Multiple Testing

Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf

Abstract

Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis. If decisions about the individual hypotheses are based on the unadjusted marginal p-values, then there is typically a large probability that some of the true null hypotheses will be rejected. Unfortunately, such a course of action is still common. In this article, we describe the problem of multiple testing more formally and discuss methods which account for the multiplicity issue. In particular, recent developments based on resampling result in an improved ability to reject false hypotheses compared to classical methods such as Bonferroni.

KEY WORDS: Multiple Testing, Familywise Error Rate, Resampling

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses. This scenario is quite common in much empirical research in economics. Some examples include: (i) one fits a multiple regression model and wishes to decide which coefficients are different from zero; (ii) one compares several forecasting strategies to a benchmark and wishes to decide which strategies are outperforming the benchmark; (iii) one evaluates a program with respect to multiple outcomes and wishes to decide for which outcomes the program yields significant effects.

If one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected by chance alone may be unduly large. Take the case of $S = 100$ hypotheses being tested at the same time, all of them being true, with the size and level of each test exactly equal to α. For $\alpha = 0.05$, one expects five true hypotheses to be rejected. Further, if all tests are mutually independent, then the probability that at least one true null hypothesis will be rejected is given by $1 - 0.95^{100} = 0.994$.

Of course, there is no problem if one focuses on a particular hypothesis, and only one of them, *a priori*. The decision can still be based on the corresponding marginal p-value. The problem only arises if one searches the list of p-values for significant results *a posteriori*. Unfortunately, the latter case is much more common.

Notation

Suppose data X is generated from some unknown probability distribution P. In anticipation of asymptotic results, we may write $X = X^{(n)}$, where n typically refers to the sample size. A model assumes that P belongs to a certain family of probability distributions, though we make no rigid requirements for this family; it may be a parametric, semiparametric, or nonparametric model.

Consider the problem of simultaneously testing a hypothesis H_s against the alternative hypothesis H'_s, for $s = 1, \ldots, S$. A multiple testing procedure (MTP) is a rule which makes some decision about each H_s. The term *false discovery* refers to the rejection of a true null hypothesis. Also, let $I(P)$ denote the set of true null hypotheses, that is, $s \in I(P)$ if and only if (iff) H_s is true.

We also assume that a test of the individual hypothesis H_s is based on a test statistic $T_{n,s}$, with large values indicating evidence against H_s. A marginal p-value for testing H_s is denoted by $\hat{p}_{n,s}$.

Familywise Error Rate

Accounting for the multiplicity of individual tests can be achieved by controlling an appropriate *error rate*. The traditional or classical *familywise error rate* (FWE) is the probability of one

or more false discoveries:

$$\text{FWE}_P = P\{\text{reject at least one hypothesis } H_s : s \in I(P)\} \ .$$

Control of the FWE means that, for a given significance level α,

$$\text{FWE}_P \leq \alpha \quad \text{for any } P \ . \tag{1}$$

Control of the FWE allows one to be $1-\alpha$ confident that there are no false discoveries among the rejected hypotheses.

Note that 'control' of the FWE is equated with 'finite-sample' control: (1) is required to hold for any given sample size n. However, such a requirement can often only be achieved under strict parametric assumptions or for special permutation set-ups. Instead, one then settles for *asymptotic* control of the FWE:

$$\limsup_{n\to\infty} \text{FWE}_P \leq \alpha \quad \text{for any } P \ . \tag{2}$$

Methods Based on Marginal p-values

MTPs falling in this category are derived from the marginal or individual p-values. They do not attempt to incorporate any information about the dependence structure between these p-values. There are two advantages to such methods. First, one might only have access to the list of p-values from a past study, but not to the underlying complete data set. Second, such methods can be very quickly implemented. On the other hand, as discussed later, such methods are generally sub-optimal in terms of power.

To show that such methods control the desired error rate, one needs a condition on the p-values corresponding to the true null hypotheses:

$$H_s \text{ true } \iff s \in I(P) \Longrightarrow P\{\hat{p}_{n,s} \leq u\} \leq u \quad \text{for any } u \in (0,1) \ . \tag{3}$$

Condition (3) merely asserts that, when testing H_s alone, the test that rejects H_s when $\hat{p}_{n,s} \leq u$ has level u, i.e., it is a proper p-value.

The classical method to control the FWE is the Bonferroni method, which rejects H_s iff $\hat{p}_{n,s} \leq \alpha/S$. More generally, the weighted Bonferroni method rejects H_s if $\hat{p}_{n,s} \leq w_s \cdot \alpha/S$, where the constants w_s, satisfying $w_s \geq 0$ and $\sum w_s = 1$, reflect the 'importance' of the individual hypotheses.

An improvement is obtained by the method of Holm (1979). The marginal p-values are ordered from smallest to largest: $\hat{p}_{n,(1)} \leq \hat{p}_{n,(2)} \leq \ldots \leq \hat{p}_{n,(S)}$ with their corresponding null hypotheses labeled accordingly: $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. Then, $H_{(s)}$ is rejected iff $\hat{p}_{n,(j)} \leq \alpha/(S-j+1)$ for $j = 1, \ldots, s$. In other words, the method starts with testing the most significant hypothesis by comparing its p-value to α/S, just as the Bonferroni method. If the hypothesis is rejected, the method moves on to the second most significant hypothesis by comparing its p-value to $\alpha/(S-1)$, and so on, until the procedure comes to a stop. Necessarily, all hypotheses

rejected by Bonferroni will also be rejected by Holm, but potentially a few more. So, trivially, the method is more powerful. But it still controls the FWE under (3).

If it is known that the p-values are suitably positive dependent, then further improvements can be obtained with the use of Simes identity; see Sarkar (1998).

So far, we have assumed 'finite-sample validity' of the null p-values expressed by (3). However, often p-values are derived by asymptotic approximations or resampling methods, only guaranteeing 'asymptotic validity' instead:

$$H_s \text{ true } \iff s \in I(P) \Longrightarrow \limsup_{n\to\infty} P\{\hat{p}_{n,s} \le u\} \le u \quad \text{for any } u \in (0,1) \ . \tag{4}$$

Under this more realistic condition, the MTPs presented in this section only provide asymptotic control of the FWE in the sense of (2).

Single-step vs. Stepwise Methods

In single-step MTPs, individual test statistics are compared to their critical values simultaneously, and after this simultaneous 'joint' comparison, the multiple testing method stops. Often there is only one common critical value, but this need not be the case. More generally, the critical value for the sth test statistic may depend on s. An example is the weighted Bonferroni method discussed above.

Often single-step methods can be improved in terms of power via stepwise methods, while still maintaining control of the desired error rate. Stepdown methods start with a single-step method but then continue by possibly rejecting further hypotheses in subsequent steps. This is achieved by decreasing the critical values for the remaining hypotheses depending on the hypotheses already rejected in previous steps. As soon as no further hypotheses are rejected, the method stops. The Holm (1979) method discussed above is a stepdown method.

Stepdown methods therefore improve upon single-step methods by possibly rejecting 'less significant' hypotheses in subsequent steps. In contrast, there also exist stepup methods that start with the least significant hypotheses, having the smallest test statistics, and then 'step up' to further examine the remaining hypotheses having larger test statistics.

More general methods to construct MTPs which control the FWE can be obtained by the closure method; see Hochberg and Tamhane (1987).

Resampling Methods Accounting for Dependence

Methods based on p-values often achieve (asymptotic) control of the FWE by assuming (i) a worst-case dependence structure or (ii) a 'convenient' dependence structure (such as mutual independence). This has two potential disadvantages. In case of (i), the method can be quite sub-optimal in terms of power if the true dependence structure is quite far away from the worst-case scenario. In case of (ii), if the convenient dependence structure does not hold, even asymptotic control may not result. As an example for case (i), consider the Bonferroni method.

If the p-values were perfectly dependent, then the cut-off value could be changed from α/S to α. While perfect dependence is rare, this example serves to make a point. In the realistic scenario of 'strong cross dependence', the cut-off value could be changed to something a lot larger than α/S while still maintaining control of the FWE. Hence, it is desirable to account for the underlying dependence structure.

Of course, this dependence structure is unknown and must be (implicitly) estimated from the available data. Consistent estimation, in general, requires that the sample size grows to infinity. Therefore, in this subsection, we will settle for asymptotic control of the FWE. In addition, we will specialize to making simultaneous inference on the elements of a parameter vector $\theta = (\theta_1, \ldots, \theta_S)^T$. Assume the individual hypotheses are one-sided of the form:

$$H_s : \theta_s \leq 0 \quad \text{vs.} \quad H_s' : \theta_s > 0 \ . \tag{5}$$

Modifications for two-sided hypotheses are straightforward.

The test statistics are of the form $T_{n,s} = \hat{\theta}_{n,s}/\hat{\sigma}_{n,s}$. Here, $\hat{\theta}_{n,s}$ is an estimator of θ_s computed from $X^{(n)}$. Further, $\hat{\sigma}_{n,s}$ is either a standard error for $\hat{\sigma}_{n,s}$ or simply equal to $1/\sqrt{n}$ in case such a standard error is not available or only very difficult to obtain.

We start by discussing a single-step method. An idealized method would reject all H_s for which $T_{n,s} \geq d_1$ where d_1 is the $1-\alpha$ quantile under P of the random variable $\max_s(\hat{\theta}_{n,s} - \theta_s)/\hat{\sigma}_{n,s}$. Naturally, the quantile d_1 does not only depend on the marginal distributions of the centered statistics $(\hat{\theta}_{n,s} - \theta_s)/\hat{\sigma}_{n,s}$ but, crucially, also on their dependence structure.

Since P is unknown, the idealized critical value d_1 is not available. But it can be estimated consistently under weak regularity conditions as follows. Take $\hat{d}_1$ as the $1-\alpha$ quantile under $\hat{P}_n$ of $\max_s(\hat{\theta}^*_{n,s} - \hat{\theta}_{n,s})/\hat{\sigma}^*_{n,s}$. Here, $\hat{P}_n$ is an *unrestricted* estimate of P. Further $\hat{\theta}^*_{n,s}$ and $\hat{\sigma}^*_{n,s}$ are the estimator of θ_s and its standard error (or simply $1/\sqrt{n}$), respectively, computed from $X^{(n),*}$ where $X^{(n),*} \sim \hat{P}_n$. In other words, we use the bootstrap to estimate d_1. The particular choice of $\hat{P}_n$ depends on the situation. In particular, if the data are collected over time a suitable time series bootstrap needs to be employed; see Davison and Hinkley (1997) and Lahiri (2003).

We have thus described a single-step MTP. However, a stepdown improvement is possible. In any given step j, one simply discards the hypotheses that have been rejected so far and applies the single-step MTP to the remaining universe of non-rejected hypotheses. The resulting critical value $\hat{d}_j$ necessarily satisfies $\hat{d}_j \leq \hat{d}_{j-1}$ so that new rejections may result; otherwise the method stops.

This bootstrap stepdown MTP provides asymptotic control of the FWE under remarkably weak regularity conditions. Mainly, it is assumed that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a (multivariate) continuous limit distribution and that the bootstrap consistently estimates this limit distribution. In addition, if standard errors are employed for $\hat{\sigma}_{n,s}$, as opposed to simply using $1/\sqrt{n}$, it is assumed that they converge to the same non-zero limiting values in probability, both in the 'real world' and in the 'bootstrap world'. Under even weaker regularity conditions, a subsampling approach could be used instead; see Romano and Wolf (2005). Furthermore,

when a randomization setup applies, randomization methods can be used as an alternative; see Romano and Wolf (2005) again.

Related methods are developed in White (2000) and Hansen (2005). However, both works only treat the special case $k = 1$ and are restricted to single-step methods. In addition, White (2000) does not consider studentized test statistics. Stepwise bootstrap methods to control the FWE are already proposed in Westfall and Young (1993). An important difference in their approach is that they bootstrap under the joint null, that is, they use a *restricted* estimate of P where the contraints of all null hypotheses jointly hold. This approach requires the so-called *subset pivotality* condition and is generally less valid than the approaches discussed so far based on an unrestricted estimate of P; e.g., see Example 4.1 of Romano and Wolf (2005).

Generalized Error Rates

So far, attention has been restricted to the FWE. Of course, this criterion is very strict; not even a single true hypothesis is allowed to be rejected. When S is very large, the corresponding multiple testing procedure (MTP) might result in low power, where we loosely define 'power' as the ability to reject false null hypotheses.

Let F denote the number of false rejections and let R denote the total number of rejections. The *false discovery proportion* (FDP) is defined as $\text{FDP} = (F/R) \cdot 1\{R > 0\}$, where $1\{\cdot\}$ denotes the indicator function. Instead of the FWE, one may consider the probability of the FDP exceeding a small, pre-specified proportion: $P_\theta\{\text{FDP} > \gamma\}$, for some $\gamma \in [0, 1)$. The special choice of $\gamma = 0$ simplifies to the traditional FWE. Another alternative to the FWE is the *false discovery rate* (FDR), defined to be the expected value of the FDP: $\text{FDR}_\theta = E_\theta(\text{FDP})$.

By allowing for a small (expected) fraction of false discoveries, one can generally gain a lot of power compared to FWE control, especially when S is large. For the discussion of MTPs to provide (asymptotic) control of the FDP and the FDR, the reader is referred to Romano et al. (2008a,b) and the references therein.

References

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economics Statistics*, 23:365–380.

Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling (with discussion). *Test*, 17(3):417–442.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008b). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(2):404–447.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of simes conjecture. *Annals of Statistics*, 26:494–504.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley, New York.

White, H. L. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.