Generalized Error Control in Multiple Hypothesis Testing

July 2007, MCP 5

Joseph P. Romano, Stanford University

http://www-stat.stanford.edu/~romano/joe.html

romano@stanford.edu

Joint work with M. Wolf

Other Collaborators: Guo, Lehmann, Shaffer, Shaikh

The Basic Setup

Observe data $X = (X_1, \ldots, X_n)$ from P.

Test hypotheses $H_1, \ldots, H_s$: $H_j \equiv P \in \omega_j$

Let $I = I(P) \subset \{1, \ldots, s\}$ denote the indices of the set of true hypotheses: $j \in I$ if and only $P \in \omega_j$. The familywise error rate (FWE_P) is the probability under P that any H_j with $j \in I$ is rejected.

Require $\mathrm{FWE}_P \leq \alpha \quad \forall P$.

Suppose H_j is rejected for large values of $T_{n,j}$, or small p-value $\hat{p}_j$.

Starting point: Stepdown methods based on marginal p-values. Given p-value $\hat{p}_j$ for testing H_j, order them as

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

with corresponding $H_{(1)}, \ldots, H_{(s)}$.

Let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s$.

Method: Let j^* be the largest j: $\hat{p}_{(1)} \leq \alpha_1, \ldots, \hat{p}_{(j)} \leq \alpha_j$

and reject $H_{(1)}, \ldots, H_{(j^*)}$.

Bonferroni: $\alpha_i = \alpha/s$ controls the FWE.

Holm: $\alpha_i = \alpha/(s - i + 1)$

While a big improvement over Bonferroni, still can be conservative.

Directions for Improving Holm

I. Incorporating or estimating the dependence structure of p-values. This is the approach taken in Westfall and Young (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.* Also see Dudoit, Pollard and van der Laan (2004) and Romano and Wolf (2005).

II. Relax control of the FWE. Given a multiple testing decision rule, let $F = \#$ false rejections, $R = \#$ rejections. Define the *false discovery proportion* (FDP) as F/R (defined to be 0 if $R = 0$).

(i) As a measure of error control, Benjamini & Hochberg (1995) popularized the *false discovery rate* (FDR) defined by

Require $E(\text{FDP}) \leq \alpha$.

(ii) Let k-FWE: the probability that $F \geq k$. Require $P\{F \geq k\} \leq \alpha$.

(iii) Given a value γ, require $P\{\text{FDP} > \gamma\} \leq \alpha$.

eg. FDP control with $\alpha = 1/2$ means

$$\text{median}(\text{FDP}) \leq \gamma$$

Given p-values for individual tests, stepdown methods exist for controlling these at level α with no assumptions about the dependence structure of the p-values; see Benjamini and Yekutieli (2001) and Romano and Shaikh (2006).

Here, we will combine I (incorporate dependence structure) and II (weaken error measure) to achieve greater power.

Goal: Derive stepwise procedures that control k-FWE and FDP which incorporate dependence structure among test statistics or p-values. Begin with k-FWE.

Theorem 1 *(Generalized Bonferroni) The method that rejects H_i if $\hat{p}_i \le k\alpha/s$ controls the k-FWE.*

Theorem 2 *(Generalized Holm) Let $\alpha_i = k\alpha/s$ if $i \le k$ and*

$$\alpha_i = \frac{k\alpha}{s+k-i} \quad \text{if } i > k \ . \tag{1}$$

The stepdown procedure with above α_i controls the k-FWE.

Above results due to Hommel and Hoffman (1987) and elaborated on in Lehmann and Romano (2005).

The above results do not incorporate dependence structure. But we now argue it is vital to do so, especially for generalized error rates.

Under independence, one can improve the constant $k\alpha/s$ dramatically. Let

$$H_{k,s}(u) = \sum_{j=k}^{s} \binom{s}{j} u^j (1-u)^{s-j} \ . \qquad (2)$$

Consider the (generalized Sidák) procedure that rejects any H_i whose corresponding p-value $\hat{p}_i$ is $\leq H_{k,s}^{-1}(\alpha)$.

This controls the k-FWE (Guo and Romano, 2007).

Further stepdown improvement: Let

$$\alpha_1 = \cdots = \alpha_k = H_{k,s}^{-1}(\alpha)$$

and, for $j > 0$,

$$\alpha_{k+j} = H_{k,s-j}^{-1}(\alpha) \ .$$

This controls the k-FWE.

How dramatic are these improvements? For $k = 1$, the ratio of critical values satisfies:

$$\lim_{s \to \infty} \frac{1 - (1-\alpha)^{1/s}}{\alpha/s} \to \frac{-\log(1-\alpha)}{\alpha} \ ,$$

which $= 1.026$ when $\alpha = 0.05$.

In general, if you use the cutoff $k\alpha/s$, then under independence,

$$k - \mathrm{FWE} = O(\alpha^k) \qquad \text{as } \alpha \to 0, s \to \infty \ .$$

Table 1: Single step constants for k-FWE control with $s = 100$ and $\alpha = 0.05$

k	$A = k\alpha/s$	$B = C_{k,s}(\alpha)$	B/A
1	0.0005	0.00051	1.026
2	0.0010	0.00353	3.530
3	0.0015	0.00806	5.376
5	0.0025	0.01913	7.653
7	0.0035	0.03140	8.972
10	0.0050	0.05062	10.124

A general construction of stepdown tests under weak assumptions Related work by Korn, Troendle, McShane and Simin (2004), van der Laan, Dudoit and Pollard (2004).

Let P be the true probability, $P \in \Omega$.

H_j specified by $\omega_j \subset \Omega$.

$j \in I(P)$ if and only $P \in \omega_j$.

Let

$$T_{n,r_1} \geq T_{n,r_2} \geq \cdots \geq T_{n,r_s}$$

denote the observed ordered test statistics, and $H_{(1)}$, $H_{(2)}, \ldots, H_{(s)}$ the corresponding hypotheses.

Motivating Example: Correlations

$X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$, with $X_i = (X_{i,1}, \ldots, X_{i,d})$.

Assume $E|X_{i,j}|^2 < \infty$ and $Var(X_{i,j}) > 0$, so that the correlation between $X_{1,i}$ and $X_{1,j}$, namely $\rho_{i,j}$ is well-defined.

$H_{i,j}: \ \rho_{i,j} = 0$, $(s = \binom{d}{2})$

Let $T_{n,i,j}$ = sample correlation between variables i and j. (Note we are indexing hypotheses and test statistics by 2 indices i and j.)

By Aitken (1969, 71), if $d = 3$, $H_{1,2}$ and $H_{1,3}$ are true but $H_{2,3}$ is false, the limiting distribution of $n^{1/2}(T_{n,1,2}, T_{n,1,3})$ is biv. normal: means 0, variances 1, and correlation $\rho_{2,3}$. Subset pivotality fails, as noted by WY (1993).

A stepdown procedure begins with the most significant test statistic. First, test all null hypotheses $H_1, \ldots, H_s$. $H_{(1)}$ is rejected if T_{n,r_1} is large. If it is not large, accept all hypotheses. Once a hypothesis is rejected, remove it and test the remaining hypotheses by rejecting for large values of the maximum of the remaining test statistics, and so on.

Problem: how to construct the critical values at each step so that the k-FWE is controlled?

Idea: Reduce the multiple testing problem of controlling the k-FWE in a stepdown procedure to that of constructing single tests which control the probability of k or more false rejections.

Notation: If $\{y_i,\ i \in K\}$ is a collection of numbers indexed by a finite set K having $|K|$ elements. Then, for $k \leq |K|$, $k\text{-}\max_{i \in K}(y_i)$ is used to denote the kth largest value of the y_i with $i \in K$.

Start with single step method. Suppose $H_i \equiv \theta_i(P) = 0$. let $K_0 = \{1, \ldots, s\}$. For any $K \subset K_0$, let $c_{n,K}(\alpha, k, P)$ denote an α-quantile of the distribution of $k\text{-}\max_{j \in K} |\hat{\theta}_{n,j} - \theta_j(P)|$ under P. (Note: can studentize here.)

Then, $\{\theta_j \in K_0 : |\hat{\theta}_{n,j} - \theta_j| \le c_{n,K_0}(1-\alpha, k, P)\}$ is a confidence region for $(\theta_j : j \in K_0)$ which contains all of the θ_j, except possibly $k-1$ of them.

By *duality*, rejecting any H_j for which $|\hat{\theta}_{n,j}|$ exceeds $c_{n,K_0}(1-\alpha, k, P)$ controls the k-FWE. Since P is unknown, a bootstrap method replaces P by $\hat{Q}_n$ and uses the critical value $c_{n,K}(1-\alpha, k, \hat{Q}_n)$, providing an asymptotic solution (under weak conditions).

In the following algorithm designed for control of the k-FWE, suppose $\hat{c}_{n,K}(1-\alpha, k)$ are used to test H_i with $i \in K$.

Algorithm 1 **Generic Stepdown Method For Control of the k-FWE** Let $A_1 = \{1, \ldots, s\}$.

1. If $\max_{i \in A_1} T_{n,i} \leq \hat{c}_{n,A_1}(1-\alpha, k)$, then accept all hypotheses and stop; otherwise, reject any H_i for which $T_{n,i} \geq \hat{c}_{n,A_1}(1-\alpha)$ and continue.

2. Let R_2 be the indices i of hypothesis H_i previously rejected, and let A_2 be the remaining hypotheses. If $R_2 < k$, stop. Otherwise, let

$$\hat{d}_{n,A_2}(1-\alpha, k) = \max\{c_{n,K}(1-\alpha, k) :$$

$$K = A_2 \bigcup I,\ I \subset R_2,\ |I| = k-1\}\ .$$

   Then, reject any $T_{n,i}$ with $i \in A_2$ satisfying $T_{n,i} > \hat{d}_{n,A_2}(1-\alpha, k)$. If no further rejections, stop.

$\vdots$

j. Let R_j be the indices i of hypotheses previously rejected, and let A_j be the remaining hypotheses. Let

$$\hat{d}_{n,A_j}(1-\alpha,k) = \max\{c_{n,K}(1-\alpha,k) :$$

$$K = A_j \bigcup I,\ I \subset R_j,\ |I| = k-1\} \ .$$

Then, reject any $T_{n,i}$ with $i \in A_j$ satisfying $T_{n,i} > \hat{d}_{n,A_j}(1-\alpha,k)$. If no further rejections, stop.

$\vdots$

And so on.

Theorem 3 *Using above algorithm with critical values* $\hat{c}_{n,K}(1-\alpha,k)$ *satisfying whenever* $I(P) \subset K$

$$\hat{c}_{n,K}(1-\alpha,k) \geq \hat{c}_{n,I(P)}(1-\alpha,k) \ ,$$

$$k\text{-}FWE_P \leq P\{k\text{-}\max(T_{n,j} : \ j \in I(P)) >$$

$$\hat{c}_{n,I(P)}(1-\alpha,k)\}$$

So, if last expression $\leq \alpha$*, then* k-$FWE_P \leq \alpha$*.*

• Resampling methods can be used to always satisfy monotonicity requirement, and the last requirement typically holds at least asymptotically. Bootstrap consistency theorems ensue.

For those unfamiliar with the bootstrap, $c_{n,K}(1-\alpha, \hat{Q}_n)$ approximated by Monte Carlo:

For $b = 1, \ldots B$, let $X^*(b)$ be a sample drawn from $\hat{Q}_n$. Based on $X^*(b)$, compute estimates $\hat{\theta}^*_{n,i}(b)$. Let

$$m_b = \max_{i \in K} \tau_n |\hat{\theta}^*_{n,i}(b) - \hat{\theta}_{n,i}|$$

A $1-\alpha$ quantile of the empirical distribution of the B values $m_1, \ldots, m_B$ approximates $c_{n,K}(1-\alpha, \hat{Q}_n)$.

- Same set of resamples for any K.

If $k = 1$, at step j, no need to consider previously rejected hypotheses.

For $k > 1$, at step j, having made R_j rejections, one has to evaluate $\binom{R_j}{k-1}$ quantiles over which one maximizes.

Asymptotically, one need only consider the subset of $k - 1$ least significant hypotheses rejected.

Operative Method: Fix $N_{max} = 50$, say, and let M be the largest integer for which $\binom{M}{k-1} \leq N_{max}$. Consider at most the M most "recently" rejected hypotheses and maximize over subsets corresponding to those M hypotheses together with those not already rejected.

Eg 1 *[Testing Correlations]* Suppose $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$, so that $X_i = (X_{i,1}, \ldots, X_{i,d})$. Assume $E|X_{i,j}|^2 < \infty$ and $Var(X_{i,j}) > 0$.

$H_{i,j}$ specifies $\rho_{i,j} = 0$. Let $T_{n,i,j}$ = sample correlation between variables i and j.

The conditions for the bootstrap hold because correlations are smooth functions of means. ■

Eg 2 *[s-variate 2-sample Problem]* $Y_1, \cdots, Y_{n_Y}$ i.i.d. P_Y,

$Z_1, \cdots, Z_{n_Z}$ i.i.d. P_Z.

P_Y and P_Z are distributions on $\mathbf{R}^s$, with jth components denoted $P_{Y,j}$ and $P_{Z,j}$. Assume H_j implies $P_{Y,j} = P_{Z,j}$. Permutation tests apply and yield exact control.

Generalizes to:

- more general hypotheses
- other resampling schemes:

(i) permutations (can lead to finite sample control)

(ii) moving blocks bootstrap (for dependent data)

(iii) subsampling (under weakest conditions)

- Also applies if $s = \infty$ (applications to underidentified econometric models).

Simulations support

• good control of the k-FWE in finite samples

• increase in "power" over generalized Holm or methods based on marginal pvalues. For example, if $k = 1$, for s in the range 10–40, the stepdown method rejects between 20% and 50% more false hypotheses than Holm. Not surprisingly, increasing k rejects many more hypotheses.

Balance and Error Allocation May be desirable to have $P\{\text{reject } H_i\}$ independent of i for all true i.

This can be achieved by using studentized statistics, or p-values.

Of course, one can use the bootstrap to convert each $T_{n,i}$ into a p-value $\hat{p}_{n,i}$ and then apply the basic algorithm to $T'_{n,i} = -\hat{p}_{n,i}$. This would involve a double bootstrap. Actually, the bootstrap can be used to achieve balance automatically, by a generalization of the basic algorithm, and by doing only a single bootstrap.

Control of the FDP

Recall $F = \#$ false rejections, $R = \#$ rejections. Define the *false discovery proportion* (FDP) as F/R (defined as 0 if $R = 0$). Given a value γ, require

$$P\{FDP > \gamma\} \leq \alpha$$

Basic idea: At step i, having rejected $i - 1$ hypotheses, we want to guarantee $F/i \leq \gamma$, i.e. $F \leq \lfloor \gamma i \rfloor$, where $\lfloor x \rfloor$ is the greatest integer $\leq x$. So, if $k = \lfloor \gamma i \rfloor + 1$, then $F \geq k$ should have probability no greater than α; that is, we must control the number of false rejections to be $\leq k$.

Therefore, we use a stepdown procedure such that at step i, we apply a k-FWE controlling procedure, where

$$k = k(i, \gamma) = \lfloor \gamma i \rfloor + 1 \ .$$

eg. Apply generalized Bonferroni/Holm constants. Leads to a stepdown method based on marginal p-values with critical values $\alpha_i = \frac{(\lfloor \gamma i \rfloor + 1)\alpha}{s + \lfloor \gamma i \rfloor + 1 - i}$.

Theorem 4 *Under weak dependence assumptions, the stepdown method with these α_i controls the FDP.*

e.g. the family of distributions is positively dependent and is characterized by the multivariate positive of order two condition. (Sarkar, 1998)

By same reasoning, apply the bootstrap method to control the k-FWE at step i, where $k = k(i, \gamma) = \lfloor \gamma i \rfloor + 1$.

Simulation results and applications presented in Romano and Wolf (Annals 07) and Romano, Shaikh and Wolf (Econometric Theory 07).

Conclusion: Asymptotic Theory and Simulations support the value of methods which account for dependence based on weaker measures of error control.

For FDR control, go to Wolf's talk.

Caveats: Asymptotics, Increase in number of true rejections.