

1 Multiple Testing

We have spent a considerable amount of time on testing a single null hypothesis, but we are typically in a setting where there is more than one hypothesis of interest to us. To this end, consider the following multiple hypothesis testing framework. We observe data X with distribution $P \in \Omega$. Let $H_i : P \in \omega_i \subseteq \Omega, i = 1, \dots, s$ be the family of null hypotheses of interest. We will assume that it is known how to test each null hypothesis individually in a way that controls the usual probability of a Type 1 error. Specifically, we will assume that for each null hypothesis there is a p -value $\hat{p}_i = \hat{p}_i(X)$. A p -value satisfies

$$\Pr_P\{\hat{p}_i \leq u\} \leq u \text{ for all } u \in (0, 1) \text{ and } P \in \omega_i . \quad (1)$$

Note that we do not require that $\Pr_P\{\hat{p}_i \leq u\} = u$ for all $u \in (0, 1)$ and $P \in \omega_i$; i.e., $\hat{p}_i \sim U(0, 1)$ for all $P \in \omega_i$. (This is so we can accommodate both situations in which the null hypothesis is composite and situations in which the underlying test statistic is discrete, as is the case with randomization tests.) A test of H_i at level α (i.e., a test for which the probability of a false rejection is no more than α) is therefore simply given by the test that rejects whenever $\hat{p}_i \leq \alpha$. Of course, if we were to test *all* of the null hypotheses in this way, then the probability of a false rejection may be much greater than α . In order to illustrate this possibility, suppose that all of the null hypotheses are true and moreover that each $\hat{p}_i \sim U(0, 1)$. Note that

$$\begin{aligned} \Pr_P\{\text{some false rejection}\} &= \Pr_P\left\{\bigcup_{1 \leq i \leq s} \{\hat{p}_i \leq \alpha\}\right\} \\ &\geq \Pr_P\{\hat{p}_1 \leq \alpha\} = \alpha , \end{aligned}$$

where the first inequality follows from the Bonferonni inequality. This is a very crude lower bound on the probability of some false rejection – its not hard to imagine situations (like the case in which the \hat{p}_i are all independent) in which the probability of some false rejection will be much greater. Hence, by testing the hypotheses in this fashion, we may very likely reject hypotheses that are in fact true.

Before proceeding, let's pause and give a few concrete examples of multiple testing problems that might arise in practice and the implications of ignoring the multiplicity in such problems (i.e., simply testing each null hypothesis in the way described above).

Example 1.1 Suppose one observes $(Y_j, X_{j,1}, \dots, X_{j,s}), j = 1, \dots, n$ with distribution P . Assume

$$Y_j = X_{j,1}\beta_1 + \dots + X_{j,s}\beta_s + \epsilon_j .$$

One would like to determine which covariates $X_{j,i}, i = 1, \dots, s$ help explain the dependent variable Y_j . To this end, one may consider the family of null hypotheses $H_i : \beta_i = 0, i = 1, \dots, s$. Ignoring multiplicity in this setting would lead one to decide that “too many” covariates helped explain the dependent variable. ■

Example 1.2 Suppose one observes $(R_{1,j}, \dots, R_{s,j}, B_j), j = 1, \dots, n$ where $R_{i,j}$ is the return from investment strategy i in period j and B_j is the return from some “benchmark” investment strategy in period j . For example, the benchmark strategy may be some proxy for the risk free rate of return. One would like to determine which strategies outperform the benchmark. To this end, one may consider the family of null hypotheses $H_i : \theta_i \leq 0, i = 1, \dots, s$, where $\theta_i = E[R_{i,j} - B_j]$. Ignoring multiplicity in this setting would lead one to decide that “too many” strategies outperformed the benchmark. ■

Example 1.3 One observes $(Y_{1,j}, \dots, Y_{s,j}, D_j), j = 1, \dots, n$, where $Y_{i,j}$ is the i th outcome for the j th individual and D_j is an indicator variable for whether the j th individual was treated. Assume that individuals are assigned at random to treatment. One would like to determine which of the s outcomes are affected by the treatment. To this end, one may consider the family of null hypotheses $H_i : Y_{i,j} \perp D_j, i = 1, \dots, s$. Ignoring multiplicity in this setting would lead one to decide that the treatment impacted “too many” of the outcomes. ■

Our goal is therefore to devise a way of testing the family of null hypotheses of interest without making “too many” false rejections. The classical definition of “too many” in the multiple testing literature is the *familywise error rate*. We may abbreviate the familywise error rate as $FWER = FWER_P$. (The notation emphasizes the fact that the familywise error rate obviously depends on the distribution of the observed data P .) We will say that a procedure controls the familywise error rate at level α if

$$FWER_P = \Pr_P\{\text{some false rejection}\} \leq \alpha \text{ for all } P \in \Omega . \quad (2)$$

(This is sometimes referred to as strong control of the familywise error rate to distinguish it from weak control of the familywise error rate, which only requires $FWER_P \leq \alpha$ for $P \in \cap_{1 \leq i \leq s} \omega_i$, that is, when all null hypotheses are true. Weak control is not particularly useful, so we won’t make any more use of it.) We will seek procedures that control the familywise error rate at level α only under the assumption that the p -values satisfy (1). In particular, we will not impose any assumptions on the joint distribution of the p -values. In most situations, it would not be sensible to assume, for example, that the p -values were independent.

Let $I(P)$ denote the set of indices corresponding to true null hypotheses; that is,

$$I(P) = \{1 \leq i \leq s : P \in \omega_i\} .$$

Suppose we were to test all of the null hypotheses by comparing each p -value with a single, common cutoff c ; that is, reject H_i if $\hat{p}_i \leq c$. Such a procedure is an example of a *single-step* multiple testing procedure. For such a procedure, we can compute

$$\begin{aligned} FWER_P &= \Pr_P\{i \in I(P) : \hat{p}_i \leq c\} = \Pr_P\left\{ \bigcup_{i \in I(P)} \{\hat{p}_i \leq c\} \right\} \\ &\leq \sum_{i \in I(P)} \Pr_P\{\hat{p}_i \leq c\} \leq \sum_{i \in I(P)} c = |I(P)|c \leq sc . \end{aligned}$$

The first inequality follows from the so-called Bonferonni inequality, the second inequality follows from (1), and the final inequality follows from the

fact that $|I(P)| \leq s$. Therefore, in order to ensure (2), it suffices to choose $c = \alpha/s$. Because of its use of the Bonferonni inequality, the resulting procedure is referred to as the Bonferonni procedure.

It is worthwhile to point out that if one uses the Bonferonni procedure, one will not reject a null hypothesis unless its corresponding p -value is exceedingly small – less than α/s ! A natural question is therefore to ask whether one could improve upon the procedure at all.

We might ask first whether the Bonferonni procedure is even best among all single step procedures – after all, the Bonferonni inequality has a reputation as being quite crude. In other words, can we decrease the cutoff $c = \alpha/s$ without violating control of the familywise error rate? To this end, let's reexamine the proof of the above bound on the familywise error rate with $c = \alpha/s$. The third inequality would be an equality if $I(P) = s$. The second inequality would be an equality if $\hat{p}_i \sim U(0, 1)$. The first inequality would be an equality if the events $\{\hat{p}_i \leq \alpha/s\}, i = 1, \dots, s$ were all disjoint. Is it possible for this to be true while still having each $\hat{p}_i \sim U(0, 1)$? The answer is yes. To see how, let divide the unit interval into s equal pieces – $(0, 1/s], (1/s, 2/s], \dots, ((s-1)/s, 1)$. For $1 \leq i \leq s$, let $U_i \sim U((i-1)/s, i/s)$. Let π be a random permutation of $1, \dots, s$. Define $\hat{p}_i = U_{\pi(i)}$. To see that $\hat{p}_i \sim U(0, 1)$, let $u \in (0, 1)$. Define j so that $u \in ((j-1)s, j/s]$. We have that

$$\begin{aligned} \Pr\{\hat{p}_i \leq u\} &= \Pr\{U_{\pi(i)} \leq u\} \\ &= \frac{1}{s} \sum_{1 \leq i \leq s} \Pr\{U_i \leq u\} \\ &= \frac{1}{s} \left((j-1) + s \left(u - \frac{j-1}{s} \right) \right) = u . \end{aligned}$$

Moreover, by construction the events $\{\hat{p}_i \leq \alpha/s\}, i = 1, \dots, s$ are disjoint. Therefore, for such a joint distribution of p -values

$$FWER = \alpha .$$

It follows that if we increased $c = \alpha/s$ at all (say, by replacing α with some $\alpha' > \alpha$), then the $FWER > \alpha$.

So, in order to improve upon the Bonferonni procedure, we must abandon the single-step paradigm. In other words, different p -values must be compared with different cutoffs – e.g., the smallest p -value might be compared with a smaller cutoff than, say, the largest p -value. This idea leads to a class of *stepwise* procedures which order the p -values from smallest to largest and compare them with different thresholds to determine which ones to reject and which ones not to reject. It turns out that by doing so, one can improve upon the Bonferonni procedure substantially one can construct a procedure that controls the familywise error rate but always rejects at least as many hypotheses (i.e., is more powerful in the sense that it can reject more false hypotheses).

One class of stepwise procedures are known as *stepdown* procedures. Stepdown procedures begin with the most significant (smallest) p -values and then “step down” to the less significant (larger) p -values. In order to describe a stepdown procedure, let

$$\hat{p}_{(1)} \leq \cdots \leq \hat{p}_{(s)}$$

denote the ordered p -values and let

$$H_{(1)}, \dots, H_{(s)}$$

denote the corresponding null hypotheses. Let $c_1 \leq \cdots \leq c_s$ be an increasing sequence of constants. A stepdown procedure determines which null hypotheses to reject in the following fashion:

Step 1: If $\hat{p}_{(1)} > c_1$, then stop (and reject no null hypotheses). Otherwise, reject $H_{(1)}$ and go to Step 2.

⋮

Step j : If $\hat{p}_{(j)} > c_j$, then stop. Otherwise, reject $H_{(j)}$ and go to Step $j+1$.

⋮

Step s : If $\hat{p}_{(s)} > c_s$, then stop. Otherwise, reject $H_{(s)}$ and stop.

Equivalently, a stepdown procedure may be described as follows: If $\hat{p}_{(1)} > c_1$, then reject no null hypotheses; otherwise reject $H_{(1)}, \dots, H_{(r)}$ where r is the largest index such that

$$\hat{p}_{(1)} \leq c_1, \dots, \hat{p}_{(r)} \leq c_r .$$

Holm (1979) proposed a stepdown procedure with c_i defined by the rule

$$c_i = \frac{\alpha}{s - i + 1} .$$

We will now show that this procedure controls the familywise error rate. Assume w.l.o.g. that $|I(P)| \geq 1$, for otherwise there is nothing to prove. Suppose that the procedure makes at least one false rejection. If so, there must be a first step at which a false rejection occurs. Let j denote this random step. This means two things: first, $\hat{p}_{(j)} \leq c_j$; second, exactly one false rejection was made in the first j steps. Hence, there are at least $j - 1$ false null hypotheses, or, put differently, there are at most $s - (j - 1) = s - j + 1$ true hypotheses, i.e. $|I(P)| \leq s - j + 1$. It follows that

$$\hat{p}_{(j)} \leq c_j = \frac{\alpha}{s - j + 1} \leq \frac{\alpha}{|I(P)|} .$$

This in turn implies that

$$\hat{q}_{(1)} \leq \frac{\alpha}{|I(P)|} ,$$

where

$$\hat{q}_{(1)} \leq \dots \leq \hat{q}_{(|I(P)|)}$$

denote the ordered values of the p -values corresponding to true null hypotheses. Therefore,

$$FWER_P \leq \Pr_P\{\hat{q}_{(1)} \leq \frac{\alpha}{|I(P)|}\} ,$$

but, by the proof for the Bonferonni procedure, we know that this probability is bounded above by α .

Again, a natural question to ask is whether it is possible to improve upon the Holm procedure. In particular, we might ask whether it is possible to

increase any of the constants c_i and preserve control of the familywise error rate. The answer is in fact ‘no’. To see this, we will exhibit for every i a joint distribution of p -values such that

$$FWER = \Pr\{\hat{p}_{(1)} \leq c_1, \dots, \hat{p}_{(i)} \leq c_i\} = \alpha$$

and the probability is strictly increasing in c_i . To this end, let $i - 1$ of the p -values correspond to false null hypotheses and $s - i + 1$ of the p -values correspond to true null hypotheses. To make our life as easy as possible, we may as well choose the $i - 1$ p -values corresponding to false null hypotheses to be identically equal to zero. Therefore, $\hat{p}_{(1)} = \dots = \hat{p}_{(i-1)} = 0$. It follows that

$$FWER = \Pr\{\hat{p}_{(1)} \leq c_1, \dots, \hat{p}_{(i)} \leq c_i\} = \Pr\{\hat{q}_{(1)} \leq c_i\},$$

where $\hat{q}_{(1)} \leq \dots \leq \hat{q}_{(s-i+1)}$ denote the ordered values of the p -values corresponding to true null hypotheses. We will specify the joint distribution of the $(\hat{q}_1, \dots, \hat{q}_{s-i+1})$ as follows. Divide the unit interval into $s - i + 1$ equal pieces $-(0, \frac{1}{s-i+1}]$, $(\frac{1}{s-i+1}, \frac{2}{s-i+1}]$, \dots , $(\frac{s-i}{s-i+1}, 1]$. For $1 \leq j \leq s - i + 1$, let $U_j \sim U(0, \frac{j}{s-i+1})$. Let π be a random permutation of $1, \dots, s - i + 1$. Define $\hat{q}_j = U_{\pi(j)}$. It follows from our earlier arguments that each $\hat{q}_j \sim U(0, 1)$. Moreover,

$$\begin{aligned} \Pr\{\hat{q}_{(1)} \leq c_i\} &= \Pr\left\{\bigcup_{1 \leq j \leq s-i+1} \{\hat{q}_j \leq c_i\}\right\} \\ &= \sum_{1 \leq j \leq s-i+1} \Pr\{\hat{q}_j \leq c_i\} \\ &= (s - i + 1)c_i = \alpha. \end{aligned}$$

For such a joint distribution of p -values, $FWER$ is strictly increasing in c_i , as desired.

Recall that the “least favorable” distribution of p -values for the Bonferroni procedure involved s p -values that satisfied (1). In contrast, there are multiple “least favorable” distributions of p -values for the Holm procedure, and these distributions involve different numbers of p -values that satisfy (1).

It is worth pointing out that the proof that it was not possible to improve upon the Bonferonni procedure involves a fairly peculiar joint distribution of p -values. To the extent that the true distribution of the p -values is different from this “least favorable” case, it may be possible to improve upon the Bonferonni procedure. To do this, one must construct a cutoff that incorporates in some way the joint distribution of the p -values. In the same way, it may be possible to improve upon the Holm procedure. We will not develop these methods here.

In some instances, one may be willing to relax control of the familywise error rate so as to gain power – the ability to reject false null hypotheses. After all, even with the Holm procedure, in order to reject any null hypotheses at all, the smallest p -value must be less than α/s . For example, one may only require control of the k -familywise error rate (k – $FWER$) at level α ; that is,

$$k - FWER_P = \Pr_P\{\geq k \text{ false rejections}\} \leq \alpha \text{ for all } P \in \Omega .$$

For large s , this may be a reasonable measure of error control. How can one construct multiple testing procedures that control the k – $FWER$? Consider first a single-step multiple testing procedure with cutoff c . For such a procedure, we have that

$$\begin{aligned} k - FWER_P &= \Pr_P\left\{ \sum_{i \in I(P)} I\{\hat{p}_i \leq c\} \geq k \right\} \leq \frac{E_P[\sum_{i \in I(P)} I\{\hat{p}_i \leq c\}]}{k} \\ &= \frac{|I(P)|\Pr_P\{\hat{p}_i \leq c\}}{k} \leq \frac{s}{k}c , \end{aligned}$$

where the first inequality follows from Markov’s inequality and the second inequality follows from (1). It therefore suffices to choose $c = \frac{k\alpha}{s}$ – a k -fold improvement over the original Bonferonni procedure.

Although this argument may seem quite crude, it is again possible to show that it is not possible to increase this cutoff at all without violating control of the k -familywise error rate. This procedure is therefore the best for control of the k -familywise error rate among all single-step procedures,

but we can improve upon it as before by considering stepwise procedures. In particular, we may consider the stepdown procedure with c_i defined by the rule

$$c_i = \begin{cases} \frac{k\alpha}{s} & \text{if } i < k \\ \frac{k\alpha}{s-i+k} & \text{if } i \geq k \end{cases} .$$

We now argue that such a procedure controls the k -familywise error rate. Assume w.l.o.g. that $|I(P)| \geq k$, for otherwise there is nothing to prove. Suppose that the procedure makes at least k false rejections. If so, there must be a first step at which k or more false rejections occur. Let j denote this random step. This means two things, first $\hat{p}_{(j)} \leq c_j$; second, exactly k false rejections were made in the first j steps. Hence, there are at least $j - k$ false null hypotheses, or, put differently, there are at most $s - (j - k) = s - j + k$ true null hypotheses, i.e., $|I(P)| \leq s - j + k$. It follows that

$$\hat{p}_{(j)} \leq c_j = \frac{k\alpha}{s - j + k} \leq \frac{k\alpha}{|I(P)|} .$$

This in turn implies that

$$\hat{q}_{(k)} \leq \frac{k\alpha}{|I(P)|} ,$$

where

$$\hat{q}_{(1)} \leq \cdots \leq \hat{q}_{(|I(P)|)}$$

denote, as before, the ordered values of the p -values corresponding to true null hypotheses. Therefore,

$$k - FWER_P \leq \Pr_P \left\{ \hat{q}_{(k)} \leq \frac{k\alpha}{|I(P)|} \right\} ,$$

but, by the proof for the single-step procedure above, we know that this probability is bounded above by α .

Of course, one can increase c_i for $i < k$ to 1 without violating control of the k -familywise error rate. To see this, simply note that these constants never entered the argument above. On the other hand, it is not possible to increase c_i for $i \geq k$ at all without violating control of the k -familywise error rate.

It is natural to allow the number of false rejections one is willing to tolerate to vary with the total number of rejections. If one makes 10 rejections, then perhaps one would be unwilling to have more than one false rejection, but if one makes 100 rejections, then perhaps one would be willing to tolerate, say, 10 false rejections. This idea leads one to control of the *false discovery proportion* (FDP). The false discovery proportion is defined to be

$$FDP = \begin{cases} \frac{F}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases},$$

where F is the total number of false rejections and R is the total number of rejections. Control of the false discovery proportion requires that

$$\Pr_P\{FDP > \gamma\} \leq \alpha \text{ for all } P \in \Omega$$

for some user-specified value of $\gamma \in [0, 1]$. Note that if $\gamma = 0$, control of the false discovery proportion reduces to control of the familywise error rate.

A popular measure of error control that is related to control of the false discovery proportion is control of the *false discovery rate* (FDR), proposed by Benjamini and Hochberg (1995). The false discovery rate is defined to be

$$FDR_P = E_P[FDP]$$

and control of the false discovery rate requires that

$$FDR_P \leq \alpha \text{ for all } P \in \Omega .$$

Finally, it is important to point out that often one may only have p -values that satisfy an asymptotic version of (1):

$$\limsup_{n \rightarrow \infty} \Pr_P\{\hat{p}_i \leq u\} \leq u \text{ for all } u \in (0, 1) \text{ and } P \in \omega_i .$$

In this case, the procedures above are still valid in the sense that they ensure that

$$\limsup_{n \rightarrow \infty} FWER_P \leq \alpha .$$