbf{y}) := \min(T_1(\mathbf{x}, \mathbf{y}), T_2(\mathbf{x}, \mathbf{y})). \tag{10}$$

The general motivation for considering more than one test here is that the two tests might be powerful against different alternative hypotheses, and so by combining them in this way $T_{\min}$ will be powerful against a wider range of alternatives than either T_1 or T_2 alone.

To define a permutation-based procedure for assessing the significance of $T_{\min}$ we need some notation. First, for any permutation π of the indices $(1, \ldots, n)$ let $\pi(\mathbf{x})$ denote the vector obtained by applying the permutation π to the vector $\mathbf{x}$:

$$\pi(\mathbf{x}) := (x_{\pi(1)}, \ldots, x_{\pi(n)}). \tag{11}$$

Now let $\pi_1, \ldots, \pi_N$ denote N different permutations of the indices $(1, \ldots, n)$ (with each π_j sampled uniformly from the set of all permutations). Finally, let $T^j_{\min} := T_{\min}(\pi_j(\mathbf{x}), \mathbf{y})$, be the result of computing $T_{\min}$ for the jth permutation. Then the permutation-based p value is

$$p := \frac{1 + \#\{T^j_{\min} \leq T_{\min}(\mathbf{x}, \mathbf{y})\}}{1 + N}. \tag{12}$$

The idea here is that the test statistics $(T^1_{\min}, \ldots, T^N_{\min})$ approximate the distribution of $T_{\min}(\mathbf{x}, \mathbf{y})$ under $H_0 : \mathbf{x}$ unassociated with $\mathbf{y}$. Therefore the expression (**??**) approximates the probability, under H_0, of seeing a more extreme (or equally extreme) value of $T_{\min}$ as the observed value $T_{\min}(\mathbf{x}, \mathbf{y})$.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64* (3), 479–498.

Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics 31* (6), 2013–2035.