The Annals of Mathematical Statistics
1972, Vol. 43, No. 2, 398-411

ON THE OPTIMALITY OF SOME MULTIPLE COMPARISON PROCEDURES[1]

BY EMIL SPJØTVOLL

University of Wisconsin[2]

Optimality criteria formulated in terms of the power functions of the individual tests are given for problems where several hypotheses are tested simultaneously. Subject to the constraint that the expected number of false rejections is less than a given constant γ when all null hypotheses are true, tests are found which maximize the minimum average power and the minimum power of the individual tests over certain alternatives. In the common situations in the analysis of variance this leads to application of multiple t-tests. In that case the resulting procedure is to use Fisher's "least significant difference," but without a preliminary F-test and with a smaller level of significance. Recommendations for choosing the value of γ are given by relating γ to the probability of no false rejections if all hypotheses are true. Based upon the optimality of the tests, a similar optimality property of joint confidence sets is also derived.

1. Introduction. Although the literature on multiple hypothesis testing and multiple comparison methods is vast (see e.g. Miller [20]), the literature on the optimality of the methods is rather scarce. An important contribution was made by Lehmann [16], [17]. Lehmann finds optimal rules among the class of unbiased rules, where optimality means minimizing the expected loss, and where the loss is the sum of the losses from the individual decisions.

It has been a common complaint that the powers of separate tests are small when using multiple tests. Therefore, in this paper attention is directed towards maximizing the power of the individual test. Instead of using the constraint that the probability of at least one false rejection is smaller than a certain number α, an upper bound γ on the expected number of false rejections is used. The latter is technically easier to work with (see [1], [6], [7], [8], [9] for the amount of numerical work connected with the former constraint) and the author personally finds it more instructive to think in terms of expected number of false rejections than in terms of the probability of at least one false rejection. Suppose a statistician uses $\gamma = .05$, then in average for every twentieth problem he makes one false statement. On the other hand if he uses $\alpha = .05$, then in average for every twentieth problem he makes false rejections, but he does not know how many false rejections he makes. The author feels

Received December 29, 1970; revised August 10, 1971.

[1] This research was supported in part by the Office of Naval Research under Contract Nonr-1202 (17), Project NR 042-222.

[2] Now at University of Oslo.

that it is important to know this. It is also easily seen that the probability of at least one false rejection is less than γ, hence one has an upper bound on probability of at least one false rejection when γ is known. The knowledge of α, however, cannot be used to give an upper bound on γ.

2. Statement of the problem. Let X be a random variable with probability distribution depending upon a parameter θ, $\theta \in \Omega$. Consider a family of hypothesis testing problems

$$H_t : \theta \in \Omega_{0t} \quad \text{against} \quad K_t : \theta \in \Omega_{1t}, \qquad t \in T, \tag{2.1}$$

where $\Omega_{it} \subset \Omega$, $i = 0, 1$, and T is finite with N elements. A test of the hypotheses (2.1) will be defined to be a vector $(\phi_1(x), \cdots, \phi_N(x))$, where the elements of the vector are ordinary test functions; when x is observed we reject H_t with probability $\phi_t(x)$, $t \in T$. The power function of a test $(\phi_1, \cdots, \phi_N)$ is defined to be the vector $(\beta_1(\theta), \cdots, \beta_N(\theta))$ where $\beta_t(\theta) = E_\theta \phi_t(X)$, $t \in T$. For a related definition of power of tests of a multiple hypothesis testing problem see Duncan [3]. Let $S(\gamma)$ be the set of all tests $(\phi_1, \cdots, \phi_N)$ such that

$$\sum_{t=1}^{N} E_\theta \phi_t(X) \leqq \gamma, \qquad \theta \in \Omega_0, \tag{2.2}$$

where $\Omega_0 = \bigcap_T \Omega_{0t}$. Hence $S(\gamma)$ is the set of tests such that the expected number of false rejections under Ω_0 is less than or equal to γ.

For each $t \in T$ we would, subject to (2.2), like to have $\beta_t(\theta)$ large when $\theta \in \Omega_{1t}$. If we make $\beta_t(\theta)$, $\theta \in \Omega_{1t}$, large for a fixed t, then $\beta_t(\theta)$, $\theta \in \Omega_{1t}$, will often have to be small for other values of t, if (2.2) is to be satisfied. Therefore we will have to compromise, and we will consider tests which maximize the minimum power over certain subsets ω_t of Ω_{1t}, $t \in T$, and tests which maximize average power over certain subsets. A test $(\phi_1, \cdots, \phi_N) \in S(\gamma)$ will be said to *maximize the minimum power over* ω_t, $t \in T$, *if it maximizes*

$$\inf_T \inf_{\omega_t} E_\theta \phi_t(X) \tag{2.3}$$

among tests $(\phi_1, \cdots, \phi_N) \in S(\gamma)$. It will be said to *maximize the minimum average power over* ω_t, $t \in T$, *if it maximizes*

$$\sum_{t=1}^{N} \inf_{\omega_t} E_\theta \phi_t(X) \tag{2.4}$$

among tests $(\phi_1, \cdots, \phi_N) \in S(\gamma)$.

Note that the above optimality criteria are more directed towards the performances of the individual tests, than towards their simultaneous performance. Let, for example, X_1 and X_2 be independent $N(\mu_i, 1)$, $i = 1, 2$, and $H_i : \mu_i = 0$ against $K_i : \mu_i > 0$, $i = 1, 2$. Furthermore let $\omega_i = \{(\mu_1, \mu_2) : \mu_i \geqq \Delta\}$, $i = 1, 2$, for some $\Delta > 0$. Then a test of the two hypotheses satisfying (2.3) is such that if one of the μ_i is greater than Δ, then we have a guaranteed smallest probability of discovering this, and this smallest probability is the largest possible. The optimality criterion does not tell us anything about the probability of rejecting

both H_1 and H_2 when both μ_1 and μ_2 are greater than Δ. (For a more traditional approach to this problem see Steffens [24]). The second criterion (2.4) is similarly directed towards maximizing the minimum average of the powers of the individual tests. The reason for studying individual powers is that a common complaint about multiple comparison methods has been that the individual powers are very small. It is the objective of this paper to find procedures which maximize the individual powers of tests.

For later reference we state the following lemma.

LEMMA. *Let ω_t be subsets of Ω_{1t}, $t \in T$. Suppose that there exists a test $(\phi_1 \cdots, \phi_N)$ such that* (I) *there exist points $\theta_t^* \in \omega_t$, $t \in T$, such that $(\phi_1, \cdots, \phi_N)$ maximizes $\inf_T E_{\theta_t^*}\phi_t(X)$ $(\sum_{t=1}^N E_{\theta_t^*}\phi_t(X))$ among tests $(\psi_1, \cdots, \psi_N) \in S(\gamma)$,* (II) *$\inf_{\theta \in \omega_t} E_\theta \phi_t(X) = \sum_{\theta_t^*}\phi_t(X)$, $t \in T$. Then $(\phi_1, \cdots, \phi_N)$ maximizes* (2.3), ((2.4)) *among tests in $S(\gamma)$.*

PROOF. Obvious, since any other test has less or equal minimum (average) power at the points θ_t^*, $t \in T$.

The following two theorems will be helpful when trying to find tests maximizing (2.3) and (2.4). In the following let $f_{01}, \cdots, f_{0N}, f_1, \cdots, f_N$ be integrable functions with respect to a σ-finite measure μ defined on a measurable space $(\mathscr{X}, \mathscr{A})$. Let $S'(\gamma)$ be the set of all tests $(\psi_1, \cdots, \psi_N)$ satisfying

$$\sum_{t=1}^N \int \psi_t(x) f_{0t}(x)\, d\mu(x) = \gamma\,. \tag{2.5}$$

THEOREM 1. *Suppose that there exists a test $(\phi_1, \cdots, \phi_N) \in S'(\gamma)$ defined by*

$$\begin{aligned} \phi_t(x) &= 1 \quad && \text{when} \quad f_t(x) > c f_{0t}(x) \\ &= a_t && \text{when} \quad f_t(x) = c f_{0t}(x) \\ &= 0 && \text{when} \quad f_t(x) = c f_{0t}(x)\,. \end{aligned} \tag{2.6}$$

Then $(\phi_1, \cdots, \phi_N)$ maximizes

$$\sum_{t=1}^N \int \psi_t(x) f_t(x)\, d\mu(x) \tag{2.7}$$

among all tests $(\psi_1, \cdots, \psi_N) \in S'(\gamma)$.

PROOF. Let $(\psi_1, \cdots, \psi_N)$ be any other test in $S'(\gamma)$. We have

$$\begin{aligned} &\sum_{t=1}^N \int \phi_t(x) f_t(x)\, d\mu(x) - \sum_{t=1}^N \int \psi_t(x) f_t(x)\, d\mu(x) \\ &\qquad = \sum_{t=1}^N \int (\phi_t(x) - \psi_t(x))(f_t(x) - c f_{0t}(x))\, d\mu(x) \geqq 0\,, \end{aligned}$$

which proves the theorem.

THEOREM 2. *Suppose that there exists a test $(\phi_1, \cdots, \phi_N) \in S'(\gamma)$ defined by*

$$\begin{aligned} \phi_t(x) &= 1 \quad && \text{when} \quad c_t f_t(x) > f_{0t}(x) \\ &= a_t && \text{when} \quad c_t f_t(x) = f_{0t}(x) \\ &= 0 && \text{when} \quad c_t f_t(x) < f_{0t}(x) \end{aligned} \tag{2.8}$$

where $a_1, \cdots, a_N, c_1, \cdots, c_N$ *are such that*

$$\int \phi_t(x) f_t(x)\, d\mu(x) = \inf_T \int \phi_t(x) f_t(x)\, d\mu(x), \qquad t \in T, \tag{2.9}$$

and $c_t \geqq 0$, $t \in T$, $\sum_{t=1}^N c_t > 0$. *Then* $(\phi_1, \cdots, \phi_N)$ *maximizes*

$$\inf_T \int \psi_t(x) f_t(x)\, d\mu(x) \tag{2.10}$$

among all tests $(\psi_1, \cdots, \psi_N) \in S'(\gamma)$.

PROOF. Let $(\psi_1, \cdots, \psi_N)$ be any other tests in $S'(\gamma)$, and let m be the value of (2.10) for $(\psi_1, \cdots, \psi_N)$. Then

$$\begin{aligned}
&\left(\sum_{t=1}^N c_t\right)\left(\int \phi_t(x) f_t(x)\, d\mu(x) - m\right) \\
&\qquad \geqq \sum_{t=1}^N \int \phi_t(x) c_t f_t(x)\, d\mu(x) - \sum_{t=1}^N \int \psi_t(x) c_t f_t(x)\, d\mu(x) \\
&\qquad = \sum_{t=1}^N \int (\phi_t(x) - \psi_t(x))(c_t f_t(x) - f_{0t}(x))\, d\mu(x) \geqq 0 .
\end{aligned}$$

Since $\sum_{t=1}^N c_t > 0$ and $\int \phi_t(x) f_t(x)\, d\mu(x)$ does not depend upon t, the theorem is proved.

COROLLARY. *Let the test* $(\phi_1, \cdots, \phi_N) \in S'(\gamma)$ *be of the form* (2.6) *with* $c > 0$, *and hence maximize the minimum average power. If* $\int \phi_t(x) f_t(x)\, d\mu(x) =$ constant, $t \in T$, *then* $(\phi_1, \cdots, \phi_N)$ *also maximizes* (2.10) *among tests in* $S'(\gamma)$.

PROOF. Follows trivially from Theorem 2 since $(\phi_1, \cdots, \phi_N)$ is of the form (2.8) and satisfies (2.9).

REMARK 2.1. If the index set T is infinite, we can formulate results similar to Theorems 1 and 2 if we have a finite measure ν defined over T. Corresponding to (2.2) we would have $\int E_\theta \phi_t(X)\, d\nu(t) \leqq \gamma$. (2.3) would stand as it is, while (2.4) would become $\int (\inf_{\omega_t} E_\theta \phi_t(X))\, d\nu(t)$. In Theorem 1 the test defined by (2.6) would maximize $\int [\int \psi_t(x) f_t(x)\, d\mu(x)]\, d\nu(t)$ and the test (2.8) would maximize (2.10) if $c_t \geqq 0$, $t \in T$, and $\int c_t\, d\nu(t) > 0$. This would, of course, require that we consider tests such that $E_\theta \phi_t(X)$ is measurable as a function on T.

REMARK 2.2. In addition to the constraint (2.2), we could also require that the expected number of false rejections is always less than γ, not only under Ω_0. A similar constraint is used by e.g. Scheffé [23] in a situation where he wants the probability of false rejections to be less than α. It will be seen that all tests derived in later sections satisfy this additional requirement.

3. Application to comparison of means of normal random variables with common known variance. Let X_{ij} be $N(\mu_i, 1)$ $j = 1, \cdots, n_i$, $i = 1, \cdots, r$, and independent. Hence we assume, without loss of generality, that the known variance is 1. Consider the following hypotheses about linear functions in the $\{\mu_i\}$

$$H_t: \sum_{i=1}^r a_{ti}\mu_i = b_t \quad \text{against} \quad K_t: \sum_{i=1}^r a_{ti}\mu_i > b_t, \qquad t \in T, \tag{3.1}$$

where the $\{a_{ti}\}$ and $\{b_t\}$ are given constants. Let ω_t be the set of all parameter

points such that $\sum_{i=1}^{r} a_{ti}\mu_i - b_t \geqq \Delta_t$, where $\Delta_t > 0$ is to be fixed later. We assume that Ω_0 is not empty.

Let $\bar{X}_i = (\sum_{j=1}^{n_i} X_{ij})/n_i$, $i = 1, \cdots, r$. For fixed t we transform $\bar{X}_1, \cdots, \bar{X}_r$ to $Y_1, \cdots, Y_r$ by a nonsingular linear transformation, $Y_1 = \sum_{i=1}^{r} a_{ti}\bar{X}_i$, $Y_j = \sum_{i=1}^{r} b_{ji}\bar{X}_i$, $j = 2, \cdots, r$, where the $\{b_{ji}\}$ are chosen so that $\mathrm{Cov}(Y_1, Y_j) = \sum_{i=1}^{r} a_{ti}b_{ji}/n_i = 0$, $j = 2, \cdots, r$. We have that Y_1 is $N(\sum_{i=1}^{r} a_{ti}^2/n_i)$ and Y_j is $N(\sum_{i=1}^{r} b_{ji}\mu_i, \sum_{i=1}^{r} b_{ji}^2/n_i)$.

Let f_0 be the density of the observations when $\mu_i = \mu_i^0$, $i = 1, \cdots, r$, where $(\mu_1^0, \cdots, \mu_r^0)$ is any given point in Ω_0. Let f_t be the density when $\mu_1, \cdots, \mu_r$ are the solution of the equations

$$\begin{aligned} \sum_{i=1}^{r} a_{ti}\mu_i &= \Delta_t + b_t \\ \sum_{i=1}^{r} b_{ji}\mu_i &= \sum_{i=1}^{r} b_{ji}\mu_i^0 \qquad j = 2, \cdots, r . \end{aligned} \tag{3.2}$$

Then

$$\log (f_t/f_0) = (\textstyle\sum_{i=1}^{r} a_{ti}^2/n_i)^{-1}[(Y_1 - b_t)\Delta_t - \frac{1}{2}\Delta_t^2] .$$

Hence $c_t f_t > f_0$ is equivalent to

$$\begin{aligned} (\textstyle\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}}(\sum_{i=1}^{r} a_{ti}\bar{X}_i - b_t) > \tfrac{1}{2}(\textstyle\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}}\Delta_t \\ + \Delta_t^{-1}k_t(\textstyle\sum_{i=1}^{r} a_{ti}^2/n_i)^{\frac{1}{2}} , \end{aligned} \tag{3.3}$$

where $k_t = -\log c_t$. Let ϕ_t be the test which rejects H_t when (3.3) holds. The power function of ϕ_t is

$$\begin{aligned} \beta_t(\mu_1, \cdots, \mu_r) = \Phi((\textstyle\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}}(\sum_{i=1}^{r} a_{ti}\mu_i - b_t - \Delta_t/2) \\ - \Delta_t^{-1}k_t(\textstyle\sum_{i=1}^{r} a_{ti}^2/n_i)^{\frac{1}{2}}) , \end{aligned} \tag{3.4}$$

where Φ is the cumulative standard normal distribution function. It is seen that the power is increasing in $\sum_{i=1}^{r} a_{ti}\mu_i$, and has its minimum over ω_t for $\sum_{i=1}^{r} a_{ti}\mu_i - b_t = \Delta_t$.

The condition (2.2) on the tests $\{\phi_t\}$ becomes

$$\textstyle\sum_t \Phi(-\frac{1}{2}\Delta_t(\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}} - \Delta_t^{-1}k_t(\sum_{i=1}^{r} a_{ti}^2/n_i)^{\frac{1}{2}}) \leqq \gamma . \tag{3.5}$$

Let θ_t^* in the Lemma correspond to the solution of (3.2). It is then easily seen by Theorem 1 and the Lemma that the test maximizing the minimum average power over the alternatives ω_t is given by the tests $\{\phi_t\}$ with $k_t = k$, $t \in T$, and where k is determined so that we have equality in (3.5). By Theorem 2 the test maximizing minimum power over the alternatives ω_t is given by the tests $\{\phi_t\}$ where the $\{k_t\}$ are determined so that we have equality in (3.5) and the powers $\beta_t(\mu_1, \cdots, \mu_r)$ in (3.4) all have the same value when $\sum_{i=1}^{r} a_{ti}\mu_i - b_t = \Delta_t$.

From the form of (3.3) it is seen that we are using a t-test for the individual hypothesis, but that the significance level of the individual hypothesis may vary depending upon $\{a_{ti}\}$, $\{\Delta_t\}$ and $\{n_i\}$. We will now consider three cases.

(a) $\qquad \sum_{i=1}^{r} a_{ti}^2/n_i = A , \qquad \Delta_t = \Delta , \qquad t \in T .$

In this case the tests maximizing the minimum average power and minimum power coincide, and H_t is rejected when

$$A^{-\frac{1}{2}}(\sum_{i=1}^{r} a_{ti}\bar{X}_i - b_t) > z_\rho \,, \tag{3.6}$$

where z_ρ is the upper ρ-point of the standard normal distribution and $\rho = \gamma/N$. Note that the result holds uniformly in Δ.

(b) $$\Delta_t = \Delta \,, \qquad t \in T \,.$$

This might seem a reasonable set of alternatives to consider. The computational work, however, will be greater than under (a). For the test maximizing minimum average power there is one constant k to determine from (3.5), but to find the test maximizing minimum power we would have to determine N constants k_t. The tests will also depend upon the alternative Δ.

(c) $$\Delta_t = \Delta(\sum_{i=1}^{r} a_{ti}^2/n_i)^{\frac{1}{2}} \,, \qquad t \in T \,.$$

The reason for choosing these values of the $\{\Delta_t\}$ is the following: Our "best" estimate of Δ_t is $\hat{\Delta}_t = \sum_{i=1}^{r} a_{ti}\bar{X}_i$ with $\operatorname{Var}\hat{\Delta}_t = \sum_{i=1}^{r} a_{ti}^2/n_i$. It seems reasonable to measure the distances from the hypotheses in terms of the standard deviations of the estimates. This leads to the above $\{\Delta_t\}$. Also in this case it is found that the tests coincide, and are given by rejecting H_t when

$$(\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}}(\sum_{i=1}^{r} a_{ti}\bar{X}_i - b_t) > z_\rho \,. \tag{3.7}$$

This holds uniformly in Δ.

Now look at various special cases.

1. *Differences between means.* Here the hypotheses are

$$H_{ij}: \mu_i = \mu_j \quad \text{against} \quad K_{ij}: \mu_i > \mu_j \,, \qquad i \neq j \,. \tag{3.8}$$

The pair (i, j) corresponds to the index t, and $N = r(r-1)$. Note that H_{ij} against K_{ij} and H_{ji} against K_{ji} are two different problems; H_{ij} is the same as H_{ji}, but the alternatives are different. If we had used alternatives $\mu_i \neq \mu_j$, T would have $\frac{1}{2}r(r-1)$ elements. But since one usually wants to know which mean is the greater when H_{ij} is rejected, the above formulation of the problem seems to be the more useful one.

If all n_i are equal and $\Delta_t = \Delta$, we have the situation in (a) above. In general we find that the test maximizing minimum average power and minimum power over the alternatives $\mu_i - \mu_j \geqq \Delta(1/n_i + 1/n_j)^{\frac{1}{2}}$, rejects H_{ij} and accepts K_{ij} when $(1/n_i + 1/n_j)^{-\frac{1}{2}}(\bar{X}_i - \bar{X}_j) > z_\rho$, where $\rho = \gamma/r(r-1)$.

2. *Comparison with a known standard.* We want to compare the means $\mu_1, \cdots, \mu_r$ with a known standard μ_0 (see [19]). More precisely our problem is

$$H_t: \mu_t = \mu_0 \quad \text{against} \quad K_t: \mu_t > \mu_0 \,, \qquad t = 1, \cdots, r \,.$$

In this case $N = r$, and the test maximizing minimum average power and

minimum power over alternatives of the form $\mu_t - \mu_0 \geqq \Delta n_t^{\frac{1}{2}}$, $t = 1, \cdots, r$, consists of rejecting H_t when $n_t^{\frac{1}{2}}(\bar{X}_t - \mu_0) > z_\rho$, where $\rho = \gamma/r$. We could, of course, also have added hypotheses with alternatives with $\mu_t < \mu_0$ if that was a possible result, and was of interest to the experimenter. In that case $N = 2r$ and $\rho = \gamma/2r$.

3. *Comparison with an unknown standard or a control.* In this case one of the means, μ_r say, is a control (see [8] and [9]). The problem is

$$H_t\colon \mu_t - \mu_r = 0 \quad \text{against} \quad K_t\colon \mu_t - \mu_r > 0\,, \qquad t = 1, \cdots, r-1\,.$$

If we consider alternatives of the form $\mu_t - \mu_r \geqq \Delta(1/n_t + 1/n_r)^{\frac{1}{2}}$, we will reject H_t when $(1/n_t + 1/n_r)^{-\frac{1}{2}}(\bar{X}_t - \bar{X}_r) > z_\rho$, where $\rho = \gamma/(r-1)$. Again we could have considered alternatives $\mu_t - \mu_r < 0$ at the expense of decreasing ρ.

4. *Ordered means.* In some situations it is known that $\mu_1 \leqq \cdots \leqq \mu_r$. Then we would consider the problem (3.8) with $i > j$. We get the same results with $\rho = 2\gamma/r(r-1)$ instead of $\rho = \gamma/r(r-1)$. Hence the power of the test is larger in this case.

REMARK 3.1. Consider the situation where we are interested in all linear functions in $\mu_1, \cdots, \mu_r$. More precisely, consider the hypotheses (3.1) where $a_{ti}, \cdots, a_{tr}$ varies over all constants satisfying

$$\sum_{i=1}^r a_{ti}^2 = 1\,. \tag{3.9}$$

The set T therefore consists of all points on the sphere (3.9). Making use of Remark 2.1 with the measure ν a probability measure proportional to the area of the sphere, we will find that the test maximizing minimum power over alternatives with

$$\sum_{i=1}^r a_{ti}\mu_i - b_t \geqq (\sum_{i=1}^r a_{ti}^2/n_i)^{\frac{1}{2}}\Delta\,,$$

rejects when (3.7) holds. In this case ρ is interpreted as the average probability of false rejections under Ω_0. The expected number of false rejections is infinite.

The tests (3.7) are the same as the ones derived from Scheffé's S-method of multiple comparison; see [22]. The test derived from T-method of multiple comparison (see [22] page 74) rejects H_t when

$$m^{\frac{1}{2}} \sum_{i=1}^r a_{ti}\bar{X}_i > k \sum_{i=1}^r |a_{ti}|$$

where $m = n_1 = \cdots = n_r$ (the T-method can be used only in the case when the $\{n_i\}$ are equal). This test is not of the form (3.7), hence it has less average power and less minimum power over the alternatives ω_t, which in this case are $m^{\frac{1}{2}} \sum_{i=1}^r a_{ti}\mu_i \geqq \Delta + b_t$.

REMARK 3.2. An interesting improvement of the power of multiple tests in the case where we are interested in all contrasts in the $\{\mu_i\}$ is derived by

Scheffé [23]. If the F-test rejects the hypothesis that all means are equal, one can reject H_t in (3.1) for smaller observed values of $|\sum_{i=1}^r a_{ti}\bar{X}_i|$ than one should expect using the S-method of multiple comparison. The modification does not change the maximum probability of at least one false rejection. The improvement cannot be used here since it will increase the average number of wrong decisions under Ω_0, even if the probability of at least one false rejection is not changed.

The same is the case with a multiple comparison method like the Newman-Keuls method (see [20] page 82). The probability of at least one false rejection is α under Ω_0. Obviously the Newman-Keuls method is more powerful than the T-method which (for a suitably chosen γ) corresponds to our result in the case $n_1 = \cdots = n_r$. The Newman-Keuls method, therefore, has a larger expected number of false rejections under Ω_0. If we were willing to use the same expected number of false rejections as the Newman-Keuls method, we could improve on it (improve means here to get a greater minimum power) by using the method derived here.

REMARK 3.3. The alternatives considered above have been one-sided. Suppose that, instead of (3.1), we have

$$H_t: \sum_{i=1}^r a_{ti}\mu_i = b_t \quad \text{against} \quad K_t: \sum_{i=1}^r a_{ti}\mu_i \neq b_t \qquad t \in T,$$

and we consider alternatives ω_t of the form $|\sum_{i=1}^r a_{ti}\mu_i - b_t| \geqq \Delta(\sum_{i=1}^r a_{ti}^2/n_i)^{\frac{1}{2}}$. Then it is easily shown by using a least favorable distribution over $|\sum_{i=1}^r a_{ti}\mu_i - b_t| = \Delta(\sum_{i=1}^r a_{ti}^2/n_i)^{\frac{1}{2}}$ (see [18] page 90) that the test which rejects H_t when

$$(\sum_{i=1}^r a_{ti}^2/n_i)^{-\frac{1}{2}}|\sum_{i=1}^r a_{ti}\bar{X}_i - b_t| > z_\rho,$$

maximizes both minimum average power and minimum power.

REMARK 3.4. We could also easily extend the results to the case where the estimates of the $\{\mu_i\}$ are correlated with known correlations (see Kramer [15] and Duncan [4]). We would get results analogous to (3.3).

4. Comparison of means of normal random variables with common unknown variance. In this situation we will restrict attention to unbiased tests, and since we are concerned with the performance of each individual test we will require that each test ϕ_t is unbiased i.e. in the notation of Section 2

$$\sup_{\Omega_{0t}} E_\theta \phi_t(X) \leqq \inf_{\Omega_{1t}} E_\theta \phi_t(X), \qquad t \in T. \tag{4.1}$$

We will then try to find the family of tests maximizing minimum average power among unbiased tests. Consider now the problem (3.1) in Section 3 when we assume that the X_{ij} have a common unknown variance σ^2. Suppose we have a test $(\psi_1, \cdots, \psi_N)$ such that (4.1) is satisfied. Then $E_{\mu_1,\cdots,\mu_r,\sigma}\psi_t(X) =$ constant, γ_t, say, for $\sum_{i=1}^r a_{ti}\mu_i = b_t$, and any value of σ. Now consider the test $(\phi_1, \cdots, \phi_N)$ where ϕ_t rejects H_t when $(\sum_{i=1}^r a_{ti}^2/n_i)^{-\frac{1}{2}}(\sum_{i=1}^r a_{ti}\bar{X}_i - b_t)/S$ is

greater then the upper γ_t-point of the t-distribution with $n - r$ degrees of freedom. Here $n = \sum_{i=1}^{r} n_i$, and $S^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/(n - r)$. The test ϕ_t is uniformly most powerful unbiased at γ_t level, hence it is at least as powerful as ψ_t. They both have the same expected value under Ω_0. It follows that we can restrict attention to tests where the test of H_t against K_t is based upon $(\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}}(\sum_{i=1}^{r} a_{ti}\bar{X}_i - b_t)/S = V_t$.

The density of V_t is

$$C(r, n) \int_0^\infty y^{(n-r-1)/2} \exp(-\tfrac{1}{2}y) \exp[-\tfrac{1}{2}(v_t(y/n - r)^{\frac{1}{2}} - \delta_t)^2]\, dy\,, \tag{4.2}$$

where $\delta_t = (\sum_{i=1}^{r} a_{ti}^2/n_i)^{\frac{1}{2}}(\sum_{i=1}^{r} a_{ti}\mu_i - b_t)/\sigma$, and $C(r, n)$ is a constant. The density (4.2) has monotone likelihood ratio in v_t (see [18] page 223). Consider the alternatives ω_t defined by $(\sum_{i=1}^{r} a_{ti}\mu_i - b_t)/\sigma \geqq \Delta(\sum_{i=1}^{r} a_{ti}^2/n_i)^{-\frac{1}{2}}$ which corresponds to $\delta_t \geqq \Delta$. Let the density (4.2) with $\delta_t = \Delta$ correspond to the function f_t in Theorems 1 and 2, and let the density (4.2) with $\sum_{i=1}^{r} a_{ti}\mu_i = b_t$ (which is a central t-distribution) correspond to the density f_{0t}. Using the fact that the density (4.2) has monotone likelihood ratio in v_t it is easily seen by using the Lemma and Theorems 1 and 2 that the test maximizing minimum average power and minimum power among unbiased tests is given by rejecting H_t when $V_t > t_\rho$, where t_ρ is the upper ρ-point of the t-distribution with $n - r$ degrees of freedom and $\rho = \gamma/N$.

If we considered other alternatives ω_t the form of the solution is easily obtained by using the Lemma and Theorems 1 and 2, but will lead to considerable numerical work since we would have to evaluate the integral (4.2).

The application to the special cases considered in Section 3 follows by replacing the tests obtained there by t-tests.

5. Comparison of variances. Let X_{ij}, be $N(\mu_i, \sigma_i^2)$ $j = 1, \cdots, n_i$, $i = 1, \cdots, r$, and independent. Consider the hypotheses

$$H_{ij}: \sigma_i^2 = \sigma_j^2 \quad \text{against} \quad K_{ij}: \sigma_i^2 > \sigma_j^2 \qquad i \neq j\,. \tag{5.1}$$

Restricting attention to unbiased tests and arguing as in Section 4, we find that the test of H_{ij} against K_{ij} can be based upon $R_{ij} = S_i^2/S_j^2$ only, where $S_i^2 = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$, $i = 1, \cdots, r$. Let $\sigma_i^2/\sigma_j^2 = \Delta_{ij}$. Then the density of R_{ij} is

$$C_{ij}\Delta_{ij}^{-\frac{1}{2}(n_i-1)} r_{ij}^{\frac{1}{2}(n_i-1)-1}(n_j - 1 + (n_i - 1)\Delta_{ij}^{-1} r_{ij})^{-\frac{1}{2}(n_i+n_j-2)}\,, \tag{5.2}$$

where C_{ij} is a constant. Call the density (5.2) for f_{ij} and let f_{0ij} be the density (5.2) when $\Delta_{ij} = 1$. The inequality $c_{ij} f_{ij} > f_{0ij}$ is equivalent to

$$R_{ij} > \Delta_{ij} \frac{(n_j - 1)(1 - (\Delta_{ij}^{\frac{1}{2}(n_i-1)} c_{ij}^{-1})^{2/(n_i+n_j-2)})}{(n_i - 1)((\Delta_{ij}^{\frac{1}{2}((n_i-1)} c_{ij}^{-1})^{2/(n_i-n_j-2)} - \Delta_{ij})}\,. \tag{5.3}$$

The power function of the test which rejects H_{ij} when (5.3) holds is

$$1 - \mathscr{F}(a_{ij}\sigma_j^2/\sigma_i^2) \tag{5.4}$$

where a_{ij} is the expression on the right-hand side of (5.3) and $\mathcal{F}$ is the cumulative F-distribution with $n_i - 1$ and $n_j - 1$ degrees of freedom.

In the case where $n_i = \cdots = n_r = m$, and we choose ω_{ij} as the set of all points such that $\sigma_i^2/\sigma_j^2 \geqq \Delta > 0$, it is easily seen that the test which rejects H_{ij} when R_{ij} is greater than the upper ρ-point ($\rho = r(r-1)$) of the F-distribution with $m - 1$ and $m - 1$ degrees of freedom maximizes both minimum average power and minimum power over the alternatives ω_{ij}. If the n_i are not all equal, we could still find the optimum tests by adjusting the constants c_{ij} in (5.3) and (5.4). The numerical work, however, will be extensive. The tests will also depend upon the alternatives ω_{ij} chosen.

If we had a problem where we wanted to compare variances with a control or standard (see Bechhofer [1]), we could use an approach analogous to the one in Section 3.

6. How to choose γ. γ is the upper bound for the expected number of wrong rejections. If γ had been a significance level, we would not have any problems since it seems that many statisticians have become used to using significance level .05 or .01 without ever questioning it. Therefore, relying upon tradition, we will relate the value of γ to the expected number of false rejections we have when using a traditional multiple comparison method with a guaranteed probability $1 - \alpha$ of no false rejections. Take problem (3.8) in Section 3, and assume $n_1 = \cdots = n_r = m$ and σ^2 unknown. Let S^2 be an estimate of σ^2 such that $\nu S^2/\sigma^2$ has a chi-square distribution with ν degrees of freedom. Suppose we base the solution upon the T-method of multiple comparison, then we reject H_{ij} if

$$(6.1) \qquad (m/2)^{\frac{1}{2}}(\bar{X}_i - \bar{X}_j)/S > 2^{-\frac{1}{2}}q_{\alpha;r,\nu} ,$$

where $q_{\alpha;r,\nu}$ is the upper α-point of the studentized range with parameters r and ν. The probability of at least one false rejection is α or smaller. The expected number of false rejections is

$$\gamma(\alpha, r, \nu) = r(r-1)(1 - G_\nu(2^{-\frac{1}{2}}q_{\alpha;r,\nu})) ,$$

where G_ν is the cumulative t-distribution with ν degrees of freedom. In Table 6.1 is given the $\gamma(\alpha, r, \nu)$ corresponding to the traditional significance levels .01 and .05. The conclusion seems to be that we should choose γ of approximately the same size as we choose α.

The expected value γ does not tell us how frequent the different numbers of wrong rejections are. We can, however, obtain a crude bound as follows. We

TABLE 6.1

r	2	3	4	5	6	7	8	9	10	11	12
$\gamma(.05, r, \infty)$	.050	.058	.061	.064	.066	.067	.068	.069	.070	.071	.072
$\gamma(.01, r, \infty)$	.0100	.0108	.0111	.0114	.0116	0.117	.0118	.0119	.0120	.0121	.0121

have $\gamma = \sum_{k=0}^{\infty} kP[k \text{ false rejections}]$, from which we easily obtain

$$P[k \text{ or more false rejections}] \leqq \gamma/k .$$

In particular, P(at least one false rejection) $\leqq \gamma$.

For a discussion of topics related to this section, see, e.g., Miller [20] and Duncan [5], who also consider the error rate which is γ/N. The present author prefers to think in terms of γ and not in terms of the error rate. Suppose, for example, that $r = 10$ and we reject H_{ij} if (6.1) holds with the right-hand side replaced by the upper 0.25 point of the t-distribution with ν degrees of freedom (this is called a multiple comparison method based upon the least significance difference, see [20]). The error rate is then .25, but $\gamma = 10 \cdot 9 \cdot .025 = 2.25$. Hence the expected number of false rejections if all H_{ij} are true is 2.25. If it turned out that two of the hypotheses H_{ij} were rejected, we would hesitate to attribute this to real departures from these hypotheses, since we have observed exactly what we should expect if all H_{ij} were true.

7. An example. To see how the method described in the previous sections can be applied, consider the following example taken from [2] pp. 96–97. The example refers to an experiment on the effects of applications of sulphur in reducing scab disease of potatoes. Three amounts of dressing were compared, 300, 600 and 1200 lbs. per acre, and both fall and spring applications were tested. Furthermore a control was used. The layout was a completely randomized design, hence we have a one-way layout with $r = 7$. The result of the experiment is given in Table 7.1. The estimate of σ^2 was 44.9 with 25 degrees of freedom. The observations are the

TABLE 7.1

Treatment	0	F3	S3	F6	S6	F12	S12
n_i	8	4	4	4	4	4	4
$\bar{x}_i$	22.6	9.5	16.8	15.5	18.2	5.8	14.2

measured scab index. Since the control has 8 replications and the other 4 replications, most multiple comparison methods based upon range cannot be used. (Though there exist approximate methods which are modifications of the traditional methods see, e.g., [4], [14], and [15]).

Suppose that we would like to make all pairwise comparisons which are 42, all pairwise comparisons between 300, 600, and 1200 lbs. and the control which are 12, all pairwise comparisons between fall application, spring application and control which are 6, and finally two comparisons of application of sulphur with the control. Hence, we get $N = 62$. We will use $\gamma = .062$, which corresponds (by Section 4) to test each comparison with a one-sided t-test at .001 significance level. Doing this, we find that F12, fall application, 1200 lbs.

application and application of sulphur is better than the control. If we had used Scheffé's S-method as .05 significance level (.062 would not have made any difference) we would have found one significant difference, namely F12 versus the control. The T-method cannot be used.

Some statisticians would perphaps, in addition to the above 62 comparisons, have wanted to compare linear and quadratic terms in the amount of sulphur. That would have added at most 8 comparisons which would have changed γ to .070.

8. Simultaneous confidence intervals. In this section we shall use the optimal tests derived in the previous sections to construct confidence sets with a corresponding optimality property. Suppose we are interested in confidence sets for certain functions $a_t(\theta)$, $t \in T$. If $S_t(x)$, $t \in T$, is a family of confidence sets such that $S_t(x)$ is a confidence set for $a_t(\theta)$ when x is observed, we will require

$$\sum_{t=1}^{N} P_\theta[a_t(\theta) \in S_t(X)] \geqq N - \gamma \,. \tag{8.1}$$

This means that the expected number of confidence sets covering only false values should be less than or equal to γ.

To construct the confidence sets we will (similar to [18] page 79) start with tests of the hypotheses

$$H_t : a_t(\theta) = \delta_t \quad \text{against} \quad K_t : a_t(\theta) \neq \delta_t \,, \qquad t \in T \,.$$

Let $A_t(\delta_1, \cdots, \delta_N)$, $t \in T$, be the acceptance regions of the test maximizing the minimum power (possibly among unbiased tests) over alternatives with $|a_t(\theta) - \delta_t| \geqq \Delta_t$, $t \in T$, (see Remark 3.3). Call these alternatives $\omega_t(\delta_t, \Delta_t)$. Suppose that there exists such a test for all $\delta_1, \cdots, \delta_N$, and suppose that $A_t(\delta_1, \cdots, \delta_N)$ depends upon δ_t alone, so we can write $A_t(\delta_t)$. (This is the crucial assumption.) We have

$$\sum_{t=1}^{N} P_\theta[X \in A_t(a_t(\theta))] \geqq N - \gamma \,, \qquad \text{for all } \theta \,. \tag{8.2}$$

Since the test maximizes the minimum power over the alternatives $\omega_t(\delta_t, \Delta_t)$,

$$\sup_T \sup_{\omega_t(\delta_t, \Delta_t)} P_\theta[X \in A_t(\delta_t)] \tag{8.3}$$

is minimized among sets $A_t(\delta_t)$ satisfying (8.2).

Define the set $S_t(x)$ by

$$S_t(x) = \{\delta_t : x \in A_t(\delta_t)\} \,. \tag{8.4}$$

We have that $\delta_t \in S_t(x)$ if and only if $x \in A_t(\delta_t)$. Hence by (8.2)

$$\sum_{t=1}^{N} P_\theta[a_t(\theta) \in S_t(X)] \geqq N - \gamma \,.$$

Hence $S_t(x)$, $t \in T$, is a family of confidence sets satisfying (8.1).

Let $S_t^*(x)$ be any other family of confidence sets such that

$$\sum_{t=1}^{N} P_\theta[a_t(\theta) \in S_t^*(X)] \geqq N - \gamma \,. \tag{8.5}$$

Define $A_t^*(\delta_t)$ by $A_t^*(\delta_t) = \{x: \delta_t \in S^*(x)\}$. Then by (8.5)

$$\sum_{t=1}^{N} P_\theta[X \in A_t^*(a_t(\theta))] \geqq N - \gamma ,$$

and hence since $A_t(\delta_t)$ minimizes (8.3) subject to conditions (8.1) and (8.5) we must have

$$\sup_T \sup_{\omega_t(\delta_t,\Delta_t)} P_\theta[X \in A_t(\delta_t)] \leqq \sup_T \sup_{\omega_t(\delta_t,\Delta_t)} P_\theta[X \in A_t^*(\delta_t)] ,$$

and thereby

$$\sup_T \sup_{\omega_t(\delta_t,\Delta_t)} P_\theta[\delta_t \in S_t(X)] \leqq \sup_T \sup_{\omega_t(\delta_t,\Delta_t)} P_\theta[\delta_t \in S_t^*(X)] .$$

This shows that the sets $\{S_t(x)\}$ minimize the maximum probability of covering a false value when this has a distance Δ_t from the true value.

9. Other applications. Multiple comparison problems which can be treated in a way similar to that of the problems in Sections 3 and 4 arise when comparing proportions in multinomial trials and contingency tables, see [10], [11], [12], [13], [21], where normal approximations to multinomial probabilities are used. For comparison of pairs of proportions it will also be possible by the methods used here to obtain exact results, so that one does not have to rely upon the large sample results.

REFERENCES

[1] BECHHOFER, R. E. (1968). Multiple comparisons with a control for multiple-classified variances of normal populations. *Technometrics* **10** 715–718.

[2] COCHRAN, W. G. and COX, G. M. (1957). *Experimental Designs*. Wiley, New York.

[3] DUNCAN, D. B. (1955). Multiple range and multiple *F*-tests. *Biometrics* **11** 1–42.

[4] DUNCAN, D. B. (1957). Multiple range tests for correlated and heteroscedastic means. *Biometrics* **13** 164–176.

[5] DUNCAN, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7** 171–222.

[6] DUNN, O. J. (1961). Multiple comparisons among means. *J. Amer. Statist Assoc.* **56** 52–64.

[7] DUNN, O. J. and MASSEY, F. J., JR. (1965). Estimation of multiple contrasts using *t*-distributions. *J. Amer. Statist. Assoc.* **60** 573–583.

[8] DUNNETT, C. W. (1955). A multiple comparisons procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* **50** 1096–1121.

[9] DUNNETT, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics* **20** 482–491.

[10] GOLD, R. Z. (1963). Tests auxiliary to χ^2 tests in a Markov chain. *Ann. Math. Statist.* **34** 56–74.

[11] GOODMAN, L. A. (1964a). Simultaneous confidence limits for cross-product ratios in contingency tables. *J. Roy. Statist. Soc. Ser. B* **26** 86–102.

[12] GOODMAN, L. A. (1964b). Simultaneous confidence intervals for contrasts among multinomial populations. *Ann. Math. Statist.* **35** 716–725.

[13] GOODMAN, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7** 247–254.

[14] KRAMER, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12** 307–310.

[15] KRAMER, C. Y. (1957). Extension of multiple range tests to group correlated adjusted means. *Biometrics* **13** 13-18.

[16] LEHMANN, E. L. (1957a). A theory of some multiple decision problems, I. *Ann. Math. Statist.* **28** 1-25.

[17] LEHMANN, E. L. (1957b). A theory of some multiple decision problems, II. *Ann. Math. Statist.* **28** 547-572.

[18] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[19] LEHMANN, E. L. (1961). Some model I problems of selection. *Ann. Math. Statist.* **32** 990-1012.

[20] MILLER, R. G., JR. (1966). *Simultaneous Statistical Inference.* McGraw-Hill, New York.

[21] QUESENBERRY, C. P. and HURST, D. C. (1964). Large sample simultaneous confidence intervals for mulitnomial proportions. *Technometrics* **6** 191-195.

[22] SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

[23] SCHEFFÉ, H. (1970). Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios. *Ann. Math. Statist.* **41** 1-19.

[24] STEFFENS, F. E. (1970). Power of bivariate Studentized maximum and minimum modulus tests. *J. Amer. Statist. Assoc.* **65** 1639-1644.