SOME REMARKS ON SIMES TYPE MULTIPLE TESTS OF SIGNIFICANCE

by

Pranab K. Sen

Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2177

Oct 1997

Some Remarks on Simes type multiple tests of significance

PRANAB K. SEN
Departments of Biostatistics and Statistics,
University of North Carolina, Chapel Hill, NC 27599-7400
e-mail: pksen@bios.unc.edu

Abstract

Simes' theorem, posed in a multiple testing setup, has its genesis in the classical ballot theorem in uniform order statistics. The purpose of this note is to comment on this rediscovery of the Ballot theorem in multiple testing theory, and with a view to enhance scope of applications, to stress its ramifications in the same context. This privides sharper error rates.

AMS Subject Classifications: 62J15, 62E15.

Keywords: Ballot Therorem; Bonferroni procedure; discrete distributions, step-down procedures, uniform order statistics.

1 Introduction

Let a null hypothesis H_0 be expressible as the intersection of $k(\geq 2)$ components $H_{01}, \ldots, H_{0k}$, whose respestive alternatives are denoted by $H_{11}, \ldots, H_{1k}$; then the alternative hypothesis to the null one is H_1 that can be expressed as the uninon of the components $H_{1j}, j = 1, \ldots, k$. Let $T_1, \ldots, T_k$ be the test statistics for testing these component (null vs. alternative) hypotheses, and let their *observed significance level* (OSL) or the so called p-values be denoted by $P_1, \ldots, P_k$ respectively. We denote the ordered p-values by $P_{(1)}, \ldots, P_{(k)}$ respectively. Then Simes' (1986) theorem, quite popular in multiple testing theory, can be stated as follows.

Theorem 1. If the T_j are independent, then under H_0, for every $\alpha \in (0, 1)$,

$$P\{P_{(j)} > j\alpha/k, \text{ for all } j = 1, \ldots, k\} = 1 - \alpha. \tag{1.1}$$

Simes (1986) has an elegant proof bsaed on the method of induction. Basically, Simes' theorem is a reinstatement of the classical ballot theorem related to uniform order statistics, as may be found in the texts by Takács (1967) or Karlin (1969). We shall make this point clear in the next section What is more in this context is the intricate connection between the two setups that generates additional multiple testing procedures some of which have already been considered by others [viz., Hochberg 1988, Hommel, 1988, Hochberg and Rom, 1995, Benjamini and Hochberg, 1996 and Samuel-Cahn 1996, among others). In Section 3, we will add some further feathers to this hat by incorporating

some exact results in this setup. Finally, in Section 4, along with the case of discrete distributions for p-values, the case of dependent tests is discussed with due emphasis on the step-down procedures.

2 The ballot theorems

Let $U_1, \ldots, U_k$ be k independent and identically distributed (i.i.d.) random variables (r.v.) having the uniform (0, 1) distribution, and let $U_{(1)} < \cdots < U_{(k)}$ be the associated ordered r.v.'s. Let $G_k(t) = k^{-1} \sum_{i=1}^{k} I(U_1 \leq t)$, $t \in (0,1)$ be the empirical distribution function (d.f.). Then note that $G_k(t) = i/k$, for $U_{(i)} \leq t < U_{(i+1)}$, $i = 0, 1, \ldots, k$, where conventionally, $U_0 = 0$, $U_{(k+1)} = 1$. The ballot theorem can then be stated as follows.

Theorem 2. For every $\gamma \geq 1$, and every $k \geq 1$,

$$P\{G_k(t) \ < \ \gamma t, \ \forall t \ \in (0,1)\} = 1 - \gamma^{-1}. \tag{2.1}$$

We refer to Karlin (1969, p.251) for an elegant proof. It is easy to verify that

$$G_k(t) < \gamma t, \forall t \in (0,1) \ \Leftrightarrow \ U_{(i)} > i/(k\gamma), \ i = 1, \ldots, k. \tag{2.2}$$

Therefore, the same probability holds for (2.2).

Let us now compare Theorems 1 and 2. Note that under the null hypothesis, the p-values $P_1, \ldots, P_k$ are i.i.d.r.v.'s with the uniform (0,1) distribution, so that the $P_{(j)}$ correspond to the uniform order statistics $U_{(j)}$. Hence, letting $\gamma = \alpha^{-1}$, we readily obtain that (1.1) and (2.1) are equivalent.

The Simes procedure provides only a test for the overall null hypothesis. Hommel (1988) considered a stagewise rejective multiple testing procedure, while Hochberg (1988) considered an extended Simes procedure wherein he compared $P_{(i)}$ against $\alpha/(k - i + 1)$, $i = 1, \ldots, k$, and showed that works out well in the multiple hypothesis testing setup. Yochberg and Rom (1995) have some further results in this direction with emphasis on *logically related hypotheses* (LRH). Keeping these in mind, we consider the following extension of the ballot theorem, where the order statistics $U_{(i)}$ are defined as in before.

Theorem 3. For every $0 < a_1 < \cdots < a_k < 1$; $k \geq 1$,

$$P\{\mathbf{a}\} \ = \ P\{ \ U_{(j)} \geq a_j, \ j = 1, \ldots, k \ \} \ = \ k! H_{kk}(1), \tag{2.3}$$

where for each $j (= 1, ..., k)$, and $u \in (a_j, 1)$,

$$H_{kj}(u) = \int_{a_j}^{u} \int_{a_{j-1}}^{u_j} \cdots \int_{a_1}^{u_2} du_1 \cdots du_j, \tag{2.4}$$

and

$$H_{kj}(u) = \int_{a_j}^{u} H_{k(j-1)}(u_j) du_j, \ j = 2, \ldots k; \ \ H_{k0}(u) = I(u \geq a_1). \tag{2.5}$$

Proof. Note that the joint density of $U_{(j)}$, $j = 1, \ldots, k$ is given by

$$k! \ I(0 < u_1 < \cdots < u_k < 1), \tag{2.6}$$

where $\{u_j \geq a_j, j = 1, \ldots, k\}$ is a subset of the cone $\{0 < u_1 < \cdots < u_k < 1\}$. Hence, the rest of the proof follows by some standard arguments.

In passing we may remark that the a_j may generally depend on k, and hence the H_{kj}, even though dependent only on $a_r, r \leq j$, may depend on k. These recursive relations are helpful for actual numerical evaluation of $P\{\mathbf{a}\}$ in specific situations that would be considered in the next section.

3 Multiple comparisons

We consider some specific values of k. For $k = 1$, we note that by definition $P\{a_1\} = 1 - a_1 = H_{11}(1)$, so that $a_1 = \alpha$, as it should be. For $k = 2$, we have

$$\begin{aligned} P\{a_1, a_2\} &= 2H_{22}(1) = 2\int_{a_2}^{1} H_{21}(u)du \\ &= 2((1-a_1)(1-a_2) - \frac{1}{2}(1-a_2)^2) = (1-a_2)(1-2a_1+a_2) \end{aligned} \tag{3.1}$$

so that $a_1 = \frac{1}{2}a_2 = \alpha/2$ leads to an exact $1-\alpha$ probability for $P\{a_1, a_2\}$. This is in agreement with Simes (1986) as well as Hochberg (1988). The attainment of the exact level $1-\alpha$ may also be accomplished with some other combinations of a_1, a_2 for which the right hand side of (3.1) is equal to $1-\alpha$. For example, allowing a_1 to approach a_2 we get the Bonferroni solution $a_2 = 1 - \sqrt{(1-\alpha)}$, and allowing a_1 to be arbirary close to 0, we have a_2 larger than α. In this way, a variety of multiple testing procedures can be considered along the same line as in Hochberg (1988). Consider next the case of $k = 3$. Using the recursion relation, we obtain that

$$\begin{aligned} P\{a_1, a_2, a_3\} &= 6H_{33}(1) = 6(1-a_1)(1-a_2)(1-a_3) \\ &\quad -3(1-a_2)^2(1-a_3) - 3(1-a_1)(1-a_3)^2 + (1-a_3)^3. \end{aligned} \tag{3.2}$$

If as in Hochberg (1988), we let $a = a_3 = 2a_2 = 3a_1$, then the above expression simplifies to $(1-a)(1+\frac{1}{4}a^2)$ that is strictly greater than $1-a$, so that choosing $a = \alpha$ the overall significance level is given by $1-(1-\alpha)(1+\frac{1}{4}\alpha^2)$ that is strictly smaller than α. Consequently, solving for $(1-a)(1+\frac{1}{4}a^2) = 1-\alpha$, we can achieve an exact significance level α. It also shows that there are other combinations of the a_j that lead to the attainment of the exact level $1-\alpha$ for $P\{a_1, a_2, a_3\}$. The recursive relations can be used successively to consider the case of $k \geq 3$, and instead of the Hochberg (1988) solution, a slightly modified one can be obtained that has an exact size. This way, Simes procedure can be adapted better to multiple hypothesis testing problems. In general, for any $k \geq 2$, if we relate the $a_j = a \cdot m(j,k)$, $j = 1, \ldots, k$ where the $m(j,k)$ are preassumed nonnegative constants (and nondecreasing in j), then we may note that $H_{kk}(1)$ is a polynomial of degree k in a with coefficients that are functions of the preassumed $m(j,k)$'s. Therefore, we are at liberty to choose a *spending function* $m(j,k), j \leq k$ that is appropriate in a particular context and leads to an attainment of the exact significance level α. Of course, such a spending function needs to be so chosen that the *closure principle* referred to in Hochberg (1988) and Hochberg and Rom (1995) holds for the subset hypothesis testing problem. The latter authors considered the LRH family (Shaffer, 1986), and our findings remain pertinent for that as well. There is, however, a basic theoretical query: Can some of these multiple testing procedures be justified on the ground of admissibility or minimaxity? As of now, the major emphasis has been laid down on the overall significance level and

distribution of P_j, though concentrated on the unit interval (0, 1), has jump-discontinuities and jump-magnitudes that depend on the underlying null distributions. Therefore, the P_j are nolonger distribution-free. In this case, for $k = 1$, the value of a_1 may not be exactly equal to α, and moreover, for a given $\alpha(0 < \alpha < 1)$, there may not be an exact solution for $P\{a\} = 1 - \alpha$. The situation may depend on whether we define the P_j by including or excluding the observed points of the test statistics; as they have discrete null distributions, these two values may not generally agree. The inclusion may result in an OSL level larger while the exclusion may lead to a smaller than the value as obtained in the continuous case. A randomization test can be used to achieve this goal, but then the tie probability explicitly enters the picture, and the Ballot theorems may no longer be tenable. Whether this results in a conservativeness or not depends on the underlying null distributions, and hence, no blanket prescription can be made to suit all cases. Moreover, for different component hypotheses, the null hypothesis distributions might not be the same, and hence, their impact on the distributions of the P_j may vary from one to other components. We refer to Samuel-Cahn (1996) for some discussion on conservativeness aspects of the Simes test, and that applies to others as well.

Let us make some comments on the dependent case where the test statistics for the component hypotheses are not stochastically independent of each other (even under the null hypothesis). Again the dependence may be either of the *positive association* type or *negative association* type, and in the case of several component hypotheses, there could be a mixed-type dependence too. Thus, the effect of dependence on the Simes-Hochberg type multiple tests may depend on the underlying dependence pattern. If the P_j are not stochastically independent, the Simes theorem or the parent ballot theorems may no longer be true. As a simple illustration consider the case where the P_i are exchangeable (i.e., symmetric dependent) r.v.'s, each having marginally the uniform (0, 1) distribution. If the dependence is of negative type the result will be different from the independence case as well as the case of positive dependfence. This can be studied as in Samuel-Cahn (1996) where the bivariate normal case has been dealt with in detail.

In multiple tests of significance, the *step-down procedure*, introduced by J. Roy (1958) for multivariate analysis and popularized by the pioneering work of late S. N. Roy and his associates (Roy et al., 1971), occupies a prominent place. The step-down procedure relates to a finite *union-intersection* principle wherein there are essentially k, a fixed positive integer, component hypotheses that are to be tested in an order of priority or importance. The main feature of this procedure is that the OSL values as obtained in the subsequent steps are quasi-independent under the null hypothesis, though the usual test statistics for the component hypotheses are generally dependent when viewed simultaneously. This finite UI-structure results in a sharper error rate for the component hypotheses, though in terms of consistency, the infinite UI-tests generally perform better. This feature was incorporated by Sen (1983) in proposing a Fisherian detour of the step-down procedure based on the p-values for the subsequent steps. It follows from the same argument that the ballot theorems (2 and 3) apply to these p-values as well, and hence, the Simes-Hochberg multiple tests of significance apply also to the step-down procedures. For some allied results on multiple testing based on OSL values, we refer to Sen (1988); the findings of the current study also pertain to those setups.

family-wise error rates for subfamilies of hypotheses, and when it comes to the question of power properties, the alternative hypotheses may be too broad to allow a simple resolution.

In many problems of practical interest, we encounter multiple tests of significance where k, the number of components, may not be very small. For example, in a clinical trial involving the group sequential methodology, one may have a value of k as large as 10. Although Theorem 3 can still be used in such a context with recursive computations, there remains the open question: Is there any optimality or desirability property of the Simes-Hochberg tests in such a context? As has been discussed in Sen (1998) that in a multiparameter setting, even in most simple parametric models, a universally optimal multiple testing procedure may not generally exit. The choice of a multiple testing procedure is largely guided by practical considerations and robustness cum validity perspectives. In that way the flexibility of choice of the a_j, granted by Theorem 3 offers a variety of OSL based tests. In this context, the Fisher method of combinining independent tests has some asymptotic optimality properties; though in finite sample cases and for broader families of alternatives, this asymptotic optimality criterion might not be very appealing. If a group-sequential plan is adopted, the stopping number relating to the stopping of the trial at an intermediate stage has a distribution that might depend on the null hypothesis being true or not. Hence, in judging the merits of a testing scheme, these factors are to be taken into account. Guided by robustness considerations, for large k, it might be appealing to choose a comparatively smaller value r, and let

$$a_1 < \cdots < a_r = a_{r+1} = \cdots = a_k. \tag{3.3}$$

Thus, effectively more importance is attached to the largest r p-values. The expression for $P\{\mathbf{a}\}$ in Theorem 3 then reduces to

$$P\{\mathbf{a}\} = k^{[r]} \int_{a_r}^{1} H_{k(r-1)}(u)(1-u)^{k-r} du, \tag{3.4}$$

where $k^{[r]} = k(k-1)\cdots(k-r+1)$, $r \geq 1$, and $k^{[0]} = 1$. When we do not want to attach too much importance to just a few components, we may choose the opposite way, namely

$$a_1 = \cdots = a_r < a_{r+1} < \cdots < a_k, \text{ for some } r < k. \tag{3.5}$$

This way we can make use of Theorem 3 in a variety of models of practical importance. For large values of k, the weak convergence of the empirical distributional process or equivalently the uniform quantile process to a Brownian bridge can be most convenioently used to provide some simple approximations; we refer to Sen (1998) for some allied results. In this specific situation at hand, since we are dealing with uniform order statistics, a value of k as low as 16 would justify such asymptotic approximations.

4 Discrete and dependent cases

There has been lot of discussions on the impact of discreteness of the underlying null hypothesis distributions of the test statistics as well as on the case where the component test statistics are not necessarily independent. In the discrete case, the P_j may not have the continuous uniform (0,1) distribution (under the null hypothesis), and as has been noted by Schmid (1958), the null hypothesis

References

Benjamini, Y., and Hochberg, Y. (1996). More on Simes' test. Preprint.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800-802.

Hochberg, Y., and Rom, D. R. (1995). Extensions of multiple testing procedures based on Simes' test. *J. Statist. Plan. Infer. 48*, 141-152.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*, 383-386.

Karlin, S. (1969). *A First Course in Stochastic Processes.* Academic Press, New York.

Roy, J. (1958). Step-down procedures in multivariate analysis. *Ann. Math. Statist. 29*, 1177-1188.

Roy, S. N., Gnanadesikan, R., and Srivastava, J. N. (1971). *Analysis and Design of Certain Quantitative Multiresponse Experiments.* Pergamon Press, New York.

Samuel-Cahn, E. (1996). Is the Simes' improved Bonferroni procedure conservative? *Biometrika, 83*, 928-933.

Schmid, P. (1958). On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions. *Ann. Math. Statist. 29*, 1011-1027.

Sen, P. K. (1983). A Fisherian detour of the step-down procedure. In *Contributions to Statistics: Essays in Honour of Norman L. Johnson* (ed. P. K. Sen), North Holland, Amsterdam, pp. 367-377.

Sen, P. K. (1988). Combination of statistical tests for multivariate hypotheses against restricted alternatives. In *Advasnces in Multivariate Statistical Analysis* (ed. S. Dasgupta and J. K. Ghosh), Indian Statistical Institute, Calcutta, pp. 377-402.

Sen, P. K. (1998). Multiple comparisons in interim analysis. Preprint

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc. 81*, 826-831.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751-754.

Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes.* John Wiley, New York.