SIGNIFICANCE LEVELS FOR MULTIPLE TESTS[1]

by

SERGIU HART[2] AND BENJAMIN WEISS[3]

Abstract

Let $X_1, ..., X_n$ be n random variables, with cumulative distribution functions $F_1, ..., F_n$. Define $\xi_i := F_i(X_i)$ for all i, and let $\xi^{(1)} \leq ... \leq \xi^{(n)}$ be the order statistics of the $(\xi_i)_i$. Let $\alpha_1 \leq ... \leq \alpha_n$ be n numbers in the interval [0,1]. We show that the probability of the event $R := \{\xi^{(i)} \leq \alpha_i$ for all $1 \leq i \leq n\})$ is at most $\min_i \{n\alpha_i/i\}$. Moreover, this bound is exact: for any given n marginal distributions $(F_i)_i$, there exists a joint distribution with these marginals such that the probability of R is exactly $\min_i \{n\alpha_i/i\}$. This result is used in analyzing the significance level of multiple hypotheses testing. In particular, it implies that the Rüger tests dominate all tests with rejection regions of type R as above.

---

[1]March 1996. Most of these results are from 1992. We thank Robert J. Aumann, Yoav Benjamini, Gerhard Hommel, Yossef Hochberg and Ester Samuel-Cahn, for useful comments.

[2]Department of Mathematics; Department of Economics; and Center for Rationality and Interactive Decision Theory, The Hebrew University of Jerusalem, Feldman building, Givat-Ram, 91904 Jerusalem, Israel.
E-mail: hart@math.huji.ac.il.

[3]Department of Mathematics, The Hebrew University of Jerusalem, Givat-Ram, 91904 Jerusalem, Israel.
E-mail: weiss@math.huji.ac.il.

SIGNIFICANCE LEVELS FOR MULTIPLE TESTS

by

SERGIU HART AND BENJAMIN WEISS

Let X_1, X_2, ..., X_n be n random variables, with cumulative distribution functions F_1, F_2, ..., F_n. Define $\xi_i := F_i(X_i)$ for all i, and let $\xi^{(1)} \le \xi^{(2)} \le ... \le \xi^{(n)}$ be the order statistics of the $(\xi_i)_i$. Let $\alpha_1 \le \alpha_2 \le ... \le \alpha_n$ be n numbers in the interval [0,1], and consider the event

$$R := \{\xi^{(i)} \le \alpha_i \text{ for all } 1 \le i \le n\}. \tag{1}$$

For n = 2, this event is depicted in Figure 1 (as the corresponding dark area). The problem is to evaluate the probability of R, for arbitrary joint distribution of the X_i's.

Events of this kind arise in "multiple hypotheses testing" (see, e.g., Bonferroni [1936], Rüger [1978], Hommel [1983], Hochberg and Hommel [1996]). The setup is as follows: There are n statistical tests, with null hypotheses H_{0i} and alternatives H_{1i} (for i = 1, ..., n), respectively. The test statistics are the random variables X_1, ..., X_n. The tests are applied to the same data, but the dependence between them is not known. Let $H_0 := \cap_{i=1,...,n} H_{0i}$ be the null hypothesis that is the conjunction of the n null hypotheses. The problem is to construct an "overall test" for H_0 that combines the n tests, i.e., it is based on the rejection regions of the n individual tests. Since H_0 is not a simple hypothesis, the significance level α of a test with rejection region R is defined as the worst-case probability of type I error; i.e., $\alpha(R) := \sup_{\theta\in\Theta_0} P_\theta(R)$, where Θ_0 is the set of all probability distributions that are consistent with H_0. In the setup above, where H_0 is the intersection of the H_{0i}'s, the set Θ_0 contains all the joint distributions with marginals $F_1, ..., F_n$.

A rejection region R of type (1) means that H_0 is rejected if the n test statistics are more significant than $\alpha_1, \alpha_2, ..., \alpha_n$, regardless of which test has passed which significance level.

Assume that all the α_j's are possible values of each F_i (this clearly holds for continuous distributions, and, if there are atoms, assume they are "split" accordingly; recall that these random variables represent statistical tests); thus $P(\xi_i \le \alpha_j) = \alpha_j$ for all i, j.

Theorem: $P(\{\xi^{(i)} \le \alpha_i \text{ for all } 1 \le i \le n\}) \le \min_{1\le i\le n} \frac{n\,\alpha_i}{i}$. Moreover, this bound is exact: for any n marginal distributions $(F_i)_{i=1,...,n}$, there exists a joint distribution P with these marginals such that the inequality becomes an equality.

In terms of the multiple test procedure discussed above, the result is that the significance level of R as defined in (1) is $\alpha(R) = \min_i \{n\alpha_i/i\}$. Equivalently, the p-value p of a test of type (1) may be expressed as a function of the p-values $p_1, p_2, ..., p_n$ of the n individual tests, namely $p = \min_i \{np_i/i\}$.

We come now to an interesting Corollary of the Theorem. Consider two tests with rejection regions R^1 and R^2, respectively. We say that R^1 is (weakly) *dominated* by R^2 if $\alpha(R^1) = \alpha(R^2)$ and $R^1 \subset R^2$; thus the two tests have the same significance level, but the probability of type II error is always at least as high for R^1 than for R^2. For each i = 1, ..., n and each α in (0,1), define $R_{i,\alpha} := \{\xi^{(i)} \leq i\alpha/n\}$; i.e., at least i among $\xi_1, ..., \xi_n$ are $\leq i\alpha/n$. Note that $R_{i,\alpha}$ obtains by taking $\alpha_j = i\alpha/n$ for $j \leq i$ and $\alpha_j = 1$ for $j > i$ in (1). The $R_{i,\alpha}$ are the Rüger [1978] tests.

> **Corollary:** Let R be the rejection region of a test of type (1), with significance level $\alpha = \alpha(R)$. Then there exists an integer i, $1 \leq i \leq n$, such that R is dominated by $R_{i,\alpha}$.

Proof: Let i be such that the minimum in the formula for $\alpha(R)$ is attained at i, i.e., $\alpha = \alpha(R) = n\alpha_i/i$; thus $\alpha_i = i\alpha/n$. By the result of the

Theorem, one can increase α_j for all $j \neq i$ without increasing the significance level; the largest set is thus obtained by taking[4] $\alpha_j = \alpha_i = i\alpha/n$ for $j < i$ and $\alpha_j = 1$ for $j > i$ -- which yields precisely $R_{i,\alpha}$. ▮

Thus, for each α there are exactly[5] n tests of type (1) to consider, namely the Rüger [1978] tests $R_{i,\alpha}$. Note that $R_{1,\alpha}$ is the Bonferroni [1936] test -- reject if at least one of the ξ_i is $\leq \alpha/n$ -- and that the Simes [1986] test is the union $\cup_{i=1,...,n} R_{i,\alpha}$ of all the[6] $R_{i,\alpha}$.

We come now to the proof of the Theorem. A few notations first: Put $I := \{1, 2, ..., n\}$ and $J := \{1, 2, ..., n+1\}$. It will be convenient to define $\alpha_0 := 0$ and $\alpha_{n+1} := 1$, and $\delta_i := \alpha_i - \alpha_{i-1}$ for all $i \in J$. Partition [0,1] into n+1 disjoint intervals $D_1 := [0,\alpha_1]$, $D_2 := (\alpha_1,\alpha_2]$, ..., $D_j := (\alpha_{j-1},\alpha_j]$, ..., $D_{n+1} := (\alpha_n,1]$. Define random variables $Y_1, Y_2, ..., Y_n$ by $Y_i = j$ whenever $\xi_i \in D_j$, and put $Y := (Y_i)_{i\in I}$. Note that the event R is actually defined in terms of Y.

Proof of Theorem: For each vector $y = (y_i)_{i\in I}$ in J^I, write p(y) for the probability $P(Y = y)$. Denote by Q the set of all values of Y corresponding to the event R; i.e., y belongs to Q if, for each $j = 1, 2, ..., n$ (note:

---

[4]Recall that $0 \leq \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_n \leq 1$.

[5]It is clear that the $R_{i,\alpha}$ for different i do not dominate one another.

[6]The significance level of the Simes test is α (1 + 1/2 + ... + 1/n), as shown by Hommel [1983].

$j \le n$), there are at least j indices i such that $y_i \le j$. Thus the upper bound of Theorem 1 is obtained from the following Linear Programming problem

$$\text{maximize} \sum_{y \in Q} p(y), \text{ subject to:}$$

$$\sum_{y \text{ s.t. } y_i = j} p(y) = \delta_j \text{ for all } i \in I \text{ and } j \in J, \tag{LP}$$

$$p(y) \ge 0 \text{ for all } y.$$

The dual of (LP) is

$$\text{minimize} \sum_{i \in I} \sum_{j \in J} \delta_j \, \lambda_{i,j}, \text{ subject to:}$$

$$\sum_{i \in I} \lambda_{i,y_i} \ge 1 \text{ for all } y \in Q, \tag{DLP}$$

$$\sum_{i \in I} \lambda_{i,y_i} \ge 0 \text{ for all } y \notin Q.$$

We will show that the optimal value of (DLP) -- and thus also of (LP) -- is indeed $\min_i \{n\alpha_i/i\}$. First, note that if $\lambda = (\lambda_{i,j})_{i \in I, j \in J}$ is feasible in (DLP), then, for any permutation π of I, $\lambda' := (\lambda_{\pi(i),j})_{i \in I, j \in J}$ is also feasible in (DLP). Moreover, the value of the objective function is the same for λ' as for λ. Averaging over all permutations π of I implies that we can assume without loss of generality that λ is "symmetric", i.e. $\lambda_{i,j} = \mu_j$ for all $i \in I$ and $j \in J$. Thus (DLP) becomes

$$\text{minimize} \quad n \sum_{j \in J} \delta_j \, \mu_j, \text{ subject to:}$$

$$\sum_{i \in I} \mu_{y_i} \geq 1 \text{ for all } y \in Q, \qquad \text{(DLP}'\text{)}$$

$$\sum_{i \in I} \mu_{y_i} \geq 0 \text{ for all } y \notin Q.$$

Let $\mu = (\mu_j)_{j \in J}$ be an optimal point for (DLP′); recall that all the δ_j are non-negative. We have:

(1) $\mu_1 \geq 1/n$ (the vector $y = (1, 1, ..., 1)$ belongs to Q).

(2) $\mu_j \geq 0$ for all $j \in J$ (the vector $y = (j, j, ..., j)$ does not belong to Q when $j \geq 2$).

(3) Without loss of generality[7], $\mu_{n+1} = 0$ (indeed, μ_{n+1} does not appear in the constraints for $y \in Q$; by (2), decreasing μ_{n+1} to 0 does not violate the constraints for $y \notin Q$, and can only decrease the value of the objective function).

(4) Without loss of generality, $\mu_j \geq \mu_{j+1}$ for all $j \in J$, $j \leq n$ (let j be the highest index such that $\mu_j < \mu_{j+1}$; put $\mu'_{j+1} := \mu_j$ and $\mu'_k := \mu_k$ for $k \neq j$, then μ' is also feasible -- if y belongs to Q and, for some i, we have $y_i =$

---

[7]If $\delta_{n+1} > 0$ then necessarily $\mu_{n+1} = 0$; if $\delta_{n+1} = 0$, then μ_{n+1} may be taken to be 0. Similarly for the following statements.

j+1, then y' belongs to Q too, where $y'_i := j$ and $y'_k := y_k$ for $k \neq i$ -- and, again, the value of the objective function could have only decreased).

(5) Without loss of generality, $\Sigma_{j\in J}\ \mu_j = \Sigma_{j=1}^{n}\ \mu_j = 1$ (the vector y = (1, 2, ..., n) belongs to Q, therefore $\Sigma_{j=1}^{n}\ \mu_j \geq 1$; by (4), this inequality implies that the constraints for all y in Q are satisfied; if the inequality were strict, we could therefore decrease some of the μ_j's, remain feasible and decrease the objective function).

(6) Without loss of generality, μ is a convex combination of the n vectors $\{m^i\}_{i\in I}$, defined by $m^i := (\underbrace{1/i,\ 1/i,\ ...,\ 1/i}_{i},\ \underbrace{0,\ 0,\ ...,\ 0}_{n+1-i})$

(indeed, $\mu = \Sigma_{i\in I}\ i(\mu_i - \mu_{i+1})\ m^i$).

Therefore the minimal value of (DLP′) is obtained as the minimum over the points $\{m^i\}_{i\in I}$, hence it equals $\min_{i\in I}\ \{n\ \Sigma_{j\leq i}\ \delta_j/i\} = \min_{i\in I}\ \{n\alpha_i/i\}$, as claimed.

To complete the proof, one needs to exhibit, for any given marginals $(F_i)_{i\in I}$, a joint distribution where the bound is attained. This is done as follows: Let $p^* = (p^*(y))_y$ be an optimal point for the primal problem (LP). This determines the distribution of Y, by $P(Y = y) := p^*(y)$ for all y. Define the X_i's to be conditionally independent given Y, specifically $P(X_i \leq x_i$ for all $i \mid Y = y) := \prod_{i\in I} P(X_i \leq x_i$ and $\xi_i \in D_{y_i}) / \delta_{y_i}$ for[8] all y. It can now

---

[8]Note that the i'th term in the product equals $(F_i(x_i)-\alpha_{y_i-1})/(\alpha_{y_i}-\alpha_{y_i-1})$ if $F_i(x_i) \in D_{y_i}$, it is 0 if $F_i(x_i) \leq \alpha_{y_i-1}$ and 1 if $F_i(x_i) \geq \alpha_{y_i}$.

be checked that, indeed, $P(\xi_i \in D_{y_i}$ for all i) = p*(y) for all y, and that the marginal of each X_i is precisely F_i. ▮

A few remarks are now in order.

Remarks: (i) The inequality $P(R) \le \min_i \{n\alpha_i/i\}$ follows directly from $R \subset R_{i,\alpha_i}$ and $P(R_{i,\beta}) \le n\beta/i$ for all β and all i. The probabilistic inequality is proved as follows: Let $B_j := \{\xi_j \le \beta\}$ and $E^k :=$ {exactly k of the $\{B_j\}_{j\in I}$}, then

$$\sum_{j\in I} \mathbf{1}_{B_j} = \sum_{k=1}^{n} k\ \mathbf{1}_{E^k} \ge \sum_{k=i}^{n} k\ \mathbf{1}_{E^k} \ge \sum_{k=i}^{n} i\ \mathbf{1}_{E^k} = i\ \mathbf{1}_{R_{i,\beta}}.$$

Here $\mathbf{1}_C$ denotes the indicator function of the event C; the first equality is due to the fact that each point in E^k is counted exactly k times in the left-most sum. Therefore

$$i\ P(R_{i,\beta}) \le \sum_{j\in I} P(B_j) = n\ \beta.$$

(ii) The linear programming approach is needed to prove the exactness of the bound for arbitrary marginals $(F_i)_i$. If all the marginal distributions are continuous, thus all the ξ_i are uniformly distributed on [0,1], a joint distribution P with $P(R) = \min_i \{n\alpha_i/i\}$ may be explicitly constructed as follows: Put $\gamma := \min_i \{\alpha_i/i\}$. Let Δ_0 be a one-dimensional "diagonal" in $[0,1]^n$ of length $1\text{-}n\gamma$:

$$\Delta_0 := \{x = (x_i)_{i\in I} \in [0,1]^n \mid x_i = n\gamma + t \text{ for all } i \in I,\ t \in [0,1\text{-}n\gamma)\}.$$

For each permutation π of I, let $\Delta(\pi)$ be a diagonal of length γ:

$$\Delta(\pi) := \{x = (x_i)_{i\in I} \in [0,1]^n \mid x_i = (\pi(i)\text{-}1)\gamma + t \text{ for all } i \in I,\ t \in [0,\gamma)\}.$$

We then put probability mass $1\text{-}n\gamma$ on Δ_0 and, for every one of the n! permutations π, mass $\gamma/(n\text{-}1)!$ on $\Delta(\pi)$; the mass is uniformly spread over each one of these diagonals. It is now easy to check that the resulting probability distribution P on $[0,1]^n$ has uniform marginals, and that P(R) = $P(\cup_\pi \Delta(\pi)) = n\gamma$.

This construction however breaks down if there are atoms (recall that we have assumed that the possible values of each ξ_i only include the α_j -- but not necessarily the multiples of γ used above).

(iii) One may consider other types of rejection regions. For instance, $R' := \{\xi^{(i)} \le \alpha_i$ for *some* $i\}$. Such tests have been considered by Hommel [1983], Simes [1986], Röhmel and Streitberg [1987], Hochberg [1988], and others. The significance level is $\alpha(R') = \min\ \{\sum_{i=1}^n n(\alpha_i\text{-}\alpha_{i\text{-}1})/i,\ 1\}$ -- see Hommel [1983]. One may again use the linear programming approach to obtain this bound. However, for such R′ the direct proof (parallel to (i) and (ii) above) works. Still, if additional linear constraints were added (which express restrictions on the set of distributions in H_0), then the linear programming approach may be useful.

(iv) Our result implies a *lower bound* for the probability of the rejection region R′ (cf. (iii) above). Indeed, R for $(-X_i)_i$ is essentially[9] the complementary event of R′ for $(X_i)_i$. Therefore,

$$P(\{\xi^{(i)} \le \alpha_i \text{ for some } i\}) \ge 1 - \min_{1\le i\le n} \frac{n\,(1 - \alpha_{n+1-i})}{i}$$
$$= \max_{1\le i\le n} \frac{n\,\alpha_{n+1-i} - n + i}{i}\,.$$

In the case of Simes' test, this can be shown to simplify to

$$P(\{\xi^{(i)} \le i\alpha/n \text{ for some } i\} \ge \max\,\{\, \frac{\alpha}{n},\ 1 - n(1-\alpha)\,\}.$$

Similarly, the result of Hommel [1983] yields a lower bound for the region R in (1), namely,

$$P(\{\xi^{(i)} \le \alpha_i \text{ for all } i\}) \ge \max\,\{1 - n \sum_{i=1}^{n} \frac{\alpha_{n+2-i} - \alpha_{n+1-i}}{i}\ ,\ 0\}.$$

---

[9]For i = 1, ..., n+1, define the intervals $D'_i := D_{n+2-i}$; thus $\delta'_i = \delta_{n+2-i}$ and $\alpha'_i = 1 - \alpha_{n+1-i}$.

One possible use for these lower bounds is to evaluate the power of the tests. However, this requires the alternative H_1 to be of a special form, namely, that all the X_i's are identically distributed under H_1 (with an arbitrary joint distribution; actually, it suffices to require that, for each θ in H_1, all the probabilities $P_\theta(X_i \le \alpha_j)$ do not depend on i).

References

Bonferroni, C. E. [1936], Theoria Statistica Classi e Calcolo delle Probabilita, *Pubbl. R. Int. Super. Sci. Econ. Comm. Firenze* 8, 1-62.

Hochberg, Y. [1988], A Sharper Bonferroni Procedure for Multiple Tests of Significance, *Biometrika* 75, 800-802.

Hochberg, Y. and Hommel, G. [1996], Simes' Test and Generalizations, *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson (eds.), Wiley (forthcoming).

Hommel, G. [1983], Tests of the Overall Hypothesis for Arbitrary Dependence Structures, *Biom. J.* 25, 423-430.

Röhmel, J. and B. Streitberg [1987], Zur Konstruktion globaler Tests, *EDV in Medizin und Biologie* 18, 7-11.

Rüger, B. [1978], Das Maximale Signifikanzniveau des Tests "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen", *Metrika* 25, 171-178.

Simes, R. J. [1986], An Improved Bonferroni Procedure for Multiple Tests of Significance, *Biometrika* 73, 751-754.