*Biometrika* (1976), **63**, 3, *pp.* 655–60
*Printed in Great Britain*



# On closed testing procedures with special reference to ordered analysis of variance

By RUTH MARCUS
*Department of Statistics, Tel-Aviv University*

ERIC PERITZ
*Department of Statistics, Hebrew University of Jerusalem*

AND K. R. GABRIEL
*Department of Statistics, University of Rochester, New York*

SUMMARY

A method of devising stepwise multiple testing procedures with fixed experimentwise error is presented. The method requires the set of hypotheses tested to be closed under intersection. The method is applied to the problem of comparing many treatments to one control and to ordered analysis of variance.

*Some key words:* Closed testing procedures; Multiple comparisons; Multiple testing procedure; One-way analysis of variance; Ordered alternatives.

## 1. Introduction

The aim of this paper is to propose a method for devising multiple testing procedures with bounded experimentwise error rates. The procedures thus obtained are sometimes more powerful that those in common use. The method is applied to the problem of comparing many treatments to one control and to Bartholomew's (1959) analysis of variance with ordered alternatives.

The idea of a closed testing procedure stems from the need to amend some multiple testing procedures in current use. Some of these methods may attain very high experimentwise error rates. Such is the case, for instance, with the procedure of Newman (1939) and Keuls (1952) in analysis of variance whenever the treatments are homogeneous within one of several very distinct sets (Hartley, 1955). Other methods, such as that of Dunnett discussed in the next section, are unduly conservative in that more inferences are possible at the same experimentwise error rate.

Our aim is to construct multiple testing procedures in which the experimentwise error rate equals the required level $\alpha$ of the overall test. The essential feature of our method is that we refer to sets of hypotheses which are closed under intersection, and that each test is of level $\alpha$. An example of a closed procedure, due to Peritz, which modifies the Newman–Keuls method is given by Einot & Gabriel (1975, § 1·8). Williams's (1971) procedure also is closed, although he does not point this out.

## 2. Closed testing procedures

Let $X$ be a random variable with distribution $P_\theta$ $(\theta \in \Omega)$. Let $W = \{\omega_\beta\}$ be a set of null hypotheses, i.e. a set of subsets of $\Omega$, closed under intersection: $\omega_i, \omega_j \in W$ implies $\omega_i \cap \omega_j \in W$.

For each $\omega_\beta$ let $\phi_\beta(X)$ be a level $\alpha$ test, that is, $\mathrm{pr}_\theta\{\phi_\beta(X) = 1\} \leqslant \alpha$ for all $\theta \in \omega_\beta$. Now consider the following procedure.

Any null hypothesis $\omega_\beta$ is tested by means of $\phi_\beta(X)$ if and only if all hypotheses $\omega$ that are included in $\omega_\beta$ ($\omega \subset \omega_\beta$) and belonging to $W$ ($\omega \in W$) have been tested and rejected. The probability of making no type I error with this procedure is at least $1-\alpha$. This is so since a type I error is committed if and only if the intersection of all true hypotheses, $\omega_\tau$ say, is tested and rejected by means of $\phi_\tau(X)$; in other words, if we denote by $A$ the event that any true $\omega_\beta$ is rejected, and by $B$ the event that $\phi_\tau(X) = 1$, then

$$\mathrm{pr}(A \cap B) = \mathrm{pr}(B)\,\mathrm{pr}(A|B) \leqslant \alpha$$

since $\phi_\tau$ is a level $\alpha$ test. However, since $A \cap B = A$, $\mathrm{pr}(A \cap B) = \mathrm{pr}(A)$ and hence $\mathrm{pr}(A) \leqslant \alpha$.

A simple example of a closed testing procedure is provided by modifying Dunnett's (1955) one-sided comparison of many treatment groups to one control group: let $X_i \sim N(\mu_i, \sigma^2 n^{-1})$ ($i = 1, \ldots, k$) and $X_0 \sim N(\mu_0, \sigma^2 m^{-1})$. Let $s^2$ be an unbiased estimate of $\sigma^2$ distributed $\sigma^2\chi^2_\nu/\nu$ and independent of $X_0, \ldots, X_k$. It is known that $\mu_i - \mu_0 \geqslant 0$ for all $i = 1, \ldots, k$. We want to test the hypotheses $\mu_i = \mu_0$ against $\mu_i > \mu_0$ for all $i$ so that the probability of making no type I error is at least $1-\alpha$.

We start by enlarging the set of hypotheses to be tested so as to include all hypotheses of the type $\omega_P$: $\mu_i = \mu_0$ for all $i \in P$, where $P$ is some subset of $\{1, \ldots, k\}$. Clearly $W = \{\omega_P\}$ is closed under intersection. Now, $\omega_P$ will be rejected if

$$\max_{i \in P} (X_i - X_0) > s d_{p,\nu,\alpha},$$

where $p$ is the number of elements in $P$ provided all hypotheses $\omega_R$ with $R \supset P$ have been rejected. Here $d_{p,\nu,\alpha}$ is the $\alpha$-critical point for Dunnett's (1955) statistic with $p$ and $\nu$ degrees of freedom, $p$ being the number of treatments in $P$. Since $d_{p,\nu,\alpha}$ is increasing with $p$, this procedure is clearly more powerful than Dunnett's original one, which uses the critical value $d_{k,\nu,\alpha}$ for all the comparisons. On the other hand this procedure does not provide one-sided confidence bounds for $\mu_i - \mu_0$ which Dunnett's procedure does. Also, unlike Dunnett's procedure, in this closed testing procedure the inferences made on $\mu_i - \mu_0$ depend not only on $X_i$, $X_0$ and $s^2$ but also on the other 'irrelevant', $X$'s.

The above procedure is consonant in the sense of Gabriel (1969): whenever a composite hypothesis is rejected at least one of its component hypotheses is rejected as well. Therefore this procedure can be written in the following simplified form: if $X_i$ is the $i$th largest $X$, reject $\omega_i$ if $X_i - X_0 > s d_{k-i+1,\nu,\alpha}$, provided the hypotheses corresponding to the $X$'s larger than $X$ have been rejected.

An alternative, nonconsonant, procedure consists in using at each stage, instead of Dunnett's test, the corresponding likelihood ratio, or $\bar{\chi}^2$ test; see Barlow, Bartholomew, Bremner & Brunk (1972, p. 145) under 'simple tree alternatives'.

It is easy to derive closed testing procedures, consonant or otherwise, for a variety of situations. Difficulties arise, however, with hypotheses that have so-called two-sided alternatives. This is readily illustrated by the case of one-way analysis of variance.

Let $\mu_1, \ldots, \mu_k$ be the population means with respect to which null hypotheses are formulated. The overall null hypothesis is, of course, $\omega_0$: $\mu_1 = \ldots = \mu_k$ and a closed set of 'interesting' null hypotheses consists of all hypotheses of the form $\omega_P$: $\mu_{i_1} = \ldots = \mu_{i_p}$, where $\{i_1, \ldots, i_p\} \subset \{1, \ldots, k\}$.

Now, whenever we reject a null hypothesis $\omega_P$ relating to exactly two means $\mu_{i_1} = \mu_{i_2}$, we accept instead of it one of two alternatives: $\mu_{i_1} > \mu_{i_2}$ or $\mu_{i_1} < \mu_{i_2}$. It seems therefore natural to require from a closed testing procedure that the probability of not rejecting any true $\omega_P$ and not accepting any alternative of the type $\mu_{i_1} > \mu_{i_2}$ when the reverse is true should be at least $1-\alpha$. Until now no closed testing procedure has been shown to have this property.

### 3. Application to the one-way analysis of variance with ordered alternatives

Let $\bar{X}_1, \ldots, \bar{X}_k$ be averages of $k$ independent samples of sizes $n_1, \ldots, n_k$, $\bar{X}_i \sim N(\mu_i, \sigma^2/n_i)$ $(i = 1, \ldots, k)$, where the means $\mu_i$ are unknown and $\sigma^2$ is known, and will be taken henceforth to equal one. Assume that the means $\mu_i$ are known *a priori* to satisfy the ordering $\Omega: \mu_1 \leqslant \ldots \leqslant \mu_k$.

The problem of testing the null hypothesis $\omega_0: \mu_1 = \ldots = \mu_k$ against the alternative $\Omega \cap \bar{\omega}_0: \mu_1 \leqslant \ldots \leqslant \mu_k$ with at least one strict inequality has been investigated by Bartholomew (1959) and others; for discussion and references, see Barlow *et al.* (1972, §§ 3·2, 3·3).

Let $\lambda_1, \ldots, \lambda_r$ be positive integers satisfying $\lambda_1 + \ldots + \lambda_r = k$. Put $\tau_0 \equiv 0$ and $\tau_j = \lambda_1 + \ldots + \lambda_j$. Let $g_j$ be the set of consecutive integers $(\tau_{j-1}+1, \ldots, \tau_j)$, and define by $g = (g_1, \ldots, g_r)$ the corresponding partition of the set $(1, \ldots, k)$. Let $\bar{\mu}(g_j) = \Sigma n_i \mu_i / \Sigma n_i$, where the summation is over all $i \in g_j$. Consider the following family of hypotheses.

$$\omega_g = \omega(g_1, \ldots, g_r): \mu_i = \bar{\mu}(g_j) \quad (i \in g_j;\ j = 1, \ldots, r).$$

It is easy to see that $\{\omega_g\}$ is closed under intersection and $\omega_0 = \cap\, \omega_g$, where the intersection goes over all partitions $g$ of $\{1, \ldots, k\}$.

The likelihood ratio statistic for testing $\omega_0$ against $\Omega \cap \bar{\omega}_0$ (Bartholomew, 1959) is

$$D^2 = \sum_{i=1}^{k} n_i (\hat{\mu}_i - \bar{X})^2,$$

where $\bar{X} = \Sigma n_i \bar{X}_i / \Sigma n_i$ and $(\hat{\mu}_1, \ldots, \hat{\mu}_k)$ are the maximum likelihood estimators of $(\mu_1, \ldots, \mu_k)$ under the model $\Omega$, and are obtained by the amalgamation process described by Brunk (1958). The null distribution of $D^2$ has been shown by Bartholomew (1959) to be

$$\operatorname{pr}(D^2 > t^2) = \sum_{m=2}^{k} p(n_1, \ldots, n_k; m; k) \operatorname{pr}(\chi^2_{m-1} > t^2),$$

where $p(n_1, \ldots, n_k; m; k)$ is the probability that the amalgamation process leads to exactly $m$ different values. We define the following statistic for testing $\omega_g$ against

$$\bar{\omega}_g: \mu_{\tau_{j-1}+1} \leqslant \ldots \leqslant \mu_{\tau_j} \quad (j = 1, \ldots, r)$$

with at least one strict inequality:

$$D_g^2 \equiv D^2(g_1, \ldots, g_r) = \sum_{j=1}^{r} \sum_{i=\tau_{j-1}+1}^{\tau_j} n_i \{\tilde{\mu}_i - \bar{X}(g_j)\}^2,$$

where $\bar{X}(g_j) = \Sigma n_i \bar{X}_i / \Sigma n_i$ and $(\tilde{\mu}_{\tau_{j-1}+1}, \ldots, \tilde{\mu}_{\tau_j})$ $(j = 1, \ldots, r)$ are those values which minimize the functions $\Sigma n_i (\bar{X}_i - \mu_i)^2$ under the restrictions $\mu_{\tau_{j-1}+1} \leqslant \ldots \leqslant \mu_{\tau_j}$. Note that the last two summations are over all $i \in g_j$. Clearly the $\tilde{\mu}_i$ are 'maximum likelihood' estimates if one agrees to ignore information derived from the order postulated by $\Omega$ for $\mu$'s belonging to different partitions. In this sense $D_g^2$ may be called 'pseudo likelihood ratio' statistic. Let $m_j$ be the number of different numerical values in the set $(\tilde{\mu}_{\tau_{j-1}+1}, \ldots, \tilde{\mu}_{\tau_j})$. Note that $1 \leqslant m_j \leqslant \lambda_j = \tau_j - \tau_{j-1}$. The conditional distribution of $D_g^2$ given $m_1, \ldots, m_r$, because of the independence of the $\bar{X}_i$'s, is $\chi^2_{M-r}$, where $M = m_1 + \ldots + m_r$. The probability of obtaining

$m_j$ different values of $(\tilde{\mu}_{\tau_{j-1}+1}, \ldots, \tilde{\mu}_{\tau_j})$ is $p(n_{\tau_{j-1}+1}, \ldots, n_{\tau_j}; m_j; \lambda_j)$. Thus, again because of the independence of the $\bar{X}_i$'s, the unconditional distribution of $D_g^2$ is

$$\mathrm{pr}\,(D_g^2 > t^2) = \sum_{M=r}^{k} \Sigma^* \prod_{j=1}^{r} p(n_{\tau_{j-1}+1}, \ldots, n_{\tau_j}; m_j; \lambda_j)\,\mathrm{pr}\,(\chi^2_{M-r} > t^2),$$

where $\Sigma^*$ denotes summation over all possible choices of $(m_1, \ldots, m_r)$ with $1 \leqslant m_j \leqslant \lambda_j$ $(j = 1, \ldots, r)$ and $m_1 + \ldots + m_r = M$.

In the special case $n_1 = \ldots = n_k$, the distribution of $D_g^2$ is given by

$$\mathrm{pr}\,(D_g^2 > t^2) = \sum_{M=r}^{k} \Sigma^* \prod_{j=1}^{r} p(m_j; \lambda_j)\,\mathrm{pr}\,(\chi^2_{M-r} > t^2).$$

Table 1. *Upper 5 % and 1 % points of the distribution of $D^2(g_1, \ldots, g_r)$ for 4 to 10 means, with $\sigma^2 n_i^{-1} = 1$, $(i = 4, \ldots, 10)$*

| $(\lambda_1, \ldots, \lambda_r)$, | $t^2_{g,0\cdot05}$ | $t^2_{g,0\cdot01}$ | $(\lambda_1, \ldots, \lambda_r)$ | $t^2_{g,0\cdot05}$ | $t^2_{g,0\cdot01}$ | $(\lambda_1, \ldots, \lambda_r)$ | $t^2_{g,0\cdot05}$ | $t^2_{g,0\cdot01}$ |
|---|---|---|---|---|---|---|---|---|
| (2, 2) | 4·231 | 7·290 | (3, 7) | 7·488 | 11·277 | (2, 3, 4) | 7·394 | 11·128 |
| (2, 3) | 5·088 | 8·352 | (4, 4) | 6·944 | 10·611 | (2, 3, 5) | 7·799 | 11·613 |
| (2, 4) | 5·686 | 9·090 | (4, 5) | 7·356 | 11·110 | (2, 4, 4) | 7·892 | 11·723 |
| (2, 5) | 6·144 | 9·653 | (4, 6) | 7·694 | 11·518 | (3, 3, 3) | 7·552 | 11·307 |
| (2, 6) | 6·513 | 10·106 | (5, 5) | 7·757 | 11·593 | (3, 3, 4) | 8·043 | 11·897 |
| (2, 7) | 6·822 | 10·484 | (2, 2, 2) | 5·435 | 8·747 | (2, 2, 2, 2) | 6·322 | 10·019 |
| (2, 8) | 7·087 | 10·808 | (2, 2, 3) | 6·184 | 9·661 | (2, 2, 2, 3) | 6·966 | 10·848 |
| (3, 3) | 5·862 | 9·295 | (2, 2, 4) | 6·723 | 10·320 | (2, 2, 2, 4) | 7·440 | 11·457 |
| (3, 4) | 6·415 | 9·970 | (2, 2, 5) | 7·144 | 10·832 | (2, 2, 3, 3) | 7·585 | 11·633 |
| (3, 5) | 6·845 | 10·494 | (2, 2, 6) | 7·487 | 11·250 | (2, 2, 2, 2, 2) | 7·248 | 11·001 |
| (3, 6) | 7·194 | 10·919 | (2, 3, 3) | 6·885 | 10·508 | | | |

$\lambda_j$ is the number of integers in $g_j$ $(j = 1, \ldots, r)$.

Upper 5% points of the null distribution of $D_g^2$ are tabulated in Table 1 for different partitions. It is worth noting that subsets $g_j$ for which $\lambda_j = 1$ contribute nothing to $D_g^2$ and hence can be neglected in calculating the distribution of $D_g^2$, or in looking up the critical values in Table 1. The closed inference procedure of the general type described in § 2 is constructed in the following way. If $D^2 \leqslant t_\alpha^2$, where $t_\alpha^2$ is the upper $\alpha$ point of the null distribution of $D^2$, then neither $\omega_0$ nor any of the hypotheses $\omega_g$ is rejected. If $D^2 > t_\alpha^2$, then we reject $\omega_0$ and proceed to test all those hypotheses $\omega_g$ which correspond to partitions $g = (g_1, g_2)$ of $\{1, \ldots, k\}$. Each such hypothesis $\omega_g$ is tested using the corresponding statistic $D_g^2$. If $D_g^2 \leqslant t_{g,\alpha}^2$, where $t_{g,\alpha}^2$ is the upper $\alpha$ point of the distribution of $D_g^2$, then neither $\omega_g$ nor any of the hypotheses $\omega_h$ which correspond to subpartitions $h$ of $g$ is rejected. If $D_g^2 > t_{g,\alpha}^2$ then we reject $\omega_g$. After testing all those $\omega_g$ with $r = 2$ we proceed to test all hypotheses $\omega_u$ which correspond to partitions $u = (u_1, u_2, u_3)$ of $\{1, \ldots, k\}$ which are not subpartitions of any $g = (g_1, g_2)$ for which $\omega_g$ has not been rejected. Each such $\omega_u$ is tested by comparing the corresponding $D_u^2$ with $t_{u,\alpha}^2$ and so on. This stepwise procedure is continued until no more hypotheses are left to be tested.

These results are readily extended to the case of unknown variance by replacing $D_g^2$ with

$$\bar{E}_g^2 = D_g^2 \Big/ \left[\nu s^2 + \sum_{j=1}^{r} \sum_{i=\tau_{j-1}+1}^{\tau_j} n_i\{\bar{X}_i - \bar{X}(g_j)\}^2\right],$$

where $s^2$ is an estimate of $\sigma^2$ independent of the $\bar{X}_i$ and distributed as chi-squared with $\nu$ degrees of freedom. The null distribution of $\bar{E}_g^2$ is analogous to that of $D_g^2$, with $\chi^2_{m-1}$ replaced by the beta variables $\beta_{\frac{1}{2}(m-1), \frac{1}{2}(\nu+k-r)}$. This distribution has been tabulated for the overall null hypotheses and equal $n_i$'s (Barlow *et al.*, 1972, p. 362, Table A.4), but not for partitions. Critical values of the distribution for partitions can be calculated from the probabilities given in Table A.5 of Barlow *et al.* (1972, p. 363) and readily available tables of the beta distribution.

Another way of defining a closed family of hypotheses is to consider all hypotheses $\{\omega_j\}$, each of which postulates $\omega_j: \mu_1 = \ldots = \mu_j \leqslant \mu_{j+1} \leqslant \ldots \leqslant \mu_k$ for some $j$ ($j = 2, \ldots, k$). Each hypothesis is tested either by Williams's statistic $W_j = \hat{\mu}_j - \bar{X}_1$ (Williams, 1971) or by the modified Williams's statistic $R_j = \hat{\mu}_j - \hat{\mu}_1$ (Marcus, 1976).

No way of constructing a simultaneous testing procedure of the general type described by Gabriel (1969) by means of the statistic $D^2$ is known. The family of hypotheses and statistics $\{\omega_g, D_g^2\}$ is a testing family which is not monotone as required by Gabriel's method.

*Numerical example.* Consider an ordered analysis of variance with six treatments and, for simplicity, let $\sigma^2 = 1$, $n_i = 1$ ($i = 1, \ldots, 6$). Let the sample averages be, in that order: 8, 10, 16, 12, 8, 8. The estimates of the $\mu_i$, as found by the amalgamation process, are given in Table 2. The inference procedure is summarized in Table 3. The inferences, in this case, are summarized by the inference from the last term, namely $\mu_1 < \mu_3$, and hence $\mu_1 < \mu_i$ for $i = 4, 5, 6$.

Table 2. *Estimates of means in the various subsets*

| Set of means | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ |
|---|---|---|---|---|---|---|
| (1, 2, 3, 4, 5, 6) | 8 | 10 | 11 | 11 | 11 | 11 |
| (1, 2, 3, 4, 5) | 8 | 10 | 12 | 12 | 12 | — |
| (1, 2, 3, 4) | 8 | 10 | 14 | 14 | — | — |
| (1, 2, 3) | 8 | 10 | 16 | — | — | — |
| (1, 2) | 8 | 10 | — | — | — | — |
| (2, 3, 4, 5, 6) | — | 10 | 11 | 11 | 11 | 11 |

In (3, 4, 5, 6) and any of its subsets all the $\hat{\mu}_i$ are equal.

Table 3. *Test statistics critical values and inferences*

| $g$ | $D_g^2$ | $t_{g, 0\cdot05}$ | Inference |
|---|---|---|---|
| (1, 2, 3, 4, 5, 6) | 7·333 | 5·460 | $\mu_1 < \mu_6$ |
| (1), (2, 3, 4, 5, 6) | 0·800 | 5·049 | — |
| (1, 2), (3, 4, 5, 6) | 2·000 | 5·686 | — |
| (1, 2, 3), (4, 5, 6) | 34·667 | 5·862 | $\mu_1 < \mu_3$ or $\mu_4 < \mu_6$ |
| (1, 2, 3, 4), (5, 6) | 27·000 | 5·686 | $\mu_1 < \mu_4$ or $\mu_5 < \mu_6$ |
| (1, 2, 3, 4, 5), (6) | 12·800 | 5·049 | $\mu_1 < \mu_5$ |
| (1, 2, 3), (4), (5, 6) | 34·667 | 5·088 | $\mu_1 < \mu_3$ or $\mu_5 < \mu_6$ |
| (1, 2, 3), (4, 5), (6) | 34·667 | 5·088 | $\mu_1 < \mu_3$ or $\mu_4 < \mu_5$ |
| (1, 2, 3, 4), (5), (6) | 27·000 | 4·528 | $\mu_1 < \mu_4$ |
| (1, 2, 3), (4), (5), (6) | 34·667 | 3·820 | $\mu_1 < \mu_3$ |

The critical points for $t_{g, 0\cdot05}$ are taken from Table 1.

This study was supported by a research grant from the U.S. National Center of Health Statistics.

## References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, D. (1972). *Statistical Inference under Order Restrictions*. New York: Wiley.

Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46**, 34–48.

Brunk, H. D. (1958). On the estimation of parameters restricted by inequalities. *Ann. Math. Statist.* **29**, 437–54.

Dunnett, C. W. (1955). A multiple comparisons procedure for comparing several treatments with a control. *J. Am. Statist. Assoc.* **60**, 573–83.

EINOT, I. & GABRIEL, K. R. (1975). A study of the powers of several methods of multiple comparisons. *J. Am. Statist. Assoc.* **70**, 574–83.

GABRIEL, K. R. (1969). Simultaneous test procedures – some theory of multiple comparisons. *Ann. Math. Statist.* **40**, 224–50.

HARTLEY, H. O. (1955). Some recent developments in analysis of variance. *Comm. Pure and Applied Math.* **8**, 47–72.

KEULS, M. (1952). The use of the 'studentized range' in connection with an analysis of variance. *Euphytica* **1**, 112–22.

MARCUS, R. (1976). The powers of some tests of the equality of normal means against an ordered alternative. *Biometrika*, **63**, 177–83.

NEWMAN, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika* **31**, 20–30.

WILLIAMS, D. A. (1971). A test for differences between population means when several dose levels are compared to a zero dose control. *Biometrics* **27**, 103–17.

[*Received August* 1975. *Revised December* 1975]