

Testing hypotheses in order

BY PAUL R. ROSENBAUM

*Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia,
Pennsylvania 19104-6340, U.S.A.*

rosenbaum@stat.wharton.upenn.edu

SUMMARY

In certain circumstances, one wishes to test one hypothesis only if certain other hypotheses have been rejected. This ordering of hypotheses simplifies the task of controlling the probability of rejecting any true hypothesis. In an example from an observational study, a treated group is shown to be further from both of two control groups than the two control groups are from each other.

Some key words: Multiparameter hypothesis; Multiple control groups; Observational study; Ordered family of hypotheses.

1. TWO METHODS FOR TESTING HYPOTHESES IN ORDER

Empirical investigations often test several hypotheses and need to take precautions to avoid frequent rejection of true hypotheses. One approach emphasizes setting a more stringent standard when more than one hypothesis is tested (Holm, 1979; Benjamini & Hochberg, 1995). A second approach emphasizes testing hypotheses in order of priority, exploiting logical relationships among hypotheses (Marcus et al., 1976; Bauer & Kieser, 1996; Hsu & Berger, 1999).

Let \mathcal{T} be a totally ordered set with strict inequality $<$ and equality-or-inequality \leq . Three such sets are (1) $\mathcal{T} = \{1, 2, \dots, k\}$ with \leq , (2) $\mathcal{T} = [0, \infty)$ with \leq and (3) the direct product, $\mathcal{T} = [0, \infty) \times \{1, 2, \dots, k\}$ with the lexicographical order, $(t_1, t_2) \leq (u_1, u_2)$ if either (i) $t_1 < u_1$ or (ii) $t_1 = u_1$ and $t_2 \leq u_2$.

Let $H_t, t \in \mathcal{T}$, be a collection of hypotheses. For each hypothesis $H_t, t \in \mathcal{T}$, there is a level- α test, yielding significance level p_t , so that $\text{pr}(p_t \leq \alpha) \leq \alpha$ if H_t is true. Only one assumption is made about the hypotheses, the ‘structure assumption,’ which says that either all $H_t, t \in \mathcal{T}$, are false or there is a first true hypothesis; that is, there is some $H_v, v \in \mathcal{T}$, that is true, and, for all $t < v$, H_t is false. The structure assumption is always true for $\mathcal{T} = \{1, 2, \dots, k\}$ or $\mathcal{T} = \{1, 2, \dots\}$, but, if \mathcal{T} were the nonnegative rational numbers with \leq , and if H_t asserted that the circumference of a circle is less than or equal to t times its diameter, then, because $\pi \notin \mathcal{T}$, H_t is true for all rational $t > \pi$, and there is not a first $t \in \mathcal{T}$ such that H_t is true; however, the problem disappears in this case if $\mathcal{T} = [0, \infty)$. The structure assumption is made throughout the following discussion; in practical situations, it is an innocuous assumption.

Method 1. For each $t \in \mathcal{T}$, test H_t at level α if and only if $H_s, s \in \mathcal{T}$, has been previously tested at level α and rejected for all $s < t$; otherwise, do not test H_t .

PROPOSITION 1. *The probability that Method 1 rejects at least one true hypothesis is at most α .*

Proof. Let \mathcal{F}_1 be the event that Method 1 tests and rejects at least one true hypothesis. If all $H_t, t \in \mathcal{T}$, are false, then there is nothing to prove. Suppose H_u is false for all $u < v$, but H_v is true, and let \mathcal{R}_v be the event $p_v \leq \alpha$, so that $\text{pr}(\mathcal{R}_v) \leq \alpha$. To reject falsely at least one true hypothesis, one must first falsely reject H_v with $p_v \leq \alpha$, so that $\mathcal{F}_1 = \mathcal{R}_v$ and $\text{pr}(\mathcal{F}_1) = \text{pr}(\mathcal{R}_v) \leq \alpha$. \square

Method 1 may perform many tests, each at level α , and yet rejects at least one true hypothesis with probability at most α . The possibility of doing this is familiar from several contexts, some of which are

directly related to Method 1 or to Method 2. Most familiar is the construction of a confidence set for a real parameter by inverting a test (Lehmann, 1959, § 3·5), so that infinitely many hypotheses are tested, only one hypothesis being true; however, this most familiar instance is more closely related to Method 2 than to Method 1. Other methods that may perform many tests at level α and yet falsely reject with probability at most α include Miller's (1966, § 3·7) special version of Fisher's 'least significant difference test,' the method of closed testing due to Marcus et al. (1976), certain methods of equivalence testing as described, for example, by Bauer & Kieser (1996) and certain methods of dose determination proposed by Bauer (1997) and Hsu & Berger (1999). Applications to clinical trials are discussed by Lehmacher et al. (1991) and Koch & Gansky (1996). The proof of Proposition 1 resembles the proof for 'closed testing' in Marcus et al. (1976); however, unlike closed testing and equivalence testing, the hypotheses H_t in Method 1 do not need to be related in any specific way. Like Method 1, the methods of Bauer (1997) and Hsu & Berger (1999) do not use closed testing, and instead test hypotheses in a given order; their concern is with a specific family of hypotheses, and Hsu & Berger (1999) develop a new stepwise confidence interval for that family.

Method 2, described below, generalizes Method 1 with a view to testing and rejecting additional hypotheses. We say that a subset $\mathcal{J} \subseteq \mathcal{T}$ is 'exclusive' if at most one $H_t, t \in \mathcal{J}$, is true, and that a subset $\mathcal{J} \subseteq \mathcal{T}$ is an 'interval' if $u \in \mathcal{J}, w \in \mathcal{J}$, and $u \preceq v \preceq w$ implies $v \in \mathcal{J}$. Every $t \in \mathcal{T}$ is contained in at least one exclusive interval, \mathcal{J} , namely $\mathcal{J} = \{t\}$. Suppose that $\mathcal{J}_\lambda, \lambda \in \Lambda$, is a collection of disjoint intervals such that each \mathcal{J}_λ is exclusive, and $\mathcal{T} = \bigcup_{\lambda \in \Lambda} \mathcal{J}_\lambda$. In a trivial sense, one such partition always exists, namely $\Lambda = \mathcal{T}, \mathcal{J}_t = \{t\}$, so the trivial partition of \mathcal{T} into disjoint exclusive intervals is $\mathcal{T} = \bigcup_{t \in \mathcal{T}} \{t\}$. In \mathcal{T} , disjoint intervals, \mathcal{J}_λ and \mathcal{J}_ω say, with $\mathcal{J}_\lambda \cap \mathcal{J}_\omega = \emptyset$, are themselves ordered by their contents: either $u < v$ for all $u \in \mathcal{J}_\lambda$ and $v \in \mathcal{J}_\omega$, or $v < u$ for all $u \in \mathcal{J}_\lambda$ and $v \in \mathcal{J}_\omega$; hence, Λ inherits an ordering from the ordering of the intervals $\mathcal{J}_\lambda, \lambda \in \Lambda$, which, with a slight abuse of notation, is again designated by $<$, and $\lambda < \omega$ means that $u < v$ for all $u \in \mathcal{J}_\lambda, v \in \mathcal{J}_\omega$, while $u < \mathcal{J}_\omega$ means that $u < v$ for all $v \in \mathcal{J}_\omega$.

Method 2. If $p_s \leq \alpha$ for all $s \in \mathcal{J}_\lambda$ for all $\lambda < \omega$, then reject all hypotheses $H_t, t \in \mathcal{J}_\omega$ with $p_t \leq \alpha$.

For the trivial partition, $\mathcal{T} = \bigcup_{t \in \mathcal{T}} \{t\}$, Method 2 is Method 1. With a nontrivial partition, $\mathcal{T} = \bigcup_{\lambda \in \Lambda} \mathcal{J}_\lambda$, Method 2 rejects at least as many, and possibly more, hypotheses compared to Method 1. Method 1 terminates with some hypothesis H_t such that $p_t > \alpha$, and hypotheses H_s with $t < s$ are not tested. In contrast, Method 2 terminates with an interval, \mathcal{J}_ω , with $p_t > \alpha$ for at least one $t \in \mathcal{J}_\omega$, but all hypotheses, $H_s, s \in \mathcal{J}_\omega$, are tested, even those with $t < s$, and possibly some such H_s is rejected. For Method 2, the ordering of hypotheses within an interval does not matter, in the sense that it does not affect which hypotheses are tested and rejected, but the ordering between intervals does matter. For a different trivial partition, Method 2 inverts a test to produce a $1 - \alpha$ confidence set for a real parameter θ when $\mathcal{T} = (-\infty, \infty)$ is the real line, $H_t : \theta = t$ for $t \in \mathcal{T}, \Lambda = \{1\}$ and $\mathcal{J}_1 = \mathcal{T}$; then $\mathcal{T} = \bigcup_{\lambda \in \Lambda} \mathcal{J}_\lambda = \mathcal{J}_1$ is itself exclusive, and Method 2 tests every $H_t : \theta = t$, excluding from the confidence set the rejected candidate values of θ .

PROPOSITION 2. *Let $\mathcal{T} = \bigcup_{\lambda \in \Lambda} \mathcal{J}_\lambda$ be a partition of \mathcal{T} into disjoint intervals where each \mathcal{J}_λ is exclusive. The probability that Method 2 rejects at least one true hypothesis is at most α .*

Proof. Let \mathcal{F}_2 be the event that Method 2 tests and rejects at least one true hypothesis. If all $H_t, t \in \mathcal{T}$, are false, then there is nothing to prove. Suppose that H_v is true for some $v \in \mathcal{J}_\omega$, but H_t is false for all $t < \mathcal{J}_\omega$. Since \mathcal{J}_ω is exclusive, H_v is the only true hypothesis in \mathcal{J}_ω ; therefore, H_v is the only hypothesis in \mathcal{J}_ω that can lead to a false rejection, and the true H_v must be falsely rejected if any true $H_t, t \in \mathcal{T}$, is to be rejected; that is, $\mathcal{F}_2 = \mathcal{R}_v$ where \mathcal{R}_v is the event $p_v \leq \alpha$, so that $\text{pr}(\mathcal{F}_2) = \text{pr}(\mathcal{R}_v) \leq \alpha$. \square

Proposition 2 required each \mathcal{J}_λ to be exclusive, but the same conclusion is available under a weaker premise. By the time the hypotheses $H_u, u \in \mathcal{J}_\lambda$, are tested, the earlier hypotheses, H_t with $t < \mathcal{J}_\lambda$, have all been tested and rejected. Is it safe to assume these earlier hypotheses are false? We say that a partition of \mathcal{T} into disjoint intervals is 'sequentially exclusive' if each \mathcal{J}_λ would be exclusive if all earlier hypotheses H_t were false; that is, $\mathcal{T} = \bigcup_{\lambda \in \Lambda} \mathcal{J}_\lambda$ is sequentially exclusive if every \mathcal{J}_λ contains at most one true hypothesis

whenever H_t is false for all $t < \mathcal{J}_\lambda$. An exclusive partition, as in Proposition 2, is sequentially exclusive, but not conversely. The proof of Proposition 3 is identical to the proof of Proposition 2.

PROPOSITION 3. *Let $\mathcal{T} = \bigcup_{\lambda \in \Lambda} \mathcal{J}_\lambda$ be a sequentially exclusive partition of \mathcal{T} into disjoint intervals. The probability that Method 2 rejects at least one true hypothesis is at most α .*

2. HYPOTHESES ABOUT TWO PARAMETERS

Consider testing the null hypothesis, $H_0 : \theta^* \leq 0$ or $\theta' \leq 0$, versus $H_1 : \theta^* > 0$ and $\theta' > 0$. Lehmann (1952), Berger (1982) and Laska & Meisner (1989) suggested testing H_0 versus H_1 at level α using two separate, consistent tests, with significance levels p^* and p' , each performed at level α , of $H_0^* : \theta^* \leq 0$ versus $H_1^* : \theta^* > 0$ and $H_0' : \theta' \leq 0$ versus $H_1' : \theta' > 0$, accepting H_0 if either $p^* > \alpha$ or $p' > \alpha$, rejecting H_0 in favour of H_1 if both $p^* \leq \alpha$ and $p' \leq \alpha$. Leon Gleser and Roger Berger refer to this as an intersection–union test. Generally, one cannot reject H_0^* if $p^* \leq \alpha$ or reject H_0' if $p' \leq \alpha$ without taking a risk of falsely rejecting a true hypothesis that is greater than α , so that the intermediate tests do not provide partial information about H_0^* and H_0' ; it is all or nothing.

Propositions 1–3 offer two alternative approaches whose only advantage is the possibility of rejecting intermediate hypotheses when H_0 is not rejected. Suppose one were willing to view H_0^* and H_1' asymmetrically, giving priority to H_0^* . Take $\mathcal{T} = \{*, '\}$ with $* < '$. Then Method 1 tests H_0^* and stops if $p^* > \alpha$; otherwise, if $p^* \leq \alpha$, it rejects H_0^* at level α , and tests H_0' , rejecting both H_0' and H_0 if $p' \leq \alpha$. Method 1 rejects H_0 if and only if $p^* \leq \alpha$ and $p' \leq \alpha$, as in the method in the previous paragraph, but it permits the intermediate conclusion that H_0^* is rejected, when H_1' and H_0 are not. By Proposition 1, the chance of falsely rejecting at least one true hypothesis is at most α . The use of Method 1 with intersection–union tests has been called ‘stepwise intersection–union testing’ by R. Berger in unpublished work; see Berger et al. (1988) and Berger & Boos (1999) for two applications.

Method 2 permits H_0^* and H_0' to be handled symmetrically. Consider the null hypothesis $H_0^+ : (\theta^* + \theta')/2 \leq 0$, which is of interest in some contexts. Let p_+ be the significance level for a level α test of H_0^+ . Let $\mathcal{T} = \{+, *, '\}$ with the order $+ < * < '$. If H_0^+ is false, then either H_0^* or H_0' must be false, so that $\mathcal{T} = \{+\} \cup \{*, '\}$ is a sequentially exclusive partition of \mathcal{T} into disjoint intervals, and Proposition 3 is applicable. Method 2 accepts H_0^+ and stops if $p_+ > \alpha$; otherwise, Method 2 rejects H_0^+ and tests both H_0^* and H_0' , rejecting H_0^* if $p^* \leq \alpha$, rejecting H_0' if $p' \leq \alpha$ and rejecting H_0 if H_0^* and H_0' are both rejected. Here, it is possible to accept all hypotheses, to reject H_0^+ alone, to reject H_0^+ and H_0^* , to reject H_0^+ and H_0' or to reject H_0^+ , H_0^* , H_0' and H_0 . In this process, the chance of falsely rejecting a true hypothesis is at most α .

3. COMPARING TREATED SUBJECTS TO TWO CONTROL GROUPS

In observational studies, a treated response is often compared to more than one type of control, in the hope that similar results among different types of control will strengthen the evidence that the treatment caused the difference between the treated response and the several types of control; see Campbell (1969) and Rosenbaum (2002, § 8) for discussion of how two control groups may provide information about biases from nonrandom treatment assignment. For instance, in the triangle design, measurements are obtained on untreated controls (c) and on treated subjects at baseline (b) before treatment and also after treatment (t). Were control and baseline responses more similar to each other than either was to the post-treatment responses?

Masjedi et al. (2000) studied possible genetic damage from the powerful drugs used to treat tuberculosis, comparing $n = 36$ patients with tuberculosis before (b) and after drug treatment (t), and 36 untreated controls (c) without tuberculosis, where the controls were matched to treated patients for age and gender. This produced 36 triples, (y_{ti}, y_{bi}, y_{ci}) , $i = 1, \dots, 36$, of post-treatment, baseline and control responses, where a measure of genetic damage was the frequency of chromosome aberrations including gaps per 100 cells, with higher numbers indicative of greater damage. A systematic difference between y_{ti} and y_{ci} could be an effect of tuberculosis rather than an effect of the drugs used to treat tuberculosis. An effect of the drugs

should produce y_{ti} that are higher than both y_{bi} and y_{ci} , with y_{bi} and y_{ci} similar to each other. Consider the model $y_{ti} = \mu_t + \pi_i + \lambda_i + \epsilon_i$, $y_{bi} = \mu_b + \pi_i + \lambda_i + \zeta_i$, $y_{ci} = \mu_c + \pi_i + \eta_i$, $i = 1, \dots, n$, where the π_i , λ_i , ϵ_i , ζ_i and η_i are mutually independent, with continuous distributions F_π , F_λ , F_ϵ , F_ζ and F_η , each of which is symmetric about zero. Here, π_i reflects the pairing of treated subjects and controls using age and gender, while λ_i reflects the possible correlation between the pre- and post-treatment measurements on the same person.

To assert that treated responses exceed baseline and control responses by more than baseline and control responses differ from each other is to assert the truth of

$$H_1: \mu_t - \max(\mu_b, \mu_c) > \max(\mu_b, \mu_c) - \min(\mu_b, \mu_c), \tag{1}$$

and to reject the null hypothesis

$$H_0: \mu_t - \max(\mu_b, \mu_c) \leq \max(\mu_b, \mu_c) - \min(\mu_b, \mu_c) \tag{2}$$

is to provide a basis for asserting H_1 . Perhaps, however, H_0 cannot be rejected, but some weaker statements can be made.

Define five hypotheses, namely $H_0^+ : \mu_t - (\mu_b + \mu_c)/2 \leq 0$, $H_0^b : \mu_t - \mu_b \leq 0$, $H_0^c : \mu_t - \mu_c \leq 0$, $H_0^* : \mu_t - \mu_c \leq \mu_c - \mu_b$ and $H_0' : \mu_t - \mu_b \leq \mu_b - \mu_c$, and let p^+ , p^b , p^c , p^* and p' be the respective significance levels from level α tests. To reject H_0^+ is to conclude that the post-treatment median μ_t exceeds the average $(\mu_b + \mu_c)/2$ of the baseline and control medians. To reject H_0^b or H_0^c is to conclude that the post-treatment median μ_t exceeds, respectively, the baseline median μ_b or the control median μ_c . If $\mu_c = \max(\mu_b, \mu_c)$, then H_0^* is H_0 , whereas if $\mu_b = \max(\mu_b, \mu_c)$ then H_0' is H_0 , so rejecting the conjunction of H_0^* and H_0' entails rejecting H_0 . Let $\mathcal{T} = \{+, b, c, *, '\}$ with the order $+$ < b < c < $*$ < $'$. If H_0^+ is false, then at most one of H_0^b and H_0^c can be true. If H_0^b and H_0^c are both false, then at most one of H_0^* and H_0' can be true, because either $\mu_c - \mu_b \leq 0$ or $\mu_b - \mu_c \leq 0$. It follows that $\mathcal{T} = \{+\} \cup \{b, c\} \cup \{*, '\}$ is a sequentially exclusive partition of \mathcal{T} into disjoint intervals, so Proposition 3 applies. Method 2 accepts H_0^+ and stops if $p^+ > \alpha$; otherwise, it rejects H_0^+ and tests both H_0^b and H_0^c , rejecting H_0^b if $p^b \leq \alpha$ and rejecting H_0^c if $p^c \leq \alpha$; then, if either $p^b > \alpha$ or $p^c > \alpha$, Method 2 stops; otherwise, it tests both H_0^* and H_0' , rejecting H_0^* if $p^* \leq \alpha$ and rejecting H_0' if $p' \leq \alpha$, and logically rejects H_0 if all five hypotheses have been rejected. In this process, the chance of falsely rejecting a true hypothesis is at most α .

The five hypotheses are tested by applying Wilcoxon's one-sided signed-rank test to $y_{ti} - (y_{bi} + y_{ci})/2$ for H_0^+ , to $y_{ti} - y_{bi}$ for H_0^b , to $y_{ti} - y_{ci}$ for H_0^c , to $y_{ti} + y_{bi} - 2y_{ci}$ for H_0^* and to $y_{ti} + y_{ci} - 2y_{bi}$ for H_0' . When this is done $p^+ = 9.5 \times 10^{-8}$ for H_0^+ , $p^b = 1.5 \times 10^{-7}$ for H_0^b , $p^c = 9.4 \times 10^{-8}$ for H_0^c , $p^* = 8.9 \times 10^{-7}$ for H_0^* and $p' = 0.0027$ for H_0' , so H_0 in (2) is rejected in favour of H_1 in (1), and there is strong evidence that the post-treatment response differs more from both baseline and control responses than the latter differ from each other. If the same procedure is applied to a different outcome in Masjedi et al. (2000), specifically the micronucleus frequency, then H_0^+ , H_0^b , H_0^c and H_0^* are rejected at the 0.001 level, but $p' = 0.14$ for H_0' , so that H_0 in (2) is not rejected, but there is, nonetheless, strong evidence that the post-treatment responses exceeded both the baseline and control responses.

Stronger inferences may be possible. For $\delta \geq 0$, replace H_0 in (2) by

$$H_{0,\delta}: \mu_t - \max(\mu_b, \mu_c) \leq \delta + \max(\mu_b, \mu_c) - \min(\mu_b, \mu_c). \tag{3}$$

If $H_{0,\delta}$ were false, then the difference between the post-treatment response and both baseline and control responses, $\mu_t - \max(\mu_b, \mu_c)$, is at least δ greater than the difference between the baseline and control responses, $\max(\mu_b, \mu_c) - \min(\mu_b, \mu_c)$. Take \mathcal{T} to be the direct product, with the lexicographical order, of the possible values of δ , namely $[0, \infty)$, and the ordered set used above, so that $\mathcal{T} = [0, \infty) \times \{+, b, c, *, '\}$; then $\mathcal{T} = \bigcup_{\delta \in [0, \infty)} \{(\delta, +)\} \cup \{(\delta, b), (\delta, c)\} \cup \{(\delta, *), (\delta, ')\}$ is a sequentially exclusive partition of \mathcal{T} into disjoint intervals, so that Proposition 3 applies to the infinite collection of hypotheses, $H_{0,\delta}^+ : (\mu_t - \delta) - (\mu_b + \mu_c)/2 \leq 0$, $H_{0,\delta}^b : (\mu_t - \delta) - \mu_b \leq 0$, $H_{0,\delta}^c : (\mu_t - \delta) - \mu_c \leq 0$, $H_{0,\delta}^* : (\mu_t - \delta) - \mu_c \leq \mu_c - \mu_b$ and $H_{0,\delta}' : (\mu_t - \delta) - \mu_b \leq \mu_b - \mu_c$. Starting with $\delta = 0$, the signed-rank test is applied to five contrasts such as $(y_{ti} - \delta) - (y_{bi} + y_{ci})/2$, yielding one-sided significance levels, p_δ^+ , p_δ^b , p_δ^c , p_δ^* and p_δ' , until a δ is reached such that one of the five significance levels is above α . In the case of chromosome aberrations, at

$\alpha = 0.05$, all five signed-rank statistics reject every $\delta \leq \frac{1}{2}$, but $p'_\delta > 0.05$ for all $\delta > \frac{1}{2}$, so that, with 95% confidence, $\mu_t - \max(\mu_b, \mu_c)$ is at least $\frac{1}{2} + \max(\mu_b, \mu_c) - \min(\mu_b, \mu_c)$.

ACKNOWLEDGEMENT

This work was supported by a grant from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation.

REFERENCES

- BAUER, P. (1997). A note on multiple testing procedures in dose finding. *Biometrics* **53**, 1125–8.
- BAUER, P. & KIESER, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83**, 934–7.
- BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate. *J. R. Statist. Soc. B* **57**, 289–300.
- BERGER, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**, 295–300.
- BERGER, R. L. & BOOS, D. D. (1999). Confidence limits for the onset and duration of treatment effect. *Biomet. J.* **41**, 517–31.
- BERGER, R. L., BOOS, D. D. & GUESS, F. M. (1988). Tests and confidence sets for comparing two mean residual life functions. *Biometrics* **44**, 103–15.
- CAMPBELL, D. T. (1969). Prospective. In *Artifact in Behavioral Research*, Ed. R. Rosenthal and R. Rosnow, pp. 351–82. New York: Academic Press.
- HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.
- HSU, J. C. & BERGER, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *J. Am. Statist. Assoc.* **94**, 468–75.
- KOCH, G. G. & GANSKY, S. A. (1996). Statistical considerations for multiplicity in confirmatory protocols. *Drug Inform. J.* **30**, 523–33.
- LASKA, E. M. & MEISNER, M. J. (1989). Testing whether an identified treatment is best. *Biometrics* **45**, 1139–51.
- LEHMACHER, W., WASSMER, G. & REITMEIR, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47**, 511–21.
- LEHMANN, E. L. (1952). Testing multiparameter hypotheses. *Ann. Math. Statist.* **23**, 541–52.
- LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
- MARCUS, R., PERITZ, E. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–60.
- MASJEDI, M. R., HEIDARY, A., MOHAMMADI, F., VELAYATI, A. A. & DOKOUHAKI, P. (2000). Chromosomal aberrations and micronuclei in lymphocytes of patients before and after exposure to anti-tuberculosis drugs. *Mutagenesis* **15**, 489–94.
- MILLER, R. G., JR. (1966). *Simultaneous Statistical Inference*. New York: Springer.
- ROSENBAUM, P. R. (2002). *Observational Studies*. New York: Springer.

[Received February 2007. Revised June 2007]