STATISTICS IN MEDICINE, VOL. 9, 811–818 (1990)

MORE POWERFUL PROCEDURES FOR MULTIPLE SIGNIFICANCE TESTING

YOSEF HOCHBERG AND YOAV BENJAMINI

Department of Statistics, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel

SUMMARY

The problem of multiple comparisons is discussed in the context of medical research. The need for more powerful procedures than classical multiple comparison procedures is indicated. To this end some new, general and simple procedures are discussed and demonstrated by two examples from the medical literature: the neuropsychologic effects of unidentified childhood exposure to lead, and the sleep patterns of sober chronic alcoholics.

1. INTRODUCTION

The problem of multiple comparisons may be viewed as stemming from an incompatibility between some formal statistical methodology and existing research practice. Confirmatory statistical methodology is based on the premise of a separate experiment for each inference assumed to be stated prior to the experiment. However, often a better scientific practice is to raise multiple related questions within the framework of a single study (experimental or observational) and also formulate research questions in view of the data. The problem is described by the following extract from Mosteller *et al.*:[1]

> If we are not limited in our choice of explanatory hypotheses, we often capitalize on some unusual features of the data. The typical clinical trial provides an enormous number of sub-groups (male smokers over 50 who are underweight, etc.) that can be compared, and many correlations and covariates. Many comparisons will seem reasonable under some hypotheses, *and some will of course appear to be statistically significant if tested as if they were predesignated items for assessment.*

Godfrey[2] and Pocock[3] show that multiple comparison problems are often not analysed by appropriate procedures in medical publications. According to Pocock[4] the effect of multiplicity and selective publications is: 'Perhaps the majority of trial reports claiming a treatment difference are false-positives.'

A simple solution of the problem is to impose stricter control on the probability of getting a false-positive result, or type I error rate, by requiring the use of classical multiple comparison procedures. One should, however, be aware of the cost: unnecessarily low type I errors without an increase in sample sizes imply higher type II error rates, that is a higher probability of not detecting a true positive treatment difference. This trade-off is often overlooked.

0277–6715/90/070811–08$05.00

Received July 1988
Revised September 1989

A similar and associated problem is that of selective publication of studies according to findings. Pocock[4] (p. 241) demonstrates how the excess of false-positives in publication arises, assuming that small ($n = 40$) non-significant experiments are not published. In his example the type I error rate more than triples from 0·05 to 54/320 = 0·17, whereas the type II error rate drops to a third from an overall of 177/270 = 0·65 to 27/120 = 0·22.

Obviously, the problem of multiple testing is of a higher complexity when both error rates are considered. The following quotation is from Meier:[5] 'Among the most difficult challenges to statistical theory arising from the field of clinical trials, are those grouped under the heading of multiplicity.' The growing awareness of the cruel trade-off between type I and type II error rates associated with multiple comparison has had a strong effect on the philosophy and methodology of multiple comparisons procedures.[6] There are researchers who advocate an across the board per-comparison approach, that is, they advocate no modification in significance levels in spite of the simultaneous consideration of multiple comparisons. They maintain that a per-comparison approach is legitimate whenever the researcher submits unselectively the results for all comparisons which were examined. The problem however lies with the need, which often arises, to properly support the *conclusions* of the study. This is not automatically satisfied by indicating significance or lack of significance for a collection of *individual comparisons*. A conclusion is often based on a *family* of comparisons, for example, recommending one treatment over another in view of the outcome of several end points, or several side-effects. In a large experiment one may want to consider several such families separately. There are situations where a careful analysis of the relation between a conclusion and the family of comparisons on which it is based indicates the need to control the *familywise error rate*, defined as the *probability of at least one error*. In such situations, a per-comparison approach may result in an excessive rate of false-positive conclusions. On the other hand, classical multiple comparison procedures are often too conservative when they properly control the familywise type I error rate.

In this paper we assume that the control of a familywise type I error rate is required and give procedures which achieve that with a lower type II error rate than that of existing procedures. We focus on the Bonferroni method which is a simple yet general procedure. It does not require any constraining assumptions on the distributions of the individual statistics, or any knowledge of the dependence structure among them. All it requires are the p-values corresponding to the various individual hypotheses. However, *the* Bonferroni method is often too conservative. We are interested in obtaining a sharper multiple comparison procedure, that is one having lower type II error rates, which can be completely specified in terms of the individual p-values. Such a procedure is referred to as *a* Bonferroni type procedure, and such are described in Sections 3 and 4. The procedures are illustrated using two examples taken from published medical literature, which are described in Section 2. The choice of the examples was dictated by the availability of appropriate information in the published analysis. We wish to emphasize that in no way does our choice reflect criticism of the original analysis.

2. TWO EXAMPLES

2.1. Example 1: Effects of exposure to lead

Needleman *et al.*[7] studied the neuropsychologic effects of unidentified childhood exposure to lead, by comparing various psychologic and classroom performances between two groups of children differing in the lead level observed in their shed teeth. In their Table 8 they also present p-values for a family of comparisons concerning verbal processing scores and reaction times. It

makes sense here to control the familywise error rate of this family separately from the other families of comparisons involving differences in intelligence and class behaviour. The ordered $p_{(i)}$s and the hypotheses to which they refer are as follows:

0·90	0·42	0·37	0·32	0·07	0·05	0·04	0·03	0·01	0·002	0·001	0·001
T-2	T-3	T-1	RT-1	R-C	T-4	SR	R-B	RT-4	R-A	RT-3	RT-2

T-1 to T-4 are the four blocks of the token test; R-A to R-C are the three subtests of the seashore rhythm test; SR is a sentence repetition test; and RT-1 to RT-4 are reaction times at varying interval of delays. The authors address the problems of multiplicity in the footnote to the table and suggest some Bonferroni correction.

2.2. Example 2: Sleep patterns

The second example is from a study by Synder and Karacan[8] on sleep patterns of sober chronic alcoholics. The somnopolygraphs of 26 sober chronic alcoholics, taken after approximately 25 days of sobriety, were compared with those of a matched control group. The comparisons included 22 different parameters, and the ordered $p_{(i)}$s are as follows:

0·56	0·50	0·44	0·40	0·34	0·28	0·26	0·24	0·18	0·18	0·16
0·08	0·04	0·04	0·02	0·004	0·001	0·001	0·001	0·001	0·001	0·001

These p-values refer to hypotheses on total sleep parameters such as time, sleep efficiency and number of awakenings, REP period parameters, sleep parameters at various stages, and latencies to various stages. The study concludes that the sleep efficiency index and latency to sleep stages were disturbed in the alcoholic subjects. There was a decrease in the number of REM episodes but slow-wave sleep was generally unaffected.

Reporting their results, Synder and Karacan address the issue of multiplicity:

> A probability of 0·05 was the limit of statistical significance. However, because of the large number of factors tested (twenty-two), it is preferable to consider α inflation when interpreting the results . . . The usual case is that a test significant at the level of 0·001 in Table I would stay significant at the 0·05 level after adjusting for α inflation, whereas 0·05 significance levels would become insignificant. It is doubtful whether significance levels of 0·01 would remain significant with the adjustment.

3. SHARPER BONFERRONI PROCEDURES

Consider simultaneously testing a family of m null hypotheses $H_1, \ldots, H_m$ in order to reach a conclusion. Let $P_1, \ldots, P_m$ be the individual P-values corresponding to these hypotheses, where upper case P is used here to denote that these are considered prior to the experimentation. Nothing needs to be assumed about the experiment, the nature of the hypotheses, or the dependence or independence among the P-values. The various P-values may correspond to different types of test-statistics (chi squares, t-tests and so on). Under the null hypothesis H_i, P_i is equally likely to fall anywhere between 0 and 1. If H_i is wrong then P_i tends to be closer to zero. The test of H_i in terms of P_i is to reject H_i when P_i is small. In a per-comparison approach *small* means smaller than a pre-specified significance level considered appropriate for H_i. If however the *familywise* error rate needs to be controlled at a specified level, α say, then the Bonferroni procedure can be used. The original Bonferroni procedure amounts to using a significance level α_i for H_i with the added requirement that the sum of the α_is equals α. Often, when there is no good reason to use different α_is, the Bonferroni method rejects H_i if $p_i \leqslant \alpha/m$. Thus, in such cases, the

Bonferroni method amounts to using a per-comparison error rate of α/m in order to achieve a familywise error rate of α. This can be very conservative when m is large and for some dependence structures.

In Example 1, controlling the familywise error at 0·05, the Bonferroni approach suggests using $0{\cdot}05/12 = 0{\cdot}004$. In this case only the three hypotheses concerning RT-2, RT-3, and R-A, corresponding to the p-values of 0·001 and 0·002, are rejected. No hypothesis will be rejected if we hold the familywise error rate at 0·01, since $0{\cdot}01/12 = 0{\cdot}0008$ is smaller than all p-values.

In Example 2 the simple Bonferroni correction for the 0·05 level is to test individually at the $0{\cdot}05/22 = 0{\cdot}0022$ level. Thus only the six hypotheses corresponding to p-values of 0·001 are rejected at this level. This calculation is reflected in the statement of Synder and Karacan quoted above regarding the significance of a p-value of 0·001. A hypothesis with corresponding p-value larger than 0·0022 will not be rejected using the simple Bonferroni procedure.

Holm[9] gave a sharper procedure, that is a procedure which always achieves the same type I familywise error rate control and lower type II error rates. Holm's procedure requires the ordering of the p_is into $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(m)}$ so that $p_{(i)}$ is the ith smallest p-value. The hypothesis corresponding to $p_{(i)}$ is denoted by $H_{(i)}$. According to Holm's procedure one rejects $H_{(i)}$ if, and only if,

$$p_{(j)} \leqslant \frac{\alpha}{m-j+1} \qquad \text{for all } j \leqslant i\,. \tag{1}$$

Thus the hypothesis $H_{(1)}$ will be rejected when $p_{(1)}$, the smallest p-value, is less than α/m. The hypothesis $H_{(2)}$ is rejected when $p_{(1)} \leqslant \alpha/m$ and $p_{(2)} \leqslant \alpha/(m-1)$, and so on. Obviously, this method is identical with the original Bonferroni for $H_{(1)}$ but sharper for all other hypotheses.

Using Holm's procedure in Example 1, with $\alpha = 0{\cdot}05$, the smallest p-value of 0·001 is compared with $0{\cdot}05/12 = 0{\cdot}004$ and the corresponding hypothesis is rejected. The second is compared with $0{\cdot}05/11 = 0{\cdot}004$ and the hypothesis is rejected. The third is 0·002 and the hypothesis is rejected after comparing with $0{\cdot}05/10 = 0{\cdot}005$. The fourth ordered hypothesis, regarding RT-4, has a p-value of 0·01. It is larger than $0{\cdot}05/9 = 0{\cdot}0057$, so the procedure terminates here rejecting only the previous 3 hypotheses. When controlling the familywise error rate at the 0·01 level, even the smallest p-value does not lead to rejection of any hypothesis.

In the sleep patterns example, since $0{\cdot}05/22 = 0{\cdot}0022 > 0{\cdot}001$, the procedure rejects the six hypotheses having p-value smaller than that. The next hypothesis is about a crucial parameter: sleep latency to any stage. The p-value is $0{\cdot}004 > 0{\cdot}05/16$, so it is not rejected even according to Holm's method, and hence all the other 15 hypotheses with larger p-values are retained.

Recently, Hochberg[10] derived an even sharper procedure which uses the ordered p_is but in a different way from Holm's procedure. This procedure starts by examining the largest p-value $p_{(m)}$. If $p_{(m)} \leqslant \alpha$, then $H_{(m)}$ and all other hypotheses are rejected. If not, $H_{(m)}$ is not rejected and one proceeds to compare $p_{(m-1)}$ with $\alpha/2$. If the former is smaller, then $H_{(m-1)}$ and all hypotheses with smaller p-values are rejected. Generally, one proceeds from highest to lower p-values, retaining $H_{(i)}$ if its p-value satisfies $p_{(i)} > \alpha/(m-i+1)$. One stops the procedure at the first ordered hypothesis when that inequality is reversed. This hypothesis is rejected and so are *all* hypotheses with lower or equal p-values. This is always a sharper procedure than Holm's.

In our first example we compare the largest p-value, that is $p_{(12)} = 0{\cdot}9$, corresponding to the hypothesis T-2 with 0·05; the second with $0{\cdot}05/2 = 0{\cdot}025$; the third with 0·0166; and so on. The p-value for RT-4 is $p_{(4)} = 0{\cdot}01$ and is compared with $0{\cdot}05/9$, and the corresponding hypothesis not rejected. However, $p_{(3)} = 0{\cdot}002 < 0{\cdot}05/10 = 0{\cdot}005$, and therefore the last three hypotheses are all rejected. No hypothesis is rejected if the 0·01 significance level is used.

In Example 2 there are twelve p-values larger than 0·05. According to Hochberg's method, the next p-value should be compared with $0{\cdot}05/13 = 0{\cdot}00385$. Since $p_{(10)} = 0.04$ is not smaller, continue in this manner. The first hypothesis to be rejected is the seventeenth, where $p_{(6)} = 0{\cdot}001$, and therefore the other five hypotheses with this level are also rejected. For this example all three approaches result in the same conclusions.

Note that both the Holm and the Hochberg procedures can be viewed as comparing 'inflated' p-values, defined by $p^*_{(i)} = (m + 1 - i)p_{(i)}$, with the desired level α. For the first example the inflated p-values are 0·9, 0·84, 1·11, 1·28, 0·35, 0·3, 0·28, 0·24, 0·09, 0·02, 0·011, 0·012, and can be compared sequentially to any desired significance level α.

4. UTILIZING A SIMPLE GRAPHICAL PROCEDURE

Let m_0 be the number of *true* (null) hypotheses among m (null) hypotheses. Schweder and Spjøtvoll[11] gave a graphical method for estimating m_0. We present here a modification of their approach which utilizes the quantities displayed in their method.

Set $Q_{(i)} = 1 - P_{(m+1-i)}$ (and respectively for the observed values, $q_{(i)} = p_{(m+1-i)}$) to be the complement of the ith largest P-value; that is $Q_{(1)} \leqslant \ldots \leqslant Q_{(i)} \leqslant \ldots \leqslant Q_{(m)}$. If all null hypotheses are true, then, $m_0 = m$ and the set of $Q_{(i)}$s behaves as an ordered sample from a uniform distribution over $[0, 1]$. The expected value of $Q_{(i)}$ is approximately $i/(m_0 + 1)$.

The plot of $q_{(i)}$ versus i, also called a quantile plot or a uniform probability plot, should exhibit a linear relationship, along a line of slope $1/(m_0 + 1)$ passing through the origin. When $m_0 < m$, the p-values corresponding to the false null hypotheses tend to be smaller than the p-values corresponding to the true null hypotheses, so the $q_{(i)}$s corresponding to the false null hypotheses concentrate on the right side of the probability plot. The relationship over the left side of the plot remains approximately linear with slope $1/(m_0 + 1)$. To implement the method, a straight line is fitted to the smallest $q_{(i)}$s using the ordinary least squares method. The estimated slope of the line $\hat{\beta}$ is an unbiased estimator of $1/(m_0 + 1)$, and can be used to estimate m_0 by $\hat{m}_0 = 1/\hat{\beta} - 1$. This estimator of m_0 is biased but the bias is upwards, leading to an estimate of m_0 which is possibly too high and thereby leaning towards a conservative procedure. Figure 1 is the quantile plot of the 12 $q_{(i)}$s in our Example 1. The first 4 or 5 $q_{(i)}$s do seem to lie on some line through the origin.

The remaining question is how many of the smallest $q_{(i)}$s should be used to fit the line and estimate its slope. If many p-values are below 0·5 then just these points can be used to estimate the slope. If this is not the case, then we can fit lines to a progressively larger number of the smallest $q_{(i)}$s. As long as all points are along the line the estimated slopes tend to vary non-systematically. As we leave that region of linearity the estimated slope will consistently decrease because of the curvature. This can serve as an indication of where to stop including additional $q_{(i)}$s. Using this set of $q_{(i)}$s the number of true null hypotheses is estimated by $\hat{m}_0$. This estimate can be used for further sharpening the procedures given in Section 3.

Holm's procedure can be sharpened as follows. First reject any hypothesis whose p-value is less than or equal to $\alpha/\hat{m}_0$. Let m_1 denote the number of hypotheses with p-values greater than $\alpha/\hat{m}_0$. If $m_1 \geqslant \hat{m}_0$, stop and retain all these m_1 (null) hypotheses. If $m_1 < \hat{m}_0$ then reject any hypothesis with p-value less than or equal to α/m_1. Let m_2 denote the number of hypotheses retained after this step. Now reject any hypothesis with p-value less than or equal to α/m_2; and so on.

Hochberg's procedure can be further sharpened as follows. First, retain any hypothesis with p-value greater than α. Let $m^{(1)}$ denote the number of hypotheses retained at this stage. Now, reject $H_{(m-m^{(1)})}$ if

$$P_{(m-m^{(1)})} \leqslant \frac{\alpha}{\min(\hat{m}_0, \hat{m}^{(1)} + 1)}. \tag{2}$$

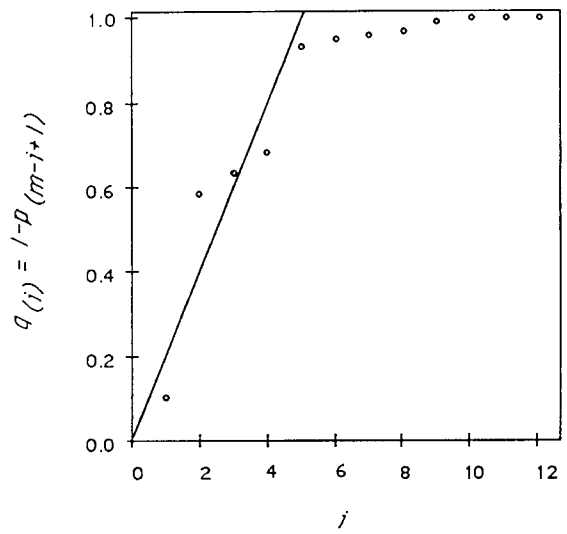

Figure 1. Neuropsychologic effects of childhood exposure to lead: the quantile plot of $q_{(i)} = 1 - p_{(m+1-i)}$ versus i. The line through the origin is fitted by least squares to the four smallest $q_{(i)}$s

If rejected, then all hypotheses with p-values less than or equal to $P_{(m-m^{(1)})}$ are rejected. If retained, let $m^{(2)} = m^{(1)} + 1$ and examine $H_{(m-m^{(2)})}$. If $P_{(m-m^{(2)})} \leqslant \alpha/\min(\hat{m}_0, m^{(2)} + 1)$ then reject it. If not, then retain $H_{(m-m^{(2)})}$ and let $m^{(3)} = m^{(2)} + 1$; and so on. With this procedure when one hypothesis is rejected, all hypotheses with less than or equal p-values are rejected.

It can be judged visually from Figure 1 that, in the effects of exposure to lead example, the estimated slope starts to decrease consistently from the third point on. The fitted slopes based on 4, 5 or 6 first $q_{(i)}$s lead to estimated m_0 of 4·1, 4·24 and 4·6 respectively, so we can set $\hat{m}_0 = 5$. At the 0·05 familywise error rate use $0{\cdot}05/5 = 0{\cdot}01$. The hypothesis corresponding to RT-4 is now also significant according to both amended procedures. At the 0·01 level, using $0{\cdot}01/5 = 0{\cdot}002$, the three last comparisons regarding RT-3, RT-2 and R-A are declared significant.

Turning to the sleep patterns example, we see in Figure 2 that a line through the origin does not seem to fit well any number of points. This is an indication of no true null hypothesis. However, successive slopes estimated from Figure 2 are 0·44, 0·288, 0·229 and 0·184, for the first 1, 2, 3 and 4 $q_{(i)}$s respectively. The slope is decreasing right from the beginning. Using the first two points for the estimate gives $\hat{m}_0 = 3$ after rounding upwards, and the threshold for significance is $0{\cdot}05/3 = 0{\cdot}0166$. Using this level, the conclusion about the significance of the individual comparison at the 0·004 level should be revised. From this a significant difference in latency to sleep stages between the two groups is recognized.

The advantage of Hochberg's procedure over Holm's procedure is reciprocal to $\hat{m}_0$. Furthermore, when $\hat{m}_0$ is small, both amended procedures are substantially improved in terms of their type II credentials relative to their original forms and become essentially equivalent.

If m_0 was known, then the amended procedures, with m_0 replacing $\hat{m}_0$, control the familywise type I error rate like the procedures in Section 3. The outline of the argument is as follows. For any subset of the hypotheses we may pose the 'subset intersection null hypothesis', that all hypotheses in this subset are true. According to the closure principle of Marcus *et al.*,[12] we control the probability of at least one erroneous rejection at α, by testing subset intersections in

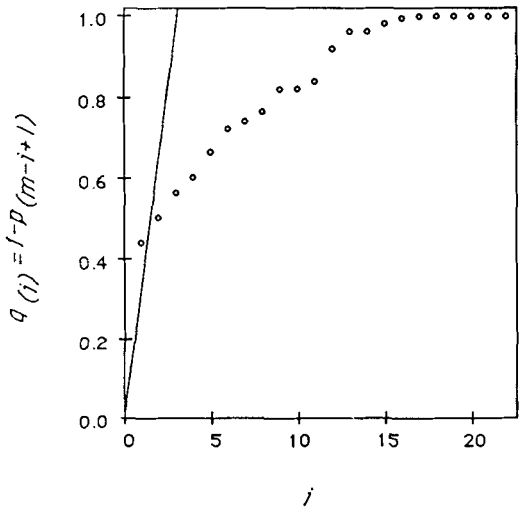

Figure 2. Sleep patterns of sober alcoholics: the quantile plot of $q_{(i)} = 1 - p_{(m+1-i)}$ versus i. The line through the origin is fitted by least squares to the two smallest $q_{(i)}$s

the following hierarchical way: test each subset intersection at level α only after all intersections of subsets containing it were tested at level α and rejected.

When it is known that there are m_0 true null hypotheses among those tested, all subsets containing more than m_0 hypotheses cannot have a true subset intersection. Therefore the above hierarchical testing can be confined to subsets of size $\leqslant m_0$, leading to the amended procedures. Thus when $\hat{m}_0$ can be considered a good estimator of m_0 and turns out to be non-negligibly smaller than m, the amended procedures discussed here should be used instead of the procedures of Section 3. Situations when $\hat{m}_0$ can be considered a good estimator of m_0 usually involve large values of m and comparisons which are strongly dichotomized as either null or substantially different from null. Research on typologies of the quantile plots is presently being pursued.

5. DISCUSSION

Biomedical investigations, whether experimental or observational, may be extremely complex. They may involve the study of many aspects of effectiveness of a therapy at various levels, for different subgroups, all adjusted for yet other factors. The results of such investigations are not reported merely as a single conclusion or set of conclusions, but rather the individual component results from which the major conclusions are drawn are also presented.

It has become accepted practice to report the individual results by giving the estimates of the size of any effect or comparison made in the form of means, percentages, survival rates and so on. These estimates are supported by evidence of their statistical significance in the form of individual p-values.

The methods presented in this article require only the p-values corresponding to the various individual results. Therefore they enable the researchers, or a reader of their report, to assess the overall *statistical significance* of the conclusions. Such an ability is especially important to the external reader who may not possess the necessary information on the inter-relationships

between the individual pieces of the structure of the whole study. The reader cannot assess the combined significance level of the conclusions through a single precise multivariate testing procedure which simultaneously controls all errors, and has to rely on the reported p-values.

For a long time the Bonferroni method for controlling the familywise type I error rate has been the only method which makes use of only the p-values. The Bonferroni procedure is simple and general, but often it is too conservative. The procedures of Holm and of Hochberg sharpen the Bonferroni procedure: both procedures have lower type II errors, while keeping the type I error rate at a level less than α. These procedures examine the sequential order of the (inflated) p-values. Hochberg's procedure, which starts with the largest p-value and works its way down, has the lower type II error.

We propose further modification which involves the graphical plot of the complements of the individual p-values versus their order. The number of the true null hypotheses may be estimated from the linear part of the plot at the left side, which displays the complements of the high p-values. This estimate is then used to amend the Holm and Hochberg procedures. It is shown that the amended procedures still control the type I error rate, up to the approximation involved in estimating – rather than knowing – the number of true null hypotheses. As the two biomedical examples demonstrate, the amended procedures gain from further reduction in the type II error rate. Initial results from a simulation study currently under way suggest that this is not merely anecdotal evidence.

With the enhanced type II error rate performance, the arsenal of multiple comparisons tools presented in this article can serve not only the reader of the report of the investigation but the researchers too. In those complex situations where a single 'overall' test of conclusion is hard to arrive at, or where unrealistic additional assumptions are needed to justify the use of such a test, the suggested methods offer attractive alternatives.

REFERENCES

1. Mosteller, F. Gilbert, I. and McPeek, Q. 'Controversies in design and analysis of clinical trials', in Shapiro, S. H. and Louis, T. A. (eds) *Clinical Trials: Issues and Approaches*, Marcel-Dekker, 1983.
2. Godfrey, K. 'Comparing the means of several groups', *New England Journal of Medicine*, **311**, 1450–1456 (1985).
3. Pocock, S. J. 'Statistical problems in the reporting of clinical trials', *New England Journal of Medicine*, **317**, 426–432 (1987).
4. Pocock, S. J. *Clinical Trials – A Practical Approach*, Wiley, 1983.
5. Meier, P. 'Statistical analysis of clinical trials', in Shapiro, S. H. and Louis, T. A. (eds) *Clinical Trials: Issues and Approaches*, Marcel-Dekker, 1983.
6. Hochberg, Y. and Tamhane, A. *Multiple Comparison Procedures*, Wiley, 1987.
7. Needleman, H., Gunnoe, C., Leviton, A., Reed, R., Presie, H., Maher, C. and Barret, P. 'Deficits in psychologic and classroom performance of children with elevated dentine lead levels', *New England Journal of Medicine*, **300**, 689–695 (1979).
8. Snyder, S. and Karacan, I. 'Sleep patterns of sober chronic alcoholics', *Neuropsychobiology*, **13**, 97–100 (1985).
9. Holm, S. 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics*, **6**, 65–70 (1979).
10. Hochberg, Y. 'A sharper Bonferroni procedure for multiple tests of significance', *Biometrika*, **75**, 800–803 (1988).
11. Schweder, T. and Spjøtvoll, E. 'Plots of p-values to evaluate many tests simultaneously', *Biometrika*, **69**, 493–502 (1982).
12. Marcus, R., Peritz, E. and Gabriel, K. R. 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika*, **63**, 655–660 (1976).